

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

Marius Durea

Optimizare practică

Desfășurarea activității în anul universitar 2025–2026

În fiecare săptămână, timp de 14 săptămâni, vor fi 2 ore de curs și 2 ore de seminar/laborator. Cursul și seminarul/laboratorul vor urmări, preponderent, materialul corespunzător postat la adresa:

www.math.uaic.ro/~durea – secțiunea teaching

Evaluarea se desfășoară după procedura de mai jos.

- În timpul semestrului va avea loc evaluarea continuă (EC) care va avea ponderea de 50% din nota finală și va cuprinde două componente:
 - un examen parțial (EP): în săptămâna 7, 8 sau 9, la seminar studenții vor susține un examen parțial scris (cu durată de 100 de minute), iar nota obținută va avea ponderea de 50% din EC; EP nu se poate reface și nu se poate susține în sesiune; exclusiv pe baza unor motive bine întemeiate, EP se poate susține la o dată ulterioară în timpul semestrului;
 - activitatea de seminar (AS): activitatea și calitatea răspunsurilor din timpul desfășurării 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 și care constă dintr-o probă scrisă și în fața calculatorului de 0,5+1=1,5 ore. O condiție necesară pentru promovare este ca nota de la EF să fie minim 4,5. Odată ce această condiție este îndeplinită se va calcula nota finală, care va fi rotunjirea la cel mai apropiat număr natural din intervalul $[1, 10]$ a valorii

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

Prefață

Scopul acestei lucrări este acela de a prezenta unele completări și dezvoltări ulterioare unui prim curs de bază de *Teoria optimizării neliniare*. Avem în vedere o aprofundare a condițiilor de optimalitate pentru probleme cu restricții funcționale și, 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 și tehnicilor prezentate. Pentru implementările propuse se pot folosi coduri Matlab/Octave.

Materialul prezentat are două părți principale. În prima dintre acestea, facem un studiu detaliat al condițiilor necesare și suficiente de ordinul întâi și de ordinul al doilea pentru probleme de optimizare cu restricții și descriem mai multe aplicații mai puțin imediate ale acestora. Motivația 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 ecuații neliniare, studiem algoritmi de optimizare și algoritmi de determinare a rădăcinilor unor funcții, 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 părți, este caracterul euristic al unor metode numerice care au motivații teoretice solide, dar nu sunt întru totul justificate aici din punct de vedere matematic, mai ales din lipsa spațiului. Pentru aceste metode existența teoriei "din spate" este ușor de acceptat, tocmai datorită faptului că metodele de bază sunt complet acoperite din acest punct de vedere.

Prin urmare, această lucrare își propune să prezinte rezultate și tehnici deja consacrate, iar aranjarea materialului, selectarea exemplelor ilustrative, ordonarea codurilor utilizate reprezintă punctul de vedere al autorului asupra acestui domeniu. Monografiile indicate la bibliografie conțin multe extensii și completări semnificative ale chestiunilor ce se regăsesc în lucrarea de față.

Cuprins

Prefață	i
1 Optimizare - exemple și elemente recapitulative	1
1.1 Exemple de probleme practice	1
1.1.1 Probleme fără restricții	1
1.1.2 Probleme cu restricții	2
1.2 Cadrul general și rezultate de existență a soluțiilor – recapitulare	5
1.3 Condiții de optimalitate – recapitulare și completări	7
2 Elemente teoretice și practice în optimizarea cu restricții	16
2.1 Teorema Karush-Kuhn-Tucker – recapitulare	16
2.2 Semnificația practică a multiplicatorilor Lagrange	21
2.3 O selecție de probleme	25
2.4 Un exemplu, două întrebări	33
2.5 Condițiile Fritz John	34
2.6 Condiții de optimalitate de ordinul al doilea	39
3 Solvabilitate și aproximare	47
3.1 Solvabilitatea problemelor de optimizare	47
3.2 Metode numerice, în general – algoritmi iterativi	53
3.3 Un algoritm reprezentativ: metoda lui Newton	58
3.4 Aplicație la funcția W , a lui Lambert	66
3.5 O metodă cvasi-Newton: metoda secantei	75
4 Algoritmi pentru probleme de optimizare fără restricții	81
4.1 Algoritmi de ordin zero pentru funcții de o variabilă	81
4.1.1 Delimitarea (bracketing) punctului de minim	82
4.1.2 Metoda secțiunii de aur	84
4.1.3 Metoda Fibonacci	86
4.2 Algoritmi de ordinul întâi (metode de tip gradient)	90
4.2.1 Descrierea metodei direcției de descreștere (gradient descent, GD)	90
4.2.2 Convergența metodei direcției de descreștere. Cazul general	93
4.2.3 Convergența metodei direcției de descreștere. Cazul convex	100
4.2.4 Posibilități de accelerare a metodei gradient descent	107

4.2.5	Algoritmi specifici <i>machine learning</i> bazați pe metoda direcției de descreștere	109
4.2.6	Convergența metodei direcției de descreștere. Cazul funcțiilor pătratice . .	114
4.2.7	Metoda gradientilor conjugați	118
4.3	Algoritmi de ordinul al doilea	123
4.3.1	Metoda lui Newton în optimizare	124
4.3.2	Metode derivate din metodele fundamentale	125
4.4	Cazul funcțiilor cu valori vectoriale: optimizare Pareto	130
4.5	Obținerea unor algoritmi pentru probleme cu restricții - o metodă de penalizare .	133
Modele de examene scrise		141
	Model examen parțial	141
	Model examen final	142
Bibliografie		143

Capitolul 1

Optimizare - exemple și elemente recapitulative

1.1 Exemple de probleme practice

1.1.1 Probleme fără restricții

Exemplul 1.1.1 Funcția rezultată folosind Principiul lui Fermat pentru modelarea problemei refracției 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 $f : \mathbb{R} \rightarrow \mathbb{R}$,

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

unde v_1 este viteza luminii în primul mediu, iar v_2 este viteza luminii în al doilea mediu. Această funcție 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 funcții. Un sportiv înoată într-un lac și se află în larg la o distanță de 5km față de mal. El trebuie să ajungă pe uscat într-un punct aflat la 7km în lungul țărmului față de poziția sa în raport cu țărmul. Știind că sportivul aleargă cu 6m/s și înoată cu 2m/s , să se determine timpul minim în care poate ajunge la destinație.

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

Exercițiul 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 soldați ce pot fi angajați: mercenari sau cetățeni aleși prin tragere la sorți. Fie $c \in [0, 1]$ proporția de cetățeni (și, evident, $1 - c$ proporția de mercenari). Fiecare dintre cele două tipuri de soldați comportă un risc. Mercenarii sunt oportuniști și pot specula unele momente pentru a obține profituri suplimentare, ceea ce diminuează averea conducătorului, notată cu w . Notăm cu $v \in [0, 1]$ "gradul de oportunism" al unui (fiecărui) mercenar. Pe de altă parte, soldații cetățeni pot avea opțiuni politice în favoarea sau împotriva conducătorului, ceea ce ar putea conduce la îndepărtarea acestuia de la putere pe calea armelor. Dacă notăm cu $g \in [0, 1]$ proporția de cetățeni favorabili din punct de vedere politic conducătorului, gradul de adversitate va fi $c(1 - g)$. Pentru a stabili ce proporție din fiecare categorie angajează, conducătorul are de rezolvat așadar problema

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

adică

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

Exercițiul 1.3 *Să se rezolve problema de mai sus și să se interpreteze rezultatele.*

1.1.2 Probleme cu restricții

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

$$\sum_i c_i x_i$$

pentru asigurarea unei alimentații potrivite. Astfel, notăm cu a_{ij} conținutul de nutrient j dintr-o unitate de produs i și b_j cantitatea necesară din acest nutrient. Deci avem restricțiile

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

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

Exemplul 1.1.6 Se dorește construirea unui bloc paralelipipedic având n nivele, o adâncime a subsolului notată cu d și o înălțime a suprasolului notată cu h . Dimensiunile amprenteii la sol sunt l și w . Se dorește ca suprafața utilă să fie de cel puțin 20000 m^2 , fiecare nivel să fie înalt de 3.5 m , iar dimensiunea terenului determină ca l, w să fie sub 50 m . Din motive estetice, se dorește ca $l/w = \varphi \simeq 1.618$ (numărul de aur). Costurile de încălzire/răcire sunt 100 u.m./an/m^2 pentru suprafață supratereană iar bugetul de 225000 u.m./an . Problema este aceea de a minimiza spațiul de excavat pentru a fi îndeplinite toate aceste condiții. Modelând problema, se ajunge la

$$\min dlw$$

cu restricțiile

$$\begin{aligned}nlw &\geq 20000 \\l &= \varphi w \leq 50 \\d + h &= 3.5n \\100(2hl + 2hw + lw) &\leq 225000.\end{aligned}$$

Folosind aceste relații și neglijând faptul că n este întreg, eliminăm l , n și ajungem la problema de minimizare cu variabilele d, w, h

$$\min dw^2$$

cu

$$\begin{aligned}(d + h)w^2 &\geq \frac{70000}{\varphi} \\w &\leq \frac{50}{\varphi} \\2(\varphi + 1)hw + \varphi w^2 &\leq 2250.\end{aligned}$$

Exemplul 1.1.7 Avem un sistem de eșafodaj 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 această ordine, cu distanțele dintre ele de 2 și 10 m;
- la capetele firelor C, D , aflate la același 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 același nivel cu firul A ;
- la capetele firelor A, B se află o platformă de masă neglijabilă.
- pe cele două platforme se așează, la mijloc, câte un corp cu greutatea x_1 și respectiv x_2 .

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

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

și 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 restricții le includ pe altele (de exemplu, ultima pe penultima) pentru că $x_1, x_2 \geq 0$ și putem să reducem problema la

$$\max(x_1 + x_2)$$

cu

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

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

Exemplul 1.1.8 O companie dorește să investească 75000 unități monetare (u.m.) pentru achiziționarea unor mașini de tip A, B și C . O mașină A costă 1500 u.m. și produce un beneficiu anual de 315 u.m.; o mașină B costă 1750 u.m. și produce un beneficiu anual de 450 u.m.; o mașină C costă 1375 u.m. și produce un beneficiu anual de 275 u.m. Toate mașinile au nevoie de întreținere și perioade de pauză, astfel: pentru o mașină A întreținerea costă 20 u.m. și e nevoie de 34 de zile de pauză anual, pentru o mașină B întreținerea costă 27 u.m. și e nevoie de 29 de zile de pauză anual, pentru o mașină C întreținerea costă 25 u.m. și e nevoie de 24 de zile de pauză anual. Bugetul maxim anual pentru întreținere este de 1000 u.m., iar compania dorește ca numărul de zile de pauză a tuturor mașinilor să nu depășească 800 pe an. Chestiunea este câte mașini de fiecare tip e de dorit să fie achiziționate pentru maximizarea profitului. Dacă notăm numărul mașinilor de fiecare tip cu a, b și respectiv c , avem de maximizat

$$315a + 450b + 275c$$

cu restricțiile

$$1500a + 1750b + 1375c \leq 75000$$

$$20a + 27b + 25c \leq 1000$$

$$34a + 29b + 24c \leq 800.$$

De asemenea, din motive evidente, trebuie să avem $a, b, c \in \mathbb{N}$. Această ultimă restricție schimbă complet natura problemei, plasând-o în categoria problemelor de optimizare cu întregi, iar în acest curs nu ne vom ocupa de această categorie de probleme.

Exercițiul 1.5 Fie A, B două surse situate în punctele $(0, 1)$ și $(2, 1)$ care emit aceeași cantitate de energie. Pe axa Ox este plasat un receptor. Dacă energia captată de receptor este invers proporțională cu pătratul distanței față de sursă, să se determine poziția acestui astfel încât raportul dintre energia captată de la sursa A și energia captată de la sursa B să fie maxim.

Exercițiul 1.6 Se dorește construirea unei cutii paralelipipedice cu baza pătrat și fără capac cu volumul de 100. Determinați dimensiunile unei astfel de cutii astfel încât materialul folosit să fie minim.

Exercițiul 1.7 O fabrică dorește să vândă n exemplare dintr-un produs astfel încât profitul să fie maxim. Cercetările de piață arată că la un preț 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 niște costuri de start a producției de 2000 u.m., iar costul de producție al fiecărui exemplar este 0.5 u.m. Să se modeleze problema și să se determine profitul maxim, prețul de vânzare și n .

Exercițiul 1.8 Determinați dreptunghiul cu aria cea mai mare care poate fi inclus în epigraful funcției $x \mapsto x^2$ și care se află sub linia de nivel $y = a$, unde $a > 0$.

Exercițiul 1.9 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 și rezultate de existență a soluțiilor – recapitulare

Fie $U \subset \mathbb{R}^p$ o mulțime nevidă și deschisă, $f : U \rightarrow \mathbb{R}$ o funcție și $M \subset U$ o mulțime nevidă. Suntem interesați să studiem problema minimizării funcției f atunci când argumentul acesteia parcurge mulțimea M . Formal, vom scrie această problemă în forma următoare:

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

Mulțimea M se va numi mulțimea punctelor fezabile ale problemei (P) sau mulțimea constrângerilor sau, încă, mulțimea restricțiilor.

Să spunem de la început că ne vom ocupa de minimizarea funcției f , dar rezultate referitoare la maximizarea sa pot fi obținute aplicând rezultatele de minimizare funcției $-f$ în baza relației $\max f = -\min(-f)$. Pentru început, să definim noțiunea de soluție asociată problemei (P) .

Definiția 1.2.1 Spunem că $\bar{x} \in M$ este soluție locală (sau, simplu, soluție) pentru problema (P) sau punct de minim pentru funcția f pe mulțimea M dacă există o vecinătate V a punctului $\bar{x}$ astfel încât $f(\bar{x}) \leq f(x)$ pentru orice $x \in M \cap V$.

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

Să remarcăm că $\bar{x} \in M$ nu este punct de minim pentru funcția f pe mulțimea M dacă și numai dacă există un șir $(x_n) \subset M$, $x_n \rightarrow \bar{x}$ astfel încât $f(x_n) < f(\bar{x})$ pentru orice n .

Observația 1.2.2 În general, vom distinge două mari situații pentru studiul problemei (P) : cazul în care $M = U$ și cazul în care M este intersecția dintre o mulțime închisă a lui $\mathbb{R}^p$ și U . În primul caz vom spune că problema de optimizare (P) este fără restricții (sau constrângeri) în timp ce în al doilea caz vorbim despre o problemă cu restricții (constrângeri). Să mai observăm că în Definiția 1.2.1, dacă $\bar{x} \in \text{int } M$, atunci $\bar{x}$ este soluție locală și a problemei fără restricții (este suficient să micșorăm vecinătatea V astfel încât $V \subset M$). Deci în cazul problemelor cu restricții cazul interesant (care nu se reduce la cazul unei probleme fără restricții) este cazul în care $\bar{x} \in \text{Fr } M$. Dacă $\bar{x} \in \text{int } M$ mai spunem și că restricția este inactivă.

Un rol important îl vor juca mulțimile de nivel ale unei funcții. Având o funcție $f : \mathbb{R}^p \rightarrow \mathbb{R}$, 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 mulțime compactă și $f : M \rightarrow \mathbb{R}$ este o funcție continuă, atunci problema minimizării lui f pe M și problema maximizării lui f pe M au soluții globale.

Teorema 1.2.4 Fie $f : \mathbb{R}^p \rightarrow \mathbb{R}$ o funcție continuă și $M \subset \mathbb{R}^p$ o mulțime nevidă și închisă. Dacă există $\nu > \inf_{x \in M} f(x)$ astfel încât mulțimea 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 își atinge minimul global pe M .

Rezultatele de mai sus asigură existența punctelor de minim în condiții de compactitate a mulțimilor de nivel. Evident, mărginirea inferioară a funcției este condiție necesară pentru existența minimului, dar este clar că mărginirea mulțimilor de nivel nu este o astfel de condiție. De exemplu funcția $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = (x - 1)^2 e^{-x}$ își atinge minimul (global) în $\bar{x} = 1$, 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, situație în care ipoteza este verificată dacă impunem funcției f o condiție de coercivitate.

