Optimizare practică
(suport de curs şi seminar – versiune în lucru)

Marius Durea

Optimizare practică

Desfăşurarea activităţii în anul universitar 2026–2027

În fiecare săptămână, timp de 14 săptămâni, vor fi 2 ore de curs şi 1 oră sau 2 ore, conform anului de studiu, de seminar/laborator. Cursul şi seminarul/laboratorul vor urmări, preponderent, materialul corespunzător postat la adresa:

$www.math.uaic.ro/\tilde{}durea$ – secţiunea teaching

Evaluarea se desfăşoară după procedura de mai jos.

- În timpul semestrului va avea loc evaluarea continuă (EC) care va avea ponderea de 50% din nota finală şi va cuprinde două componente:
  - un examen parţial (EP): în săptămâna 7, 8 sau 9, studenţii vor susţine un examen parţial scris (cu durata de 100 de minute), iar nota obţinută va avea ponderea de 50% din EC; EP nu se poate reface şi nu se poate susţine în sesiune; exclusiv pe baza unor motive bine întemeiate, EP se poate susţine la o dată ulterioară în timpul semestrului;
  - activitatea de seminar (AS): activitatea şi calitatea răspunsurilor din timpul desfăşurării seminarelor vor genera o notă ce va avea o pondere de 50% din EC.
- În sesiune va avea loc evaluarea finală (EF) care va avea o pondere de 50% din nota finală şi care constă dintr-o probă scrisă de jumătate de oră şi o probă practică (la calculator) de o oră. Proba scrisă valorează 50% din EF iar cea practică de asemenea 50%. O condiţie necesară pentru promovare este ca nota de la EF să fie minim 4,5. Odată ce această condiţie este îndeplinită se va calcula nota finală, care va fi rotunjirea la cel mai apropiat număr natural din intervalul $[1, 10]$ a valorii

$$\frac{1}{2}EF + \frac{1}{2}EC = \frac{1}{2}EF + \frac{1}{4}EP + \frac{1}{4}AS.$$

Prefaţă

Scopul acestei lucrări este acela de a prezenta unele completări şi dezvoltări ulterioare unui prim curs de bază de *Teoria optimizării neliniare*. Avem în vedere o aprofundare a condiţiilor de optimalitate pentru probleme cu restricţii funcţionale şi, mai ales, descrierea câtorva algoritmi pentru aproximarea punctelor de extrem în probleme de optimizare. De asemenea, un accent special este pus pe aspectele de ordin practic ale rezultatelor şi tehnicilor prezentate. Pentru implementările propuse se pot folosi coduri Matlab/Octave.

Materialul prezentat are două părţi principale. În prima dintre acestea, facem un studiu detaliat al condiţiilor necesare şi suficiente de ordinul întâi şi de ordinul al doilea pentru probleme de optimizare cu restricţii şi descriem mai multe aplicaţii mai puţin imediate ale acestora. Motivaţia unui studiu care poate părea prea teoretic într-un curs care are cuvântul „practic" în titlu este oferită de dictonul *O teorie corectă este cel mai practic lucru care există* (F.W. Dörpfeld, 1873).

În a doua parte, motivată de faptul că nu toate problemele de optimizare pot fi complet rezolvate, din lipsa instrumentelor necesare rezolvării unor ecuaţii neliniare, studiem algoritmi de optimizare şi algoritmi de determinare a rădăcinilor unor funcţii, din nou pe baza unor fundamente teoretice solide. Astfel, din această perspectivă, practica este teorie aplicată. Un alt aspect des întâlnit, mai ales spre finalul celei de-a doua părţi, este caracterul euristic al unor metode numerice care au motivaţii teoretice solide, dar nu sunt întru totul justificate aici din punct de vedere matematic, mai ales din lipsa spaţiului. Pentru aceste metode existenţa teoriei „din spate" este uşor de acceptat, tocmai datorită faptului că metodele de bază sunt complet acoperite din acest punct de vedere.

Prin urmare, această lucrare îşi propune să prezinte rezultate şi tehnici deja consacrate, iar aranjarea materialului, selectarea exemplelor ilustrative, ordonarea codurilor utilizate reprezintă punctul de vedere al autorului asupra acestui domeniu. Monografii clasice ale subiectului tratat sau lucrări mai recente care oferă un punct de vedere modern corelat cu tendinţele curente ale teoriei optimizării sunt prezentate la final şi au constituit bazele pe care acest curs a fost construit. De asemenea, monografiile indicate la bibliografie conţin multe extensii şi completări semnificative ale chestiunilor ce se regăsesc în lucrarea de faţă.

Cuprins

Prefaţă **i**

1 Optimizare - exemple şi elemente recapitulative **1**
1.1 Exemple de probleme practice 1
1.1.1 Probleme fără restricţii 1
1.1.2 Probleme cu restricţii 2
1.2 Cadrul general şi rezultate de existenţă a soluţiilor – recapitulare 4
1.3 Condiţii de optimalitate – recapitulare şi completări 5

2 Elemente teoretice şi practice în optimizarea cu restricţii **15**
2.1 Teorema Karush-Kuhn-Tucker – recapitulare 15
2.2 Semnificaţia practică a multiplicatorilor Lagrange 20
2.3 O selecţie de probleme 24
2.4 Un exemplu, două întrebări 34
2.5 Condiţiile Fritz John 34
2.6 Condiţii de optimalitate de ordinul al doilea 40

3 Solvabilitate şi aproximare **47**
3.1 Solvabilitatea problemelor de optimizare 47
3.2 Principiul variaţional al lui Ekeland 48
3.3 Penalizare 51
3.4 Generalităţi privind metodele numerice 53
3.5 Algoritmi iterativi 59
3.6 Un algoritm reprezentativ: metoda lui Newton 64
3.7 Aplicaţie la funcţia W, a lui Lambert 72
3.8 O metodă cvasi-Newton: metoda secantei 81

4 Algoritmi pentru probleme de optimizare fără restricţii **87**
4.1 Algoritmi de ordin zero pentru funcţii de o variabilă 87
4.1.1 Delimitarea (bracketing) punctului de minim 88
4.1.2 Metoda secţiunii de aur 90
4.1.3 Metoda Fibonacci 92
4.2 Algoritmi de ordinul întâi (metode de tip gradient) 96
4.2.1 Descrierea metodei direcţiei de descreştere (gradient descent, GD) 96

4.2.2 Convergenţa metodei direcţiei de descreştere. Cazul general 99
4.2.3 Convergenţa metodei direcţiei de descreştere. Cazul convex 107
4.2.4 Convergenţa metodei direcţiei de descreştere. Cazul funcţiilor pătratice . . 113
4.2.5 Legătura dintre metoda gradient descent şi o ecuaţie diferenţială ordinară . 117
4.2.6 Posibilităţi de accelerare a metodei gradient descent 120
4.2.7 Algoritmi specifici *machine learning* bazaţi pe metoda direcţiei de descreştere 123
4.2.8 Metoda gradienţilor conjugaţi 127
4.3 Algoritmi de ordinul al doilea 132
4.3.1 Metoda lui Newton în optimizare 132
4.3.2 Metode derivate din metodele fundamentale 134
4.4 Cazul funcţiilor cu valori vectoriale: optimizare Pareto 139
4.5 Obţinerea unor algoritmi pentru probleme cu restricţii - o metodă de penalizare . 142

Modele de examene scrise **151**
Model examen parţial 151
Model examen final 152

Bibliografie **153**

Capitolul 1

Optimizare - exemple şi elemente recapitulative

1.1 Exemple de probleme practice

1.1.1 Probleme fără restricţii

Exemplul 1.1.1 Funcţia rezultată folosind Principiul lui Fermat pentru modelarea problemei refracţiei luminii care trece din punctul $(0, a)$ (cu $a > 0$) dintr-un mediu (aer) în punctul (b, c) (cu $b > 0$, $c < 0$) din alt mediu (apă) este $f : \mathbb{R} \to \mathbb{R}$,

$$f(x) = \frac{\sqrt{a^2 + x^2}}{v_1} + \frac{\sqrt{(b-x)^2 + c^2}}{v_2},$$

unde v_1 este viteza luminii în primul mediu, iar v_2 este viteza luminii în al doilea mediu. Această funcţie trebuie minimizată.

Exemplu numeric (ce va fi studiat): $a = 10$; $(b, c) = (30, -40)$; $v_1 = 299702$ km/s, $v_2 = 228849$ km/s.

Exemplul 1.1.2 Modelarea problemei de mai jos conduce de asemenea la minimizarea unei funcţii. Un sportiv înoată într-un lac şi se află în larg la o distanţă de $5km$ faţă de mal. El trebuie să ajungă pe uscat într-un punct aflat la $7km$ în lungul ţărmului faţă de poziţia sa în raport cu ţărmul. Ştiind că sportivul aleargă cu $6m/s$ şi înoată cu $2m/s$, să se determine timpul minim în care poate ajunge la destinaţie.

Exerciţiul 1.1 *Să se modeleze problema de mai sus.*

Exemplul 1.1.3 O bacterie are forma unei elipse de dimensiuni $a, b > 0$, $a > b$. Coeficientul de frecare cu lichidul în care se află este dat de

$$\tau = \frac{4\pi\rho a}{\ln\left(\frac{2a}{b}\right) - \frac{1}{2}}.$$

Luăm $b = 1$ şi $\rho = 1$ (apă). Se cere determinarea lui a astfel încât τ să fie minim.

Exerciţiul 1.2 *Să se rezolve problema de mai sus.*

Exemplul 1.1.4 Un conducător al unui oras-stat italian din secolul 14 trebuie să angajeze o armată. Există două tipuri de soldaţi ce pot fi angajaţi: mercenari sau cetăţeni aleşi prin tragere la sorţi. Fie $c \in [0,1]$ proporţia de cetăţeni (şi, evident, $1-c$ proporţia de mercenari). Fiecare dintre cele două tipuri de soldaţi comportă un risc. Mercenarii sunt oportunişti şi pot specula unele momente pentru a obţine profituri suplimentare, ceea ce diminuează avuţia conducătorului, notată cu w. Notăm cu $v \in [0,1]$ "gradul de oportunism" al unui (fiecărui) mercenar. Pe de altă parte, soldaţii cetăţeni pot avea opţiuni politice în favoarea sau împotriva conducătorului, ceea ce ar putea conduce la îndepărtarea acestuia de la putere pe calea armelor. Dacă notăm cu $g \in [0,1]$ proporţia de cetăţeni favorabili din punct de vedere politic conducătorului, gradul de adversitate va fi $c(1-g)$. Pentru a stabili ce proporţie din fiecare categorie angajează, conducătorul are de rezolvat aşadar problema

$$\max_{c\in[0,1]} (1-c(1-g))(w-wv(1-c)),$$

adică

$$\max_{c\in[0,1]} (1-c(1-g))(1-v(1-c)).$$

Exerciţiul 1.3 *Să se rezolve problema de mai sus şi să se interpreteze rezultatele.*

1.1.2 Probleme cu restricţii

Exemplul 1.1.5 (problema dietei) Fie x_i cantitatea de produs de alimentaţie de tipul i. Pentru fiecare tip de produs costul pe unitate este c_i. Se doreşte minimizarea costului total

$$\sum_i c_i x_i$$

pentru asigurarea unei alimentaţii potrivite. Astfel, notăm cu a_{ij} conţinutul de nutrient j dintr-o unitate de produs i şi b_j cantitatea necesară din acest nutrient. Deci avem restricţiile

$$\sum_i a_{ij} x_i \geq b_j, \ \forall j.$$

Evident, trebuie să avem şi $x_i \geq 0$ pentru orice i.

Exemplul 1.1.6 Avem un sistem de eşafodaj cu patru fire A, B, C, D care suportă respectiv tensiunile 200, 100, 200, 100 construit astfel:

- de un plafon sunt prinse trei fire, A, C, D, în acestă ordine, cu distanţele dintre ele de 2 şi 10 m;
- la capetele firelor C, D, aflate la acelaşi nivel se află o platformă de masă neglijabilă;
- de această platformă la 2 m de D înspre C este prins un fir B care coboară până la acelaşi nivel cu firul A;
- la capetele firelor A, B se află o platformă de masă neglijabilă.
- pe cele două platforme se aşează, la mijloc, câte un corp cu greutăţile x_1 şi respectiv x_2.

Problema este de a determina masa maximă ce poate fi suportată de sistem, deci trebuie să maximizăm $(x_1 + x_2)$ ţinând cont de ecuaţiile de echilibru. Dacă notăm cu T_A, T_B, T_C, T_D tensiunile din cele patru fire, ecuaţiile devin

$$\begin{aligned} x_2 &= T_A + T_B \\ x_1 + T_B &= T_C + T_D \\ 5x_2 &= 10T_B \\ 5x_1 + 8T_B &= 10T_D. \end{aligned}$$

Modelarea problemei conduce deci la

$$\max (x_1 + x_2)$$

cu

$$\begin{aligned} x_2 &= T_A + T_B \\ x_1 + T_B &= T_C + T_D \\ 5x_2 &= 10T_B \\ 5x_1 + 8T_B &= 10T_D, \\ T_A \leq 200,\ T_B &\leq 100, T_C \leq 200,\ T_D \leq 100. \end{aligned}$$

Se observă că tensiunile pot fi eliminate pentru că

$$\begin{aligned} T_A = T_B &= \frac{x_2}{2} \\ T_C &= \frac{x_2}{10} + \frac{x_1}{2} \\ T_D &= \frac{2x_2}{5} + \frac{x_1}{2} \end{aligned}$$

şi se ajunge la problema

$$\max (x_1 + x_2)$$

cu

$$\begin{aligned} x_2 &\leq 200 \\ x_1 &\leq 400 \\ 5x_1 + x_2 &\leq 2000 \\ 5x_1 + 4x_2 &\leq 1000. \end{aligned}$$

Mai mult, unele din aceste restricţii le includ pe altele (de exemplu, ultima pe penultima) pentru că $x_1, x_2 \geq 0$ şi putem să reducem problema la

$$\max (x_1 + x_2)$$

cu

$$\begin{aligned} 0 \leq x_2 &\leq 200 \\ 0 &\leq x_1 \\ 5x_1 + 4x_2 &\leq 1000. \end{aligned}$$

Exerciţiul 1.4 *Să se rezolve problema de mai sus.*

Exerciţiul 1.5 *Se doreşte construirea unei cutii paralelipipedice cu baza pătrat şi fără capac cu volumul de 100. Determinaţi dimensiunile unei astfel de cutii astfel încât materialul folosit să fie minim.*

Exerciţiul 1.6 *O fabrică doreşte să vândă n exemplare dintr-un produs astfel încât profitul să fie maxim. Cercetările de piaţă arată că la un preţ de 1,5 u.m./exemplar se vând 5000 de exemplare, iar pentru fiecare reducere de 0,1 u.m./exemplar se vor vinde în plus 1000 de exemplare. Presupunem că există nişte costuri de start a producţiei de 2000 u.m., iar costul de producţie al fiecărui exemplar este 0.5 u.m. Să se modeleze problema şi să se determine profitul maxim, preţul de vânzare şi n.*

Exerciţiul 1.7 *Determinaţi dreptunghiul cu aria cea mai mare care poate fi inclus în epigraful funcţiei $x \mapsto x^2$ şi care se află sub linia de nivel $y = a$, unde $a > 0$.*

Exerciţiul 1.8 *Într-o sferă dată se include cel mai mare con circular drept posibil. Să se determine raportul volumelor celor două corpuri.*

1.2 Cadrul general şi rezultate de existenţă a soluţiilor – recapitulare

Fie $U \subset \mathbb{R}^p$ o mulţime nevidă şi deschisă, $f : U \to \mathbb{R}$ o funcţie şi $M \subset U$ o mulţime nevidă. Suntem interesaţi să studiem problema minimizării funcţiei f atunci când argumentul acesteia parcurge mulţimea M. Formal, vom scrie această problemă în forma următoare:

$$(P)\ \min f(x),\ x \in M.$$

Mulţimea M se va numi mulţimea punctelor fezabile ale problemei (P) sau mulţimea constrângerilor sau, încă, mulţimea restricţiilor.

Să spunem de la început că ne vom ocupa de minimizarea funcţiei f, dar rezultate referitoare la maximizarea sa pot fi obţinute aplicând rezultatele de minimizare funcţiei $-f$ în baza relaţiei $\max f = -\min(-f)$. Pentru început, să definim noţiunea de soluţie asociată problemei (P).

Definiţia 1.2.1 *Spunem că $\overline{x} \in M$ este soluţie locală (sau, simplu, soluţie) pentru problema (P) sau punct de minim pentru funcţia f pe mulţimea M dacă există o vecinătate V a punctului $\overline{x}$ astfel încât $f(\overline{x}) \leq f(x)$ pentru orice $x \in M \cap V$.*

Dacă $V = \mathbb{R}^p$, se mai spune că $\overline{x}$ este soluţie globală pentru problema (P) sau punct de minim global pentru f pe M.

Să remarcăm că $\overline{x} \in M$ nu este punct de minim pentru funcţia f pe mulţimea M dacă şi numai dacă există un şir $(x_n) \subset M$, $x_n \to \overline{x}$ astfel încât $f(x_n) < f(\overline{x})$ pentru orice n.

Observaţia 1.2.2 *În general, vom distinge două mari situaţii pentru studiul problemei (P) : cazul în care $M = U$ şi cazul în care M este intersecţia dintre o mulţime închisă a lui $\mathbb{R}^p$ şi U. În primul*

caz vom spune că problema de optimizare (P) este fără restricţii (sau constrângeri) în timp ce în al doilea caz vorbim despre o problemă cu restricţii (constrângeri). Să mai observăm că în Definiţia 1.2.1, dacă $\overline{x} \in \operatorname{int} M$, atunci $\overline{x}$ este soluţie locală şi a problemei fără restricţii (este suficient să micşorăm vecinătatea V astfel încât $V \subset M$). Deci în cazul problemelor cu restricţii cazul interesant (care nu se reduce la cazul unei probleme fără restricţii) este cazul în care $\overline{x} \in \operatorname{Fr} M$. Dacă $\overline{x} \in \operatorname{int} M$ mai spunem şi că restricţia este inactivă.

Un rol important îl vor juca mulţimile de nivel ale unei funcţii. Având o funcţie $f : \mathbb{R}^p \to \mathbb{R}$, dacă $\nu \in \mathbb{R}$,

$$N_\nu f := \{x \in \mathbb{R}^p \mid f(x) \leq \nu\} = f^{-1}((-\infty, \nu]).$$

Să observăm că dacă $\nu > \inf_{x\in\mathbb{R}^p} f(x)$ atunci $N_\nu f \neq \emptyset$ iar dacă f este continuă atunci $N_\nu f$ este închisă.

Teorema 1.2.3 (Teorema lui Weierstrass) *Dacă $M \subset \mathbb{R}^p$ este o mulţime compactă şi $f : M \to \mathbb{R}$ este o funcţie continuă, atunci problema minimizării lui f pe M şi problema maximizării lui f pe M au soluţii globale.*

Teorema 1.2.4 *Fie $f : \mathbb{R}^p \to \mathbb{R}$ o funcţie continuă şi $M \subset \mathbb{R}^p$ o mulţime nevidă şi închisă. Dacă există $\nu > \inf_{x\in M} f(x)$ astfel încât mulţimea de nivel a lui f relativ la M, adică $M \cap N_\nu f = \{x \in M \mid f(x) \leq \nu\}$, este mărginită, atunci f îşi atinge minimul global pe M.*

Rezultatele de mai sus asigură existenţa punctelor de minim în condiţii de compactitate a mulţimilor de nivel. Evident, mărginirea inferioară a funcţiei este condiţie necesară pentru existenţa minimului, dar este clar că mărginirea mulţimilor de nivel nu este o astfel de condiţie. De exemplu funcţia $f : \mathbb{R} \to \mathbb{R}$, $f(x) = (x-1)^2 e^{-x}$ îşi atinge minimul (global) în $\overline{x} = 1$, valoarea minimă este 0, dar $N_\nu f$ nu este mărginită pentru nicio valoare $\nu > 0 = \inf\{f(x) \mid x \in \mathbb{R}\}$.

Evident, în teorema de mai sus, dacă M este mărginită, atunci ipoteza este automat verificată. Cazul interesant este acela în care M este nemărginită, situaţie în care ipoteza este verificată dacă impunem funcţiei f o condiţie de coercivitate.

Propoziţia 1.2.5 *Fie $f : \mathbb{R}^p \to \mathbb{R}$ o funcţie şi $M \subset \mathbb{R}^p$ o mulţime nevidă, închisă şi nemărginită. Dacă $\lim_{x\in M, \|x\|\to\infty} f(x) = \infty$ (i.e., pentru orice $(x_n) \subset M, \|x_n\| \to \infty$ are loc $f(x_n) \to \infty$) atunci mulţimea $N_\nu f \cap M$ este mărginită pentru orice $\nu > \inf_{x\in M} f(x)$.*

1.3 Condiţii de optimalitate – recapitulare şi completări

Definiţia 1.3.1 *Fie $M \subset \mathbb{R}^p$ o mulţime nevidă şi $\overline{x} \in \operatorname{cl} M$. Un vector $u \in \mathbb{R}^p$ se numeşte tangent la mulţimea M în $\overline{x}$ dacă există $(t_n) \subset (0, \infty), t_n \to 0$ şi $(u_n) \to u$ astfel încât pentru orice $n \in \mathbb{N}$,*

$$\overline{x} + t_n u_n \in M.$$

Evident este suficient ca relaţia de mai sus să aibă loc pentru orice $n \in \mathbb{N}$ de la un loc încolo (pentru n suficient de mare).

Teorema 1.3.2 *Mulţimea, notată $T(M, \overline{x})$, a tuturor vectorilor tangenţi la M în $\overline{x}$ este un con închis, numit conul tangent (în sens Bouligand) la M în $\overline{x}$.*

Propoziţia 1.3.3 *Dacă* $\emptyset \neq M \subset \mathbb{R}^p$ *şi* $\overline{x} \in \operatorname{cl} M$, *atunci are loc relaţia* $T(M,\overline{x}) = T(\overline{M},\overline{x})$. *Dacă* $\overline{x} \in \operatorname{int} M$, *atunci* $T(M,\overline{x}) = \mathbb{R}^p$.

Vom nota cu $N(M,\overline{x})$ polara lui $T(M,\overline{x})$ (i.e. $N(M,\overline{x}) := T(M,\overline{x})^-$) şi vom numi acest con conul normal la M în $\overline{x}$.

Dacă mulţimea M este convexă, atunci conul Bouligand are următoarea formă specială.

Propoziţia 1.3.4 *Fie* $\emptyset \neq M \subset \mathbb{R}^p$ *convexă şi* $\overline{x} \in M$. *Atunci*

$$T(M,\overline{x}) = \operatorname{cl} \mathbb{R}_+(M - \overline{x}),$$

iar

$$N(M,\overline{x}) = \{u \in \mathbb{R}^p \mid \langle u, c - \overline{x}\rangle \leq 0, \forall c \in M\}.$$

Teorema 1.3.5 (Condiţia necesară de ordinul I) *Dacă* $\overline{x}$ *este soluţie locală pentru problema* (P) *şi* f *este diferenţiabilă în* $\overline{x}$, *atunci* $\nabla f(\overline{x})(u) \geq 0$ *pentru orice* $u \in T(M,\overline{x})$.

Observaţia 1.3.6 *Condiţia din concluzia teoremei precedente se mai scrie, echivalent,*

$$-\nabla f(\overline{x}) \in N(M,\overline{x}).$$

Observaţia 1.3.7 *Ţinând cont de Propoziţia 1.3.3, dacă în teorema de mai sus,* $\overline{x} \in \operatorname{int} M$ *(restricţie inactivă), obţinem* $\nabla f(\overline{x})(u) \geq 0$ *pentru orice* $u \in \mathbb{R}^p$. *Din liniaritatea lui* $\nabla f(\overline{x})$, *deducem* $\nabla f(\overline{x}) = 0$, *adică Teorema lui Fermat.*

Exemplul 1.3.8 (restricţii geometrice) Fie $f : \mathbb{R}^2 \to \mathbb{R}$,

$$f(x) = -x_1 - 2x_2 - 2x_1x_2 + \frac{x_1^2}{2} + \frac{x_2^2}{2}$$

şi mulţimea restricţiilor

$$M := \left\{x \in \mathbb{R}^2 \mid x_1 + x_2 \leq 1,\ x_1 \geq 0,\ x_2 \geq 0\right\}.$$

Să considerăm, ca în discuţia teoretică, problema (P) a minimizării lui f pe M.

Funcţia f este continuă, iar M este compactă, deci problema (P) admite cel puţin o soluţie globală.

Datorită faptului că inegalităţile ce definesc mulţimea M sunt liniare, avem o imagine geometrică clară a mulţimii (un triunghi dreptunghic cu vârfurile în punctele $(0,0)$, $(0,1)$, $(1,0)$).

Dacă ar exista un punct de minim $\overline{x}$ în interiorul lui M atunci acel punct ar fi un minim fără restricţii (conform Observaţiei 1.2.2), deci, din Teorema lui Fermat, $\nabla f(\overline{x}) = 0$. Dar

$$\nabla f(\overline{x}) = (-1 + \overline{x}_1 - 2\overline{x}_2, -2 - 2\overline{x}_1 + \overline{x}_2)$$

şi rezolvând sistemul găsim soluţia $\overline{x} = (-\frac{5}{3}, -\frac{4}{3})$ care nu aparţine lui M. Deci problema nu are soluţii în $\operatorname{int} M$.

Observăm că putem calcula relativ simplu conurile tangent şi normal la M în punctele de pe frontieră şi apoi să verificăm îndeplinirea condiţiei necesare de optimalitate: $-\nabla f(\overline{x}) \in N(M,\overline{x})$ (Teorema 1.3.5).

Astfel, dacă punctul $\overline{x}$ este

– pe segmentul deschis de capete $(0,1)$, $(1,0)$:

$$T(M,\overline{x}) = \{u \in \mathbb{R}^2 \mid u_1 + u_2 \leq 0\};\ N(M,\overline{x}) = \mathbb{R}_+\{(1,1)\};$$

– pe segmentul deschis de capete $(0,0)$, $(0,1)$:

$$T(M,\overline{x}) = \{u \in \mathbb{R}^2 \mid u_1 \geq 0\};\ N(M,\overline{x}) = \mathbb{R}_+\{(-1,0)\};$$

– pe segmentul deschis de capete $(0,0)$, $(1,0)$:

$$T(M,\overline{x}) = \{u \in \mathbb{R}^2 \mid u_2 \geq 0\};\ N(M,\overline{x}) = \mathbb{R}_+\{(0,-1)\};$$

– punctul $(0,1)$:

$$\begin{aligned} T(M,\overline{x}) &= \{u \in \mathbb{R}^2 \mid u_1 + u_2 \leq 0,\ u_1 \geq 0\} \\ N(M,\overline{x}) &= \{a(1,1) + b(-1,0) \mid a,b \geq 0\}; \end{aligned}$$

– punctul $(1,0)$:

$$\begin{aligned} T(M,\overline{x}) &= \{u \in \mathbb{R}^2 \mid u_1 + u_2 \leq 0,\ u_2 \geq 0\} \\ N(M,\overline{x}) &= \{a(1,1) + b(0,-1) \mid a,b \geq 0\}; \end{aligned}$$

– punctul $(0,0)$:

$$\begin{aligned} T(M,\overline{x}) &= \{u \in \mathbb{R}^2 \mid u_1 \geq 0,\ u_2 \geq 0\}; \\ N(M,\overline{x}) &= \{a(-1,0) + b(0,-1) \mid a,b \geq 0\}. \end{aligned}$$

Prin calcul direct, se verifică faptul că un singur punct satisface condiţia necesară de optimalitate: $\overline{x} = \left(\frac{1}{3}, \frac{2}{3}\right)$. Prin urmare, conform discuţiei precedente, acesta este singurul punct de minim al problemei.

Putem, de asemenea, să ne punem problema maximului global al lui f pe M (a cărui existenţă este asigurată de Teorema lui Weierstrass), problemă echivalentă cu găsirea minimului lui $-f$ pe M. Repetând discuţia precedentă, găsim două puncte ce verifică condiţia necesară de optimalitate (i.e. $\nabla f(\overline{x}) \in N(M,\overline{x})$): $\overline{x} = (0,0)$ şi $\overline{x} = (1,0)$. Dar $f(0,0) = 0$, iar $f(1,0) = -2^{-1}$, deci $(0,0)$ este punctul de maxim.

Exemplul 1.3.9 Să calculăm conurile tangente şi conurile normale în diferite puncte la mulţimea $M \subset \mathbb{R}^p$,

$$M = \left\{ x = (x_1, x_2, ..., x_p) \in \mathbb{R}^p \mid x_i \geq 0,\ \forall i \in \overline{1,p},\ \sum_{i=1}^{p} x_i = 1 \right\},$$

numită simplexul unitate. Această mulţime este, evident, convexă şi închisă. Conform celor de mai sus, pentru fiecare $\overline{x} \in M$,

$$\begin{aligned} T(M,\overline{x}) &= \operatorname{cl} \mathbb{R}_+(M - \overline{x}) \\ &= \operatorname{cl} \{u \in \mathbb{R}^p \mid \exists \alpha \geq 0,\ x \in M,\ u = \alpha(x - \overline{x})\}. \end{aligned}$$

Fie u din mulţimea din membrul drept. Este clar că, pe de o parte, $\sum_{i=1}^{p} u_i = 0$ iar, pe de altă parte, dacă $\overline{x}_i = 0$, atunci $u_i \geq 0$. Notăm cu $I(\overline{x}) := \left\{ i \in \overline{1,p} \mid \overline{x}_i = 0 \right\}$. Deducem că

$$T(M,\overline{x}) \subset \left\{ u \in \mathbb{R}^p \mid \sum_{i=1}^{p} u_i = 0 \text{ şi } u_i \geq 0,\ \forall i \in I(\overline{x}) \right\}.$$

Să demonstrăm incluziunea inversă. Este simplu de verificat că mulţimea din dreapta este închisă. Fie u din această mulţime. Dacă $u = 0$, atunci, evident, $u \in T(M,\overline{x})$. Dacă $u \neq 0$ atunci arătăm că există $\alpha > 0$ cu $\overline{x} + \alpha u \in M$. Pe de o parte, faptul că $\sum_{i=1}^{p}(\overline{x}_i + \alpha u_i) = 1$ este clar pentru orice α. Dacă nu există indici i cu $u_i < 0$, atunci şi faptul că $\overline{x}_i + \alpha u_i \geq 0$, pentru orice $i \in \overline{1,p}$ este evident şi deci $u \in T(M,\overline{x})$. Presupunem acum că mulţimea J a indicilor pentru care $u_j < 0$ este nevidă. Atunci $J \subset \overline{1,p} \setminus I(\overline{x})$, deci $\overline{x}_j > 0$ pentru orice $j \in J$. Alegem atunci α strict pozitiv cu

$$\alpha < \min\{-u_j^{-1}\overline{x}_j \mid j \in J\}$$

şi avem din nou că $\overline{x}_i + \alpha u_i \geq 0$, pentru orice $i \in \overline{1,p}$. Prin urmare şi în acest caz $u \in T(M,\overline{x})$, deci are loc egalitatea.

În continuare, arătăm că

$$\begin{aligned} N(M,\overline{x}) = &\{(a,a,...,a) \in \mathbb{R}^p \mid a \in \mathbb{R}\} \\ &+ \left\{ v \in \mathbb{R}^p \mid v_i \leq 0,\ \forall i \in I(\overline{x}),\ v_i = 0,\ i \notin I(\overline{x}) \right\}. \end{aligned}$$

Pentru aceasta considerăm elementele

$$a_0 = (1,1,...,1),\ a_1 = -(1,0,...,0),\ ...,\ a_n = -(0,0,...,1)$$

şi observăm următoarea scriere echivalentă a lui $T(M,\overline{x})$:

$$T(M,\overline{x}) = \left\{ u \in \mathbb{R}^p \mid \langle a_0,u\rangle \leq 0,\ \langle -a_0,u\rangle \leq 0,\ \langle a_i,u\rangle \leq 0,\ \forall i \in I(\overline{x}) \right\}.$$

Polara acestei mulţimi este

$$N(M,\overline{x}) = \left\{ \alpha a_0 - \beta a_0 + \sum_{i \in I(\overline{x})} \alpha_i a_i \mid \alpha,\beta,\alpha_i \geq 0,\ \forall i \in I(\overline{x}) \right\}.$$

Într-adevăr, faptul că mulţimea din dreapta este inclusă în conul normal este evident, iar incluziunea inversă rezultă din Lema lui Farkas, întrucât:

$$\begin{aligned} v \in N(M,\overline{x}) &\iff \langle u,v\rangle \leq 0, \forall u \in T(M,\overline{x}) \\ &\iff \forall u \in \mathbb{R}^p : [\langle a_0,u\rangle \leq 0,\ \langle -a_0,u\rangle \leq 0,\ \langle a_i,u\rangle \leq 0,\ \forall i \in I(\overline{x})] \Rightarrow \langle u,v\rangle \leq 0. \end{aligned}$$

Astfel, se ajunge la forma anunţată a conului normal.

Exemplul 1.3.10 (optimizare pe simplexul unitate) Fie $f : \mathbb{R}^p \to \mathbb{R}$ convexă şi diferenţiabilă. Dacă dorim să minimizăm această funcţie pe simplexul unitate (a se vedea Exemplul 1.3.9),

atunci, conform Propoziţiei 1.3.15, $\overline{x} \in M$ este punct de minim pentru f pe M dacă şi numai dacă $-\nabla f(\overline{x}) \in N(M, \overline{x})$. Din forma particulară a lui $N(M, \overline{x})$, deducem că această condiţie se scrie

$$\frac{\partial f}{\partial x_i}(\overline{x}) = c, \text{ (constantă)}, \forall i \notin I(\overline{x})$$
$$\frac{\partial f}{\partial x_i}(\overline{x}) \geq c, \ \forall i \in I(\overline{x}).$$

Fie $n_1, ..., n_p \in \mathbb{N}^*$, $\sum_{i=1}^p n_i = N > 0$ şi fie $f : \mathbb{R}^p \to \mathbb{R}$, $f(x) = -x_1^{n_1} x_2^{n_2} ... x_p^{n_p}$. Să minimizăm această funcţie pe simplexul unitate din $\mathbb{R}^p$. Evident că problema are soluţie, f fiind continuă, iar M compactă. Cum f este nulă dacă măcar una dintre componentele argumentului este zero, soluţiile se vor găsi în mulţimea

$$\left\{ x \in \mathbb{R}^p \mid x_i > 0, \ \forall i \in \overline{1,p}, \ \sum_{i=1}^p x_i = 1 \right\}.$$

Cu notaţiile din Exemplul 1.3.9, aceasta înseamnă $I(\overline{x}) = \emptyset$. Mai întâi, condiţia necesară de optimalitate $-\nabla f(\overline{x}) \in N(M, \overline{x})$, se scrie, ţinând cont de expresia conului normal (Exemplul 1.3.9)

$$\frac{n_i}{\overline{x}_i} f(\overline{x}) = c, \text{ constantă}, \forall i \in \overline{1,p},$$

adică

$$\frac{n_i}{\overline{x}_i} = c', \text{ constantă}, \forall i \in \overline{1,p}.$$

Cum $\sum_{i=1}^p \overline{x}_i = 1$ şi $\sum_{i=1}^p n_i = N$, găsim

$$\overline{x}_i = \frac{n_i}{N}, \ \forall i \in \overline{1,p}.$$

Cum problema are soluţie şi un singur punct verifică condiţia necesară, deducem că acel punct este soluţia căutată.

Observaţia 1.3.11 *O altă abordare în cazul Exemplelor 1.3.8 şi 1.3.10 constă în transformarea restricţiilor geometrice în restricţii funcţionale şi aplicarea Teoremei Karush-Kuhn-Tucker pe care o vom reaminti într-un capitol ulterior.*

O prima întrebare este dacă reciproca Teoremei 1.3.5 este adevărată. Răspunsul este negativ. Astfel, este suficient să considerăm acelaşi exemplul funcţiei $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^3$ şi $\overline{x} = 0$.

Se pot însă impune condiţii suplimentare aşa încât să avem unele echivalenţe şi cel mai important caz este cel al funcţiilor convexe.

Teorema 1.3.12 *Fie $U \subset \mathbb{R}^p$ o mulţime convexă şi deschisă şi fie $f : U \to \mathbb{R}$ o funcţie convexă şi diferenţiabilă pe U. Următoarele afirmaţii sunt echivalente:*

(i) $\overline{x}$ este un minim global al lui f;

(ii) $\overline{x}$ este un minim local al lui f;

(iii) $\overline{x}$ este punct critic al lui f (i.e. $\nabla f(\overline{x}) = 0$).

Aşa cum se observă, în cazul funcţiilor convexe, condiţia de ordinul I (în cazul fără constrângeri) este necesară şi suficientă pentru optimalitate.

Privitor la natura punctelor de extrem pentru o funcţie convexă, dăm rezultatele de mai jos.

Propoziţia 1.3.13 *Fie $M \subset \mathbb{R}^p$ o mulţime convexă şi fie $f : M \to \mathbb{R}$ o funcţie convexă. Dacă $\overline{x} \in M$ este punct de minim local pentru f pe M atunci $\overline{x}$ este punct de minim global pentru f pe M.*

Propoziţia 1.3.14 *Fie $M \subset \mathbb{R}^p$ o mulţime convexă şi fie $f : M \to \mathbb{R}$ o funcţie convexă. Mulţimea punctelor de minim ale lui f pe M este convexă. Dacă, în plus, f este strict convexă, atunci această mulţime are cel mult un element.*

În cazul cu restricţii, pentru funcţiile convexe condiţia necesară de optimalitate de ordinul I se scrie într-o formă specială şi, în plus, este şi condiţie suficientă.

Propoziţia 1.3.15 *Fie $U \subset \mathbb{R}^p$ o mulţime convexă şi deschisă şi fie $f : U \to \mathbb{R}$ o funcţie convexă şi diferenţiabilă pe U. Fie $M \subset U$ convexă. Elementul $\overline{x} \in M$ este punct de minim pentru f pe M dacă şi numai dacă*

$$-\nabla f(\overline{x}) \in N(M, \overline{x}).$$

Revenind la Teoremele 1.3.5 şi 1.3.18, întărind condiţiile din concluziile rezultatelor precedente, obţinem condiţii suficiente de optimalitate.

Rezultatele anunţate sunt următoarele.

Teorema 1.3.16 *Presupunem că f este diferenţiabilă în $\overline{x} \in M$ şi*

$$\nabla f(\overline{x})(u) > 0, \ \forall u \in T(M, \overline{x}) \setminus \{0\}.$$

Atunci $\overline{x}$ este soluţie locală pentru problema (P). □

Observaţia 1.3.17 *În Problema 1.3.23 de mai jos, vom vedea că se pot deduce mai multe concluzii în cadrul teoremei anterioare. Deocamdată, am fost interesaţi doar de precizarea unei condiţii suficiente de optimalitate.*

Prezentăm acum o cunoscută condiţie necesară de optimalitate de ordinul al doilea, pentru problema fără restricţii.

Teorema 1.3.18 (Condiţia necesară de ordinul al II-lea) *Dacă f este de clasă C^2, iar $\overline{x} \in \operatorname{int} M$ este punct de minim local pentru f, atunci $\nabla f(\overline{x}) = 0$ şi $\nabla^2 f(\overline{x})$ este pozitiv semidefinită (adică $\nabla^2 f(\overline{x})(u, u) \geq 0$ pentru orice $u \in \mathbb{R}^p$).*

Evident, cu o demonstraţie similară, se constată că dacă $\overline{x} \in \operatorname{int} M$ este punct de maxim local pentru f, atunci $\nabla f(\overline{x}) = 0$ şi $\nabla^2 f(\overline{x})$ este negativ semidefinită (adică $\nabla^2 f(\overline{x})(u, u) \leq 0$ pentru orice $u \in \mathbb{R}^p$). Se poate cu uşurinţă observa, din rezultatul de mai sus şi din demonstraţia sa, că dacă $\nabla^2 f(\overline{x})$ nu este nici pozitiv semidefinită nici negativ semidefinită (caz în care spunem că este nedefinită) atunci $\overline{x}$ nu este punct de extrem.

În privinţa condiţiilor suficiente de ordinul al doilea, are loc următorul rezultat.

Teorema 1.3.19 *Presupunem că f este de clasă C^2, $\nabla f(\overline{x}) = 0$ şi*

$$\nabla^2 f(\overline{x})(u,u) > 0,\ \forall u \in T(M,\overline{x}) \setminus \{0\}.$$

Atunci $\overline{x}$ este soluţie locală pentru problema (P).

În cazul fără restricţii, rezultatul de mai sus se reduce la următorul corolar.

Corolarul 1.3.20 *Fie f de clasă C^2. Dacă $\overline{x} \in \operatorname{int} M$ este punct critic pentru f şi $\nabla^2 f(\overline{x})$ este pozitiv definită (i.e. $\nabla^2 f(\overline{x})(u,u) > 0$ pentru orice $u \in \mathbb{R}^p \setminus \{0\}$), atunci $\overline{x}$ este soluţie locală pentru f.*

De fapt, în Teoremele 1.3.16 şi 1.3.19 se obţin relaţii mai tari decât simpla minimalitate. Introducem mai întâi o definiţie.

Definiţia 1.3.21 *Fie $\alpha > 0$. Spunem că $\overline{x} \in M$ este o soluţie strictă locală de ordin α pentru (P) sau punct de minim local strict de ordin α pentru f pe M dacă există $\varepsilon, \ell > 0$ astfel încât pentru orice $x \in M \cap B(\overline{x}, \varepsilon)$,*

$$f(x) \geq f(\overline{x}) + \ell \|x - \overline{x}\|^{\alpha}.$$

Observaţia 1.3.22 *Să observăm că pentru funcţii diferenţiabile noţiunea de soluţie strictă locală de ordin 1 este specifică cazului cu restricţii active (adică $\overline{x} \in M \setminus \operatorname{int} M$): dacă f este diferenţiabilă în $\overline{x} \in \operatorname{int} M$, atunci $\overline{x}$ nu poate fi soluţie strictă locală de ordin 1. Într-adevăr, dacă $\overline{x} \in \operatorname{int} M$ ar fi soluţie strictă locală de ordin 1, atunci, pe de o parte, $\nabla f(\overline{x}) = 0$ (din Teorema lui Fermat), iar, pe de altă parte, $\nabla f(\overline{x}) \neq 0$ din definiţia soluţiei stricte.*

O altă caracteristică importantă a unei soluţii este aceea de a fi izolată, adică proprietatea ca pe o întreagă vecinătate a sa să nu mai existe altă soluţie. Acest lucru este important din punct de vedere practic pentru aplicarea eficientă a algoritmilor de căutare a soluţiei. Au loc rezultatele cuprinse în rezultatele de mai jos.

Teorema 1.3.23 *Presupunem că f este diferenţiabilă în $\overline{x} \in M$ şi considerăm următoarele afirmaţii:*

(i) $\nabla f(\overline{x})(u) > 0$, pentru orice $u \in T(M,\overline{x}) \setminus \{0\}$;

(ii) $-\nabla f(\overline{x}) \in \operatorname{int} N(M,\overline{x})$;

(iii) $\overline{x}$ este soluţie strictă locală de ordinul $\alpha = 1$ pentru problema (P);

(iv) $\overline{x}$ este soluţie izolată pentru problema (P).

Atunci $(i) \iff (ii) \iff (iii)$. Dacă, în plus, M este convexă şi f este continuu diferenţiabilă în $\overline{x}$, atunci $(i) \implies (iv)$.

Demonstraţie Arătăm echivalenţa primelor două condiţii. Începem cu implicaţia $(i) \implies (ii)$. Cum $\nabla f(\overline{x})$ este continuă şi $T(M,\overline{x}) \cap S(0,1)$ este compactă, folosind ipoteza, există $\mu > 0$ astfel încât $\nabla f(\overline{x})(u) \geq \mu$ pentru orice $T(M,\overline{x}) \cap S(0,1)$. Deducem că pentru orice $u \in T(M,\overline{x})$, $\nabla f(\overline{x})(u) \geq \mu \|u\|$. Fie $v \in D(-\nabla f(\overline{x}), \mu) = -\nabla f(\overline{x}) + D(0,\mu)$ şi $u \in T(M,\overline{x})$. Atunci există $z \in D(0,1)$ astfel încât $v = -\nabla f(\overline{x}) + \mu z$ şi avem

$$v(u) = (-\nabla f(\overline{x}) + \mu z)(u) = -\nabla f(\overline{x})(u) + \mu z(u) \leq \mu \|u\| - \mu \|u\| = 0,$$

deci $v \in N(M, \overline{x})$. Aşadar (ii) are loc.

Demonstrăm implicaţia $(ii) \implies (i)$. Fie $\varepsilon > 0$ astfel încât $-\nabla f(\overline{x}) + D(0, \varepsilon) \subset N(M, \overline{x})$. Pentru orice $u \in T(M, \overline{x})$ şi orice $z \in D(0, \varepsilon)$, avem $(-\nabla f(\overline{x}) + z)(u) \leq 0$, adică $\nabla f(\overline{x})(u) \geq z(u)$, de unde, $\nabla f(\overline{x})(u) \geq \varepsilon \|u\|$. Deci (i) are loc.

Arătăm implicaţia $(i) \implies (iii)$. Presupunem, prin reducere la absurd, că $\overline{x}$ nu este soluţie strictă de ordin 1. Atunci, negând definiţia, deducem că există un şir $(x_n) \to \overline{x}$, $(x_n) \subset M$ astfel încât pentru orice $n \in \mathbb{N}^*$,

$$f(x_n) < f(\overline{x}) + n^{-1} \|x_n - \overline{x}\|.$$

În virtutea acestei inegalităţi,

$$x_n \neq \overline{x}, \ \forall n \in \mathbb{N}^*.$$

Cum f este diferenţiabilă, există un şir de numere reale $(\gamma_n) \to 0$ astfel încât pentru orice $n \in \mathbb{N}$,

$$f(x_n) = f(\overline{x}) + \nabla f(\overline{x})(x_n - \overline{x}) + \gamma_n \|x_n - \overline{x}\|.$$

Combinând cele două relaţii, avem

$$n^{-1} \|x_n - \overline{x}\| > \nabla f(\overline{x})(x_n - \overline{x}) + \gamma_n \|x_n - \overline{x}\|,$$

de unde, prin împărţire la numărul nenul $\|x_n - \overline{x}\|$ deducem

$$n^{-1} > \nabla f(\overline{x}) \left(\frac{x_n - \overline{x}}{\|x_n - \overline{x}\|} \right) + \gamma_n, \ \forall n \in \mathbb{N}^*. \tag{1.1}$$

Cum şirul $\left(\frac{x_n - \overline{x}}{\|x_n - \overline{x}\|} \right)$ este mărginit, există un subşir convergent al acestuia. Limita, notată u, a respectivului subşir este nenulă (chiar de normă 1) şi, în plus, din faptul că $\|x_n - \overline{x}\| \to 0$, deducem că $u \in T(M, \overline{x})$. Prin urmare $u \in T(M, \overline{x}) \setminus \{0\}$ şi trecând la limită în relaţia (1.1) avem

$$0 \geq \nabla f(\overline{x})(u),$$

ceea ce contrazice ipoteza.

Arătăm că $(iii) \implies (i)$. Fie $u \in T(M, \overline{x}) \setminus \{0\}$. Conform definiţiei vectorului tangent, există $(t_n) \subset (0, \infty)$, $t_n \to 0$ şi $(u_n) \to u$ astfel încât pentru orice n, $\overline{x} + t_n u_n \in M$. Atunci, pentru n suficient de mare,

$$f(\overline{x} + t_n u_n) \geq f(\overline{x}) + \ell \|t_n u_n\|.$$

Cum f este diferenţiabilă în $\overline{x}$, există $(\beta_n) \to 0$ astfel încât

$$f(\overline{x} + t_n u_n) - f(\overline{x}) = t_n \nabla f(\overline{x})(u_n) + \beta_n t_n \|u_n\|, \ \forall n.$$

Obţinem, pentru n suficient de mare,

$$\nabla f(\overline{x})(u_n) \geq (\ell - \beta_n) \|u_n\| \geq 2^{-1} \ell \|u_n\|.$$

Prin trecere la limită,

$$\nabla f(\overline{x})(u) \geq 2^{-1} \ell \|u\| > 0,$$

adică concluzia.

În final, arătăm că $(ii) \implies (iv)$ în ipoteza suplimentară că M este convexă şi f este continuu diferenţiabilă în $\overline{x}$. Fie $\varepsilon > 0$ astfel încât $-\nabla f(\overline{x}) + D(0,\varepsilon) \subset N(M,\overline{x})$. Presupunem, prin reducere la absurd, că există un şir de soluţii $(x_n) \subset M \setminus \{\overline{x}\}$ cu $x_n \to \overline{x}$. Conform unei consecinţe binecunoscute a Teoremei Hahn-Banach, pentru fiecare n există z_n cu $\|z_n\| = 1$ şi $z_n(\overline{x} - x_n) = \|\overline{x} - x_n\|$. Fie $(v_n) = (-\nabla f(x_n) - 2^{-1}\varepsilon z_n)$. Din continuitatea diferenţialei în $\overline{x}$, deducem că $v_n \in N(M,\overline{x})$ pentru orice n suficient de mare. Din condiţia necesară de optimalitate de ordinul I, $-\nabla f(x_n) \in N(M,x_n)$. Cum M este convexă, $\overline{x} - x_n \in T(M,x_n)$ şi $x_n - \overline{x} \in T(M,\overline{x})$. Atunci $-\nabla f(x_n)(\overline{x} - x_n) \leq 0$ şi $v_n(x_n - \overline{x}) \leq 0$, deci pentru n mare $(v_n + \nabla f(x_n))(\overline{x} - x_n) \geq 0$. Obţinem $-2^{-1}\varepsilon z_n(\overline{x} - x_n) \geq 0$, de unde deducem că $-2^{-1}\varepsilon\|\overline{x} - x_n\| \geq 0$. În mod evident, ultima relaţie reprezintă o contradicţie. □

Teorema 1.3.24 *Presupunem că f este de clasă C^2, $\nabla f(\overline{x}) = 0$ şi*

$$\nabla^2 f(\overline{x})(u,u) > 0, \ \forall u \in T(M,\overline{x}) \setminus \{0\}.$$

Atunci $\overline{x}$ este soluţie strictă locală de ordinul $\alpha = 2$ pentru problema (P).

Demonstraţie Ca mai sus, presupunem prin reducere la absurd că nu are loc concluzia. Atunci, există $(x_n) \to \overline{x}$, $(x_n) \subset M \setminus \{\overline{x}\}$ astfel încât pentru orice $n \in \mathbb{N}^*$,

$$f(x_n) < f(\overline{x}) + n^{-1}\|x_n - \overline{x}\|^2.$$

Din Teorema lui Taylor, pentru orice $n \in \mathbb{N}$ există c_n pe segmentul ce uneşte $\overline{x}$ cu x_n astfel încât

$$\begin{aligned} f(x_n) - f(\overline{x}) &= \nabla f(\overline{x})(x_n - \overline{x}) + \frac{1}{2}\nabla^2 f(c_n)(x_n - \overline{x}, x_n - \overline{x}) \\ &= \frac{1}{2}\nabla^2 f(c_n)(x_n - \overline{x}, x_n - \overline{x}). \end{aligned}$$

Obţinem

$$n^{-1}\|x_n - \overline{x}\|^2 > \frac{1}{2}\nabla^2 f(c_n)(x_n - \overline{x}, x_n - \overline{x}),$$

de unde, pentru finalizarea demonstraţiei, împărţim la $\|x_n - \overline{x}\|^2$ şi repetăm argumentele de la implicaţia $(i) \implies (iii)$ de mai sus. □

Definiţia 1.3.25 *Considerăm problema (P). Spunem că $u \in \mathbb{R}^p \setminus \{0\}$ este direcţie fezabilă în $x \in M$ dacă există $\alpha > 0$ astfel încât $x + \beta u \in M$ pentru orice $\beta \in [0,\alpha]$. Notăm cu $F(M,\overline{x})$ mulţimea direcţiilor fezabile în $\overline{x}$.*

Exerciţiul 1.9 *În notaţiile definiţiei precedente, să se arate că:*
(i) mulţimea $F(M,\overline{x}) \cup \{0\}$ este un con, nu neaparat închis;
(ii) $\mathrm{cl}(F(M,\overline{x}) \cup \{0\}) \subset T(M,\overline{x})$, *iar incluziunea poate fi strictă;*
(iii) dacă M este convexă, atunci $\mathrm{cl}(F(M,\overline{x}) \cup \{0\}) = T(M,\overline{x})$.

Problema 1.10 *Să se arate următoarea condiţie necesară de optimalitate de ordinul al doilea pentru problema (P) : dacă f este de clasă C^2 şi $\overline{x} \in M$ este soluţie pentru (P) atunci pentru orice u, direcţie fezabilă în $\overline{x}$ pentru care $\nabla f(\overline{x})(u) = 0$ avem $\nabla^2 f(\overline{x})(u,u) \geq 0$.*

Exerciţiul 1.11 *Să se determine punctele de extrem local ale funcţiilor de mai jos (probleme fără restricţii):*

(i) $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x_1, x_2) = (8x_1^2 - 6x_1x_2 + 3x_2^2)e^{2x_1+3x_2}$;
(ii) $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x_1, x_2) = x_1x_2^2e^{x_1-x_2}$;
(iii) $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x_1, x_2) = x_1x_2\left(x_1^2 + x_2^2 - 4\right)$;
(iii) $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x_1, x_2, x_3) = 2x_1^2 + x_2^2 + x_3^2 - 4x_2 + 8x_3 - 5$;
(iv) $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x_1, x_2, x_3) = x_2 \ln(1 + x_1^2 + x_3^2)$;
(v) $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x_1, x_2, x_3) = x_1x_2e^{x_3}$;
(vi) $f : \mathbb{R}^3 \setminus \{0\} \to \mathbb{R}$, $f(x_1, x_2, x_3) = \frac{x_1}{x_2} + \frac{x_2}{4} + \frac{x_3}{x_1} + \frac{1}{x_3}$.

Exerciţiul 1.12 *Fie* $f : \mathbb{R}^2 \to \mathbb{R}$. *Considerăm problema* (P) *a minimizării lui* f *cu restricţiile* $x_1 \geq 0$, $x_2 \geq 0$. *Ştiind că* $\nabla f(0) \neq 0$, $\frac{\partial f}{\partial x_1}(0) \leq 0$ *şi* $\frac{\partial f}{\partial x_2}(0) \leq 0$ *să se arate că* 0 *nu este soluţie a problemei* (P).

Capitolul 2

Elemente teoretice şi practice în optimizarea cu restricţii

2.1 Teorema Karush-Kuhn-Tucker – recapitulare

Problema generală (P) considerată anterior are restricţia $x \in M$. De cele mai multe ori în practică această mulţime M a punctelor fezabile este definită prin intermediul unor funcţii. Fie aşadar $g : \mathbb{R}^p \to \mathbb{R}^n$ şi $h : \mathbb{R}^p \to \mathbb{R}^m$ funcţii de clasă C^1. Evident, g şi h pot fi gândite ca fiind de forma $g = (g_1, g_2, ..., g_n)$, respectiv $h = (h_1, h_2, ..., h_m)$ unde unde $g_i : \mathbb{R}^p \to \mathbb{R}$ $(i \in \overline{1,n})$ şi $h_j : \mathbb{R}^p \to \mathbb{R}$ $(j \in \overline{1,m})$ sunt de clasă C^1.

Considerăm că mulţimea punctelor fezabile este

$$M := \{x \in U \mid g(x) \leq 0,\ h(x) = 0\} \subset \mathbb{R}^p.$$

Se observă că avem două tipuri de restricţii: restricţii cu inegalităţi şi restricţii cu egalităţi. Fie $x \in M$. Dacă pentru un indice $i \in \overline{1,n}$, $g_i(x) < 0$, atunci, din continuitatea lui g, există o întreagă vecinătate V a lui x astfel încât $g_i(y) < 0$ pentru orice $y \in V$. Aceasta face ca, în cazul în care cercetăm dacă x este soluţie locală a problemei (P), restricţia $g_i \leq 0$ să nu influenţeze efectiv mulţimea punctelor u pentru care trebuie să comparăm $f(x)$ cu $f(u)$. De aceea, în acest caz, spunem că restricţia $g_i \leq 0$ este inactivă în x. Astfel, ne interesează ca astfel de restricţii să fie eliminate din discuţie. Pentru $\overline{x} \in M$, notăm mulţimea indicilor restricţiilor inegalităţi active prin

$$A(\overline{x}) = \{i \in \overline{1,n} \mid g_i(\overline{x}) = 0\}.$$

Pentru a reaminti condiţiile necesare de optimalitate de ordinul I pentru această formulare a problemei este necesară recapitularea unor condiţii numite condiţii de calificare. O condiţie de calificare este orice condiţie care poate fi folosită pentru a avea loc rezultatul de mai jos.

Teorema 2.1.1 (Teorema Karush-Kuhn-Tucker) *Fie $\overline{x} \in M$ soluţie a problemei (P). Presupunem că are loc o condiţie de calificare. Atunci există $\lambda = (\lambda_1, \lambda_2, ..., \lambda_n) \in \mathbb{R}^n$, $\mu = (\mu_1, \mu_2, ..., \mu_m) \in \mathbb{R}^m$, astfel încât*

$$\nabla f(\overline{x}) + \sum_{i=1}^{n} \lambda_i \nabla g_i(\overline{x}) + \sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x}) = 0 \tag{2.1}$$

şi

$$\lambda_i \geq 0,\ \lambda_i g_i(\overline{x}) = 0,\ \textit{pentru orice } i \in \overline{1,n}. \tag{2.2}$$

În principiu, condiţiile de calificare trebuie să fie legate de punctul de referinţă ($\overline{x}$ în cazul nostru). De fiecare dată când nu vor fi dubii cu privire la punctul de referinţă vom evita precizarea acestuia, pentru uşurarea expunerii.

Observaţia 2.1.2 *Două dintre cele mai importante condiţii de calificare sunt prezentate mai jos.*

Prima dintre acestea se numeşte condiţia de liniară independenţă (în $\overline{x}$) şi se formulează astfel: mulţimea $\{\nabla g_i(\overline{x}) \mid i \in A(\overline{x})\} \cup \{\nabla h_j(\overline{x}) \mid j \in \overline{1,m}\}$ este liniar independentă.

A doua se numeşte condiţia de calificare Mangasarian-Fromovitz (în $\overline{x}$): mulţimea $\{\nabla h_j(\overline{x}) \mid j \in \overline{1,m}\}$ este liniar independentă şi există $u \in \mathbb{R}^p$ astfel încât

$$\nabla h(\overline{x})(u) = 0 \text{ şi } \nabla g_i(\overline{x})(u) < 0,\ \forall i \in A(\overline{x}).$$

Este uşor de observat că dacă are loc condiţia de liniară independenţă în $\overline{x} \in M$, atunci are loc condiţia Mangasarian-Fromovitz în $\overline{x}$. Totuşi, cele două condiţii nu sunt echivalente.

Există şi alte condiţii de calificare, mai slabe, dar care sunt mai greu de verificat în practică pentru că presupun, printre altele, calculul conului tangent la M în $\overline{x}$.

