

Metode de optimizare numerică

Ion Necoară

Departamentul de Automatică și Ingineria Sistemelor

Universitatea Politehnica din București

Email: ion.necoara@acse.pub.ro

2013

Prefață

Lucrarea de față este o sinteză, care tratează în mod concis, dar și riguros din punct de vedere matematic, principalele metode numerice de rezolvare a problemelor de optimizare neliniară. Optimizarea este un proces de minimizare sau maximizare a unei funcții obiectiv și, în același timp, de satisfacere a unor constrângeri. Natura abundă de exemple unde un nivel optim este dorit și în multe aplicații din inginerie, economie, biologie și numeroase alte ramuri ale științei se caută regulatorul, portofoliul sau compoziția optim(ă).

Lucrarea este construită pe structura cursului de *Tehnici de Optimizare*, predat de autor la Facultatea de Automatică și Calculatoare a Universității Politehnica din București. Lucrarea se adresează studenților din facultățile cu profil tehnic sau economic, dar în aceeași măsură și celor de la programele de master și doctorat cu tematici adiacente. Studenții ce urmează acest curs necesită cunoștințe de algebră liniară (teoria matricelor, concepte din teoria spațiilor vectoriale) și analiză matematică (noțiuni de funcții diferențiabile, convergența șirurilor). Scopul lucrării este prezentarea unei introduceri în metodele numerice de rezolvare a problemelor de optimizare care servește la pregătirea studenților pentru dezvoltarea și adaptarea acestor metode la aplicații specifice ingineriei și altor domenii. Tematica include elemente de *optimizare continuă* ce se concentrează în special pe *programarea neliniară*. În acest sens, structura lucrării este divizată în trei părți majore:

Partea I: *Introducere* - se prezintă formularea matematică a unei probleme generale de optimizare cu noțiunile asociate și se introduc principalele concepte legate de convexitate (caracterizări și proprietăți ale mulțimilor și funcțiilor convexe).

Partea II: *Optimizare fără constrângeri* - se prezintă proprietățile de bază ale soluțiilor și metodelor numerice, condițiile necesare și suficiente de optimalitate pentru o soluție în cazul problemelor convexe și nonconvexe, analiza metodelor de căutare de-a lungul unei direcții de descreștere (metoda gradient, Newton), reguli de selectare a mărimii pasului (condițiile Wolfe, algoritmul de backtracking), aplicații în estimare și metoda celor mai mici pătrate.

Partea III: *Optimizare cu constrângeri* - se prezintă condițiile necesare și suficiente de optimalitate pentru o soluție în cazul problemelor convexe și nonconvexe având constrângeri (condițiile de tip Karush-Kuhn-Tucker). Se analizează metode bazate pe funcții de penalitate și barieră, metode de punct interior și metode de programare pătratică secvențială pentru probleme generale de optimizare cu constrângeri sub formă de egalități și inegalități. În final, se prezintă formularea sub forma unei probleme de optimizare a unei probleme de control optimal cu orizont finit și rezolvarea numerică cu metodele prezentate.

Prin teoria și exemplele expuse în această lucrare se urmărește familiarizarea studenților cu formularea corectă a unei probleme de optimizare, identificarea tipului unei probleme (convexă sau nonconvexă, cu sau fără constrângeri) și rezolvarea numerică a unei probleme. Sunt descriși principalii algoritmi numerici de optimizare (algoritmi de tip direcție de descădere, e.g. gradient sau Newton, de punct interior) pentru care se analizează convergența și necesarul de calcul. Algoritmii prezentați sunt apoi testați pe aplicații practice, în particular din inginerie. Pe tot parcursul cărții se prezintă multe exemple de aplicații și exerciții rezolvate, tratate amănunțit, pentru a face mai accesibilă înțelegerea teoriei și a conceptelor introduse, și pentru a ajuta studentul să înțeleagă subtilitățile inerente unei discipline avându-și originea în matematica aplicată.

Deși lucrarea de față adresează tehnici și algoritmi de optimizare standard, regăsiți în majoritatea literaturii de specialitate [2, 9, 11, 13], am intenționat de asemenea ca ea să reflecte punctul de vedere modern în acest domeniu. În acest sens, unul din principalele aporturi îl reprezintă conexiunea dintre caracterul analitic al unei probleme de optimizare, exprimat prin condițiile de optimalitate, și analiza algoritmilor numerici de optimizare folosiți pentru rezolvarea problemei. Mai mult, condițiile de optimalitate și algoritmii numerici corespunzători problemelor de optimizare constrânse sunt prezentate ca o generalizare a condițiilor de optimalitate și respectiv a algoritmilor numerici dezvoltați pentru probleme de optimizare fără constrângeri. Această abordare oferă lucrării o structură simplă care facilitează înțelegerea teoriei prezentate dar și posibilitatea dezvoltării de noi rezultate în acest domeniu.

Ca o concluzie, lucrarea de față reprezintă un mixt reușit între rigurozitatea matematică și practicalitatea inginerească ce conduce la

insușirea cu ușurință și manipularea eficace a unor tehnici de optimizare neliniară moderne. De aceea, considerăm că apariția acestei cărți este extrem de utilă, în primul rând studenților, datorită conținutului și stilului adecvat acestei audiențe.

Autorul le este recunoscător referenților științifici prof. univ. dr. C. Oară și prof. univ. dr. B. Dumitrescu pentru sugestiile și observațiile extrem de valoroase pe care le-au oferit pe parcursul conceperii acestui material. Mulțumiri sunt aduse și studenților care de-a lungul timpului au contribuit cu observații pertinente la îmbunătățirea expunerii acestei lucrări.

*Ion Necoară
mai 2013*

Listă de notații**• Vectori**

\mathbb{R} mulțimea numerelor reale; $\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$

\mathbb{R}^n spațiul Euclidian n dimensional al vectorilor coloană x cu n componente reale $x_i \in \mathbb{R}$ pentru orice $i = 1, \dots, n$

$\text{Span}(S)$ subspațiul liniar generat de vectorii din mulțimea $S \subset \mathbb{R}^n$

e_i , cu $i = 1, \dots, n$, baza standard a lui \mathbb{R}^n și e vectorul din \mathbb{R}^n cu toate intrările 1, anume $e = \sum_{i=1}^n e_i$

$\langle x, y \rangle = x^T y = \sum_{i=1}^n x_i y_i$ produsul scalar a doi vectori $x, y \in \mathbb{R}^n$

$\|x\| = \langle x, x \rangle^{1/2} = (\sum_{i=1}^n x_i^2)^{1/2}$ norma Euclidiană a vectorului $x \in \mathbb{R}^n$

$\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ norma p a vectorului $x \in \mathbb{R}^n$, unde $p \geq 1$; în calcule se utilizează în special normele $\|x\|_1 = \sum_{i=1}^n |x_i|$, $\|x\|_2 = \|x\|$ și $\|x\|_\infty = \max_{i=1, \dots, n} |x_i|$

\forall oricare ar fi și \exists există

• Matrice

$\mathbb{R}^{m \times n}$ spațiul Euclidian al matricelor cu m linii și n coloane cu elemente $a_{ij} \in \mathbb{R}$ pentru orice $i = 1, \dots, m$ și $j = 1, \dots, n$

S^n spațiul matricelor simetrice cu n linii și coloane

$A \succ 0$ ($A \succcurlyeq 0$) matrice pozitiv (semi)definită

S_+^n spațiul matricelor pozitiv semidefinite

I_n matricea identitate de ordinul n

a_{ij} sau A_{ij} elementul matricei A situat în linia i și coloana j

A^T transpusa matricei A , iar A^{-1}/A^+ inversa/pseudoinversa matricei A ; $A^{-T} = (A^{-1})^T = (A^T)^{-1}$

$\text{Tr}(A)$ urma matricei pătrate A , anume suma elementelor de pe diagonala principală

$\det(A)$ determinantul matricei pătrate A

$\lambda_i(A)$ valorile proprii ale unei matrice simetrice de dimensiune n ordonate crescător; $\lambda_{\min} = \lambda_1$ și $\lambda_{\max} = \lambda_n$

$\sigma_i(A)$ valorile singulare ale matricei A ordonate crescător

$\text{rang}(A)$ rangul matricei A (numărul valorilor singulare nenule)

$\text{kern}(A)$ nucleul matricei A

$\langle A, B \rangle = \text{Tr}(A^T B)$ produsul scalar a două matrice reale

$\|A\|_F = \langle A, A \rangle^{1/2}$ norma Frobenius; avem relația $\|A\|_F = \left(\sum_{i=1}^r \sigma_i^2\right)^{1/2}$, unde r este rangul lui A

• Funcții

\mathcal{C}^k clasa funcțiilor diferentiabile de k ori, cu derivata de ordinul k continuă

funcția scalară $f : \mathbb{R} \rightarrow \bar{\mathbb{R}}$ are domeniul efectiv $\text{dom} f = \{x \in \mathbb{R}^n : f(x) < \infty\}$

$\inf_{x \in X} f(x) = \inf\{f(x) : x \in X\}$ infimul funcției f peste mulțimea X ; când infimul se atinge, atunci folosim “min” în loc de “inf”

$\nabla f(x) \in \mathbb{R}^n$ gradientul funcției f , anume $(\nabla f(x))_i = \frac{\partial f}{\partial x_i}(x)$ pentru orice $i = 1, \dots, n$

matricea simetrică $\nabla^2 f(x) \in S^n$ este Hessiana funcției f , anume $(\nabla^2 f(x))_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$ pentru orice $i, j = 1, \dots, n$

$\nabla h(x) \in \mathbb{R}^{p \times n}$ Jacobianul funcției vectoriale $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$ cu $h = [h_1 \dots h_p]^T$, anume $(\nabla h(x))_{ij} = \frac{\partial h_i}{\partial x_j}(x)$ pentru orice $i = 1, \dots, p$ și $j = 1, \dots, n$

$\mathcal{L}(x, \lambda, \mu)$ Lagrangianul și $\nabla_x \mathcal{L}(x, \lambda, \mu) = \frac{\partial \mathcal{L}}{\partial x}(x, \lambda, \mu)$ derivata parțială în raport cu x

$q(\lambda, \mu)$ funcția duală

- **Prescurtări**

e.g. de exemplu (*exempli gratia*)

i.e. adică (*id est*)

NLP programare neliniară (*NonLinear Programming*)

UNLP programare neliniară fără constrângeri (*Unconstrained NonLinear Programming*)

QP programare pătratică (*Quadratic Programming*)

LP programare liniară (*Linear Programming*)

CP programare convexă (*Convex Programming*)

KKT Karush-Kuhn-Tucker

DVS descompunerea valorilor singulare

Cuprins

Prefață	5
I Introducere	15
1 Teorie convexă	17
1.1 Teoria mulțimilor convexe	18
1.1.1 Mulțimi convexe	18
1.1.2 Conuri	22
1.1.3 Operații ce conservă convexitatea mulțimilor . . .	24
1.2 Teoria funcțiilor convexe	26
1.2.1 Funcții convexe	26
1.2.2 Condiții de ordinul I pentru funcții convexe . . .	28
1.2.3 Condiții de ordinul II pentru funcții convexe . . .	29
1.2.4 Operații ce conservă convexitatea funcțiilor . . .	31
2 Concepte fundamentale din teoria optimizării	33
2.1 Evoluția teoriei optimizării	33
2.2 Caracteristicile unei probleme de optimizare	36
2.3 Tipuri de probleme de optimizare	41
2.3.1 Programare neliniară (NLP)	41
2.3.2 Programare liniară (LP)	42
2.3.3 Programare pătratică (QP)	43
2.3.4 Optimizare convexă (CP)	45
2.3.5 Probleme de optimizare neconstrânsă (UNLP) . .	48
2.3.6 Programare mixtă cu întregi (MIP)	48
II Optimizare fără constrângeri	51
3 Metode de optimizare unidimensională	53

3.1	Metoda forward-backward pentru funcții unimodale . . .	54
3.2	Metode de căutare	55
3.2.1	Metoda secțiunii de aur	56
3.2.2	Metoda lui Fibonacci	58
3.3	Metode de interpolare	60
3.3.1	Metode de interpolare pătratică	60
3.3.2	Metode de interpolare cubică	65
4	Condiții de optimalitate pentru (UNLP)	68
4.1	Condiții de ordinul I pentru (UNLP)	70
4.2	Condiții de ordinul II pentru (UNLP)	73
4.3	Condiții de optimalitate pentru probleme convexe	75
4.4	Analiza perturbațiilor	77
5	Convergența metodelor de descreștere	79
5.1	Metode numerice de optimizare	80
5.2	Convergența metodelor numerice	84
5.3	Metode de descreștere	85
5.3.1	Strategii de alegere a lungimii pasului	85
5.3.2	Convergența metodelor de descreștere	87
6	Metode de ordinul I pentru (UNLP)	90
6.1	Metoda gradient	91
6.1.1	Convergența globală a metodei gradient	93
6.1.2	Rata de convergență globală a metodei gradient . .	95
6.1.3	Rata de convergență locală a metodei gradient . .	98
6.2	Metoda direcțiilor conjugate	99
6.2.1	Metoda direcțiilor conjugate pentru QP	100
6.2.2	Metoda gradientilor conjugați pentru QP	102
6.2.3	Metoda gradientilor conjugați pentru UNLP . . .	105
7	Metode de ordinul II pentru (UNLP)	108
7.1	Metoda Newton	109
7.1.1	Rata de convergență locală a metodei Newton . .	112
7.1.2	Convergența globală a metodei Newton	114
7.2	Metode quasi-Newton	117
7.2.1	Actualizări de rang unu	118
7.2.2	Actualizări de rang doi	119
7.2.3	Convergența locală a metodelor quasi-Newton . .	121

8 Probleme de estimare și fitting	123
8.1 Problema celor mai mici pătrate: cazul liniar	124
8.1.1 Probleme CMMP liniare prost condiționate . . .	127
8.1.2 Formularea statistică a problemelor CMMP liniare	129
8.2 Problema celor mai mici pătrate: cazul neliniar	130
8.2.1 Metoda Gauss-Newton (GN)	130
8.2.2 Metoda Levenberg-Marquardt	132
8.3 Aplicație: identificarea unui sistem Hammerstein	134
 III Optimizare cu constrângeri	 138
9 Teoria dualității	140
9.1 Funcția Lagrange	144
9.2 Problema duală	146
9.3 Programare liniară (LP)	153
10 Condiții de optimalitate pentru (NLP)	160
10.1 Condiții de ordinul I pentru (NLP) având constrângeri de egalitate	162
10.2 Condiții de ordinul II pentru (NLP) având constrângeri de egalitate	167
10.3 Condiții de ordinul I pentru (NLP) generale	172
10.4 Condiții de ordinul II pentru (NLP) generale	175
11 Metode de ordinul I și II pentru (NLP) având constrângeri convexe	181
11.1 Metode de direcții de descreștere	182
11.1.1 Metoda gradient condițional	185
11.1.2 Metoda gradient proiectat	186
11.2 Metoda Newton proiectat	190
12 Metode de optimizare pentru (NLP) având constrângeri de egalitate	194
12.1 Metode pentru QP cu constrângeri de egalitate	196
12.2 Metode Lagrange	199
12.2.1 Metoda Lagrange de ordinul I	201
12.2.2 Metoda Lagrange-Newton	204
12.3 Metoda Newton pentru probleme convexe având constrângeri de egalitate	207

13 Metode de optimizare pentru (NLP) generale	211
13.1 Metoda mulțimilor active	213
13.1.1 Metoda mulțimilor active pentru (QP)	216
13.2 Metoda pătratică secvențială	217
13.3 Metode de penalitate și barieră	222
13.3.1 Metode de penalitate	222
13.3.2 Metode de barieră	225
13.4 Metode de punct interior	227
13.4.1 Metode de punct interior pentru probleme convexe	230
13.4.2 Metode de punct interior pentru probleme neconvexe	232
14 Studii de caz din inginerie	236
14.1 Control optimal liniar	236
14.1.1 Formularea (QP) rară fără eliminarea stărilor . .	237
14.1.2 Formularea (QP) densă cu eliminarea stărilor . .	239
14.1.3 Control optimal pentru urmărirea traiectoriei cu un robot E-Puck	242
14.1.4 Control optimal pentru pendulul invers	246
14.2 Control optimal neliniar	249
14.2.1 Formularea (NLP) rară și densă	249
14.2.2 Control optimal aplicat unei instalații cu patru rezervoare	251
14.3 Stabilitatea sistemelor dinamice	254
14.3.1 Calcularea valorilor proprii ale unei matrice . . .	255
14.4 Problema Google (ierarhizarea paginilor web)	257
14.5 Învățare automată și clasificare	259
A Noțiuni de algebră liniară și analiză matematică	265
A.1 Noțiuni de analiză matriceală	265
A.2 Noțiuni de analiză matematică	269
Bibliografie	275

Partea I

Introducere

Capitolul 1

Teorie convexă

În acest capitol vom prezenta noțiunile de bază din teoria mulțimilor convexe și a funcțiilor convexe. Aceste noțiuni vor fi utilizate deseori în capitolele următoare. Prezentarea realizată în acest capitol este foarte concisă, dar în același timp suficientă pentru scopul declarat al lucrării. O detaliere a teoriei convexității poate fi găsită în cartea clasică a lui R.T. Rockafellar, *Convex Theory* [14].

Definim mai întâi conceptele algebrice de bază, cum ar fi mulțimile convexe, hiperplane, conuri, inegalități de matrice, dar și concepte topologice cu privire la conservarea convexității sau separarea prin hiperplane. Apoi continuăm cu teoria funcțiilor convexe și în special cu acele condiții de caracterizare a funcțiilor convexe.

În cadrul acestei lucrări fixăm simpla convenție de a considera vectorii $x \in \mathbb{R}^n$ vectori coloană, adică $x = [x_1 \dots x_n]^T \in \mathbb{R}^n$. În spațiul Euclidian \mathbb{R}^n produsul scalar este definit după cum urmează:

$$\langle x, y \rangle = x^T y = \sum_{i=1}^n x_i y_i.$$

Unde nu se specifică, norma considerată pe spațiul Euclidian \mathbb{R}^n este norma Euclidiană standard (adică norma indusă de acest produs scalar):

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{i=1}^n x_i^2}.$$

Conceptele standard din algebra liniară și analiza matematică folosite sunt prezentate în Apendice.

1.1 Teoria mulțimilor convexe

1.1.1 Mulțimi convexe

Mulțimile convexe au un rol important în teoria optimizării. Începem acest capitol cu prezentarea noțiunilor fundamentale din teoria mulțimilor convexe.

Definiția 1.1.1 O mulțime $S \subseteq \mathbb{R}^n$ este afină dacă pentru oricare doi vectori $x_1, x_2 \in S$ și orice scalar $\alpha \in \mathbb{R}$ avem $\alpha x_1 + (1 - \alpha)x_2 \in S$ (i.e. dreapta generată de oricare două puncte din S este inclusă în S).

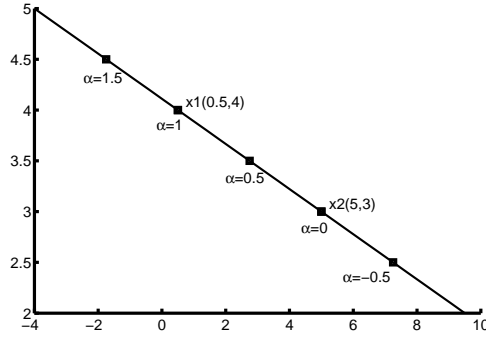


Figura 1.1: Mulțime afină (dreapta) generată de două puncte $x_1 = [0.5 \ 4]^T$ și $x_2 = [5 \ 3]^T$.

De exemplu, dreapta generată de două puncte este mulțime afină (vezi Fig. 1.1). Mulțimea soluțiilor unui sistem liniar $Ax = b$, unde $A \in \mathbb{R}^{m \times n}$ și $b \in \mathbb{R}^m$, este mulțime afină, i.e. mulțimea $\{x \in \mathbb{R}^n : Ax = b\}$ este afină.

O combinație afină de p vectori $\{x_1, \dots, x_p\} \subseteq \mathbb{R}^n$ este definită astfel:

$$\sum_{i=1}^p \alpha_i x_i, \quad \text{unde} \quad \sum_{i=1}^p \alpha_i = 1, \alpha_i \in \mathbb{R}.$$

Acoperirea afină a mulțimii $S \subseteq \mathbb{R}^n$, notată $\text{Aff}(S)$, reprezintă mulțimea ce conține toate combinațiile afine finite posibile ale punctelor din S :

$$\text{Aff}(S) = \left\{ \sum_{i \in \mathcal{I}, \mathcal{I} \text{ finit}} \alpha_i x_i : x_i \in S, \sum_{i \in \mathcal{I}} \alpha_i = 1, \alpha_i \in \mathbb{R} \right\}.$$

Cu alte cuvinte, $\text{Aff}(S)$ este mulțimea afină cea mai mică ce o conține pe mulțimea dată S .

Definiția 1.1.2 Mulțimea $S \subseteq \mathbb{R}^n$ se numește convexă dacă pentru oricare două puncte $x_1, x_2 \in S$ și un scalar $\alpha \in [0, 1]$ avem $\alpha x_1 + (1 - \alpha)x_2 \in S$ (i.e. segmentul generat de oricare două puncte din S este inclus în S , vezi Fig. 1.2).

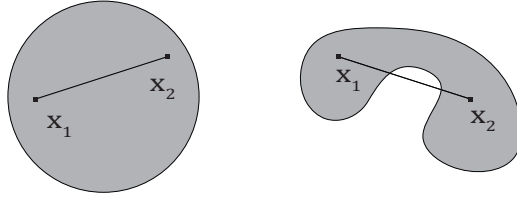


Figura 1.2: Exemplu de mulțime convexă (stânga) și mulțime neconvexă (dreapta).

Rezultă imediat că orice mulțime afină este mulțime convexă. Mai departe, o *combinație convexă* de p vectori $\{x_1, \dots, x_p\} \subset \mathbb{R}^n$ este definită de expresia:

$$\sum_{i=1}^p \alpha_i x_i, \quad \text{unde} \quad \sum_{i=1}^p \alpha_i = 1, \alpha_i \geq 0.$$

Acoperirea convexă a mulțimii S , notată $\text{Conv}(S)$, reprezintă mulțimea ce conține toate combinațiile convexe finite posibile ale punctelor mulțimii S , i.e. mulțimea:

$$\text{Conv}(S) = \left\{ \sum_{i \in \mathcal{I}, \mathcal{I} \text{ finit}} \alpha_i x_i : x_i \in S, \sum_{i \in \mathcal{I}} \alpha_i = 1, \alpha_i \geq 0 \right\}.$$

Se observă că acoperirea convexă a unei mulțimi este cea mai mică mulțime convexă ce conține mulțimea dată (vezi Fig. 1.3). Rezultă că dacă S este convexă, atunci acoperirea convexă a lui S coincide cu S .

Teorema 1.1.1 (Teorema lui Caratheodory) Dacă $S \subseteq \mathbb{R}^n$ este o mulțime convexă, atunci orice element din S este o combinație convexă de cel mult $n + 1$ vectori din S .

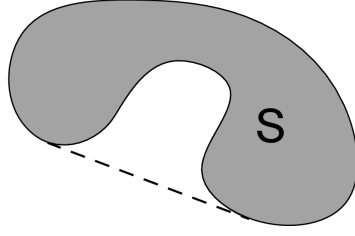


Figura 1.3: Acoperirea convexă a unei mulțimi neconvexe S .

Un *hiperplan* este o mulțime convexă definită de relația (vezi Fig. 1.4):

$$\{x \in \mathbb{R}^n : a^T x = b\} \quad a \neq 0, a \in \mathbb{R}^n, b \in \mathbb{R}.$$

Un *semiplan* este mulțimea convexă definită de relația (vezi Fig. 1.4):

$$\{x \in \mathbb{R}^n : a^T x \geq b\} \quad \text{sau} \quad \{x \in \mathbb{R}^n : a^T x \leq b\},$$

în care $a \neq 0, a \in \mathbb{R}^n$ și $b \in \mathbb{R}$.

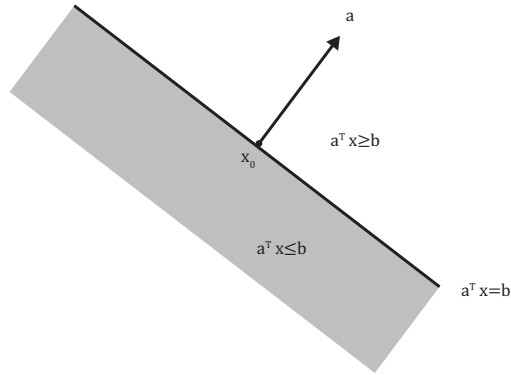


Figura 1.4: Hiperplanul definit de $a^T x = b$ și semiplanele corespunzătoare.

Un *poliedru* este mulțimea convexă definită de un număr p de hiperplane și/sau un număr m de semiplane:

$$\{x \in \mathbb{R}^n : a_i^T x = b_i \quad \forall i = 1, \dots, p, \quad c_j^T x \leq d_j \quad \forall j = 1, \dots, m\},$$

sau într-o formă compactă

$$\{x \in \mathbb{R}^n : Ax = b, \quad Cx \leq d\}.$$

O altă reprezentare a poliedrului poate fi dată de vârfurile (punctele de extrem) sale:

$$\left\{ \sum_{i=1}^{n_1} \alpha_i v_i + \sum_{j=1}^{n_2} \beta_j r_j : \sum_{i=1}^{n_1} \alpha_i = 1, \alpha_i \geq 0, \beta_j \geq 0 \quad \forall i, j \right\},$$

unde v_i se numesc vârfuri (numite și noduri), iar r_j sunt raze afine. Un *politop* reprezintă un poliedru mărginit și în acest caz el este definit numai de vârfuri (vezi Fig. 1.5).

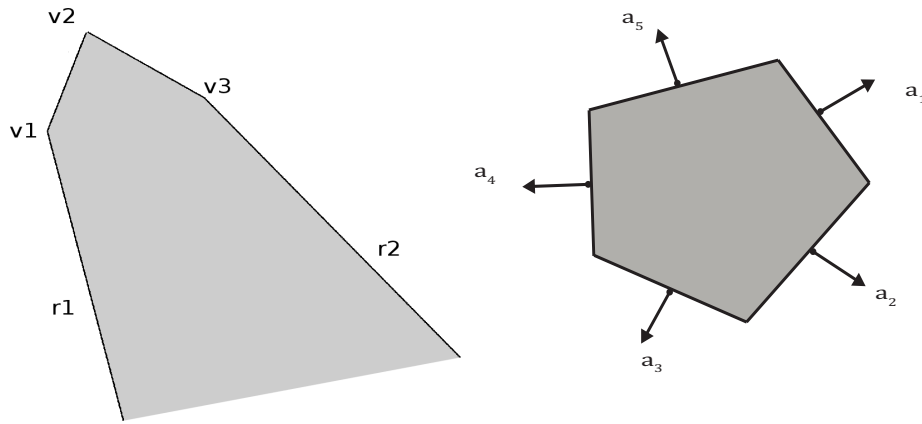


Figura 1.5: Poliedru nemărginit generat de trei vârfuri și două raze afine (stânga); poliedru mărginit (*politop*) format din intersecția a cinci semiplane (dreapta).

O *bilă* cu centrul în punctul $x_0 \in \mathbb{R}^n$ și raza $r > 0$ definită de norma Euclidiană $\|\cdot\|$ este o mulțime convexă dată de relația:

$$B(x_0, r) = \{x \in \mathbb{R}^n : \|x - x_0\| \leq r\}$$

sau, în mod echivalent:

$$B(x_0, r) = \{x \in \mathbb{R}^n : x = x_0 + ru, \|u\| \leq 1\}.$$

Un *elipsoid* este mulțimea convexă definită astfel:

$$\{x \in \mathbb{R}^n : (x - x_0)^T Q^{-1} (x - x_0) \leq 1\} = \{x_0 + Lu : \|u\| \leq 1\},$$

unde Q este o matrice simetrică pozitiv definită (notație $Q \succ 0$) și $Q = L^T L$, pentru o anumită matrice L .

1.1.2 Conuri

Definiția 1.1.3 O mulțime K se numește con dacă pentru orice $x \in K$ și $\alpha \in \mathbb{R}_+$ avem $\alpha x \in K$. Conul K se numește con convex dacă în plus K este mulțime convexă.

Combinatia conică de p vectori $\{x_1, \dots, x_p\} \subset \mathbb{R}^n$ este definită în felul următor:

$$\sum_{i=1}^p \alpha_i x_i, \quad \text{unde } \alpha_i \geq 0 \forall i.$$

Acoperirea conică a unei mulțimi S , notată $\text{Con}(S)$, reprezintă mulțimea ce conține toate combinațiile conice finite posibile cu elemente din S :

$$\text{Con}(S) = \left\{ \sum_{i \in \mathcal{I}, \mathcal{I} \text{ finit}} \alpha_i x_i : x_i \in S, \alpha_i \geq 0 \right\}.$$

Se observă că acoperirea conică a unei mulțimi reprezintă cel mai mic con ce conține mulțimea dată (vezi Fig. 1.6). Pentru un con K

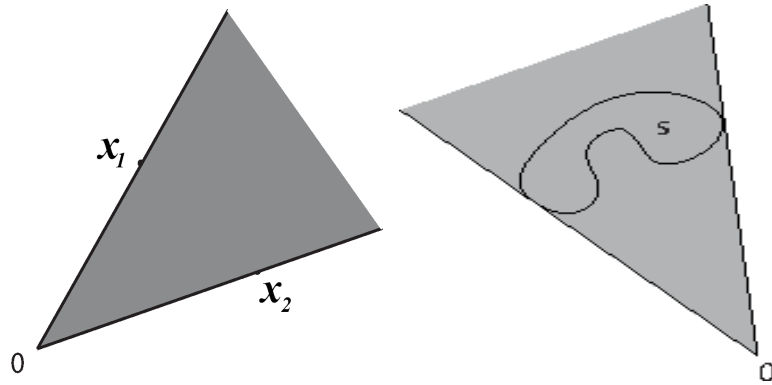


Figura 1.6: Acoperirea conică generată de doi vectori x_1 și x_2 (stânga); acoperirea conică generată de mulțimea S (dreapta).

dintr-un spațiu Euclidian înzestrat cu un produs scalar $\langle \cdot, \cdot \rangle$, conul dual corespunzător, notat K^* , este definit astfel:

$$K^* = \{y : \langle x, y \rangle \geq 0 \forall x \in K\}.$$

Observăm că conul dual este întotdeauna o mulțime închisă. Folosind relația $\langle x, y \rangle = \|x\| \cdot \|y\| \cos \angle(x, y)$, ajungem la concluzia că unghiul

dintre un vector ce aparține conului K și unul ce aparține conului K^* este mai mic decât $\frac{\pi}{2}$. Dacă conul K satisface condiția $K = K^*$, atunci mulțimea K se numește *con auto-dual*.

Exemplul 1.1.1 *Prezentăm în cele ce urmează câteva exemple de conuri:*

1. Mulțimea \mathbb{R}^n este un con, iar conul său dual este $(\mathbb{R}^n)^* = \{0\}$.
2. $\mathbb{R}_+^n = \{x \in \mathbb{R}^n : x \geq 0\}$ se numește conul orthant și este auto-dual în raport cu produsul scalar uzual $\langle x, y \rangle = x^T y$, i.e. $(\mathbb{R}_+^n)^* = \mathbb{R}_+^n$.
3. $\mathcal{L}^n = \{[x^T \ t]^T \in \mathbb{R}^{n+1} : \|x\| \leq t\}$ se numește conul Lorentz sau conul de înghețată și este, de asemenea, auto-dual în raport cu produsul scalar $\langle [x^T \ t]^T, [y^T \ v]^T \rangle = x^T y + tv$, i.e. $(\mathcal{L}^n)^* = \mathcal{L}^n$ (vezi Fig. 1.7).
4. $S_+^n = \{X \in S^n : X \succeq 0\}$ reprezintă conul semidefinit și este auto-dual în raport cu produsul scalar $\langle X, Y \rangle = \text{Tr}(XY)$, i.e. $(S_+^n)^* = S_+^n$.

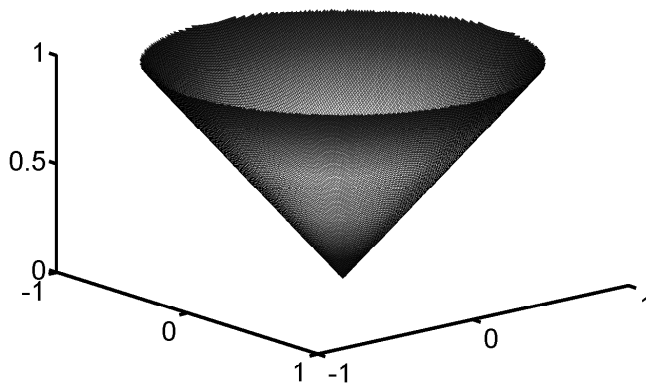


Figura 1.7: Conul Lorentz pentru $n = 2$.

1.1.3 Operații ce conservă convexitatea mulțimilor

Enumerăm câteva operații pe mulțimi care conservă proprietatea de convexitate:

1. Intersecția de mulțimi convexe este o mulțime convexă, i.e. dacă familia de mulțimi $\{S_i\}_{i \in \mathcal{I}}$ este convexă, atunci și $\bigcap_{i \in \mathcal{I}} S_i$ este convexă.
2. Suma a două mulțimi convexe S_1 și S_2 este de asemenea convexă, i.e. mulțimea $S_1 + S_2 = \{x + y : x \in S_1, y \in S_2\}$ este convexă. Mai mult, mulțimea $\alpha S = \{\alpha x : x \in S\}$ este convexă dacă mulțimea S este convexă și $\alpha \in \mathbb{R}$.
3. Translația unei mulțimi convexe S este de asemenea convexă, i.e. fie o funcție afină $f(x) = Ax + b$, atunci imaginea lui S prin f , $f(S) = \{f(x) : x \in S\}$, este convexă. Similar, preimaginea: $f^{-1}(S) = \{x : f(x) \in S\}$ este și ea convexă.
4. Definim o funcție $p: \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$, cu $\text{dom } p = \mathbb{R}^n \times \mathbb{R}_{++}$ prin $p(z, t) = z/t$, numită și funcție de perspectivă. Această funcție scalează (normalizează) vectorii astfel încât ultima componentă să fie 1, care apoi este eliminată (funcția returnează doar primele n componente din vectorul normalizat). Dacă $C \subseteq \text{dom } p$ este o mulțime convexă, atunci imaginea sa prin p , $p(C) = \{p(x) : x \in C\}$ este o mulțime convexă.
5. O funcție liniar-fracțională este formată prin compunerea funcției perspectivă cu o funcție afină. Fie o funcție afină $g: \mathbb{R}^n \rightarrow \mathbb{R}^{m+1}$, anume:

$$g(x) = \begin{bmatrix} A \\ c^T \end{bmatrix} x + \begin{bmatrix} b \\ d \end{bmatrix}$$

unde $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ și $d \in \mathbb{R}$. Funcția $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ dată de $f = p \circ g$, i.e.

$$f(x) = (Ax + b)/(c^T x + d), \quad \text{dom } f = \{x \in \mathbb{R}^n : c^T x + d > 0\}$$

se numește funcție liniar-fracțională. Astfel, dacă mulțimea C este convexă și aparține domeniului lui f , i.e. $c^T x + d > 0$ pentru $x \in C$, atunci imaginea sa prin f , $f(C)$, este convexă.

Inegalități Matriceale Liniare (*Linear Matrix Inequalities (LMI)*): Se poate arăta ușor că mulțimea matricelor pozitiv semidefinite (notație S_+^n) este convexă. Considerăm o funcție $G : \mathbb{R}^m \rightarrow S_+^n$, $G(x) = A_0 + \sum_{i=1}^m x_i A_i$, unde x_i sunt componentele unui vector $x \in \mathbb{R}^m$, iar matricele $A_0, \dots, A_m \in S^n$ sunt simetrice. Expresia

$$G(x) \succcurlyeq 0$$

se numește *inegalitate matriceală liniară* (LMI). Aceasta definește o mulțime convexă $\{x \in \mathbb{R}^m : G(x) \succcurlyeq 0\}$, cu rolul de preimagine a lui S_+^n prin $G(x)$.

Stabilitatea sistemelor: Fie un sistem liniar discret invariant în timp

$$z_{t+1} = Az_t,$$

unde $A \in \mathbb{R}^{n_z \times n_z}$ și z_t reprezintă starea sistemului la pasul t . Acest sistem este asimptotic stabil (adică $\lim_{t \rightarrow \infty} z_t = 0$ pentru orice stare inițială $z_0 \in \mathbb{R}^n$) dacă și numai dacă există o funcție Lyapunov pătratică $V(z) = z^T P z$ astfel încât:

$$V(z) > 0 \quad \forall z \in \mathbb{R}^n \quad \text{și} \quad V(z_{t+1}) - V(z_t) < 0 \quad \forall t \geq 0.$$

Aceste inegalități de matrice pot fi exprimate ca (LMI):

$$A^T P A - P \prec 0 \quad \text{și} \quad P \succ 0.$$

În mod echivalent, sistemul este asimptotic stabil dacă $\max_{i=1, \dots, n} |\lambda_i(A)| < 1$ (i.e. toate valorile proprii ale matricei A sunt incluse strict în cercul unitate). În reglarea automată (control) întâlnim adesea și inegalități matriceale cu necunoscutele P și R , de forma:

$$P - A^T R^{-1} A \succ 0, \quad P \succ 0$$

ce pot fi scrise (prin folosirea complementului Schur), în mod echivalent ca un LMI:

$$\begin{bmatrix} P & A^T \\ A & R \end{bmatrix} \succ 0.$$

Teorema 1.1.2 (Teorema de separare cu hiperplane) Fie S_1 și S_2 două mulțimi convexe astfel încât $S_1 \cap S_2 = \emptyset$. Atunci, există un hiperplan ce separă aceste mulțimi, i.e. există $a \neq 0, a \in \mathbb{R}^n$ și $b \in \mathbb{R}$ astfel încât $a^T x \geq b$ oricare ar fi $x \in S_1$ și $a^T x \leq b$ oricare ar fi $x \in S_2$.

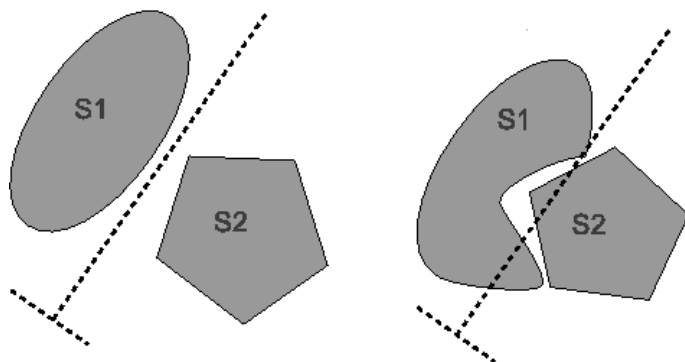


Figura 1.8: *Teorema de separare cu hiperplane.*

În Fig. 1.8 putem observa că pentru exemplul din dreapta mulțimea S_1 nu poate satisface teorema de separare cu hiperplane deoarece nu este convexă.

Teorema 1.1.3 (Teorema de suport cu un hiperplan)

Fie o mulțime convexă S și $x_0 \in bd(S) = cl(S) - int(S)$. Atunci, există un hiperplan de suport pentru S în punctul x_0 , adică există $a \neq 0, a \in \mathbb{R}^n$ astfel încât $a^T x \geq a^T x_0$ oricare ar fi $x \in S$.

1.2 Teoria funcțiilor convexe

1.2.1 Funcții convexe

În cadrul acestei lucrări ne vom concentra atenția preponderent asupra conceptelor, relațiilor și rezultatelor ce implică funcții al căror codomeniu este inclus în $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$. Pentru început, o observație importantă pentru rigurozitatea rezultatelor ce urmează este aceea că domeniul efectiv al unei funcții scalare f se poate extinde (prin echivalență) la întreg spațiul \mathbb{R}^n prin atribuirea valorii $+\infty$ funcției în toate punctele din afara domeniului său. În cele ce urmează considerăm că toate funcțiile sunt extinse implicit. O funcție scalară $f : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ are *domeniul efectiv* descris de mulțimea:

$$\text{dom } f = \{x \in \mathbb{R}^n : f(x) < +\infty\}.$$

Definiția 1.2.1 Funcția f se numește convexă dacă domeniul său efectiv $\text{dom} f$ este o mulțime convexă și următoarea relație are loc:

$$f(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha f(x_1) + (1 - \alpha)f(x_2),$$

pentru orice $x_1, x_2 \in \text{dom} f$ și $\alpha \in [0, 1]$. Dacă în plus

$$f(\alpha x_1 + (1 - \alpha)x_2) < \alpha f(x_1) + (1 - \alpha)f(x_2),$$

pentru orice $x_1 \neq x_2 \in \text{dom} f$ și $\alpha \in (0, 1)$, atunci f se numește funcție strict convexă.

Dacă există o constantă $\sigma > 0$ astfel încât

$$f(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha f(x_1) + (1 - \alpha)f(x_2) - \frac{\sigma}{2}\alpha(1 - \alpha)\|x_1 - x_2\|^2,$$

pentru orice $x_1, x_2 \in \text{dom} f$ și $\alpha \in [0, 1]$, atunci f se numește funcție tare convexă.

O funcție f se numește concavă dacă $-f$ este convexă.

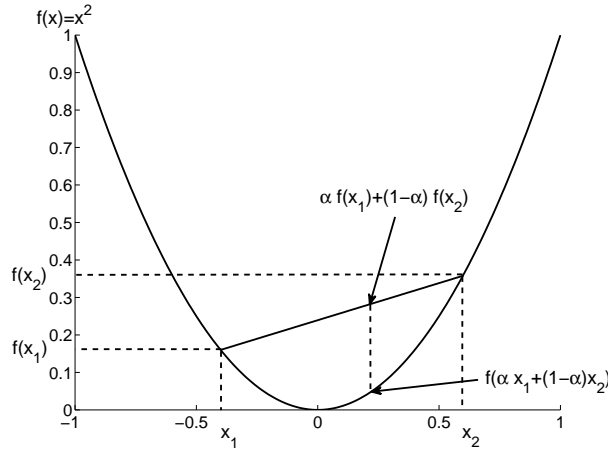


Figura 1.9: Exemplu de funcție convexă $f(x) = x^2$.

Exemplul 1.2.1 Dăm în continuare câteva exemple de funcții convexe:

1. Funcția definită de orice normă este convexă, i.e. $f(x) = \|x\|$ este convexă pe \mathbb{R}^n .

2. Funcția max definită astfel $f(x) = \max\{x_1, \dots, x_n\}$ este convexă pe \mathbb{R}^n .
3. Funcția $f(X) = -\log \det(X)$ este convexă pe spațiul matricelor pozitiv definite S_{++}^n .
4. Funcția dată de media geometrică $f(x) = (\prod_{i=1}^n x_i)^{1/n}$ este concavă pe \mathbb{R}_{++}^n .

Inegalitatea lui Jensen este o generalizare a definiției anterioare și ne spune că f este o funcție convexă dacă și numai dacă $\text{dom} f$ este mulțime convexă și

$$f\left(\sum_{i=1}^p \alpha_i x_i\right) \leq \sum_{i=1}^p \alpha_i f(x_i)$$

pentru orice $x_i \in \text{dom} f$ și $\sum_{i=1}^p \alpha_i = 1$ cu $\alpha_i \in [0, 1]$ pentru orice $i = 1, \dots, p$. Interpretarea geometrică a convexității este foarte simplă. Pentru o funcție convexă, fie două puncte din domeniul său $x_1, x_2 \in \text{dom} f$, atunci valorile funcției evaluate în punctele din intervalul $[x_1, x_2]$ sunt mai mici decât cele de pe segmentul cu capetele $(x_1, f(x_1))$ și $(x_2, f(x_2))$. Cu alte cuvinte, valorile funcției (convexe) în punctele $\alpha x_1 + (1 - \alpha)x_2$, pentru $\alpha \in [0, 1]$, sunt mai mici sau egale cu înălțimea corzii ce unește coordonatele $(x_1, f(x_1))$ și $(x_2, f(x_2))$ (vezi Fig. 1.9).

Remarca 1.2.1 O funcție $f : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ este convexă dacă și numai dacă restricția domeniului său la o dreaptă (care intersectează domeniul) este, de asemenea, convexă. Cu alte cuvinte, f este convexă dacă și numai dacă oricare ar fi $x \in \text{dom} f$ și o direcție $d \in \mathbb{R}^n$, funcția scalară $g(t) = f(x + td)$ este convexă pe domeniul $\{t \in \mathbb{R} : x + td \in \text{dom} f\}$. Această proprietate este utilă în anumite probleme în care se dorește să se arate convexitatea unei funcții.

1.2.2 Condiții de ordinul I pentru funcții convexe

În această secțiune prezentăm condițiile de convexitate de ordinul întâi pentru funcții diferențiabile:

Teorema 1.2.1 (Convexitatea funcțiilor de clasă C^1)

Presupunem că funcția $f : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ este continuu diferențiabilă și $\text{dom} f$ este o mulțime convexă. Atunci, f este convexă dacă și numai dacă:

$$f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) \quad \forall x_1, x_2 \in \text{dom} f. \quad (1.1)$$

Demonstrație: Mai întâi demonstrăm că dacă funcția este convexă atunci inegalitatea precedentă are loc. Din convexitatea lui f rezultă că pentru orice $x_1, x_2 \in \text{dom} f$ și oricare $\alpha \in [0, 1]$ avem:

$$f(x_1 + \alpha(x_2 - x_1)) - f(x_1) \leq \alpha(f(x_2) - f(x_1)).$$

Pe de altă parte, avem că:

$$\nabla f(x_1)^T(x_2 - x_1) = \lim_{\alpha \rightarrow +0} \frac{f(x_1 + \alpha(x_2 - x_1)) - f(x_1)}{\alpha} \leq f(x_2) - f(x_1).$$

de unde rezultă inegalitatea (1.1).

Pentru implicația inversă, se observă că pentru orice $z = x_1 + \alpha(x_2 - x_1) = (1 - \alpha)x_1 + \alpha x_2$ se satisface relația $f(z) \leq (1 - \alpha)f(x_1) + \alpha f(x_2)$. De aceea, prin aplicarea relației (1.1) de două ori în punctul z se obțin următoarele inegalități: $f(x_1) \geq f(z) + \nabla f(z)^T(x_1 - z)$ și $f(x_2) \geq f(z) + \nabla f(z)^T(x_2 - z)$. Prin înmulțirea cu ponderile $(1 - \alpha)$ și α și apoi, însumarea celor două relații avem:

$$(1 - \alpha)f(x_1) + \alpha f(x_2) \geq f(z) + \nabla f(z)^T \underbrace{[(1 - \alpha)(x_1 - z) + \alpha(x_2 - z)]}_{=(1 - \alpha)x_1 + \alpha x_2 - z = 0}.$$

□

Interpretarea relației anterioare este foarte simplă: tangenta la graficul unei funcții convexe în orice punct se află sub grafic. O consecință imediată a acestei teoreme este următoarea inegalitate: fie $f : \mathbb{R}^n \rightarrow \mathbb{R}$ o funcție convexă de clasă C^1 , atunci

$$\langle \nabla f(x_1) - \nabla f(x_2), x_1 - x_2 \rangle \geq 0 \quad \forall x_1, x_2 \in \text{dom} f.$$

1.2.3 Condiții de ordinul II pentru funcții convexe

Teorema 1.2.2 (Convexitatea funcțiilor de clasă C^2)

Fie o funcție $f : \mathbb{R}^n \rightarrow \mathbb{R}$ de două ori continuu diferențiabilă și presupunem că $\text{dom} f$ este mulțime convexă. Atunci f este convexă dacă și numai dacă pentru orice $x \in \text{dom} f$ matricea Hessiană este pozitiv semidefinită, adică:

$$\nabla^2 f(x) \succeq 0 \quad \forall x \in \text{dom} f. \quad (1.2)$$

Demonstrație: Mai întâi arătăm că dacă funcția este convexă atunci inegalitatea anterioară are loc. Folosim aproximarea Taylor de ordin II a lui f în punctul x într-o direcție arbitrară $d \in \mathbb{R}^n$:

$$f(x + td) = f(x) + t\nabla f(x)^T d + \frac{1}{2}t^2 d^T \nabla^2 f(x) d + \mathcal{R}(t^2 \|d\|^2).$$

De aici obținem:

$$d^T \nabla^2 f(x) d = \lim_{t \rightarrow 0} \frac{2}{t^2} \underbrace{(f(x + td) - f(x) - t\nabla f(x)^T d)}_{\geq 0, \text{ datorită (1.1)}} + \underbrace{\lim_{t \rightarrow 0} \frac{\mathcal{R}(t^2 \|d\|^2)}{t^2}}_{=0} \geq 0,$$

și deci $\nabla f(x)^2 \succcurlyeq 0$. Pe de altă parte, pentru demonstrația implicației inverse, folosim expresia restului Taylor pentru un parametru $\alpha \in [0, 1]$:

$$\begin{aligned} f(x_2) &= f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \\ &\quad \underbrace{\frac{1}{2}(x_2 - x_1)^T \nabla^2 f(x_1 + \alpha(x_2 - x_1))(x_2 - x_1)}_{\geq 0, \text{ datorită (1.2)}} \\ &\geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) \end{aligned}$$

și apoi utilizăm condițiile de ordinul I pentru funcții convexe. □

Exemplul 1.2.2

1. Funcția $f(x) = -\log(x)$ este convexă pe $\mathbb{R}_{++} = \{x \in \mathbb{R} : x > 0\}$ deoarece $\nabla^2 f(x) = \frac{1}{x^2} > 0$ oricare ar fi $x > 0$.
2. Funcția pătratică $f(x) = \frac{1}{2}x^T Qx + q^T x + r$ este convexă pe \mathbb{R}^n dacă și numai dacă $Q \succcurlyeq 0$, deoarece pentru orice $x \in \mathbb{R}^n$ Hessiana $\nabla^2 f(x) = Q$ (dacă Q este simetrică).
3. Se observă că orice funcție afină este convexă și, de asemenea, concavă.
4. Funcția $f(x, t) = \frac{x^T x}{t}$ este convexă pe $\mathbb{R}^n \times (0, \infty)$ deoarece matricea Hessiană

$$\nabla^2 f(x, t) = \begin{bmatrix} \frac{2}{t} I_n & -\frac{2}{t^2} x \\ -\frac{2}{t^2} x^T & \frac{2}{t^3} x^T x \end{bmatrix}$$

este pozitiv definită pe această mulțime. Pentru a scoate în evidență acest lucru, se înmulțește la dreapta și la stânga cu un vector $v = [z^T \ s]^T \in \mathbb{R}^{n+1}$ de unde rezultă $v^T \nabla^2 f(x, t) v = \frac{2}{t^3} \|tz - sx\|^2 \geq 0$ dacă $t > 0$.

5. Condiția $\text{dom} f$ mulțime convexă impusă în toate teoremele precedente este necesară. De exemplu, considerăm funcția $f(x) = 1/x^2$ având $\text{dom} f = \mathbb{R} \setminus \{0\}$ mulțime neconvexă. Observăm că Hessiana $\nabla^2 f(x) = \frac{6}{x^4} \succ 0$, dar funcția nu este convexă.

Teorema 1.2.3 (Convexitatea mulțimilor subnivel) Pentru un scalar $c \in \mathbb{R}$, mulțimea subnivel $\{x \in \text{dom} f : f(x) \leq c\}$ a unei funcții convexe $f : \mathbb{R}^n \rightarrow \mathbb{R}$ este convexă.

Demonstrație: Dacă $f(x_1) \leq c$ și $f(x_2) \leq c$ atunci pentru orice $\alpha \in [0, 1]$ funcția f satisface de asemenea:

$$f((1 - \alpha)x_1 + \alpha x_2) \leq (1 - \alpha)f(x_1) + \alpha f(x_2) \leq (1 - \alpha)c + \alpha c = c.$$

□

Mulțimile izonivel (contururile) unei funcții $f : \mathbb{R}^n \rightarrow \mathbb{R}$ sunt mulțimile de forma $\{x \in \mathbb{R}^n : f(x) = c\}$.

Epigraful funcției: Fie o funcție $f : \mathbb{R}^n \rightarrow \mathbb{R}$, atunci *epigraful* (numit și *supragrafic*) funcției este definit ca fiind următoarea mulțime:

$$\text{epi} f = \{[x^T \ t]^T \in \mathbb{R}^{n+1} : x \in \text{dom} f, \ f(x) \leq t\}.$$

Teorema 1.2.4 (Proprietatea de convexitate a epigrafului)

O funcție $f : \mathbb{R}^n \rightarrow \mathbb{R}$ este convexă dacă și numai dacă epigraful său este o mulțime convexă.

1.2.4 Operații ce conservă convexitatea funcțiilor

În cele ce urmează prezentăm câteva operații pe funcții convexe care conservă proprietatea de convexitate:

1. Dacă f_1 și f_2 sunt funcții convexe și $\alpha_1, \alpha_2 \geq 0$ atunci $\alpha_1 f_1 + \alpha_2 f_2$ este de asemenea convexă.
2. Dacă f este convexă atunci $g(x) = f(Ax + b)$ (adică compunerea unei funcții convexe cu o funcție afină) este de asemenea convexă.
3. Fie $f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ astfel încât funcția $f(\cdot, y)$ este convexă pentru orice $y \in S \subseteq \mathbb{R}^m$. Atunci următoarea funcție este convexă:

$$g(x) = \sup_{y \in S} f(x, y).$$

4. Compunerea cu o funcție convexă monotonă unidimensională: dacă $f : \mathbb{R}^n \rightarrow \mathbb{R}$ este convexă și $g : \mathbb{R} \rightarrow \mathbb{R}$ este convexă și monoton crescătoare, atunci funcția $g \circ f : \mathbb{R}^n \rightarrow \mathbb{R}$ este de asemenea convexă.
5. Dacă g și f sunt multidimensionale, i.e. $g : \mathbb{R}^k \rightarrow \mathbb{R}$, iar $f : \mathbb{R}^n \rightarrow \mathbb{R}^k$, atunci pentru funcția $h : \mathbb{R}^n \rightarrow \mathbb{R}$, $h = g \circ f$, i.e. $h(x) = g(f(x)) = g(f_1(x), \dots, f_k(x))$, unde $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$, putem afirma:
 - (i) h este convexă dacă g este convexă, g este monoton crescătoare, în fiecare argument, iar toate funcțiile f_i sunt convexe
 - (ii) h este convexă dacă g este convexă, g este monoton crescătoare în fiecare argument, iar toate funcțiile f_i sunt concave.

Funcții conjugate: Fie funcția $f : \mathbb{R}^n \rightarrow \mathbb{R}$, atunci funcția *conjugată*, notată cu f^* , se definește prin

$$f^*(y) = \sup_{x \in \text{dom } f} \underbrace{y^T x - f(x)}_{F(x,y)}.$$

Din discuția precedentă rezultă că funcția conjugată f^* este convexă indiferent de proprietățile lui f . Mai mult, $\text{dom } f^* = \{y \in \mathbb{R}^n : f^*(y) \text{ finit}\}$. O altă consecință evidentă a definiției este *inegalitatea Fenchel*:

$$f(x) + f^*(y) \geq y^T x \quad \forall x \in \text{dom } f, y \in \text{dom } f^*.$$

Exemplul 1.2.3 Pentru funcția pătratică convexă $f(x) = \frac{1}{2}x^T Qx$, unde $Q \succ 0$, funcția conjugată are expresia:

$$f^*(y) = \frac{1}{2}y^T Q^{-1}y.$$

Pentru funcția $f(x) = -\log x$, conjugata sa este dată de expresia:

$$f^*(y) = \sup_{x>0} (xy + \log x) = \begin{cases} -1 - \log(-y) & \text{dacă } y < 0 \\ \infty & \text{altfel.} \end{cases}$$

Capitolul 2

Concepte fundamentale din teoria optimizării

În cadrul acestui capitol prezentăm o scurtă istorie a evoluției optimizării și apoi introducem conceptele fundamentale ale unei probleme de optimizare (valoarea optimă, puncte de extrem, mulțimea fezabilă) împreună cu principalele clase de probleme de optimizare (probleme de optimizare neliniare, liniare, convexe, pătratice).

2.1 Evoluția teoriei optimizării

Optimizarea are aplicații în extrem de numeroase domenii, dintre care putem aminti următoarele (aplicații concrete din inginerie sunt prezentate în Capitolul 14):

- *economie*: alocarea resurselor în logistică, investiții, calcularea unui portofoliu optim;
- *științele exacte*: estimarea și proiectarea de modele pentru seturi de date măsurate, proiectarea de experimente;
- *inginerie*: proiectarea și operarea în domeniul sistemelor tehnologice (poduri, autovehicule, dispozitive electronice), optimizarea motoarelor de căutare, control optimal, procesarea de semnale.

Apariția teoriei optimizării în problemele de extrem (minimum/maximum) datează cu câteva secole înaintea lui Hristos.

Matematicienii din Antichitate manifestau interes pentru un număr de probleme de tip izoperimetric: de exemplu care este curba închisă de lungime fixată ce înconjoară suprafața de arie maximă? În această perioadă au fost folosite abordări *geometrice* pentru rezolvarea problemelor de optimizare și determinarea punctului de optim. Cu toate acestea, o soluție riguroasă pentru aceste tipuri de probleme nu a fost găsită până în secolul al XIX-lea. Problema izoperimetrică își are originile în legenda reginei Dido, descrisă de Virgil în *Eneida*. În primul capitol, Virgil ne povestește cum, ținută în captivitate de propriul său frate, regina evadează și stabilește fundația viitorului oraș al Cartaginei prin delimitarea sa cu fâșii din piele de bizon. În acest fel ia naștere problema înconjurării unei suprafețe de arie maximă având constrângerea ca perimetrul figurii rezultate să fie constant. Legenda spune că fenicienii au tăiat pielea de bizon în fâșii subțiri și, în acest fel, au reușit să îngrădească o suprafață foarte mare. Nu este exclus faptul ca supușii reginei să fi rezolvat o versiune practică a problemei. Fundația Cartaginei datează din secolul al IX-lea î.H. când nu exista nici o urmă a geometriei Euclidiene. Problema reginei Dido are o soluție unică în clasa figurilor convexe cu condiția ca partea fixată a frontierei să fie o linie convexă poligonală.

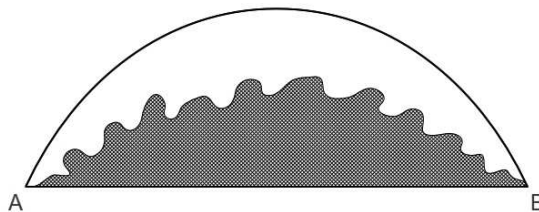


Figura 2.1: Problema reginei Dido (problema izoperimetrică).

Există și alte metode pe care matematicienii din perioada ce precedă calculul diferențial le foloseau pentru a rezolva probleme de optimizare, și anume abordările *algebrice*. Una dintre cele mai elegante este inegalitatea mediilor:

$$\frac{x_1 + \cdots + x_n}{n} \geq (x_1 \cdots x_n)^{1/n} \quad \forall x_i \geq 0, n \geq 1,$$

satisfăcută ca egalitate dacă și numai dacă $x_1 = \cdots = x_n$. O simplă aplicație este următoarea: pentru a arăta că din mulțimea tuturor dreptunghiurilor cu arie fixă, pătratul are cel mai mic perimetru, putem

folosi această simplă inegalitate algebrică: dacă notăm cu x și y laturile dreptunghiului, atunci problema se reduce la a determina anumite valori pentru x și y astfel încât să se minimizeze perimetrul $2(x + y)$ cu constrângerea $xy = A$, unde A este aria dată. Din inegalitatea mediilor avem

$$\frac{x + y}{2} \geq \sqrt{xy} = \sqrt{A},$$

cu egalitate dacă $x = y = \sqrt{A}$, adică figura este un pătrat.

Optimizarea deciziilor a devenit o știință începând cu a doua jumătate a secolului XIX-lea, perioadă în care *calculul diferențial* a fost puternic dezvoltat. Folosirea metodei gradient (adică utilizarea derivatei funcției obiectiv) pentru minimizare a fost prezentată de Cauchy în 1847. Metodele de optimizare moderne au fost pentru prima dată propuse într-o lucrare de Courant (1943) unde a fost introdusă noțiunea de funcție penalitate, în lucrarea lui Dantzig (1951) unde a fost prezentată metoda simplex pentru programare liniară și la Karush-Kuhn-Tucker care derivează condițiile de optimalitate KKT pentru probleme de optimizare constrânsă (1939,1951). Apoi, în anii 1960 au fost propuse foarte multe metode pentru a rezolva probleme de optimizare neliniară: metode pentru optimizare fără constrângeri, cum ar fi metoda gradientilor conjugați dată de Fletcher și Reeves (1964), metode de tip quasi-Newton dată de Davidon-Fletcher-Powell (1959). Metode de optimizare cu constrângeri au fost propuse de Rosen (metoda gradientului proiectat), Zoutendijk a propus metoda direcțiilor fezabile (1960), Fiacco și McCormick propun încă din anii 1970 metodele de punct interior și exterior. Metodele de programare pătratică secvențială (SQP) au fost de asemenea propuse în anii 1970. Dezvoltarea de metode de punct interior pentru programarea liniară a început cu lucrarea lui Karmarkar (1984). Această lucrare și brevetarea ei ulterioară au determinat comunitatea academică să se reorienteze iarăși în direcția metodelor de punct interior, culminând cu apariția cărții lui Nesterov și Nemirovski în 1994. Pe lângă metodele de tip gradient, au fost dezvoltate și alte tipuri de metode care nu se bazează pe informația de gradient. În această direcție putem aminti metoda simplex a lui Nelder și Meade (1965). Metode speciale care exploatează structura particulară a unei probleme au fost de asemenea dezvoltate încă din anii 1960. A apărut și programarea dinamică, ce se bazează pe rezultatele lui Bellman (1952). Lasdon a atras atenția asupra problemelor de dimensiuni mari prin cartea publicată în 1970. Optimalitatea Pareto a fost dezvoltată

pentru optimizarea multiobiectiv. Totodata, au fost dezvoltate și metode heuristice: algoritmi genetici (1975).

Un exemplu de problemă simplă de inginerie civilă ce poate fi rezolvată folosind calculul diferențial este prezentat în cele ce urmează. Fie două orașe localizate pe maluri diferite ale unui râu cu lățime constantă w ; orașele se află la distanța a și respectiv b de râu, cu o separare laterală d (Fig. 2.2). Problema constă în a afla locul de construcție a unui pod pentru a face cât mai scurtă posibil călătoria între cele două orașe. Această problemă se poate pune ca o problemă de optimizare:

$$\min_x f(x),$$

unde $f(x) = \sqrt{x^2 + a^2} + w + \sqrt{b^2 + (d - x)^2}$. Considerând condițiile de optimalitate prezentate ulterior, impunem derivata $f'(x) = 0$ și obținem locația optimă: $x^* = \frac{a}{a+b}d$.

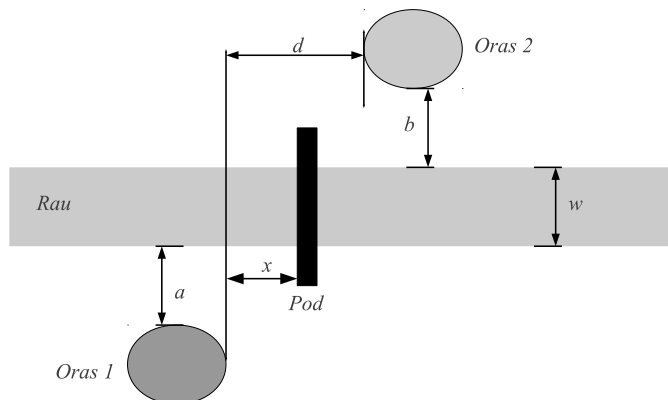


Figura 2.2: Aplicație a localizării optime.

2.2 Caracteristicile unei probleme de optimizare

O problemă de optimizare conține următoarele trei ingrediente:

- (i) o funcție obiectiv, $f(x)$, ce va fi minimizată sau maximizată;
- (ii) variabile de decizie, x , care se pot alege dintr-o anumită mulțime;

- (iii) constrângeri (numite și restricții) ce vor fi respectate, de forma $g(x) \leq 0$ (constrângeri de inegalitate) și/sau $h(x) = 0$ (constrângeri de egalitate).

Formularea matematică standard a unei probleme de optimizare este următoarea:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.l.:} \quad & g_1(x) \leq 0, \dots, g_m(x) \leq 0 \\ & h_1(x) = 0, \dots, h_p(x) = 0. \end{aligned}$$

Dacă introducem notațiile compacte $g(x) = [g_1(x) \dots g_m(x)]^T$ și $h(x) = [h_1(x) \dots h_p(x)]^T$, atunci în forma compactă problema de optimizare anterioară se scrie ca:

$$\begin{aligned} (NLP) : \quad & \min_{x \in \mathbb{R}^n} f(x) \\ \text{s.l.:} \quad & g(x) \leq 0, \quad h(x) = 0. \end{aligned}$$

În această problemă, funcția obiectiv $f : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$, funcția vectorială ce definește constrângerile de inegalitate $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ și funcția vectorială ce definește constrângerile de egalitate $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$ se presupun de obicei a fi diferențiabile. Observăm că o problemă de maxim se poate reduce la una de minim datorită echivalenței

$$\max_x f(x) = - \min_x -f(x).$$

Exemplul 2.2.1 *Considerăm următoarea problemă de optimizare:*

$$\begin{aligned} \min_{x \in \mathbb{R}^2} \quad & x_1^2 + x_2^2 \\ \text{s.l.:} \quad & x \geq 0, \quad x_1^2 + x_2 - 1 \leq 0 \\ & x_1 x_2 - 1 = 0. \end{aligned}$$

În acest exemplu avem:

- variabila de decizie este $x = [x_1 \ x_2]^T \in \mathbb{R}^2$ și funcția obiectiv $f(x) = x_1^2 + x_2^2$ este funcție pătratică convexă;

- avem trei constrângeri de inegalitate: $g : \mathbb{R}^2 \rightarrow \mathbb{R}^3$, unde $g_1(x) = -x_1$, $g_2(x) = -x_2$ și $g_3(x) = x_1^2 + x_2 - 1$. Observăm că funcția g are toate cele trei componente funcții convexe;
- o singură constrângere de egalitate definită de funcția $h(x) = x_1x_2 - 1$. Observăm că funcția h nu este convexă.

Definiția 2.2.1

1. Mulțimea $\{x \in \mathbb{R}^n : f(x) = c\}$ este mulțimea nivel (conturul) a funcției f pentru valoarea $c \in \mathbb{R}$.
2. Mulțimea fezabilă a problemei de optimizare (NLP) este:

$$X = \{x \in \mathbb{R}^n : g(x) \leq 0, h(x) = 0\}.$$

3. Punctul $x^* \in \mathbb{R}^n$ este un punct de minim global (adesea denumit minim global) dacă și numai dacă $x^* \in X$ și $f(x^*) \leq f(x)$ oricare ar fi $x \in X$.
4. Punctul $x^* \in \mathbb{R}^n$ este un punct strict de minim global dacă și numai dacă $x^* \in X$ și $f(x^*) < f(x)$ oricare ar fi $x \in X \setminus \{x^*\}$.
5. Punctul $x^* \in \mathbb{R}^n$ este minim local dacă și numai dacă $x^* \in X$. și există o vecinătate \mathcal{N} a lui x^* (e.g. o bilă deschisă cu centrul în x^*) astfel încât $f(x^*) \leq f(x)$ oricare ar fi $x \in X \cap \mathcal{N}$.
6. Punctul $x^* \in \mathbb{R}^n$ este un punct strict de minim local dacă și numai dacă $x^* \in X$ și există o vecinătate \mathcal{N} a lui x^* astfel încât $f(x^*) < f(x)$ oricare ar fi $x \in (X \cap \mathcal{N}) \setminus \{x^*\}$.

Exemplul 2.2.2 Pentru următoarea problemă unidimensională având constrângeri de tip box (vezi Fig. 2.3)

$$\begin{aligned} \min_{x \in \mathbb{R}} \quad & \frac{\cos 5\pi x}{x} \\ \text{s.l.:} \quad & x \geq 0.1, x \leq 1.1 \end{aligned}$$

- funcția obiectiv este $f(x) = \frac{\cos 5\pi x}{x}$, iar mulțimea fezabilă este un politop (interval) $X = \{x \in \mathbb{R} : x \geq 0.1, x \leq 1.1\} = [0.1, 1.1]$;

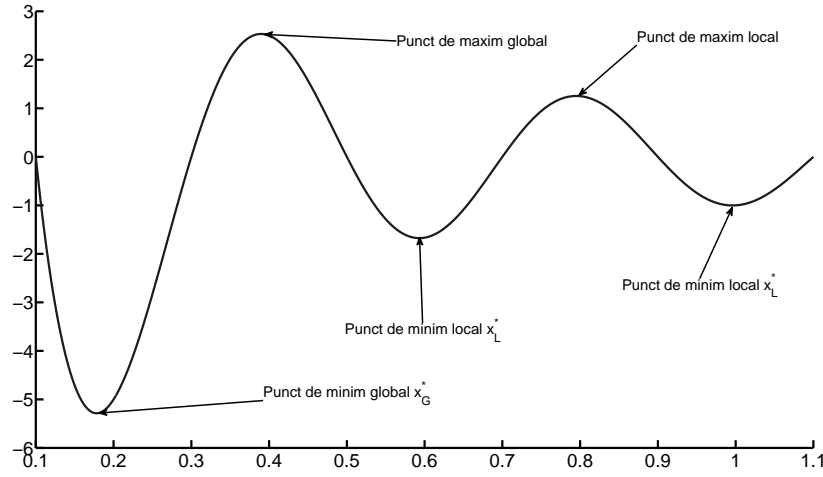


Figura 2.3: Puncte de minim local (x_L^*) și punctul de minim global (x_G^*) pentru $f(x) = \cos(5\pi x)/x$ în intervalul $[0.1, 1.1]$.

- reprezentând grafic funcția în Matlab (Fig. 2.3) putem identifica trei puncte de minim local, dintre care unul singur este de minim global.

Exemplul 2.2.3 Considerăm următoarea problemă de optimizare:

$$\begin{aligned} & \min_{x \in \mathbb{R}^2} (x_1 - 3)^2 + (x_2 - 2)^2 \\ & \text{s.l.: } x_1^2 - x_2 - 3 \leq 0, \quad x_2 - 1 \leq 0, \quad -x_1 \leq 0. \end{aligned}$$

Funcția obiectiv și cele trei constrângeri de inegalitate sunt: $f(x_1, x_2) = (x_1 - 3)^2 + (x_2 - 2)^2$ și respectiv $g_1(x_1, x_2) = x_1^2 - x_2 - 3$, $g_2(x_1, x_2) = x_2 - 1$, $g_3(x_1, x_2) = -x_1$.

Fig 2.4 ilustrează mulțimea fezabilă și contururile funcției obiectiv. Problema se reduce la a găsi un punct în mulțimea fezabilă în care funcția obiectiv, $(x_1 - 3)^2 + (x_2 - 2)^2$, ia cea mai mică valoare. Observăm că punctele $[x_1 \ x_2]^T$ cu $(x_1 - 3)^2 + (x_2 - 2)^2 = c$ sunt cercuri de rază c și centru în $[3 \ 2]^T$. Aceste cercuri se numesc contururile funcției obiectiv având valoarea c . Pentru a minimiza f trebuie să găsim cercul cu cea mai mică rază c care intersectează mulțimea fezabilă. După cum se observă din Fig. 2.4, cel mai mic cerc corespunde lui $c = 2$ și intersectează mulțimea fezabilă în punctul de optim $x^* = [2 \ 1]^T$.

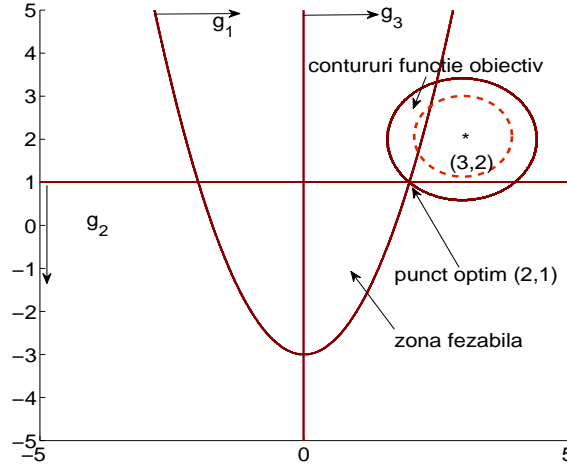


Figura 2.4: Soluția grafică a problemei de optimizare.

În teoria optimizării, un aspect important îl reprezintă existența punctelor de minim. Următoarea teoremă ne arată când astfel de puncte de optim există:

Teorema 2.2.1 (Weierstrass) *Dacă mulțimea fezabilă $X \subset \mathbb{R}^n$ este compactă (adică mărginită și închisă) și $f : X \rightarrow \mathbb{R}$ este continuă atunci există un punct de minim global pentru problema de minimizare $\min_{x \in X} f(x)$.*

Demonstrație: Observăm că graficul funcției f poate fi reprezentat prin $G = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} : x \in X, f(x) = t\}$. Mulțimea G este o mulțime compactă și proiecția lui G pe ultima sa coordonată este de asemenea compactă. Mai exact, mulțimea $\text{Proj}_{\mathbb{R}} G = \{t \in \mathbb{R} : \exists x \in \mathbb{R}^n \text{ astfel încât } (x, t) \in G\}$ este un interval compact $[f_{\min}, f_{\max}] \subset \mathbb{R}$. Prin construcție, există cel puțin un $x^* \in X$ astfel încât $(x^*, f_{\min}) \in G$. \square

Din teorema anterioară concluzionăm că punctele de minim există în condiții relativ generale, de aceea în această lucrare utilizăm *min* în loc de *inf*. Cu toate că demonstrația a fost constructivă, nu conduce către un algoritm eficient pentru a găsi punctul de minim. Scopul acestei lucrări este de a prezenta principalii algoritmi numerici de optimizare care determină punctele de optim.

2.3 Tipuri de probleme de optimizare

Pentru alegerea algoritmului potrivit pentru o problemă practică, avem nevoie de o clasificare a algoritmilor existenți și informații despre structurile matematice exploatate de ei. Înlocuirea unui algoritm inadecvat cu unul eficient poate scurta găsirea soluției cu mai multe ordine de magnitudine în ceea ce privește complexitatea aritmetică (numărul total de flopi).

2.3.1 Programare neliniară (NLP)

Acest curs tratează în principal algoritmi proiectați pentru rezolvarea de probleme generale de Programare Neliniară (*NLP - NonLinear Programming*) de forma:

$$(NLP) : \quad \min_{x \in \mathbb{R}^n} f(x) \quad (2.1)$$

$$\text{s.l.: } g(x) \leq 0, \quad h(x) = 0,$$

în care funcțiile $f : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$, $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ și $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$ se presupun a fi continuu diferențiabile cel puțin o dată, iar în unele cazuri de două sau de mai multe ori diferențiabile.

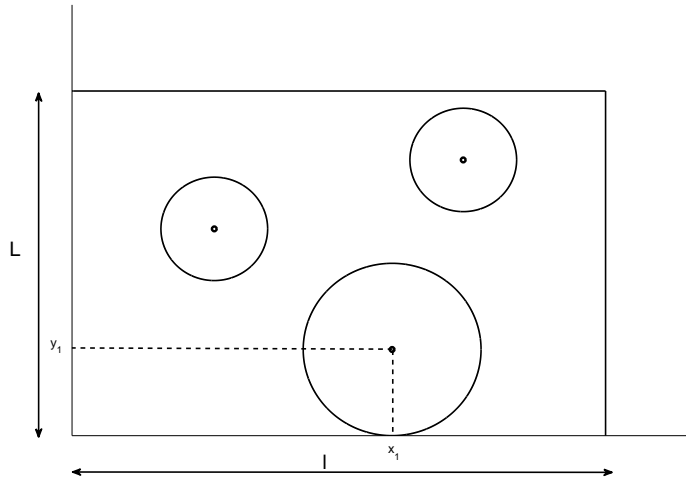


Figura 2.5: Problema de împachetare.

Minimizarea dimensiunii unui pachet - Care sunt dimensiunile celui mai mic pachet ce conține trei obiecte rotunde, de raze r_1, r_2 și r_3 date? Considerăm problema în \mathbb{R}^2 , extensia în \mathbb{R}^3 este imediată. Notăm cu (x_i, y_i) coordonatele plane ale celor trei obiecte și cu l și L laturile pachetului. Dorim să minimizăm aria $l \cdot L$ astfel încât au loc următoarele constrângeri:

- fiecare obiect se află în pachet:

$$x_i \geq r_i, y_i \geq r_i, \quad x_i \leq l - r_i, y_i \leq L - r_i \quad \forall i = 1, 2, 3$$

- dimensiunile sunt numere pozitive:

$$x_i \geq 0, y_i \geq 0, l \geq 0, L \geq 0 \quad \forall i = 1, 2, 3$$

- cele trei obiecte nu se suprapun:

$$(x_i - x_j)^2 + (y_i - y_j)^2 \geq (r_i + r_j)^2 \quad \forall i \neq j = 1, 2, 3.$$

În acest caz, problema precedentă se poate pune ca o problemă de optimizare (NLP) unde variabila de decizie este $x = [x_1 \ y_1 \ x_2 \ y_2 \ x_3 \ y_3 \ l \ L]^T$:

$$\begin{aligned} \min_{x \in \mathbb{R}^8} \quad & l \cdot L \\ \text{s.l.:} \quad & x \geq 0, x_i \geq r_i, y_i \geq r_i, x_i \leq l - r_i, y_i \leq L - r_i \\ & (x_i - x_j)^2 + (y_i - y_j)^2 \geq (r_i + r_j)^2 \quad \forall i \neq j = 1, 2, 3. \end{aligned}$$

Observăm că în problema anterioară funcția obiectiv $f(x) = l \cdot L$ nu este convexă. În anumite situații însă, multe dintre probleme prezintă structuri particulare, care pot fi exploatate pentru o rezolvare mai rapidă a acestora. În cele ce urmează enumerăm cele mai importante clase de probleme de optimizare cu structură specială.

2.3.2 Programare liniară (LP)

În cazul în care funcțiile f, g și h din formularea generală (2.1) sunt afine, problema (NLP) devine un Program Liniar (*LP - Linear Program*). Mai exact, un (LP) poate fi definit ca:

$$\begin{aligned} (LP) : \quad & \min_{x \in \mathbb{R}^n} c^T x \\ \text{s.l.:} \quad & Cx - d \leq 0, \quad Ax - b = 0. \end{aligned} \tag{2.2}$$

Datele problemei sunt: $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{p \times n}$, $b \in \mathbb{R}^p$, $C \in \mathbb{R}^{m \times n}$ și $d \in \mathbb{R}^m$. Se observă că putem adăuga o constantă la funcția obiectiv, adică $f(x) = c^T x + c_0$, însă aceasta nu schimbă punctul de minim x^* .