Propoziția 1.2.5 Fie $f : \mathbb{R}^p \rightarrow \mathbb{R}$ o funcție și $M \subset \mathbb{R}^p$ o mulțime nevidă, închisă și nemărginită. Dacă $\lim_{x \in M, \|x\| \rightarrow \infty} f(x) = \infty$ (i.e., pentru orice $(x_n) \subset M$, $\|x_n\| \rightarrow \infty$ are loc $f(x_n) \rightarrow \infty$) atunci mulțimea $N_\nu f \cap M$ este mărginită pentru orice $\nu > \inf_{x \in M} f(x)$.

1.3 Condiții de optimalitate – recapitulare și completări

Definiția 1.3.1 Fie $M \subset \mathbb{R}^p$ o mulțime nevidă și $\bar{x} \in \text{cl } M$. Un vector $u \in \mathbb{R}^p$ se numește tangent la mulțimea M în $\bar{x}$ dacă există $(t_n) \subset (0, \infty)$, $t_n \rightarrow 0$ și $(u_n) \rightarrow u$ astfel încât pentru orice $n \in \mathbb{N}$,

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

Evident este suficient ca relația 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 Mulțimea, notată $T(M, \bar{x})$, a tuturor vectorilor tangenți la M în $\bar{x}$ este un con închis, numit conul tangent (în sens Bouligand) la M în $\bar{x}$.

Propoziția 1.3.3 Dacă $\emptyset \neq M \subset \mathbb{R}^p$ și $\bar{x} \in \text{cl } M$, atunci are loc relația $T(M, \bar{x}) = T(\overline{M}, \bar{x})$. Dacă $\bar{x} \in \text{int } M$, atunci $T(M, \bar{x}) = \mathbb{R}^p$.

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

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

Propoziția 1.3.4 Fie $\emptyset \neq M \subset \mathbb{R}^p$ convexă și $\bar{x} \in M$. Atunci

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

iar

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

Teorema 1.3.5 (Condiția necesară de ordinul I) Dacă $\bar{x}$ este soluție locală pentru problema (P) și f este diferențiabilă în $\bar{x}$, atunci $\nabla f(\bar{x})(u) \geq 0$ pentru orice $u \in T(M, \bar{x})$.

Observația 1.3.6 Condiția din concluzia teoremei precedente se mai scrie, echivalent,

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

Observația 1.3.7 Ținând cont de Propoziția 1.3.3, dacă în teorema de mai sus, $\bar{x} \in \text{int } M$ (restricție inactivă), obținem $\nabla f(\bar{x})(u) \geq 0$ pentru orice $u \in \mathbb{R}^p$. Din liniaritatea lui $\nabla f(\bar{x})$, deducem $\nabla f(\bar{x}) = 0$, adică Teorema lui Fermat.

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

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

și mulțimea restricțiilor

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

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

Funcția f este continuă, iar M este compactă, deci problema (P) admite cel puțin o soluție globală.

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

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

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

și rezolvând sistemul găsim soluția $\bar{x} = (-\frac{5}{3}, -\frac{4}{3})$ care nu aparține lui M . Deci problema nu are soluții în int M .

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

Astfel, dacă punctul $\bar{x}$ este

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

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

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

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

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

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

– punctul $(0, 1)$:

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

– punctul $(1, 0)$:

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

– punctul $(0, 0)$:

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

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

Putem, de asemenea, să ne punem problema maximului global al lui f pe M (a cărei existență este asigurată de Teorema lui Weierstrass), problemă echivalentă cu găsirea minimului lui $-f$ pe M . Repetând discuția precedentă, găsim două puncte ce verifică condiția necesară de optimalitate (i.e. $\nabla f(\bar{x}) \in N(M, \bar{x})$): $\bar{x} = (0, 0)$ și $\bar{x} = (1, 0)$. 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 și conurile normale în diferite puncte la mulțimea $M \subset \mathbb{R}^p$,

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

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

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

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

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

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

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

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

În continuare, arătăm că

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

Pentru aceasta considerăm elementele

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

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

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

Polara acestei mulțimi este

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

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

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

Astfel, se ajunge la forma anunțată a conului normal.

Exemplul 1.3.10 (optimizare pe simplexul unitate) Fie $f : \mathbb{R}^p \rightarrow \mathbb{R}$ convexă și diferențibilă. Dacă dorim să minimizăm această funcție pe simplexul unitate (a se vedea Exemplul 1.3.9), atunci, conform Propoziției 1.3.15, $\bar{x} \in M$ este punct de minim pentru f pe M dacă și numai dacă $-\nabla f(\bar{x}) \in N(M, \bar{x})$. Din forma particulară a lui $N(M, \bar{x})$, deducem că această condiție se scrie

$$\begin{aligned} \frac{\partial f}{\partial x_i}(\bar{x}) &= c, \text{ (constantă), } \forall i \notin I(\bar{x}) \\ \frac{\partial f}{\partial x_i}(\bar{x}) &\geq c, \forall i \in I(\bar{x}). \end{aligned}$$

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

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

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

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

adică

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

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

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

Cum problema are soluție și un singur punct verifică condiția necesară, deducem că acel punct este soluția căutată.

Observația 1.3.11 O altă abordare în cazul Exemplelor 1.3.8 și 1.3.10 constă în transformarea restricțiilor geometrice în restricții funcționale și 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 același exemplul funcției $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^3$ și $\bar{x} = 0$.

Se pot însă impune condiții suplimentare așa încât să avem unele echivalențe și cel mai important caz este cel al funcțiilor convexe.

Teorema 1.3.12 *Fie $U \subset \mathbb{R}^p$ o mulțime convexă și deschisă și fie $f : U \rightarrow \mathbb{R}$ o funcție convexă și diferențiabilă pe U . Următoarele afirmații sunt echivalente:*

- (i) $\bar{x}$ este un minim global al lui f ;
- (ii) $\bar{x}$ este un minim local al lui f ;
- (iii) $\bar{x}$ este punct critic al lui f (i.e. $\nabla f(\bar{x}) = 0$).

Așa cum se observă, în cazul funcțiilor convexe, condiția de ordinul I (în cazul fără constrângeri) este necesară și suficientă pentru optimalitate.

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

Propoziția 1.3.13 *Fie $M \subset \mathbb{R}^p$ o mulțime convexă și fie $f : M \rightarrow \mathbb{R}$ o funcție convexă. Dacă $\bar{x} \in M$ este punct de minim local pentru f pe M atunci $\bar{x}$ este punct de minim global pentru f pe M .*

Propoziția 1.3.14 *Fie $M \subset \mathbb{R}^p$ o mulțime convexă și fie $f : M \rightarrow \mathbb{R}$ o funcție convexă. Mulțimea punctelor de minim ale lui f pe M este convexă. Dacă, în plus, f este strict convexă, atunci această mulțime are cel mult un element.*

În cazul cu restricții, pentru funcțiile convexe condiția necesară de optimalitate de ordinul I se scrie într-o formă specială și, în plus, este și condiție suficientă.

Propoziția 1.3.15 *Fie $U \subset \mathbb{R}^p$ o mulțime convexă și deschisă și fie $f : U \rightarrow \mathbb{R}$ o funcție convexă și diferențiabilă pe U . Fie $M \subset U$ convexă. Elementul $\bar{x} \in M$ este punct de minim pentru f pe M dacă și numai dacă*

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

Revenind la Teoremele 1.3.5 și 1.3.18, întărind condițiile din concluziile rezultatelor precedente, obținem condiții suficiente de optimalitate.

Rezultatele anunțate sunt următoarele.

Teorema 1.3.16 *Presupunem că f este diferențiabilă în $\bar{x} \in M$ și*

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

Atunci $\bar{x}$ este soluție locală pentru problema (P).

□

Observația 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 interesați doar de precizarea unei condiții suficiente de optimalitate.*

Prezentăm acum o cunoscută condiție necesară de optimalitate de ordinul al doilea, pentru problema fără restricții.

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

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

În privința condițiilor suficiente de ordinul al doilea, are loc următorul rezultat.

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

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

Atunci $\bar{x}$ este soluție locală pentru problema (P) .

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

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

De fapt, în Teoremele 1.3.16 și 1.3.19 se obțin relații mai tari decât simpla minimalitate. Introducem mai întâi o definiție.

Definiția 1.3.21 Fie $\alpha > 0$. Spunem că $\bar{x} \in M$ este o soluție 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 $x \in M \cap B(\bar{x}, \varepsilon)$,

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

Observația 1.3.22 Să observăm că pentru funcții diferențiabile noțiunea de soluție strictă locală de ordin 1 este specifică cazului cu restricții active (adică $\bar{x} \in M \setminus \text{int } M$): dacă f este diferențiabilă în $\bar{x} \in \text{int } M$, atunci $\bar{x}$ nu poate fi soluție strictă locală de ordin 1. Într-adevăr, dacă $\bar{x} \in \text{int } M$ ar fi soluție strictă locală de ordin 1, atunci, pe de o parte, $\nabla f(\bar{x}) = 0$ (din Teorema lui Fermat), iar, pe de altă parte, $\nabla f(\bar{x}) \neq 0$ din definiția soluției stricte.

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

Teorema 1.3.23 Presupunem că f este diferențiabilă în $\bar{x} \in M$ și considerăm următoarele afirmații:

- (i) $\nabla f(\bar{x})(u) > 0$, pentru orice $u \in T(M, \bar{x}) \setminus \{0\}$;
- (ii) $-\nabla f(\bar{x}) \in \text{int } N(M, \bar{x})$;

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

(iv) $\bar{x}$ este soluție izolată pentru problema (P).

Să se arate că (i) $\iff$ (ii) $\iff$ (iii). Dacă, în plus, M este convexă și f este continuu diferențiabilă în $\bar{x}$, să se arate că (i) $\implies$ (iv).

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

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

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

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

Arătăm implicația (i) $\implies$ (iii). Presupunem, prin reducere la absurd, că $\bar{x}$ nu este soluție strictă de ordin 1. Atunci, negând definiția, deducem că există un șir $(x_n) \rightarrow \bar{x}$, $(x_n) \subset M$ astfel încât pentru orice $n \in \mathbb{N}^*$,

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

În virtutea acestei inegalități,

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

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

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

Combinând cele două relații, avem

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

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

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

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

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

ceea ce contrazice ipoteza.

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

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

Cum f este diferențiabilă în $\bar{x}$, există $(\beta_n) \rightarrow 0$ astfel încât

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

Obținem, pentru n suficient de mare,

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

Prin trecere la limită,

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

adică concluzia.

În final, arătăm că (ii) $\implies$ (iv) în ipoteza suplimentară că M este convexă și f este continuu diferențiabilă în $\bar{x}$. Fie $\varepsilon > 0$ astfel încât $-\nabla f(\bar{x}) + D(0, \varepsilon) \subset N(M, \bar{x})$. Presupunem, prin reducere la absurd, că există un șir de soluții $(x_n) \subset M \setminus \{\bar{x}\}$ cu $x_n \rightarrow \bar{x}$. Conform unei consecințe binecunoscute a Teoremei Hahn-Banach, pentru fiecare n există z_n cu $\|z_n\| = 1$ și $z_n(\bar{x} - x_n) = \|\bar{x} - x_n\|$. Fie $(v_n) = (-\nabla f(x_n) - 2^{-1} \varepsilon z_n)$. Din continuitatea diferențialei în $\bar{x}$, deducem că $v_n \in N(M, \bar{x})$ pentru orice n suficient de mare. Din condiția necesară de optimalitate de ordinul I, $-\nabla f(x_n) \in N(M, x_n)$. Cum M este convexă, $\bar{x} - x_n \in T(M, x_n)$ și $x_n - \bar{x} \in T(M, \bar{x})$. Atunci $-\nabla f(x_n)(\bar{x} - x_n) \leq 0$ și $v_n(x_n - \bar{x}) \leq 0$, deci pentru n mare $(v_n + \nabla f(x_n))(\bar{x} - x_n) \geq 0$. Obținem $-2^{-1} \varepsilon z_n(\bar{x} - x_n) \geq 0$, de unde deducem că $-2^{-1} \varepsilon \|\bar{x} - x_n\| \geq 0$. În mod evident, ultima relație reprezintă o contradicție. $\square$

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

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

Să se arate că $\bar{x}$ este soluție strictă locală de ordinul $\alpha = 2$ pentru problema (P).

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

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

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

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

Obținem

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

de unde, pentru finalizarea demonstrației, împărțim la $\|x_n - \bar{x}\|^2$ și repetăm argumentele de la implicația (i) $\implies$ (iii) de mai sus. $\square$

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

Exercițiul 1.10 În notațiile definiției precedente, să se arate că:

- (i) mulțimea $F(M, \bar{x}) \cup \{0\}$ este un con, nu neapărat închis;
- (ii) $\text{cl}(F(M, \bar{x}) \cup \{0\}) \subset T(M, \bar{x})$, iar incluziunea poate fi strictă;
- (iii) dacă M este convexă, atunci $\text{cl}(F(M, \bar{x}) \cup \{0\}) = T(M, \bar{x})$.

Problema 1.11 Să se arate următoarea condiție necesară de optimalitate de ordinul al doilea pentru problema (P) : dacă f este de clasă C^2 și $\bar{x} \in M$ este soluție pentru (P) atunci pentru orice u , direcție fezabilă în $\bar{x}$ pentru care $\nabla f(\bar{x})(u) = 0$ avem $\nabla^2 f(\bar{x})(u, u) \geq 0$.

Exercițiul 1.12 Să se determine punctele de extrem local ale funcțiilor de mai jos (probleme fără restricții):

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

Exercițiul 1.13 Fie $f : \mathbb{R}^2 \rightarrow \mathbb{R}$. Considerăm problema (P) a minimizării lui f cu restricțiile $x_1 \geq 0$, $x_2 \geq 0$. Știind că $\nabla f(0) \neq 0$, $\frac{\partial f}{\partial x_1}(0) \leq 0$ și $\frac{\partial f}{\partial x_2}(0) \leq 0$ să se arate că 0 nu este soluție a problemei (P).

Capitolul 2

Elemente teoretice și practice în optimizarea cu restricții

2.1 Teorema Karush-Kuhn-Tucker – recapitulare

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

Considerăm că mulțimea 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 restricții: restricții cu inegalități și restricții cu egalități. 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 soluție locală a problemei (P) , restricția $g_i \leq 0$ să nu influențeze efectiv mulțimea punctelor u pentru care trebuie să comparăm $f(x)$ cu $f(u)$. De aceea, în acest caz, spunem că restricția $g_i \leq 0$ este inactivă în x . Astfel, ne interesează ca astfel de restricții să fie eliminate din discuție. Pentru $\bar{x} \in M$, notăm mulțimea indicilor restricțiilor inegalități active prin

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

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

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

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

și

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

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

Observația 2.1.2 Două dintre cele mai importante condiții de calificare sunt prezentate mai jos.

Prima dintre acestea se numește condiția de liniară independentă (în $\bar{x}$) și se formulează astfel: mulțimea $\{\nabla g_i(\bar{x}) \mid i \in A(\bar{x})\} \cup \{\nabla h_j(\bar{x}) \mid j \in \overline{1, m}\}$ este liniar independentă.

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

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

Este ușor de observat că dacă are loc condiția de liniară independentă în $\bar{x} \in M$, atunci are loc condiția Mangasarian-Fromovitz în $\bar{x}$. Totuși, cele două condiții nu sunt echivalente.

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

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

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

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

Definiția 2.1.5 Spunem că sistemul de restricții asociat problemei (P) este de tip convex dacă restricțiile cu inegalități sunt exprimate prin funcții convexe, în timp ce restricțiile cu egalități sunt date cu funcții afine, adică $(g_i)_{i \in \overline{1, n}}$ sunt convexe și h este afină. În acest context, spunem că are loc condiția Slater dacă există $u \in \mathbb{R}^p$ astfel încât $h(u) = 0$ și $g(u) < 0$.

Teorema 2.1.6 Dacă sistemul de restricții asociat problemei (P) este de tip convex, atunci condiția Slater este o condiție de calificare.

Definiția 2.1.7 Spunem că sistemul de restricții asociat problemei (P) este de tip afin dacă atât restricțiile cu inegalități, cât și cele cu egalități sunt date prin funcții afine.