Cele două condiţii de calificare de mai sus se referă la sisteme generale de restricţii. Să discutăm acum condiţii speciale referitoare la cazuri particulare ale datelor problemei. Mai întâi, precizăm faptul că în anumite ipoteze, condiţia Karush-Kuhn-Tucker este şi suficientă.

Definiţia 2.1.3 *Spunem că problema (P) este convexă dacă U este convexă, f este convexă pe U, funcţiile $g_i, i \in \overline{1,n}$ sunt convexe, iar funcţia h este afină.*

Teorema 2.1.4 *Presupunem că problema (P) este convexă. Fie $\overline{x} \in M$. Dacă există $(\lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^m$ astfel încât au loc relaţiile (2.1) şi (2.2) atunci $\overline{x}$ este soluţie pentru (P) (sau minim al lui f pe M).*

Definiţia 2.1.5 *Spunem că sistemul de restricţii asociat problemei (P) este de tip convex dacă restricţiile cu inegalităţi sunt exprimate prin funcţii convexe, în timp ce restricţiile cu egalităţi sunt date cu funcţii afine, adică $(g_i)_{i\in\overline{1,n}}$ sunt convexe şi h este afină. În acest context, spunem că are loc condiţia Slater dacă există $u \in \mathbb{R}^p$ astfel încât $h(u) = 0$ şi $g(u) < 0$.*

Teorema 2.1.6 *Dacă sistemul de restricţii asociat problemei (P) este de tip convex, atunci condiţia Slater este o condiţie de calificare.*

Definiţia 2.1.7 *Spunem că sistemul de restricţii asociat problemei (P) este de tip afin dacă atât restricţiile cu inegalităţi, cât şi ce restricţiile cu egalităţi sunt date prin funcţii afine.*

Să considerăm acum cazul unui sistem de restricţii de tip afin. Astfel, vom considera o matrice A de tip $n \times p$, o matrice B de tip $m \times p$ şi $b \in \mathbb{R}^n$, $c \in \mathbb{R}^m$. Astfel mulţimea M devine $M = \{x \in \mathbb{R}^p \mid Ax \leq b,\ Bx = c\}$, unde relaţia "$\leq$" este înţeleasă ca având loc pe componente. Deci $g(x) = Ax - b$, $h(x) = Bx - c$.

Teorema 2.1.8 *Dacă sistemul de restricţii este de tip afin, atunci este satisfăcută o condiţie de calificare, deci Teorema 2.1.1 are loc fără verificarea explicită a unei condiţii de calificare.*

Funcţia $L : U \times \mathbb{R}^{n+m} \to \mathbb{R}$,

$$L(x, (\lambda, \mu)) := f(x) + \sum_{i=1}^{n} \lambda_i g_i(x) + \sum_{j=1}^{m} \mu_j h_j(x)$$

se numeşte lagrangianul problemei (P). Astfel, concluzia din relaţia (2.1) se poate scrie

$$\nabla_x L(\overline{x}, (\lambda, \mu)) = 0,$$

iar elementele $(\lambda, \mu) \in \mathbb{R}^n_+ \times \mathbb{R}^m$ se numesc multiplicatori Lagrange. Denumirea se datorează faptului că pentru prima dată o astfel de metodă de a converti o problemă de optimizare cu restricţii într-o problemă fără restricţii prin intermediul unor noi nedeterminate a apărut în unele dintre lucrările lui Lagrange legate de probleme de calculul variaţiilor. Este clar că teorema precedentă nu asigură unicitatea acestor multiplicatori.

Teorema 2.1.1 asigură condiţii necesare de optimalitate pentru problema (P). Dacă în loc de minimizare dorim să maximizăm funcţia obiectiv sub aceleaşi restricţii, atunci, din faptul că $\max f = -\min(-f)$, condiţia necesară (2.1) se scrie

$$-\nabla f(\overline{x}) + \sum_{i=1}^{n} \lambda_i \nabla g_i(\overline{x}) + \sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x}) = 0.$$

Să mai observăm că în lipsa constrângerilor inegalităţi, ţinând seama de faptul că $h(x) = 0$ este totuna cu $-h(x) = 0$, condiţia necesară se poate scrie, atât pentru minime cât şi pentru maxime, în forma

$$\nabla f(\overline{x}) + \sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x}) = 0.$$

Revenind la rezultatul principal, să observăm două lucruri. Mai întâi, dacă problema nu are restricţii (de exemplu, $U = M = \mathbb{R}^p$), atunci relaţia (2.1) se reduce la condiţia necesară de optimalitate de ordinul I (Teorema lui Fermat): $\nabla f(\overline{x}) = 0$. A doua observaţie este că relaţia (2.1) nu are loc în general fără o condiţie de calificare.

Exemplul 2.1.9 Fie $a, b, x \in \mathbb{R}^p$ şi $c, d \in \mathbb{R}$. Rezolvăm problema minimizării normei euclidiene a lui x cu restricţiile $\langle a, x\rangle \leq c$, $\langle b, x\rangle \leq d$. Funcţia obiectiv este strict convexă şi coercivă, restricţiile sunt liniare. Deci, dacă mulţimea punctelor fexabile este nevidă, problema are exact o soluţie, iar condiţiile Karush-Kuhn-Tucker sunt necesare şi suficiente. Deci soluţia x este caracterizată de faptul că există λ_1, $\lambda_2 \geq 0$ astfel încât

$$\begin{aligned} &2x + \lambda_1 a + \lambda_2 b = 0 \\ &\lambda_1 (\langle a, x\rangle - c) = 0 \\ &\lambda_2 (\langle b, x\rangle - d) = 0. \end{aligned}$$

Avem mai multe cazuri de luat în considerare.

Cazul 1: $\lambda_1 = \lambda_2 = 0$. Atunci dacă $c, d \geq 0$, $x = 0$ este punct fezabil şi soluţie a problemei.

Cazul 2: $\lambda_1 = 0$, $\lambda_2 \neq 0$. Atunci ecuaţia $\langle b, x\rangle - d = 0$ se scrie

$$\left\langle b, x - \frac{db}{\|b\|^2} \right\rangle = 0,$$

adică $x \in \frac{db}{\|b\|^2} + \{b\}^{\perp}$, iar minimalitatea normei conduce la $x = \frac{db}{\|b\|^2}$. Atunci $\lambda_2 = -\frac{2d}{\|b\|^2}$. Astfel, pentru ca x să fie soluţie a sistemului, trebuie ca $d < 0$ şi $d\langle a, b\rangle \leq c\|b\|^2$.

Cazul 3: $\lambda_1 \neq 0$, $\lambda_2 = 0$. Atunci, ca în cazul anterior, $x = \frac{da}{\|a\|^2}$. Avem $\lambda_1 = -\frac{2c}{\|a\|^2}$ şi trebuie să aibă loc relaţiile $c < 0$, $c\langle a, b\rangle \leq d\|a\|^2$.

Cazul 4: $\lambda_1 \neq 0$, $\lambda_2 \neq 0$. Determinăm multiplicatorii folosind relaţiile $\langle a, x\rangle = c$, $\langle b, x\rangle = d$ şi înmulţind scalar în mod corespunzător prima ecuaţie. Obţinem

$$\begin{aligned} \|a\|^2 \lambda_1 + \langle a, b\rangle \lambda_2 &= -2c \\ \langle a, b\rangle \lambda_1 + \|b\|^2 \lambda_2 &= -2d. \end{aligned}$$

Determinantul acestui sistem este $\|a\|^2\|b\|^2 - \langle a, b\rangle^2$ care este nenul dacă a, b nu sunt coliniari, caz în care găsim

$$\begin{aligned} \lambda_1 &= \frac{2\left(d\langle a, b\rangle - c\|b\|^2\right)}{\|a\|^2\|b\|^2 - \langle a, b\rangle^2} \\ \lambda_2 &= \frac{2\left(c\langle a, b\rangle - d\|a\|^2\right)}{\|a\|^2\|b\|^2 - \langle a, b\rangle^2}. \end{aligned}$$

Astfel,

$$x = -\frac{1}{2}\left(\lambda_1 a + \lambda_2 b\right).$$

Pentru ca acesta să fie soluţie trebuie ca $d\langle a, b\rangle > c\|b\|^2$, $c\langle a, b\rangle > d\|a\|^2$.

Încheiem această secţiune cu câteva exerciţii recapitulative.

Exerciţiul 2.1 *Rezolvaţi problema minimizării globale a expresiei* $x_1^2 + x_2^2 - 3x_1x_2$ *cu restricţia* $x_1^2 + x_2^2 - 6 \leq 0$.

Exerciţiul 2.2 *Rezolvaţi problema minimizării globale a expresiei* $x_1^2 + x_2^2 - 4x_1 - 6x_2$ *cu restricţiile* $x_1 + x_2 \geq 7$, $x_1 - x_2 \leq 2$.

Exerciţiul 2.3 *Rezolvaţi problema*

$$\min\left(2x_1^2 - 3x_2^2 - 2x_1\right) \text{ cu } x_1^2 + x_2^2 \leq 1.$$

Exerciţiul 2.4 *Rezolvaţi problema*

$$\max\left(-x_1^2 + x_1x_2 - 2x_2^2 + x_1 + x_2\right) \text{ cu } 2x_1 + x_2 \leq 1.$$

Exerciţiul 2.5 *Rezolvaţi problema*

$$\min\left(3x_1 + x_2\right) \ \ cu\ x_1^2 + x_2^2 - 5 \leq 0.$$

Exerciţiul 2.6 *Rezolvaţi problema*

$$\min\left(x_1^2 + x_2^2 - 14x_1 - 6x_2 - 7\right) \ \ cu\ x_1 + x_2 \leq 2,\ x_1 + 2x_2 \leq 3.$$

Exerciţiul 2.7 *Rezolvaţi problema*

$$\min\left(x_1^2 + 2\left(x_2 + 1\right)^2\right) \ \ cu\ -x_1 + x_2 = 2,\ -x_1 - x_2 - 1 \leq 0.$$

Exerciţiul 2.8 *Rezolvaţi problema*

$$\min\left(x_1^2 + x_2^2 - x_1\left(30 - x_1\right) - x_2\left(35 - x_2\right)\right) \ \ cu\ x_1^2 + x_2^2 \leq 250,\ x_1 + x_2 \leq 16,\ x_1 \geq 0,\ x_2 \geq 0.$$

Exerciţiul 2.9 *Fie problema* $\min\left(x_1^2 + x_2^2 - x_2x_3 + x_3^2\right)$ *cu restricţiile* $x_1 - x_2 \geq 1, x_1 = x_3$ *şi* $x_3 \geq 0$. *Determinaţi soluţia sau soluţiile problemei.*

Exerciţiul 2.10 *Rezolvaţi problema*

$$\min\left(x_1^2 + x_2^2 + x_3^2\right) \ \ cu\ x_1 + x_2 + x_3 = 0,\ x_1 + 2x_2 + 3x_3 - 1 = 0.$$

Exerciţiul 2.11 *Găsiţi extremele globale ale expresiei*

$$\left(x_1 + x_2 + x_3\right) \ \ cu\ x_1^2 + x_2^2 + x_3^2 = 4,\ x_3 \leq 1.$$

Exerciţiul 2.12 *Găsiţi extremele globale ale expresiei*

$$\left(x_1x_2 + x_2x_3 + x_1x_3\right)$$

astfel încât:
(i) $x_1^2 + x_2^2 + x_3^2 = 1$;
(ii) $x_1^2 + x_2^2 + x_3^2 = 1,\ x_1 - x_2 + x_3 = 0.$

Exerciţiul 2.13 *Fie* $a \in \mathbb{R}^3$. *Să se găsească valorile extreme ale produsului scalar* $\langle a, x\rangle$ *cu restricţiile* $x_1^2 + x_2^2 + x_3^2 = 1,\ x_1x_2 + x_2x_3 + x_1x_3 = 0.$

Exerciţiul 2.14 *Să se găsească valorile extreme ale expresiei* $x_1^2 + x_2^2$ *pe elipsa* $x_1^2 + 2x_2^2 = 1.$

Exerciţiul 2.15 *Fie* $f : \mathbb{R}^2 \to \mathbb{R},\ f(x_1, x_2) = 5x_1^2 + 4x_1x_2 + x_2^2$ *şi* $h : \mathbb{R}^2 \to \mathbb{R},\ h(x_1, x_2) = 3x_1 + 2x_2 + 5$. *Considerăm problema minimizării lui* f *cu restricţia* $h(x) = 0$. *Să se arate că problema are soluţie unică şi apoi să se determine soluţia precum şi multiplicatorul Lagrange* $\bar{\mu}$ *asociat. Să se arate că apoi că soluţia determinată este minim global fără restricţii pentru funcţia* $x \to L(x, \bar{\mu})$. *Se păstrează aceste concluzii pentru problema minimizării lui* $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x_1, x_2) = x_1^2 - x_2^2 - 3x_2$ *cu restricţia egalitate* $h(x) = 0$, *unde* $h : \mathbb{R}^2 \to \mathbb{R},\ h(x_1, x_2) = x_2$*?*

Exerciţiul 2.16 *Fie $n \in \mathbb{N}^*$ şi $a_i > 0$, pentru orice $i \in \overline{1,n}$. Să se determine minimul expresiei*

$$\sum_{i=1}^{n} a_i x_i^2$$

supusă la constrângerea

$$\sum_{i=1}^{n} x_i = c,$$

unde c este o constantă dată. Care este maximul expresiei sub aceeaşi restricţie?

Exerciţiul 2.17 *Să se studieze problema minimizării expresiei $x_1 + x_2 + ... + x_n$ $(n \geq 2)$ pentru $x_1, x_2, ..., x_n > 0$ cu restricţia $x_1 \cdot x_2 \cdot ... \cdot x_n = 1$ şi apoi să se deducă inegalitatea mediilor.*

2.2 Semnificaţia practică a multiplicatorilor Lagrange

În rezolvarea problemelor de optimizare cu restricţii funcţionale considerate anterior, rolul multiplicatorilor a fost unul auxiliar. Uneori, aflarea punctelor de extrem din sistemului Karush-Kuhn-Tucker necesită determinarea acestora, alteori nu. Totuşi, determinarea efectivă a multiplicatorilor este importantă din perspectiva unei semnificaţii practice pe care o vom investiga mai jos.

Exemplul 2.2.1 (Planificare economică - 1) Trei generatoare electrice deservesc un obiectiv pentru care trebuie să genereze o energie totală de 952 Mw. Fiecare dintre cele trei generatoare contribuie cu o energie x_i, iar costurile specifice sunt respectiv

$$1 + 0.0625x_1 \text{ Euro/Mw}$$
$$1 + 0.0125x_2 \text{ Euro/Mw}$$
$$1 + 0.0250x_3 \text{ Euro/Mw}.$$

Dorim să determinăm cantitatea de energie care trebuie livrată de fiecare generator în parte pentru un cost minim.

Modelarea problemei conduce la

$$\min \left(x_1 + 0.0625x_1^2 + x_2 + 0.0125x_2^2 + x_3 + 0.0250x_3^2\right)$$
$$\text{cu restricţia } x_1 + x_2 + x_3 = 952.$$

Este uşor de văzut că problema este convexă iar sistemul de restricţii este afin. Aşadar, (x_1, x_2, x_3) este soluţie dacă şi numai dacă există $\mu \in \mathbb{R}$ astfel încât

$$1 + 2 \cdot 0.0625x_1 + \mu = 0$$
$$1 + 2 \cdot 0.0125x_2 + \mu = 0$$
$$1 + 2 \cdot 0.0250x_3 + \mu = 0$$
$$x_1 + x_2 + x_3 = 952.$$

Acest sistem se scrie

$$\begin{aligned}
0.125x_1 &= -1 - \mu \\
0.025x_2 &= -1 - \mu \\
0.05x_3 &= -1 - \mu \\
x_1 + x_2 + x_3 &= 952.
\end{aligned}$$

Obţinem $\mu = -15$, $x_1 = 112$, $x_2 = 560$, $x_3 = 280$. Valoarea minimă a funcţiei de cost (i.e., valoarea în acest punct) este 7616.

În continuare, dăm o interpretare (care va fi pe alocuri euristică), pe o problemă abstractă, a valorii multiplicatorilor, pentru ca apoi să revenim la problema concretă de mai sus.

Să presupunem că $f, v : \mathbb{R}^2 \to \mathbb{R}$ (dimensiunea 2 nu este esenţială) sunt două funcţii de clasă C^1 şi că avem problema

$$\min f(x, y) \text{ cu } v(x, y) = c.$$

În ipoteza că problema are soluţie şi Teorema Karush-Kuhn-Tucker este aplicabilă, dacă $(\overline{x}, \overline{y})$ este soluţie a acestei probleme, există $\mu \in \mathbb{R}$ astfel încât

$$\nabla f(\overline{x}, \overline{y}) + \mu \nabla v(\overline{x}, \overline{y}) = 0.$$

În particular, obţinem că

$$-\mu = \frac{\frac{\partial f}{\partial x}(\overline{x}, \overline{y})}{\frac{\partial v}{\partial x}(\overline{x}, \overline{y})} = \frac{\frac{\partial f}{\partial y}(\overline{x}, \overline{y})}{\frac{\partial v}{\partial y}(\overline{x}, \overline{y})},$$

(atunci când împărţirile sunt permise) ceea ce sugerează că μ reprezintă o cuantificare a variaţiei lui f atunci când v variază. Acest lucru poate fi justificat ceva mai riguros după cum urmează. Presupunem că variază valoarea lui c. Atunci soluţia problemei şi multiplicatorul μ pot fi interpretate ca funcţii de c. De asemenea, definim $F(c) = f(\overline{x}(c), \overline{y}(c))$ şi presupunem că toate aceste funcţii sunt diferenţiabile (lucru care, în anumite ipoteze, are loc). Atunci, pe baza regulii lanţului şi a faptului că

$$\frac{\partial f}{\partial x}(\overline{x}(c), \overline{y}(c)) = -\mu(c)\frac{\partial v}{\partial x}(\overline{x}(c), \overline{y}(c)); \frac{\partial f}{\partial y}(\overline{x}(c), \overline{y}(c)) = -\mu(c)\frac{\partial v}{\partial y}(\overline{x}(c), \overline{y}(c))$$

avem

$$\begin{aligned}
F'(c) &= \frac{\partial f}{\partial x}(\overline{x}(c), \overline{y}(c))\overline{x}'(c) + \frac{\partial f}{\partial y}(\overline{x}(c), \overline{y}(c))\overline{y}'(c) \\
&= -\mu(c)\frac{\partial v}{\partial x}(\overline{x}(c), \overline{y}(c))\overline{x}'(c) - \mu(c)\frac{\partial v}{\partial y}(\overline{x}(c), \overline{y}(c))\overline{y}'(c) \\
&= -\mu(c)\left(\frac{\partial v}{\partial x}(\overline{x}(c), \overline{y}(c))\overline{x}'(c) + \frac{\partial v}{\partial y}(\overline{x}(c), \overline{y}(c))\overline{y}'(c)\right).
\end{aligned}$$

Dar $v(\overline{x}(c), \overline{y}(c)) = c$, prin derivare, conduce la

$$\frac{\partial v}{\partial x}(\overline{x}(c), \overline{y}(c))\overline{x}'(c) + \frac{\partial v}{\partial y}(\overline{x}(c), \overline{y}(c))\overline{y}'(c) = 1.$$

Prin urmare,

$$F'(c) = -\mu(c),$$

adică ceea ce am intuit anterior.

Revenind la problema concretă discutată, dacă sarcina de alimentare totală creşte cu o unitate, atunci, cum $-\mu = 15$, funcţia obiectiv îşi creşte valoarea minimă cu aproximativ 15 Euro (conform Teoremei lui Lagrange). Din acest motiv, valorile multiplicatorilor în astfel de probleme se numesc "shadow prices" în teoria economică.

Exemplul 2.2.2 (Planificare economică - 2) O mică fabrică produce două tipuri din acelaşi produs (de exemplu, chipsuri de cartofi). Pentru fiecare sortiment sunt necesare trei operaţiuni principale: tăierea (felierea), prăjirea, împachetarea. Următorul tabel prezintă minutele necesare pentru procesarea unui kilogram din fiecare sortiment.

	tăiere	prăjire	împachetare
tipul 1	2	4	4
tipul 2	4	5	2

Presupunem că pentru un ciclu de producţie, maşinile care fac cele trei operaţiuni au autonomii de 345, 480 şi respectiv 330 de minute. Ştiind că profitul este de 2 Euro şi respectiv $1,5$ Euro per fiecare kilogram din cele două sortimente, să se determine câte kilograme din fiecare sortiment trebuie produse în fiecare ciclu pentru maximizarea profitului.

Modelăm matematic acest enunţ şi obţinem problema de optimizare

$$\max\left(2x + \frac{3}{2}y\right) \text{ cu } 2x + 4y \leq 345, 4x + 5y \leq 480, 4x + 2y \leq 330,$$

ceea ce este echivalent cu determinarea soluţiilor problemei de minimizare

$$\min\left(-2x - \frac{3}{2}y\right) \text{ cu } 2x + 4y \leq 345, 4x + 5y \leq 480, 4x + 2y \leq 330.$$

Problema este convexă, iar mulţimile de nivel relativ la mulţimea punctelor fezabile sunt compacte, deci problema are soluţie globală. Sistemul de restricţii este afin, deci putem scrie: $(x, y) \in M$ este soluţie a problemei dacă şi numai dacă există $\alpha, \beta, \gamma \geq 0$ astfel încât

$$\begin{aligned}
&\left(-2, -\frac{3}{2}\right) + \alpha\,(2, 4) + \beta\,(4, 5) + \gamma\,(4, 2) = (0, 0)\\
&\alpha\,(2x + 4y - 345) = 0\\
&\beta\,(4x + 5y - 480) = 0\\
&\gamma\,(4x + 2y - 330) = 0
\end{aligned}$$

adică

$$\begin{aligned}
&2\alpha + 4\beta + 4\gamma = 2\\
&4\alpha + 5\beta + 2\gamma = \frac{3}{2}\\
&\alpha\left(2x + 4y - 345\right) = 0\\
&\beta\left(4x + 5y - 480\right) = 0\\
&\gamma\left(4x + 2y - 330\right) = 0\\
&\alpha, \beta, \gamma \geq 0\\
&2x + 4y \leq 345\\
&4x + 5y \leq 480\\
&4x + 2y \leq 330.
\end{aligned}$$

Considerăm mai multe cazuri.

1. Situaţia $\alpha = \beta = \gamma = 0$ este imposibilă (de exemplu, pe baza primei ecuaţii).

2. Pe baza primelor două ecuaţii deducem că şi situaţia în care doi dintre aceşti multiplicatori sunt nuli este imposibilă.

3. Dacă toţi multiplicatorii sunt nenuli, obţinem din ecuaţiile $3, 4, 5$, un sistem incompatibil.

4. Dacă $\alpha = 0, \beta, \gamma \neq 0$ obţinem

$$\beta = \frac{1}{6}; \gamma = \frac{1}{3}; x = \frac{115}{2}; y = 50.$$

Aceste valori verifică toate ecuaţiile şi inecuaţiile sistemului de mai sus, deci $\left(\frac{115}{2}; 50\right)$ este soluţie a problemei. Observăm că valoarea funcţiei obiectiv pentru problema iniţială (de maxim) în acest punct este 190 (deci valoarea profitului este de 190 de Euro).

5. Dacă $\beta = 0, \alpha, \gamma \neq 0$ obţinem valorile $x = \frac{105}{2}, y = 60$ care nu este punct fezabil pentru că nu satisface a doua restricţie.

6. Dacă $\gamma = 0, \alpha, \beta \neq 0$ obţinem o valoare negativă pentru α, deci nu avem soluţie în acest caz.

Prin urmare, unica soluţie a problemei este $x = \frac{115}{2}, y = 50$. Pentru această soluţie, multiplicatorii corespunzători celor trei restricţii sunt $\left(0, \frac{1}{6}, \frac{1}{3}\right)$.

Observaţii de felul celei de după Problema 1 relativ la shadow prices sunt valabile şi pentru probleme cu mai multe restricţii. Astfel, multiplicatorii $\alpha = 0, \beta = \frac{1}{6}, \gamma = \frac{1}{3}$ corespunzători soluţiei $(\frac{115}{2}, 50)$ reprezintă (aproximativ) valorile cu care se modifică funcţia obiectiv (în particular şi profitul) dacă de modifică cu o unitate valoarea din dreapta restricţiilor respective. Astfel, dacă maşina de tăiere îşi creşte timpul de funcţionare cu un minut, valoarea profitului nu se modifică. Dacă maşina de prăjire îşi creşte timpul de funcţionare cu un minut, valoarea profitului creşte cu $\frac{1}{6} \approx 0,17$ Euro ($-\beta < 0$, deci scade funcţia de minimizat, adică creşte cea de maximizat), iar maşina de împachetare îşi creşte timpul de funcţionare cu un minut, valoarea profitului creşte cu $\frac{1}{3} \approx 0,34$ Euro. Deci, în principiu, fabricantul va fi interesat să crească în primul rând timpul de funcţionare al acestei maşini.

Exemplul 2.2.3 Fie problema maximizării expresiei $2xy + 3y$ cu restricţiile $4x + y = 10$, $x, y > 0$. Se verifică faptul că punctul de maxim este $\left(2^{-1}, 8\right)$, valoarea maximă este 32, iar multiplicatorul

asociat restricţiei egalitate (singura activă) este $\mu = 4$. Deci, dacă se înlocuieşte 10 cu 11 în formularea problemei ar trebui ca valoarea maximă să fie aproximativ $32 + 4 = 36$. Se constată prin calcul direct că de fapt această valoare este 36.125.

Exerciţiul 2.18 *Determinaţi dimensiunile unei cutii de arie totală minimă având volumul $v > 0$. Interpretaţi valoarea multiplicatorului asociat restricţiei în punctul soluţie.*

Exerciţiul 2.19 *Determinaţi dimensiunile unui vas cilindric care are volumul $v > 0$ şi pentru care aria este minimă. Daţi o interpretare valorii multiplicatorului asociat restricţiei în punctul soluţie.*

2.3 O selecţie de probleme

În această secţiune prezentăm o selecţie de probleme care sunt ivite din chestiuni aplicative şi care pot fi tratate cu metodele studiate până în acest moment.

Problema 2.20 *Se dă un elipsoid de ecuaţie*

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1,$$

unde $a, b, c > 0$. Să se determine paralelipipedul de volum maxim care poate fi inlus în interiorul geometric al elipsoidului precum şi raportul volumelor celor două corpuri.

Soluţie Notăm cu $2x, 2y, 2z$ laturile paralelipipedului. Problema poate fi formalizată astfel:

$$\max 8xyz$$

cu restricţiile

$$\begin{aligned} \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} &\leq 1, \\ x, y, z &\geq 0. \end{aligned}$$

Evident, problema are soluţie, iar dacă (x, y, z) e soluţie, $x, y, z > 0$. De asemenea, Teorema Karush-Kuhn-Tucker este aplicabilă, deci dacă (x, y, z) e soluţie, cu notaţiile evidente, există $\lambda \geq 0$ astfel încât

$$-\nabla f(x, y, z) + \lambda \nabla g(x, y, z) = 0.$$

Avem

$$8yz = \lambda \frac{2x}{a^2},\ 8xz = \lambda \frac{2y}{b^2},\ 8xy = \lambda \frac{2z}{c^2}.$$

Înmulţind două câte două ecuaţiile deducem

$$z = \lambda \frac{1}{4ab}; y = \lambda \frac{1}{4ac}; x = \lambda \frac{1}{4bc}.$$

Este clar că restricţia este activă în punctul soluţie (altfel, gradientul ar trebui sa fie zero în acest punct, ceea ce atrage faptul că avem coordonate nule), deci

$$\lambda = \frac{4abc}{\sqrt{3}},$$

adică

$$x = \frac{a}{\sqrt{3}}; y = \frac{b}{\sqrt{3}}; z = \frac{c}{\sqrt{3}}.$$

Raportul volumelor este

$$\frac{V_p}{V_e} = \frac{8 \cdot 3}{3\sqrt{3} \cdot (4\pi)} = \frac{2}{\sqrt{3}\pi},$$

adică aproximativ $0,3675$. □

Problema 2.21 *Minimizarea unei energii de tip Coulomb. Considerăm N puncte în $\mathbb{R}^3$. Acestea determină C_N^3 triunghiuri (posibil genenerate) şi deci $2^{-1}N(N-1)(N-2)$ unghiuri, notate θ_i. Funcţie de energie locală care trebuie minimizată are forma*

$$E_N = -\left[\frac{N(N-1)}{2} + F_N\right],$$

unde F_N suma tuturor $\cos\theta_i$. Deci problema este de a maximiza global această sumă. Să se rezolve această problemă pentru cazurile $N = 3$ şi $N = 4$.

Soluţie Pentru $N = 3$, trebuie determinate soluţiile globale ale problemei

$$\max(\cos\theta_1 + \cos\theta_2 + \cos\theta_3)$$

cu restricţiile $\theta_1 + \theta_2 + \theta_3 = \pi$, $\theta_i \geq 0$, $i \in \overline{1,3}$. Evident, problema are soluţie. Dacă $(\theta_1, \theta_2, \theta_3)$ este soluţie, atunci toate coordonatele sunt strict pozitive. Într-adevăr, de exemplu $3\cos\frac{\pi}{3} = \frac{3}{2}$, iar dacă $\theta_1 = 0$, atunci $\cos\theta_1 + \cos\theta_2 + \cos\theta_3 = 1 + \cos\theta_2 + \cos(\pi - \theta_2) = 1$. Este aplicabilă Teorema Karush-Kuhn-Tucker şi făcând calculele obţinem că pentru soluţie toate sinusurile sunt egale. Dacă un unghi, de exemplu θ_1, ar fi mai mare sau egal decât $\pi/2$, at $\cos\theta_1 + \cos\theta_2 + \cos\theta_3 \leq \cos\theta_2 + \cos\theta_3 \leq \sqrt{2}$, ceea ce nu e posibil pentru o soluţie globală. Deci toate unghiurile sunt egale cu $\pi/3$. Deci tringhiul trebuie să fie echilateral.

Pentru cazul $N = 4$, bazându-ne pe cazul deja studiat, deducem că configuraţia optimală este un tetraedru regulat. Este de notat şi faptul că pentru $N \geq 5$ problema nu este complet rezolvată încă. □

Problema 2.22 *(Metoda celor mai mici pătrate - un caz neliniar) Metoda celor mai mici pătrate constă în a determina valorile optimale ale parametrilor unui model astfel încât acesta să se potrivească cât mai bine unor date măsurate. Astfel să presupunem că avem două seturi de date $(x_i)_{i\in\overline{1,n}}$ şi $(y_i)_{i\in\overline{1,n}}$, unde $n \in \mathbb{N} \setminus \{0\}$ şi că este natural să ne gândim că ele sunt legate printr-o relaţie pătratică de forma $y = a_1 + a_2x + a_3x^2$, unde $a_1, a_2, a_3 \in \mathbb{R}$. Dorim atunci să determinăm a_1, a_2, a_3 care minimizează $f : \mathbb{R}^3 \to \mathbb{R}$,*

$$f(a) = \sum_{i=1}^{n} \left(y_i - \left(a_1 + a_2x_i + a_3x_i^2\right)\right)^2.$$

Să se studieze această problemă.

Soluţie Dacă notăm matricea

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ ... & ... & ... \\ 1 & x_n & x_n^2 \end{pmatrix},$$

atunci pentru orice $a \in \mathbb{R}^3$ avem $f(a) = \|Xa^t - y\|^2$, unde $y = (y_1, ..., y_n)^t$. Astfel, f este compunerea dintre pătratul normei şi funcţia afină, notată φ, $a \mapsto Xa^t - y$. Evident, f este de clasă C^∞ şi

$$\begin{aligned} \nabla f(a)(u) &= \nabla \|\cdot\|^2 (\varphi(a)) (\nabla h(a)(u)) = 2 \left\langle Xa^t - y, Xu^t \right\rangle = 2 \left\langle X^t \left(Xa^t - y \right), u \right\rangle \\ &= 2 \left\langle a, X^t X u^t \right\rangle - 2 \left\langle y, Xu^t \right\rangle, \ \forall u \in \mathbb{R}^3. \end{aligned}$$

Apoi,

$$\nabla^2 f(a)(u, v) = 2 \left\langle v, X^t X u^t \right\rangle, \ \forall u, v \in \mathbb{R}^3.$$

Astfel,

$$\nabla^2 f(a)(u, u) = 2 \left\langle u, X^t X u^t \right\rangle = 2 \left\langle Xu^t, Xu^t \right\rangle = \left\| Xu^t \right\|^2 \geq 0, \ \forall u \in \mathbb{R}^3,$$

deci f este convexă. Cum problema este fără restricţii iar f este convexă, soluţiile sunt punctulele critice al gradientului, adică punctele a pentru care $X^t(Xa^t - y) = 0$, adică $X^t X a^t = X^t y$.

De fapt,

$$X^t X = \begin{pmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^4 \end{pmatrix}.$$

Dacă $X^t X$ este inversabilă, atunci soluţia este unică: $a^t = (X^t X)^{-1} X^t y$. □

Problema 2.23 *(Minimizarea unei energii electrostatice) Fie $n \geq 2$ un număr natural. Notăm cu $\mathcal{D}$ mulţimea elementelor din $\mathbb{R}^n$ cu toate coordonatele distincte şi definim $f : \mathcal{D} \to \mathbb{R}$ prin*

$$f(x) = \sum_{i=1}^n x_i^2 - \sum_{1 \leq i < j \leq n} \ln(|x_i - x_j|),$$

care modelează un tip de energie electrostatică a unui sistem.

(i) Să se arate că f îşi atinge minimul global pe $\mathcal{D}$.

(ii) Fie $a \in \mathcal{D}$ un punct de minim global al lui f pe $\mathcal{D}$. Notăm cu H funcţia polinomială $t \mapsto \Pi_{i=1}^n (t - a_i)$. Să se arate că

$$H''(a_i) - 4a_i H'(a_i) = 0, \ \forall i \in \overline{1, n}$$

şi

$$H'' - 4tH' = -4nH.$$

(iii) Să se precizeze o modalitate de determinare a punctelor de minim global.

(iv) Particularizaţi pentru $n = 2$ şi $n = 3$.

Soluţie (i) Mulţimea $\mathcal{D}$ este deschisă $\mathbb{R}^n$, fiind intersecţia deschişilor $\{x \in \mathbb{R}^n \mid x_i \neq x_j\}$ cu $1 \leq i < j \leq n$. Frontiera lui $\mathcal{D}$ este formată din cele $2^{-1}n(n-1)$ hiperplane închise de ecuaţii $x_i = x_j$ cu $1 \leq i < j \leq n$.

Dacă $x \in \mathcal{D}$ şi $\|x\| \to \infty$, este clar că $f(x) \to \infty$. Dacă $x \in \mathcal{D}$ şi $x \to y \in \operatorname{Fr}\mathcal{D}$ din nou, $f(x) \to \infty$. Deci, a minimiza pe f pe $\mathcal{D}$ revine la a minimiza pe f pe o mulţime de forma

$$\left\{x \in \mathbb{R}^n \mid \max_{i \in \overline{1,n}} |x_i| \leq r, \ \min_{1 \leq i < j \leq n} |x_i - x_j| \geq \varepsilon\right\},$$

unde $r, \varepsilon > 0$. Cum aceasta este compactă şi f este continuă, deducem că f îşi atinge minimul global pe $\mathcal{D}$. Observăm de asemenea că dacă $a = (a_1, ..., a_n)$ este un astfel de punct atunci orice alt punct având aceleaşi coordonate dar în altă ordene este minim globla la rândul său. Deci odată cu un minim avem de fapt $n!$ minime.

(ii) Pentru $i \in \overline{1,n}$, notăm

$$Q_i(t) = \prod_{j \neq i} (t - a_j).$$

Este clar că pentru $i \in \overline{1,n}$, $H'(a_i) = Q_i(a_i)$, iar derivarea relaţiei $H(t) = (t - a_i) Q_i(t)$ conduce la

$$H''(t) = 2Q_i'(t) + Q_i''(t)(t - a_i),$$

deci $H''(a_i) = 2Q_i'(a_i)$. Derivând Q_i avem

$$Q_i'(t) = \sum_{j \neq i} \prod_{k \neq i, k \neq j} (t - a_k) = \sum_{j \neq i} \frac{Q_i(t)}{t - a_j}, \ \forall t \neq a_j, \ j \in \overline{1,n} \setminus \{i\}.$$

deci

$$Q_i'(a_i) = Q_i(a_i) \sum_{j \neq i} \frac{1}{a_i - a_j}.$$

Teorema lui Fermat pentru a, conduce la

$$2a_i - \sum_{j \neq i} \frac{1}{a_i - a_j} = 0, \ \forall i \in \overline{1,n}.$$

Combinând toate aceste relaţii găsim

$$H''(a_i) - 4a_i H'(a_i) = 0, \ \forall i \in \overline{1,n}.$$

Pentru a justifica următoare relaţie, constatăm că funcţiile polinomiale $H'' - 4tH'$ şi H sunt de grad n şi se anulează în $(a_i)_{i \in \overline{1,n}}$. Aşadar, există o constantă reală c astfel încât $H'' - 4tH' = cH$, iar identificarea coeficienţilor lui t^n duce la $c = -4n$.

(iii) Se observă că polinomul H este caracterizat de relaţia $H'' - 4tH' = -4nH$. Dacă scriem $H(t) = t^n + c_{n-1}t^{n-1} + ... + c_1 t + c_0$, atunci relaţia de mai sus permite determinarea coeficienţilor lui H. Apoi, rezolvarea ecuaţiei $H(t) = 0$, dă cele n componente ale soluţiei a. Aşadar problema are $n!$ soluţii ce se obţin de maniera indicată.

(iv) Pentru $n = 2$, găsim $H(t) = t^2 - 4^{-1}$ care se anulează în 2^{-1} şi -2^{-1}. Deci minimele sunt $(2^{-1}, -2^{-1})$ şi $(-2^{-1}, 2^{-1})$, iar valoarea minimă a lui f este 2^{-1}.

Pentru $n = 3$, $H(t) = t^3 - 3 \cdot 4^{-1} t$ care se anulează în 0 şi $-2^{-1}\sqrt{3}$ şi $-2^{-1}\sqrt{3}$. Deci minimele sunt tripletele obţinute prin permutarea acestor numere. Valoarea minimă a lui f este $2^{-1} \cdot 3 - \ln\left(3 \cdot 4^{-1} \cdot \sqrt{3}\right)$. □

Problema 2.24 *(Minimizarea unei energii într-un volum dat) O formă de energie a unei particule într-un paralelipiped de laturi $a, b, c > 0$ are forma*

$$E(a,b,c) = C\left(\frac{1}{a^2} + \frac{1}{b^2} + \frac{1}{c^2}\right),$$

unde C este o constantă (definită pe baza constantei lui Plank şi a masei particulei). Să se determine, folosind Teorema Karush-Kuhn-Tucker, paralelipipedul de volum fixat $v > 0$ care minimizează această energie. Apoi, să se verifice că funcţia $v \mapsto \inf\{E(a,b,c) \mid abc = v\}$ este derivabilă şi să se comenteze legătura dintre derivata acestei funcţii şi valoarea multiplicatorului Lagrange al problemei.

Soluţie Folosind metoda obişnuită se obţine soluţia $(\sqrt[3]{v}, \sqrt[3]{v}, \sqrt[3]{v})$, deci cubul de volum v. Multiplicatorul este $\mu = 2Cv^{-\frac{5}{3}}$. Apoi, funcţia $v \mapsto \inf\{E(a,b,c) \mid abc = v\}$, notată φ este $\varphi(v) = 2Cv^{\frac{2}{3}}$, este derivabilă şi $\varphi'(v) = -\mu$, adică rezultatul aşteptat, având în vederew semnificaţia practică a multiplicatorilor. □

Problema 2.25 *Se înfăşoară o cutie paralelipipedică cu o panglică de lungime $\ell > 0$. Se trece panglica de câte două ori pe două dintre feţele paralelipipedului şi de 4 ori pe celelalte două. Să se determine cutia de volum maximal care poate fi înfăşurată în acest fel. Studiaţi de asemenea stabilitatea valorii optimale în raport cu data ℓ.*

Soluţie Cu notaţiile evidente, avem de maximizat abc cu restricţiile $2a + 2b + 4c \leq \ell$, $a, b, c \geq 0$. Ultimele restricţii pot fi considerate inactive şi din nou discuţia este cea obişnuită. Se obţine soluţia

$$(a,b,c) = \left(\frac{\ell}{6}, \frac{\ell}{6}, \frac{\ell}{12}\right),$$

şi $\mu = 144^{-1}\ell^2$. Valoarea optimală este $432^{-1}\ell^3$. Constatăm şi aici legătura dintre derivata acestei funcţii şi μ. □

Problema 2.26 *Se consideră tetraedrele de bază un triunghi fixat şi de înălţime fixată. Să se determine, dacă există tetraedrele de arie laterală minimă.*

Soluţie Notăm triunghiul de la bază cu ABC şi înălţimea cu $h > 0$. Notăm de asemenea cu a, b, c laturile triunghiului ABC, cu H piciorul înălţimii pe baza ABC şi cu h_1, h_2, h_3 distanţele de la H la a, b, c, respectiv. Problema se scrie ca minimizarea lui

$$a\sqrt{h^2 + h_1^2} + b\sqrt{h^2 + h_2^2} + c\sqrt{h^2 + h_3^2}$$

cu restricţia

$$ah_1 + bh_2 + ch_3 = 0.$$

O analiză de tipul celor deja utilizate arată că problema are soluţie şi că soluţia satisface $h_1 = h_2 = h_3$. Deci piramida optimă este cea în care vârful se proiectează în centrul cercului înscris bazei.

De altfel, problema poate fi formulată cu un convex compact C ca bază în locul unui triunghi, iar în acest caz general se ştie că există un unic punct în C care realizează minimul cerut, însă determinarea şi caracterizarea acestuia reprezintă o problemă deschisă. □

Problema 2.27 *În spaţiul afin euclidian de dimensiune* 3 *cu reperul* $(O; i, j, k)$ *consideră tetraedre de tip* $OABC$*, unde* A, B, C *se află pe cele trei axe având coordonata nenulă strict pozitivă. Fie* $P = (a, b, c)$ *un punct cu toate coordonatele strict pozitive.*

(i) Să se arate că pătratul ariei triunghiului ABC *este egal cu suma pătratelor ariilor triunghiurilor* OAB*,* OAC*,* OBC*.*

(ii) Construim triunghiuri ABC *care conţin acest punct în interior. Considerăm problema de determinării triunghiului de arie minimă. Să se arate că problema are soluţie, să se determine teoretic soluţia şi să se calculeze în cazurile* $(a, b, c) = (1, 1, 1)$ *şi* $(a, b, c) = (1, 4, 6)$*.*

Soluţie Fie $A = (u, 0, 0)$, $B = (0, v, 0)$, $C = (0, 0, w)$ cu $u, v, w > 0$. Planul ABC are ecuaţia

$$\frac{x}{u} + \frac{y}{v} + \frac{z}{w} = 1.$$

Fie Ω proiecţia lui O pe acest plan. Atunci

$$|O\Omega| = \frac{1}{\sqrt{\frac{1}{u^2} + \frac{1}{v^2} + \frac{1}{w^2}}}.$$

Atunci, volumul tetraedrului $OABC$ este

$$\mathcal{V} = 3^{-1}\mathcal{A}_{ABC}\,|O\Omega| = \frac{uvw}{6}.$$

Deci

$$\mathcal{A}_{ABC} = \frac{1}{2}uvw\sqrt{\frac{1}{u^2} + \frac{1}{v^2} + \frac{1}{w^2}}$$

deci

$$(\mathcal{A}_{ABC})^2 = \left(\frac{uv}{2}\right)^2 + \left(\frac{uw}{2}\right)^2 + \left(\frac{vw}{2}\right)^2,$$

adică suma pătratelor ariilor triunghiurilor OAB, OAC, OBC.

(ii) Conform punctului anterior problema de minimizare se poate scrie sub forma minimizării funcţiei $f : \mathbb{R}^3 \to \mathbb{R}$, $f(u, v, w) = u^2v^2 + u^2w^2 + v^2w^2$ cu restricţiile

$$\frac{a}{u} + \frac{b}{v} + \frac{c}{w} = 1,$$
$$u > 0, v > 0, w > 0.$$

Observăm că dacă (u, v, w) satisface restricţiile, atunci $u \geq a$, $v \geq b$, $w \geq c$, deci

$$f(u, v, w) \geq \min\left(a^2, b^2, c^2\right)\left(u^2 + v^2 + w^2\right),$$

astfel condiţia de coercivitate este satisfăcută. Deci problema admite soluţie globală.

Evident, avem doar o singură restricţie activă, $h(u,v,w)=0$, unde $h:\mathbb{R}^3\to\mathbb{R}$,

$$h(u,v,w)=\frac{a}{u}+\frac{b}{v}+\frac{c}{w}-1.$$

Este clar că are loc condiţia de liniară independenţă în toate punctele fezabile. Problema nu e convexă, deci pentru o soluţie (u,v,w) există $\mu\in\mathbb{R}$ astfel încât

$$\nabla f(u,v,w)+\mu\nabla h(u,v,w)=0.$$

Se obţine sistemul

$$\begin{aligned}2u\left(v^2+w^2\right)-\mu\frac{a}{u^2}&=0\\ 2v\left(u^2+w^2\right)-\mu\frac{b}{v^2}&=0\\ 2w\left(u^2+v^2\right)-\mu\frac{c}{w^2}&=0.\end{aligned}$$

Putem scrie

$$2u^2\left(v^2+w^2\right)=\mu\frac{a}{u},\ 2v^2\left(u^2+w^2\right)=\mu\frac{b}{v},\ 2w^2\left(u^2+v^2\right)=\mu\frac{c}{w},$$

iar prin adunare, pentru că $h(u,v,w)=0$ obţinem $4f(u,v,w)=\mu$. Rescriem sistemul astfel:

$$\begin{aligned}u\left(u^2+v^2+w^2\right)&=2a\left(u^2v^2+u^2w^2+v^2w^2\right)\\ v\left(u^2+v^2+w^2\right)&=2b\left(u^2v^2+u^2w^2+v^2w^2\right)\\ w\left(u^2+v^2+w^2\right)&=2c\left(u^2v^2+u^2w^2+v^2w^2\right).\end{aligned}$$

Dar $u\neq 2a$, $v\neq 2b$, $w\neq 2c$, deci

$$\begin{aligned}u^2v^2+u^2w^2&=\frac{2a}{u-2a}v^2w^2\\ v^2w^2+u^2v^2&=\frac{2b}{v-2b}u^2w^2\\ w^2v^2+u^2w^2&=\frac{2c}{w-2c}v^2u^2.\end{aligned}$$

Scăzând primele două ecuaţii găsim

$$u^2w^2-v^2w^2=\frac{2a}{u-2a}v^2w^2-\frac{2b}{v-2b}u^2w^2,$$

de unde

$$\left(1+\frac{2b}{v-2b}\right)u^2w^2=\left(1+\frac{2a}{u-2a}\right)v^2w^2,$$

care se reduce la

$$u(u-2a)=v(v-2b).$$

Astfel, găsim

$$u\left(u-2a\right)=v\left(v-2b\right)=w\left(w-2c\right).$$

Notăm cu $t>0$ această valoare comună şi avem

$$u=a+\sqrt{t+a^2}=b+\sqrt{t+b^2}=c+\sqrt{t+c^2}.$$

Cum $h\left(u,v,w\right)=0$, $t>0$ este soluţie a ecuaţiei

$$\frac{a}{a+\sqrt{t+a^2}}+\frac{b}{b+\sqrt{t+b^2}}+\frac{c}{c+\sqrt{t+c^2}}=1.$$

Este uşor de văzut că această ecuaţie are o singură rădăcină strict pozitivă. Deci determinarea soluţiei problemei depinde de determinarea acestei rădăcini.

Evident, pentru $\left(a,b,c\right)=\left(1,1,1\right)$, se observă că $t=3$ este rădăcina căutată şi $u=v=w=3$ este soluţia problemei.

Pentru $\left(a,b,c\right)=\left(1,4,6\right)$ trebuie găsită rădăcina t cu aproximaţie (a se Exerciţiul 3.14). □

Problema 2.28 *Demonstrarea Teoremei fundamentale a algebrei (Orice polinom cu coeficienţi complecşi admite o rădăcină complexă) folosind metodele Teoriei optimizării.*

Soluţie Fie

$$P(z)=a_nz^n+a_{n-1}z_{n-1}+...+a_1z+a_0,$$

unde $n\in\mathbb{N}\setminus\{0\}$, $a_i\in\mathbb{C}$ $(i\in\overline{1,n})$, $a_n\neq0$.

Observăm că nu se pierde din generalitate dacă presupunem coeficienţii reali: polinomul $Q(z)=P(z)\overline{P(\bar{z})}$ are coeficienţi reali şi dacă z^* este o radăcină a sa, atunci z^* sau $\overline{z^*}$ este rădăcină pentru P.

Considerăm funcţia $f:\mathbb{C}=\mathbb{R}^2\to\mathbb{R}$,

$$f(z)=\left|P(z)\right|^2.$$

Evident, f este o funcţie continuă. Apoi, f este coercivă, adică $\lim_{|z|\to\infty}f(z)=+\infty$, lucru care se observă din scrierea

$$f(z)=\left|\sum_{k=0}^{\infty}a_kz^k\right|^2\geq\left(\left|a_n\right|\cdot\left|z\right|^n-\sum_{k=0}^{n-1}\left|a_k\right|\cdot\left|z\right|^k\right)^2.$$

Această condiţie de coercivitate, împreună cu continuitatea asigură faptul că f este mărginită inferior şi îşi atinge marginea inferioară: există z^* such that $f(z^*)=\inf_{\mathbb{C}}f$. Acum este suficient să arătăm că $f(z^*)=0$. Fie polinomul

$$R(z)=P(z^*+z),$$

la rândul său un polinom de grad n,

$$R(z)=b_nz^n+...+b_1z+b_0,\ b_n\neq0.$$

Trebuie să arătăm că 0 este rădăcină a acestui polinom, adică $b_0 = 0$. Presupunem că nu este cazul şi notăm cu s primul număr natural nenul pentru care $b_s \neq 0$, deci $R(z) = b_0 + b_s z^s + ... + b_n z^n$. Pentru fiecare $\theta \in [0, 2\pi)$. considerăm funcţia $\varphi_\theta : \mathbb{R} \to \mathbb{R}$,

$$\varphi_\theta(t) = \left|R(te^{i\theta})\right|^2,$$

adică funcţia f evaluată în lungul dreptei care trece prin z^* având direcţia $e^{i\theta}$. Câteva calcule simple arată că putem scrie

$$\varphi_\theta(t) = |b_0|^2 + 2t^s \operatorname{Re}(b_s\overline{b_0}e^{is\theta}) + o(t^s).$$

În plus, ştim că pentru orice $\theta \in [0, 2\pi)$, $t = 0$ este punct de minim pentru φ_θ, deci, în particular, $\varphi'_\theta(0) = 0$, adică $s \geq 2$. De asemenea, $\varphi_\theta^{(k)}(0) = 0$, pentru $k \in \overline{1, s-1}$ şi $\varphi_\theta^{(s)}(0) = 2s! \operatorname{Re}(b_s\overline{b_0}e^{is\theta})$. Din condiţiile de optimalitate cunoscute, trebuie ca primul ordin k pentru care $\varphi_\theta^{(k)}(0) \neq 0$ să fie par şi $\varphi_\theta^{(k)}(0) > 0$. Aceasta înseamnă că s este par şi că $\operatorname{Re}(b_s\overline{b_0}e^{is\theta}) > 0$ pentru orice $\theta \in [0, 2\pi)$. Cum $b_s\overline{b_0} \neq 0$, putem scrie acest număr sub forma $u + iv$ cu $u, v \in \mathbb{R}$ şi $u^2 + v^2 \neq 0$. Atunci

$$\begin{aligned}\operatorname{Re}(b_s\overline{b_0}e^{is\theta}) &= \operatorname{Re}\left[(u + iv)(\cos s\theta + i \sin s\theta)\right] \\ &= u \cos s\theta - v \sin s\theta = \sqrt{u^2 + v^2}\cos(s\theta + \rho),\end{aligned}$$

unde ρ este argumentul lui $u + iv$. Evident, atunci când θ variază în intervalul $[0, 2\pi)$, această expresie ia toate valorile dintre $-\sqrt{u^2 + v^2}$ şi $\sqrt{u^2 + v^2}$, deci nu este mereu pozitivă, ceea ce reprezintă o contradicţie. Presupunerea făcută este deci falsă, ceea ce înseamnă că $b_0 = 0$, adică $P(z^*) = 0$. □

Problema 2.29 *Fie funcţia $f : \mathbb{R}^3_+ \to \mathbb{R}$ definită prin*

$$f(x, y, z) = \begin{cases} 0, & \text{dacă } (x, y, z) = 0 \in \mathbb{R}^3, \\ \frac{xyz}{(x+y+z)^2}, & \text{dacă } (x, y, z) \neq 0. \end{cases}$$

Fie $a > 0$. Să se arate că îşi atinge maximul global pe mulţimea

$$M = \{(x, y, z) \in \mathbb{R}^3 : x > 0, y > 0, z > 0, x^2 + y^2 + z^2 = a^2\}$$

şi să se determine acest maxim. Să se deducă următoarea inegalitate: pentru orice $x, y, z \in \mathbb{R}$,

$$|xyz| \leq \frac{1}{9\sqrt{3}} \left(|x| + |y| + |z|\right)^2 \sqrt{x^2 + y^2 + z^2}.$$

Soluţie Funcţia f este evident continuă pe $\mathbb{R}^3 \setminus \{0\}$ ca raport de funcţii continue. Mai mult, pentru $(x, y, z) \in \mathbb{R}^3_+$:

$$xyz \leq (\max(x, y, z))^3$$

şi

$$(x + y + z)^2 \geq (\max(x, y, z))^2,$$

deci $f(x, y, z) \leq \max(x, y, z)$, ceea ce probează continuitatea lui f în 0.

Mulţimea $M \cup \{0\}$ este compactă, aşadar f îşi atinge maximul global pe $M \cup \{0\}$. Cum acest maxim nu poate fi în 0, rezultă că punctul de maxim este în M. Din nou, este evident că f este de clasă C^1 pe $\mathbb{R}^3_+$. Este aplicabilă Teorema Karush-Kuhn-Tucker pentru problema maximizării lui f cu restricţia $x^2+y^2+z^2=a^2$ şi cu $(x,y,z) \in \mathbb{R}^3_+ \setminus \{0\}$. Astfel, dacă $(x,y,z) \in \mathbb{R}^3_+ \setminus \{0\}$ este punct de maxim pentru această problemă, există $\mu \in \mathbb{R}$ astfel încât

$$\begin{aligned}
&\frac{yz}{(x+y+z)^2} - 2\frac{xyz}{(x+y+z)^3} + 2\mu x = 0\\
&\frac{xz}{(x+y+z)^2} - 2\frac{xyz}{(x+y+z)^3} + 2\mu y = 0\\
&\frac{xy}{(x+y+z)^2} - 2\frac{xyz}{(x+y+z)^3} + 2\mu z = 0\\
&x^2+y^2+z^2 = a^2.
\end{aligned}$$

Deducem

$$\begin{aligned}
&\frac{xyz}{(x+y+z)^2} - 2x\frac{xyz}{(x+y+z)^3} + 2\mu x^2 = 0\\
&\frac{xyz}{(x+y+z)^2} - 2y\frac{xyz}{(x+y+z)^3} + 2\mu y^2 = 0\\
&\frac{xyz}{(x+y+z)^2} - 2z\frac{xyz}{(x+y+z)^3} + 2\mu z^2 = 0\\
&x^2+y^2+z^2 = a^2,
\end{aligned}$$

de unde

$$0 < f(x,y,z) = 3\frac{xyz}{(x+y+z)^2} - 2\frac{xyz}{(x+y+z)^2} = -2\mu a^2.$$

În particular, $\mu < 0$ şi scăzând primele două ecuaţii din sistemul Karush-Kuhn-Tucker, avem

$$\frac{z(y-x)}{(x+y+z)^2} + 2\mu(x-y) = 0,$$

deci

$$\left(2\mu - \frac{z}{(x+y+z)^2}\right)(x-y) = 0.$$

Cum expresia din prima paranteză este strict mai mică decât 0, găsim $x=y$. Similar, $x=z$, deci

$$x = y = z = \frac{a}{\sqrt{3}}$$

şi

$$\sup_{(x,y,z)\in M} f(x,y,z) = \frac{a}{9\sqrt{3}}.$$

Fie acum $x,y,z \in \mathbb{R}$. Notăm $a = \sqrt{x^2+y^2+z^2}$ şi din cele de mai sus,

$$f(|x|,|y|,|z|) \le \frac{\sqrt{x^2+y^2+z^2}}{9\sqrt{3}},$$

iar concluzia se obţine imediat. □

Problema 2.30 *Fie $n \in \mathbb{N}\setminus\{0\}$ şi $(\alpha_i)_{i\in\overline{1,n}}$, β numere reale strict pozitive. Să se rezolve problema maximizării expresiei*

$$\prod_{i\in\overline{1,n}} (x_i + \alpha_i)$$

cu restricţiile $x_i \geq 0$ pentru orice $i \in \overline{1,n}$ şi

$$\sum_{i\in\overline{1,n}} x_i = \beta.$$

2.4 Un exemplu, două întrebări

Să considerăm următoarea problemă de optimizare cu restricţii.

Exerciţiul 2.31 *Să se determine minimul global al expresiei*

$$(x_1 + x_2 + x_3) \ \text{ cu } x_1^2 + x_2^2 + x_3^2 = 4, \ x_3 \geq 1.$$

Se parcurg următorii paşi:

I. se arată că există soluţie (Teorema lui Weierstrass);

II. se arată că are loc condiţia de liniară independenţă în toate punctele fezabile studiind separat cazul în care restricţia inegalitate este activă, respectiv inactivă;

III. se aplică Teorema Krush-Kuhn-Tucker şi se determină punctele care sunt potenţiale soluţii, iar odată cu ele, multiplicatorii Lagrange asociaţi, obţinându-se $x = \left(\frac{2}{\sqrt{3}}, \frac{2}{\sqrt{3}}, \frac{2}{\sqrt{3}}\right)$ *cu* $\lambda = 0$, $\mu = -\frac{\sqrt{3}}{4}$, $x = \left(\sqrt{\frac{3}{2}}, \sqrt{\frac{3}{2}}, 1\right)$ *cu* $\lambda = 1 - \sqrt{\frac{2}{3}}$, $\mu = -\frac{1}{\sqrt{6}}$, $x = \left(-\sqrt{\frac{3}{2}}, -\sqrt{\frac{3}{2}}, 1\right)$ *cu* $\lambda = 1 + \sqrt{\frac{2}{3}}$, $\mu = \frac{1}{\sqrt{6}}$.

IV. se determină soluţia globală, $x = \left(-\sqrt{\frac{3}{2}}, -\sqrt{\frac{3}{2}}, 1\right)$.

Studiind aceşti paşi şi calculele implicate, ne punem următoarele întrebări:

- este imperios necesară etapa (relativ dificilă din punct de vedere tehnic) a verificării unei condiţii de calificare?