Aplicație financiară: Considerăm un număr n de active financiare și x_i reprezintă suma investită în activul i . Notăm cu $r_i(t)$ rata de rentabilitate într-un an, iar cu c_i rata de rentabilitate medie peste o perioadă de T ani a produsului i (i.e. $c_i = \frac{1}{T} \sum_{t=1}^T r_i(t)$). Dorim să maximizăm profitul:

$$\begin{aligned} \max_{x \in \mathbb{R}^n} \quad & \sum_{i=1}^n c_i x_i \\ \text{s.l.:} \quad & x \geq 0, \quad \sum_{i=1}^n x_i = 1. \end{aligned}$$

LP-urile pot fi rezolvate foarte eficient. Încă din anii 1940 aceste probleme au putut fi rezolvate cu succes, odată cu apariția *metodei simplex* dezvoltată de G. Dantzig. Metoda simplex este o metodă de tip *mulțime activă* și care este în competiție cu o clasă la fel de eficientă de algoritmi numiți *algoritmi de punct interior*. În zilele noastre se pot rezolva LP-uri chiar și cu milioane de variabile și constrângeri, orice student din domeniul finanțelor având în curriculum principalele metode de rezolvare a acestora. Algoritmii specializați pentru LP nu sunt tratați în detaliu în acest curs, însă trebuie recunoscuți atunci când sunt întâlniți în practică având la dispoziție mai multe produse software: CPLEX, Soplex, lp_solve, lingo, MATLAB (linprog), SeDuMi, YALMIP, CVX.

2.3.3 Programare pătratică (QP)

Dacă în formularea generală (NLP) dată în (2.1) constrângerile g și h sunt afine (ca și în cazul problemei LP), însă funcția obiectiv este o funcție pătratică, problema care rezultă se numește Problemă de Programare Pătratică (*QP - Quadratic Program*). O problemă generală QP poate fi formulată după cum urmează:

$$\begin{aligned} (QP) : \quad & \min_{x \in \mathbb{R}^n} \quad \frac{1}{2} x^T Q x + q^T x + r \\ \text{s.l.:} \quad & Cx - d \leq 0, \quad Ax - b = 0. \end{aligned} \tag{2.3}$$

Aici, în plus față de datele problemei LP, dacă $Q = Q^T$, avem *matricea Hessiană* $Q \in \mathbb{R}^{n \times n}$. Numele său provine din relația $\nabla^2 f(x) = Q$, unde $f(x) = \frac{1}{2} x^T Q x + q^T x + r$.

Dacă matricea Hessiană Q este pozitiv semidefinită (i.e. $Q \succeq 0$) atunci numim problema QP (2.3) o problemă *QP convexă*. QP-urile convexe sunt cu mult mai ușor de rezolvat global decât *QP-urile neconvexe* (i.e. unde matricea Hessiană Q este indefinită). După cum vom arăta în capitolele următoare, o problemă convexă are numai minime globale în timp ce una neconvexă poate avea multe minime locale. Dacă matricea Hessiană Q este pozitiv definită (i.e. $Q \succ 0$) numim problema QP (2.3) o problemă *QP strict convexă*. QP-urile strict convexe sunt o subclasă a problemelor QP convexe, dar de cele mai multe ori mai ușor de rezolvat decât QP-urile care nu sunt strict convexe.

Exemplul 2.3.1 *Exemplu de QP care nu este convex:*

$$\begin{aligned} \min_{x \in \mathbb{R}^2} \quad & \frac{1}{2} x^T \begin{bmatrix} 5 & 0 \\ 0 & -1 \end{bmatrix} x + \begin{bmatrix} 0 \\ 2 \end{bmatrix}^T x \\ \text{s.l.:} \quad & -1 \leq x_1 \leq 1, \quad -1 \leq x_2 \leq 10. \end{aligned}$$

Această problemă are minime locale în $x_1^* = [0 \ -1]^T$ și $x_2^* = [0 \ 10]^T$, însă doar x_2^* este punct de minim global pentru această problemă.

Exemplu de QP strict convex:

$$\begin{aligned} \min_{x \in \mathbb{R}^2} \quad & \frac{1}{2} x^T \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix} x + \begin{bmatrix} 0 \\ 2 \end{bmatrix}^T x \\ \text{s.l.:} \quad & -1 \leq x_1 \leq 1, \quad -1 \leq x_2 \leq 10. \end{aligned}$$

Problema anterioară are un punct de minim local (strict) unic în $x^* = [0 \ -1]^T$ care este de asemenea minim global deoarece Hessiana este pozitiv definită.

Aplicație financiară - continuare: Observăm că în problema considerată în secțiunea 2.3.2 nu s-a luat în considerare riscul. Riscul este dat de fluctuația ratei de rentabilitate $r_i(t)$ de-a lungul celor T ani. Minimizarea riscului este echivalentă cu minimizarea varianței investiției (*risk averse*). În acest caz, matricea de covarianță Q se exprimă astfel:

$$Q_{ij} = \sigma_{ij}^2 = \frac{1}{T} \sum_{t=1}^T (r_i(t) - c_i)(r_j(t) - c_j) \quad \forall i, j = 1, \dots, n.$$

Problema minimizării riscului poate fi formulată ca o problemă QP:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \frac{1}{2} x^T Q x \\ \text{s.l.:} \quad & x \geq 0, \quad \sum_{i=1}^n c_i x_i \geq R, \quad \sum_{i=1}^n x_i = 1. \end{aligned}$$

Constrângerea $\sum_{i=1}^n c_i x_i \geq R$ se impune pentru a asigura cel puțin un profit R .

În practică avem la dispoziție mai multe produse software pentru rezolvarea de probleme QP: MOSEK, MATLAB (quadprog), SeDuMi, CVX, YALMIP.

2.3.4 Optimizare convexă (CP)

Ambele tipuri de probleme LP și QP aparțin unei clase mai largi de probleme de optimizare, și anume probleme de optimizare convexe. O problemă de optimizare cu o mulțime fezabilă X convexă și o funcție obiectiv f convexă se numește *problemă de optimizare convexă* (CP - *Convex Programming*), i.e.

$$\begin{aligned} (CP) : \quad & \min_{x \in \mathbb{R}^n} f(x) \\ \text{s.l.:} \quad & g(x) \leq 0, \quad Ax - b = 0, \end{aligned} \tag{2.4}$$

unde $f : \mathbb{R}^n \rightarrow \mathbb{R}$ și componentele lui $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ sunt funcții convexe și constrângerile de egalitate sunt descrise de funcții afine $h(x) = Ax - b$, unde $A \in \mathbb{R}^{p \times n}$ și $b \in \mathbb{R}^p$. Rezultă deci că mulțimea fezabilă asociată $X = \{x \in \mathbb{R}^n : g(x) \leq 0, Ax = b\}$ este mulțime convexă.

Exemplul 2.3.2 *Programare pătratică cu constrângeri pătratice (QCQP - Quadratically Constrained Quadratic Program): O problemă de optimizare convexă de forma (2.4) cu funcțiile f și componentele lui g pătratice convexe se numește problemă pătratică cu constrângeri pătratice:*

$$\begin{aligned} (QCQP) : \quad & \min_{x \in \mathbb{R}^n} \frac{1}{2} x^T Q x + q^T x + r \\ \text{s.l.:} \quad & \frac{1}{2} x^T Q_i x + q_i^T x + r_i \leq 0 \quad i = 1, \dots, m \\ & Ax - b = 0. \end{aligned}$$

Alegând $Q_1 = \dots = Q_m = 0$ obținem o problemă uzuală QP , iar dacă în plus alegem $Q = 0$ obținem un LP . De aceea, clasa problemelor $QCQP$ conține și clasa LP -urilor și pe cea a QP -urilor. Dacă matricele Q și Q_i cu $i = 1, \dots, m$ sunt pozitiv semidefinite atunci problema ($QCQP$) este convexă.

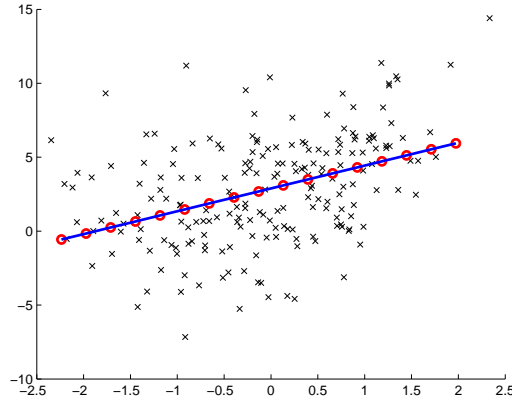


Figura 2.6: Analiza statistică a datelor.

Analiza statistică: Analiza datelor și interpretarea acestora într-un sens cât mai corect este una din preocupările principale din domeniul statisticii. Problema se formulează în următorul mod: pe baza unei colecții de date cunoscute (reprezentate în Fig. 2.6 prin puncte), să se realizeze predicția cu o eroare cât mai mică a unui alt set de date parțial cunoscut. În termeni matematici, această problemă presupune determinarea unei direcții de-a lungul căreia elementele date (punctele) tind să se alinieze, astfel încât să se poată prezice zona de apariție a punctelor viitoare. S-a constatat că direcția de căutare este dată de vectorul singular corespunzător celei mai mici valori singulare al matricei formate din colecția de puncte date, ce poate fi găsit prin intermediul unei probleme de optimizare convexă:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \frac{1}{2} x^T A^T A x \\ \text{s.l.:} \quad & x^T x \leq 1, \end{aligned}$$

unde $A \in \mathbb{R}^{m \times n}$ reprezintă matricea ale cărei coloane sunt vectorii (punctele) cunoscute inițial a_1, \dots, a_n .

Exemplul 2.3.3 *Programare semidefinită (SDP - SemiDefinite Programming)* O clasă importantă de probleme de optimizare convexă folosește inegalități liniare matriceale (LMI) pentru a descrie mulțimea fezabilă. Datorită naturii constrângerilor ce impun ca anumite matrice să rămână pozitiv semidefinite, această clasă de probleme se numește programare semidefinită (SDP). O problemă generală SDP poate fi formulată după cum urmează:

$$\begin{aligned} (\text{SDP}) : \quad & \min_{x \in \mathbb{R}^n} c^T x \\ \text{s.l.:} \quad & A_0 + \sum_{i=1}^n A_i x_i \preceq 0, \quad Ax - b = 0, \end{aligned}$$

unde matricele $A_i \in S^m$ oricare ar fi $i = 0, \dots, n$. Remarcăm că problemele LP, QP, și QCQP pot fi de asemenea formulate ca probleme SDP. Programarea Semidefinită este un instrument des utilizat în teoria sistemelor și reglare (control).

Minimizarea valorii proprii maxime: a unei matrice poate fi formulată ca o problemă SDP. Avem o matrice simetrică $G(x)$ care depinde afin de anumite variabile structurale $x \in \mathbb{R}^n$, i.e. $G(x) = A_0 + \sum_{i=1}^n A_i x_i$ cu $A_i \in S^m$ oricare ar fi $i = 0, \dots, n$. Dacă dorim să minimizăm valoarea proprie maximă a lui $G(x)$ în funcție de x , i.e. să rezolvăm

$$\min_{x \in \mathbb{R}^n} \lambda_{\max}(G(x))$$

putem formula această problemă ca un SDP, după cum urmează: adăugând o variabilă auxiliară $t \in \mathbb{R}$ și ținând cont că $t \geq \lambda_{\max}(G(x))$ este echivalent cu un LMI $tI_m \succcurlyeq G(x)$, obținem:

$$\begin{aligned} \min_{t \in \mathbb{R}, x \in \mathbb{R}^n} \quad & t \\ \text{s.l.:} \quad & tI_m - \sum_{i=1}^n A_i x_i - A_0 \succcurlyeq 0. \end{aligned}$$

Două produse software excelente pentru formularea și rezolvarea problemelor de optimizare convexă în mediul de programare MATLAB sunt YALMIP și CVX, ce se pot găsi open-source și sunt foarte ușor de instalat. Un software comercial des utilizat este MOSEK.

2.3.5 Probleme de optimizare neconstrânsă (UNLP)

Orice problemă NLP fără constrângeri se numește *problemă de optimizare neconstrânsă* (UNLP - *Unconstrained NonLinear Programming*). Are forma generală:

$$(UNLP) : \min_{x \in \mathbb{R}^n} f(x). \quad (2.5)$$

Metodele numerice de optimizare neliniară fără constrângeri vor forma subiectul Părții a II-a a acestei lucrări, în timp ce algoritmi pentru probleme generale constrânse vor fi studiați în Partea a III-a. Cel mai utilizat software pentru programare neconstrânsă este programul Matlab cu funcțiile incluse `fminunc` și `fminsearch`.

Probleme de optimizare nediferențiabilă: dacă una sau mai multe funcții f, g și h din structura problemei NLP (2.1) nu sunt diferenziabile, atunci avem o problemă de optimizare *nediferențiabilă*. Problemele de optimizare nediferențiabilă sunt mult mai greu de rezolvat decât problemele NLP generale. Există un număr mai redus de algoritmi pentru a rezolva astfel de probleme: metoda subgradient, metoda Nelder-Mead, căutare aleatorie, algoritmi genetici, etc. De obicei, acești algoritmi sunt mult mai slabi din punct de vedere numeric decât algoritmi bazați pe informație de tip gradient și Hessiană, și care sunt subiectul acestei lucrări.

2.3.6 Programare mixtă cu întregi (MIP)

O problemă de programare mixtă cu întregi este o problemă în care anumite variabile de decizie sunt constrânse la o mulțime de numere întregi. Un MIP poate fi formulat după cum urmează:

$$(MIP) : \min_{x \in \mathbb{R}^n, z \in \mathbb{Z}^m} f(x, z) \\ \text{s.l.: } g(x, z) \leq 0, \quad h(x, z) = 0.$$

În general, aceste probleme sunt foarte greu de rezolvat, datorită naturii combinatoriale a variabilei z . Cu toate acestea, dacă *problema relaxată*, unde variabilele z nu mai sunt restrânse la întregi, ci la mulțimi de numere reale, este convexă, de regulă există algoritmi eficienți pentru rezolvarea lor. Algoritmi eficienți de găsire a soluției sunt adesea bazați pe tehnica *branch-and-bound*, care folosește probleme parțial relaxate unde unele

variabile din z sunt fixate la anumite valori întregi și altele sunt relaxate exploatând proprietatea că soluția problemelor relaxate este întotdeauna mai bună decât orice soluție cu componente întregi. În acest fel, căutarea poate avea loc mult mai eficient decât o pură verificare a elementelor mulțimii fezabile. Două exemple importante de asemenea probleme sunt date în cele ce urmează:

(i) **Program liniar mixt cu întregi** (*MILP - Mixed Integer Linear Programming*): dacă funcțiile f , g și h sunt afine în ambele variabile x și z obținem un program liniar mixt cu întregi. O problemă faimoasă din această clasă este *problema comis-voiajorului*.

(ii) **Program pătratic mixt cu întregi** (*MIQP - Mixed Integer Quadratic Programming*): dacă g și h sunt funcții afine și f pătratică convexă în ambele variabile x și z rezultă un program pătratic mixt cu întregi (MIQP).

Probleme (MILP)/(MIQP) de dimensiuni mici/medii (i.e. dimensiunea variabilei $n, m < 100$) pot fi rezolvate eficient de pachete de software comerciale CPLEX, TOMLAB sau `lp_solve`.

Partea II

Optimizare fără constrângeri

Capitolul 3

Metode de optimizare unidimensională

După cum vom vedea în capitolele următoare, metodele bazate pe direcții de descreștere presupun găsirea unui pas care, ideal, trebuie ales optim. Astfel de metode se mai numesc și metode de căutare exactă. În această situație, trebuie să calculăm parametrul optim α^* ce determină valoarea minimă a funcției obiectiv f în direcția d , cu alte cuvinte minimizarea funcției $\phi(\alpha) = f(x + \alpha d)$. Din acest motiv, în acest capitol analizăm metode numerice de optimizare unidimensională, adică pentru funcții de o singură variabilă $f : \mathbb{R} \rightarrow \mathbb{R}$:

$$\min_{\alpha \in \mathbb{R}} f(\alpha). \quad (3.1)$$

Metodele de optimizare unidimensională se bazează fie pe căutare directă fie pe aproximarea funcției f cu un polinom ce se determină prin interpolare folosind valorile funcției și/sau derivatele funcției obiectiv în anumite puncte. În metodele de căutare principiul de bază este următorul: se identifică intervalul $[a, b] \subset \mathbb{R}$ ce include punctul de minim α^* , numit și intervalul de căutare sau intervalul de incertitudine, urmat apoi de o reducere iterativă a lungimii acestuia până la o valoare ce coboară sub toleranța impusă pentru a localiza α^* . Eficiența acestei abordări depinde de strategia de construcție a șirului de intervale $[a_k, b_k]$, $k = 1, 2, \dots$, ce îl conțin pe α^* . Cele mai renumite metode de căutare unidimensională sunt: metoda secțiunii de aur și metoda lui Fibonacci, pe care le vom prezenta în acest capitol. Metoda clasică Newton-Raphson și metoda secantei sunt de asemenea considerate membre ale aceleiași clase. Pe de altă parte, metodele de interpolare găsesc o aproximare a lui

α^* folosind valori ale funcției obiectiv $f(\alpha)$ în puncte din intervalul inițial de căutare $[a, b]$, sau pot folosi valoarea funcției obiectiv și derivata sa $f'(\alpha)$ în anumite puncte din $[a, b]$. Prin intermediul acestor valori se formează polinomul de interpolare de gradul doi sau mai mare, $q(\alpha)$, al funcției $f(\alpha)$ și este determinat punctul de minim $\hat{\alpha}$ al funcției $q(\alpha)$. Printre cele mai renumite metode de interpolare se numără cea pătratică (în două sau trei puncte) și cea cubică.

3.1 Metoda forward-backward pentru funcții unimodale

O metodă simplă de determinare a unui interval inițial de căutare, adică determinarea unui interval care conține punctul de optim α^* , este dată de metoda forward-backward. Ideea de bază este următoarea: dându-se un punct inițial și o lungime a pasului, se încearcă determinarea a două puncte pentru care funcția are o formă geometrică convexă pe acel interval cu capetele în cele două puncte. Metoda presupune următorii pași: fie un punct inițial α_0 și lungimea pasului $h_0 > 0$

- dacă $f(\alpha_0 + h_0) < f(\alpha_0)$, atunci se începe din punctul $\alpha_0 + h_0$ și se continuă cu o lungime a pasului mai mare cât timp valoarea funcției crește;

- dacă $f(\alpha_0 + h_0) > f(\alpha_0)$, atunci ne deplasăm din α_0 înapoi până când valoarea funcției crește.

În acest fel vom obține un interval inițial ce conține valoarea optimă α^* . Metoda forward-backward se bazează pe proprietățile de unimodalitate ale funcțiilor.

Definiția 3.1.1 Fie funcția $f: \mathbb{R} \rightarrow \mathbb{R}$ și un interval $[a, b] \subset \mathbb{R}$. Dacă există $\alpha^* \in [a, b]$ astfel încât f este strict descrescătoare pe intervalul $[a, \alpha^*]$ și strict crescătoare pe intervalul $[\alpha^*, b]$, atunci f se numește funcție unimodală pe intervalul $[a, b]$. Intervalul $[a, b]$ se numește interval de unimodalitate pentru f .

Se observă imediat că funcțiile unimodale nu implică continuitate și diferențiabilitate. Următoarea teoremă arată că dacă f este unimodală, atunci intervalul de incertitudine poate fi redus comparând valorile lui f în doar două puncte ale intervalului.

Teorema 3.1.1 Fie funcția unimodală $f: \mathbb{R} \rightarrow \mathbb{R}$ pe intervalul $[a, b]$ și $\alpha_1, \alpha_2 \in [a, b]$ cu $\alpha_1 < \alpha_2$. În acest caz:

- (i) dacă $f(\alpha_1) \leq f(\alpha_2)$, atunci $[a, \alpha_2]$ este interval de unimodalitate pentru f .
(ii) dacă $f(\alpha_1) \geq f(\alpha_2)$, atunci $[\alpha_1, b]$ este interval de unimodalitate pentru f .

Pentru o expunere mai ușoară presupunem că punctul de minim se găsește în \mathbb{R}_+ . Metoda forward-backward presupune următorii pași:

Pas 1. Fie un $\alpha_0 \in [0, \infty)$, $h_0 > 0$ și coeficientul multiplicativ $t > 1$ (adesea se alege $t = 2$). Evaluăm $f(\alpha_0) = f_0$ și $k = 0$.

Pas 2. Comparăm valorile funcției obiectiv. Actualizăm $\alpha_{k+1} = \alpha_k + h_k$ și evaluăm $f_{k+1} = f(\alpha_{k+1})$. Dacă $f_{k+1} < f_k$, sărim la Pas 3; altfel, sărim la Pas 4.

Pas 3. Pas forward. Actualizăm $h_{k+1} = th_k$, $\alpha = \alpha_k$, $\alpha_k = \alpha_{k+1}$, $f_k = f_{k+1}$ și $k = k + 1$, sărim la Pas 2.

Pas 4. Pas backward. Dacă $k = 0$, inversăm direcția de căutare. Luăm $h_k = -h_k$, $\alpha_k = \alpha_{k+1}$, sărim la Pas 2; altfel, considerăm

$$a = \min \{ \alpha, \alpha_{k+1} \}, \quad b = \max \{ \alpha, \alpha_{k+1} \},$$

returnăm intervalul $[a, b]$ ce conține punctul de minim α^* și ne oprim.

3.2 Metode de căutare

Metoda secțiunii de aur și metoda lui Fibonacci sunt metode de tip partiționare. Ideea din spatele acestor metode de minimizare a funcțiilor unimodale pe intervalul $[a, b] \subset \mathbb{R}$ constă în reducerea iterativă a intervalului de incertitudine doar prin compararea valorilor luate de funcția obiectiv. Odată ce lungimea intervalului de incertitudine este mai mică decât o acuratețe prestabilită, atunci punctele din acest interval pot fi considerate aproximări ale valorii minime a funcției. Această clasă de metode folosește doar valoarea funcției obiectiv și are un rol important în algoritmi de optimizare, în special când ne confruntăm cu funcții obiectiv nediferențiabile sau funcții obiectiv ale căror derivate prezintă forme complicate.

3.2.1 Metoda secțiunii de aur

Considerăm funcția f unimodală pe intervalul $[a, b]$ și definim $a_1 = a$ și $b_1 = b$. La iterația k , metoda secțiunii de aur determină intervalul $[a_{k+1}, b_{k+1}]$ astfel încât $\alpha^* \in [a_{k+1}, b_{k+1}]$. În acest moment considerăm două puncte $\lambda_k, \mu_k \in [a_k, b_k]$ unde $\lambda_k < \mu_k$ și calculăm $f(\lambda_k)$ și $f(\mu_k)$ (vezi Fig. 5.2). Din teorema precedentă rezultă:

- (i) Dacă $f(\lambda_k) \leq f(\mu_k)$ atunci $a_{k+1} = a_k$ și $b_{k+1} = \mu_k$.
- (ii) Dacă $f(\lambda_k) > f(\mu_k)$ atunci $a_{k+1} = \lambda_k$ și $b_{k+1} = b_k$.

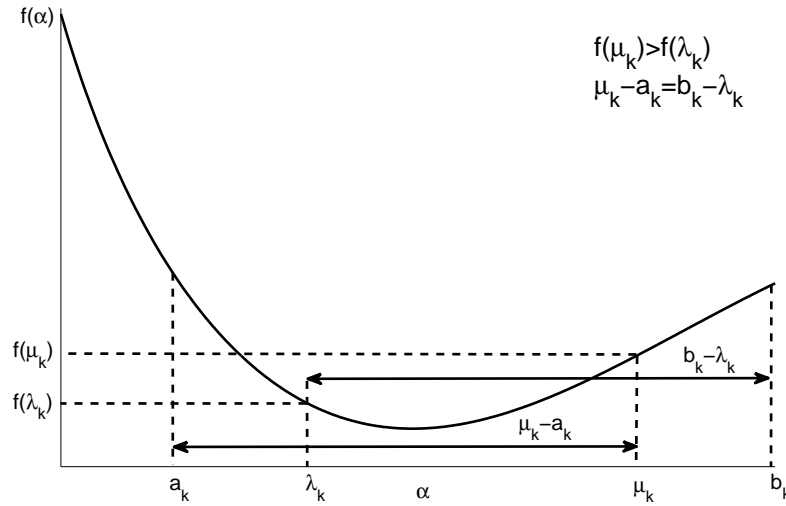


Figura 3.1: Exemplu de pas pentru metoda secțiunii de aur

Rămâne să discutăm alegerea punctelor λ_k și μ_k . În acest scop impunem următoarele trei condiții:

1. distanțele de la λ_k și respectiv μ_k la capetele intervalului $[a_k, b_k]$ sunt egale:

$$b_k - \lambda_k = \mu_k - a_k. \quad (3.2)$$

2. rata de micșorare a lungimii intervalului de incertitudine la fiecare iterație este aceeași, rezultând

$$b_{k+1} - a_{k+1} = \tau(b_k - a_k), \quad \text{unde } \tau \in (0, 1). \quad (3.3)$$

3. este necesară o singură evaluare a funcției obiectiv pentru o nouă iterație.

Dacă substituim valorile ce constituie cazul (i) în (3.3) obținem $\mu_k - a_k = \tau(b_k - a_k)$ și prin combinarea cu (3.2) avem $b_k - \lambda_k = \mu_k - a_k$. Prin rearanjarea acestor egalități avem:

$$\lambda_k = a_k + (1 - \tau)(b_k - a_k) \quad (3.4)$$

$$\mu_k = a_k + \tau(b_k - a_k). \quad (3.5)$$

În acest caz, noul interval este $[a_{k+1}, b_{k+1}] = [a_k, \mu_k]$. Pentru a reduce intervalul de incertitudine este necesară selecția parametrilor λ_{k+1} și μ_{k+1} . Din (3.5) rezultă

$$\begin{aligned} \mu_{k+1} &= a_{k+1} + \tau(b_{k+1} - a_{k+1}) = a_k + \tau(\mu_k - a_k) \\ &= a_k + \tau(a_k + \tau(b_k - a_k) - a_k) = a_k + \tau^2(b_k - a_k). \end{aligned} \quad (3.6)$$

Considerând

$$\tau^2 = 1 - \tau \quad (3.7)$$

rezultă

$$\mu_{k+1} = a_k + (1 - \tau)(b_k - a_k) = \lambda_k.$$

Astfel, μ_{k+1} coincide cu λ_k și funcția obiectiv nu necesită o evaluare deoarece valoarea sa este stocată în λ_k . Cazul (ii) poate fi demonstrat într-o manieră similară, din care rezultă $\lambda_{k+1} = \mu_k$ astfel încât nu este necesară evaluarea funcției obiectiv. Metoda secțiunii de aur constă în următorii pași:

Pas 1. Determinăm intervalul inițial $[a_1, b_1]$ și alegem precizia $\delta > 0$. Calculăm primele două puncte λ_1 și μ_1 :

$$\begin{aligned} \lambda_1 &= a_1 + 0.382(b_1 - a_1) \\ \mu_1 &= a_1 + 0.618(b_1 - a_1) \end{aligned}$$

și evaluăm $f(\lambda_1)$ și $f(\mu_1)$, inițializăm $k = 1$.

Pas 2. Comparăm valorile funcțiilor. Dacă $f(\lambda_k) > f(\mu_k)$, trecem la Pas 3; dacă $f(\lambda_k) \leq f(\mu_k)$, trecem la Pas 4.

Pas 3. Dacă $b_k - \lambda_k \leq \delta$, ne oprim și returnăm μ_k ; altfel iterăm:

$$\begin{aligned} a_{k+1} &= \lambda_k, \quad b_{k+1} = b_k, \quad \lambda_{k+1} = \mu_k \\ f(\lambda_{k+1}) &= f(\mu_k), \quad \mu_{k+1} = a_{k+1} + 0.618(b_{k+1} - a_{k+1}). \end{aligned}$$

Evaluăm $f(\mu_{k+1})$ și trecem la Pas 5.

Pas 4. Dacă $\mu_k - a_k \leq \delta$, ne oprim și returnăm λ_k ; altfel iterăm:

$$\begin{aligned} a_{k+1} &= a_k, \quad b_{k+1} = \mu_k, \quad \mu_{k+1} = \lambda_k, \\ f(\mu_{k+1}) &= f(\lambda_k), \quad \lambda_{k+1} = a_{k+1} + 0.382(b_{k+1} - a_{k+1}). \end{aligned}$$

Evaluăm $f(\lambda_{k+1})$ și trecem la Pas 5.

Pas 5. Iterăm $k = k + 1$, revenim la Pas 2.

Observăm că acest algoritm produce un șir de intervale $[a_k, b_k]$ astfel încât punctul de minim α^* al funcției f se află în fiecare din aceste intervale. Mai departe ne concentrăm spre analiza ratei de reducere a intervalului de incertitudine. Rezolvând ecuația (3.7) obținem:

$$\tau = \frac{-1 \pm \sqrt{5}}{2}.$$

Deoarece $\tau > 0$ considerăm

$$\tau = \frac{b_{k+1} - a_{k+1}}{b_k - a_k} = \frac{\sqrt{5} - 1}{2} \cong 0.618$$

Înlocuind valoarea lui τ în (3.4) și (3.5) avem

$$\begin{aligned} \lambda_k &= a_k + 0.382(b_k - a_k) \\ \mu_k &= a_k + 0.618(b_k - a_k). \end{aligned}$$

Deoarece rata de reducere este fixă la fiecare iterație, $\tau = 0.618$, considerând un interval inițial $[a_1, b_1]$, după k iterații lungimea intervalului este $\tau^{k-1}(b_1 - a_1)$, ceea ce arată că rata de convergență a metodei secțiunii de aur este liniară.

3.2.2 Metoda lui Fibonacci

În metoda lui Fibonacci principala diferență față de metoda secțiunii de aur constă în definiția legii de reducere a intervalului de incertitudine în acord cu șirul lui Fibonacci. Cu alte cuvinte, rata de reducere nu este fixă în această metodă, ci variază de la un interval la altul. Șirul lui Fibonacci F_k este definit de următoarea lege:

$$\begin{aligned} F_0 &= F_1 = 1 \\ F_{k+1} &= F_k + F_{k-1} \quad \forall k = 1, 2, \dots \end{aligned}$$

Dacă în (3.4) și (3.5) înlocuim τ cu $\frac{F_{k-j}}{F_{k-j+1}}$ atunci

$$\begin{aligned}\lambda_j &= a_j + \left(1 - \frac{F_{k-j}}{F_{k-j+1}}\right)(b_j - a_j) = a_j + \frac{F_{k-j-1}}{F_{k-j+1}} \quad \forall j = 1, \dots, k-1 \\ \mu_j &= a_j + \frac{F_{k-j}}{F_{k-j+1}}(b_j - a_j) \quad \forall j = 1, \dots, k-1.\end{aligned}\quad (3.8)$$

Dacă $f(\lambda_j) \leq f(\mu_j)$ atunci noul interval de incertitudine este dat de $[a_{j+1}, b_{j+1}] = [a_j, \mu_j]$. Astfel, prin (3.8) obținem:

$$b_{j+1} - a_{j+1} = \frac{F_{k-j}}{F_{k-j+1}}(b_j - a_j),$$

ceea ce arată reducția la fiecare iterație. Poate fi ușor de observat că această ecuație este de asemenea valabilă pentru $f(\lambda_j) > f(\mu_j)$. Mai departe, impunem ca lungimea intervalului final de incertitudine să nu depășească o toleranță dată $\delta > 0$, adică $b_k - a_k \leq \delta$. Luând în considerare:

$$b_k - a_k = \frac{F_1}{F_2}(b_{k-1} - a_{k-1}) = \frac{F_1}{F_2} \frac{F_2}{F_3} \dots \frac{F_{k-1}}{F_k}(b_1 - a_1) = \frac{1}{F_k}(b_1 - a_1),$$

avem

$$F_k \geq \frac{b_1 - a_1}{\delta}. \quad (3.9)$$

De aceea, având intervalul inițial $[a_1, b_1]$ și marginea superioară δ putem calcula numărul Fibonacci F_k și valoarea k din (3.9). Căutarea are loc până la iterația k . O observație importantă în legătură cu ratele de convergență ale metodelor studiate este că odată ce $k \rightarrow \infty$ metoda Fibonacci și metoda secțiunii de aur au aceeași rată de reducere a intervalului de incertitudine. Considerând $F_k = r^k$, atunci din definiția șirului Fibonacci avem $r^2 - r + 1 = 0$ cu rădăcinile:

$$r_1 = \frac{1 + \sqrt{5}}{2}, r_2 = \frac{1 - \sqrt{5}}{2}.$$

Soluția generală a ecuației $F_{k+1} = F_k + F_{k-1}$ este $F_k = Ar_1^k + Br_2^k$. De aceea, din condițiile inițiale $F_0 = F_1 = 1$ avem $A = 1/\sqrt{5}$, $B = -1/\sqrt{5}$ și

$$F_k = \frac{1}{\sqrt{5}} \left\{ \left(\frac{1 + \sqrt{5}}{2} \right)^k - \left(\frac{1 - \sqrt{5}}{2} \right)^k \right\}.$$

Cu aceste relații deducem că:

$$\lim_{k \rightarrow \infty} \frac{F_{k-1}}{F_k} = \frac{\sqrt{5} - 1}{2} = \tau. \quad (3.10)$$

De aceea, ambele metode împărtășesc aceeași rată de convergență când $k \rightarrow \infty$, însă metoda lui Fibonacci este optimă în clasa metodelor de căutare.

3.3 Metode de interpolare

Metodele de interpolare pentru minimizare unidimensională sunt o alternativă foarte eficientă la metoda secțiunii de aur și cea a lui Fibonacci. Cea mai importantă din această clasă de metode aproximează funcția f cu un polinom de ordin doi sau trei, ce are valori identice cu derivatele funcției în anumite puncte și în final, calculează valoarea α ce minimizează polinomul. În cazul general în care funcția obiectiv prezintă proprietăți analitice *bune*, cum ar fi diferențiabilitatea continuă, atunci metodele de interpolare sunt cu mult superioare metodei secțiunii de aur și celei a lui Fibonacci.

3.3.1 Metode de interpolare pătratică

Metoda de interpolare în două puncte (prima variantă): Fie două puncte α_1 și α_2 . Presupunem cunoscute valorile funcției f în punctele corespunzătoare $f(\alpha_1)$ și $f(\alpha_2)$ și derivatele de ordinul I în aceleași puncte: $f'(\alpha_1)$ și $f'(\alpha_2)$. Construim polinomul de interpolare de ordinul II:

$$q(\alpha) = a\alpha^2 + b\alpha + c,$$

ce satisface următoarele condiții:

$$\begin{aligned} q(\alpha_1) &= a\alpha_1^2 + b\alpha_1 + c = f(\alpha_1) \\ q(\alpha_2) &= a\alpha_2^2 + b\alpha_2 + c = f(\alpha_2) \\ q'(\alpha_1) &= 2a\alpha_1 + b = f'(\alpha_1). \end{aligned} \quad (3.11)$$

Dacă notăm $f_1 = f(\alpha_1)$, $f_2 = f(\alpha_2)$, $f'_1 = f'(\alpha_1)$ și $f'_2 = f'(\alpha_2)$, atunci din (3.11) avem:

$$a = \frac{f_1 - f_2 - f'_1(\alpha_1 - \alpha_2)}{-(\alpha_1 - \alpha_2)^2}$$

$$b = f'_1 + 2\alpha_1 \frac{f_1 - f_2 - f'_1(\alpha_1 - \alpha_2)}{(\alpha_1 - \alpha_2)^2}.$$

Mai departe, rezultă că punctul de minim al polinomului de interpolare este:

$$\begin{aligned} \bar{\alpha} &= -\frac{b}{2a} = \alpha_1 + \frac{1}{2} \frac{f'_1(\alpha_1 - \alpha_2)^2}{\alpha_1 - \alpha_2 - f'_1(\alpha_1 - \alpha_2)} \\ &= \alpha_1 - \frac{1}{2} \frac{f'_1(\alpha_1 - \alpha_2)}{f'_1 - \frac{f_1 - f_2}{\alpha_1 - \alpha_2}}. \end{aligned} \quad (3.12)$$

Și în final, obținem formula de interpolare pătratică:

$$\alpha_{k+1} = \alpha_k - \frac{1}{2} \frac{f'_k(\alpha_k - \alpha_{k-1})}{f'_k - \frac{f_k - f_{k-1}}{\alpha_k - \alpha_{k-1}}}, \quad (3.13)$$

unde $f_k = f(\alpha_k)$, $f_{k-1} = f(\alpha_{k-1})$, $f'_k = f'(\alpha_k)$. Algoritmul este foarte simplu: cât timp α_{k+1} este determinat, se compară cu α_k și α_{k-1} , rezultând o reducere a lungimii intervalului de incertitudine. Acest proces se repetă până când lungimea intervalului scade sub un anumit prag.

Metoda de interpolare în două puncte (a doua variantă): Fie două puncte α_1 și α_2 , evaluările funcției f în punctele corespunzătoare $f(\alpha_1)$ și $f(\alpha_2)$ și derivatele de ordinul I în aceleași puncte $f'(\alpha_1)$ și $f'(\alpha_2)$. Construim polinomul de interpolare

$$q(\alpha) = a\alpha^2 + b\alpha + c,$$

ce satisface următoarele condiții:

$$\begin{aligned} q(\alpha_1) &= a\alpha_1^2 + b\alpha_1 + c = f(\alpha_1) \\ q'(\alpha_1) &= 2a\alpha_1 + b = f'(\alpha_1) \\ q'(\alpha_2) &= 2a\alpha_2 + b = f'(\alpha_2). \end{aligned} \quad (3.14)$$

De aici se obține:

$$\bar{\alpha} = -\frac{b}{2a} = \alpha_1 - \frac{1}{2} \frac{\alpha_1 - \alpha_2}{f'_1 - f'_2} f'_1 \quad (3.15)$$

și formula iterativă:

$$\alpha_{k+1} = \alpha_k - \frac{1}{2} \frac{\alpha_k - \alpha_{k-1}}{f'_k - f'_{k-1}} f'_k. \quad (3.16)$$

Teorema următoare ilustrează rata de convergență foarte rapidă a acestei metode:

Teorema 3.3.1 *Dacă $f: \mathbb{R} \rightarrow \mathbb{R}$ este de trei ori continuu diferențiabilă și există un α^* astfel încât $f'(\alpha^*) = 0$ și $f''(\alpha^*) \neq 0$, atunci șirul α_k generat de iterația (3.16) converge la α^* cu rata $(1 + \sqrt{5})/2 \cong 1.618$, adică $\lim_{k \rightarrow \infty} \frac{|\alpha_{k+1} - \alpha^*|}{|\alpha_k - \alpha^*|^{(1+\sqrt{5})/2}} = \rho$ cu $\rho > 0$.*

Demonstrație: Relația (3.15) poate fi scrisă și prin intermediul formulei de interpolare Lagrange:

$$L(\alpha) = \frac{(\alpha - \alpha_1)f'_2 - (\alpha - \alpha_2)f'_1}{\alpha_2 - \alpha_1},$$

dacă luăm $L(\alpha) = 0$. Acum, termenul rezidual al formulei de interpolare Lagrange se consideră ca fiind:

$$f'(\alpha) - L(\alpha) = \frac{1}{2} f'''(\xi)(\alpha - \alpha_k)(\alpha - \alpha_{k-1}) \quad \text{cu } \xi \in \{\alpha, \alpha_{k-1}, \alpha_k\}.$$

Dacă luăm $\alpha = \alpha_{k+1}$ și observând că $L(\alpha_{k+1}) = 0$, avem:

$$f'(\alpha_{k+1}) = \frac{1}{2}(\alpha_{k+1} - \alpha_k)(\alpha_{k+1} - \alpha_{k-1}) \quad \text{cu } \xi \in \{\alpha_{k-1}, \alpha_k, \alpha_{k+1}\}, \quad (3.17)$$

Înlocuind (3.16) în (3.17) avem:

$$f'(\alpha_{k+1}) = \frac{1}{2} f'''(\xi) f'_k f'_{k-1} \frac{(\alpha_k - \alpha_{k-1})^2}{(f'_k - f'_{k-1})^2}.$$

Din teorema valorii medii știm că:

$$\frac{(f'_k - f'_{k-1})}{\alpha_k - \alpha_{k-1}} = f''(\xi_0) \quad \text{cu } \xi_0 \in [\alpha_{k-1}, \alpha_k],$$

iar drept urmare:

$$f'_i = f'_i - f'(\alpha^*) = (\alpha_i - \alpha^*) f''(\xi_i), \quad \text{unde } \xi_i \in [\alpha_i, \alpha^*], i = k-1, k, k+1.$$

Astfel, din ultimele trei ecuații rezultă:

$$\alpha_{k+1} - \alpha^* = \frac{1}{2} \frac{f'''(\xi)f''(\xi_k)f''(\xi_{k-1})}{f''(\xi_{k+1})f''(\xi_0)^2} (\alpha_k - \alpha^*)(\alpha_{k-1} - \alpha^*).$$

Dacă notăm distanțele de la α_i la punctul optim α^* prin $e_i = |\alpha_i - \alpha^*|$, cu $i = k-1, k, k+1$ și considerăm valorile m_1, M_1, m_2, M_2 și K_1, K astfel încât

$$0 < m_2 \leq |f'''(\alpha)| \leq M_2, \quad 0 < m_1 \leq |f''(\alpha)| \leq M_1 \\ K_1 = m_2 m_1^2 / (2M_1^3), \quad K = M_2 M_1^2 / (2m_1^3).$$

Atunci:

$$K_1 e_k e_{k-1} \leq e_{k+1} \leq K e_k e_{k-1}.$$

Observând că $f''(\alpha)$ și $f'''(\alpha)$ sunt continue în α^* , avem:

$$\frac{\alpha_{k+1} - \alpha^*}{(\alpha_k - \alpha^*)(\alpha_{k-1} - \alpha^*)} \rightarrow \frac{1}{2} \frac{f'''(\alpha^*)}{f''(\alpha^*)}$$

și obținem următoarea relație între distanțele până la punctul de optim:

$$e_{k+1} = M e_k e_{k-1},$$

unde $M = |f'''(\eta_1)/2f''(\eta_2)|$, iar $\eta_1 \in \{\alpha_{k-1}, \alpha_k, \alpha^*\}$ și $\eta_2 \in \{\alpha_{k-1}, \alpha_k\}$. Dacă există o precizie $\delta > 0$ astfel încât punctele inițiale $\alpha_0, \alpha_1 \in (\alpha^* - \delta, \alpha^* + \delta)$ și $\alpha_0 \neq \alpha_1$, atunci se poate observa din relațiile anterioare că șirul α_k va converge către α^* . Am demonstrat că α_k converge la α^* și ne mai rămâne să demonstrăm rata de convergență. În acest scop notăm $\epsilon_i = M e_i$, $y_i = \ln(\epsilon_i)$, $i = k-1, k, k+1$, iar conform relațiilor anterioare avem:

$$\epsilon_{k+1} = \epsilon_k \epsilon_{k-1}, \quad y_{k+1} = y_k + y_{k-1}. \quad (3.18)$$

Este evident că ecuația (3.18) reprezintă o secvență Fibonacci. Ecuația caracteristică a secvenței Fibonacci și rădăcinile aferente sunt:

$$r^2 - r - 1 = 0, \quad r_1 = (1 + \sqrt{5})/2, \quad r_2 = (1 - \sqrt{5})/2,$$

Astfel, secvența Fibonacci y_k poate fi scrisă ca:

$$y_k = A r_1^k + B r_2^k \quad k = 0, 1, \dots,$$

unde A și B sunt coeficienți ce pot fi determinați. Din moment ce progresăm cu algoritmul și $k \rightarrow \infty$ atunci evident $y_k = \ln(\epsilon_k) \approx At_1^k$, de unde rezultă că:

$$\frac{\epsilon_{k+1}}{\epsilon_k^{t_1}} \approx \frac{e^{(At_1^{k+1})}}{(e^{(At_1^k)})^{t_1}} = 1$$

și $e_{k+1}/e_k^{t_1} \approx M^{t_1-1}$, i.e:

$$\lim_{k \rightarrow \infty} \frac{|\alpha_{k+1} - \alpha^*|}{|\alpha_k - \alpha^*|^{t_1}} = M^{t_1-1}$$

și demonstrația este completă. \square

Metoda de interpolare în trei puncte: Această metodă presupune cunoașterea a trei puncte α_i , $i = 1, 2, 3$ și de asemenea, evaluarea funcției f în aceste puncte. Condițiile de interpolare sunt:

$$q(\alpha_i) = a\alpha_i^2 + b\alpha_i + c = f(\alpha_i) = f_i \quad i = 1, 2, 3. \quad (3.19)$$

Rezolvând acest sistem avem:

$$a = -\frac{(\alpha_2 - \alpha_3)f_1 + (\alpha_3 - \alpha_1)f_2 + (\alpha_1 - \alpha_2)f_3}{(\alpha_1 - \alpha_2)(\alpha_2 - \alpha_3)(\alpha_3 - \alpha_1)}$$

$$b = -\frac{(\alpha_2^2 - \alpha_3^2)f_1 + (\alpha_3^2 - \alpha_1^2)f_2 + (\alpha_1^2 - \alpha_2^2)f_3}{(\alpha_1 - \alpha_2)(\alpha_2 - \alpha_3)(\alpha_3 - \alpha_1)}.$$

De aici rezultă:

$$\bar{\alpha} = -\frac{b}{2a}$$

$$= \frac{1}{2} \frac{(\alpha_2^2 - \alpha_3^2)f_1 + (\alpha_3^2 - \alpha_1^2)f_2 + (\alpha_1^2 - \alpha_2^2)f_3}{(\alpha_2 - \alpha_3)f_1 + (\alpha_3 - \alpha_1)f_2 + (\alpha_1 - \alpha_2)f_3}. \quad (3.20)$$

$$= \frac{1}{2}(\alpha_1 + \alpha_2) + \frac{1}{2} \frac{(f_1 - f_2)(\alpha_2 - \alpha_3)(\alpha_3 - \alpha_1)}{(\alpha_2 - \alpha_3)f_1 + (\alpha_3 - \alpha_1)f_2 + (\alpha_1 - \alpha_2)f_3} \quad (3.21)$$

și formula iterativă:

$$\alpha_{k+1} = \frac{1}{2}(\alpha_k + \alpha_{k-1}) + \frac{1}{2} \frac{(f_k - f_{k-1})(\alpha_{k-1} - \alpha_{k-2})(\alpha_{k-2} - \alpha_k)}{(\alpha_{k-1} - \alpha_{k-2})f_k + (\alpha_{k-2} - \alpha_k)f_{k-1} + (\alpha_k - \alpha_{k-1})f_{k-2}}. \quad (3.22)$$

Metoda interpolării în trei puncte are următorii pași:

Pas 1. Fie o toleranță dată ϵ . Găsește un interval de căutare $\{\alpha_1, \alpha_2, \alpha_3\}$ astfel încât să-l conțină pe α^* ; Calculează $f(\alpha_i)$, $i = 1, 2, 3$.

Pas 2. Utilizează formula (3.20) pentru a calcula $\bar{\alpha}$.

Pas 3. Dacă $(\bar{\alpha} - \alpha_1)(\bar{\alpha} - \alpha_3) \geq 0$, trecem la Pas 4.; altfel, trecem la Pas 5.

Pas 4. Construim un nou interval de căutare $\{\alpha_1, \alpha_2, \alpha_3\}$ utilizând $\alpha_1, \alpha_2, \alpha_3$ și $\bar{\alpha}$. Revenim la Pas 2.

Pas 5. Dacă $|\bar{\alpha} - \alpha_2| < \epsilon$, ne oprim; altfel, trecem la Pas 4.

Teorema 3.3.2 *Fie o funcție f care este de cel puțin patru ori continuu diferențiabilă și α^* astfel încât $f(\alpha^*) = 0$ și $f''(\alpha^*) \neq 0$. Atunci șirul $\{\alpha_k\}$ generat de (3.22) converge la α^* cu rata de ordinul 1.32.*

Observăm că metoda celor trei puncte are o rată de convergență mai mică decât cea a metodelor ce folosesc formula secantei. Explicația constă în faptul că metoda celor trei puncte nu folosește informație dată de derivatele funcției f în punctele intervalului de căutare. Cu alte cuvinte, metoda nu ține cont de curbura funcției f . În general, implementările avansate folosesc informație de secantă.

3.3.2 Metode de interpolare cubică

Aceste metode aproximează funcția obiectiv $f(\alpha)$ cu un polinom cubic. Procedura de aproximare implică patru condiții de interpolare. În cazul general, interpolarea cubică are o rată de convergență mai bună decât interpolarea pătratică, însă presupune evaluarea derivatelor funcției și de aceea este mai costisitoare din punctul de vedere al complexității. Mai pe larg, fie două puncte α_1 și α_2 pentru care cunoaștem valorile funcției obiectiv, $f(\alpha_1)$ și $f(\alpha_2)$ și de asemenea, derivatele $f'(\alpha_1)$ și $f'(\alpha_2)$. Construim polinomul de interpolare cubică:

$$p(\alpha) = c_1(\alpha - \alpha_1)^3 + c_2(\alpha - \alpha_1)^2 + c_3(\alpha - \alpha_1) + c_4, \quad (3.23)$$

unde $c_i, i = 1, 2, 3, 4$, sunt coeficienții determinați din următoarele condiții:

$$\begin{aligned} p(\alpha_1) &= c_4 = f(\alpha_1) \\ p(\alpha_2) &= c_1(\alpha_2 - \alpha_1)^3 + c_2(\alpha_2 - \alpha_1)^2 + c_3(\alpha_2 - \alpha_1) + c_4 = f(\alpha_2) \\ p'(\alpha_1) &= c_3 = f'(\alpha_1) \\ p'(\alpha_2) &= 3c_1(\alpha_2 - \alpha_1)^2 + 2c_2(\alpha_2 - \alpha_1) + c_3 = f'(\alpha_2). \end{aligned}$$

După cum știm, condițiile de optimalitate suficiente sunt:

$$p'(\alpha) = 3c_1(\alpha - \alpha_1)^2 + 2c_2(\alpha - \alpha_1) + c_3 = 0 \quad (3.24)$$

și

$$p''(\alpha) = 6c_1(\alpha - \alpha_1) + 2c_2 > 0. \quad (3.25)$$

Rezolvând (3.24) avem:

$$\alpha = \alpha_1 + \frac{-c_2 \pm \sqrt{c_2^2 - 3c_1c_3}}{3c_1}, \text{ dacă } c_1 \neq 0 \quad (3.26)$$

$$\alpha = \alpha_1 - \frac{c_3}{2c_2}, \text{ dacă } c_1 = 0. \quad (3.27)$$

Pentru a satisface (3.25) considerăm rădăcina corespunzătoare semnelui + din (3.26), care împreună cu (3.27) conduce la:

$$\alpha - \alpha_1 = \frac{-c_2 + \sqrt{c_2^2 - 3c_1c_3}}{3c_1} = \frac{-c_3}{c_2 + \sqrt{c_2^2 - 3c_1c_3}}. \quad (3.28)$$

În cazul în care $c_1 = 0$, (3.28) se transformă în (3.27). Atunci valoarea minimă a lui $p(\alpha)$ este:

$$\bar{\alpha} = \alpha_1 - \frac{c_3}{c_2 + \sqrt{c_2^2 - 3c_1c_3}}, \quad (3.29)$$

exprimată în funcție de c_1, c_2 și c_3 . Problema se reduce la exprimarea sa în funcție de $f(\alpha_1), f(\alpha_2), f'(\alpha_2)$ și $f'(\alpha_1)$. Pentru aceasta notăm:

$$s = 3 \frac{f(\alpha_2) - f(\alpha_1)}{\alpha_2 - \alpha_1}, z = s - f'(\alpha_1) - f'(\alpha_2), w^2 = z^2 - f'(\alpha_1)f'(\alpha_2).$$

Din condițiile de interpolare avem:

$$s = 3[c_1(\alpha_2 - \alpha_1)^2 + c_2(\alpha_2 - \alpha_1) + c_3]$$

$$\begin{aligned} z &= c_2(\alpha_2 - \alpha_1) + c_3 \\ w^2 &= (\alpha_2 - \alpha_1)^2(c_2^2 - 3c_1c_3). \end{aligned}$$

Deci rezultă:

$$(\alpha_2 - \alpha_1)c_2 = z - c_3, \sqrt{c_2^2 - 3c_1c_3} = \frac{w}{\alpha_2 - \alpha_1}$$

și

$$c_2 + \sqrt{c_2^2 - 3c_1c_3} = \frac{z + w - c_3}{\alpha_2 - \alpha_1}. \quad (3.30)$$

Însă $c_3 = f'(\alpha)$ și substituind (3.30) în (3.29) avem relația

$$\bar{\alpha} - \alpha_1 = \frac{-(\alpha_2 - \alpha_1)f'(\alpha_1)}{z + w - f'(\alpha_1)},$$

ce poate fi rescrisă în următoarea formă:

$$\begin{aligned} \bar{\alpha} - \alpha_1 &= \frac{-(\alpha_2 - \alpha_1)f'(\alpha_1)f'(\alpha_2)}{(z + w - f'(\alpha_1))f'(\alpha_2)} = \frac{-(\alpha_2 - \alpha_1)(z^2 - w^2)}{f'(\alpha_2)(z + w) - (z^2 - w^2)} \\ &= \frac{(\alpha_2 - \alpha_1)(w - z)}{f'(\alpha_2) - z + w}. \end{aligned}$$

Această relație este lipsită de utilitate pentru calcularea lui $\bar{\alpha}$ deoarece numitorul este foarte mic sau chiar se poate anula. De aceea, considerăm o formă alternativă favorabilă:

$$\begin{aligned} \bar{\alpha} - \alpha_1 &= \frac{-(\alpha_2 - \alpha_1)f'(\alpha_1)}{z + w - f'(\alpha_1)} = \frac{(\alpha_2 - \alpha_1)(w - z)}{f'(\alpha_2) - z + w} \\ &= \frac{(\alpha_2 - \alpha_1)(-f'(\alpha_1) + w - z)}{f'(\alpha_2) - f'(\alpha_1) + 2w} \\ &= (\alpha_2 - \alpha_1) \left(1 - \frac{f'(\alpha_2) + z + w}{f'(\alpha_2) - f'(\alpha_1) + 2w} \right), \end{aligned} \quad (3.31)$$

sau

$$\bar{\alpha} = \alpha_1 + (\alpha_2 - \alpha_1) \frac{w - f'(\alpha_1) - z}{f'(\alpha_2) - f'(\alpha_1) + 2w}. \quad (3.32)$$

În (3.31) sau (3.32) numitorul $f'(\alpha_2) - f'(\alpha_1) + 2w \neq 0$. De fapt, din moment ce $f'(\alpha_1) < 0$ și $f'(\alpha_2) > 0$, atunci $w^2 = z^2 - f'(\alpha_1)f'(\alpha_2) > 0$ și dacă luăm $w > 0$, atunci $f'(\alpha_2) - f'(\alpha_1) + 2w > 0$. Se poate arăta de asemenea că această metodă produce un șir α_k ce converge cu rată de ordinul 2 la α^* , adică $\lim_{k \rightarrow \infty} \frac{|\alpha_{k+1} - \alpha^*|}{|\alpha_k - \alpha^*|^2} = \rho$ cu $\rho > 0$.

Capitolul 4

Condiții de optimalitate pentru (UNLP)

Multe probleme din inginerie, economie sau fizică se formulează ca probleme de optimizare fără constrângeri. Astfel de probleme apar în găsirea punctului de echilibru al unui sistem prin minimizarea energiei acestuia, potrivirea unei funcții la un set de date folosind cele mai mici pătrate sau determinarea parametrilor unei distribuții de probabilitate corespunzătoare unui set de date. Probleme de optimizare fără constrângeri apar de asemenea când constrângerile sunt eliminate sau mutate în cost prin folosirea unei funcții de penalitate adecvate, după cum vom vedea în Partea a III-a a acestei lucrări. În concluzie, în această parte a lucrării ne concentrăm analiza asupra problemelor de optimizare neconstrânsă de forma:

$$(UNLP) : \quad \min_{x \in \mathbb{R}^n} f(x). \quad (4.1)$$

În acest capitol discutăm condițiile necesare și suficiente de optimalitate pentru problema generală (UNLP) și apoi particularizăm la cazul problemelor convexe, adică atunci când funcția obiectiv f este convexă. Pentru o expunere mai ușoară presupunem că funcția obiectiv $f : \mathbb{R}^n \rightarrow \mathbb{R}$ are domeniul efectiv $\text{dom} f \subseteq \mathbb{R}^n$ mulțime deschisă. După cum am precizat și în capitolele anterioare, în această lucrare considerăm extensia funcției f la întreg spațiul \mathbb{R}^n atribuindu-i valoarea $+\infty$ în punctele din afara domeniului efectiv. De aceea, căutăm punctele de minim ce fac parte din mulțimea deschisă $\text{dom} f$. Putem avea $\text{dom} f = \mathbb{R}^n$, dar de cele mai multe ori nu este cazul, la fel ca în următorul exemplu unde considerăm $\text{dom} f = (0, \infty)$ și presupunem că în afara mulțimii $\text{dom} f$

funcția ia valoarea $+\infty$:

$$\min_{x \in \mathbb{R}} \frac{1}{x} + x,$$

Reamintim că un punct x^* se numește punct de minim global pentru problema (UNLP) precedentă dacă $f(x^*) \leq f(x)$ pentru orice $x \in \text{dom} f$. Mai, mult, $f^* = f(x^*)$ se numește valoarea optimă a problemei de optimizare (UNLP). De asemenea, x^* este punct de minim local dacă există un $\delta > 0$ astfel încât $f(x^*) \leq f(x)$ pentru orice $x \in \text{dom} f$ cu $\|x - x^*\| \leq \delta$.

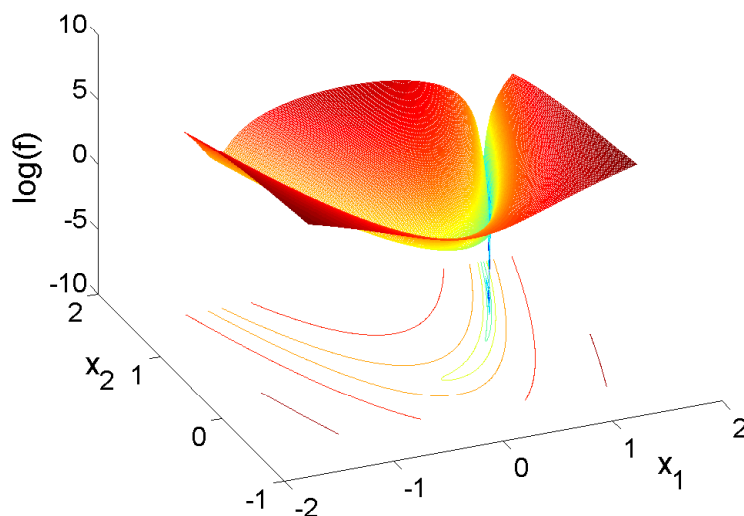


Figura 4.1: Funcția Rosenbrock.

Exemplul 4.0.1 În optimizarea fără constrângeri o funcție obiectiv des utilizată pentru a testa performanța algoritmilor este funcția Rosenbrock. Aceasta este o funcție neconvexă având următoarea formă:

$$f(x) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2.$$

Se poate observa ușor că punctul de minim global este $x^* = [1 \ 1]^T$ unde funcția ia valoarea optimă $f^* = f(1, 1) = 0$. Acest punct de minim se găsește într-o vale lungă dar îngustă (vezi Fig. 4.1), ceea ce face dificilă determinarea acestui punct de minim cu algoritmi numerici de optimizare.

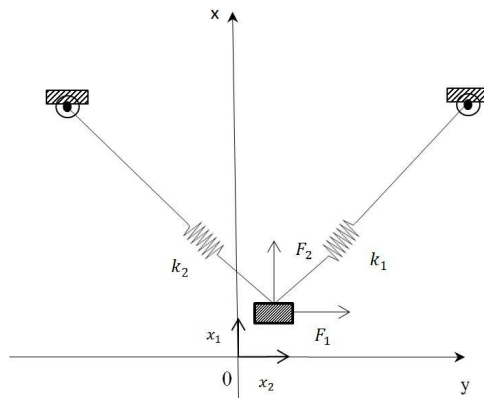


Figura 4.2: Un sistem neliniar cu două resorturi.

Exemplul 4.0.2 Considerăm un sistem neliniar compus din două resorturi (Fig. 4.2). Dislocarea x_1 și x_2 sub o anumită greutate aplicată poate fi obținută prin minimizarea energiei potențiale dată de expresia:

$$f(x_1, x_2) = \frac{1}{2}k_1 E_1^2(x_1, x_2) + \frac{1}{2}k_2 E_2^2(x_1, x_2) - F_1 x_1 - F_2 x_2,$$

în care extensiile resorturilor ca funcție de dislocările lor au următoarele expresii:

$$E_1(x_1, x_2) = \sqrt{(x_1 + 10)^2 + (x_2 - 10)^2} - 10\sqrt{2}$$

$$E_2(x_1, x_2) = \sqrt{(x_1 - 10)^2 + (x_2 - 10)^2} - 10\sqrt{2}.$$

Problema se reduce la a găsi (x_1, x_2) ce minimizează

$$\min_{x \in \mathbb{R}^2} f(x_1, x_2).$$

Folosind tehnici numerice de optimizare ce vor fi discutate în capitolele următoare, obținem că pentru $k_1 = k_2 = 1$ și $F_1 = 0, F_2 = 2$, soluția optimă este $x_1^* = 0$ și $x_2^* = 2.55$.

4.1 Condiții de ordinul I pentru (UNLP)

Mai întâi definim noțiunea de direcție de descreștere. O direcție $d \in \mathbb{R}^n$ se numește *direcție de descreștere* pentru funcția $f \in \mathcal{C}^1$ în punctul $x \in \text{dom} f$ dacă

$$\nabla f(x)^T d < 0.$$

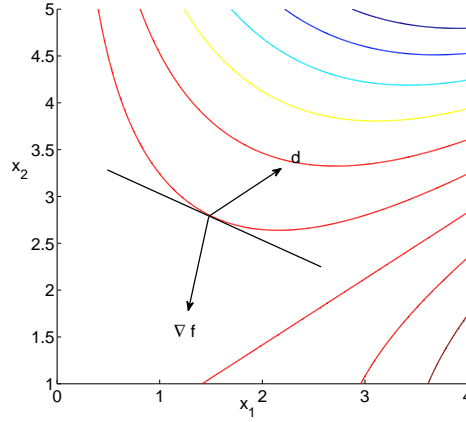


Figura 4.3: Mulțimile nivel (contururile) și o direcție de descreștere pentru funcția $f(x) = x_1^3 - 2x_1x_2^2$.

De exemplu, o direcție de descreștere este dată de următoarea expresie: $d = -B\nabla f(x)$, unde matricea B este pozitiv definită, adică $B \succ 0$ (vezi Fig. 4.3).

Avem următoarea interpretare: dacă d este direcție de descreștere în $x \in \text{dom} f$ atunci funcția obiectiv descrește în vecinătatea lui x . Într-adevăr, mulțimea $\text{dom} f$ fiind deschisă, putem găsi un $t > 0$ suficient de mic astfel încât oricare ar fi $\tau \in [0, t]$ avem $x + \tau d \in \text{dom} f$ și $\nabla f(x + \tau d)^T d < 0$ (datorită continuității lui $\nabla f(\cdot)$ într-o vecinătate a lui x). Din teorema lui Taylor, există un $\theta \in [0, t]$ astfel încât

$$f(x + td) = f(x) + t \underbrace{\nabla f(x^* + \theta d)^T d}_{<0} < f(x).$$

Teorema 4.1.1 (Condiții necesare de ordinul I) Fie f o funcție diferențiabilă cu gradientul continuu (i.e. $f \in \mathcal{C}^1$) și $x^* \in \text{dom} f$ un punct de minim local al problemei de optimizare (UNLP). Atunci gradientul funcției satisface relația:

$$\nabla f(x^*) = 0. \quad (4.2)$$

Demonstrație: Presupunem prin contradicție că $\nabla f(x^*) \neq 0$. Atunci putem arăta că $d = -\nabla f(x^*)$ este o direcție de descreștere, i.e. funcția obiectiv poate lua o valoare mai mică în jurul lui x^* . Într-adevăr, putem găsi un $t > 0$ suficient de mic astfel încât oricare ar fi $\tau \in [0, t]$ avem

$\nabla f(x^* + \tau d)^T d = -\nabla f(x^* - \tau \nabla f(x^*))^T \nabla f(x^*) < 0$. Mai mult, există un $\theta \in [0, t]$ ce satisface:

$$f(x^* - t \nabla f(x^*)) = f(x^*) - t \nabla f(x^* + \theta \nabla f(x^*))^T \nabla f(x^*) < f(x^*).$$

Aceasta este o contradicție cu ipoteza că x^* este un punct de minim local.

□

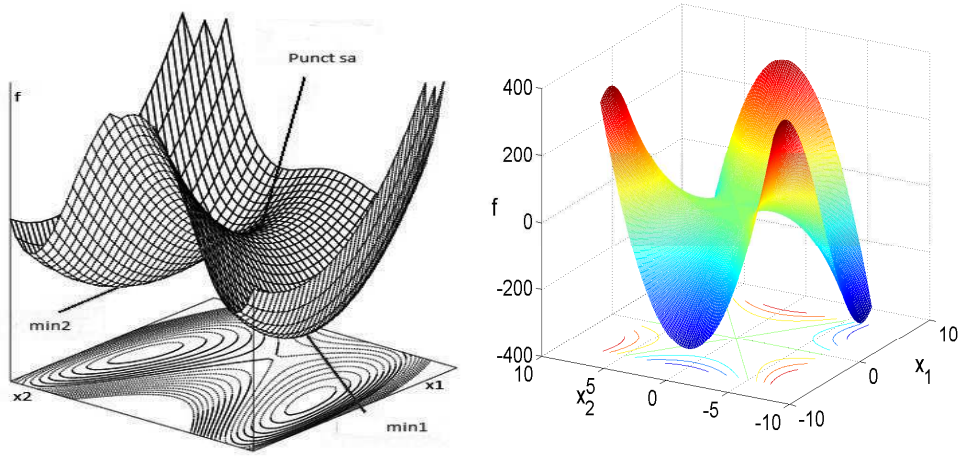


Figura 4.4: Exemple de puncte staționare.

Orice punct $x^* \in \text{dom} f$ ce satisface condițiile necesare de ordinul întâi $\nabla f(x^*) = 0$ se numește *punct staționar* al problemei de optimizare fără constrângeri (UNLP). În Fig. 4.4 distingem mai multe tipuri de puncte staționare: puncte de minim, puncte de maxim și puncte sa. Din condițiile necesare de ordinul I concluzionăm că pentru a găsi punctele staționare ale unei probleme (UNLP) trebuie să rezolvăm un sistem neliniar $\nabla f(x) = 0$ de n ecuații cu n necunoscute. Acest sistem va fi rezolvat în capitolele următoare cu metode numerice (algoritmi iterativi).

Exemplul 4.1.1 Considerăm problema de optimizare (vezi Fig. 4.3)

$$\min_{x \in \mathbb{R}^2} f(x) \quad (= x_1^3 - 2x_1x_2^2).$$

Pentru a găsi punctele staționare rezolvăm sistemul neliniar de două ecuații $\nabla f(x) = 0$, i.e.:

$$3x_1^2 - 2x_2^2 = 0 \quad \text{și} \quad 4x_1x_2 = 0.$$

Singura soluție a acestui sistem este $x_1^* = 0$ și $x_2^* = 0$. Soluția $x^* = [0 \ 0]^T$ nu este punct de minim sau maxim, este punct șa pentru funcția $f(x) = x_1^3 - 2x_1x_2^2$. Într-adevăr, dacă luăm $x_1 = x_2$ observăm că funcția obiectiv evaluată în $f(x_1, x_1) = -x_1^3$ poate lua atât valori pozitive cât și valori negative în orice vecinătate a lui 0.

4.2 Condiții de ordinul II pentru (UNLP)

Pentru a arăta că un punct staționar este punct de extrem (minim sau maxim local) trebuie să folosim informație despre Hessiana funcției obiectiv. În cele ce urmează dăm condiții necesare și suficiente pentru caracterizarea punctelor de minim local utilizând Hessiana.

Teorema 4.2.1 (Condiții necesare de ordinul II) *Fie f o funcție diferențiabilă de două ori cu Hessiana continuă (i.e. $f \in \mathcal{C}^2$) și $x^* \in \text{dom} f$ un punct de minim local al problemei (UNLP). Atunci Hessiana în x^* este pozitiv semidefinită, i.e.:*

$$\nabla^2 f(x^*) \succcurlyeq 0. \quad (4.3)$$

Demonstrație: Dacă condiția (4.3) nu este satisfăcută, atunci există o direcție $d \in \mathbb{R}^n$ astfel încât $d^T \nabla^2 f(x^*) d < 0$. Atunci, datorită continuității lui $\nabla^2 f(\cdot)$ în jurul lui x^* putem alege un parametru suficient de mic $t > 0$ astfel încât oricare ar fi $\tau \in [0, t]$ următoarea relație are loc:

$$d^T \nabla^2 f(x^* + \tau d) d < 0.$$

Din teorema lui Taylor rezultă că există un $\theta \in [0, t]$ astfel încât:

$$f(x^* + td) = f(x^*) + \underbrace{t \nabla f(x^*)^T d}_{=0} + \frac{1}{2} t^2 \underbrace{d^T \nabla^2 f(x^* + \theta d) d}_{<0} < f(x^*),$$

ceea ce intră în contradicție cu faptul că x^* este un punct de minim local. \square

Condiția necesară de ordinul II (4.3) nu este suficientă pentru ca un punct staționar x^* să fie punct de minim. Acest lucru este ilustrat de funcția $f(x) = x^3$ pentru care punctul staționar $x^* = 0$ satisface condițiile necesare de ordinul doi, dar nu este punct de minim/maxim local. Observăm că $x^* = 0$ este punct șa. În secțiunea următoare enunțăm condițiile suficiente de optimalitate.

Teorema 4.2.2 (Condiții suficiente de ordinul II) Fie f o funcție diferențiabilă de două ori cu Hessiana continuă (i.e. $f \in \mathcal{C}^2$) și $x^* \in \text{dom} f$ un punct staționar (i.e. $\nabla f(x^*) = 0$) astfel încât Hessiana este pozitiv definită (i.e. $\nabla^2 f(x^*) \succ 0$). Atunci x^* este un punct strict de minim local al problemei (UNLP).

Demonstrație: Fie λ_{\min} valoarea proprie minimă a matricei $\nabla^2 f(x^*)$. Evident, $\lambda_{\min} > 0$ din moment ce este satisfăcută relația $\nabla^2 f(x^*) \succ 0$ și mai mult,

$$d^T \nabla^2 f(x^*) d \geq \lambda_{\min} \|d\|^2 \quad \forall d \in \mathbb{R}^n.$$

Din aproximarea Taylor avem:

$$\begin{aligned} f(x^* + d) - f(x^*) &= \nabla f(x^*)^T d + \frac{1}{2} d^T \nabla^2 f(x^*) d + \mathcal{R}(\|d\|^2) \\ &\geq \frac{\lambda_{\min}}{2} \|d\|^2 + \mathcal{R}(\|d\|^2) = \left(\frac{\lambda_{\min}}{2} + \frac{\mathcal{R}(\|d\|^2)}{\|d\|^2} \right) \|d\|^2. \end{aligned}$$

Știind că $\lambda_{\min} > 0$ atunci există $\epsilon > 0$ și $\delta > 0$ astfel încât $\frac{\lambda_{\min}}{2} + \frac{\mathcal{R}(\|d\|^2)}{\|d\|^2} \geq \delta$ pentru orice $\|d\| \leq \epsilon$, ceea ce conduce la concluzia că x^* este un punct strict de minim local. \square

Exemplul 4.2.1 Considerăm următoarea problemă de optimizare

$$\min_{x \in \mathbb{R}^2} f(x) \quad (= x_1^3 - x_1^2 x_2 + 2x_2^2).$$

Punctele staționare sunt soluțiile sistemului nelinier $\nabla f(x) = 0$, i.e.:

$$3x_1^2 - 2x_1 x_2 = 0 \quad \text{și} \quad -x_1^2 + 4x_2 = 0.$$

Acest sistem are două soluții $[0 \ 0]^T$ și $[6 \ 9]^T$. Hessiana are expresia:

$$\nabla^2 f(x) = \begin{bmatrix} 6x_1 - 2x_2 & -2x_1 \\ -2x_1 & 4 \end{bmatrix}.$$

Această matrice este pozitiv semidefinită în $[0 \ 0]^T$ și indefinită în $[6 \ 9]^T$. În concluzie, punctul staționar $[6 \ 9]^T$ nu este punct de extrem (minim sau maxim local).

4.3 Condiții de optimalitate pentru probleme convexe

În acest subcapitol discutăm condițiile suficiente de optimalitate în cazul convex, adică funcția obiectiv f în problema de optimizare (UNLP) este convexă. Primul rezultat se referă la următoarea problemă de optimizare constrânsă:

Teorema 4.3.1 (Condiții de optimalitate generale) *Fie o mulțime convexă X și funcția $f \in \mathcal{C}^1$ (nu neapărat convexă). Pentru problema de optimizare constrânsă*

$$\min_{x \in X} f(x)$$

următoarele condiții sunt satisfăcute:

- (i) *dacă x^* este minim local, atunci $\nabla f(x^*)^T(x - x^*) \geq 0 \quad \forall x \in X$;*
- (ii) *dacă f este funcție convexă, atunci x^* este punct de minim dacă și numai dacă $\nabla f(x^*)^T(x - x^*) \geq 0 \quad \forall x \in X$.*

Demonstrație: (i) Presupunem că există un $y \in X$ astfel încât

$$\nabla f(x^*)^T(y - x^*) < 0.$$

Din teorema lui Taylor rezultă că pentru un $t > 0$ există un $\theta \in [0, 1]$ astfel încât:

$$f(x^* + t(y - x^*)) = f(x^*) + t\nabla f(x^* + \theta t(y - x^*))^T(y - x^*).$$

Din continuitatea lui ∇f , alegând un t suficient de mic avem $\nabla f(x^* + \theta t(y - x^*))^T(y - x^*) < 0$ și de aceea $f(x^* + t(y - x^*)) < f(x^*)$ care este în contradicție cu faptul că x^* este minim local al problemei de optimizare cu constrângeri din enunțul teoremei.

(ii) Dacă f este convexă, utilizând condițiile de convexitate de ordinul întâi avem: $f(x) \geq f(x^*) + \nabla f(x^*)^T(x - x^*)$ pentru orice $x \in X$. Întrucât $\nabla f(x^*)^T(x - x^*) \geq 0$ rezultă că $f(x) \geq f(x^*)$ pentru orice $x \in X$, adică x^* este punct de minim global. \square

Teorema 4.3.2 *Pentru o problemă de optimizare convexă $\min_{x \in X} f(x)$ (adică X este mulțime convexă și f funcție convexă), orice minim local este de asemenea minim global.*

Demonstrație: Fie x^* un minim local pentru problema de optimizare convexă precedentă. Arătăm că pentru orice punct $y \in X$ dat avem $f(y) \geq f(x^*)$. Într-adevăr, întrucât x^* este minim local, există o vecinătate \mathcal{N} a lui x^* astfel încât pentru orice $\tilde{x} \in X \cap \mathcal{N}$ avem $f(\tilde{x}) \geq f(x^*)$. Considerând segmentul cu capetele în x^* și y . Acest segment este conținut în X datorită proprietății de convexitate a lui X . Mai departe, alegem un \tilde{x} pe acest segment în vecinătatea \mathcal{N} , însă diferit de x^* , adică alegem $\tilde{x} = x^* + t(y - x^*)$, unde $t \in (0, 1)$ astfel încât $\tilde{x} \in X \cap \mathcal{N}$. Datorită optimalității locale, avem $f(x^*) \leq f(\tilde{x})$, și datorită convexității lui f avem:

$$f(\tilde{x}) = f(x^* + t(y - x^*)) \leq f(x^*) + t(f(y) - f(x^*)).$$

Rezultă că $t(f(y) - f(x^*)) \geq 0$, implicând $f(y) - f(x^*) \geq 0$ ceea ce conduce la concluzia că x^* este punct de minim global. \square

Teorema 4.3.3 (Condiții suficiente de ordinul întâi pentru cazul convex) Fie $f \in \mathcal{C}^1$ o funcție convexă. Dacă x^* este punct staționar al lui f (i.e. $\nabla f(x^*) = 0$), atunci x^* este punct de minim global al problemei de optimizare convexe fără constrângeri $\min_{x \in \mathbb{R}^n} f(x)$.

Demonstrație: Întrucât f este convexă avem:

$$f(x) \geq f(x^*) + \underbrace{\nabla f(x^*)(x - x^*)}_{=0} = f(x^*) \quad \forall x \in \mathbb{R}^n$$

ceea ce arată că x^* este minim global. \square

În concluzie, pentru o problemă de optimizare neconstrânsă $\min_{x \in \mathbb{R}^n} f(x)$, unde $f \in \mathcal{C}^1$, o condiție necesară pentru ca punctul x^* să fie punct de extrem local este:

$$\nabla f(x^*) = 0. \quad (4.4)$$

În general, dacă funcția obiectiv nu este convexă, se rezolvă sistemul neliniar de ecuații $\nabla f(x^*) = 0$ și se verifică dacă soluția este punct de minim local sau nu, folosind condițiile de optimalitate suficiente de ordinul doi. Pentru problemele convexe neconstrânse, adică f este convexă, o condiție necesară și suficientă pentru ca punctul x^* să fie minim global este dată de relația $\nabla f(x^*) = 0$.

4.4 Analiza perturbațiilor

În domeniile numerice ale matematicii nu există posibilitatea de a evalua o funcție cu o precizie mai mare decât cea oferită de mașinile de calcul. De aceea, de cele mai multe ori se calculează doar soluții pentru probleme ale căror date sunt perturbate, iar interesul se îndreaptă către minimele stabile la apariția perturbațiilor. Acesta este cazul punctelor de minim strict locale ce satisfac condițiile suficiente de ordinul doi.

Considerăm funcții obiectiv de forma $f(x, a)$ ce depind nu doar de variabila de decizie $x \in \mathbb{R}^n$, dar și de un *parametru de perturbație* $a \in \mathbb{R}^m$. Suntem interesați de familia parametrică de probleme:

$$\min_{x \in \mathbb{R}^n} f(x, a)$$

ce produce minime de forma $x^*(a)$ ce depind de parametrul a .

Teorema 4.4.1 (Stabilitatea soluțiilor parametrice)

Presupunem că funcția $f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ este de clasă \mathcal{C}^2 și considerăm minimizarea funcției $f(\cdot, \bar{a})$ pentru o valoare fixată a parametrului $\bar{a} \in \mathbb{R}^m$. Dacă punctul de minim corespunzător \bar{x} satisface condițiile suficiente de ordinul doi, i.e. $\nabla_x f(\bar{x}, \bar{a}) = 0$ și $\nabla_x^2 f(\bar{x}, \bar{a}) \succ 0$, atunci există o vecinătate $\mathcal{N} \subset \mathbb{R}^m$ în jurul lui \bar{a} astfel încât funcția parametrică de minim $x^(a)$ este bine definită pentru orice $a \in \mathcal{N}$, este diferențiabilă pe \mathcal{N} și $x^*(\bar{a}) = \bar{x}$. Derivata sa în punctul \bar{a} este dată de:*

$$\frac{\partial(x^*(\bar{a}))}{\partial a} = -\frac{\partial(\nabla_x f(\bar{x}, \bar{a}))}{\partial a} \left(\nabla_x^2 f(\bar{x}, \bar{a}) \right)^{-1}. \quad (4.5)$$

Mai mult, fiecare $x^(a)$ cu $a \in \mathcal{N}$ satisface condițiile suficiente de ordinul doi și deci este un punct de minim strict local.*

Demonstrație: Existența funcției diferențiabile $x^* : \mathcal{N} \rightarrow \mathbb{R}^n$ rezultă din teorema funcțiilor implicite (dată în Apendice) aplicată condiției de staționaritate $\nabla_x f(x^*(\bar{a}), \bar{a}) = 0$. Pentru derivarea ecuației (4.5) se folosesc regulile standard de diferențiere:

$$\begin{aligned} 0 &= \frac{\partial(\nabla_x f(x^*(a), a))}{\partial a} \\ &= \frac{\partial x^*(a)}{\partial a} \cdot \underbrace{\frac{\partial(\nabla_x f(x^*(a), a))}{\partial x}}_{=\nabla_x^2 f} + \frac{\partial(\nabla_x f(x^*(a), a))}{\partial a}. \end{aligned}$$

Pentru a arăta că punctele de minim $x^*(a)$ satisfac condițiile suficiente de ordinul doi, se observă că Hessiana este continuă și se ține seama de faptul că $\nabla_x^2 f(\bar{x}, \bar{a}) \succ 0$. \square

O analiză extinsă a sensibilității soluțiilor la perturbații și o prezentare detaliată a teoriei optimizării parametrice se poate găsi în [6].