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

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

Funcția $L : U \times \mathbb{R}^{n+m} \rightarrow \mathbb{R}$,

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

se numește lagrangianul problemei (P) . Astfel, concluzia din relația (2.1) se poate scrie

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

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 restricții într-o problemă fără restricții prin intermediul unor noi nedeterminate a apărut în unele dintre lucrările lui Lagrange legate de probleme de calculul variațiilor. Este clar că teorema precedentă nu asigură unicitatea acestor multiplicatori.

Teorema 2.1.1 asigură condiții necesare de optimalitate pentru problema (P) . Dacă în loc de minimizare dorim să maximizăm funcția obiectiv sub aceleași restricții, atunci, din faptul că $\max f = -\min(-f)$, condiția necesară (2.1) se scrie

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

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

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

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

Exemplul 2.1.9 Fie $a, b, x \in \mathbb{R}^p$ și $c, d \in \mathbb{R}$. Rezolvăm problema minimizării normei euclidiene a lui x cu restricțiile $\langle a, x \rangle \leq c$, $\langle b, x \rangle \leq d$. Funcția obiectiv este strict convexă și coercivă, restricțiile sunt liniare. Deci, dacă mulțimea punctelor fezabile este nevidă, problema are exact o soluție, iar condițiile Karush-Kuhn-Tucker sunt necesare și suficiente. Deci soluția x este caracterizată de faptul că există $\lambda_1, \lambda_2 \geq 0$ astfel încât

$$2x + \lambda_1 a + \lambda_2 b = 0$$

$$\lambda_1 (\langle a, x \rangle - c) = 0$$

$$\lambda_2 (\langle b, x \rangle - d) = 0.$$

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 și soluție a problemei.

Cazul 2: $\lambda_1 = 0$, $\lambda_2 \neq 0$. Atunci ecuația $\langle b, x \rangle - d = 0$ 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 soluție a sistemului, trebuie ca $d < 0$ și $d \langle a, b \rangle \leq c \|b\|^2$.

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

Cazul 4: $\lambda_1 \neq 0$, $\lambda_2 \neq 0$. Determinăm multiplicatorii folosind relațiile $\langle a, x \rangle = c$, $\langle b, x \rangle = d$ și înmulțind scalar în mod corespunzător prima ecuație. Obținem

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

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

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

Astfel,

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

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

Încheiem această secțiune cu câteva exerciții recapitulative.

Exercițiul 2.1 Rezolvați problema minimizării globale a expresiei $x_1^2 + x_2^2 - 3x_1x_2$ cu restricția $x_1^2 + x_2^2 - 6 \leq 0$.

Exercițiul 2.2 Rezolvați problema minimizării globale a expresiei $x_1^2 + x_2^2 - 4x_1 - 6x_2$ cu restricțiile $x_1 + x_2 \geq 7$, $x_1 - x_2 \leq 2$.

Exercițiul 2.3 Rezolvați problema

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

Exercițiul 2.4 Rezolvați problema

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

Exercițiul 2.5 *Rezolvați problema*

$$\min (3x_1 + x_2) \text{ cu } x_1^2 + x_2^2 - 5 \leq 0.$$

Exercițiul 2.6 *Rezolvați problema*

$$\min (x_1^2 + x_2^2 - 14x_1 - 6x_2 - 7) \text{ cu } x_1 + x_2 \leq 2, x_1 + 2x_2 \leq 3.$$

Exercițiul 2.7 *Rezolvați problema*

$$\min (x_1^2 + 2(x_2 + 1)^2) \text{ cu } -x_1 + x_2 = 2, -x_1 - x_2 - 1 \leq 0.$$

Exercițiul 2.8 *Rezolvați problema*

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

Exercițiul 2.9 *Fie problema $\min (x_1^2 + x_2^2 - x_2x_3 + x_3^2)$ cu restricțiile $x_1 - x_2 \geq 1, x_1 = x_3$ și $x_3 \geq 0$. Determinați soluția sau soluțiile problemei.*

Exercițiul 2.10 *Rezolvați problema*

$$\min (x_1^2 + x_2^2 + x_3^2) \text{ cu } x_1 + x_2 + x_3 = 0, x_1 + 2x_2 + 3x_3 - 1 = 0.$$

Exercițiul 2.11 *Găsiți extremele globale ale expresiei*

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

Exercițiul 2.12 *Găsiți extremele globale ale expresiei*

$$(x_1x_2 + x_2x_3 + x_1x_3)$$

astfel încât:

$$(i) x_1^2 + x_2^2 + x_3^2 = 1;$$

$$(ii) x_1^2 + x_2^2 + x_3^2 = 1, x_1 - x_2 + x_3 = 0.$$

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

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

Exercițiul 2.15 *Fie $f : \mathbb{R}^2 \rightarrow \mathbb{R}, f(x_1, x_2) = 5x_1^2 + 4x_1x_2 + x_2^2$ și $h : \mathbb{R}^2 \rightarrow \mathbb{R}, h(x_1, x_2) = 3x_1 + 2x_2 + 5$. Considerăm problema minimizării lui f cu restricția $h(x) = 0$. Să se arate că problema are soluție unică și apoi să se determine soluția precum și multiplicatorul Lagrange $\bar{\mu}$ asociat. Să se arate că apoi că soluția determinată este minim global fără restricții pentru funcția $x \rightarrow L(x, \bar{\mu})$. Se păstrează aceste concluzii pentru problema minimizării lui $f : \mathbb{R}^2 \rightarrow \mathbb{R}, f(x_1, x_2) = x_1^2 - x_2^2 - 3x_2$ cu restricția egalitate $h(x) = 0$, unde $h : \mathbb{R}^2 \rightarrow \mathbb{R}, h(x_1, x_2) = x_2$?*

Exercițiul 2.16 Fie $n \in \mathbb{N}^*$ și $a_i > 0$, 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 aceeași restricție?

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

2.2 Semnificația practică a multiplicatorilor Lagrange

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

Exemplul 2.2.1 (Planificare economică - 1) Trei generatoare electrice deservește 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

$$\begin{aligned} \min & (x_1 + 0.0625x_1^2 + x_2 + 0.0125x_2^2 + x_3 + 0.0250x_3^2) \\ \text{cu restricția} & x_1 + x_2 + x_3 = 952. \end{aligned}$$

Este ușor de văzut că problema este convexă iar sistemul de restricții este afin. Așadar, (x_1, x_2, x_3) este soluție dacă și 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

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

Obținem $\mu = -15$, $x_1 = 112$, $x_2 = 560$, $x_3 = 280$. Valoarea minimă a funcției 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 $f, v : \mathbb{R}^2 \rightarrow \mathbb{R}$ (dimensiunea 2 nu este esențială) sunt două funcții de clasă C^1 și că avem problema

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

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

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

În particular, obținem că

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

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

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

avem

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

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

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

Prin urmare,

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

adică ceea ce am intuit anterior.

Revenind la problema concretă discutată, dacă sarcina de alimentare totală crește cu o unitate, atunci, cum $-\mu = 15$, funcția obiectiv își crește 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 același produs (de exemplu, chipsuri de cartofi). Pentru fiecare sortiment sunt necesare trei operațiuni 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 producție, mașinile care fac cele trei operațiuni au autonomii de 345, 480 și respectiv 330 de minute. Știind că profitul este de 2 Euro și 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 enunț și obținem 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 soluțiilor 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 mulțimile de nivel relativ la mulțimea punctelor fezabile sunt compacte, deci problema are soluție globală. Sistemul de restricții este afin, deci putem scrie: $(x, y) \in M$ este soluție a problemei dacă și 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

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

Considerăm mai multe cazuri.

1. Situația $\alpha = \beta = \gamma = 0$ este imposibilă (de exemplu, pe baza primei ecuații).
2. Pe baza primelor două ecuații deducem că și situația în care doi dintre acești multiplicatori sunt nuli este imposibilă.
3. Dacă toți multiplicatorii sunt nenuli, obținem din ecuațiile 3, 4, 5, un sistem incompatibil.
4. Dacă $\alpha = 0, \beta, \gamma \neq 0$ obținem

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

Aceste valori verifică toate ecuațiile și inecuațiile sistemului de mai sus, deci $(\frac{115}{2}; 50)$ este soluție a problemei. Observăm că valoarea funcției obiectiv pentru problema inițială (de maxim) în acest punct este 190 (deci valoarea profitului este de 190 de Euro).

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

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

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

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

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

asociat restricției egalitate (singura activă) este $\mu = 4$. Deci, dacă se înlocuiește 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.

Exercițiul 2.18 *Determinați dimensiunile unei cutii de arie totală minimă având volumul $v > 0$. Interpretați valoarea multiplicatorului asociat restricției în punctul soluție.*

Exercițiul 2.19 *Determinați dimensiunile unui vas cilindric care are volumul $v > 0$ și pentru care aria este minimă. Dați o interpretare valorii multiplicatorului asociat restricției în punctul soluție.*

2.3 O selecție de probleme

În această secțiune prezentăm o selecție de probleme care sunt ivite din chestiuni aplicative și care pot fi tratate cu metodele studiate până în acest moment.

Problema 2.20 *Se dă un elipsoid de ecuație*

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

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

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

$$\max 8xyz$$

cu restricțiile

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

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

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

Avem

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

Înmulțind două câte două ecuațiile deducem

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

Este clar că restricția este activă în punctul soluție (altfel, gradientul ar trebui să 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 generate) și deci $2^{-1}N(N-1)(N-2)$ unghiuri, notate θ_i . Funcție 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 $N = 3$ și $N = 4$.

Soluție Pentru $N = 3$, trebuie determinate soluțiile globale ale problemei

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

cu restricțiile $\theta_1 + \theta_2 + \theta_3 = \pi$, $\theta_i \geq 0$, $i \in \overline{1,3}$. Evident, problema are soluție. Dacă $(\theta_1, \theta_2, \theta_3)$ este soluție, 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 și făcând calculele obținem că pentru soluție 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 soluție 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 configurația optimală este un tetraedru regulat. Este de notat și 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 optime 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 $(x_i)_{i \in \overline{1,n}}$ și $(y_i)_{i \in \overline{1,n}}$, unde $n \in \mathbb{N} \setminus \{0\}$ și că este natural să ne gândim că ele sunt legate printr-o relație 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 $f : \mathbb{R}^3 \rightarrow \mathbb{R}$,*

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

Să se studieze această problemă.

Soluție Dacă notăm matricea

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

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

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

Apoi,

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

Astfel,

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

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

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 soluția 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 $\mathcal{D}$ mulțimea elementelor din $\mathbb{R}^n$ cu toate coordonatele distincte și definim $f : \mathcal{D} \rightarrow \mathbb{R}$ 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 își 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 funcția polinomială $t \mapsto \prod_{i=1}^n (t - a_i)$. Să se arate că

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

și

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

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

(iv) Particularizați pentru $n = 2$ și $n = 3$.

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

Dacă $x \in \mathcal{D}$ și $\|x\| \rightarrow \infty$, este clar că $f(x) \rightarrow \infty$. Dacă $x \in \mathcal{D}$ și $x \rightarrow y \in \text{Fr } \mathcal{D}$ din nou, $f(x) \rightarrow \infty$. Deci, a minimiza pe f pe $\mathcal{D}$ revine la a minimiza pe f pe o mulțime 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 și f este continuă, deducem că f își atinge minimumul global pe $\mathcal{D}$. Observăm de asemenea că dacă $a = (a_1, \dots, a_n)$ este un astfel de punct atunci orice alt punct având aceleași coordonate dar în altă ordine este minim global 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 relației $H(t) = (t - a_i)Q_i(t)$ 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

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

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

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

Combinând toate aceste relații găsim

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

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

(iii) Se observă că polinomul H este caracterizat de relația $H'' - 4tH' = -4nH$. Dacă scriem $H(t) = t^n + c_{n-1}t^{n-1} + \dots + c_1t + c_0$, atunci relația de mai sus permite determinarea coeficienților lui H . Apoi, rezolvarea ecuației $H(t) = 0$, dă cele n componente ale soluției a . Așadar problema are $n!$ soluții ce se obțin de maniera indicată.

(iv) Pentru $n = 2$, găsim $H(t) = t^2 - 4^{-1}$ care se anulează în 2^{-1} și -2^{-1} . Deci minimele sunt $(2^{-1}, -2^{-1})$ și $(-2^{-1}, 2^{-1})$, 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 și $-2^{-1}\sqrt{3}$ și $-2^{-1}\sqrt{3}$. Deci minimele sunt tripletele obținute prin permutarea acestor numere. Valoarea minimă a lui f este $2^{-1} \cdot 3 - \ln(3 \cdot 4^{-1} \cdot \sqrt{3})$. $\square$

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 și 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 funcția $v \mapsto \inf \{E(a, b, c) \mid abc = v\}$ este derivabilă și să se comenteze legătura dintre derivata acestei funcții și valoarea multiplicatorului Lagrange al problemei.

Soluție Folosind metoda obișnuită se obține soluția $(\sqrt[3]{v}, \sqrt[3]{v}, \sqrt[3]{v})$, deci cubul de volum v . Multiplicatorul este $\mu = 2Cv^{-\frac{5}{3}}$. Apoi, funcția $v \mapsto \inf \{E(a, b, c) \mid abc = v\}$, notată φ este $\varphi(v) = 2Cv^{\frac{2}{3}}$, este derivabilă și $\varphi'(v) = -\mu$, adică rezultatul așteptat, având în vedere semnificația practică a multiplicatorilor. $\square$

Problema 2.25 Se înfășoară o cutie paralelipipedică cu o panglică de lungime $\ell > 0$. Se trece panglica de câte două ori pe două dintre fețele paralelipipedului și de 4 ori pe celelalte două. Să se determine cutia de volum maximal care poate fi înfășurată în acest fel. Studiați de asemenea stabilitatea valorii optime în raport cu data ℓ .

Soluție Cu notațiile evidente, avem de maximizat abc cu restricțiile $2a + 2b + 4c \leq \ell$, $a, b, c \geq 0$. Ultimele restricții pot fi considerate inactice și din nou discuția este cea obișnuită. Se obține soluția

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

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

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

Soluție Notăm triunghiul de la bază cu ABC și înălțimea cu $h > 0$. Notăm de asemenea cu a, b, c laturile triunghiului ABC , cu H piciorul înălțimii pe baza ABC și cu h_1, h_2, h_3 distanțele 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 restricția

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

O analiză de tipul celor deja utilizate arată că problema are soluție și că soluția 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 știe că există un unic punct în C care realizează minimumul cerut, însă determinarea și caracterizarea acestuia reprezintă o problemă deschisă. $\square$

Problema 2.27 În spațiul afin euclidian de dimensiune 3 cu reperul $(O; i, j, k)$ considerăm 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 conțin acest punct în interior. Considerăm problema de determinării triunghiului de arie minimă. Să se arate că problema are soluție, să se determine teoretic soluția și să se calculeze în cazurile $(a, b, c) = (1, 1, 1)$ și $(a, b, c) = (1, 4, 6)$.

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

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

Fie Ω proiecția 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 funcției $f : \mathbb{R}^3 \rightarrow \mathbb{R}$, $f(u, v, w) = u^2v^2 + u^2w^2 + v^2w^2$ cu restricțiile

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

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

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

astfel condiția de coercivitate este satisfăcută. Deci problema admite soluție globală.

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

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

Este clar că are loc condiția de liniară independență în toate punctele fezabile. Problema nu e convexă, deci pentru o soluție (u, v, w) există $\lambda \geq 0$ astfel încât

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

Se obține sistemul

$$2u(v^2 + w^2) - \lambda \frac{a}{u^2} = 0$$

$$2v(u^2 + w^2) - \lambda \frac{b}{v^2} = 0$$

$$2w(u^2 + v^2) - \lambda \frac{c}{w^2} = 0.$$

Putem scrie

$$2u^2(v^2 + w^2) = \lambda \frac{a}{u}, \quad 2v^2(u^2 + w^2) = \lambda \frac{b}{v}, \quad 2w^2(u^2 + v^2) = \lambda \frac{c}{w},$$

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

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

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 ecuații 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

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

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

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

Cum $h(u, v, w) = 0$, $t > 0$ este soluție a ecuației

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

Este ușor de văzut că această ecuație are o singură rădăcină strict pozitivă. Deci determinarea soluției problemei depinde de determinarea acestei rădăcini.

Evident, pentru $(a, b, c) = (1, 1, 1)$, se observă că $t = 3$ este rădăcina căutată și $u = v = w = 3$ este soluția problemei.