- putem preciza natura punctelor care satisfac Teorema Krush-Kuhn-Tucker, dar nu sunt minime globale?

În continuare, încercăm să dăm, pe rând, un răspuns acestor chestiuni.

2.5 Condiţiile Fritz John

Aşa cum am văzut, pentru problema (P) cu restricţii funcţionale, avem condiţiile de optimalitate date de Teorema Karush-Kuhn-Tucker (Teorema 2.1.1) care pentru aplicare necesită verificarea unei condiţii de calificare. În multe cărţi (cu profil mai cu seamă ingineresc) această teoremă este aplicată fără verificarea unei asemenea condiţii. Evident, acest lucru este greşit, dar există unele motive, pe care le discutăm mai jos, pentru care o asemenea abordare conduce totuşi, în multe situaţii, la rezultate corecte.

Prezentăm deci unele condiţii de optimalitate, numite condiţiile Fritz John, în care funcţia obiectiv şi funcţiile ce definesc restricţiile sunt implicate nediferenţiat. Aceste condiţii sunt apropiate ca formă de condiţiile Karush-Kuhn-Tucker, dar diferă printr-un aspect care se dovedeşte esenţial în determinarea punctelor de extrem.

Rezultatul de mai jos referă la condiţii necesare de optimalitate pentru problema (P) cu restricţii funcţionale fără nicio ipoteză în plus în afara cadrului general precizat mai sus. Aceste condiţii au fost obţinute de către Fritz John în 1948, motiv pentru care poartă numele acestui matematician.

Teorema 2.5.1 *Fie $\overline{x} \in M$ soluţie a problemei (P). Atunci există $\lambda_0 \in \mathbb{R}$, $\lambda_0 \geq 0$, $\lambda = (\lambda_1, \lambda_2, ..., \lambda_n) \in \mathbb{R}^n$, $\mu = (\mu_1, \mu_2, ..., \mu_m) \in \mathbb{R}^m$, cu proprietatea că $\lambda_0 + \|\lambda\| + \|\mu\| \neq 0$ astfel încât*

$$\lambda_0 \nabla f(\overline{x}) + \sum_{i=1}^{n} \lambda_i \nabla g_i(\overline{x}) + \sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x}) = 0$$

şi

$$\lambda_i \geq 0, \ \lambda_i g_i(\overline{x}) = 0, \ \textit{pentru orice } i \in \overline{1,n}.$$

Demonstraţie Fie $\delta > 0$ astfel încât $D(\overline{x}, \delta) \subset U$ şi pentru orice $x \in M \cap D(\overline{x}, \delta)$, $f(\overline{x}) \leq f(x)$. Pentru fiecare $k \in \mathbb{N} \setminus \{0\}$ considerăm funcţia $\varphi_k : D(\overline{x}, \delta) \to \mathbb{R}$ dată prin

$$\varphi_k(x) = f(x) + \frac{k}{2} \sum_{i=1}^{n} \left(g_i^+(x)\right)^2 + \frac{k}{2} \sum_{j=1}^{m} \left(h_j(x)\right)^2 + \frac{1}{2} \|x - \overline{x}\|^2,$$

unde $g_i^+(x) = \max\{g_i(x), 0\}$. Evident, φ_k îşi atinge minimul pe $D(\overline{x}, \delta)$ şi notăm cu x_k un astfel de punct de minim. De asemenea,

$$\varphi_k(x_k) = f(x_k) + \frac{k}{2} \sum_{i=1}^{n} \left(g_i^+(x_k)\right)^2 + \frac{k}{2} \sum_{j=1}^{m} \left(h_j(x_k)\right)^2 + \frac{1}{2} \|x_k - \overline{x}\|^2 \leq \varphi_k(\overline{x}) = f(\overline{x}).$$

Din faptul că (x_k) este un şir mărginit, iar f este continuă pe $D(\overline{x}, \delta)$, deducem că $(f(x_k))$ este un şir mărginit. Făcând $k \to \infty$ în relaţia de mai sus, obţinem

$$\lim_{k\to\infty} \sum_{i=1}^{n} \left(g_i^+(x_k)\right)^2 = 0$$

$$\lim_{k\to\infty} \sum_{j=1}^{m} \left(h_j(x_k)\right)^2 = 0.$$

Tot mărginirea lui (x_k) asigură faptul că putem extrage un subşir convergent al acestui şir. Fără a mai modifica indicii, putem scrie $x_k \to x^* \in D(\overline{x}, \delta)$, iar relaţiile anterioare asigură faptul că $x^* \in M$. Mai mult, inegalitatea de mai sus ne permite să scriem

$$f(x_k) + \frac{1}{2} \|x_k - \overline{x}\|^2 \leq f(\overline{x}),$$

de unde, prin trecere la limită,

$$f(x^*) + \frac{1}{2}\|x^* - \overline{x}\|^2 \leq f(\overline{x}).$$

Dar, pe de altă parte, $f(\overline{x}) \leq f(x^*)$, deci $\|x^* - \overline{x}\| = 0$, adică $x^* = \overline{x}$. Prin urmare, $x_k \to \overline{x}$.

O observaţie esenţială este că φ_k este diferenţiabilă întrucât funcţiile (scalare) care ar putea pune probleme, i.e., $g_i^+(x)$, sunt ridicate la pătrat şi $\nabla\left(g_i^+(\cdot)\right)^2(x) = 2g_i^+(x)\nabla g_i(x)$. De fapt, $\left(g_i^+(x)\right)^2$ este compunerea dintre funcţia scalară derivabilă $t \to (t^+)^2$ şi funcţia g_i.

Cum x_k este minim pentru φ_k pe $D(\overline{x}, \delta)$ şi pentru k suficient de mare x_k se află în interiorul bilei $D(\overline{x}, \delta)$, concluzionăm că pentru acele numere k putem aplica Teorema lui Fermat, adică

$$\nabla\varphi_k(x_k) = 0.$$

Punând aceste lucruri împreună, putem scrie

$$\nabla f(x_k) + k\sum_{i=1}^{n} g_i^+(x_k)\nabla g_i(x_k) + k\sum_{j=1}^{m} h_j(x_k)\nabla h_j(x_k) + x_k - \overline{x} = 0, \tag{2.3}$$

pentru orice k suficient de mare. Notăm pentru $i \in \overline{1,n}$, $j \in \overline{1,m}$, $\alpha_i^k := kg_i^+(x_k)$, $\beta_j^k := kh_j(x_k)$ şi $\gamma^k = \sqrt{1 + \sum_{i=1}^n \left(\alpha_i^k\right)^2 + \sum_{j=1}^m \left(\beta_i^k\right)^2}$. Este clar că $\gamma^k > 0$ şi notăm $\lambda_0^k := \frac{1}{\gamma^k}$, $\lambda_i^k := \frac{\alpha_i^k}{\gamma^k}$, $\mu_j^k := \frac{\beta_j^k}{\gamma^k}$. Se observă că

$$\left(\lambda_0^k\right)^2 + \sum_{i=1}^{n}\left(\lambda_i^k\right)^2 + \sum_{j=1}^{m}\left(\mu_j^k\right)^2 = 1,$$

deci şirurile (λ_0^k), (λ_i^k), (μ_j^k) ($i \in \overline{1,n}$, $j \in \overline{1,m}$) sunt mărginite. Prin urmare, vor exista subşiruri (indexate la fel) convergente în $\mathbb{R}$ la nişte scalari pe care îi notăm

$$\lambda_0, \lambda_1, \lambda_2, ..., \lambda_n,\ \ \mu_1, \mu_2, ..., \mu_m$$

şi care nu pot fi simultan zero. Pozitivitatea şirurilor (λ_0^k), (λ_i^k) ($i \in \overline{1,n}$) atrage pozitivitatea limitelor λ_0, $\lambda_1, \lambda_2, ..., \lambda_n$. Acum, împărţind relaţia (2.3) la γ^k, se obţine

$$\lambda_0^k\nabla f(x_k) + \sum_{i=1}^{n}\lambda_i^k\nabla g_i(x_k) + \sum_{j=1}^{m}\mu_j^k\nabla h_j(x_k) + \frac{1}{\gamma^k}(x_k - \overline{x}) = 0.$$

Prin trecere la limită ($k \to \infty$) se obţine prima relaţie căutată. Arătăm acum a doua relaţie. Fie $i \in \overline{1,n}$. Dacă $\lambda_i = 0$, atunci nu avem nimic de demonstrat. Contrar, dacă $\lambda_i > 0$, din definiţia lui λ_i deducem că pentru k suficient de mare $g_i^+(x_k) > 0$, deci $g_i^+(x_k) = g_i(x_k)$. Relaţia

$$0 < g_i(x_k) \to g_i(\overline{x}) \leq 0$$

conduce la concluzia $g_i(\overline{x}) = 0$. Deci, şi cea de-a doua concluzie are loc şi teorema este complet demonstrată. □

Dacă comparăm Teorema 2.1.1 cu Teorema 2.5.1 observăm următoarea diferenţă: scalarul asociat funcţiei obiectiv este de această dată nenul, iar acesta este inconvenientul major al teoremei

de mai sus şi anume că nu elimină posibilitatea ca scalarul asociat funcţiei obiectiv să fie nul. Astfel, există riscul ca prea multe puncte fezabile să verifice condiţiile Fritz John, iar aplicarea teoremei să nu aducă precizări importante. De exemplu, dacă un punct fezabil x satisface $\nabla g_i(x) = 0$ pentru un $i \in A(x)$ sau $\nabla h_j(x) = 0$ pentru un $j \in \overline{1,m}$, atunci el satisface condiţiile Fritz-John (cu $\lambda_0 = 0$), funcţia obiectiv nemaiavând niciun rol de jucat. Condiţiile de calificare prezente în Teorema 2.1.1 evită tocmai situaţia $\lambda_0 = 0$. De exemplu, având în vedere Observaţia 2.1.2, pentru a demonstra că cele două condiţii din respectivul rezultat sunt de calificare, este suficient să arătăm acest lucru pentru condiţia Mangasarian-Fromovitz. Acest lucru este evident dacă aplicăm Teorema 2.5.1 şi raţionăm prin reducere la absurd: dacă $\lambda_0 = 0$ atunci

$$\sum_{i \in A(\overline{x})} \lambda_i \nabla g_i(\overline{x}) + \sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x}) = 0.$$

Înmulţind scalar cu vectorul u din condiţia Mangasarian-Fromovitz deducem că

$$\sum_{i \in A(\overline{x})} \lambda_i \langle \nabla g_i(\overline{x}), u \rangle = 0$$

de unde se obţine $\lambda_i = 0$ pentru orice $i \in A(\overline{x})$. Deci

$$\sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x}) = 0,$$

iar condiţia de liniară independenţă asupra gradienţilor $\{\nabla h_j(\overline{x}) \mid j \in \overline{1,m}\}$ atrage faptul că $\mu_j = 0$ pentru orice $j \in \overline{1,m}$. Reunind aceste deducţii ajungem la contradicţie cu relaţia $|\lambda_0| + \|\lambda\| + \|\mu\| \neq 0$. Deci $\lambda_0 \neq 0$.

Ilustrăm modul de funcţionare a Teoremei 2.5.1 prin câteva exerciţii concrete.

Exerciţiul 2.32 *Fie problema de minimizare a funcţiei* $f : \mathbb{R}^2 \to \mathbb{R}$,

$$f(x_1, x_2) = (x_1 - 3)^2 + (x_2 - 2)^2$$

cu restricţia $g(x) \leq 0$, *unde* $g : \mathbb{R}^2 \to \mathbb{R}^4$,

$$g(x_1, x_2) = (x_1^2 + x_2^2 - 5, x_1 + 2x_2 - 4, -x_1, -x_2).$$

(i) Să se arate, folosind un desen, că $(2,1)$ *este soluţie a problemei.*
(ii) Să se arate că au loc condiţiile Fritz John în acest punct.
(iii) Studiaţi îndeplinirea condiţiilor Fritz John în $(0,0)$ *şi concluzionaţi.*

Soluţie (i) Se poate observa grafic că $\overline{x} = (2,1)$ este soluţie a problemei desenând mulmea punctelor fezabile şi interpretând funcţia obiectiv ca fiind pătratul distanţei de la (x_1, x_2) la $(3,2)$.

(ii) Pentru $\overline{x} = (2,1)$ avem $A(\overline{x}) = \{1,2\}$. Dorim să verificăm condiţiile Fritz John în acest punct. Astfel, din a doua condiţie, cum $3,4 \notin A(\overline{x})$, deducem că $\lambda_3 = \lambda_4 = 0$. Cum $\nabla f(\overline{x}) = (-2,-2)$, $\nabla g_1(\overline{x}) = (4,2)$, $\nabla g_2(\overline{x}) = (1,2)$, trebuie să găsim numere reale pozitive $\lambda_0, \lambda_1, \lambda_2 \geq 0$, nu toate nule, astfel încât

$$\lambda_0(-2,-2) + \lambda_1(4,2) + \lambda_2(1,2) = (0,0).$$

Obţinem $\lambda_1 = \frac{1}{3}\lambda_0$ şi $\lambda_2 = \frac{2}{3}\lambda_0$, deci alegând $\lambda_0 > 0$, prima condiţie Fritz John este îndeplinită.

(iii) Să vedem acum dacă punctul $x = (0,0)$ satisface condiţiile. De această dată $A(x) = \{3,4\}$, deci $\lambda_1 = \lambda_2 = 0$. Avem $\nabla f(x) = (-6,-4)$, $\nabla g_3(x) = (-1,0)$, $\nabla g_4(x) = (0,-1)$. Un calcul simplu arată că ecuaţia

$$\lambda_0(-6,-4) + \lambda_3(-1,0) + \lambda_4(0,-1) = (0,0)$$

nu admite soluţie $(\lambda_0, \lambda_3, \lambda_4)$ nenulă cu toate componentele pozitive. Deci x nu îndeplineşte condiţiile Fritz John, aşadar nu este soluţie pentru problema dată. □

Exerciţiul 2.33 *Fie problema de minimizare a funcţiei* $f : (0,\infty)\times(0,\infty) \to \mathbb{R}$, $f(x_1,x_2) = -2x_2$ *cu restricţia* $g(x) \leq 0$, *unde* $g : \mathbb{R}^2 \to \mathbb{R}^3$, $g(x_1,x_2) = (x_1 - x_2 - 2, -x_1 + x_2 + 2, x_1 + x_2 - 6)$.

(i) Desenaţi mulţimea punctelor fezabile şi determinaţi grafic soluţia.

(ii) Să se arate că toate punctele fezabile satisfac condiţiile Fritz John.

(iii) Care sunt punctele în care au loc condiţiile Karush-Kuhn-Tucker. Comentaţi

Soluţie (i) Mulţimea M a punctelor fezabile este segmentul $[(2,0),(4,2)] \setminus \{(2,0)\}$ şi soluţia este $\overline{x} = (4,2)$.

(ii), (iii) Este uşor de observat că orice punct fezabil satisface condiţiile Fritz John, dar soluţia este singurul pentru care putem alege $\lambda_0 \neq 0$. Într-adevăr, dacă x este un punct fezabil diferit de soluţie, atunci $\lambda_3 = 0$, $\lambda_1 = \lambda_2$ şi $\lambda_0 = 0$. Deci, în acest caz, aplicarea condiţiilor Karush-Kuhn-Tucker determină soluţia, ceea ce nu se întâmplă pentru condiţiile Fritz John. Observăm că toate restricţiile sunt afine, deci se poate aplica Teorema Karush-Kuhn-Tucker. □

Exerciţiul 2.34 *Fie problema minimizării funcţiei* $(x_1^2 + x_2^2)$ *cu restricţia* $x_1^2 - (x_2 - 1)^3 = 0$.

(i) Să se arate că problema are soluţie.

(ii) Care sunt punctele în care au loc condiţiile Fritz John? Comentaţi

Soluţie (i) Cum funcţia obiectiv este coercivă, problema are soluţie globală.

(ii) Se poate cu uşurinţă verifica că singurul punct care satisface condiţiile Fritz John este $(0,1)$. Deci acest punct este soluţia problemei. □

Exerciţiul 2.35 *Fie* $f : \mathbb{R}^2 \to \mathbb{R}$ *şi* $g : \mathbb{R}^2 \to \mathbb{R}^2$ *definite prin* $f(x_1,x_2) = x_1$ *şi* $g(x_1,x_2) = (-x_2 + (1-x_1)^3, x_2)$.

(i) Determinaţi soluţia folosind structura particulară a problemei.

(ii) Să se arate că în punctul soluţie au loc condiţiile Fritz John, dar nu au loc condiţiile Karush-Kuhn-Tucker.

Soluţie (i) Se verifică faptul că $\overline{x} = (1,0)$ este soluţie pentru problema asociată.

(ii) Condiţiile Fritz John sunt îndeplinite cu $\lambda_0 = 0$. Condiţia (2.1) nu are loc. □

Punem acum în lumină şi o situaţie în care putem asigura condiţii suficiente de optimalitate.

Teorema 2.5.2 *Fie* $\overline{x} \in M$ *un punct fezabil pentru problema* (P). *Presupunem că au loc condiţiile din concluzia Teoremei 2.5.1, adică există* $\lambda_0 \in \mathbb{R}$, $\lambda_0 \geq 0$, $\lambda = (\lambda_1, \lambda_2, ..., \lambda_n) \in \mathbb{R}^n$, $\mu = (\mu_1, \mu_2, ..., \mu_m) \in \mathbb{R}^m$, *cu proprietatea că* $\lambda_0 + \|\lambda\| + \|\mu\| \neq 0$ *astfel încât*

$$\lambda_0 \nabla f(\overline{x}) + \sum_{i=1}^{n} \lambda_i \nabla g_i(\overline{x}) + \sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x}) = 0$$

şi

$$\lambda_i \geq 0,\ \lambda_i g_i(\overline{x}) = 0,\ \textit{pentru orice } i \in \overline{1,n}.$$

Dacă mulţimea de vectori

$$\left\{\lambda_0 \nabla f(\overline{x}), \left(\lambda_i \nabla g_i(\overline{x})\right)_{i\in A(\overline{x})}, \left(\mu_j \nabla h_j(\overline{x})\right)_{j=\overline{1,m}}\right\}$$

formează un sistem de generatori pentru $\mathbb{R}^p$*, atunci* $\overline{x}$ *este soluţie pentru problema* (P)*.*

Demonstraţie. Presupunem, prin reducere la absurd, că $\overline{x}$ nu este soluţie (locală) a problemei (P). Deci, există un şir (x_k) de puncte fezabile astfel încât $x_k \to \overline{x}$ şi $f(x_k) < f(\overline{x})$ pentru orice k. Atunci, pentru orice k, $x_k \neq \overline{x}$ şi

$$x_k = \overline{x} + \|x_k - \overline{x}\| \frac{1}{\|x_k - \overline{x}\|}(x_k - \overline{x}).$$

Notăm cu (t_k) şirul de numere strict pozitive convergent la 0, $(\|x_k - \overline{x}\|)$ şi cu (d_k) şirul de vectori de normă 1, $\left(\frac{1}{\|x_k-\overline{x}\|}(x_k - \overline{x})\right)$. Fără a restrânge generalitatea, putem presupune că (d_k) este convergent la un vector d de normă 1. Folosind faptul că funcţiile implicate sunt de clasă C^1, există nişte şiruri (α_k), $(\beta_k^i)_{i\in A(\overline{x})}$, $\left(\gamma_k^j\right)_{j\in\overline{1,m}}$ cu limita 0 astfel încât pentru orice k

$$\begin{aligned}
&0 > f(\overline{x} + t_k d_k) - f(\overline{x}) = t_k \nabla f(\overline{x})(d_k) + t_k \alpha_k,\\
&0 \geq g_i(\overline{x} + t_k d_k) - g_i(\overline{x}) = t_k \nabla g_i(\overline{x})(d_k) + t_k \beta_k^i,\ \forall i \in A(\overline{x}),\\
&0 = h_j(\overline{x} + t_k d_k) - h_j(\overline{x}) = t_k \nabla h_j(\overline{x})(d_k) + t_k \gamma_k^j,\ \forall i \in \overline{1,m}.
\end{aligned}$$

Împărţim în fiecare relaţie de mai sus cu (t_k) şi prin trecere la limită avem

$$\nabla f(\overline{x})(d) \leq 0,\ \nabla g_i(\overline{x})(d) \leq 0,\ \forall i \in A(\overline{x}),\ \nabla h_j(\overline{x})(d) = 0,\ \forall j \in \overline{1,m}.$$

Cum, din condiţiile Fritz John, avem

$$\lambda_0 \nabla f(\overline{x})(d) + \sum_{i\in A(\overline{x})} \lambda_i \nabla g_i(\overline{x})(d) + \sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x})(d) = 0,$$

deducem că

$$\lambda_0 \nabla f(\overline{x})(d) = 0,\ \lambda_i \nabla g_i(\overline{x})(d) = 0,\ \forall i \in A(\overline{x}),\ \mu_j \nabla h_j(\overline{x})(d) = 0,\ \forall j \in \overline{1,m}.$$

Din ipoteza că mulţimea de vectori

$$\left\{\lambda_0 \nabla f(\overline{x}), \left(\lambda_i \nabla g_i(\overline{x})\right)_{i\in A(\overline{x})}, \left(\mu_j \nabla h_j(\overline{x})\right)_{j=\overline{1,m}}\right\}$$

formează un sistem de generatori pentru $\mathbb{R}^p$, obţinem că $d = 0$, ceea ce este imposibil, întrucât $\|d\| = 1$. □

2.6 Condiţii de optimalitate de ordinul al doilea

Condiţiile de optimalitate studiate până acum în cazul cu restricţii sunt condiţii de ordinul I întrucât implică doar diferenţialele de ordinul întâi ale funcţiilor ce definesc problema. Uneori, aceste condiţii nu rezolvă complet problema mai ales atunci când avem puncte care îndeplinesc condiţiile necesare Karush-Kuhn-Tucker, dar nu sunt soluţii globale. A decide dacă respectivele puncte sunt sau nu soluţii (locale) este o chestiune la care condiţiile menţionate nu pot răspunde. Studiem mai întâi un exemplu care ilustrează aceste comentarii.

Exemplul 2.6.1 Fie problema minimizării funcţiei de două variabile reale $f(x,y) = -\frac{1}{2}(x+1)^2 - \frac{1}{2}(y+1)^2$ cu restricţiile $g_1(x,y) = x^2+y^2-2 \leq 0$, $g_2(x,y) = y-1 \leq 0$. Se constată cu uşurinţă că mulţimea punctelor fezabile este compactă, deci problema are soluţie globală. Sistemul de restricţii este convex şi are loc condiţia Slater. Se aplică Teorema Karush-Kuhn-Tucker şi obţinem că posibilele soluţii ale problemei sunt

$$\begin{aligned}&(1,1) \text{ cu multiplicatorii } (\lambda_1,\lambda_2) = (1,0)\\ &(-1,1) \text{ cu multiplicatorii } (\lambda_1,\lambda_2) = (0,2)\\ &(-1,-1) \text{ cu multiplicatorii } (\lambda_1,\lambda_2) = (0,0).\end{aligned}$$

Calculul valorilor funcţiei în aceste puncte arată că $(1,1)$ este soluţie globală.

Pentru celelate două puncte nu putem însă decide.

Să încercăm să aplicăm Teorema 2.5.2 pentru $(-1,1)$. Avem:

$$\begin{aligned}A(-1,1) = \{1,2\}, \lambda_0 = 1, \lambda_1 = 0, \lambda_2 = 2,\\ \nabla f(-1,1) = (0,-2), \nabla g_1(-1,1) = (-2,2), \nabla g_2(-1,1) = (0,1).\end{aligned}$$

Este uşor de văzut că vectorii $\lambda_0 \nabla f(-1,1), \lambda_1 \nabla g_1(-1,1), \lambda_2 \nabla g_2(-1,1)$ nu formează un sistem de generatori pentru $\mathbb{R}^2$, deci teorema nu se poate aplica.

Pentru $(-1,-1)$ avem:

$$\begin{aligned}A(-1,-1) = \{1\}, \lambda_0 = 1, \lambda_1 = 0,\\ \nabla f(-1,-1) = (0,0), \nabla g_1(-1,-1) = (-2,-2),\end{aligned}$$

şi nici de această dată $\lambda_0 \nabla f(-1,-1), \lambda_1 \nabla g_1(-1,-1)$ nu formează un sistem de generatori pentru $\mathbb{R}^2$.

Astfel, pentru cele două puncte în discuţie va trebui să aplicăm alte rezultate.

În continuarea acestei secţiuni obţinem condiţii de optim de ordinul al doilea pentru problema cu restricţii funcţionale, aşa încât vom presupune că funcţiile sunt de clasă C^2.

Teorema 2.6.2 (Condiţie necesară de ordinul II) *Presupunem că $\overline{x}$ este punct soluţie pentru problema (P) şi că au loc condiţiile Karush-Kuhn-Tucker cu multiplicatorii λ, μ. Dacă gradienţii activi,*

$$\nabla g_i(\overline{x}),\ i \in A(\overline{x}),\ \nabla h_j(\overline{x}),\ j \in \overline{1,m}$$

sunt liniar independenţi, atunci pentru orice direcţie $u \in \mathbb{R}^p$ *ce satisface*

$$\nabla g_i(\overline{x})\,(u) = 0,\ \forall i \in A(\overline{x}), \quad \nabla h_j(\overline{x})\,(u) = 0,\ \forall j \in \overline{1,m} \tag{2.4}$$

avem

$$\nabla_x^2 L(\overline{x},\lambda,\mu)(u,u) \geq 0.$$

Demonstraţie. Considerăm funcţia $\Phi : \mathbb{R}^p \to \mathbb{R}^{\operatorname{card} A(\overline{x})+m}$ ce are drept componente funcţiile g_i, $i \in A(\overline{x})$, h_j, $j = \overline{1,m}$. În baza condiţiei de liniară independenţă a gradienţilor activi, rezultă că $\nabla\Phi(\overline{x})$ este surjectiv, deci în baza Teoremei lui Lyusternik vom avea că

$$T(\Phi^{-1}(0),\overline{x}) = \operatorname{Ker}\nabla\Phi(\overline{x}).$$

Atunci orice vector u ce satisface (2.4), fiind din $\operatorname{Ker}\nabla\Phi(\overline{x})$, va fi din $T(\Phi^{-1}(0),\overline{x})$, adică vor exista $(t_k) \subset (0,\infty)$, $t_k \to 0$ şi $u_k \to u$ astfel încât $\overline{x} + t_k u_k \in \Phi^{-1}(0)$ pentru orice k, deci $g_i(\overline{x} + t_k u_k) = 0$, pentru orice $i \in A(\overline{x})$ şi $h_j(\overline{x} + t_k u_k) = 0$, pentru orice $j = \overline{1,m}$. În particular, pentru orice k suficient de mare, $\overline{x} + t_k u_k$ este punct fezabil. Dar

$$L(\overline{x},\lambda,\mu) = f(\overline{x}) + \sum_{i=1}^{n}\lambda_i g_i(\overline{x}) + \sum_{j=1}^{m}\mu_j h_j(\overline{x}) = f\,(\overline{x})$$

$$L(\overline{x}+t_k u_k,\lambda,\mu) = f(\overline{x}+t_k u_k) + \sum_{i=1}^{n}\lambda_i g_i(\overline{x}+t_k u_k) + \sum_{j=1}^{m}\mu_j h_j(\overline{x}+t_k u_k) = f\,(\overline{x}+t_k u_k)\,,\ \forall k.$$

Deci, există un şir (α_k) cu limita 0 astfel încât pentru k suficient de mare,

$$\begin{aligned}
0 &\leq f(\overline{x}+t_k u_k) - f(\overline{x}) \\
&= L(\overline{x}+t_k u_k,\lambda,\mu) - L(\overline{x},\lambda,\mu) \\
&= t_k \nabla_x L(\overline{x},\lambda,\mu)\,(u_k) + \frac{t_k^2}{2}\nabla_x^2 L(\overline{x},\lambda,\mu)(u_k,u_k) + t_k^2\left\|u_k\right\|^2\alpha_k \\
&= t_k^2\left[\frac{1}{2}\nabla_x^2 L(\overline{x},\lambda,\mu)(u_k) + \left\|u_k\right\|^2\alpha_k\right].
\end{aligned}$$

Împărţind la t_k^2 şi trecând la la limită, având în vedere şi faptul că $\left\|u_k\right\|^2\alpha_k \to 0$, obţinem concluzia dorită. □

Putem obţine chiar o rafinare a acestui rezultat făcând observaţia că pot exista restricţii cu inegalităţi active pentru care multiplicatorul să fie 0. Considerăm mulţimea indicilor activi strict dată prin

$$A_s\,(\overline{x}) = \{i \in A\,(\overline{x}) \mid \lambda_i > 0\}\,.$$

Corolarul 2.6.3 *Presupunem că* $\overline{x}$ *este punct soluţie pentru problema* (P) *şi că au loc condiţiile Karush-Kuhn-Tucker cu multiplicatorii* λ,μ. *Dacă vectorii*

$$\nabla g_i(\overline{x}),\ i \in A_s(\overline{x}),\ \nabla h_j(\overline{x}),\ j \in \overline{1,m}$$

sunt liniar independenţi, atunci pentru orice direcţie u ce satisface

$$\nabla g_i(\overline{x})\,(u) = 0,\ \forall i \in A_s(\overline{x}), \quad \nabla h_j(\overline{x})\,(u) = 0,\ \forall j \in \overline{1,m}$$

avem

$$\nabla_x^2 L(\overline{x}, \lambda, \mu)(u,u) \geq 0.$$

Demonstraţie. Se repetă demonstraţia anterioară, considerând doar acele restricţii g_i pentru care multiplicatorii sunt strict pozitivi. □

Exemplul 2.6.4 Revenim la discutarea punctelor ce satisfac condiţiile Karusk-Kuhn-Tucker din Exemplul 2.6.1 pentru care nu am putut decide dacă sunt soluţii. Lagrangianul este

$$L\left((x,y),\lambda_1,\lambda_2\right) = -\frac{1}{2}(x+1)^2 - \frac{1}{2}(y+1)^2 + \lambda_1\left(x^2+y^2-2\right) + \lambda_2(y-1).$$

Avem că $\nabla_x^2 L((x,y),\lambda_1,\lambda_2)$ se identifică cu matricea hessiană corespunzătoare care este $(2\lambda_1 - 1)\,I_2$.

Începem cu punctul $(-1,1)$ pentru care avem $A(-1,1) = \{1,2\}, \lambda_1 = 0, \lambda_2 = 2$.

Dorim să aplicăm Teorema 2.6.2. Cum $\nabla g_1(-1,1) = (-2,2), \nabla g_2(-1,1) = (0,1)$ sunt liniar independenţi, căutăm $u = (u_1,u_2) \in \mathbb{R}^2$ pentru care

$$\nabla g_1(-1,1)\,(u) = 0, \nabla g_2(-1,1)\,(u) = 0.$$

Singurul vector care satisface aceste condiţii este vectorul nul. Evident, pentru acest vector, are loc condiţia din concluzia teoremei, deci în continuare nu putem să decidem.

Încercăm aplicarea Corolarului 2.6.3. Avem $A_s(-1,1) = \{2\}$ şi $\nabla g_2(-1,1) = (0,1) \neq (0,0)$ satisface condiţia de liniară independenţă. Căutăm $u = (u_1,u_2) \in \mathbb{R}^2$ pentru care $\nabla g_2(-1,1)\,(u) = 0$ şi găsim mulţimea vectorilor $\{u \in \mathbb{R}^2 \mid u_1 \in \mathbb{R} \text{ şi } u_2 = 0\}$. Cum

$$\nabla^2_{(x,y)} L((-1,1),\lambda_1,\lambda_2)\,(u,u) = -u_1^2 - u_2^2,$$

este evident că nu are loc concluzia din Teorema 2.6.3. Astfel, punctul în cauză nu este soluţie.

Pentru $(-1,-1)$. Avem $A(-1,-1) = \{1\}$, $\nabla g_1(-1,-1) \neq (0,0)$, iar vectorii $u = (u_1,u_2) \in \mathbb{R}^2$ pentru care $\nabla g_1(-1,-1)\,(u) = 0$ sunt $\{u \in \mathbb{R}^2 \mid u_1 + u_2 = 0\}$. Cum şi în acest caz

$$\nabla^2_{(x,y)} L((-1,-1),\lambda_1,\lambda_2)\,(u,u) = -u_1^2 - u_2^2,$$

este evident că nu are loc concluzia din Teorema 2.6.2. Astfel, punctul $(-1,-1)$ nu este soluţie.

Exemplul 2.6.5 Similar, punctele rămase în discuţie la Exemplul 2.31 nu sunt puncte de minim local pentru că nu verifică condiţiile necesare de ordinul al doilea.

Prezentăm în continuare o condiţie suficientă de ordinul al doilea.

Teorema 2.6.6 (Condiţie suficientă de ordinul II) *Presupunem că $\overline{x}$ este un punct fezabil pentru problema (P) ce satisface condiţiile Karush-Kuhn-Tucker cu multiplicatorii λ, μ. Dacă*

$$\nabla_x^2 L(\overline{x}, \lambda, \mu)(u,u) > 0$$

pentru orice $u \in \mathbb{R}^p \setminus \{0\}$ *ce satisface*

$$\begin{aligned} &\nabla g_i(\overline{x})(u) \leq 0,\ \forall i \in A(\overline{x}) \backslash A_s(\overline{x}), \\ &\nabla g_i(\overline{x})(u) = 0,\ \forall i \in A_s(\overline{x}), \\ &\nabla h_j(\overline{x})(u) = 0,\ \forall j \in \overline{1,m}, \end{aligned} \tag{2.5}$$

atunci $\overline{x}$ *este soluţie strictă de ordin* 2 *pentru* (P).

Demonstraţie. Să presupunem, prin reducere la absurd, că $\overline{x}$ nu este soluţie strictă de ordin 2 pentru (P). Atunci, pentru orice $k \in \mathbb{N}^*$, vom găsi $x_k \in B\left(\overline{x}, \frac{1}{k}\right) \cap M$ astfel încât

$$f(x_k) < f(\overline{x}) + \frac{1}{k} \|x_k - \overline{x}\|^2.$$

Observăm că $x_k \neq \overline{x}$ pentru orice k. Definind $u_k = \dfrac{x_k - \overline{x}}{\|x_k - \overline{x}\|}$, putem presupune, trecând eventual la un subşir, că $u_k \to u$, cu $\|u\| = 1$. Cum $\nabla_x L(\overline{x}, \lambda, \mu) = 0$ din condiţiile Karush-Kuhn-Tucker, raţionând ca mai sus, găsim un şir (α_k) cu limita 0 astfel încât pentru orice k suficient de mare,

$$\begin{aligned} \frac{1}{k} \|x_k - \overline{x}\|^2 &> f(x_k) - f(\overline{x}) \geq L(x_k, \lambda, \mu) - L(\overline{x}, \lambda, \mu) \\ &= \nabla_x L(\overline{x}, \lambda, \mu)(x_k - \overline{x}) + \frac{1}{2} \nabla_x^2 L(\overline{x}, \lambda, \mu)(x_k - \overline{x}, x_k - \overline{x}) + \|x_k - \overline{x}\|^2 \alpha_k \\ &= \frac{1}{2} \nabla_x^2 L(\overline{x}, \lambda, \mu)(x_k - \overline{x}, x_k - \overline{x}) + \|x_k - \overline{x}\|^2 \alpha_k, \end{aligned}$$

de unde, împărţind la $\|x_k - \overline{x}\|^2$ şi trecând la limită pentru $k \to \infty$, obţinem

$$\nabla_x^2 L(\overline{x}, \lambda, \mu)(u, u) \leq 0.$$

Pe de altă parte, există din nou nişte şiruri (β_k), (γ_k^i), $\left(\varepsilon_k^j\right)$ $\left(\text{unde } i \in A(\overline{x}),\ j \in \overline{1,m}\right)$ cu limita 0 astfel încât

$$\begin{aligned} &\frac{1}{k} \|x_k - \overline{x}\|^2 > f(x_k) - f(\overline{x}) = \nabla f(\overline{x})(x_k - \overline{x}) + \|x_k - \overline{x}\| \beta_k, \\ &0 \geq g_i(x_k) - g_i(\overline{x}) = \nabla g_i(\overline{x})(x_k - \overline{x}) + \|x_k - \overline{x}\| \gamma_k,\ \forall i \in A(\overline{x}), \\ &0 = h_j(x_k) - h_j(\overline{x}) = \nabla h_j(\overline{x})(x_k - \overline{x}) + \|x_k - \overline{x}\| \varepsilon_k,\ \forall j \in \overline{1,m}. \end{aligned}$$

Împărţind inegalităţile prin $\|x_k - \overline{x}\|$ şi trecând la limită pentru $k \to \infty$, obţinem

$$\begin{aligned} &\nabla f(\overline{x})(u) \leq 0, \\ &\nabla g_i(\overline{x})(u) \leq 0,\ \forall i \in A(\overline{x}), \\ &\nabla h_j(\overline{x})(u) = 0,\ \forall j \in \overline{1,m}. \end{aligned}$$

Dacă ar exista un $i \in A_s(\overline{x})$ astfel încât $\nabla g_i(\overline{x})\,(u) < 0$, atunci, multiplicând relaţiile anterioare respectiv cu $1, \lambda_i, i \in A(\overline{x}), \mu_j, j \in \overline{1,m}$ şi adunându-le, obţinem

$$\begin{aligned} 0 = \nabla_x L(\overline{x},\lambda,\mu)(u) &= \left(\nabla f(\overline{x}) + \sum_{i=1}^{n}\lambda_i \nabla g_i(\overline{x}) + \sum_{j=1}^{m}\mu_j \nabla h_j(\overline{x})\right)(u) \\ &= \nabla f(\overline{x})\,(u) + \sum_{i\in A(\overline{x})}\lambda_i \nabla g_i(\overline{x})\,(u) < 0, \end{aligned}$$

ceea ce este imposibil. Deci u satisface (2.5), iar $\nabla_x^2 L(\overline{x},\lambda,\mu)(u,u) \leq 0$, ceea ce ne dă o contradicţie cu ipotezele teoremei. Prin urmare, presupunerea făcută este falsă, iar teorema este complet demonstrată. □

Obţinem următorul corolar.

Corolarul 2.6.7 *Presupunem că $\overline{x}$ este un punct fezabil pentru problema (P) ce satisface condiţiile Karush-Kuhn-Tucker cu multiplicatorii λ, μ, iar $\lambda_i > 0$ pentru orice $i \in A(\overline{x})$. Dacă $\nabla_x^2 L(\overline{x};\lambda,\mu)$ e pozitiv definită pe subspaţiul*

$$\left\{u \mid \nabla g_i(\overline{x})\,(u) = 0,\ i \in A(\overline{x}),\ \nabla h_j(\overline{x})\,(u) = 0,\ j \in \overline{1,m}\right\},$$

atunci $\overline{x}$ este soluţie strictă de ordin 2 pentru (P).

Exemplul 2.6.8 (Utilizarea condiţiilor de optimalitate) Fie problema minimizării funcţiei de două variabile reale $f(x,y) = -(x+1)^2 - (y+1)^2$ cu restricţiile $g_1(x,y) = x^2 + y^2 - 3 \leq 0$, $g_2(x,y) = -x^2 + 2y \leq 0$. Se constată cu uşurinţă că mulţimea punctelor fezabile este compactă, deci problema are soluţie globală. O imagine geometrică asupra problemei oferă unele informaţii importante, dar dorim să utilizăm condiţiile analitice pe care le avem la dispoziţie şi care funcţionează şi în contexte în care nu putem vizualiza structura problemei sau nu putem extrage informaţii din aceasta. Se verifică condiţia de liniară independenţă în toate punctele fezabile. Se aplică Teorema Karush-Kuhn-Tucker şi obţinem că posibilele soluţii are problemei sunt

$$\begin{aligned} &(-1,-1) \text{ cu multiplicatorii } (\lambda_1,\lambda_2) = (0,0), \\ &\left(\sqrt{2},1\right) \text{ cu multiplicatorii } (\lambda_1,\lambda_2) = \left(\frac{3}{2} + \frac{1}{2\sqrt{2}}, \frac{1}{2} - \frac{1}{2\sqrt{2}}\right), \\ &\left(-\sqrt{2},1\right) \text{ cu multiplicatorii } (\lambda_1,\lambda_2) = \left(\frac{3}{2} - \frac{1}{2\sqrt{2}}, \frac{1}{2} + \frac{1}{2\sqrt{2}}\right), \\ &\left(-\sqrt{\frac{3}{2}},-\sqrt{\frac{3}{2}}\right) \text{ cu multiplicatorii } (\lambda_1,\lambda_2) = \left(1 - \sqrt{\frac{2}{3}}, 0\right), \\ &(a,b) \text{ cu multiplicatorii } (\lambda_1,\lambda_2) = (0,\theta), \\ &\text{unde } \theta \text{ este singura rădăcină reală a ecuaţiei } \theta^3 + \theta^2 - \theta - \frac{3}{2} = 0, \\ &a = -\frac{1}{\theta+1},\ b = \theta - 1. \end{aligned}$$

Studiind ecuaţia $\theta^3+\theta^2-\theta-\frac{3}{2}=0$ se poate constata că unica sa soluţie reală este în intervalul $(1,1+8^{-1})$. La fel, $b\in(0,8^{-1})$, $a\in\left(-2^{-1},-\frac{8}{17}\right)$.

Prin calculul direct a valorilor funcţiei obiectiv în aceste puncte, deducem că $\left(\sqrt{2},1\right)$ este soluţia globală.

Considerăm acum punctul $\left(-\sqrt{2},1\right)$ şi încercăm să verificăm concluzia Teoremei 2.5.2. Ambele restricţii sunt active, $\lambda_0=1$, $\lambda_1=\frac{3}{2}-\frac{1}{2\sqrt{2}}$, $\lambda_2=\frac{1}{2}+\frac{1}{2\sqrt{2}}$. Astfel, e evident că vectorii

$$\begin{aligned}\nabla f\left(-\sqrt{2},1\right)&=\left(-2\left(-\sqrt{2}+1\right),-4\right),\\ \lambda_1\nabla g_1\left(-\sqrt{2},1\right)&=2\left(\frac{3}{2}-\frac{1}{2\sqrt{2}}\right)\left(-\sqrt{2},1\right),\\ \lambda_2\nabla g_2\left(-\sqrt{2},1\right)&=\left(\frac{1}{2}+\frac{1}{2\sqrt{2}}\right)\left(2\sqrt{2},2\right)\end{aligned}$$

formează un sistem de generatori pentru $\mathbb{R}^2$, deci sunt satisfăcute condiţiile suficiente. Astfel, punctul $\left(-\sqrt{2},1\right)$ este soluţie locală a problemei.

Continuăm cu punctul $(-1,-1)$. Cum multiplicatorii sunt nuli, nu vom avea suficienţi vectori nenuli printre cei ce trebuie consideraţi în Teorema 2.5.2 (ar trebui cel puţin doi) care să formeze un sistem de generatori pentru $\mathbb{R}^2$. Deci, în acest caz nu putem decide.

Pentru punctul $\left(-\sqrt{\frac{3}{2}},-\sqrt{\frac{3}{2}}\right)$ doar prima inegalitate este activă. Avem vectorii

$$\begin{aligned}\nabla f\left(-\sqrt{\frac{3}{2}},-\sqrt{\frac{3}{2}}\right)&=-2\left(1-\sqrt{\frac{3}{2}},1-\sqrt{\frac{3}{2}}\right)\\ \lambda_1\nabla g_1\left(-\sqrt{\frac{3}{2}},-\sqrt{\frac{3}{2}}\right)&=2\left(1-\sqrt{\frac{2}{3}}\right)\left(-\sqrt{\frac{3}{2}},-\sqrt{\frac{3}{2}}\right)\end{aligned}$$

care nu sunt liniari independenţi, deci nu formează sistem de generatori. Nici pentru acest punct nu putem decide deocamdată.

În sfârşit, luăm punctul (a,b) de mai sus pe care îl cunoaştem doar cu aproximaţie. A doua restricţie este activă în acest punct. Ţinând seama de relaţiile dintre a,b şi θ avem vectorii

$$\begin{aligned}\nabla f(a,b)&=-2(a+1,\theta)\\ \lambda_2\nabla g_2(a,b)&=\theta(-2a,2)=2\theta(-a,1)\end{aligned}$$

care nu sunt liniari independenţi. Din nou, nu putem decide.

Trecem la investigarea condiţiilor de ordinul al doilea pentru punctele de mai sus.

Pentru $(-1,-1)$ ambele restricţii sunt inactive şi deci ne interesează hessiana lui f în acest punct care este $-2I_2$. Pe baza condiţiilor necesare de ordinul al doilea pentru probleme fără restricţii concluzionăm că de fapt $(-1,-1)$ este punct de maxim pentru f, deci nu este soluţie a problemei considerate.

Pentru $\left(-\sqrt{\frac{3}{2}},-\sqrt{\frac{3}{2}}\right)$, cum $\nabla g_1\left(-\sqrt{\frac{3}{2}},-\sqrt{\frac{3}{2}}\right)\neq(0,0)$ are loc condiţia de liniară independenţă din Teorema 2.6.2. Mulţimea direcţiilor ce trebuie considerate este

$$\left\{u\in\mathbb{R}^2\mid\nabla g_1\left(-\sqrt{\frac{3}{2}},-\sqrt{\frac{3}{2}}\right)(u)=0\right\}=\left\{u\in\mathbb{R}^2\mid u_1+u_2=0\right\}.$$

Matricea asociată lui $\nabla^2_{(x,y)} L\left(\left(-\sqrt{\frac{3}{2}}, -\sqrt{\frac{3}{2}}\right), \left(1-\sqrt{\frac{2}{3}}, 0\right)\right)$ este $-2\sqrt{\frac{2}{3}} I_2$, deci nu este pozitiv semidefinită pe mulţimea de mai sus. Deducem că $\left(-\sqrt{\frac{3}{2}}, -\sqrt{\frac{3}{2}}\right)$ nu este soluţie.

Pentru (a,b), $\nabla g_2(a,b) \neq (0,0)$, deci loc condiţia de liniară independenţă din Teorema 2.6.2. Mulţimea direcţiilor pe care le avem în vedere este

$$\left\{u \in \mathbb{R}^2 \mid \nabla g_2(a,b)(u) = 0\right\} = \left\{u \in \mathbb{R}^2 \mid -au_1 + u_2 = 0\right\}.$$

Matricea asociată lui $\nabla^2 L_{(x,y)}((a,b),(0,\theta))$ este

$$\begin{pmatrix} 2(-\theta-1) & 0 \\ 0 & -2 \end{pmatrix},$$

deci nu este pozitiv semidefinită pe mulţimea direcţiilor de mai sus. Deducem că (a,b) nu este soluţie.

În sfârşit mai facem observaţia că punctelor ce sunt soluţie, adică $\left(\sqrt{2}, 1\right)$ şi $\left(-\sqrt{2}, 1\right)$, li se poate aplica Teorema 2.6.6 pentru a concluziona că sunt de fapt soluţii stricte de ordin 2.

Exemplul 2.6.9 Fie problema minimizării expresiei $-x^3y$ cu restricţia $x + y \leq 6$. Mai întâi, se observă că problema nu are soluţie globală. În orice caz, Teorema Karush-Kuhn-Tucker este aplicabilă. Rezolvând sistemul Karush-Kuhn-Tucker obţinem următoarele puncte: $(0,y)$ cu $y \leq 6$ şi $\lambda = 0$; $\left(\frac{9}{2}, \frac{3}{2}\right)$ cu $\lambda = \left(\frac{9}{2}\right)^3$. Pentru punctele de forma $(0,y)$ cu $y \leq 6$, valoarea funcţiei obiectiv este 0 şi se observă că mereu găsim un şir de puncte fezabile convergent la punctul în cauză pentru care valorile funcţiei obiectiv sunt pozitive. Deci acestea nu sunt puncte de extrem. Cercetăm acum natura punctului $\left(\frac{9}{2}, \frac{3}{2}\right)$ prin studiul condiţiei suficiente de ordinul al doilea: Teorema 2.6.6. Mulţimea direcţiilor $u \neq 0$ care satisfac ipotezele acestei teorema este $\{u \in \mathbb{R}^2 \setminus \{0\} \mid u_1 + u_2 = 0\}$. Avem pentru orice x şi orice astfel de direcţie

$$\nabla^2_x L(x,\lambda)(u,u) = -6xyu_1^2 - 6x^2u_1u_2 = 6xu_1^2(x-y),$$

deci

$$\nabla^2_x L(\left(\frac{9}{2}, \frac{3}{2}\right), \lambda)(u,u) > 0.$$

Astfel $\left(\frac{9}{2}, \frac{3}{2}\right)$ este soluţie locală a problemei.

Exerciţiul 2.36 *Să se studieze problema minimizării şi problema maximizării locale a expresiei xyz cu restricţia $xy + xz + yz = 8$.*

Exerciţiul 2.37 *Să se studieze problema minimizării şi problema maximizării locale a expresiei xyz cu restricţia $x + y + z = 1$.*

Exerciţiul 2.38 *Să se studieze problema minimizării şi problema maximizării locale a expresiei $y^4 - x^4$ cu restricţia $x - y^2 \leq 0$.*

Capitolul 3

Solvabilitate şi aproximare

3.1 Solvabilitatea problemelor de optimizare

Fie $P \in \mathbb{R}[X]$ un polinom de grad mai mare sau egal decât 5. Ecuaţia $P(x) = 0$ cu necunoscuta $x \in \mathbb{R}$ poate fi redusă la problema de optimizare fără restricţii $\min P^2(x)$ cu $x \in \mathbb{R}$. Aşa cum ştim, ecuaţia algebrică nu poate fi rezolvată în general, deci nici problema de optimizare asociată nu poate fi rezolvată.

De fapt, în multe cazuri, rezolvarea unei probleme de optimizare se rezumă la rezolvarea unei ecuaţii algebrice sau a unui sistem de ecuaţii (de exemplu, pe baza Teoremei lui Fermat sau a Teoremei Karush-Kuhn-Tucker) care nu pot fi rezolvate prin metode directe.

Să discutăm un alt model general. În multe probleme, în multe domenii ale matematicii, avem următorul model: considerăm $f_i : \mathbb{R}^p \to \mathbb{R}$ şi $a_i \in \mathbb{R}$ cu $i \in \overline{1,q}$; cu aceste date, problema este aceea de a găsi $x \in \mathbb{R}^p$ astfel încât

$$f_i(x) = a_i, \ \forall i \in \overline{1,q}. \tag{3.1}$$

Din nou, această problemă poate fi uşor convertită într-o problemă de optimizare fără restricţii:

$$\min \sum_{i=1}^{q} \left(f_i(x) - a_i\right)^2.$$

Bineînţeles, dacă ar exista o metodă generală de a rezolva problema de optimizare de mai sus atunci toate problemele de forma (3.1) ar fi rezolvabile. Dar ştim că aceasta este imposibil. Deci, în general, problemele de optimizare nu sunt rezolvabile. Clasa problemelor de tipul celor de mai sus este prea vastă pentru exista o metodă universală care să le rezolve pe toate. În afară de aceste aserţiuni care sunt mai degrabă euristice, vom arăta riguros că, într-adevăr, aşa stau lucrurile.

Pentru a depăşi această dificultate de principiu vom descrie două posibile abordări. Prima dintre ele este de matură teoretică şi va servi drept bază pentru pregătirea celei de-a doua. Astfel, discutăm mai întâi despre puncte de extrem aproximativ iar apoi vom începe studiul unor proceduri numerice care, prin excelenţă, se referă la astfel de puncte.

3.2 Principiul variaţional al lui Ekeland

Definiţia 3.2.1 *Fie $M \subset \mathbb{R}^p$ o mulţime închisă, $f : M \to \mathbb{R}$ o funcţie mărginită inferior şi $\varepsilon > 0$. Un punct x_ε se numeşte $\varepsilon-$minim pentru f pe M dacă*

$$\inf_{x\in M} f(x) \leq f(x_\varepsilon) \leq \inf_{x\in M} f(x) + \varepsilon.$$

Cum spuneam, faptul că $\inf_{x\in M} f(x) \in \mathbb{R}$ asigură existenţa punctelor de minim aproximativ. Dăm acum o foarte importantă teoremă, numită Principiul Variaţional al lui Ekeland, care, în esenţă, spune că în apropierea unui punct de minim aproximativ putem găsi un punct de minim veritabil pentru o perturbare a funcţiei iniţiale. Acest rezultat a fost stabilit de către matematicianul francez Ivar Ekeland în 1974.

Teorema 3.2.2 (Principiul Variaţional al lui Ekeland) *Fie $M \subset \mathbb{R}^p$ o mulţime închisă şi $f : M \to \mathbb{R}$ o funcţie continuă şi mărginită inferior. Fie $\varepsilon > 0$ şi x_ε un $\varepsilon-$minim pentru f pe M Atunci, pentru orice $\delta > 0$ există $\overline{x}_\varepsilon \in M$ cu proprietăţile:*

$$\begin{aligned} &f(\overline{x}_\varepsilon) \leq f(x_\varepsilon) \\ &\|\overline{x}_\varepsilon - x_\varepsilon\| \leq \delta \\ &f(\overline{x}_\varepsilon) \leq f(x) + \varepsilon\delta^{-1} \|x - \overline{x}_\varepsilon\|, \ \forall x \in M. \end{aligned}$$

Demonstraţie Fie $g : M \to \mathbb{R}$

$$g(x) = f(x) + \varepsilon\delta^{-1} \|x - x_\varepsilon\|.$$

Folosind ipoteza, deducem că g este continuă şi mărginită inferior. Mai mult, dacă M este nemărginită

$$\lim_{\|x\|\overset{M}{\to}\infty} g(x) = +\infty.$$

Aşadar, funcţia g admite un punct de minim global pe M pe care îl notăm cu $\overline{x}_\varepsilon$. Deci,

$$f(\overline{x}_\varepsilon) + \varepsilon\delta^{-1} \|\overline{x}_\varepsilon - x_\varepsilon\| \leq f(x) + \varepsilon\delta^{-1} \|x - x_\varepsilon\|, \ \forall x \in M. \tag{3.2}$$

Pentru $x = x_\varepsilon$, obţinem

$$f(\overline{x}_\varepsilon) \leq f(x_\varepsilon),$$

adică prima relaţie căutată. Pe de altă parte, din faptul că

$$f(x_\varepsilon) \leq \inf_{x\in M} f(x) + \varepsilon,$$

relaţia (3.2), pentru $x = x_\varepsilon$, implică

$$\delta^{-1} \|\overline{x}_\varepsilon - x_\varepsilon\| \leq 1,$$

a doua dintre relaţiile anunţate. Tot din (3.2), scriem succesiv, pentru orice $x \in M$,

$$\begin{aligned} f(\overline{x}_\varepsilon) &\leq f(x) + \varepsilon\delta^{-1}(\|x - x_\varepsilon\| - \|\overline{x}_\varepsilon - x_\varepsilon\|) \\ &\leq f(x) + \varepsilon\delta^{-1} \|x - \overline{x}_\varepsilon\|, \end{aligned}$$

adică ultima parte a concluziei. □

În cazul $M = \mathbb{R}^p$ obţinem următoarea consecinţă.

Corolarul 3.2.3 *Fie $f : \mathbb{R}^p \to \mathbb{R}$ o funcţie continuă şi mărginită inferior. Fie $\varepsilon > 0$ şi x_ε un $\varepsilon-$minim pentru f (pe $\mathbb{R}^p$) Atunci, pentru orice $\delta > 0$ există $\overline{x}_\varepsilon \in \mathbb{R}^p$ cu proprietăţile:*

$$\begin{aligned} &f(\overline{x}_\varepsilon) \leq f(x_\varepsilon) \\ &\|\overline{x}_\varepsilon - x_\varepsilon\| \leq \delta \\ &f(\overline{x}_\varepsilon) \leq f(x) + \varepsilon\delta^{-1} \|x - \overline{x}_\varepsilon\|, \ \forall x \in \mathbb{R}^p. \end{aligned}$$

Evident, în corolarul de mai sus $\overline{x}_\varepsilon$ este punct de minim global pentru funcţia

$$x \mapsto f(x) + \varepsilon\delta^{-1} \|x - \overline{x}_\varepsilon\|.$$

Pe de altă parte, dacă dorim ca $\overline{x}_\varepsilon$ să fie apropiat de x_ε (adică δ să fie mic), atunci perturbarea este mare. O cale de mijloc ar fi să alegem $\delta = \sqrt{\varepsilon}$, caz în care se obţine următorul corolar.

Corolarul 3.2.4 *Fie $f : \mathbb{R}^p \to \mathbb{R}$ o funcţie continuă şi mărginită inferior. Fie $\varepsilon > 0$ şi x_ε un $\varepsilon-$minim pentru f. Atunci există $\overline{x}_\varepsilon \in \mathbb{R}^p$ cu proprietăţile:*

$$\begin{aligned} &f(\overline{x}_\varepsilon) \leq f(x_\varepsilon) \\ &\|\overline{x}_\varepsilon - x_\varepsilon\| \leq \sqrt{\varepsilon} \\ &f(\overline{x}_\varepsilon) \leq f(x) + \sqrt{\varepsilon} \|x - \overline{x}_\varepsilon\|, \ \forall x \in \mathbb{R}^p. \end{aligned}$$

Principiul Variaţional al lui Ekeland are foarte multe aplicaţii. Unele dintre ele se referă la aceeaşi problematică a obţinerii de informaţii despre punctele de extrem. Următorul exemplu afirmă că orice funcţie diferenţiabilă admite puncte critice aproximative (pentru care norma diferenţialei este oricât de mică).

Teorema 3.2.5 *Fie $f : \mathbb{R}^p \to \mathbb{R}$ o funcţie diferenţiabilă şi mărginită inferior. Atunci pentru orice $\varepsilon, \delta > 0$ există $\overline{x}_\varepsilon \in \mathbb{R}^p$ cu $f(\overline{x}_\varepsilon) \leq \inf_{x\in\mathbb{R}^p} f(x) + \varepsilon$ şi $\|\nabla f(\overline{x}_\varepsilon)\| \leq \varepsilon\delta^{-1}$. În particular, există $(x_n) \subset \mathbb{R}^p$ cu*

$$f(x_n) \to \inf_{x\in\mathbb{R}^p} f(x), \ \nabla f(x_n) \to 0.$$

Demonstraţie Fie x_ε un $\varepsilon-$minim pentru f. Conform Teoremei 3.2.2, există $\overline{x}_\varepsilon \in \mathbb{R}^p$ cu proprietăţile respective. Din $f(\overline{x}_\varepsilon) \leq f(x_\varepsilon)$ deducem că $f(\overline{x}_\varepsilon) \leq \inf_{x\in\mathbb{R}^p} f(x) + \varepsilon$. Fie $x = \overline{x}_\varepsilon + tu$ cu $u \in \mathbb{R}^p$ şi $t > 0$. Din relaţia $f(\overline{x}_\varepsilon) \leq f(x) + \varepsilon\delta^{-1} \|x - \overline{x}_\varepsilon\|$, valabilă pentru orice $x \in \mathbb{R}^p$, avem

$$\frac{f(\overline{x}_\varepsilon + tu) - f(\overline{x}_\varepsilon)}{t} \geq -\varepsilon\delta^{-1} \|u\|.$$

Prin trecere la limită cu $t \to 0$, deducem

$$\nabla f(\overline{x}_\varepsilon)(u) \geq -\varepsilon\delta^{-1} \|u\|, \ \forall u \in \mathbb{R}^p,$$

adică

$$-\nabla f(\overline{x}_\varepsilon)(u) \leq \varepsilon\delta^{-1} \|u\|, \ \forall u \in \mathbb{R}^p.$$

Schimbând u în $-u$,

$$\nabla f(\overline{x}_\varepsilon)(u) \leq \varepsilon\delta^{-1} \|u\|, \ \forall u \in \mathbb{R}^p,$$

deci

$$|\nabla f(\overline{x}_\varepsilon)(u)| \leq \varepsilon\delta^{-1} \|u\|, \ \forall u \in \mathbb{R}^p,$$

ceea ce implică $\|\nabla f(\overline{x}_\varepsilon)\| \leq \varepsilon\delta^{-1}$. Pentru $\varepsilon := n^{-1}$, $\delta = \sqrt{n^{-1}}$, $n \in \mathbb{N}^*$ se obţine a doua parte a concluziei. □

În privinţa problemei (P) cu restricţii funcţionale investigată în secţiunile anterioare, pe baza Principiului variaţional al lui Ekeland se poate arăta rezultatul următor.