Capitolul 5

Convergența metodelor de descreștere

În Capitolul 4 s-a demonstrat că pentru aflarea unui punct de minim local/global corespunzător unei probleme de optimizare neconstrânsă:

$$(UNLP) : \quad f^* = \min_{x \in \mathbb{R}^n} f(x), \quad (5.1)$$

este nevoie de rezolvarea unui sistem neliniar de n ecuații cu n necunoscute:

$$\nabla f(x) = 0.$$

În unele cazuri acest sistem poate fi rezolvat analitic:

Exemplul 5.0.1 (QP neconstrâns) *Considerăm următoarea problemă de tip QP neconstrânsă:*

$$\min_{x \in \mathbb{R}^n} f(x) \quad \left(= \frac{1}{2} x^T Q x - q^T x \right), \quad (5.2)$$

unde matricea Q este inversabilă (de exemplu, dacă $Q \succ 0$ atunci funcția obiectiv este convexă și deci problema de optimizare este convexă). Din condiția $0 = \nabla f(x) = Qx - q$, unicul punct staționar este $x^* = Q^{-1}q$. Dacă $Q \succ 0$, atunci $x^* = Q^{-1}q$ este punct de minim global și valoarea optimă a problemei (5.2) este dată de următoarea expresie:

$$f^* = \min_{x \in \mathbb{R}^n} \frac{1}{2} x^T Q x - q^T x = -\frac{1}{2} q^T Q^{-1} q.$$

Cu toate acestea, în majoritatea cazurilor $\nabla f(x) = 0$ este un sistem de ecuații neliniare ce nu poate fi rezolvat analitic, ci este nevoie de metode iterative pentru rezolvarea lui. Cea mai mare parte a acestei lucrări este dedicată prezentării și analizei diferiților algoritmi pentru rezolvarea problemelor de optimizare (UNLP) sau a sistemului neliniar $\nabla f(x) = 0$. Toți algoritmi prezentați în această lucrare sunt iterativi. Prin *iterativ* înțelegem că acești algoritmi generează un șir de puncte, fiecare punct fiind calculat pe baza punctelor găsite anterior. De asemenea, majoritatea algoritmilor sunt de descreștere, adică în fiecare punct nou generat de către algoritm valoarea funcției obiectiv este mai mică decât în punctul generat anterior. În cele mai multe cazuri, vom arăta că șirul de puncte generate în acest mod de către algoritm converge într-un număr finit sau infinit de pași la o soluție a problemei originale.

Un algoritm iterativ pornește de la un punct inițial. Dacă pentru orice punct inițial putem garanta că algoritmul produce un șir de puncte convergente la o soluție, atunci pentru acel algoritm spunem că este *convergent global*. În multe situații, algoritmi dezvoltați nu pot garanta convergența globală și numai inițializați în apropierea unui punct de optim vor produce un șir de puncte convergente la acel punct de optim. Atunci spunem că algoritmul este *convergent local*. În general, convergența algoritmilor de optimizare poate fi tratată printr-o analiză a unei teorii generale a algoritmilor dezvoltată în anii 1960 de Zangwill. Vom prezenta această teorie în cele ce urmează.

5.1 Metode numerice de optimizare

Considerăm problema de optimizare fără constrângeri (5.1). Există diferite metode iterative pentru rezolvarea unei astfel de probleme, iar în capitolele următoare vom discuta despre cele mai importante dintre ele. Presupunem că o problemă de optimizare aparține unei anumite clase de probleme \mathcal{F} . În general, o metodă numerică este dezvoltată în scopul rezolvării diferitelor probleme ce împărtășesc caracteristici similare (e.g. continuitate, convexitate, etc.). Datele cunoscute din structura problemei se regăsesc sub numele de *model* (i.e. formularea problemei, funcțiile ce descriu problema, etc.). Pentru rezolvarea problemei, o metodă numerică va trebui să colecteze informația specifică, iar procesul de colectare a datelor se realizează cu ajutorul unui *oracol* (i.e. unitate de calcul ce returnează date sub forma unor răspunsuri la

întrebări succesive din partea metodei). În concluzie, metoda numerică rezolvă problema prin colectarea datelor și manipularea răspunsurilor oracolului. Există diferite tipuri de oracole:

0. oracole de ordinul zero \mathcal{O}_0 ce furnizează informație bazată doar pe evaluarea funcției obiectiv, i.e. $f(x)$;
1. oracole de ordinul întâi \mathcal{O}_1 ce furnizează informație bazată pe evaluarea funcției și gradientului său, i.e. $f(x)$ și $\nabla f(x)$;
2. oracole de ordinul doi \mathcal{O}_2 ce furnizează informație bazată pe evaluarea funcției, gradientului și Hessienei, i.e. $f(x)$, $\nabla f(x)$ și $\nabla^2 f(x)$.

Eficiența unei metode numerice constă în efortul numeric necesar metodei pentru rezolvarea unei anumite clase de probleme. Rezolvarea unei probleme, în unele cazuri, constă în aflarea unei soluții exacte, însă în cele mai multe dintre cazuri este posibilă doar aproximarea soluției. De aceea, pentru rezolvare este suficientă aflarea unei soluții aproximative cu o acuratețe prestabilită ϵ . În general, această acuratețe reprezintă de asemenea criteriul de oprire pentru metoda numerică aleasă. Pentru cazul particular al problemelor de optimizare neconstrânse (UNLP), adică (5.1), criteriul de oprire în general utilizat este următorul:

$$\|\nabla f(x)\| \leq \epsilon.$$

Pe lângă acest criteriu, se mai folosește și criteriul referitor la apropierea valorii funcției de minimizat față de valoarea sa optimă:

$$|f(x) - f^*| \leq \epsilon.$$

Anumite implementări utilizează și alte criterii de oprire a iterațiilor, cum ar fi de exemplu distanța dintre estimațiile variabilelor:

$$\|x_{k+1} - x_k\| \leq \epsilon.$$

Schema generală a unui algoritm numeric de optimizare iterativ constă în următorii pași:

O metodă generică de optimizare numerică:

1. se începe cu un punct inițial dat x_0 , acuratețea $\epsilon > 0$ și contorul $k = 0$;

2. la pasul k notăm cu \mathcal{I}_k mulțimea ce conține toată informația acumulată de la oracol până la iterația k :
 - 2.1 se apelează oracolul \mathcal{O} în punctul x_k ;
 - 2.2 se actualizează informația $\mathcal{I}_{k+1} = \mathcal{I}_k \cup \mathcal{O}(x_k)$;
 - 2.3 se aplică regulile metodei numerice folosind ultimele informații \mathcal{I}_{k+1} pentru calculul următorului punct x_{k+1} ;
3. se verifică criteriul de oprire; dacă criteriul de oprire nu este satisfăcut, se repetă pasul 2.

Complexitatea unei metode numerice poate fi exprimată în următoarele forme:

- (i) complexitate *analitică*, dată de numărul total de apeluri ale oracolului;
- (ii) complexitate *aritmetică*, dată de numărul total de operații aritmetice.

Rata de convergență se referă la viteza cu care șirul x_k se apropie de soluție x^* . Ordinul de convergență este cel mai mare număr pozitiv q ce satisface următoarea relație:

$$0 \leq \overline{\lim}_{k \rightarrow \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^q} < \infty,$$

în care precizăm că limita superioară a unui șir z_k *sup lim* este definită de:

$$\overline{\lim}_{k \rightarrow \infty} z_k = \lim_{n \rightarrow \infty} y_n, \quad \text{unde} \quad y_n = \sup_{k \geq n} z_k.$$

Presupunând că limita șirului x_k există, atunci q indică comportamentul șirului. Când q este mare, atunci rata de convergență este mare, deoarece distanța până la x^* este redusă cu q zecimale într-un singur pas:

$$\|x_{k+1} - x^*\| \approx \beta \|x_k - x^*\|^q.$$

Convergență liniară: dacă există $\beta \in (0, 1)$ și $q = 1$ astfel încât

$$\|x_{k+1} - x^*\| \leq \beta \|x_k - x^*\|$$

și deci $\|x_k - x^*\| \approx c\beta^k$, unde $c = \|x_0 - x^*\|$. De exemplu, șirul $x_k = \beta^k$, unde $\beta \in (0, 1)$, converge liniar la $x^* = 0$.

Convergență superliniară: dacă $\lim_{k \rightarrow \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0$, aici de asemenea $q = 1$, sau echivalent

$$\|x_{k+1} - x^*\| \leq \beta_k \|x_k - x^*\| \quad \text{cu} \quad \beta_k \rightarrow 0.$$

De exemplu, șirul $x_k = \frac{1}{k!}$ converge superliniar la $x^* = 0$, întrucât $\frac{x_{k+1}}{x_k} = \frac{1}{k+1}$.

Convergență pătratică: dacă $\lim_{k \rightarrow \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} = \beta$, unde $\beta \in (0, \infty)$ și $q = 2$, sau echivalent

$$\|x_{k+1} - x^*\| \leq \beta \|x_k - x^*\|^2.$$

De exemplu șirul $x_k = \frac{1}{2^{2^k}}$ converge pătratic la $x^* = 0$, deoarece $\frac{x_{k+1}}{(x_k)^2} = \frac{2^{2^{k+1}}}{(2^{2^k})^2} = 1 < \infty$. Pentru $k = 6$, $x^k = \frac{1}{2^{64}} \approx 0$, de aceea, în practică, convergența la acuratețea mașinii de calcul se realizează după aproximativ șase iterații.

Întâlnim adesea și convergență *subliniară* definită astfel:

$$\|x_k - x^*\| \leq \frac{\beta}{k^q},$$

unde $q > 0$.

R-convergență: Dacă șirul de norme $\|x^k - x^*\|$ este mărginit superior de șirul $y_k \rightarrow 0$, adică $\|x_k - x^*\| \leq y_k$ și dacă y_k converge cu o rată dată, i.e. liniară, superliniară sau pătratică, atunci x_k converge *R-liniar*, *R-superliniar* sau *R-pătratic* la x^* . Aici, R indică *root*, deoarece convergența R-liniară poate fi de asemenea definită prin criteriul rădăcinii $\lim_{k \rightarrow \infty} \sqrt[k]{\|x_k - x^*\|} < 1$.

Exemplul 5.1.1 *Considerăm șirul de numere reale convergent la zero:*

$$x_k = \begin{cases} \frac{1}{2^k} & \text{dacă } k \text{ este par} \\ 0 & \text{altfel.} \end{cases}$$

Acest șir are o convergență R-liniară, dar nu regulată ca a unui șir ce converge liniar.

Remarca 5.1.1 *Cele trei convergențe și ratele de R-convergență corespunzătoare satisfac anumite relații între ele. În continuare, $X \Rightarrow Y$*

are semnificația: dacă șirul converge cu rata X , atunci aceasta implică că șirul de asemenea converge cu rata Y .

$$\begin{array}{ccccc} \text{patratic} & \Rightarrow & \text{superliniar} & \Rightarrow & \text{liniar} \\ \Downarrow & & \Downarrow & & \Downarrow \\ R - \text{patratic} & \Rightarrow & R - \text{superliniar} & \Rightarrow & R - \text{liniar} \end{array}$$

Se observă că rata pătratică asigură cea mai rapidă convergență.

5.2 Convergența metodelor numerice

Considerând spațiul metric (X, ρ) , o metodă numerică poate fi privită ca o aplicație punct-multime $M : X \rightarrow 2^X$, definită de $x_{k+1} \in M(x_k)$. Modul cum se alege $x_{k+1} \in M(x_k)$ este dat de metoda dezvoltată. Cu toate acestea, o metodă numerică nu este un proces aleatoriu deoarece aceasta generează același șir x_k când se pornește din același punct inițial x_0 . Definiția metodei în această manieră oferă posibilitatea analizării ei cu instrumente matematice mai laborioase.

Exemplul 5.2.1 Considerăm următoarea aplicație punct-multime

$$x_{k+1} \in \left[-\frac{|x_k|}{n}, \frac{|x_k|}{n} \right],$$

pentru care o instanță particulară este definită de un punct inițial x_0 și iterația

$$x_{k+1} = \frac{|x_k|}{n}.$$

Definiția 5.2.1 Fie spațiul metric (X, ρ) , o submultime $S \subseteq X$ și o metodă descrisă de aplicația punct-multime $M : X \rightarrow 2^X$. Definim funcția descrescătoare $\phi : X \rightarrow \mathbb{R}$ pentru perechea (S, M) , o funcție ce satisface următoarele condiții:

- (i) pentru orice $x \in S$ și $y \in M(x)$ avem $\phi(y) \leq \phi(x)$;
- (ii) pentru orice $x \notin S$ și $y \in M(x)$ avem $\phi(y) < \phi(x)$.

Exemplul 5.2.2 Fie problema de optimizare $\min_{x \in X} f(x)$, unde X este o multime convexă și f este o funcție diferențiabilă. Definim $S = \{x^* \in \mathbb{R}^n : \langle \nabla f(x^*), x - x^* \rangle \geq 0 \quad \forall x \in X\}$ mulțimea punctelor staționare (adică mulțimea tuturor soluțiilor posibile – minime locale, maxime locale, puncte ș.a). Se observă de asemenea că în general alegem $\phi = f$, adică metoda alege x_{k+1} astfel încât $f(x_{k+1}) \leq f(x_k)$.

Definiția 5.2.2 *O aplicație punct-multime $M : X \rightarrow 2^X$ este închisă în punctul x_0 dacă pentru orice două șiruri $x_k \rightarrow x_0$ și $y_k \rightarrow y_0$ cu $y_k \in M(x_k)$, avem $y_0 \in M(x_0)$. Aplicația M este închisă dacă este închisă în toate punctele din X .*

Teorema 5.2.1 (Teorema de convergența generală) *Fie o metodă numerică M pe spațiul metric (X, ρ) , șirul $x_{k+1} \in M(x_k)$, iar S mulțimea soluțiilor. Presupunem următoarele condiții satisfăcute:*

- (i) *șirul x_k se află într-o mulțime compactă;*
- (ii) *M este o aplicație punct-multime închisă pe $X \setminus S$;*
- (iii) *există o funcție continuă ϕ decrescătoare pentru perechea (M, S) .*

Atunci toate punctele limită ale șirului x_k aparțin mulțimii S .

5.3 Metode de descreștere

În continuare, considerăm o metodă iterativă de optimizare de forma:

$$x_{k+1} = x_k + \alpha_k d_k,$$

în care presupunem că d_k este o *direcție de descreștere* pentru f în x_k , iar $\alpha_k \in (0, 1]$ este *lungimea pasului* (vezi Fig. 5.1). Precizăm că dacă d_k este o direcție de descreștere pentru f în x_k atunci există $\alpha_k > 0$ suficient de mic astfel încât $f(x_{k+1}) < f(x_k)$. În cele ce urmează vom analiza felul cum putem alege pasul α_k și direcția de descreștere d_k astfel încât metoda respectivă să producă un șir de puncte convergente la un punct staționar al problemei (UNLP).

5.3.1 Strategii de alegere a lungimii pasului

Prezentăm în această secțiune cele mai des întâlnite proceduri de alegere a lungimii pasului $\alpha_k \in (0, 1]$. Ideea de bază constă în alegerea adecvată a pasului α_k astfel încât să garantăm o descreștere suficientă în funcția obiectiv, adică $f(x_{k+1}) < f(x_k)$, și în același timp să și facem avans sensibil către soluția problemei, adică $\lim_{k \rightarrow \infty} x_k = x^*$. În calcularea lungimii pasului trebuie să facem un compromis între o reducere substanțială a funcției obiectiv f și calculul numeric necesar determinării pasului.

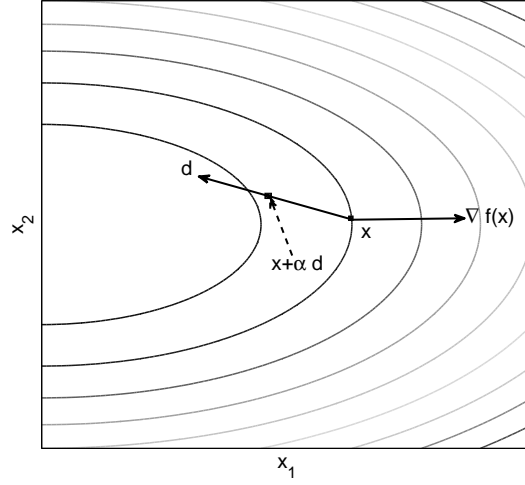


Figura 5.1: Metoda direcțiilor de descreștere.

În cazul *ideal* alegem lungimea pasului după următoarea relație:

$$\alpha_k = \arg \min_{0 \leq \alpha \leq 1} \phi(\alpha) \quad (= f(x_k + \alpha d_k)).$$

După cum am menționat, există diferite metode eficiente pentru determinarea unui punct de optim al unei probleme de optimizare unidimensională. În cele ce urmează numim această procedură *metoda ideală de alegere a lungimii pasului*.

Cu toate acestea, în multe situații problema de optimizare unidimensională corespunzătoare alegerii ideale a lungimii pasului este foarte dificil de rezolvat. De aceea, alte modalități de alegere a lungimii pasului α_k au fost dezvoltate, iar printre acestea cea mai cunoscută este definită de *condițiile Wolfe*: se caută α_k astfel încât următoarele două condiții sunt satisfăcute (vezi Fig. 5.2)

$$(W1) \quad f(x_k + \alpha_k d_k) \leq f(x_k) + c_1 \alpha_k \nabla f(x_k)^T d_k, \quad \text{unde } c_1 \in (0, 1)$$

$$(W2) \quad \nabla f(x_k + \alpha_k d_k)^T d_k \geq c_2 \nabla f(x_k)^T d_k, \quad \text{unde } 0 < c_1 < c_2 < 1.$$

În general, condiția (W1) de una singură nu este suficientă pentru a garanta că algoritmul de optimizare face un progres rezonabil de-a lungul direcției de căutare. Însă dacă lungimea pasului se alege în mod

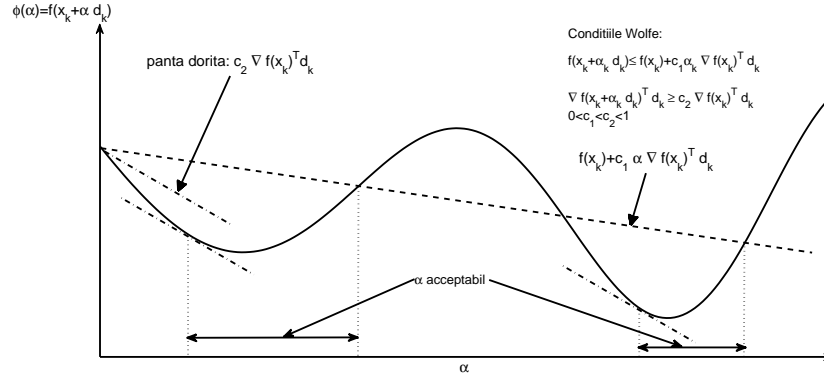


Figura 5.2: Condițiile Wolfe.

adecvat astfel încât să nu fie prea scurt, condiția (W1) este suficientă. De aceea, definim o a treia posibilitate de căutare a pasului α_k , mai puțin costisitoare decât primele două metode prezentate anterior, care se bazează pe *backtracking*:

0. se alege $\alpha > 0$ și $\rho, c_1 \in (0, 1)$

1. cât timp

$$f(x_k + \alpha d_k) > f(x_k) + c_1 \alpha \nabla f(x_k)^T d_k$$

se actualizează $\alpha = \rho \alpha$

2. ieșirea: $\alpha_k = \alpha$.

În general, se consideră valoarea inițială $\alpha = 1$, dar în alte cazuri această valoare trebuie aleasă cu grijă. Se observă că prin tehnica *backtracking* putem găsi α_k într-un număr finit de pași. Mai mult, α_k găsit prin această metodă nu este prea mic întrucât α_k are o valoare apropiată de $\frac{\alpha_k}{\rho}$, valoare respinsă la iterația precedentă datorită faptului că inegalitatea (W1) nu avea loc deoarece pasul era prea lung.

5.3.2 Convergența metodelor de descreștere

În această secțiune analizăm convergența globală a metodelor de descreștere.

Teorema 5.3.1 (Convergența metodelor de descreștere)

Fie problema de optimizare fără constrângeri $\min_{x \in \mathbb{R}^n} f(x)$, unde $f \in \mathcal{C}^1$ este funcție mărginită inferior și gradientul ∇f este Lipschitz continuu. Considerăm metoda iterativă $x_{k+1} = x_k + \alpha_k d_k$, unde d_k este o direcție de descreștere pentru orice $k \geq 0$ și pasul α_k este ales astfel încât cele două condiții Wolfe (W1)-(W2) sunt satisfăcute. Atunci

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \|\nabla f(x_k)\|^2 < \infty,$$

unde θ_k este unghiul făcut de direcția d_k cu gradientul $\nabla f(x_k)$.

Demonstrație: Din condiția Wolfe (W2) avem:

$$(\nabla f(x_{k+1}) - \nabla f(x_k))^T d_k \geq (c_2 - 1) \nabla f(x_k)^T d_k.$$

Utilizând inegalitatea Cauchy-Schwartz obținem:

$$\|\nabla f(x_{k+1}) - \nabla f(x_k)\| \|d_k\| \geq (c_2 - 1) \nabla f(x_k)^T d_k.$$

Mai departe, din proprietatea de Lipschitz a gradientului avem că există constanta Lipschitz $L > 0$ astfel încât are loc următoarea inegalitate:

$$\|\nabla f(x_{k+1}) - \nabla f(x_k)\| \leq L \|x_{k+1} - x_k\| = L \alpha_k \|d_k\|$$

care, înlocuită în relația anterioară conduce la

$$L \alpha_k \|d_k\|^2 \geq (c_2 - 1) \nabla f(x_k)^T d_k$$

adică:

$$\alpha_k \geq \frac{c_2 - 1}{L} \cdot \frac{\nabla f(x_k)^T d_k}{\|d_k\|^2}.$$

Pe de altă parte, din condiția Wolfe (W1) avem:

$$f(x_{k+1}) \leq f(x_k) + c_1 \frac{(\nabla f(x_k)^T d_k)^2}{\|d_k\|^2} \cdot \frac{c_2 - 1}{L}$$

ce conduce la:

$$f(x_{k+1}) \leq f(x_k) - c_1 \frac{1 - c_2}{L} \cdot \frac{(\nabla f(x_k)^T d_k)^2 \|\nabla f(x_k)\|^2}{\|d_k\|^2 \|\nabla f(x_k)\|^2}.$$

În concluzie, notând $c = c_1 \frac{1-c_2}{L}$ obținem:

$$f(x_{k+1}) \leq f(x_k) - c \cos^2 \theta_k \|\nabla f(x_k)\|^2$$

și deci însumând aceste inegalități de la $k = 0, \dots, N-1$ avem:

$$f(x_N) \leq f(x_0) - c \sum_{j=0}^{N-1} \cos^2 \theta_j \|\nabla f(x_j)\|^2.$$

Întrucât f este marginită inferior, pentru $N \rightarrow \infty$

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \|\nabla f(x_k)\|^2 < \infty.$$

Se observă de asemenea că șirul $\cos^2 \theta_k \|\nabla f(x_k)\|^2 \rightarrow 0$. □

Dacă în metoda direcțiilor de descreștere alegem direcția d_k astfel încât $\theta_k \in [\frac{\pi}{2} + \delta, \frac{3\pi}{2} - \delta]$, cu $\delta > 0$ pentru orice $k \geq 0$, atunci $\cos^2 \theta_k \neq 0$ și deci $\|\nabla f(x_k)\| \rightarrow 0$, adică șirul x_k converge la un punct staționar al problemei de optimizare (UNLP).

Capitolul 6

Metode de ordinul I pentru (UNLP)

În acest capitol prezentăm metodele numerice de optimizare de ordinul întâi (i.e. metode bazate pe informația provenită din evaluarea funcției și a gradientului său) pentru rezolvarea problemei neconstrânse de optimizare:

$$(UNLP) : \quad f^* = \min_{x \in \mathbb{R}^n} f(x),$$

unde presupunem că funcția obiectiv $f \in \mathcal{C}^1$. În particular, ne concentrăm pe două metode clasice: metoda gradient și metoda direcțiilor conjugate. În general, orice metodă de minimizare a funcțiilor diferențiale își are originea în metoda gradient. Aceasta are caracteristici care sunt de dorit în cadrul oricărui algoritm de optimizare, cum ar fi simplitatea ei și memoria utilizată foarte redusă (această metodă presupune o singură evaluare a gradientului funcției la fiecare iterație și constă doar în operații cu vectori). Din aceste considerente, de cele mai multe ori noii algoritmi dezvoltați încearcă să modifice această metodă în așa fel încât să posede rate de convergență superioare. De aceea, prezentarea și studiul metodei gradient constituie o cale ideală de ilustrare a metodelor moderne de minimizare fără restricții. O altă metodă importantă care folosește numai informația de gradient este metoda direcțiilor conjugate. Această metodă este de asemenea foarte simplă, bazându-se pe modificarea (devierea) direcției antigradientului cu direcția precedentă și, totodată, cere memorie redusă (de exemplu, în anumite implementări discutate în acest capitol este nevoie doar de memorarea a trei vectori).

6.1 Metoda gradient

Metoda gradient este una din cele mai vechi și mai cunoscute metode iterative în optimizare, fiind propusă pentru prima dată de Cauchy în 1847. Metoda gradient mai este cunoscută și sub numele de *metoda celei mai abrupte descreșteri*. Ea este foarte importantă din punct de vedere teoretic, deoarece este una din cele mai simple metode pentru care există o analiză satisfăcătoare cu privire la convergență.

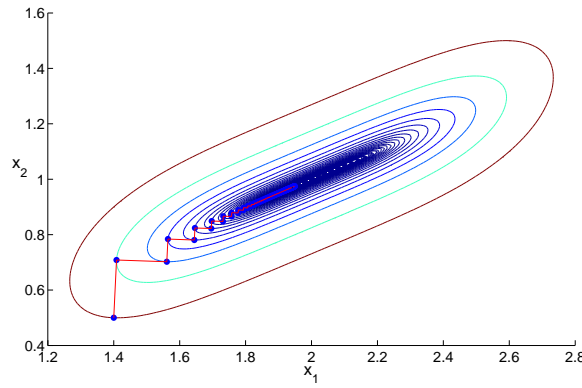


Figura 6.1: Metoda gradient aplicată funcției $f(x_1, x_2) = (x_1 - 2)^4 + (x_1 - 2x_2)^2$ cu alegerea pasului prin metoda ideală.

Metoda gradient se bazează pe următoarea iterație:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),$$

unde lungimea pasului $\alpha_k \geq 0$ se poate alege în funcție de una dintre cele trei proceduri prezentate în capitolul precedent: cea ideală, condițiile Wolfe sau backtracking (vezi Fig. 6.1 și 6.2). Cu alte cuvinte, din punctul x_k căutăm de-a lungul direcției opuse gradientului un punct de minim, iar acest punct de minim este x_{k+1} .

Metoda gradient are diferite interpretări pe care le enumerăm în continuare:

1. direcția în metoda gradient (numită adesea și antigradientul) $d = -\nabla f(x)$ este o direcție de descreștere întrucât expresia $\nabla f(x)^T d = -\|\nabla f(x)\|^2 < 0$ pentru orice x care nu este punct staționar, i.e. orice punct ce satisface $\nabla f(x) \neq 0$;

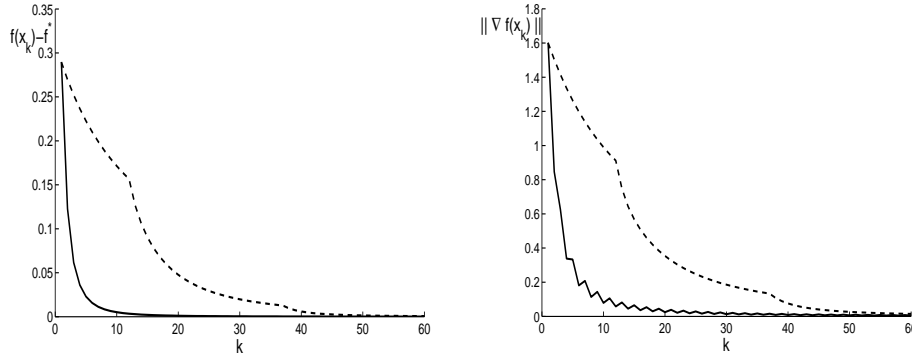


Figura 6.2: Metoda gradient aplicată funcției

$f(x_1, x_2) = (x_1 - 2)^4 + (x_1 - 2x_2)^2$ cu alegerea pasului prin metoda ideală (linie continuă) și backtracking (linie punctată). Evoluția de-a lungul iterațiilor a lui $f(x_k) - f^*$ (stânga) și $\|\nabla f(x_k)\|$ (dreapta).

2. iterația x_{k+1} se obține prin rezolvarea următoarei probleme pătratice (QP) convexe:

$$x_{k+1} = \arg \min_{y \in \mathbb{R}^n} f(x_k) + \nabla f(x_k)^T (y - x_k) + \frac{1}{2\alpha_k} \|y - x_k\|^2,$$

adică aproximăm local funcția obiectiv f în jurul lui x_k printr-un model pătratic cu Hessiana $\nabla^2 f(x) = \frac{1}{\alpha_k} I_n$ și apoi următoarea iterație este dată de punctul optim al aproximării pătratice (vezi Fig. 6.3);

3. metoda gradient prezintă cea mai rapidă descreștere locală, motiv pentru care aceasta se mai numește și *metoda celei mai abrupte pante*: într-adevăr pentru orice direcție d cu $\|d\| = 1$ avem

$$f(x + \alpha d) = f(x) + \alpha \nabla f(x)^T d + \mathcal{R}(\alpha).$$

Din inegalitatea Cauchy-Schwartz obținem:

$$\nabla f(x)^T d \geq -\|\nabla f(x)\| \|d\| = -\|\nabla f(x)\|$$

ceea ce conduce la următoarea inegalitate:

$$f(x + \alpha d) \geq f(x) - \alpha \|\nabla f(x)\| + \mathcal{R}(\alpha).$$

Pe de altă parte, considerând următoarea direcție particulară $\bar{d} = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$ obținem:

$$f(x + \alpha \bar{d}) = f(x) - \alpha \|\nabla f(x)\| + \mathcal{R}(\alpha).$$

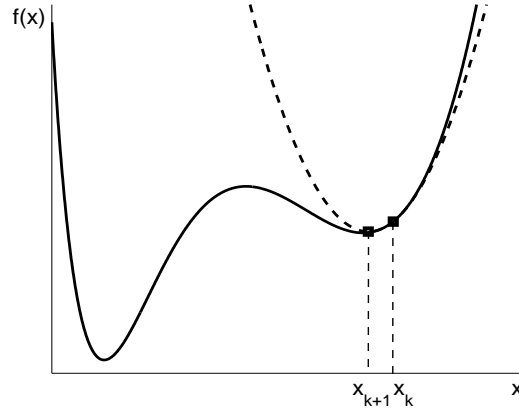


Figura 6.3: Iterația metodei gradient folosind aproximarea pătratică în x_k pentru funcția $f(x) = x^3 - x^2 - 6x + \exp(-x)/2$.

Ultimele două relații ne permit să concluzionăm că cea mai mare descreștere se obține pentru direcția antigradient \bar{d} .

6.1.1 Convergența globală a metodei gradient

În cele ce urmează analizăm proprietățile de convergență globală și locală a metodei gradient. Mai întâi prezentăm un rezultat general de convergență globală pentru metoda gradient aplicată unei probleme (UNLP) pentru care funcția obiectiv trebuie să fie doar de clasă \mathcal{C}^1 .

Teorema 6.1.1 *Dacă următoarele condiții sunt satisfăcute:*

- (i) f este diferențiabilă cu ∇f continuu (i.e. $f \in \mathcal{C}^1$);
- (ii) mulțimea subnivel $S_{f(x_0)} = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ este compactă pentru orice punct inițial x_0 ;
- (iii) lungimea pasului α_k satisface prima condiție Wolfe (W1).

Atunci orice punct limită al șirului x_k generat de metoda gradient este punct staționar pentru problema (UNLP).

Demonstrație: Demonstrația se bazează pe teorema de convergență generală prezentată în capitolul precedent (vezi Teorema 5.2.1). Definim aplicația:

$$M(x) = x - \alpha \nabla f(x).$$

Întrucât funcția obiectiv f este diferențiabilă cu gradientul ∇f continuu rezultă că $M(x)$ este o aplicație continuă punct-punct și deci închisă. Definim $S = \{x^* \in \mathbb{R}^n : \nabla f(x^*) = 0\}$, mulțimea soluțiilor (adică mulțimea punctelor staționare). Mai mult, șirul $\{x_k\}_{k \geq 0} \subseteq S_f(x_0)$, adică șirul generat de metoda gradient este inclus într-o mulțime compactă. De asemenea, definim $\phi = f$ o funcție descrescătoare întrucât prima condiție Wolfe este satisfăcută, ceea ce implică faptul că funcția obiectiv descrește strict de-a lungul iterațiilor generate de metoda gradient. În concluzie, teorema de convergență generală poate fi aplicată și deci orice punct limită al șirului se va regăsi în S . Mai mult, se observă că din condiția ca șirul x_k să fie mărginit, rezultă că există cel puțin un subșir convergent. \square

Acum prezentăm o analiză a convergenței metodei gradient pentru funcții obiectiv f ce posedă în plus față de ipotezele teoremei precedente, proprietatea că gradientul ∇f este Lipschitz continuu.

Teorema 6.1.2 *Fie f o funcție diferențiabilă cu gradientul Lipschitz (constantă Lipschitz $L > 0$) și mărginită inferior. Mai mult, lungimea pasului α_k se alege pentru a satisface cele două condiții Wolfe. Atunci șirul x_k generat de metoda gradient satisface proprietatea: $\lim_{k \rightarrow \infty} \nabla f(x_k) = 0$.*

Demonstrație: Se observă că în acest caz particular unghiul dintre gradient și direcția considerată în metoda gradient (antigradientul) este

$$\theta_k = \pi.$$

În concluzie, din teorema de convergență pentru metodele de descreștere (vezi Teorema 5.3.1) avem:

$$\sum_{k \geq 0} \cos^2 \theta_k \|\nabla f(x_k)\|^2 = \sum_{k \geq 0} \|\nabla f(x_k)\|^2 < \infty.$$

Rezultă că șirul x_k satisface proprietatea: $\nabla f(x_k) \rightarrow 0$ când $k \rightarrow \infty$. \square

Remarca 6.1.1 *Remarcăm faptul că din prima teoremă de convergență a metodei gradient (vezi Teorema 6.1.1) am obținut că un subșir al șirului x_k converge la punctul staționar x^* , în timp ce din a doua teoremă (vezi Teorema 6.1.2) avem rezultatul mai conservativ că $\nabla f(x_k) \rightarrow 0$.*

6.1.2 Rata de convergență globală a metodei gradient

În cazul în care lungimea pasului este constantă pentru toate iterațiile, adică alegem un α astfel încât $x_{k+1} = x_k - \alpha \nabla f(x_k)$, suntem interesați în aflarea unui α optim ce garantează cea mai rapidă convergență. Presupunem că funcția obiectiv are gradientul ∇f Lipschitz cu constanta Lipschitz $L > 0$. Avem atunci următoarea relație (vezi Apendice):

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|x - y\|^2 \quad \forall x, y \in \text{dom} f.$$

Mai departe, aplicând această inegalitate iterației de gradient, i.e. $y = x_{k+1}$, rezultă:

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \alpha \|\nabla f(x_k)\|^2 + \frac{L}{2}\alpha^2 \|\nabla f(x_k)\|^2 \\ &= f(x_k) - \alpha \left(1 - \frac{L}{2}\alpha\right) \|\nabla f(x_k)\|^2. \end{aligned}$$

Lungimea pasului ce garantează cea mai mare descreștere per iterație se obține din condiția:

$$\max_{\alpha > 0} \alpha \left(1 - \frac{L}{2}\alpha\right)$$

adică:

$$\alpha^* = \frac{1}{L}.$$

Metodei gradient cu pas constant îi corespunde o lungime optimală a pasului dată de $\alpha = \frac{1}{L}$. În acest caz, descreșterea la fiecare pas este ilustrată de relația:

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2,$$

iar dacă însumăm aceste inegalități de la $k = 0$ la $k = N - 1$ obținem:

$$f(x_N) \leq f(x_0) - \frac{1}{2L} \sum_{k=0}^{N-1} \|\nabla f(x_k)\|^2$$

adică:

$$\frac{1}{2L} \sum_{k=0}^{N-1} \|\nabla f(x_k)\|^2 \leq f(x_0) - f(x_N) \leq f(x_0) - f^*. \quad (6.1)$$

În continuare definim:

$$\|\nabla f_N\| = \arg \min_{k=0, \dots, N-1} \|\nabla f(x_k)\|.$$

Din inegalitatea (6.1) și inegalitatea precedentă rezultă:

$$\frac{1}{2L} N \|\nabla f_N\|^2 \leq f(x_0) - f^*.$$

În concluzie, după $N = k$ pași se obține următoarea rată de convergență:

$$\|\nabla f_k\| \leq \frac{1}{\sqrt{k}} \sqrt{2L(f(x_0) - f^*)},$$

adică metoda gradient are, în acest caz, o rată de convergență *subliniară*.
□

Din demonstrația teoremei se observă că orice pas α pentru metoda gradient în intervalul

$$\alpha \in \left(0, \frac{2}{L}\right)$$

asigură descreșterea funcției obiectiv și, în consecință, o rată de convergență subliniară. Mai mult, pentru $N \rightarrow \infty$ în inegalitatea (6.1) obținem că $\|\nabla f(x_k)\| \rightarrow 0$ când $k \rightarrow \infty$.

Observăm că nu se poate spune nimic în acest caz despre convergența șirului x_k la punctul staționar x^* sau al lui $f(x_k)$ la valoarea optimă f^* . Acest tip de convergență poate fi derivată în cazul convex.

Mai departe considerăm problema de optimizare convexă neconstrânsă $\min_{x \in \mathbb{R}^n} f(x)$, unde funcția obiectiv $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ ($\mathcal{F}_L^{1,1}(\mathbb{R}^n)$ reprezintă clasa de funcții diferențiabile, convexe, cu gradient Lipschitz de constantă L) pentru care avem inegalitatea (vezi Apendice):

$$\frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle \quad \forall x, y \in \text{dom} f.$$

De asemenea, pentru un punct de optim global x^* notăm $R_k = \|x_k - x^*\|$ distanța de la punctul x_k la punctul de optim x^* . Atunci, din relația precedentă și $\nabla f(x^*) = 0$ obținem:

$$\begin{aligned} R_{k+1}^2 &= \|x_k - x^* - \alpha \nabla f(x_k)\|^2 \\ &= R_k^2 - 2\alpha \langle \nabla f(x_k), x_k - x^* \rangle + \alpha^2 \|\nabla f(x_k)\|^2 \\ &\leq R_k^2 - \alpha \left(\frac{2}{L} - \alpha\right) \|\nabla f(x_k)\|^2. \end{aligned}$$

Observăm că pentru pas constant $\alpha \in (0, \frac{2}{L})$ avem $R_k \leq R_0$ pentru orice $k \geq 0$. Notând $\Delta_k = f(x_k) - f^*$, din convexitatea lui f și inegalitatea Cauchy-Schwartz, avem:

$$\Delta_k \leq \langle \nabla f(x_k), x_k - x^* \rangle \leq R_k \|\nabla f(x_k)\| \leq R_0 \|\nabla f(x_k)\|. \quad (6.2)$$

Din proprietatea Lipschitz avem:

$$f(x_{k+1}) \leq f(x_k) - \alpha(1 - \frac{L}{2}\alpha) \|\nabla f(x_k)\|^2,$$

de unde scăzând în ambele părți f^* și combinând cu inegalitatea (6.2) obținem:

$$\Delta_{k+1} \leq \Delta_k - \frac{\omega}{R_0^2} \Delta_k^2,$$

unde $\omega = \alpha(1 - \frac{L}{2}\alpha)$. Deci,

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\omega}{R_0^2} \frac{\Delta_k}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\omega}{R_0^2}.$$

Prin însumarea acestor inegalități de la $k = 0$ la $k = N - 1$ rezultă următoarea inegalitate:

$$\frac{1}{\Delta_N} \geq \frac{1}{\Delta_0} + \frac{\omega}{R_0^2} N.$$

Mai departe, dacă alegem $\alpha = \alpha^* = \frac{1}{L}$ obținem următoarea rată de convergență:

$$f(x_N) - f^* \leq \frac{2L(f(x_0) - f^*) \|x_0 - x^*\|^2}{2L \|x_0 - x^*\|^2 + N(f(x_0) - f^*)}.$$

Din proprietatea de Lipschitz avem:

$$\begin{aligned} f(x_0) &\leq f^* + \langle \nabla f(x^*), x_0 - x^* \rangle + \frac{L}{2} \|x_0 - x^*\|^2 \\ &= f^* + \frac{L}{2} \|x_0 - x^*\|^2. \end{aligned}$$

Obținem atunci (înlocuind $N = k$) următoarea rată de convergență subliniară a metodei gradient pentru probleme de optimizare convexe neconstrânse:

$$f(x_k) - f^* \leq \frac{2L \|x_0 - x^*\|^2}{k + 4}.$$

6.1.3 Rata de convergență locală a metodei gradient

În acest subcapitol analizăm rata de convergență locală a metodei gradient. Pentru simplitatea expunerii studiem mai întâi cazul problemelor pătratice:

$$f^* = \min_{x \in \mathbb{R}^n} f(x) \quad \left(= \frac{1}{2} x^T Q x - q^T x \right),$$

unde Q este matrice simetrică pozitiv definită. În acest caz problema de optimizare pătratică convexă precedentă are un singur punct de minim global x^* ce satisface relația $Qx^* - q = 0$. De asemenea, dând x_k la iterația k , definim reziduul $r_k = Qx_k - q = \nabla f(x_k)$. Atunci, pasul optim (obținut prin metoda ideală de alegere a pasului) se obține explicit din minimizarea funcției pătratice unidimensionale în $\alpha \in (0, \infty)$: $\phi(\alpha) = f(x_k - \alpha r_k)$:

$$\alpha_k = \frac{r_k^T r_k}{r_k^T Q r_k}.$$

Astfel, metoda gradient are următoarea iterație pentru cazul pătratic convex:

$$x_{k+1} = x_k - \frac{r_k^T r_k}{r_k^T Q r_k} \nabla f(x_k).$$

Ca să evaluăm rata de convergență, introducem următoarea funcție ce măsoară eroarea:

$$e(x) = \frac{1}{2} (x - x^*)^T Q (x - x^*) = f(x) - f^*,$$

unde am folosit ca $Qx^* - q = 0$ și $f^* = f(x^*)$. Observăm că eroarea $e(x)$ este zero dacă și numai dacă $x = x^*$. Prin calcule simple se poate arăta că

$$\frac{e(x_k) - e(x_{k+1})}{e(x_k)} = \frac{2\alpha_k r_k^T Q y_k - \alpha_k^2 r_k^T Q r_k}{y_k^T Q y_k},$$

unde $y_k = x_k - x^*$. Ținând cont că $Qy_k = r_k$, obținem:

$$e(x_{k+1}) = \left(1 - \frac{(r_k^T r_k)^2}{(r_k^T Q r_k)(r_k^T Q^{-1} r_k)} \right) e(x_k).$$

Putem arăta ușor utilizând inegalitatea lui Kantorovich:

$$\min_{y \neq 0} \frac{(y^T y)^2}{(y^T Q y)(y^T Q^{-1} y)} = \frac{4\lambda_{\min}\lambda_{\max}}{(\lambda_{\min} + \lambda_{\max})^2}$$

că următoarea relație are loc:

$$e(x_{k+1}) \leq \frac{(\kappa - 1)^2}{(\kappa + 1)^2} e(x_k),$$

unde κ este numărul de condiționare al matricei Q , adică $\kappa = \lambda_{\max}/\lambda_{\min}$, cu $\lambda_{\min} > 0$ valoarea proprie minimă, iar λ_{\max} valoarea proprie maximă a matricei pozitiv definite Q . Din această relație rezultă că $e(x_k) = f(x_k) - f^* \rightarrow 0$ cu o rată liniară mărginită de constanta $\beta = (\kappa - 1)^2/(\kappa + 1)^2$. Observăm că rata de convergență este lentă dacă numărul de condiționare κ este mare și depinde de punctul de pornire x_0 . Acest rezultat pentru probleme QP strict convexe, unde pasul se alege cu metoda ideală, ne arată că valorile funcției $f(x_k)$ converg la valoarea optimă f^* cu o rată liniară.

Pentru funcții diferențiabile nepătratice rezultate similare au loc:

Teorema 6.1.3 *Presupunem că funcția $f \in \mathcal{C}^2$ și că iterațiile metodei gradient generate cu procedura de căutare ideală a pasului converge la un punct x^* pentru care $\nabla^2 f(x^*)$ este matrice pozitiv definită. Fie un scalar $\beta \in ((\kappa - 1)^2/(\kappa + 1)^2 - 1)$, unde κ este numărul de condiționare al matricei $\nabla^2 f(x^*)$. Atunci, pentru k suficient de mare avem că:*

$$f(x_{k+1}) - f(x^*) \leq \beta(f(x_k) - f(x^*)).$$

Demonstrația se bazează pe folosirea Hessienei funcției obiectiv în punctul x^* în locul matricei Q corespunzătoare cazului pătratic. În acest caz, dacă x^* este un punct de minim local astfel încât Hessiana $\nabla^2 f(x^*)$ este pozitiv definită cu un număr de condiționare κ , putem arăta că șirul x_k convergent la x^* produs de metoda gradient satisface următoarea proprietate: $f(x_k) \rightarrow f(x^*)$ cu o rată de convergență liniară al cărei coeficient este mărginit inferior de $(\kappa - 1)^2/(\kappa + 1)^2$. Observăm că în cazul neconvex șirul x_k generat de metoda gradient converge către x^* dacă punctul de pornire x_0 este suficient de aproape de x^* , adică avem convergență locală.

6.2 Metoda direcțiilor conjugate

Metoda direcțiilor conjugate poate fi privită ca o metodă intermediară între metoda gradient (ce folosește informație de ordinul întâi) și metoda Newton (ce folosește informație de ordinul doi). Această metodă este

motivată de dorința de a accelera rata de convergență lentă a metodei gradient și în același timp de a evita folosirea Hessienei precum se întâmplă în metoda Newton. Un caz particular al metodei direcțiilor conjugate este metoda gradientilor conjugăți, care a fost inițial dezvoltată pentru probleme pătratice. Această tehnică a fost extinsă la probleme de optimizare generale, prin aproximare, deoarece se poate argumenta că în apropierea unui punct de minim local funcția obiectiv este aproximativ pătratică.

6.2.1 Metoda direcțiilor conjugate pentru QP

Metoda direcțiilor conjugate este de asemenea o metodă de ordinul întâi, adică folosește informația extrasă din valoarea funcției și a gradientului acesteia (oracol de ordinul întâi), însă prezintă o rată de convergență mai bună decât a metodei gradient cel puțin pentru cazul pătratic. Să presupunem următoarea problemă pătratică (QP) strict convexă:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T Q x - q^T x,$$

unde $Q \succ 0$ (i.e. matrice pozitiv definită). Soluția optimă a acestei probleme de optimizare este echivalentă cu rezolvarea următorului sistem de ecuații liniare

$$Qx = q.$$

Întrucât Q este inversabilă, soluția problemei de optimizare sau soluția sistemului liniar este $x^* = Q^{-1}q$. În cele mai multe cazuri, calculul inversei este foarte costisitor, și, de obicei, complexitatea aritmetică a unei metode de calcul numeric matriceal pentru găsirea soluției este de ordinul $\mathcal{O}(n^3)$. În cele ce urmează vom prezenta o metodă numerică de optimizare mai simplă și de obicei mai puțin costisitoare numeric pentru calculul soluției x^* .

Definiția 6.2.1 *Doi vectori d_1 și d_2 se numesc Q -ortogonali dacă satisfac condiția $d_1^T Q d_2 = 0$. O mulțime de vectori $\{d_1, d_2, \dots, d_k\}$ se numește Q -ortogonală dacă $d_i^T Q d_j = 0$ pentru orice $i \neq j$.*

Se observă că dacă matricea $Q \succ 0$ și dacă mulțimea $\{d_1, d_2, \dots, d_k\}$ este Q -ortogonală, iar vectorii sunt nenuli, atunci acești vectori sunt liniar independenți. Mai mult, în cazul în care $k = n$, vectorii formează o bază pentru \mathbb{R}^n . În concluzie, dacă $\{d_1, d_2, \dots, d_n\}$ este Q -ortogonală,

iar vectorii sunt nenuli, există $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ astfel încât $x^* = \alpha_1 d_1 + \alpha_2 d_2 + \dots + \alpha_n d_n$ (adică x^* este combinație liniară a vectorilor bazei). Pentru aflarea parametrilor α_i avem relația:

$$\alpha_i = \frac{d_i^T Q x^*}{d_i^T Q d_i} = \frac{d_i^T q}{d_i^T Q d_i}.$$

Concluzionăm că:

$$x^* = \sum_{i=1}^n \frac{d_i^T q}{d_i^T Q d_i} \cdot d_i$$

și deci x^* poate fi obținut printr-un proces iterativ în care la pasul i adăugăm termenul $\alpha_i d_i$.

Teorema 6.2.1 *Fie $\{d_0, d_1, \dots, d_{n-1}\}$ o mulțime Q -ortogonală de vectori cu elemente nenule. Pentru orice $x_0 \in \mathbb{R}^n$ șirul x_k generat de metoda iterativă:*

$$\begin{aligned} x_{k+1} &= x_k + \alpha_k d_k \\ \alpha_k &= -\frac{r_k^T d_k}{d_k^T Q d_k}, \quad r_k = Qx_k - q \end{aligned}$$

converge la x^* după n pași, adică $x_n = x^*$.

Demonstrație: Deoarece vectorii $\{d_0, d_2, \dots, d_{n-1}\}$ sunt liniar independenți putem scrie:

$$x^* - x_0 = \alpha'_0 d_0 + \alpha'_1 d_1 + \dots + \alpha'_{n-1} d_{n-1}$$

pentru anumiți scalari α'_k . Multiplicând această relație cu Q și apoi luând produsul scalar cu d_k obținem:

$$\alpha'_k = \frac{d_k^T Q (x^* - x_0)}{d_k^T Q d_k}.$$

Folosind iterația precedentă pentru calcularea lui x_{k+1} obținem:

$$x_k - x_0 = \alpha_0 d_0 + \alpha_1 d_1 + \dots + \alpha_{k-1} d_{k-1}$$

și din Q -ortogonalitatea lui d_k derivăm următoarea relație:

$$d_k^T Q (x_k - x_0) = 0.$$

În concluzie,

$$\alpha'_k = \frac{d_k^T Q(x^* - x_k)}{d_k^T Q d_k} = -\frac{r_k^T d_k}{d_k^T Q d_k}$$

coincide cu α_k . De aici rezultă că $x_n - x_0 = x^* - x_0$ și deci $x_n = x^*$. \square

Se observă că reziduul $r_k = Qx_k - q$ coincide cu gradientul funcției obiectiv pătratică. Folosind argumente similare teoremei precedente se poate arăta următorul rezultat:

Teorema 6.2.2 *Fie $\{d_0, d_1, \dots, d_{n-1}\}$ o mulțime Q -ortogonală de vectori cu elemente nenule, definim subspațiul $S_k = \text{Span}\{d_0, d_1, \dots, d_k\}$. Atunci pentru orice $x_0 \in \mathbb{R}^n$ șirul $x_{k+1} = x_k + \alpha_k d_k$, unde $\alpha_k = -\frac{r_k^T d_k}{d_k^T Q d_k}$ are următoarele proprietăți:*

$$(i) \ x_{k+1} = \arg \min_{x \in x_0 + S_k} \frac{1}{2} x^T Q x - q^T x;$$

(ii) reziduul la pasul k este ortogonal cu toate direcțiile precedente, adică:

$$r_k^T d_i = 0 \quad \forall i < k.$$

Din teorema precedentă, proprietatea (ii), avem că gradientul este ortogonal pe subspațiul S_{k-1} :

$$\nabla f(x_k) \perp S_{k-1}.$$

Această teoremă se mai numește și *minimizarea peste subspațiul de extindere*.

6.2.2 Metoda gradientilor conjugați pentru QP

Metoda gradientilor conjugați aparține clasei de metode de direcții conjugate cu o proprietate specială: în generarea mulțimii de vectori conjugați, noul vector d_k poate fi calculat folosind numai direcția anterioară d_{k-1} . În această metodă, pentru a calcula d_k nu trebuie să știm toate direcțiile conjugate anterioare d_0, d_1, \dots, d_{k-1} . Din construcție, noua direcție d_k va fi automat ortogonală pe aceste direcții anterioare. Această proprietate a metodei gradientilor conjugați este foarte importantă, deoarece necesită puțină memorie și calcule. Metoda gradientilor conjugați pentru rezolvarea unui QP strict convex cuprinde următorii pași:

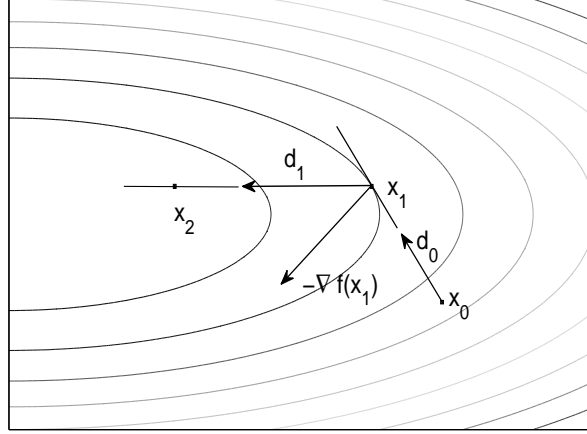


Figura 6.4: Metoda gradientilor conjugați.

1. Fie vectorul $x_0 \in \mathbb{R}^n$ dat, definim $d_0 = -\nabla f(x_0) = -r_0 = -(Qx_0 - q)$;
2. actualizăm iterația $x_{k+1} = x_k + \alpha_k d_k$, unde $\alpha_k = -\frac{r_k^T d_k}{d_k^T Q d_k}$;
3. actualizăm direcția $d_{k+1} = -r_{k+1} + \beta_k d_k$, unde $\beta_k = \frac{r_{k+1}^T Q d_k}{d_k^T Q d_k}$.

Am utilizat notația $r_k = \nabla f(x_k)$. Se observă că la fiecare pas o nouă direcție este aleasă ca o combinație liniară între gradientul curent și direcția precedentă. Metoda gradientilor conjugați are o complexitate scăzută întrucât folosește formule de actualizare simple (adică operații cu vectori).

Teorema 6.2.3 (Proprietăți ale metodei gradientilor conjugați)

Metoda gradientilor conjugați satisface următoarele proprietăți:

- (i) $\text{Span}\{d_0, \dots, d_k\} = \text{Span}\{r_0, \dots, r_k\} = \text{Span}\{r_0, Qr_0, \dots, Q^k r_0\}$;
- (ii) $d_k^T Q d_i = 0$ pentru orice $i < k$;
- (iii) $\alpha_k = \frac{r_k^T r_k}{d_k^T Q d_k}$;
- (iv) $\beta_k = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$.

Demonstrație: Relația (i) se poate demonstra prin inducție. Este evident că (i) este adevărată la pasul $k = 0$. Presupunem acum că egalitățile de mulțimi sunt valide la pasul k și demonstrăm relația pentru $k + 1$. Din iterația metodei gradientilor conjugați se observă:

$$r_{k+1} = r_k + \alpha_k Qd_k.$$

Din ipoteza de inducție avem că $r_k, Qd_k \in \text{Span}\{r_0, Qr_0, \dots, Q^{k+1}r_0\}$. Drept urmare, $r_{k+1} \in \text{Span}\{r_0, Qr_0, \dots, Q^{k+1}r_0\}$. Considerând acest rezultat și iterația:

$$d_{k+1} = -r_{k+1} + \beta_k d_k,$$

rezultă de asemenea că $d_{k+1} \in \text{Span}\{r_0, Qr_0, \dots, Q^{k+1}r_0\}$.

Pentru a demonstra (ii) folosim iarăși inducția și luăm în considerare faptul că:

$$d_{k+1}^T Qd_i = -r_{k+1}^T Qd_i + \beta_k d_k^T Qd_i.$$

Pentru $i = k$, membrul drept este zero datorită definiției lui β_k . Pentru $i < k$, ambii termeni dispar. Primul termen dispare datorită faptului că $Qd_i \in \text{Span}\{d_0, d_1, \dots, d_{i+1}\}$, a inducției ce garantează că metoda este o metodă de direcții conjugate până la pasul x_{k+1} , și datorită Teoremei 6.2.2, ce garantează că r_{k+1} este ortogonal pe $\text{Span}\{d_0, d_1, \dots, d_{i+1}\}$. Al doilea termen dispare prin aplicarea inducției asupra relației (ii).

Pentru a demonstra (iii) observăm că:

$$-r_k^T d_k = r_k^T r_k - \beta_{k-1} r_k^T d_{k-1}$$

și apoi utilizăm relația $\alpha_k = -\frac{r_k^T d_k}{d_k^T Qd_k}$, iar al doilea termen este zero, conform Teoremei 6.2.2.

În final, pentru a demonstra (iv) observăm că $r_{k+1}^T r_k = 0$, deoarece $r_k \in \text{Span}\{d_0, d_1, \dots, d_k\}$ și r_{k+1} este ortogonal pe $\text{Span}\{d_0, d_1, \dots, d_k\}$. Mai mult, din relația

$$Qd_k = \frac{1}{\alpha_k}(r_{k+1} - r_k)$$

avem că $r_{k+1}^T Qd_k = \frac{1}{\alpha_k} r_{k+1}^T r_{k+1}$. □

Din proprietatea (ii) a teoremei precedente se observă că metoda gradientilor conjugați produce direcțiile $\{d_0, d_1, \dots, d_{n-1}\}$, care sunt direcții Q -ortogonale și deci metoda converge la soluția optimă x^* în exact n pași, conform Teoremei 6.2.1.

6.2.3 Metoda gradientilor conjugați pentru UNLP

Pentru o problemă generală neconstrânsă $\min_{x \in \mathbb{R}^n} f(x)$, putem aplica metoda gradientilor conjugați folosind aproximări adecvate. Prezentăm în cele ce urmează câteva abordări în această direcție.

În abordarea aproximării pătratice se repetă aceleași iterații ca și în cazul pătratic, folosindu-se o aproximare pătratică cu următoarele identificări:

$$Q = \nabla^2 f(x_k), \quad r_k = \nabla f(x_k).$$

Aceste asocieri sunt reevaluate la fiecare pas al metodei. Dacă funcția f este pătratică, atunci aceste asocieri sunt identități, și deci algoritmul este o generalizare la cazul pătratic neconvex. Când o aplicăm la probleme nepătratice, atunci metoda gradientilor conjugați nu va produce o soluție în n pași. În acest caz se continuă procedura, găsind noi direcții și terminând atunci când un anumit criteriu este satisfăcut (de exemplu, $\|\nabla f(x_k)\| \leq \epsilon$). Este, de asemenea, posibil ca după n sau $n + 1$ pași să reinițializăm algoritmul cu $x_0 = x_n$ și să începem metoda gradientilor conjugați cu un pas de gradient.

Metoda gradientilor conjugați pentru probleme UNLP generale constă în următorii pași:

1. $r_0 = \nabla f(x_0)$ și $d_0 = -\nabla f(x_0)$;
2. $x_{k+1} = x_k + \alpha_k d_k$ pentru orice $k = 0, 1, \dots, n-1$, unde $\alpha_k = -\frac{r_k^T d_k}{d_k^T \nabla^2 f(x_k) d_k}$;
3. $d_{k+1} = -\nabla f(x_{k+1}) + \beta_k d_k$, unde $\beta_k = \frac{r_{k+1}^T \nabla^2 f(x_k) d_k}{d_k^T \nabla^2 f(x_k) d_k}$;
4. după n iterații înlocuim x_0 cu x_n și repetăm întregul proces.

O proprietate atractivă a metodei gradientilor conjugați este aceea că nu este nevoie de căutarea pe o direcție, adică nu trebuie să găsim o lungime a pasului. Pe de altă parte, această abordare are dezavantajul că necesită evaluarea Hessienei funcției obiectiv la fiecare pas, care de obicei este costisitoare. De asemenea, se observă că în cazul general această metodă nu este convergentă.

Se poate evita însă folosirea directă a Hessienei $\nabla^2 f(x)$. De exemplu, în locul formulei pentru α_k dat mai sus, putem găsi lungimea pasului α_k prin metoda de căutare ideală. Expresia corespunzătoare va coincide cu cea anterioară în cazul pătratic. De asemenea, algoritmul

se poate transforma într-unul convergent prin modificarea adecvată a parametrului β_k . Avem la dispoziție următoarele reguli de actualizare:

$$\begin{aligned} \text{Fletcher-Reeves :} \quad \beta_k &= \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}; \\ \text{Polak-Ribiere :} \quad \beta_k &= \frac{(r_{k+1} - r_k)^T r_{k+1}}{r_k^T r_k}. \end{aligned}$$

Observăm că aceste formule coincid cu β_k dat în algoritmul precedent în cazul pătratic. Din simulări s-a observat că de obicei metoda Polak-Ribiere are un comportament mai bun față de metoda Fletcher-Reeves. Obținem următoarea metodă modificată a gradientilor conjugați pentru probleme UNLP ce constă în următorii pași:

1. $r_0 = \nabla f(x_0)$ și $d_0 = -\nabla f(x_0)$;
2. $x_{k+1} = x_k + \alpha_k d_k$ pentru orice $k = 0, 1, \dots, n-1$, unde α_k minimizează funcția unidimensională $f(x_k + \alpha d_k)$;
3. $d_{k+1} = -\nabla f(x_{k+1}) + \beta_k d_k$, unde β_k este ales cu una din cele două formule anterioare;
4. după n iterații înlocuim x_0 cu x_n și repetăm întregul proces.

Convergența globală a metodei gradientilor conjugați modificată se poate demonstra din simpla observație că la fiecare n pași a acestei metode se realizează o iterație de gradient pur și că la ceilalți pași funcția obiectiv nu crește, ci de fapt se speră să descrească strict. De aceea, repornirea algoritmului este importantă pentru analiza convergenței globale a metodei, deoarece pentru direcțiile d_k produse de metodă nu putem garanta că sunt direcții de descreștere.

Proprietățile de convergență locală a metodei descrise anterior pot fi arătate folosindu-ne iarăși de analiza cazului pătratic. Presupunând că la soluția x^* Hessiana este pozitiv definită, ne așteptăm la o rată de convergență cel puțin la fel de bună ca a metodei gradient. Mai mult, în general metoda converge pătratic în raport cu fiecare ciclu de n pași. Cu alte cuvinte, observăm că în fiecare ciclu se rezolvă o problemă pătratică la fel cum metoda Newton rezolvă într-un pas această problemă pătratică și deci ne așteptăm ca

$$\|x_{k+n} - x^*\| \leq c \|x_k - x^*\|^2$$

pentru o anumită constantă $c > 0$ și $k = 1, n, 2n, \dots$. În concluzie, metoda gradientilor conjugati modificată posedă o convergență superioară metodei gradient, iar pe de altă parte are o implementare simplă. De aceea, este adesea preferată în favoarea metodei gradient pentru rezolvarea problemelor de optimizare fără constrângeri.

Capitolul 7

Metode de ordinul II pentru (UNLP)

Metodele de ordinul doi sunt cele mai complexe metode numerice de optimizare, deoarece folosesc informație despre curbura funcției obiectiv sau matricea Hessiană. De obicei, aceste metode converg mult mai rapid decât metodele de ordinul întâi, dar sunt în general dificil de implementat deoarece calcularea și memorarea matricei Hessiane poate fi costisitoare din punct de vedere numeric. Metoda Newton este un exemplu de metodă de ordinul doi care constă în devierea direcției antigradientului prin premultiplicarea lui cu inversa matricei Hessiane. Această operație este motivată prin găsirea unei direcții adecvate pentru aproximarea Taylor de ordinul doi a funcției obiectiv. În general, pentru probleme de dimensiuni mari, se preferă implementarea unei metode de ordinul întâi care ia în calcul structura funcției obiectiv. Adesea un șir de gradienti pot fi folosiți la aproximarea curburii de ordinul doi a funcției obiectiv. Metode ce se bazează pe această procedură se numesc metode quasi-Newton. În acest capitol ne ocupăm de rezolvarea unei probleme generale neliniare de optimizare neconstrânsă (UNLP) cu metode numerice ce utilizează informație furnizată de gradient și Hessiană (informație de ordin doi) sau o aproximare a acesteia (adică ce se bazează pe informație de ordinul întâi):

$$(UNLP) : \quad f^* = \min_{x \in \mathbb{R}^n} f(x), \quad (7.1)$$

unde funcția obiectiv este de două ori diferențiabilă cu Hessiana continuă (i.e. $f \in \mathcal{C}^2$).

7.1 Metoda Newton

În analiza numerică și optimizare, metoda lui Newton (sau cunoscută și sub numele de metoda Newton-Raphson) este o metodă de calcul al rădăcinilor unui sistem de ecuații. Considerăm sistemul neliniar:

$$F(y) = 0,$$

unde $y \in \mathbb{R}^n$ și $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ o funcție diferențiabilă. Acest sistem poate fi rezolvat prin metoda Newton care constă în liniarizarea ecuației în punctul curent y_k :

$$F(y_k) + \frac{\partial F}{\partial y}(y_k)(y - y_k) = 0. \quad (7.2)$$

Această procedură, ce constă în rezolvarea sistemului liniar în y , conduce la următoarea iterație Newton, sub ipoteza că Jacobianul (notat $\nabla F(y_k) = \frac{\partial F}{\partial y}(y_k)$) este matrice inversabilă:

$$y_{k+1} = y_k - \left(\frac{\partial F}{\partial y}(y_k) \right)^{-1} F(y_k).$$

Considerăm acum condițiile necesare de optimalitate de ordinul întâi pentru probleme UNLP, ce se reduc la un sistem de ecuații neliniare:

$$\nabla f(x) = 0,$$

cu gradientul $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$. Acest sistem are numărul de ecuații egale cu numărul de variabile. Metoda Newton va liniariza sistemul neliniar în punctul x_k pentru a găsi următorul punct x_{k+1} :

$$\nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k) = 0,$$

iar din această relație putem deriva *metoda Newton* care constă în următoarea iterație (vezi Fig. 7.1):

$$x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k).$$

Direcția în metoda Newton este dată de expresia

$$d_k = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k),$$

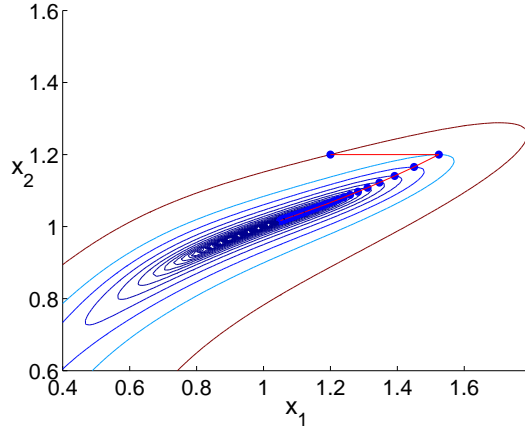


Figura 7.1: Metoda Newton aplicată funcției
 $f(x_1, x_2) = (x_1 - x_2^3)^2 + 3(x_1 - x_2)^4$.

numită și *direcția Newton*. Observăm că dacă $\nabla^2 f(x_k) \succ 0$, atunci direcția Newton d_k este direcție de descreștere. Reamintim că direcția în metoda gradient este dată de antigradientul $-\nabla f(x_k)$, adică în locul matricei $(\nabla^2 f(x_k))^{-1}$ din metoda Newton, în metoda gradient se folosește matricea identitate I_n .

O altă interpretare a metodei numerice de optimizare Newton poate fi obținută din aproximarea Taylor de ordinul doi a funcției obiectiv f . Reamintim condițiile de optimalitate suficiente de ordinul doi ce se definesc astfel: dacă există un x^* ce satisface:

$$\nabla f(x^*) = 0 \quad \text{și} \quad \nabla^2 f(x^*) \succ 0$$

atunci x^* este un minim local. Dacă punctul x^* satisface condițiile precedente, atunci există o vecinătate a lui x^* notată \mathcal{N} astfel încât pentru orice $x \in \mathcal{N}$ avem $\nabla^2 f(x) \succ 0$. Din aproximarea Taylor obținem că (Fig. 7.2):

$$f(x_{k+1}) \approx f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{1}{2} (x_{k+1} - x_k)^T \nabla^2 f(x_k) (x_{k+1} - x_k)$$

și deci iterația Newton este dată de (vezi Fig. 7.2):

$$x_{k+1} = \arg \min_y f(x_k) + \nabla f(x_k)^T (y - x_k) + \frac{1}{2} (y - x_k)^T \nabla^2 f(x_k) (y - x_k).$$

Se observă că dacă x_k este suficient de aproape de x^* atunci $\nabla^2 f(x_k) \succ 0$ și din condițiile de optimalitate corespunzătoare unei probleme QP

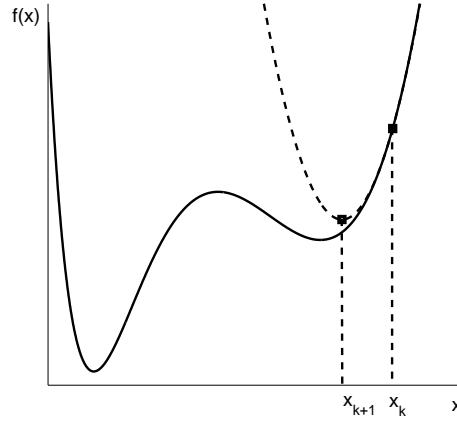


Figura 7.2: Iterația metodei Newton folosind aproximarea pătratică Taylor în x_k pentru funcția $f(x) = x^3 - x^2 - 6x + \exp(-x)/2$.

strict convexe obținem din nou $x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$, adică aceeași formulă, însă cu o interpretare diferită. Făcând analogia cu interpretarea metodei gradient, observăm că în ambele metode iterația x_{k+1} se generează din rezolvarea unei aproximări pătratice în care termenul liniar este același, dar termenul pătratic în metoda gradient este $(y-x_k)^T I_n (y-x_k)$ în timp ce în metoda Newton este $(y-x_k)^T \nabla^2 f(x_k) (y-x_k)$. Este clar că aproximarea pătratică a funcției f folosită în metoda Newton este mai bună decât cea folosită în metoda gradient și deci ne așteptăm ca metoda Newton să performeze mai bine față de metoda gradient.

Putem într-o manieră similară celei anterioare să interpretăm derivarea direcției Newton, și anume din aproximarea Taylor avem:

$$f(x_k + d) \approx f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T \nabla^2 f(x_k) d$$

și deci definim direcția Newton:

$$d_k = \arg \min_d f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T \nabla^2 f(x_k) d.$$

Se observă că dacă $\nabla^2 f(x_k) \succ 0$, din condițiile de optimalitate corespunzătoare unei probleme QP strict convexe obținem din nou direcția Newton $d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$.

7.1.1 Rata de convergență locală a metodei Newton

În acest subcapitol vom analiza convergența locală a metodei Newton în forma standard (adică pasul $\alpha_k = 1$):

$$x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k).$$

Vom arăta în următoarea teoremă că această metodă converge local cu rata pătratică, adică există $\beta > 0$ astfel încât $\|x_{k+1} - x^*\| \leq \beta \|x_k - x^*\|^2$ pentru orice $k \geq 0$, sub ipoteza că x_0 este suficient de aproape de x^* .

Teorema 7.1.1 (Convergența locală pătratică a metodei Newton)

Fie $f \in C^2$ și x^* un minim local ce satisface condițiile suficiente de ordinul II (i.e. $\nabla f(x^*) = 0$ și $\nabla^2 f(x^*) \succ 0$). Fie o constantă $l > 0$ astfel încât:

$$\nabla^2 f(x^*) \succeq lI_n.$$

Mai mult, presupunem că $\nabla^2 f(x)$ este Lipschitz, i.e.

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq M \|x - y\| \quad \forall x, y \in \text{dom} f,$$

unde $M > 0$. Dacă x_0 este suficient de aproape de x^* , adică:

$$\|x_0 - x^*\| \leq \frac{2}{3} \cdot \frac{l}{M},$$

atunci iterația Newton $x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$ are proprietatea că șirul x_k generat converge la x^* cu rata pătratică, adică $\|x_{k+1} - x^*\| \leq \frac{3M}{2l} \|x_k - x^*\|^2$ pentru orice $k \geq 0$.

Demonstrație: Întrucât x^* este un minim local atunci $\nabla f(x^*) = 0$. Mai mult, din teorema lui Taylor în forma integrală avem:

$$\nabla f(x_k) = \nabla f(x^*) + \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) (x_k - x^*) d\tau.$$

Se obține:

$$\begin{aligned} x_{k+1} - x^* &= x_k - x^* - (\nabla^2 f(x_k))^{-1} \nabla f(x_k) \\ &= (\nabla^2 f(x_k))^{-1} [\nabla^2 f(x_k)(x_k - x^*) - \nabla f(x_k) + \nabla f(x^*)] \\ &= (\nabla^2 f(x_k))^{-1} [\nabla^2 f(x_k)(x_k - x^*) - \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) (x_k - x^*) d\tau] \\ &= (\nabla^2 f(x_k))^{-1} \int_0^1 [\nabla^2 f(x_k)(x_k - x^*) - \nabla^2 f(x^* + \tau(x_k - x^*)) (x_k - x^*)] d\tau \\ &= (\nabla^2 f(x_k))^{-1} \int_0^1 [\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*))] (x_k - x^*) d\tau. \end{aligned}$$

Întrucât

$$\|\nabla^2 f(x_k) - \nabla^2 f(x^*)\| \leq M\|x_k - x^*\|$$

rezultă că (vezi Apendice):

$$-M\|x_k - x^*\|I_n \preceq \nabla^2 f(x_k) - \nabla^2 f(x^*) \preceq M\|x_k - x^*\|I_n.$$

Mai mult, vom avea:

$$\nabla^2 f(x_k) \succeq \nabla^2 f(x^*) - M\|x_k - x^*\|I_n \succeq lI_n - M\|x_k - x^*\|I_n \succ 0,$$

sub ipoteza că $\|x_k - x^*\| \leq \frac{2}{3} \frac{l}{M}$, de unde rezultă:

$$0 \prec (\nabla^2 f(x_k))^{-1} \preceq \frac{1}{l - M\|x_k - x^*\|} I_n.$$

Concluzionăm următoarele:

$$\begin{aligned} & \|x_{k+1} - x^*\| \\ &= \|(\nabla^2 f(x_k))^{-1}\| \cdot \left\| \int_0^1 \nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right\| \cdot \|x_k - x^*\| \\ &\leq \frac{1}{l - M\|x_k - x^*\|} \int_0^1 M(1 - \tau)\|x_k - x^*\| d\tau \|x_k - x^*\| \\ &\leq \frac{1}{l - M\|x_k - x^*\|} \int_0^1 M(1 - \tau) d\tau \|x_k - x^*\|^2 \\ &\leq \frac{1}{l - M\|x_k - x^*\|} \frac{M}{2} \|x_k - x^*\|^2. \end{aligned}$$

Prin inducție se arată ușor că dacă $\|x_0 - x^*\| \leq 2l/3M$, atunci $\|x_k - x^*\| \leq 2l/3M$ pentru orice $k \geq 0$. Observăm că $\frac{1}{l - M\|x_k - x^*\|} \frac{M}{2} \leq \frac{3M}{2l} < \infty$ și deci $\|x_{k+1} - x^*\| \leq \frac{3M}{2l} \|x_k - x^*\|^2$. \square

Remarca 7.1.1 Din teorema anterioară putem concluziona că metoda Newton are o rată de convergență foarte rapidă în apropierea punctului de optim local. Mai mult, observăm că metoda Newton converge într-un singur pas pentru probleme pătratice convexe. Deci în comparație cu metodele de ordinul întâi unde în cel mai bun caz convergența se atinge în n pași, în metoda Newton obținem convergența în exact un pas pentru problemele pătratice convexe. Principalul dezavantaj al acestei metode este necesitatea de a calcula Hessiana funcției f și inversarea acestei matrice. Aceste operații sunt costisitoare, complexitatea fiind de ordinul $\mathcal{O}(n^3)$, și deci pentru dimensiuni mari ale problemei de optimizare (UNLP), de exemplu $n > 10^3$, aceste operații sunt foarte greu de realizat pe un calculator obișnuit.

7.1.2 Convergența globală a metodei Newton

Dacă pornim dintr-un punct x_0 ce nu este într-o vecinătate a lui x^* , atunci metoda Newton va trebui modificată pentru a garanta convergența ei către un punct staționar. Modificarea constă în alegerea dimensiunii pasului $\alpha_k \neq 1$, astfel încât de exemplu condiția Wolfe (W1) să fie satisfăcută. În acest caz, metoda Newton (numită și *metoda Newton cu pas variabil*) devine:

$$x_{k+1} = x_k - \alpha_k (\nabla^2 f(x_k))^{-1} \nabla f(x_k).$$

În general, lungimea pasului α_k se alege cu procedura ideală (adică se obține din minimizarea funcției unidimensionale $\min_{\alpha \geq 0} f(x_k + \alpha d_k)$, unde $d_k = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$) sau cu procedura backtracking. Bazat pe una din aceste proceduri, observăm că dacă x_k este suficient de apropiat de x^* , pasul α_k va deveni 1 (vezi Fig. 7.3).