Pentru $(a, b, c) = (1, 4, 6)$ trebuie găsită rădăcina t cu aproximație (a se Exercițiul 3.14). $\square$

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

Soluție Fie

$$P(z) = a_n z^n + a_{n-1} z_{n-1} + \dots + a_1 z + a_0,$$

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

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

Considerăm funcția

$$f(z) = |P(z)|^2.$$

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

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

Această condiție de coercivitate, împreună cu continuitatea asigură faptul că f este mărginită inferior și își 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 ,

$$R(z) = b_n z^n + \dots + b_1 z + b_0, \quad b_n \neq 0.$$

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

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

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

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

În plus, știm 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 $k \in \overline{1, s-1}$ și $\varphi_\theta^{(s)}(0) = 2s! \operatorname{Re}(b_s \overline{b_0} e^{is\theta})$. Din condițiile de optimalitate cunoscute, trebuie ca primul ordin k pentru care $\varphi_\theta^{(k)}(0) \neq 0$ să fie par și $\varphi_\theta^{(k)}(0) > 0$. Aceasta înseamnă că s este par și că $\operatorname{Re}(b_s \overline{b_0} e^{is\theta}) \geq 0$ pentru orice $\theta \in [0, 2\pi)$. Cum $b_s \overline{b_0} \neq 0$, putem scrie acest număr sub forma $u + iv$ cu $u, v \in \mathbb{R}$ și $u^2 + v^2 \neq 0$. Atunci

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

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

Problema 2.29 Fie $n \in \mathbb{N} \setminus \{0\}$ și $(\alpha_i)_{i \in \overline{1, n}}$, β numere reale strict pozitive. Să se rezolve problema maximizării expresiei

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

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

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

2.4 Un exemplu, două întrebări

Să considerăm următoarea problemă de optimizare cu restricții.

Exercițiul 2.30 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 pași:

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

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

III. se aplică Teorema Krush-Kuhn-Tucker și se determină punctele care sunt potențiale soluții, iar odată cu ele, multiplicatorii Lagrange asociați, obținându-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 soluția globală, $x = \left(-\sqrt{\frac{3}{2}}, -\sqrt{\frac{3}{2}}, 1\right)$.

Studiind acești pași și calculele implicate, ne punem următoarele întrebări:

- este imperios necesară etapa (relativ dificilă din punct de vedere tehnic) a verificării unei condiții 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 Condițiile Fritz John

Așa cum am văzut, pentru problema (P) cu restricții funcționale, avem condițiile de optimalitate date de Teorema Karush-Kuhn-Tucker (Teorema 2.1.1) care pentru aplicare necesită verificarea unei condiții de calificare. În multe cărți (cu profil mai cu seamă ingineresc) această teoremă este aplicată fără verificarea unei asemenea condiții. Evident, acest lucru este greșit, dar există unele motive, pe care le discutăm mai jos, pentru care o asemenea abordare conduce totuși, în multe situații, la rezultate corecte.

Prezentăm deci unele condiții de optimalitate, numite condițiile Fritz John, în care funcția obiectiv și funcțiile ce definesc restricțiile sunt implicate nediferențiat. Aceste condiții sunt apropiate ca formă de condițiile Karush-Kuhn-Tucker dar diferă printr-un aspect care se dovedește esențial în determinarea punctelor de extrem.

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

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

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

și

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

Demonstrație Fie $\delta > 0$ astfel încât $D(\bar{x}, \delta) \subset U$ și pentru orice $x \in M \cap D(\bar{x}, \delta)$, $f(\bar{x}) \leq f(x)$. Pentru fiecare $k \in \mathbb{N}^*$ considerăm funcția $\varphi_k : D(\bar{x}, \delta) \rightarrow \mathbb{R}$ dată prin

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

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

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

Din faptul că (x_k) este un șir mărginit, iar f este continuă pe $D(\bar{x}, \delta)$, deducem că $(f(x_k))$ este un șir mărginit. Făcând $k \rightarrow \infty$ în relația de mai sus, obținem

$$\begin{aligned} \lim_{k \rightarrow \infty} \sum_{i=1}^n (g_i^+(x_k))^2 &= 0 \\ \lim_{k \rightarrow \infty} \sum_{j=1}^m (h_j(x_k))^2 &= 0. \end{aligned}$$

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

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

de unde, prin trecere la limită,

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

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

O observație esențială este că φ_k este diferențiabilă întrucât funcțiile (scalare) care ar putea pune probleme, i.e., $g_i^+(x)$, sunt ridicate la pătrat și $\nabla (g_i^+(\cdot))^2(x) = 2g_i^+(x)\nabla g_i(x)$. De fapt, $(g_i^+(x))^2$ este compunerea dintre funcția scalară derivabilă $t \rightarrow (t^+)^2$ și funcția g_i .

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

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

Punând aceste lucruri împreună, putem scrie

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

pentru orice k suficient de mare. Notăm pentru $i \in \overline{1, n}$, $j \in \overline{1, m}$, $\alpha_i^k := kg_i^+(x_k)$, $\beta_j^k := kh_j(x_k)$ și $\gamma^k = \sqrt{1 + \sum_{i=1}^n (\alpha_i^k)^2 + \sum_{j=1}^m (\beta_j^k)^2}$. Este clar că $\gamma^k > 0$ și 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

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

deci șirurile (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 subșiruri (indexate la fel) convergente în $\mathbb{R}$ la niște scalari pe care îi notăm

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

și care nu pot fi simultan zero. Pozitivitatea șirurilor (λ_0^k) , (λ_i^k) ($i \in \overline{1, n}$) atrage pozitivitatea limitelor $\lambda_0, \lambda_1, \lambda_2, \dots, \lambda_n$. Acum, împărțind relația (2.3) la γ^k , se obține

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

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

$$0 < g_i(x_k) \rightarrow g_i(\bar{x}) \leq 0$$

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

Dacă comparăm Teorema 2.1.1 cu Teorema 2.5.1 observăm următoarea diferență: scalarul asociat funcției obiectiv este de această dată nenul, iar acesta este inconvenientul major al teoremei de mai sus și anume că nu elimină posibilitatea ca scalarul asociat funcției obiectiv să fie nul. Astfel, există riscul ca prea multe puncte fezabile să verifice condițiile 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 condițiile Fritz-John (cu $\lambda_0 = 0$), funcția obiectiv nemaiaivând niciun rol de jucat. Condițiile de calificare prezente în Teorema 2.1.1 evită tocmai situația $\lambda_0 = 0$. De exemplu, având în vedere Observația 2.1.2, pentru a demonstra că cele două condiții din respectivul rezultat sunt de calificare, este suficient să arătăm acest lucru pentru condiția Mangasarian-Fromovitz. Acest lucru este evident dacă aplicăm Teorema 2.5.1 și raționăm prin reducere la absurd: dacă $\lambda_0 = 0$ atunci

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

Înmulțind scalar cu vectorul u din condiția Mangasarian-Fromovitz deducem că

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

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

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

iar condiția de liniară independență asupra gradientilor $\{\nabla h_j(\bar{x}) \mid j \in \overline{1, m}\}$ atrage faptul că $\mu_j = 0$ pentru orice $j \in \overline{1, m}$. Reunind aceste deducții ajungem la contradicție cu relația $|\lambda_0| + \|\lambda\| + \|\mu\| \neq 0$. Deci $\lambda_0 \neq 0$.

Ilustrăm modul de funcționare a Teoremei 2.5.1 prin câteva exerciții concrete.

Exercițiul 2.31 Fie problema de minimizare a funcției $f : \mathbb{R}^2 \rightarrow \mathbb{R}$,

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

cu restricția $g(x) \leq 0$, unde $g : \mathbb{R}^2 \rightarrow \mathbb{R}^4$,

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

- (i) Să se arate, folosind un desen, că $(2, 1)$ este soluție a problemei.
- (ii) Să se arate că au loc condițiile Fritz John în acest punct.
- (iii) Studiați îndeplinirea condițiilor Fritz John în $(0, 0)$ și concluzionați.

Soluție (i) Se poate observa grafic că $\bar{x} = (2, 1)$ este soluție a problemei desenând mulmea punctelor fezabile și interpretând funcția obiectiv ca fiind pătratul distanței de la (x_1, x_2) la $(3, 2)$.

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

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

Obținem $\lambda_1 = \frac{1}{3}\lambda_0$ și $\lambda_2 = \frac{2}{3}\lambda_0$, deci alegând $\lambda_0 > 0$, prima condiție Fritz John este îndeplinită.

(iii) Să vedem acum dacă punctul $x = (0, 0)$ satisface condițiile. 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 ecuația

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

nu admite soluție $(\lambda_0, \lambda_3, \lambda_4)$ nenulă cu toate componentele pozitive. Deci x nu îndeplinește condițiile Fritz John, așadar nu este soluție pentru problema dată. $\square$

Exercițiul 2.32 Fie problema de minimizare a funcției $f : (0, \infty) \times (0, \infty) \rightarrow \mathbb{R}$, $f(x_1, x_2) = -2x_2$ cu restricția $g(x) \leq 0$, unde $g : \mathbb{R}^2 \rightarrow \mathbb{R}^3$, $g(x_1, x_2) = (x_1 - x_2 - 2, -x_1 + x_2 + 2, x_1 + x_2 - 6)$.

- (i) Desenați mulțimea punctelor fezabile și determinați grafic soluția.
- (ii) Să se arate că toate punctele fezabile satisfac condițiile Fritz John.
- (iii) Care sunt punctele în care au loc condițiile Karush-Kuhn-Tucker. Comentați

Soluție (i) Mulțimea M a punctelor fezabile este segmentul $[(2, 0), (4, 2)] \setminus \{(2, 0)\}$ și soluția este $\bar{x} = (4, 2)$.

(ii), (iii) Este ușor de observat că orice punct fezabil satisface condițiile Fritz John, dar soluția este singurul pentru care putem alege $\lambda_0 \neq 0$. Într-adevăr, dacă x este un punct fezabil diferit de soluție, atunci $\lambda_3 = 0$, $\lambda_1 = \lambda_2$ și $\lambda_0 = 0$. Deci, în acest caz, aplicarea condițiilor Karush-Kuhn-Tucker determină soluția, ceea ce nu se întâmplă pentru condițiile Fritz John. Observăm că toate restricțiile sunt afine, deci se poate aplica Teorema Karush-Kuhn-Tucker. $\square$

Exercițiul 2.33 Fie problema minimizării funcției $(x_1^2 + x_2^2)$ cu restricția $x_1^2 - (x_2 - 1)^3 = 0$.

(i) Să se arate că problema are soluție.

(ii) Care sunt punctele în care au loc condițiile Fritz John? Comentati

Soluție (i) Cum funcția obiectiv este coercivă, problema are soluție globală.

(ii) Se poate cu ușurință verifica că singurul punct care satisface condițiile Fritz John este $(0, 1)$. Deci acest punct este soluția problemei. $\square$

Exercițiul 2.34 Fie $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ și $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ definite prin $f(x_1, x_2) = x_1$ și $g(x_1, x_2) = (-x_2 + (1 - x_1)^3, x_2)$.

(i) Determinați soluția folosind structura particulară a problemei.

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

Soluție (i) Se verifică faptul că $\bar{x} = (1, 0)$ este soluție pentru problema asociată

(ii) Condițiile Fritz John sunt îndeplinite cu $\lambda_0 = 0$. Condiția (2.1) nu are loc. $\square$

Punem acum în lumină și o situație în care putem asigura condiții suficiente de optimalitate.

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

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

și

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

Dacă mulțimea de vectori

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

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

Demonstrație. Presupunem, prin reducere la absurd, că $\bar{x}$ nu este soluție (locală) a problemei (P). Deci, există un șir (x_k) de puncte fezabile astfel încât $x_k \rightarrow \bar{x}$ și $f(x_k) < f(\bar{x})$ pentru orice k . Atunci, pentru orice k , $x_k \neq \bar{x}$ și

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

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

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

Împărțim în fiecare relație de mai sus cu (t_k) și prin trecere la limită avem

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

Cum, din condițiile Fritz John, avem

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

deducem că

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

Din ipoteza că mulțimea de vectori

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

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

2.6 Condiții de optimalitate de ordinul al doilea

Condițiile de optimalitate studiate până acum în cazul cu restricții sunt condiții de ordinul I întrucât implică doar diferențialele de ordinul întâi ale funcțiilor ce definesc problema. Uneori, aceste condiții nu rezolvă complet problema mai ales atunci când avem puncte care îndeplinesc condițiile necesare Karush-Kuhn-Tucker, dar nu sunt soluții globale. A decide dacă respectivele puncte sunt sau nu soluții (locale) este o chestiune la care condițiile menționate nu pot răspunde. Studiem mai întâi un exemplu care ilustrează aceste comentarii.

Exemplul 2.6.1 Fie problema minimizării funcției de două variabile reale $f(x, y) = -\frac{1}{2}(x+1)^2 - \frac{1}{2}(y+1)^2$ cu restricțiile $g_1(x, y) = x^2 + y^2 - 2 \leq 0$, $g_2(x, y) = y - 1 \leq 0$. Se constată cu ușurință că mulțimea punctelor fezabile este compactă, deci problema are soluție globală. Sistemul de restricții este convex și are loc condiția Slater. Se aplică Teorema Karush-Kuhn-Tucker și obținem că posibilele soluții 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 funcției în aceste puncte arată că $(1, 1)$ este soluție globală.

Pentru celelalte 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 ușor 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:

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

și 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 discuție va trebui să aplicăm alte rezultate.

În continuarea acestei secțiuni obținem condiții de optim de ordinul al doilea pentru problema cu restricții funcționale, așa încât vom presupune că funcțiile sunt de clasă C^2 .

Teorema 2.6.2 (Condiție necesară de ordinul II) Presupunem că $\bar{x}$ este punct soluție pentru problema (P) și că au loc condițiile Karush-Kuhn-Tucker cu multiplicatorii λ, μ . Dacă gradientii activi,

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

sunt liniar independenți, atunci pentru orice direcție $u \in \mathbb{R}^p$ ce satisface

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

avem

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

Demonstrație. Considerăm funcția $\Phi : \mathbb{R}^p \rightarrow \mathbb{R}^{\text{card } A(\bar{x})+m}$ ce are drept componente funcțiile g_i , $i \in A(\bar{x})$, h_j , $j = \overline{1, m}$. În baza condiției de liniară independență a gradientilor activi, rezultă că $\nabla\Phi(\bar{x})$ este surjectiv, deci în baza Teoremei lui Lyusternik vom avea că

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

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

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

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

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

Împărțind la t_k^2 și trecând la la limită, având în vedere și faptul că $\|u_k\|^2 \alpha_k \rightarrow 0$, obținem concluzia dorită. $\square$

Putem obține chiar o rafinare a acestui rezultat făcând observația că pot exista restricții cu inegalități active pentru care multiplicatorul să fie 0. Considerăm mulțimea indicilor activi strict dată prin

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

Corolarul 2.6.3 *Presupunem că $\bar{x}$ este punct soluție pentru problema (P) și că au loc condițiile Karush-Kuhn-Tucker cu multiplicatorii λ, μ . Dacă vectorii*

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

sunt liniar independenți, atunci pentru orice direcție u ce satisface

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

avem

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

Demonstrație. Se repetă demonstrația anterioară, considerând doar acele restricții g_i pentru care multiplicatorii sunt strict pozitivi. $\square$

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

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

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 independenți, 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 condiții este vectorul nul. Evident, pentru acest vector, are loc condiția din concluzia teoremei, deci în continuare nu putem să decidem.

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

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

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

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 și în acest caz

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

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

Exemplul 2.6.5 Similar, punctele rămase în discuție la Exemplul 2.30 nu sunt puncte de minim local pentru că nu verifică condițiile necesare de ordinul al doilea.

Prezentăm în continuare o condiție suficientă de ordinul al doilea.