Teorema 3.2.6 *Dacă are loc condiţia de liniară independenţă iar f este mărginită inferior pe mulţimea punctelor fezabile M, atunci pentru orice $\varepsilon > 0$, există $x_\varepsilon \in M$ astfel încât $f(x_\varepsilon) \leq \inf_{x\in M} f(x) + \varepsilon^2$ şi există $\lambda = (\lambda_1, \lambda_2, ..., \lambda_n) \in \mathbb{R}^n_+$, $\mu = (\mu_1, \mu_2, ..., \mu_m) \in \mathbb{R}^m$, astfel încât*

$$\left\| \nabla f(\overline{x}) + \sum_{i=1}^{n} \lambda_i \nabla g_i(\overline{x}) + \sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x}) \right\| < \varepsilon$$

şi

$$\lambda_i \geq 0, \ \lambda_i g_i(\overline{x}) = 0, \ \textit{pentru orice } i \in \overline{1,n}.$$

Mai sus am văzut câteva rezultate care asigură existenţa punctelor de minim în condiţii de compactitate a mulţimilor de nivel. Evident, mărginirea inferioară a funcţiei este condiţie necesară pentru existenţa minimului, dar este clar că mărginirea mulţimilor de nivel nu este o astfel de condiţie. De exemplu funcţia $f : (0, \infty) \to \mathbb{R}$, $f(x) = (x-1)^2 e^{-x}$ îşi atinge minimul în $\overline{x} = 1$, valoarea minimă este 0, dar $N_\nu f$ nu este mărginită pentru nicio valoare $\nu > 0 = \inf\{f(x) \mid x \in (0, \infty)\}$. Tot din Principiul lui Ekeland, obţinem rezultatul următor, care precizează echivalenţa mai multor condiţii de existenţă a punctelor de minim.

Teorema 3.2.7 *Fie $f : \mathbb{R}^p \to \mathbb{R}$ o funcţie diferenţiabilă şi mărginită inferior. Următoarele afirmaţii sunt echivalente:*

(i) $\lim_{\|x\|\to\infty} f(x) = \infty$;

(ii) $N_\nu f$ este mărginită pentru orice $\nu > \inf_{x\in\mathbb{R}^p} f(x)$;

(iii) orice şir (x_n) pentru care $(f(x_n))$ este convergent şi $\nabla f(x_n) \to 0$ admite un subşir convergent.

Demonstraţie Implicaţia $(i) \Rightarrow (ii)$ a fost arătată în Propoziţia 1.2.5, în timp ce implicaţia $(ii) \Rightarrow (iii)$ rezultă în virtutea faptului că orice şir mărginit admite un subşir convergent. Să arătăm $(iii) \Rightarrow (i)$. Presupunem, prin reducere la absurd, că există $c \in \mathbb{R}$ şi un şir $(x_n) \subset \mathbb{R}^p$ cu $\|x_n\| \to \infty$, $f(x_n) \leq c$, pentru orice $n \in \mathbb{N}$. Evident, $c \geq \inf_{x\in\mathbb{R}^p} f(x)$. Pentru fiecare $n \in \mathbb{N}^*$ alegem

$$\varepsilon_n := c + n^{-1} - \inf_{x\in\mathbb{R}^p} f(x) > 0,$$

deci

$$f(x_n) < \inf_{x\in\mathbb{R}^p} f(x) + \varepsilon_n.$$

Fie $\delta_n := 2^{-1}\|x_n\| > 0$. Ca în Teorema 3.2.5 (şi demonstraţia sa) există $\overline{x}_n$ cu

$$\begin{aligned} &f(\overline{x}_n) \leq f(x_n) \leq \inf_{x\in\mathbb{R}^p} f(x) + \varepsilon_n \\ &\|\overline{x}_n - x_n\| \leq \delta_n \\ &\|\nabla f(\overline{x}_n)\| \leq \varepsilon_n \delta_n^{-1}. \end{aligned}$$

Dar

$$\|\overline{x}_n\| \geq \|x_n\| - \|\overline{x}_n - x_n\| \geq \|x_n\| - 2^{-1}\|x_n\| = 2^{-1}\|x_n\|,$$

deci $\|\overline{x}_n\| \to \infty$. Pe de altă parte,

$$\|\nabla f(\overline{x}_n)\| \leq \frac{2}{\|x_n\|}(c + n^{-1} - \inf_{x\in\mathbb{R}^p} f(x)) \to 0.$$

Dar $(f(\overline{x}_n))$ este mărginit, deci admite un subşir convergent. Din (iii), deducem că şi $(\overline{x}_n)$ trebuie să admită un astfel de subşir, ceea ce este imposibil. Deci (i) are loc. □

Condiţia de la (iii) din rezultatul de mai sus se numeşte condiţia Palais-Smale.

3.3 Penalizare

Printr-o metodă de penalizare înţelegem o metodă prin care o problemă de optimizare cu restricţii este transformată într-o problemă de optimizare fără restricţii prin adăugarea la funcţia obiectiv a unui termen care înglobează restricţiile. În această secţiune prezentăm pe scurt o astfel de metoda, urmând ca într-un capitol ulterior să dezvoltăm această idee.

Propoziţia 3.3.1 (Penalizarea Clarke) *Fie $f : \mathbb{R}^p \to \mathbb{R}$ o funcţie şi $M \subset \mathbb{R}^p$ o mulţime închisă. Considerăm problema*

$$\min f(x),\ x \in M.$$

Presupunem că $\overline{x}$ este o soluţie (locală) a acestei probleme şi f este local Lipschitz de constantă $L \geq 0$ în jurul lui $\overline{x}$. Să se arate că $\overline{x}$ este soluţie locală fără constrângeri pentru funcţia $x \mapsto f(x) + Ld_M(x)$.

Demonstraţie Fie U vecinătatea lui $\overline{x}$ astfel încât f este $L-$Lipschitz pe U şi $f(\overline{x}) \leq f(x)$ pentru orice $x \in U \cap M$. Atunci există $\alpha > 0$ a.î. $B(\overline{x}, \alpha) \subset U$. Considerăm $V = B(\overline{x}, \alpha/3)$ and luăm $x \in V$ arbitrar. Mai întâi, dacă $x \in V \cap M \subset U \cap M$, este clar că

$$f(\overline{x}) + Ld_M(\overline{x}) \leq f(x) + Ld_M(x).$$

Apoi, dacă $x \in V \setminus M$, atunci pentru orice $\varepsilon \in (0, \alpha/3)$ există $x_\varepsilon \in M$ a.î.

$$\|x - x_\varepsilon\| < d_M(x) + \varepsilon \leq \|x - \overline{x}\| + \varepsilon \leq \alpha/3 + \varepsilon < 2\alpha/3.$$

Deci,

$$\|x_\varepsilon - \overline{x}\| \leq \|x_\varepsilon - x\| + \|x - \overline{x}\| < 2\alpha/3 + \alpha/3 = \alpha.$$

Prin urmare, $x_\varepsilon \in U \cap M$, deci,

$$\begin{aligned} f(\overline{x}) + Ld_M(\overline{x}) &\leq f(x_\varepsilon) \leq f(x) + L\|x - x_\varepsilon\| \\ &\leq f(x) + L(d_M(x) + \varepsilon) = f(x) + Ld_M(x) + L\varepsilon. \end{aligned}$$

Făcând $\varepsilon \to 0$, se obţine concluzia. □

Propoziţia 3.3.2 (Alternativa de optimalitate) *Fie $f : \mathbb{R}^p \to \mathbb{R}$ o funcţie şi $M \subset \mathbb{R}^p$ o mulţime închisă. Consideăm problema*

$$\min f(x),\ x \in M.$$

Presupunem că $\overline{x}$ este o soluţie (locală) a acestei probleme şi f este local Lipschitz de constantă $L \geq 0$ în jurul lui $\overline{x}$. Fie $\varphi : \mathbb{R}^p \to [0, \infty)$ cu $\varphi(\overline{x}) = 0$ o funcţie inferior semicontinuă. Să se arate că are loc una şi numai una dintre următoarele posibilităţi:

(i) există $\lambda > 0$ a.î. $\overline{x}$ este o soluţie locală fără constrângeri pentru funcţia $f + \lambda\varphi$;

(ii) există $(z_n)_{n\in\mathbb{N}^*} \subset \mathbb{R}^p \setminus M$, $z_n \to \overline{x}$ a.î. pentru orice $n \in \mathbb{N}^*$ funcţia

$$x \mapsto \varphi(x) + n^{-1}\|x - z_n\|$$

îşi atinge minimul în z_n.

Demonstraţie Putem avea una şi numai una dintre situaţiile:

- există o vecinătate V a lui $\overline{x}$ şi $a > 0$ a.î. pentru orice $x \in V$

$$a\varphi(x) \geq d(x, M);$$

- există $x_n \to \overline{x}$ a.î.

$$2n\varphi(x_n) < d(x_n, M).$$

În prima situaţie, fie $\alpha > 0$ astfel încât $B(\overline{x}, \alpha) \subset U \cap V$, unde U este vecinătatea lui $\overline{x}$ pe care are loc condiţia Lipschitz şi pe care $\overline{x}$ este punct de minim pentru problema dată.

Pentru orice $x \in B(\overline{x}, \alpha/3)$ există $u \in M$ cu

$$\|x - u\| \leq 2d(x, M) \leq 2\|x - \overline{x}\|.$$

Deci, în particular,

$$\|u - \overline{x}\| \leq \|u - x\| + \|x - \overline{x}\| \leq 3\|x - \overline{x}\| < \alpha.$$

Atunci

$$f(x) \geq f(u) - L\|u - x\| \geq f(\overline{x}) - 2L\alpha\varphi(x)$$

şi deci suntem în prima din cele două posibilităţi.

În a doua situaţie, cum φ are valori pozitive, $(x_n) \subset \mathbb{R}^p \setminus M$ şi

$$\varphi(x_n) \leq \inf_{x\in\mathbb{R}^p} \varphi(x) + 2^{-1}n^{-1}d(x_n, M).$$

Luăm $\varepsilon := 2^{-1}n^{-1}d(x_n, M) > 0$ şi $\delta := 2^{-1}d(x_n, M)$ şi aplicăm Principiul Variaţional al lui Ekeland funcţiei φ pentru $\varepsilon-$minimul x_n. Deducem că există $z_n \in \mathbb{R}^p$ cu

$$\begin{aligned}
&\varphi(z_n) \leq \varphi(x_n) < n^{-1}d(x_n, M)\\
&\|z_n - x_n\| \leq 2^{-1}d(x_n, M)\\
&\varphi(z_n) \leq \varphi(x) + n^{-1}\|x - z_n\|, \ \forall x \in \mathbb{R}^p.
\end{aligned}$$

Ultima relaţie relaţie probează că pentru orice $n \in \mathbb{N}^*$ funcţia $x \mapsto \varphi(x) + n^{-1}\|x - z_n\|$ îşi atinge minimul în z_n. Apoi,

$$\begin{aligned}
\|z_n - \overline{x}\| &\leq \|z_n - x_n\| + \|x_n - \overline{x}\|\\
&\leq d(x_n, M) + \|x_n - \overline{x}\| \leq 2\|x_n - \overline{x}\|,
\end{aligned}$$

deci $z_n \to \overline{x}$. Dacă am avea $z_n \in M$, atunci am putea scrie

$$\|z_n - x_n\| \leq 2^{-1}d(x_n, M) < d(x_n, M) \leq \|z_n - x_n\|,$$

ceea ce constituie o contradicţie. Soluţia este completă. □

Să remarcăm că Problema 3.3.2 este o generalizare a Problemei 3.3.1.

Observaţia 3.3.3 *Toate funcţiile perturbate care apar atât în Principiul variaţional al lui Ekeland, cât şi în metodele de penalizare, înglobează un termen de tip distanţă, funcţie care nu este peste tot diferenţiabilă.*

3.4 Generalităţi privind metodele numerice

De foarte multe ori pot apărea probleme de optimizare pentru care nu putem rezolva sistemele ce dau punctele critice ale funcţiei obiectiv (în cazul problemelor fără restricţii) sau punctele critice ale lagrangianului (la problemele cu restricţii). Pentru astfel de probleme, sunt necesari algoritmi pentru aproximarea soluţiilor. Astfel, vom începe în capitolul următor studiul unor proceduri numerice. Ca de obicei, şi pentru proiectarea algoritmilor, există o netă diferenţă între problemele cu restricţii şi cele fără restricţii.

Trebuie precizat de la început că nu există o metodă numerică universală, ci o multitudine de metode, fiecare adaptată caracteristicilor unei clase particulare de probleme. Dacă avem o problemă (P), o vom include într-o clasă (C) de probleme având caracteristici comune. Pentru implementarea oricărei scheme numerice, trebuie să ştim o parte a problemei (P) pe care o tratăm, parte pe care o numim model şi care constă de obicei din formularea problemei, a clasei de componente funcţionale ş.a. Pentru a încadra problema (P) într-o clasă (C) şi a o rezolva, metoda trebuie să colecteze informaţii specifice despre (P). Procesul de colectare a datelor se numeşte oracol. Un oracol (O) răspunde unor întrebări succesive ale metodei. Astfel, metoda (M) încearcă să rezolve problema colectând şi prelucrând răspunsurile oracolului.

În general, fiecare problemă poate fi descrisă prin mai multe modele, aşa cum se poate de asemenea ca pentru fiecare problemă să avem diferite tipuri de oracole.

Pentru o metodă numerică, accepţiunea sintagmei "rezolvare a problemei" este aceea de determinare a unei soluţii aproximative cu o acurateţe $\varepsilon > 0$ dinainte stabilită. Această înseamnă că pentru fiecare metodă trebuie să definim un criteriu de oprire (notat (T_ε)).

Schema iterativă generală de aplicare a unei metode (M) unei probleme (P) este următoarea:

Datele de intrare: se dau un punct de start x_0 (sau mai multe date, informaţii iniţiale) şi o acurateţe (toleranţă) $\varepsilon > 0$.

Iniţializare: $k = 0$, $I_{-1} = \emptyset$, unde k este contorul de iteraţii, iar I_k este mulţimea de informaţii acumulate la pasul k.

Bucla principală: 1. Se invocă oracolul în punctul x_k, notat $O(x_k)$. 2. Se actualizează informaţiile: $I_k = I_{k-1} \cup (x_k, O(x_k))$. 3. Se aplică metoda (M) mulţimii I_k şi se obţine punctul x_{k+1}. 4. Se verifică criteriul (T_ε) : dacă are loc, se scrie răspunsul final $\overline{x} = x_{k+1}$; dacă nu, se scrie $k = k+1$ şi se merge din nou la pasul 1.

Din punct de vedere computaţional, efortul cel mai mare este în bucla principală la pasul 1 (invocarea oracolului) şi la pasul 3 (generarea punctului următor).

Astfel putem vorbi despre:

– complexitatea analitică: numărul de invocări ale oracolului necesar pentru rezolvarea problemei cu acurateţea ε;

– complexitatea aritmetică: numărul de operaţii aritmetice necesar pentru rezolvarea problemei cu acurateţea ε.

Evident, complexitatea aritmetică este mai realistică, dar de obicei este corelată cu cea analitică prin gradul de complexitate a oracolului şi metodei.

În general, pentru un model funcţional al unei probleme de optimizare putem aplica diferite tipuri de oracole:

- oracol de ordin zero: returnează valoarea funcţiilor (derivative-free, în terminologie engleză);

- oracol de ordin întâi: returnează valoarea funcţiilor şi a gradienţilor de ordinul întâi (first-order gradient methods);

- oracol de ordinul al doilea: returnează valoarea funcţiilor şi a gradienţilor de ordinul întâi şi al doilea (second-order gradient methods).

Vom aplica acest limbaj formal unei clase particulare de probleme. Fie $D \subset \mathbb{R}^p$,

$$D := \left\{ x \in \mathbb{R}^p \mid x_i \in [0,1],\ \forall i \in \overline{1,p} \right\} = [0,1]^p .$$

şi $f : \mathbb{R}^p \to \mathbb{R}$ o funcţie $L-$Lipschitz pe D în raport cu norma ∞ (reamintim că $\|x\|_\infty = \max_{i \in \overline{1,p}} |x_i|$), adică

$$|f(x) - f(y)| \leq L \|x - y\|_\infty ,\ \forall x, y \in D.$$

Considerăm o metodă foarte simplă de a rezolva problema

$$(P_L)\ \min f(x)\ \text{ cu } x \in D.$$

Evident, problema are soluţie globală (Teorema lui Weierstrass), iar metoda pe care o folosim este metoda reţelei (grilei) uniforme care constă din următoarele:

1. Pentru n dat, calculăm punctele:

$$\left(\frac{i_1}{n}, \frac{i_2}{n}, ..., \frac{i_p}{n} \right),$$

unde $(i_1, i_2, ..., i_p) \in \{0, 1, ..., n\}^p$. Sunt $(n+1)^p$ astfel de puncte.

2. Calculăm f în toate aceste puncte şi îl alegem pe cel (sau pe unul din cele) pentru care f are valoarea cea mai mică.

3. Returnăm punctul selectat, notat $\overline{x}$, şi $f(\overline{x})$.

Aşadar, această metodă împarte mulţimea punctelor fezabile într-o reţea uniformă (echidistantă), calculează minimul funcţiei în nodurile acestei reţele şi returnează această valoare. Deci, este o metodă cu oracol de ordin zero şi fără influenţa informaţei acumulate în punctele testate. Dorim să determinăm eficienţa acestei metode.

Teorema 3.4.1 *Dacă f^* este valoarea minimă globală a problemei (P_L) şi $\overline{x}$ este punctul returnat de metoda reţelei uniforme, atunci*

$$f(\overline{x}) - f^* \leq \frac{L}{2n}.$$

Demonstraţie. Fie x^* un punct de minim global pentru f, adică $f(x^*) = f^*$. Atunci x^* se află într-unul dintre cuburile generalizate ale reţelei, adică există $(i_1, i_2, ..., i_p) \in \{0, 1, ..., n\}^p$ astfel încât, pe componente, au loc relaţiile

$$x = x_{(i_1, i_2, ..., i_p)} \leq x^* \leq x_{(i_1+1, i_2+1, ..., i_p+1)} = y.$$

Evident, $y_i - x_i = \frac{1}{n}$ şi $x_i^* \in [x_i, y_i]$ pentru orice $i \in \overline{1, p}$. Considerăm punctul $u = \frac{1}{2}(x + y)$ şi definim punctul v prin

$$v_i = \begin{cases} y_i, \text{ dacă } x_i^* \geq u_i \\ x_i, \text{ în caz contrar.} \end{cases}$$

Din nou este simplu de observat că

$$|v_i - x_i^*| \leq \frac{1}{2n}, \ \forall i \in \overline{1, p}.$$

Cum v este nod al reţelei, avem

$$f(\overline{x}) - f(x^*) \leq f(v) - f(x^*) \leq L \|v - x^*\|_\infty \leq \frac{L}{2n},$$

adică concluzia. □

Cu notaţiile din teorema anterioară, pentru metoda de mai sus scopul este ca

$$f(\overline{x}) - f^* < \varepsilon.$$

Obţinem consecinţa următoare.

Corolarul 3.4.2 *Complexitatea analitică a problemei (P_L) pentru metoda reţelei uniforme este cel mult*

$$\left(\left[\frac{L}{2\varepsilon} \right] + 2 \right)^p.$$

Demonstraţie. Fie $n = \left[\frac{L}{2\varepsilon}\right] + 1$. Deci $n > \frac{L}{2\varepsilon}$ şi din teorema de mai sus avem

$$f(\overline{x}) - f^* \leq \frac{L}{2n} < \varepsilon.$$

Deci, invocând oracolul de cel mult $(n+1)^p = \left(\left[\frac{L}{2\varepsilon}\right] + 2\right)^p$ ori obţinem soluţia problemei cu acurateţea dorită. □

Este foarte interesant că are loc şi rezultatul de mai jos.

Teorema 3.4.3 *Fie $\varepsilon < \frac{L}{2}$. Atunci, pentru orice metodă cu oracol de ordin zero, complexitatea analitică a problemei (P_L) este cel puţin $\left[\frac{L}{2\varepsilon}\right]^p$.*

Demonstraţie. Fie $n = \left[\frac{L}{2\varepsilon}\right]$. Din ipoteză deducem că $n \geq 1$. Presupunem că există o metodă cu oracol de ordin zero pentru care sunt suficiente $N < n^p$ invocări ale oracolului pentru a rezolva orice problemă de tip (P_L). Fie $u \in D$ astfel încât

$$u + \frac{1}{n}e \in D,$$

unde $e = (1, 1, ..., 1)$ şi astfel încât nu există puncte de control (în care să fie invocat oracolul) în mulţimea

$$E = \left\{ x \in D \mid u \leq x \leq u + \frac{1}{n}e \right\}.$$

Fie $v = u + \frac{1}{2n}e$. Atunci

$$E = \left\{ x \in D \mid \|x - v\|_\infty \leq \frac{1}{2n} \right\}.$$

Definim funcţia $g : D \to \mathbb{R}$,

$$g(x) = \min\left\{0, L\|x - v\|_\infty - \varepsilon\right\}.$$

Observăm că are valoarea minimă $-\varepsilon$ şi diferă de 0 doar în mulţimea

$$F = \left\{ x \in D \mid \|x - v\|_\infty < \frac{\varepsilon}{L} \right\}.$$

Cum $\frac{\varepsilon}{L} \leq \frac{1}{2n}$, deducem că

$$F \subset E.$$

În particular, v nu este punct de control, iar $g(v) = -\varepsilon$. Deci, chiar dacă oracolul indică valoarea exactă a lui f în toate punctele de control ale metodei (adică 0), acurateţea nu poate fi mai bună de ε.

În final, arătăm şi că g este $L-$Lipschitz. Într-adevăr, dacă $x, y \in F$ sau $x, y \notin F$, inegalitatea din proprietatea Lipschitz este evidentă. Presupunem că $x \in F$ şi $y \notin F$. Atunci

$$\begin{aligned} |g(x) - g(y)| &= \varepsilon - L\|x - v\|_\infty = L\left(\frac{\varepsilon}{L} - \|x - v\|_\infty\right) \leq L\left(\|y - v\|_\infty - \|x - v\|_\infty\right) \\ &\leq L\|x - y\|_\infty, \end{aligned}$$

ceea ce completează demonstraţia afirmaţiei făcute.

Prin urmare, pentru a fi siguri de acurateţea dorită pentru toate problemele de tip (P_L) nu putem lua mai puţin de $\left[\frac{L}{2\varepsilon}\right]^p$ puncte. □

Exemplul 3.4.4 Evaluăm consecinţele rezultatului de mai sus în cazul în care $L = 2$, $p = 10$, $\varepsilon = 0.01$. Pentru a rezolva problema cu acurateţea ε trebuie să invocăm oracolul (să evaluăm funcţia) în cel puţin $100^{10} = 10^{20}$ puncte. Având în vedere că funcţia obiectiv are $p = 10$ funcţii coordonate, pentru calculul fiecăreia trebuind cel puţin o operaţie algebrică, înseamnă că avem de făcut cel puţin 10^{21} operaţii. Spre exemplu, dacă computerul pe care lucrăm este capabil să facă 10^{12} operaţii pe secundă, înseamnă că e nevoie de 10^9 secunde. Un an are mai puţin de $3,2 \cdot 10^7$ secunde, astfel că timpul de calcul ar fi de peste 31 de ani.

Exerciţiul 3.1 *Să se testeze metoda grilei uniforme printr-un program Matlab/Octave pentru funcţia $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^6 - 4x^5 + 3x^3 + 6x$ desenând mai întâi graficul pentru a observa poziţionarea punctelor de minim.*

Pentru funcţii de mai multe variabile, am văzut că algoritmii de ordin zero pot fi foarte costisitori din punct de vedere computaţional. Prezentăm totuşi un algoritm de tip grilă uniformă, studiul putând fi, o dată plus, instructiv.

Mai întâi prezentăm nişte funcţii test şi ne vom ocupa de reprezentări grafice. Astfel, vom ilustra codurile preponderent pe următoarele funcţii

- Funcţia lui Rosenbrock, $f : \mathbb{R}^2 \to \mathbb{R}$,
$$f(x, y) = 100(y - x^2)^2 + (1 - x)^2;$$
- $f : \mathbb{R}^2 \to \mathbb{R}$,
$$f(x, y) = x^6 + y^4 + 4x^2y^2 - 3xy + 2x + y;$$
- $f : \mathbb{R}^2 \to \mathbb{R}$,
$$f(x, y) = (1 + y)^3x^2 + y^2;$$
- $f : \mathbb{R}^2 \to \mathbb{R}$,
$$f(x, y) = 0.6 * y^4 + 5x^2 2 - 7y^2 + \sin(xy) - 5y;$$
- $f : \mathbb{R}^2 \to \mathbb{R}$,
$$f(x, y) = 100\left(\sqrt{x^2 + (y + 1)^2} - 1\right)^2 + 90\left(\sqrt{x^2 + (y - 1)^2} - 1\right)^2 - (20x + 40y).$$

Implementare 3.2 De exemplu, pentru a reprezenta a treia funcţie şi liniile sale de nivel, folosim codul de mai jos.

desem_m_m_var.m (fişier de rulare)

```
[x,y]=meshgrid(-5:.01:5);
z=x.^6+y.^4+4.*x.^2.*y.^2-3.*x.*y+2.*x+y;
mesh(x,y,z)
%iar pentru a desena curbele de nivel:
[c,h]=contour(x,y,z,70);
clabel(c,h);
```

Implementare 3.3 O altă variantă este următoarea:

```
% Rosenbrock's function
clear all
clc
[x1,x2] = meshgrid(-2:.03:2,-2:.03:2);
z = 100*(x2-x1.^2).^2+(1-x1).^2;
surf(x1,x2,z);
shading interp
view (170,20)
xlabel('x1')
ylabel('x2')
zlabel('f(x1,x2)')
```

Exerciţiul 3.4 *Să se implementeze un algoritm de tip grilă unifomă pentru aceste funcţii.*

Discutatea teoretică a punctelor de minim pentru aceste funcţii poate fi dificilă. Dacă pentru funcţia Rosenbrock există un singur astfel de punct, $(1,1)$, şi acesta este exact determinat teoretic, pentru celelalte discuţia poate fi mai complicată, iar în unele cazuri imposibil de finalizat.

Algoritmii de tip grilă uniformă prezentaţi se referă la determinarea punctelor de minim. În acest curs, vom discuta de asemenea algoritmi de determinare a zerourilor unor funcţii, tocmai pentru că există o strânsă legătură între această chestiune şi problemele de extrem. De altfel, orice algoritm de aproximare a zerourilor unei funcţii se poate trasforma într-un algoritm de aproximare a punctelor de optim prin aplicarea sa asupra gradientului funcţiei (conform Teoremei lui Fermat).

Un algoritm simplu de aproximare a zerourilor unei funcţii reale de o variabilă, este oferit de metoda înjumătăţirii intervalului. Să presupunem că avem o funcţie continuă f şi două numere reale $a < b$ pentru care $f(a)f(b) < 0$. Atunci, f are o rădăcină în (a,b). Pentru uşurinţa expunerii, presupunem că această soluţie este unică. Vom genera două şiruri (a_k) şi (b_k) astfel: $a_0 = a$, $b_0 = b$. Fie $x_0 = 2^{-1}(a_0 + b_0)$. Dacă $f(x_0) = 0$ atunci x_0 este soluţia căutată şi iterarea se opreşte. Astfel, dacă $f(a_0)f(x_0) < 0$ alegem $a_1 = a_0$ şi $b_1 = x_0$, iar dacă $f(x_0)f(b_0) < 0$ alegem $a_1 = x_0$ şi $b_1 = b_0$. Continuăm procedeul luând $x_1 = 2^{-1}(a_1 + b_1)$. Procedând recurent, ne apropiem cu (x_k) de soluţie, înjumătăţind la fiecare pas intervalul în care se găseşte soluţia. În general, această convergenţă nu este foarte rapidă, dar în unele cazuri este eficientă.

Exemplul de mai jos reflectă aceste observaţii.

Exemplul 3.4.5 Considerăm funcţia $f : \mathbb{R}^2 \to \mathbb{R}$

$$f(x,y) = x^4 - 16x^2 + 5x + y^4 - 16y^2 + 5y.$$

Desenând graficul constatăm că funcţia are patru puncte de minim local şi un punct de maxim local. Vrem să găsim aceste puncte. Primul pas este să determinăm punctele critice, ceea ce, datorită simetriei în x şi y, revine la rezolvarea ecuaţiei $4u^3 - 32u + 5 = 0$. Asociem funcţia $\varphi : \mathbb{R} \to \mathbb{R}$, $\varphi(u) = 4u^3 - 32u + 5$ şi constatăm că aceasta are trei zerouri $u_1 \in \left(-\infty, -\frac{2\sqrt{2}}{\sqrt{3}}\right)$, $u_2 \in \left(-\frac{2\sqrt{2}}{\sqrt{3}}, \frac{2\sqrt{2}}{\sqrt{3}}\right)$, $u_3 \in \left(\frac{2\sqrt{2}}{\sqrt{3}}, \infty\right)$. Cum $\varphi'(u_1) > 0$, $\varphi'(u_2) < 0$, $\varphi'(u_3) > 0$, constituind matricea hessiana a funcţiei iniţiale, deducem că următoarele puncte sunt de minim local: (u_1, u_1), (u_1, u_3), (u_3, u_1), (u_3, u_3). De asemenea, (u_2, u_2) este maxim local.

În plus, funcţia fiind coercivă, ea admite un punct de minim global. O aproximare a punctelor u_1, u_3 (de exemplu $u_1 \in \left(-3, -\frac{2\sqrt{2}}{\sqrt{3}}\right)$, $u_3 \in \left(\frac{2\sqrt{2}}{\sqrt{3}}, 3\right)$) şi variaţia lui $\psi : \mathbb{R} \to \mathbb{R}$, $\psi(v) = v^4 - 16v^2 + 5v$ arată că (u_1, u_1) este punctul de minim global pentru că $\psi(u_1) < \psi(u_3)$.

Deci pentru a aproxima punctele de extrem ale funcţiei date, trebuie să aproximăm soluţiile ecuaţiei $4u^3 - 32u + 5 = 0$, lucru care se poate face cu metoda descrisă mai sus.

Exerciţiul 3.5 *Să se implementeze metoda înjumătăţirii intervalului pentru finalizarea exemplului anterior.*

Exerciţiul 3.6 *Să se implementeze metoda înjumătăţirii intervalului pentru funcţia* $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^3 - x - 2$.

3.5 Algoritmi iterativi

Toţi algoritmii pe care îi vom studia sunt iterativi (adică soluţia (i.e., aproximarea punctului căutat) este găsită calculând termenii unui şir până la un rang dat de un criteriu de oprire a algoritmului) şi cer precizarea unui punct de start pe care îl notăm cu x_0. Este bine ca acest punct să fie el însuşi o cât mai bună aproximare a soluţiei căutate (mai ales dacă aceasta nu este unică). De exemplu, funcţia $f : \mathbb{R} \to \mathbb{R}$,

$$f(x) = \frac{x^6}{6} - \frac{x^3}{3} - x^2$$

are două puncte de minim şi două rădăcini reale nenule. Dacă se pleacă cu o valoare x_0 apropiată de unul dintre aceste puncte, atunci este probabil ca algoritmul (pentru rezolvarea ecuaţiei sau pentru găsirea punctelor de extrem) să găsească o aproximare a acelui punct.

Revenind la discuţia generală, după alegerea lui x_0, algoritmul generează un şir de iteraţii $(x_k)_{k \in \mathbb{N}}$ care au scopul de a se apropia de soluţie (soluţie a unei ecuaţii sau punct de optim). Generarea acestui şir se va opri atunci când nu se mai pot face progrese în încercarea de apropiere de soluţie (conform regulii interne a algoritmului de generare a iteraţiilor) sau când a fost atinsă o anumită acurateţe dinainte stabilită (care se defineşte printr-un criteriu de oprire). Orice algoritm trebuie să implementeze o regulă internă de generare a unei noi iteraţii utilizând iteraţiile deja existente. În general, fiecare nouă iteraţie trebuie să realizeze o apropiere de soluţie, dar există şi algoritmi nemonotoni, pentru care descreşterea nu trebuie să se producă neapărat la fiecare pas.

În termeni matematici, un algorithm iterativ poate fi descris ca având forma unei recurenţe de tipul

$$\begin{aligned} &x_{k+1} = A(x_k), \ \forall k \geq 0, \\ &x_0 \in \mathbb{R}^p, \end{aligned}$$

unde $A : \mathbb{R}^p \to \mathbb{R}^p$ este o funcţie.Totuşi, în practică, datorită implementării pe diferite computere, de către diferiţi programatori care pot interpreta diferit secvenţele sintactice ale programului ce implementează algoritmul, datorită acumulării rotunjirilor numerice care şi ele pot diferi în cazul unor software-uri diferite, este mai realistic să considerăm A ca fiind o multifuncţie. De fapt, în realitate, dacă A este funcţia de mai sus, relaţia dintre x_{k+1} şi x_k este $x_{k+1} = A(x_k) + \varepsilon_k$, unde ε_k este un element provenit din diferenţele de aproximare descrise. Datorită faptului că acest ε_k

are un caracter aleatoriu, este mai firesc să îl încorporăm în $A(x_k)$ care devine astfel o mulţime. În acest fel, proprietăţile generale ce pot fi deduse pentru un algoritm astfel modelat se păstrează pentru diferite implementări ale acestuia şi chiar pentru algoritmi similari. Astfel, cu această abordare, algoritmul general de mai sus capătă forma

$$\begin{aligned} &x_{k+1} \in A(x_k), \ \forall k \geq 0, \\ &x_0 \in \mathbb{R}^p. \end{aligned} \tag{3.3}$$

Definiţia 3.5.1 *O multifuncţie $A : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ se numeşte închisă în $\overline{x} \in D$ dacă pentru orice două şiruri convergente (x_k), (y_k) cu $x_k \to \overline{x}$, $y_k \to \overline{y}$ şi $y_k \in A(x_k)$ pentru orice k, are loc $\overline{y} \in A(\overline{x})$. O multifuncţie închisă în toate punctele unei mulţimi se numeşte închisă pe acea mulţime.*

Exemplul 3.5.2 1. Multifuncţia $A : \mathbb{R} \rightrightarrows \mathbb{R}$ dată prin

$$A(x) = \begin{cases} \left\{\frac{1}{2}(x+2)\right\}, & \text{dacă } x > 1 \\ \left\{\frac{1}{4}x\right\}, & \text{dacă } x \leq 1 \end{cases}$$

nu este închisă în $\overline{x} = 1$. Pentru a observa acest lucru este suficient să analizăm cazul şirului

$$x_k = 1 + \frac{1}{2^k}, \ \forall k \geq 0.$$

2. Multifuncţia $A : \mathbb{R} \rightrightarrows \mathbb{R}$, $A(x) = [x^2, \infty)$ este închisă pe $\mathbb{R}$.

În cadrul oricărui algoritm iterativ o funcţie specifică (de obicei funcţia obiectiv în cazul problemelor de optimizare) cunoaşte o descreştere până în momentul când convergenţa (aproximarea) dorită se realizează. Aceasta sugerează definiţia de mai jos.

Definiţia 3.5.3 *Fie $S \subset \mathbb{R}^p$ o mulţime. Considerăm un algoritm de forma (3.3). O funcţie continuă $\varphi : \mathbb{R}^p \to \mathbb{R}$ se numeşte funcţie de descreştere pentru S şi algoritmul A dacă sunt satisfăcute următoarele proprietăţi:*

(i) pentru orice $x \in S$ şi $y \in A(x)$, avem $\varphi(y) \leq \varphi(x)$;

(ii) pentru orice $x \notin S$ şi $y \in A(x)$, avem $\varphi(y) < \varphi(x)$.

Exemplul 3.5.4 Funcţia $\varphi : \mathbb{R} \to \mathbb{R}$, $\varphi(x) = |x|$ este o funcţie de descreştere pentru $S = \{0\}$ şi $A : \mathbb{R} \to \mathbb{R}$, $A(x) = 4^{-1}x$.

Teorema 3.5.5 *Fie un algoritm de forma (3.3). Fie S mulţimea soluţiilor problemei studiate. Presupunem că au loc următoarele condiţii:*

(i) şirul (x_k) este conţinut într-o mulţime compactă;

(ii) A este închisă pe $\mathbb{R}^p \setminus S$;

(iii) există o funcţie continuă $\varphi : \mathbb{R}^p \to \mathbb{R}$ de descreştere pentru S şi A.

Atunci orice punct limită al lui (x_k) se află în S.

Demonstraţie Fie $\overline{x}$ un punct limită al lui (x_k) şi $(x_{k_r})_r$ subşir al lui (x_k) cu limita $\overline{x}$. Cum φ este continuă,

$$\lim_r \varphi(x_{k_r}) = \varphi(\overline{x}).$$

Cum şirul $(\varphi(x_k))$ este monoton descrescător iar un subşir al său este convergent, deducem că întreg şirul este convergent la aceeaşi limită, $\varphi(\overline{x})$.

Presupunem, prin reducere la absurd că $\overline{x} \notin S$. Din prima ipoteză, există un punct limită, notat u, al şirului $(x_{k+1})_k$. Ca mai sus, $\lim_k \varphi(x_{k+1}) = \varphi(u)$, deci $\varphi(\overline{x}) = \varphi(u)$. Pe de altă parte, ipoteza (ii) asigură $u \in A(\overline{x})$, deci, conform (iii), $\varphi(u) < \varphi(\overline{x})$, ceea ce reprezintă o contradicţie. □

Observaţia 3.5.6 *În rezultatul anterior data iniţială poate fi arbitrar aleasă, deci în ipotezele date algoritmul are proprietăţi globale de convergenţă.*

Corolarul 3.5.7 *În ipotezele Teoremei 3.5.5, dacă S este formată dintr-un singur punct $\overline{x}$, atunci $(x_k) \to \overline{x}$.*

Demonstraţie Din prima ipoteză a teoremei, şirul are un punct limită care, conform teoremei şi ipotezei rezultatului de faţă coincide cu $\overline{x}$. Dacă şirul nu ar converge la $\overline{x}$, ar mai avea un punct limită diferit de $\overline{x}$ care ar trebui să fie din S ceea ce nu este posibil. □

Exemplul 3.5.8 Considerăm problem minimizării lui $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$ şi algoritmul definit de $A : \mathbb{R} \rightrightarrows \mathbb{R}$ dată prin

$$A(x) = \begin{cases} \left\{\frac{1}{2}(x+2)\right\}, & \text{dacă } x > 1 \\ \left\{\frac{1}{4}x\right\}, & \text{dacă } x \leq 1. \end{cases}$$

Acest algoritm nu este convergent global (pentru orice dată iniţială). Într-adevăr, dacă luăm $x_0 = 4$, atunci $x_k \to 2$. Se observă că pentru $x_0 = -4$, $x_k \to 2$. Am văzut mai sus (Exemplul 3.5.2) că A nu este închisă în 1.

Exemplul 3.5.9 Considerăm problem minimizării lui $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^3$ şi algoritmul definit de $A : \mathbb{R} \rightrightarrows \mathbb{R}$ dată prin

$$A(x) = \left\{-\left(x^2+1\right)\right\}$$

Acest algoritm nu este convergent pentru că pentru orice dată iniţială şirul generat are limita $-\infty$.

Exemplul 3.5.10 Considerăm problem minimizării lui $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x-1|$ şi algoritmul definit de $A : \mathbb{R} \rightrightarrows \mathbb{R}$ dată prin

$$A(x) = \left\{-\sqrt{|x|}\right\}$$

Acest algoritm este convergent pentru orice dată iniţială $x_0 \in \mathbb{R}$.

O problemă foarte importantă din punct de vedere practic atunci când este studiat un algoritm este viteza sa de convergenţă.

Definiţia 3.5.11 (viteză de convergenţă) *Fie $(x_k)_{k\in\mathbb{N}^*} \subset \mathbb{R}^p$ un şir convergent la $\overline{x} \in \mathbb{R}^p$ cu $x_k \neq \overline{x}$ pentru orice $k \in \mathbb{N}^*$. Spunem că:*

(i) (x_k) converge liniar dacă există $r \in (0,1)$ (numit rată de convergenţă) astfel încât pentru orice k suficient de mare,

$$\frac{\|x_{k+1} - \overline{x}\|}{\|x_k - \overline{x}\|} \leq r;$$

(ii) (x_k) *converge superliniar dacă*

$$\lim_k \frac{\|x_{k+1} - \overline{x}\|}{\|x_k - \overline{x}\|} = 0;$$

(iii) (x_k) *converge pătratic dacă există* $M > 0$ *astfel încât pentru orice* k *suficient de mare,*

$$\frac{\|x_{k+1} - \overline{x}\|}{\|x_k - \overline{x}\|^2} \leq M.$$

Evident, convergenţa pătratică implică convergenţa superliniară care, la rândul său implică convergenţa liniară. Pentru convergenţa liniară rata r este importantă pentru viteza de convergenţă, în timp ce pentru convergenţa pătratică dependenţa de M este mai puţin importantă. Folosind modelul oferit de definiţia convergenţei pătratice, putem defini convergenţe de orice ordin; spunem că avem o convergenţă de ordin $\theta > 1$ dacă există $M > 0$ astfel încât pentru orice k suficient de mare

$$\frac{\|x_{k+1} - \overline{x}\|}{\|x_k - \overline{x}\|^\theta} \leq M.$$

Exemplul 3.5.12 (i) Şirul $\left(1 + \left(\frac{1}{2}\right)^k\right)$ este liniar convergent la 1 cu rata $r = \frac{1}{2}$.

(ii) Şirul $\left(1 + \frac{1}{k^k}\right)$ converge superliniar la 1.

(iii) Şirul $\left(1 + \left(\frac{1}{2}\right)^{2^k}\right)$ converge pătratic la 1.

Exemplul 3.5.13 Reamintim că dacă $I \subset \mathbb{R}$ este un interval închis $f : I \to I$ este o funcţie derivabilă astfel încât derivata sa este mărginită (în modul) pe I de o constantă strict subunitară, atunci pentru orice dată iniţială $x_0 \in I$ iteraţia Picard definită de $x_{k+1} = f(x_k)$, $k \geq 0$ este convergentă către unicul punct fix $\overline{x}$ al lui f din I. Presupunem că f este de clasă C^1.

Se poate observa fără dificultate că dacă punctul fix nu este atins (şirul este nestaţionar), atunci

$$\frac{x_{k+1} - \overline{x}}{x_k - \overline{x}} = \frac{f(x_k) - \overline{x}}{x_k - \overline{x}} \overset{k\to\infty}{\to} f'(\overline{x}) \in (-1, 1).$$

Astfel, observăm că, în general avem de-a face cu o convergenţă cel puţin liniară de rată $|f'(\overline{x})|$.

În cazul cel mai bun în care $f'(x) = 0$ putem avea convergenţe mai bune decât convergenţa liniară. În general, în contextul de mai sus, dacă $f'(\overline{x}) = 0$ şi f este de clasă C^2 atunci, aplicând de două ori regula lui L'Hôpital avem:

$$\lim_{x\to\overline{x}} \frac{f(x) - \overline{x}}{(x - \overline{x})^2} = \frac{f''(\overline{x})}{2},$$

deci pentru orice iteraţie Picard nestaţionară

$$\lim_{k\to\infty} \frac{x_{k+1} - \overline{x}}{(x_k - \overline{x})^2} = \frac{f''(\overline{x})}{2},$$

adică o convergenţă pătratică.

Să considerăm funcţia $f : [0,1] \to [0,1]$ dată prin $f(x) = \frac{1}{1+x^2}$. Aceasta este o contracţie şi are un singur punct fix care este unica soluţie pozitivă a ecuaţiei $x^3 + x - 1 = 0$ (care are valoarea aproximativă $\overline{x} \approx 0.6823$), iar şirul iteraţiilor Picard satisface:

$$\frac{x_{k+1} - \overline{x}}{x_k - \overline{x}} \overset{n\to\infty}{\to} f'(\overline{x}) = \frac{-2\overline{x}}{(1+\overline{x}^2)^2} = -2\overline{x}^3 \approx -0.6353.$$

Deci, practic pentru k suficient de mare, la fiecare pas al iteraţiei, eroarea se multiplică (în valoare absolută) cu aproximativ 0.6353.

Să luăm şi cazul funcţiei $f : [\sqrt{2}, \infty) \to [\sqrt{2}, \infty)$ dată prin

$$f(x) = \frac{x}{2} + \frac{1}{x}.$$

Este uşor de verificat că f este bine definită (inegalitatea mediilor). În plus,

$$|f'(x)| = \left|\frac{1}{2} - \frac{1}{x^2}\right| \leq \frac{1}{2},$$

deci f este contracţie şi are ca unic punct fix $\overline{x} = \sqrt{2}$. Se observă că $f'(\overline{x}) = 0$ şi, deci, pentru orice iteraţie Picard nestaţionară

$$\frac{x_{k+1} - \overline{x}}{(x_k - \overline{x})^2} = \frac{1}{2x_k} \to \frac{1}{2\sqrt{2}} = \frac{f''(\overline{x})}{2},$$

deci avem o convergenţă pătratică.

Exerciţiul 3.7 *Să se implementeze acest algoritm pentru funcţiile de mai sus.*

O altă clasă de viteze de convergenţă este definită mai jos.

Definiţia 3.5.14 ($r-$convergenţă) *Fie $(x_k)_{k\in\mathbb{N}^*} \subset \mathbb{R}^p$ un şir convergent la $\overline{x} \in \mathbb{R}^p$. Spunem că (x_k) converge $r-$liniar ($r-$superliniar, $r-$pătratic) dacă există un şir $(y_k)_{k\in\mathbb{N}^*}$ care este convergent liniar (superliniar, pătratic) la 0 astfel încât pentru orice k*

$$\|x_k - \overline{x}\| \leq y_k.$$

Exemplul 3.5.15 Şirul

$$x_k = \begin{cases} 1 + \left(\frac{1}{2}\right)^k, & \text{dacă } k \text{ este par} \\ 1 + \left(\frac{1}{3}\right)^k, & \text{dacă } k \text{ este impar} \end{cases}$$

este $r-$liniar convergent la 1, dar nu este liniar convergent la 1.

Observaţia 3.5.16 *Vitezele de convergenţă în sensul Definiţiei 3.5.11 implică $r-$convergenţele corespunzătoare, în timp ce reciprocele sunt, în general, false, aşa cum arată exemplul anterior.*

Exerciţiul 3.8 *Studiaţi viteza de convergenţă a şirului $\left(\frac{1}{k!}\right)$.*

Exerciţiul 3.9 *Fie* $f:\mathbb{R}^2\to\mathbb{R}$ *definită prin* $f(x)=\|x\|^2$. *Arătaţi că iteraţiile*

$$x_k=\left(1+\frac{1}{2^k}\right)(\cos k,\sin k)$$

satisfac $f(x_{k+1})<f(x_k)$ *pentru orice* $k\geq 0$. *Arătaţi că orice punct de pe sfera (cercul) unitate este punct limită al lui* (x_k).

Exerciţiul 3.10 *Studiaţi viteza de convergenţă şi* $r-$*convergenţa şirului*

$$x_k=\begin{cases}\left(\frac{1}{4}\right)^{2^k}, & \text{dacă } k \text{ este par}\\ \frac{x_{k-1}}{k}, & \text{dacă } k \text{ este impar.}\end{cases}$$

3.6 Un algoritm reprezentativ: metoda lui Newton

Această secţiune se ocupă cu studiul unei celebre metode, atribuită lui Newton, de a determina cu aproximaţie rădăcinile unor ecuaţii neliniare. În plus, această metodă ilustrează foarte bine ideile generale expuse despre algoritmii iterativi în secţiunea precedentă.

Să remarcăm că aplicarea condiţiilor de optimalitate transformă o problemă de optimizare în problema rezolvării unei ecuaţii sau a unui sistem de ecuaţii (de cele mai multe ori neliniare şi chiar imposibil de rezolvat exact). Un alt argument pentru studiul acestui algoritm în cadrul unui curs de Teoria optimizării ar fi acela că, în fond, problema rezolvării unei ecuaţii de forma $f(x)=0$ este echivalentă cu problema determinării soluţiilor globale ale problemei de optimizare fără restricţii $\min f^2$.

Metoda lui Newton este una dintre cele mai cunoscute metode iterative de a aproxima rădăcinile funcţiilor care au proprietăţi de diferenţiabilitate suficient de bune. Vom vedea că acest algoritm este unul local pentru că data iniţială trebuie aleasă suficient de aproape de soluţie, dar convergenţa este pătratică, adică foarte bună.

Fie $f:\mathbb{R}^p\to\mathbb{R}^p$ o funcţie de clasă C^1 şi fie $\overline{x}$ o rădăcină simplă a lui f (i.e., $f(\bar{x})=0$ şi $\nabla f(\overline{x})$ este nesingulară). Considerăm o valoare x_0 suficient de apropiată de $\overline{x}$ (vom discuta sensul acestei apropieri ceva mai târziu).

Şirul iteraţiilor Newton porneşte de la ecuaţia

$$0=f(x_k)+\nabla f(x_k)(x_{k+1}-x_k). \tag{3.4}$$

Această ecuaţie, din care se obţine valoarea lui x_{k+1} arată de ce este important să avem soluţie simplă şi de ce trebuie să pornim din apropierea lui $\overline{x}$: trebuie ca x_0 să fie într-o vecinătate a lui $\overline{x}$ în care ∇f este inversabilă, iar o astfel de vecinătate există tocmai pentru că ∇f este continuă şi nesingulară în $\overline{x}$. Astfel, definim formal iteraţia Newton prin:

$$x_{k+1}=x_k-\nabla f(x_k)^{-1}(f(x_k)). \tag{3.5}$$

Aşa cum deja se vede din relaţiile anterioare şi se va vedea şi în continuare, unele dificultăţi în aplicarea acestei metode pot fi rezumate astfel:

- dacă $\nabla f(x_k)$ este singulară, nu putem defini x_{k+1};

- calculul exact al lui $\nabla f(x_k)^{-1}$ poate fi costisitor din punct de vedere computaţional;
- e dificil de ştiut a priori că $\nabla f(\overline{x})$ este nesingulară.

După prezentarea riguroasă a metodei vom reveni asupra unor aspecte de acest tip. Înainte de a prezenta rezultatul principal privind convergenţa metodei lui Newton prezentăm o proprietate ajutătoare a funcţiilor de clasă C^1.

Lema 3.6.1 *Fie $D \subset \mathbb{R}^p$ o mulţime deschisă şi convexă, $f : D \to \mathbb{R}^p$ o funcţie de clasă C^1 pe D şi $x, y \in D$. Atunci*

$$f(y) = f(x) + \int_0^1 \nabla f(x + t\,(y - x))(y - x)dt,$$

unde integrarea şi egalitatea se înţeleg pe componente.

Demonstraţie. Reamintim că pentru orice $x, y \in D$, mulţimea

$$I_{y,x} = \{t \in \mathbb{R} \,|ty + (1 - t)\, x \in D\}$$

este un interval care conţine intervalul $[0, 1]$.

Pentru fiecare funcţie coordonată f_i cu $i \in \overline{1, p}$ definim funcţia scalară $\alpha_i : I_{y,x} \to \mathbb{R}$,

$$\alpha_i\,(t) = f_i\,(ty + (1 - t)\, x) = f_i\,(x + t\,(y - x))\,.$$

Evident,

$$\alpha_i'\,(t) = \nabla f_i\,(x + t\,(y - x))\,(y - x)\,.$$

Cum $[0, 1] \subset I_{y,x}$ şi

$$\alpha_i\,(1) - \alpha_i\,(0) = \int_0^1 \alpha_i'\,(t)\,dt,$$

deducem

$$f_i\,(y) - f_i\,(x) = \int_0^1 \nabla f_i\,(x + t\,(y - x))\,(y - x)\,dt,$$

ceea ce reprezintă concluzia. □

În cele ce urmează privim o matrice de tip $p \times p$ (cum ar fi diferenţiala lui f) ca operator liniar de la $\mathbb{R}^p$ la $\mathbb{R}^p$ înzestrat cu norma operatorială generată de $\|\cdot\|_2$ pe $\mathbb{R}^p$. Prezentăm o proprietate a normei $\|\cdot\|_2$ pe $\mathbb{R}^p$ (pe care o notăm $\|\cdot\|$) de care avem nevoie în continuare.

Lema 3.6.2 *Fie $g : \mathbb{R} \to \mathbb{R}^p$, $g = (g_i)_{i \in \overline{1,p}}$ astfel încât toate funcţiile g_i sunt Riemann integrabile pe $\mathbb{R}$ şi fie $a, b \in \mathbb{R}$ cu $a < b$. Atunci*

$$\left\| \int_a^b g\,(t)\,dt \right\| \leq \int_a^b \|g\,(t)\|\,dt.$$

Demonstraţie. Fie

$$v = \left(\int_a^b g_i(t)\, dt \right)_{i \in \overline{1,p}} \in \mathbb{R}^p.$$

Avem

$$\|v\|^2 = \sum_{i=1}^p v_i^2 = \sum_{i=1}^p v_i \int_a^b g_i(t)\, dt = \sum_{i=1}^p \int_a^b v_i g_i(t)\, dt = \int_a^b \left(\sum_{i=1}^p v_i g_i(t) \right) dt$$
$$= \int_a^b \langle v, g(t) \rangle\, dt \leq \int_a^b \|v\| \, \|g(t)\|\, dt = \|v\| \int_a^b \|g(t)\|\, dt$$

şi prin simplificare cu $\|v\|$ (cazul $v = 0$ este evident) obţinem concluzia. □

Teorema 3.6.3 *Fie $D \subset \mathbb{R}^p$ o mulţime deschisă şi convexă, $f : D \to \mathbb{R}^p$ o funcţie de clasă C^1 pe D. Fie $\overline{x}$ o soluţie simplă a ecuaţiei $f(x) = 0$ şi fie (x_k) un şir de iteraţii generate de ecuaţia (3.5). Atunci există $\varepsilon > 0$ astfel încât dacă $x_0 \in D(\overline{x}, \varepsilon) \subset D$ toate elementele şirului sunt bine definite şi rămân în $D(\overline{x}, \varepsilon)$ şi (x_k) converge superliniar la $\overline{x}$, adică*

$$\lim_{k \to \infty} \frac{\|x_{k+1} - \overline{x}\|}{\|x_k - \overline{x}\|} = 0. \tag{3.6}$$

În plus, dacă ∇f este Lipschitz pe $D(\overline{x}, \varepsilon)$ atunci convergenţa este pătratică, adică există $K > 0$ astfel încât

$$\frac{\|x_{k+1} - \overline{x}\|}{\|x_k - \overline{x}\|^2} \leq K, \ \forall k. \tag{3.7}$$

Demonstraţie. Fie $M > 0$ şi $\varepsilon > 0$ astfel încât $D(\overline{x}, \varepsilon) \subset D$ şi pentru orice $x \in D(\overline{x}, \varepsilon)$ au loc simultan următoarele:

$$\|\nabla f(x) - \nabla f(\overline{x})\| \leq \frac{1}{4M}$$
$$\nabla f(x) \text{ este inversabilă}$$
$$\left\|\nabla f(x)^{-1}\right\| < M.$$

Astfel de numere pozitive există, pe baza ipotezelor considerate.

Dacă $x_0 = \overline{x}$, atunci iteraţia Newton este staţionară şi concluziile sunt evidente.

Fie $x_0 \in D(\overline{x}, \varepsilon) \setminus \{\overline{x}\}$ şi x_1 definit de iteraţia Newton, adică

$$x_1 = x_0 - \nabla f(x_0)^{-1}(f(x_0)).$$

Atunci

$$x_1 - \overline{x} = x_0 - \overline{x} - \nabla f(x_0)^{-1}(f(x_0)) = \nabla f(x_0)^{-1} \left(\nabla f(x_0)(x_0 - \overline{x}) - f(x_0) \right).$$

Pe baza lemei anterioare putem scrie

$$x_1 - \overline{x} = \nabla f(x_0)^{-1} \left(\nabla f(x_0)(x_0 - \overline{x}) - \int_0^1 \nabla f(\overline{x} + t(x_0 - \overline{x}))(x_0 - \overline{x}) dt \right)$$
$$= \nabla f(x_0)^{-1} \left(\int_0^1 \left(\nabla f(x_0) - \nabla f(\overline{x} + t(x_0 - \overline{x})) \right)(x_0 - \overline{x})\, dt \right).$$

Deci, aplicând inclusiv Lema 3.6.2,

$$\begin{aligned}\|x_1-\overline{x}\| &\leq M\left\|\int_0^1\left(\nabla f(x_0)-\nabla f(\overline{x}+t\left(x_0-\overline{x}\right)\right)\right)\left(x_0-\overline{x}\right)dt\Big\| \\ &\leq M\int_0^1\left\|\left(\nabla f(x_0)-\nabla f(\overline{x}+t\left(x_0-\overline{x}\right)\right)\right)\left(x_0-\overline{x}\right)\right\|dt \\ &\leq M\left\|x_0-\overline{x}\right\|\int_0^1\left\|\nabla f(x_0)-\nabla f(\overline{x}+t\left(x_0-\overline{x}\right))\right\|dt \\ &\leq M\cdot\frac{1}{2M}\left\|x_0-\overline{x}\right\|=\frac{1}{2}\left\|x_0-\overline{x}\right\|.\end{aligned}$$

În particular, $x_1\in D\left(\overline{x},\varepsilon\right)$ şi repetând inductiv calculele de mai sus deducem că $\left(x_k\right)\subset D\left(\overline{x},\varepsilon\right)$ şi pentru orice k

$$\left\|x_{k+1}-\overline{x}\right\|\leq M\left\|x_k-\overline{x}\right\|\int_0^1\left\|\nabla f(x_k)-\nabla f(\overline{x}+t\left(x_k-\overline{x}\right))\right\|dt$$

Dacă unul dintre punctele şirului este $\overline{x}$, atunci şirul devine staţionar şi deci concluziile sunt clare.