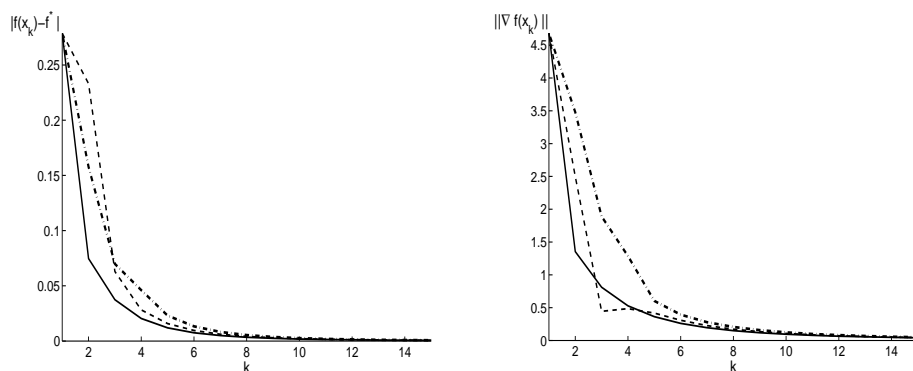


Figura 7.3: Metoda Newton aplicată funcției

$f(x_1, x_2) = (x_1 - x_2^3)^2 + 3(x_1 - x_2)^4$ cu alegerea pasului prin metoda ideală (linie continuă), backtracking (linie întreruptă-punctată) și $\alpha_k = 1$ (linie întreruptă). Evoluția de-a lungul iterațiilor $f(x_k) - f^*$ (stânga) și a $\|\nabla f(x_k)\|$ (dreapta).

Următoarea teoremă furnizează o serie de condiții suficiente ce garantează convergența globală a metodei Newton către un punct staționar.

Teorema 7.1.2 (Convergența globală a metodei Newton) Fie funcția obiectiv $f \in \mathcal{C}^2$ cu gradientul ∇f Lipschitz. Considerăm metoda Newton cu pas variabil $x_{k+1} = x_k - \alpha_k (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$, unde α_k se alege folosind backtracking. Mai departe presupunem că Hessiana

satisfacă condiția $\beta_1 I_n \preceq (\nabla^2 f(x_k))^{-1} \preceq \beta_2 I_n$ pentru orice $k \geq 0$, unde $0 < \beta_1 \leq \beta_2$. Atunci, metoda Newton produce un șir x_k cu proprietatea că fie $\nabla f(x_k) \rightarrow 0$, ori $f(x_k) \rightarrow -\infty$ (adică funcția f nu este marginită inferior).

Demonstrație: Presupunem că algoritmul Newton produce un șir x_k astfel încât șirul $f(x_k)$ este mărginit inferior. Din moment ce alegem α_k cu metoda backtracking, avem că condiția Wolfe (W1) este satisfăcută și deci $f(x_{k+1}) \leq f(x_k)$. Știm că un șir descrescător și mărginit inferior este convergent, deci există f^* astfel încât $f(x_k) \rightarrow f^*$. Aceasta implică de asemenea că $[f(x_k) - f(x_{k+1})] \rightarrow 0$. Din condiția Wolfe (W1) avem că:

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq -c_1 \alpha_k \nabla f(x_k)^T d_k \\ &= c_1 \alpha_k \nabla f(x_k)^T (\nabla^2 f(x_k))^{-1} \nabla f(x_k) \\ &\geq c_1 \alpha_k \beta_1 \|\nabla f(x_k)\|^2. \end{aligned}$$

Trecând la limita pentru $k \rightarrow \infty$ obținem că:

$$c_1 \alpha_k \beta_1 \|\nabla f(x_k)\|^2 \rightarrow 0. \quad (7.3)$$

Este suficient să arătăm că $\alpha_k \geq \alpha_{\min} > 0$ pentru orice $k \geq 0$. Vom arăta în cele ce urmează că atunci când lungimea pasului α_k este aleasă conform procedurii backtracking avem că $\alpha_k \geq \alpha_{\min}$, unde $\alpha_{\min} = \min\{1, \frac{(1-c_1)\rho}{L\beta_2}\} > 0$ și L este constanta Lipschitz pentru gradientul ∇f , adică $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$.

Pentru pas complet $\alpha_k = 1$ avem în mod evident satisfăcută relația $\alpha_k \geq \alpha_{\min}$. În celălalt caz, datorită procedurii de backtracking pentru alegerea lungimii pasului, avem că pasul anterior $\alpha = \frac{\alpha_k}{\rho}$ nu satisface condiția Wolfe (W1), pentru că în caz contrar ar fi fost folosit acest pas, și deci:

$$\begin{aligned} f(x_k + \frac{\alpha_k}{\rho} d_k) &> f(x_k) + c_1 \frac{\alpha_k}{\rho} \nabla f(x_k)^T d_k \\ f(x_k + \frac{\alpha_k}{\rho} d_k) - f(x_k) &> c_1 \frac{\alpha_k}{\rho} \nabla f(x_k)^T d_k. \end{aligned}$$

Folosind Taylor avem că există $\tau \in (0, \frac{\alpha_k}{\rho})$ astfel încât $f(x_k + \frac{\alpha_k}{\rho} d_k) - f(x_k) = \frac{\alpha_k}{\rho} \nabla f(x_k + \tau d_k)^T d_k$. Mai departe, obținem:

$$\begin{aligned} \nabla f(x_k + \tau d_k)^T d_k &> c_1 \nabla f(x_k)^T d_k \\ \underbrace{(\nabla f(x_k + \tau d_k) - \nabla f(x_k))^T d_k}_{\leq \tau L \|d_k\|^2} &> (1 - c_1) \underbrace{(-\nabla f(x_k)^T d_k)}_{= d_k^T \nabla^2 f(x_k) d_k}. \end{aligned}$$

În concluzie, ținând seama că $\tau \leq \frac{\alpha_k}{\rho}$, avem:

$$\frac{\alpha_k}{\rho} L \|d_k\|^2 > (1 - c_1) d_k^T \nabla^2 f(x_k) d_k \geq \frac{1 - c_1}{\beta_2} \|d_k\|^2,$$

ceea ce conduce la $\alpha_k > \frac{(1-c_1)\rho}{\beta_2 L} > 0$, adică șirul α_k este mărginit inferior de o constantă pozitivă $\alpha_{\min} = \frac{(1-c_1)\rho}{\beta_2 L}$. Am folosit faptul că valorile proprii ale matricei Hessiene satisfac condiția: $\frac{1}{\beta_1} \geq \lambda(\nabla^2 f(x_k)) \geq \frac{1}{\beta_2}$. Din faptul că $\alpha_k \geq \alpha_{\min} > 0$ și relația (7.3) obținem că $\|\nabla f(x_k)\| \rightarrow 0$ pentru $k \rightarrow \infty$. \square

O observație importantă legată de această metodă ține de faptul că direcția Newton este direcție de descreștere dacă $\nabla^2 f(x_k) \succ 0$. Dacă această condiție nu este satisfăcută, atunci în locul matricei $\nabla^2 f(x_k)$ vom considera matricea $\epsilon_k I_n + \nabla^2 f(x_k)$ pentru o valoare adecvată $\epsilon_k > 0$ astfel încât noua matrice să devină pozitiv definită. În acest caz, metoda Newton cu pas variabil devine:

$$x_{k+1} = x_k - \alpha_k (\epsilon_k I_n + \nabla^2 f(x_k))^{-1} \nabla f(x_k), \quad (7.4)$$

unde ca și mai înainte lungimea pasului α_k se alege prin metoda ideală sau backtracking. În cele ce urmează definim o procedură de alegere a constantei ϵ_k . Observăm că dacă alegem ϵ_k prea mic pentru ca iterația anterioară să fie aproximativ similară cu cea a metodei Newton (care are rată de convergență pătratică locală), atunci putem avea probleme numerice datorită faptului că Hessiana este prost condiționată când este aproape singulară. Pe de altă parte, dacă ϵ_k este foarte mare atunci matricea $\epsilon_k I_n + \nabla^2 f(x_k)$ este diagonal dominantă și deci metoda va avea un comportament similar cu algoritmul gradient (care are rată de convergență liniară). Dând o iterație x_k și $\epsilon_k > 0$, încercăm să calculăm factorizarea Cholesky LL^T a matricei $\epsilon_k I_n + \nabla^2 f(x_k)$. Dacă această factorizare nu este posibilă atunci multiplicăm ϵ_k cu un factor β (de exemplu, putem alege $\beta = 4$) și repetăm până când această factorizare este posibilă. Odată ce această factorizare este posibilă, o folosim în aflarea direcției Newton, adică o folosim în rezolvarea sistemului $LL^T \cdot d_k = -\nabla f(x_k)$.

Convergența globală a metodei Newton modificată (7.4) se demonstrează în aceeași manieră ca în Teorema 7.1.2.

7.2 Metode quasi-Newton

După cum am menționat anterior, principalul dezavantaj al metodei Newton constă în faptul că la fiecare iterație este nevoie să calculăm Hessiana și inversa sa, operații costisitoare, în general de ordinul $\mathcal{O}(n^3)$. Metodele quasi-Newton au scopul de a înlocui inversa Hessienei $\nabla^2 f(x_k)^{-1}$ cu o matrice H_k ce poate fi calculată mult mai ușor dar în același timp de a păstra rata de convergență rapidă a metodei Newton. În metodele quasi-Newton se construiește de asemenea o aproximare pătratică a funcției obiectiv unde Hessiana funcției pătratice se aproximează pe baza diferențelor de gradient, deci aceste metode folosesc numai informație de ordinul întâi. Mai mult, costul per iterație la metodele quasi-Newton este de ordinul $\mathcal{O}(n^2)$. Considerăm următoarea iterație:

$$x_{k+1} = x_k - \alpha_k H_k \nabla f(x_k).$$

Se observă că direcția $d_k = -H_k \nabla f(x_k)$ este una de descreștere dacă matricea $H_k \succ 0$:

$$\nabla f(x_k)^T d_k = -\nabla f(x_k)^T H_k \nabla f(x_k) < 0.$$

În metoda quasi-Newton pasul α_k se alege de obicei cu procedura ideală sau pe bază de backtracking. În general, ca și în cazul metodei Newton, dacă x_k este suficient de apropiat de soluția x^* alegem $\alpha_k = 1$.

Obiectivul nostru este de a găsi reguli de actualizare pentru matricea H_k astfel încât aceasta să convergă asimptotic la adevărata inversă a Hessienei, adică:

$$H_k \rightarrow \nabla^2 f(x^*)^{-1}.$$

Din aproximarea Taylor avem:

$$\nabla f(x_{k+1}) \approx \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k).$$

În concluzie, din aproximarea adevăratei Hessiene $\nabla^2 f(x_k)$ cu matricea B_{k+1} obținem următoarea relație:

$$\nabla f(x_{k+1}) - \nabla f(x_k) = B_{k+1}(x_{k+1} - x_k) \quad (7.5)$$

sau echivalent, notând $H_{k+1} = B_{k+1}^{-1}$ obținem:

$$H_{k+1}(\nabla f(x_{k+1}) - \nabla f(x_k)) = x_{k+1} - x_k. \quad (7.6)$$

Ecuția (7.5) sau (7.6) este cunoscută sub numele de *ecuația secantei*.

Pentru $H_{k+1}^{-1} = \nabla^2 f(x_k)$ recuperăm metoda Newton. Se observă că avem o interpretare similară celei a metodei Newton, și anume că la fiecare iterație considerăm o aproximare pătratică convexă a funcției obiectiv (i.e. $B_k \succcurlyeq 0$) și o minimizăm pentru a obține direcția de la următorul pas:

$$d_k = \arg \min_{d \in \mathbb{R}^n} f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T B_k d. \quad (7.7)$$

Întrucât Hessiana este simetrică este necesar ca matricele B_{k+1} și H_{k+1} să fie simetrice de asemenea. În concluzie avem n ecuații (din relația (7.6)) cu $\frac{n(n+1)}{2}$ necunoscute (prin impunerea simetriei asupra matricei H_{k+1}) și deci se obține un număr infinit de soluții. În continuare vom enunța diferite reguli de actualizare a matricei B_{k+1} sau H_{k+1} ce satisfac ecuația secantei (7.5) sau (7.6) și simetria.

7.2.1 Actualizări de rang unu

Cea mai simplă actualizare posibilă pentru matricea B_k sau echivalent pentru matricea H_k este cea de rang unu. În acest caz considerăm o matrice simetrică pozitiv definită inițială B_0 dată și apoi actualizăm matricea B_{k+1} prin următoarea formulă:

$$B_{k+1} = B_k + \beta_k u_k u_k^T,$$

în care $\beta_k \in \mathbb{R}$ și $u_k \in \mathbb{R}^n$ sunt alese astfel încât ecuația secantei (7.5) să fie satisfăcută. Observăm că dacă matricea simetrică $B_0 \succ 0$ și $\beta_k \geq 0$ atunci matricele B_k sunt simetrice și pozitiv definite pentru orice $k \geq 0$. Introducem acum următoarele notații:

$$\Delta_k = x_{k+1} - x_k \quad \text{and} \quad \delta_k = \nabla f(x_{k+1}) - \nabla f(x_k).$$

Impunem asupra matricei B_{k+1} condiția (7.5):

$$B_{k+1} \Delta_k = \delta_k.$$

Aceasta conduce la

$$\delta_k = B_k \Delta_k + \beta_k (u_k^T \Delta_k) u_k.$$

Concluzionăm că $u_k = \gamma(\delta_k - B_k \Delta_k)$ pentru un anumit scalar γ și înlocuind această expresie în egalitatea precedentă obținem:

$$\delta_k - B_k \Delta_k = \beta_k \gamma^2 [(\delta_k - B_k \Delta_k)^T \Delta_k] (\delta_k - B_k \Delta_k).$$

Din această relație rezultă că β_k și γ trebuie alese astfel încât:

$$\beta_k = \text{sgn}((\delta_k - B_k \Delta_k)^T \Delta_k), \quad \gamma = \pm |(\delta_k - B_k \Delta_k)^T \Delta_k|^{-1/2}.$$

În concluzie, obținem următoarea formulă pentru actualizarea lui B_{k+1} :

$$B_{k+1} = B_k + \frac{1}{(\delta_k - B_k \Delta_k)^T \Delta_k} (\delta_k - B_k \Delta_k)(\delta_k - B_k \Delta_k)^T.$$

Aplicând formula Sherman-Morrison pentru $H_{k+1} = B_{k+1}^{-1}$ obținem următoarea actualizare pentru H_{k+1} :

$$H_{k+1} = H_k + \frac{1}{\delta_k^T (\Delta_k - H_k \delta_k)} (\Delta_k - H_k \delta_k)(\Delta_k - H_k \delta_k)^T.$$

Pentru a garanta că $H_{k+1} \succ 0$, este necesară satisfacerea următoarei inegalități:

$$\delta_k^T (\Delta_k - H_k \delta_k) > 0.$$

Însă în practică se poate observa că există și cazuri când $\delta_k^T (\Delta_k - H_k \delta_k) = 0$. O strategie folosită în această situație este următoarea: dacă $\delta_k^T (\Delta_k - H_k \delta_k)$ este mic, de exemplu $|\delta_k^T (\Delta_k - H_k \delta_k)| < r \|\delta_k\| \cdot \|\Delta_k - H_k \delta_k\|$, pentru un $r < 1$ suficient de mic, atunci considerăm $H_{k+1} = H_k$.

7.2.2 Actualizări de rang doi

În actualizările de rang doi pornim de asemenea de la ecuația secantei (7.5). Din moment ce avem o infinitate de matrice simetrice ce satisfac această ecuație, determinăm B_{k+1} în mod unic prin impunerea condiției ca această matrice să fie cât mai aproape posibil de matricea de la iterația precedentă B_k :

$$B_{k+1} = \arg \min_{B=B^T, B\Delta_k=\delta_k} \|B - B_k\|,$$

unde $\|\cdot\|$ este o anumită normă pe matrice. Pentru o rezolvare explicită a problemei de optimizare anterioare putem considera norma:

$$\|A\|_W = \|W^{1/2} A W^{1/2}\|_F,$$

adică norma Frobenius, unde matricea W este aleasă astfel încât să fie pozitiv definită satisfăcând condiția $W\delta_k = \Delta_k$. În acest caz, soluția B_{k+1} a problemei de optimizare anterioare este dată de următoarea formulă:

$$B_{k+1} = (I_n - \beta_k \delta_k \Delta_k^T) B_k (I_n - \beta_k \Delta_k \delta_k^T) + \beta_k \delta_k \delta_k^T,$$

în care $\beta_k = \frac{1}{\Delta_k^T \delta_k}$. Folosind formula Sherman-Morrison-Woodbury obținem că actualizarea pentru $H_{k+1} = B_{k+1}^{-1}$ este dată de expresia:

$$H_{k+1} = H_k + \frac{1}{\Delta_k^T \delta_k} \Delta_k \Delta_k^T - \frac{1}{\delta_k^T H_k \delta_k} (H_k \delta_k)(H_k \delta_k)^T.$$

Metoda de optimizare quasi-Newton bazată pe această actualizare a matricei H_{k+1} se numește metoda *Davidon-Fletcher-Powell* (DFP).

Ca și mai înainte, alegem matricea inițială $H_0 \succ 0$. Metoda (DFP) satisface următoarele proprietăți

- (i) toate matricele $H_k \succ 0$ pentru orice $k \geq 0$;
- (ii) dacă $f(x) = \frac{1}{2}x^T Qx - q^T x$ este pătratică și strict convexă atunci metoda (DFP) furnizează direcții conjugate, adică vectorii $d_k = -H_k \nabla f(x_k)$ sunt direcții Q -conjugate. Mai mult, $H_n = Q^{-1}$ și în particular dacă $H_0 = I_n$ atunci direcțiile d_k coincid cu direcțiile din metoda gradientilor conjugati. De aceea, putem găsi soluția unei probleme de optimizare pătratice în maximum n pași cu ajutorul metodei quasi-Newton (DFP).

Dacă în locul problemei de optimizare anterioare considerăm problema:

$$H_{k+1} = \arg \min_{H=H^T, H\delta_k=\Delta_k} \|H - H_k\|,$$

unde folosim aceeași normă pe matrice ca și înainte, dar de data aceasta matricea pozitiv definită W satisface condiția $W\Delta_k = \delta_k$. Obținem următoarea soluție, numită și metoda *Broyden-Fletcher-Goldfarb-Shanno* (BFGS):

$$H_{k+1} = H_k - \frac{1}{\Delta_k^T \delta_k} ((H_k \delta_k) \Delta_k^T + \Delta_k (H_k \delta_k)^T) + \beta_k (\Delta_k \Delta_k^T)$$

$$\beta_k = \frac{1}{\Delta_k^T \delta_k} \left[1 + \frac{\delta_k^T \delta_k}{\Delta_k^T \delta_k} \right].$$

Aceleași proprietăți sunt valide și pentru metoda (BFGS) ca și în cazul metodei (DFP). Cu toate acestea, din punct de vedere numeric, (BFGS) este considerată cea mai stabilă metodă. Se observă că metodele quasi-Newton necesită doar informație de ordinul întâi (adică avem nevoie de informație de tip gradient). Observăm de asemenea că numărul de operații aritmetice pentru actualizarea matricelor H_{k+1} și apoi pentru

calcularea noului punct x_{k+1} este de ordinul $\mathcal{O}(n^2)$, mult mai mic decât în cazul metodei Newton care are complexitate de ordinul $\mathcal{O}(n^3)$. Mai mult, direcțiile generate de metodele quasi-Newton sunt direcții de descreștere dacă asigurăm satisfacerea condiției $H_k \succ 0$. În general, $H_k \rightarrow (\nabla^2 f(x^*))^{-1}$ pentru $k \rightarrow \infty$, iar în anumite condiții vom arăta că aceste metode au rată de convergență superliniară.

7.2.3 Convergența locală a metodelor quasi-Newton

În acest subcapitol analizăm rata de convergență locală a metodelor quasi-Newton, în particular arătăm convergența superliniară pentru aceste metode:

Teorema 7.2.1 *Fie x^* un punct ce satisface condițiile suficiente de optimalitate de ordinul II. Presupunem iterația quasi-Newton de forma $x_{k+1} = x_k - H_k \nabla f(x_k)$, unde H_k este inversabilă pentru orice $k \geq 0$ și satisface următoarea condiție Lipschitz:*

$$\|H_k(\nabla^2 f(x_k) - \nabla^2 f(y))\| \leq M \|x_k - y\| \quad \forall y \in \mathbb{R}^n,$$

și condiția de compatibilitate:

$$\|H_k(\nabla^2 f(x_k) - H_k^{-1})\| \leq \gamma_k \quad (7.8)$$

cu $0 < M < \infty$ și $\gamma_k \leq \gamma < 1$. De asemenea, presupunem că:

$$\|x_0 - x^*\| \leq \frac{2(1 - \gamma)}{M}. \quad (7.9)$$

Atunci x_k converge la x^* cu rată superliniară sub ipoteza că $\gamma_k \rightarrow 0$ sau rată liniară dacă $\gamma_k > \bar{\gamma} > 0$.

Demonstrație: Arătăm că $\|x_{k+1} - x^*\| \leq \beta_k \|x_k - x^*\|$, unde $\beta_k < \infty$. În acest scop, avem următoarele relații:

$$\begin{aligned} x_{k+1} - x^* &= x_k - x^* - H_k \nabla f(x_k) \\ &= x_k - x^* - H_k(\nabla f(x_k) - \nabla f(x^*)) \\ &= H_k(H_k^{-1}(x_k - x^*)) - H_k \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) (x_k - x^*) d\tau \\ &= H_k(H_k^{-1} - \nabla^2 f(x_k))(x_k - x^*) \\ &\quad - H_k \int_0^1 [\nabla^2 f(x^* + \tau(x_k - x^*)) - \nabla^2 f(x_k)] (x_k - x^*) d\tau. \end{aligned}$$

Aplicând norma în ambele părți, rezultă:

$$\begin{aligned}
 \|x_{k+1} - x^*\| &\leq \gamma_k \|x_k - x^*\| + \int_0^1 M \|x^* + \tau(x_k - x^*) - x_k\| d\tau \|x_k - x^*\| \\
 &= \left(\gamma_k + M \int_0^1 (1 - \tau) d\tau \|x_k - x^*\| \right) \|x_k - x^*\| \\
 &= \left(\gamma_k + \frac{M}{2} \|x_k - x^*\| \right) \|x_k - x^*\|.
 \end{aligned}$$

Avem următoarea rată de convergență:

$$\|x_{k+1} - x^*\| \leq \beta_k \|x_k - x^*\|,$$

unde $\beta_k = \gamma_k + \frac{M}{2} \|x_k - x^*\|$. Observăm că obținem rată de convergență superliniară dacă $\beta_k \rightarrow 0$, care are loc când $\gamma_k \rightarrow 0$. \square

Convergența globală a metodelor quasi-Newton se poate arăta în aceeași manieră ca în Teorema 7.1.2 corespunzătoare cazului metodei Newton. Metodele quasi-Newton care se bazează pe aproximarea inversei Hessienei reprezintă clasa de metode cea mai sofisticată pentru rezolvarea problemelor de optimizare fără constrângeri și constituie punctul culminant în dezvoltarea de algoritmi eficienți pentru aceste tipuri de probleme. Deși folosesc numai informație de gradient, în special prin măsurarea schimbărilor în gradient, metodele quasi-Newton construiesc o aproximare pătratică a funcției obiectiv suficient de bună pentru a produce convergență superliniară. Aceste metode sunt implementate în majoritatea pachetelor software existente: Matlab, IPOPT, etc.

Capitolul 8

Probleme de estimare și fitting

Problemele de estimare și fitting sunt probleme de optimizare având funcții obiectiv cu structura specială, și anume de tipul celor mai mici pătrate:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|\eta - \mathcal{M}(x)\|^2. \quad (8.1)$$

În această problemă de optimizare, $\eta \in \mathbb{R}^m$ sunt m măsurători și $\mathcal{M} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ este un *model*, iar $x \in \mathbb{R}^n$ se numesc *parametrii modelului*. Dacă adevărata valoare a lui x ar fi cunoscută, am putea evalua modelul $\mathcal{M}(x)$ pentru a obține predicțiile corespunzătoare măsurătorilor. Calculul lui $\mathcal{M}(x)$, ce poate reprezenta o funcție foarte complexă și de exemplu include în structura sa soluția unei ecuații diferențiale, se numește uneori *problema forward*: pentru intrări date ale modelului, se determină ieșirile corespunzătoare.

În problemele de estimare și fitting se caută setul de parametri ai modelului x ce realizează o predicție $\mathcal{M}(x)$ cât mai exactă pentru măsurătorile η date. Această problemă este denumită uzual *problemă inversă*: pentru un vector de ieșiri ale modelului η , se caută intrările corespunzătoare folosind un model ce depinde de setul de parametri $x \in \mathbb{R}^n$. Această clasă de probleme de optimizare (8.1) este frecvent întâlnită în cadrul unor aplicații cum ar fi:

- aproximare de funcții și estimare de parametri;
- estimare online pentru controlul proceselor dinamice;
- prognoză meteo (asimilare de date meteorologice).

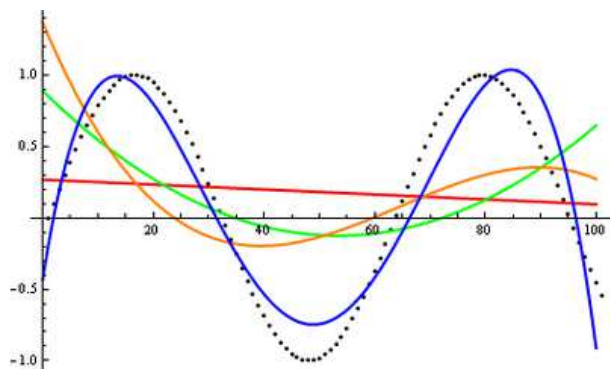


Figura 8.1: Aproximarea funcției $\sin(x)$ cu polinoame de grad unu până la gradul patru.

8.1 Problema celor mai mici pătrate: cazul liniar

Reamintim mai întâi definiția pseudo-inversei unei matrice:

Definiția 8.1.1 (Pseudo-Inversa Moore-Penrose) Fie matricea $J \in \mathbb{R}^{m \times n}$ cu $\text{rang}(J) = r$, iar descompunerea valorilor singulare (DVS) corespunzătoare lui J dată de $J = U\Sigma V^T$. Atunci, pseudo-inversa Moore-Penrose J^+ are expresia:

$$J^+ = V\Sigma^+U^T,$$

în care $\sigma_1, \dots, \sigma_r$ sunt valorile singulare ale matricei J și pentru

$$\Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & 0 \end{bmatrix}, \quad \text{definim} \quad \Sigma^+ = \begin{bmatrix} \sigma_1^{-1} & & & \\ & \ddots & & \\ & & \sigma_r^{-1} & \\ & & & 0 \end{bmatrix}.$$

Teorema 8.1.1 Dacă $\text{rang}(J) = n$, atunci

$$J^+ = (J^T J)^{-1} J^T.$$

Dacă $\text{rang}(J) = m$, atunci

$$J^+ = J^T (J J^T)^{-1}.$$

Demonstrație: Observăm următoarele relații:

$$\begin{aligned}(J^T J)^{-1} J^T &= (V \Sigma^T U^T U \Sigma V^T)^{-1} V \Sigma^T U^T = V (\Sigma^T \Sigma)^{-1} V^T V \Sigma^T U^T \\ &= V (\Sigma^T \Sigma)^{-1} \Sigma^T U^T = V \Sigma^+ U^T.\end{aligned}$$

Se urmează un raționament similar și în cel de-al doilea caz. \square

Se observă că dacă $\text{rang}(J) = n$, i.e. coloanele lui J sunt liniar independente atunci $J^T J$ este inversabilă.

Întâlnim frecvent în aplicații de estimare și fitting modele descrise de funcții liniare în x . Dacă \mathcal{M} este liniar, adică $\mathcal{M}(x) = Jx$, atunci funcția obiectiv devine $f(x) = \frac{1}{2} \|\eta - Jx\|^2$, ceea ce reprezintă o funcție convexă pătratică datorită faptului că Hessiana $\nabla^2 f(x) = J^T J \succ 0$. În acest caz definim problema celor mai mici pătrate (CMMP) liniară ca:

$$\min_{x \in \mathbb{R}^n} f(x) \quad \left(= \frac{1}{2} \|\eta - Jx\|^2 \right).$$

Deoarece în problema CMMP liniară funcția obiectiv f este pătratică convexă, condițiile de optimalitate de ordinul întâi $\nabla f(x) = 0$ sunt suficiente pentru determinarea unui punct de optim global. În concluzie, orice soluție a sistemului $J^T Jx - J^T \eta = 0$ este punct de minim global al problemei CMMP. Presupunând că $\text{rang}(J) = n$, punctul de minim global este unic și se determină prin următoarea relație:

$$J^T Jx^* - J^T \eta = 0 \quad \Longleftrightarrow \quad x^* = (J^T J)^{-1} J^T \eta = J^+ \eta. \quad (8.2)$$

Exemplul [Problema mediei]: Fie următoarea problemă simplă de optimizare:

$$\min_{x \in \mathbb{R}} \frac{1}{2} \sum_{i=1}^m (\eta_i - x)^2.$$

Observăm că ea se încadrează în clasa de probleme liniare de tip CMMP, unde vectorul η și matricea $J \in \mathbb{R}^{m \times 1}$ sunt date de

$$\eta = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_m \end{bmatrix}, \quad J = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}. \quad (8.3)$$

Deoarece $J^T J = m$, se observă ușor că:

$$J^+ = (J^T J)^{-1} J^T = \frac{1}{m} [1 \quad 1 \quad \cdots \quad 1]$$

și din acest motiv concluzionăm că punctul de minim este egal cu media $\hat{\eta}$ a punctelor date η_i , adică:

$$x^* = J^+ \eta = \frac{1}{m} \sum_{i=1}^m \eta_i = \hat{\eta}.$$

Exemplul [Regresie liniară]: Se dă setul de date $\{t_1, \dots, t_m\}$ cu valorile corespunzătoare $\{\eta_1, \dots, \eta_m\}$. Dorim să determinăm vectorul parame- trilor $x = (x_1, x_2)$, astfel încât polinomul de ordinul întâi $p(t; x) = x_1 + x_2 t$ realizează predicția lui η la momentul t . Problema de optimizare se prezintă sub forma:

$$\min_{x \in \mathbb{R}^2} \frac{1}{2} \sum_{i=1}^m (\eta_i - p(t_i; x))^2 = \min_{x \in \mathbb{R}^2} \frac{1}{2} \left\| \eta - J \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right\|^2,$$

în care η este același vector ca și în cazul (8.3), iar matricea J are forma:

$$J = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}.$$

Punctul de minim local este determinat de ecuația (8.2), unde calculul matricei $(J^T J)$ este trivial:

$$J^T J = \left[\begin{array}{c|c} m & \sum t_i \\ \hline \sum t_i & \sum t_i^2 \end{array} \right] = m \begin{bmatrix} 1 & \hat{t} \\ \hat{t} & \hat{t}^2 \end{bmatrix}.$$

Pentru a obține x^* , în primul rând se calculează $(J^T J)^{-1}$:

$$(J^T J)^{-1} = \frac{1}{m(\hat{t}^2 - (\hat{t})^2)} \begin{bmatrix} \hat{t}^2 & -\hat{t} \\ -\hat{t} & 1 \end{bmatrix}. \quad (8.4)$$

În al doilea rând, calculăm $J^T \eta$ după cum urmează:

$$J^T \eta = \begin{bmatrix} 1 & \cdots & 1 \\ t_1 & \cdots & t_m \end{bmatrix} \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_m \end{bmatrix} = \begin{bmatrix} \sum \eta_i \\ \sum \eta_i t_i \end{bmatrix} = m \begin{bmatrix} \hat{\eta} \\ \hat{\eta} \hat{t} \end{bmatrix}. \quad (8.5)$$

Deci, punctul de minim local este determinat de combinarea expresiilor (8.4) și (8.5). Se observă că:

$$\hat{t}^2 - (\hat{t})^2 = \frac{1}{m} \sum (t_i - \hat{t})^2 = \sigma_t^2,$$

unde ultima relație rezultă din definiția standard a varianței σ_t . Coeficientul de corelație ρ este definit similar de expresia:

$$\rho = \frac{\sum(\eta_i - \hat{\eta})(t_i - \hat{t})}{m\sigma_t\sigma_\eta} = \frac{\hat{t}\eta - \hat{\eta}\hat{t}}{\sigma_t\sigma_\eta}.$$

Vectorul parametrilor $x = (x_1, x_2)$ este determinat de:

$$x^* = \frac{1}{\sigma_t^2} \begin{bmatrix} \hat{t}^2 & \hat{\eta} - \hat{t}\hat{\eta} \\ -\hat{t} & \hat{\eta} + \hat{\eta}\hat{t} \end{bmatrix} = \begin{bmatrix} \hat{\eta} - \hat{t}\frac{\sigma_\eta}{\sigma_t}\rho \\ \frac{\sigma_\eta}{\sigma_t}\rho \end{bmatrix}.$$

În final, expresia poate fi formulată ca un polinom de gradul întâi:

$$p(t; x^*) = \hat{\eta} + (t - \hat{t})\frac{\sigma_\eta}{\sigma_t}\rho.$$

8.1.1 Probleme CMMP liniare prost condiționate

Dacă $J^T J$ este inversabilă, mulțimea soluțiilor optime X^* conține un singur punct de optim x^* , determinat de ecuația (8.2): $X^* = \{(J^T J)^{-1} J\eta\}$. Dacă $J^T J$ nu este inversabilă, mulțimea soluțiilor X^* este dată de:

$$X^* = \{x \in \mathbb{R}^n : \nabla f(x) = 0\} = \{x \in \mathbb{R}^n : J^T Jx - J^T \eta = 0\}.$$

Pentru alegerea celei mai bune soluții din această mulțime, se caută *soluția cu norma minimă*, adică vectorul x^* cu norma minimă ce satisface condiția $x^* \in X^*$.

$$\min_{x \in X^*} \frac{1}{2} \|x\|^2. \quad (8.6)$$

Arătăm mai departe că această soluție cu normă minimă este dată de *pseudo-inversa Moore-Penrose*, adică soluția optimă a problemei de optimizare (8.6) este dată de $x^* = J^+ \eta$. Soluția cu norma minimă, adică soluția problemei de optimizare (8.6), poate fi determinată dintr-o *problemă regularizată* CMMP liniară, și anume:

$$\min_{x \in \mathbb{R}^n} \bar{f}(x) \quad \left(= \frac{1}{2} \|\eta - Jx\|^2 + \frac{\beta}{2} \|x\|^2 \right), \quad (8.7)$$

cu o constantă $\beta > 0$ suficient de mică. Se știe că problema de optimizare (8.7) este echivalentă cu problema de optimizare (8.6) cu condiția că β

suficient de mic este ales în mod adecvat. Condițiile de optimalitate pentru problema pătratică convexă (8.7) sunt:

$$\begin{aligned}\nabla \bar{f}(x) &= J^T J x - J^T \eta + \beta x = (J^T J + \beta I_n) x - J^T \eta = 0 \\ \Rightarrow x^* &= (J^T J + \beta I_n)^{-1} J^T \eta.\end{aligned}\quad (8.8)$$

q

Lema 8.1.1 *Următoarea relație are loc pentru o matrice $J \in \mathbb{R}^{m \times n}$:*

$$\lim_{\beta \rightarrow 0} (J^T J + \beta I_n)^{-1} J^T = J^+.$$

Demonstrație: Din descompunerea DVS corespunzătoare matricei $J = U \Sigma V^T$ avem că matricea $(J^T J + \beta I_n)^{-1} J^T$ poate fi scrisă în forma:

$$\begin{aligned}(J^T J + \beta I_n)^{-1} J^T &= (V \Sigma^T U^T U \Sigma V^T + \beta \underbrace{I_n}_{V V^T})^{-1} \underbrace{J^T}_{U \Sigma^T V^T} \\ &= V (\Sigma^T \Sigma + \beta I_n)^{-1} V^T V \Sigma^T U^T \\ &= V (\Sigma^T \Sigma + \beta I_n)^{-1} \Sigma^T U^T.\end{aligned}$$

Partea dreaptă a ecuației are expresia:

$$V \begin{bmatrix} \sigma_1^2 + \beta & & & \\ & \ddots & & \\ & & \sigma_r^2 + \beta & \\ & & & \beta \end{bmatrix}^{-1} \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & 0 \end{bmatrix} U^T$$

Calcularea produsului de matrice conduce la:

$$V \begin{bmatrix} \frac{\sigma_1}{\sigma_1^2 + \beta} & & & \\ & \ddots & & \\ & & \frac{\sigma_r}{\sigma_r^2 + \beta} & \\ & & & \frac{0}{\beta} \end{bmatrix} U^T.$$

Se observă ușor că pentru $\beta \rightarrow 0$ fiecare element diagonal are forma:

$$\lim_{\beta \rightarrow 0} \frac{\sigma_i}{\sigma_i^2 + \beta} = \begin{cases} \frac{1}{\sigma_i} & \text{daca } \sigma_i \neq 0 \\ 0 & \text{daca } \sigma_i = 0. \end{cases}$$

□

Am arătat ca pseudo-inversa Moore-Penrose J^+ rezolvă problema (8.7) pentru o constantă suficient de mică $\beta > 0$. Din acest motiv, se realizează selecția unei soluții $x^* \in X^*$ cu norma minimă.

8.1.2 Formularea statistică a problemelor CMMP liniare

O problemă CMMP liniară (8.1) poate fi interpretată în sensul determinării unui set de parametri $x \in \mathbb{R}^n$ ce *explică* măsurătorile perturbate η în cel mai *bun* mod. Fie η_1, \dots, η_m valorile observate ale unei variabile aleatorii având densitatea $P(\eta|x)$ ce depinde de setul de parametri x . Presupunem $\eta_i = M_i(\bar{x}) + \beta_i$, cu \bar{x} valoarea *adevărată* a parametrului și β_i *zgomot Gaussian* cu media $\mathbb{E}(\beta_i) = 0$ și varianța $\mathbb{E}(\beta_i \beta_i) = \sigma_i^2$. Mai mult, presupunem că β_i și β_j sunt independente. Atunci definim *funcția de verosimilitate*:

$$P(\eta|x) = \prod_{i=1}^m P(\eta_i | x) = \prod_{i=1}^m \exp\left(\frac{-(\eta_i - M_i(x))^2}{2\sigma_i^2}\right). \quad (8.9)$$

Metoda verosimilității maxime (introdusă de Fischer în 1912) presupune că estimatorul x^* al adevăratului set de parametri \bar{x} este egal cu valoarea optimă ce maximizează funcția de verosimilitate. Estimatorul astfel obținut se numește *estimator de verosimilitate maximă*.

În general, funcțiile $P(\eta|x)$ și $\log P(\eta|x)$ își ating maximul în același punct x^* . Pentru a determina deci punctul de maxim al funcției de verosimilitate $P(\eta|x)$ determinăm punctul de maxim al funcției $\log P(\eta|x)$:

$$\log P(\eta|x) = \sum_{i=1}^m -\frac{(\eta_i - M_i(x))^2}{2\sigma_i^2}.$$

Deci parametrul ce maximizează $P(\eta|x)$ este dat de:

$$\begin{aligned} x^* &= \arg \max_{x \in \mathbb{R}^n} P(\eta|x) = \arg \min_{x \in \mathbb{R}^n} -\log(P(\eta|x)) \\ &= \arg \min_{x \in \mathbb{R}^n} \sum_{i=1}^m \frac{(\eta_i - M_i(x))^2}{2\sigma_i^2} = \arg \min_{x \in \mathbb{R}^n} \frac{1}{2} \|S^{-1}(\eta - M(x))\|^2, \end{aligned}$$

unde $S = \text{diag}(\sigma_1^2, \dots, \sigma_m^2)$. Deci, concluzionăm că problema CMMP are o interpretare statistică. Se observă că datorită faptului că putem avea diferite deviații standard σ_i pentru diferite măsurători η_i , se recomandă scalarea măsurătorilor și funcțiilor modelului pentru a obține o funcție

obiectiv în formă uzuală CMMP $\|\hat{\eta} - \hat{M}(x)\|_2^2$, după cum urmează:

$$\begin{aligned} \min_x \frac{1}{2} \sum_{i=1}^n \left(\frac{\eta_i - M_i(x)}{\sigma_i} \right)^2 &= \min_x \frac{1}{2} \|S^{-1}(\eta - M(x))\|^2 \\ &= \min_x \frac{1}{2} \|S^{-1}\eta - S^{-1}M(x)\|^2. \end{aligned}$$

8.2 Problema celor mai mici pătrate: cazul neliniar

Problemele CMMP liniare se pot rezolva ușor folosind metode numerice matriceale clasice, cum ar fi factorizarea QR. Pe de altă parte, rezolvarea globală a problemelor neliniare CMMP este în general NP-hard, dar pentru determinarea unui minim local se poate realiza iterativ. Principiul de bază constă în faptul că la fiecare iterație aproximăm problema originală cu propria liniarizare în punctul curent. În acest fel se obține o *apreciere* mai bună pentru următoarea iterație, folosind același procedeu prin care metoda Newton determină rădăcinile unui polinom dat. În mod uzual, pentru probleme neliniare CMMP de forma:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|\eta - M(x)\|^2$$

se aplică *metoda Gauss-Newton* sau *metoda Levenberg-Marquardt*. Pentru a descrie aceste metode, introducem mai întâi câteva notații convenabile:

$$F(x) = \eta - M(x)$$

și redefinim funcția obiectiv prin:

$$f(x) = \frac{1}{2} \|F(x)\|^2,$$

unde $F(x)$ este o funcție neliniară $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$, cu $m > n$ (adică considerăm un număr mai mare de măsurători decât parametri).

8.2.1 Metoda Gauss-Newton (GN)

Metoda Gauss-Newton este o metodă specializată pentru a rezolva problema CMMP neliniară:

$$\min_{x \in \mathbb{R}^n} f(x) \quad \left(= \frac{1}{2} \|F(x)\|^2 \right). \quad (8.10)$$

Într-un punct dat x_k la iterația k , $F(x)$ este liniarizat:

$$F(x) \approx F(x_k) + J(x_k)(x - x_k),$$

unde $J(x)$ este Jacobianul lui $F(x)$ definit de:

$$J(x) = \frac{\partial F(x)}{\partial x},$$

iar următoarea iterație x_{k+1} se obține prin rezolvarea unei probleme liniare CMMP. În concluzie, x_{k+1} poate fi determinat ca o soluție a următoarei probleme liniare CMMP:

$$x_{k+1} = \arg \min_{x \in \mathbb{R}^n} \frac{1}{2} \|F(x_k) + J(x_k)(x - x_k)\|^2.$$

Pentru simplitate, în locul notației $J(x_k)$ folosim J_k , iar în locul lui $F(x_k)$ folosim F_k și dacă presupunem că $J_k^T J_k$ este inversabilă atunci:

$$\begin{aligned} x_{k+1} &= \arg \min_{x \in \mathbb{R}^n} \frac{1}{2} \|F_k + J_k(x - x_k)\|^2 \\ &= x_k + \arg \min_{d \in \mathbb{R}^n} \frac{1}{2} \|F_k + J_k d\|^2 \\ &= x_k - (J_k^T J_k)^{-1} J_k^T F_k. \end{aligned}$$

Observăm că în iterația metodei Gauss-Newton direcția

$$d_k = -(J_k^T J_k)^{-1} J_k^T F_k = -J_k^+ F_k = \arg \min_{d \in \mathbb{R}^n} \frac{1}{2} \|F_k + J_k d\|^2$$

este o direcție de descreștere pentru funcția f , deoarece gradientul $\nabla f(x_k)$

$= \left(\frac{\partial F(x_k)}{\partial x} \right)^T F(x_k) = J_k^T F_k$ și matricea $J_k^T J_k$ este pozitiv definită. Pentru a asigura convergența metodei Gauss-Newton, de obicei introducem și un pas de lungime α_k , adică:

$$x_{k+1} = x_k - \alpha_k (J_k^T J_k)^{-1} J_k^T F_k,$$

unde α_k se alege cu una din procedurile descrise în capitolele anterioare (ideală, satisfăcând condițiile Wolfe sau backtracking). Se poate observa că în apropierea punctului de minim local lungimea pasului devine 1, adică $\alpha_k = 1$.

8.2.2 Metoda Levenberg-Marquardt

Această metodă reprezintă generalizarea metodei Gauss-Newton, ce se aplică în cazurile particulare când $J_k^T J_k$ nu este inversabilă și poate conduce la o convergență mai robustă pornind dintr-o regiune îndepărtată față de soluție. Metoda Levenberg-Marquardt realizează un avans mai redus prin penalizarea normei acestuia. Direcția în această metodă este dată de următoarea expresie:

$$\begin{aligned} d_k &= \arg \min_d \frac{1}{2} \|F_k + J_k d\|_2^2 + \frac{\beta_k}{2} \|d\|_2^2 \\ &= -(J_k^T J_k + \beta_k I_n)^{-1} J_k^T F_k \end{aligned}$$

cu scalarul $\beta_k > 0$ ales astfel încât matricea $J_k^T J_k + \beta_k I_n$ este pozitiv definită. Utilizând această direcție, iterația în metoda Levenberg-Marquardt este dată de următoarea expresie:

$$x_{k+1} = x_k - \alpha_k (J_k^T J_k + \beta_k I_n)^{-1} J_k^T F_k,$$

în care α_k se alege iarăși cu una din procedurile descrise în capitolele anterioare. În mod similar, în apropierea punctului de minim local lungimea pasului devine 1, adică $\alpha_k = 1$.

Observăm că dacă valoarea scalarului β_k se consideră foarte mare, nu aplicăm nici o corecție punctului curent x_k pentru că dacă $\beta_k \rightarrow \infty$ atunci direcția în metoda Levenberg-Marquardt satisface $d_k \rightarrow 0$. Mai precis, în acest caz $d_k \approx \frac{1}{\beta_k} J_k^T F_k \rightarrow 0$. Pe de altă parte, pentru valori mici ale lui β_k , adică pentru $\beta_k \rightarrow 0$ avem că direcția în metoda Levenberg-Marquardt satisface $d_k \rightarrow -J_k^+ F_k$ (conform Lemmei 8.1.1) și deci coincide cu direcția din metoda Gauss-Newton.

În cele ce urmează arătăm că aceste două metode au legătură strânsă cu metoda Newton. Este interesant de observat că gradientul funcției obiectiv aferent problemei CMMP neliniare $f(x) = \frac{1}{2} \|F(x)\|_2^2$ este dat de relația:

$$\nabla f(x) = J(x)^T F(x),$$

unde reamintim că $J(x)$ este Jacobianul funcției $F(x)$. În mod evident acest gradient se regăsește în iterațiile metodelor Gauss-Newton sau Levenberg-Marquardt. Deci, dacă gradientul este nul, atunci direcțiile în cele două metode sunt de asemenea nule. Aceasta este o condiție necesară pentru convergența la puncte staționare a unei metode: ambele

metode Gauss-Newton și Levenberg-Marquardt nu avansează dintr-un punct staționar x_k cu $\nabla f(x_k) = 0$. Mai departe notăm cu F_i componenta i a funcției multivectoriale F . Utilizând calculul diferențial standard observăm că Hessiana funcției obiectiv f este dată de următoarea expresie:

$$\nabla^2 f(x) = J(x)^T J(x) + \sum_{i=1}^m F_i(x) \cdot \nabla^2 F_i(x).$$

În concluzie, în cele două metode Gauss-Newton și Levenberg-Marquardt neglijăm cel de-al doilea termen al Hessiane funcției obiectiv f , adică termenul $\sum_{i=1}^m F_i(x) \cdot \nabla^2 F_i(x)$. Deci în aceste metode salvăm calcule prin neluarea în calcul a acestui termen $\sum_{i=1}^m F_i(x) \cdot \nabla^2 F_i(x)$, ceea ce în principiu conduce la o deteriorare a ratei de convergență a acestor metode față de rata de convergență a metodei Newton. Pe de altă parte, dacă acest termen $\sum_{i=1}^m F_i(x) \cdot \nabla^2 F_i(x)$ este mic în apropierea unei soluții locale, atunci rata de convergență a acestor două metode este comparabilă cu cea a metodei Newton. Observăm că acest termen este mic în apropierea unei soluții dacă funcția $F(x)$ este aproape liniară sau dacă componentele $F_i(x)$ sunt mici în apropiere de soluție. De exemplu, dacă se caută o soluție a sistemului neliniar $F(x) = 0$, cu $m = n$, atunci termenul neglijat este nul la soluție. Mai mult, dacă matricea $J_k = J(x_k) \in \mathbb{R}^{n \times n}$ este inversabilă, atunci direcția în metoda Gauss-Newton este dată de:

$$-(J_k^T J_k)^{-1} J_k^T F_k = -J_k^{-1} F_k.$$

Deci iterația în această metodă devine:

$$x_{k+1} = x_k - (J(x_k))^{-1} F(x_k),$$

și deci coincide cu iterația din metoda Newton standard pentru rezolvarea sistemului $F(x) = 0$. În acest caz, de obicei rata de convergență este superliniară.

Convergența globală și locală a metodelor Gauss-Newton și Levenberg-Marquardt poate fi derivată utilizând argumente similare celor din capitolul precedent pentru metoda (quasi-) Newton. Pentru convergență locală avem următorul rezultat:

Teorema 8.2.1 *Fie x^* un punct ce satisface condițiile suficiente de ordinul doi. Iterația metodelor Gauss-Newton sau Levenberg-Marquardt în apropierea punctului x^* are forma $x_{k+1} = x_k - H_k \nabla f(x_k)$, unde*

matricea H_k este dată fie de $H_k = (J_k^T J_k)^{-1}$ fie de $H_k = (J_k^T J_k + \beta_k I_n)^{-1}$. Pentru matricea inversabilă pozitiv definită H_k presupunem satisfăcută următoarea condiție Lipschitz:

$$\|H_k(\nabla^2 f(x_k) - \nabla^2 f(y))\| \leq M \|x_k - y\| \quad \forall k \in \mathbb{N}, \quad y \in \mathbb{R}^n$$

și, de asemenea, condiția de compatibilitate:

$$\|H_k(\nabla^2 f(x_k) - H_k^{-1})\| \leq \gamma_k \quad \forall k \in \mathbb{N} \quad (8.11)$$

cu $0 < M < \infty$ și $\gamma_k \leq \gamma < 1$. Presupunem de asemenea că

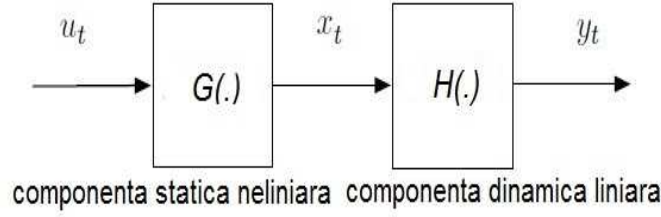
$$\|x_0 - x^*\| \leq \frac{2(1 - \gamma)}{M}. \quad (8.12)$$

Atunci x_k converge la x^* cu rata liniară dacă $\gamma_k > \bar{\gamma} > 0$.

Demonstrație: Vezi demonstrația Teoremei 7.2.1. □

8.3 Aplicație: identificarea unui sistem Hammerstein

Un sistem dinamic Hammerstein este un sistem discret de tip cascadă format dintr-o componentă neliniară, urmată de o dinamică liniară (vezi Fig. 8.2). Multe sisteme practice pot fi modelate folosindu-se acest tip de dinamică neliniară, e.g. modelarea liniilor de transmisie sau amplificatoarelor de putere înaltă, etc. Componenta statică neliniară este reprezentată de o funcție neliniară $G(\cdot)$, în timp ce componenta dinamică liniară este descrisă de o funcție de transfer $H(q) = q^{-1} \frac{A_p(q^{-1})}{B_p(q^{-1})}$, unde A_p și B_p sunt polinoame. Problema constă în găsirea coeficienților a_i și b_j pentru orice $i = 1, \dots, n_A$ și $j = 0, \dots, n_B$, unde n_A și n_B sunt gradele polinoamelor $A_p(q^{-1})$ și $B_p(q^{-1})$ și aproximarea componentei neliniare $G(\cdot)$, folosind numai date de intrări, u_t , și ieșiri, y_t . Considerăm de asemenea că e_t reprezintă erori de măsurare. Principala caracteristică a acestei probleme este aceea că semnalul intermediar, x_t , nu este accesibil pentru măsurători, și deci componenta neliniară $G(\cdot)$ nu se poate determina direct din teoria de aproximare a funcțiilor.

**Figura 8.2:** Sistem dinamic discret Hammerstein.

Ecuatiile ce descriu sistemul Hammerstein sunt următoarele:

$$\begin{cases} A(q^{-1})y_t = B(q^{-1}) \cdot x_{t-1} + A(q^{-1})e_t \\ x_t = G(u_t) \\ A(q^{-1}) = 1 + a_1q^{-1} + a_2q^{-2} + \dots + a_{n_A}q^{-n_A} \\ B(q^{-1}) = b_0 + b_1q^{-1} + b_2q^{-2} + \dots + b_{n_B}q^{-n_B}. \end{cases} \quad (8.13)$$

Pentru problema aproximării componenteii neliniare $G(\cdot)$ utilizăm aproximarea bazată pe funcții de bază:

$$G(u_t) = \sum_{i \in \Gamma} \gamma_i \cdot f_i(u_t) + \epsilon_t, \quad (8.14)$$

unde $f_i(\cdot)$ sunt funcții de bază (e.g. funcții polinomiale) și ϵ_t reprezintă eroarea de aproximare. Înlocuind ecuația (8.14) în (8.13), obținem:

$$\begin{aligned} y_t = & -a_1y_{t-1} - a_2y_{t-2} - \dots - a_{n_A}y_{t-n_A} \\ & + b_0 \sum_{i \in \Gamma} \gamma_i \cdot f_i(u_{t-1}) + \dots + b_{n_B} \sum_{i \in \Gamma} \gamma_i \cdot f_i(u_{t-n_B-1}) + v_t, \end{aligned} \quad (8.15)$$

pentru orice $t = N_0, \dots, N_{\max}$, unde $v_t = A_p(q^{-1})e_t + B_p(q^{-1})\epsilon_{t-1}$ reprezintă eroarea de aproximare totală. Suntem interesați în găsirea coeficienților polinoamelor A_p și B_p și a ponderilor γ_i corespunzătoare funcțiilor de bază f_i folosind metoda celor mai mici pătrate pentru relația (8.15). Introducem următoarele notații:

$$\begin{cases} x_a = [a_1 \ a_2 \ \dots \ a_{n_A}]^T \\ x_b = [b_0 \ b_1 \ \dots \ b_{n_B}]^T \\ x_\gamma = [\gamma_1 \ \gamma_2 \ \dots \ \gamma_n]^T, \end{cases} \quad (8.16)$$

unde pentru consistență notăm γ_i pentru $i \in \Gamma$ cu $\gamma_1, \gamma_2, \dots, \gamma_m$ (m reprezintă cardinalitatea mulțimii Γ) și cu f_i , funcțiile de bază

corespunzătoare ponderilor γ_i pentru orice $i = 1, \dots, m$. Ecuația (8.15) ne conduce la forma vectorială clasică a problemei celor mai mici pătrate:

$$y_t = \varphi^T(t)x + v_t, \quad (8.17)$$

unde vectorii din ecuația (8.17) au următoarele expresii:

$$\begin{aligned} x &= [x_a^T \ \theta_{\gamma_1}^T \ \theta_{\gamma_2}^T \ \dots \ \theta_{\gamma_m}^T]^T, \quad \theta_{\gamma_j} = [b_0\gamma_j \ b_1\gamma_j \ \dots \ b_{n_B}\gamma_j]^T \\ \varphi(t) &= [\varphi_a^T(t) \ \varphi_1^T(t) \ \varphi_2^T(t) \ \dots \ \varphi_m^T(t)]^T \\ \varphi_a(t) &= [-y_{t-1} \ -y_{t-2} \ \dots \ -y_{t-n_A}]^T \\ \varphi_i(t) &= [f_i(u_{t-1}) \ \dots \ f_i(u_{t-n_B-1})]^T \end{aligned} \quad (8.18)$$

pentru orice $i = 1, \dots, m$ și $t = N_0, \dots, N_{\max}$. Dorim să găsim parametrii x_a, x_b și x_γ prin rezolvarea unei probleme de tipul celor mai mici pătrate ce rezultă din relația (8.15) cu metoda Gauss-Newton. Ecuația (8.15) poate fi scrisă compact, după cum urmează:

$$y_t = \varphi_a(t)^T x_a + x_b^T Q_t x_\gamma, \quad (8.19)$$

unde matricea Q_t este dată de:

$$Q_t = [\varphi_1(t) \ \dots \ \varphi_m(t)],$$

pentru toți $t = N_0, \dots, N_{\max}$. Definim funcția biliniară:

$$F(x_a, x_b, x_\gamma) = \begin{bmatrix} y_{N_0} - \varphi_a(N_0)^T x_a - x_b^T Q_{N_0} x_\gamma \\ \dots \\ y_{N_{\max}} - \varphi_a(N_{\max})^T x_a - x_b^T Q_{N_{\max}} x_\gamma \end{bmatrix}$$

Atunci, putem estima parametrii x_a, x_b și x_γ prin rezolvarea unei probleme de tipul celor mai mici pătrate neliniare în forma:

$$(x_a^*, x_b^*, x_\gamma^*) = \arg \min_{x_a, x_b, x_\gamma} \|F(x_a, x_b, x_\gamma)\|^2, \quad (8.20)$$

unde F are o structură biliniară, $x = [x_a^T \ x_b^T \ x_\gamma^T]^T \in \mathbb{R}^n$ și $n = n_A + n_B + m$. Observăm că funcția $F(x_a, x_b, x_\gamma)$ este diferențiabilă:

$$\nabla F(x_a, x_b, x_\gamma) = \begin{bmatrix} \varphi_a(N_0)^T & x_\gamma^T B_{N_0}^T & x_b^T B_{N_0} \\ \dots & \dots & \dots \\ \varphi_a(N_{\max})^T & x_\gamma^T B_{N_{\max}}^T & x_b^T B_{N_{\max}} \end{bmatrix}.$$

Din Fig. 8.3 se observă o urmărire foarte bună a traiectoriei măsurate y_m de către sistemul Hammerstein ai cărui parametri au fost obținuți ca soluție a problemei celor mai mici pătrate neliniare rezolvată cu metoda Gauss-Newton.

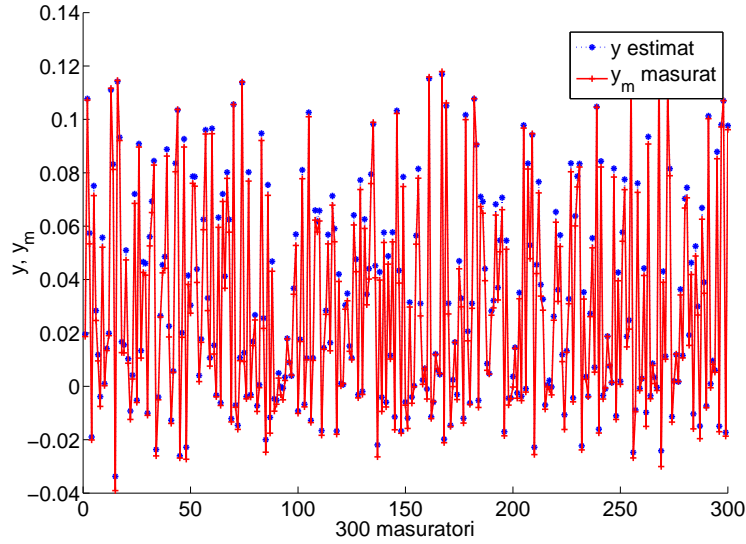


Figura 8.3: Pentru $N_{\max} = 300$ date de intrare și ieșire reprezentăm ieșirea y_m măsurată și ieșirea y produsă de sistemul Hammerstein ai cărui parametri au fost identificați prin procedura descrisă anterior.

Comentarii finale: Cu aceste două metode prezentate în acest capitol, Gauss-Newton și Levenberg-Marquardt, încheiem Partea a II-a a acestei lucrări dedicate metodelor numerice de optimizare pentru probleme fără constrângeri (UNLP): $\min_{x \in \mathbb{R}^n} f(x)$. Mai multe detalii despre metodele prezentate aici cât și alte metode care nu au fost abordate se pot găsi în cărțile clasice de optimizare neliniară ale lui Bertsekas [2], Luenberger [9], Nesterov [11] și Nocedal și Wright [13]. Teoria optimizării dezvoltată aici urmează în linii mari prezentarea din [9]. Pentru cazul convex, o analiză completă a metodelor de optimizare existente se găsește în [11]. Dintre cărțile dedicate implementării numerice a acestor metode de optimizare amintim de exemplu lucrarea lui Gill, Murray și Wright [7]. O descriere detaliată a pachetelor software existente pe piață este dată în More și Wright [10].

Partea III

Optimizare cu constrângeri

Capitolul 9

Teoria dualității

În această parte finală a lucrării ne îndreptăm din nou atenția asupra problemelor de optimizare cu constrângeri (NLP). Expunerea noastră va prezenta cazul constrâns ca o generalizare a cazului neconstrâns: vom defini condițiile de optimalitate pentru cazul constrâns (de ordinul întâi și doi), apoi vom arăta cum metodele numerice de optimizare de ordinul întâi și doi pentru probleme de optimizare neconstrânse pot fi extinse la cazul în care avem constrângeri. În final vom discuta algoritmi specializați pentru cazul particular al problemelor convexe constrânse. Începem expunerea noastră cu teoria dualității, fundamentală în înțelegerea algoritmilor de optimizare prezentați ulterior. Reamintim că o problemă neliniară cu constrângeri (NLP) (*NonLinear Programming*) în forma standard este descrisă de:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.l.:} \quad & g_1(x) \leq 0, \dots, g_m(x) \leq 0 \\ & h_1(x) = 0, \dots, h_p(x) = 0. \end{aligned}$$

Dacă introducem următoarele notații $g(x) = [g_1(x) \dots g_m(x)]^T$ și $h(x) = [h_1(x) \dots h_p(x)]^T$, atunci în formă compactă problema de optimizare amintită se rescrie astfel:

$$\begin{aligned} (NLP) : \quad & \min_{x \in \mathbb{R}^n} f(x) \\ \text{s.l.:} \quad & g(x) \leq 0, \quad h(x) = 0, \end{aligned}$$

unde funcția obiectiv $f : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$, funcția vectorială ce definește constrângerile de inegalitate $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ și funcția vectorială ce

definește constrângerile de egalitate $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$ se presupune a fi de două ori diferențiabile. În acest caz, mulțimea fezabilă asociată problemei (NLP) este:

$$X = \{x \in \mathbb{R}^n : g(x) \leq 0, h(x) = 0\}$$

și astfel putem rescrie problema (NLP) și sub forma:

$$\min_{x \in X} f(x).$$

Exemplul 9.0.1 (Optimizarea rutării în rețea de comunicație)

În cele ce urmează considerăm o rețea de comunicație a datelor, modelată de un graf direcționat $G = (V, E)$, unde V este mulțimea nodurilor și E mulțimea perechilor ordonate $e = (i, j)$ (vezi Fig. 9.1). Nodul i se numește origine și nodul j destinație. Pentru orice pereche e considerăm

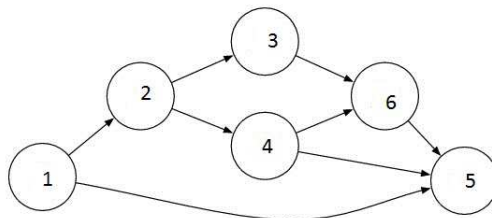


Figura 9.1: Optimizare în rețeaua de comunicație.

scalarul r_e reprezentând traficul de intrare în e . În contextul rutării de date într-o rețea, r_e este rata de trafic ce intră și iese din rețea prin originea și destinația lui e (măsurată în unități de date pe secundă). Obiectivul de rutare este acela de a împărți fiecare trafic r_e între diferitele rute existente de la originea i la destinația j în așa fel încât fluxul total rezultat minimizează o funcție cost adecvată. Notăm cu P_e mulțimea tuturor rutelor existente între originea i și destinația j a lui e și cu x_c partea de trafic din r_e atribuită rutei $c \in P_e$, numită de asemenea fluxul rutei c . Colecția tuturor fluxurilor de date $\{x_c : c \in P_e, e \in E\}$ trebuie să satisfacă următoarea constrângere:

$$\sum_{c \in P_e} x_c = r_e \quad \forall e \in E$$

și de asemenea $x_c \geq 0$ pentru orice $c \in P_e$ și $e \in E$. Fluxul total t_{ij} corespunzător arcului (i, j) este suma tuturor fluxurilor traversând arcul:

$$t_{ij} = \sum_{c: (i,j) \in c} x_c.$$

Putem defini o funcție cost de forma: $\sum_{(i,j) \in E} f_{ij}(t_{ij})$. Problema este să găsim toate fluxurile x_c care minimizează această funcție cost cu constrângerile anterioare:

$$\begin{aligned} \min_{x_c, t_{ij}} \quad & \sum_{(i,j) \in E} f_{ij}(t_{ij}) \\ \text{s.l. : } \quad & x_c \geq 0 \quad \forall c \in P_e, e \in E \\ & \sum_{c \in P_e} x_c = r_e \quad \forall e \in E, \quad t_{ij} = \sum_{c: (i,j) \in c} x_c \quad \forall (i, j) \in E. \end{aligned}$$

Se observă că putem elimina variabila t_{ij} din problema precedentă folosind egalitatea $t_{ij} = \sum_{c: (i,j) \in c} x_c$, adică putem obține o problemă de

optimizare doar în variabila x_c și cu mai puține constrângeri de egalitate, dar în acest caz funcția obiectiv nu mai are structura separabilă de mai sus (e.g. după eliminare, Hessiana funcției obiectiv nu mai este diagonală).

Exemplul 9.0.2 (Proiecția Euclideană) O noțiune fundamentală în geometrie și optimizare este proiecția Euclideană a unui vector $x_0 \in \mathbb{R}^n$ pe mulțimea $X \subseteq \mathbb{R}^n$ definită de vectorul din X aflat la cea mai mică distanță Euclidiană de x_0 (vezi Fig. 9.2). Matematic, această problemă se formulează sub forma unei probleme de optimizare constrânsă:

$$\min_{x \in X} \|x - x_0\|^2.$$

Se observă că funcția obiectiv pentru problema proiecției Euclidiene este pătratică, convexă, iar Hessiana este definită de matricea identitate. Când mulțimea X este nevidă, convexă și închisă, se poate arăta că există o singură soluție a problemei de optimizare anterioare, i.e. proiecția este unică. Mai mult, dacă X este convexă, atunci problema precedentă este problemă de optimizare convexă. În particular, dacă mulțimea X este un poliedru, adică $X = \{x \in \mathbb{R}^n : Ax = b, Cx \leq d\}$ atunci proiecția este o problemă de optimizare QP strict convexă. De exemplu, dacă presupunem

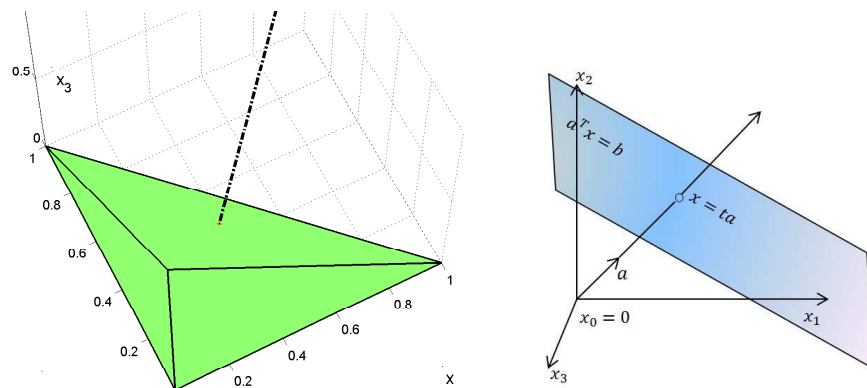


Figura 9.2: Proiecția vectorului $x_0 = [1 \ 1 \ 1]^T$ pe politopul $X = \{x \in \mathbb{R}^3 : x \geq 0, x_1 + x_2 + x_3 \leq 1\}$ (stânga) și a originii ($x_0 = 0$) pe un hiperplan $X = \{x \in \mathbb{R}^3 : a^T x = b\}$ (dreapta).

că mulțimea X este un hiperplan $X = \{x \in \mathbb{R}^n : a^T x = b\}$, unde $b \neq 0$, atunci proiecția originii $x_0 = 0$ devine o problemă pătratică cu o formă simplă:

$$\min_{x \in \{x \in \mathbb{R}^n : a^T x = b\}} \|x\|^2.$$

Se știe că vectorul a este perpendicular pe hiperplan, deci proiecția lui $x_0 = 0$ pe X este coliniară cu a , adică $x = ta$ pentru un scalar t (vezi Fig. 9.2). Înlocuind $x = ta$ în ecuația ce definește hiperplanul și rezolvând pentru scalarul t obținem $t = b/(a^T a)$, iar proiecția este dată de expresia:

$$x^* = \frac{b}{a^T a} a.$$

Exemplul 9.0.3 (Problema localizării) Problema localizării are foarte multe aplicații în inginerie, cum ar fi localizarea unei ținte, localizarea unui robot, etc. Considerăm că avem un număr m de senzori având locațiile cunoscute $s_i \in \mathbb{R}^3$ și cunoaștem, de asemenea, distanțele R_i de la acești senzori la obiectul necunoscut a cărui poziție trebuie determinată. Geometric (vezi Fig. 9.3), din datele cunoscute avem că obiectul se găsește la intersecția a m sfere de centre s_i și raze R_i . Dorim să estimăm poziția obiectului și de asemenea să măsurăm mărimea/volumul intersecției acestor sfere. Putem considera problema găsirii celei mai mari sfere incluse în această intersecție. Este ușor de observat că o sferă de centru x și rază R este conținută într-o sferă de centru s_i și rază R_i dacă și numai dacă diferența dintre raze este mai

mare decât distanța dintre centre. Putem atunci formula următoarea problemă de optimizare convexă cu constrângeri pătratice:

$$\begin{aligned} & \max_{x \in \mathbb{R}^3, R > 0} R \\ \text{s.l. : } & R_i \geq R + \|s_i - x\| \quad \forall i = 1, \dots, m. \end{aligned}$$

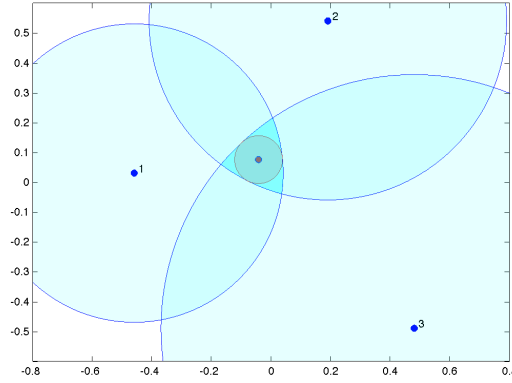


Figura 9.3: Problema localizării.

9.1 Funcția Lagrange

Funcția Lagrange, denumită astfel după matematicianul Joseph Louis Lagrange, este foarte importantă în teoria dualității. Începem prin a defini noțiuni standard din teoria dualității.

Definiția 9.1.1 (Problema de optimizare primală) Vom nota valoarea optimă globală a problemei de optimizare (NLP) cu f^* și o vom numi valoarea optimă primală:

$$f^* = \left\{ \min_{x \in \mathbb{R}^n} f(x) : g(x) \leq 0, h(x) = 0 \right\}. \quad (9.1)$$

Vom numi problema de optimizare (NLP) ca problema de optimizare primală, iar variabila de decizie x variabila primală. Notăm de asemenea cu

$$X = \{x \in \mathbb{R}^n : g(x) \leq 0, h(x) = 0\}$$

mulțimea fezabilă primală a problemei (NLP).

În concluzie, problema de optimizare primală este definită astfel:

$$(NLP) : \quad f^* = \min_{x \in X} f(x).$$

Se observă că putem determina relativ ușor o margine superioară pentru valoarea optimă f^* : selectăm un punct fezabil $\tilde{x} \in X$ și atunci avem $f^* \leq f(\tilde{x})$. Desigur, ne putem întreba cum se determină o margine inferioară pentru f^* . Vom arăta în cele ce urmează că această margine inferioară se poate determina folosind teoria dualității. Vom vedea de asemenea, că anumite probleme de optimizare pot fi rezolvate folosind teoria dualității. Cea mai populară formă a dualității pentru problemele de optimizare constrânse este *dualitatea Lagrange*. Deși dualitatea Lagrange poate fi dezvoltată pentru probleme generale de optimizare constrânsă, cele mai interesante rezultate se dau pentru cazul problemelor convexe constrânse. Reamintim că problema (NLP) precedentă este problemă de optimizare convexă dacă funcțiile f și g_1, \dots, g_m sunt funcții convexe, iar funcțiile h_1, \dots, h_p sunt funcții affine. Începem prin a defini funcția Lagrange (sau Lagrangianul):

Definiția 9.1.2 (Funcția Lagrange și multiplicatorii Lagrange)

Definim funcția Lagrange sau Lagrangianul, $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \bar{\mathbb{R}}$, ca fiind:

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \lambda^T g(x) + \mu^T h(x). \quad (9.2)$$

În această funcție am introdus două variabile noi, vectorii $\lambda \in \mathbb{R}^m$ și $\mu \in \mathbb{R}^p$, numiți multiplicatorii Lagrange sau variabile duale.

Funcția Lagrange joacă un rol principal atât în optimizarea convexă cât și în cea neconvexă. În mod obișnuit, se impune ca multiplicatorii pentru constrângerile de inegalitate λ să nu fie negativi, adică $\lambda \geq 0$, în timp ce multiplicatorii de egalitate μ sunt arbitrari. Aceste cerințe sunt motivate de următoarea leamnă:

Lema 9.1.1 (Mărginirea superioară a funcției Lagrange) Pentru orice variabilă primală \tilde{x} fezabilă pentru problema de optimizare (NLP) (i.e. $g(\tilde{x}) \leq 0$ și $h(\tilde{x}) = 0$) și pentru orice variabilă duală $(\tilde{\lambda}, \tilde{\mu}) \in \mathbb{R}^m \times \mathbb{R}^p$ satisfăcând $\tilde{\lambda} \geq 0$, următoarea inegalitate are loc:

$$\mathcal{L}(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) \leq f(\tilde{x}). \quad (9.3)$$

Demonstrație: Demonstrația urmează imediat din definiția funcției Lagrange și din faptul că $\tilde{\lambda} \geq 0$, $g(\tilde{x}) \leq 0$ și $h(\tilde{x}) = 0$:

$$\mathcal{L}(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) = f(\tilde{x}) + \tilde{\lambda}^T g(\tilde{x}) + \tilde{\mu}^T h(\tilde{x}) \leq f(\tilde{x}).$$

□

9.2 Problema duală

Din lema precedentă se observă că putem determina o margine inferioară pentru valoarea optimă primală a problemei (NLP). Mai mult, suntem interesați în determinarea celei mai bune margini inferioare pentru f^* . Pentru aceasta introducem mai întâi funcția duală.

Definiția 9.2.1 (Funcția duală) *Definim funcția duală $q : \mathbb{R}^m \times \mathbb{R}^p \rightarrow \bar{\mathbb{R}}$ ca infimul neconstrâns al Lagrangianului în funcție de variabila x , pentru multiplicatorii λ și μ fixați:*

$$q(\lambda, \mu) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda, \mu). \quad (9.4)$$

Această funcție va lua adesea valoarea $-\infty$, caz în care spunem că perechea (λ, μ) este *dual infeasibilă*. Funcția duală are proprietăți foarte interesante, pe care le demonstrăm în cele ce urmează:

Lema 9.2.1 (Mărginirea superioară a funcției duale) *Pentru orice pereche duală $(\tilde{\lambda}, \tilde{\mu})$ fezabilă, adică $\tilde{\lambda} \geq 0$ și $\tilde{\mu} \in \mathbb{R}^p$, următoarea inegalitate are loc:*

$$q(\tilde{\lambda}, \tilde{\mu}) \leq f^*. \quad (9.5)$$

Demonstrație: Această leamnă este o consecință directă a ecuației (9.3) și a definiției funcției duale: pentru \tilde{x} fezabil (i.e. $g(\tilde{x}) \leq 0$ și $h(\tilde{x}) = 0$) avem

$$q(\tilde{\lambda}, \tilde{\mu}) \leq \mathcal{L}(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) \leq f(\tilde{x}) \quad \forall \tilde{x} \in X, \tilde{\lambda} \in \mathbb{R}_+^m, \tilde{\mu} \in \mathbb{R}^p.$$

Această inegalitate este satisfăcută în particular pentru punctul de minim global x^* (care este de asemenea fezabil, adică $x^* \in X$), ceea ce conduce la: $q(\tilde{\lambda}, \tilde{\mu}) \leq f(x^*) = f^*$. □

Teorema 9.2.1 (Concavitățile funcției duale) *Funcția duală $q : \mathbb{R}^m \times \mathbb{R}^p \rightarrow \bar{\mathbb{R}}$ este întotdeauna funcție concavă.*

Demonstrație: Se observă că Lagrangianul $\mathcal{L}(x, \cdot, \cdot)$ este o funcție afină în multiplicatorii (λ, μ) pentru x fixat. Fie $\alpha \in [0, 1]$, atunci pentru (λ_1, μ_1) și (λ_2, μ_2) avem:

$$\begin{aligned} & q(\alpha\lambda_1 + (1 - \alpha)\lambda_2, \alpha\mu_1 + (1 - \alpha)\mu_2) \\ &= \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \alpha\lambda_1 + (1 - \alpha)\lambda_2, \alpha\mu_1 + (1 - \alpha)\mu_2) \\ &= \inf_{x \in \mathbb{R}^n} \alpha\mathcal{L}(x, \lambda_1, \mu_1) + (1 - \alpha)\mathcal{L}(x, \lambda_2, \mu_2) \\ &\geq \alpha \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda_1, \mu_1) + (1 - \alpha) \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda_2, \mu_2) \\ &= \alpha q(\lambda_1, \mu_1) + (1 - \alpha)q(\lambda_2, \mu_2). \end{aligned}$$

Din definiția concavității rezultă că funcția duală q este concavă. \square

O întrebare naturală ar fi următoarea: care este cea mai bună margine inferioară ce poate fi obținută dintr-o funcție duală? Răspunsul este simplu: se obține prin maximizarea dualei după toate valorile fezabile ale multiplicatorilor, rezultând astfel așa-numita *problemă duală*.

Definiția 9.2.2 (Problema duală) *Problema duală este definită ca fiind problema de maximizare concavă a funcției duale:*

$$q^* = \max_{\lambda \geq 0, \mu \in \mathbb{R}^p} q(\lambda, \mu), \quad (9.6)$$

unde notăm cu q^* valoarea optimă duală.

Este interesant de observat că problema duală este întotdeauna problemă convexă chiar dacă problema primală (UNLP) nu este convexă. Definim *mulțimea fezabilă duală*

$$\Omega = \mathbb{R}_+^m \times \mathbb{R}^p.$$

Ca o consecință imediată a ultimei leme, obținem următorul rezultat fundamental numit dualitatea slabă:

Teorema 9.2.2 (Dualitate slabă) *Următoarea inegalitate are loc pentru orice problemă de optimizare (NLP):*

$$q^* \leq f^*. \quad (9.7)$$

Se observă că dacă există x^* fezabil pentru problema primală și (λ^*, μ^*) fezabil pentru problema duală astfel încât $q(\lambda^*, \mu^*) = f(x^*)$ atunci x^* este punct de minim global pentru problema primală și (λ^*, μ^*) este punct de maxim global pentru problema duală. Mai mult, dacă problema primală este nemărginită inferior (adică $f^* = -\infty$), atunci $q(\lambda, \mu) = -\infty$ pentru orice $(\lambda, \mu) \in \Omega$ (adică pentru orice pereche duală fezabilă). De asemenea, dacă $q^* = \infty$, atunci problema primală este infeasibilă.