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

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

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

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

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

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

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

Observăm că $x_k \neq \bar{x}$ pentru orice k . Definind $u_k = \frac{x_k - \bar{x}}{\|x_k - \bar{x}\|}$, putem presupune, trecând eventual la un subsir, că $u_k \rightarrow u$, cu $\|u\| = 1$. Cum $\nabla_x L(\bar{x}, \lambda, \mu) = 0$ din condițiile Karush-Kuhn-Tucker, raționând ca mai sus, găsim un șir (α_k) cu limita 0 astfel încât pentru orice k suficient de mare,

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

de unde, împărțind la $\|x_k - \bar{x}\|^2$ și trecând la limită pentru $k \rightarrow \infty$, obținem

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

Pe de altă parte, există din nou niște șiruri (β_k) , (γ_k^i) , (ε_k^j) (unde $i \in A(\bar{x})$, $j \in \overline{1, m}$) cu limita 0 astfel încât

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

Împărțind inegalitățile prin $\|x_k - \bar{x}\|$ și trecând la limită pentru $k \rightarrow \infty$, obținem

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

Dacă ar exista un $i \in A_s(\bar{x})$ astfel încât $\nabla g_i(\bar{x})(u) < 0$, atunci, multiplicând relațiile anterioare respectiv cu 1, λ_i , $i \in A(\bar{x})$, μ_j , $j \in \overline{1, m}$ și adunându-le, obținem

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

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

Obținem următorul corolar.

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

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

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

Exemplul 2.6.8 (Utilizarea condițiilor de optimalitate) Fie problema minimizării funcției de două variabile reale $f(x, y) = -(x+1)^2 - (y+1)^2$ cu restricțiile $g_1(x, y) = x^2 + y^2 - 3 \leq 0$, $g_2(x, y) = -x^2 + 2y \leq 0$. Se constată cu ușurință că mulțimea punctelor fezabile este compactă, deci problema are soluție globală. O imagine geometrică asupra problemei oferă unele informații importante, dar dorim să utilizăm condițiile analitice pe care le avem la dispoziție și care funcționează și în contexte în care nu putem vizualiza structura problemei sau nu putem extrage informații din aceasta. Se verifică condiția de liniară independență în toate punctele fezabile. Se aplică Teorema Karush-Kuhn-Tucker și obținem că posibilele soluții are problemei sunt

$$(-1, -1) \text{ cu multiplicatorii } (\lambda_1, \lambda_2) = (0, 0),$$

$$(\sqrt{2}, 1) \text{ cu multiplicatorii } (\lambda_1, \lambda_2) = \left(\frac{3}{2} + \frac{1}{2\sqrt{2}}, \frac{1}{2} - \frac{1}{2\sqrt{2}}\right),$$

$$(-\sqrt{2}, 1) \text{ cu multiplicatorii } (\lambda_1, \lambda_2) = \left(\frac{3}{2} - \frac{1}{2\sqrt{2}}, \frac{1}{2} + \frac{1}{2\sqrt{2}}\right),$$

$$\left(-\sqrt{\frac{3}{2}}, -\sqrt{\frac{3}{2}}\right) \text{ cu multiplicatorii } (\lambda_1, \lambda_2) = \left(1 - \sqrt{\frac{2}{3}}, 0\right),$$

$$(a, b) \text{ cu multiplicatorii } (\lambda_1, \lambda_2) = (0, \theta),$$

$$\text{unde } \theta \text{ este singura rădăcină reală a ecuației } \theta^3 + \theta^2 - \theta - \frac{3}{2} = 0,$$

$$a = -\frac{1}{\theta + 1}, \ b = \theta - 1.$$

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

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

Considerăm acum punctul $(-\sqrt{2}, 1)$ și încercăm să verificăm concluzia Teoremei 2.5.2. Ambele restricții 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

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

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

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

Pentru punctul $(-\sqrt{\frac{3}{2}}, -\sqrt{\frac{3}{2}})$ 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 independenți, deci nu formează sistem de generatori. Nici pentru acest punct nu putem decide deocamdată.

În sfârșit, luăm punctul (a, b) de mai sus pe care îl cunoaștem doar cu aproximație. A doua restricție este activă în acest punct. Ținând seama de relațiile dintre a, b și 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 independenți. Din nou, nu putem decide.

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

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

Pentru $(-\sqrt{\frac{3}{2}}, -\sqrt{\frac{3}{2}})$, cum $\nabla g_1(-\sqrt{\frac{3}{2}}, -\sqrt{\frac{3}{2}}) \neq (0, 0)$ are loc condiția de liniară independență din Teorema 2.6.2. Mulțimea direcțiilor ce trebuie considerate este

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

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 mulțimea de mai sus. Deducem că $(-\sqrt{\frac{3}{2}}, -\sqrt{\frac{3}{2}})$ nu este soluție.

Pentru (a, b) , $\nabla g_2(a, b) \neq (0, 0)$, deci loc condiția de liniară independență din Teorema 2.6.2. Mulțimea direcțiilor pe care le avem în vedere este

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

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 mulțimea direcțiilor de mai sus. Deducem că (a, b) nu este soluție.

În sfârșit mai facem observația că punctelor ce sunt soluție, adică $(\sqrt{2}, 1)$ și $(-\sqrt{2}, 1)$, li se poate aplica Teorema 2.6.6 pentru a concluziona că sunt de fapt soluții stricte de ordin 2.

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

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

deci

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

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

Exercițiul 2.35 Să se studieze problema minimizării și problema maximizării locale a expresiei xyz cu restricția $xy + xz + yz = 8$.

Exercițiul 2.36 Să se studieze problema minimizării și problema maximizării locale a expresiei xyz cu restricția $x + y + z = 1$.

Exercițiul 2.37 Să se studieze problema minimizării și problema maximizării locale a expresiei $y^4 - x^4$ cu restricția $x - y^2 \leq 0$.

Capitolul 3

Solvabilitate și aproximare

3.1 Solvabilitatea problemelor de optimizare

Fie $P \in \mathbb{R}[X]$ un polinom de grad mai mare sau egal decât 5. Ecuația $P(x) = 0$ cu necunoscuta $x \in \mathbb{R}$ poate fi redusă la problema de optimizare fără restricții $\min P^2(x)$ cu $x \in \mathbb{R}$. Așa cum știm, ecuația 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 ecuații algebrice sau a unui sistem de ecuații (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 $f_i : \mathbb{R}^p \rightarrow \mathbb{R}$ și $a_i \in \mathbb{R}$ cu $i \in \overline{1, q}$; cu aceste date, problema este aceea de a găsi $x \in \mathbb{R}^p$ astfel încât

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

Din nou, această problemă poate fi ușor convertită într-o problemă de optimizare fără restricții:

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

Bineînțeles, 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 știm 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 a exista o metodă universală care să le rezolve pe toate. În afară de aceste aserțiuni care sunt mai degrabă euristice, vom arăta riguros că, într-adevăr, așa stau lucrurile.

Astfel, vom începe în capitolul următor studiul unor proceduri numerice. 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 știm o parte a problemei (P) pe care o tratăm, parte pe care o numim model și care constă de obicei din formularea problemei, a clasei de componente funcționale ș.a. Pentru a încadra problema (P) într-o clasă (C) și a o rezolva, metoda trebuie să colecteze informații

specifice despre (P) . Procesul de colectare a datelor se numește oracol. Un oracol (O) răspunde unor întrebări succesive ale metodei. Astfel, metoda (M) încearcă să rezolve problema colectând și prelucrând răspunsurile oracolului.

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

Pentru o metodă numerică, accepțiunea sintagmei "rezolvare a problemei" este aceea de determinare a unei soluții aproximative cu o acuratețe $\varepsilon > 0$ 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, informații inițiale) și o acuratețe (toleranță) $\varepsilon > 0$.

Inițializare: $k = 0$, $I_{-1} = \emptyset$, unde k este contorul de iterații, iar I_k este mulțimea de informații acumulate la pasul k .

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

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

Astfel putem vorbi despre:

- complexitatea analitică: numărul de invocări ale oracolului necesar pentru rezolvarea problemei cu acuratețea ε ;
- complexitatea aritmetică: numărul de operații aritmetice necesar pentru rezolvarea problemei cu acuratețea ε .

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

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

- oracol de ordin zero: returnează valoarea funcțiilor (derivative-free, în terminologie engleză);
- oracol de ordin întâi: returnează valoarea funcțiilor și a gradientilor de ordinul întâi (first-order gradient methods);
- oracol de ordinul al doilea: returnează valoarea funcțiilor și a gradientilor de ordinul întâi și al doilea (second-order gradient methods).

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

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

și $f : \mathbb{R}^p \rightarrow \mathbb{R}$ o funcție 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 soluție globală (Teorema lui Weierstrass), iar metoda pe care o folosim este metoda rețelei (grilei) uniforme care constă din următoarele:

1. Pentru n dat, calculăm punctele:

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

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

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

3. Returnăm punctul selectat, notat $\bar{x}$, și $f(\bar{x})$.

Așadar, această metodă împarte mulțimea punctelor fezabile într-o rețea uniformă (echidistantă), calculează minimul funcției în nodurile acestei rețele și returnează această valoare. Deci, este o metodă cu oracol de ordin zero și fără influența informației acumulate în punctele testate. Dorim să determinăm eficiența acestei metode.

Teorema 3.1.1 *Dacă f^* este valoarea minimă globală a problemei (P_L) și $\bar{x}$ este punctul returnat de metoda rețelei uniforme, atunci*

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

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

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

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

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

Cum v este nod al rețelei, avem

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

adică concluzia. □

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

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

Obținem consecința următoare.

Corolarul 3.1.2 *Complexitatea analitică a problemei (P_L) pentru metoda rețelei uniforme este cel mult*

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

Demonstrație. Fie $n = \left\lceil \frac{L}{2\varepsilon} \right\rceil + 1$. Deci $n > \frac{L}{2\varepsilon}$ și din teorema de mai sus avem

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

Deci, invocând oracolul de cel mult $(n+1)^p = \left(\left\lceil \frac{L}{2\varepsilon} \right\rceil + 2 \right)^p$ ori obținem soluția problemei cu acuratețea dorită. $\square$

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

Teorema 3.1.3 *Fie $\varepsilon < \frac{L}{2}$. Atunci, pentru orice metodă cu oracol de ordin zero, complexitatea analitică a problemei (P_L) este cel puțin $\left\lceil \frac{L}{2\varepsilon} \right\rceil^p$.*

Demonstrație. Fie $n = \left\lceil \frac{L}{2\varepsilon} \right\rceil$. 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 $e = (1, 1, \dots, 1)$ și astfel încât nu există puncte de control (în care să fie invocat oracolul) în mulțimea

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

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

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

Definim funcția $g : D \rightarrow \mathbb{R}$,

$$g(x) = \min \{0, L\|x - v\|_\infty - \varepsilon\}.$$

Observăm că are valoarea minimă $-\varepsilon$ și diferă de 0 doar în mulțimea

$$F = \left\{ x \in D \mid \|x - v\|_\infty < \frac{\varepsilon}{L} \right\}.$$

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), acuratețea nu poate fi mai bună de ε .

În final, arătăm și 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 $x \in F$ și $y \notin F$. Atunci

$$\begin{aligned} |g(x) - g(y)| &= \varepsilon - L \|x - v\|_\infty = L \left(\frac{\varepsilon}{L} - \|x - v\|_\infty \right) \leq L (\|y - v\|_\infty - \|x - v\|_\infty) \\ &\leq L \|x - y\|_\infty, \end{aligned}$$

ceea ce completează demonstrația afirmației făcute.

Prin urmare, pentru a fi siguri de acuratețea dorită pentru toate problemele de tip (P_L) nu putem lua mai puțin de $\left[\frac{L}{2\varepsilon}\right]^p$ puncte. $\square$

Exemplul 3.1.4 Evaluăm consecințele rezultatului de mai sus în cazul în care $L = 2$, $p = 10$, $\varepsilon = 0.01$. Pentru a rezolva problema cu acuratețea ε trebuie să invocăm oracolul (să evaluăm funcția) în cel puțin $100^{10} = 10^{20}$ puncte. Având în vedere că funcția obiectiv are $p = 10$ funcții coordonate, pentru calculul fiecăreia trebuind cel puțin o operație algebrică, înseamnă că avem de făcut cel puțin 10^{21} operații. Spre exemplu, dacă computerul pe care lucrăm este capabil să facă 10^{12} operații pe secundă, înseamnă că e nevoie de 10^9 secunde. Un an are mai puțin de $3,2 \cdot 10^7$ secunde, astfel că timpul de calcul ar fi de peste 31 de ani.

Exercițiul 3.1 Să se testeze metoda grilei uniforme printr-un program Matlab/Octave pentru funcția $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^6 - 4x^5 + 3x^3 + 6x$ desenând mai întâi graficul pentru a observa poziționarea punctelor de minim.

Pentru funcții de mai multe variabile, am văzut că algoritmi de ordin zero pot fi foarte costisitori din punct de vedere computațional. Prezentăm totuși un algoritm de tip grilă uniformă, studiul putând fi, o dată plus, instructiv.

Mai întâi prezentăm niște funcții test și ne vom ocupa de reprezentări grafice. Astfel, vom ilustra codurile preponderent pe următoarele funcții

- Funcția lui Rosenbrock, $f : \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$f(x, y) = 100(y - x^2)^2 + (1 - x)^2;$$

- $f : \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$f(x, y) = x^6 + y^4 + 4x^2y^2 - 3xy + 2x + y;$$

- $f : \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$f(x, y) = (1 + y)^3 x^2 + y^2;$$

- $f : \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$f(x, y) = 0.6 * y^4 + 5x^2y - 7y^2 + \sin(xy) - 5y;$$

- $f : \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$f(x, y) = 100 \left(\sqrt{x^2 + (y + 1)^2} - 1 \right)^2 + 90 \left(\sqrt{x^2 + (y - 1)^2} - 1 \right)^2 - (20x + 40y).$$

Implementare 3.2 De exemplu, pentru a reprezenta a treia funcție și liniile sale de nivel, folosim codul de mai jos.

```
desem_m_m_var.m (fișier 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)')
```

Exercițiul 3.4 Să se implementeze un algoritm de tip grilă uniformă pentru aceste funcții.

Discutarea teoretică a punctelor de minim pentru aceste funcții poate fi dificilă. Dacă pentru funcția Rosenbrock există un singur astfel de punct, $(1, 1)$, și acesta este exact determinat teoretic, pentru celelalte discuția poate fi mai complicată, iar în unele cazuri imposibil de finalizat.

Algoritmii de tip grilă uniformă prezentați se referă la determinarea punctelor de minim. În acest curs, vom discuta de asemenea algoritmi de determinare a zerourilor unor funcții, tocmai pentru că există o strânsă legătură între această chestiune și problemele de extrem. De altfel, orice algoritm de aproximare a zerourilor unei funcții se poate transforma într-un algoritm de aproximare a punctelor de optim prin aplicarea sa asupra gradientului funcției (conform Teoremei lui Fermat).

Un algoritm simplu de aproximare a zerourilor unei funcții reale de o variabilă, este oferit de metoda înjumătățirii intervalului. Să presupunem că avem o funcție continuă f și două numere reale $a < b$ pentru care $f(a)f(b) < 0$. Atunci, f are o rădăcină în (a, b) . Pentru ușurința expunerii, presupunem că această soluție este unică. Vom genera două șiruri (a_k) și (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 soluția căutată și iterarea se oprește. Astfel, dacă $f(a_0)f(x_0) < 0$ alegem $a_1 = a_0$ și $b_1 = x_0$, iar dacă $f(x_0)f(b_0) < 0$ alegem $a_1 = x_0$ și $b_1 = b_0$. Continuăm procedeul luând $x_1 = 2^{-1}(a_1 + b_1)$. Procedând recurent, ne apropiem cu (x_k) de soluție, înjumătățind la fiecare pas intervalul în care se găsește soluția. În general, această convergență nu este foarte rapidă, dar în unele cazuri este eficientă.

Exemplul de mai jos reflectă aceste observații.