Contrar, deducem, pentru orice k,

$$\left\|x_{k+1}-\overline{x}\right\|\leq M\left\|x_k-\overline{x}\right\|\int_0^1\left\|\nabla f(x_k)-\nabla f(\overline{x}+t\left(x_k-\overline{x}\right))\right\|dt\leq\frac{1}{2}\left\|x_k-\overline{x}\right\|,$$

de unde $x_k\to\overline{x}$ şi

$$\frac{\left\|x_{k+1}-\overline{x}\right\|}{\left\|x_k-\overline{x}\right\|}\leq M\int_0^1\left\|\nabla f(x_k)-\nabla f(\overline{x}+t\left(x_k-\overline{x}\right))\right\|dt.$$

Cum, pe baza faptului că $x_k\to\overline{x}$ şi a continuităţii lui ∇f, $\nabla f(x_k)-\nabla f(\overline{x}+t\left(x_k-\overline{x}\right))\to 0$, pentru orice $t\in[0,1]$, iar funcţiile de sub integrală sunt uniform mărginite pe $[0,1]$, deducem că

$$\lim\frac{\left\|x_{k+1}-\overline{x}\right\|}{\left\|x_k-\overline{x}\right\|}=0.$$

Presupunem acum că ∇f este Lipschitz pe $D\left(\overline{x},\varepsilon\right)$. Atunci, există $L>0$ astfel încât pentru orice $x,y\in D\left(\overline{x},\varepsilon\right)$,

$$\left\|\nabla f\left(y\right)-\nabla f\left(x\right)\right\|\leq L\left\|y-x\right\|.$$

Aşa cum am demonstrat mai sus, în cazul nestaţionar, pentru orice k,

$$\begin{aligned}\frac{\left\|x_{k+1}-\overline{x}\right\|}{\left\|x_k-\overline{x}\right\|} &\leq M\int_0^1\left\|\nabla f(x_k)-\nabla f(\overline{x}+t\left(x_k-\overline{x}\right))\right\|dt \\ &\leq M\int_0^1 L\left\|x_k-(\overline{x}+t\left(x_k-\overline{x}\right))\right\|dt \\ &= ML\left\|x_k-\overline{x}\right\|\int_0^1(1-t)\,dt=\frac{ML}{2}\left\|x_k-\overline{x}\right\|.\end{aligned}$$

Aşadar, pentru $K = \frac{ML}{2}$ obţinem

$$\frac{\|x_{k+1} - \overline{x}\|}{\|x_k - \overline{x}\|^2} \leq K,$$

ceea ce demonstrează rezultatul. □

Observaţia 3.6.4 *Algoritmul lui Newton este local, data iniţială trebuind aleasă aproape de soluţie.*

În continuare, implementăm metoda lui Newton şi discutăm consecinţe ale acesteia. Începem cu cazul funcţiilor reale de o variabilă reală, adică situaţia $p = 1$.

În cazul $p = 1$ ecuaţia (3.5) devine

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

Astfel, iteraţia x_{k+1} este punctul în care tangenta la graficul lui f în punctul $(x_k, f(x_k))$ intersectează axa Ox. Discutăm unele posibilităţi de a alege punctul x_0 suficient de aproape de soluţie astfel încât metoda lui Newton să conveargă la punctul căutat. O primă posibilitate (empirică) este aceea de a studia graficul funcţiei şi de a alege o valoare x_0 care pare a fi destul de apropiată soluţiei. O altă variantă ar fi de a aplica o altă metodă de aproximare a rădăcinilor unei funcţii a cărei convergenţă este mai lentă, dar care totuşi după câteva iteraţii ne duce în apropierea soluţiei, moment din care putem alege x_0 şi aplica metoda lui Newton pentru accelerarea convergenţei.

Implementăm acum metoda lui Newton.

Implementare 3.11 Testăm metoda lui Newton pentru funcţia $f : \mathbb{R} \to \mathbb{R}$ dată de relaţia

$$f(x) = x + e^x + \frac{10}{1 + x^2} - 5$$

care are o rădăcină simplă în intervalul $(-2, 0)$ după cum se poate constata din studiul graficului său. Plecând cu data iniţială $x_0 = 1.5$ aproximăm, folosind metoda lui Newton, această soluţie cu rapiditate, aşa cum arată programul de mai jos:

```
functie=@(x) [x+exp(x)+10/(1+x^2)-5];
%desen
fplot(functie,[-3,3]);
functie_der=@(x) [1+exp(x)-20*x/(1+x^2)^2];
%Newton
x=-1.5;x_precedent=-1;n=0;eps=10^(-6);maxiter=50;
while abs(x-x_precedent)>eps && n<maxiter
x_precedent=x
x=x-functie(x)/functie_der(x);
n=n+1;
end
x
```

```
n
functie(x)
```
care returnează:
```
x = -0.90456
n = 34
ans = 8.8818e-16
```
Dacă se pleacă cu data inițială $u = -1.5$, atunci se obține valoarea de mai sus după doar 5 iterații.

Observația 3.6.5 *Fie $f : \mathbb{R} \to \mathbb{R}$ o funcție ce clasă C^3, strict convexă și coercivă. Atunci f are un unic punct de minim global $\overline{x}$ care este soluția ecuației $f'(\overline{x}) = 0$. Astfel, pentru aproximarea sa, putem folosi metoda lui Newton. Astfel, iterația devine*

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}.$$

Vom reveni asupra generalizărilor acestei observații într-o secțiune ulterioară.

Observația 3.6.6 *Metoda lui Newton aplicată problemelor de optimizare (Observația 3.6.5) are un oracol de ordinul al doilea.*

Exercițiul 3.12 *Fie $f : \mathbb{R} \to \mathbb{R}$, $f(x) = 2x^2 + 3e^{-2x}$. Să se arate că f este strict convexă și admite punct de minim. Să se testeze metoda descrisă mai sus pentru a aproxima punctul de minim.*

Totuși, aplicarea metodei lui Newton pentru determinarea unui punct de minim presupune calculul primelor două derivate, proces ce poate fi anevoios. Pe de altă parte, în cadrul implementării niciodată metoda nu este lipsită de aproximări a calculelor de pe parcursul generării iterațiilor. Din acest motiv ne putem gândi la o variantă de aproximare a derivatelor la fiecare pas pe baza relațiilor

$$\begin{aligned} &\lim_{t\to 0} \frac{f(x+t) - f(x-t)}{2t} = f'(x), \\ &\lim_{t\to 0} \frac{f(x+t) + f(x-t) - 2f(x)}{t^2} = f''(x). \end{aligned} \tag{3.8}$$

Astfel de înlocuiri conduc la metode numite de tip cvasi-Newton. Este intuitiv rezonabil să credem că aceste metode păstrează proprietăți bune de convergență (deși poate nu de viteză pătratică). Nu vom demonstra efectiv convergența metodei derivate pe baza înlocuirilor anterioare, dar o vom testa și vom constata comportarea ei foarte bună. În secțiunea următoare vom studia în detaliu o astfel de metodă.

Mai întâi implementăm aceste observații pentru a (re)găsi rădăcina funcției $f(x) = x + e^x + \frac{10}{1+x^2} - 5$, fără a mai calcula efectiv derivata acesteia.

Implementare 3.13
```
delx=0.01;
functie=@(x) [x+exp(x)+10/(1+x^2)-5];
x=-1.5;x_precedent=-1;n=0;eps=10^(-6);maxiter=50;
```

```
while abs(x-x_precedent)>eps && n<maxiter
 x_precedent=x;
 derivata=(functie(x+delx) - functie(x-delx) )/(2*delx);
 x=x-functie(x)/derivata;
 n=n+1;
end
x
n
```

Exerciţiul 3.14 *Să se rezolve al doilea caz numeric de la Problema 2.27.*

Exerciţiul 3.15 *Să se determine minimul funcţiei* $f : [30, 70] \to \mathbb{R}$*,*

$$f(x) = \frac{204165.5}{330 - 2x} + \frac{10400}{x - 20},$$

aproximând numeric ambele derivate.

Implementăm acum metoda lui Newton pentru dimensiuni mai mari decât 1. Căutăm soluţii pentru sistemul

$$\begin{cases} x_1 + x_2 + x_3 = 3 \\ x_1^2 + x_2^2 + x_3^2 = 5 \\ e^{x_1} + x_1 x_2 - x_1 x_3 = 1. \end{cases}$$

Implementare 3.16 Exemplificăm prin codul următor:

Newton_sistem (fişier funcţie)

```
function a=Newton_sistem(f,J,x0,tol)
xvechi=x0;
n=0;
xnou=x0-J(x0)^(-1)*(f(x0));
while norm(xnou-xvechi)>tol %&& n<nrit
 xvechi=xnou;
 xnou=xvechi-J(xvechi)^(-1)*f(xvechi);
 n=n+1;
end
disp(xnou)
disp(n)
end
```

folosit cu comanda

```
f=@(x) [x(1)+x(2)+x(3)-3;x(1)^2+x(2)^2+x(3)^2-5; exp(x(1))+x(1)*x(2)-x(1)*x(3)-1];
f_der=@(x) [1,1,1;2*x(1),2*x(2),2*x(3);exp(x(1))+x(2)-x(3),x(1),-x(1)];
Newton_sistem(f,f_der,[1;0;1],10^(-6))
```

care returnează

```
 1.224394
 -0.093133
 1.868739
5
```

Deci, în acest caz, metoda lui Newton returnează după 5 iteraţii, folosind data iniţială $(1, 0, 1)$ şi toleranţa 10^{-6}, soluţia aproximativă $(1.224394, -0.093133, 1.868739)$ pentru sistemul menţionat. De remarcat că soluţia $(0, 2, 1)$ se obţine cu o altă alegere a datei iniţiale.

Exerciţiul 3.17 *Să se scrie nişte funcţii* `newton, newton_optim, cvasinewton, cvasinewton_opti` *pentru metodele Newton şi cvasi-Newton discutate mai sus, după modelul Implementare 3.16.*

Exerciţiul 3.18 *Să se determine minimul funcţiei de la Exemplul 1.1.1.*

Exerciţiul 3.19 *Deteminaţi minimele funcţiei* $f : \mathbb{R} \to \mathbb{R}$,

$$f(x) = \frac{1}{2}x^5 - \frac{3}{40}x^4 - \frac{3}{5}x^3 - 7x.$$

Exerciţiul 3.20 *Fie şirul de numere reale dat prin* $x_0 = 1$ *şi*

$$x_{n+1} = \frac{2x_n^3 + 3}{3x_n^2},\ \forall n \geq 0.$$

Să se arate că $(x_n)_n$ *este o iteraţie Newton pentru o funcţie care se va determina. Ce se poate spune despre convergenţa lui* $(x_n)_n$?

Exerciţiul 3.21 *Fie şirul de numere reale dat prin* $x_0 = 1$ *şi*

$$x_{n+1} = \frac{x_n^2 + 2}{2x_n},\ \forall n \geq 0.$$

Să se arate că $(x_n)_n$ *este o iteraţie Newton pentru o funcţie care se va determina. Studiaţi convergenţa şirului* (x_n) *şi, în caz afirmativ, determinaţi-i limita.*

Exerciţiul 3.22 *Să se aproximeze minimul funcţiei* $f : (0, \infty) \to \mathbb{R}$

$$f(x) = \frac{-0.75}{1 + x^2} - 0.65x \operatorname{arctg} \frac{1}{x}.$$

Exerciţiul 3.23 *Să se aproximeze punctele de extrem local ale funcţiei de la Exemplul 3.4.5 de mai sus folosind metoda lui Newton pentru ecuaţia* $4u^3 - 32u + 5 = 0$ *cu diverse date iniţiale.*

Exerciţiul 3.24 *Să se determine soluţii aproximative pentru sistemele de mai jos cu datele iniţiale indicate:*
(i) $(x_0, y_0) = (1, 2)$,

$$\begin{cases} 1 + y^2 - 4x^2 = 0 \\ 3 + 2y - x^2 - y^2 = 0 \end{cases};$$

(ii) $(x_0, y_0) = (1, 2)$,

$$\begin{cases} y - x^3 = 0 \\ 36 - 4x^2 - 9y^2 = 0 \end{cases};$$

(iii) $(x_0, y_0) = (-0.5, -1.5)$,

$$\begin{cases} y - x^3 + 3x^2 - 4x = 0 \\ y^2 - x - 2 = 0 \end{cases};$$

(iv) $(x_0, y_0) = (0, 1)$,

$$\begin{cases} x^2 - 2x - y + 2^{-1} = 0 \\ x^2 + 4y^2 - y = 0 \end{cases};$$

(v) $(x_0, y_0) = (2, 1.5)$,

$$\begin{cases} 3 + y - x^2 = 0 \\ 3 - xy = 0. \end{cases}$$

3.7 Aplicaţie la funcţia W, a lui Lambert

Fie $f : \mathbb{R} \to \mathbb{R}$ dată prin $f(x) = xe^x$. Este uşor de văzut că această funcţie este indefinit derivabilă şi, ţinând cont de dezvoltatea lui e^x în serie de puteri, chiar analitică pe $\mathbb{R}$. De asemenea, pentru orice $x \in \mathbb{R}$,

$$f'(x) = (x+1)e^x$$
$$f''(x) = (x+2)e^x,$$

deci:

(i) f are un punct de extrem, $x = -1$, care este punct de minim global (e de observat că $\lim_{x\to-\infty} f(x) = 0$ şi $\lim_{x\to\infty} f(x) = \infty$);

(ii) f este strict descrescătoare pe $(-\infty, -1]$ şi strict crescătoare pe $[-1, \infty)$;

(iii) $\operatorname{Im} f = \left[-\frac{1}{e}, \infty\right)$, $f((-\infty, -1]) = \left[-\frac{1}{e}, 0\right)$, $f([-1, \infty)) = \left[-\frac{1}{e}, \infty\right)$;

(iv) f are un singur punct de inflexiune, $x = -2$, este strict concavă pe $(-\infty, -2]$ şi strict convexă pe $[-2, \infty)$.

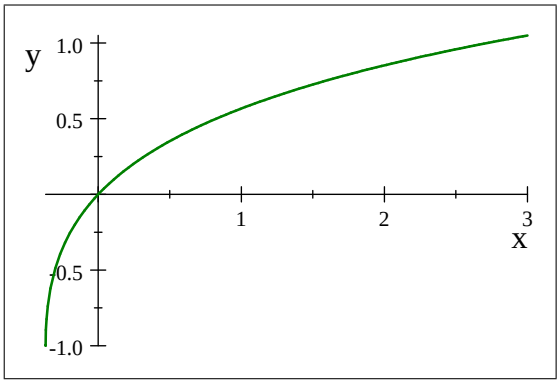
Considerăm două restricţii bijective ale lui f, astfel:

$$f_1 : [-1, \infty) \to \left[-\tfrac{1}{e}, \infty\right),\ f_1(x) = xe^x,$$
$$f_2 : (-\infty, -1] \to \left[-\tfrac{1}{e}, 0\right),\ f_2(x) = xe^x.$$

Inversa lui f_1 se numeşte funcţia lui Lambert şi se notează în mod tradiţional prin W, deci

$$W : \left[-\frac{1}{e}, \infty\right) \to [-1, \infty),\ W(y) = x,\ \text{unde } y = xe^x.$$

Din cauza faptului că, în general, ecuaţia $xe^x = y$ nu poate fi rezolvată, nu se poate indica o formă explicită pentru W. Graficul lui W este simetric faţă de prima bisectoare în raport cu graficul lui f_1.



Graficul lui W

Unele valori ale lui W pot fi calculate. De exemplu,

$$W\left(-\frac{1}{e}\right)=-1;\ W(0)=0;\ W\left(-\frac{1}{2}\ln 2\right)=-\ln 2;\ W(e)=1.$$

Putem de asemenea scrie $W(y)\,e^{W(y)}=y$, pentru orice $y\geq -\frac{1}{e}$. Cum $f_1'(x)\neq 0$ pentru orice $x>-1$, funcţia W este derivabilă pe $\left(-\frac{1}{e},\infty\right)$ şi pentru orice punct y din acest interval deducem următoarele formule:

$$W'(y)=\frac{1}{f'(x)}=\frac{1}{(x+1)\,e^{x}},\ \text{unde } y=xe^{x};$$
$$W'(y)=\frac{e^{-W(y)}}{1+W(y)};$$
$$W'(y)=\frac{W(y)}{y\,(1+W(y))},\ \text{dacă } y\neq 0.$$

Cum f_1 este indefinit derivabilă pe $(-1,\infty)$, deducem că W este indefinit derivabilă pe $\left(-\frac{1}{e},\infty\right)$. Derivând în formulele anterioare, avem:

$$W''(y)=-\frac{e^{-2W(y)}\,(W(y)+2)}{(W(y)+1)^{3}};$$
$$W''(y)=-\frac{W^{2}(y)\,(W(y)+2)}{y^{2}\,(W(y)+1)^{3}},\ \text{dacă } y\neq 0.$$

Din acest calcul deducem stricta concavitate a funcţiei W.

O metodă de a aproxima valorile lui W (pe tot domeniul său de definiţie) este să rezolvăm aproximativ ecuaţia (în necunoscuta x) $xe^{x}-y=0$, de exemplu prin metoda lui Newton. Astfel, cu o dată apropiată de soluţie (pentru y aproape de 0 alegem data inţială aproape de 0, pentru y mare alegem data iniţială $\ln y$, de exemplu) folosim iteraţia Newton

$$x_{n+1}=x_n-\frac{x_ne^{x_n}-y}{(xe^{x}-y)'_{x}(x_n)}$$
$$=\frac{x_n^{2}+ye^{-x_n}}{x_n+1},$$

pentru a aproxima cu viteză pătratică valoarea lui x, adică a lui $W(y)$.

Exerciţiul 3.25 *Să se scrie un program care să determine toate valorile lui* $W(n)$ *cu* $n \in \overline{1,100}$.

Cum am precizat, f_1 este analitică. De asemenea, $f_1'(0) = 1 \neq 0$ şi pe baza unui rezultat numit Teorema lui Lagrange de inversiune, se obţine că inversa W a lui f_1 este local analitică în jurul lui 0, iar dezvoltarea sa este

$$W(y) = \sum_{n=1}^{\infty} w_n \frac{y^n}{n!},$$

unde

$$w_n = \lim_{x \to 0} \left(\left(\frac{x}{f(x)} \right)^n \right)^{(n-1)} = \lim_{x \to 0} \left(e^{-nx} \right)^{(n-1)} = (-n)^{n-1}.$$

Deci,

$$W(y) = \sum_{n=1}^{\infty} \frac{(-n)^{n-1}}{n!} y^n.$$

Este uşor de văzut că raza de convergenţă a acestei serii este $\frac{1}{e}$, deci dezvoltarea e valabilă pe intervalul $\left(-\frac{1}{e}, \frac{1}{e}\right)$.

Să calculăm o primitivă lui W, cu schimbarea de variabilă $y = f_1(x)$ (adică $y = xe^x$ şi $x = W(y)$)

$$\int W(y)\, dy = \int x(x+1) e^x dx = e^x \left(x^2 - x + 1\right) + C = yW(y) - y + e^{W(y)} + C.$$

Astfel, de exemplu,

$$\int_0^e W(y)\, dy = e - e + e - 1 = e - 1.$$

Având în vedere imaginile funcţiilor f_1 şi f_2, este clar că putem vorbi despre $W(f(x))$ pentru orice $x \in \mathbb{R}$. Evident, dacă $x \geq -1$, $W(f(x)) = W(f_1(x)) = x$. Pentru $x < -1$, $W(f(x)) = W(f_2(x)) \neq x$. Este clar că

$$\lim_{x \searrow -1} \frac{W(f(x)) + 1}{x + 1} = \lim_{x \searrow -1} \frac{x+1}{x+1} = 1.$$

Ne punem problem existenţei (şi a valorii) limitei

$$\lim_{x \nearrow -1} \frac{W(f(x)) + 1}{x + 1}.$$

Avem următorul rezultat mai general.

Teorema 3.7.1 *Fie* $\varphi : \mathbb{R} \to \mathbb{R}$ *şi* $\overline{x} \in \mathbb{R}$ *un punct de extrem local pentru* φ. *Presupunem că există o vecinătate* V *a lui* $\overline{x}$ *astfel încât* φ *este analitică pe* V *şi* $\varphi'(x) \neq 0$ *pentru orice* $x \in V \setminus \{\overline{x}\}$. *Atunci:*

(i) există $\varepsilon > 0$ *astfel încât* φ *este strict monotonă pe* $[\overline{x} - \varepsilon, \overline{x}]$ *şi pe* $[\overline{x}, \overline{x} + \varepsilon]$;

(ii) dacă ψ este inversa lui φ pe $[\overline{x}, \overline{x}+\varepsilon]$, există $\delta > 0$ astfel încât $\psi \circ \varphi$ este corect definită pe $[\overline{x}-\delta, \overline{x}]$ şi

$$\lim_{x \nearrow \overline{x}} \frac{(\psi \circ \varphi)(x) - \overline{x}}{x - \overline{x}} = -1.$$

Demonstraţie. Facem demonstraţia ambelor puncte. Ştim că, pe V, φ se scrie ca

$$\varphi(x) = \sum_{n=0}^{\infty} a_n (x - \overline{x})^n .$$

Cum f este neconstantă, există N, cel mai mic număr natural nenul pentru care $a_N \neq 0$. Deci

$$\varphi(x) = a_0 + \sum_{n=N}^{\infty} a_n (x - \overline{x})^n .$$

Definim, pe V, funcţia $F(x) = \sum_{n=N}^{\infty} a_n (x - \overline{x})^{n-N}$ şi avem

$$\varphi(x) = a_0 + (x - \overline{x})^N F(x).$$

Cum seria ce defineşte F are aceeaşi rază de convergenţă ca şi φ, deducem că F este continuă în $\overline{x}$. În plus, $F(\overline{x}) = a_N \neq 0$ şi prin urmare F are semn constant în apropierea lui $\overline{x}$, deci există $\varepsilon > 0$ astfel încât F are semn constant pe $[\overline{x}-\varepsilon, \overline{x}+\varepsilon]$ şi $[\overline{x}-\varepsilon, \overline{x}+\varepsilon] \subset V$. Cum $\overline{x}$ este extrem pentru φ, N este număr par. Din ipoteza $\varphi'(x) \neq 0$ pentru $x \in V \setminus \{\overline{x}\}$, pentru că $\overline{x}$ este punct de extrem, deducem că φ' este pozitivă şi respectiv negativă pe cele două zone ale lui V : $V \cap (-\infty, \overline{x})$ şi $V \cap (\overline{x}, \infty)$. Deci φ este strict monotonă, de monotonii diferite pe $[\overline{x}-\varepsilon, \overline{x}]$ şi pe $[\overline{x}, \overline{x}+\varepsilon]$. Fie ψ inversa lui φ pe $[\overline{x}, \overline{x}+\varepsilon]$. Din monotonia descrisă anterior, există $\delta \in (0, \varepsilon)$ astfel încât $\psi \circ \varphi$ este corect definită pe $[\overline{x}-\delta, \overline{x}]$. Fie $x \in (\overline{x}-\delta, \overline{x})$. Atunci notăm $z = (\psi \circ \varphi)(x)$. Avem $z \in (\overline{x}, \overline{x}+\varepsilon)$ şi

$$\begin{aligned} \varphi(z) &= (\varphi \circ (\psi \circ \varphi))(x) = ((\varphi \circ \psi) \circ \varphi)(x) \\ &= (\varphi \circ \psi)(\varphi(x)) = \varphi(x). \end{aligned}$$

Deci

$$a_0 + (z - \overline{x})^N F(z) = a_0 + (x - \overline{x})^N F(x),$$

adică

$$0 > \frac{z - \overline{x}}{x - \overline{x}} = -\sqrt[N]{\frac{F(x)}{F(z)}}.$$

Pentru $x \to \overline{x}$ avem $z \to \overline{x}$ şi continuitatea lui F ne conduce la concluzia privind limita. □

Observaţia 3.7.2 *Rezultatul nu are loc dacă φ este doar derivabilă în $\overline{x}$. De exemplu, funcţia $\varphi : \mathbb{R} \to \mathbb{R}$ dată prin*

$$\varphi(x) = \begin{cases} x^2, & \text{pentru } x < 0 \\ x^3, & \text{pentru } x \geq 0 \end{cases}$$

este derivabilă, $\overline{x} = 0$ este punct de minim, iar inversa ramurii din dreapta este funcţia $\sqrt[3]{x}$. Dar

$$\lim_{x \nearrow 0} \frac{\sqrt[3]{x^2}}{x} = -\infty.$$

Corolarul 3.7.3 *Are loc egalitatea*

$$\lim_{x \nearrow -1} \frac{W\left(f\left(x\right)\right)+1}{x+1} = -1.$$

Prezentăm câteva aplicaţii ale funcţiei W.

Aplicaţia 1. Prima aplicaţie a funcţiei W pe care o menţionăm este exprimarea soluţiilor unor ecuaţii ce nu pot fi rezolvate explicit. Un astfel de exemplu este ecuaţia

$$ax + b + ce^{dx} = 0$$

unde $a, b, c, d \in \mathbb{R}$ cu $a, d \neq 0$. Atunci ecuaţia

- are soluţia unică

$$x = -\frac{b}{a} - \frac{1}{d} W\left(\frac{cde^{-\frac{bd}{a}}}{a}\right).$$

  dacă numărul $\frac{cde^{-\frac{bd}{a}}}{a} \geq 0$;

- are soluţiile

$$x = -\frac{b}{a} - \frac{1}{d} W\left(\frac{cde^{-\frac{bd}{a}}}{a}\right)$$
$$x = -\frac{b}{a} - \frac{1}{d} W_{-1}\left(\frac{cde^{-\frac{bd}{a}}}{a}\right)$$

  dacă $\frac{cde^{-\frac{bd}{a}}}{a} \in \left[-\frac{1}{e}, 0\right)$;

- nu are soluţie dacă $\frac{cde^{-\frac{bd}{a}}}{a} < -\frac{1}{e}$.

Într-adevăr, ecuaţia iniţială se scrie echivalent astfel

$$\begin{aligned}
\left(ax+b\right)e^{-dx} &= -c \\
\left(-dx - \frac{bd}{a}\right)e^{-dx} &= \frac{cd}{a} \\
\left(-dx - \frac{bd}{a}\right)e^{-dx-\frac{bd}{a}} &= \frac{cd}{a}e^{-\frac{bd}{a}}.
\end{aligned}$$

Deci obţinem concluziile prezentate mai sus.

Aplicaţia 2. Prezentăm acum o aplicaţie în mecanică. Considerăm un proiectil care este aruncat de la sol (din origine) sub un unghi θ cu viteza v. Dorim să ştim distanţa R (pe axa ox pozitivă) la care proiectilul cade pe sol.

Considerăm mai întâi cazul fără frecare (rezistenţă). Ecuaţiile diferenţiale care descriu mişcarea sunt

$$x'' = 0,$$
$$y'' = -g.$$

Integrăm de două ori şi ţinem seama de condiţiile iniţiale, $x(0) = y(0) = 0$, $x'(0) = v\cos\theta$, $y'(0) = v\sin\theta$, pentru a găsi

$$x(t) = (v\cos\theta)\, t$$
$$y(t) = -\frac{1}{2}gt^2 + (v\sin\theta)\, t.$$

E clar că traiectoria este o parabolă şi impunând condiţia $y(t) = 0$ soluţia nenulă ne conduce la

$$R = \frac{2v^2}{g}\sin\theta\cos\theta = \frac{v^2}{g}\sin 2\theta.$$

Deci distanţa maximă va fi când unghiul θ este $\frac{\pi}{4}$.

Considerăm acum cazul în care mişcarea proiectilului întâmpină o rezistenţă proporţională cu viteza sa. Constanta de proporţionalitate se notează cu k. Acum ecuaţiile mişcării sunt

$$x'' = -kx',$$
$$y'' = -g - ky'.$$

Folosind $x'(0) = v\cos\theta$, $y'(0) = v\sin\theta$ avem

$$x'(t) = (v\cos\theta)\, e^{-kt},$$
$$y'(t) = \frac{1}{k}\left(-g + (g + kv\sin\theta)\, e^{-kt}\right).$$

Integrăm şi utilizăm $x(0) = y(0) = 0$ pentru a găsi

$$x(t) = \frac{1}{k}\left(v\cos\theta\left(1 - e^{-kt}\right)\right),$$
$$y(t) = \frac{1}{k^2}\left(-ktg + g + kv\sin\theta - e^{-kt}\left(g + kv\sin\theta\right)\right).$$

Ca mai sus, trebuie să rezolvăm ecuaţia $y(t) = 0$, care de data aceasta nu mai este atât de simplă. Observăm că această ecuaţie este de tipul celei de mai sus. Mai precis,

$$y(t) = 0 \iff -kt + 1 + \frac{kv}{g}\sin\theta - e^{-kt}\left(1 + \frac{kv}{g}\sin\theta\right) = 0.$$

Notăm, pentru simplitate, $u = -\left(1 + \frac{kv}{g}\sin\theta\right)$ şi ecuaţia devine

$$-kt - u + e^{-kt}u = 0.$$

Cum numărul care decidea asupra soluţilor acestui tip de ecuaţie este acum $ue^u \in \left[-\frac{1}{e}, 0\right)$, aplicând formula de mai sus găsim soluţiile

$$t_1 = \frac{-u}{k} + \frac{1}{k} W\left(ue^u\right),$$
$$t_2 = \frac{-u}{k} + \frac{1}{k} W_{-1}\left(ue^u\right).$$

Este esenţial de remarcat faptul că $u < -1$, deci $W_{-1}\left(ue^u\right) = u$, adică $t_2 = 0$ (timpul de start). Pe de altă parte, din acelaşi motiv, $W\left(ue^u\right) \neq u$, deci t_1 corespunde distanţei căutate. Înlocuind în x avem

$$R = \frac{1}{k} v \cos\theta \left(1 - e^{u - W\left(ue^u\right)}\right).$$

Cum $e^{-W\left(ue^u\right)} = \frac{W\left(ue^u\right)}{ue^u}$, avem

$$R = \frac{1}{k} v \cos\theta \left(1 - \frac{W\left(ue^u\right)}{u}\right).$$

Deci, în final,

$$R = \frac{1}{k} v \cos\theta \left(1 + \frac{W\left(\left(-1 - \frac{kv}{g}\sin\theta\right) e^{-1 - \frac{kv}{g}\sin\theta}\right)}{1 + \frac{kv}{g}\sin\theta}\right).$$

Observaţia 3.7.4 *Remarcăm că atunci când $k \to 0$ valoarea lui R tinde la valoarea din cazul fără rezistenţă. Cu notaţiile de mai sus, $u = -1 - \frac{kv}{g}\sin\theta$, avem*

$$\frac{1}{k} = -\frac{v \sin\theta}{g\left(u + 1\right)}$$

deci, conform calculului anterior,

$$\begin{aligned} R\left(k\right) &= -\frac{v^2 \sin\theta\cos\theta}{g\left(u+1\right)} \left(1 - \frac{W\left(ue^u\right)}{u}\right) \\ &= \frac{v^2 \sin 2\theta}{2g} \frac{W\left(ue^u\right) - u}{u\left(u+1\right)}. \end{aligned}$$

Cum $u \nearrow -1$, pentru $k \searrow 0$ avem

$$\begin{aligned} \lim_{k \searrow 0} R\left(k\right) &= \frac{v^2 \sin 2\theta}{2g} \lim_{u \nearrow -1} \frac{W\left(ue^u\right) - u}{u\left(u+1\right)} \\ &= -\frac{v^2 \sin 2\theta}{2g} \lim_{u \nearrow -1} \frac{W\left(ue^u\right) - u - 1 + 1}{u+1} \\ &= -\frac{v^2 \sin 2\theta}{2g} \lim_{u \nearrow -1} \left(\frac{W\left(ue^u\right) + 1}{u+1} - 1\right) \\ &= \frac{v^2 \sin 2\theta}{g}, \end{aligned}$$

adică rezultatul aşteptat. Menţionăm că pentru ultima limită s-a folosit Corolarul 3.7.3.

Aplicaţia 3. A treia aplicaţie se referă la un aspect din teoria reţelelor aleatorii şi, în particular, la o estimare a vitezei de propagare a unei pandemii. Fie o populaţie de N persoane. Presupunem că fiecare persoană se află în contact cu un număr de a persoane la întâmplare (a poate fi o expectanţă, neconstantă pentru toate persoanele, şi nu neaparat număr întreg). Dacă o persoană este infectată cu o boală transmisibilă şi boala se transmite tuturor celor cu care este în contanct, ne interesează numărul total de persoane care se vor infecta. Sau, mai bine spus, care este probabilitatea ca întreaga populaţie să fie infectată?

Pentru a transpune problema în limbajul iniţial, N reprezintă un număr de neuroni şi fiecare are a axoni ce pot forma sinapse cu alţi neuroni. Conexiunile axonilor sunt echiprobabile (în particular, un neuron poate face mai multe sinapse cu un alt neuron). Spunem că un neuron B este la t axoni de A dacă t este cel mai mic număr de axoni prin care A şi B sunt conectaţi. Exident, se poate ca A şi B să nu fie conectaţi, caz în care t nu există. În descrierea modelului, a neuroni, aleatoriu, sunt la un axon de fiecare neuron A.

Fie A un neuron. Notăm cu $x(N,a,t)$ numărul de neuroni aflaţi la cel mult t axoni de A. Conectivitatea slabă a reţelei se defineşte prin

$$\gamma(N,a)=\frac{x(N,a,N)}{N}.$$

Cum lucrăm cu N,a fixate, notăm $x(N,a,t)$ prin $x(t)$ şi scrien formula de mai sus prin

$$\gamma=\frac{x(N)}{N}.$$

Vom descrie în continuare un argument euristic pentru deducerea valorii lui γ. Menţionăm că, pe baza unui efort considerabil, se poate arăta că argumentul poate fi făcut riguros.

Căutăm să exprimăm numărul de neuroni aflaţi la exact t axoni de A, adică valoarea $x(t+1)-x(t)$.

Considerăm experimentul de a plasa la întâmplare s bile în N cutii. Valoarea aşteptată a numărului de cutii ocupate de măcar o bilă este dată de

$$N\left(1-\left(1-\frac{1}{N}\right)^{s}\right). \tag{3.9}$$

La fiecare pas $a(x(t)-x(t-1))$ noi neuroni sunt conectaţi, deci la pasul $(t+1)$ vor fi conectaţi

$$N\left(1-\left(1-\frac{1}{N}\right)^{a(x(t)-x(t-1))}\right)$$

noi neuroni. Deci

$$x(t+1)-x(t)=(N-x(t))\left(1-\left(1-\frac{1}{N}\right)^{a(x(t)-x(t-1))}\right).$$

Notăm

$$y(t)=N-x(t)$$

(numărul de neuroni neconectaţi de A după t paşi) şi ecuaţia se scrie

$$y(t+1) = y(t)\left(1-\frac{1}{N}\right)^{a(y(t-1)-y(t))},$$

sau

$$y(t+1)\left(1-\frac{1}{N}\right)^{ay(t)} = y(t)\left(1-\frac{1}{N}\right)^{ay(t-1)},$$

deci

$$y(t+1)\left(1-\frac{1}{N}\right)^{ay(t)}$$

este o constantă pe care o notăm prin K. Astfel,

$$y(t+1) = K\left(1-\frac{1}{N}\right)^{-ay(t)}. \tag{3.10}$$

Dar

$$y(0) = N-1,$$

deci, conform (3.9),

$$y(1) = (N-1)^{a+1} N^{-a}.$$

Obţinem

$$K = N^{-aN}(N-1)^{aN+1}.$$

Evident $y(t)$ este o cantitate descrescătoare în t, deci atunci când N creşte şi t creşte, intuim din (3.10) că limita y a acestei cantităţi va satisface ecuaţia

$$y = (N-1)\left(1-\frac{1}{N}\right)^{a(N-y)}.$$

Pentru N mare această valoare poate fi aproximată prin

$$Ne^{a\left(\frac{y}{N}-1\right)},$$

deci

$$\gamma = 1-\frac{y(t)}{N}$$

poate fi aproximată prin soluţia ecuaţiei

$$\gamma = 1-e^{-a\gamma}.$$

Astfel, conform Aplicaţiei 1,

$$\gamma = 1+\frac{W(-ae^{-a})}{a}.$$

Astfel, revenind la problema pandemiei, dacă în medie o persoană are contact aleatoriu cu a persoane, atunci numărul de persoane infectate va fi

$$N\gamma = N\left(1 + \frac{W\left(-ae^{-a}\right)}{a}\right).$$

Folosind software-ul de calcul ştiinţific Octave, putem calcula procentajul de infecţii în cadrul populaţiei.

Astfel, funcţia W se află în pachetul `symbolic` care trebuie instalat prin comanda `pkg install -forge symbolic`. De exemplu, folosind codul

```
a=0;
for k=2:10
    a=1+lambertw(-k*exp(-k))/k
endfor
```

obţinem valorile `0.79681, 0.94048, 0.98017, 0.99302, 0.99748, 0.99908, 0.99966, 0.99988, 0.99995`.

Deci,

- pentru $a = 2$, vor fi infectaţi $79,68\%$ din membrii populaţiei;
- pentru $a = 3$, vor fi infectaţi $94,04\%$ din membrii populaţiei;
- pentru $a = 4$, vor fi infectaţi $98,01\%$ din membrii populaţiei;
- pentru $a = 5$, vor fi infectaţi $99,30\%$ din membrii populaţiei;
- pentru $a = 10$, vor fi infectaţi $99,99\%$ din membrii populaţiei.

Exerciţiul 3.26 *Să se scrie un program propriu de determinare aproximativă a valorilor de mai sus.*

3.8 O metodă cvasi-Newton: metoda secantei

Discutăm în cazul funcţiilor de o variabilă o alternativă la metoda lui Newton despre care am văzut că are o viteză de convergenţă foarte bună, dar necesită la fiecare pas calculul derivatei într-un punct, calcul ce în unele situaţii poate fi costisitor din punct de vedere computaţional. Ideea noii metode, intitulată metoda secantei (pentru motive pe care le vom înţelege în curând), este de a înlocui derivata din calculul iteraţiei Newton

$$x_{k+1} = x_k - \frac{f\left(x_k\right)}{f'\left(x_k\right)}$$

cu o aproximare a sa (deci este o metodă de tip cvasi-Newton). Astfel, obţinem o iteraţie de forma

$$x_{k+1} = x_k - \frac{f(x_k)}{d_k},$$

unde

$$d_k = \frac{f(x_k + h_k) - f(x_k)}{h_k},$$

raport care pentru h_k mic aproximează $f'(x_k)$. Pentru alegerea lui h_k se poate proceda în diverse moduri, dar în metoda secantei se alege $h_k = x_{k-1} - x_k$. Astfel, iteraţia devine

$$x_{k+1} = x_k - f(x_k) \frac{x_{k-1} - x_k}{f(x_{k-1}) - f(x_k)}.$$

Evident, pentru ca recurenţa să fie definită, este nevoie de două date iniţiale x_0, x_1. O ilustrare grafică a semnificaţiei iteraţiei de mai sus justifică numele metodei. Este de asemenea uşor de intuit că limita $h_k \to 0$ nu este automat satisfăcută, ci depinde de unele detalii tehnice legate de alegerea datelor iniţiale. Pe de altă parte, noua iteraţie nu face decât o singură evaluare a lui f la fiecare pas.

Prezentăm acum detaliile matematice ale convergenţei acestei metode.

Lema 3.8.1 *Fie $f : [a,b] \to \mathbb{R}$ o funcţie derivabilă astfel încât*

$$\mu = \inf_{x \in [a,b]} |f'(x)| > 0.$$

Atunci există $H > 0$ astfel încât pentru orice $x \in [a,b]$ şi orice $h \in [-H, H]$ astfel încât $x+h \in [a,b]$ avem

$$\left| \frac{f(x+h) - f(x)}{h} \right| \geq \frac{\mu}{2}.$$

Demonstraţie. Presupunem, prin reducere la absurd, că există $(x_k) \subset [a,b]$, $h_k \to 0$ astfel încât $x_k + h_k \in [a,b]$ cu

$$\left| \frac{f(x_k + h_k) - f(x_k)}{h_k} \right| < \frac{\mu}{2}, \ \forall k.$$

Conform Teoremei lui Lagrange, pentru fiecare k, există ν_k între $x_k + h_k$ şi x_k astfel încât

$$\frac{f(x_k + h_k) - f(x_k)}{h_k} = f'(\nu_k).$$

Cum $(\nu_k) \subset (a,b)$, această relaţie contrazice ipoteza. □

Următoarea lemă este conţinută în demonstraţia Teoremei 3.6.3, dar o prezentăm explicit.

Lema 3.8.2 *Fie $f : [a,b] \to \mathbb{R}$ o funcţie derivabilă astfel încât f' este Lipschitz de constantă $L > 0$ pe $[a,b]$. Atunci pentru orice $x, y \in [a,b]$*

$$|f(y) - f(x) - f'(x)(y-x)| \leq \frac{L}{2}(x-y)^2.$$

Demonstraţie. Scrierea

$$|f(y) - f(x) - f'(x)(y-x)| = \left| \int_x^y (f'(t) - f'(x))\, dt \right|$$

şi ipoteza asigură concluzia. □

Teorema 3.8.3 *Fie $f : [a,b] \to \mathbb{R}$ o funcţie derivabilă astfel încât*
(i) f se anulează într-un punct $\overline{x} \in (a,b)$;
(ii) $\mu = \inf_{x\in[a,b]} |f'(x)| > 0$;
(iii) f' este Lipschitz de constantă $L > 0$ pe $[a,b]$.
Atunci există $\delta, \eta > 0$ astfel încât pentru orice dată iniţială $x_0 \in (\overline{x} - \delta, \overline{x} + \delta)$ şi orice şir $\{h_k\} \subset (-\eta, \eta) \setminus \{0\}$, iteraţiile

$$x_{k+1} = x_k - f(x_k) \frac{h_k}{f(x_k + h_k) - f(x_k)}$$

sunt bine definite şi rămân în intervalul $(\overline{x} - \delta, \overline{x} + \delta)$ şi, mai mult, $x_k \to \overline{x}$.

Demonstraţie. Pentru buna definire este suficient ca $x_k + h_k, x_k \in (a,b)$ pentru orice k (cf. Teoremei lui Lagrange şi ipotezei (ii)), lucru care se va realiza pentru alegeri potrivite ale constantelor δ şi η, după cum se va vedea mai jos.

Definim $\varepsilon_k = \overline{x} - x_k$ pentru orice k. Din Lema 3.8.1,

$$d_k = \left| \frac{f(x_k + h_k) - f(x_k)}{h_k} \right| \geq \frac{\mu}{2}$$

dacă $|h_k| \leq H$ şi $x_k, x_k + h_k \in (a,b)$. Alegem deci $\{h_k\} \subset (-H, H)$. Presupunem pentru moment că x_k şi $x_k + h_k$ sunt în (a,b). Atunci,

$$\begin{aligned}
\varepsilon_{k+1} = \overline{x} - x_{k+1} &= \overline{x} - x_k + \frac{f(x_k)}{d_k} \\
&= \overline{x} - x_k + \frac{f(x_k)}{d_k} - \frac{f(\overline{x})}{d_k} \\
&= \varepsilon_k + \frac{f(x_k) - f(\overline{x})}{d_k} \\
&= \frac{1}{d_k} (f(x_k) - f(\overline{x}) + \varepsilon_k d_k) \\
&= \frac{1}{d_k} (f(x_k) - f(\overline{x}) + f'(x_k)\varepsilon_k + \varepsilon_k d_k - f'(x_k)\varepsilon_k),
\end{aligned}$$

deci, folosind şi Lema 3.8.2,

$$\begin{aligned}
|\varepsilon_{k+1}| &\leq \frac{1}{|d_k|} (|f(\overline{x}) - f(x_k) - f'(x_k)\varepsilon_k| + |d_k - f'(x_k)|\,|\varepsilon_k|) \\
&\leq \frac{2}{\mu} \left(\frac{L|\varepsilon_k|^2}{2} + \frac{L|h_k|\,|\varepsilon_k|}{2} \right).
\end{aligned}$$

Am găsit astfel estimarea,

$$|\varepsilon_{k+1}| \leq \frac{L}{\mu}\left(|\varepsilon_k| + |h_k|\right)|\varepsilon_k|$$

în ipotezele mai sus menţionate.

Trecem acum la demonstraţia efectivă a concluziei. Fie $\theta \in (0,1)$, fie $\rho = \min\{|\overline{x} - a|, |\overline{x} - b|\}$ şi $\delta = \min\left\{\frac{\rho}{2}, \frac{\theta\mu}{2L}\right\}$, $\eta = \min\{H, \delta\}$. Fie (x_k) un şir de iteraţii generat ca în enunţul rezultatului. Pentru a obţine concluziile este suficient să arătăm că

$$|\varepsilon_{k+1}| \leq \theta\,|\varepsilon_k|,\ \forall k.$$

Pentru $k = 0$, $x_0 \in (a,b)$, $x_0 + h_0 \in (a,b)$ (din alegerea constantelor) şi putem aplica estimarea anterioară

$$|\varepsilon_1| \leq \frac{L}{\mu}\left(|\varepsilon_0| + |h_0|\right)|\varepsilon_0| \leq \frac{L}{\mu} \cdot 2\delta \cdot |\varepsilon_0| \leq \frac{L}{\mu} \cdot 2\frac{\theta\mu}{2L} \cdot |\varepsilon_0| = \theta\,|\varepsilon_0|.$$

Inductiv, se obţine inegalitatea anunţată şi demonstraţia este încheiată. □

Corolarul 3.8.4 *În ipotezele Teoremei 3.8.3, dacă $h_k \to 0$, atunci convergenţa lui (x_k) este superliniară.*

Demonstraţie. Din demonstraţia Teoremei 3.8.3 deducem

$$|\varepsilon_{k+1}| \leq \frac{2L}{\mu}\max\{|\varepsilon_k|, |h_k|\}\,|\varepsilon_k|,\ \forall k \geq 0.$$

Cum $\varepsilon_k \to 0$ (conform Teoremei 3.8.3) şi $h_k \to 0$ (conform ipotezei), avem şi $\max\{|\varepsilon_k|, |h_k|\}$. Astfel, are loc convergenţa superliniară. □

În continuare, precizăm ordinul de convergenţă al metodei secantei. Din nou, începem cu un rezultat ajutător.

Lema 3.8.5 *Fie $f : [a,b] \to \mathbb{R}$ de două ori derivabilă şi $x, y, z \in [a,b]$ puncte distincte. Atunci există $u \in (\min\{x,y,z\}, \max\{x,y,z\})$ astfel încât*

$$e(x,y,z) := \frac{\frac{f(y)-f(z)}{y-z} - \frac{f(x)-f(y)}{x-y}}{z-x} = \frac{f''(u)}{2}.$$

Demonstraţie Considerăm funcţia polinomială

$$p(u) = f(x) + \frac{f(y) - f(x)}{y - x}(u - x) + e(x,y,z)(u-x)(u-y)$$

şi constatăm că $f(x) = p(x)$, $f(y) = p(y)$, $f(z) = p(z)$. Aplicând Teorema lui Rolle funcţiei $f - p$, găsim două zerouri are derivatei sale şi tot din Teorema lui Rolle pentru $(f-p)'$ găsim u astfel încât $(f-p)''(u) = 0$, ceea ce conduce la concluzie. □

Teorema 3.8.6 *Fie $f \in C^2([a,b])$ astfel încât*

(i) f se anulează într-un punct $\overline{x} \in (a,b)$;

(ii) $\mu = \inf_{x\in[a,b]} |f'(x)| > 0$;

(iii) $L = \sup_{x\in[a,b]} |f''(x)| > 0$.

Dacă metoda secantei cu datele iniţiale $x_0, x_1 \in [a,b]$ converge la $\overline{x}$ atunci există $C > 0$ astfel încât

$$|\varepsilon_{k+1}| \leq C |\varepsilon_k|^{\varphi}, \ \forall k,$$

unde $\varphi = \frac{1+\sqrt{5}}{2}$ este numărul de aur.

Demonstraţie. Pentru orice k avem

$$\begin{aligned}
-\varepsilon_{k+1} = x_{k+1} - \overline{x} &= x_k - \overline{x} - f(x_k) \frac{x_{k-1} - x_k}{f(x_{k-1}) - f(x_k)} \\
&= -\varepsilon_k - f(x_k) \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \\
&= \varepsilon_k \varepsilon_{k-1} \frac{\frac{f(\overline{x})-f(x_k)}{\overline{x}-x_k} - \frac{f(x_k)-f(x_{k-1})}{x_k - x_{k-1}}}{\frac{f(x_k)-f(x_{k-1})}{x_k - x_{k-1}}}. \\
&= \varepsilon_k \varepsilon_{k-1} \frac{f''(v_k)}{2f'(u_k)},
\end{aligned}$$

unde $u_k, v_k \in (a,b)$. Existenţa lui v_k rezultă din Teorema lui Lagrange, în timp ce existenţa lui (v_k) rezultă din Lema 3.8.5. De asemenea, este clar că $(u_k), (v_k) \to \overline{x}$.

Din ipoteză,

$$|\varepsilon_{k+1}| \leq \frac{2L}{\mu} |\varepsilon_k| |\varepsilon_{k-1}|, \ \forall k \geq 0.$$

Notăm $M := \frac{2L}{\mu}$. Cum $|\varepsilon_{k-1}| \to 0$, deducem că ordinul de convergenţă ar putea fi mai mare decât 1.

Arătăm că ordinul de convergenţă este φ, soluţia supraunitară a ecuaţiei $p^2 = p + 1$. Pentru determina ordinul de convergenţă, căutăm două constante C şi $p > 1$ astfel încât pentru orice k suficient de mare să avem

$$|\varepsilon_{k+1}| \leq C |\varepsilon_k|^p.$$

Atunci,

$$|\varepsilon_{k+1}| \leq C |\varepsilon_k|^p \leq C^{p+1} |\varepsilon_{k-1}|^{p^2}$$

şi

$$|\varepsilon_{k+1}| \leq M |\varepsilon_k| |\varepsilon_{k-1}| \leq MC |\varepsilon_{k-1}|^p.$$

Este suficient ca

$$\begin{aligned}
C^{p+1} &= MC \\
p^2 &= p + 1,
\end{aligned}$$

de unde $p = \varphi$ şi $C = M^{\frac{1}{\varphi}}$. □

Observaţia 3.8.7 *Aşadar, în anumite ipoteze metoda secantei are ordinul de convergenţă φ. Cum $\varphi^2 > 2$, iar uneori două iteraţii ale acestei metode sunt mai rapide decât o iteraţie a metodei lui Newton, în astfel de cazuri, metoda secantei este mai eficientă.*

Exerciţiul 3.27 *Să se implementeze metoda secantei (eventual, pentru una dintre funcţiile studiate la metoda lui Newton).*

Observaţia 3.8.8 *Metoda secantei aplicată derivatei conduce la o metodă de ordin 1 pentru aproximarea punctelor critice (de minim, în unele cazuri).*

Exerciţiul 3.28 *Să se folosească metoda secantei pentru determinarea minimului funcţiei de la Exemplul 1.1.1 (cu derivata calculată exact şi/sau numeric).*

Exerciţiul 3.29 *Folosiţi metoda înjumătăţirii intervalului, metoda lui Newton, metoda cvasi-Newton şi metoda secantei pentru a aproxima rădăcini ale ecuaţiilor de mai jos.*

(i) $\cos x - xe^x = 0$; *(ii)* $x^4 - x - 10 = 0$; *(iii)* $x - e^{-x} = 0$; *(iv)* $e^{-x}\left(x^2 - 5x + 2\right) + 1 = 0$;
(v) $x - \sin x - 2^{-1} = 0$; *(vi)* $e^{-x} = 3\ln x$; *(vii)* $e^{x^2-1} + 10\sin 2x - 5 = 0$;
(viii) $\sin 2x - e^{x-1} = 0$; *(ix)* $3x + \sin x - e^x = 0$ *(x)* $\operatorname{tg} x - x - 1 = 0$; *(xi)* $-5x^2 + e^x = 0$.

Capitolul 4

Algoritmi pentru probleme de optimizare fără restricţii

Aşa cum am putut constata, multe dintre problemele de optimizare nu pot fi rezolvate exact. De asemenea, nu există o metodă universală de aproximare a soluţiilor, ci doar unele principii generale privind designul majorităţii algoritmilor pe care îi considerăm. Pe baza acestor principii vom încerca să construim clase de algoritmi ce sunt eficienţi pentru clase specifice (particulare) de probleme. În acest capitol ne ocupăm de algoritmi pentru probleme fără restricţii.

Până în acest moment, am discutat deja câţiva algoritmi pentru aproximarea punctelor de minim, dar cei mai eficienţi dintre aceştia sunt bazaţi pe metode de determinare a zerourilor unei funcţii aplicate diferenţialei funcţiei de minimizat (metoda lui Newton şi metoda secantei). În acest capitol vom studia algoritmi care sunt de la bun început concepuţi pentru minimizare. Astfel, începem cu algoritmi de ordin zero pentru funcţii de o variabilă, ştiut fiind faptul că astfel de algoritmi sunt greoi din punct de vedere computaţional pentru dimensiuni mari (a se vedea Exemplul 3.4.4) şi vom continua cu algoritmi de ordin superior în cazul general.

4.1 Algoritmi de ordin zero pentru funcţii de o variabilă

Prima categorie de algoritmi de care ne ocupăm este cu oracol de ordin zero, adică fără diferenţiale, baza reprezentând-o cunoaşterea funcţiei obiectiv în punctele de control. Deci, în principiu, aceşti algoritmi funcţionează şi pentru funcţii nediferenţiabile.

Aşa cum am văzut (Teorema 3.4.3), pentru mai funcţii de mai multe variabile, complexitatea metodelor de ordin zero este descurajantă, motiv pentru care ne vom concentra, mai ales, pe cazul funcţiilor de o variabilă reală. Mai mult, funcţiile de care ne vom ocupa vor fi continue şi definite pe intervale compacte. În plus, vom impune o proprietate foarte importantă, numită proprietatea de unimodalitate.

Fie deci $f : [a, b] \to \mathbb{R}$ o funcţie continuă ($a, b \in \mathbb{R}$, $a < b$). În virtutea Teoremei lui Weierstrass, există $\overline{x} \in [a, b]$, un punct în care f îşi atinge minimul global pe $[a, b]$. Cum căutăm metode de a aproxima un astfel de punct, este rezonabil să considerăm că $\overline{x} \in (a, b)$, pentru că, în caz contrar, $\overline{x}$ este cunoscut cu exactitate. Spunem că f este unimodală dacă este strict descrescătoare pe $[a, \overline{x}]$ şi strict crescătoare pe $[\overline{x}, b]$. Evident, în această ipoteză, $\overline{x}$ este unicul punct de minim global pentru f.

4.1.1 Delimitarea (bracketing) punctului de minim

Fie deci $f : [a, b] \to \mathbb{R}$ continuă şi unimodală. Fie $\overline{x} \in (a, b)$ unicul său punct de minim. Ne interesează să determinăm un interval mai restrâns în care se găseşte $\overline{x}$. Fie $x \in (a, b)$. Dacă $f(a) < f(x)$, cum f este unimodală, putem spune că $\overline{x} \in (a, x)$. Altfel, dacă $f(a) > f(x)$, nu se poate spune nimic cu privire la poziţionarea lui $\overline{x}$ în raport cu cele două intervale delimitate de $\overline{x}$. Ideea este aceea de a creşte suficient de încet punctul x plecând din a astfel încât să nu îl depăşim pe $\overline{x}$.

Implementare 4.1 Ilustrăm această discuţie cu următorul cod Matlab/Octave:

```
functie=@(x) [x*x-sqrt(17)*x+3];
fplot(functie, [-2,7]);
a=-2; b=7; s=0.0001; x=0; k=0;
while functie(a)>functie(a+s)
 a=a+s; k=k+1;
endwhile
a
functie(a)
k
```

care returnează: `a = 2.0616; ans = -1.2500; k = 40616`

Exerciţiul 4.2 *Să se deseneze gaficul funcţiei $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^6 - 4x^5 + 3x^3 + 6x$ (a se vedea şi Exerciţiul 3.1) pe diferite intervale şi să de deducă faptul că există două puncte de minim. Apoi, să de delimiteze în jurul fiecărui minim un interval pe care funcţia este unimodală şi să se aproximeze aceste puncte prin metoda de mai sus.*

Chiar dacă este un procedeu simplu şi oarecum euristic, această metodă ilustrează o idee ce se întâlneşte des şi la algoritmi mai sofisticaţi, şi anume ideea unei mişcări progresive în direcţia potrivită cu un pas mic (dar totuşi nu prea mic, pentru a evita un calcul prea costisitor).

Putem îmbunătăţi metoda de mai sus după principiul reţelei echidistante, astfel: alegem un număr natural nenul $n > 1$ de subdiviziuni egale ale intervalului $[a, b]$, deci lungimea fiecărei subdiviziuni este $\delta := (b - a)/n$. Calculăm valoarea lui f în toate numerele $a + i\delta$ cu $i \in \overline{0, n}$ şi alegem punctul în care se obţine cea mai mică valoare. Obţinem un punct $x_{\min}$ şi o valoare $f(x_{\min})$, iar intervalul de căutare pentru pasul următor este $[x_{\min} - \delta, x_{\min} + \delta]$. Pe acest nou interval de lungime 2δ folosim acelaşi procedeu: recalculăm noul δ şi aşa mai departe. Procesul de căutare se va termina când lungimea intervalului de căutare este mai mică decât o valoare $\varepsilon > 0$ prestabilită.

Exerciţiul 4.3 *Să se implementeze această metodă pe funcţia de la exerciţiul anterior.*

Aşa cum am văzut, acest algoritm este destul de costisitor din punct de vedere computaţional pentru că presupune calcularea funcţiei (invocarea oracolului de ordin zero) în multe puncte.

În orice caz, pentru orice algoritm de acest tip apare problema determinării punctelor a şi b între care facem căutarea. Aceasta poate fi rezolvată astfel: alegem u un punct arbitrar şi $\delta > 0$ o constantă. Dacă

$$f(u-\delta) \geq f(u) \geq f(u+\delta),$$

atunci, cu siguranţă, pe baza unimodalităţii, punctul de minim se află la dreapta lui u. În acest caz, ne deplasăm la dreapta cu constanta δ atât timp cât este respectată condiţia $f(u) \geq f(u+\delta)$. Dacă acestă condiţie nu este respectată, înseamnă că minimul se află în $(u-\delta, u+\delta)$. Un raţionament similar se face dacă

$$f(u-\delta) \leq f(u) \leq f(u+\delta).$$

Dacă niciuna dintre condiţiile de mai sus nu se întâmplă pentru alegerea u, cerem o altă dată iniţială (unele raţionamente ar permite să spunem că punctul de minim e relativ aproape de u, deci noua dată iniţială nu trebuie să fie foarte departe de alegerea anterioară).

Exerciţiul 4.4 *Să se implementeze această idee pentru funcţia* $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2 - \sqrt{17}x + 3$.