Interpretarea geometrică: Dăm o interpretare simplă a funcției duale și a dualității slabe în termeni geometrici. Pentru a vizualiza grafic, considerăm un caz particular al problemei (NLP) de forma $\min_{x \in \mathbb{R}^n} \{f(x) : g(x) \leq 0\}$, având o singură constrângere de inegalitate. Definim mulțimea

$$S = \{(u, t) : \exists x \in \mathbb{R}^n, f(x) = t, g(x) = u\}.$$

Deoarece fezabilitatea cere ca $g(x) \leq 0$, problema primală presupune găsirea celui mai de jos punct al lui S situat în partea stângă a axei verticale (vezi Fig. 9.4).

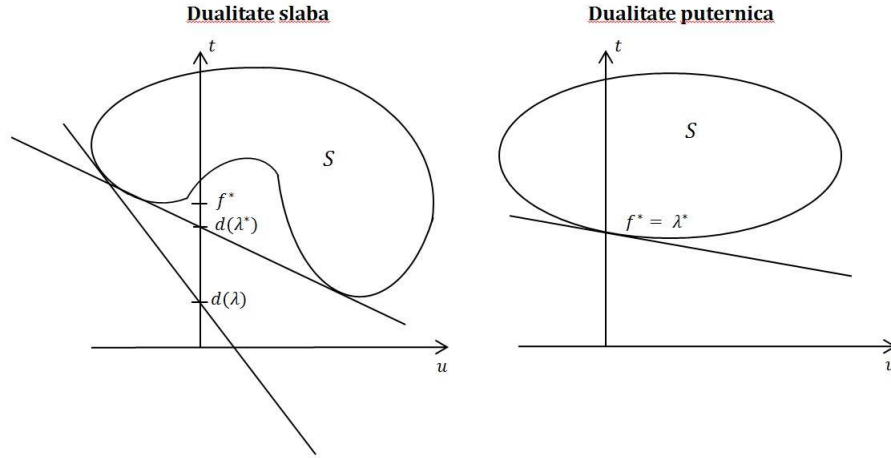


Figura 9.4: Interpretarea geometrică a dualității: dualitatea slabă (stânga) și dualitatea puternică (dreapta).

Este clar că pentru un scalar λ dat funcția duală se obține din următoarea problemă de minimizare:

$$q(\lambda) = \min_{(u,t) \in S} \lambda u + t.$$

În concluzie, observăm că inegalitatea:

$$\lambda^T u + t \geq q(\lambda)$$

definește un hiperplan suport pentru mulțimea S definit de vectorul $[\lambda \ 1]^T$ și mai mult, intersecția acestui hiperplan cu axa verticală (i.e. pentru $u = 0$) dă $q(\lambda)$.

Diferența $f^* - q^*$ se numește *duality gap*. Se observă că dualitatea slabă este valabilă pentru orice problemă de optimizare (NLP), însă în anumite cazuri (e.g. optimizarea convexă în care mulțimea fezabilă îndeplinește condiții speciale) există o versiune mai puternică a dualității, numită dualitatea puternică. Pentru a obține dualitatea puternică avem nevoie de anumite proprietăți de convexitate pentru problema (NLP):

Condiția Slater: Presupunem că problema (NLP) este convexă (i.e. f și g_1, \dots, g_m sunt funcții convexe, iar h_1, \dots, h_p sunt funcții afine) și că există $\bar{x} \in \mathbb{R}^n$ fezabil astfel încât $g(\bar{x}) < 0$ și $h(\bar{x}) = 0$. Pentru a demonstra dualitatea puternică vom utiliza teorema de separare prin hiperplane:

Teorema 9.2.3 (Dualitatea puternică) *Dacă problema de optimizare convexă primală (NLP) satisface condiția Slater, atunci valorile optime pentru problemele primale și duale sunt egale, adică:*

$$q^* = f^* \quad (9.8)$$

și mai mult $(\lambda^*)^T g(x^*) = 0$, unde x^* este punct de minim global pentru problema primală și (λ^*, μ^*) este punct de maxim global pentru problema duală.

Demonstrație: Introducem următoarea mulțime convexă $S_1 \subseteq \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}$ definită explicit sub forma:

$$S_1 = \{(u, v, t) : \exists x \in \mathbb{R}^n, g_i(x) \leq u_i \forall i, h_j(x) = v_j \forall j, f(x) \leq t\}.$$

Deoarece h este funcție afină există matricea $A \in \mathbb{R}^{p \times n}$ și $b \in \mathbb{R}^p$ astfel încât $h(x) = Ax - b$. Presupunem de asemenea că $\text{rang}(A) = p$ și că f^* este finit. Definim o a doua mulțime convexă:

$$S_2 = \{(0, 0, s) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R} : s < f^*\}.$$

Se observă imediat că mulțimile S_1 și S_2 sunt convexe și nu se intersectează. Din teorema de separare prin hiperplane avem că există $(\tilde{\lambda}, \tilde{\mu}, \nu) \neq 0$ și $\alpha \in \mathbb{R}$ astfel încât:

$$\tilde{\lambda}^T u + \tilde{\mu}^T v + \nu t \geq \alpha \quad \forall (u, v, t) \in S_1$$

și

$$\tilde{\lambda}^T u + \tilde{\mu}^T v + \nu t \leq \alpha \quad \forall (u, v, t) \in S_2.$$

Din prima inegalitate se observă că $\tilde{\lambda} \geq 0$ și $\nu \geq 0$ (altfel $\tilde{\lambda}^T u + \nu t$ este nemărginită inferior peste mulțimea S_1). A doua inegalitate implică $\nu t \leq \alpha$ pentru orice $t < f^*$ și deci $\nu f^* \leq \alpha$. Din această discuție putem concluziona că pentru orice $x \in \mathbb{R}^n$:

$$\nu f(x) + \tilde{\lambda}^T g(x) + \tilde{\mu}^T (Ax - b) \geq \alpha \geq \nu f^*.$$

Presupunem că $\nu > 0$. În acest caz, împărțind ultima inegalitate prin ν obținem:

$$\mathcal{L}(x, \tilde{\lambda}/\nu, \tilde{\mu}/\nu) \geq f^* \quad \forall x \in \mathbb{R}^n.$$

Introducând notațiile $\lambda = \tilde{\lambda}/\nu, \mu = \tilde{\mu}/\nu$ și minimizând după x în inegalitatea precedentă obținem $q(\lambda, \mu) \geq f^*$, ceea ce implică $q(\lambda, \mu) = f^*$, adică dualitatea puternică are loc în acest caz.

Dacă $\nu = 0$ avem că pentru orice $x \in \mathbb{R}^n$:

$$\tilde{\lambda}^T g(x) + \tilde{\mu}^T (Ax - b) \geq 0.$$

Aplicând această relație pentru vectorul Slater \bar{x} avem:

$$\tilde{\lambda}^T g(\bar{x}) \geq 0.$$

Dar știm că $g_i(\bar{x}) < 0$ și $\tilde{\lambda} \geq 0$, ceea ce conduce la $\tilde{\lambda} = 0$. Dar avem $(\tilde{\lambda}, \tilde{\mu}, \nu) \neq 0$, ceea ce implică $\tilde{\mu} \neq 0$. În concluzie, obținem că pentru orice $x \in \mathbb{R}^n$ avem $\tilde{\mu}^T (Ax - b) \geq 0$. Dar pentru vectorul Slater \bar{x} avem $\tilde{\mu}^T (A\bar{x} - b) = 0$ și deci există vectori $x \in \mathbb{R}^n$ astfel încât $\tilde{\mu}^T (Ax - b) < 0$, excepție făcând cazul în care $A^T \tilde{\mu} = 0$. Dar $A^T \tilde{\mu} = 0$ nu este posibil deoarece $\text{rang}(A) = p$ și $\tilde{\mu} \neq 0$. În concluzie, cazul $\nu = 0$ nu poate avea loc. \square

Condiția Slater poate fi relaxată când anumite constrângeri de inegalitate g_i sunt funcții afine. De exemplu, dacă primele $r \leq m$ constrângeri de inegalitate sunt descrise de funcțiile g_1, \dots, g_r afine, atunci dualitatea puternică are loc dacă următoarea condiție Slater relaxată este satisfăcută: funcțiile f și g_{r+1}, \dots, g_m sunt funcții convexe, iar funcțiile g_1, \dots, g_r și h_1, \dots, h_p sunt funcții afine și există \bar{x} astfel încât $g_\ell(\bar{x}) \leq 0$ pentru $\ell = 1, \dots, r$, $g_\ell(\bar{x}) < 0$ pentru $\ell = r + 1, \dots, m$ și $h(\bar{x}) = 0$.

Interpretarea minimax: Se observă că dualitatea puternică poate fi prezentată utilizând teorema minimax (vezi Apendice): dacă următoarea relație are loc

$$\inf_{x \in \mathbb{R}^n} \sup_{\lambda \geq 0} \mathcal{L}(x, \lambda, \mu) = \sup_{\lambda \geq 0} \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda, \mu),$$

atunci $q^* = f^*$. Într-adevăr, observăm că partea dreaptă a acestei relații este problema duală. Pe de altă parte, expresia din stânga, la prima vedere, nu are legătură cu problema primală. Dar se observă că funcția în x definită ca valoarea optimă a problemei de maximizare $\sup_{\lambda \geq 0} \mathcal{L}(x, \lambda, \mu)$ este finită dacă $g(x) \leq 0$ și $h(x) = 0$, iar în acest caz $\sup_{\lambda \geq 0} \mathcal{L}(x, \lambda, \mu) = f(x)$, ceea ce ne conduce la problema primală. În concluzie, dacă relația minimax anterioară este validă avem dualitate puternică. Deci dualitatea puternică poate avea loc și pentru probleme (NLP) neconvexe care satisfac egalitatea minimax precedentă.

Exemplul 9.2.1 *Un exemplu de problema de optimizare neconvexă, des întâlnită în teoria sistemelor și control, pentru care dualitatea puternică are loc este următorul:*

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \frac{1}{2} x^T Q x + q^T x + r \\ \text{s.l.:} \quad & \frac{1}{2} x^T Q_1 x + q_1^T x + r_1 \leq 0, \end{aligned}$$

unde matricele simetrice Q și Q_1 nu sunt pozitiv semidefinite. Deci această problemă de optimizare cu funcție obiectiv pătratică și o singură constrângere de inegalitate descrisă de asemenea de o funcție pătratică nu este convexă. Se poate arăta că dualitatea puternică are loc pentru această problemă sub ipoteza că există \bar{x} pentru care inegalitatea este strictă, adică $\frac{1}{2} \bar{x}^T Q_1 \bar{x} + q_1^T \bar{x} + r_1 < 0$.

În concluzie, dualitatea puternică are loc și pentru probleme particulare neconvexe (NLP). În toate aceste situații, dualitatea puternică ne permite să reformulăm o problemă de optimizare (NLP) într-o problemă echivalentă duală, dar care este întotdeauna problemă convexă (deoarece duala este funcție concavă). Pentru a înțelege mai bine reformularea duală a unei probleme (NLP), vom prezenta următorul exemplu:

Exemplul 9.2.2 (Duala unei probleme QP strict convexă) Fie problema QP strict convexă de forma:

$$\begin{aligned} f^* &= \min_{x \in \mathbb{R}^n} \frac{1}{2} x^T Q x + q^T x \\ \text{s.l.: } & Cx - d \leq 0, \quad Ax - b = 0. \end{aligned}$$

Presupunem că $Q \succ 0$ și că mulțimea fezabilă $X = \{x \in \mathbb{R}^n : Cx - d \leq 0, Ax - b = 0\}$ este nevidă. Din expunerea anterioară avem că în acest caz dualitatea puternică are loc. Lagrangianul este dat de următoarea expresie:

$$\begin{aligned} \mathcal{L}(x, \lambda, \mu) &= \frac{1}{2} x^T Q x + q^T x + \lambda^T (Cx - d) + \mu^T (Ax - b) \\ &= -\lambda^T d - \mu^T b + \frac{1}{2} x^T Q x + (q + C^T \lambda + A^T \mu)^T x. \end{aligned}$$

Funcția duală este infimumul neconstrâns al Lagrangianului în funcție de variabila x , Lagrangian ce este o funcție pătratică de x . Obținem că duala are forma:

$$\begin{aligned} q(\lambda, \mu) &= -\lambda^T d - \mu^T b + \inf_{x \in \mathbb{R}^n} \left(\frac{1}{2} x^T Q x + (q + C^T \lambda + A^T \mu)^T x \right) \\ &= -\lambda^T d - \mu^T b - \frac{1}{2} (q + C^T \lambda + A^T \mu)^T Q^{-1} (q + C^T \lambda + A^T \mu), \end{aligned}$$

unde în ultima egalitate am utilizat rezultate de bază pentru optimizarea neconstrânsă convexă pătratică. Se observă că funcția duală este de asemenea pătratică în variabilele duale (λ, μ) . Mai mult, funcția duală este concavă deoarece Hessiana este negativ semidefinită. Astfel, problema de optimizare duală a unui QP strict convex este dată de expresia:

$$\begin{aligned} q^* &= \max_{\lambda \geq 0, \mu \in \mathbb{R}^p} -\frac{1}{2} \begin{bmatrix} \lambda \\ \mu \end{bmatrix}^T \begin{bmatrix} C \\ A \end{bmatrix} Q^{-1} \begin{bmatrix} C \\ A \end{bmatrix}^T \begin{bmatrix} \lambda \\ \mu \end{bmatrix} \\ &\quad - \begin{bmatrix} d + C Q^{-1} q \\ b + A Q^{-1} q \end{bmatrix}^T \begin{bmatrix} \lambda \\ \mu \end{bmatrix} - \frac{1}{2} q^T Q^{-1} q. \end{aligned} \tag{9.9}$$

Datorită faptului că funcția obiectiv este concavă, această problemă duală este ea însăși un QP convex, dar în general nu este strict convex (adică Hessiana nu mai este pozitiv definită). Însă formularea QP duală dată

de (9.9) are constrângeri mult mai simple, adică mulțimea fezabilă este descrisă de constrângeri foarte simple: $\lambda \geq 0$ și $\mu \in \mathbb{R}^p$. Observăm că ultimul termen în funcția duală este o constantă, care trebuie însă păstrată pentru ca $q^* = f^*$, adică dualitatea puternică să fie menținută.

9.3 Programare liniară (LP)

Programarea liniară ocupă un loc deosebit de important, atât în teorie cât și în aplicațiile practice din inginerie, economie, etc. Reamintim că o problemă de programare liniară (LP) are următoarea formă:

$$\begin{aligned} f^* &= \min_{x \in \mathbb{R}^n} c^T x \\ \text{s.l.: } & Cx - d \leq 0, \quad Ax - b = 0. \end{aligned}$$

Exemplul 9.3.1 (Dieta economică) Dorim să determinăm o dietă cât mai puțin costisitoare care să acopere însă în totalitate substanțele nutritive necesare organismului uman (această aplicație aparține clasei de probleme de alocare a resurselor, de exemplu dieta unei armate). Presupunem că există pe piață n alimente care se vând la prețul c_i pe bucată și, de asemenea, există m ingrediente nutriționale de bază pe care fiecare om trebuie să le consume într-o cantitate de minim d_j unități. Mai știm, de asemenea, că fiecare aliment i conține c_{ji} unități din elementul nutrițional j .

Problema este să se determine numărul de unități din alimentul i , notat x_i , care să minimizeze costul total și în același timp să satisfacă constrângerile nutriționale, adică avem următoarea problemă de optimizare (LP):

$$\begin{aligned} \min_{x \in \mathbb{R}^n} & c_1 x_1 + \cdots + c_n x_n \\ \text{s.l.: } & c_{11} x_1 + \cdots + c_{1n} x_n \geq d_1 \\ & \dots \\ & c_{m1} x_1 + \cdots + c_{mn} x_n \geq d_m \\ & x_i \geq 0 \quad \forall i = 1, \dots, n. \end{aligned}$$

Exemplul 9.3.2 La o problemă de programare operativă a producției restricțiile se referă la o serie de mașini (utilaje) cu care se execută produsele dorite, d_i fiind timpul disponibil utilajului i pe perioada analizată, iar c_{ij} timpul necesar prelucrării unui produs de tipul j pe

utilajul i , scopul fiind maximizarea producției. Ca urmare, problema se pune ca un (LP), unde x_i reprezintă numărul de unități de produs i pe perioada analizată:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & x_1 + \cdots + x_n \\ \text{s.l.:} \quad & c_{j1}x_1 + \cdots + c_{jn}x_n \leq d_j \quad \forall j = 1, \dots, m \\ & x_i \geq 0 \quad \forall i = 1, \dots, n. \end{aligned}$$

Ideea de bază în programarea liniară este că trebuie să căutăm soluția problemei într-o mulțime cu un număr finit de soluții de bază care sunt punctele de extrem (vârfurile) ale poliedrului ce definește mulțimea fezabilă:

$$X = \{x \in \mathbb{R}^n : Cx - d \leq 0, \quad Ax - b = 0\}.$$

Enunțăm această teoremă pentru cazul în care mulțimea fezabilă este mărginită, adică este un politop:

Teorema 9.3.1 *Presupunem că mulțimea fezabilă X este un politop, atunci există un punct de minim al problemei (LP) într-unul din vârfurile politopului.*

Demonstrație: Dacă mulțimea fezabilă X este politop, atunci X este acoperirea convexă generată de vârfurile politopului:

$$X = \text{Conv}(\{v_1, \dots, v_q\}).$$

Mai mult, din faptul că X este mărginită avem că un punct de minim x^* există pentru problema (LP). Deoarece x^* este fezabil avem:

$$x^* = \sum_{i=1}^q \alpha_i v_i,$$

unde $\alpha_i \geq 0$ și $\sum_{i=1}^q \alpha_i = 1$. Este clar că $c^T v_i \geq f^*$, deoarece v_i este fezabil pentru orice i . Notăm cu \mathcal{I} mulțimea de indecși definită astfel: $\mathcal{I} = \{i : \alpha_i > 0\}$. Dacă există $i_0 \in \mathcal{I}$ astfel încât $c^T v_{i_0} > f^*$, atunci:

$$f^* = c^T x^* = \alpha_{i_0} c^T v_{i_0} + \sum_{i \in \mathcal{I} \setminus \{i_0\}} \alpha_i c^T v_i > \sum_{i \in \mathcal{I}} \alpha_i f^* = f^*$$

și deci obținem o contradicție. Aceasta implică că orice vârf pentru care $\alpha_i > 0$ este un punct de minim. \square

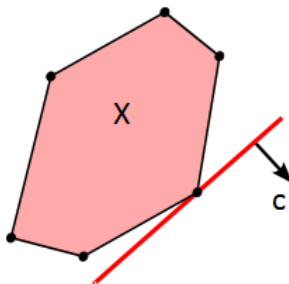


Figura 9.5: Soluția unui LP.

Din teorema anterioară se poate observa că pentru a găsi o soluție optimă pentru problema (LP) este suficient să determinăm vârfurile politopului ce descriu mulțimea fezabilă X , să evaluăm apoi funcția obiectiv în aceste vârfuri și să considerăm soluția corespunzătoare celei mai mici valori (vezi Fig. 9.5). Se poate observa că în anumite cazuri această metodă nu este eficientă, deoarece există multe clase de mulțimi de tip politop des întâlnite în aplicații pentru care numărul de vârfuri este exponențial, de exemplu politopul $X = \{x \in \mathbb{R}^n : \|x\|_\infty \leq 1\}$ are 2^n vârfuri. Pentru o astfel de problemă, la $n = 100$ de variabile, avem nevoie să căutăm soluția printre $2^{100} \simeq 10^{30}$ vârfuri, ceea ce presupune un efort de calcul aproape imposibil de realizat de către calculatoarele actuale. Există însă metode alternative mai eficiente pentru rezolvarea unui (LP).

O metodă modernă fundamentală de rezolvare a problemelor de optimizare (LP) este *algoritmul simplex*. Acest algoritm a fost introdus pentru prima dată de către matematicianul George B. Dantzig, în 1947. Algoritmul se bazează pe noțiunea de *soluție fundamentală a unui sistem de ecuații*. Se poate arăta că un (LP) general poate fi întotdeauna scris în *forma standard*:

$$\min_{x \in \mathbb{R}^n} \{c^T x : Ax = b, x \geq 0\},$$

prin folosirea de *variabile suplimentare* (numite și *variabile artificiale*). Într-adevar, observăm următoarele:

(i) orice restricție de inegalitate poate fi transformată în egalitate, prin introducerea unei variabile suplimentare care nu este negativă și folosind relațiile:

$$x \leq d \iff x + y = d, y \geq 0 \quad \text{și} \quad x \geq d \iff x - y = d, y \geq 0.$$

(ii) orice variabilă fără restricție de semn poate fi înlocuită cu două variabile cu restricție de semn pozitiv, folosind relația:

$$x \text{ oarecare} \iff x = y - z, \quad y \geq 0, \quad z \geq 0.$$

Folosind aceste două transformări putem aduce orice problemă (LP) în forma (LP) standard anterioară.

Pentru problema (LP) standard presupunem că matricea $A \in \mathbb{R}^{p \times n}$ are rangul $p < n$ (adică numărul de ecuații este mai mic decât numărul de variabile și deci avem suficiente grade de libertate pentru a optimiza). Fie o matrice $B \in \mathbb{R}^{p \times p}$ nesingulară (numită și matrice de bază) formată din coloanele lui A și fie x_k soluția unică a sistemului de ecuații $Bx_k = d$. Definim *soluția fundamentală* a sistemului $Ax = b$, vectorul $\bar{x}_k \in \mathbb{R}^n$ obținut extinzând x_k cu zerourile corespunzătoare componentelor ce nu sunt asociate coloanelor lui B . Definim de asemenea *soluțiile fundamentale fezabile*, adică soluțiile fundamentale \bar{x}_k care satisfac în plus constrângerea $\bar{x}_k \geq 0$. Observăm că vârfurile mulțimii fezabile pentru problema (LP) standard, adică vârfurile poliedrului $X = \{x : \mathbb{R}^n : Ax = b, x \geq 0\}$, sunt de fapt soluțiile fundamentale fezabile și reciproc.

Exemplul 9.3.3 Considerăm polipul (numit adesea și simplex) $X = \{x \in \mathbb{R}^3 : x \geq 0, x_1 + x_2 + x_3 = 1\}$. Observăm că vârfurile acestui polip coincid cu cele trei soluții de bază ale ecuației $x_1 + x_2 + x_3 = 1$ (vezi Fig. 9.6).

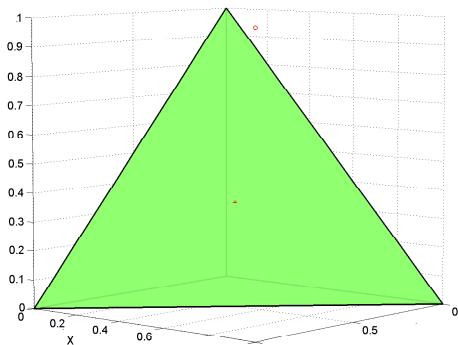


Figura 9.6: Vârfurile unui simplex în \mathbb{R}^3 .

Se poate arăta că o soluție optimală a problemei (LP) în forma standard (în cazul în care aceasta există) se găsește printre *soluțiile fundamentale*

fezabile. Acest rezultat este consecința Teoremei 9.3.1, observând că vârfurile mulțimii fezabile $X = \{x : \mathbb{R}^n : Ax = b, x \geq 0\}$ coincid cu soluțiile fundamentale fezabile. Aceasta ne permite să căutăm soluția optimă a problemei (LP) în submulțimea soluțiilor fundamentale care sunt cel mult $\frac{n!}{p!(n-p)!}$ la număr (corespunzătoare diverselor modalități de a alege p coloane din n coloane). Ideea de bază în metoda simplex este următoarea: pornind de la o soluție fundamentală fezabilă găsim o nouă soluție fundamentală fezabilă în care funcția obiectiv să descrească și această căutare se face folosind *tabelul simplex*, care deși necesită o matematică extrem de simplă nu se poate exprima ușor într-o formă matriceală compactă.

A durat mult timp până s-a demonstrat că algoritmul simplex standard nu are complexitate polinomială. Un exemplu fiind clasa de probleme de mai jos, găsită de Klee și Minty în 1972, în care algoritmul trebuie să analizeze 2^n baze (n numărul de necunoscute) până la găsirea celei optime:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \sum_{i=1}^n 10^{n-i} x_i \\ \text{s.l.:} \quad & \left(2 \sum_{j=1}^{i-1} 10^{i-j} x_j \right) + x_i \leq 100^{i-1} \quad \forall i = 1, \dots, n \\ & x_i \geq 0 \quad \forall i = 1, \dots, n. \end{aligned}$$

Pentru o astfel de problemă, la 100 de variabile, algoritmul va avea $2^{100} \simeq 10^{30}$ iterații, și chiar la o viteză de un miliard iterații pe secundă (mult peste puterea unui calculator actual) va termina în 10^{13} ani. Nu se știe încă dacă există sau nu o altă modalitate de trecere de la o bază la alta, folosind tabelele simplex, prin care algoritmul simplex standard să devină polinomial. Au fost însă găsiți algoritmi alternativi care nu se bazează pe tabele simplex, primul de acest gen fiind algoritmul de punct interior al lui Karmakar, despre care s-a demonstrat că are complexitate polinomială.

În ciuda dezavantajelor amintite, algoritmul simplex rămâne și în zilele noastre cel mai eficient algoritm în ceea ce privește viteza de lucru, simplitatea și implementarea pe calculator. Mai mult, folosirea acestuia aduce informații mult mai ample decât găsirea soluției propriu-zise, este mult mai maleabil în cazul modificărilor ulterioare ale datelor problemei și se pretează mult mai bine la interpretări economice. Un argument

în plus în favoarea acestui algoritm este acela că încă nu a apărut o problemă practică în fața căruia să clacheze. Algoritmii de punct interior rămân doar ca alternative teoretice sau pentru cazurile în care algoritmul simplex este lent, dar ei nu-l pot înlocui complet.

Funcția Lagrange asociată unui (LP) general este dată de expresia:

$$\begin{aligned}\mathcal{L}(x, \lambda, \mu) &= c^T x + \lambda^T (Cx - d) + \mu^T (Ax - b) \\ &= -\lambda^T d - \mu^T b + (c + C^T \lambda + A^T \mu)^T x.\end{aligned}$$

Observăm că Lagrangianul este de asemenea liniar în variabila x . Atunci funcția duală corespunzătoare este dată de expresia:

$$\begin{aligned}q(\lambda, \mu) &= -\lambda^T d - \mu^T b + \inf_{x \in \mathbb{R}^n} (c + C^T \lambda + A^T \mu)^T x \\ &= -\lambda^T d - \mu^T b + \begin{cases} 0 & \text{dacă } c + C^T \lambda + A^T \mu = 0 \\ -\infty & \text{altfel.} \end{cases}\end{aligned}$$

Astfel, funcția obiectiv duală $q(\lambda, \mu)$ este de asemenea liniară și ia valoarea $-\infty$ în toate punctele ce nu satisfac egalitatea liniară $c + C^T \lambda + A^T \mu = 0$. Din moment ce vrem să maximizăm funcția duală, aceste puncte pot fi privite ca puncte nefezabile a problemei duale (de aceea, le numim *dual nefezabile*), și putem scrie în mod explicit duala LP-ului precedent ca:

$$\begin{aligned}q^* &= \max_{\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p} \begin{bmatrix} -d \\ -b \end{bmatrix}^T \begin{bmatrix} \lambda \\ \mu \end{bmatrix} \\ \text{s.l.: } &\lambda \geq 0, \quad c + C^T \lambda + A^T \mu = 0.\end{aligned}$$

Se observă că problema duală este de asemenea un (LP). În anumite situații, problema (LP) duală este mai simplă decât problema (LP) primală (e.g. constrângerile problemei duale sunt mai simple decât ale problemei primale) și deci în acest caz este de preferat rezolvarea dualei. Raționând ca mai înainte, problema primală și duală (LP) standard au următoarea formă:

$$\textbf{Primala:} \min_{x \in \mathbb{R}^n} \{c^T x : Ax = b, x \geq 0\} \quad \textbf{Duala:} \max_{\mu \in \mathbb{R}^n} \{b^T \mu : A^T \mu \leq c\}$$

Din dualitatea slabă avem valabilă inegalitatea:

$$b^T \mu \leq c^T x$$

pentru orice x și μ fezabile pentru problema primală și respectiv duală. De asemenea, se poate arăta următoarea teoremă, cunoscută sub numele de *teorema de dualitate pentru programarea liniară*:

Teorema 9.3.2 (Teorema de dualitate pentru (LP)) *Dacă una dintre problemele (LP), primală sau duală, are soluție optimă atunci și cealaltă problemă are soluție optimă și valorile optime corespunzătoare sunt egale. Mai mult, dacă una dintre probleme, primală sau duală, are funcție obiectiv nemărginită atunci cealaltă problemă nu are puncte fezabile.*

O consecință imediată a acestei teoreme și dualitatea slabă este lema Farkas (sau alternativei): fie $A \in \mathbb{R}^{p \times n}$ și $b \in \mathbb{R}^p$, atunci una și numai una din următoarele relații are loc:

- (i) există $x \in \mathbb{R}^n$ astfel încât $Ax = b$ și $x \geq 0$;
- (ii) există $\mu \in \mathbb{R}^p$ astfel încât $A^T \mu \geq 0$ și $b^T \mu < 0$.

Lema Farkas are foarte multe aplicații, e.g. poate fi folosită în programarea liniară, în teoria jocurilor sau în derivarea condițiilor de optimalitate de ordinul întâi pentru probleme de optimizare (NLP) generale.

Capitolul 10

Condiții de optimalitate pentru (NLP)

În acest capitol vom defini condițiile necesare și suficiente de optimalitate pentru cazul problemelor constrânse. Vom arăta că aceste condiții de optimalitate pot fi privite ca o generalizare a cazului neconstrâns la cel constrâns în care în locul funcției obiectiv folosim Lagrangianul. Reamintim problema (NLP) în forma standard:

$$(NLP) : \quad \min_{x \in \mathbb{R}^n} f(x) \quad (10.1) \\ \text{s.l.: } g(x) \leq 0, \quad h(x) = 0,$$

în care funcțiile $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ și $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$ sunt funcții diferențiabile de două ori. Mulțimea fezabilă a problemei (NLP) este mulțimea punctelor ce satisfac contrângerile aferente, adică $X = \{x \in \mathbb{R}^n : g(x) \leq 0, h(x) = 0\}$. Cu aceste notații, putem rescrie problema (NLP) într-o formă compactă:

$$\min_{x \in X} f(x).$$

Pentru această problemă constrânsă (NLP) vom defini condițiile necesare și suficiente de optimalitate. Primul rezultat se referă la următoarea problemă de optimizare constrânsă (demonstrația acestui rezultat a fost dată în Capitolul 4, Teorema 4.3.1):

Teorema 10.0.3 (Condiții de ordinul I pentru (NLP) având constrângeri convexe) Fie X o mulțime convexă și o funcție $f \in \mathcal{C}^1$ (nu neapărat convexă). Pentru problema de optimizare constrânsă

$\min_{x \in X} f(x)$ următoarele condiții de optimalitate sunt satisfăcute: dacă x^* este minim local atunci:

$$\nabla f(x^*)^T (x - x^*) \geq 0 \quad \forall x \in X.$$

Dacă în plus f este funcție convexă atunci x^* este punct de minim dacă și numai dacă:

$$\nabla f(x^*)^T (x - x^*) \geq 0 \quad \forall x \in X.$$

Punctele x^* ce satisfac inegalitatea anterioară se numesc *puncte staționare* pentru problema (NLP) cu constrângeri convexe. Înainte să continuăm cu definirea altor condiții de optimalitate mai generale, vom avea nevoie să introducem noțiunea de constrângere activă/inactivă.

Definiția 10.0.1 (Constrângere activă/inactivă) O constrângere de inegalitate $g_i(x) \leq 0$ se numește activă în punctul fezabil $x \in X$ dacă și numai dacă $g_i(x) = 0$, altfel ea se numește inactivă. Desigur, orice constrângere de egalitate $h_i(x) = 0$ este activă într-un punct fezabil.

Definiția 10.0.2 (Mulțimea activă) Mulțimea de indecși, notată $\mathcal{A}(x) \subseteq \{1, \dots, m\}$, corespunzătoare constrângerilor active este numită mulțimea activă în punctul $x \in X$.

Considerarea constrângerilor active este esențială deoarece într-un punct fezabil x acestea restricționează domeniul de fezabilitate aflat într-o vecinătate a lui x , în timp ce constrângerile inactive nu influențează această vecinătate. În particular, se poate observa ușor că dacă x^* este un punct de minim local al problemei (NLP), atunci x^* este de asemenea minim local pentru probleme de optimizare numai cu constrângeri de egalitate:

$$\begin{aligned} & \min_{x \in \mathbb{R}^n} f(x) \\ \text{s.l.: } & g_i(x) = 0 \quad \forall i \in \mathcal{A}(x^*), \quad h(x) = 0. \end{aligned}$$

Astfel, pentru studierea proprietăților unui punct de minim local ne putem rezuma la studierea constrângerilor active. Prezentăm mai întâi condițiile de optimalitate de ordinul întâi și doi pentru cazul când problema (NLP) are numai constrângeri de egalitate și apoi extindem aceste condiții la cazul general.

10.1 Condiții de ordinul I pentru (NLP) având constrângeri de egalitate

Pentru a defini condițiile de optimalitate necesare și suficiente de ordinul I și II pentru probleme NLP generale, mai întâi studiem condițiile de optimalitate pentru probleme NLP care au doar constrângeri de egalitate:

$$(NLPe) : \min_{x \in \mathbb{R}^n} f(x) \quad (10.2)$$

$$\text{s.l.: } h(x) = 0.$$

Observațiile obținute din această categorie de probleme, în care toate constrângerile sunt considerate active, vor fi utilizate ulterior pentru problemele (NLP) generale. Mai întâi însă, trebuie să definim anumite noțiuni ce se vor dovedi esențiale în analiza noastră. O curbă pe o suprafață S este o mulțime de puncte $x(t) \in S$ continuu parametrizate în t , pentru $a \leq t \leq b$. O curbă este diferențiabilă dacă $\dot{x}(t) = \frac{dx(t)}{dt}$ există și este de două ori diferențiabilă dacă $\ddot{x}(t)$ există. O curbă $x(t)$ trece prin punctul x^* dacă $x^* = x(t^*)$ pentru un t^* ce satisface $a \leq t^* \leq b$. Derivata curbei în x^* este desigur definită ca $\dot{x}(t^*)$. Acum, considerăm toate curbele diferențiabile aflate pe suprafața S , ce trec printr-un punct x^* . Planul tangent în $x^* \in S$ este definit ca mulțimea tuturor derivatelor acestor curbe diferențiabile în t^* , adică mulțimea tuturor vectorilor de forma $\dot{x}(t^*)$ definite de curbele $x(t) \in S$.

Pentru o funcție $h: \mathbb{R}^n \rightarrow \mathbb{R}^p$, cu $h(x) = [h_1(x) \dots h_p(x)]^T$ notăm Jacobianul sau prin $\nabla h(x)$, unde reamintim că $\nabla h(x)$ este o matrice $p \times n$ cu elementul $\frac{\partial h_i(x)}{\partial x_j}$ pe poziția (i, j) :

$$\nabla h(x) = \begin{bmatrix} \frac{\partial h_1(x)}{\partial x_1} & \dots & \frac{\partial h_1(x)}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial h_p(x)}{\partial x_1} & \dots & \frac{\partial h_p(x)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla h_1(x)^T \\ \vdots \\ \nabla h_p(x)^T \end{bmatrix} \quad (10.3)$$

Introducem acum un subspațiu:

$$M = \{d \in \mathbb{R}^n : \nabla h(x^*)d = 0\}$$

și investigăm acum sub ce condiții acest subspațiu M este egal cu planul tangent în x^* la suprafața $S = \{x \in \mathbb{R}^n : h(x) = 0\}$. În acest scop trebuie să introducem noțiunea de punct regulat.

Definiția 10.1.1 (Punct regulat) *Un punct x^* ce satisface constrângerea $h(x^*) = 0$ se numește punct regulat dacă gradientii componentelor lui h , $\nabla h_1(x^*), \dots, \nabla h_p(x^*)$, sunt liniar independenți.*

De exemplu, dacă $h(x)$ este afină, adică $h(x) = Ax - b$ cu $A \in \mathbb{R}^{p \times n}$, atunci condiția de regularitate este echivalentă cu matricea A să aibă rangul egal cu p .

Teorema 10.1.1 *Într-un punct regulat x^* al suprafeței S definită de constrângerile de egalitate $h(x) = 0$, planul tangent este egal cu:*

$$M = \{d \in \mathbb{R}^n : \nabla h(x^*)d = 0\}.$$

Demonstrație: Notăm prin T planul tangent în x^* . Pentru ca $T = M$ trebuie să demonstrăm incluziunea dublă $T \subseteq M$ și $M \subseteq T$. Este clar că $T \subseteq M$, chiar dacă x^* este regulat sau nu, deoarece orice curbă $x(t)$ ce trece prin x^* la $t = t^*$, având derivata $\dot{x}(t^*)$ astfel încât $\nabla h(x^*)\dot{x}(t^*) \neq 0$ nu ar fi în S (ținem seama că $h(x(t)) = 0$ pentru orice $a \leq t \leq b$). Pentru a demonstra că $M \subseteq T$, trebuie să arătăm că pentru un $d \in M$ există o curbă în S ce trece prin x^* cu derivata d în t^* . Pentru a construi o asemenea curbă, considerăm ecuația:

$$h(x^* + td + \nabla h(x^*)^T u(t)) = 0$$

în care pentru un t fixat, considerăm $u(t) \in \mathbb{R}^p$ ca fiind necunoscută. Această ecuație este un sistem de p ecuații și p necunoscute, parametrizat în mod continuu prin t . La $t = 0$ avem soluția $u(0) = 0$. Jacobianul sistemului în funcție de u la $t = 0$ este matricea:

$$\nabla h(x^*)\nabla h(x^*)^T \in \mathbb{R}^{p \times p},$$

ce este nesingulară din moment ce x^* este un punct regulat și astfel $\nabla h(x^*)$ este de rang maxim. Ca urmare, prin teorema funcție implicite (vezi Apendice), există o soluție continuu diferențiabilă $u(t)$ într-o regiune $-a \leq t \leq a$. Curba $x(t) = x^* + td + \nabla h(x^*)^T u(t)$ este astfel, prin construcție, o curbă în S . Prin derivarea sistemului la $t = 0$ avem:

$$0 = \left. \frac{d}{dt} h(x(t)) \right|_{t=0} = \nabla h(x^*)d + \nabla h(x^*)\nabla h(x^*)^T \dot{u}(0).$$

Din definiția lui d avem $\nabla h(x^*)d = 0$ și astfel, din moment ce matricea $\nabla h(x^*)\nabla h(x^*)^T$ este nesingulară, tragem concluzia că $\dot{x}(0) = 0$. Astfel

$$\dot{x}(0) = d + \nabla h(x^*)^T \dot{x}(0) = d$$

iar curba construită are în x^* derivata d . \square

Exemplul 10.1.1 (Plan tangent) Fie constrângerea dată de funcția $h : \mathbb{R}^3 \rightarrow \mathbb{R}$, $h(x) = x_1^2 + x_2^2 + 3x_1 + 3x_2 + x_3 - 1$ și punctul $x^* = [0 \ 0 \ 1]^T$ astfel încât $h(x^*) = 0$. Jacobianul lui $h(x)$ va fi:

$$\nabla h(x) = [2x_1 + 3 \quad 2x_2 + 3 \quad 1]$$

iar $\nabla h(x^*) = [3 \ 3 \ 1]$, ceea ce arată că x^* este punct regulat. Astfel, din definiția planului tangent (conform teoremei anterioare), orice direcție tangentă $d = [d_1 \ d_2 \ d_3]^T$ va trebui să satisfacă

$$\nabla h(x^*)d = 0,$$

și anume $3d_1 + 3d_2 + d_3 = 0$. În Fig. 10.1 am reprezentat suprafața definită de $h(x) = 0$ și planul tangent în punctul $x^* = [0 \ 0 \ 1]^T$.

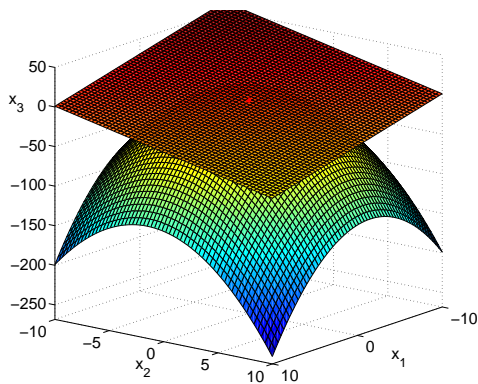


Figura 10.1: Suprafața pentru $h(x) = 0$ și planul tangent aferent punctului $x^* = [0 \ 0 \ 1]^T$.

Prin cunoașterea reprezentării planului tangent, derivarea condițiilor necesare și suficiente pentru ca un punct să fie un punct de minim local pentru probleme cu constrângeri de egalitate este destul de simplă.

Lema 10.1.1 Fie x^* un punct regulat al constrângerilor $h(x) = 0$ și punct de extrem local (minim sau maxim local) al problemei de optimizare (NLPe). Atunci, orice $d \in \mathbb{R}^n$ ce satisface:

$$\nabla h(x^*)d = 0$$

trebuie să satisfacă și:

$$\nabla f(x^*)^T d = 0.$$

Demonstrație: Fie un vector d în planul tangent în x^* și $x(t)$ fie orice curbă netedă pe suprafața de constrângere ce trece prin x^* cu derivata d , i.e. $x(0) = x^*$, $\dot{x}(0) = d$ și $h(x(t)) = 0$, cu $-a \leq t \leq a$ pentru un $a > 0$. Din moment ce x^* este un punct regulat, planul tangent este identic cu mulțimea de d -uri ce satisfac $\nabla h(x^*)d = 0$. Astfel, din moment ce x^* este punct de extrem local constrâns al lui f avem:

$$\left. \frac{d}{dt} f(x(t)) \right|_{t=0} = 0,$$

sau, în mod echivalent

$$\nabla f(x^*)^T d = 0.$$

□

Teorema 10.1.2 (Condiții necesare de ordinul I pentru NLPe)

Fie x^* un punct de extrem al funcției obiectiv f supusă la constrângerile $h(x) = 0$, i.e. al problemei de optimizare (NLPe), și presupunem că x^* este un punct regulat pentru aceste constrângeri. Atunci, există un multiplicator Lagrange $\mu^* \in \mathbb{R}^p$ astfel încât:

$$(\mathbf{KKT} - \mathbf{NLPe}) : \quad \nabla f(x^*) + \nabla h(x^*)^T \mu^* = 0 \quad \text{și} \quad h(x^*) = 0.$$

Demonstrație: Din Lema 10.1.1 tragem concluzia că valoarea optimă a LP-ului:

$$\begin{aligned} & \max_{d \in \mathbb{R}^n} \nabla f(x^*)^T d \\ & \text{s.l.: } \nabla h(x^*)d = 0 \end{aligned}$$

este zero. Astfel, din moment ce LP-ul are o valoare optimă finită, atunci din teorema dualității pentru LP, duala ei va fi fezabilă. În particular, există un $\mu^* \in \mathbb{R}^p$ astfel încât $\nabla f(x^*) + \nabla h(x^*)^T \mu^* = 0$. □

Punctele x^* pentru care există μ^* astfel încât condițiile (KKT-NLPe) sunt satisfăcute se numesc *puncte staționare* pentru problema (NLPe). Observăm că dacă exprimăm Lagrangianul asociat problemei constrânse:

$$\mathcal{L}(x, \mu) = f(x) + \mu^T h(x)$$

atunci condițiile necesare de ordinul I pot fi rescrise sub forma:

$$\begin{aligned} \nabla_x \mathcal{L}(x^*, \mu^*) &= 0 \\ \nabla_\mu \mathcal{L}(x^*, \mu^*) &= 0 \end{aligned}$$

sau echivalent sub forma:

$$\nabla \mathcal{L}(x^*, \mu^*) = 0.$$

După cum observăm, aceste condiții seamănă foarte mult cu condițiile de optimalitate de ordinul I pentru cazul neconstrâns (i.e. $\nabla f(x^*) = 0$). Pentru cazul constrâns, în locul funcției obiectiv f se consideră Lagrangianul \mathcal{L} . Condițiile de ordinul I se reduc la rezolvarea unui sistem $\nabla \mathcal{L}(x^*, \mu^*) = 0$ de $n + p$ ecuații (de obicei neliniare) cu $n + p$ necunoscute. Deci, acest sistem de ecuații ar trebui să permită, cel puțin local, determinarea unei soluții. Dar ca și în cazul neconstrâns, o soluție a sistemului dat de condițiile necesare de ordinul I nu este neapărat un minim (local) al problemei de optimizare; poate fi la fel de bine un maxim (local) sau un punct șa.

Exemplul 10.1.2 *Considerăm problema:*

$$\min_{x \in \mathbb{R}^2: h(x)=x_1^2+x_2^2-2=0} x_1 + x_2.$$

Mai întâi observăm că orice punct fezabil este regulat (punctul $x = [0 \ 0]^T$ nu este fezabil). În concluzie, orice minim local al acestei probleme satisface sistemul $\nabla \mathcal{L}(x, \mu) = 0$ care se poate scrie explicit astfel:

$$\begin{aligned} 2\mu x_1 &= -1 \\ 2\mu x_2 &= -1 \\ x_1^2 + x_2^2 &= 2. \end{aligned}$$

Aceste sistem de trei ecuații cu trei necunoscute x_1, x_2 și μ are următoarele două soluții: $(x_1^*, x_2^*, \mu^*) = (-1, -1, 1/2)$ și $(x_1^*, x_2^*, \mu^*) = (1, 1, -1/2)$. Se poate observa (vezi Fig. 10.2) că prima soluție este un minim local, în timp ce cealaltă soluție este un maxim local.

Este important să observăm că pentru ca un punct de minim să satisfacă condițiile de ordinul I este necesar să avem regularitate. Cu alte cuvinte, condițiile de optimalitate de ordinul I pot să nu aibă loc la un punct de minim local care nu este regulat.

Exemplul 10.1.3 *Considerăm problema:*

$$\begin{aligned} \min_{x \in \mathbb{R}^2} \quad & -x_1 \\ \text{s.l.:} \quad & h_1(x) = (1 - x_1)^3 + x_2 = 0, \quad h_2(x) = (1 - x_1)^3 - x_2 = 0. \end{aligned}$$

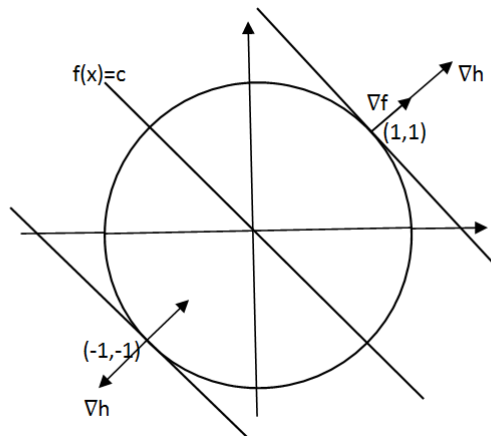


Figura 10.2: Condițiile de ordinul I.

Se observă că această problemă are un singur punct fezabil $x^* = [1 \ 0]^T$ care este de asemenea și minimul global. Pe de altă parte avem că $\nabla f(x^*) = [-1 \ 0]^T$, $\nabla h_1(x^*) = [0 \ 1]^T$ și $\nabla h_2(x^*) = [0 \ -1]^T$ și deci x^* nu este punct regulat. Se observă că în acest caz condițiile de optimalitate de ordinul I nu pot fi satisfăcute, adică nu există μ_1 și μ_2 astfel încât:

$$\mu_1 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \mu_2 \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Acest exemplu ilustrează că un punct de minim e posibil să nu satisfacă condițiile de staționaritate pentru Lagrangian dacă punctul nu este regulat.

10.2 Condiții de ordinul II pentru (NLP) având constrângeri de egalitate

În mod asemănător condițiilor de optimalitate de ordinul II definite pentru probleme de optimizare fără constrângeri, putem deriva condițiile corespunzătoare pentru probleme constrânse. Considerăm din nou probleme având constrângeri de tip egalitate (10.2):

Teorema 10.2.1 (Condiții necesare de ordinul II pentru NLPe)

Presupunem că x^* este un punct de minim local al problemei (NLPe) definită în (10.2) și un punct regulat pentru constrângerile aferente. Atunci există un $\mu^* \in \mathbb{R}^P$ astfel încât

$$\nabla f(x^*) + \nabla h(x^*)^T \mu^* = 0 \quad \text{si} \quad h(x^*) = 0.$$

În plus, dacă notăm prin M planul tangent în x^* $M = \{d \in \mathbb{R}^n : \nabla h(x^*)d = 0\}$, atunci matricea Hessiană a Lagrangianului în raport cu x

$$\nabla_x^2 \mathcal{L}(x^*, \mu^*) = \nabla^2 f(x^*) + \sum_{i=1}^p \mu_i^* \nabla^2 h_i(x^*)$$

este pozitiv semidefinită pe M , adică $d^T \nabla_x^2 \mathcal{L}(x^*, \mu^*) d \geq 0$ pentru orice $d \in M$.

Demonstrație Este clar că pentru orice curbă de două ori diferențiabilă pe suprafața de constrângeri S ce trece prin x^* (cu $x(0) = x^*$) avem:

$$\left. \frac{d^2}{dt^2} f(x(t)) \right|_{t=0} \geq 0. \quad (10.4)$$

Prin definiție avem:

$$\left. \frac{d^2}{dt^2} f(x(t)) \right|_{t=0} = \dot{x}(0)^T \nabla^2 f(x^*) \dot{x}(0) + \nabla f(x^*)^T \ddot{x}(0). \quad (10.5)$$

Mai mult, dacă derivăm relația $h(x(t))^T \mu^* = 0$ de două ori, obținem:

$$\dot{x}(0)^T \left(\sum_{i=1}^p \mu_i^* \nabla^2 h_i(x^*) \right) \dot{x}(0) + (\mu^*)^T \nabla h(x^*) \ddot{x}(0) = 0. \quad (10.6)$$

Adăugând (10.6) la (10.5) și ținând cont de (10.4), obținem:

$$\left. \frac{d^2}{dt^2} f(x(t)) \right|_{t=0} = \dot{x}(0)^T \nabla_x^2 \mathcal{L}(x^*, \mu) \dot{x}(0) \geq 0.$$

Din moment ce $\dot{x}(0)$ este arbitrar în M , atunci demonstrația este completă. \square

Exemplul 10.2.1 Considerăm problema de optimizare dată în Exemplul 10.1.2. Se observă că matricea Hessiană a funcției Lagrange în variabilă x este:

$$\nabla_x^2 \mathcal{L}(x, \mu) = \nabla^2 f(x) + \mu \nabla^2 h(x) = \mu \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix},$$

și o bază a planului tangent la un punct $x \neq 0$ este de forma $D(x) = [-x_2 \ x_1]^T$. Atunci, avem relația:

$$D(x)^T \nabla_x^2 \mathcal{L}(x, \mu) D(x) = 2\mu(x_1^2 + x_2^2).$$

Pentru prima soluție a sistemului rezultat din condiția de staționaritate a Lagrangianului avem $d_1^T \nabla_x^2 \mathcal{L}(-1, -1, 1/2) d_1 = 2 > 0$, unde $d_1 = D(-1, -1)$, și deci această soluție satisface condițiile necesare de ordinul II. Pe de altă parte, pentru cea de-a doua soluție avem expresia $d_2^T \nabla_x^2 \mathcal{L}(1, 1, -1/2) d_2 = -2 < 0$, unde $d_2 = D(1, 1)$, și deci această soluție nu poate fi minim local.

Teorema 10.2.2 (Condiții suficiente de ordinul II pentru NLPe)

Presupunem un punct $x^* \in \mathbb{R}^n$ și un $\mu^* \in \mathbb{R}^p$ astfel încât:

$$\nabla f(x^*) + \nabla h(x^*)^T \mu^* = 0 \quad \text{și} \quad h(x^*) = 0. \quad (10.7)$$

Presupunem de asemenea că matricea dată de Hessiana Lagrangianului $\nabla_x^2 \mathcal{L}(x^*, \mu^*) = \nabla^2 f(x^*) + \sum_{i=1}^p \mu_i^* \nabla^2 h_i(x^*)$ este pozitiv definită pe planul tangent $M = \{d : \nabla h(x^*) d = 0\}$. Atunci x^* este punct de minim local strict al problemei având constrângeri de egalitate (NLPe) definită în (10.2).

Demonstrație: Dacă x^* nu ar fi un punct de minim local strict, atunci ar exista un șir de puncte fezabile z_k ce converge către x^* astfel încât $f(z_k) \leq f(x^*)$. Putem scrie $z_k = x^* + \delta_k s_k$, unde $s_k \in \mathbb{R}^n$, $\|s_k\| = 1$, și $\delta_k > 0$. În mod clar $\delta_k \rightarrow 0$, iar șirul s_k fiind mărginit, va trebui să convergă către un $s^* \neq 0$. Avem de asemenea că $h(z_k) - h(x^*) = 0$ și prin împărțirea cu δ_k vom observa, pentru $k \rightarrow \infty$, că $\nabla h(x^*) s^* = 0$, adică s^* este vector tangent. Acum, prin Teorema lui Taylor, avem pentru orice j :

$$0 = h_j(z_k) - h_j(x^*) = \delta_k \nabla h_j(x^*) s_k + \frac{\delta_k^2}{2} s_k^T \nabla^2 h_j(\eta_j) s_k \quad (10.8)$$

și

$$0 \geq f(z_k) - f(x^*) = \delta_k \nabla f(x^*) s_k + \frac{\delta_k^2}{2} s_k^T \nabla^2 f(\eta_0) s_k, \quad (10.9)$$

unde η_j sunt puncte pe segmentul de dreaptă dintre x^* și z_k , și deci converge la x^* . Înmulțind acum ecuațiile (10.8) cu μ_j^* , adăugându-le la (10.9), și ținând cont de (10.7), obținem:

$$s_k^T \left(\nabla^2 f(\eta_0) + \sum_{i=1}^p \mu_i^* \nabla^2 h_i(\eta_i) \right) s_k \leq 0,$$

relație contradictorie pentru $k \rightarrow \infty$, deoarece s_k converge la vectorul tangent $s^* \neq 0$. Am ținut cont de faptul că η_j sunt convergente la x^* . \square

Putem iarăși concluziona că aceste condiții de ordinul II pentru cazul constrâns sunt similare celor corespunzătoare cazului neconstrâns. În cazul problemelor constrânse însă, în locul funcției obiectiv se folosește Lagrangianul.

Exemplul 10.2.2 *Considerăm problema:*

$$\min_{x \in \mathbb{R}^3: x_1+x_2+x_3=3} -x_1x_2 - x_1x_3 - x_2x_3.$$

Condițiile de ordinul I conduc la un sistem liniar de patru ecuații cu patru necunoscute:

$$\begin{aligned} -(x_2 + x_3) + \mu &= 0 \\ -(x_1 + x_3) + \mu &= 0 \\ -(x_1 + x_2) + \mu &= 0 \\ x_1 + x_2 + x_3 &= 3. \end{aligned}$$

Se poate observa ușor că $x_1^ = x_2^* = x_3^* = 1$ și $\mu^* = 2$ satisface acest sistem. Mai mult Hessiana Lagrangianului în orice punct x are forma:*

$$\nabla_x^2 \mathcal{L}(x, \mu) = \nabla^2 f(x) = \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix},$$

și o bază a planului tangent la suprafața definită de constrângerea $h(x) = x_1 + x_2 + x_3 - 3 = 0$ în orice punct x fezabil este:

$$D(x) = \begin{bmatrix} 0 & 2 \\ 1 & -1 \\ -1 & -1 \end{bmatrix}.$$

Obținem:

$$D(x^*)^T \nabla_x^2 \mathcal{L}(x^*, \mu^*) D(x^*) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \succ 0,$$

adică este pozitiv definită. În concluzie punctul x^* este punct de minim strict local. Interesant de observat este faptul că Hessiana funcției obiectiv evaluată în x^* este matrice indefinită.

Interpretarea multiplicatorilor Lagrange folosind senzitivitatea:

După cum am văzut, cu ajutorul multiplicatorilor Lagrange putem muta constrângerile în funcția obiectiv. O interpretare interesantă a multiplicatorilor Lagrange este dată cu ajutorul senzitivității (pentru mai multe detalii se poate consulta cartea clasică [5]). Pentru simplitate, considerăm o problemă (NLP) definită numai de egalități:

$$\min_{x \in \mathbb{R}^n} \{f(x) : h(x) = 0\}$$

și apoi asociem acesteia problema perturbată:

$$v(y) = \min_{x \in \mathbb{R}^n} \{f(x) : h(x) = y\}.$$

Fie x^* soluția optimă a problemei originale și fie $\chi(y)$ soluția optimă a problemei perturbate. Atunci avem că $v(0) = f(x^*)$ și $\chi(0) = x^*$. Mai mult, din identitatea $h(\chi(y)) = y$ pentru orice y , avem :

$$\nabla_y h(\chi(y)) = I_p = \nabla_x h(\chi(y))^T \nabla_y \chi(y).$$

Fie μ^* multiplicatorul Lagrange optim (soluția optimă duală) pentru problema originală neperturbată. Atunci,

$$\nabla_y v(0) = \nabla_x f(x^*) \nabla_y \chi(0) = -\mu^* \nabla_x h(x^*)^T \nabla_y \chi(0) = -\mu^*.$$

În concluzie, multiplicatorul Lagrange optim μ^* poate fi interpretat ca senzitivitatea funcției obiectiv f în raport cu constrângerea $h(x) = 0$. Altfel spus, μ^* indică cât de mult valoarea optimă s-ar schimba dacă constrângerea ar fi perturbată. Această interpretare poate fi extinsă la probleme generale (NLP) definite și de constrângeri de inegalitate. Multiplicatorii Lagrange optimi λ^* corespunzători unei constrângeri active $g(x) \leq 0$ pot fi interpretați ca senzitivitatea lui $f(x^*)$ în raport cu o perturbație în constrângeri de forma $g(x) \leq y$. În acest caz, pozitivitatea multiplicatorilor Lagrange urmează din faptul că prin creșterea lui y ,

mulțimea fezabilă este relaxată și deci valoarea optimă nu poate crește. Pentru inegalitățile inactive, interpretarea în termeni de senzitivitate explică de asemenea de ce multiplicatorii Lagrange sunt zero, pentru că o perturbație foarte mică în aceste constrângeri lasă valoarea optimă neschimbată.

10.3 Condiții de ordinul I pentru (NLP) generale

În această secțiune extindem condițiile de optimalitate de la cazul problemelor având constrângeri de egalitate la cel al problemelor de optimizare generale:

$$(NLP) : \quad \min_{x \in \mathbb{R}^n} f(x) \\ \text{s.l.: } g(x) \leq 0, \quad h(x) = 0.$$

Definiția 10.3.1 Fie un punct x^* ce satisface constrângerile problemei (NLP), adică avem $h(x^*) = 0$, $g(x^*) \leq 0$ și $\mathcal{A}(x^*)$ mulțimea constrângerilor active. Numim punctul x^* punct regulat dacă gradientii funcțiilor de constrângere, $\nabla h_i(x^*)$ pentru $i = 1, \dots, p$, și $\nabla g_j(x^*)$ pentru $j \in \mathcal{A}(x^*)$ sunt liniari independenți.

Teorema 10.3.1 (Condiții necesare de ordinul I pentru NLP)

Fie x^* un punct de minim local pentru problema NLP generală și presupunem că x^* este și regulat. Atunci există un vector $\lambda^* \in \mathbb{R}^m$ și un vector $\mu^* \in \mathbb{R}^p$ astfel încât condițiile Karush-Kuhn-Tucker (KKT) au loc:

$$(KKT) : \quad \nabla f(x^*) + \nabla h(x^*)^T \mu^* + \nabla g(x^*)^T \lambda^* = 0 \quad (10.10)$$

$$g(x^*)^T \lambda^* = 0 \quad (10.11)$$

$$g(x^*) \leq 0, \quad h(x^*) = 0$$

$$\mu^* \in \mathbb{R}^p, \quad \lambda^* \geq 0.$$

Demonstrație: Observăm mai întâi, că din moment ce $\lambda^* \geq 0$ și $g(x^*) \leq 0$, relația (10.11) este echivalentă cu a spune că o componentă λ_i^* a vectorului λ^* poate fi nenulă doar dacă constrângerea sa corespunzătoare $g_i(x^*)$ este activă. Astfel, faptul că $g_i(x^*) < 0$

implică $\lambda_i^* = 0$, iar $\lambda_i^* > 0$ implică $g_i(x^*) = 0$. Din moment ce x^* este punct de minim pentru problema (NLP) definită de mulțimea de constrângeri $X = \{x : g(x) \leq 0, h(x) = 0\}$, atunci este un punct de minim și pentru problema de optimizare având submulțimea mulțimii X de constrângeri definită prin setarea constrângerilor active la zero. Drept urmare, pentru problema având constrângeri de egalitate ce ar rezulta, definită pentru o vecinătate a lui x^* , există multiplicatori Lagrange. Astfel, tragem concluzia că relația (10.10) este satisfăcută pentru $\lambda_i^* = 0$ dacă $g_i(x^*) \neq 0$, iar drept urmare și relația (10.11) este satisfăcută. Mai trebuie să arătăm că $\lambda_i^* \geq 0$ pentru constrângerile active $g_i(x^*) = 0$. Presupunem o componentă $\lambda_k^* < 0$. Fie S_k și M_k suprafața și respectiv planul tangent definit de toate celelalte constrângeri active în x^* cu excepția constrângerii active $g_k(x^*) = 0$. Din moment ce am presupus că x^* este un punct regulat, atunci există un $d \in M_k$ astfel încât $\nabla g_k(x^*)d < 0$. Fie $x(t)$ o curbă în S_k ce trece prin x^* la $t = 0$, cu $\dot{x}(0) = d$. Atunci, pentru un $t \geq 0$ suficient de mic, $x(t)$ este fezabil și folosind prima relație (KKT) obținem:

$$\left. \frac{df(x(t))}{dt} \right|_{t=0} = \nabla f(x^*)d < 0$$

ce ar contrazice minimalitatea lui x^* . □

Punctele x^* ce satisfac condițiile (KKT) se numesc *puncte staționare* pentru problema (NLP) generală. Observăm că prima relație din condițiile (KKT) exprimă că x^* este punct staționar pentru funcția Lagrange, adică condiția de optimalitate de ordinul I:

$$\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0.$$

Cea de-a doua relație din condițiile (KKT) este *complementaritatea*: deoarece $\lambda^* \geq 0$ și $g(x^*) \leq 0$, atunci $g(x^*)^T \lambda^* = 0$ implică că dacă $g_i(x^*) < 0$ atunci $\lambda_i^* = 0$, iar dacă $\lambda_i^* > 0$ atunci $g_i(x^*) = 0$. Ultimele două relații din condițiile (KKT) exprimă fezabilitatea primală și duală, adică x^* este fezabil pentru problema primală și perechea (λ^*, μ^*) este fezabilă pentru problema duală. Condițiile (KKT) sunt numite după Karush, a cărui teză de master nepublicată din 1939 a fost introdusă în cartea publicată de Kuhn și Tucker în 1951.

Exemplul 10.3.1 *Considerăm problema de optimizare*

$$\begin{aligned} \min_{x \in \mathbb{R}^3} f(x) &= \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) \\ \text{s.l.: } g_1(x) &= x_1 + x_2 + x_3 + 3 \leq 0, \quad g_2(x) = x_1 \leq 0. \end{aligned}$$

Dorim să calculăm punctele KKT asociate problemei. Evident, vom avea $\lambda \in \mathbb{R}^2$. Observăm că orice punct fezabil pentru această problemă este un punct regulat, iar din condiția de optimalitate $\nabla f(x^*) + \sum_{i=1}^2 \lambda_i^* \nabla g_i(x^*) = 0$ avem:

$$x_1^* + \lambda_1^* + \lambda_2^* = 0$$

$$x_2^* + \lambda_1^* = 0$$

$$x_3^* + \lambda_1^* = 0.$$

Din condiția de complementaritate putem distinge patru cazuri:

1. presupunem constrângerile g_1 și g_2 sunt amândouă inactive, adică $x_1^* + x_2^* + x_3^* < -3$ și $x_1^* < 0$, de unde rezultă că $\lambda_1^* = 0$ și $\lambda_2^* = 0$. Din condițiile de optimalitate avem $x_1^* = x_2^* = x_3^* = 0$, ceea ce contrazice faptul că g_1 și g_2 ar fi inactive;
2. presupunem g_1 inactivă iar g_2 activă, adică $x_1^* + x_2^* + x_3^* < -3$ și $x_1^* = 0$, iar $\lambda_1^* = 0$, $\lambda_2^* \geq 0$. Din condițiile de optimalitate avem că $\lambda_1^* = -\lambda_2^* = 0$, iar implicit $x_2^* = x_3^* = 0$ ce conduce din nou la o contradicție;
3. presupunem g_1 activă și g_2 inactivă, adică $x_1^* + x_2^* + x_3^* = -3$ și $x_1^* < 0$, iar $\lambda_1^* \geq 0$ și $\lambda_2^* = 0$. Astfel, putem lua $x_1^* = x_2^* = x_3^* = -1$ și $\lambda_1^* = 1$ ce satisfac condițiile KKT, deci această soluție este un punct KKT;
4. presupunem că ambele constrângeri sunt active, adică $x_1^* + x_2^* + x_3^* = -3$ și $x_1^* = 0$, iar $\lambda_1^*, \lambda_2^* \geq 0$. Atunci obținem $x_2^* = x_3^* = -\frac{3}{2}$, iar $\lambda_1^* = \frac{3}{2}$ și $\lambda_2^* = -\frac{3}{2}$, ce contrazice condiția $\lambda_2^* \geq 0$.

Exemplul 10.3.2 *Considerăm problema de optimizare:*

$$\begin{aligned} \min_{x \in \mathbb{R}^2} 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2 \\ \text{s.l.: } x_1^2 + x_2^2 \leq 5, \quad 3x_1 + x_2 \leq 6. \end{aligned}$$

Condițiile (KKT) sunt în acest caz următoarele:

$$\begin{aligned} 4x_1 + 2x_2 - 10 + 2\lambda_1 x_1 + 3\lambda_2 &= 0, & 2x_1 + 2x_2 - 10 + 2\lambda_1 x_2 + \lambda_2 &= 0 \\ \lambda_1(x_1^2 + x_2^2 - 5) &= 0, & \lambda_2(3x_1 + x_2 - 6) &= 0 \\ x_1^2 + x_2^2 &\leq 5, & 3x_1 + x_2 &\leq 6, & \lambda_1 \geq 0, & \lambda_2 \geq 0. \end{aligned}$$

Pentru a găsi o soluție încercăm diferite combinații de constrângeri active și verificăm semnul multiplicatorilor Lagrange rezultați. Pentru acest exemplu putem considera două constrângeri active, una sau nici una. Presupunem că prima constrângere este activă și a doua este inactivă și rezolvăm sistemul de trei ecuații corespunzător:

$$\begin{aligned} 4x_1 + 2x_2 - 10 + 2\lambda_1 x_1 &= 0 \\ 2x_1 + 2x_2 - 10 + 2\lambda_1 x_2 &= 0 \\ x_1^2 + x_2^2 - 5 &= 0. \end{aligned}$$

Obținem soluția: $x_1^* = 1, x_2^* = 2$ și $\lambda_1^* = 1, \lambda_2^* = 0$. Observăm că această soluție verifică $3x_1 + x_2 \leq 6$ și $\mu_1 \geq 0$ și deci satisface condițiile (KKT).

10.4 Condiții de ordinul II pentru (NLP) generale

Condițiile de optimalitate de ordinul II, atât necesare cât și suficiente pentru probleme (NLP) generale, sunt derivate în mod esențial prin considerarea doar a problemei având constrângeri de egalitate echivalentă ce este implicată de constrângerile active. Planul tangent în x^* corespunzător pentru aceste probleme generale (NLP) este planul tangent pentru constrângerile active:

$$M = \{d : \nabla g_j(x^*)^T d = 0 \quad \forall j \in \mathcal{A}(x^*), \quad \nabla h_i(x^*)^T d = 0 \quad \forall i = 1, \dots, p\}.$$

Teorema 10.4.1 (Condiții necesare de ordinul II pentru NLP)

Fie f, g și h funcții continuu diferențiabile de două ori și un punct x^* punct regulat pentru constrângerile din problema (NLP) generală. Dacă x^* este un punct de minim local pentru problema (NLP), atunci există un $\lambda^* \in \mathbb{R}^m$ și un $\mu^* \in \mathbb{R}^p$, astfel încât condițiile (KKT) sunt satisfăcute, iar în plus Hessiana Lagrangianului în raport cu x :

$$\nabla_x^2 \mathcal{L}(x^*, \lambda^*, \mu^*) = \nabla^2 f(x^*) + \sum_{i=1}^p \mu_i^* \nabla^2 h_i(x^*) + \sum_{i=1}^m \lambda_i^* \nabla^2 g_i(x^*)$$

este pozitiv semidefinită pe subspațiul tangent al constrângerilor active în x^* , adică avem $d^T \nabla_x^2 \mathcal{L}(x^*, \lambda^*, \mu^*) d \geq 0$ pentru orice $d \in M$.

Demonstrație: Din moment ce x^* este punct de minim pentru constrângerile din problema (NLP) generală, atunci este punct de minim și pentru problema în care constrângerile active sunt luate drept constrângeri de egalitate și neglijate constrângerile inactive. În acest fel, demonstrația urmează imediat din condițiile necesare de ordinul II pentru probleme având constrângeri numai de egalitate. \square

Ca și în teoria minimizării neconstrânse, putem formula pentru problema (NLP) generală în mod asemănător condiții suficiente de ordinul II. Prin analogie cu rezultatul din cazul neconstrâns, condiția necesară este ca matricea $\nabla_x^2 \mathcal{L}(x^*, \lambda^*, \mu^*)$ să fie pozitiv definită pe planul tangent M corespunzător constrângerilor active. Acest fapt este într-adevăr suficient în majoritatea cazurilor, mai exact în cazurile nedegenerate.

Teorema 10.4.2 (Condiții suficiente de ordinul II pentru NLP)

Fie f, g și h funcții continuu diferențiabile de două ori. Fie de asemenea un punct regulat $x^ \in \mathbb{R}^n$ și variabilele duale $\lambda^* \in \mathbb{R}^m$ și $\mu^* \in \mathbb{R}^p$ pentru care condițiile (KKT) sunt satisfăcute și pentru care nu avem constrângeri de inegalitate degenerate, adică $\lambda_j^* > 0$ pentru orice $j \in \mathcal{A}(x^*)$. Dacă, de asemenea, Hessiana Lagrangianului $\nabla_x^2 \mathcal{L}(x^*, \lambda^*, \mu^*)$ este pozitiv definită pe subspațiul tangent:*

$$M = \{d : \nabla h(x^*)d = 0, \nabla g_j(x^*)^T d = 0 \ \forall j \in \mathcal{A}(x^*)\},$$

atunci x^ este un punct de minim local strict pentru problema (NLP) generală.*

Demonstrație: Similar cu demonstrația din cazul problemelor având constrângeri de egalitate, presupunem că x^* nu este un punct de minim strict. Fie astfel un șir de puncte fezabile z_k ce converge la x^* și pentru care $f(z_k) \leq f(x^*)$. Putem scrie $z_k = x^* + \delta_k s_k$, cu $\|s_k\| = 1$ și $\delta_k > 0$. Putem presupune că $\delta_k \rightarrow 0$ și s_k converge către un punct finit, adică $s_k \rightarrow s^*$. Vom avea astfel $\nabla f(x^*)^T s^* \leq 0$ și $\nabla h_i(x^*)^T s^* = 0$ pentru toți $i = 1, \dots, p$. De asemenea, pentru fiecare constrângere activă g_j avem $g_j(z_k) - g_j(x^*) \leq 0$, iar drept urmare:

$$\nabla g_j(x^*)^T s^* \leq 0,$$

dacă $\nabla g_j(x^*)^T s^* = 0$ pentru toți $j \in \mathcal{A}(x^*)$. Atunci, demonstrația ar continua ca și în cazul problemelor doar cu constrângeri de egalitate. În schimb, dacă $\nabla g_j(x^*)^T s^* < 0$ pentru cel puțin un $j \in \mathcal{A}(x^*)$, atunci:

$$0 \geq \nabla f(x^*)^T s^* = -(s^*)^T \nabla g(x^*)^T \lambda^* - (s^*)^T \nabla h(x^*)^T \mu^* > 0,$$

relație de altfel contradictorie. \square

De remarcat este faptul că dacă avem *constrângeri de inegalitate degenerate*, adică constrângeri de inegalitate active $g_i(x^*) = 0$ cu multiplicatorul Lagrange asociat $\lambda_i^* = 0$, atunci în teorema precedentă trebuie să cerem ca Hessiana Lagrangianului $\nabla_x^2 \mathcal{L}(x^*, \lambda^*, \mu^*)$ să fie pozitiv definită pe un subspațiu mai mare decât M , și anume pe subspațiul:

$$M' = \left\{ d : \nabla h(x^*)d = 0, \nabla g_j(x^*)^T d = 0 \quad \forall j \in \mathcal{A}_+(x^*), \right. \\ \left. \nabla g_j(x^*)^T d \leq 0 \quad \forall j \in \mathcal{A}_0(x^*) \right\},$$

unde am definit mulțimile de indecși $\mathcal{A}_+(x^*) = \{j : g_j(x^*) = 0, \lambda_j^* > 0\}$ și $\mathcal{A}_0(x^*) = \{j : g_j(x^*) = 0, \lambda_j^* = 0\}$.

Exemplul 10.4.1 *Considerăm problema:*

$$\min_{x \in \mathbb{R}^2 : x_1^2 + x_2^2 - 1 \leq 0} x_2.$$

În mod evident, punctul de minim global al acestei probleme este $x^* = [0 \ -1]^T$. Vom arăta că acesta este de fapt punct de minim strict. Pentru aceasta observăm că prima condiție (KKT) are forma:

$$2\lambda x_1 = 0, \quad 1 + 2\lambda x_2 = 0,$$

de unde rezultă că $\lambda > 0$ și deci constrângerea este activă. Din sistemul de trei ecuații dat de cele două ecuații de mai sus și $x_1^2 + x_2^2 - 1 = 0$ obținem soluția $x^* = [0 \ -1]^T$ și $\lambda^* = 1/2$. Mai departe, planul tangent în x^* este dat de $M = \{d : [0 \ 2]d = 0\} = \{d : d_2 = 0\}$. Observăm de asemenea că Hessiana Lagrangianului este $\nabla_x^2 \mathcal{L}(x^*, \lambda^*) = 2\lambda^* I_2$, care bineînțeles este pozitiv definită pe M . Aceasta arată că x^* este punct de minim strict.

În final, analizăm condițiile de optimalitate pentru cazul convex. Reamintim că problema (NLP) generală este o problemă convexă (CP)

dacă funcțiile f și g_1, \dots, g_m sunt funcții convexe, iar funcțiile h_1, \dots, h_p sunt funcții afine. Dacă funcția h este afină atunci există $A \in \mathbb{R}^{p \times n}$ și $b \in \mathbb{R}^p$ astfel încât $h(x) = Ax - b$. În acest caz, condițiile de ordinul I (KKT) sunt necesare și suficiente. De remarcat este faptul că în cazul convex condițiile necesare de ordinul I au loc sub condiția: x^* punct de minim global și regulat.

Teorema 10.4.3 (Condiții suficiente de ordinul I pentru probleme convexe) Fie o problema convexă (CP) de forma:

$$(CP): \quad f^* = \min_{x \in \mathbb{R}^n} f(x) \\ \text{s.l.: } g(x) \leq 0, \quad Ax = b,$$

în care funcțiile f și g_1, \dots, g_m sunt funcții convexe. Dacă următoarele condiții (KKT) sunt satisfăcute pentru tripletul (x^*, λ^*, μ^*) :

$$(KKT-CP): \quad \nabla f(x^*) + \nabla g(x^*)^T \lambda^* + A^T \mu^* = 0 \\ g(x^*)^T \lambda^* = 0 \\ g(x^*) \leq 0, \quad Ax^* = b \\ \mu^* \in \mathbb{R}^p, \lambda^* \geq 0,$$

atunci x^* este punct de minim global pentru problema convexă (CP), (λ^*, μ^*) este punct de maxim global pentru problema duală și strong dualitatea are loc, adică $f^* = q^*$.

Demonstrație: Deoarece funcția Lagrange este convexă în variabila x și ținând cont că prima relație din condițiile (KKT-CP) înseamnă $\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0$, implică $x^* = \arg \min_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda^*, \mu^*)$ și $q(\lambda^*, \mu^*) = \mathcal{L}(x^*, \lambda^*, \mu^*)$. Combinând proprietățile funcției duale cu aceste condiții (KKT-CP) avem:

$$f^* \geq q(\lambda^*, \mu^*) = \mathcal{L}(x^*, \lambda^*, \mu^*) \\ = f(x^*) + g(x^*)^T \lambda^* + (Ax^* - b)^T \mu^* = f(x^*) \geq f^*.$$

În concluzie, avem $f(x^*) = f^*$ și cum x^* este fezabil pentru problema convexă (CP) atunci este punct de minim global pentru această problemă. Mai departe, din dualitatea slabă și faptul că $f^* = q(\lambda^*, \mu^*)$ obținem $f^* = q^*$. \square

Exemplul 10.4.2 Considerăm problema proiecției originii $x_0 = 0$ pe subspațiul $X = \{x \in \mathbb{R}^n : Ax = b\}$, unde $A \in \mathbb{R}^{p \times n}$ și are rangul p cu $p < n$. Această problemă se formulează ca o problema convexă (CP):

$$\min_{x: Ax=b} \|x\|^2.$$

Condițiile (KKT-CP) pentru această problemă devin:

$$x^* + A^T \mu^* = 0, \quad Ax^* = b.$$

De aici obținem $-AA^T \mu^* = b$ și ținând seama că AA^T este matrice inversabilă, ajungem la $x^* = A^T(AA^T)^{-1}b = A^+b$, unde $A^+ = A^T(AA^T)^{-1}$ este pseudoinversa lui A .