Exemplul 3.1.5 Considerăm funcția $f : \mathbb{R}^2 \rightarrow \mathbb{R}$

$$f(x, y) = x^4 - 16x^2 + 5x + y^4 - 16y^2 + 5y.$$

Desenând graficul constatăm că funcția are patru puncte de minim local și un punct de maxim local. Vrem să găsim aceste puncte. Primul pas este să determinăm punctele critice, ceea ce, datorită simetriei în x și y , revine la rezolvarea ecuației $4u^3 - 32u + 5 = 0$. Asociem funcția $\varphi : \mathbb{R} \rightarrow \mathbb{R}$, $\varphi(u) = 4u^3 - 32u + 5$ și 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 hessiană a funcției inițiale, 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, funcția fiind coercivă, ea admite un punct de minim global. O aproximare a punctelor u_1, u_3 (de exemplu $u_1 \in \left(-3, -\frac{2\sqrt{2}}{\sqrt{3}}\right)$, $u_3 \in \left(\frac{2\sqrt{2}}{\sqrt{3}}, 3\right)$) și variația lui $\psi : \mathbb{R} \rightarrow \mathbb{R}$, $\psi(v) = v^4 - 16v^2 + 5v$ 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 funcției date, trebuie să aproximăm soluțiile ecuației $4u^3 - 32u + 5 = 0$, lucru care se poate face cu metoda descrisă mai sus.

Exercițiul 3.5 Să se implementeze metoda înjumătățirii intervalului pentru finalizarea exemplului anterior.

Exercițiul 3.6 Să se implementeze metoda înjumătățirii intervalului pentru funcția $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^3 - x - 2$.

3.2 Metode numerice, în general – algoritmi iterativi

De foarte multe ori pot apărea probleme de optimizare pentru care nu putem rezolva sistemele ce dau punctele critice ale funcției obiectiv (în cazul problemelor fără restricții) sau punctele critice ale lagrangianului (la problemele cu restricții). Pentru astfel de probleme, sunt necesari algoritmi pentru aproximarea soluțiilor. Ca de obicei, și pentru proiectarea algoritmilor, există o netă diferență între problemele cu restricții și cele fără restricții.

Toți algoritmi pe care îi vom studia sunt iterativi (adică soluția (i.e., aproximarea punctului căutat) este găsită calculând termenii unui șir până la un rang dat de un criteriu de oprire a algoritmului) și cer precizarea unui punct de start pe care îl notăm cu x_0 . Este bine ca acest punct să fie el însuși o cât mai bună aproximare a soluției căutate (mai ales dacă aceasta nu este unică). De exemplu, funcția $f : \mathbb{R} \rightarrow \mathbb{R}$,

$$f(x) = \frac{x^6}{6} - \frac{x^3}{3} - x^2$$

are două puncte de minim și 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 ecuației sau pentru găsirea punctelor de extrem) să găsească o aproximare a aceluia punct.

Revenind la discuția generală, după alegerea lui x_0 , algoritmul generează un șir de iterații $(x_k)_{k \in \mathbb{N}}$ care au scopul de a se apropia de soluție (soluție a unei ecuații sau punct de optim). Generarea acestui șir se va opri atunci când nu se mai pot face progrese în încercarea de apropiere

de soluție (conform regulii interne a algoritmului de generare a iterațiilor) sau când a fost atinsă o anumită acuratețe dinainte stabilită (care se definește printr-un criteriu de oprire). Orice algoritm trebuie să implementeze o regulă internă de generare a unei noi iterații utilizând iterațiile deja existente. În general, fiecare nouă iterație trebuie să realizeze o apropiere de soluție, dar există și algoritmi nemonotoni, pentru care descreșterea nu trebuie să se producă neapărat la fiecare pas.

În termeni matematici, un algorithm iterativ poate fi descris ca având forma unei recurențe de tipul

$$\begin{aligned}x_{k+1} &= A(x_k), \quad \forall k \geq 0, \\x_0 &\in \mathbb{R}^p,\end{aligned}$$

unde $A : \mathbb{R}^p \rightarrow \mathbb{R}^p$ este o funcție. Totuși, în practică, datorită implementării pe diferite computere, de către diferiți programatori care pot interpreta diferit secvențele sintactice ale programului ce implementează algoritmul, datorită acumulării rotunjirilor numerice care și ele pot diferi în cazul unor software-uri diferite, este mai realist să considerăm A ca fiind o multifuncție. De fapt, în realitate, dacă A este funcția de mai sus, relația dintre x_{k+1} și x_k este $x_{k+1} = A(x_k) + \varepsilon_k$, unde ε_k este un element provenit din diferențele 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 mulțime. În acest fel, proprietățile generale ce pot fi deduse pentru un algoritm astfel modelat se păstrează pentru diferite implementări ale acestuia și chiar pentru algoritmi similari. Astfel, cu această abordare, algoritmul general de mai sus capătă forma

$$\begin{aligned}x_{k+1} &\in A(x_k), \quad \forall k \geq 0, \\x_0 &\in \mathbb{R}^p.\end{aligned}\tag{3.2}$$

Definiția 3.2.1 O multifuncție $A : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ se numește închisă în $\bar{x} \in D$ dacă pentru orice două șiruri convergente $(x_k), (y_k)$ cu $x_k \rightarrow \bar{x}$, $y_k \rightarrow \bar{y}$ și $y_k \in A(x_k)$ pentru orice k , are loc $\bar{y} \in A(\bar{x})$. O multifuncție închisă în toate punctele unei mulțimi se numește închisă pe acea mulțime.

Exemplul 3.2.2 1. Multifuncția $A : \mathbb{R} \rightrightarrows \mathbb{R}$ 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 $\bar{x} = 1$. Pentru a observa acest lucru este suficient să analizăm cazul șirului

$$x_k = 1 + \frac{1}{2^k}, \quad \forall k \geq 0.$$

2. Multifuncția $A : \mathbb{R} \rightrightarrows \mathbb{R}$, $A(x) = [x^2, \infty)$ este închisă pe $\mathbb{R}$.

În cadrul oricărui algoritm iterativ o funcție specifică (de obicei funcția obiectiv în cazul problemelor de optimizare) cunoaște o descreștere până în momentul când convergența (aproximarea) dorită se realizează. Aceasta sugerează definiția de mai jos.

Definiția 3.2.3 Fie $S \subset \mathbb{R}^p$ o mulțime. Considerăm un algoritm de forma (3.2). O funcție continuă $\varphi : \mathbb{R}^p \rightarrow \mathbb{R}$ se numește funcție de descreștere pentru S și algoritmul A dacă sunt satisfăcute următoarele proprietăți:

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

Exemplul 3.2.4 Funcția $\varphi : \mathbb{R} \rightarrow \mathbb{R}$, $\varphi(x) = |x|$ este o funcție de descreștere pentru $S = \{0\}$ și $A : \mathbb{R} \rightarrow \mathbb{R}$, $A(x) = 4^{-1}x$.

Teorema 3.2.5 Fie un algoritm de forma (3.2). Fie S mulțimea soluțiilor problemei studiate. Presupunem că au loc următoarele condiții:

- (i) șirul (x_k) este conținut într-o mulțime compactă;
 - (ii) A este închisă pe $\mathbb{R}^p \setminus S$;
 - (iii) există o funcție continuă $\varphi : \mathbb{R}^p \rightarrow \mathbb{R}$ de descreștere pentru S și A .
- Atunci orice punct limită al lui (x_k) se află în S .

Demonstrație Fie $\bar{x}$ un punct limită al lui (x_k) și $(x_{k_r})_r$ subșir al lui (x_k) cu limita $\bar{x}$. Cum φ este continuă,

$$\lim_r \varphi(x_{k_r}) = \varphi(\bar{x}).$$

Cum șirul $(\varphi(x_k))$ este monoton descrescător iar un subșir al său este convergent, deducem că întreg șirul este convergent la aceeași limită, $\varphi(\bar{x})$.

Presupunem, prin reducere la absurd că $\bar{x} \notin S$. Din prima ipoteză, există un punct limită, notat u , al șirului $(x_{k+1})_k$. Ca mai sus, $\lim_k \varphi(x_{k+1}) = \varphi(u)$, deci $\varphi(\bar{x}) = \varphi(u)$. Pe de altă parte, ipoteza (ii) asigură $u \in A(\bar{x})$, deci, conform (iii), $\varphi(u) < \varphi(\bar{x})$, ceea ce reprezintă o contradicție. $\square$

Observația 3.2.6 În rezultatul anterior data inițială poate fi arbitrar aleasă, deci în ipotezele date algoritmul are proprietăți globale de convergență.

Corolarul 3.2.7 În ipotezele Teoremei 3.2.5, dacă S este formată dintr-un singur punct $\bar{x}$, atunci $(x_k) \rightarrow \bar{x}$.

Demonstrație Din prima ipoteză a teoremei, șirul are un punct limită care, conform teoremei și ipotezei rezultatului de față coincide cu $\bar{x}$. Dacă șirul nu ar converge la $\bar{x}$, ar mai avea un punct limită diferit de $\bar{x}$ care ar trebui să fie din S ceea ce nu este posibil. $\square$

Exemplul 3.2.8 Considerăm problemă de minimizare a lui $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = |x|$ și algoritmul definit de $A : \mathbb{R} \Rightarrow \mathbb{R}$ 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 inițială). Într-adevăr, dacă luăm $x_0 = 4$, atunci $x_k \rightarrow 2$. Se observă că pentru $x_0 = -4$, $x_k \rightarrow 2$. Am văzut mai sus (Exemplul 3.2.2) că A nu este închisă în 1.

Exemplul 3.2.9 Considerăm problemă de minimizare a lui $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^3$ și algoritmul definit de $A : \mathbb{R} \Rightarrow \mathbb{R}$ dată prin

$$A(x) = \{-(x^2 + 1)\}$$

Acest algoritm nu este convergent pentru că pentru orice dată inițială șirul generat are limita $-\infty$.

Exemplul 3.2.10 Considerăm problem minimizării lui $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = |x - 1|$ și algoritmul definit de $A : \mathbb{R} \Rightarrow \mathbb{R}$ dată prin

$$A(x) = \left\{ -\sqrt{|x|} \right\}$$

Acest algoritm este convergent pentru orice dată inițială $x_0 \in \mathbb{R}$.

O problemă foarte importantă din punct de vedere practic atunci când este studiat un algoritm este viteza sa de convergență.

Definiția 3.2.11 (viteză de convergență) Fie $(x_k)_{k \in \mathbb{N}^*} \subset \mathbb{R}^p$ un șir convergent la $\bar{x} \in \mathbb{R}^p$ cu $x_k \neq \bar{x}$ pentru orice $k \in \mathbb{N}^*$. Spunem că:

(i) (x_k) converge liniar dacă există $r \in (0, 1)$ (numit rată de convergență) astfel încât pentru orice k suficient de mare,

$$\frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|} \leq r;$$

(ii) (x_k) converge superliniar dacă

$$\lim_k \frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|} = 0;$$

(iii) (x_k) converge pătratic dacă există $M > 0$ astfel încât pentru orice k suficient de mare,

$$\frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|^2} \leq M.$$

Evident, convergența pătratică implică convergența superliniară care, la rândul său implică convergența liniară. Pentru convergența liniară rata r este importantă pentru viteza de convergență, în timp ce pentru convergența pătratică dependența de M este mai puțin importantă. Folosind modelul oferit de definiția convergenței pătratice, putem defini convergențe de orice ordin; spunem că avem o convergență de ordin $\theta > 1$ dacă există $M > 0$ astfel încât pentru orice k suficient de mare

$$\frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|^\theta} \leq M.$$

Exemplul 3.2.12 (i) Șirul $\left(1 + \left(\frac{1}{2}\right)^k\right)$ este liniar convergent la 1 cu rata $r = \frac{1}{2}$.

(ii) Șirul $\left(1 + \frac{1}{k^k}\right)$ converge superliniar la 1.

(iii) Șirul $\left(1 + \left(\frac{1}{2}\right)^{2^k}\right)$ converge pătratic la 1.

Exemplul 3.2.13 Reamintim că dacă $I \subset \mathbb{R}$ este un interval închis $f : I \rightarrow I$ este o funcție derivabilă astfel încât derivata sa este mărginită (în modul) pe I de o constantă strict subunitară, atunci pentru orice dată inițială $x_0 \in I$ iterația Picard definită de $x_{k+1} = f(x_k)$, $k \geq 0$ este convergentă către unicul punct fix $\bar{x}$ 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 (șirul este nestaționar), atunci

$$\frac{x_{k+1} - \bar{x}}{x_k - \bar{x}} = \frac{f(x_k) - \bar{x}}{x_k - \bar{x}} \xrightarrow{k \rightarrow \infty} f'(\bar{x}) \in (-1, 1).$$

Astfel, observăm că, în general avem de-a face cu o convergență cel puțin liniară de rată $|f'(\bar{x})|$.

În cazul cel mai bun în care $f'(x) = 0$ putem avea convergențe mai bune decât convergența liniară. În general, în contextul de mai sus, dacă $f'(\bar{x}) = 0$ și f este de clasă C^2 atunci, aplicând de două ori regula lui L'Hôpital avem:

$$\lim_{x \rightarrow \bar{x}} \frac{f(x) - \bar{x}}{(x - \bar{x})^2} = \frac{f''(\bar{x})}{2},$$

deci pentru orice iterație Picard nestăționară

$$\lim_{k \rightarrow \infty} \frac{x_{k+1} - \bar{x}}{(x_k - \bar{x})^2} = \frac{f''(\bar{x})}{2},$$

adică o convergență pătratică.

Să considerăm funcția $f : [0, 1] \rightarrow [0, 1]$ dată prin $f(x) = \frac{1}{1+x^2}$. Aceasta este o contracție și are un singur punct fix care este unica soluție pozitivă a ecuației $x^3 + x - 1 = 0$ (care are valoarea aproximativă $\bar{x} \approx 0.6823$), iar șirul iterațiilor Picard satisface:

$$\frac{x_{k+1} - \bar{x}}{x_k - \bar{x}} \xrightarrow{n \rightarrow \infty} f'(\bar{x}) = \frac{-2\bar{x}}{(1 + \bar{x}^2)^2} = -2\bar{x}^3 \approx -0.6353.$$

Deci, practic pentru k suficient de mare, la fiecare pas al iterației, eroarea se multiplică (în valoare absolută) cu aproximativ 0.6353.

Să luăm și cazul funcției $f : [\sqrt{2}, \infty) \rightarrow [\sqrt{2}, \infty)$ dată prin

$$f(x) = \frac{x}{2} + \frac{1}{x}.$$

Este ușor 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 contracție și are ca unic punct fix $\bar{x} = \sqrt{2}$. Se observă că $f'(\bar{x}) = 0$ și, deci, pentru orice iterație Picard nestăționară

$$\frac{x_{k+1} - \bar{x}}{(x_k - \bar{x})^2} = \frac{1}{2x_k} \rightarrow \frac{1}{2\sqrt{2}} = \frac{f''(\bar{x})}{2},$$

deci avem o convergență pătratică.

Exercițiul 3.7 Să se implementeze acest algoritm pentru funcțiile de mai sus.

O altă clasă de viteze de convergență este definită mai jos.

Definiția 3.2.14 (r -convergență) Fie $(x_k)_{k \in \mathbb{N}^*} \subset \mathbb{R}^p$ un șir convergent la $\bar{x} \in \mathbb{R}^p$. Spunem că (x_k) converge r -liniar (r -superliniar, r -pătratic) dacă există un șir $(y_k)_{k \in \mathbb{N}^*}$ care este convergent liniar (superliniar, pătratic) la 0 astfel încât pentru orice k

$$\|x_k - \bar{x}\| \leq y_k.$$

Exemplul 3.2.15 Șirul

$$x_k = \begin{cases} 1 + \left(\frac{1}{2}\right)^k, & \text{dacă } k \text{ este par} \\ 1 + \left(\frac{1}{3}\right)^k, & \text{dacă } k \text{ este impar} \end{cases}$$

este r -liniar convergent la 1, dar nu este liniar convergent la 1.

Observația 3.2.16 Vitezele de convergență în sensul Definiției 3.2.11 implică r -convergențele corespunzătoare, în timp ce reciprocele sunt, în general, false, așa cum arată exemplul anterior.

Exercițiul 3.8 Studiați viteza de convergență a șirului $\left(\frac{1}{k!}\right)$.

Exercițiul 3.9 Fie $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ definită prin $f(x) = \|x\|^2$. Arătați că iterațiile

$$x_k = \left(1 + \frac{1}{2^k}\right) (\cos k, \sin k)$$

satisfac $f(x_{k+1}) < f(x_k)$ pentru orice $k \geq 0$. Arătați că orice punct de pe sfera (cercul) unitate este punct limită al lui (x_k) .