O altă metodă de delimitare a punctului (punctelor) de minim, dar şi a unor intervale pe care funcţia este unimodulară, e bazată pe o vizualizare grafică. Pentru codul aferent reprezentărilor grafice ale funcţiilor de o variabilă, dar şi o listă de funcţii pe care se pot testa codurile din această secţiune prezentăm liniile Matlab/Octave de mai jos.

Implementare 4.5 Codul desen_o_var.m (fişier de rulare):

```
%functie=@(x) [sqrt(abs(x))];a=-2;b=4
%functie=@(x) [x^2+27/x];a=2;b=2.5;
%functie=@(x) [x^2-sqrt(17)*x+3];a=1;b=4;
%functie=@(x) [204165.5./(330-2*x) + 10400./(x-20)];a=40;b=90;
%functie=@(x) [x^6-4*x^5+3*x^3+6*x];a=-1;b=-0.5;
%functie=@(x) [3*exp(x)-x^3+5*x];a=-3;b=3;
%functie=@(x) [-5*x^2+exp(x)];a=1;b=4;
%functie=@(x) [sqrt(x^2+25)/2+(7-x)/6];a=0;b=7;
%functie=@(x) [max((x^6),(3*x-1)^2)];a=-2;b=2;
%functie=@(x) [sqrt(100+x^2)/299702+sqrt((30-x)^2+1600)/228849];a=0,b=40;
functie=@(x) [4*pi/(log(2*x)-1/2)*x];a=1;b=7;
%functie=@(x) [exp(x)-2*x^2];a=0;b=4;
fplot(functie,[a,b]);
set(gca, 'XAxisLocation', 'origin')
set(gca, 'YAxisLocation', 'origin')
```

Exemplul 4.1.1 Unele dintre funcţiile de mai sus provin din chestiuni cu caracter practic ce nu pot fi tratate cu uşurinţă prin metodele clasice de determinare cu exactitate a punctelor de extrem (a se vedea prima secţiune a lucrării).

4.1.2 Metoda secţiunii de aur

Prima metodă mai elaborată pe care o prezentăm este bazată pe o observaţie simplă şi pe un detaliu tehnic, nu chiar simplu la rândul său.

Ideea este una similară celei de mai sus şi constă în micşorarea intervalului ce conţine punctul de minim până la atingerea unui criteriu de oprire. Avem în vedere două lucruri: o reducere optimă a lungimii intervalului ce delimitează punctul de minim şi un număr cât mai mic de invocări ale oracolului.

Fie $f : [a, b] \to \mathbb{R}$ o funcţie continuă unimodală, ca mai sus. Fie $x_1, x_2 \in [a, b]$ cu $a < x_1 < x_2 < b$. Dacă $f(x_1) \leq f(x_2)$, atunci, cu siguranţă, $\overline{x} \in (a, x_2]$, iar altfel $\overline{x} \in (x_1, b)$. Deci, în orice caz putem micşora intervalul. Ideea este de a alege judicios punctele x_1 şi x_2 în aşa fel încât să avem o descreştere semnificativă a lungimii intervalului, indiferent de care parte a intervalului iniţial este îndepărtată. O altă cerinţă este de a proceda suficient de eficient astfel încât acel punct dintre x_1 şi x_2 care rămâne în interiorul noului interval ce delimitează minimul să poată servi ca punct intermediar pentru următorul pas al iteraţiei. Dacă această a doua cerinţă este îndeplinită, atunci la noul pas este necesară evaluarea lui f doar într-un punct.

De exemplu, ne-am putea gândi la împărţirea intervalului iniţial în trei părţi egale, adică să luăm $x_1 = a + \frac{1}{3}(b - a)$ şi $x_2 = a + \frac{2}{3}(b - a)$. Procedând astfel, dacă $f(x_1) < f(x_2)$ atunci trebuie să reţinem intervalul (a, x_2) iar acum x_1 este punctul din mijloc al acestui interval. Deci, pentru următorul pas, dacă alegem un singur punct (pentru a economisi efort computaţional) acest punct nu va oferi simetrie împărţirii intervalului. Deci, această împărţire nu este optimală din acest punct de vedere.

Fie $\rho \in \left(\frac{1}{2}, 1\right)$ şi $x_1 = \rho a + (1 - \rho) b$, $x_2 = (1 - \rho) a + \rho b$. Această diviziune permite ca proporţia din lungimea iniţială care va fi eliminată să fie aceeaşi în oricare dintre cele două cazuri. Considerăm, la fiecare pas, această împărţire a intervalului. Astfel, dorim ca punctul ce rămâne în interval după primul pas să fie unul din punctele ce respectă această regulă. Dacă $f(x_1) \leq f(x_2)$ atunci trebuie să reţinem intervalul $(a, x_2]$. Pe baza cerinţelor formulate:

$$x_1 = (1 - \rho) a + \rho x_2,$$

(x_1 va fi al doilea punct pentru pasul următor, pentru că, altfel, $x_1 = \rho a + (1 - \rho) x_2$, ceea ce conduce la $\rho a + (1 - \rho) b = \rho a + (1 - \rho) x_2$, adică $x_2 = b$, ceea ce este fals), deci

$$\rho a + (1 - \rho) b = (1 - \rho) a + \rho ((1 - \rho) a + \rho b),$$

adică

$$\rho^2 + \rho - 1 = 0,$$

ceea ce conduce la

$$\rho = \frac{\sqrt{5} - 1}{2}$$

(pentru că $\rho > 2^{-1}$).

Dacă $f(x_1) > f(x_2)$ trebuie să reţinem intervalul (x_1, b). Din nou, trebuie ca:

$$x_2 = \rho x_1 + (1 - \rho) b,$$

(x_2 va fi al primul punct la pasul următor) deci

$$(1-\rho)\,a+\rho b=\rho\,(\rho a+(1-\rho)\,b)+(1-\rho)\,b,$$

şi din nou

$$\rho=\frac{\sqrt{5}-1}{2}\simeq 0.6180339.$$

În primul caz trebuie să alegem primul punct la pasul al doilea ca fiind $(1-\rho)\,a+\rho x_2$, iar în al doilea caz alegem al doilea punct ca fiind $\rho x_1+(1-\rho)\,b$.

Numele acestei metode provine din faptul că, de exemplu,

$$\frac{x_2-a}{b-x_2}=\frac{\rho}{1-\rho}=\frac{\sqrt{5}+1}{2}\simeq 1.618034,$$

acest număr fiind numit raportul de aur şi fiind tradiţional notat cu φ.

Această metodă asigură faptul că intervalul ce delimitează minimul se micşorează la fiecare pas astfel încât noua lungime este aproximativ 0.618034 din lungimea celei precedente. Mai mult, începând de la al doilea pas o singură nouă evaluare a funcţiei este necesară.

Din nou, căutarea se termină când lungimea intervalului este mai mică decât o valoare prescrisă. Aproximarea minimului este mijlocul intervalului final.

Implementare 4.6 Ilustrăm metoda prin codul de mai jos:

```
functie=@(x) [3*exp(x)-x^3+5*x];a=-3;b=3;
fplot(functie,[-3,3]);
eps=0.00001;rho=0.618034;intervallength=b-a;i=0;
while (intervallength>eps)
 x1=rho*a+(1-rho)*b;x2=(1-rho)*a+rho*b;
 if functie(x1)<=functie(x2)
 b=x2;
else
 a=x1;
endif
intervallength=b-a;
i=i+1;
endwhile
x=(a+b)/2
i
functie(x)
```

cere returnează: `x = -1.3846, i = 28, ans = -3.5173.`

Codul de mai sus nu este optimal pentru că evaluează funcţia de două ori la fiecare pas.

Exerciţiul 4.7 *Să se implemeteze un cod care să remedieze această problemă.*

Vom dori ca pe viitor să utilizăm această metodă pentru diverse funcţii, motiv pentru care vom prefera unele variante care să implementeze o funcţie Matlab/Octave.

Exerciţiul 4.8 *Să se scrie o funcţie care să implementeze acest algoritm.*

4.1.3 Metoda Fibonacci

Următoarea metodă pe care o prezentăm este apropiată de cea precedentă, dar de data aceasta, raportul de diviziune nu este constant la toţi paşii, ci se bazează pe termenii şirului lui Fibonacci pentru construirea acestuia. Amintim principalele elemente despre acest şir necesare în continuare.

Propoziţia 4.1.2 *(Şirul lui Fibonacci) Fie şirul* $(F_k)_{k\geq 0}$, *este definit cu termenii iniţiali* $F_0 = 0, F_1 = 1$ *şi prin relaţia de recurenţă:*

$$F_k = F_{k-1} + F_{k-2},\ \forall k \geq 2.$$

Atunci:

(i) pentru orice $k \geq 0$,

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

(ii)

$$\lim \frac{F_k}{F_{k-1}} = \varphi.$$

(iii) Seria

$$\sum_{k=1}^{\infty} \frac{1}{F_k}$$

este convergentă.

Demonstraţie. (i) O demonstraţie se poate face prin inducţie.

O altă demonstraţie, prin care să se ajungă la forma anunţată a şirului este următoarea. Observăm că puterile rădăcinilor ecuaţiei

$$x^2 - x - 1 = 0.$$

satisfac relaţia de recurenţă. Rădăcinile acestei ecuaţii sunt

$$\varphi = \frac{1+\sqrt{5}}{2}$$
$$\psi = \frac{1-\sqrt{5}}{2}.$$

Deducem că pentru orice numere reale a, b, termenii şirului $\left(a\varphi^k + b\psi^k\right)_k$ satisfac relaţia de recurenţă. Cum datele iniţiale împreună cu relaţia de recurenţă definesc în mod unic termenii şirului,

înseamnă că prin determinarea valorilor a, b pentru respectarea datelor iniţiale, şirul obţinut este şirul lui Fibonacci. Condiţiile

$$\begin{aligned} a+b &= 0 \\ a\varphi + b\psi &= 1 \end{aligned}$$

conduc la

$$a = -b = \frac{1}{\sqrt{5}},$$

de unde deducem formula de demonstrat.

(ii) Folosind punctul anterior putem scrie

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

(iii) Pe baza punctului de mai sus,

$$\lim \frac{F_k}{F_{k+1}} = \frac{1}{\varphi} < 1$$

şi din criteriul raportului pentru serii cu termeni pozitivi obţinem concluzia. □

Prezentăm acum metoda anunţată. Cum am spus, procedăm ca la metoda secţiunii de aur, dar fără a avea un raport constant, dar ca mai înainte, la pasul cu numărul k funcţia este testată în două puncte x_k^1 şi x_k^2 din interiorul intervalului de delimitare $[a_k, b_k]$ şi dorim să păstrăm unul dintre aceste puncte drept capăt al noului interval, iar pe celălalt ca punct interior. La iteraţia k intervalul are o lungime aflată în proporţia $r_k \in \left[\frac{1}{2}, 1\right)$ cu intervalul precedent. Astfel pentru orice k,

$$\begin{aligned} x_k^1 &= r_k a_k + (1 - r_k) b_k \\ x_k^2 &= (1 - r_k) a_k + r_k b_k. \end{aligned}$$

Presupunem, fără a restrânge generalitatea, că $f\left(x_k^1\right) \leq f\left(x_k^2\right)$, deci $a_{k+1} = a_k$ şi $b_{k+1} = x_k^2$. Renotăm punctul x_k^1 prin x_{k+1}^2 (a se vedea cazul corespunzător de la metoda secţiunii de aur). Atunci, lungimea noului interval de căutare este

$$d_{k+1} = r_k \left(b_k - a_k\right).$$

Mai mult,

$$x_k^2 - x_k^1 = b_{k+1} - x_{k+1}^2$$

şi avem

$$\begin{aligned} x_k^2 - x_k^1 &= (1 - r_k) a_k + r_k b_k - r_k a_k - (1 - r_k) b_k \\ &= (2r_k - 1)\left(b_k - a_k\right). \end{aligned}$$

În plus,

$$b_{k+1} - x_{k+1}^2 = b_{k+1} - (1 - r_{k+1})\, a_{k+1} - r_{k+1} b_{k+1} = (1 - r_{k+1})\,(b_{k+1} - a_{k+1})$$
$$= (1 - r_{k+1})\, r_k\,(b_k - a_k)\,.$$

Astfel, obţinem

$$2r_k - 1 = (1 - r_{k+1})\, r_k,$$

adică

$$r_{k+1} = \frac{1 - r_k}{r_k}.$$

În acest moment utilizăm termenii şirului lui Fibonacci.

Fixăm $n > 2$ şi definim raportul r_k ca fiind

$$r_k = \frac{F_{n-k}}{F_{n-k+1}} \in \left[\frac{1}{2}, 1\right),\ \forall k \in \{1, 2, ..., n-2\}\,.$$

Atunci

$$r_{k+1} = \frac{1 - \frac{F_{n-k}}{F_{n-k+1}}}{\frac{F_{n-k}}{F_{n-k+1}}} = \frac{F_{n-k-1}}{F_{n-k}}.$$

Deci,

$$r_1 = \frac{F_{n-1}}{F_n}$$

şi

$$r_2 = \frac{F_{n-2}}{F_{n-1}}, ..., r_{n-2} = \frac{F_2}{F_3} = \frac{1}{2}.$$

La pasul următor nu mai putem adăuga un nou punct pentru că r_{n-1} ar fi $\frac{F_1}{F_2} = 1$. Astfel, algoritmul se termină după $n-2$ paşi, unde n este numărul natural fixat mai sus. Avem

$$d_{n-1} = r_{n-2} d_{n-2} = \frac{F_2}{F_3} d_{n-2} = \frac{F_2}{F_3}\frac{F_3}{F_4} d_{n-3} = ... = \frac{1}{F_n} d_1 = \frac{1}{F_n}\,(b - a)\,,$$

deci putem fixa n astfel încât $\frac{1}{F_n}\,(b-a)$ este mai mic decât o constantă pozitivă dată.

Cum limita lui $\left(\frac{F_k}{F_{k+1}}\right)$ este $\frac{1}{\varphi} = \rho$, pentru n mare acest algoritm este practic echivalent cu algoritmul secţiunii de aur.

Implementare 4.9 Prezentăm o ilustrare de bază pentru această metodă.

```
functie=@(x) [3*exp(x)-x^3+5*x];a=-3;b=3;
n=40;
fib=zeros(1,n);
fib(1)=1;
fib(2)=1;
for i=3:n;
 fib(i)=fib(i-2)+fib(i-1);
```

```
endfor
for i=n:-1:3
 r=fib(i-1)/fib(i);
 x1=r*a+(1-r)*b;
 x2=(1-r)*a+r*b;
 if functie(x1)<=functie(x2)
 b=x2;
 else
 a=x1;
 endif
endfor
(a+b)/2
functie((a+b)/2)
```

Exerciţiul 4.10 *Să se modifice codul de mai sus astfel încât oracolul să fie la fiecare pas invocat o singură dată (deci cu păstrarea punctului intermediar).*

Observaţia 4.1.3 *Avantajul acestor algoritmi este acela că, oracolul fiind de ordin* 0*, se aplică şi funcţiilor nederivabile, aşa cum se vede din lista funcţiilor pentru care deja le-am implementat.*

Exerciţiul 4.11 *Să se implementeze metoda secantei pentru derivată în scopul aproximării punctelor de minim pentru funcţiile studiate mai sus.*

Exerciţiul 4.12 *Să se implementeze metoda înjumătăţirii intervalului pentru derivată în scopul aproximării punctelor de minim pentru funcţiile studiate mai sus.*

Exerciţiul 4.13 *Să se recapituleze şi să se compare pe exemplele concrete de mai jos metodele de ordin* 0, 1, 2 *pentru optimizarea funcţiilor de o variabilă studiate.*
(i) $f(x) = 2(x-3)^2 + e^{0.5x^2}$, $x \in [0,4]$;
(ii) $f(x) = 3x^4 + (x-1)^2$, $x \in [0,4]$;
(iii) $f(x) = -4x \sin x$, $x \in [0,\pi]$;
(iv) $f(x) = 3x^2 + 12x^{-3} - 5$, $x \in [0.1,4]$;
(v) $f(x) = 2x^2 + 16x^{-1}$, $x \in [1,5]$.

Exerciţiul 4.14 *Reprezentând grafic funcţiile de mai jos, să se observe că sunt unimodale pe intervalele indicate şi să se determine punctul de minim prin metodele anterioare.*
(i) $f(x) = -5x^5 + 4x^4 - 12x^3 + 11x^2 - 2x + 1$, $x \in [-0.5, 0.5]$;
(ii) $f(x) = \ln^2(x-2) + \ln^2(10-x) - x^{0.2}$, $x \in [6, 9.9]$;
(iii) $f(x) = -3x \sin(0.75x) + e^{-2x}$, $x \in [0, 2\pi]$;
(iv) $f(x) = e^{3x} + 5e^{-2x}$, $x \in [0,1]$;
(v) $f(x) = 0.2x \ln x + (x-2.3)^2$, $x \in [0.5, 2.5]$.

Pentru funcţii de mai multe variabile, am văzut că algoritmii de ordin zero pot fi foarte costisitori din punct de vedere computaţional. Pentru a obţine algoritmi eficienţi pentru astfel de funcţii sunt necesare metode de ordin superior.

4.2 Algoritmi de ordinul întâi (metode de tip gradient)

Există mai multe modele generale de proiectare a algoritmilor cu oracol de ordinul întâi pentru optimizarea liberă de constrângeri. În această secţiune vom studia câteva dintre aceste metode. Avantajul considerării diferenţialei este că putem renunţa la unimodalitate, întrucât semnul derivatei ne indică dacă în vecinătatea punctului funcţia creşte sau scade, iar dezavantajele sunt că nu putem trata funcţii nediferenţiabile şi metodele ne indică, în general, că ne îndreptăm către puncte critice care, în absenţa convexităţii, ar putea să nu fie puncte de extrem.

4.2.1 Descrierea metodei direcţiei de descreştere (gradient descent, GD)

Pentru început, ne concentrăm pe modelul numit "căutarea direcţiei de descreştere".

Scopul acestui algoritm este acela de a realiza la fiecare pas o descreştere a funcţiei obiectiv $f : \mathbb{R}^p \to \mathbb{R}$, de clasă C^2 (i.e. $f(x_{k+1}) < f(x_k)$). Algoritmul calculează la fiecare pas k, o direcţie p_k (un vector de pe sfera unitate) şi un pas $\alpha_k > 0$ de mişcare pe direcţia p_k. Astfel, plecând de la iteraţia x_k, noua iteraţie va fi

$$x_{k+1} = x_k + \alpha_k p_k. \tag{4.1}$$

Primul punct al iteraţiei, adică x_0, se alege arbitrar (vom vedea ulterior că, şi de această dată, este indicat ca x_0 să satisfacă anumite cerinţe). În cadrul acestei scheme sunt importante atât alegerea direcţiei cât şi a pasului.

Alegerea direcţiei Din formula lui Taylor, pentru α, p fixate, există $t \in (0, \alpha)$ astfel încât

$$f(x_k + \alpha p) = f(x_k) + \alpha \nabla f(x_k)(p) + \frac{1}{2}\alpha^2 \nabla^2 f(x_k + tp)(p, p),$$

adică

$$f(x_k + \alpha p) - f(x_k) = \alpha \left[\nabla f(x_k)(p) + \frac{1}{2}\alpha \nabla^2 f(x_k + tp)(p, p) \right]. \tag{4.2}$$

Oprindu-ne cu evaluarea la diferenţiala de ordinul întâi, direcţia pe care f descreşte cel mai mult este dată de soluţia problemei de minimizare pe bila unitate închisă a funcţiei (în variabila p), $p \mapsto \nabla f(x_k)(p)$. Cum

$$\nabla f(x_k)(p) = \|p\| \, \|\nabla f(x_k)\| \cos\theta = \|\nabla f(x_k)\| \cos\theta$$

unde θ este unghiul dintre p şi $\nabla f(x_k)$, este clar că minimul este atins pentru

$$p = -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}$$

dacă $\|\nabla f(x_k)\| \neq 0$. Deci dacă s-a atins un punct critic, atunci nu putem înainta. În caz contrar, alegerea lui p_k de forma de mai sus se numeşte metoda celei mai bune (sau celei mai abrupte) descreşteri (steepest descent method, în limba engleză). Pe de altă parte, orice direcţie pentru care unghiul cu $\nabla f(x_k)$ este mai mare de $\frac{\pi}{2}$ (i.e. $\cos\theta < 0$) produce o descreştere a funcţiei f dacă α este suficient de mic (conform (4.2)), întrucât ultimul termen conţine α, deci tinde mai rapid

către zero. O astfel de direcţie (pentru care $\nabla f(x_k)(p_k) < 0$) se numeşte direcţie de descreştere (gradient descent).

Alegerea pasului Apare acum problema alegerii lui α_k. Există mai multe variante pentru această alegere. Le vom enumera pe cele mai importante dintre ele şi vom detalia consecinţele celor mai populare dintre aceste posibilităţi.

Regula minimizării exacte Alegerea ideală este minimul, pentru $\alpha > 0$, al funcţiei $\alpha \mapsto f(x_k + \alpha p_k)$, dar, din punct de vedere practic, problema aceasta nu este neapărat una simplu de rezolvat. Totuşi, aplicarea algoritmului secţiunii de aur conduce la o soluţie aproximativă a acestei probleme.

Regula factorului constant Alegem

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

unde $\beta > 0$ este dat (adică $\alpha_k = \beta \|\nabla f(x_k)\|$. Aceasta este o variaţie a metodei celei mai bune direcţii de descreştere, dar vom vedea că funcţionează pentru funcţii f având unele proprietăţi suplimentare. În plus, în general, ea trebuie combinată şi cu alte idei. Dacă $f(x_k + \beta \nabla f(x_k)) \geq f(x_k)$ atunci factorul β trebuie micşorat (procedură ce se numeşte backtracking).

Regula Armijo Fie $\alpha > 0$, $\beta \in (0,1)$ şi $c \in (0,1)$. Se testează inegalitatea

$$f(x_k) + c\left(\beta^i \alpha\right) \nabla f(x_k)(p_k) > f\left(x_k + \beta^i \alpha p_k\right), \ i = 0, 1, 2, ... \tag{4.3}$$

iterativ pornind cu $i = 0$ (sau, mai practic, cu valoarea lui i obţinută la precedenta iteraţie). Ideea este că dorim ca descreşterea funcţiei să fie suficient de mare. Primul test, pentru $i = 0$ (sau pentru i specificat mai sus) adică pentru pasul α (respectiv pasul α_k) este

$$f(x_k) + c\alpha \nabla f(x_k)(p_k) > f(x_k + \alpha p_k).$$

Dacă inegalitatea de mai sus este satisfăcută alegem $x_{k+1} = x_k + \alpha p_k$ (adică $\alpha_k = \alpha$). Altfel, pasul se micşorează la $\beta\alpha$ şi efectuăm testul (pentru $i = 1$)

$$f(x_k) + c\beta\alpha \nabla f(x_k)(p_k) > f(x_k + \beta\alpha p_k).$$

Continuând procedeul în acelaşi fel, obţinem o metodă foarte practică.

Observaţia 4.2.1 *Iteraţiile prevăzute de Regula Armijo pentru determinarea pasului se termină la un moment dat. Demonstrăm această aserţiune. Evident, atunci când i este foarte mare, $t := \beta^i \alpha$ este foarte apropiat de 0. Presupunem că*

$$f(x_k + tp_k) \geq f(x_k) + ct \nabla f(x_k)(p_k), \ \forall t.$$

Cum f este de clasă C^1, există o funcţie θ continuă în 0, cu $\lim_{t \to 0} \frac{\theta(t)}{t} = 0$ astfel încât

$$f(x_k + tp_k) = f(x_k) + t \nabla f(x_k)(p_k) + \theta(t).$$

Obţinem

$$(c-1)\,\nabla f(x_k)(p_k) + \frac{\theta(t)}{t} \leq 0.$$

Pentru $t \to 0$ (adică pentru $i \to \infty$), deducem

$$(c-1)\,\nabla f(x_k)(p_k) \leq 0,$$

ceea ce este fals.

Regula Wolfe Fie $0 < c_1 < c_2 < 1$. Alegem α_k ca fiind un număr $\alpha > 0$ satisfăcând următoarele două condiţii

$$\begin{cases} f(x_k + \alpha p_k) < f(x_k) + c_1 \alpha \nabla f(x_k)(p_k), \\ c_2 \nabla f(x_k)(p_k) \leq \nabla f(x_k + \alpha p_k)(p_k). \end{cases} \tag{4.4}$$

În general, prima condiţie (care este condiţia Armijo de mai sus) poate fi satisfăcută de valori α foarte mici, lucru împiedicat de a doua condiţie.

Consistenţa acestor condiţii este arătată riguros mai jos.

Propoziţia 4.2.2 *Fie $f : \mathbb{R}^p \to \mathbb{R}$ de clasă C^2, p_k o direcţie de descreştere în x_k şi $0 < c_1 < c_2 < 1$. Dacă f este mărginită inferior pe mulţimea $\{x_k + \lambda p_k \mid \lambda > 0\}$ atunci există $\alpha > 0$ care satisface condiţiile din regula Wolfe.*

Demonstraţie Conform ipotezei, funcţia $\alpha \mapsto f(x_k + \alpha p_k)$ este mărginită inferior pe $(0, \infty)$.

Pentru orice $\alpha > 0$ există $t \in (0, \alpha)$ astfel încât

$$\begin{aligned} f(x_k + \alpha p_k) &= f(x_k) + \alpha \nabla f(x_k)(p_k) + \frac{1}{2}\alpha^2 \nabla^2 f(x_k + t p_k)(p_k, p_k) \\ &= f(x_k) + c_1 \alpha \nabla f(x_k)(p_k) + \alpha \left[(1 - c_1)\,\nabla f(x_k)(p_k) + \frac{1}{2}\alpha \nabla^2 f(x_k + t p_k)(p_k, p_k) \right], \end{aligned}$$

Cum $c_1 \in (0,1)$ şi $\nabla f(x_k)(p_k) < 0$, pentru orice α suficient de mic avem

$$f(x_k + \alpha p_k) < f(x_k) + c_1 \alpha \nabla f(x_k)(p_k),$$

deci, ţinând cont de proprietatea de mărginire, ecuaţia (în α)

$$f(x_k + \alpha p_k) = f(x_k) + c_1 \alpha \nabla f(x_k)(p_k)$$

are cel puţin o soluţie strict pozitivă. Din continuitatea (în α) a funcţiilor implicate, există o cea mai mică soluţie pozitivă pe care o notăm cu α'. Evident, pentru orice $\alpha \in (0, \alpha')$ are loc prima condiţie. Aplicând din nou Formula lui Taylor, există $\alpha'' \in (0, \alpha')$ cu

$$f(x_k + \alpha' p_k) = f(x_k) + \alpha' \nabla f(x_k + \alpha'' p_k)(p_k),$$

deci

$$\nabla f(x_k + \alpha'' p_k)(p_k) = c_1 \nabla f(x_k)(p_k) > c_2 \nabla f(x_k)(p_k).$$

Aşadar, pentru α'' are loc şi a doua condiţie. Cum pentru α'' inegalităţile din ambele condiţii sunt stricte, există chiar un interval în jurul acestui punct pe care aceste condiţii sunt îndeplinite. □

Vom vedea ulterior unele criterii de oprire pentru algoritmi bazaţi pe Regula Wolfe.

4.2.2 Convergenţa metodei direcţiei de descreştere. Cazul general

Discutăm acum convergenţa metodei celei mai bune descreşteri pentru cele mai importante dintre regulile de mai sus: Regula Armijo şi Regula Wolfe.

O primă cerinţă pentru a avea rezultate de convergenţă semnificative este legată de unghiul dintre $\nabla f(x_k)$ şi direcţia (vectorul de normă 1) p_k. Dacă notăm cu θ_k acest unghi, atunci impunem să existe $\varepsilon > 0$ astfel încât pentru orice k

$$\cos\theta_k = \frac{\nabla f(x_k)(p_k)}{\|\nabla f(x_k)\|} \in [-1, -\varepsilon). \tag{4.5}$$

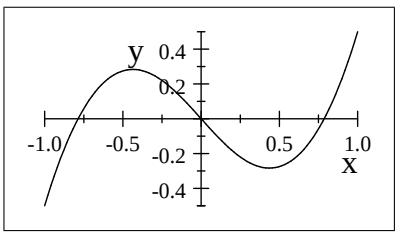
Astfel, chiar dacă p_k nu este cea mai bună direcţie de descreştere, cerem totuşi ca unghiul să nu se apropie oricât de mult de $\frac{\pi}{2}$. Evident, în cazul alegerii celei mai bune direcţii de descreştere $\cos\theta_k = -1$ pentru orice k, deci condiţia are loc.

Observaţia 4.2.3 *Să considerăm funcţia $f : \mathbb{R} \to \mathbb{R}$ dată prin*

$$f(x) = x^3 - \frac{x}{x^2+1}$$

al cărei grafic este redat mai jos.

Este clar că folosind chiar cea mai bună direcţie de descreştere pentru o dată iniţială $x_0 < -0.6$ (de exemplu) vom obţine o descreştere infinită, deci metoda nu converge la punctul de minim. De aceea, şi pentru această metodă, ideea de a aproxima (grafic sau prin altă tehnică) în primă instanţă punctul de minim pentru o alegere potrivită a datei iniţiale este importantă.



Teorema 4.2.4 *Fie (x_k) şir de vectori generaţi după formula de recurenţă*

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

unde direcţiile (p_k) satisfac (4.5) şi paşii (α_k) satisfac Regula Armijo (4.3) cu parametrii α, β, c. Dacă $\overline{x}$ este un punct limită al şirului (x_k), atunci $\overline{x}$ este punct critic al lui f (adică $\nabla f(\overline{x}) = 0$).

Demonstraţie. Fără a restrânge generalitatea, putem presupune că (x_k) este convergent la $\overline{x}$, iar (p_k) este convergent la o direcţie p. Dacă pasul α_k a fost validat la încercarea i în (4.3), atunci avem

$$f(x_k) - f(x_k + \alpha_k p_k) > -c\alpha_k \nabla f(x_k)(p_k) \geq \varepsilon c \alpha_k \|\nabla f(x_k)\|, \tag{4.6}$$

iar dacă încercarea $i-1$ este invalidă

$$f(x_k) - f\left(x_k + \frac{\alpha_k}{\beta} p_k\right) \leq -c\frac{\alpha_k}{\beta}\nabla f(x_k)(p_k). \tag{4.7}$$

Presupunem, prin reducere la absurd, că $\nabla f(\overline{x}) \neq 0$. Cum $(f(x_k))$ este descrescător la $f(\overline{x})$ deducem că

$$f(x_k) - f(x_k + \alpha_k p_k) = f(x_k) - f(x_{k+1}) \to 0,$$

deci

$$\alpha_k \nabla f(x_k)(p_k) \to 0.$$

Dar, pe baza (4.5), prin trecere la limită, $\nabla f(\overline{x})(p) < 0$, deci $\alpha_k \to 0$. Aceasta implică faptul că pentru k mare alegerea pasului trece prin procedura de ajustare (backtracking), adică se întâmplă (4.7). Folosind Teorema lui Lagrange (pentru funcţii de mai multe variabile), există $z_k \in \left(x_k, x_k + \frac{\alpha_k}{\beta} p_k\right)$ astfel încât

$$-\frac{\alpha_k}{\beta} \nabla f(z_k)(p_k) \leq -c \frac{\alpha_k}{\beta} \nabla f(x_k)(p_k).$$

Cum $z_k \to \overline{x}$, prin simplificare şi trecere la limită obţinem contradicţia $\nabla f(\overline{x})(p) \geq 0$. Deci $\nabla f(\overline{x}) = 0$. □

Corolarul 4.2.5 *Presupunem că direcţiile de căutare satisfac* (4.5). *Dacă* (α_k) *sunt aleşi după Regula minimizării exacte, atunci orice punct limită al lui* (x_k) *este punct critic pentru* f.

Demonstraţie. Presupunem că un subşir (x_{k_r}) converge la un punct $\overline{x}$ care nu e critic. Fie $\overline{x}_{k_r} = x_{k_r} + \bar{\alpha}_{k_r} p_{k_r}$ punctul ales după Regula Armijo. Atunci, pentru orice r,

$$f(x_{k_r}) - f(x_{k_r+1}) \geq f(x_{k_r}) - f(\overline{x}_{k_r+1}) \geq -c\bar{\alpha}_{k_r} \langle f(x_{k_r}), p_{k_r} \rangle.$$

Cum $f(x_k) \to f(\overline{x})$ avem $f(x_{k_r}) - f(x_{k_r+1}) \to 0$ şi ca în demonstraţia teoremei precedente, $\bar{\alpha}_{k_r} \to 0$.

Pe baza relaţiei (4.7),

$$f(x_{k_r}) - f\left(x_{k_r} + \frac{\bar{\alpha}_{k_r}}{\beta} p_{k_r}\right) \leq -c \frac{\bar{\alpha}_{k_r}}{\beta} \nabla f(x_{k_r})(p_{k_r}).$$

Urmând raţionamentul din rezultatul precedent, ajungem la aceeaşi contradicţie. □

Discutăm acum convergenţa algoritmului de căutare a direcţiei de descreştere pentru Regula Wolfe.

Teorema 4.2.6 *(Zoutendijk) Fie iteraţia (4.1) în care* (p_k) *sunt direcţii de descreştere, iar* (α_k) *sunt paşi ce satisfac condiţiile Wolfe. Presupunem că* f *este de clasă* C^1 *şi mărginită inferior, iar* ∇f *este Lipschitz. Atunci seria*

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

este convergentă.

În particular, dacă este satisfăcută condiţia (4.5), *atunci* $\nabla f(x_k) \to 0$ *şi orice punct limită* $\overline{x}$ *al lui* (x_k) *este punct critic pentru* f.

Demonstraţie Din condiţiile Wolfe, pentru fiecare $k \in \mathbb{N}^*$,

$$\nabla f(x_k + \alpha_k p_k)(p_k) - \nabla f(x_k)(p_k) \geq (c_2 - 1)\nabla f(x_k)(p_k),$$

adică

$$\nabla f(x_{k+1})(p_k) - \nabla f(x_k)(p_k) \geq (c_2 - 1)\nabla f(x_k)(p_k),$$

iar din condiţia Lipschitz asupra diferenţialei există o constantă L astfel încât,

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

de unde,

$$(\nabla f(x_{k+1}) - \nabla f(x_k))(p_k) \leq \alpha_k L \|p_k\|^2.$$

Deducem

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

Ţinând din nou seama de inegalitatea $\nabla f(x_k)(p_k) < 0$,

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

Dar

$$\frac{(\nabla f(x_k)(p_k))^2}{\|p_k\|^2} = \cos^2 \theta_k \|\nabla f(x_k)\|^2,$$

deci

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

Sumând relaţiile deducem

$$f(x_{k+1}) \leq f(x_0) + c_1 \frac{c_2 - 1}{L} \sum_{i=0}^{k} \cos^2 \theta_i \|\nabla f(x_i)\|^2.$$

Cum $c_2 - 1 < 0$, din mărginirea inferioară a funcţiei f, obţinem convergenţa seriei din enunţ.

Convergenţa demonstrată asigură faptul că

$$\cos^2 \theta_k \|\nabla f(x_k)\|^2 \to 0.$$

Dacă alegerea lui p_k se face astfel încât are loc condiţia (4.5), atunci $\nabla f(x_k) \to 0$. Acum este evident că dacă $\overline{x}$ este punct limită pentru (x_k), atunci $\overline{x}$ este punct critic. □

Observaţia 4.2.7 *Algoritmii de mai sus se aplică şi funcţiilor convexe, determinându-se puncte de minim global (pentru că punctele critice au această proprietate) dar, cum vom vedea, forma particulară a convexităţii permite elaborarea unor algoritmi specifici puternici.*

Implementare 4.15 Implementăm acum metoda celei mai bune descreşteri cu regula Armijo pentru o funcţie de o variabilă reală. Aşdar, direcţia se alege

$$p_k = -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}$$

dacă $\|\nabla f(x_k)\| \neq 0$ iar, în ceea ce priveşte pasul α_k, vom utiliza doar condiţia de tip Armijo

$$f(x_k + \alpha_k p_k) < f(x_k) + c\alpha_k \nabla f(x_k)(p_k),$$

unde $c \in (0,1)$. Astfel, vom testa la fiecare pas dacă se îndeplineşte condiţia, iar în caz contrar, îl vom înmulţi pe α_k cu un factor subunitar (β) pentru a scădea pasul. Ca de obicei, criteriul de oprire utilizează valoarea absolută a derivatei.

Pentru exemplificare alegem funcţia $f : \mathbb{R} \to \mathbb{R}$, $f(x) = 3e^x - x^3 + 5x$ care a fost discutată şi mai sus dintr-o altă perspectivă. Astfel, avem codul de mai jos

```
funct=@(x) 3*exp(x)-x^3+5*x;
der=@(x) 3*exp(x)-3*x^2+5;
tol=1e-3;u=5; u_vechi=1; n=0; beta=0.5;
c=0.01; alpha=1;
%steepest descent
while abs(der(u))>tol
 u_vechi=u;
while funct(u_vechi-alpha*der(u_vechi)/abs(der(u_vechi)))...
    >=funct(u_vechi)-c*alpha*abs(der(u_vechi))
 alpha=beta*alpha;
end
u=u-alpha*der(u)/abs(der(u));
n=n+1;
end
u
n
alpha
der(u)
u = -1.3845; n = 12; alpha = 0.00024414; ans = 0.00063161
```

Deci algoritmul se opreşte atunci când valoarea absolută a derivatei în punctul curent este suficient de mică. Evoluţia descreşterii pasului şi a gradientului poate fi de asemenea elocventă.

Exerciţiul 4.16 *Să se implementeze metoda celei mai bune descreşteri cu regula Armijo pentru care derivata se va calcula numeric (a se vedea (3.8)).*

Implementare 4.17 Scriem acum un cod pentru aplicarea metodei funcţiei Rosenbrock, funcţie-test de două variabile reale pentru care ştim că unicul punct de minim este $(1,1)$.

```
funct=@(x) 100.0*(x(2)-x(1)^2)^2 + (1-x(1))^2;
dif_fct=@(x) [-2*(1-x(1))-400*x(1)*(x(2)-x(1)^2),200*(x(2)-x(1)^2)];
n=0; beta=0.5;
```

```
c=0.01; alpha=1;
u_vechi=[0, 0];u=[-1.5, 2];v=u;
%steepest descent
while norm(dif_fct(u))>10^(-7)
 u_vechi=u;
% u=u-alpha*dif_fct(u)/norm(dif_fct(u));
v=[v;u];
 while funct(u_vechi-alpha*dif_fct(u_vechi)/norm(dif_fct(u_vechi)))...
    >=funct(u_vechi)-c*alpha*norm(dif_fct(u_vechi))
 alpha=beta*alpha;
end
u=u-alpha*dif_fct(u)/norm(dif_fct(u));v=[v;u];
n=n+1;
end
u
n
alpha
plot(v(:,1),v(:,2),'-');hold on;
plot(1,1,'r*');hold on;[x,y] = meshgrid (-2:0.2:2,-1:0.2:4);
z1=100*(y-x.^2).^2 + (1-x).^2;
[c,h] = contour(x,y,z1,100);
```

În cazul acesei metode se poate observa că sunt necesare foarte multe iteraţii, dar punctul de start poate fi îndepărtat de soluţie.

Exerciţiul 4.18 *Să se implementeze algoritmul celei mai bune direcţii de descreştere pentru funcţia* $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x - 2x^2$.

Exerciţiul 4.19 *Fie* $f : \mathbb{R} \to \mathbb{R}$, $f(x) = 2x^2 + 3e^{-2x}$. *Să se arate că* f *este strict convexă şi admite punct de minim. Să se determine acest punct cu aproximaţie folosind metoda lui Newton (aplicată lui* $f'(x) = 0$*) şi metoda celei mai bune descreşteri.*

Exerciţiul 4.20 *Să se arate că funcţia* $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^4 - x^2 + x - 1$ *admite un singur punct de minim. Să se aproximeze acest punct folosind metoda celei mai bune descreşteri.*

Exerciţiul 4.21 *Să se testeze metoda căutării direcţiei de descreştere cu data iniţială* $(-1, -1)$ *pentru funcţia* $f : \mathbb{R}^2 \to \mathbb{R}$,

$$f(x, y) = x^4 - 16x^2 + 5x + y^4 - 16y^2 + 5y.$$

Exerciţiul 4.22 *Să se implementeze metoda pentru* $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = 4x^2 - 3xy + y^2$. *Să se observe faptul că minimul (care este originea) găsindu-se într-o zonă mai accesibilă a graficului, numărul de iteraţii este semnificativ mai mic.*

Exerciţiul 4.23 *Să se determine teoretic punctul de minim pentru* $f : \mathbb{R}^2 \to \mathbb{R}$ *dată prin* $f(x, y) = x^2(1 + y)^3 + y^2$ *şi să se aplice metoda de mai sus.*

Exerciţiul 4.24 *Să se aplice metoda de mai sus pentru* $f : \mathbb{R}^2 \to \mathbb{R}$,

$$f(x, y) = x^6 + y^4 + 4x^2y^2 - 3xy + 2x + y.$$

Încheiem această secţiune folosind şi Regula minimizării exacte pe care o implementăm cu ajutorul algoritmului raportului de aur. De asemenea, folosim un cod care calculează aproximativ gradientul funcţiei de minimizat.

Implementare 4.25 fişier funcţie

```
function fx = func_multivar(x)
fx= x(1)^6+x(2)^4+4*x(1)^2*x(2)^2-3*x(1)*x(2)+2*x(1)+x(2);
```

fişier funcţie

```
function deriv = grad_vec(x,delx)
xvec = x;
xvec1 = x;
for i = 1:length(x)
 xvec = x;
 xvec1 = x;
 xvec(i) = x(i) + delx;
 xvec1(i) = x(i) - delx;
deriv(i) = (func_multivar(xvec) - func_multivar(xvec1) )/(2*delx);
end
```

fişier funcţie

function [alpha1,falpha1] = golden_funct(x,search)

```
a = -5; b = 5; rho = 0.618034; epsilon = 1e-5;
alpha1 = rho*a+(1-rho)*b;
alpha2 =(1-rho)*a+rho*b;
falpha1 = func_multivar(x+alpha1*search);
falpha2 = func_multivar(x+alpha2*search);
while abs(alpha1-alpha2)>epsilon
 if falpha1 > falpha2
 a = alpha1;
 alpha1 = alpha2;
 falpha1 = falpha2;
 alpha2 =(1-rho)*a+rho*b;
 falpha2 = func_multivar(x+alpha2*search);
     else
 b = alpha2;
 alpha2 = alpha1;
 falpha2 = falpha1;
 alpha1 = rho*a+(1-rho)*b;
 falpha1 = func_multivar(x+alpha1*search);
     end
end
alpha1=(alpha1+alpha2)/2;
```

```
falpha1=func_multivar(x+alpha1*search);
```
fişier de rulare
```
u_vechi=[0, 0];u=[-1, 2];v=u;
n=0;
delx = 10.^(-3);
pk=-grad_vec(u,delx)/norm(grad_vec(u,delx));
[alpha,val_alpha]=golden_funct(u,pk);
%exact minimization rule
while norm(grad_vec(u,delx))>10^(-4)
u_vechi=u;
pk=-grad_vec(u,delx)/norm(grad_vec(u,delx));
[alpha,val_alpha]=golden_funct(u,pk);
u=u+alpha*pk;
v=[v;u];
n=n+1;
end
u
n
alpha
norm(grad_vec(u,delx))
```
Se poate observa că această metodă funcţionează foarte greoi pe funcţia Rosenbrock din cauza specificităţii acestei funcţii şi e de recomandat înlocuirea valorii 10^{-4} cu 10^{-2} în testul de stop.

Exerciţiul 4.26 *Fie $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^7 + x^3 + 1$. Să se arate că ecuaţia $f(x) = 0$ admite o singură rădăcină reală, $\overline{x}$. Să se justifice faptul că $\overline{x}$ este punct de minim global pentru $F : \mathbb{R} \to \mathbb{R}$, $F(x) = \int_0^x f(s)ds$. Să se aproximeze $\overline{x}$ folosind metoda celei mai bune descreşteri. Să se compare cu metoda lui Newton aplicată direct ecuaţiei. Similar pentru funcţia $f : (0, \infty) \to \mathbb{R}$, $f(x) = \ln x + 6x$.*

Exerciţiul 4.27 *Fie $f : \mathbb{R}^2 \to \mathbb{R}$,*

$$f(x, y) = 0.6y^4 + 5x^2 - 7y^2 + \sin(xy) - 5y.$$

(i) Să se reprezinte grafic funcţia f pe pătratul $[-5, 5]^2$ şi să se deducă existenţa a două puncte de minim local.

(ii) Să se aproximeze punctele de minim local folosind metodele de mai sus.

Exerciţiul 4.28 *Fie $f : \mathbb{R}^2 \to \mathbb{R}$,*

$$f(x, y) = 0.7x^4 + 6y^2 - 8x^2 + \cos(xy) - 8x.$$

(i) Să se reprezinte grafic funcţia f pe pătratul $[-5, 5]^2$.

(ii) Să se aproximeze punctele de minim local folosind metodele de mai sus.

Exerciţiul 4.29 *Fie $f : \mathbb{R}^2 \to \mathbb{R}$ dată prin*

$$f(x) = x_1^2 + 2x_2^2 + 4x_1 + 4x_2.$$

Considerăm metoda celei mai bune descreşteri cu regula minimizării exacte şi data iniţială $(0,0)$. *Să se arate că*

$$x_k = \left(\frac{2}{3^k} - 2, \left(-\frac{1}{3} \right)^k - 1 \right), \ \forall k$$

şi să se deducă punctul de minim.

Exerciţiul 4.30 *Fie* $f : \mathbb{R}^2 \to \mathbb{R}$ *dată prin*

$$f(x) = 5x_1^2 - 9x_1x_2 + 4.075x_2^2 + x_1.$$

(i) Să se rezolve problema minimizării lui f *folosind metoda direcţiei de descreştere (cu diverse reguli de determinare a pasului).*

(ii) Să se explice de ce metoda menţionată necesită un număr relativ mare de iteraţii pentru aproximarea minimului.

Exerciţiul 4.31 *Să se rezolve problema minimizării funcţiei* $f : \mathbb{R}^3 \to \mathbb{R}$ *dată prin*

$$f(x) = (x_1 + 5)^2 + (x_2 + 8)^2 + (x_3 + 7)^2 + 2x_1^2x_2^2 + 4x_1^2x_3^2$$

prin metodele studiate. Să se testeze metodele pentru datele iniţiale $(1,1,1)$, $(-2.3,0,0)$, $(0,2,-12)$.

Exerciţiul 4.32 *Fie* $f : \mathbb{R}^2 \to \mathbb{R}$ *dată prin*

$$f(x) = \left(x_1^2 + x_2^2 - 1\right)^2 + (x_1 + x_2 - 1)^2.$$

Să se reprezinte grafic, să se găsească punctele de minim global şi să se arate că $\left(\sqrt[3]{4^{-1}}, \sqrt[3]{4^{-1}}\right)$ *este un punct critic care nu este extrem local. Să se utilizeze metoda căutării direcţiei de descreştere (cu diverse reguli de determinare a pasului) pentru datele iniţiale* $(4,4)$, $(4,-4)$, $(-4,4)$, $(-4,-4)$ *şi să se justifice rezultatele obţinute.*

Exerciţiul 4.33 *Fie* $f : \mathbb{R}^2 \to \mathbb{R}$ *dată prin*

$$f(x) = x_1^3 e^{x_2 - x_1^2 - 10(x_1 - x_2)^2}.$$

Să se reprezinte grafic pe $[-1.5,-0.5]^2$ *şi să se utilizeze metoda căutării direcţiei de descreştere (cu diverse reguli de determinare a pasului) pentru datele iniţiale* $(-3,-3)$, $(-3,3)$, $(3,-3)$, $(3,3)$.

Exerciţiul 4.34 *Fie* $f : \mathbb{R}^2 \to \mathbb{R}$ *dată prin*

$$f(x) = -x_2^2 e^{1 - x_1^2 - 20(x_1 - x_2)^2}.$$

Să se reprezinte grafic şi să se utilizeze metoda căutării direcţiei de descreştere (cu diverse reguli de determinare a pasului) pentru datele iniţiale $(0.1,0.1)$, $(0.8,0.1)$, $(1.1,0.1)$.

4.2.3 Convergenţa metodei direcţiei de descreştere. Cazul convex

Atunci când adăugăm proprietăţi suplimnetare funcţiilor pe care le studiem, putem demonstra că algoritmii sunt mai eficienţi decât în cazul general. Exemplificăm cu algoritmul direcţiei de decreştere pentru funcţii convexe.

Astfel, în această secţiune, în cadrul rezultatelor principale considerăm $f : \mathbb{R}^p \to \mathbb{R}$ o funcţie convexă, diferenţiabilă cu diferenţiala Lipschitz. Presupunem deci, pe lângă convexitate, că există $L > 0$ astfel încât

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \ \forall x, y \in \mathbb{R}^p.$$

Această din urmă ipoteză a fost întâlnită şi la Teorema lui Zoutendijk (Teorema 4.2.6). Începem cu un rezultat tehnic care pune în evidenţă rolul proprietăţii Lipschitz a diferenţialei în obţinerea unei majorări a funcţiei date (nu neaparat convexă în acest caz) printr-o funcţie pătratică.

Lema 4.2.8 *Dacă $f : \mathbb{R}^p \to \mathbb{R}$ este diferenţiabilă, cu diferenţiala $L-$Lipschitz, atunci*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2, \ \forall x, y \in \mathbb{R}^p. \tag{4.8}$$

Demonstraţie. Fie $x, y \in \mathbb{R}^p$ şi $\varphi_{y,x} : \mathbb{R} \to \mathbb{R}$,

$$\varphi_{y,x}(\lambda) = f(\lambda y + (1 - \lambda) x).$$

Avem

$$\varphi_{y,x}(1) - \varphi_{y,x}(0) = \int_0^1 \varphi'_{y,x}(\lambda) \, d\lambda = \varphi'_{y,x}(0) + \int_0^1 \left(\varphi'_{y,x}(\lambda) - \varphi'_{y,x}(0) \right) d\lambda.$$

Înlocuind şi folosind ipoteza, obţinem

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= \int_0^1 \langle \nabla f(x + \lambda(y - x)) - \nabla f(x), y - x \rangle \, d\lambda \\ &\leq \|y - x\| \int_0^1 \|\nabla f(x + \lambda(y - x)) - \nabla f(x)\| \, d\lambda \\ &\leq \|y - x\|^2 \int_0^1 \lambda d\lambda = \frac{L}{2} \|y - x\|^2. \end{aligned}$$

Aceasta este concluzia. □

Exerciţiul 4.35 *Să se arate că condiţia (4.8) este echivalentă cu convexitatea funcţiei $\mathbb{R}^p \ni x \mapsto \frac{L}{2} \|x\|^2 - f(x)$.*

Exerciţiul 4.36 *Folosind funcţia $\mathbb{R} \ni x \mapsto -e^x$ să se arate că condiţia (4.8) nu implică faptul că diferenţiala lui f este Lipschitz.*

Exerciţiul 4.37 *Considerăm $f : \mathbb{R} \to \mathbb{R}$ dată prin*

$$f(x) = \begin{cases} \frac{3}{4}x^2, \ \text{dacă} \ |x| < \frac{2}{3} \\ -\frac{3}{2}x^2 + 3x - 1, \ \text{dacă} \ x \in \left[\frac{2}{3}, 1\right] \\ -\frac{3}{2}x^2 - 3x - 1, \ \text{dacă} \ x \in \left[-1, -\frac{2}{3}\right] \\ |x| - \frac{x^2}{1}, \ \text{dacă} \ |x| > 1. \end{cases}$$

Să se arate că f este diferenţiabilă (derivabilă), iar cea mai mică constantă Lipschitz pentru derivată este 3. Să se arate că f satisface condiţia (4.8) pentru $L = \frac{3}{2}$ (folosind, eventual, un grafic).

Observaţia 4.2.9 *Aşadar, pentru o funcţie $f : \mathbb{R}^p \to \mathbb{R}$ o funcţie convexă, diferenţiabilă cu diferenţiala $L-$Lipschitz, pentru orice $x, y \in \mathbb{R}^p$*

$$f(y) + \langle \nabla f(y), x - y \rangle \leq f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|y - x\|^2 . \tag{4.9}$$

Privită în raport cu x, adică pentru y fixat, estimarea inferioară (din partea stângă) este o funcţie afină, în timp ce estimarea superioară (din partea dreaptă) este o funcţie pătratică. Această funcţie se scrie ca

$$\begin{aligned} h(x) &= f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|y - x\|^2 \\ &= \frac{L}{2} \left\| x - \left(y - \frac{1}{L} \nabla f(y) \right) \right\|^2 - \frac{1}{2L} \|\nabla f(y)\|^2 + f(y) \end{aligned}$$

şi are minimul global în

$$\overline{x} = y - \frac{1}{L} \nabla f(y) .$$

Aceasta sugerează următoare abordare algoritmică: dată fiind iteraţia x_k, la pasul următor, minimizăm estimarea superioară pentru care luăm $y = x_k$. Aceasta conduce deci la

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

adică la metoda celei mai bune descreşteri cu pas

$$\alpha_k = \frac{\|\nabla f(x_k)\|}{L}.$$

Această observaţie este ilustrată prin exerciţiul de mai jos, iar consistenţa acestei metode este demonstrată riguros după aceea.

Exerciţiul 4.38 *Fie $f : \mathbb{R} \to \mathbb{R}$ dată prin*

$$f(x) = \frac{1}{2} (x - 1)^2 - \frac{1}{2} \ln \left(x^2 + 1 \right) + 3.$$

Avem

$$\begin{aligned} f'(x) &= x - 1 - \frac{x}{x^2 + 1}, \\ f''(x) &= \frac{x^2 (x^2 + 3)}{(x^2 + 1)^2}, \ \forall x \in \mathbb{R}. \end{aligned}$$

Deci f este strict convexă pe $\mathbb{R}$. Se poate cu uşurinţă constata că $|f''(x)| \leq 1.25$ pentru orice x, deci f' este Lipschitz de constantă $L = 1.25$.

(i) Pentru $y = 3$, să se reprezinte pentru $x \in [0, 6]$ în acelaşi grafic cele trei funcţii din estimarea (4.9) .

(ii) Rezolvând ecuaţia $f'(x) = 0$ să se aproximeze punctul de minim.

(iii) Folosind iteraţia

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$

cu data iniţială $x_0 = 3$ să se determine x_7 şi să se compare cu punctul de minim determinat anterior.

Cum spuneam, demonstrăm acum faptul că iteraţia intuită mai sus şi apoi ilustrată prin exemplul anterior conduce la un punct de minim. Începem cu un rezultat ajutător care pune în evidenţă caracterizări ale proprietăţii Lipschitz pentru diferenţială în cazul funcţiilor convexe.

Lema 4.2.10 *Dacă $f : \mathbb{R}^p \to \mathbb{R}$ este convexă şi diferenţiabilă, atunci următoarele afirmaţii sunt echivalente:*

(i) diferenţiala lui f este $L-$Lipschitz, adică

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \ \forall x, y \in \mathbb{R}^p;$$

(ii)

$$\frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{L}{2}\|y - x\|^2, \ \forall x, y \in \mathbb{R}^p;$$

(iii)

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

Demonstraţie. $(i) \implies (ii)$ Ca în lema de mai sus se obţine a doua inegalitate de la (ii) .

Pentru a proba prima inegalitate de la (ii) , fixăm x şi definim funcţia convexă diferenţiabilă $g : \mathbb{R}^p \to \mathbb{R}$ prin

$$g(y) = f(y) - \langle \nabla f(x), y\rangle .$$

Evident, pentru orice $y \in \mathbb{R}^p$

$$\nabla g(y) = \nabla f(y) - \nabla f(x)$$

şi cum $\nabla g(x) = 0$ şi g este convexă, deducem că x este punct de minim global pentru g. În plus, este simplu de observat că diferenţiala lui g este $L-$Lipschitz. Astfel, folosind şi lema anterioară, avem, pentru orice $y \in \mathbb{R}^p$,

$$\begin{aligned} g(x) &\leq g\left(y - \frac{1}{L}\nabla g(y)\right) \leq g(y) + \left\langle \nabla g(y), -\frac{1}{L}\nabla g(y)\right\rangle + \frac{L}{2}\left\|\frac{1}{L}\nabla g(y)\right\|^2 \\ &= g(y) - \frac{1}{2L}\|\nabla g(y)\|^2 . \end{aligned}$$

Înlocuim şi avem

$$f(x) - \langle \nabla f(x), x\rangle \leq f(y) - \langle \nabla f(x), y\rangle - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 ,$$

şi (ii) este complet demonstrată.

$(ii) \Longrightarrow (iii)$ Scriem relaţiile de la (ii) pentru cuplurile (x, y) şi (y, x) şi le adunăm.

$(iii) \Longrightarrow (i)$ Evident. □

Teorema 4.2.11 *Fie $f : \mathbb{R}^p \to \mathbb{R}$ o funcţie convexă, diferenţiabilă cu diferenţiala $L-$Lipschitz. Presupunem că f admite punct de minim şi considerăm următoarea metodă a căutării direcţiei de descreştere*

$$x_0 \in \mathbb{R}^p,$$
$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k), \ \forall k \geq 0.$$

Atunci, dacă $\nabla f(x_k) \neq 0$, pentru orice k, (x_k) converge la un punct de minim $\overline{x}$ al lui f, $(\|x_k - \overline{x}\|)$ este strict descrescător la 0 şi au loc următoarele estimări

$$f(x_k) - f(\overline{x}) \leq \frac{5L}{2\sum_{i=0}^{k-1} \|x_i - \overline{x}\|^{-2}}, \ \forall k$$
$$\lim_k k\left(f(x_k) - f(\overline{x})\right) = 0.$$

Demonstraţie. Conform Lemei 4.2.8,

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

În particular, $f(x_{k+1}) < f(x_k)$ pentru orice k întrucât am presupus că $\nabla f(x_k) \neq 0$, pentru orice k.

Fie x^* un punct de minim (global, pentru că f este convexă) al lui f.

Fără a restrânge generalitatea, presupunem că $x_k \neq x^*$, pentru orice k (în caz contrar, şirul este staţionar şi de la un loc încolo $\nabla f(x_k) = 0$, ceea ce am presupus că nu se întâmplă).