Acest raționament poate fi extins la funcții pătratice convexe generale:

$$\min_{x: Ax=b} \frac{1}{2} x^T Q x + q^T x,$$

unde $Q \succ 0$. Pentru această problemă convexă condițiile (KKT-CP) devin:

$$Qx^* + q + A^T \mu^* = 0, \quad Ax^* = b.$$

Într-o manieră similară obținem că soluțiile optime primale și duale sunt date de expresiile:

$$\mu^* = -(AQ^{-1}A^T)^{-1}[AQ^{-1}q + b] \quad \text{și} \quad x^* = -Q^{-1}A^T \mu^* - Q^{-1}q.$$

Exemplul 10.4.3 Considerăm următoarea problemă convexă:

$$\begin{aligned} \min_{x \in \mathbb{R}^2} & (x_1 - 5)^2 + (x_2 - 5)^2 \\ \text{s.l.: } & g_1(x) = x_1^2 + x_2^2 - 5 \leq 0, \quad g_2(x) = x_1 + 2x_2 - 4 \leq 0 \\ & g_3(x) = -x_1 \leq 0, \quad g_4(x) = -x_2 \leq 0. \end{aligned} \quad (10.12)$$

Condițiile (KKT-CP) devin:

$$\begin{aligned} & \begin{bmatrix} 2(x_1^* - 5) \\ 2(x_2^* - 5) \end{bmatrix} + \begin{bmatrix} 2x_1^* & 1/2 & -1 & 0 \\ 2x_2^* & 1 & 0 & -1 \end{bmatrix} \lambda^* = 0 \\ & g_i(x^*) \leq 0, \quad \lambda_i^* \geq 0, \quad \lambda_i^* g_i(x^*) = 0 \quad \forall i = 1, \dots, 4. \end{aligned}$$

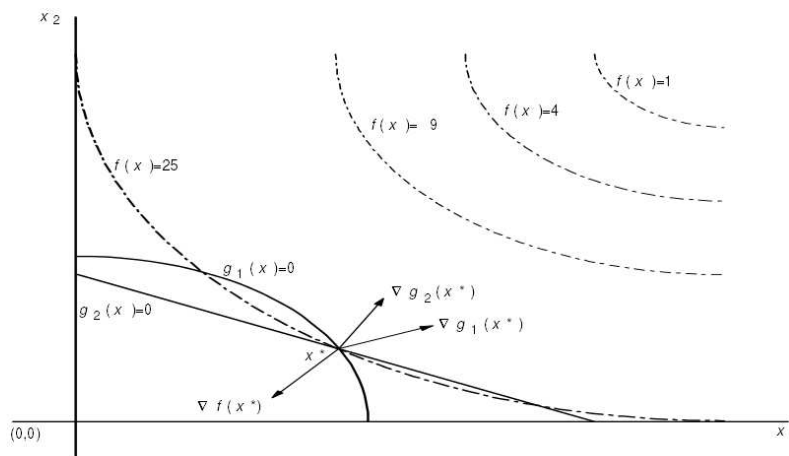


Figura 10.3: Condițiile KKT pentru problemă convexă definită în (10.12).

Presupunem (vezi Fig. 10.3) că primele două inegalități sunt active. Atunci avem $\lambda_3^* = \lambda_4^* = 0$ și deci condițiile (KKT-CP) se reduc la un sistem de patru ecuații cu patru necunoscute:

$$\begin{bmatrix} 2(x_1^* - 5) \\ 2(x_2^* - 5) \end{bmatrix} + \begin{bmatrix} 2x_1^* & 1/2 \\ 2x_2^* & 1 \end{bmatrix} \begin{bmatrix} \lambda_1^* \\ \lambda_2^* \end{bmatrix} = 0$$

$$(x_1^*)^2 + (x_2^*)^2 - 5 = 0, \quad x_1^* + 2x_2^* - 4 = 0,$$

care are soluția $x^* = [2 \ 1]^T$, și deci x^* este punct de minim global pentru problema convexă considerată.

Capitolul 11

Metode de ordinul I și II pentru (NLP) având constrângeri convexe

În acest capitol vom analiza metode numerice de optimizare pentru probleme (NLP) unde mulțimea fezabilă este convexă, adică:

$$\min_{x \in X} f(x), \quad (11.1)$$

unde f este o funcție diferentiabilă (nu neapărat convexă), dar mulțimea X este nevidă, închisă și convexă. Reamintim un rezultat fundamental, condițiile de optimalitate necesare de ordinul I, pentru problema de optimizare de forma (11.1) (vezi Teorema (10.0.3)): dacă x^* este punct de minim local atunci:

$$\nabla f(x^*)^T (x - x^*) \geq 0 \quad \forall x \in X.$$

Reamintim că un punct x^* satisfăcând relația precedentă se numește *punct staționar* pentru problema de optimizare (11.1).

Exemplul 11.0.4 (Proiecția Euclideană) *Revenim la problema proiecției unui vector $x_0 \in \mathbb{R}^n$ pe mulțimea convexă $X \subseteq \mathbb{R}^n$ ce se formulează matematic astfel:*

$$\min_{x \in X} \|x - x_0\|^2.$$

Notăm cu $[x_0]_{(I_n, X)}$ soluția optimă a acestei probleme, adică proiecția $[x_0]_{(I_n, X)}$ este acel vector din X care se află la cea mai mică distanță de

x_0 . Din condițiile de optimalitate ale problemei precedente avem relația:

$$([x_0]_{(I_n, X)} - x_0)^T (x - [x_0]_{(I_n, X)}) \geq 0 \quad \forall x \in X. \quad (11.2)$$

O altă proprietate importantă a proiecției este cea de nonexpansivitate:

$$\|[x_0]_{(I_n, X)} - [y_0]_{(I_n, X)}\| \leq \|x_0 - y_0\| \quad \forall x_0, y_0 \in \mathbb{R}^n. \quad (11.3)$$

Această proprietate se derivează ușor din condițiile de optimalitate. Într-adevăr, aplicând condiția de optimalitate pentru x_0 cu $x = [y_0]_{(I_n, X)} \in X$ obținem:

$$([x_0]_{(I_n, X)} - x_0)^T ([y_0]_{(I_n, X)} - [x_0]_{(I_n, X)}) \geq 0.$$

Aplicând același procedeu pentru y_0 cu $x = [x_0]_{(I_n, X)} \in X$ obținem:

$$([y_0]_{(I_n, X)} - y_0)^T ([x_0]_{(I_n, X)} - [y_0]_{(I_n, X)}) \geq 0.$$

Adunând aceste două relații, aranjând termenii și apoi aplicând inegalitatea Cauchy-Schwartz obținem rezultatul dorit.

Metodele prezentate în acest capitol vor fi generalizarea la cazul constrâns al metodelor direcțiilor de descreștere (de exemplu gradient și Newton) dezvoltate pentru cazul problemelor neconstrânse. Algoritmii ce vor fi analizați în această parte a lucrării aparțin clasei de metode bazate pe direcții fezabile.

11.1 Metode de direcții de descreștere

Pentru un punct fezabil $x \in X$, o *direcție fezabilă* la x este un vector d pentru care $x + \alpha d$ este fezabil pentru orice α suficient de mic. O metodă a direcțiilor fezabile pornește dintr-un punct fezabil $x_0 \in X$ și generează un șir de vectori pe baza următoarei iterații:

$$x_{k+1} = x_k + \alpha_k d_k,$$

unde, dacă x_k nu este punct staționar, d_k este o direcție fezabilă la x_k , adică:

$$\nabla f(x_k)^T d_k < 0$$

și pasul $\alpha_k > 0$ este ales astfel încât $x_k + \alpha_k d_k \in X$. În particular considerăm metode de direcții fezabile care sunt și metode de descreștere, adică pasul α_k se alege astfel încât:

$$f(x_k + \alpha_k d_k) < f(x_k) \quad \forall k.$$

Deoarece X este mulțime convexă, atunci direcțiile fezabile la x_k sunt vectori de forma

$$d_k = \gamma(\bar{x}_k - x_k),$$

în care $\gamma > 0$ și \bar{x}_k este un vector fezabil. În acest caz avem iterația:

$$x_{k+1} = x_k + \alpha_k(\bar{x}_k - x_k),$$

unde $\alpha_k \in [0, 1]$ și dacă x_k nu este punct staționar atunci știm că există $\bar{x}_k \in X$ astfel încât $\nabla f(x_k)^T(\bar{x}_k - x_k) < 0$. Este evident că dacă mulțimea fezabilă X este convexă atunci $x_k + \alpha_k(\bar{x}_k - x_k) \in X$ pentru orice $\alpha_k \in [0, 1]$. Procedurile de alegere a pasului α_k sunt aceleași ca și în cazul neconstrâns: putem alege α_k prin procedura ideală, prin backtracking sau pas constant $\alpha_k = 1$.

Observăm că dacă x_k este punct staționar, atunci metoda se oprește, adică $x_{k+1} = x_k$. În concluzie, un posibil criteriu de oprire pentru aceste metode de direcții de descreștere ce vor fi prezentate în acest capitol este următorul:

$$\|x_{k+1} - x_k\| \leq \epsilon,$$

pentru o acuratețe fixată $\epsilon > 0$.

În cele ce urmează presupunem că direcțiile d_k sunt alese astfel încât ele sunt *conectate prin gradient* la x_k : adică pentru orice șir x_k care converge la un punct nestaționar, șirul corespunzător d_k este mărginit și satisface:

$$\limsup_{k \rightarrow \infty} \sup_{k \geq 0} \nabla f(x_k)^T d_k < 0.$$

Teorema 11.1.1 *Fie un șir x_k generat de metoda direcțiilor fezabile $x_{k+1} = x_k + \alpha_k d_k$. Presupunem că direcțiile d_k sunt conectate prin gradient la x_k și pasul α_k este ales prin metoda ideală sau backtracking. Atunci orice punct limită al șirului x_k este punct staționar.*

Demonstrație: Folosim metoda reducerii la absurd pentru a demonstra această teoremă. Considerăm cazul când alegem α_k prin backtracking (cazul când se alege prin metoda ideală se tratează într-o

manieră similară). Pentru simplitate presupunem că întreg șirul x_k este convergent la \bar{x} care nu este punct staționar. Deoarece șirul $f(x_k)$ este descrescător, atunci el este convergent și datorită continuității lui f converge la $f(\bar{x})$. Din această relație avem:

$$f(x_k) - f(x_{k+1}) \rightarrow 0.$$

Conform strategiei de backtracking pentru alegerea pasului, avem

$$f(x_k) - f(x_{k+1}) \geq -c_1 \alpha_k \nabla f(x_k)^T d_k.$$

Deci și $\alpha_k \nabla f(x_k)^T d_k \rightarrow 0$. Pe de altă parte, am presupus că d_k sunt conectate prin gradient la x_k și deci d_k sunt mărginite și următoarea relație are loc $\lim_{k \rightarrow \infty} \sup_{k \geq 0} \nabla f(x_k)^T d_k < 0$, ceea ce implică:

$$\alpha_k \rightarrow 0.$$

Din această relație și din definiția procedurii de backtracking, rezultă că există un index \bar{k} suficient de mare astfel încât:

$$f(x_k) - f(x_k + (\alpha_k/\rho)d_k) < -c_1(\alpha_k/\rho)\nabla f(x_k)^T d_k \quad \forall k \geq \bar{k}.$$

Din mărginirea șirului d_k , înseamnă că există un subșir convergent. Pentru simplitate presupunem că întreg șirul d_k este convergent și are punctul limită \bar{d} . Din inegalitatea precedentă avem:

$$\frac{f(x_k) - f(x_k + \bar{\alpha}_k d_k)}{\bar{\alpha}_k} < c_1 \nabla f(x_k)^T d_k \quad \forall k \geq \bar{k},$$

unde $\bar{\alpha}_k = \alpha_k/\rho$. Din teorema valorii medii avem că există scalarul $\tilde{\alpha}_k \in [0, \bar{\alpha}_k]$ astfel încât

$$-\nabla f(x_k + \tilde{\alpha}_k d_k)^T d_k < -c_1 \nabla f(x_k)^T d_k \quad \forall k \geq \bar{k}.$$

Evaluând limita în ambele părți obținem:

$$-\nabla f(\bar{x})^T \bar{d} \leq -c_1 \nabla f(\bar{x})^T \bar{d}$$

sau echivalent

$$0 \leq (1 - c_1) \nabla f(\bar{x})^T \bar{d}.$$

Dar $c_1 < 1$ și deci

$$0 \leq \nabla f(\bar{x})^T \bar{d},$$

ceea ce contrazice presupunerea noastră că d_k este conectat prin gradient la x_k . \square

În cele ce urmează vom da câteva metode de a alege direcțiile fezabile $d_k = \gamma(\bar{x}_k - x_k)$, unde $\bar{x}_k \in X$.

11.1.1 Metoda gradient condițional

Cea mai evidentă modalitate de a alege o direcție fezabilă este următoarea: alegem \bar{x}_k ca fiind soluția optimă a subproblemei de optimizare:

$$\bar{x}_k = \arg \min_{x \in X} \nabla f(x_k)^T (x - x_k).$$

Metoda direcțiilor de descreștere corespunzătoare acestei alegeri a direcției fezabile se numește *metoda gradient condițional*. Observăm că problema de optimizare ce trebuie să fie rezolvată la fiecare pas este o problemă convexă (funcția obiectiv este liniară). Bineînțeles că această metodă face sens atâta timp cât costul pe iterație (complexitatea numerică pentru rezolvarea subproblemei) este mult mai mic decât costul total pentru rezolvarea problemei originale (11.1). Acest lucru se întâmplă de exemplu atunci când funcția obiectiv f este neconvexă și mulțimea fezabilă este simplă (e.g. dacă X este un hipercub în \mathbb{R}^n sau simplex atunci subproblema devine un (LP) ce poate fi rezolvat eficient).

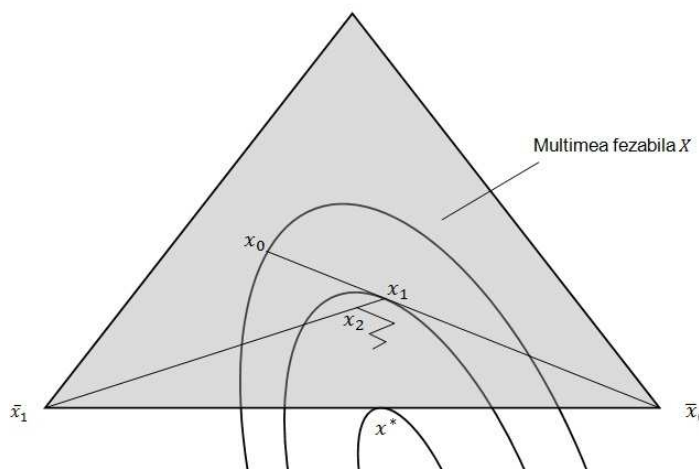


Figura 11.1: Iteratiile metodei gradient condițional.

Folosind teoria anterioară de convergență, putem arăta că metoda gradient condițional are convergență asimptotică. Presupunem că X este mulțime compactă. Într-adevăr este suficient să arătăm că această metodă produce direcții conectate prin gradient. Presupunem că șirul x_k este convergent la un punct \bar{x} care nu este punct staționar. Atunci

trebuie să arătăm că:

$$\limsup_{k \rightarrow \infty} \sup_{k \geq 0} \nabla f(x_k)^T (\bar{x}_k - x_k) < 0, \quad \limsup_{k \rightarrow \infty} \sup_{k \geq 0} \|\bar{x}_k - x_k\| < \infty.$$

Deoarece X este compactă, atunci este evident că cea de-a doua relație are loc. Pentru a demonstra prima relație, observăm mai întâi

$$\nabla f(x_k)^T (\bar{x}_k - x_k) \leq \nabla f(x_k)^T (x - x_k) \quad \forall x \in X$$

și evaluând limita obținem:

$$\limsup_{k \rightarrow \infty} \sup_{k \geq 0} \nabla f(x_k)^T (\bar{x}_k - x_k) \leq \nabla f(\bar{x})^T (x - \bar{x}) \quad \forall x \in X.$$

Evaluând minimum peste $x \in X$ și ținând cont de faptul că \bar{x} nu este punct staționar obținem:

$$\limsup_{k \rightarrow \infty} \sup_{k \geq 0} \nabla f(x_k)^T (\bar{x}_k - x_k) \leq \min_{x \in X} \nabla f(\bar{x})^T (x - \bar{x}) < 0$$

ceea ce ne conduce la prima relație care trebuia demonstrată.

În ceea ce privește rata de convergență, metoda gradientului condițional are o rată de convergență slabă. De exemplu, se poate arăta că pentru anumite tipuri de mulțimi X (e.g. politop) șirul $f(x_k) - f^*$ sau $\|x_k - x^*\|$ nu converge liniar. Explicația este următoarea: vectorul \bar{x}_k folosit în algoritm coincide de obicei cu vârfurile politopului X și astfel direcțiile fezabile folosite de acest algoritm pot să fie ortogonale pe direcția ce conduce la punctul de minim (conform Fig. 11.1).

11.1.2 Metoda gradient proiectat

Metoda gradient condițional folosește o direcție fezabilă obținută prin rezolvarea unei subprobleme cu funcție obiectiv liniară. *Metoda gradientului proiectat* folosește în locul acesteia o subproblemă cu funcție obiectiv pătratică. Deși în general această subproblemă poate fi mai complexă, rata de convergență va fi mai bună, după cum vom arăta în cele ce urmează. Metoda gradientului proiectat aparține de asemenea clasei de metode de direcții de descreștere de forma:

$$x_{k+1} = x_k + \alpha_k (\bar{x}_k - x_k),$$

în care

$$\bar{x}_k = [x_k - s_k \nabla f(x_k)]_{(I_n, X)}.$$

În această metodă, pașii sunt aleși astfel: $\alpha_k \in (0, 1]$ și $s_k > 0$. Reamintim că notația $[z]_{(I_n, X)}$ reprezintă proiecția Euclideană pe mulțimea X a vectorului z . Deci pentru a obține \bar{x}_k luăm un pas s_k în direcția antigradientului $-\nabla f(x_k)$, ca în metoda gradient pentru cazul neconstrâns, apoi proiectăm (folosind norma Euclidiană) pe X vectorul $x_k - s_k \nabla f(x_k)$. Putem vedea s_k de asemenea ca un pas: de exemplu, dacă luăm $\alpha_k = 1$ pentru orice k , atunci $x_{k+1} = \bar{x}_k$, ceea ce conduce la gradientul proiectat clasic (vezi Fig. 11.2):

$$x_{k+1} = [x_k - s_k \nabla f(x_k)]_{(I_n, X)},$$

sau astfel spus:

$$\begin{aligned} x_{k+1} &= \arg \min_{x \in X} \|x - x_k + s_k \nabla f(x_k)\|^2 \\ &= \arg \min_{x \in X} \nabla f(x_k)^T (x - x_k) + \frac{1}{2s_k} (x - x_k)^T I_n (x - x_k). \end{aligned}$$

Observăm că dacă $x_k - \alpha_k \nabla f(x_k) \in X$ atunci iterația acestei metode coincide cu iterația metodei gradient pentru cazul neconstrâns. Mai mult, avem $x^* = [x^* - s \nabla f(x^*)]_{(I_n, X)}$ dacă și numai dacă x^* este punct staționar pentru problema (11.1). În mod evident, metoda gradient proiectat este eficientă dacă proiecția se calculează ușor (e.g. când mulțimea X este un hiperplan sau hiperplan).

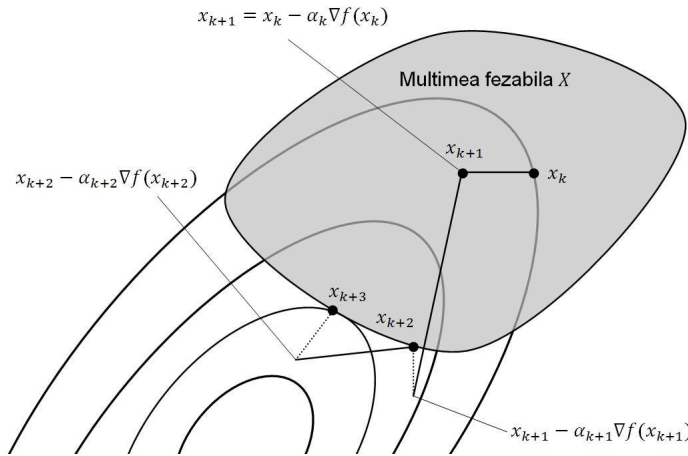


Figura 11.2: Iterațiile metodei gradient proiectat.

Avem diferite posibilități de alegere a pasului α_k și s_k :

- (i) fixăm $s_k = s$ constant și α_k este ales cu metoda ideală pe baza direcției fezabile $d_k = \bar{x}_k - x_k$, adică:

$$\alpha_k = \arg \min_{\alpha \in [0, 1]} f(x_k + \alpha(\bar{x}_k - x_k))$$

- (ii) fixăm $s_k = s$ constant și α_k este ales cu procedura de backtracking pe baza direcției fezabile $d_k = \bar{x}_k - x_k$. În particular alegem $c_1 > 0$ și $\rho \in (0, 1)$ și apoi luăm $\alpha_k = \rho^{m_k}$, unde m_k este primul număr natural pentru care:

$$f(x_k + \rho^m(\bar{x}_k - x_k)) \leq f(x_k) + c_1 \rho^m \nabla f(x_k)^T (\bar{x}_k - x_k)$$

- (iii) fixăm $\alpha_k = 1$ constant și alegem s_k cu procedura de backtracking. În particular, definim $x_k(s) = [x_k - s \nabla f(x_k)]_{(I_n, X)}$ și alegem $\bar{s} > 0$, $c_1 > 0$ și $\rho \in (0, 1)$. Atunci definim $s_k = \rho^{m_k} \bar{s}$, unde m_k este primul număr natural pentru care:

$$f(x_k(\rho^m \bar{s})) \leq f(x_k) + c_1 \nabla f(x_k)^T (x_k - x_k(\rho^m \bar{s}))$$

- (iv) fixăm $\alpha_k = 1$ și $s_k = s$ constante
- (v) fixăm $\alpha_k = 1$ și alegem $s_k \rightarrow 0$ astfel încât $\sum_{k=0}^{\infty} s_k = \infty$, de exemplu $s_k = 1/k$.

Rezultatele de convergență pentru metoda gradient proiectat sunt similare celor corespunzătoare metodei gradient pentru cazul neconstrâns. În cele ce urmează enunțăm câteva teoreme de convergență corespunzătoare diferitelor posibilități de alegere a pașilor.

Teorema 11.1.2 *Fie x_k șirul generat de metoda gradient proiectat cu pasul s_k constant și α_k ales prin metoda ideală sau backtracking de-a lungul direcțiilor fezabile. Atunci orice punct limită al șirului x_k este punct staționar.*

Demonstrație: Vom arăta că direcțiile $\bar{x}_k - x_k$ sunt conectate prin gradient la x_k . Într-adevăr, presupunem că șirul x_k este convergent la un punct \tilde{x} care nu este punct staționar. Trebuie să arătăm că:

$$\limsup_{k \rightarrow \infty} \sup_{k \geq 0} \nabla f(x_k)^T (\bar{x}_k - x_k) < 0, \quad \limsup_{k \rightarrow \infty} \sup_{k \geq 0} \|\bar{x}_k - x_k\| < \infty.$$

Folosind continuitatea proiecției avem

$$\lim_{k \rightarrow \infty} \bar{x}_k = [\tilde{x} - s \nabla f(\tilde{x})]_{(I_n, X)}$$

și deci cea de-a doua relație are loc deoarece $\|\bar{x}_k - x_k\|$ converge la $\|[\tilde{x} - s \nabla f(\tilde{x})]_{(I_n, X)} - \tilde{x}\|$. Pentru a arăta și prima relație folosim proprietățile proiecției, și anume condițiile de optimalitate pentru proiecție implică:

$$(x_k - s \nabla f(x_k) - \bar{x}_k)^T (x - \bar{x}_k) \leq 0 \quad \forall x \in X.$$

Luând în această relație $x = x_k$ obținem:

$$\nabla f(x_k)^T (\bar{x}_k - x_k) \leq -\frac{1}{s} \|x_k - \bar{x}_k\|^2. \quad (11.4)$$

Aplicând limita în relația anterioară obținem:

$$\lim_{k \rightarrow \infty} \sup_{k \geq 0} \nabla f(x_k)^T (\bar{x}_k - x_k) \leq -\frac{1}{s} \|\tilde{x} - [\tilde{x} - s \nabla f(\tilde{x})]_{(I_n, X)}\|^2.$$

Deoarece \tilde{x} nu este punct staționar, partea dreaptă a acestei inegalități este strict negativă ceea ce implică $\lim_{k \rightarrow \infty} \sup_{k \geq 0} \nabla f(x_k)^T (\bar{x}_k - x_k) < 0$, adică prima relație care trebuia demonstrată. \square

Teorema 11.1.3 *Fie x_k șirul generat de metoda gradient proiectat cu pasul $s_k = s$ constant și $\alpha_k = 1$. Presupunem de asemenea că funcția obiectiv are gradientul Lipschitz (i.e. există $L > 0$ astfel încât $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ pentru orice $x, y \in X$). Dacă $s \in (0, 2/L)$, atunci orice punct limită al șirului x_k este punct staționar.*

Demonstrație: Din proprietatea de Lipschitz avem:

$$f(x_{k+1}) - f(x_k) = f(\bar{x}_k) - f(x_k) \leq \nabla f(x_k)^T (\bar{x}_k - x_k) + \frac{L}{2} \|\bar{x}_k - x_k\|^2.$$

Folosind această relație și inegalitatea (11.4) obținem:

$$f(x_{k+1}) - f(x_k) \leq \left(\frac{L}{2} - \frac{1}{s}\right) \|\bar{x}_k - x_k\|^2.$$

Dacă $s \in (0, 2/L)$, partea dreaptă a acestei relații este negativă și deci dacă șirul x_k este convergent, partea stângă tinde la 0. În concluzie, șirul

$\|\bar{x}_k - x_k\|$ tinde la 0 și deci pentru orice punct limită \tilde{x} al șirului x_k avem $\tilde{x} = [\tilde{x} - s\nabla f(\tilde{x})]_{(I_n, X)}$, i.e. \tilde{x} este punct staționar. \square

Rata de convergență a metodei gradient proiectat este în esență similară celei corespunzătoare cazului neconstrâns. De exemplu, considerăm cazul pătratic:

$$f(x) = \frac{1}{2}x^T Qx - q^T x,$$

unde Q este matrice pozitiv definită. Fie x^* unicul punct de minim al lui f peste mulțimea fezabilă X . Considerăm cazul când $\alpha_k = 1$ și $s_k = s$. Folosind proprietățile proiecției (în particular proprietatea de nonexpansiune) avem:

$$\begin{aligned} \|x_{k+1} - x^*\| &= \|[x_k - s\nabla f(x_k)]_{(I_n, X)} - [x^* - s\nabla f(x^*)]_{(I_n, X)}\| \\ &\leq \|x_k - s\nabla f(x_k) - (x^* - s\nabla f(x^*))\| = \|(I_n - sQ)(x_k - x^*)\| \\ &\leq \max\{|1 - s\lambda_{\min}|, |1 - s\lambda_{\max}|\} \|x_k - x^*\|, \end{aligned}$$

adică rată de convergență liniară. De aceea, în partea finală a acestui capitol vom considera metode bazate pe scalare, adică aplicarea metodei gradient proiectat într-un sistem de coordonate diferit. Un caz particular al acestor metode este metoda Newton.

11.2 Metoda Newton proiectat

Presupunem că f este de două ori diferențiabilă și Hessiana $\nabla^2 f(x)$ este pozitiv definită pentru orice $x \in X$. Considerăm următoarea iterație pentru metoda Newton proiectat:

$$x_{k+1} = x_k + \alpha_k(\bar{x}_k - x_k),$$

unde

$$\bar{x}_k = \arg \min_{x \in X} \nabla f(x_k)^T (x - x_k) + \frac{1}{2s_k} (x - x_k)^T \nabla^2 f(x_k) (x - x_k). \quad (11.5)$$

Putem interpreta problema de optimizare (11.5) ca o problemă de proiecție generalizată. În particular, \bar{x}_k este vectorul din X care se află la distanța minimă de vectorul $x_k - s_k(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$, dar cu distanța măsurată în funcție de norma $\|z\|_{\nabla^2 f(x_k)}^2 = z^T \nabla^2 f(x_k) z$, după cum vom

arăta în cele ce urmează. Într-adevăr, pentru cazul când $\alpha_k = 1$, obținem următoarea iterație:

$$\begin{aligned} x_{k+1} = \bar{x}_k &= \arg \min_{x \in X} \nabla f(x_k)^T (x - x_k) + \frac{1}{2s_k} (x - x_k)^T \nabla^2 f(x_k) (x - x_k) \\ &= \arg \min_{x \in X} \|x - x_k + s_k (\nabla^2 f(x_k))^{-1} \nabla f(x_k)\|_{\nabla^2 f(x_k)}^2 \\ &= [x_k - s_k (\nabla^2 f(x_k))^{-1} \nabla f(x_k)]_{(\nabla^2 f(x_k), X)}, \end{aligned}$$

unde am considerat norma $\|z\|_{\nabla^2 f(x_k)}^2 = z^T \nabla^2 f(x_k) z$ și $[y]_{(\nabla^2 f(x_k), X)}$ reprezintă proiecția vectorului y pe mulțimea X în raport cu această normă. Putem interpreta această iterație și în felul următor: metoda Newton proiectat se obține prin aplicarea metodei gradient proiectat într-un alt sistem de coordonate. Într-adevăr, la iteratia k fie matricea pozitiv definită H_k și considerăm următoarea schimbare de variabilă:

$$x = H_k^{-1/2} y.$$

Atunci, problema originală (11.1) se rescrie în variabila y astfel:

$$\min_{y \in Y_k} f_k(y) \quad \left(= f(H_k^{-1/2} y) \right),$$

unde mulțimea Y_k este definită în felul următor:

$$Y_k = \{y : H_k^{-1/2} y \in X\}.$$

Aplicăm metoda gradient proiectat pentru această nouă problemă de optimizare:

$$y_{k+1} = y_k + \alpha_k (\bar{y}_k - y_k), \quad \bar{y}_k = [y_k - s_k \nabla f_k(y_k)]_{(I_n, Y_k)}.$$

Atunci, \bar{y}_k este soluția problemei pătratice:

$$\bar{y}_k = \arg \min_{y \in Y_k} \nabla f_k(y_k)^T (y - y_k) + \frac{1}{2s_k} \|y - y_k\|^2.$$

Folosind schimbarea de variabilă precedentă avem:

$$x_k = H_k^{-1/2} y_k, \quad \bar{x}_k = H_k^{-1/2} \bar{y}_k, \quad \nabla f_k(y_k) = H_k^{-1/2} \nabla f(x_k).$$

Pe baza iterației gradient proiectat pentru y_k obținem următoarea iterație în x_k :

$$x_{k+1} = x_k + \alpha_k (\bar{x}_k - x_k),$$

unde \bar{x}_k se obține din:

$$\begin{aligned}\bar{x}_k &= \arg \min_{x \in X} \nabla f(x_k)^T (x - x_k) + \frac{1}{2s_k} (x - x_k)^T H_k (x - x_k) \\ &= [x_k - s_k H_k^{-1} \nabla f(x_k)]_{(H_k, X)}.\end{aligned}\quad (11.6)$$

De obicei, ne referim la această iterație, ce depinde de felul cum alegem matricea H_k , prin *metoda gradient scalat proiectat*.

Metoda Newton proiectat se obține pentru $H_k = \nabla^2 f(x_k)$. Observăm că dacă $s_k = 1$ atunci funcția pătratică din (11.5) este aproximarea Taylor de ordinul II în jurul lui x_k a lui f . În particular, dacă $\alpha_k = 1$ și $s_k = 1$, atunci x_{k+1} este vectorul care minimizează aproximarea Taylor de ordinul II în jurul lui x_k al funcției f supus la constrângerea $x \in X$. În concluzie, ne așteptăm la următorul comportament pentru această metodă: dacă punctul de pornire x_0 este suficient de aproape de punctul de minim local, atunci metoda Newton proiectat cu pașii $\alpha_k = s_k = 1$ converge la x^* superliniar.

Principala dificultate în folosirea metodei Newton proiectat constă în faptul că subproblema ce trebuie rezolvată la fiecare pas este complexă. De aceea, au fost dezvoltate metode care să păstreze rata de convergență rapidă a metodei Newton proiectat, dar care în același timp să aibă o iterație mai puțin costisitoare numeric. De exemplu, putem înlocui Hessiana $\nabla^2 f(x_k)$ cu o matrice pozitiv definită H_k , așa cum am arătat mai înainte, adică metoda gradient scalat proiectat, unde direcția se calculează pe baza relației (11.6). O posibilă alegere pentru H_k este matricea diagonală cu elementele diagonale identice cu cele ale matricei Hessiane $\nabla^2 f(x_k)$. Proprietatea de convergență asimptotică a metodei Newton proiectat este prezentată în următoarea teoremă a cărei demonstrație este aproape identică cu cea a Teoremei 11.1.2.

Teorema 11.2.1 *Fie x_k șirul generat de metoda Newton proiectat cu pasul s_k constant și α_k ales prin metoda ideală sau backtracking de-a lungul direcțiilor fezabile. Presupunem de asemenea că există scalarii pozitivi β_1 și β_2 astfel încât:*

$$\beta_1 I_n \preceq \nabla^2 f(x) \preceq \beta_2 I_n \quad \forall x \in X.$$

Atunci orice punct limită al șirului x_k este punct staționar.

În ceea ce privește rata de convergență locală a metodei Newton proiectat, avem următorul rezultat:

Teorema 11.2.2 *Fie f de două ori diferențiabilă cu Hessiana pozitiv definită și Lipschitz continuă. Fie, de asemenea, x^* un punct de minim local pentru problema (11.1). Atunci există $\gamma > 0$ astfel încât dacă $\|x_0 - x^*\| < \gamma$, șirul x_k produs de metoda Newton proiectat cu $\alpha_k = s_k = 1$ satisface $\|x_k - x^*\| < \gamma$ și x_k converge la x^* cu rată de convergență superliniară.*

Demonstrație: Pentru simplitate, notăm cu $H_k = \nabla^2 f(x_k)$ și considerăm norma vectorială $\|z\|_{H_k}^2 = z^T H_k z$ pentru orice $z \in \mathbb{R}^n$. Pe baza acestei norme definim și norma indusă pentru matrice: $\|A\|_{H_k} = \sup_{z \neq 0} \|Az\|_{H_k} / \|z\|_{H_k}$ pentru orice matrice $A \in \mathbb{R}^{n \times n}$. Folosind proprietățile proiecției putem deriva următorul șir de inegalități:

$$\begin{aligned} \|x_{k+1} - x^*\|_{H_k} &= \|[x_k - \alpha_k (\nabla^2 f(x_k))^{-1} \nabla f(x_k)]_{(H_k, X)} - x^*\|_{H_k} \\ &\leq \|x_k - \alpha_k (\nabla^2 f(x_k))^{-1} \nabla f(x_k) - x^*\|_{H_k}. \end{aligned}$$

Folosind aceleași argumente ca în demonstrația convergenței metodei Newton pentru cazul neconstrâns, putem arăta inegalitatea:

$$\|x_{k+1} - x^*\|_{H_k} \leq M \int_0^1 \|\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*))\|_{H_k} d\tau \cdot \|x_k - x^*\|_{H_k},$$

pentru un $M > 0$. Datorită continuității lui $\nabla^2 f$ putem alege $\gamma > 0$ suficient de mic pentru a asigura $\|x_k - x^*\|_{H_k} < \gamma$ și termenul de sub integrală devine arbitrar de mic. Din această proprietate, convergența superliniară a lui x_k la x^* rezultă imediat. \square

Capitolul 12

Metode de optimizare pentru (NLP) având constrângeri de egalitate

În acest capitol considerăm metode numerice de optimizare pentru problema (NLP) având constrângeri de tip egalitate:

$$(NLPe) : \min_{x \in \mathbb{R}^n} f(x) \quad (12.1)$$

$$\text{s.l.: } h(x) = 0, \quad (12.2)$$

unde $f : \mathbb{R}^n \rightarrow \mathbb{R}$ și $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$ sunt funcții de două ori diferențiabile. Reamintim condițiile de optimalitate de ordinul I pentru această problemă (condițiile KKT): fie x^* punct de minim atunci există $\mu^* \in \mathbb{R}^p$ astfel încât

$$(KKT - NLPe) : \quad \begin{aligned} \nabla_x \mathcal{L}(x^*, \mu^*) &= 0 \\ h(x^*) &= 0, \end{aligned}$$

unde $\mathcal{L}(x, \mu) = f(x) + \mu^T h(x)$ este Lagrangianul. Reamintim că aceste condiții de optimalitate au loc sub presupunerea că x^* este punct regulat: adică rangul matricei $\nabla h(x^*)$ este p . Observăm că $h(x) = \nabla_\mu \mathcal{L}(x, \mu)$, deci condițiile precedente de optimalitate se pot scrie compact în forma:

$$\nabla \mathcal{L}(x^*, \mu^*) = 0.$$

Avem un sistem neliniar de $n + p$ ecuații cu $n + p$ necunoscute formate din x^* și μ^* . Acest sistem se numește *sistemul Lagrange*.

Exemplul 12.0.1 (Control optimal) Considerăm sistemul dinamic liniar discret:

$$z_{t+1} = A_t z_t + B_t u_t \quad \forall t \geq 0,$$

unde $z_t \in \mathbb{R}^{n_z}$ este starea și $u_t \in \mathbb{R}^{n_u}$ intrarea sistemului. De exemplu considerăm un rezervor cu apă ce este folosit la fabricarea unui produs. Notăm cu z_t volumul de apă din rezervor și cu u_t apa utilizată în perioada t . Atunci volumul de apă din rezervor evoluează astfel:

$$z_{t+1} = z_t - u_t.$$

Presupunem de asemenea un cost pe etapă asociat stărilor $\ell_t^z(z_t)$ și intrărilor $\ell_t^u(u_t)$, unde $\ell_t^z : \mathbb{R}^{n_z} \rightarrow \mathbb{R}$ și $\ell_t^u : \mathbb{R}^{n_u} \rightarrow \mathbb{R}$. De exemplu, putem considera $\ell_t^z(z_t) = 1/2 \|z_t - z_t^{ref}\|_{Q_t}^2$ și $\ell_t^u(u_t) = 1/2 \|u_t - u_t^{ref}\|_{R_t}^2$, unde z_t^{ref} și u_t^{ref} sunt referințe dorite pentru nivelul rezervorului și pentru volumul de apă utilizat în perioada t . De asemenea, introducem un orizont de control (predicție) N pentru sistemul dinamic dat. Atunci problema de control optimal devine:

$$\begin{aligned} \min_{z_t, u_t} \sum_{t=1}^N \ell_t^z(z_t) + \sum_{t=0}^{N-1} \ell_t^u(u_t) \\ \text{s.l.: } z_{t+1} = A_t z_t + B_t u_t \quad \forall t = 0, \dots, N-1. \end{aligned} \quad (12.3)$$

Problema de control optimal (12.3) se reduce la o problemă de optimizare cu constrângeri de egalitate în forma (NLPe). Într-adevăr, definim variabila de optimizare:

$$x = [u_0^T \ z_1^T \ u_1^T \ \dots \ u_{N-1}^T \ z_N^T]^T \in \mathbb{R}^{N(n_z+n_u)}$$

și funcția obiectiv

$$f(x) = \sum_{t=1}^N \ell_t^z(z_t) + \sum_{t=0}^{N-1} \ell_t^u(u_t).$$

Observăm că funcția obiectiv este bloc separabilă și deci Hessiana lui f este bloc diagonală:

$$\nabla^2 f(x) = \text{diag}(R_0(x), Q_1(x), \dots, R_{N-1}(x), Q_N(x)),$$

unde $R_t(x) = \nabla^2 \ell_t^u(u_t)$ și $Q_t(x) = \nabla^2 \ell_t^z(z_t)$. Colectăm toate constrângerile de egalitate date de dinamici pentru $t = 0, \dots, N-1$ în

$Ax = b$ (i.e. $h(x) = Ax - b$), unde $A \in \mathbb{R}^{Nn_z \times N(n_z+n_u)}$ și $b \in \mathbb{R}^{Nn_z}$ sunt date de următoarele expresii:

$$A = \begin{bmatrix} -B_0 & I_{n_z} & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & -A_1 & -B_1 & I_{n_z} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -A_{N-1} & -B_{N-1} & I_{n_z} \end{bmatrix} \text{ și } b = \begin{bmatrix} A_0 z_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

12.1 Metode pentru QP cu constrângeri de egalitate

Fie problema de optimizare pătratică:

$$(QPe) : \min_{x \in \mathbb{R}^n} \frac{1}{2} x^T Q x - q^T x \\ \text{s.l.: } Ax = b,$$

unde matricea $Q \in \mathbb{R}^{n \times n}$ este simetrică (posibil indefinită) și $A \in \mathbb{R}^{p \times n}$. Condițiile KKT conduc la sistemul de ecuații liniare (sistemul Lagrange):

$$Qx - q + A^T \mu = 0 \\ Ax = b,$$

sau în notație matriceală:

$$\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \mu \end{bmatrix} = \begin{bmatrix} q \\ b \end{bmatrix}.$$

Definim matricea $K = \begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix}$ numită *matricea (KKT)*.

Lema 12.1.1 *Observăm că matricea (KKT) este întotdeauna indefinită. Dacă matricea $A \in \mathbb{R}^{p \times n}$ are rangul p și pentru orice $d \in \text{kernel}(A)$ cu $d \neq 0$ avem $d^T Q d > 0$, atunci matricea (KKT) este inversabilă.*

Demonstrație: Reamintim definiția subspațiului $\text{kernel}(A) = \{x \in \mathbb{R}^n : Ax = 0\}$. O formulare echivalentă a proprietății de inversabilitate pentru o matrice pătratică implică faptul că singură soluție a sistemului:

$$\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} y = 0, \quad (12.4)$$

este vectorul 0. Dacă partiționăm soluțiile sistemului astfel: $y = \begin{bmatrix} u \\ v \end{bmatrix}$, rămâne de arătat că singurele instanțe ale vectorilor u și v care satisfac sistemul (12.4) sunt nule. Din (12.4) avem:

$$Qu + A^T v = 0 \quad \text{și} \quad Au = 0.$$

A doua ecuație indică $u \in \text{kernel}(A)$. Înmulțind la stânga în prima ecuație cu u^T obținem:

$$u^T Qu + u^T A^T v = 0. \quad (12.5)$$

Ținând cont că $u \in \text{kernel}(A)$, iar în enunț am presupus că $u^T Qu > 0$, concluzionăm că relația (12.5) este imposibil de satisfăcut pentru orice $u \neq 0$. Pe de altă parte, considerând $u = 0$ și $v \neq 0$, prima ecuație a sistemului (12.4) devine $A^T v = 0$. Deoarece matricea A are rang maxim pe linii, matricea A^T are rang maxim pe coloane, și deci singurul vector ce satisface $A^T v = 0$ este $v = 0$. În final, am arătat că unicul vector ce satisface sistemul (12.4) este vectorul nul.

Dacă A are rang p , atunci condiția de regularitate este îndeplinită pentru orice punct $x \in \mathbb{R}^n$ în problema (QPe) precedentă. Mai departe, dacă pentru orice $d \in \text{kernel}(A)$ cu $d \neq 0$ avem $d^T Q d > 0$, atunci condițiile suficiente de ordinul II sunt satisfăcute pentru (QPe). În concluzie, pentru o problemă pătratică cu constrângeri de egalitate în forma (QPe), existența unui minim local este echivalentă cu inversabilitatea matricei (KKT). Există multe modalități de rezolvare a sistemului (KKT) anterior. Dintre aceste metode, enumerăm:

- (i) *metoda factorizării LU* se bazează pe factorizarea LU a matricei KKT și apoi rezolvarea celor două sisteme triunghiulare. Datorită faptului că K este matrice indefinită, nu putem folosi factorizare Cholesky. În schimb, putem utiliza eliminarea gaussiană cu pivotare parțială pentru a obține factorii L și U . În această factorizare nu luăm în calcul simetria. De aceea, în general se utilizează o factorizare Cholesky indefinită: $P^T K P = LDL^T$, unde P este o matrice de permutare, L este inferior triunghiulară și D este o matrice bloc diagonală cu blocuri de dimensiune 1 sau 2. Putem de asemenea utiliza metode iterative din algebra liniară sau optimizare (e.g. metoda gradientilor conjugați sau metode Krylov).
- (ii) *metoda complementului Schur* presupune că matricea Q este inversabilă și A are rangul p și se bazează pe eliminarea lui x din

prima ecuație:

$$x = -Q^{-1}(A^T\mu - q)$$

și apoi introducerea acestei expresii în cea de-a doua ecuație obținând μ

$$AQ^{-1}A^T\mu = AQ^{-1}q - b.$$

Rezolvăm sistemul liniar în μ cu matricea $AQ^{-1}A^T$ pozitiv definită și apoi recuperăm x . Această metodă necesită ca Q să fie inversabilă, ceea ce nu este întotdeauna valabil, și apoi calcularea factorizării matricei $AQ^{-1}A^T$, care este complementul Schur al lui Q în matricea K .

- (iii) *metoda spațiului nul* se bazează pe găsirea unei baze $Z \in \mathbb{R}^{n \times (n-p)}$ pentru $\text{kernel}(A)$ și apoi se definește $x = Zv + y$, unde y este o soluție particulară a sistemului $Ay = b$. Orice $x = Zv + y$ satisface $Ax = b$, astfel încât trebuie să luăm în considerare doar prima ecuație din sistemul (KKT). Aceasta se poate reformula ca o problemă de minimizare neconstrânsă:

$$\min_{v \in \mathbb{R}^{n-p}} \frac{1}{2}(Zv + y)^T Q(Zv + y) - q^T(Zv + y).$$

Condițiile de ordinul I pentru probleme de optimizare fără constrângeri conduc la:

$$Z^T Q Z v + Z^T Q y - Z^T q = 0 \quad \Longleftrightarrow \quad v = (Z^T Q Z)^{-1}(Z^T q - Z^T Q y).$$

Matricea $Z^T Q Z$ este numită *Hessiana redusă*. Dacă v^* este o soluție a problemei QP fără constrângeri, atunci $x^* = Zv^* + y$. Această metodă poate fi aplicată dacă condițiile suficiente de ordinul II sunt satisfăcute.

Exemplul 12.1.1 *Considerăm problema de optimizare pătratică:*

$$\begin{aligned} \min_{x \in \mathbb{R}^3} \quad & 3x_1^2 + 2x_1x_2 + x_1x_3 + 2.5x_2^2 + 2x_2x_3 + 2x_3^2 - 8x_1 - 3x_2 - 3x_3 \\ \text{s.l.:} \quad & x_1 + x_2 = 3, \quad x_2 + x_3 = 0. \end{aligned}$$

În acest caz, matricele corespunzătoare acestui (QP) sunt:

$$Q = \begin{bmatrix} 6 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 4 \end{bmatrix}, \quad q = \begin{bmatrix} 8 \\ 3 \\ 3 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 3 \\ 0 \end{bmatrix}.$$

Observăm că matricea ce definește spațiul nul al lui A este $Z = [-1 \ -1 \ 1]^T$. În acest caz, problema redusă neconstrânsă în variabila v este unidimensională. Putem găsi imediat soluția $x^* = [2 \ -1 \ 1]^T$ și $\mu^* = [3 \ -2]^T$ care, datorită faptului că Q este matrice pozitiv definită, este soluția de minim global al problemei (QP).

În cele ce urmează vom prezenta diferite metode pentru rezolvarea sistemului Lagrange pe cazul general (KKT-NLPe).

12.2 Metode Lagrange

Metodele Lagrange se bazează pe condițiile (KKT-NLPe):

$$\nabla_x \mathcal{L}(x, \mu) = 0, \quad h(x) = 0 \quad \Longleftrightarrow \quad \nabla \mathcal{L}(x, \mu) = 0.$$

Definim variabila:

$$y = \begin{bmatrix} x \\ \mu \end{bmatrix} \quad \text{și} \quad F(y) = \nabla \mathcal{L}(x, \mu) = \begin{bmatrix} \nabla_x \mathcal{L}(x, \mu) \\ h(x) \end{bmatrix},$$

unde $y \in \mathbb{R}^{n+p}$ și $F : \mathbb{R}^{n+p} \rightarrow \mathbb{R}^{n+p}$, astfel încât soluțiile problemei de optimizare (NLPe) se găsesc printre rădăcinile sistemului neliniar:

$$F(y) = 0,$$

Acest sistem poate fi rezolvat prin metode Lagrange de tip gradient sau Newton. Metodele Lagrange au următoarea formă generică:

$$(ML) : \quad \begin{aligned} x_{k+1} &= L(x_k, \mu_k) \\ \mu_{k+1} &= H(x_k, \mu_k), \end{aligned}$$

în care funcțiile $L : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^n$ și $H : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^p$ sunt funcții diferențiabile. În plus, iterația anterioară converge la un (x^*, μ^*) dacă aceste funcții satisfac condițiile $x^* = L(x^*, \mu^*)$ și $\mu^* = H(x^*, \mu^*)$. Mai mult, pentru a asigura convergența globală a acestor metode, iterațiile pot fi dependente și de un pas α_k . Există diferite posibilități de alegere a pasului în metodele Lagrange, însă cea mai des utilizată se bazează pe o funcție merit asociată problemei (NLPe). Un exemplu de funcție merit este definit de:

$$\mathcal{M}(x, \mu) = \frac{1}{2} \|\nabla_x \mathcal{L}(x, \mu)\|^2 + \frac{1}{2} \|h(x)\|^2.$$

Remarcăm că $\mathcal{M}(x, \mu) \geq 0$ și $\mathcal{M}(x, \mu) = 0$ dacă și numai dacă $\nabla_x \mathcal{L}(x, \mu) = 0$ și $h(x) = 0$, adică această funcție măsoară cât de aproape este un punct (x, μ) de soluție. Așadar, punctele de minim global ale funcției merit $\mathcal{M}(x, \mu)$ satisfac condițiile necesare de ordinul I (KKT-NLPe) pentru problema cu constrângeri de egalitate (NLPe) studiată în acest capitol. În concluzie, putem minimiza funcția merit fără constrângeri:

$$\min_{x \in \mathbb{R}^n, \mu \in \mathbb{R}^p} \mathcal{M}(x, \mu) \quad \left(= \frac{1}{2} \|\nabla_x \mathcal{L}(x, \mu)\|^2 + \frac{1}{2} \|h(x)\|^2 \right) \quad (12.6)$$

și orice punct de minim global (x^*, μ^*) al acestei probleme fără constrângeri satisface relația $\nabla \mathcal{L}(x^*, \mu^*) = 0$. Însă, funcția merit precedentă poate avea și minime locale, care în general nu sunt folositoare în găsirea soluției optime a problemei (NLPe). Suntem interesați în găsirea condițiilor suficiente care garantează că un punct de minim local al problemei fără constrângeri (12.6) este minim global.

Lema 12.2.1 *Presupunem că (x^*, μ^*) este punct de minim local al problemei (12.6) care satisface condițiile necesare de ordinul I. Presupunem de asemenea că rangul lui $\nabla h(x^*)$ este p și Hessiana $\nabla_x^2 \mathcal{L}(x^*, \mu^*)$ este pozitiv definită. Atunci (x^*, μ^*) este punct de minim global pentru (12.6), adică $\mathcal{M}(x^*, \mu^*) = 0$.*

Demonstrație: Cum (x^*, μ^*) satisface condițiile necesare de ordinul I pentru o problemă de optimizare fără constrângeri avem că gradientul lui \mathcal{M} în acest punct este 0:

$$\begin{aligned} \nabla_x^2 \mathcal{L}(x^*, \mu^*) \nabla_x \mathcal{L}(x^*, \mu^*) + (\nabla h(x^*))^T h(x^*) &= 0 \\ \nabla h(x^*) \nabla_x \mathcal{L}(x^*, \mu^*) &= 0. \end{aligned}$$

Multiplicând prima relație cu $\nabla_x \mathcal{L}(x^*, \mu^*)$ și folosind apoi cea de-a doua relație, obținem $(\nabla_x \mathcal{L}(x^*, \mu^*))^T \nabla_x^2 \mathcal{L}(x^*, \mu^*) \nabla_x \mathcal{L}(x^*, \mu^*) = 0$. Deoarece $\nabla_x^2 \mathcal{L}(x^*, \mu^*)$ este pozitiv definită, obținem că $\nabla_x \mathcal{L}(x^*, \mu^*) = 0$. Folosind iarăși prima relație de optimalitate avem următoarea identitate în punctul x^* $(\nabla h(x^*))^T h(x^*) = 0$ și cum rangul lui $\nabla h(x^*)$ este p rezultă $h(x^*) = 0$.

Din condițiile de optimalitate de ordinul I observăm că un posibil criteriu de oprire pentru metodele ce vor fi dezvoltate în acest capitol este următorul:

$$\|\nabla \mathcal{L}(x_k, \mu_k)\| \leq \epsilon,$$

pentru o acuratețe fixată $\epsilon > 0$.

12.2.1 Metoda Lagrange de ordinul I

Cea mai simplă metodă Lagrange, numită și *metoda Lagrange de ordinul I*, este dată de iterația:

$$\begin{aligned} (ML - I) : \quad x_{k+1} &= x_k - \alpha_k \nabla_x \mathcal{L}(x_k, \mu_k) \\ \mu_{k+1} &= \mu_k + \alpha_k h(x_k), \end{aligned}$$

unde $\alpha_k > 0$ este un pas. Dacă utilizăm notația de mai înainte $y = [x^T \mu^T]^T$, atunci iterația precedentă se scrie sub forma:

$$y_{k+1} = y_k - \alpha_k F(y_k) \quad \Longleftrightarrow \quad y_{k+1} = y_k - \alpha_k \nabla \mathcal{L}(y_k),$$

care bineînțeles poate fi interpretată ca metoda gradient pentru sistemul de ecuații neliniare $F(y^*) = \nabla \mathcal{L}(y^*) = 0$.

Pe de altă parte putem interpreta iterația (ML-I) ca o metodă de descreștere pentru problema fără constrângeri (12.6) dacă presupunem că Hessiana Lagrangianului $\nabla_x^2 \mathcal{L}(x, \mu)$ este pozitiv definită într-o regiune de interes. Într-adevăr se poate arăta că direcția dată de următoarea expresie $(-\nabla_x \mathcal{L}(x_k, \mu_k), h(x_k))$ este o direcție de descreștere pentru funcția \mathcal{M} în punctul (x_k, μ_k) . Pentru o mai simplă expunere, introducem notațiile:

$$L_k = \nabla_x^2 \mathcal{L}(x_k, \mu_k), \quad l_k = \nabla_x \mathcal{L}(x_k, \mu_k), \quad h_k = h(x_k) \text{ și } A_k = \nabla h(x_k).$$

Atunci, produsul scalar dintre această direcție și gradientul lui \mathcal{M} capătă următoarea formă:

$$\begin{aligned} [(L_k l_k + A_k^T h_k)^T \quad (A_k l_k)^T] [-l_k^T \quad h_k^T]^T &= -l_k^T L_k l_k - h_k^T A_k l_k + l_k^T A_k^T h_k \\ &= -l_k^T L_k l_k < 0, \end{aligned}$$

sub ipoteza că Hessiana $L_k = \nabla_x^2 \mathcal{L}(x_k, \mu_k)$ este pozitiv definită și $l_k = \nabla_x \mathcal{L}(x_k, \mu_k) \neq 0$. În concluzie, putem alege pasul α_k care minimizează funcția merit de-a lungul acestei direcții de descreștere:

$$\alpha_k = \arg \min_{\alpha \geq 0} \mathcal{M}(x_k - \alpha \nabla_x \mathcal{L}(x_k, \mu_k), \mu_k + \alpha h(x_k)).$$

Folosind rezultatele de convergență ale metodelor de descreștere din partea a doua a acestei lucrări, putem concluziona că iterația (ML-I) va converge către un punct ce satisface $\nabla_x \mathcal{L}(x^*, \mu^*) = 0$. Dar, nu putem garanta că $h(x^*) = 0$. Putem îmbunătăți proprietățile de convergență

ale acestei metode prin alegerea altei funcții merit. De exemplu, putem utiliza următoarea funcție merit:

$$\mathcal{M}_\gamma(x, \mu) = \frac{1}{2} \|\nabla_x \mathcal{L}(x, \mu)\|^2 + \frac{1}{2} \|h(x)\|^2 - \gamma \mathcal{L}(x, \mu),$$

unde $\gamma > 0$ este suficient de mic. În aceeași manieră ca mai înainte se poate arăta că direcția $(-\nabla_x \mathcal{L}(x_k, \mu_k), h(x_k))$ este o direcție de descreștere pentru \mathcal{M}_γ , și anume:

$$\begin{aligned} & [(L_k l_k + A_k^T h_k - \gamma l_k)^T \quad (A_k l_k - \gamma h_k)^T] [-l_k^T \quad h_k^T]^T \\ & = -l_k^T (L_k - \gamma I_n) l_k - \gamma h_k^2 < 0, \end{aligned}$$

sub ipoteza că Hessiana $L_k = \nabla_x^2 \mathcal{L}(x_k, \mu_k)$ este pozitiv definită și $l_k = \nabla_x \mathcal{L}(x_k, \mu_k) \neq 0$ sau $h_k = h(x_k) \neq 0$. În concluzie, dacă alegem α_k care minimizează funcția merit \mathcal{M}_γ de-a lungul acestei direcții de descreștere se poate arăta că iterația (ML-I) converge către un punct ce satisface $\nabla_x \mathcal{L}(x^*, \mu^*) = 0$ și $h(x^*) = 0$, adică către un punct ce satisface condițiile (KKT-NLPe) pentru problema originală (NLPe).

Vom analiza acum convergența metodei (ML-I) pentru pas constant $\alpha_k = \alpha$. Pentru aceasta introducem noțiunea de *punct de atracție*. O pereche (x^*, μ^*) se numește punct de atracție pentru iterația (ML) dacă există o mulțime deschisă $V \subset \mathbb{R}^{n+p}$ astfel încât pentru orice $(x_0, \mu_0) \in S$ șirul (x_k, μ_k) generat de iterație rămâne în S și converge la (x^*, μ^*) .

Teorema 12.2.1 Fie $L : \mathbb{R}^{n+p} \rightarrow \mathbb{R}^n$ și $H : \mathbb{R}^{n+p} \rightarrow \mathbb{R}^p$ două funcții diferentiabile astfel încât $x^* = L(x^*, \mu^*)$ și $\mu^* = H(x^*, \mu^*)$. Mai mult, presupunem că toate valorile proprii ale matricei de dimensiune $(n+p) \times (n+p)$

$$R^* = \begin{bmatrix} \nabla_x L(x^*, \mu^*) & \nabla_x H(x^*, \mu^*) \\ \nabla_\mu L(x^*, \mu^*) & \nabla_\mu H(x^*, \mu^*) \end{bmatrix}$$

sunt în interiorul cercului unitate. Atunci, (x^*, μ^*) este un punct de atracție al iterației (ML) și când șirul generat (x_k, μ_k) converge la (x^*, μ^*) , rata de convergență este liniară.

Demonstrație: Notăm cu $y = [x^T \quad \mu^T]^T$ și cu $M(y) = [L(x, \mu) \quad H(x, \mu)]$. Din teorema valorii medii avem că pentru orice doi vectori y și \tilde{y} :

$$M(\tilde{y}) - M(y) = R^T(\tilde{y} - y),$$

unde R este o matrice având coloana i definită de gradientul $\nabla M_i(\hat{y}_i)$ al componentei i a lui M evaluat într-un vector \hat{y}_i ce se găsește pe segmentul având capetele în \tilde{y} și y . Luând \tilde{y} și y suficient de aproape de y^* , putem face această matrice R să fie apropiată de R^* și deci putem asigura existența valorilor proprii ale matricei R sau echivalent R^T în cercul unitate. În concluzie, există o vecinătate S a lui $y^* = [(x^*)^T (\mu^*)^T]^T$ astfel încât în această vecinătate norma matricei R^T este mai mică decât $1 - \epsilon$, unde ϵ este un scalar pozitiv. Aplicând norma în relația anterioară obținem:

$$\|M(\tilde{y}) - M(y)\| \leq \|R^T\| \|\tilde{y} - y\|.$$

În concluzie, în vecinătatea S a lui (x^*, μ^*) operatorul M este o contracție și rezultatul urmează apoi din teorema de contracție (vezi Apendice).

Demonstrăm acum convergența locală a metodei Lagrange de ordinul I:

Teorema 12.2.2 *Presupunem că f și h sunt funcții de două ori diferențiabile, x^* este punct de minim local pentru care există μ^* satisfăcând condițiile (KKT-NLPe). Presupunem de asemenea că x^* este regulat și Hessiana $\nabla_x^2 \mathcal{L}(x^*, \mu^*)$ este pozitiv definită. Atunci există $\bar{\alpha} > 0$ astfel încât pentru orice $\alpha \in (0, \bar{\alpha}]$, (x^*, μ^*) este un punct de atracție al iterației (ML-I) și dacă șirul generat (x_k, μ_k) converge la (x^*, μ^*) , atunci rata de convergență este liniară.*

Demonstrație: Vom arăta că pentru α suficient de mic ipotezele teoremei anterioare sunt satisfăcute. Considerăm funcția:

$$M_\alpha(x, \mu) = \begin{bmatrix} x - \alpha \nabla_x \mathcal{L}(x, \mu) \\ \lambda + \alpha \nabla_\mu \mathcal{L}(x, \mu) \end{bmatrix}.$$

Este clar că $M_\alpha(x^*, \mu^*) = [(x^*)^T (\mu^*)^T]^T$ și $\nabla M_\alpha(x^*, \mu^*) = I_{n+p} - \alpha B$, unde:

$$B = \begin{bmatrix} \nabla_x^2 \mathcal{L}(x^*, \mu^*) & \nabla h(x^*) \\ -\nabla h(x^*) & 0 \end{bmatrix}.$$

Arătăm că partea reală a fiecărei valori proprii a matricei B este strict pozitivă. Pentru orice vector y , notăm \bar{y} conjugatul lui și cu $\text{Re}(\cdot)$ partea reală a unui număr complex. Fie γ o valoare proprie a lui B având vectorul propriu corespunzător $y = [z^T w^T]^T \neq 0$. Avem:

$$\text{Re}(\bar{y}^T B y) = \text{Re}(\gamma)(\|z\|^2 + \|w\|^2),$$

și în același timp:

$$\text{Re}(\bar{y}^T B y) = \text{Re}(\bar{z}^T \nabla_x^2 \mathcal{L}(x^*, \mu^*) z + \bar{z}^T \nabla h(x^*) w - \bar{w}^T \nabla h(x^*) z).$$

Deoarece $\operatorname{Re}(\bar{z}^T Dw) = \operatorname{Re}(\bar{w}^T Dz)$ pentru orice matrice D cu intrări numere reale, obținem:

$$\operatorname{Re}(\bar{y}^T By) = \operatorname{Re}(\bar{z}^T \nabla_x^2 \mathcal{L}(x^*, \mu^*) z) = \operatorname{Re}(\gamma)(\|z\|^2 + \|w\|^2).$$

Dar pentru orice matrice D pozitiv definită avem $\operatorname{Re}(\bar{z}^T Dz) > 0$ pentru orice $z \neq 0$. În concluzie, din relația precedentă avem fie $\operatorname{Re}(\gamma) > 0$ sau $z = 0$. Dar dacă $z = 0$, din definiția valorii proprii $By = \gamma y$ obținem $\nabla h(x^*)w = 0$. Însă x^* este punct regulat și deci $\nabla h(x^*)$ are rangul p , ceea ce implică că $w = 0$. Aceasta contrazice faptul că vectorul propriu $y \neq 0$. Obținem că $\operatorname{Re}(\gamma) > 0$. \square

12.2.2 Metoda Lagrange-Newton

Dorim acum să aplicăm metoda Newton pentru a rezolva sistemul de ecuații dat de condițiile (KKT - NLPe):

$$\begin{aligned} \nabla \mathcal{L}(x, \mu) &= 0 \\ h(x) &= 0. \end{aligned}$$

Utilizând notațiile de mai înainte, dorim să rezolvăm sistemul neliniar

$$F(y) = 0,$$

unde $y = \begin{bmatrix} x \\ \mu \end{bmatrix} \in \mathbb{R}^{n+p}$ și $F : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^{n+m}$, $F(y) = \begin{bmatrix} \nabla \mathcal{L}(x, \lambda) \\ h(x) \end{bmatrix}$. Acest sistem poate fi rezolvat prin metoda Newton:

$$F(y_k) + \frac{\partial F}{\partial y}(y_k)(y - y_k) = 0 \quad (12.7)$$

care conduce la următoarea iterație Newton, sub ipoteza că Jacobianul $\nabla F(y_k) = \frac{\partial F}{\partial y}(y_k)$ este matrice inversabilă:

$$y_{k+1} = y_k - \left(\frac{\partial F}{\partial y}(y_k) \right)^{-1} F(y_k).$$

Iterația (12.7) poate fi scrisă în termeni de gradienti astfel:

$$\begin{aligned} \nabla_x \mathcal{L}(x_k, \mu_k) + \nabla_x^2 \mathcal{L}(x_k, \mu_k)(x - x_k) + (\nabla h(x_k))^T(\mu - \mu_k) &= 0 \\ h(x_k) + \nabla h(x_k)(x - x_k) &= 0. \end{aligned}$$

Scrisă în forma matriceală, obținem următorul sistem liniar:

$$\underbrace{\begin{bmatrix} \nabla_x^2 \mathcal{L}(x_k, \mu_k) & (\nabla h(x_k))^T \\ \nabla h(x_k) & 0 \end{bmatrix}}_{\text{matricea KKT}} \begin{bmatrix} x - x_k \\ \mu - \mu_k \end{bmatrix} = \begin{bmatrix} -\nabla_x \mathcal{L}(x_k, \mu_k) \\ -h(x_k) \end{bmatrix}.$$

Este clar că pentru orice soluție (x^*, μ^*) pentru care condițiile suficiente de ordinul II sunt satisfăcute, matricea KKT este inversabilă într-o vecinătate a lui (x^*, μ^*) . În acest caz, avem soluție unică pentru sistemul liniar anterior. Obținem următoarea iterație Newton:

$$(ML - N) : \quad \begin{aligned} x_{k+1} &= x_k + d_k \\ \mu_{k+1} &= \mu_k + d_k^\mu, \end{aligned}$$

unde direcțiile (d_k, d_k^μ) sunt soluția sistemului:

$$\begin{bmatrix} \nabla_x^2 \mathcal{L}(x_k, \mu_k) & (\nabla h(x_k))^T \\ \nabla h(x_k) & 0 \end{bmatrix} \begin{bmatrix} d_k \\ d_k^\mu \end{bmatrix} = \begin{bmatrix} -\nabla_x \mathcal{L}(x_k, \mu_k) \\ -h(x_k) \end{bmatrix}. \quad (12.8)$$

Rezultatele standard de convergență locală ale metodei Newton aplicate unui sistem de ecuații neliniare sunt aplicabile și pentru sistemul neliniar $F(y) = 0$. Aceste rezultate afirmă că dacă sistemul liniarizat la soluția (x^*, μ^*) este inversabil (în cazul nostru această condiție este îndeplinită dacă condițiile suficiente de ordinul II în (x^*, μ^*) au loc pentru problema (NLPe)) și dacă punctul inițial (x_0, μ_0) este suficient de aproape de soluția (x^*, μ^*) , atunci șirul (x_k, μ_k) generat de metoda Newton (ML - N) este convergent și converge către soluție cu rată cel puțin pătratică.

Putem arăta că direcția generată de metoda Newton este o direcție de descreștere pentru funcția merit:

$$\mathcal{M}(x, \mu) = \frac{1}{2} \|\nabla_x \mathcal{L}(x, \mu)\|^2 + \frac{1}{2} \|h(x)\|^2.$$

Reamintim notațiile: $L_k = \nabla_x^2 \mathcal{L}(x_k, \mu_k)$, $l_k = \nabla_x \mathcal{L}(x_k, \mu_k)$, $h_k = h(x_k)$ și $A_k = \nabla h(x_k)$. Atunci, produsul scalar dintre direcțiile Newton (d_k, d_k^μ) și gradientul lui \mathcal{M} capătă următoarea formă:

$$\begin{aligned} [L_k l_k + A_k^T h_k \quad A_k l_k] [d_k^T \quad (d_k^\mu)^T]^T &= l_k^T L_k d_k + h_k^T A_k d_k + l_k^T A_k^T d_k^\mu \\ &= -\|l_k\|^2 - \|h_k\|^2. \end{aligned}$$

Această expresie este strict negativă, cu excepția cazului când $l_k = 0$ și $h_k = 0$ (adică condițiile (KKT-NLPe)). Din interpretarea metodei

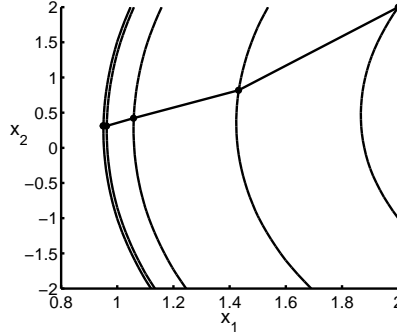


Figura 12.1: *Iterația metodei (ML-N) cu punctul inițial $x_0 = [2 \ 2]^T$ pentru problema neconvexă $\min_{x \in \mathbb{R}^2: x_1^2 + x_2^2 = 1} (x_1 - 6)^4 + (x_1 - 4x_2)^2$.*

Lagrange-Newton ca o metodă de descreștere pentru o funcție merit, putem concluziona că metoda Newton are proprietăți de convergență globală când iterația se ia cu un pas variabil. Definim metoda Lagrange-Newton generală prin următoarea iterație:

$$(ML - N_\alpha) : \quad \begin{aligned} x_{k+1} &= x_k + \alpha_k d_k \\ \mu_{k+1} &= \mu_k + \alpha_k d_k^\mu, \end{aligned}$$

unde (d_k, d_k^μ) sunt direcțiile Newton din sistemul (12.8) și α_k este ales să minimizeze funcția merit, i.e.

$$\alpha_k = \arg \min_{\alpha \geq 0} \mathcal{M}(x_k + \alpha d_k, \mu_k + \alpha d_k^\mu).$$

Folosind rezultatele de convergență a metodelor de descreștere pentru cazul problemelor de optimizare neconstrânsă (Capitolul 4) se poate arăta că orice punct limită al șirului generat de metoda Lagrange-Newton cu pas variabil $(ML - N_\alpha)$ satisface condițiile necesare de ordinul I pentru problema cu constrângeri de egalitate (NLPe), adică condițiile (KKT-NLPe).

Interpretarea metodei Lagrange-Newton folosind formularea QP secvențială: Dacă adăugăm expresia $(\nabla h(x_k))^T \mu_k$ în prima ecuație a sistemului (12.8), atunci acest sistem se rescrie în forma:

$$\begin{bmatrix} \nabla_x^2 \mathcal{L}(x_k, \mu_k) & (\nabla h(x_k))^T \\ \nabla h(x_k) & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \mu_{k+1} \end{bmatrix} = \begin{bmatrix} -\nabla f(x_k) \\ -h(x_k) \end{bmatrix}.$$

Sistemul liniar (12.8) se scrie explicit astfel:

$$\nabla f(x_k) + \nabla_x^2 \mathcal{L}(x_k, \mu_k) d_k + (\nabla h(x_k))^T \mu_{k+1} = 0, \quad h(x_k) + \nabla h(x_k) d_k = 0.$$

Aceste două relații sunt de fapt condițiile (KKT) pentru o problemă pătratică, adică d_k și μ_{k+1} sunt soluțiile optime (punctele staționare) obținute din rezolvarea unui QP de forma:

$$\begin{aligned} \min_{d \in \mathbb{R}^n} \quad & \nabla f(x_k)^T d + \frac{1}{2} d^T \nabla_x^2 \mathcal{L}(x_k, \mu_k) d \\ \text{s.l.:} \quad & h(x_k) + \nabla h(x_k) d = 0. \end{aligned} \quad (12.9)$$

Această metodă se mai numește și *metoda pătratică secvențială*, deoarece pentru problema originală (NLP) se construiește un model pătratic pentru funcția obiectiv și constrângerile de egalitate se linarizează în punctul curent. Mai general, putem înlocui Hessiana Lagrangianului $\nabla_x^2 \mathcal{L}(x_k, \mu_k)$ printr-o aproximare B_k , unde $B_k = B_k^T$ și adesea $B_k \succeq 0$. Matricea B_k poate fi aleasă în diferite moduri. O posibilitate ar fi să alegem matricea B_k constantă de-a lungul tuturor iterațiilor. O altă posibilitate ar fi să alegem B_k folosind informație din $\nabla_x^2 \mathcal{L}(x_k, \mu_k)$, de exemplu B_k este egală cu diagonală lui $\nabla_x^2 \mathcal{L}(x_k, \mu_k)$. În final, matricea B_k se poate actualiza folosind updatări cvasi Newton de rang unu sau doi pe baza ecuației secantei:

$$B_{k+1}(x_{k+1} - x_k) = \nabla_x \mathcal{L}(x_{k+1}, \mu_{k+1}) - \nabla_x \mathcal{L}(x_k, \mu_{k+1}).$$

12.3 Metoda Newton pentru probleme convexe având constrângeri de egalitate

O problemă de optimizare convexă având constrângeri de egalitate are forma:

$$(CPe) : \quad \min_{x \in \mathbb{R}^n} f(x) \quad (12.10)$$

$$\text{s.l.:} \quad Ax = b, \quad (12.11)$$

în care $f : \mathbb{R}^n \rightarrow \mathbb{R}$ este funcție convexă și $A \in \mathbb{R}^{p \times n}$ are rangul p . În acest caz, condițiile (KKT) sunt necesare și suficiente, adică x^* este punct de minim dacă și numai dacă există $\mu^* \in \mathbb{R}^p$ astfel încât:

$$(KKT - CPe) : \quad \nabla f(x^*) + A^T \mu^* = 0, \quad Ax^* = b.$$

În problema (CPe) putem elimina constrângerile de egalitate pe baza observației:

$$\{x \in \mathbb{R}^n : Ax = b\} = \{x \in \mathbb{R}^n : x = Zv + y, v \in \mathbb{R}^{n-p}\},$$

unde coloanele lui Z reprezintă o bază pentru $\text{kernel}(A)$ și y este o soluție particulară a sistemului $Ax = b$. În acest caz putem rezolva problema de optimizare convexă fără constrângeri:

$$\min_{v \in \mathbb{R}^{n-p}} \bar{f}(v) \quad (= f(Zv + y)). \quad (12.12)$$

Dacă v^* este o soluție a problemei convexe fără constrângeri, atunci $x^* = Zv^* + y$ este o soluție a problemei originale.

Exemplul 12.3.1 Considerăm problema de alocare optimă având constrângeri pe resurse:

$$\min_{x \in \mathbb{R}^n : x_1 + \dots + x_n = b} f(x).$$

Putem elimina de exemplu $x_n = b - x_1 - \dots - x_{n-1}$ care corespunde lui $y = be_n$ și $Z = \begin{bmatrix} I_{n-1} \\ -e^T \end{bmatrix} \in \mathbb{R}^{n \times n-1}$. Problema fără constrângeri devine:

$$\min_{x \in \mathbb{R}^{n-1}} f(x_1, \dots, x_{n-1}, b - x_1 - \dots - x_{n-1}).$$

Prezentăm în cele ce urmează o extensie a metodei Newton pentru cazul neconstrâns la cel cu constrângeri de egalitate pentru probleme convexe: fie x_k fezabil, următorul punct al iterației Newton se calculează folosind direcția Newton d_k ce rezultă din aproximarea pătratică de ordinul II a lui f în jurul lui x_k , i.e.

$$d_k = \arg \min_{d \in \mathbb{R}^n} f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T \nabla^2 f(x_k) d \quad (12.13)$$

s.l.: $A(x_k + d) = b$.

Subproblema (12.13) este o problemă QP convexă având constrângeri de egalitate $Ad = 0$. Așadar, direcția Newton este caracterizată de sistemul liniar:

$$\begin{bmatrix} \nabla^2 f(x_k) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \mu_{k+1} \end{bmatrix} = \begin{bmatrix} -\nabla f(x_k) \\ 0 \end{bmatrix},$$

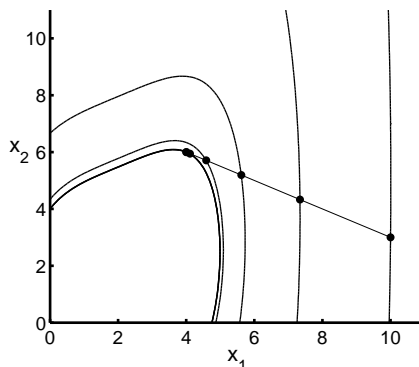


Figura 12.2: Iterațiile metodei Newton cu pasul $\alpha_k = 1$ pentru problema convexă $\min_{x \in \mathbb{R}^2: x_1 + 2x_2 = 16} (x_1 - 2)^4 + (x_1 - 2x_2)^2$.

unde μ_{k+1} este multiplicatorul Lagrange optim asociat problemei pătratice convexe (12.13). Introducem *decrementul Newton*:

$$\nu(x_k) = (d_k^T \nabla^2 f(x_k) d_k)^{1/2}.$$

Putem observa că:

$$f(x_k) - \min_d \left\{ f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T \nabla^2 f(x_k) d : Ad = 0 \right\} = \nu(x_k)^2 / 2,$$

ceea ce arată că decrementul Newton poate fi folosit drept criteriu de oprire. De asemenea,

$$\left. \frac{d}{dt} f(x_k + td_k) \right|_{t=0} = \nabla f(x_k)^T d_k = -\nu(x_k)^2,$$

adică direcția Newton este direcție de descreștere dacă alegem t suficient de mic. În concluzie, metoda Newton generală pentru cazul convex (CPe) este dată de iterația:

$$(MNe) : \quad x_{k+1} = x_k + \alpha_k d_k,$$

unde direcția Newton d_k este soluția problemei QP (12.13), pasul α_k se alege prin procedura backtracking în raport cu direcția d_k și criteriul de oprire folosit este:

$$\nu(x_k)/2 \leq \epsilon,$$

pentru o acuratețe fixată ϵ . Alegerea pasului α_k prin backtracking constă în alegerea unui $\rho \in (0, 1)$ și respectiv $c_1 \in (0, 1)$ și atunci $\alpha_k = \rho^{m_k}$, unde m_k este primul întreg ne-negativ ce satisface:

$$f(x_k + \rho^m d_k) \leq f(x_k) + c_1 \rho^m \nabla f(x_k)^T d_k.$$

În ceea ce privește convergența acestei metode, se poate arăta ușor că iterațiile metodei Newton (MNe) pentru o problemă convexă cu constrângeri de egalitate (CPe) coincid cu iterațiile metodei Newton pentru problema convexă redusă fără constrângeri (12.12). În concluzie, performanța metodei Newton pentru cazul constrângerilor de egalitate este similară cu cea a metodei Newton pentru cazul neconstrâns. În particular, dacă x_k este aproape de soluția optimă x^* , convergența este cel puțin pătratică.