Exercițiul 3.10 Studiați viteza de convergență și r -convergența șirului

$$x_k = \begin{cases} \left(\frac{1}{4}\right)^{2^k}, & \text{dacă } k \text{ este par} \\ \frac{x_{k-1}}{k}, & \text{dacă } k \text{ este impar.} \end{cases}$$

3.3 Un algoritm reprezentativ: metoda lui Newton

Această secțiune se ocupă cu studiul unei celebre metode, atribuită lui Newton, de a determina cu aproximație rădăcinile unor ecuații neliniare. În plus, această metodă ilustrează foarte bine ideile generale expuse despre algoritmii iterativi în secțiunea precedentă.

Să remarcăm că aplicarea condițiilor de optimalitate transformă o problemă de optimizare în problema rezolvării unei ecuații sau a unui sistem de ecuații (de cele mai multe ori neliniare și 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 ecuații de forma $f(x) = 0$ este echivalentă cu problema determinării soluțiilor globale ale problemei de optimizare fără restricții min f^2 .

Metoda lui Newton este una dintre cele mai cunoscute metode iterative de a aproxima rădăcinile funcțiilor care au proprietăți de diferențiabilitate suficient de bune. Vom vedea că acest algoritm este unul local pentru că data inițială trebuie aleasă suficient de aproape de soluție, dar convergența este pătratică, adică foarte bună.

Fie $f : \mathbb{R}^p \rightarrow \mathbb{R}^p$ o funcție de clasă C^1 și fie $\bar{x}$ o rădăcină simplă a lui f (i.e., $f(\bar{x}) = 0$ și $\nabla f(\bar{x})$ este nesingulară). Considerăm o valoare x_0 suficient de apropiată de $\bar{x}$ (vom discuta sensul acestei apropieri ceva mai târziu).

Șirul iterațiilor Newton pornește de la ecuația

$$0 = f(x_k) + \nabla f(x_k)(x_{k+1} - x_k). \quad (3.3)$$

Această ecuație, din care se obține valoarea lui x_{k+1} arată de ce este important să avem soluție simplă și de ce trebuie să pornim din apropierea lui $\bar{x}$: trebuie ca x_0 să fie într-o vecinătate a lui $\bar{x}$ în care ∇f este inversabilă, iar o astfel de vecinătate există tocmai pentru că ∇f este continuă și neregulară în $\bar{x}$. Astfel, definim formal iterația Newton prin:

$$x_{k+1} = x_k - \nabla f(x_k)^{-1}(f(x_k)). \quad (3.4)$$

Așa cum deja se vede din relațiile anterioare și se va vedea și în continuare, unele dificultăți 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 computațional;
- e dificil de știut a priori că $\nabla f(\bar{x})$ este neregulară.

După prezentarea riguroasă a metodei vom reveni asupra unor aspecte de acest tip. Înainte de a prezenta rezultatul principal privind convergența metodei lui Newton prezentăm o proprietate ajutătoare a funcțiilor de clasă C^1 .

Lema 3.3.1 Fie $D \subset \mathbb{R}^p$ o mulțime deschisă și convexă, $f : D \rightarrow \mathbb{R}^p$ o funcție de clasă C^1 pe D și $x, y \in D$. Atunci

$$f(y) = f(x) + \int_0^1 \nabla f(x + t(y - x))(y - x) dt,$$

unde integrarea și egalitatea se înțeleg pe componente.

Demonstrație. Reamintim că pentru orice $x, y \in D$, mulțimea

$$I_{y,x} = \{t \in \mathbb{R} \mid ty + (1 - t)x \in D\}$$

este un interval care conține intervalul $[0, 1]$.

Pentru fiecare funcție coordonată f_i cu $i \in \overline{1, p}$ definim funcția scalară $\alpha_i : I_{y,x} \rightarrow \mathbb{R}$,

$$\alpha_i(t) = f_i(ty + (1 - t)x) = f_i(x + t(y - x)).$$

Evident,

$$\alpha'_i(t) = \nabla f_i(x + t(y - x))(y - x).$$

Cum $[0, 1] \subset I_{y,x}$ și

$$\alpha_i(1) - \alpha_i(0) = \int_0^1 \alpha'_i(t) dt,$$

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 diferențiala 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 3.3.2 Fie $g : \mathbb{R} \rightarrow \mathbb{R}^p$, $g = (g_i)_{i \in \overline{1,p}}$ astfel încât toate funcțiile g_i sunt Riemann integrabile pe $\mathbb{R}$ și fie $a, b \in \mathbb{R}$ cu $a < b$. Atunci

$$\left\| \int_a^b g(t) dt \right\| \leq \int_a^b \|g(t)\| dt.$$

Demonstrație. Fie

$$v = \left(\int_a^b g_i(t) dt \right)_{i \in \overline{1,p}} \in \mathbb{R}^p.$$

Avem

$$\begin{aligned} \|v\|^2 &= \sum_{i=1}^p v_i^2 = \sum_{i=1}^p v_i \int_a^b g_i(t) dt = \sum_{i=1}^p \int_a^b v_i g_i(t) dt \\ &= \int_a^b \left(\sum_{i=1}^p v_i g_i(t) \right) dt = \int_a^b \langle v, g(t) \rangle dt \\ &\leq \int_a^b \|v\| \|g(t)\| dt = \|v\| \int_a^b \|g(t)\| dt \end{aligned}$$

și prin simplificare cu $\|v\|$ (cazul $v = 0$ este evident) obținem concluzia. $\square$

Teorema 3.3.3 Fie $D \subset \mathbb{R}^p$ o mulțime deschisă și convexă, $f : D \rightarrow \mathbb{R}^p$ o funcție de clasă C^1 pe D . Fie $\bar{x}$ o soluție simplă a ecuației $f(x) = 0$ și fie (x_k) un șir de iterații generate de ecuația (3.4). Atunci există $\varepsilon > 0$ astfel încât dacă $x_0 \in D(\bar{x}, \varepsilon) \subset D$ toate elementele șirului sunt bine definite și rămân în $D(\bar{x}, \varepsilon)$ și (x_k) converge superliniar la $\bar{x}$, adică

$$\lim_{k \rightarrow \infty} \frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|} = 0. \quad (3.5)$$

În plus, dacă ∇f este Lipschitz pe $D(\bar{x}, \varepsilon)$ atunci convergența este pătratică, adică există $K > 0$ astfel încât

$$\frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|^2} \leq K, \quad \forall k. \quad (3.6)$$

Demonstrație. Fie $M > 0$ și $\varepsilon > 0$ astfel încât $D(\bar{x}, \varepsilon) \subset D$ și pentru orice $x \in D(\bar{x}, \varepsilon)$ au loc simultan următoarele:

$$\begin{aligned} \|\nabla f(x) - \nabla f(\bar{x})\| &\leq \frac{1}{4M} \\ \nabla f(x) &\text{ este inversabilă} \\ \|\nabla f(x)^{-1}\| &< M. \end{aligned}$$

Astfel de numere pozitive există, pe baza ipotezelor considerate.

Dacă $x_0 = \bar{x}$, atunci iterația Newton este staționară și concluziile sunt evidente.

Fie $x_0 \in D(\bar{x}, \varepsilon) \setminus \{\bar{x}\}$ și x_1 definit de iterația Newton, adică

$$x_1 = x_0 - \nabla f(x_0)^{-1}(f(x_0)).$$

Atunci

$$\begin{aligned} x_1 - \bar{x} &= x_0 - \bar{x} - \nabla f(x_0)^{-1}(f(x_0)) \\ &= \nabla f(x_0)^{-1}(\nabla f(x_0)(x_0 - \bar{x}) - f(x_0)). \end{aligned}$$

Pe baza lemei anterioare putem scrie

$$\begin{aligned} x_1 - \bar{x} &= \nabla f(x_0)^{-1} \left(\nabla f(x_0)(x_0 - \bar{x}) - \int_0^1 \nabla f(\bar{x} + t(x_0 - \bar{x}))(x_0 - \bar{x}) dt \right) \\ &= \nabla f(x_0)^{-1} \left(\int_0^1 (\nabla f(x_0) - \nabla f(\bar{x} + t(x_0 - \bar{x}))) (x_0 - \bar{x}) dt \right). \end{aligned}$$

Deci, aplicând inclusiv Lema 3.3.2,

$$\begin{aligned} \|x_1 - \bar{x}\| &\leq M \left\| \int_0^1 (\nabla f(x_0) - \nabla f(\bar{x} + t(x_0 - \bar{x}))) (x_0 - \bar{x}) dt \right\| \\ &\leq M \int_0^1 \|(\nabla f(x_0) - \nabla f(\bar{x} + t(x_0 - \bar{x}))) (x_0 - \bar{x})\| dt \\ &\leq M \|x_0 - \bar{x}\| \int_0^1 \|\nabla f(x_0) - \nabla f(\bar{x} + t(x_0 - \bar{x}))\| dt \\ &\leq M \cdot \frac{1}{2M} \|x_0 - \bar{x}\| = \frac{1}{2} \|x_0 - \bar{x}\|. \end{aligned}$$

În particular, $x_1 \in D(\bar{x}, \varepsilon)$ și repetând inductiv calculele de mai sus deducem că $(x_k) \subset D(\bar{x}, \varepsilon)$ și pentru orice k

$$\|x_{k+1} - \bar{x}\| \leq M \|x_k - \bar{x}\| \int_0^1 \|\nabla f(x_k) - \nabla f(\bar{x} + t(x_k - \bar{x}))\| dt$$

Dacă unul dintre punctele șirului este $\bar{x}$, atunci șirul devine staționar și deci concluziile sunt clare.

Contrar, deducem, pentru orice k ,

$$\|x_{k+1} - \bar{x}\| \leq M \|x_k - \bar{x}\| \int_0^1 \|\nabla f(x_k) - \nabla f(\bar{x} + t(x_k - \bar{x}))\| dt \leq \frac{1}{2} \|x_k - \bar{x}\|,$$

de unde $x_k \rightarrow \bar{x}$ și

$$\frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|} \leq M \int_0^1 \|\nabla f(x_k) - \nabla f(\bar{x} + t(x_k - \bar{x}))\| dt.$$

Cum, pe baza faptului că $x_k \rightarrow \bar{x}$ și a continuității lui ∇f , $\nabla f(x_k) - \nabla f(\bar{x} + t(x_k - \bar{x})) \rightarrow 0$, deducem că

$$\lim \frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|} = 0.$$

Presupunem acum că ∇f este Lipschitz pe $D(\bar{x}, \varepsilon)$. Atunci, există $L > 0$ astfel încât pentru orice $x, y \in D(\bar{x}, \varepsilon)$,

$$\|\nabla f(y) - \nabla f(x)\| \leq L \|y - x\|.$$

Aşa cum am demonstrat mai sus, în cazul nestaţionar, pentru orice k ,

$$\begin{aligned}\frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|} &\leq M \int_0^1 \|\nabla f(x_k) - \nabla f(\bar{x} + t(x_k - \bar{x}))\| dt \\ &\leq M \int_0^1 L \|x_k - (\bar{x} + t(x_k - \bar{x}))\| dt \\ &= ML \|x_k - \bar{x}\| \int_0^1 (1-t) dt = \frac{ML}{2} \|x_k - \bar{x}\|.\end{aligned}$$

Aşadar, pentru $K = \frac{ML}{2}$ obţinem

$$\frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|^2} \leq K,$$

ceea ce demonstrează rezultatul. □

Observaţia 3.3.4 *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 (3.4) 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 convergă 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 $f : \mathbb{R} \rightarrow \mathbb{R}$ 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 3.3.5 Fie $f : \mathbb{R} \rightarrow \mathbb{R}$ o funcție ce clasă C^3 , strict convexă și coercivă. Atunci f are un unic punct de minim global $\bar{x}$ care este soluția ecuației $f'(\bar{x}) = 0$. 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 3.3.6 Metoda lui Newton aplicată problemelor de optimizare (Observația 3.3.5) are un oracol de ordinul al doilea.

Exercițiul 3.12 Fie $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = 2x^2 + 3e^{-2x}$. 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

$$\begin{aligned} \lim_{t \rightarrow 0} \frac{f(x+t) - f(x-t)}{2t} &= f'(x), \\ \lim_{t \rightarrow 0} \frac{f(x+t) + f(x-t) - 2f(x)}{t^2} &= f''(x). \end{aligned} \tag{3.7}$$

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

Exercițiul 3.14 *Să se rezolve al doilea caz numeric de la Problema 2.27.*

Exercițiul 3.15 *Să se determine minimul funcției $f : [30, 70] \rightarrow \mathbb{R}$,*

$$f(x) = \frac{204165.5}{330 - 2x} + \frac{10400}{x - 20},$$

aproximând numeric ambele derivate.

Implementăm acum metoda lui Newton pentru dimensiuni mai mari decât 1. Căutăm soluții 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 (fișier funcție)
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 iterații, folosind data inițială $(1, 0, 1)$ și toleranța 10^{-6} , soluția aproximativă $(1.224394, -0.093133, 1.868739)$ pentru sistemul menționat. De remarcat că soluția $(0, 2, 1)$ se obține cu o altă alegere a datei inițiale.

Exercițiul 3.17 Să se scrie niște funcții *newton*, *newton_optim*, *cvasinewton*, *cvasinewton_optim* pentru metodele Newton și cvasi-Newton discutate mai sus, după modelul Implementare 3.16.

Exercițiul 3.18 Să se determine minimul funcției de la Exemplul 1.1.1.

Exercițiul 3.19 Determinați minimele funcției $f : \mathbb{R} \rightarrow \mathbb{R}$,

$$f(x) = \frac{1}{2}x^5 - \frac{3}{40}x^4 - \frac{3}{5}x^3 - 7x.$$

Exercițiul 3.20 Fie șirul de numere reale dat prin $x_0 = 1$ și

$$x_{n+1} = \frac{2x_n^3 + 3}{3x_n^2}, \quad \forall n \geq 0.$$

Să se arate că $(x_n)_n$ este o iterație Newton pentru o funcție care se va determina. Ce se poate spune despre convergența lui $(x_n)_n$?

Exercițiul 3.21 Fie șirul de numere reale dat prin $x_0 = 1$ și

$$x_{n+1} = \frac{x_n^2 + 2}{2x_n}, \quad \forall n \geq 0.$$

Să se arate că $(x_n)_n$ este o iterație Newton pentru o funcție care se va determina. Studiați convergența șirului (x_n) și, în caz afirmativ, determinați-i limita.

Exercițiul 3.22 Să se aproximeze minimul funcției $f : (0, \infty) \rightarrow \mathbb{R}$

$$f(x) = \frac{-0.75}{1+x^2} - 0.65x \operatorname{arctg} \frac{1}{x}.$$

Exercițiul 3.23 Să se aproximeze punctele de extrem local ale funcției de la Exemplul 3.1.5 de mai sus folosind metoda lui Newton pentru ecuația $4u^3 - 32u + 5 = 0$ cu diverse date inițiale.

Exercițiul 3.24 Să se determine soluții aproximative pentru sistemele de mai jos cu datele inițiale 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) $(x_0, y_0) = (2, 1.5)$,

$$\begin{cases} 3 + y - x^2 = 0 \\ 3 - xy = 0. \end{cases}$$

3.4 Aplicație la funcția W , a lui Lambert

Fie $f : \mathbb{R} \rightarrow \mathbb{R}$ dată prin $f(x) = xe^x$. Este ușor de văzut că această funcție este indefinit derivabilă și, ținând cont de dezvoltarea lui e^x în serie de puteri, chiar analitică pe $\mathbb{R}$. De asemenea, pentru orice $x \in \mathbb{R}$,

$$\begin{aligned} f'(x) &= (x+1)e^x \\ f''(x) &= (x+2)e^x, \end{aligned}$$

deci:

(i) f are un punct de extrem, $x = -1$, care este punct de minim global (e de observat că $\lim_{x \rightarrow -\infty} f(x) = 0$ și $\lim_{x \rightarrow \infty} f(x) = \infty$);

(ii) f este strict descrescătoare pe $(-\infty, -1]$ și strict crescătoare pe $[-1, \infty)$;

(iii) $\text{Im } f = [-\frac{1}{e}, \infty)$, $f((-\infty, -1]) = [-\frac{1}{e}, 0)$, $f([-1, \infty)) = [-\frac{1}{e}, \infty)$;

(iv) f are un singur punct de inflexiune, $x = -2$, este strict concavă pe $(-\infty, -2]$ și strict convexă pe $[-2, \infty)$.