Pentru a scurta scrierea, notăm, pentru orice k

$$w_k = f(x_k) - \inf_{x \in \mathbb{R}^p} f(x) = f(x_k) - f(x^*).$$

Din cele de mai sus şi din Lema 4.2.8 avem

$$w_{k+1} - w_k \leq -\frac{L}{2}\|x_{k+1} - x_k\|^2, \ \forall k,$$

şi

$$w_{k+1} < w_k \leq \langle \nabla f(x^*), x_k - x^* \rangle + \frac{L}{2}\|x_k - x^*\|^2 = \frac{L}{2}\|x_k - x^*\|^2, \ \forall k.$$

Avem, tot pe baza convexităţii,

$$f(x^*) \geq f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle, \ \forall k,$$

deci

$$w_k \leq -\langle \nabla f(x_k), x^* - x_k \rangle \leq \|\nabla f(x_k)\| \|x_k - x^*\| = L \|x_{k+1} - x_k\| \|x_k - x^*\|, \ \forall k.$$

Deducem

$$\begin{aligned} f(x_k) - f(x_{k+1}) = w_k - w_{k+1} &\geq \frac{L}{2} \|x_{k+1} - x_k\|^2 \\ &\geq \frac{L}{2} \left(\frac{w_k}{L \|x_k - x^*\|} \right)^2 = \frac{w_k^2}{2L \|x_k - x^*\|^2} \geq \frac{w_{k+1}^2}{2L \|x_k - x^*\|^2}, \ \forall k. \end{aligned}$$

Rescriem inegalitatea de mai sus astfel:

$$\frac{1}{w_k} \leq \frac{1}{w_{k+1}} \left(1 + \frac{w_{k+1}}{2L \|x_k - x^*\|^2} \right)^{-1}, \ \forall k.$$

Cum

$$w_{k+1} \leq \frac{L}{2} \|x_k - x^*\|^2,$$

obţinem că

$$\frac{w_{k+1}}{2L \|x_k - x^*\|^2} \leq \frac{1}{4}.$$

Pe de altă parte, este simplu de observat că

$$\forall t \in \left[0, \frac{1}{4}\right], \ \frac{1}{1+t} \leq 1 - \frac{4t}{5},$$

deci

$$\frac{1}{w_k} \leq \frac{1}{w_{k+1}} - \frac{2}{5L \|x_k - x^*\|^2}, \ \forall k.$$

Sumăm aceste relaţii de la 0 la $k-1$ şi obţinem

$$0 \leq \frac{1}{w_0} \leq \frac{1}{w_k} - \frac{2}{5L} \sum_{i=0}^{k-1} \|x_i - x^*\|^{-2},$$

ceea ce ne conduce la relaţia

$$f(x_k) - \inf_{x \in \mathbb{R}^p} f(x) \leq \frac{5L}{2 \sum_{i=0}^{k-1} \|x_i - x^*\|^{-2}}, \ \forall k.$$

Arătăm în continuare că $(\|x_k - x^*\|)$ este strict descrescător. Avem, folosind faptul că $\nabla f(x^*) = 0$ şi Lema 4.2.10,

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

adică ceea ce dorim.

În particular, (x_k) este mărginit, deci are un subşir (x_{k_r}) convergent la un punct pe care îl notăm $\overline{x}$. Dar, am văzut mai sus,

$$f\left(x_{k_{r+1}}\right) - f\left(x_{k_r}\right) \leq f\left(x_{k_r+1}\right) - f\left(x_{k_r}\right) \leq -\frac{1}{2L}\left\|\nabla f\left(x_{k_r}\right)\right\|^2, \ \forall r$$

şi trecând la limită deducem

$$0 \leq -\frac{1}{2L}\left\|\nabla f\left(\overline{x}\right)\right\|^2,$$

adică $\nabla f\left(\overline{x}\right) = 0$ şi $\overline{x}$ este punct de minim global. Cum x^* de mai sus este un punct de minim global ales la întâmplare, deducem că toate calculele sunt valabile şi pentru $\overline{x}$ în loc de x^*. În particular, $(\|x_k - \overline{x}\|)$ este strict descrescător şi are loc inegalitatea din concluzie. Aceasta înseamnă că şirul (x_k) este convergent la $\overline{x}$.

Demonstrăm acum ultima concluzie. Definim $(\alpha_k) = \left(\|x_k - \overline{x}\|^{-2}\right)$, şir care este crescător cu limita $+\infty$. Relaţia

$$f\left(x_k\right) - f\left(\overline{x}\right) \leq \frac{5L}{2\sum_{i=0}^{k-1}\|x_i - \overline{x}\|^{-2}}, \ \forall k$$

se rescrie în forma

$$k\left(f\left(x_k\right) - f\left(\overline{x}\right)\right) \leq \frac{5L}{2}\left(\frac{1}{k}\sum_{i=0}^{k-1}\alpha_i\right)^{-1}, \ \forall k.$$

Pentru a obţine concluzia, este suficient să arătăm că

$$\frac{1}{k}\sum_{i=0}^{k-1}\alpha_i \to +\infty.$$

Fie $M > 0$ şi $n \in \mathbb{N}$ astfel încât $\alpha_k \geq M$ pentru orice $k \geq n$. Atunci

$$\frac{1}{2n}\sum_{i=0}^{2n-1}\alpha_i \geq \frac{1}{2n}\sum_{i=n}^{2n-1}\alpha_i \geq \frac{M}{2}.$$

În plus, este uşor de verificat, pe baza monotoniei lui $(\|x_k - \overline{x}\|)$ că şirul $\left(\frac{1}{k}\sum_{i=0}^{k-1}\alpha_i\right)$ este crescător. Aceste observaţii încheie demonstraţia. □

Ilustrăm această teorie.

Implementare 4.39 Chiar dacă nu determinăm exact constanta L, putem încerca mai multe variante (dacă o constantă este bună, atunci orice constantă mai mare este de asemenea bună).

```
L=50;
f_m=@(x) [20*x(1)^4+5*x(1)^2+ x(2)^4+x(2)^2+x(2)+3*x(1)*x(2)];
f_m_g=@(x) [80*x(1)^3+10*x(1)+3*x(2),4*x(2)^3+2*x(2)+1+3*x(1)];
u=[-0.9 0.9];v=u;
while norm(f_m_g(u))>10^(-3) % i=1:maxiter
u=u-(1/L)*f_m_g(u);v=[v;u];
```

```
end
u
f_m_g(u)
plot(v(:,1),v(:,2),'-');hold on; hold on;
[x,y] = meshgrid (-1:0.2:1,-1:0.2:1);
z1=20.*x.^4+5*x.^2+y.^4+y.^2+y+3.*x.*y;
[c,h] = contour(x,y,z1,100);
```

Exerciţiul 4.40 *Să se implementeze metoda de mai sus pentru* $f : \mathbb{R}^2 \to \mathbb{R}$ *dată prin*

$$f(x, y) = 20x^2 + y^4 + y.$$

4.2.4 Convergenţa metodei direcţiei de descreştere. Cazul funcţiilor pătratice

Să analizăm cazul funcţiei $f : \mathbb{R}^p \to \mathbb{R}$ definită prin $f(x) = \frac{1}{2}\langle Qx, x\rangle + \langle b, x\rangle$ unde Q este o matrice pătratică de dimensiune p, simetrică, pozitiv definită, iar $b \in \mathbb{R}^p$.

Evident, funcţia este strict convexă şi are mulţimile de nivel mărginite, deci există un unic punct de minim dat de ecuaţia $\nabla f(x) = 0$, adică $Qx + b = 0$. Deci minimul este $\overline{x} = -Q^{-1}b$.

Pentru probleme de mari dimensiuni calculul lui Q^{-1} poate fi costisitor din punct de vedere computaţional.

Să studiem acum comportarea algoritmului celei mai bune direcţii de descreştere cu pasul dat de Regula minimizării exacte.

Astfel, $x_0 \in \mathbb{R}^p$ şi pentru orice $k \geq 0$,

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

unde

$$p_k = -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|} = -\frac{1}{\|Qx_k + b\|}\left(Qx_k + b\right),$$

iar α_k este minimul funcţiei scalare $\alpha \mapsto f(x_k + \alpha p_k)$.

Presupunem că gradientul nu se anulează în punctele iteraţiei (adică $Qx_k + b \neq 0$), ceea ce, ţinând cont de convexitatea problemei, este echivalent cu faptul că $f(x_k) > \overline{f}$, unde $\overline{f}$ este valoarea minimă a funcţiei.

Astfel,

$$f(x_k + \alpha p_k) = f(x_k) + \frac{1}{2}\alpha^2 \langle Qp_k, p_k\rangle + \alpha \langle Qx_k + b, p_k\rangle .$$

iar această funcţie are punctul de minim în

$$\alpha_k = -\frac{\langle Qx_k + b, p_k\rangle}{\langle Qp_k, p_k\rangle} = \frac{\|Qx_k + b\|^3}{\langle Q\left(Qx_k + b\right), Qx_k + b\rangle}.$$

Deci iteraţia este

$$\begin{aligned} x_{k+1} &= x_k + \frac{\|Qx_k + b\|^3}{\langle Q\left(Qx_k + b\right), Qx_k + b\rangle} p_k \\ &= x_k + \frac{\|Qx_k + b\|^2}{\langle Q\left(Qx_k + b\right), Qx_k + b\rangle} \left(-Qx_k - b\right) \\ &= x_k - \frac{\|Qx_k + b\|^2}{\langle Q\left(Qx_k + b\right), Qx_k + b\rangle} \left(Qx_k + b\right) \end{aligned}$$

Observaţia 4.2.12 *Dacă* $f : \mathbb{R}^2 \to \mathbb{R}$ *are forma*

$$f(x, y) = \alpha x^2 + \beta xy + \gamma y^2 + \delta x + \varepsilon y,$$

atunci pentru a fi scrisă în forma standard de mai sus, avem

$$Q = \begin{pmatrix} 2\alpha & \beta \\ \beta & 2\gamma \end{pmatrix} \text{ şi } b = (\delta, \varepsilon).$$

Exerciţiul 4.41 *Să se implementeze algoritmul descris mai sus.*

Este clar că rezultatele generale de convergenţă pentru metoda celei mai bune descreşteri se aplică şi acestei iteraţiei de mai sus. Particularităţile problemei conduc la un rezultat de convergenţă specific acestui caz. Pentru o discuţie a acestor aspecte este nevoie să utilizăm mai multe rezultate de algebră liniară pe care le trecem în revistă aici. Mai întâi, reamintim că dacă A este o matrice reală pătratică (de dimensiune p), atunci un vector nenul $x \in \mathbb{R}^p$ se numeşte vector propriu pentru A dacă există $\lambda \in \mathbb{C}$ astfel încât $Ax = \lambda x$. Scalarul λ se numeşte valoare proprie a matricii A. În general, A are n valori proprii (dintre care unele pot coincide). Este bine-cunoscut faptul că în cazul unei matrici pătratice toate valorile proprii sunt numere reale. Dacă, în plus, matricea este pozitiv definită, acestea sunt toate strict pozitive. Are loc următoarea teoremă de descompunere ortogonală.

Teorema 4.2.13 *Fie A o matrice pătratică de dimensiune p. Atunci există o matrice U de dimensiune p cu proprietatea $U^tU = UU^t = I$ şi o matrice diagonală $D = \operatorname{diag}(d_1, ..., d_p)$ astfel încât $U^tAU = D$.*

Mai mult, coloanele matricii U reprezintă o bază ortonormală a spaţiului $\mathbb{R}^p$ formată din vectori proprii şi elementele lui D sunt valorile proprii corespunzătoare.

O consecinţă a Teoremei 4.2.13 este că urma matricii A este suma valorilor sale proprii, iar determinamtul lui A este produsul valorilor sale proprii.

Pentru început este nevoie de o inegalitate (numită Inegalitatea lui Kantorovici) privind valorile proprii ale unei matrici de tipul celei de mai sus.

Lema 4.2.14 *Fie Q o matrice pătratică, simetrică, pozitiv definită de dimensiune p. Atunci pentru orice $x \in \mathbb{R}^p$ are loc*

$$\|x\|^4 \leq \langle Qx, x\rangle \cdot \left\langle Q^{-1}x, x\right\rangle \leq \frac{1}{4}\left(\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} + \sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}}\right)^2 \|x\|^4,$$

unde $\lambda_{\max}$ şi $\lambda_{\min}$ reprezintă cea mai mare şi respectiv cea mai mică valoare proprie a lui Q.

Demonstraţie Este suficient să demonstrăm inegalitatea pentru orice x de normă 1. Aşa cum am amintit în introducere, valorile proprii ale lui Q sunt reale strict pozitive şi, fără a restrânge generalitatea, le ordonăm descrescător: $\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p = \lambda_{\min}$. Notăm matricea diagonală având valorile proprii pe diagonala principală, în ordinea precizată, cu $D := \text{diag}(\lambda_1, \lambda_2, ..., \lambda_p)$. În plus, conform Teoremei 4.2.13, există o matrice ortogonală B cu proprietatea că $Q = B^t DB$. Atunci $Q^{-1} = (B^t DB)^{-1} = B^t D^{-1} B$, iar $D^{-1} = \text{diag}(\lambda_1^{-1}, \lambda_2^{-1}, ..., \lambda_p^{-1})$. Deci

$$\langle Qx, x\rangle = \left\langle B^t DBx, x\right\rangle = \langle DBx, Bx\rangle$$

şi

$$\left\langle Q^{-1}x, x\right\rangle = \left\langle B^t D^{-1}Bx, x\right\rangle = \left\langle D^{-1}Bx, Bx\right\rangle.$$

Pe de altă parte, aplicaţia $x \mapsto (Bx)$ este o bijecţie de la sfera unitate a lui $\mathbb{R}^p$ în ea însăşi deci, pentru a obţine relaţia din enunţ, este suficient să demonstrăm că pentru orice $u \in \mathbb{R}^p$ de normă 1 are loc

$$1 \leq \langle Du, u\rangle \cdot \left\langle D^{-1}u, u\right\rangle \leq \frac{1}{4}\left(\sqrt{\frac{\lambda_1}{\lambda_p}} + \sqrt{\frac{\lambda_p}{\lambda_1}}\right)^2 \|u\|^4.$$

Dacă $\lambda_1 = \lambda_p$ atunci are loc egalitatea. Presupunem că $\lambda_p < \lambda_1$. Avem

$$\langle Du, u\rangle = \sum_{i=1}^p u_i^2 \lambda_i; \;\; \langle Du, u\rangle = \sum_{i=1}^p u_i^2 \frac{1}{\lambda_i}.$$

Atunci prima inegalitate revine la

$$1 \leq \left(\sum_{i=1}^p u_i^2 \frac{1}{\lambda_i}\right)\left(\sum_{i=1}^p u_i^2 \lambda_i\right),$$

adică

$$\frac{1}{\left(\sum_{i=1}^p u_i^2 \lambda_i\right)} \leq \sum_{i=1}^p u_i^2 \frac{1}{\lambda_i}.$$

Cum $\sum_{i=1}^p u_i^2 = 1$, inegalitatea anterioară rezultă din inegalitatea lui Jensen aplicată funcţiei convexe $(0, \infty) \ni x \mapsto \frac{1}{x}$.

Pentru a doua inegalitate, să observăm că pentru orice $i \in \overline{1, p}$,

$$\frac{1}{\lambda_i} \leq \frac{1}{\lambda_1} + \frac{1}{\lambda_p} - \frac{\lambda_i}{\lambda_1 \lambda_p},$$

întrucât $\lambda_p \leq \lambda_i \leq \lambda_1$. Deci

$$\sum_{i=1}^p u_i^2 \frac{1}{\lambda_i} \leq \frac{1}{\lambda_1} + \frac{1}{\lambda_p} - \frac{\sum_{i=1}^p u_i^2 \lambda_i}{\lambda_1 \lambda_p}.$$

Obţinem

$$\begin{aligned}\left(\sum_{i=1}^p u_i^2 \frac{1}{\lambda_i}\right)\left(\sum_{i=1}^p u_i^2 \lambda_i\right) &\leq \left(\sum_{i=1}^p u_i^2 \lambda_i\right)\left(\frac{1}{\lambda_1} + \frac{1}{\lambda_p} - \frac{\sum_{i=1}^p u_i^2 \lambda_i}{\lambda_1 \lambda_p}\right) \\ &= \frac{\left(\sum_{i=1}^p u_i^2 \lambda_i\right)\left(\lambda_1 + \lambda_p - \sum_{i=1}^p u_i^2 \lambda_i\right)}{\lambda_1 \lambda_p}.\end{aligned}$$

Funcţia de gradul al doilea

$$\lambda \mapsto \frac{\lambda(\lambda_1 + \lambda_p - \lambda)}{\lambda_1 \lambda_p}$$

îşi atinge maximul pentru $\lambda = \frac{\lambda_1+\lambda_p}{2}$, deci

$$\left(\sum_{i=1}^{p} u_i^2 \frac{1}{\lambda_i}\right)\left(\sum_{i=1}^{p} u_i^2 \lambda_i\right) \leq \left(\frac{\lambda_1 + \lambda_p}{2}\right)^2 \frac{1}{\lambda_1 \lambda_p}$$
$$= \frac{1}{4}\left(\sqrt{\frac{\lambda_1}{\lambda_p}} + \sqrt{\frac{\lambda_p}{\lambda_1}}\right)^2.$$

Demonstraţia este completă. □

Teorema 4.2.15 *Cu notaţiile de mai sus are loc inegalitatea*

$$f(x_k) - \min f \leq (f(x_0) - \min f)\left(\frac{c-1}{c+1}\right)^{2k}, \ \forall k \geq 0,$$

unde $c = \lambda_{\max}/\lambda_{\min}$, *unde* $\lambda_{\max}$ *şi* $\lambda_{\min}$ *reprezintă cea mai mare şi respectiv cea mai mică valoare proprie a lui* Q.

Demonstraţie. Notăm, pentru orice k,

$$e_k = f(x_k) - \min f,$$
$$r_k = Qx_k + b.$$

Folosind calculele de mai sus, avem

$$p_k = -\frac{r_k}{\|r_k\|},$$
$$\alpha_k = \frac{\|r_k\|^3}{\langle Qr_k, r_k\rangle},$$

$$\begin{aligned} e_{k+1} &= e_k + \frac{1}{2}\alpha_k^2 \langle Qp_k, p_k\rangle + \alpha_k \langle Qx_k + b, p_k\rangle \\ &= e_k + \frac{1}{2}\frac{\|r_k\|^6}{\langle Qr_k, r_k\rangle^2}\frac{\langle Qr_k, r_k\rangle}{\|r_k\|^2} + \frac{\|r_k\|^3}{\langle Qr_k, r_k\rangle}\left\langle r_k, -\frac{r_k}{\|r_k\|}\right\rangle \\ &= e_k - \frac{1}{2}\frac{\|r_k\|^4}{\langle Qr_k, r_k\rangle}, \ \forall k. \end{aligned}$$

Similar,

$$\begin{aligned} e_k &= f(x_k) - f(\overline{x}) = \langle \nabla f(\overline{x}), x_k - \overline{x}\rangle + \frac{1}{2}\langle Q(x_k - \overline{x}), x_k - \overline{x}\rangle \\ &= \frac{1}{2}\langle Q(x_k - \overline{x}), x_k - \overline{x}\rangle = \frac{1}{2}\left\langle Qx_k + b, x_k + Q^{-1}b\right\rangle \\ &= \frac{1}{2}\left\langle Q^{-1}r_k, r_k\right\rangle, \ \forall k. \end{aligned}$$

Astfel, prin aplicarea Inegalităţii lui Kantorovici, obţinem

$$\begin{aligned}\frac{e_{k+1}}{e_k} &= 1-\frac{1}{2}\frac{1}{e_k}\frac{\|r_k\|^4}{\langle Qr_k,r_k\rangle} = 1-\frac{1}{2}\frac{2}{\langle Q^{-1}r_k,r_k\rangle}\frac{\|r_k\|^4}{\langle Qr_k,r_k\rangle}\\ &= 1-\frac{\|r_k\|^4}{\langle Q^{-1}r_k,r_k\rangle\langle Qr_k,r_k\rangle} \le 1-4\left(\sqrt{\frac{\lambda_1}{\lambda_p}}+\sqrt{\frac{\lambda_p}{\lambda_1}}\right)^{-2}\\ &= 1-4\left(\frac{\lambda_{\max}+\lambda_{\min}}{\sqrt{\lambda_{\max}\cdot\lambda_{\min}}}\right)^{-2}\\ &= 1-4\frac{\lambda_{\max}\cdot\lambda_{\min}}{(\lambda_{\max}+\lambda_{\min})^2} = \frac{(\lambda_{\max}-\lambda_{\min})^2}{(\lambda_{\max}+\lambda_{\min})^2} = \left(\frac{c-1}{c+1}\right)^2,\ \forall k.\end{aligned}$$

Aceasta asigură validitatea concluziei. □

Observaţia 4.2.16 *Dacă numărul c este foarte mare, atunci valoarea $\frac{c-1}{c+1}$ este aproape de* 1*, ceea ce face ca viteza de convergenţă a metodei să fie lentă, în general.*

Exerciţiul 4.42 *Fie $f:\mathbb{R}^2\to\mathbb{R}$ dată prin*

$$f(x) = x_1^2+x_2^2-0.2x_1x_2-2.2x_1+2.2x_2.$$

(i) Să se determine punctul de minim al lui f.

(ii) Aplicând metoda celei mai bune descreşteri pentru $x=(2,3)$ să se determine teoretic şi practic câte iteraţii sunt necesare pentru ca $f(x_k)-\min f\le 10^{-10}$.

4.2.5 Legătura dintre metoda gradient descent şi o ecuaţie diferenţială ordinară

În această secţiune arătăm modul în care metoda GD cu factor constant (aplicabilă, cum am văzut mai sus, unei clase de funcţii convexe) este strâns este de fapt o schemă de discretizare (de tip Euler) a unei ecuaţii diferenţiale ordinare (numită în literatură "gradient flow ODE"). Această abordare permite rediscutarea convergenţei metodei.

Fie deci $f:\mathbb{R}^p\to\mathbb{R}$ diferenţiabilă. Metoda GD cu factor constant $\beta>0$ implementează iteraţia

$$x_{k+1} = x_k-\beta\nabla f(x_k),\ \forall k\ge 0$$

cu $x_0\in\mathbb{R}^p$.

Ecuaţia diferenţială ordinară pe care o asociem este

$$x'(t) = -\nabla f(x(t)),\quad x(0)=x_0. \tag{EDO}$$

Scrierea

$$x(t+\beta)\approx x(t)+\beta x'(t) = x(t)-\beta\nabla f(x(t)),$$

arată legătura dintre o soluţie a ecuaţiei şi metoda GD, legătura pe care o detaliem mai jos.

Studiem mai întâi condiţii de existenţă şi unicitate a soluţiei pentru ecuaţia (EDO).

Propoziţia 4.2.17 *Presupunem că $f \in C^2(\mathbb{R}^p)$, că gradientul său este $L-$Lipschitz, adică*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \ \forall x, y \in \mathbb{R}^p$$

şi că f este coercivă. Atunci ecuaţia (EDO) are soluţie unică definită pe $[0, \infty)$.

Demonstraţie Existenţa locală şi unicitatea soluţiei x sunt consecinţe ale Teoremei lui Picard. Pe de altă parte,

$$(f(x(t)))' = \nabla f(x(t)) \cdot x'(t),$$

de unde obţinem

$$(f(x(t)))' = \nabla f(x(t)) \cdot (-\nabla f(x(t))) = -\|\nabla f(x(t))\|^2 \leq 0.$$

Astfel,

$$f(x(t)) \leq f(x_0),$$

deci x rămâne în mulţimea de nivel mărginită $\{x \mid f(x) \leq f(x_0)\}$. Pentru că ∇f este mărginită pe mulţimi mărginite, $\|x'\| = \|\nabla f(x(\cdot))\|$ rămâne mărginită. Deducem că nu are loc explozia în timp finit, deci soluţia există pentru orice $t \geq 0$. □

Fixăm acum un orizont finit de timp $T > 0$ şi notăm cu $x(t)$ soluţia ecuaţiei diferenţiale de mai sus cu data iniţială $x(0) = x_0$. Fie x_k iteraţiile metodei GD cu data iniţială x_0 şi factorul β.

Definim funcţia scară

$$\overline{x}^{\beta}(t) = x_k, \ \forall t \in [k\beta, (k+1)\beta).$$

Propoziţia 4.2.18 *Există o constantă C (depinzând de T, L, x_0) astfel încât pentru orice $\beta > 0$ suficient de mic are loc inegalitatea*

$$\sup_{0 \leq t \leq T} \left\|\overline{x}^{\beta}(t) - x(t)\right\| \leq C\beta.$$

Astfel, $\overline{x}^{\beta} \overset{\beta \to 0}{\to} x$ uniform pe $[0, T]$.

Demonstraţie Fixăm $k \in \mathbb{N}$ şi definim $t_k = k\beta$. Soluţia x a ecuaţiei (EDO) satisface

$$x(t_{k+1}) = x(t_k) + \int_0^{\beta} x'(t_k + s)\, ds = x(t_k) + \beta x'(t_k) + r_k,$$

unde restul r_k este

$$r_k = \int_0^{\beta} (x'(t_k + s) - x'(t_k)) ds.$$

Având în vedere că $x' = -\nabla f(x)$ şi că ∇f este Lipschitz,

$$\|x'(t_k + s) - x'(t_k)\| = \|\nabla f(x(t_k)) - \nabla f(x(t_k + s))\| \leq L\|x(t_k + s) - x(t_k)\|.$$

În plus,

$$\|x(t_k + s) - x(t_k)\| \leq \int_0^{s} \|x'(t_k + u)\| du.$$

Prin combinarea acestor inegalităţi şi prin integrare, deducem că

$$\|r_k\| \leq \int_0^\beta \|x'(t_k+s)-x'(t_k)\|\,ds \leq L\int_0^\beta \left(\int_0^s \|x'(t_k+u)\|du\right)ds$$
$$\leq L \max_{t\in[t_k,t_{k+1}]}\|x'(t)\|\int_0^\beta sds = \frac{L}{2}\beta^2 \max_{t\in[t_k,t_{k+1}]}\|x'(t)\|.$$

Metoda GD implementează recurenţa

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

Astfel,

$$x_{k+1}-x(t_{k+1}) = (x_k - x(t_k)) - \beta(\nabla f(x_k) - \nabla f(x(t_k))) - r_k.$$

Trecem la norme şi utilizăm din nou proprietatea Lipschitz şi avem

$$\|x_{k+1}-x(t_{k+1})\| \leq (1+\beta L)\|x_k - x(t_k)\| + \|r_k\|.$$

Definim $e_k = \|x_k - x(t_k)\|$ pentru $k \geq 0$ şi din cele de mai sus avem

$$e_{k+1} \leq (1+\beta L)e_k + c\beta^2,$$

unde $c = \frac{L}{2}\max_{t\in[0,T]}\|x'(t)\|$.

Cum relaţia de mai sus este valabilă pentru orice $k \in \mathbb{N}$, prin sumare, ţinând cont şi că $e_0 = 0$, găsim

$$e_k \leq c\beta^2 \sum_{j=0}^{k-1}(1+\beta L)^j.$$

Dar, pentru $k \leq \frac{T}{\beta}$, $(1+\beta L)^k \leq e^{LT}$, deci există o constantă $C > 0$ astfel încât

$$\|x_k - x(t_k)\| \leq C\beta,$$

care se extinde la $t \in [0,T]$ pentru funcţia scară $\overline{x}^\beta$. □

Problema 4.43 *Presupunem că ecuaţia (EDO) admite soluţie unică, notată cu x, pe $[0,\infty)$. Dacă, în plus, f este mărginită inferior, atunci*

$$\lim_{t\to\infty}\nabla f(x(t)) = 0.$$

Demonstraţie Prin integrarea relaţiei

$$(f(x(t)))' = -\|\nabla f(x(t))\|^2 \leq 0,$$

obţinem că pentru orice $T > 0$ are loc egalitatea.

$$f(x_0) - f(x(T)) = \int_0^T \|\nabla f(x(t))\|^2 dt.$$

Cum f este mărginită inferior, $\inf f \in \mathbb{R}$ şi

$$\int_0^T \|\nabla f(x(t))\|^2 dt \leq f(x_0) - \inf f, \ \forall T > 0,$$

deci integrala

$$\int_0^\infty \|\nabla f(x(t))\|^2 dt$$

este convergentă, ceea ce arată că $\lim_{t\to\infty} \nabla f(x(t)) = 0$. □

Observaţia 4.2.19 *Din rezultatele de mai sus, deducem că în ipotezele menţionate,* $\lim_k \nabla f(x_k) = 0$

Propoziţia 4.2.20 *Presupunem că suntem în condiţiile Propoziţiei 4.2.17, ecuaţia (EDO) având soluţie unică x definită pe $[0,\infty)$. În plus, presupunem că f este strict convexă şi notăm cu $\overline{x}$ punctul de minim global al lui f iar cu $\bar{f}$ valoarea $f(\overline{x})$. Atunci pentru orice $t > 0$*

$$f(x(t)) - f(\overline{x}) \leq \frac{\|x_0 - \overline{x}\|^2}{2t}.$$

Demonstraţie Considerăm funcţia $E : [0,\infty) \to \mathbb{R}$ dată prin

$$E(t) = t\left(f(x(t)) - \bar{f}\right) + \frac{\|x(t) - \overline{x}\|^2}{2}.$$

Derivata este

$$\begin{aligned} E'(t) &= f(x(t)) - \bar{f} + t\nabla f(x(t)) x'(t) + \langle x'(t), (x(t) - \overline{x})\rangle \\ &= f(x(t)) - \bar{f} - t\|\nabla f(x(t))\|^2 - \nabla f(x(t))(x(t) - \overline{x}) \\ &= -t\|\nabla f(x(t))\|^2 - \left(\bar{f} - f(x(t)) - \nabla f(x(t))(\overline{x} - x(t))\right). \end{aligned}$$

Pe baza convexităţii lui f deducem că $E'(t) \leq 0$ pentru orice $t > 0$. Aşadar

$$t\left(f(x(t)) - \bar{f}\right) \leq E(t) \leq E(0) = \|x_0 - \overline{x}\|^2, \ \forall t > 0$$

şi deducem concluzia. □

4.2.6 Posibilităţi de accelerare a metodei gradient descent

Algoritmul gradient descent poate avansa foarte lent în regiuni plate ale graficului funcţiei obiectiv, după cum am văzut deja la unele funcţii considerate anterior (de exemplu, funcţia Rosenbrock). Câteva soluţii pentru evitarea acestei dificultăţi sunt prezentate prezentate în continuare, într-o manieră euristică

O euristică simplă, cunoscută sub denumirea de *metoda impulsului* (sau *momentum method,* sau, încă, *heavy ball method*), constă în a avansa mai repede în direcţii care s-au dovedit eficiente

anterior şi a încetini în direcţii unde gradientul s-a schimbat brusc, similar cu o bilă care se rostogoleşte la vale. Aceasta poate fi implementată astfel:

$$m_{k+1} = \gamma m_k + \nabla f(x_k),$$
$$x_{k+1} = x_k - \beta m_{k+1},$$

unde m_k se numeşte impuls, $m_0 = 0$ şi $0 < \gamma < 1$. O valoare tipică pentru γ este 0.9. Pentru $\gamma = 0$, metoda devine gradient descent obişnuită.

Deoarece actualizăm parametrii folosind media gradientului m_{k+1}, şi nu ultimul gradient, $\nabla f(x_k)$, observăm ca gradienţii anteriori pot influenţa iteraţia curentă. În plus, atunci când vom combina această metodă cu Stochastic Gradient Descent (a se vedea mai jos), se vor putea observa efectele unui minibatch mai mare, fără costuri computaţionale suplimentare.

O problemă care poate apărea în momentum method este că poate să nu încetinească suficient în partea de jos a unei "văi", cauzând oscilaţii. Metoda *gradientului accelerat Nesterov* (sau *Nesterov accelerated gradient*) modifică metoda de tip gradient descent pentru a include un pas de extrapolare, astfel:

$$m_{k+1} = \gamma m_k + \nabla f(x_k - \gamma\beta m_k),$$
$$x_{k+1} = x_k - \beta m_{k+1}.$$

Aceasta explică de ce metoda gradientului accelerat Nesterov este numită uneori *Nesterov momentum*. De asemenea, arată cum această metodă poate fi mai rapidă decât momentum method standard: vectorul de impuls indică deja, în linii mari, în direcţia corectă, astfel încât măsurarea gradientului la noua locaţie, $x_k - \beta m_k$, şi nu la locaţia curentă, x_k, poate fi mai precisă.

Metoda gradientului accelerat Nesterov este demonstrabil mai rapidă decât gradientul descendent pentru funcţii convexe, pentru γ şi β sunt alese corespunzător. Se numeşte "accelerat" datorită acestei îmbunătăţiri, care este optimă pentru metodele de tip gradient descent folosind doar informaţii de ordinul întâi, atunci când funcţia obiectiv este convexă şi are gradienţi Lipschitz continui. În practică, totuşi, această metodă poate fi mai lentă decât steepest descent, şi poate deveni instabilă pentru alegeri nepotrivite ale lui γ şi β.

Implementare 4.44 Considerăm următoarea implementare a GD pentru funcţia Rosenbrock.

```
% GD pentru minimizarea functiei Rosenbrock (convexa)
clc; clear;
tic
f = @(x, y) (1 - x).^2 + 100 * (y - x.^2).^2;
grad_f = @(x, y) [-2 * (1 - x) - 400 * x .* (y - x.^2); 200 * (y - x.^2)];
% data initiala
x = -1.5;
y =2;
% parametri GD
alpha = 0.001; % Learning rate
tolerance = 1e-6; % toleranta
max_iters = 10000; % numarul maxim de iteratii
```

```
   iter = 0;
   x_vals = x;
   y_vals = y;
   while iter < max_iters
    grad = grad_f(x, y);
    x_new = x - alpha * grad(1);
    y_new = y - alpha * grad(2);
    % criteriul de oprire
    if norm([x_new - x, y_new - y]) < tolerance
    break;
    end
    x = x_new;
    y = y_new;
    x_vals = [x_vals; x];
    y_vals = [y_vals; y];
    iter = iter + 1;
   end
   % rezultatele
   fprintf('Punct de minim (x, y) = (%f, %f) after %d iterations\n', x, y, iter);
   fprintf('Valoare minima = %f\n', f(x, y));
   % Plot functie
   [X, Y] = meshgrid(-3:0.1:3, -3:0.1:3);
   Z = f(X, Y);
   figure;
   contour(X, Y, Z, 50);
   hold on;
   plot(x_vals, y_vals, 'r.-', 'LineWidth', 1.5, 'MarkerSize', 10);
   plot(x_vals(1), y_vals(1), 'go', 'MarkerSize', 10, 'LineWidth', 2); % punctul de
start
   plot(x_vals(end), y_vals(end), 'bo', 'MarkerSize', 10, 'LineWidth', 2); % punctul
final
   title('GD veritabil');
   xlabel('x');
   ylabel('y');
   legend('Contururi', 'GD traiectorie', 'Punct start', 'Punct final');
   grid on;
   hold off;
   toc
```

care returnează

```
Minimum found at (x, y) = (0.990202, 0.980461) after 10000 iterations
Minimum value of the Rosenbrock function = 0.000096
```

Să se implementeze pe acest model GD cu impuls şi GD cu impuls Nesterov şi să se compare rezultatul cu metoda GD originală de mai sus. Să se exerseze şi pe alte funcţii test.

4.2.7 Algoritmi specifici *machine learning* bazaţi pe metoda direcţiei de descreştere

Descriem acum, fără demonstraţii privind convergenţa, câţiva algoritmi derivaţi din metoda direcţiei de descreştere care sunt adaptaţi specificităţii unor probleme ce apar recurent în machine learning. Încercăm să prezentă la nivel euristic, intuitiv motivele pentru care aceste variaţii sunt mai potrivite anumitor probleme de mari dimensiuni decât algoritmul original.

Stochastic Gradient Descent

Se consideră un set de date de antrenament, $\{(a_i, y_i)\}_{i=\overline{1,n}}$, cu $a_i \in \mathbb{R}$ şi $y_i \in \mathbb{R}^p$, ce urmează o distribuţie necunoscută. Funcţia obiectiv este:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(h(x, a_i), y_i\right).$$

Exemplu de funcţie de pierdere ("loss functions"): "square loss":

$$\ell\left(h(x, a_i), y_i\right) = \sum_{i\in\overline{1,n}} \|y_i - h(x, a_i)\|^2.$$

Vom considera algoritmul de tipul steepest descent method (or gradient descent, GD):

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

unde gradientul funcţiei obiectiv este

$$\nabla f(x) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x).$$

Factorul $\beta > 0$ se numeşte rată de învăţare (*learning rate*). Dacă folosim acest algoritm, costul computaţional creşte liniar cu n. Aşadar, cu cât setul de date de antrenament este mai mare, cu atât costul computaţional pentru fiecare iteraţie este mai mare.

Stochastic Gradient Descent (SGD) reduce costul computaţional la fiecare iteraţie. Fie $m < n$. La fiecare pas se alege o mulţime $\mathcal{I}$ de indici (adică un lot, *minibatch*) având cardinalul m şi actualizăm iteraţia după formula

$$x_{k+1} = x_k - \beta \cdot \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \nabla f_i(x_k).$$

Acest algoritm se bazează pe observaţia (ce poate fi riguros demonstrată) conform căreia, dacă mulţimile $\mathcal{I}$ sunt aleator alese, atunci gradientul stochastic este un estimator nedeplasat al gradientului complet, i.e.,

$$\mathbb{E}\left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \nabla f_i(x)\right) = \nabla f(x), \ \forall x \in \mathbb{R}^p.$$

Implementare 4.45 Considerăm exemplul $f : \mathbb{R}^2 \to \mathbb{R}$,

$$f(x,y) = (x-2)^2 + (y-3)^2 + 2(x+1)^2 + (y-1)^2 + (x-1)^2 + 2(y+2)^2,$$

adică suma funcţiilor $f_i : \mathbb{R}^2 \to \mathbb{R}$ $(i \in \overline{1,3})$,

$$f_1(x,y) = (x-2)^2 + (y-3)^2,\ f_2(x,y) = 2(x+1)^2 + (y-1)^2,\ f_3(x,y) = (x-1)^2 + 2(y+2)^2.$$

Să se scrie un program care să implementeze SGD cu rata de învătare constantă 0.05, rulează 100 de iteraţii şi alege câte unul din cei trei gradienţi cu egală probabilitate (minibatch de cardinal 1). Pentru selecţia gradienţilor se poate utiliza codul:

```
r = randi(3); % selectează random 1, 2 sau 3
if r == 1
    grad = grad_f1(x, y);
elseif r == 2
    grad = grad_f2(x, y);
else
    grad = grad_f3(x, y);
end
```

Să se compare cu algoritmul direcţiei de descreştere corespunzător.

Implementare 4.46 Să se implementeze SGD cu minibatch de cardinal 2 pentru minimizarea funcţiei de mai sus. Să se compare cu metoda GD originală.

Deşi rata teoretică de convergenţă a SGD este mai lentă decât cea a GD pe întregul set de date de antrenament $\{1, 2, ..., n\}$, în practică SGD este adesea mai rapid, deoarece timpul pentru fiecare pas este mult mai scăzut (într-un program Matlab/Octave timpul de execuţie este afişat prin includerea intrucţiunilor `tic`...`toc` la începutul, respectiv sfârşitul secvenţei avute în vedere). Pe de altă parte, se pot considera rate de învăţare dinamice. Pentru a înţelege de ce SGD poate progresa mai rapid decât GD pe întregul set de date, să presupunem că avem un set de date format dintr-un singur exemplu duplicat de K ori. Antrenarea pe întregul set de date va fi (cel puţin) de K ori mai lentă decât SGD, deoarece va pierde timp calculând gradientul pentru exemplele repetate. Chiar şi în absenţa duplicatelor, antrenarea pe întregul set de date poate fi ineficientă, deoarece, la începutul antrenării, parametrii nu sunt bine estimaţi, ceea ce face ca evaluarea în detaliu a gradientului poate să nu merite efortul.

SGD precondiţionat

Se poare considera de asemenea un SGD precondiţionat, care presupune următoarea actualizare:

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

unde M_k este o matrice de precondiţionare, de obicei aleasă pozitiv definită. Din păcate, zgomotul din estimările gradientului face dificilă estimarea fiabilă a matricei hessiene, ceea ce complică utilizarea metodelor discutate anterior. De asemenea, este costisitor să aflăm direcţia de descreştere folosind o matrice de precondiţionare completă. Prin urmare, în practică vom folosi de obicei o matrice diagonală M_k. Aceste precondiţionări nu folosesc neapărat informaţii de ordinul doi, dar adesea duc la accelerări în comparaţie cu SGD simplu.

AdaGrad

Metoda AdaGrad (de la "gradient adaptiv") a fost concepută iniţial pentru optimizarea funcţiilor obiectiv convexe unde multe componente ale gradientului sunt zero; aceasta corespunde situaţiei când anumite caracteristici sunt foarte rar întâlnite în input, cum ar fi cuvintele rare. Actualizarea are forma pe componente, pentru fiecare $d = \overline{1,p}$:

$$x_{k+1,d} = x_{k,d} - \frac{\alpha_k}{\sqrt{s_{k,d} + \varepsilon}} g_{k,d},$$

unde $g_{k,d}$ reprezintă componenta d a vectorului gradient $g_k = \nabla f\left(x_k\right)$,

$$s_{k,d} = \sum_{i=1}^{k} g_{i,d}^2$$

reprezintă suma componentelor gradienţilor la pătrat, iar $\varepsilon > 0$ este un termen pozitiv mic introdus pentru a evita împărţirea la zero.

Văzută ca SGD precondiţionat, aceasta este echivalentă cu

$$M_k = \operatorname{diag}\left(s^{(k)} + \varepsilon\right)^{\frac{1}{2}}$$

şi este un exemplu de *rată de învăţare adaptativă*.

RMSProp şi AdaDelta

O caracteristică definitorie a algoritmului AdaGrad este că termenul de la numitor devine mare în timp, deci rata efectivă de învăţare scade. Deşi necesară pentru a asigura convergenţa, poate influenţa în mod negativ performanţa dacă numitorul devine mare prea repede.

O metodă alternativă este de a utiliza o medie ponderată exponenţial a pătratelor gradienţilor precedenţi, în locul sumei acestora:

$$s_{k+1,d} = \beta s_{k,d} + (1-\beta) g_{k,d}^2.$$

În practică, se foloseşte de obicei $\beta \sim 0.9$, care dă mai multă importanţă iteraţiilor recente. În acest caz,

$$\sqrt{s_{k,d}} \approx \operatorname{RMS}\left(g_{1:k,d}\right) = \sqrt{\frac{1}{k}\sum_{i=1}^{k} g_{i,d}^2},$$

unde RMS provine de la "root mean squared". Actualizarea generală a metodei *RMSProp* este:

$$x_{k+1,d} - x_{k,d} = -\frac{\alpha_k}{\sqrt{s_{k,d} + \varepsilon}} g_{k,d}.$$

Metoda *AdaDelta* extinde această idee adăugând o medie ponderată exponenţial a actualizărilor anterioare astfel:

$$x_{k+1,d} - x_{k,d} = -\alpha_k \frac{\sqrt{\delta_{k-1,d} + \varepsilon}}{\sqrt{s_{k,d} + \varepsilon}} g_{k,d},$$

unde

$$\delta_{k,d} = \beta\delta_{k-1,d} + (1-\beta)(x_{k+1,d} - x_{k,d})^2,$$

iar $s_{k,d}$ are aceeaşi expresie ca la metoda RMSProp de mai sus. Metoda are avantajul că numărătorul şi numitorul se pot simplifica printr-un ordin de mărime, deci rămânem doar cu multiplicarea pe componente a gradientului cu un scalar. Aceasta elimină nevoia de a adapta rata de învăţare α_k, pe care o putem alege 1, însă deoarece aceste rate adaptative nu descresc, aceste metode nu converg în mod sigur la o soluţie.

Adam

Este posibil să combinăm RMSProp cu metoda impulsului (momentum method). În particular, calculăm EWMA, o medie ponderată exponenţială a gradienţilor (ca în metoda impulsului) şi a gradienţilor la pătrat (ca în RMSProp):

$$m_k = \beta_1 m_{k-1} + (1-\beta_1) g_k,$$
$$s_k = \beta_2 s_{k-1} + (1-\beta_2)\, g_k^2,$$

unde în a doua ecuaţie egalitatea se realizează pe componente. Facem apoi următoarea actualizare (unde egalitatea este înţeleasă pe componente):

$$x_{k+1} - x_k = -\alpha_k \frac{1}{\sqrt{s_k + \varepsilon}} m_k.$$

Metoda rezultată este cunoscută sub numele de *Adam*, nume care provine de la "adaptive moment estimation".

Valorile standard pentru constante sunt $\beta_1 = 0.9$, $\beta_2 = 0.999$ şi $\varepsilon = 10^{-6}$. (Dacă setăm $\beta_1 = 0$ şi nu folosim corecţia de bias, recuperăm RMSProp, care nu utilizează momentum). Pentru rata de învăţare generală, se obişnuieşte să folosim o valoare fixă, cum ar fi $\alpha_k = 0.001$. Din nou, deoarece rata de învăţare adaptivă poate să nu scadă în timp, convergenţa nu este garantată.

Dacă iniţializăm $m_0 = s_0 = 0$, atunci estimările iniţiale vor fi conduse către valori mici. Se recomandă, prin urmare, utilizarea momentelor corectate pentru bias, care cresc valorile mai devreme în procesul de optimizare. Aceste estimări sunt date de:

$$\hat{m}_k = \frac{m_k}{1-\beta_1^k}, \quad \hat{s}_k = \frac{s_k}{1-\beta_2^k}.$$

Yogi

Când se utilizează metode de scalare diagonală, rata generală de învăţare este determinată de $\alpha_0 M_k^{-1}$, care se schimbă în timp. Prin urmare, aceste metode sunt adesea numite metode cu *rată de învăţare adaptivă*. Totuşi, ele necesită în continuare setarea unei rate de învăţare iniţiale α_0.

Deoarece metodele EWMA sunt utilizate de obicei în cadru stochastic, unde estimările gradientului au zgomot, adaptarea ratei de învăţare poate duce la pierderea convergenţei chiar şi pentru probleme convexe. Au fost propuse diverse soluţii la această problemă. De exemplu, *actualizarea Yogi* modifică Adam prin înlocuirea:

$$s_k = \beta_2 s_{k-1} + (1-\beta_2)\, g_k^2 = s_{k-1} + (1-\beta_2)(g_k^2 - s_{k-1})$$

cu

$$s_k = s_{k-1} + (1-\beta_2)(g_k^2 - s_{k-1}) \odot \operatorname{sgn}(g_k^2 - s_{k-1}),$$

unde sgn reprezintă funcţia signum, având valoarea 1 pentru numere strict pozitive, -1 pentru numere strict negative şi 0 pentru 0.

Aşadar, am obţinut următorul algoritm.

ADAM (Adaptive Moment Estimation)

Pasul 1**:** Introducere pas α, introducere $\beta_1, \beta_2 \in [0,1)$, introducere constantă $\varepsilon > 0$, introducere parametru iniţial $x^{(0)}$

Introducere $m_0 = 0$, $s_0 = 0$, $k = 0$

Pasul 2: Calculează gradient $g_k = \nabla f(x_k)$

Calculează biased first moment estimate: $m_k = \beta_1 m_{k-1} + (1-\beta_1) g_k$

Calculează biased second moment estimate: $s_k = s_{k-1} + (1-\beta_2)(g_k^2 - s_{k-1}) \odot \operatorname{sgn}(g_k^2 - s_{k-1})$

Calculează bias-corrected first moment: $\hat{m}_k = \frac{m_k}{1-\beta_1^k}$

Calculează bias-corrected second moment: $\hat{s}_k = \frac{s^{(k)}}{1-\beta_2^k}$

Calculează noul x_{k+1}: $x_{k+1} = x_k - \alpha_k \frac{1}{\sqrt{s_k+\varepsilon}} m_k$

Pasul 3: metoda converge; se afişează x_{k+1} şi $f(x_{k+1})$.

Comparativ cu celelalte metode prezentate, algoritmiul Adam se comportă mai bine, în special aplicat pentru reţele neuronale de tip multilayer.

4.2.8 Metoda gradienţilor conjugaţi

Metoda pe care o prezentăm acum se referă la rezolvarea unui sistem de ecuaţii liniare de forma

$$Qx = b,$$

unde Q este o matrice pătratică de dimensiune p, simetrică şi pozitiv definită. Este de observat că rezolvarea sistemului de mai sus este echivalentă cu minimizarea funcţiei pătratice strict convexe $q : \mathbb{R}^p \to \mathbb{R}$,

$$q(x) = \frac{1}{2}\langle Qx, x\rangle - \langle b, x\rangle$$

pentru că

$$\nabla q(x) = Qx - b.$$

Notăm

$$r(x) := \nabla q(x) = Qx - b.$$

Definiţia 4.2.21 *Fie Q o matrice pătratică de dimensiune p, simetrică şi pozitiv definită. Definim pe $\mathbb{R}^p \times \mathbb{R}^p$ aplicaţia cu valori reale*

$$\langle x, y\rangle_Q = \langle Qx, y\rangle,$$

numit $Q-$produs scalar.

Observaţia 4.2.22 *Se verifică uşor că aplicaţia de mai sus este într-adevăr un produs scalar, iar norma corespunzătoare, numită $Q-$normă, este*

$$\|x\|_Q = \sqrt{\langle Qx, x\rangle}.$$

Definiţia 4.2.23 *O mulţime de direcţii (i.e., vectori nenuli) $(d_i)_{i\in\overline{1,k}}$ (unde $k \in \mathbb{N}\setminus\{0\}$) se numeşte $Q-$ortogonală, iar elementele sale $Q-$conjugate dacă*

$$\|d_i\|_Q \neq 0,\ \forall i \in \overline{1,k} \text{ şi } \langle d_i, d_j\rangle_Q = 0,\ \forall i,j \in \overline{1,k},\ i \neq j.$$

Lema 4.2.24 *O mulţime de direcţii $Q-$conjugate este liniar independentă.*

Demonstraţie. Fie $(d_i)_{i\in\overline{1,k}}$ (unde $k \in \mathbb{N}\setminus\{0\}$) o mulţime de direcţii $Q-$conjugate şi fie $(\alpha_i)_{i\in\overline{1,n}}$ numere reale. Dacă $\sum_{i=1}^k \alpha_i d_i = 0$, atunci, pentru orice $j \in \overline{1,k}$,

$$0 = \left\langle \sum_{i=1}^k \alpha_i d_i, Qd_j \right\rangle = \sum_{i=1}^k \alpha_i \langle d_i, Qd_j\rangle = \alpha_j \|d_j\|_Q^2.$$

Obţinem astfel că toţi scalarii sunt nuli. □

Observaţia 4.2.25 *Dacă $(d_i)_{i\in\overline{1,k}}$ ($k \in \mathbb{N}\setminus\{0\}$) este o mulţime de direcţii $Q-$conjugate şi $x_0 \in \mathbb{R}^p$, atunci a minimiza funcţia q pe subspaţiul afin*

$$A = x_0 + \operatorname{span}\left((d_i)_{i\in\overline{1,k}}\right)$$

revine la următorul argument: pentru $A \ni x = x_0 + \sum_{i=1}^k a_i d_i$ (unde $(a_i)_{i\in\overline{1,k}}$ sunt numere reale) avem, folosind $Q-$conjugarea,

$$\begin{aligned} q(x) &= q(x_0) + \sum_{i=1}^k a_i \langle r_0, d_i\rangle + \frac{1}{2}\sum_{i,j=1}^k a_i a_j \langle Qd_i, d_j\rangle \\ &= q(x_0) + \sum_{i=1}^k \left[a_i \langle r_0, d_i\rangle + \frac{1}{2}a_i^2 \langle Qd_i, d_i\rangle\right], \end{aligned}$$

deci minimul lui q pe A se obţine minimizând separat pentru fiecare a_i; avem minimul

$$\overline{x} = x_0 + \sum_{i=1}^k a_i d_i,\ \text{unde } a_i = -\frac{\langle r_0, d_i\rangle}{\langle Qd_i, d_i\rangle}.$$

Observaţia 4.2.26 *Având o mulţime de vectori liniar independenţi $(v_i)_{i\in\overline{1,k}}$ putem genera, prin procedeul Gram-Schmidt, o mulţime de direcţii $Q-$conjugate $(d_i)_{i\in\overline{1,k}}$ astfel încât* $\operatorname{span}\left((v_i)_{i\in\overline{1,k}}\right) = \operatorname{span}\left((d_i)_{i\in\overline{1,k}}\right)$. *Procedeul este descris mai jos.*

Alegem $d_1 = v_1$. Dacă am construit $(d_j)_{i\in\overline{1,i-1}}$, definim

$$d_i = v_i + \sum_{j=1}^{i-1} \beta_j d_j,$$

unde coeficienţii β_j sunt determinaţi astfel: cum dorim ca $\langle d_i, d_l\rangle_Q = 0$ pentru orice $l < i$, trebuie ca

$$0 = \langle d_i, d_l\rangle_Q = \left\langle v_i + \sum_{j=1}^{i-1} \beta_j d_j, d_l \right\rangle_Q$$
$$= \langle v_i, d_l\rangle_Q + \sum_{j=1}^{i-1} \beta_j \langle d_j, d_l\rangle_Q = \langle v_i, d_l\rangle_Q + \beta_l \langle d_l, d_l\rangle_Q, \ \forall l < i$$

deci

$$\beta_l = -\frac{\langle v_i, d_l\rangle_Q}{\langle d_l, d_l\rangle_Q}, \ \forall l < i$$

şi astfel

$$d_i = v_i - \sum_{j=1}^{i-1} \frac{\langle v_i, d_j\rangle_Q}{\langle d_j, d_j\rangle_Q} d_j = v_i - \sum_{j=1}^{i-1} \frac{\langle Qv_i, d_j\rangle}{\langle Qd_j, d_j\rangle} d_j.$$

Considerăm metoda de mai jos, numită metoda direcţiilor conjugate.

Fie $x_0 \in \mathbb{R}^p$ şi $(d_i)_{i\in\overline{0,n-1}}$ ($n \in \mathbb{N}^*$) este o mulţime de direcţii $Q-$conjugate. Pentru $k \in \overline{0, n-1}$, definim

$$x_{k+1} = x_k - \frac{\langle r_k, d_k\rangle}{\langle Qd_k, d_k\rangle} d_k, \text{ unde } r_k = \nabla q(x_k) = Qx_k - b. \tag{4.10}$$

Astfel, conform Observaţiei 4.2.25, x_{k+1} este minimul lui q pe subspaţiul $1-$dimensional

$$\{x_k + \alpha d_k \mid \alpha \in \mathbb{R}\}.$$

Are loc rezultatul următor care deschide calea către metoda gradienţilor conjugaţi.

Teorema 4.2.27 *Fie $x_0 \in \mathbb{R}^p$ şi $(d_i)_{i\in\overline{0,p-1}}$ este o mulţime de direcţii $Q-$conjugate. Fie $(x_i)_{i\in\overline{1,p}}$ punctele generate pe baza metodei direcţiilor conjugate de mai sus. Atunci pentru orice $k \in \overline{1, p}$, x_k este minimum global al funcţiei q pe subspaţiul afin $k-$dimensional $x_0 + \operatorname{span}\left((d_i)_{i\in\overline{0,k-1}}\right)$. În particular, soluţia ecuaţiei $Qx = b$ poate fi găsită în cel mult p paşi. Mai mult, pentru orice $k \in \overline{1, p}$ şi $i \in \overline{0, k-1}$, $\langle r_k, d_i\rangle_Q = 0$.*

Demonstraţie. Fie $k \in \overline{1, p}$. Conform Observaţiei 4.2.25,

$$x_k := x_0 + \sum_{i=0}^{k-1} -\frac{\langle r_0, d_i\rangle}{\langle Qd_i, d_i\rangle} d_i,$$

este minimul lui q pe $x_0 + \operatorname{span}\left((d_i)_{i\in\overline{0,k-1}}\right)$. Arătăm că aceste elemente coincid cu cele definite prin relaţia (4.10). Dacă, pentru $i \in \overline{0, k-1}$, $x \in x_0 + \operatorname{span}\left((d_j)_{j\in\overline{0,i-1}}\right)$, atunci există $(\gamma_j)_{i\in\overline{0,i-1}}$ astfel încât

$$x = x_0 + \sum_{j=0}^{i-1} \gamma_j d_j$$

şi atunci

$$\nabla q(x) = r_0 + \sum_{j=0}^{i-1} \gamma_j Q d_j$$

şi

$$\langle \nabla q(x), d_i \rangle = \langle r_0, d_i \rangle + \sum_{j=0}^{i-1} \gamma_j \langle Q d_j, d_i \rangle = \langle r_0, d_i \rangle .$$

Alegem $x = x_i \in x_0 + \text{span}\left((d_j)_{j \in \overline{0,i-1}} \right)$ şi deducem

$$\langle r_i, d_i \rangle = \langle r_0, d_i \rangle .$$

Astfel deducem că pentru orice $i \in \overline{0, p-1}$

$$\frac{\langle r_0, d_i \rangle}{\langle Q d_i, d_i \rangle} = \frac{\langle r_i, d_i \rangle}{\langle Q d_i, d_i \rangle},$$

de unde obţinem prima concluzie.

Acum, a doua concluzie este evidentă.

Pentru ultima concluzie, observăm că pentru orice $k \in \overline{1, p}$, x_k minimizează (fără restricţii) funcţia $h : \mathbb{R}^k \to \mathbb{R}$

$$h(\gamma_0, \gamma_1, ..., \gamma_{k-1}) = q\left(x_0 + \sum_{i=0}^{k-1} \gamma_i d_i \right).$$

ceea ce înseamnă că pentru orice $i \in \overline{0, k-1}$

$$\frac{\partial h}{\partial \gamma_i}(\gamma_0, \gamma_1, ..., \gamma_{k-1}) = \langle \nabla q(x_k), d_i \rangle = 0.$$

Demonstraţia este completă. □

Metoda gradienţilor conjugaţi este o metodă a direcţiilor conjugate în care se construiesc la fiecare pas direcţii $Q-$conjugate folosind procedeul Gram-Schmidt aplicat vectorilor $(-\nabla q(x_i))_{i \in \overline{0,k}}$. Se observă faptul că modul de definire, la fiecare pas, al elementelor x_i asigură ipoteza de liniară independenţă necesară aplicării procedeului menţionat. În versiunea sa preliminară, metoda este următoarea.

Se alege $x_0 \in \mathbb{R}^p$ şi se defineşte $d_0 = -\nabla q(x_0) = -r_0$.

Atâta timp cât $r_k \neq 0$ definim

$$d_k = -r_k + \sum_{j=0}^{k-1} \frac{\langle Q r_k, d_j \rangle}{\langle Q d_j, d_j \rangle} d_j,$$
$$\alpha_k = \frac{\|r_k\|^2}{\langle Q d_k, d_k \rangle},$$
$$x_{k+1} = x_k + \alpha_k d_k.$$

Alegerea lui α_k provine din Metoda direcţiilor conjugate şi din faptul că, pe baza Teoremei 4.2.27,

$$\langle r_k, d_k\rangle = \left\langle r_k, -r_k + \sum_{j=0}^{k-1} \frac{\langle Qr_k, d_j\rangle}{\langle Qd_j, d_j\rangle} d_j \right\rangle = -\left\|r_k\right\|^2, \ \forall k.$$

Aşa cum e descrisă mai sus, metoda este costisitoare din punct de vedere computaţional din cauza formulei direcţiilor d_k. Următorul rezultat permite simplificarea semnificativă a acestei formule.