Capitolul 13

Metode de optimizare pentru (NLP) generale

În acest capitol vom prezenta metode numerice de optimizare pentru cazul general al problemelor (NLP) în forma standard:

$$\begin{aligned} (NLP) : \quad & \min_{x \in \mathbb{R}^n} f(x) \\ \text{s.l.:} \quad & g(x) \leq 0, \quad h(x) = 0, \end{aligned} \tag{13.1}$$

în care funcțiile $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ și $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$ sunt funcții diferențiabile de două ori. Metodele dezvoltate în acest capitol au la bază condițiile de optimalitate de ordinul I: dacă x^* este minim local pentru (NLP) și regulat, atunci există un vector $\lambda^* \in \mathbb{R}^m$ și un vector $\mu^* \in \mathbb{R}^p$ astfel încât condițiile (KKT) au loc:

$$\begin{aligned} (\text{KKT}) : \quad & \nabla f(x^*) + \nabla h(x^*)^T \mu^* + \nabla g(x^*)^T \lambda^* = 0 \\ & g(x^*)^T \lambda^* = 0 \\ & g(x^*) \leq 0, \quad h(x^*) = 0 \\ & \mu^* \in \mathbb{R}^p, \quad \lambda^* \geq 0. \end{aligned}$$

Exemplul 13.0.2 (Alocarea optimă de combustibil în generatoare electrice) În două generatoare de electricitate se folosește atât petrol cât și gaz pentru producerea de energie electrică. Energia care trebuie produsă de cele două generatoare este de 50 MW. Cantitatea de gaz disponibilă este de 10 unități/ph. Se dorește selectarea de combustibil (petrol și gaz) pentru fiecare generator astfel încât să se minimizeze cantitatea de petrol utilizată. Din datele de

operare ale fiecărui generator s-a construit o curbă de fitting astfel încât combustibilul necesar în generatorul 1 pentru producerea de x_1 MW poate fi exprimată sub forma:

$$\begin{aligned} w_1(x) &= 1.46 + 0.15x_1 + 0.0014x_1^2 \\ w_2(x) &= 1.57 + 0.16x_1 + 0.0013x_1^2, \end{aligned}$$

unde w_1 și w_2 sunt cantitățile de petrol, respectiv de gaz în unități pe oră. Similar pentru generatorul 2; pentru a produce cantitatea de x_2 MW cerințele de combustibil sunt următoarele:

$$\begin{aligned} v_1(x) &= 0.8 + 0.2x_2 + 0.0009x_2^2 \\ v_2(x) &= 0.72 + 0.22x_2 + 0.0007x_2^2, \end{aligned}$$

unde v_1 și v_2 sunt cantitățile de petrol, respectiv de gaz în unități pe oră. De asemenea, generatorul 1 poate produce putere în intervalul $[18, 30]$ MW și generatorul 2 în intervalul $[14, 25]$ MW. De asemenea, combustibilul poate fi combinat aditiv, adică pentru a produce o cantitate x_1 de energie orice combinație liniară de rate de combustibil utilizat va produce aceeași cantitate de electricitate:

$$x_3w_1(x_1) + (1 - x_3)w_2(x_1),$$

unde $x_3 \in [0, 1]$. La fel pentru generatorul 2. Problema de optimizare se pune astfel: să se determine ratele de producere a energiei x_1 și x_2 cât și fracțiunea de combustibil mixat x_3 și x_4 astfel încât să se minimizeze consumul total de petrol, i.e.

$$\begin{aligned} \min_{x \in \mathbb{R}^4} \quad & x_3w_1(x) + x_4v_1(x) \\ \text{s.l.:} \quad & x_1 + x_2 = 50 \\ & (1 - x_3)w_2(x) + (1 - x_4)v_2(x) \leq 10 \\ & 18 \leq x_1 \leq 30, \quad 14 \leq x_2 \leq 25, \quad 0 \leq x_3 \leq 1, \quad 0 \leq x_4 \leq 1. \end{aligned}$$

Acesta este un exemplu de problemă neconvexă (NLP) având constrângeri de egalitate liniare și constrângeri de inegalitate de tip box și neconvexe.

Exemplul 13.0.3 (Compresia multi-nivel a unui gaz) O anumită cantitate de gaz cu un flux de ϕ moli pe oră la presiunea de 1 atmosferă se urmărește a fi comprimat la 64 atmosfere, folosind un compresor cu trei niveluri. Presupunem că procesul de compresie are loc reversibil

și adiabatic; după fiecare nivel de compresie gazul este răcit până la temperatura inițială T . De cele mai multe ori, se urmărește în mod crucial procesul de compresie cu un consum de energie minim. De aceea, o problemă de optimizare (NLP) intervine în alegerea optimă a presiunilor inter-nivel astfel încât consumul de energie să fie minim. Pentru compresia reversibilă adiabatică a gazului pe un singur nivel, consumul de energie este dat de expresia:

$$Ener = \phi RT \frac{\kappa}{\kappa - 1} \left(\frac{P_{out}}{P_{in}} \right)^{(\kappa-1)/\kappa} - \phi RT \frac{\kappa}{\kappa - 1},$$

în care κ reprezintă raportul capacităților de încălzire ale gazului, T temperatura, R constanta ideală a gazului, iar P_{in} și P_{out} reprezintă presiunea gazului la intrarea, respectiv ieșirea dintr-un nivel. Pentru compresia pe trei niveluri, efortul energetic este dat de:

$$Ener_{totala} = \phi RT \frac{\kappa}{\kappa - 1} \left\{ x_1^\alpha + \left(\frac{x_2}{x_1} \right)^\alpha + \left(\frac{64}{x_2} \right)^\alpha - 3 \right\}.$$

unde x_1 este presiunea gazului la ieșirea din primul nivel, x_2 presiunea la ieșirea din cel de-al doilea nivel și $\alpha = (\kappa - 1)/\kappa$. Dacă considerăm cazul particular când $\alpha = \frac{1}{4}$, iar parametrii ϕ și T sunt fixați, presiunile optime impuse x_1 și x_2 inter-nivel sunt soluțiile următoarelor probleme de optimizare având constrângeri de inegalitate (NLP):

$$\begin{aligned} \min_{x \in \mathbb{R}^2} f(x) &= x_1^{\frac{1}{4}} + \left(\frac{x_2}{x_1} \right)^{\frac{1}{4}} + \left(\frac{64}{x_2} \right)^{\frac{1}{4}} \\ s.l.: x_1 &\geq 1, \quad x_2 \geq x_1, \quad 64 \geq x_2, \end{aligned}$$

unde constrângerile sunt impuse pentru a asigura o presiune a gazului crescătoare de la intrare la ieșire în cele trei niveluri.

13.1 Metoda mulțimilor active

Ideea de bază în metoda mulțimilor active constă în partiționarea constrângerilor de inegalitate în două grupuri: cele care sunt tratate ca constrângeri active și, respectiv, constrângeri inactive. Constrângerile inactive sunt apoi în esență ignorate. Condițiile (KKT) anterioare pot fi exprimate, folosind această partiționare, într-o formă mai simplă: fie $\mathcal{A}(x^*)$ mulțimea de indecși ai constrângerilor active în punctul x^* , adică

conține mulțimea de indecși i pentru care $g_i(x^*) = 0$. Atunci condițiile (KKT) devin:

$$\begin{aligned} \text{(KKT)} : \quad & \nabla f(x^*) + \nabla h(x^*)^T \mu^* + \sum_{i \in \mathcal{A}} \lambda_i^* \nabla g_i(x^*) = 0 \\ & g_i(x^*) < 0 \quad i \notin \mathcal{A}(x^*), \quad g_i(x^*) = 0 \quad i \in \mathcal{A}(x^*), \quad h(x^*) = 0 \\ & \lambda_i^* \geq 0 \quad i \in \mathcal{A}(x^*), \quad \lambda_i^* = 0 \quad i \notin \mathcal{A}(x^*), \quad \mu^* \in \mathbb{R}^p. \end{aligned}$$

Este evident că dacă mulțimea activă în x^* ar fi cunoscută, problema originală (NLP) ar putea fi înlocuită cu o problemă de optimizare corespunzătoare având numai constrângeri de egalitate. Cum mulțimea activă nu e cunoscută a priori, vom alege o mulțime de indecși și vom rezolva problema cu constrângeri de egalitate rezultată cu una din metodele descrise în capitolul anterior. Apoi vom verifica pentru soluția obținută dacă celelalte constrângeri sunt satisfăcute și multiplicatorii Lagrange sunt ne-negativi; în caz afirmativ, această soluție va fi cea căutată pentru problema originală (NLP).

În concluzie, în metoda mulțimilor active vom defini la fiecare iterație un set de constrângeri active, numit *set de lucru*, care va fi tratat ca mulțimea activă. Setul de lucru se alege dintr-un subset al constrângerilor care sunt active în punctul curent și deci punctul curent este fezabil pentru setul de lucru. Algoritmul va produce un nou punct cu performanțe mai bune, ce se găsește pe această suprafață definită de setul de lucru. Noul punct va fi găsit prin utilizarea unei metode de optimizare descrisă în capitolul anterior pentru rezolvarea problemelor neliniare cu constrângeri de egalitate. Presupunem că pentru o mulțime activă de indecși \mathcal{A} rezolvăm probleme de optimizare cu constrângeri de egalitate corespunzătoare:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} f(x) \\ h(x) = 0, \quad g_i(x) = 0 \quad \forall i \in \mathcal{A} \end{aligned} \tag{13.2}$$

și obținem soluția $x_{\mathcal{A}}$. Presupunem, de asemenea, că punctul $x_{\mathcal{A}}$ verifică inegalitățile $g_i(x_{\mathcal{A}}) < 0$ pentru orice $i \notin \mathcal{A}$. Atunci, acest punct satisface condițiile necesare de ordinul I:

$$\nabla f(x_{\mathcal{A}}) + \nabla h(x_{\mathcal{A}})^T \mu + \sum_{i \in \mathcal{A}} \lambda_i \nabla g_i(x_{\mathcal{A}}) = 0.$$

Dacă $\lambda_i \geq 0$ pentru orice $i \in \mathcal{A}$, atunci punctul $x_{\mathcal{A}}$ este o soluție locală (punct staționar) pentru problema (NLP) generală. Pe de altă

parte, dacă există $i \in \mathcal{A}$ astfel încât $\lambda_i < 0$, atunci valoarea funcției obiectiv poate fi micșorată prin relaxarea constrângerii i (aceasta rezultă din interpretarea sensibilității multiplicatorilor Lagrange). Astfel, prin eliminarea din setul de lucru a constrângerii i se poate obține o soluție îmbunătățită. Adesea se întâmplă că în timp ce ne mișcăm pe suprafața de lucru o nouă constrângere de inegalitate devine activă. În acest caz este preferabil să adăugăm această constrângere la setul de lucru și să continuăm căutarea pe această suprafață cu o dimensiune mai mică decât cea anterioară. O strategie pentru metodele de mulțimi active se bazează pe aceste două principii. Începem cu un set de lucru și minimizăm peste suprafața de lucru corespunzătoare. Dacă noile constrângeri devin active în timpul minimizării problemei cu constrângeri de egalitate, atunci ele pot fi adăugate la suprafața de lucru, dar nici o constrângere nu este scoasă din setul de lucru. În final, se obține un punct care minimizează funcția f peste această suprafață. Apoi, multiplicatorii Lagrange sunt determinați și se verifică semnul lor. Dacă toți sunt ne-negativi, atunci soluția obținută este optimă pentru problema originală (NLP). Altfel, una sau mai multe constrângeri corespunzătoare multiplicatorilor negativi se elimină din setul de lucru. Procedura se repetă cu acest nou set de lucru. O astfel de metodă converge, după cum este demonstrat mai jos:

Teorema 13.1.1 *Dacă pentru orice mulțime activă \mathcal{A} , problema cu constrângeri de egalitate (13.2) are o singură soluție nedegenerată (adică pentru orice $i \in \mathcal{A}$ avem $\lambda_i \neq 0$), atunci șirul generat de metoda mulțimilor active converge la un punct staționar (punct ce satisface condițiile (KKT)) pentru problema generală (NLP).*

Demonstrație: După ce o soluție corespunzătoare unui set de lucru este găsită, o descreștere strictă în funcția obiectiv este realizată și deci nu este posibilă reîntoarcerea la această mulțime activă. Cum există un număr finit de mulțimi active, procesul se termină după un număr finit de iterații. \square

13.1.1 Metoda mulțimilor active pentru (QP)

Considerăm problema (QP) având constrângeri de inegalitate (extensia la cazul general când avem și constrângeri de egalitate este evidentă):

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \frac{1}{2} x^T Q x - q^T x \\ \text{s.l.:} \quad & Cx \leq d. \end{aligned}$$

Presupunem un (QP) convex (i.e. matricea $Q \succeq 0$). Cazul în care Q este matrice indefinită ridică multe probleme în algoritmi numerici de optimizare și nu va fi tratat în această carte. Condițiile (KKT) sunt, în acest caz, necesare și suficiente pentru optimalitatea globală:

$$\begin{aligned} Qx^* - q + C^T \lambda^* &= 0 \\ (Cx^* - d)^T \lambda^* &= 0 \\ Cx^* &\leq d, \quad \lambda^* \geq 0. \end{aligned}$$

Definim o mulțime de indecși \mathcal{A} corespunzătoare unor constrângeri active și apoi introducem și complementara acestei mulțimi $\mathcal{I} = \{1, \dots, m\} \setminus \mathcal{A}$ (corespunzătoare constrângerilor inactive). Partiționăm matricele și vectorii corespunzători astfel:

$$C = \begin{bmatrix} C_{\mathcal{A}} \\ C_{\mathcal{I}} \end{bmatrix}, \quad d = \begin{bmatrix} d_{\mathcal{A}} \\ d_{\mathcal{I}} \end{bmatrix}$$

și deci putem partiționa constrângerile de inegalitate după cum urmează:

$$C_{\mathcal{A}} x = d_{\mathcal{A}}, \quad C_{\mathcal{I}} x < d_{\mathcal{I}}.$$

Lema 13.1.1 *Punctul x^* este minim global pentru problema (QP) convexă dacă și numai dacă există o mulțime de indecși \mathcal{A}^* și \mathcal{I}^* și un vector $\lambda_{\mathcal{A}^*}^*$ astfel încât:*

$$\begin{aligned} Qx^* - q + C_{\mathcal{A}^*}^T \lambda_{\mathcal{A}^*}^* &= 0 \\ C_{\mathcal{I}^*} x^* - d_{\mathcal{I}^*} &< 0, \quad C_{\mathcal{A}^*} x^* - d_{\mathcal{A}^*} = 0 \\ \lambda_{\mathcal{A}^*}^* &\geq 0. \end{aligned}$$

În plus, $\lambda^* = [(\lambda_{\mathcal{A}^*}^*)^T \ (\lambda_{\mathcal{I}^*}^*)^T]^T$, unde $\lambda_{\mathcal{I}^*}^* = 0$.

Principalii pași în metoda mulțimilor active pentru problemele (QP) sunt:

1. alegem un punct inițial fezabil x_0 cu mulțimea activă corespunzătoare \mathcal{A}_0 , pentru $k \geq 0$ repetăm următorii pași:
2. rezolvăm sistemul liniar în (\bar{x}_k, λ_k) :

$$\begin{aligned} Q\bar{x}_k + C_{\mathcal{A}_k}^T \lambda_k &= q \\ C_{\mathcal{A}_k} \bar{x}_k &= d_{\mathcal{A}_k} \end{aligned}$$

3. definim $x_{k+1} = x_k + \alpha_k(\bar{x}_k - x_k)$, unde pasul $\alpha_k \in [0, 1]$ astfel încât x_{k+1} să fie fezabil pentru problema originală. Dacă $t_k < 1$ adăugăm la mulțimea de indecși \mathcal{A}_k un index nou i_k corespunzător unei constrângeri pentru care \bar{x}_k nu este fezabil, i.e. $\mathcal{A}_{k+1} = \mathcal{A}_k \cup \{i_k\}$. Dacă $t_k = 1$, atunci \bar{x}_k este fezabil și verificăm dacă $\lambda_k \geq 0$. Dacă această inegalitate are loc, atunci ne oprim. Altfel, eliminăm indexul \bar{i}_k din \mathcal{A}_k corespunzător inegalității $(\lambda_k)_{\bar{i}_k} < 0$, iar $\mathcal{A}_{k+1} = \mathcal{A}_k \setminus \{\bar{i}_k\}$

13.2 Metoda pătratică secvențială

În această secțiune prezentăm o importantă clasă de metode pentru optimizarea constrânsă, numită *metoda pătratică secvențială* (*Sequential Quadratic Programming (SQP)*). Considerăm problema (NLP) generală și liniarizăm această problemă atât în constrângerile de egalitate cât și în cele de inegalitate în punctul curent și construim un model pătratic pentru funcția obiectiv pentru a obține o problemă (QP):

$$\begin{aligned} \min_{d \in \mathbb{R}^n} \quad & \nabla f(x_k)^T d + \frac{1}{2} d^T \nabla_x^2 \mathcal{L}(x_k, \lambda_k, \mu_k) d \\ \text{s.l.:} \quad & g(x_k) + \nabla g(x_k) d \leq 0 \\ & h(x_k) + \nabla h(x_k) d = 0. \end{aligned} \tag{13.3}$$

Se observă că această metodă este generalizarea metodei Lagrange-Newton descrisă în capitolul anterior în (12.9) pentru probleme cu constrângeri de egalitate la cazul în care avem și constrângeri de tip inegalități. De asemenea, putem folosi algoritmul de mulțimi active descris anterior pentru rezolvarea acestei probleme pătratice (13.3). Local, într-o vecinătate a unei soluții (x^*, λ^*, μ^*) putem defini iterația metodei pătratice secvențiale în forma:

$$(x_{k+1} = x_k + d_k, \lambda_{k+1}, \mu_{k+1}),$$

în care d_k și $(\lambda_{k+1}, \mu_{k+1})$ sunt soluția și respectiv multiplicatorii Lagrange pentru inegalități și egalități ai problemei pătratice anterioare (13.3). În această abordare, mulțimea de constrângeri active \mathcal{A}_k la soluția problemei (13.3) o putem presupune a fi cea corespunzătoare problemei generale (NLP) atunci când x_k este suficient de apropiat de x^* . Următoarea teoremă furnizează condițiile în care această proprietate are loc.

Teorema 13.2.1 *Presupunem că x^* este un punct de minim local al problemei (NLP) la care condițiile (KKT) sunt satisfăcute (și deci x^* este punct regulat). De asemenea, presupunem că condițiile suficiente de ordinul II au loc și nu există soluție degenerată (adică $\lambda_j^* > 0$ pentru orice $j \in \mathcal{A}(x^*)$). Atunci dacă (x_k, λ_k, μ_k) este suficient de aproape de (x^*, λ^*, μ^*) , există o soluție locală a problemei (13.3) pentru care mulțimea activă de indici satisface $\mathcal{A}_k = \mathcal{A}(x^*)$, adică soluția problemei pătratice are aceeași mulțime activă ca și problema (NLP).*

Demonstrație: Pentru această demonstrație folosim teorema funcțiilor implicite (vezi Apendice). Definim $\mathcal{A}^* = \mathcal{A}(x^*)$ și complementara acestei mulțimi \mathcal{I}^* . Considerăm funcția F în variabila $(x, d, \lambda_{\mathcal{A}^*}, \mu)$ dată de următoarea expresie:

$$\begin{aligned} \nabla f(x) + \nabla_x^2 \mathcal{L}(x, \lambda_{\mathcal{A}^*}, \mu)d + \nabla h(x)^T \mu + \nabla g_{\mathcal{A}^*}(x)^T \lambda_{\mathcal{A}^*} &= 0 \\ h(x) + \nabla h(x)d &= 0 \\ g_{\mathcal{A}^*}(x) + \nabla g_{\mathcal{A}^*}(x)d &= 0. \end{aligned}$$

Ipotezele din teorema funcțiilor implicite sunt satisfăcute și deci aceasta definește o funcție implicită de forma:

$$\chi(x) = \begin{bmatrix} d(x) \\ \mu(x) \\ \lambda_{\mathcal{A}^*}(x) \end{bmatrix},$$

cu $d(x^*) = 0$, $\mu(x^*) = \mu^*$ și $\lambda_{\mathcal{A}^*}(x) = \lambda_{\mathcal{A}^*}^*$ și ținem cont că $\lambda_{\mathcal{I}^*}^* = 0$. Într-adevăr, următoarele relații au loc:

$$\begin{aligned} \nabla f(x^*) + \nabla_x^2 \mathcal{L}(x, \lambda_{\mathcal{A}^*}^*, \mu^*)0 + \nabla h(x^*)^T \mu^* + \nabla g_{\mathcal{A}^*}(x^*)^T \lambda_{\mathcal{A}^*}^* &= 0 \\ h(x^*) + \nabla h(x^*)0 &= 0 \\ g_{\mathcal{A}^*}(x^*) + \nabla g_{\mathcal{A}^*}(x^*)0 &= 0. \end{aligned}$$

deoarece $\nabla_x \mathcal{L}(x^*, \mu^*, \lambda^*) = 0$, $h(x^*) = 0$, $g_{\mathcal{A}^*}(x^*) = 0$, $g_{\mathcal{I}^*}(x^*) < 0$ și $\lambda_{\mathcal{I}^*}^* = 0$. Observăm că $\lambda_{\mathcal{A}^*}^* > 0$ datorită complementarității stricte. Pentru $x \simeq x^*$, datorită continuității lui $d(x)$ și $\lambda_{\mathcal{A}^*}(x)$, avem în continuare $g_{\mathcal{I}^*}(x) < 0$ și $\lambda_{\mathcal{A}^*}(x) > 0$. Mai mult:

$$g_{\mathcal{I}^*}(x) + \nabla g_{\mathcal{I}^*}(x)d(x) < 0.$$

Astfel, o soluție a problemei pătratice (13.3) are aceeași mulțime activă ca și problema (NLP) și satisface complementaritatea strictă. \square

În concluzie, metoda pătratică secvențială va identifica corect mulțimea activă la optim $\mathcal{A}(x^*)$ și deci va avea un comportament asemănător metodei Lagrange-Newton pentru problemele de optimizare cu constrângeri de egalitate, adică va converge local foarte rapid. Este de asemenea remarcabil că departe de soluție metoda pătratică secvențială este capabilă să îmbunătățească estimarea mulțimii active și dirijează iterațiile către soluție. Departe de soluție, iterația metodei pătratice secvențiale are nevoie de un pas α_k variabil. Alegerea pasului se face pe baza următoarei funcții merit:

$$\mathcal{M}(x, \tau) = f(x) + \tau \|h(x)\|_1 + \tau \sum_{i=1}^m |\max\{0, g_i(x)\}|.$$

Mai exact, fie d_k și $(\bar{\lambda}_{k+1}, \bar{\mu}_{k+1})$ soluția și respectiv multiplicatorii Lagrange pentru inegalități și egalități ai problemei pătratice (13.3). Definim de asemenea $d_k^\mu = \bar{\mu}_{k+1} - \mu_k$ și $d_k^\lambda = \bar{\lambda}_{k+1} - \lambda_k$. Atunci algoritmul general are următoarea iterație:

$$(SQP): \quad x_{k+1} = x_k + \alpha_k d_k, \quad \mu_{k+1} = \mu_k + \alpha_k d_k^\mu, \quad \lambda_{k+1} = \lambda_k + \alpha_k d_k^\lambda,$$

unde pasul α_k se alege prin metoda de backtracking pentru funcția merit de mai sus, adică $\alpha_k = \rho^{m_k}$ pentru un $\rho \in (0, 1)$ și $c_1 \in (0, 1)$, unde m_k este primul întreg ne-negativ care satisface:

$$\mathcal{M}(x_k + \rho^m d_k, \tau_k) \leq \mathcal{M}(x_k, \tau_k) + c_1 \rho^m \mathcal{M}'(x_k, \tau_k; d_k).$$

Reamintim că $\mathcal{M}'(x_k, \tau_k; d_k)$ reprezintă derivata direcțională a lui \mathcal{M} în punctul (x_k, τ_k) de-a lungul direcției d_k . Mai mult, τ_k este un parametru care se ajustează la fiecare iterație, de exemplu putem alege:

$$\tau_k \geq \frac{\nabla f(x_k)^T d_k + 1/2 d_k^T \nabla_x^2 \mathcal{L}(x_k, \lambda_k, \mu_k) d_k}{(1 - \beta)(\|h(x_k)\|_1 + \sum_{i=1}^m |\max\{0, g_i(x_k)\}|)},$$

unde $\beta \in (0, 1)$. Funcția **fmincon** din Matlab utilizează această metodă pentru găsirea unui punct de minim local al problemei generale (NLP). Ca și în capitolul anterior, putem aproxima Hessiana Lagrangianului $\nabla_x^2 \mathcal{L}(x_k, \lambda_k, \mu_k)$ cu o matrice B_k , unde B_k se poate actualiza folosind updatări cvasi-Newton de rang unu sau doi derivate din ecuația secantei:

$$B_{k+1}(x_{k+1} - x_k) = \nabla_x \mathcal{L}(x_{k+1}, \mu_{k+1}, \lambda_{k+1}) - \nabla_x \mathcal{L}(x_k, \mu_{k+1}, \lambda_{k+1}).$$

Exemplul 13.2.1 *Considerăm următoarea problemă:*

$$\begin{aligned} \min_{x \in \mathbb{R}^2} f(x) &= 6 \frac{x_1}{x_2} + \frac{x_2}{x_1} \\ \text{s.l.: } h(x) &= x_1 x_2 - 2 = 0, \quad g(x) = 1 - x_1 - x_2 \leq 0. \end{aligned}$$

Rezolvăm această problemă cu metoda pătratică secvențială pornind din punctul inițial $x_0 = [2 \ 1]^T$, $\lambda_0 = 0$ și $\mu_0 = 0$. Derivatele necesare pentru construcția subproblemei QP sunt:

$$\begin{aligned} \nabla f(x) &= \begin{bmatrix} \frac{6}{x_2} - \frac{2x_2}{x_1^3} & -\frac{6x_1}{x_2^2} + \frac{1}{x_1^2} \end{bmatrix}^T, \quad \nabla^2 f(x) = \begin{bmatrix} \frac{6x_2}{x_1^4} & -\frac{6}{x_2^2} - \frac{2}{x_1^3} \\ -\frac{6}{x_2^2} - \frac{2}{x_1^3} & \frac{12x_1}{x_2^3} \end{bmatrix} \\ \nabla h(x) &= [x_2 \ x_1]^T, \quad \nabla g(x) = [-1 \ -1]^T. \end{aligned}$$

În punctul x^0 , $f(x^0) = 12.25$, $h(x^0) = 0$ și $g(x^0) = -2 < 0$. De aici rezultă prima subproblemă QP:

$$\begin{aligned} \min_{d \in \mathbb{R}^2} q(d; x_0) &= \begin{pmatrix} \left[\frac{23}{4} \quad -\frac{47}{4} \right] d + \frac{1}{2} d^T \begin{bmatrix} \frac{3}{8} & -\frac{25}{4} \\ -\frac{25}{4} & 24 \end{bmatrix} d \end{pmatrix} \\ \text{s.l.: } [1 \ 2]d &= 0, \quad [-1 \ -1]d - 2 \leq 0. \end{aligned}$$

Eliminând prima constrângere de egalitate, problema se reduce la una unidimensională ce se poate rezolva analitic obținând soluția următoare $d_0 = [0.92079 \ 0.4604]^T$. Mai departe, prima iterație a metodei pătratice secvențiale cu pas unitar este:

$$x_1 = x_0 + d_0 = [1.07921 \ 1.4604]^T.$$

Din moment ce constrângerea de inegalitate este satisfăcută strict în această soluție (adică $[-1 \ -1]d_0 - 2 < 0$), considerăm $\lambda_1 = 0$. Multiplicatorul corespunzător constrângerii de egalitate poate fi calculat

prin rezolvarea condițiilor de optimalitate necesare ale subproblemei. Mai exact $\nabla q(d; x_0) + \mu \nabla h(x_0) = 0$, sau explicit

$$\begin{bmatrix} \frac{23}{4} \\ -\frac{47}{4} \end{bmatrix} + \begin{bmatrix} \frac{3}{8} & -\frac{25}{4} \\ -\frac{25}{4} & 24 \end{bmatrix} d_0 + \mu \begin{bmatrix} 1 \\ 2 \end{bmatrix} = 0,$$

de unde avem $\mu_1 = -2.52723$. În concluzie, prima iterație și valorile aferente sunt date de:

$$x_1 = [1.07921 \ 1.4604]^T, \ f(x_1) = 5.68779, \ h(x_1) = -0.42393, \ g(x_1) < 0.$$

Observăm că funcția a progresat substanțial însă constrângerea de egalitate nu este satisfăcută. A doua subproblemă necesită următoarele derivate:

$$\begin{aligned} \nabla f(x_1) &= [1.78475 \quad -2.17750]^T & \nabla h(x_1) &= [1.4604 \quad 1.07921]^T \\ \nabla^2 f(x_1) &= \begin{bmatrix} 6.45944 & -4.40442 \\ -4.40442 & 4.15790 \end{bmatrix}, & \nabla^2 h(x_1) &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \end{aligned}$$

iar Hessiana termenului pătratic va fi (reamintim că $\lambda_1 = 0$):

$$\nabla_x^2 \mathcal{L} = \nabla^2 f + \mu \nabla^2 h = \begin{bmatrix} 6.45924 & -6.93165 \\ -6.93165 & 4.15790 \end{bmatrix}.$$

În concluzie, a doua subproblemă este dată de:

$$\begin{aligned} \min_{x \in \mathbb{R}^2} q(d; x_1) & \left(= [1.78475 \quad -2.17750] d + \frac{1}{2} d^T \begin{bmatrix} 6.45924 & -6.93165 \\ -6.93165 & 4.15790 \end{bmatrix} d \right) \\ \text{s.l.: } & 1.4604d_1 + 1.07921d_2 = 0.42393, \quad [-1 \quad -1]d - 1.539604 \leq 0, \end{aligned}$$

cu soluția $d_1 = [0.00614 \ 0.38450]^T$ și apoi calculăm $x_2 = x_1 + d_1 = [1.08535 \ 1.8449]^T$. Din nou, având constrângerea de inegalitate strict satisfăcută $[-1 \quad -1]d_1 - 1.539604 < 0$, atunci $\lambda_2 = 0$. Mai mult, μ_2 se obține din condițiile de optimalitate ale subproblemei rezultând în $\mu_2 = 0.5757$. În final avem $f(x_2) = 5.09594$, $g(x_2) < 0$ și $h(x_2) = 2.36 \times 10^{-3}$. Repetând aceiași pași pentru următoarele două iterații obținem:

$$\begin{aligned} \lambda_3 &= 0, \quad \mu_3 = 0.44046, \quad x_3 = [0.99266 \ 2.00463]^T \\ f(x_3) &= 4.99056, \quad g(x_3) < 0, \quad h(x_3) = -1.008 \times 10^{-2} \\ \lambda_4 &= 0, \quad \mu_4 = 0.49997, \quad x_4 = [0.99990 \ 2.00017]^T \\ f(x_4) &= 5.00002, \quad g(x_4) < 0, \quad h(x_4) = -3.23 \times 10^{-5}. \end{aligned}$$

13.3 Metode de penalitate și barieră

Metodele de penalitate și barieră sunt proceduri de aproximare a unei probleme de optimizare constrânsă cu o problemă neconstrânsă. Aproximarea se realizează în cazul metodelor de penalitate prin adăugarea unui termen la funcția obiectiv care atribuie un cost mare violării constrângerilor, în timp ce în metodele de tip barieră se adaugă un termen la funcția obiectiv care favorizează punctele din interiorul mulțimii fezabile față de cele de pe frontiera acestei mulțimi. Problema neconstrânsă se rezolvă apoi cu metode standard (de tip gradient sau Newton) din optimizarea neconstrânsă prezentate în Partea a II-a a lucrării.

13.3.1 Metode de penalitate

Considerăm problema de optimizare:

$$\min_{x \in X} f(x), \quad (13.4)$$

unde f este funcție diferențiabilă și $X \subseteq \mathbb{R}^n$ este mulțimea fezabilă. În general, mulțimea X este descrisă de un set de egalități și inegalități. Definim o funcție $P : \mathbb{R}^n \rightarrow \mathbb{R}$, numită *penalitate* cu următoarele proprietăți:

- (i) P este continuă și ne-negativă, adică $P(x) \geq 0$ pentru orice $x \in \mathbb{R}^n$;
- (ii) $P(x) = 0$ dacă și numai dacă $x \in X$.

În aceste condiții înlocuim problema de optimizare constrânsă (13.4) cu una fără constrângeri de forma:

$$\min_{x \in \mathbb{R}^n} F(x, \tau) \quad (= f(x) + \tau P(x)),$$

în care $\tau > 0$ este un *parametru de penalitate*.

Exemplul 13.3.1 *Presupunem că mulțimea fezabilă $X = \{x \in \mathbb{R}^n : g(x) \leq 0, h(x) = 0\}$. O funcție de penalitate în acest caz este următoarea:*

$$P(x) = \|h(x)\|_q^q + \sum_{i=1}^m |\max\{0, g_i(x)\}|^q, \quad (13.5)$$

unde $q > 0$ și $\|\cdot\|_q$ reprezintă norma vectorială q . Pentru $q = 2$ funcția din (13.5) se numește funcție de penalitate pătratică.

Procedura de bază în metodele de penalitate constă în alegerea unui șir τ_k ce tinde la ∞ astfel încât $\tau_{k+1} > \tau_k \geq 0$ pentru orice k . La fiecare iterație k , rezolvăm problema fără constrângeri folosind metode de ordinul I sau II din optimizarea neconstrânsă:

$$\min_{x \in \mathbb{R}^n} F(x, \tau_k), \quad (13.6)$$

a cărei soluție (punct de minim global) o notăm cu x_k .

Lema 13.3.1 *Fie funcția de penalitate $P(x)$. Dacă definim șirul x_k ca punctul de minim al șirului de probleme (13.6), atunci următoarele relații au loc:*

$$F(x_k, \tau_k) \leq F(x_{k+1}, \tau_{k+1}), \quad P(x_k) \geq P(x_{k+1}) \quad \text{și} \quad f(x_k) \leq f(x_{k+1}).$$

Demonstrație: Deducem ușor următoarele inegalități:

$$\begin{aligned} F(x_{k+1}, \tau_{k+1}) &= f(x_{k+1}) + \tau_{k+1}P(x_{k+1}) \geq f(x_{k+1}) + \tau_k P(x_{k+1}) \\ &\geq f(x_k) + \tau_k P(x_k) = F(x_k, \tau_k), \end{aligned}$$

deci prima relație din lema este demonstrată. De asemenea:

$$\begin{aligned} f(x_k) + \tau_k P(x_k) &\leq f(x_{k+1}) + \tau_k P(x_{k+1}) \\ f(x_{k+1}) + \tau_{k+1} P(x_{k+1}) &\leq f(x_k) + \tau_{k+1} P(x_k). \end{aligned}$$

Adunând aceste două relații obținem:

$$(\tau_{k+1} - \tau_k)P(x_{k+1}) \leq (\tau_{k+1} - \tau_k)P(x_k)$$

ceea ce conduce la a doua relație din lema. În final,

$$f(x_{k+1}) + \tau_k P(x_{k+1}) \geq f(x_k) + \tau_k P(x_k),$$

adică ultima inegalitate din lema. □

Fie x^* un punct de minim local al problemei (13.4). Atunci,

$$f(x^*) = f(x^*) + \tau_k P(x^*) \geq f(x_k) + \tau_k P(x_k) = F(x_k, \tau_k) \geq f(x_k) \quad \forall k \geq 0.$$

Convergența globală a metodelor de penalitate este demonstrată în următoarea teoremă:

Teorema 13.3.1 *Orice punct de acumulare al șirului x_k generat de o metodă de penalitate este un punct de minim global al problemei de optimizare (13.4).*

Demonstrație: Pentru simplitatea expoziției presupunem că întreg șirul x_k converge la \bar{x} . Din continuitatea lui f avem că $f(x_k)$ converge la $f(\bar{x})$. Notăm cu f^* valoarea optimă globală a problemei (13.4). Atunci, din lema precedentă urmează că șirul $F(x_k, \tau_k)$ este nedescrescător și mărginit superior și deci acest șir are limita $\lim_{k \rightarrow \infty} F(x_k, \tau_k) = F^* \leq f^*$. Obținem atunci $\lim_{k \rightarrow \infty} \tau_k P(x_k) = F^* - f(\bar{x})$. Cum $P(x_k) \geq 0$ și $\tau_k \rightarrow \infty$ obținem $\lim_{k \rightarrow \infty} P(x_k) = 0$. Folosind continuitatea lui P obținem $P(\bar{x}) = 0$ și deci \bar{x} este fezabil pentru problema (13.4). Pe de altă parte, $f(\bar{x}) = \lim_{k \rightarrow \infty} f(x_k) \leq f^*$ și deci \bar{x} este minim global pentru problema de optimizare constrânsă (13.4). \square

Exemplul 13.3.2 Considerăm problema de optimizare cu o singură constrângere de egalitate:

$$\min_{x_1+x_2-5=0} (x_1-4)^2 + (x_2-4)^2$$

și folosim funcția de penalitate pătratică pentru egalități (i.e. $q = 2$ în (13.5))

$$F(x, \tau) = (x_1-4)^2 + (x_2-4)^2 + \tau(x_1+x_2-5)^2.$$

Căutăm punctele staționare ale acestei funcții, i.e. $\nabla_x F(x, \tau) = 0$:

$$2(x_1-4) + 2\tau(x_1+x_2-5) = 0$$

$$2(x_2-4) + 2\tau(x_1+x_2-5) = 0$$

cu soluția $x_1 = x_2 = \frac{5\tau+4}{2\tau+1}$. Observăm că pentru $\tau \rightarrow \infty$ obținem $x_1 = x_2 = 2.5$. Pe de altă parte, se observă că $x^* = [2.5 \ 2.5]^T$ este și soluția problemei originale. Punctele staționare ale problemei neconstrânse împreună cu valorile corespunzătoare funcțiilor f, h și F pentru diferite valori ale lui τ sunt date în Tabelul 13.1.

Exemplul 13.3.3 Considerăm acum problema de optimizare cu o singură inegalitate:

$$\min_{x_1+x_2-5 \leq 0} (x_1-4)^2 + (x_2-4)^2$$

și folosim funcția de penalitate pătratică pentru inegalități (i.e. $q = 2$ în (13.5))

$$F(x, \tau) = (x_1-4)^2 + (x_2-4)^2 + \tau \max\{0, x_1+x_2-5\}^2.$$

τ_k	$(x_k)_1 = (x_k)_2$	$f(x_k)$	$h(x_k)$	$F(x_k, \tau_k)$
0	4	0	3	0
0.1	3.75	0.125	2.5	0.75
1	3	2	1	3
10	2.5714	4.0818	0.1428	4.2857
100	2.5075	4.4551	0.015	4.4776
∞	2.5	4.5	0	4.5

Tabelul 13.1: Punctele staționare și valorile funcțiilor f, h și F pentru diferite valori τ .

Căutăm punctele staționare ale acestei funcții, adică $\nabla_x F(x, \tau = 0)$:

$$2(x_1 - 4) + 2\tau \max\{0, x_1 + x_2 - 5\} = 0$$

$$2(x_2 - 4) + 2\tau \max\{0, x_1 + x_2 - 5\} = 0$$

cu soluția $x_1 = x_2$ ca și în exemplul precedent. În concluzie, $(x_1 - 4) + \tau \max\{0, 2x_1 - 5\} = 0$. Avem trei cazuri: $2x_1 - 5$ este zero, pozitiv sau negativ. Presupunem că $2x_1 \geq 5$ și atunci obținem $x_1 = x_2 = \frac{5\tau+4}{2\tau+1}$. Observăm că pentru $\tau \rightarrow \infty$ obținem $x_1 = x_2 = 2.5$. Când τ variază de la 0 la ∞ , punctele staționare ale problemei neconstrânse se mișcă pe segmentul determinat de capetele $[4 \ 4]^T$ (soluția neconstrânsă) și $[2.5 \ 2.5]^T$ (soluția problemei originale). Observăm de asemenea că pentru toate valorile lui $\tau < \infty$ punctele staționare ale problemei neconstrânse de penalitate sunt nefezabile pentru problema originală.

13.3.2 Metode de barieră

În continuare considerăm problema de optimizare de forma:

$$\min_{x \in X} f(x), \quad (13.7)$$

în care f este funcție diferențiabilă și mulțimea fezabilă $X \subseteq \mathbb{R}^n$ are interiorul nevid și este posibil să ajungem la orice punct de pe frontieră din interiorul mulțimii. În general, o astfel de mulțime X este descrisă de un set de inegalități. Definim o funcție $B : \mathbb{R}^n \rightarrow \mathbb{R}$, numită *barieră*, cu următoarele proprietăți:

- (i) B este continuă și ne-negativă, adică $B(x) \geq 0$ pentru orice $x \in \text{dom} B$;

(ii) $B(x) \rightarrow \infty$ dacă x se apropie de frontiera lui X .

Atunci înlocuim problema de optimizare cu constrângeri (13.7) cu una fără constrângeri de forma:

$$\min_{x \in \mathbb{R}^n} F(x, \tau) \quad (= f(x) + \tau B(x)),$$

în care $\tau > 0$ este un *parametru de barieră*.

Exemplul 13.3.4 Presupunem că mulțimea fezabilă este definită de $X = \{x \in \mathbb{R}^n : g(x) \leq 0\}$. Funcții de barieră pot fi în acest caz e.g.:

$$B(x) = -\sum_{i=1}^m \frac{1}{g_i(x)} \quad \text{sau} \quad B(x) = -\sum_{i=1}^m \ln(-g_i(x)).$$

Cea de-a doua funcție se numește *bariera logaritmică* și este cea mai des utilizată în algoritmi de optimizare.

Procedura fundamentală abordată în metodele de barieră este similară celei de la metodele de penalitate: fie τ_k un șir ce tinde la 0 astfel încât $0 < \tau_{k+1} < \tau_k$ pentru orice k . La fiecare iterației k , rezolvăm folosind metode de ordinul I sau II din optimizarea neconstrânsă următoarea problemă fără constrângeri

$$\min_{x \in \mathbb{R}^n} F(x, \tau_k)$$

a cărei soluție o notăm cu x_k . Se observă că metodele de tip barieră prezintă un comportament similar celui de la metodele de penalitate. În particular, orice punct de acumulare al șirului x_k este soluție a problemei (13.7).

Exemplul 13.3.5 Considerăm problema de optimizare cu o singură inegalitate:

$$\min_{x_1+x_2-5 \leq 0} (x_1-4)^2 + (x_2-4)^2$$

și folosim funcția barieră logaritmică:

$$F(x, \tau) = (x_1-4)^2 + (x_2-4)^2 - \tau \log(5 - x_1 - x_2).$$

Căutăm punctele staționare ale acestei funcții, adică $\nabla_x F(x, \tau) = 0$:

$$\begin{aligned} 2(x_1-4) + \tau \frac{1}{5-x_1-x_2} &= 0 \\ 2(x_2-4) + 2\tau \frac{1}{5-x_1-x_2} &= 0 \end{aligned}$$

cu soluția $x_1 = x_2$. Obținem $2x_1^2 - 13x_1 + 20 - \tau/2 = 0$ cu singura rădăcină fezabilă $x_1 = 13/4 - 1/4\sqrt{9 + 4\tau}$. Observăm că pentru $\tau \rightarrow 0$ obținem $x_1 = x_2 = 2.5$. Pe de altă parte, se observă că $x^* = [2.5 \ 2.5]^T$ este și soluția problemei originale. Punctele staționare ale problemei

Tabelul 13.2: Punctele staționare și valorile funcțiilor f, g și F pentru diferite valori τ .

τ_k	$(x_k)_1 = (x_k)_2$	$f(x_k)$	$g(x_k)$	$-\tau \log(-g(x_k))$	$F(x_k, \tau_k)$
100	-1.80	67.41	8.61	-215.31	-147.89
10	1.5	12.5	2	-6.93	5.56
1	2.34	5.45	0.30	1.19	6.64
0.1	2.483	4.59	0.0034	0.34	4.94
0.01	2.498	4.51	0.0034	0.056	4.566
0	2.5	4.5	0	0	4.5

neconstrânse împreună cu valorile corespunzătoare funcțiilor f, h și F pentru diferite valori ale lui τ sunt date în Tabelul 13.2.

13.4 Metode de punct interior

Metodele de punct interior reprezintă o alternativă modernă a metodei mulțimilor active și a altor metode prezentate anterior pentru rezolvarea problemei de optimizare constrânsă (NLP). Metodele anterioare întâmpină probleme deoarece condițiile (KKT) sunt ne-netede (non-smooth), în particular condiția de complementaritate $\lambda_i g_i(x) = 0$ împreună cu cele de fezabilitate $g(x) \leq 0$ și $\lambda \geq 0$ sunt dificil de rezolvat datorită faptului că nu sunt netede. Ideea centrală în jurul căreia s-au dezvoltat metodele de punct interior este înlocuirea condițiilor ne-netede cu un set de condiții netede (ce reprezintă o aproximare a celor originale), și anume: $\lambda_i g_i(x) = \tau$, unde $\tau > 0$ dar arbitrar de mic. Condițiile KKT devin acum o problemă netedă de găsire a rădăcinilor sistemului:

$$\begin{aligned}
 (KKT - IP) : \quad \nabla f(x) + \nabla h(x)^T \mu + \nabla g(x) \lambda &= 0 \\
 \Lambda g(x) &= -\tau e \\
 h(x) &= 0
 \end{aligned}$$

împreună cu constrângerile de inegalitate:

$$g(x) \leq 0, \quad \lambda \geq 0.$$

Am folosit notația $e = [1 \dots 1] \in \mathbb{R}^m$ și $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_m)$. Condițiile (KKT-IP) se numesc *condițiile (KKT) perturbate*. Este clar că pentru $\tau = 0$ obținem condițiile (KKT) pentru problema generală (NLP). Procedura de bază în metodele de punct interior constă în rezolvarea (aproximativă) a sistemului (KKT) perturbat (folosind de exemplu metoda Newton), adică sistemul de ecuații dat în (KKT-IP), cu soluția corespunzătoare $(x(\tau), \lambda(\tau), \mu(\tau))$ și apoi se dorește ca în limita pentru $\tau \rightarrow 0$ aceste soluții să convergă la o soluție (x^*, λ^*, μ^*) a problemei (NLP). Traectoria descrisă de soluția sistemului perturbat $(x(\tau), \lambda(\tau), \mu(\tau))$ se numește *calea centrală (central path)*.

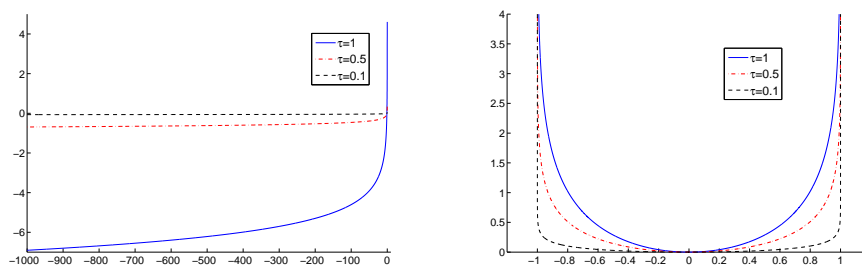


Figura 13.1: Aproximarea funcției indicator pentru mulțimea $(-\infty, 0]$ (stânga) și $[-1, 1]$ (dreapta) cu bariera logaritmică $(-\tau \log(-x))$ și respectiv $-\tau(\log(1+x) + \log(1-x))$ pentru diferite valori ale lui τ .

Metodele de punct interior pot fi interpretate și ca metode de tip barieră. Mai exact, putem reformula problema (NLP) ca o problemă având constrângeri de egalitate, prin mutarea constrângerilor de inegalitate în funcția obiectiv cu ajutorul funcției indicator:

$$\begin{aligned} \min_x \quad & f(x) + \sum_{i=1}^m I_-(g_i(x)) \\ \text{s.l.:} \quad & h(x) = 0, \end{aligned}$$

unde $I_- : \mathbb{R} \rightarrow \mathbb{R}$ este funcția indicator pentru mulțimea numerelor nepozitive $\mathbb{R}_- = (-\infty, 0]$:

$$I_-(y) = \begin{cases} 0, & \text{dacă } y \leq 0 \\ \infty, & \text{dacă } y > 0. \end{cases} \quad (13.8)$$

Prin această reformulare am eliminat constrângerile de inegalitate, însă apare problema majoră a nediferențiabilității noii funcții obiectiv. În

acest scop, funcția indicator I_- se aproximează cu una diferentiabilă folosindu-se o funcție barieră (vezi Fig. 13.1), mai exact considerăm aproximarea:

$$\begin{aligned} \min_x \quad & f(x) - \tau \sum_{i=1}^m \log(-g_i(x)) \\ \text{s.l:} \quad & h(x) = 0. \end{aligned} \quad (13.9)$$

unde τ se numește *parametru de barieră* și am utilizat funcția $B(x) = -\sum_{i=1}^m \log(-g_i(x))$, numită și *bariera logaritmică* pentru problema (NLP). Această aproximare a problemei originale poate fi rezolvată acum prin metoda Lagrange-Newton pentru probleme având constrângeri de egalitate descrisă în capitolul anterior. Pentru a simplifica calculele viitoare, menționăm că gradientul și Hessiana barierei logaritmice sunt date de următoarele expresii:

$$\begin{aligned} \nabla B(x) &= \sum_{i=1}^m \frac{1}{-g_i(x)} \nabla g_i(x) \\ \nabla^2 B(x) &= \sum_{i=1}^m \frac{1}{g_i(x)^2} \nabla g_i(x) \nabla g_i(x)^T - \frac{1}{g_i(x)} \nabla^2 g_i(x). \end{aligned}$$

Condițiile (KKT) pentru această formulare (13.9) sunt:

$$\begin{aligned} \nabla f(x) - \sum_{i=1}^m \frac{\tau}{g_i(x)} \nabla g_i(x) + \nabla h(x)^T \mu &= 0 \\ h(x) &= 0. \end{aligned}$$

Datorită domeniului de definiție al funcției $\log(\cdot)$, avem automat îndeplinită condiția $g(x) < 0$ și dacă notăm cu $\lambda_i = -\frac{\tau}{g_i(x)} > 0$, observăm că aceste condiții conduc la sistemul (KKT-IP). În concluzie, metodele de punct interior rezolvă problema (13.9) (folosind metode de ordinul I sau II pentru probleme cu constrângeri de egalitate) pentru un șir de parametri $\tau_{k+1} < \tau_k$ astfel încât $\tau_k \rightarrow 0$.

13.4.1 Metode de punct interior pentru probleme convexe

Considerăm problema convexă (CP) generală:

$$(CP) : \min_{x \in \mathbb{R}^n} f(x) \\ \text{s.l.: } g(x) \leq 0, Ax = b,$$

unde funcția obiectiv $f : \mathbb{R}^n \rightarrow \mathbb{R}$ și funcția vectorială ce definește constrângerile de inegalitate $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ sunt convexe și de două ori diferențiabile, iar $A \in \mathbb{R}^{p \times n}$ cu rangul $p < n$. Metodele de punct interior rezolvă problema (CP) sau condițiile KKT corespunzătoare prin aplicarea metodei Newton unei secvențe de probleme supuse numai la constrângeri de egalitate, sau asupra unei secvențe de condiții (KKT) perturbate. În acest scop, considerăm aproximarea problemei originale (CP) cu o problemă ce conține doar constrângeri liniare de egalitate:

$$\min_x f(x) - \tau \sum_{i=1}^m \log(-g_i(x)) \quad (13.10) \\ \text{s.l.: } Ax = b.$$

unde am utilizat funcția $B(x) = -\sum_{i=1}^m \log(-g_i(x))$, numită și bariera logaritmică pentru problema (CP). Această aproximare a problemei originale este de asemenea o problemă convexă și poate fi rezolvată prin metoda Newton pentru probleme având constrângeri de egalitate descrisă în capitolul anterior. Un concept esențial în metodele de punct interior este acela de cale centrală: punctele $x(\tau)$ se află pe calea centrală dacă sunt strict fezabile, și anume satisfac:

$$Ax(\tau) = b, \quad g(x(\tau)) < 0,$$

și există un $\hat{\mu} \in \mathbb{R}^p$ astfel încât:

$$\nabla f(x(\tau)) + \tau \nabla B(x(\tau)) + A^T \hat{\mu} = 0.$$

Ca urmare, putem deriva o proprietate importantă a punctelor de pe calea centrală: orice punct de pe această cale produce un punct dual fezabil, și astfel o limită inferioară a lui f^* . În mod specific, considerând

$$\lambda_i(\tau) = -\frac{\tau}{g_i(x(\tau))} \text{ și } \mu(\tau) = \tau \hat{\mu},$$

se poate demonstra ușor că perechea $\lambda(\tau)$ și $\mu(\tau)$ este dual fezabilă. Astfel, funcția duală $q(\lambda(\tau), \mu(\tau))$ este finită iar:

$$\begin{aligned} q(\lambda(\tau), \mu(\tau)) &= f(x(\tau)) + \sum_{i=1}^m \lambda_i(\tau) g_i(x(\tau)) + \mu(\tau)^T (Ax(\tau) - b) \\ &= f(x(\tau)) - m\tau. \end{aligned}$$

În mod particular, diferența de dualitate dintre funcțiile f și q , asociată cu punctul $x(\tau)$ și perechea duală fezabilă $(\lambda(\tau), \mu(\tau))$ este simplu cantitatea $m\tau$. Drept urmare, avem:

$$f(x(\tau)) - f^* \leq m\tau,$$

ceea ce confirmă ideea intuitivă că $x(\tau)$ converge către punctul de optim când $\tau \rightarrow 0$, adică avem o aproximare cât mai bună a problemei (CP). Iterația metodei de punct interior este definită în următorul mod: fie un punct inițial x_0 strict fezabil, $\tau_0 > 0$, $\sigma < 1$ și acuratețea fixată $\epsilon > 0$

Cât timp $m\tau_k \geq \epsilon$ se repetă următorii pași:

1. calculăm soluția $x_{k+1} = x(\tau_k)$ a problemei convexe cu constrângeri de egalitate (13.10) pornind din punctul inițial x_k (*warm start*);
2. descreștem parametrul $\tau_{k+1} = \sigma\tau_k$.

După cum se observă și în practică, pentru această metodă un aspect esențial este selectarea unei actualizări corespunzătoare pentru τ la fiecare pas, în particular este esențial felul cum alegem σ . Un aspect important al metodei de punct interior este strategia de *warm start*: metoda folosită în rezolvarea problemei convexe (13.10) la τ_k pornește din soluția problemei precedente corespunzătoare parametrului τ_{k-1} .

Analiza convergenței metodei de punct interior pentru cazul convex este evidentă. Presupunând că problema perturbată (13.10) se rezolvă cu metoda Lagrange-Newton pentru $\tau = \tau_0, \sigma\tau_0, \sigma^2\tau_0, \dots, \sigma^k\tau_0$, atunci după k pași distanța de la funcția obiectiv la valoarea optimă este mai mică decât $m\tau_0\sigma^k$. Pe de altă parte, pentru anumite clase de probleme convexe (de exemplu, pentru probleme (CP) cu funcția obiectiv auto-concordantă) se poate determina riguros o margine superioară asupra numărului total de iterații Newton necesare pentru rezolvarea problemei, în particular se poate arăta că metoda de punct interior are complexitate polinomială [3, 12].

Exemplul 13.4.1 Fie problema de optimizare convexă:

$$\begin{aligned} \min_{x \in \mathbb{R}^2} & (x_1 - 4)^4 + (x_1 - 6x_2)^2 \\ \text{s.l.: } & x_1^2 + x_2^2 \leq 25, \quad Ax = b \end{aligned}$$

unde $A = \begin{bmatrix} 2 & 3 \end{bmatrix}$ și $b = 12$. Rezolvăm această problemă prin metoda

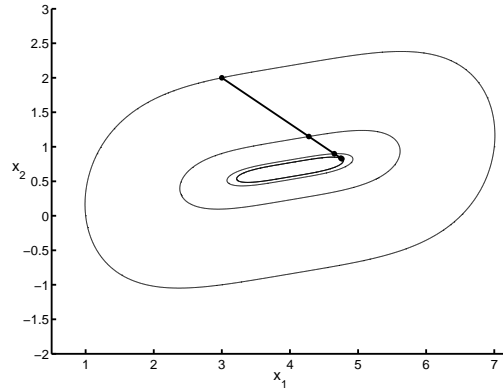


Figura 13.2: Liniile de contur și punctele obținute prin metoda de punct interior.

de punct interior. Pornim dintr-un punct inițial fezabil, exemplu $x_0 = [3 \ 2]^T$. În Fig. 13.2 putem observa convergența foarte rapidă a metodei.

13.4.2 Metode de punct interior pentru probleme neconvexe

În cazul neconvex se preferă următoarea reformulare pentru problema (NLP):

$$\begin{aligned} \min_{x \in \mathbb{R}^n, s \in \mathbb{R}^m} & f(x) \\ \text{s.l.: } & g(x) + s = 0, \quad h(x) = 0, \quad s \geq 0. \end{aligned}$$

Într-o manieră similară cazului convex mutăm constrângerile de inegalitate în cost printr-o funcție barieră de tip logaritm:

$$\begin{aligned} \min_{x, s} & f(x) - \tau \sum_{i=1}^m \log(s_i) \\ \text{s.l.: } & g(x) + s = 0, \quad h(x) = 0. \end{aligned} \tag{13.11}$$

Condițiile (KKT) perturbate (adică condițiile de optimalitate pentru problema (13.11)) au în acest caz următoarea formulare neliniară:

$$\begin{aligned} (KKT - IPs): \quad \nabla f(x) + \nabla h(x)^T \mu + \nabla g(x) \lambda &= 0 \\ \Lambda s - \tau e &= 0 \\ g(x) + s = 0, \quad h(x) &= 0, \end{aligned}$$

împreună cu $s \geq 0$ și $\lambda \geq 0$. Pentru o anumită valoare τ dorim să rezolvăm sistemul neliniar perturbat (KKT-IPs) cu ajutorul metodei Newton. Aplicând metoda Newton sistemului neliniar (KKT-IPs) în variabilele (x, s, λ, μ) obținem:

$$\begin{bmatrix} \nabla_x^2 \mathcal{L} & 0 & \nabla h^T & \nabla g^T \\ 0 & \Lambda & 0 & S \\ \nabla h & 0 & 0 & 0 \\ \nabla g & I & 0 & 0 \end{bmatrix} \begin{bmatrix} d^x \\ d^s \\ d^\mu \\ d^\lambda \end{bmatrix} = - \begin{bmatrix} \nabla f + \nabla h^T \mu + \nabla g^T \lambda \\ \Lambda s - \tau e \\ h \\ g + s \end{bmatrix}, \quad (13.12)$$

unde $S = \text{diag}(s_1, \dots, s_m)$ și $\mathcal{L}(x, s, \lambda, \mu) = f(x) + (g(x) + s)^T \lambda + h(x)^T \mu$. Sistemul precedent se numește *sistemul liniar primal-dual*. Pentru o rezolvare numerică mai eficientă, sistemul primal-dual se aduce într-o formă simetrică. De obicei, în această metodă pentru a defini criteriul de oprire introducem următoarea funcție:

$$\begin{aligned} E(x, s, \lambda, \mu; \tau) \\ = \max\{\|\nabla f(x) + \nabla h^T(x) \mu + \nabla g^T(x) \lambda\|, \|\Lambda s - \tau e\|, \|h(x)\|, \|g(x) + s\|\}. \end{aligned}$$

Metoda Newton pentru rezolvarea sistemului perturbat (KKT-IPs) corespunzător unei valori fixate τ_k are următoarea iterație:

$$\begin{aligned} (MNs): \quad x_{k+1} &= x_k + \alpha_k^x d_k^x, & s_{k+1} &= s_k + \alpha_k^s d_k^s \\ \mu_{k+1} &= \mu_k + \alpha_k^\mu d_k^\mu, & \lambda_{k+1} &= \lambda_k + \alpha_k^\lambda d_k^\lambda, \end{aligned}$$

unde direcțiile $(d_k^x, d_k^s, d_k^\lambda, d_k^\mu)$ sunt soluția sistemului primal-dual (13.12) în $(x_k, s_k, \lambda_k, \mu_k)$. De asemenea, pasul α_k^μ se ia într-un interval de forma $(0, \alpha_{\max}^\mu]$, iar pasul α_k^x se alege pe baza unei funcții merit. De obicei, se consideră următoarea funcție merit:

$$\mathcal{M}_\nu(x, s, \tau) = f(x) - \tau \sum_{i=1}^m \log(s_i) + \nu \|h(x)\|_1 + \nu \sum_{i=1}^m |\max\{0, g_i(x)\}|.$$

Mai exact, pasul α_k^x se alege prin metoda de backtracking pentru funcția merit de mai sus, adică $\alpha_k^x = \rho^{m_k}$ pentru un $\rho \in (0, 1)$ și $c_1 \in (0, 1)$, unde m_k este primul întreg ne-negativ care satisface:

$$\mathcal{M}_\nu(x_k + \rho^m d_k^x, s_k + \rho^m d_k^s, \tau_k) \leq \mathcal{M}_\nu(x_k, s_k, \tau_k) + c_1 \rho^m \mathcal{M}'_\nu(x_k, s_k, \tau_k; d_k^x, d_k^s).$$

Reamintim că $\mathcal{M}'_\nu(x_k, s_k, \tau_k; d_k^x, d_k^s)$ reprezintă derivata direcțională a lui \mathcal{M}_ν în punctul (x_k, s_k, τ_k) de-a lungul direcției (d_k^x, d_k^s) .

Metoda de punct interior constă în acest caz în rezolvarea unui șir de sisteme neliniare perturbate corespunzătoare unui șir de valori τ_k . Mai precis, pentru fiecare τ_k se rezolvă sistemul neliniar (KKT-IPs) aproximativ folosind iterația de mai sus (MNs). Criteriul de oprire folosit pentru rezolvarea aproximativă a sistemului perturbat (KKT-IPs) poate fi $E(x_k, s_k, \lambda_k, \mu_k; \tau_k) \leq \tau_k$. Apoi se actualizează $\tau_{k+1} = \sigma \tau_k$ pentru un $\sigma \in (0, 1)$ și iarăși se rezolvă sistemul perturbat până când criteriul de oprire

$$E(x_k, s_k, \lambda_k, \mu_k; 0) \leq \epsilon$$

este satisfăcut pentru o acuratețe ϵ dorită. O caracteristică importantă a metodei de punct interior este aceea că se bazează pe *warm start*: punctul de pornire în metoda Newton pentru rezolvarea sistemului perturbat (KKT-IPs) corespunzător lui τ_{k+1} coincide cu soluția aproximativă a sistemului perturbat corespunzător lui τ_k .

Pentru cazul neconvex, analiza convergenței metodei de punct interior este mult mai dificilă și rezultatele sunt mai slabe decât pentru cazul convex. Totuși, sub anumite condiții se poate arăta că metoda de punct interior converge la un punct staționar (soluție a sistemului (KKT)) al problemei (NLP) generale și local avem convergență superliniară.

Comentarii finale: Cu metodele de punct interior, încheiem Partea a III-a a acestei lucrări dedicată metodelor numerice de optimizare pentru probleme cu constrângeri (NLP). Metodele de punct interior sunt cele mai eficiente pentru rezolvarea problemelor (NLP) convexe sau neconvexe, fiind implementate în majoritatea pachetelor software de optimizare: CVX, IPOPT, MOSEK, etc. Analiza complexității polinomiale ale metodelor de punct interior pentru cazul convex poate fi găsită în [3, 12]. Mai multe detalii despre metodele prezentate în această parte a lucrării cât și alte metode care nu au fost prezentate aici se pot găsi în cărțile clasice de optimizare neliniară ale lui Bertsekas [2], Luenberger [9], Nesterov [11] și Nocedal și Wright [13]. Dintre lucrările dedicate

implementării numerice a acestor metode de optimizare o amintim de exemplu pe cea a lui Gill, Murray și Wright [7]. O descriere detaliată a pachetelor software existente pe piață este dată de More și Wright în [10].

În ultimul capitol al acestei lucrări prezentăm în detaliu câteva aplicații moderne din inginerie (control optimal, stabilitatea sistemelor, clasificare, învățare automată, ierarhizarea paginilor web) și arătăm că ele pot fi formulate ca probleme de optimizare pe care le vom rezolva cu metodele numerice prezentate aici.

Capitolul 14

Studii de caz din inginerie

În acest capitol final prezentăm câteva studii de caz ce implică optimizarea unor sisteme din domeniul ingineriei. În primul studiu de caz analizăm problema de control optimal al unui sistem dinamic supus constrângerilor, în particular urmărirea unei referințe impuse pentru un robot și o instalație cu patru rezervoare. O alta aplicație importantă din teoria sistemelor este analiza stabilității unui sistem liniar dinamic pe care o vom formula ca o problemă de optimizare. În final, vom analiza problema Google (ierarhizarea paginilor web) și problema învățării automate (sau clasificarea de obiecte). Fiecare studiu de caz ilustrează formulări specifice și strategii de pregătire a modelului matematic de optimizare pentru sistemul respectiv. Acest capitol demonstrează astfel aplicabilitatea metodelor numerice de optimizare prezentate în capitolele anterioare la exemple reale și actuale din inginerie.

14.1 Control optimal liniar

Fie un sistem liniar cu dinamica discretă:

$$z_{t+1} = A_z z_t + B_u u_t$$

unde $z_t \in \mathbb{R}^{n_z}$ reprezintă vectorul de stare al sistemului, $u_t \in \mathbb{R}^{n_u}$ vectorul de intrări al sistemului, iar matricele $A_z \in \mathbb{R}^{n_z \times n_z}$ și $B_u \in \mathbb{R}^{n_z \times n_u}$ descriu dinamica sistemului. Considerăm de asemenea constrângeri de inegalitate liniare pe stare și intrare de forma:

$$lb_z \leq z_t \leq ub_z, \quad C_u u_t \leq d_u \quad \forall t \geq 0,$$

unde $C_u \in \mathbb{R}^{n_i \times n_z}$ și $d_u \in \mathbb{R}^{n_i}$. Formulăm acum problema de control optimal de urmărire a referinței pe un orizont finit N , în care utilizăm funcții de cost pe etapă pătratice:

$$\begin{aligned} \min_{z_t, u_t} & \frac{1}{2} \sum_{t=1}^N \|z_t - z_t^{ref}\|_{Q_t}^2 + \sum_{t=0}^{N-1} \|u_t - u_t^{ref}\|_{R_t}^2 \\ \text{s.l: } & z_0 = z, \quad z_{t+1} = A_z z_t + B_u u_t \\ & lb_z \leq z_t \leq ub_z, \quad C_u u_t \leq d_u \quad \forall t = 0, \dots, N-1, \end{aligned} \quad (14.1)$$

unde presupunem cunoscută starea inițială a sistemului $z_0 = z$ și definim $\|z - z^{ref}\|_Q^2 = (z - z^{ref})^T Q (z - z^{ref})$. Mai mult, presupunem că matricele Q_t și R_t sunt pozitiv definite pentru orice t și z_t^{ref} și respectiv u_t^{ref} reprezintă anumite referințe impuse peste orizontul de predicție pentru starea și intrarea sistemului. Vom arăta mai întâi că această problemă de control optimal pe orizont finit poate fi formulată ca o problemă pătratică convexă de optimizare și apoi aplicăm acest tip de control pe aplicații practice.

14.1.1 Formularea (QP) rară fără eliminarea stărilor

În acest caz, vom defini variabila de decizie $x \in \mathbb{R}^{N(n_z+n_u)}$ care să cuprindă variabilele de stare și intrare peste întreg orizontul de predicție N , i.e.:

$$x = [u_0^T \ z_1^T \ u_1^T \ z_2^T \ \dots \ u_{N-1}^T \ z_N^T]^T.$$

Din problema de control optimal (14.1) vor rezulta constrângeri de egalitate și de inegalitate liniare. Într-adevăr, constrângerile de egalitate vor rezulta din faptul că variabilele de stare și intrare trebuie să respecte dinamica procesului $z_{t+1} = A_z z_t + B_u u_t$ peste întreg orizontul de predicție, adică pentru $t = 0, \dots, N-1$. Luând în calcul forma variabilei x , putem concatena aceste constrângeri într-o constrângere de forma $Ax = b$, unde matricea $A \in \mathbb{R}^{Nn_z \times N(n_z+n_u)}$ și vectorul $b \in \mathbb{R}^{Nn_z}$ vor fi de forma:

$$A = \begin{bmatrix} -B_u & I_{n_z} & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & -A_z & -B_u & I_{n_z} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -A_z & -B_u & I_{n_z} \end{bmatrix}, \quad b = \begin{bmatrix} A_z z_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Fiecare linie de blocuri din sistemul $Ax = b$ reprezintă satisfacerea constrângerilor rezultate din dinamici la un pas k , adică prima linie va asigura $z_1 = A_z z_0 + B_u u_0$ sau rescris $-B_u u_0 + I_{n_z} z_1 = A_z z_0$, a doua linie va asigura $-A_z z_1 - B_u u_1 + I_{n_z} z_2 = 0$, etc. Observăm că matricea A ce descrie constrângerile de egalitate are o structură bloc tridiagonală. În ceea ce privește constrângerile de inegalitate, observăm că avem constrângeri de tip box pe stare și constrângeri poliedrale pe intrare. Dorim să concatenăm toate aceste constrângeri peste întreg orizontul de predicție într-o singură constrângere de forma $Cx \leq d$. Constrângerea de tip box pentru stare poate fi rescrisă ca:

$$\underbrace{\begin{bmatrix} I_{n_z} \\ -I_{n_z} \end{bmatrix}}_{C_z} z_t \leq \underbrace{\begin{bmatrix} ub_z \\ -lb_z \end{bmatrix}}_{d_z}$$

Matricea $C \in \mathbb{R}^{N(2n_z+n_i) \times N(n_z+n_u)}$ și vectorul d vor avea astfel următoarea formă:

$$C = \begin{bmatrix} C_u & 0 & 0 & 0 & 0 \\ 0 & C_z & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & C_u & 0 \\ 0 & 0 & 0 & 0 & C_z \end{bmatrix}, \quad d = \begin{bmatrix} d_u \\ d_z \\ \vdots \\ d_u \\ d_z \end{bmatrix}.$$

Observăm că în acest caz matricea C este bloc diagonală. În privința funcției obiectiv, putem lua un vector care să concateneze referințele pentru stare și intrare peste întreg orizontul de predicție:

$$x^{ref} = \left[(u_0^{ref})^T (z_1^{ref})^T \dots (u_{N-1}^{ref})^T (z_N^{ref})^T \right]^T \in \mathbb{R}^{N(n_z+n_u)}.$$

Cu x și x^{ref} definiți anterior putem rescrie întreaga funcție cost din (14.1) sub forma:

$$\frac{1}{2} \|x - x^{ref}\|_Q^2 = \frac{1}{2} (x - x^{ref})^T Q (x - x^{ref}),$$

în care $Q \in \mathbb{R}^{N(n_z+n_u) \times N(n_z+n_u)}$ va fi bloc diagonală de forma:

$$Q = \begin{bmatrix} R_0 & 0 & \dots & 0 & 0 \\ 0 & Q_1 & \dots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \dots & R_{N-1} & 0 \\ 0 & 0 & \dots & 0 & Q_N \end{bmatrix}. \quad (14.2)$$

Mai mult, matricea Q este pozitiv definită deoarece am presupus că toate matricele Q_t și R_t sunt pozitiv definite. Pentru a aduce în final problema la forma (QP), observăm:

$$\begin{aligned} & \frac{1}{2}(x - x^{ref})^T Q(x - x^{ref}) \\ &= \frac{1}{2}(x^T Qx - x^T Qx^{ref} - (x^{ref})^T Qx + (x^{ref})^T Qx^{ref}) \\ &= \frac{1}{2}x^T Qx - x^T Qx^{ref} + \frac{1}{2}(x^{ref})^T Qx^{ref}. \end{aligned}$$

Dacă luăm $q = -Qx^{ref}$ și ignorăm termenul constant $(x^{ref})^T Qx^{ref}$ din moment ce nu depinde de variabila x , atunci problema (14.1) poate fi rescrisă ca o problemă (QP) convexă având matricele modelului rare (i.e. aceste matrice au foarte multe intrări nule):

$$\begin{aligned} \min_{x \in \mathbb{R}^{N(n_z+n_u)}} \quad & \frac{1}{2}x^T Qx + q^T x \\ \text{s.l.:} \quad & Ax = b, \quad Cx \leq d. \end{aligned} \tag{14.3}$$

14.1.2 Formularea (QP) densă cu eliminarea stărilor

Pentru a elimina stările din problema (14.1), utilizăm dinamica sistemului $z_{t+1} = A_z z_t + B_u u_t$ pentru a exprima stările de-a lungul întregului orizont de predicție N în funcție de starea inițială z_0 și intrările sistemului $(u_0 \dots u_{N-1})$:

$$\begin{aligned} z_1 &= A_z z_0 + B_u u_0 \\ z_2 &= A_z z_1 + B_u u_1 = A_z^2 z_0 + A_z B_u u_0 + B_u u_1 \\ z_3 &= A_z z_2 + B_u u_2 = A_z^3 z_0 + A_z^2 B_u u_0 + A_z B_u u_1 + B_u u_2 \\ &\vdots \\ z_N &= A_z z_{N-1} + B_u u_{N-1} \\ &= A_z^N z_0 + A_z^{N-1} B_u u_0 + A_z^{N-2} B_u u_1 + \dots + A_z B_u u_{N-2} + B_u u_{N-1}. \end{aligned}$$

Dacă eliminăm stările, atunci singurele variabile de decizie rămân intrările. Notăm astfel:

$$x = [u_0^T \dots u_{N-1}^T]^T \in \mathbb{R}^{Nn_u}$$

și de asemenea introducem notația $\bar{z} = [z_1^T \dots z_N^T]^T \in \mathbb{R}^{Nn_z}$. Ecuațiile anterioare pot fi scrise sub forma $\bar{z} = \bar{A}Bx + A_p z_0$, unde matricele $\bar{A}B \in \mathbb{R}^{Nn_z \times Nn_u}$ și $A_p \in \mathbb{R}^{Nn_z \times n_z}$ sunt definite astfel:

$$\bar{A}B = \begin{bmatrix} B_u & 0 & 0 & 0 & \dots & 0 \\ A_z B_u & B_u & 0 & 0 & \dots & 0 \\ A_z^2 B_u & A_z B_u & B_u & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_z^{N-1} B_u & A_z^{N-2} B_u & A_z^{N-3} B_u & A_z^{N-4} B_u & \dots & B_u \end{bmatrix}, \quad A_p = \begin{bmatrix} A_z \\ A_z^2 \\ A_z^3 \\ \vdots \\ A_z^N \end{bmatrix}.$$

Dacă rescriem constrângerile de tip box pentru stare sub forma $C_z z_t \leq d_z$, similar cazului în care nu eliminăm stările, și le concatenăm peste întreg orizontul de predicție astfel încât să avem $\bar{C}_z \bar{z} \leq \bar{d}_z$, în care $\bar{C}_z = \text{diag}(C_z, C_z, \dots, C_z)$ și $\bar{d}_z = [d_z^T \ d_z^T \ \dots \ d_z^T]^T$, atunci constrângerile de inegalitate pentru stare se transformă în constrângeri poliedrale pentru intrare de forma $C'_x x \leq d'_x$, în care $C'_x \in \mathbb{R}^{2Nn_z \times Nn_u}$ și $d'_x \in \mathbb{R}^{2Nn_z}$ sunt date de expresiile:

$$\bar{C}_z \bar{z} \leq \bar{d}_z \iff \bar{C}_z (\bar{A}Bx + A_p z_0) \leq \bar{d}_z \iff \underbrace{\bar{C}_z \bar{A}B}_{=C'_x} x \leq \underbrace{\bar{d}_z - \bar{C}_z A_p z_0}_{=d'_x}.$$

Pentru constrângerile de inegalitate pentru intrare luăm $C''_x x \leq d''_x$, în care definim matricea $C''_x = \text{diag}(C_u, C_u, \dots, C_u) \in \mathbb{R}^{Nn_i \times Nn_u}$ și vectorul $d''_x = [d_u^T \ d_u^T \ \dots \ d_u^T]^T \in \mathbb{R}^{Nn_i}$. Concatenăm acum cele două constrângeri pe stare și intrare peste orizontul de predicție într-una singură de forma $Cx \leq d$ în care:

$$C = \begin{bmatrix} C'_x \\ C''_x \end{bmatrix}, \quad d = \begin{bmatrix} d'_x \\ d''_x \end{bmatrix}.$$

Observăm că matricea $C \in \mathbb{R}^{N(2n_z+n_i) \times Nn_u}$ este bloc inferior triunghiulară deoarece matricea $\bar{A}B$ este bloc inferior triunghiulară și matricea \bar{C}_z este bloc diagonală.

Privind funcția obiectiv, observăm că funcțiile de cost corespunzătoare stărilor sistemului pot fi rescrise sub forma:

$$\begin{aligned} \frac{1}{2} \sum_{t=1}^N \|z_t - z_t^{ref}\|_{Q_t}^2 &= \frac{1}{2} \|\bar{z} - \bar{z}^{ref}\|_{\bar{Q}}^2 = \frac{1}{2} \|\bar{A}Bx + A_p z_0 - \bar{z}^{ref}\|_{\bar{Q}}^2 \\ &= \frac{1}{2} x^T \bar{A}B^T \bar{Q} \bar{A}B x + (z_0^T A_p^T \bar{Q} \bar{A}B - (\bar{z}^{ref})^T \bar{Q} \bar{A}B) x, \end{aligned}$$

în care am folosit notațiile $\bar{Q} = \text{diag}(Q_1, \dots, Q_N) \in \mathbb{R}^{Nn_z \times Nn_z}$ și $\bar{z}^{ref} = [(z_1^{ref})^T \dots (z_N^{ref})^T]^T \in \mathbb{R}^{Nn_z}$ și am neglijat termenii constanți. Se observă că \bar{Q} este matrice pozitiv definită deoarece am presupus că matricele Q_t sunt pozitiv definite peste orizontul de predicție. Pe de altă parte funcțiile de cost corespunzătoare intrărilor sistemului pot fi rescrise astfel:

$$\begin{aligned} \sum_{t=0}^{N-1} \|u_t - u_t^{ref}\|_{R_t}^2 &= \|x - \bar{x}^{ref}\|_{\bar{R}}^2 \\ &= \frac{1}{2}x^T \bar{R}x - (\bar{x}^{ref})^T \bar{R}x + \frac{1}{2}(\bar{x}^{ref})^T \bar{R}\bar{x}^{ref}, \end{aligned}$$

unde am notat cu $\bar{R} = \text{diag}(R_0, \dots, R_{N-1}) \in \mathbb{R}^{Nn_u \times Nn_u}$ și definim vectorul $\bar{x}^{ref} = [(u_0^{ref})^T \dots (u_{N-1}^{ref})^T]^T \in \mathbb{R}^{Nn_u}$. Matricea \bar{R} este pozitiv definită deoarece am presupus că matricele R_t sunt pozitiv definite peste orizontul de predicție. Ignorând termenii constanți, funcția obiectiv devine pătratică convexă:

$$\frac{1}{2}x^T Qx + q^T x$$

unde matricea $Q \in \mathbb{R}^{Nn_u \times Nn_u}$ în acest caz este dată de expresia:

$$Q = \bar{R} + \bar{A}B^T \bar{Q} \bar{A}B.$$

Se observă imediat că matricea Q este pozitiv definită, deoarece matricele \bar{Q} și \bar{R} sunt pozitiv definite, însă are o structură densă datorită termenului $\bar{A}B^T \bar{Q} \bar{A}B$, unde reamintim că $\bar{A}B$ este bloc inferior triunghiulară. Mai mult, vectorul q are următoarea expresie:

$$q = \bar{A}B^T \bar{Q} A_p z_0 - \bar{A}B^T \bar{Q} \bar{z}^{ref} - \bar{R} \bar{x}^{ref}.$$

În final, obținem următoarea problemă pătratică convexă având numai constrângeri de inegalitate:

$$\begin{aligned} \min_{x \in \mathbb{R}^{Nn_u}} \quad & \frac{1}{2}x^T Qx + q^T x \\ \text{s.l.:} \quad & Cx \leq d. \end{aligned} \tag{14.4}$$

Această problemă de optimizare (14.4) are matricele Q și C dense, în particular, matricea C este inferior bloc triunghiulară, iar matricea Q este complet densă. Pe de altă parte, matricele ce descriu problema (QP) din

(14.3) ce se obține când nu se elimină stările au o structură foarte rară, în particular matricea Hessiană corespunzătoare funcției obiectiv este bloc diagonală. Totuși, dimensiunea problemei (QP) din (14.3) este $N(n_z + n_u)$ mult mai mare decât dimensiunea Nn_u a problemei (QP) din (14.4). Deși ambele formulări sunt folosite în aplicații, în anumite situații (e.g. pentru orizont de predicție N mare sau sistem dinamic de dimensiune (n_z, n_u) mare) se preferă formularea rară (14.3) datorită avantajelor pe care acest model de optimizare îl oferă în calculele numerice.