Teorema 4.2.28 *Dacă elementele $(r_i)_{i\in\overline{0,k}}$ nu sunt zero, atunci ele sunt mutual $Q-$ortogonale şi*

$$d_k = -r_k + \beta_k d_{k-1}, \ \textit{unde } \beta_k = \frac{\left\|r_k\right\|^2}{\left\|r_{k-1}\right\|^2}.$$

Demonstraţie. Cum $(d_i)_{i\in\overline{0,k-1}}$ sunt generaţi prin procedeul Gram-Schmidt din $(-r_i)$ avem

$$\operatorname{span}\left((d_i)_{i\in\overline{0,k-1}}\right) = \operatorname{span}\left((r_i)_{i\in\overline{0,k-1}}\right).$$

Cum $(d_i)_{i\in\overline{0,k-1}}$ sunt $Q-$conjugaţi, din Teorema 4.2.27, r_k este $Q-$ortogonal pe toţi vectorii $(d_i)_{i\in\overline{0,k-1}}$, deci şi pe vectorii $(r_i)_{i\in\overline{0,k-1}}$, ceea ce implică prima parte a concluziei.

În particular, $\alpha_i \neq 0$ pentru orice $i \in \overline{0, k-1}$. Atunci

$$Qd_i = Q\left(\frac{x_{i+1} - x_i}{\alpha_i}\right) = \frac{(Qx_{i+1} - b) - (Qx_i - b)}{\alpha_i} = \frac{r_{i+1} - r_i}{\alpha_i}, \ \forall i \in \overline{0, k-1}.$$

Deci,

$$\langle Qr_k, d_i\rangle = \langle r_k, Qd_i\rangle = \frac{1}{\alpha_i}\langle r_k, r_{i+1} - r_i\rangle, \ \forall i \in \overline{0, k-1}.$$

Astfel,

$$\langle Qr_k, d_i\rangle = 0, \ \forall i \in \overline{0, k-2}$$
$$\langle Qr_k, d_{k-1}\rangle = \frac{\left\|r_k\right\|^2}{\alpha_{k-1}}.$$

Substituim în expresia lui d_k şi avem

$$d_k = -r_k + \frac{\left\|r_k\right\|^2}{\alpha_{k-1}\langle Qd_{k-1}, d_{k-1}\rangle} = -r_k + \frac{\left\|r_k\right\|^2}{\left\|r_{k-1}\right\|^2} d_{k-1},$$

ceea ce încheie demonstraţia. □

Pe baza teoremei tocmai demonstrate, putem formula Metoda gradienţilor conjugaţi astfel:
Se alege $x_0 \in \mathbb{R}^p$ şi se defineşte $d_0 = -\nabla q(x_0) = -r_0$.

Atâta timp cât $r_k \neq 0$ definim

$$\begin{aligned}
\alpha_k &= \frac{\|r_k\|^2}{\langle Qd_k, d_k\rangle},\\
x_{k+1} &= x_k + \alpha_k d_k,\\
r_{k+1} &= r_k + \alpha_k Q d_k,\\
\beta_{k+1} &= \frac{\|r_{k+1}\|^2}{\|r_k\|^2},\\
d_{k+1} &= -r_{k+1} + \beta_{k+1} d_k.
\end{aligned}$$

Teoretic, Metoda gradienţilor conjugaţi determină exact soluţia în cel mult p paşi. Totuşi, din cauza rotunjirilor din cadrul calculelor, aceasta este o situaţie rară în practică. În schimb, este o metodă eficientă pentru sisteme cu dimensiune mare.

Implementare 4.47 Să se scrie o funcţe `function x=conj_grad(A,b,x,maxiter)` care să ilustreze această metodă pe un sistem liniar generat aleatoriu după modelul de mai jos:

```
clear
dim=12;
b=10*rand(dim,1)
B=10*rand(dim,dim)
A=B*B'
x=zeros(dim,1);
maxiter=dim+3;
conj_grad(A,b,x,maxiter);
```

4.3 Algoritmi de ordinul al doilea

Metodele de ordinul al doilea (adică având oracol de ordinul al doilea) sunt, în comparaţie cu metodele de ordin zero sau unu, mai costisitoare din punct de vedere computaţional, dar au, în general, o convergenţă mai bună. Cea mai reprezentativă metodă de acest tip este metoda lui Newton adaptată problemelor de optimizare pe care o detaliem mai jos. Există apoi unele metode care combină această metodă cu unele tehnici pe care le-am studiat anterior.

4.3.1 Metoda lui Newton în optimizare

Exerciţiul 4.48 *Pentru a reaminti metoda lui Newton, să se implementeze pentru rezolvarea aproximativă a sistemului*

$$\begin{cases} 3x^2y + y^2 = 1 \\ x^4 + xy^3 = 1. \end{cases}$$

Să se testeze cu datele iniţiale $(1,1)$, $(-1,-1)$, $(-1,1)$.

Aşa cum am arătat pe scurt în Observaţia 3.6.5, metoda lui Newton poate fi folosită şi la probleme de optimizare, fiind în acest context o metodă cu oracol de ordinul al doilea. Aplicată

ecuaţiei $\nabla f(x) = 0$, metoda lui Newton returnează o aproximare a unui punct critic. Ca de obicei, în condiţii de convexitate obţinem convergenţa către un punct de minim global. Pentru o funcţie $f : \mathbb{R}^p \to \mathbb{R}$ iteraţia devine

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

unde $H_f(x)$ notează matricea Hessiană a lui f în x.

Metoda lui Newton poate fi rezumată astfel:

Pasul 1: se dau x_0 (punctul de start), $\varepsilon_1, \varepsilon_2 > 0$ (toleranţe), $\Delta x > 0$ (pentru calculul gradientului);

Pasul al 2-lea: fiind calculat x_k, se calculează $\nabla f(x_k), H_f(x_k)$, gradientul şi Hessiana în x_k;

$p_k = -H_f^{-1}(x_k) \nabla f(x_k)$ (direcţia de căutare);

se actualizează $x_{k+1} = x_k + p_k$;

dacă $\|x_{k+1} - x_k\| > \varepsilon_1$ sau $\|\nabla f(x_{k+1})\| > \varepsilon_2$

atunci se merge la Pasul 2

altfel se merge la Pasul 3

Pasul al 3-lea: metoda converge. Se afişează x_{k+1} şi $f(x_{k+1})$.

Prezentăm două modele de aplicare a acestei metode.

Exerciţiul 4.49 *Să se implementeze această metodă pentru funcţiile test deja utilizate.*

Implementare 4.50 Putem folosi de asemenea un program care să calculeze (cu aproximaţie) hessiana, aşa încât să nu fim nevoiţi să calculăm prin derivare directă această matrice.

```
function sec_deriv = hessian(x,delx)
for i = 1:length(x)
 for j = 1:length(x)
 if i == j
 temp = x;
 temp(i) = x(i) + delx;
 term1 = func_multivar(temp);
 temp = x;
 temp(i) = x(i) - delx;
 term2 = func_multivar(temp);
 temp = x;
 term3 = func_multivar(x);
 sec_deriv(i,j) = (term1-2*term3+term2)/(delx^2);
 else
 temp = x;
 temp(i) = x(i) + delx;
 temp(j) = x(j) + delx;
 term1 = func_multivar(temp);

 temp = x;
 temp(i) = x(i) + delx;
 temp(j) = x(j) - delx;
```

```
    term2 = func_multivar(temp);

    temp = x;
    temp(i) = x(i) - delx;
    temp(j) = x(j) + delx;
    term3 = func_multivar(temp);

    temp = x;
    temp(i) = x(i) - delx;
    temp(j) = x(j) - delx;
    term4 = func_multivar(temp);
    sec_deriv(i,j) = (term1-term2-term3+term4)/(4*delx^2);
    end
    end
  end
```

Exerciţiul 4.51 *Folosind această funcţie, să se implementeze din nou metoda lui Newton (a se vedea* funct_multivar *şi* grad_vec *de mai sus) pentru funcţiile test deja considerate.*

4.3.2 Metode derivate din metodele fundamentale

Metoda căutării direcţiei de descreştere, metoda gradienţilor conjugaţi şi metoda lui Newton sunt metode fundamentale în optimizarea fără restricţii. Pe baza lor au fost derivate şi alte metode care fie sunt hibride, în sensul că prin construcţia lor combină caracteristici ale celor trei metode de bază, fie sunt aproximative, în sensul că oferă posibilităţi de evitare a calculului elementelor celor mai costisitoare din punct de vedere computaţional ale acestora (ca de exemplu, calculul Hessianei în cazul metodei lui Newton).

Prezentăm mai jos, fără demonstraţii privind convergenţa, principalele metode derivate. În orice caz, fiind vorba de variaţii ale metodelor de bază, considerăm că, prin însăşi natura lor, aceşti algoritmi pot fi înţeleşi, acceptaţi şi folosiţi cu uşurinţă.

1. Metoda lui Newton modificată

Pasul 1: se dau x_0 (punctul de start), $\varepsilon_1, \varepsilon_2 > 0$ (toleranţe), $\Delta x > 0$ (pentru calculul gradientului);

Pasul al 2-lea: fiind calculat x_k, se calculează $\nabla f(x_k)$, $H_f(x_k)$, gradientul şi Hessiana în x_k;

$p_k = -H_f^{-1}(x_k)\nabla f(x_k)$ (direcţia de căutare);

se minimizează $\alpha \to f(x_k + \alpha p_k)$ (utilizând, de exemplu, metoda secţiunii de aur) şi se obţine α;

se actualizează $x_{k+1} = x_k + \alpha p_k$;

dacă $\|x_{k+1} - x_k\| > \varepsilon_1$ sau $\|\nabla f(x_{k+1})\| > \varepsilon_2$

atunci se merge la Pasul 2

altfel se merge la Pasul 3

Pasul al 3-lea: metoda converge. Se afişează x_{k+1} şi $f(x_{k+1})$.

În unele situaţii această metodă este mai rapidă decât metoda lui Newton. Spre comparaţie, această metodă mai caută un pas α suplimentar faţă de metoda lui Newton, deci combină metoda clasică cu metoda căutării direcţiei de descreştere.

Exerciţiul 4.52 *Să se exemplifice această metodă folosindu-se funcţiile Matlab/Octave deja prezentate.*

2. Metoda Levenberg–Marquardt

Pasul 1: se dau x_0 (punctul de start), $\varepsilon_1, \varepsilon_2 > 0$ (toleranţe), $\Delta x > 0$ (pentru calculul gradientului), $\lambda > 0$ (o valoare mare);

Pasul al 2-lea: fiind calculat x_k, se calculează $\nabla f(x_k)$, $H_f(x_k)$, gradientul şi Hessiana în x_k;

$p_k = -\left[H_f(x_k) + \lambda I\right]^{-1} \nabla f(x_k)$ (direcţia de căutare);

se actualizează $x_{k+1} = x_k + p_k$;

dacă $f(x_{k+1}) < f(x_k)$

atunci λ devine $\lambda/2$

altfel λ devine 2λ

dacă $\|x_{k+1} - x_k\| > \varepsilon_1$ sau $\|\nabla f(x_{k+1})\| > \varepsilon_2$

atunci se merge la Pasul 2

altfel se merge la Pasul 3

Pasul al 3-lea: metoda converge. Se afişează x_{k+1} şi $f(x_{k+1})$.

Se poate observa că şi această metodă este una hibrid între metoda lui Newton şi metoda căutării direcţiei de descreştere.

Exerciţiul 4.53 *Să se exemplifice această metodă folosindu-se funcţiile Matlab/Octave deja prezentate.*

3. Metoda Fletcher–Reeves a gradienţilor conjugaţi

Pasul 1: se dau x_0 (punctul de start), $\varepsilon_1, \varepsilon_2 > 0$ (toleranţe), $\Delta x > 0$ (pentru calculul gradientului);

se minimizează $\alpha \to f(x_0 - \alpha \nabla f(x_0))$ (utilizând, de exemplu, metoda secţiunii de aur) şi se obţine α;

se calculează $x_1 = x_0 - \alpha \nabla f(x_0)$

Pasul al 2-lea: fiind calculate valorile lui x_i până la rangul k, se procedează astfel:

se calculează $p_k = -\nabla f(x_k) + \frac{\|\nabla f(x_k)\|^2}{\|\nabla f(x_{k-1})\|^2} (-\nabla f(x_{k-1}))$

se minimizează $\alpha \to f(x_k + \alpha p_k)$ (utilizând, de exemplu, metoda secţiunii de aur) şi se obţine α;

se actualizează $x_{k+1} = x_k + \alpha p_k$;

dacă $\|x_{k+1} - x_k\| > \varepsilon_1$ sau $\|\nabla f(x_{k+1})\| > \varepsilon_2$

atunci se merge la Pasul 2

altfel se merge la Pasul 3

Pasul al 3-lea: metoda converge. Se afişează x_{k+1} şi $f(x_{k+1})$.

Această metodă este o extindere la funcţii nu neaparat pătratice a metodei gradienţilor conjugaţi. În particular, este o metodă de ordinul întâi.

Exerciţiul 4.54 *Să se exemplifice această metodă folosindu-se funcţiile Matlab/Octave deja prezentate.*

4. Metoda Davidon–Fletcher–Powell (Metoda DFP)

Ideea acestei metode este de a evita calculul Hessianei în metoda lui Newton, folosind în schimb o aproximare a inversei acesteia. Este o metodă de ordinul întâi cu convergenţă pătratică, mai fiind numită şi metoda quasi-Newton.

Aproximarea inversei Hessianei în punctul curent se face printr-un şir de matrici (A_k), cu primul termen identitatea care respectă recurenţa

$$A_{k+1} = A_k + \frac{(x_{k+1} - x_k)^t (x_{k+1} - x_k)}{(x_{k+1} - x_k)(\nabla f(x_{k+1}) - \nabla f(x_k))^t} - \frac{A_k (\nabla f(x_{k+1}) - \nabla f(x_k))^t (\nabla f(x_{k+1}) - \nabla f(x_k)) A_k}{(\nabla f(x_{k+1}) - \nabla f(x_k)) A_k (\nabla f(x_{k+1}) - \nabla f(x_k))^t}.$$

Metoda poate fi descrisă astfel:

Pasul 1: se dau x_0 (punctul de start), $\varepsilon_1, \varepsilon_2 > 0$ (toleranţe), $\Delta x > 0$ (pentru calculul gradientului), matricea A iniţializată cu identitatea;

se minimizează $\alpha \to f(x_0 - \alpha \nabla f(x_0))$ (utilizând, de exemplu, metoda secţiunii de aur) şi se obţine α;

se calculează $x_1 = x_0 - \alpha \nabla f(x_0)$

Pasul al 2-lea: Fiind calculate valorile lui x_i până la rangul k şi ale lui A_j până la rangul $k-1$, se procedează astfel:

se calculează A_k după formula de mai sus; se calculează $p_k = -A_k \nabla f(x_k)$;

se minimizează $\alpha \to f(x_k + \alpha p_k)$ (utilizând, de exemplu, metoda secţiunii de aur) şi se obţine α;

se actualizează $x_{k+1} = x_k + \alpha p_k$;

dacă $\|x_{k+1} - x_k\| > \varepsilon_1$ sau $\|\nabla f(x_{k+1})\| > \varepsilon_2$

atunci se merge la Pasul 2

altfel se merge la Pasul 3

Pasul al 3-lea: metoda converge. Se afişează x_{k+1} şi $f(x_{k+1})$.

Exerciţiul 4.55 *Să se exemplifice această metodă folosindu-se funcţiile Matlab/Octave deja prezentate.*

5. Metoda Broyden–Fletcher–Goldfarb–Shanno (Metoda BFGS)

Din nou, ideea este de a evita calculul Hessianei în metoda lui Newton, folosind în schimb o aproximare a acesteia. Deci de data aceasta se aproximează Hessiana (nu inversa sa) în punctul curent prin şirul de matrici (A_k) cu primul termen identitatea care respectă recurenţa

$$A_{k+1} = A_k + \frac{(\nabla f(x_{k+1}) - \nabla f(x_k))^t (\nabla f(x_{k+1}) - \nabla f(x_k))}{(\nabla f(x_{k+1}) - \nabla f(x_k))(x_{k+1} - x_k)^t} - \frac{\nabla f(x_k)^t \nabla f(x_k)}{\nabla f(x_k) \nabla f(x_{k+1})^t},$$

unde p_k este definit în descrierea de mai jos a metodei:

Pasul 1: se dau x_0 (punctul de start), $\varepsilon_1, \varepsilon_2 > 0$ (toleranţe), $\Delta x > 0$ (pentru calculul gradientului), matricea A iniţializată cu identitatea;

se minimizează $\alpha \to f(x_0 - \alpha \nabla f(x_0))$ (utilizând, de exemplu, metoda secţiunii de aur) şi se obţine α;

se calculează $x_1 = x_0 - \alpha \nabla f(x_0)$

Pasul al 2-lea: Fiind calculate valorile lui x_i până la rangul k şi ale lui A_j până la rangul $k-1$, se procedează astfel:

se calculează A_k după formula de mai sus; se calculează $p_k = -A_k^{-1} \nabla f(x_k)$;

se minimizează $\alpha \to f(x_k + \alpha p_k)$ (utilizând, de exemplu, metoda secţiunii de aur) şi se obţine α;

se actualizează $x_{k+1} = x_k + \alpha p_k$;

dacă $\|x_{k+1} - x_k\| > \varepsilon_1$ sau $\|\nabla f(x_{k+1})\| > \varepsilon_2$

atunci se merge la Pasul 2

altfel se merge la Pasul 3

Pasul al 3-lea: metoda converge. Se afişează x_{k+1} şi $f(x_{k+1})$.

Exerciţiul 4.56 *Să se exemplifice această metodă folosindu-se funcţiile Matlab/Octave deja prezentate.*

Exerciţiul 4.57 *Să se testeze şi să se compare metodele pe funcţia Rosenbrock.*

Exerciţiul 4.58 *Fie* $f : \mathbb{R}^2 \to \mathbb{R}$ *dată prin*

$$f(x, y) = e^{x^2+y^2} + x + y - 3 - \sin(3(x + y)).$$

Să se arate deseneze graficul şi liniile de contur pe $[-1, 1]^2$ *şi să se deducă faptul că* f *are două puncte de minim local. Să se verifice că* $(0, 0)$ *nu este un astfel de punct. Să se folosească metodele anterioare pentru a aproxima minimele.*

Exerciţiul 4.59 *Fie sistemul*

$$\begin{cases} 3x - \cos yz - \frac{3}{2} = 0 \\ 3x^2 - 625y^2 + 2y - 1 = 0 \\ e^{-xy} + 20z + 9.1 = 0. \end{cases}$$

Să se aproximeze o soluţie a acestui sistem folosind metodele anterioare pentru funcţia $f : \mathbb{R}^3 \to \mathbb{R}$,

$$f(x, y, z) = \left(3x - \cos yz - \frac{3}{2}\right)^2 + \left(3x^2 - 625y^2 + 2y - 1\right)^2 + \left(e^{-xy} + 20z + 9.1\right)^2.$$

Exerciţiul 4.60 *Fie sistemul (supradeterminat)*

$$\begin{cases} x^2 - y^2 + x - 3y = 2 \\ x^3 - y^4 = -2 \\ x^2 + y^3 + 2x - y = -1.1. \end{cases}$$

Să se aproximeze soluţii a acestui sistem folosind metodele anterioare pentru funcţia construită ca la exerciţiul anterior pentru datele iniţiale $(1.5, -1.75)$, $(1.5, 1.5)$.

Exerciţiul 4.61 *Să se compare metodele studiate pe cazul minimizării funcţiei lui Himmelblau,* $f : \mathbb{R}^2 \to \mathbb{R}$ *date prin*

$$f(x) = \left(x_1^2 + x_2 - 11\right)^2 + \left(x_2^2 + x_1 - 7\right)^2$$

luând diverse date iniţiale (se va reprezenta grafic şi se va constata că funcţia are patru puncte de minim local).

Exerciţiul 4.62 *Să se verifice pe cazul funcţiei* $f : \mathbb{R}^2 \to \mathbb{R}$ *date prin*

$$f(x) = x_1^2 + 3x_1x_2 + 5x_2^2$$

luând data iniţială $x = (1, 1)$ *că*

(i) iteraţia pentru generarea matricilor $(A_k)_k$ *din metoda DFP conduce la o aproximare a inversei hessianei funcţiei;*

(ii) iteraţia pentru generarea matricilor $(A_k)_k$ *din metoda BFGS conduce la o aproximare a hessianei funcţiei.*

Exerciţiul 4.63 *Fie* $f : \mathbb{R}^4 \to \mathbb{R}$,

$$f(x) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^2 + 100(x_1 - x_4)^4.$$

Verificaţi că $0 \in \mathbb{R}^4$ *este singurul punct de minim global şi că nu există alte puncte de minim local. Testaţi pe funcţia* f

(i) metoda celei mai bune descreşteri cu datele iniţiale $(-2, -1, 1, 2)$, $(200, -200, 100, -100)$;

(ii) metodele derivate din metoda lui Newton.

Exerciţiul 4.64 *Să presupunem că din observarea unui fenomen fizic la momentele* t_i *rezultă datele* m_i*, conform tabelului următor:*

i	1	2	3	4	5	6
t_i	0.2	0.4	0.6	0.8	0.9	0.95
m_i	0.7123	1.754	4.852	22.27	94.91	388.2

.

În plus, presupunem că se observă un comportament de tip t^{x_1} *în jurul lui* 0 *şi un comportament de tip* $\frac{1}{(1-t)^{x_2}}$ *în jurul lui* 1*. Legea sugerată de comportare la fiecare moment* t *depinde de trei parametri* x_1, x_2, x_3 *şi este modelată de funcţia* $f : \mathbb{R}^3 \to \mathbb{R}$,

$$f(x) = x_3 t^{x_1} \frac{1}{(1-t)^{x_2}}.$$

Funcţia de minimizat pentru a găsi modelul cel mai apropiat de aceste date este, conform metodei celor mai mici pătrate (a se vedea şi Problema 2.22),

$$x \mapsto \sum_{i=1}^{6} \left[m_i - x_3 t_i^{x_1} \frac{1}{(1-t_i)^{x_2}} \right]^2.$$

Determinaţi parametrii modelului şi valoarea totală a pătratelor abaterilor minimizând funcţia de mai sus, prin diverse metode.

Exerciţiul 4.65 *Să se testeze metodele studiate pe următoarele funcţii test:*
(i) funcţia Beale, $f : \mathbb{R}^2 \to \mathbb{R}$ *dată prin*

$$f(x) = (1.5 - x_1 + x_1x_2)^2 + \left(2.25 - x_1 + x_1x_2^2\right)^2 + \left(2.625 - x_1 + x_1x_2^3\right)^2,$$

pentru care punctul de minim global este $(3, 0.5)$, *iar valoarea minimă globală este* 0;
(ii) funcţia „cămilă cu trei cocoaşe", $f : \mathbb{R}^2 \to \mathbb{R}$ *dată prin*

$$f(x) = 2x_1^2 - 1.05x_1^4 + \frac{x_1^6}{6} + x_1x_2 + x_2^2,$$

pentru care punctul de minim global este $(0, 0)$, *iar valoarea minimă globală este* 0;
(iii) funcţia Branin, $f : \mathbb{R}^2 \to \mathbb{R}$ *dată prin*

$$f(x) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10,$$

pentru care, printre minimele globale se regăsesc punctele $(-\pi, 12.275)$, $(\pi, 2.275)$, $(9.42478, 2.475)$, *iar valoarea minimă globală este* 0.397887.

4.4 Cazul funcţiilor cu valori vectoriale: optimizare Pareto

Optimizarea Pareto (după numele economistului italian W. F. Pareto) se ocupă de situaţia în care minimizarea (sau maximizarea) se face după mai multe criterii, unele fiind în concurenţă cu altele. Formalizarea matematică a acestei idei este prezentată succint mai jos.

Ca de obicei, considerăm $p \in \mathbb{N} \setminus \{0\}$.

Definiţia 4.4.1 *O submulţime nevidă* $C \subset \mathbb{R}^p$ *se numeşte con dacă are loc proprietatea*

$$\forall x \in C, \forall \lambda \in \mathbb{R}_+ := [0, \infty) : \lambda x \in C.$$

Conform acestei definiţii, orice con conţine elementul 0. Dacă $A \subset \mathbb{R}^p$ este o mulţime nevidă, atunci mulţimea

$$\mathbb{R}_+A := \{\lambda a \mid \lambda \in \mathbb{R}_+, a \in C\}$$

este con, numit conul generat de A.

Propoziţia 4.4.2 *Un con* C *este mulţime convexă dacă şi numai dacă* $C + C \subset C$.

Prin urmare, un con convex este o mulţime nevidă închisă la adunare şi la înmulţirea cu scalari pozitivi.

Definiţia 4.4.3 *(i) Un con* C *se numeşte propriu dacă* $C \neq \{0\}$ *şi* $C \neq \mathbb{R}^p$.
(ii) Un con C *se numeşte punctat dacă* $C \cap (-C) = \{0\}$.

Următorul rezultat arată că un con introduce pe $\mathbb{R}^p$ o relaţie de ordine (parţială) compatibilă cu structura de spaţiu liniar.

Teorema 4.4.4 *Fie $\mathbb{R}^p$ un spaţiu liniar şi $C \subset \mathbb{R}^p$ un con. Atunci relaţia*

$$\mathcal{R}_C := \{(x_1, x_2) \in \mathbb{R}^p \times \mathbb{R}^p \mid x_2 - x_1 \in C\}$$

este reflexivă şi

$$\forall x_1, x_2 \in \mathbb{R}^p, \forall \lambda \in \mathbb{R}_+,\ x_1 \mathcal{R}_C x_2 \Rightarrow \lambda x_1 \mathcal{R}_C \lambda x_2 \tag{4.11}$$

$$\forall x_1, x_2, x \in \mathbb{R}^p,\ x_1 \mathcal{R}_C x_2 \Rightarrow (x_1 + x)\mathcal{R}_C(x_2 + x). \tag{4.12}$$

Mai mult, C este convex dacă şi numai dacă $\mathcal{R}_C$ este tranzitivă şi, respectiv, C este punctat dacă şi numai dacă $\mathcal{R}_C$ este antisimetrică.

Invers, dacă $\mathcal{R}$ este o relaţie reflexivă pe $\mathbb{R}^p$ care satisface (4.11) şi (4.12), atunci $C := \{x \in \mathbb{R}^p \mid 0\mathcal{R}x\}$ este con şi $\mathcal{R} = \mathcal{R}_C$.

Demonstraţie. Toate afirmaţiile se verifică direct, pe baza definiţiilor. □

Observaţia 4.4.5 *Pentru $p = 1$ şi $C = \mathbb{R}_+$, $\mathcal{R}_C$ coincide cu ordinea naturală pe $\mathbb{R}$.*

Având în vedere aceasta teoremă, dacă C este con convex punctat vom scrie $x_1 \leq_C x_2$ în loc de $x_1 \mathcal{R}_C x_2$.

Exerciţiul 4.66 *Fie $C \subset \mathbb{R}^p$ un con convex închis punctat cu interior nevid. Să se arate că:*
(i) $C + \operatorname{int} C = \operatorname{int} C$;
(ii) $\mathbb{R}^p \setminus -\operatorname{int} C + C = \mathbb{R}^p \setminus -\operatorname{int} C$ şi $(\mathbb{R}^p \setminus -C) \cup \{0\} + C = (\mathbb{R}^p \setminus -C) \cup \{0\}$.

Exerciţiul 4.67 *Fie $C \subset \mathbb{R}^p$ un con închis convex cu interior nevid care nu coincide cu $\mathbb{R}^p$ şi $e \in \operatorname{int} K$. Să se arate că:*
(i) $C + [0, \infty)e \subset C$;
(ii) $C + (0, \infty)e = \operatorname{int} C$;
(iii) $\mathbb{R}e - C = \mathbb{R}^p$;
(iv) pentru orice $x \in \mathbb{R}^p$, $x + \mathbb{R}e \not\subset C$.

Definiţia 4.4.6 *Considerăm $\mathbb{R}^p$ ca fiind ordonat cu ordinea parţială indusă de un con convex, punctat, propriu C şi $A \subset \mathbb{R}^p$ o mulţime nevidă.*

(i) Un punct $\overline{x} \in A$ se numeşte minim tare al mulţimii A (în raport cu conul C) dacă $\overline{x} \leq_C x$ pentru orice $x \in A$, sau, echivalent, $A \subset \overline{x} + C$.

(ii) Un punct $\overline{x} \in A$ se numeşte minim Pareto al mulţimii A (în raport cu conul C) dacă $x \in A$, $x \leq_C \overline{x}$ implică $x = \overline{x}$, sau, echivalent, $A \cap (\overline{x} - C) = \{\overline{x}\}$.

(iii) Presupunem că C are interior nevid. Un punct $\overline{x} \in A$ se numeşte minim slab al mulţimii A (în raport cu conul C) dacă $A \cap (\overline{x} - \operatorname{int} C) = \emptyset$.

Observaţia 4.4.7 *Este evident că definiţia minimului Pareto este echivalentă cu oricare dintre relaţiile*

$$(A - \overline{x}) \cap (-C) = \{0\},$$
$$A \cap (\overline{x} - C) = \{\overline{x}\}$$

şi

$$(A + C) \cap (\overline{x} - C) = \{\overline{x}\}.$$

Având în vedere incluziunea $C + \operatorname{int} C \subset \operatorname{int} C$, *definiţia minimului slab este echivalentă cu*

$$(A - \overline{x}) \cap (-\operatorname{int} C) = \emptyset,$$

$$A \cap (\overline{x} - \operatorname{int} C) = \emptyset,$$

$$(A + C) \cap (\overline{x} - \operatorname{int} C) = \emptyset$$

şi cu

$$\overline{x} \notin A + \operatorname{int} C.$$

Mulţimea punctelor de minim tare (minim Pareto, minim slab) ale mulţimii A va fi notată cu $IMin(A, C)$ (respectiv, $Min(A, C)$, $WMin(A, C)$). Evident, dacă $\operatorname{int} C \neq \emptyset$, $IMin(A, C) \subset Min(A, C) \subset WMin(A, C)$.

Observaţia 4.4.8 *Pentru* $p = 1$ *şi* $C = \mathbb{R}_+$, *toate noţiunile de mai sus coincid cu minimalitatea scalară standard.*

Exerciţiul 4.68 *Să se ilustreze minimalitatea Pareto, căutând mulţimea punctelor optimale Pareto în raport cu* $\mathbb{R}^2_+$ *pentru un grup de puncte generate aleator în primul cadran.*

Exerciţiul 4.69 *Să se genereze* 50 *puncte echidistante pe un cerc şi să se identifice dintre aceste punctele optimale în sens Pareto (din nou, în raport cu* $\mathbb{R}^2_+$*).*

Exerciţiul 4.70 *Fie* $C \subset \mathbb{R}^p$ *un con închis convex cu interior nevid şi* $\emptyset \neq A \subset \mathbb{R}^p$. *Să se arate că următoarele afirmaţii sunt echivalente:*
(i) $A \cap -\operatorname{int} C = \emptyset$;
(ii) $\operatorname{cl} A \cap -\operatorname{int} C = \emptyset$;
(iii) $(A + C) \cap -\operatorname{int} C = \emptyset$;
(iv) $\operatorname{cl}(\operatorname{con}(A + C)) \cap -\operatorname{int} C = \emptyset$.

Definiţia 4.4.9 *Fie* $f : \mathbb{R}^n \to \mathbb{R}^p$ *unde* $n, p \in \mathbb{N} \setminus \{0\}$, $C \subset \mathbb{R}^p$ *un con convex închis punctat popriu şi* $A \subset \mathbb{R}^n$ *o mulţime închisă. Spunem că* $\overline{x} \in A$ *este minim (slab) Pareto pentru* f *pe* A *dacă* $f(\overline{x})$ *este minim (slab) Pareto al mulţimii* $f(A)$.

Următorul rezultat prezintă o situaţie în care găsirea unui minim Pareto pentru o funcţie vectorială se reduce la găsirea unui minim pentru o funcţie scalară.

Propoziţia 4.4.10 *Cu notaţiile din Definiţia 4.4.9, considerăm* $C = \mathbb{R}^p_+$, $w \in \operatorname{int} \mathbb{R}^p_+$ *cu* $\sum_{i=1}^p w_i = 1$. *Fie* $g : \mathbb{R}^n \to \mathbb{R}$,

$$g(x) = \sum_{i=1}^{p} w_i f_i(x).$$

Dacă $\overline{x} \in A$ *este punct de minim global pe* A *pentru* g, *atunci* $\overline{x}$ *este minim Pareto pentru* f *pe* A.

Demonstraţie. Presupunem, prin reducere la absurd, că $\overline{x}$ nu este minim Pareto pentru f pe A. Atunci există $x \in A$ astfel încât pentru orice $i \in \overline{1,n}$, $f_i(x) \leq f_i(\overline{x})$ şi există $j \in \overline{1,n}$ astfel încât $f_j(x) < f_j(\overline{x})$. Înmulţind aceste inegalităţi cu scalarii $w_i > 0$ corespunzători şi adunându-le, obţinem $g(x) < g(\overline{x})$, ceea ce reprezintă o contradicţie. □

Exerciţiul 4.71 *Fie* $f : \mathbb{R}^2 \to \mathbb{R}^2$,

$$f(x) = \left(x_1^2 + x_2^2, (x_1 - 1)^2 + (x_2 - 3)^2\right).$$

Să se scrie un program care să determine puncte de minim Pareto pentru f *(pe* $\mathbb{R}^2$*) prin metoda descrisă în Propoziţia 4.4.10 (de exemplu, prin modificarea corespunzătoare a metodei DFP).*

4.5 Obţinerea unor algoritmi pentru probleme cu restricţii - o metodă de penalizare

În cazul problemelor cu restricţii, cei mai importanţi algoritmi derivă din metodele studiate pentru probleme fără restricţii pe baza unor rezultate teoretice care transformă problemele din prima categorie în probleme din a doua categorie.

Vom considera două probleme concrete pentru ilustrarea teoriei.

Mai întâi, fie $f, g_1, g_2 : \mathbb{R}^2 \to \mathbb{R}$ date respectiv prin

$$\begin{aligned} f(x) &= (x_1 - 1)^2 + (x_2 - 5)^2 \\ g_1(x) &= -x_1^2 + x_2 - 4 \\ g_2(x) &= -(x_1 - 2)^2 + x_2 - 3. \end{aligned}$$

Studiem problema

$$(P_1) \ \min f(x) \ \text{ cu } g_1(x) \leq 0, g_2(x) \leq 0.$$

Este evident că problema admite soluţie. O analiză geometrică, combinată cu una analitică, arată că soluţia este

$$x = (x_1, x_2) = \left(\frac{3}{4}, 4 + \left(\frac{3}{4}\right)^2\right) = (0.75, 4.5625).$$

Apoi, considerăm problema (P_2) a minimizării lui $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2 - 8x + 3$ cu restricţia $g(x) \leq 0$, unde $g : \mathbb{R} \to \mathbb{R}$, $g(x) = x - 1$. Evident, soluţia este $x = 1$.

Evident, există şi algoritmi direcţi cum ar fi algoritmul grilei uniforme care pot fi adaptaţi foarte uşor.

Exerciţiul 4.72 *Să se implementeze un algoritm de tipul grilei uniforme adaptat pentru problema* (P_1).

Vom constata că metodele de aproximare a soluţiilor problemelor cu restricţii se reduc, prin diverse procedee, la minimizarea unor familii de funcţii fără restricţii pentru care se pot aplica algoritmii studiaţi anterior.

O metodă mai elaborată pe care o descriem este bazată pe o tehnică numită penalizare. Ideea este de a îngloba restricţiile problemei într-o funcţie, numită termen de penalizare, astfel încât rezolvarea problemei fără restricţii care constă din minimizarea funcţiei obiectiv adunată cu acest termen să conducă la aproximarea soluţiei problemei iniţiale. De regulă, trebuie considerată o familie parametrică de termeni de penalizare pentru a obţine o aproximare bună a soluţiei pe care o căutăm. Prezentăm în continuare aspectele matematice cele mai relevante ale acestei abordări.

Considerăm problema generală cu restricţii

$$\min f(x) \text{ cu } x \in M,$$

unde $f : \mathbb{R}^p \to \mathbb{R}$ este o funcţie netedă şi $M \subset \mathbb{R}^p$ este o mulţime închisă. Dorim să aproximăm soluţiile acestei probleme aceasta prin intermediul soluţiilor problemelor fără restricţii de forma

$$\min \left(f(x) + \gamma P(x) \right),$$

unde $\gamma > 0$ şi $P : \mathbb{R}^p \to \mathbb{R}$ este o funcţie. Constanta γ se numeşte parametru de penalizare, iar funcţia P se numeşte funcţie de penalizare, adică o funcţie care satisface definiţia de mai jos.

Definiţia 4.5.1 *O funcţie $P : \mathbb{R}^p \to \mathbb{R}$ se numeşte funcţie de penalizare pentru problema generală cu restricţii dacă este continuă, cu valori nenegative şi*

$$P(x) = 0 \iff x \in M.$$

Ideea este că o astfel de funcţie "penalizează" punctele din afara mulţimii M astfel încât acestea nu pot fi soluţii pentru problema fără restricţii asociată mai sus.

Exemplul 4.5.2 Fie

$$M = \{x \in \mathbb{R}^p \mid g(x) \leq 0\}$$

unde $g : \mathbb{R}^p \to \mathbb{R}^n$ este o funcţie diferenţiabilă, iar inegalitatea este înţeleasă pe coordonate. Atunci funcţia $P : \mathbb{R}^p \to \mathbb{R}$ dată prin

$$P(x) = \sum_{i=1}^{n} \max \{0, g_i(x)\}$$

este o funcţie de penalizare. Evident, această funcţie este una nediferenţiabilă, în general.

O altă posibilitate este să alegem $P : \mathbb{R}^p \to \mathbb{R}$ dată prin

$$P(x) = \sum_{i=1}^{n} \left(\max \{0, g_i(x)\} \right)^2,$$

iar aceasta este o funcţie de penalizare diferenţiabilă.

Exemplul 4.5.3 Fie

$$M = \{x \in \mathbb{R}^p \mid h(x) = 0\}$$

unde $h : \mathbb{R}^p \to \mathbb{R}^m$ este o funcţie diferenţiabilă. Atunci funcţia $P : \mathbb{R}^p \to \mathbb{R}$ dată prin

$$P(x) = \sum_{i=1}^{n} h_i^2(x)$$

este o funcţie de penalizare.

Exerciţiul 4.73 *Să se scrie prima variantă a funcţiei de penalizare din exemplul anterior pentru* $g : \mathbb{R} \to \mathbb{R}^2$,

$$g(x) = \left(x - 2, -(x+1)^3\right).$$

Studiaţi diferenţiabilitatea şi graficul funcţiei de penalizare.

Exemplul 4.5.4 Fie Q o matrice pătratică simetrică de dimensiune p. Considerăm aplicaţia $f : \mathbb{R}^p \to \mathbb{R}$,

$$f(x) = \langle Qx, x\rangle$$

şi problema minimizării şi maximizării acesteia pe sfera unitate din $\mathbb{R}^p$, adică pe mulţimea

$$\{x \in \mathbb{R}^p \mid \|x\| = 1\}.$$

Evident, în ambele cazuri există soluţie, sfera fiind compactă.

Să remarcăm că într-un punct x de pe sferă conul tangent Bouligand la acesta este mulţimea

$$\{u \in \mathbb{R}^p \mid \langle x, u\rangle = 0\}$$

(Teorema lui Lyusternik), deci conul normal corespunzător este

$$\{\alpha x \mid x \in \mathbb{R}\}.$$

Atunci, conform teoremei care dă condiţii necesare de optimalitate de ordinul I, un punct $\overline{x}$ de pe sferă este soluţie a problemei de minimizare sau de maximizare dacă există $\mu \in \mathbb{R}$ cu

$$Q\overline{x} = \mu\overline{x},$$

adică

$$\mu = \langle Q\overline{x}, \overline{x}\rangle.$$

(Aceeaşi concluzie rezultă prin aplicarea Teoremei Karush-Kuhn-Tucker.) Deducem de aici că punctele ce satisfac condiţiile de optimalitate pentru problemele noastre corespund vectorilor proprii de normă 1 ai lui Q, iar valorile funcţiei f în acele puncte sunt valorile proprii ale lui Q. Concluzionăm că cea mai mare valoare proprie a lui f este $\lambda_{\max} = \max_{\|x\|=1} \langle Qx, x\rangle$, în timp ce cea mai mică este $\lambda_{\min} = \min_{\|x\|=1} \langle Qx, x\rangle$. Prin normalizare se obţine

$$\lambda_{\max} = \max_{x \in \mathbb{R}^p \setminus \{0\}} \frac{\langle Qx, x\rangle}{\|x\|^2} \text{ şi } \lambda_{\min} = \min_{x \in \mathbb{R}^p \setminus \{0\}} \frac{\langle Qx, x\rangle}{\|x\|^2}.$$

Să considerăm acum pentru problema de mai sus funcţia de penalizare $P(x) = \left(\|x\|^2 - 1\right)^2$ şi un parametru $\gamma > 0$. Conform ideii generale, problema fără restricţii asociată este

$$\min\left(\langle Qx, x\rangle + \gamma\left(\|x\|^2 - 1\right)^2\right).$$

O soluţie $\overline{x}_\gamma$ a acestei probleme trebuie să satisfacă, conform Teoremei lui Fermat, ecuaţia

$$2Q\overline{x}_\gamma + 4\gamma\left(\|\overline{x}_\gamma\|^2 - 1\right)\overline{x}_\gamma = 0,$$

adică

$$Q\overline{x}_\gamma = 2\gamma\left(1 - \|\overline{x}_\gamma\|^2\right)\overline{x}_\gamma,$$

ceea ce înseamnă în particular că $\overline{x}_\gamma$ este vector propriu al lui Q.

În plus, valoarea proprie asociată este $2\gamma\left(1 - \|\overline{x}_\gamma\|^2\right)$, deci

$$\lambda_{\min} \leq 2\gamma\left(1 - \|\overline{x}_\gamma\|^2\right) \leq \lambda_{\max}.$$

Astfel,

$$\frac{\lambda_{\min}}{2\gamma} \leq 1 - \|\overline{x}_\gamma\|^2 \leq \frac{\lambda_{\max}}{2\gamma}$$

şi pentru $\gamma \to \infty$, norma lui $\overline{x}_\gamma$ tinde la 1, deci $\overline{x}_\gamma$ tinde (măcar pe un subşir) la un punct fezabil al problemei iniţiale.

Lema 4.5.5 *Fie P o funcţie de penalizare pentru problema minimizării lui $f(x)$ cu $x \in M$. Fie $\overline{x}$ o soluţie globală a acestei probleme şi (γ_k) un şir strict crescător de numere reale pozitive. Pentru fiecare $k \in \mathbb{N}$ considerăm funcţia $q_k : \mathbb{R}^p \to \mathbb{R}$ dată prin*

$$q_k(x) = f(x) + \gamma_k P(x).$$

Presupunem că fiecare dintre aceste funcţii admite un punct de minim global x_k. Atunci, pentru orice k au loc următoarele:

(i) $q_{k+1}(x_{k+1}) \geq q_k(x_k)$;

(ii) $P(x_{k+1}) \leq P(x_k)$;

(iii) $f(x_{k+1}) \geq f(x_k)$;

(iv) $f(\overline{x}) \geq q_k(x_k) \geq f(x_k)$.

Demonstraţie. (i) Pe baza monotoniei lui (γ_k) avem

$$q_{k+1}(x_{k+1}) = f(x_{k+1}) + \gamma_{k+1}P(x_{k+1}) \geq f(x_{k+1}) + \gamma_k P(x_{k+1}).$$

Pe de altă parte, cum x_k e punct de minim global pentru q_k, avem

$$q_k(x_k) = f(x_k) + \gamma_k P(x_k) \leq f(x_{k+1}) + \gamma_k P(x_{k+1}).$$

Astfel obţinem prima concluzie.

(ii) Avem, ca mai sus,

$$q_k(x_k) = f(x_k) + \gamma_k P(x_k) \leq f(x_{k+1}) + \gamma_k P(x_{k+1})$$

şi

$$q_{k+1}(x_{k+1}) = f(x_{k+1}) + \gamma_{k+1}P(x_{k+1}) \leq f(x_k) + \gamma_{k+1}P(x_k),$$

inegalităţi care adunate conduc la

$$\gamma_k P(x_k) + \gamma_{k+1}P(x_{k+1}) \leq \gamma_k P(x_{k+1}) + \gamma_{k+1}P(x_k).$$

Deducem că

$$(\gamma_{k+1} - \gamma_k)P(x_{k+1}) \leq (\gamma_{k+1} - \gamma_k)P(x_k).$$

Cum $\gamma_{k+1} > \gamma_k$ obţinem că $P(x_{k+1}) \leq P(x_k)$.

(iii) Putem scrie din nou

$$q_k(x_k) = f(x_k) + \gamma_k P(x_k) \leq f(x_{k+1}) + \gamma_k P(x_{k+1}),$$

deci, folosind (ii),

$$f(x_{k+1}) \geq f(x_k) + \gamma_k (P(x_k) - P(x_{k+1})) \geq f(x_k).$$

(iv) Avem

$$f(\overline{x}) + \gamma_k P(\overline{x}) \geq q_k(x_k) = f(x_k) + \gamma_k P(x_k).$$

Dar $\overline{x}$ este punct fezabil pentru problema iniţială, deci $P(\overline{x}) = 0$. Astfel,

$$f(\overline{x}) \geq q_k(x_k) = f(x_k) + \gamma_k P(x_k) \geq f(x_k).$$

Demonstraţia este completă. □

Rezultatul următor este esenţial pentru fundamentarea metodei de penalizare.

Teorema 4.5.6 *Cu notaţiile de mai sus, în ipotezele Lemei 4.5.5, dacă $\gamma_k \to \infty$, atunci orice punct limită al şirului (x_k) este soluţie a problemei cu restricţii.*

Demonstraţie. Fie (x_{k_l}) un subşir al lui (x_k) convergent la un punct x^*. Conform Lemei 4.5.5, şirul $(q_k(x_k))$ este crescător şi mărginit superior de $f(\overline{x})$. Deci există limita reală $\overline{q}$ al lui $(q_k(x_k))$ şi $\overline{q} \leq f(\overline{x})$. Cum f este continuă, avem

$$\lim_l f(x_{k_l}) = f(x^*) \leq f(\overline{x}).$$

Dar şirul cu termenii

$$\gamma_{k_l} P(x_{k_l}) = q_{k_l}(x_{k_l}) - f(x_{k_l})$$

este convergent la $\overline{q} - f(x^*)$. Tot din lema precedentă, şirul $(P(x_k))$ este descrescător şi mărginit inferior de 0, deci este convergent. Cum $\gamma_{k_l} \to \infty$, deducem că $P(x_{k_l}) \to 0$, iar pe baza continuităţii lui P,

$$0 = \lim P(x_{k_l}) = P(\lim x_{k_l}) = P(x^*).$$

Deci x^* e punct fezabil pentru problema cu restricţii şi cum $f(x^*) \leq f(\overline{x})$, deducem că x^* este la rândul său o soluţie globală a acestei probleme. □

Observaţia 4.5.7 *Conform rezultatului precedent, pentru convergenţă trebuie să considerăm un şir de parametri $\gamma_k \to \infty$.*

Ideal ar fi să găsim o soluţie a problemei iniţiale minimizând doar o funcţie de tip

$$f(\cdot) + \gamma P(\cdot)$$

cu $\gamma > 0$, caz în care funcţia de penalizare se numeşte exactă. Cum vom vedea în exemplul următor, o astfel de funcţie nu poate fi diferenţiabilă, în general.

Exemplul 4.5.8 Fie problema minimizării funcţiei $f:\mathbb{R}\to\mathbb{R}$, $f(x)=5-3x$ pentru $x\in[0,1]$. Evident, $\overline{x}=1$ este soluţia problemei. Presupunem că utilizăm o funcţie de penalizare diferenţiabilă. Cum P este nulă pe $[0,1]$, $P'(1)=0$. Atunci, pentru orice $\gamma>0$,

$$(f+\gamma P)'(1)=-3\neq 0,$$

deci punctul $\overline{x}$ nu satisface condiţia necesară de minimalitate fără restricţii pentru $(f+\gamma P)$. Deci P nu este o funcţie de penalizare exactă.

Pe baza acestei teorii, putem prezenta o metodă de aproximare a soluţiilor unei probleme cu restricţii funcţionale de tipul

$$\min f(x)\ \text{ cu } g(x)\leq 0,\ h(x)=0,$$

unde $f:\mathbb{R}^p\to\mathbb{R}$, $g:\mathbb{R}^p\to\mathbb{R}^n$, $h:\mathbb{R}^p\to\mathbb{R}^m$, cu notaţiile standard.

Astfel, în notaţiile de mai sus, considerăm funcţia de penalizare

$$P(x)=\sum_{i=1}^{n}\left(\max\{0,g_i(x)\}\right)^2+\sum_{j=1}^{m}h_j^2(x)$$

şi şirul de parametri

$$\gamma_k=\max\left\{1,\frac{1}{\|\max\{0,g(x_{k-1})\},h(x_{k-1})\|}\right\},$$

unde $\max\{0,g(x_{k-1})\}$ e înţeles pe componente.

O altă posibilitate de alegere a lui γ_k este $10\log(k+1)$.

Concret, funcţia $f+\gamma_k P$ este la fiecare pas minimizată folosind o metodă cât mai rapidă din cazul problemelor fără restricţii, iar un bun candidat este metoda DFP întrucât nu face apel la diferenţiala a doua care ar putea să nu existe în cazul lui P.

Implementare 4.74 Mai întâi, evidenţiem pe baza unui desen comportamentul funcţiei de penalizare în cazul problemei (P_2) la varierea (creşterea) parametrului γ.

```
f=@(x) [x^2-8*x+3];
a=-7;b=5;
g=@(x) [x-1];
fplot(f,[a,b]);
set(gca, 'XAxisLocation', 'origin')
set(gca, 'YAxisLocation', 'origin')
hold on;
f1=@(x) [f(x)+4*max(0,g(x))^2];
fplot(f1,[a,b]);
hold on;
f2=@(x) [f(x)+8*max(0,g(x))^2];
fplot(f2,[a,b]);
hold on;
f3=@(x) [f(x)+12*max(0,g(x))^2];
fplot(f3,[a,b]);
```

Exerciţiul 4.75 *Să se scrie fişierele care determină soluţia problemei* (P_2) *folosind algoritmul secţiunii de aur pentru funcţia penalizată (prin creşterea progresivă a parametrului* γ*).*

Exerciţiul 4.76 *Să se scrie fişierele care adaptează metoda DFP pentru determinarea soluţiei problemei* (P_1)*.*

Observaţia 4.5.9 *O metodă foarte simplă, dar eficientă în unele situaţii, bazată de asemenea pe ideea penalizării, este prezentată în continuare. Considerăm problema generală cu restricţii funcţionale şi date netede (în notaţia standard):*

$$\min f(x) \ \ cu\ g(x) \leq 0,\ h(x) = 0,$$

unde $f : \mathbb{R}^p \to \mathbb{R}$, $g : \mathbb{R}^p \to \mathbb{R}^n$, $h : \mathbb{R}^p \to \mathbb{R}^m$*. Pentru metoda pe care o vom prezenta, o inegalitate de tipul* $h_j(x) = 0$ *se scrie echivalent în forma* ($h_j(x) \leq 0$ *şi* $-h_j(x) \leq 0$*), de aceea ne vom ocupa de problema cu inegalităţi:*

$$\min f(x) \ \ cu\ g(x) \leq 0,$$

unde $g = (g_1, ..., g_n)$*. Algoritmul propus este următorul:*

Pasul 1. Se alege $y_0 = (1, ..., 1) \in \mathbb{R}^n$ *(un parametru) şi* $x_{-1} = 0 \in \mathbb{R}^p$ *(o dată iniţială),* $\varepsilon > 0$ *(o toleranţă)*

Pasul 2. Date fiind $y_k = (y_k^i)_{i \in \overline{1,n}} \in \mathbb{R}^n$ *şi* $x_{k-1} \in \mathbb{R}^p$ *minimizăm funcţia (fără restricţii)*

$$f(x) + \sum_{i=1}^{n} e^{y_k^i g_i(x)}$$

cu data iniţială x_{k-1}*. Fie* x_k *punct de minim pentru această funcţie.*

Dacă toate produsele $y_k^i g_i(x)$ ($i \in \overline{1,n}$) *sunt mai mici în valoare absolută decât* ε *atunci* x_j *e aproximarea pe care o căutăm.*

Altfel, actualizăm pentru $i \in \overline{1,n}$

$$y_{k+1}^i = e^{y_k^i g_i(x_k)} y_k^i.$$

Rezultatul de convergenţă a acestui algoritm este următorul: dacă şirurile obţinute pe baza paşilor anteriori satisfac

$$y_k^i g_i(x_k) \to 0, \forall i \in \overline{1,n} \text{ şi } x_k \to \overline{x},$$

atunci $\overline{x}$ *este soluţie a problemei cu restricţii.*

Pentru a demonstra, este suficient să avem în vedere cazul $n = 1$*. Astfel, considerăm* $g : \mathbb{R}^p \to \mathbb{R}$*.*

Mai întâi arătăm că $\overline{x}$ *este punct fezabil. Presupunând contrariul,* $g(\overline{x}) > 0$*, deci* $g(x_k) > 0$ *pentru* k *suficient de mare. Obţinem:*

$$y_{k+1} = e^{y_k g(x_k)} y_k \geq y_k.$$

Deci (y_k) *este un şir crescător şi cum* $y_0 > 0$*, acesta nu poate avea limita* 0*. Această observaţie este în contradicţie cu ipoteza* $y_k g_i(x_k) \to 0$*. Deducem că* $g(\overline{x}) \leq 0$*, deci* $\overline{x}$ *este fezabil.*

Presupunem, tot prin reducere la absurd, că pentru orice $\rho > 0$ *există* $x_\rho \in B(\overline{x}, \rho) \cap M$ *astfel încât* $f(x_\rho) < f(\overline{x})$. *Atunci pentru orice* k.

$$f(x_\rho) + e^{y_k g(x_\rho)} \leq f(x_\rho) + 1 < f(\overline{x}) + 1.$$

Cum $y_k g(x_k) \to 0$ *şi* $x_k \to \overline{x}$, *pentru* k *suficient de mare avem*

$$f(x_\rho) + e^{y_k g(x_\rho)} \leq f(x_\rho) + 1 < f(x_k) + e^{y_k g(x_k)}.$$

Ultima inegalitate contrazice minimalitatea lui x_k.

Aşadar, convergenţa menţionată are loc.

Modele de examene scrise

Model examen parţial

Subiectul 1. Fie mulţimea $M = \{x \in \mathbb{R}^2 \mid x + y \leq 1\}$. Să se scrie expresia conului tangent la M în $(2, -1)$. Fie $f : \mathbb{R}^2 \to \mathbb{R}$ o funcţie de clasă C^1. Să se scrie condiţia necesară de optimalitate de ordinul I astfel încât $(2, -1)$ să fie soluţie locală a problemei de optimizare $\min f(x, y)$ cu $(x, y) \in M$.

Subiectul 2. Se doreşte construirea unei cutii parelelipipedice cu baza pătrat şi fără capac cu volumul de 100. Determinaţi dimensiunile unei astfel de cutii astfel încât materialul folosit să fie minim.

Subiectul 3. Fie $f, g : \mathbb{R}^2 \to \mathbb{R}$ date prin $f(x, y) = (x + 1)^2 + y^2$ şi $g(x, y) = -x^3 + y^2$. Considerăm problema minimizării lui $f(x, y)$ cu restricţia $g(x, y) \leq 0$.

(i) Să se verifice dacă are loc condiţia Mangasarian-Fromovitz în $(0, 0)$.

(ii) Să se determine punctele care satisfac condiţiile de optimalitate Fritz John.

(iii) Să se determine soluţiile problemei.

Subiectul 4. Fie $f, g, h : \mathbb{R}^2 \to \mathbb{R}$ date prin $f(x, y) = (x - 1)^2 + y - 2$, $g(x, y) = x + y - 2$ şi $h(x, y) = y - x - 1$. Considerăm problema minimizării lui $f(x, y)$ cu restricţiile $g(x, y) \leq 0$, $h(x, y) = 0$.

(i) Să se găsească punctele care satisfac condiţiile de optimalitate Karush-Kuhn-Tucker.

(ii) Să se verifice dacă condiţiile suficiente de ordinul al doilea au loc în aceste puncte şi să se concluzioneze.

Barem de notare:

1. 1p – baza; conul tangent 4p; condiţia necesară de optimalitate 5p.
2. 1p – baza; modelarea 2p; rezolvarea problemei 5p; concluzia 2p.
3. 1p – baza; condiţia MF: 2p; punctele Fritz John 5; soluţiile 2p.
4. 1p – baza; punctele Karush-Kuhn-Tucker 4p; condiţia suficientă de ordinul al doilea 4p; concluzia 1p.

Timp de lucru: 100 minute

Model examen final

Partea I (30') (scris) Metoda celei mai bune descreşteri cu regula minimizării exacte. Descriere, convergenţă

Partea a II-a (60') (scris+calculator) Fie sistemul (supradeterminat)

$$\begin{cases} x^2 - y^2 + x - 3y = 2 \\ x^3 - y^4 = -2 \\ x^2 + y^3 + 2x - y = -1.1. \end{cases}$$

Să se aproximeze soluţii ale acestui sistem folosind pentru funcţia $f : \mathbb{R}^2 \to \mathbb{R}$,

$$f(x, y) = \left(x^2 - y^2 + x - 3y - 2\right)^2 + \left(x^3 - y^4 + 2\right)^2 + \left(x^2 + y^3 + 2x - y + 1.1\right)^2,$$

cu data iniţială $(1.5, -1.75)$, una dintre metodele de mai jos:
(1) metoda lui Newton;
(2) metoda celei mai bune direcţii de descreştere cu regula minimizării exacte;
(3) metoda Levenberg–Marquardt;
(4) metoda Davidon–Fletcher–Powell.

Să se descrie pe scurt, în scris, metoda folosită.

Barem de notare:
1p – oficiu;
Partea I: enunţ 1 p, demonstraţie 2 p.
Partea a II-a: implementare 4p, descrierea metodei 2p.

Bibliografie

[1] F.J. Aragón-Artacho, M.A. Goberna, *Mathematics in Politics and Governance*, Springer, Cham, 2024.

[2] M. Ancău, *Practical optimization with MATLAB*, Cambridge Scholars Publishing, 2019.

[3] A. Antoniou, W.-S. Lu, *Practical Optimization. Algorithms and Engineering Applications*, Springer, New York, 2021.

[4] R.K. Arora, *Optimization. Algorithms and applications*, CRC Press, 2015.

[5] D. Azé, J.-B. Hiriart-Urruty, *Analyse variationelle et optimization*, Editions Cépanduès, Toulouse, 2010.

[6] A. Beck, *First-order methods in optimization*, SIAM, 2017.

[7] M. Durea, *O introducere în teoria optimizarii neliniare*, Tehnopress, 2012.

[8] M. Durea, R. Strugariu, *An introduction to nonlinear optimization theory*, De Gruyter, 2014.

[9] O. Güler, *Foundations of optimization*, Springer, 2010.

[10] Y. Nesterov, *Introductory lectures on convex optimization*, Springer, 2004.

[11] J. Nocedal, S.J. Wright, *Numerical optimization*, Springer, 2006.

[12] P. Pedregal, *Introduction to optimization*, Springer, 2004.

[13] P. Pedregal, *Optimization and approximation*, Springer, 2017.