14.1.3 Control optimal pentru urmărirea traiectoriei cu un robot E-Puck

O aplicație des întâlnită, simplă și favorabilă pentru testarea algoritmilor de optimizare și control este robotul E-Puck (vezi Fig. 14.1). Acest sistem robotic reprezintă un ansamblu electronic mobil ce suportă implementarea numerică și experimentarea cu algoritmi de optimizare de complexitate relativ ridicată. Mai exact, structura mecanică a robotului E-Puck este susținută de două motoare pas-cu-pas atașate ambelor roți, iar cea electronică este definită de următoarele componente: microcontroler dsPIC30 (16-bit), dispozitiv de comunicație Bluetooth (folosit în simularea sistemelor de tip rețea), senzori infraroșu, cameră video CMOS (rezoluție 640×480), senzor ultrasunete, accelerometru 3D, etc.

Un exemplu de test simplu și eficient al unei metode de optimizare cunoscute este problema *Rendez-Vous*: pentru o repartizare inițială a unui colectiv de roboți pe o suprafață plană, să se determine traiectoria optimă a fiecărui robot până la un punct comun de întâlnire. Această problemă se formulează ușor în termeni de optimizare, și poate fi definită ca un suport de test pentru diferiți algoritmi numerici de optimizare. Chiar și pentru cele mai simple probleme ce implică sisteme multi-robot, modelul matematic al unui sistem robotic este crucial în proiectarea de algoritmi numerici.

În acest subcapitol considerăm un model simplificat al robotului E-Puck, și anume cel restricționat doar la deplasarea înainte (fără a considera posibilitatea de deplasare înapoi). Modelul dinamic simplificat este



Figura 14.1: *Robot e-Puck.*

liniar, continuu și este definit de următoarele ecuații:

$$\begin{aligned}\dot{y} &= \frac{ru_1}{2} + \frac{ru_2}{2} \\ \dot{\theta} &= \frac{ru_1}{2l} - \frac{ru_2}{2l},\end{aligned}$$

unde y reprezintă distanța parcursă în direcția înainte, θ unghiul de viraj, r raza roților, l distanța de la roată la centrul de greutate al robotului, iar u_1 și u_2 reprezintă viteza unghiulară a primei roți și respectiv, a celei de-a doua roți. Pentru a respecta consistența notațiilor, notăm starea sistemului cu $z = \begin{bmatrix} y \\ \theta \end{bmatrix}$ și intrarea cu $u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$. Mai departe, rescrierea modelului precedent va avea următoarea formă:

$$\dot{z} = \bar{A}_z z + \bar{B}_u u,$$

în care $\bar{A}_z = 0 \in \mathbb{R}^{2 \times 2}$ și $\bar{B}_u = \begin{bmatrix} \frac{r}{2} & \frac{r}{2} \\ \frac{r}{2l} & -\frac{r}{2l} \end{bmatrix}$. Pentru a putea să efectuăm experimente numerice, avem nevoie de discretizarea sistemului liniar continuu anterior. Una dintre metodele cele mai vechi și mai simple este *metoda Euler* de discretizare, ce presupune aproximarea derivatei unei funcții diferențiabile $f(t)$ cu următoarea expresie:

$$\frac{df}{dt}(t) \approx \frac{f(t + \Delta t) - f(t)}{\Delta t},$$

în care intervalul Δt se determină în funcție de viteza de evoluție a procesului. Obținem aproximarea discretă a modelului pentru robot dată de următoarea relație de recurență:

$$z_{t+1} = (I_2 - \Delta t \bar{A}_z) z_t + \Delta t \bar{B}_u u_t,$$

în care I_2 este matricea identitate de ordin 2. Alegând $\Delta t = 0.5$ s, obținem sistemul dinamic și matricele sistemului de forma:

$$z_{t+1} = A_z z_t + B_u u_u, \quad \text{unde } A_z = I_2 - \frac{1}{2} \bar{A}_z, \quad B_u = \frac{1}{2} \bar{B}_u.$$

Un exemplu simplu de problemă de control optimal poate fi definit de urmărirea unei traiectorii sinusoidale pe o suprafață plană de către robotul E-Puck. În acest caz, definim traiectoria discretă $z_t^{\text{ref}} = (y_t^{\text{ref}}, \theta_t^{\text{ref}})^T$ ce se dorește a fi urmărită. Caracteristica discretă a referinței impune eșantionarea funcției sinus continue cu o anumită perioadă T (în simulări am considerat $T = 0.1$). Acuratețea cu care robotul urmărește o curbă sinusoidală variază în funcție de perioada de eșantionare a funcției sinus, orizontul de predicție considerat și de constrângerile aplicate problemei de control optimal.

În cel mai simplu caz, considerăm că robotul pleacă din origine și dorim urmărirea unui șir de puncte $(x_t, \sin x_t)$, unde $x_{t+1} - x_t = T$. În acest caz, nu putem considera că referința este definită de șirul propriu-zis de puncte, deoarece mărimile x_t și $\sin x_t$ diferă de mărimile stării sistemului date de distanța parcursă y_t^{ref} și unghiul de orientare θ_t^{ref} . Pentru a realiza conversia mărimilor facem următoarele observații:

- observăm că orice punct de pe graficul funcției $\sin x$ se află la un unghi $\theta = \arctan \cos x$ față de orizontală;
- distanța dintre doua puncte din șirul definit anterior este dată de $y = \sqrt{(x_{t+1} - x_t)^2 + (\sin x_{t+1} - \sin x_t)^2}$.

În concluzie, putem realiza conversia șirului $(x_t, \sin x_t)$ și obținem referința:

$$z_t^{\text{ref}} = (y_t^{\text{ref}}, \theta_t^{\text{ref}}) = \left(\sqrt{(x_{t+1} - x_t)^2 + (\sin x_{t+1} - \sin x_t)^2}, \arctan \cos x_t \right)$$

și $u_t^{\text{ref}} = 0$. Alegând orizontul de predicție $N = 2$, problema de control optimal ce rezultă din urmărirea traiectoriei sinusoidale se poate enunța astfel:

$$\begin{aligned} \min_{z_t, u_t} \quad & \frac{1}{2} [(z_1 - z_1^{\text{ref}})^T Q_1 (z_1 - z_1^{\text{ref}}) + (z_2 - z_2^{\text{ref}})^T Q_2 (z_2 - z_2^{\text{ref}}) + u_0^T R_0 u_0 + u_1^T R_1 u_1] \\ \text{s.l.:} \quad & z_0 = z, \quad z_1 = A_z z_0 + B_u u_0, \quad z_2 = A_z z_1 + B_u u_1, \\ & u_{\min} \leq u_0 \leq u_{\max}, \quad u_{\min} \leq u_1 \leq u_{\max}, \end{aligned}$$

în care considerăm $Q_1 = Q_2 = I_2$ și $R_0 = R_1 = 0.1I_2$. Mai mult, considerăm $r = 2$ cm, $l = 1$ cm, $u_{\max} = [1 \ 1]^T$ și $u_{\min} = [-1 \ -1]^T$. Aplicăm metoda de punct interior pentru rezolvarea acestei probleme (QP). Folosim de asemenea procedura de control bazată pe *orizontul alunecător* (i.e. la fiecare pas se măsoară/estimează starea sistemului și se rezolvă problema de control optimal cu orizont finit de mai înainte; se obține o secvență de N intrări optimale, dar se aplică doar primele $N_c \leq N$ intrări din această secvență, după care procedura se repetă). Metoda de control bazată pe principiul orizontului alunecător se numește *control predictiv* (MPC - *Model Predictive Control*). Rescriem compact problema de control optimal pentru starea inițială $z_0 = z$ sub forma:

$$\begin{aligned} \min_{x \in \mathbb{R}^8} \quad & \frac{1}{2} x^T Q x + q^T x \\ \text{s.l.:} \quad & Ax = b, \quad Cx \leq d, \end{aligned} \quad (14.5)$$

unde matricele și vectorii corespunzători problemei sunt date de:

$$\begin{aligned} x &= \begin{bmatrix} u_0 \\ z_1 \\ u_1 \\ z_2 \end{bmatrix}, \quad Q = \begin{bmatrix} R_0 & 0 & 0 & 0 \\ 0 & Q_1 & 0 & 0 \\ 0 & 0 & R_1 & 0 \\ 0 & 0 & 0 & Q_2 \end{bmatrix}, \quad A = \begin{bmatrix} -B_u & I_2 & 0 & 0 \\ 0 & -A_z & -B_u & I_2 \end{bmatrix}, \\ q &= \begin{bmatrix} 0 \\ -Q_1 z_1^{ref} \\ 0 \\ -Q_2 z_2^{ref} \end{bmatrix}, \quad b = \begin{bmatrix} A_z z_0 \\ 0 \end{bmatrix}, \quad C = \begin{bmatrix} I_2 & 0 & 0 & 0 \\ -I_2 & 0 & 0 & 0 \\ 0 & 0 & I_2 & 0 \\ 0 & 0 & -I_2 & 0 \end{bmatrix}, \quad d = \begin{bmatrix} u_{\max} \\ -u_{\min} \\ u_{\max} \\ -u_{\min} \end{bmatrix}. \end{aligned}$$

Reamintim că primul pas din metoda de punct interior pentru o problemă convexă presupune transformarea echivalentă a problemei (14.5) într-una fără inegalități:

$$\begin{aligned} \min_x \quad & \frac{1}{2} x^T Q x + q^T x - \tau \sum_{i=1}^8 \log(d_i - C_i x) \\ \text{s.l.:} \quad & Ax = b, \end{aligned} \quad (14.6)$$

unde C_i reprezintă linia i a matricei C . Apoi, aplicăm algoritmul propriu-zis:

1. se dau un punct inițial x strict fezabil, $\tau > 0$, $\sigma < 1$, toleranța $\epsilon > 0$ și parametrul m numărul de inegalități;

2. cât timp $m\tau \geq \epsilon$ repetă:

- (i) calculează $x(\tau)$ ca soluție a problemei (14.6) pornind din x ;
- (ii) actualizează $x = x(\tau)$ și $\tau = \sigma\tau$.

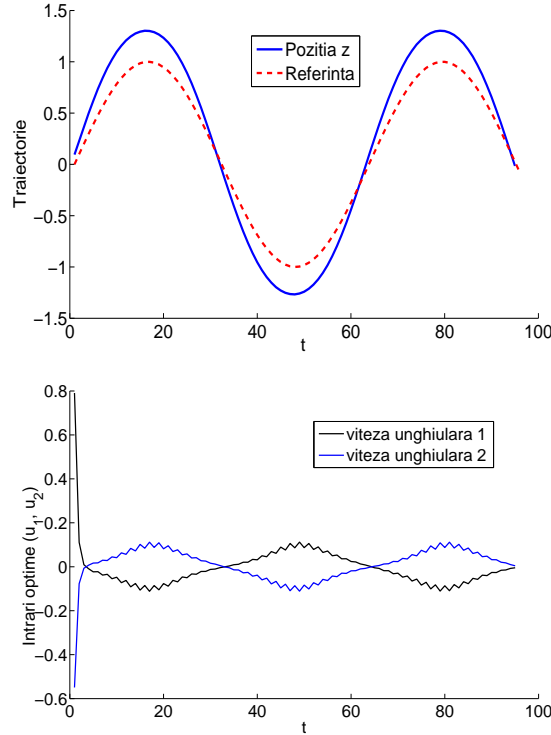


Figura 14.2: Traectoria robotului folosind controlul optimal cu orizont alunecător (MPC): evoluția stărilor sistemului și intrărilor optimale.

Rezultatele obținute pe baza strategiei de control predictiv, unde la fiecare pas problema de control optimal se rezolvă cu metoda de punct interior pentru probleme convexe, sunt prezentate în Fig. 14.2. Se observă o urmărire bună a traiectoriei impuse robotului. În cea de-a doua figură reprezentăm traiectoria optimă a intrărilor peste orizontul de simulare.

14.1.4 Control optimal pentru pendulul invers

Considerăm problema de control optimal al aducerii pendulului invers în poziție verticală și menținerea lui în această stare. Pendulul invers este

compus dintr-un cărucior de masă M care alunecă unidimensional de-a lungul axei Ox pe o suprafață orizontală, și un pendul format dintr-o bilă de masă m aflată la capătul unei tije de lungime l , considerată imponderabilă (vezi Fig. 14.3). Vectorul de stare al sistemului este $z \in \mathbb{R}^4$, unde $z_1 = \theta$ este unghiul făcut de tijă cu verticala, $z_2 = \dot{\theta}$ este viteza unghiulară, z_3 reprezintă poziția pendulului (căruciorului) pe axa Ox , iar z_4 este viteza sa.

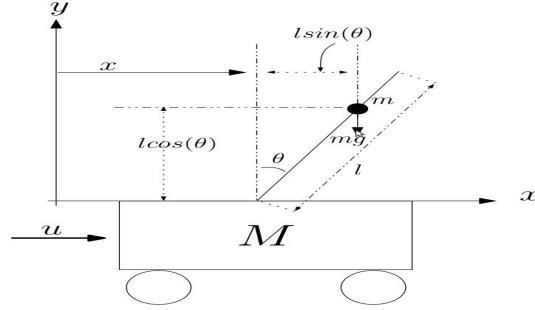


Figura 14.3: Pendulul invers.

Dinamica liniară discretă a sistemului este exprimată prin $z_{t+1} = A_z z_t + B_u u_t$, unde $u_t \in \mathbb{R}$ este intrarea sistemului, reprezentând o corecție de deplasare orizontală aplicată căruciorului. Matricele dinamicilor în acest caz sunt:

$$A_z = \begin{bmatrix} 1.0259 & 0.504 & 0 & 0 \\ 1.0389 & 1.0259 & 0 & 0 \\ -0.0006 & 0 & 1 & 0.05 \\ -0.0247 & -0.0006 & 0 & 1 \end{bmatrix} \quad \text{și} \quad B_u = \begin{bmatrix} -0.0013 \\ -0.0504 \\ 0.0006 \\ 0.025 \end{bmatrix}.$$

La pendulul invers, obiectivul de control este să menținem tija suficient de aproape de verticală, anume să menținem starea $z_1 = \theta$ într-un interval admisibil centrat în 0° , adică $\theta_{min} \leq z_1 \leq \theta_{max}$, ce reprezintă constrângeri pe stare ale sistemului, unde $\theta_{max} = -\theta_{min} = 10^\circ$. Astfel, formulăm problema de control optimal pe un orizont finit N , cu scopul menținerii tije în poziție verticală (în acest caz $z_t^{ref} = 0$ și $u_t^{ref} = 0$ pentru orice $t \geq 0$):

$$\min_{z_t, u_t} \sum_{k=1}^N \frac{1}{2} z_k^T Q_0 z_k + \sum_{t=0}^{N-1} \frac{1}{2} R_0 u_t^2 \quad (14.7)$$

$$\text{s.l.: } z_0 = z, \quad z_{t+1} = A_z z_t + B_u u_t$$

$$\theta_{min} \leq (z_t)_1 \leq \theta_{max} \quad \forall t = 0, \dots, N-1,$$

unde z este starea inițială a pendulului, iar matricele din costurile de etapă sunt:

$$Q_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0.01 \end{bmatrix} \quad \text{și} \quad R_0 = 10.$$

După cum am arătat, această problemă (14.7) se formulează ca o problemă (QP) convexă. Considerăm formularea (QP) rară fără eliminarea stărilor din (14.3) și rezolvăm cu algoritmul de punct interior. Observăm că nu avem constrângeri pe intrare. Considerând că $z \in \mathbb{R}^4$, constrângerea că prima componentă a vectorului de stare z_t se află în intervalul corespunzător, anume $\theta_{\min} \leq (z_t)_1 \leq \theta_{\max}$, poate fi rescrisă matriceal sub forma:

$$C_z z_t \leq d_z, \quad \text{unde} \quad C_z = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix} \quad \text{și} \quad d_z = \begin{bmatrix} \theta_{\max} \\ -\theta_{\min} \end{bmatrix}.$$

Pentru întreg orizontul de predicție vom avea $Cx \leq d$ unde C și d vor fi de forma:

$$C = \begin{bmatrix} 0 & C_z & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & C_z & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & C_z \end{bmatrix}, \quad d = \begin{bmatrix} d_z \\ d_z \\ \vdots \\ d_z \end{bmatrix}$$

Deoarece referința este 0, atunci funcția obiectiv a problemei (QP) va fi de forma $f(x) = x^T Q x$, adică în acest caz $q = 0$, unde Q este bloc diagonală, formată din matricele ce definesc costurile pe etapă pentru stare și intrare:

$$Q = \text{diag}(R_0, Q_0, \dots, R_0, Q_0).$$

Problema QP finală va fi de forma:

$$\begin{aligned} \min_x \quad & \frac{1}{2} x^T Q x \\ \text{s.l.:} \quad & Ax = b, \quad Cx \leq d. \end{aligned}$$

Problema este rezolvată prin metoda de punct interior pentru probleme convexe și traiectoria unghiului $(z_t)_1 = \theta_t$ pentru $t = 0, 1, \dots, N$, unde $N = 25$, este prezentată în Fig. 14.4. Se observă că la pasul $t = 2$ constrângerea pe unghi este activă și după pasul $t = 15$ unghiul format de tijă cu axa verticală devine $\theta = 0$. În concluzie, strategia de control optimal este capabilă să stabilizeze pendulul și în același timp să satisfacă constrângerile impuse pendulului.

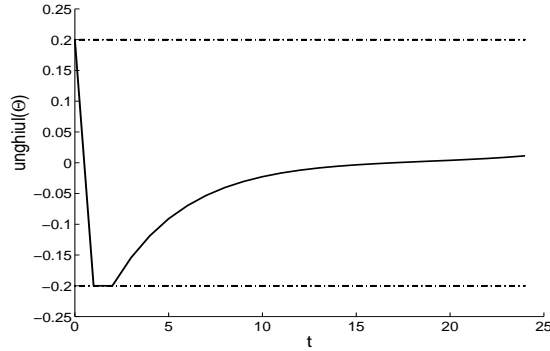


Figura 14.4: Traiectoria unghiului θ pentru un orizont de predicție $N = 25$.

14.2 Control optimal neliniar

Considerăm un sistem dinamic discret cu dinamici neliniare:

$$z_{t+1} = \phi(z_t, u_t) \quad (14.8)$$

unde $u_t \in \mathbb{R}^{n_u}$ este intrarea și $z_t \in \mathbb{R}^{n_z}$ starea sistemului la pasul t , iar $\phi : \mathbb{R}^{n_z} \rightarrow \mathbb{R}^{n_z}$. Pentru simplitate nu presupunem constrângeri pe stare și intrare, deși astfel de constrângeri pot fi încorporate ușor în problema de control optimal într-o manieră similară celei prezentate la cazul sistemelor liniare. Problema de control optimal neliniar peste un orizont de predicție N se definește astfel:

$$\min_{z_t, u_t} \sum_{t=1}^N \ell_t^z(z_t) + \sum_{t=0}^{N-1} \ell_t^u(u_t) \quad (14.9)$$

$$\text{s.l.: } z_0 = z, \quad z_{t+1} - \phi(z_t, u_t) = 0 \quad \forall t = 0, \dots, N-1, \quad (14.10)$$

unde $\ell_t^z : \mathbb{R}^{n_z} \rightarrow \mathbb{R}$ și $\ell_t^u : \mathbb{R}^{n_u} \rightarrow \mathbb{R}$ sunt costurile pe etapa t pentru stări și intrări.

14.2.1 Formularea (NLP) rară și densă

Ca și în cazul liniar, în formularea rară fără eliminarea stărilor definim variabila de decizie:

$$x = [u_0^T \ z_1^T \ u_1^T \ z_2^T \ \dots \ u_{N-1}^T \ z_N^T]^T \in \mathbb{R}^{N(n_z+n_u)}.$$

Problema de control optimal neliniar se poate aduce la o problemă de optimizare neconvexă de forma:

$$\begin{aligned} \min_{x \in \mathbb{R}^{N(n_z+n_u)}} \quad & f(x) \\ \text{s.l:} \quad & h(x) = 0, \end{aligned} \quad (14.11)$$

în care funcția obiectiv este $f(x) = \sum_{t=1}^N \ell_t^z(z_t) + \sum_{t=0}^{N-1} \ell_t^u(u_t)$, iar funcția $h : \mathbb{R}^{N(n_z+n_u)} \rightarrow \mathbb{R}^{Nn_z}$ ce descrie constrângerile de egalitate este dată de expresia:

$$h(x) = \begin{bmatrix} z_1 - \phi(z_0, u_0) \\ z_2 - \phi(z_1, u_1) \\ \vdots \\ z_N - \phi(z_{N-1}, u_{N-1}) \end{bmatrix}.$$

Funcția Lagrange pentru multiplicatorii $\mu = [\mu_1^T \dots \mu_N^T]^T$ are forma:

$$\begin{aligned} \mathcal{L}(x, \mu) &= f(x) + \mu^T h(x) \\ &= \sum_{t=1}^N \ell_t^z(z_t) + \sum_{t=0}^{N-1} \ell_t^u(u_t) + \sum_{t=0}^{N-1} \mu_{t+1}^T (z_{t+1} - \phi(z_t, u_t)). \end{aligned}$$

Condițiile KKT ale problemei sunt:

$$\nabla_x \mathcal{L}(x, \mu) = 0, \quad h(x) = 0.$$

Mai detaliat, derivata lui \mathcal{L} în funcție de z_t sau u_t are o structură specială. De exemplu, pentru $t = 1, \dots, N-1$ obținem următoarele expresii:

$$\begin{aligned} \nabla_{z_t} \mathcal{L}(x, \mu) &= \nabla \ell_t^z(z_t) + \mu_t - \frac{\partial \phi}{\partial z_t}(z_t, u_t)^T \mu_{t+1} = 0 \\ \nabla_{u_t} \mathcal{L}(x, \mu) &= \nabla \ell_t^u(u_t) - \frac{\partial \phi}{\partial u_t}(z_t, u_t)^T \mu_{t+1} = 0. \end{aligned}$$

Sistemul Lagrange poate fi rezolvat cu metodele prezentate în capitolele anterioare, e.g. metoda Lagrange-Newton. În această metodă iterația are forma:

$$x_{k+1} = x_k + d_k, \quad \mu_{k+1} = \mu_k^{QP}$$

în care direcția d_k este soluția optimă a problemei (QP) din (14.12) cu multiplicatorul Lagrange optim asociat constrângerilor μ_k^{QP} :

$$\begin{aligned} \min_d \quad & \nabla f(x_k)^T d + \frac{1}{2} d^T B_k d \\ \text{s.l.:} \quad & h(x_k) + \nabla h(x_k) d = 0, \end{aligned} \quad (14.12)$$

unde $B_k = \nabla_x^2 \mathcal{L}(x_k, \mu_k)$. Această problemă pătratică (14.12) are o structură rară unde matricea B_k este bloc diagonală, iar matricea ce definește egalitățile (Jacobianul $\nabla h(x_k)$) este tridiagonală (i.e. structura acestei probleme (QP) din (14.12) este similară cu problema (QP) rară (14.3) corespunzătoare cazului liniar).

Altă abordare ar fi să eliminăm stările într-o manieră similară cazului liniar. În acest caz, obținem o problemă de optimizare fără constrângeri. Ideea de bază constă în a păstra doar z_0 și a defini variabila de decizie $x = [u_0^T, \dots, u_{N-1}^T]^T \in \mathbb{R}^{Nn_u}$. Stările z_1, \dots, z_N sunt eliminate în mod recursiv prin relațiile:

$$\begin{aligned} \psi_0(z_0, x) &= z \\ \psi_{t+1}(z_0, x) &= \phi(\psi_t(z_0, x), u_t). \end{aligned}$$

Astfel, problema de control optimal este echivalentă cu o problemă fără constrângeri și cu mai puține variabile:

$$\min_{x \in \mathbb{R}^{Nn_u}} \sum_{t=1}^N \ell_t^z(\psi_t(z_0, x)) + \sum_{t=0}^{N-1} \ell_t^u(u_t).$$

Această problemă este numită problema de control optimal redusă. Poate fi rezolvată eficient prin metode de tip gradient sau Newton pentru cazul neconstrâns prezentate în Partea a II-a a lucrării.

14.2.2 Control optimal aplicat unei instalații cu patru rezervoare

În acest subcapitol, considerăm o *instalație cu patru rezervoare interconectate* prezentată în Fig. 14.5, dispusă cu două pompe de împingere a apei. Modelul matematic corespunzător instalației este

neliniar, dat de următoarele ecuații diferențiale:

$$\begin{aligned}\frac{dh_1}{dt} &= -\frac{a_1}{S}\sqrt{2gh_1} + \frac{a_4}{S}\sqrt{2gh_4} + \frac{\gamma_a}{S}q_a, \\ \frac{dh_2}{dt} &= -\frac{a_2}{S}\sqrt{2gh_2} + \frac{a_3}{S}\sqrt{2gh_3} + \frac{\gamma_b}{S}q_b, \\ \frac{dh_3}{dt} &= -\frac{a_3}{S}\sqrt{2gh_3} + \frac{(1-\gamma_a)}{S}q_a, \\ \frac{dh_4}{dt} &= -\frac{a_4}{S}\sqrt{2gh_4} + \frac{(1-\gamma_b)}{S}q_b,\end{aligned}$$

în care funcțiile h_1, h_2, h_3, h_4 reprezintă dinamicile nivelurilor lichidului în rezervoare, cu rolul de stări ale sistemului, iar q_a, q_b reprezintă debitele de intrare, cu rolul de comenzi (intrări). Pentru similaritate cu subcapitolul anterior, vom renota nivelul h_i cu z_i pentru $i = 1, \dots, 4$, iar debitele (q_a, q_b) cu (u_1, u_2) .

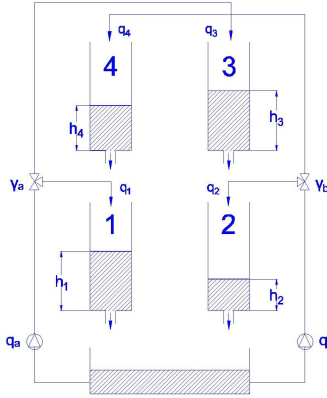


Figura 14.5: Structura instalației cu patru rezervoare.

Pentru discretizare, utilizăm metoda Euler pentru care putem alege perioada de eșantionare $\Delta t = 5$ s, deoarece procesul este unul lent. Pentru o prezentare simplificată, scriem compact sistemul neliniar discret anterior prin intermediul următoarelor notații: $z_{t+1} = \phi(z_t) + B_u u_t$, unde $\phi : \mathbb{R}^4 \rightarrow \mathbb{R}^4$ cu

$$\phi(z) = \begin{bmatrix} z_1 - \frac{5a_1}{S}\sqrt{2gz_1} + \frac{5a_4}{S}\sqrt{2gz_4} \\ z_2 - \frac{5a_2}{S}\sqrt{2gz_2} + \frac{5a_3}{S}\sqrt{2gz_3} \\ z_3 - \frac{5a_3}{S}\sqrt{2gz_3} \\ z_4 - \frac{5a_4}{S}\sqrt{2gz_4} \end{bmatrix} \quad B_u = \begin{bmatrix} \frac{5\gamma_a}{S} & 0 \\ 0 & \frac{5\gamma_b}{S} \\ \frac{5(1-\gamma_a)}{S} & 0 \\ 0 & \frac{5(1-\gamma_b)}{S} \end{bmatrix}.$$

Valorile parametrilor din cadrul modelului se pot identifica experimental prin diferite tehnici. Valorile aproximative identificate în laborator sunt prezentate în Tabelul 14.1.

Parametrii	S	a_1	a_2	a_3	a_4	γ_a	γ_b
Valori	0.02	$5.8e-5$	$6.2e-5$	$2e-5$	$3.6e-5$	0.58	0.54
Unitate	m^2	m^2	m^2	m^2	m^2		

Tabelul 14.1: Parametrii procesului cu patru rezervoare.

Formulăm o problemă de control optimal pentru modelul neliniar al instalației, considerând un orizont de predicție N și referințe pentru intrare și stare date z_t^{ref} și u_t^{ref} :

$$\min_{z_t, u_t} \frac{1}{2} \sum_{t=1}^N \|z_t - z_t^{ref}\|_{Q_0}^2 + \frac{1}{2} \sum_{t=0}^{N-1} \|u_t - u_t^{ref}\|_{R_0}^2$$

$$\text{s.l.: } z_0 = z, \quad z_{t+1} = \phi(z_t) + B_u u_t \quad \forall t = 0, \dots, N-1.$$

Observăm că problema de optimizare neconvexă ce rezultă din problema de control optimal fără eliminarea stărilor are forma:

$$\min_{x \in \mathbb{R}^{6N}} \frac{1}{2} x^T Q x + q^T x \quad (14.13)$$

$$\text{s.l.: } h(x) = 0,$$

în care matricea Q și vectorul q este definit în secțiunea 14.1.1, iar funcția h este definită în secțiunea 14.2.1. Rezolvăm problema neconvexă cu constrângeri de egalitate prin metoda Newton-Lagrange. Reamintim că metoda Newton-Lagrange se bazează pe iterația Newton pentru rezolvarea sistemului de ecuații:

$$Qx + q + \nabla h(x)^T \mu = 0, \quad h(x) = 0. \quad (14.14)$$

În Fig. 14.6 considerăm două referințe pentru fiecare rezervor. Mai exact, considerăm referințe constantă pentru 500 s și apoi modificăm aceste referințe la alte valori constante pentru încă 1000 s. Se observă că sistemul urmărește foarte bine aceste referințe prin strategia de control optimal folosind principiul de orizont alunecător cu $N = 5$.

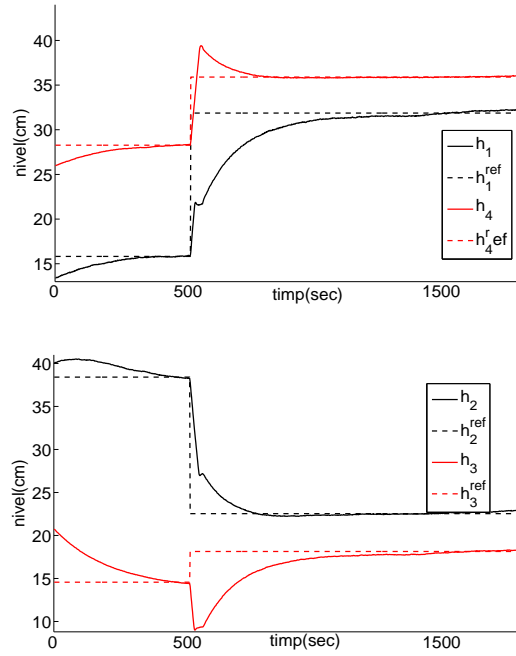


Figura 14.6: Curba nivelului de apa din cele patru rezervoare.

14.3 Stabilitatea sistemelor dinamice

În această aplicație facem conexiunea între teoria sistemelor cu cea a optimizării. Considerăm, de exemplu, clasa de sisteme liniare discrete autonome:

$$z_{t+1} = Az_t, \quad (14.15)$$

unde matricea $A \in \mathbb{R}^{n \times n}$. În particular, analizăm clasa *sistemelor liniare pozitive*. Pentru a defini matematic această clasă de sisteme liniare pozitive, introducem mai întâi noțiunea de *matrice ne-negativă*: o matrice $A \in \mathbb{R}^{n \times n}$ se numește ne-negativă, dacă pentru orice $x \in \mathbb{R}^n$ ce satisface $x \geq 0$ avem $Ax \geq 0$. Pe baza acestei definiții putem introduce noțiunea de *sistem liniar pozitiv*: un sistem liniar discret se numește pozitiv dacă matricea de stare A din definiția (14.15) este matrice ne-negativă.

Sistemele pozitive se regăsesc în foarte multe domenii din economie, biologie, probabilistică, rețelistică (modele ce includ protocolul TCP), controlul traficului, probleme de sincronizare și control în rețele wireless, etc. În toate aceste aplicații starea sistemului este reprezentată numeric

de numere ne-negative, adică traiectoria unui astfel de sistem evoluează în \mathbb{R}_+^n . Analiza stabilității sistemelor pozitive liniare se verifică prin calcularea valorilor proprii extreme corespunzătoare spectrului matricei de stare. Un astfel de sistem este asimptotic stabil dacă raza spectrală a matricei A este strict mai mică decât 1. De aceea, în continuare ne propunem să calculăm valorile proprii ale unei matrice ne-negative.

14.3.1 Calcularea valorilor proprii ale unei matrice

Proprietățile unei matrice pătratice sunt definite fundamental de *spectrul* acesteia (setul de valori proprii). Complexitatea calculării spectrului reprezintă subiectul central de studiu în domeniul algebrei liniare. Majoritatea algoritmilor ce tratează problema calculării spectrului unei matrice pătratice sunt *metode iterative* (e.g. algoritmul QR). Însă, în anumite cazuri, experimentele numerice indică o eficiență numerică vizibil mai ridicată a abordării calculării spectrului prin intermediul metodelor de optimizare.

O tehnică des folosită pentru calcularea întregului spectru a unei matrice simetrice presupune calcularea valorii proprii maxime, reducerea spectrului prin operația de *deflație* și apoi, reiterarea întregului proces până la eliminarea tuturor valorilor proprii. Problema calculării valorii proprii maxime a unei matrice simetrice $A \in \mathbb{R}^{n \times n}$ se poate formula prin următoarea problemă neconvexă și neconstrânsă:

$$\max_{x \in \mathbb{R}^n, x \neq 0} \frac{x^T A x}{x^T x}. \quad (14.16)$$

Observăm că funcția obiectiv din cadrul problemei (14.16) este neconvexă, iar punctele staționare (vectorii ce satisfac condițiile de ordinul I) sunt date de vectorii proprii ai matricei A . Presupunând că matricea A are n vectori proprii distincți, orice metodă de ordin I sau II prezentată în această lucrare va converge la unul dintre acești vectori proprii. Din acest motiv, apar dificultăți în a garanta că metoda aleasă de noi converge la vectorul propriu maximal al matricei A . Însă, se pot observa proprietăți excepționale ale matricelor ne-negative în legătură cu vectorul propriu maximal. Din teorema Perron-Frobenius se poate deduce ușor că vectorul propriu maximal al unei matrice ne-negative ireductibile este singurul din setul de vectori proprii ce are componente ne-negative. În concluzie, dacă adăugăm un set de constrângeri $x \geq 0$ problemei (14.16), putem garanta că metodele de

ordinul I sau II prezentate în această lucrare converg la punctul de maxim global (vectorul propriu maximal). Deoarece subspațiul invariant definit de vectorii proprii ai unei matrice este liniar, pentru a garanta o soluție finită a problemei (14.16) considerăm vectorul normalizat și rezolvăm următoarea problemă de optimizare neconvexă și constrânsă:

$$\begin{aligned} \max_{x \in \mathbb{R}^n} \frac{x^T A x}{x^T x} \\ \text{s.l.: } e^T x = 1, \quad x \geq 0. \end{aligned} \quad (14.17)$$

Punctul de optim (global) al acestei probleme reprezintă vectorul propriu corespunzător valorii proprii maxime a matricei A ne-negative și ireductibile, iar valoarea optimă este dată de valoarea proprie maximă a matricei A . Pentru un scalar $\tau > 0$, aplicăm mai întâi metoda de penalitate pentru a muta constrângerile de egalitate în cost și apoi rezolvăm problema corespunzătoare cu metoda gradient proiectat, i.e.:

$$\begin{aligned} \max_{x \in \mathbb{R}^n} \frac{x^T A x}{x^T x} + \tau(e^T x - 1)^2 \\ \text{s.l.: } x \geq 0. \end{aligned} \quad (14.18)$$

Diferența dintre problemele (14.17) și (14.18) este penalizarea constrângerii de egalitate cu parametrul τ . Pentru a aplica forma standard a metodei gradient proiectat, aducem modelul (14.17) la forma unei probleme de minimizare observând următoarea echivalență:

$$\max_{x \in \mathbb{R}^n, x \neq 0} \frac{x^T A x}{x^T x} = \min_{x \in \mathbb{R}^n, x \neq 0} \frac{x^T x}{x^T A x}.$$

În concluzie, pentru un $\tau > 0$ suficient de mare putem considera problema de optimizare:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} F(x, \tau) \quad \left(= \frac{x^T x}{x^T A x} + \tau(e^T x - 1)^2 \right) \\ \text{s.l.: } x \geq 0. \end{aligned} \quad (14.19)$$

Deoarece mulțimea fezabilă a problemei (14.19) este convexă și simplă, proiecția pe această mulțime se poate obține analitic. Mai exact, fie un vector $x \in \mathbb{R}^n$, proiecția pe mulțimea vectorilor cu elemente ne-negative $\mathbb{R}_+^n = \{x \in \mathbb{R}^n : x \geq 0\}$ este dată de:

$$[x]_{(I_n, \mathbb{R}_+^n)} = [\max\{0, x_1\} \dots \max\{0, x_n\}]^T.$$

Aplicăm metoda de gradient proiectat pentru rezolvarea problemei neconvexe (14.19) cu pas constant $\alpha > 0$ pentru anumite valori ale parametrului de penalitate τ :

$$x_{k+1} = [x_k - \alpha \nabla F(x_k, \tau)]_{(I_n, \mathbb{R}_+^n)}.$$

Rezultatele obținute sunt prezentate în Fig. 14.7. Se observă o rată de convergență relativ rapidă (liniară) a valorilor funcției $f(x_k)$ către valoarea optimă f^* indiferent de valoarea parametrului de penalitate τ . Din simulări observăm că nu trebuie să luăm valori foarte mari pentru parametrul de penalitate τ .

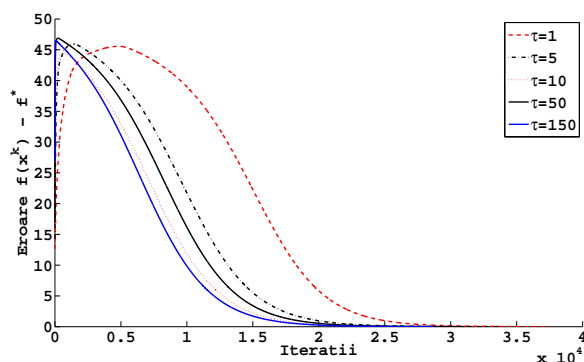


Figura 14.7: Convergenta metodei gradient proiectat pentru diferite valori ale parametrului de penalitate τ pe o problemă de dimensiune $n = 10^3$.

14.4 Problema Google (ierarhizarea paginilor web)

Internetul este dominat progresiv de *motoare de căutare*, în sprijinul selecției și găsirii surselor de informații cu relevanță maximă. Unul dintre cele mai vechi și eficiente motoare de căutare este Google, care se află în continuă dezvoltare pe măsură ce progresele în domeniul algoritmilor avansează. Tehnica folosită de Google pentru căutarea și clasificarea paginilor web se numește *PageRank*, iar pasul central din această tehnică presupune clasificarea (ranking-ul) unui număr uriaș de pagini web. În acest fel, rezultă o listă ordonată de site-uri în sensul descrescător al relevanței în legătură cu subiectul căutat.

Datorită conexiunilor permanente dintre paginile web în rețeaua internet-ului, putem să reprezentăm structura legăturilor dintre pagini prin intermediul unui graf ponderat orientat. Nodurile grafului reprezintă paginile, iar muchiile au rolul link-urilor. Ponderea p_{ij} (corespunzătoare muchiei dintre nodurile i și j) reprezintă probabilitatea ca la o navigare aleatorie în rețeaua de pagini să se ajungă din pagina i în pagina j . În plus, putem atribui grafului o matrice de adiacență $E \in \mathbb{R}^{n \times n}$, cu componenta $E_{ij} \neq 0$ dacă între nodurile i și j există muchie, iar $E_{ij} = 0$ dacă nodurile i și j nu sunt legate de o muchie. Numărul de muchii din graf se reflectă în numărul de elemente nenule ale matricei E ; de aceea, pentru un graf rar (cu puține muchii), matricea de adiacență va fi rară (va conține preponderent zerouri, vezi Fig. 14.8).

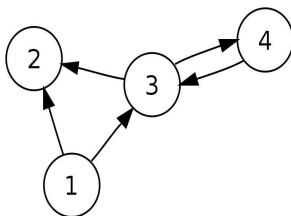


Figura 14.8: *Exemplu de graf orientat.*

Pentru a analiza mai îndeaproape proprietățile matricei de adiacență rezultate din graful paginilor web, introducem următoarele noțiuni:

Definiția 14.4.1 O matrice $E \in \mathbb{R}^{m \times n}$ se numește stocastică pe linii dacă are elemente nenegative (i.e. $E_{ij} \geq 0$), iar suma pe fiecare linie este egală cu 1. O matrice $E \in \mathbb{R}^{m \times n}$ se numește stocastică pe coloane dacă are elemente nenegative (i.e. $E_{ij} \geq 0$), iar suma pe fiecare coloană este egală cu 1.

Deoarece componentele nenule ale matricei de adiacență E au rolul de probabilități, acestea sunt nenegative, iar matricea E este stocastică pe coloane. Forma algebrică a problemei Google se reduce la a găsi vectorul propriu corespunzător valorii proprii maxime 1, adică soluția următorului sistem liniar supus constrângerilor:

$$\begin{cases} Ex = x \\ e^T x = 1, \quad x \geq 0. \end{cases}$$

Problema găsirii soluției sistemului anterior se poate formula ușor în termeni de optimizare:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} f(x) \quad & \left(= \frac{1}{2} \|Ex - x\|^2 \right) \\ \text{s.l.: } & e^T x = 1, \quad x \geq 0, \end{aligned} \quad (14.20)$$

unde $e = [1 \dots 1]^T$, matricea $E \in \mathbb{R}^{n \times n}$ este rară (elementele au valori preponderent nule). Observăm că problema rezultată este constrânsă, însă dacă alegem un parametru $\tau > 0$ suficient de mare, putem obține o formulare echivalentă fără constrângeri folosind funcția de penalitate pătratică:

$$\min_{x \in \mathbb{R}^n} F(x, \tau) \quad \left(= \frac{1}{2} \|Ex - x\|^2 + \frac{\tau}{2} (e^T x - 1)^2 \right). \quad (14.21)$$

Pe baza teoremei Peron-Frobenius, observăm că putem elimina constrângerile de inegalitate $x \geq 0$, deoarece soluția optimă globală a problemei de optimizare (14.21) satisface automat această constrângere. Datorită dimensiunilor foarte mari ale ambelor probleme considerate (14.20) și (14.21), ne orientăm atenția către algoritmi de ordinul I, deoarece au o *complexitate scăzută per iterație*:

- (i) metoda de gradient proiectat pentru cazul constrâns (14.20):
 $x_{k+1} = [x_k - \alpha \nabla f(x_k)]_{(I_n, \Delta_n)}$;
- (ii) metoda gradient pentru cazul neconstrâns (14.21): $x_{k+1} = x_k - \alpha \nabla F(x_k, \tau)$,

în care $\Delta_n = \{x \in \mathbb{R}^n : e^T x = 1, x \geq 0\}$ este mulțimea numită simplex și $\alpha > 0$ este un pas constant.

Rezultatele obținute sunt prezentate în Fig. 14.9 and 14.10. Se observă o convergență rapidă în ambele metode. De asemenea, observăm că metoda de penalitate produce o soluție optimă pentru problema originală pentru valori relativ mici ale parametrului de penalitate τ .

14.5 Învățare automată și clasificare

Tehnicile de *învățare automată* și *clasificare* sunt noțiuni centrale în domeniul statisticii, calculatoarelor, prelucrării semnalelor, etc. Ambele se ocupă în mod fundamental cu *problema recunoașterii tiparelor* (*pattern*

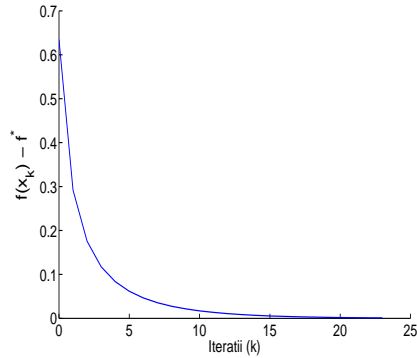


Figura 14.9: Curba de convergență a metodei de gradient proiectat aplicată unei probleme Google cu dimensiunea $n = 10^2$.

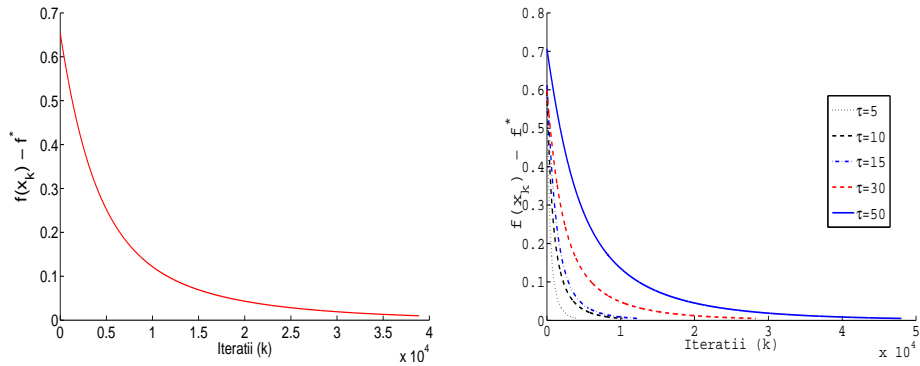


Figura 14.10: Curba de convergență a metodei de gradient pentru problema Google cu $n = 10^3$, $\tau = 50$ (stânga); dependența convergenței metodei gradient de parametrul τ aplicată problemei Google pe $n = 10^3$ (dreapta).

recognition) prin dezvoltarea de modele matematice ce suportă o etapă preliminară de *antrenare* (experiență), pe baza căreia realizează operații de clasificare/regresie de obiecte și funcții.

O parte din numeroasele aplicații ale acestor tehnici cuprinde:

1. recunoașterea email-urilor de tip spam sau malware;
2. recunoașterea vocii/feței;
3. compresia cantităților uriașe de date;
4. detecția de tipare în cadrul unei imagini;

5. recunoașterea scrisului de mână.

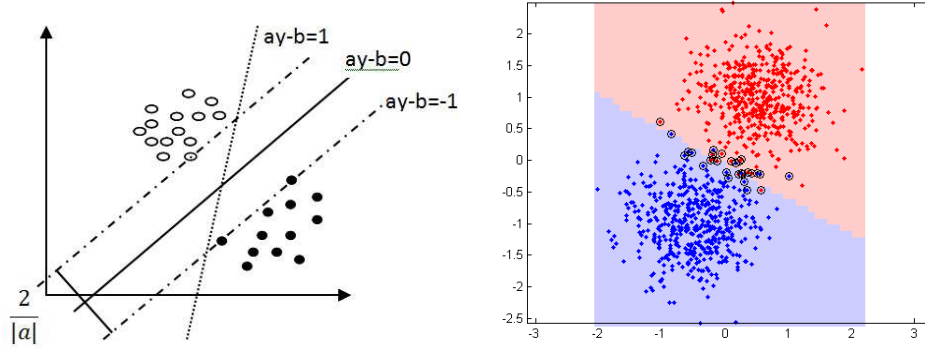


Figura 14.11: Hiperplan de separare a două clase de obiecte.

Una dintre cele mai renumite tehnici de recunoaștere/clasificare este *SVM* - *Support Vector Machine*. Această tehnică presupune determinarea unui model matematic ce separă două sau mai multe clase de obiecte cu o anumită acuratețe. În vederea clasificării sau recunoașterii unui obiect necunoscut, se introduc datele obiectului în modelul matematic, iar la ieșire se primește id-ul clasei din care face parte. În cele mai simple cazuri, modelul matematic căutat este reprezentat de un hiperplan $H = \{y \in \mathbb{R}^n : a^T y = b\}$ caracterizat de parametrii $a \in \mathbb{R}^n$ și $b \in \mathbb{R}$. De aceea, problema se reduce la a găsi parametrii optimi (a, b) care să separe cât mai bine clasele de obiecte.

Fie setul de puncte recunoscute *a priori* y_i cu $i = 1, \dots, m$, reprezentate în Fig. 14.11 de puncte având două culori diferite. Termenul *recunoscute* denotă că pentru fiecare punct y_i cunoaștem clasa din care face parte. Dacă luăm ca exemplu Fig. 14.11, putem argumenta că se cunoaște un parametru auxiliar c_i cu valoarea $+1$ dacă obiectul y_i este e.g. de culoare roșie, iar $c_i = -1$ dacă obiectul este de culoare albastră.

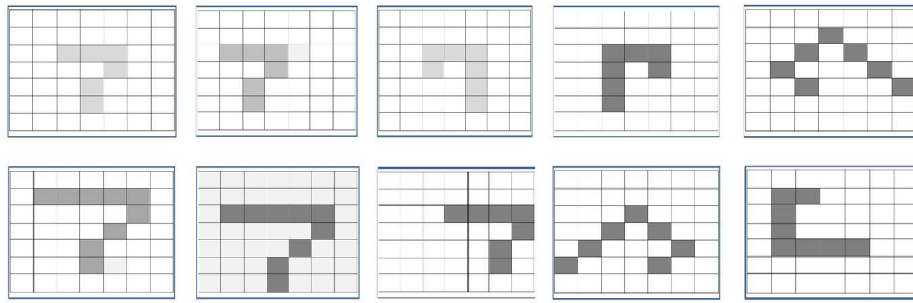
Dacă datele de antrenare sunt liniar separabile, atunci putem selecta două hiperplane în așa fel încât ele separe datele, nu conțin puncte între ele și maximizează distanța între cele două hiperplane. Aceste două hiperplane pot fi descrise de ecuațiile: $a^T y - b = 1$ și $a^T y - b = -1$. Regiunea dintre aceste două hiperplane se numește *margină* și este descrisă de expresia $2/\|a\|$.

În termenii teoriei optimizării, problema se formulează după cum

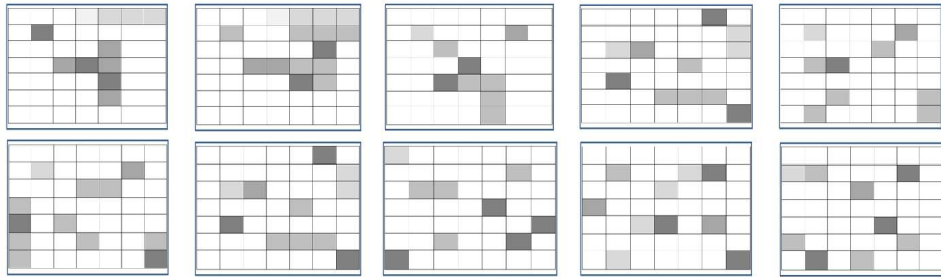
urmează:

$$\begin{aligned} \min_{a \in \mathbb{R}^n, b \in \mathbb{R}} \quad & \frac{1}{2} \|a\|^2 \\ \text{s.l.: } \quad & c_i (a^T y_i - b) \geq 1 \quad \forall i = 1, \dots, m, \end{aligned} \quad (14.22)$$

unde a și b reprezintă parametrii hiperplanului, iar c_i indică clasa din care face parte obiectul y_i . Variabilele de decizie $x = [a^T \ b]^T$ reprezintă parametrii unui hiperplan de separare a claselor de obiecte/imagini, așa cum se observă în Fig. 14.11. Problema de optimizare convexă pătratică având numai constrângeri de inegalitate (14.22) o rezolvăm prin metoda de punct interior aplicată problemelor convexe (CP), metodă descrisă în capitolul anterior.



(a)



(b)

Figura 14.12: Mulțimea de antrenare a modelului matematic de separare:

(a) imagini ce fac parte din clasa I; (b) imagini ce fac parte din clasa II.

În continuare, exemplificăm o aplicație practică a tehnicii SVM prin problema recunoașterii cifrei 7 dintr-o imagine. Se cunoaște că orice imagine poate fi reprezentată sub forma unei serii de pixeli, unde fiecare

pixel la rândul său este definit de o valoare (e.g. între 0 și 256) dată de culoarea acestuia. Pentru a simplifica exemplul, considerăm imagini mono-culore compuse din 49 de pixeli, în care pixelii sunt reprezentați de nivele de gri cu valori între 0 și 5 (vezi Fig. 14.12). În etapa de inițializare a tehnicii SVM se fixează o mulțime de antrenare compusă din diferite imagini ce conțin variante ale cifrei 7 (ce fac parte din clasa I de obiecte) și imagini aleatorii complet diferite de cifra 7 (ce fac parte din clasa II de obiecte). Deoarece această etapă se mai numește și *antrenare*, se cunoaște pentru fiecare imagine clasa din care face parte. Fiecărei imagini i i se asociază un vector de 49 de componente (fiecare componentă luând valori întregi între 0 și 5) și un parametru c ce reprezintă indexul clasei din care face parte imaginea respectivă (dacă $c = 1$ atunci imaginea conține cifra 7, dacă $c = -1$ atunci imaginea este aleatorie). Pe baza acestei mulțimi de antrenare, urmărim realizarea unui hiperplan de separare a acestor două clase.

Dorim să rezolvăm problema SVM (14.22) în contextul prezentat anterior și, de asemenea, să testăm eficiența soluției (hiperplanului) obținute prin evaluarea ratei de succes în recunoașterea cifrei 7. Pentru aceasta alegem un set de imagini ale cifrei 7 (vezi Fig. 14.12 (a)) și un set de imagini aleatorii (vezi Fig. 14.12 (b)) ce reprezintă *mulțimea de antrenare* a hiperplanului de separare.

Transformăm aceste imagini din Fig. 14.12 în vectori de pixeli după cum am descris mai înainte, apoi aceștia vor fi introduși într-o funcție Matlab și folosiți în rezolvarea problemei (14.22). În final, pentru a testa soluția găsită $x^* = [(a^*)^T \ b^*]^T$ din rezolvarea problemei convexe pătratice (14.22), calculăm pentru anumite *puncte de test* (imagini de test date în Fig. 14.13) valoarea hiperplanului:

$$a^T y - b \begin{cases} < 0, & \text{atunci imaginea dată de } y \text{ nu conține cifra 7;} \\ \geq 0, & \text{atunci imaginea dată de } y \text{ conține cifra 7.} \end{cases}$$

Putem trage următoarele concluzii:

dacă testăm hiperplanul cu diferite imagini aleatorii cu densitate mare de pixeli gri (vezi Fig. 14.13) și respectiv, imagini cu cifra 7 transformată în diverse moduri (translație la stânga/dreapta, înclinare, etc.) atunci rezultă o rată de succes (recunoaștere corectă) de aproximativ 80%;

dacă pentru testare considerăm imagini aleatorii cu densitate mică (vezi Fig. 14.12 (b)) și respectiv, imagini cu cifra 7 transformată



Figura 14.13: *Exemple de imagini de test aleatorii, ce conțin cifra 7 sau cu densitate ridicată de pixeli gri.*

în diverse moduri (translație la stânga/dreapta, înclinare, din Fig. 14.13), atunci rezultă o rată de succes de aproximativ 52%.

Motivația ratei mici de succes în cel de-al doilea caz este dată de doi factori: (i) similaritatea ridicată între imaginile cu densitate mică de pixeli și cele ce conțin cifra 7; (ii) numărul relativ mic de imagini de antrenare, în cazul nostru 20 de imagini test. Evident, cu cât mulțimea de antrenare conține mai multe date (imagini) cu atât hiperplanul rezultat este mai eficient în recunoașterea noilor obiecte.

Apendice A

Noțiuni de algebră liniară și analiză matematică

În acest capitol reamintim pe scurt noțiunile de bază din algebra liniară și analiza matematică ce se vor fi utilizate în această lucrare. Pentru mai multe detalii și demonstrații ale rezultatelor prezentate în acest capitol se pot consulta cărțile [4, 8, 17] pentru algebra liniară și [15, 16] pentru analiza matematică.

A.1 Noțiuni de analiză matriceală

În cadrul acestei lucrări fixăm simpla convenție de a considera vectorii $x \in \mathbb{R}^n$ vectori coloană, i.e. $x = [x_1 \cdots x_n]^T \in \mathbb{R}^n$. În spațiul Euclidian \mathbb{R}^n produsul scalar este definit după cum urmează:

$$\langle x, y \rangle = x^T y = \sum_{i=1}^n x_i y_i.$$

Unde nu se specifică, norma considerată pe spațiul Euclidian \mathbb{R}^n este norma Euclidiană standard (i.e. norma indusă de acest produs scalar):

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{i=1}^n x_i^2}.$$

Alte norme vectoriale des întâlnite sunt:

$$\|x\|_1 = \sum_{i=1}^n |x_i| \quad \text{și} \quad \|x\|_\infty = \max_{i=1, \dots, n} |x_i|.$$

Unghiul $\theta \in [0, \pi]$ dintre doi vectori nenuli x și y din \mathbb{R}^n este definit de:

$$\cos \theta = \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|}.$$

Orice normă $\|\cdot\|$ în \mathbb{R}^n are o *normă duală* corespunzătoare $\|\cdot\|^*$ definită de:

$$\|y\|^* = \max_{x \in \mathbb{R}^n: \|x\|=1} \langle x, y \rangle \quad \forall y \in \mathbb{R}^n.$$

Se poate arăta că $\|x\|_\infty = \|x\|_1^*$ și $\|x\|_1 = \|x\|_\infty^*$ pentru orice vector $x \in \mathbb{R}^n$.

O relație fundamentală ce se folosește intens în acest curs este inegalitatea Cauchy-Schwarz definită de următoarea relație între produsul scalar dintre doi vectori și normele duale corespunzătoare:

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\|^* \quad \forall x, y \in \mathbb{R}^n,$$

egalitatea având loc dacă și numai dacă vectorii x și y sunt vectori liniar dependenți. Observăm că această inegalitate este o consecință imediată a definiției normei duale.

Spațiul matricelor de dimensiune (m, n) este notat cu $\mathbb{R}^{m \times n}$. *Urma* unei matrice pătratică $Q = [Q_{ij}]_{ij} \in \mathbb{R}^{n \times n}$ este definită de relația:

$$\text{Tr}(Q) = \sum_{i=1}^n Q_{ii}.$$

În spațiul matricelor de dimensiune (m, n) definim produsul scalar folosind noțiunea de urmă:

$$\langle Q, P \rangle = \text{Tr}(Q^T P) = \text{Tr}(QP^T) \quad \forall Q, P \in \mathbb{R}^{m \times n}.$$

Din proprietățile produsului scalar rezultă:

$$\text{Tr}(QPR) = \text{Tr}(RQP) = \text{Tr}(PRQ),$$

oricare ar fi matricele Q, P și R de dimensiuni compatibile. În consecință, pentru matricele pătratică $Q \in \mathbb{R}^{n \times n}$ avem de asemenea relația:

$$x^T Q x = \text{Tr}(Q x x^T) \quad \forall x \in \mathbb{R}^n.$$

Pentru o matrice pătratică $Q \in \mathbb{R}^{n \times n}$, un scalar $\lambda \in \mathbb{C}$ și un vector nenul x ce satisfac ecuația $Qx = \lambda x$ se numesc *valoare proprie* și respectiv,

vector propriu al matricei Q . O relație echivalentă ce descrie perechea valoare-vector propriu este dată de:

$$(\lambda I_n - Q)x = 0, \quad x \neq 0,$$

i.e. matricea $\lambda I_n - Q$ este singulară, de aceea,

$$\det(\lambda I_n - Q) = 0.$$

În acest scop, *polinomul caracteristic* al matricei Q este definit de:

$$p_Q(\lambda) = \det(\lambda I_n - Q).$$

Evident, mulțimea de soluții ale ecuației $p_Q(\lambda) = 0$ coincide cu mulțimea de valori proprii ale lui Q . Mulțimea tuturor valorilor proprii corespunzătoare matricei Q este denumită *spectrul* matricei Q și se notează cu $\Lambda(Q) = \{\lambda_1, \dots, \lambda_n\}$. Folosind această notație avem:

$$p_Q(\lambda) = (\lambda - \lambda_1) \cdots (\lambda - \lambda_n)$$

și rezultă $p_Q(0) = \prod_{i=1}^n (-\lambda_i)$. Din discuția precedentă se obține următorul rezultat:

Lema A.1.1 *Următoarele relații au loc pentru orice matrice pătratică $Q \in \mathbb{R}^{n \times n}$:*

$$\det(Q) = \prod_{i=1}^n \lambda_i \quad \text{și} \quad \text{Tr}(Q) = \sum_{i=1}^n \lambda_i$$

$$\lambda_i(Q^k) = \lambda_i^k \quad \text{și} \quad \lambda_i(\alpha I_n + \beta Q) = \alpha + \beta \lambda_i \quad \forall \alpha, \beta \in \mathbb{R} \quad \text{și} \quad i = 1, \dots, n.$$

Notăm cu S^n spațiul matricelor simetrice:

$$S^n = \{Q \in \mathbb{R}^{n \times n} : Q = Q^T\}.$$

Pentru o matrice simetrică $Q \in S^n$ valorile proprii corespunzătoare sunt reale, i.e. $\Lambda(Q) \subset \mathbb{R}$. O matrice simetrică $Q \in S^n$ este *pozitiv semidefinită* (notație $Q \succeq 0$) dacă

$$x^T Q x \geq 0 \quad \forall x \in \mathbb{R}^n$$

și *pozitiv definită* (notație $Q \succ 0$) dacă

$$x^T Q x > 0 \quad \forall x \in \mathbb{R}^n, x \neq 0.$$

Precizăm că $Q \succeq P$ dacă $Q - P \succeq 0$. Notăm mulțimea matricelor pozitiv (semi)definite cu $(S_+^n)_{++}^n$. Mai departe, avem următoarea caracterizare a unei matrice pozitiv semidefinite:

Lema A.1.2 Următoarele echivalențe au loc pentru orice matrice simetrică $Q \in S^n$:

- (i) matricea Q este pozitiv semidefinită;
- (ii) toate valorile proprii ale matricei Q sunt ne-negative (adică $\lambda_i \geq 0 \forall i = 1, \dots, n$);
- (iii) toți minorii principali ai lui Q sunt ne-negativi;
- (iv) există o matrice L astfel încât $Q = L^T L$.

În continuare, folosim notația λ_{\min} și λ_{\max} pentru cea mai mică și respectiv, cea mai mare valoare proprie a unei matrice simetrice $Q \in S^n$. Atunci,

$$\lambda_{\min} = \min_{x \in \mathbb{R}^n: x \neq 0} \frac{x^T Q x}{x^T x} = \min_{x \in \mathbb{R}^n: \|x\|=1} x^T Q x$$

$$\lambda_{\max} = \max_{x \in \mathbb{R}^n: x \neq 0} \frac{x^T Q x}{x^T x} = \max_{x \in \mathbb{R}^n: \|x\|=1} x^T Q x.$$

În concluzie avem:

$$\lambda_{\min} I_n \preceq Q \preceq \lambda_{\max} I_n.$$

Putem defini norme matriceale utilizând norme vectoriale. Fie normele vectoriale $\|\cdot\|'$ pe \mathbb{R}^n și $\|\cdot\|''$ pe \mathbb{R}^m , atunci putem defini o normă matricială indusă pe spațiul matricelor $\mathbb{R}^{m \times n}$ prin următoarea relație:

$$\|Q\|_{',''} = \sup_{x \in \mathbb{R}^n: x \neq 0} \frac{\|Qx\|''}{\|x\|'} = \sup_{x \in \mathbb{R}^n: \|x\|=1} \|Qx\|'' \quad \forall Q \in \mathbb{R}^{m \times n}.$$

Pentru norma vectorială Euclidiană norma matricială indusă este dată de:

$$\|Q\| = (\lambda_{\max}(Q^T Q))^{1/2}.$$

De asemenea, norma Frobenius a unei matrice este definită prin relația:

$$\|Q\|_F = \left(\sum_{i=1}^m \sum_{j=1}^n Q_{ij}^2 \right)^{1/2}.$$

Reamintim de asemenea o formulă pentru inversarea de matrici, numită formula *Sherman-Morrison-Woodbury*: fie o matrice $A \in \mathbb{R}^{n \times n}$ inversabilă și două matrice U și V în $\mathbb{R}^{n \times p}$, cu $p \leq n$. Atunci matricea $A + UV^T$ este inversabilă dacă și numai dacă matricea $I_n + V^T A^{-1} U$ este inversabilă și în acest caz avem:

$$(A + UV^T)^{-1} = A^{-1} - A^{-1} U (I_n + V^T A^{-1} U)^{-1} V^T A^{-1}.$$

Un caz particular al acestei formule este următorul: pentru doi vectori $u, v \in \mathbb{R}^n$

$$(A + uv^T)^{-1} = A^{-1} - \frac{1}{1 + v^T A^{-1} u} A^{-1} uv^T A^{-1}.$$

A.2 Noțiuni de analiză matematică

În cadrul acestei lucrări ne vom concentra atenția preponderent asupra conceptelor, relațiilor și rezultatelor ce implică funcții al căror codomeniu este inclus în $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$. Pentru început, o observație importantă pentru rigurozitatea rezultatelor ulterioare este aceea că domeniul efectiv al unei funcții scalare f se poate extinde (prin echivalență) la întreg spațiul \mathbb{R}^n prin atribuirea valorii $+\infty$ funcției în toate punctele din afara domeniului său. În cele ce urmează considerăm că toate funcțiile sunt extinse implicit. O funcție scalară $f : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ are *domeniul efectiv* descris de mulțimea:

$$\text{dom } f = \{x \in \mathbb{R}^n : f(x) < +\infty\}.$$

Funcția f se numește *diferențiabilă* în punctul $x \in \text{dom } f$ dacă există un vector $s \in \mathbb{R}^n$ astfel încât următoarea relație are loc:

$$f(x + y) = f(x) + \langle s, y \rangle + \mathcal{R}(\|y\|) \quad \forall y \in \mathbb{R}^n,$$

unde $\lim_{y \rightarrow 0} \frac{\mathcal{R}(\|y\|)}{\|y\|} = 0$ și $\mathcal{R}(0) = 0$. Vectorul s se numește derivata sau gradientul funcției f în punctul x și se notează cu $\nabla f(x)$. Cu alte cuvinte, funcția este diferențiabilă în x dacă admite o aproximare liniară de ordinul I în punctul x . Observăm că gradientul este unic determinat și este definit de vectorul cu componentele:

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

Funcția f se numește diferențiabilă pe mulțimea $X \subseteq \text{dom } f$ dacă este diferențiabilă în toate punctele din X .

Expresia (în condițiile în care limita de mai jos există):

$$f'(x; d) = \lim_{t \rightarrow +0} \frac{f(x + td) - f(x)}{t}$$

se numește *derivata direcțională* a funcției f în punctul $x \in \text{dom} f$ de-a lungul direcției $d \in \mathbb{R}^n$. Precizăm că derivata direcțională poate exista de asemenea pentru funcții nediferențiabile, după cum observăm din următorul exemplu:

Exemplul A.2.1 Pentru funcția $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $f(x) = \|x\|_1$ avem că derivata direcțională în punctul $x = 0$ de-a lungul oricărei direcții $d \in \mathbb{R}^n$ este dată de expresia $f'(0; d) = \|d\|_1$, însă f nu este diferentiabilă în punctul $x = 0$.

În cazul în care funcția este diferentiabilă, atunci:

$$f'(x; d) = \langle \nabla f(x), d \rangle.$$

O funcție scalară $f : \mathbb{R}^n \rightarrow \mathbb{R}$ se numește *diferentiabilă de două ori* în punctul $x \in \text{dom} f$ dacă este diferentiabilă în x și există o matrice simetrică $H \in \mathbb{R}^{n \times n}$ astfel încât:

$$f(x + y) = f(x) + \langle \nabla f(x), y \rangle + \frac{1}{2} y^T H y + \mathcal{R}(\|y\|^2) \quad \forall y \in \mathbb{R}^n,$$

unde $\lim_{y \rightarrow 0} \frac{\mathcal{R}(\|y\|^2)}{\|y\|^2} = 0$. Matricea H se numește matricea *Hessiană* și se notează cu $\nabla^2 f(x)$. În concluzie, o funcție este diferentiabilă de două ori în punctul x dacă admite o aproximare pătratică de ordin doi în vecinătatea lui x . Ca și în cazul gradientului, matricea Hessiană este unică în cazurile în care există și este simetrică având componentele:

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial^2 x_1} & \dots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \dots & \dots & \dots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \dots & \frac{\partial^2 f(x)}{\partial^2 x_n} \end{bmatrix}.$$

Funcția f se numește diferentiabilă de două ori pe mulțimea $X \subseteq \text{dom} f$ dacă este diferentiabilă de două ori în fiecare punct din X . Matricea Hessiană poate fi considerată derivata vectorului ∇f :

$$\nabla f(x + y) = \nabla f(x) + \nabla^2 f(x) y + \mathcal{R}(\|y\|).$$

Exemplul A.2.2 Fie f o funcție pătratică:

$$f(x) = \frac{1}{2} x^T Q x + q^T x + r,$$

unde $Q \in \mathbb{R}^{n \times n}$ este matrice simetrică. Atunci, este evident că gradientul lui f în orice punct $x \in \mathbb{R}^n$ este:

$$\nabla f(x) = Qx + q$$

iar matricea Hessiană în punctul x este:

$$\nabla^2 f(x) = Q.$$

O funcție diferențiabilă cel puțin o dată se numește *funcție netedă* (*smooth*). O funcție diferențiabilă de k ori, cu derivata de ordinul k continuă aparține clasei de funcții \mathcal{C}^k .

Pentru o funcție diferențiabilă $g : \mathbb{R} \rightarrow \mathbb{R}$, avem aproximarea Taylor de ordinul I exprimată în termeni de valoare medie sau integrală:

$$g(b) - g(a) = g'(\alpha)(b - a) = \int_a^b g'(\tau) d\tau,$$

pentru un anumit $\alpha \in [a, b]$. Aceste egalități pot fi extinse la orice funcție diferențiabilă $f : \mathbb{R}^n \rightarrow \mathbb{R}$ folosind relațiile precedente adaptate pentru funcția $g(t) = f(x + t(y - x))$ și utilizând regulile de diferențiere:

$$g'(\tau) = \langle \nabla f(x + \tau(y - x)), y - x \rangle$$

și deci pentru orice $x, y \in \text{dom} f$:

$$f(y) = f(x) + \langle \nabla f(x + \alpha(y - x)), y - x \rangle \quad \alpha \in [0, 1]$$

$$f(y) = f(x) + \int_0^1 \langle \nabla f(x + \tau(y - x)), y - x \rangle d\tau.$$

Următoarele extensii sunt posibile:

$$\nabla f(y) = \nabla f(x) + \int_0^1 \langle \nabla^2 f(x + \tau(y - x)), y - x \rangle d\tau$$

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}(y - x)^T \nabla^2 f(x + \alpha(y - x))(y - x)$$

pentru un $\alpha \in [0, 1]$. O funcție diferențiabilă $f : \mathbb{R}^n \rightarrow \mathbb{R}$ are *gradient Lipschitz continuu* dacă există o constantă $L > 0$ astfel încât

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y \in \text{dom} f.$$

Folosind aproximarea Taylor se obține următorul rezultat:

Lema A.2.1 (i) O funcție diferențiabilă de două ori $f : \mathbb{R}^n \rightarrow \mathbb{R}$ are gradient Lipschitz continuu dacă și numai dacă următoarea inegalitate are loc:

$$\|\nabla^2 f(x)\| \leq L \quad \forall x \in \text{dom} f.$$

(ii) Dacă o funcție diferențiabilă f are gradientul Lipschitz continuu, atunci următoarea inegalitate are loc:

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2} \|y - x\|^2 \quad \forall x, y \in \text{dom} f.$$

Din Lema A.2.1 rezultă că funcțiile diferențiabile cu gradient Lipschitz continuu sunt mărginite superior de o funcție pătratică, cu formă specială careia îi corespunde o matrice Hessiană $L \cdot I_n$:

$$f(y) \leq \frac{L}{2} \|y - x\|^2 + \langle \nabla f(x), y - x \rangle + f(x) \quad \forall x, y \in \text{dom} f.$$

Notăm cu $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$ clasa de funcții diferențiabile, convexe, cu gradient Lipschitz continuu. Pentru o funcție f din această clasă, următoarea inegalitate are loc:

$$\frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle \quad \forall x, y \in \text{dom} f.$$

O funcție diferențiabilă de două ori are Hessiana Lipschitz continuă dacă există o constantă $M > 0$ astfel încât:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq M \|x - y\| \quad \forall x, y \in \text{dom} f.$$

Pentru această clasă de funcții avem următoarea caracterizare:

Lema A.2.2 Pentru o funcție diferențiabilă de două ori $f : \mathbb{R}^n \rightarrow \mathbb{R}$ cu Hessiana Lipschitz continuă avem:

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x)\| \leq \frac{M}{2} \|y - x\|^2 \quad \forall x, y \in \text{dom} f.$$

Mai mult, următoarea inegalitate are loc:

$$-M \|x - y\| I_n \preceq \nabla^2 f(x) - \nabla^2 f(y) \preceq M \|x - y\| I_n \quad \forall x, y \in \text{dom} f.$$

Pentru o funcție $h: \mathbb{R}^n \rightarrow \mathbb{R}^p$, cu $h(x) = [h_1(x) \dots h_p(x)]^T$, notăm Jacobianul său prin $\nabla h(x)$, unde $\nabla h(x)$ este o matrice $p \times n$ cu elementul $\frac{\partial h_i(x)}{\partial x_j}$ pe poziția (i, j) :

$$\nabla h(x) = \begin{bmatrix} \frac{\partial h_1(x)}{\partial x_1} & \dots & \frac{\partial h_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_p(x)}{\partial x_1} & \dots & \frac{\partial h_p(x)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla h_1(x)^T \\ \vdots \\ \nabla h_p(x)^T \end{bmatrix}.$$

Teorema *funcțiilor implicite* se folosește des în optimizare și în alte domenii ale matematicii.

Teorema A.2.1 (Teorema funcțiilor implicite) Fie $F: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ o funcție continuă astfel încât:

- (i) $F(x^*, 0) = 0$ pentru un $x^* \in \mathbb{R}^n$;
- (ii) funcția F este de clasă C^1 într-o vecinătate a lui $(x^*, 0)$;
- (iii) $\nabla_x F(x, u)$ este inversabilă în punctul $(x, u) = (x^*, 0)$.

Atunci există o vecinătate \mathcal{N}_1 a lui x^* , o vecinătate \mathcal{N}_2 a lui 0 și o funcție continuă $\chi: \mathcal{N}_1 \rightarrow \mathcal{N}_2$ astfel încât $\chi(0) = x^*$ și $F(\chi(u), u) = 0$ pentru orice $u \in \mathcal{N}_2$. Mai mult, χ este definită în mod unic și dacă F este în clasa C^k pentru un $k > 0$, atunci și funcția implicită χ este în clasa C^k cu Jacobianul dat de expresia:

$$\nabla \chi(u) = -(\nabla_x F(\chi(u), u))^{-1} \nabla_u F(\chi(u), u).$$

Presentăm, de asemenea *teorema minimax* care are foarte multe aplicații în teoria jocurilor, dar după cum vom vedea se aplică și în teoria optimizării. Această teoremă a fost formulată și analizată de von Neumann în 1928 pentru funcții biliniare și apoi extinsă la funcții mai generale. Teorema tratează o clasă de probleme de optim care implică o combinație între maximizare și minimizare. Considerăm o funcție $F: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ și două mulțimi convexe $X \subseteq \mathbb{R}^n$ și $\Omega \subseteq \mathbb{R}^m$. Pentru orice $u \in \Omega$ putem considera minimumul funcției $F(u, x)$ pe $x \in X$ și apoi lua supremul acestui infimum ca funcție pe Ω , i.e.:

$$\sup_{u \in \Omega} \inf_{x \in X} F(u, x).$$

Pe de altă parte, putem considera și

$$\inf_{x \in X} \sup_{u \in \Omega} F(u, x).$$

Dacă valorile optime ale celor două probleme sunt egale, i.e. $\sup \inf$ și $\inf \sup$ sunt egale, atunci valoarea optimă comună se numește *valoarea minimax* sau *valoarea șa*. Se pune problema determinării de condiții când valoarea minimax există. Se poate arăta ușor că următoarea inegalitate are loc:

$$\sup_{u \in \Omega} \inf_{x \in X} F(u, x) \leq \inf_{x \in X} \sup_{u \in \Omega} F(u, x).$$

Se observă de asemenea că valoarea minimax este atinsă dacă există o pereche (u^*, x^*) astfel încât $(u^*, x^*) \in \Omega \times X$ și:

$$F(u, x^*) \leq F(u^*, x^*) \leq F(u^*, x) \quad \forall u \in \Omega, x \in X.$$

Numim o astfel de pereche (u^*, x^*) *punct șa*.

Teorema A.2.2 (Teorema minimax) *Fie Ω și X mulțimi convexe și cel puțin una din ele compactă și presupunem că funcția F este continuă și concavă în variabila u și convexă în variabila x . Atunci:*

$$\sup_{u \in \Omega} \inf_{x \in X} F(u, x) = \inf_{x \in X} \sup_{u \in \Omega} F(u, x).$$

Mulți algoritmi iterativi de optimizare pot fi scriși sub forma:

$$x_{k+1} = M(x_k) \quad \forall k \geq 0,$$

în care funcția $M : X \rightarrow X$ cu $X \subseteq \mathbb{R}^n$. Un vector $x^* \in X$ ce satisface condiția $x^* = M(x^*)$ se numește *punct fix* pentru M și pentru iterație. De asemenea, dacă există $0 \leq \beta < 1$ astfel încât operatorul M satisface condiția:

$$\|M(x) - M(y)\| \leq \beta \|x - y\| \quad \forall x, y \in X,$$

atunci acest operator se numește *contracție*. Următoarea teoremă furnizează proprietăți importante pentru o contracție:

Teorema A.2.3 (Teorema contracției) *Presupunem că M este o contracție și X este mulțime închisă. Atunci, există un punct fix unic $x^* \in X$ pentru M și pentru orice $x_0 \in X$ șirul generat de iterația $x_{k+1} = M(x_k)$ converge la punctul fix x^* . În particular, rata de convergență este liniară: $\|x_k - x^*\| \leq \beta^k \|x_0 - x^*\|$ pentru orice $k \geq 0$.*

Probleme rezolvate dar și multe exerciții propuse se găsesc în culegerea de probleme [1], care poate și trebuie să fie consultată pentru aprofundarea bazelor teoretice prezentate în lucrarea de față.

Bibliografie

- [1] I. Necoară, D. Clipici, and A. Pătraşcu. *Metode de Optimizare Numerică: Culegere de Probleme*. Editura Politehnica Press, 2013.
- [2] D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.
- [3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [4] B. Dumitrescu, C. Popeea, and B. Jora. *Metode de calcul numeric matriceal: algoritmi fundamentali*. Editura All, 1998.
- [5] A.V. Fiacco. *Introduction to sensitivity and stability analysis in nonlinear programming*. Academic Press, 1983.
- [6] T. Gal. *Postoptimal analyses, parametric programming, and related topics*. de Gruyter, 1995.
- [7] P.E. Gill, W. Murray, and M.H. Wright. *Practical Optimization*. Academic Press, 1981.
- [8] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.
- [9] D.G. Luenberger. *Linear and nonlinear programming*. Kluwer, 1994.
- [10] J. Moore and S. Wright. *Optimization Software Guide*. SIAM Studies in Applied Mathematics, 1993.
- [11] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer, 2004.
- [12] Y. Nesterov and A. Nemirovskii. *Interior Point Polynomial Algorithms in Convex Programming*. SIAM Studies in Applied Mathematics, 1994.

- [13] J. Nocedal and S. Wright. *Numerical Optimization*. Springer Verlag, 2006.
- [14] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
- [15] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 1976.
- [16] O. Stanasila. *Analiza matematică*. Editura Didactică și Pedagogică, 1981.
- [17] G. Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 1993.