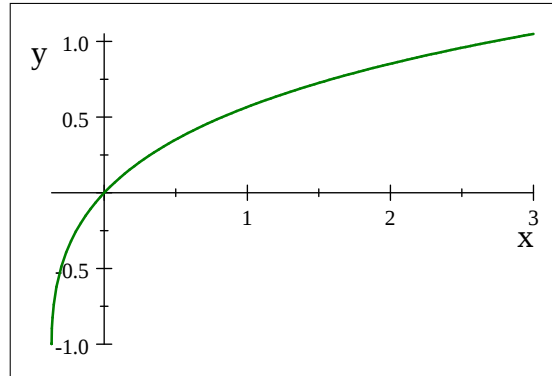
Considerăm două restricții bijective ale lui f , astfel:

$$\begin{aligned} f_1 : [-1, \infty) &\rightarrow [-\frac{1}{e}, \infty), \quad f_1(x) = xe^x, \\ f_2 : (-\infty, -1] &\rightarrow [-\frac{1}{e}, 0), \quad f_2(x) = xe^x. \end{aligned}$$

Inversa lui f_1 se numește funcția lui Lambert și se notează în mod tradițional prin W , deci

$$W : \left[-\frac{1}{e}, \infty\right) \rightarrow [-1, \infty), \quad W(y) = x, \quad \text{unde } y = xe^x.$$

Din cauza faptului că, în general, ecuația $xe^x = y$ nu poate fi rezolvată, nu se poate indica o formă explicită pentru W . Graficul lui W este simetric față de prima bisectoare în raport cu graficul lui f_1 .



Graficul lui W

Unele valori ale lui W pot fi calculate. De exemplu,

$$W\left(-\frac{1}{e}\right) = -1; \quad W(0) = 0; \quad W\left(-\frac{1}{2} \ln 2\right) = -\ln 2; \quad W(e) = 1.$$

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 $x > -1$, funcția W este derivabilă pe $(-\frac{1}{e}, \infty)$ și pentru orice punct y din acest interval deducem următoarele formule:

$$\begin{aligned} W'(y) &= \frac{1}{f'(x)} = \frac{1}{(x+1)e^x}, \text{ unde } y = xe^x; \\ W'(y) &= \frac{e^{-W(y)}}{1+W(y)}; \\ W'(y) &= \frac{W(y)}{y(1+W(y))}, \text{ dacă } y \neq 0. \end{aligned}$$

Cum f_1 este indefinit derivabilă pe $(-1, \infty)$, deducem că W este indefinit derivabilă pe $(-\frac{1}{e}, \infty)$. Derivând în formulele anterioare, avem:

$$\begin{aligned} W''(y) &= -\frac{e^{-2W(y)}(W(y)+2)}{(W(y)+1)^3}; \\ W''(y) &= -\frac{W^2(y)(W(y)+2)}{y^2(W(y)+1)^3}, \text{ dacă } y \neq 0. \end{aligned}$$

Din acest calcul deducem stricta concavitate a funcției W .

O metodă de a aproxima valorile lui W (pe tot domeniul său de definiție) este să rezolvăm aproximativ ecuația (în necunoscuta x) $xe^x - y = 0$, de exemplu prin metoda lui Newton. Astfel, cu o dată apropiată de soluție (pentru y aproape de 0 alegem data inițială aproape de 0, pentru y

mare alegem data inițială $\ln y$, de exemplu) folosim iterația Newton

$$\begin{aligned} x_{n+1} &= x_n - \frac{x_n e^{x_n} - y}{(x e^x - y)'_x(x_n)} \\ &= \frac{x_n^2 + y e^{-x_n}}{x_n + 1}, \end{aligned}$$

pentru a aproxima cu viteză pătratică valoarea lui x , adică a lui $W(y)$.

Exercițiul 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, $f'_1(0) = 1 \neq 0$ și pe baza unui rezultat numit Teorema lui Lagrange de inversiune, se obține 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

$$w_n = \lim_{x \rightarrow 0} \left(\left(\frac{x}{f(x)} \right)^n \right)^{(n-1)} = \lim_{x \rightarrow 0} (e^{-nx})^{(n-1)} = (-n)^{n-1}.$$

Deci,

$$W(y) = \sum_{n=1}^{\infty} \frac{(-n)^{n-1}}{n!} y^n.$$

Este ușor de văzut că raza de convergență a acestei serii este $\frac{1}{e}$, deci dezvoltarea e valabilă pe intervalul $(-\frac{1}{e}, \frac{1}{e})$.

Să calculăm o primitivă lui W , cu schimbarea de variabilă $y = f_1(x)$ (adică $y = x e^x$ și $x = W(y)$)

$$\int W(y) dy = \int x(x+1) e^x dx = e^x (x^2 - x + 1) + C = yW(y) - y + e^{W(y)} + C.$$

Astfel, de exemplu,

$$\int_0^e W(y) dy = e - e + e - 1 = e - 1.$$

Având în vedere imaginile funcțiilor f_1 și 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 existenței (și a valorii) limitei

$$\lim_{x \nearrow -1} \frac{W(f(x)) + 1}{x + 1}.$$

Avem următorul rezultat mai general.

Teorema 3.4.1 Fie $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ și $\bar{x} \in \mathbb{R}$ un punct de extrem local pentru φ . Presupunem că există o vecinătate V a lui $\bar{x}$ astfel încât φ este analitică pe V și $\varphi'(x) \neq 0$ pentru orice $x \in V \setminus \{\bar{x}\}$. Atunci:

- (i) există $\varepsilon > 0$ astfel încât φ este strict monotonă pe $[\bar{x} - \varepsilon, \bar{x}]$ și pe $[\bar{x}, \bar{x} + \varepsilon]$;
- (ii) dacă ψ este inversa lui φ pe $[\bar{x}, \bar{x} + \varepsilon]$, există $\delta > 0$ astfel încât $\psi \circ \varphi$ este corect definită pe $[\bar{x} - \delta, \bar{x}]$ și

$$\lim_{x \nearrow \bar{x}} \frac{(\psi \circ \varphi)(x) - \bar{x}}{x - \bar{x}} = -1.$$

Demonstrație. Facem demonstrația ambelor puncte. Știm că, pe V , φ se scrie ca

$$\varphi(x) = \sum_{n=0}^{\infty} a_n (x - \bar{x})^n.$$

Cum f este neconstantă, există N , cel mai mic număr natural nenul pentru care $a_N \neq 0$. Deci

$$\varphi(x) = a_0 + \sum_{n=N}^{\infty} a_n (x - \bar{x})^n.$$

Definim, pe V , funcția $F(x) = \sum_{n=N}^{\infty} a_n (x - \bar{x})^{n-N}$ și avem

$$\varphi(x) = a_0 + (x - \bar{x})^N F(x).$$

Cum seria ce definește F are aceeași rază de convergență ca și φ , deducem că F este continuă în $\bar{x}$. În plus, $F(\bar{x}) = a_N \neq 0$ și prin urmare F are semn constant în apropierea lui $\bar{x}$, deci există $\varepsilon > 0$ astfel încât F are semn constant pe $[\bar{x} - \varepsilon, \bar{x} + \varepsilon]$ și $[\bar{x} - \varepsilon, \bar{x} + \varepsilon] \subset V$. Cum $\bar{x}$ este extrem pentru φ , N este număr par. Din ipoteza $\varphi'(x) \neq 0$ pentru $x \in V \setminus \{\bar{x}\}$, pentru că $\bar{x}$ este punct de extrem, deducem că φ' este pozitivă și respectiv negativă pe cele două zone ale lui $V : V \cap (-\infty, \bar{x})$ și $V \cap (\bar{x}, \infty)$. Deci φ este strict monotonă, de monotonii diferite pe $[\bar{x} - \varepsilon, \bar{x}]$ și pe $[\bar{x}, \bar{x} + \varepsilon]$. Fie ψ inversa lui φ pe $[\bar{x}, \bar{x} + \varepsilon]$. Din monotonia descrisă anterior, există $\delta \in (0, \varepsilon)$ astfel încât $\psi \circ \varphi$ este corect definită pe $[\bar{x} - \delta, \bar{x}]$. Fie $x \in (\bar{x} - \delta, \bar{x})$. Atunci notăm $z = (\psi \circ \varphi)(x)$. Avem $z \in (\bar{x}, \bar{x} + \varepsilon)$ și

$$\begin{aligned} \varphi(z) &= (\varphi \circ (\psi \circ \varphi))(x) = ((\varphi \circ \psi) \circ \varphi)(x) \\ &= (\varphi \circ \psi)(\varphi(x)) = \varphi(x). \end{aligned}$$

Deci

$$a_0 + (z - \bar{x})^N F(z) = a_0 + (x - \bar{x})^N F(x),$$

adică

$$0 > \frac{z - \bar{x}}{x - \bar{x}} = -\sqrt[N]{\frac{F(x)}{F(z)}}.$$

Pentru $x \rightarrow \bar{x}$ avem $z \rightarrow \bar{x}$ și continuitatea lui F ne conduce la concluzia privind limita. □

Observația 3.4.2 Rezultatul nu are loc dacă φ este doar derivabilă în $\bar{x}$. De exemplu, funcția $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ dată prin

$$\varphi(x) = \begin{cases} x^2, & \text{pentru } x < 0 \\ x^3, & \text{pentru } x \geq 0 \end{cases}$$

este derivabilă, $\bar{x} = 0$ este punct de minim, iar inversa ramurii din dreapta este funcția $\sqrt[3]{x}$. Dar

$$\lim_{x \nearrow 0} \frac{\sqrt[3]{x^2}}{x} = -\infty.$$

Corolarul 3.4.3 Are loc egalitatea

$$\lim_{x \nearrow -1} \frac{W(f(x)) + 1}{x + 1} = -1.$$

Prezentăm câteva aplicații ale funcției W .

Aplicația 1. Prima aplicație a funcției W pe care o menționăm este exprimarea soluțiilor unor ecuații ce nu pot fi rezolvate explicit. Un astfel de exemplu este ecuația

$$ax + b + ce^{dx} = 0$$

unde $a, b, c, d \in \mathbb{R}$ cu $a, d \neq 0$. Atunci ecuația

- are soluția 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 soluțiile

$$x = -\frac{b}{a} - \frac{1}{d} W\left(\frac{cde^{-\frac{bd}{a}}}{a}\right)$$

$$x = -\frac{b}{a} - \frac{1}{d} W_{-1}\left(\frac{cde^{-\frac{bd}{a}}}{a}\right)$$

dacă $\frac{cde^{-\frac{bd}{a}}}{a} \in \left[-\frac{1}{e}, 0\right)$;

- nu are soluție dacă $\frac{cde^{-\frac{bd}{a}}}{a} < -\frac{1}{e}$.

Într-adevăr, ecuația inițială se scrie echivalent astfel

$$(ax + b)e^{-dx} = -c$$

$$\left(-dx - \frac{bd}{a}\right)e^{-dx} = \frac{cd}{a}$$

$$\left(-dx - \frac{bd}{a}\right)e^{-dx - \frac{bd}{a}} = \frac{cd}{a}e^{-\frac{bd}{a}}.$$

Deci obținem concluziile prezentate mai sus.

Aplicația 2. Prezentăm acum o aplicație în mecanică. Considerăm un proiectil care este aruncat de la sol (din origine) sub un unghi θ cu viteza v . Dorim să știm distanța R (pe axa ox pozitivă) la care proiectilul cade pe sol.

Considerăm mai întâi cazul fără frecare (rezistență). Ecuațiile diferențiale care descriu mișcarea sunt

$$\begin{aligned}x'' &= 0, \\y'' &= -g.\end{aligned}$$

Integrăm de două ori și ținem seama de condițiile inițiale, $x(0) = y(0) = 0$, $x'(0) = v \cos \theta$, $y'(0) = v \sin \theta$, pentru a găsi

$$\begin{aligned}x(t) &= (v \cos \theta) t \\y(t) &= -\frac{1}{2}gt^2 + (v \sin \theta) t.\end{aligned}$$

E clar că traiectoria este o parabolă și impunând condiția $y(t) = 0$ soluția nenulă ne conduce la

$$R = \frac{2v^2}{g} \sin \theta \cos \theta = \frac{v^2}{g} \sin 2\theta.$$

Deci distanța maximă va fi când unghiul θ este $\frac{\pi}{4}$.

Considerăm acum cazul în care mișcarea proiectilului întâmpină o rezistență proporțională cu viteza sa. Constanta de proporționalitate se notează cu k . Acum ecuațiile mișcării sunt

$$\begin{aligned}x'' &= -kx', \\y'' &= -g - ky' .\end{aligned}$$

Folosind $x'(0) = v \cos \theta$, $y'(0) = v \sin \theta$ avem

$$\begin{aligned}x'(t) &= (v \cos \theta) e^{-kt}, \\y'(t) &= \frac{1}{k} (-g + (g + kv \sin \theta) e^{-kt}).\end{aligned}$$

Integrăm și utilizăm $x(0) = y(0) = 0$ pentru a găsi

$$\begin{aligned}x(t) &= \frac{1}{k} (v \cos \theta (1 - e^{-kt})), \\y(t) &= \frac{1}{k^2} (-ktg + g + kv \sin \theta - e^{-kt} (g + kv \sin \theta)).\end{aligned}$$

Ca mai sus, trebuie să rezolvăm ecuația $y(t) = 0$, care de data aceasta nu mai este atât de simplă. Observăm că această ecuație 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, $u = -\left(1 + \frac{kv}{g} \sin \theta\right)$ și ecuația devine

$$-kt - u + e^{-kt}u = 0.$$

Cum numărul care decidea asupra soluțiilor acestui tip de ecuație este acum $ue^u \in \left[-\frac{1}{e}, 0\right)$, aplicând formula de mai sus găsim soluțiile

$$\begin{aligned} t_1 &= \frac{-u}{k} + \frac{1}{k} W(ue^u), \\ t_2 &= \frac{-u}{k} + \frac{1}{k} W_{-1}(ue^u). \end{aligned}$$

Este esențial de remarcat faptul că $u < -1$, deci $W_{-1}(ue^u) = u$, adică $t_2 = 0$ (timpul de start). Pe de altă parte, din același motiv, $W(ue^u) \neq u$, deci t_1 corespunde distanței căutate. Înlocuind în x avem

$$R = \frac{1}{k} v \cos \theta \left(1 - e^{u - W(ue^u)}\right).$$

Cum $e^{-W(ue^u)} = \frac{W(ue^u)}{ue^u}$, avem

$$R = \frac{1}{k} v \cos \theta \left(1 - \frac{W(ue^u)}{u}\right).$$

Deci, în final,

$$R = \frac{1}{k} v \cos \theta \left(1 + \frac{W\left(\left(-1 - \frac{kv}{g} \sin \theta\right) e^{-1 - \frac{kv}{g} \sin \theta}\right)}{1 + \frac{kv}{g} \sin \theta}\right).$$

Observația 3.4.4 Remarcăm că atunci când $k \rightarrow 0$ valoarea lui R tinde la valoarea din cazul fără rezistență. Cu notațiile de mai sus, $u = -1 - \frac{kv}{g} \sin \theta$, avem

$$\frac{1}{k} = -\frac{v \sin \theta}{g(u+1)}$$

deci, conform calculului anterior,

$$\begin{aligned} R(k) &= -\frac{v^2 \sin \theta \cos \theta}{g(u+1)} \left(1 - \frac{W(ue^u)}{u}\right) \\ &= \frac{v^2 \sin 2\theta}{2g} \frac{W(ue^u) - u}{u(u+1)}. \end{aligned}$$

Cum $u \nearrow -1$, pentru $k \searrow 0$ avem

$$\begin{aligned} \lim_{k \searrow 0} R(k) &= \frac{v^2 \sin 2\theta}{2g} \lim_{u \nearrow -1} \frac{W(ue^u) - u}{u(u+1)} \\ &= -\frac{v^2 \sin 2\theta}{2g} \lim_{u \nearrow -1} \frac{W(ue^u) - u - 1 + 1}{u+1} \\ &= -\frac{v^2 \sin 2\theta}{2g} \lim_{u \nearrow -1} \left(\frac{W(ue^u) + 1}{u+1} - 1\right) \\ &= \frac{v^2 \sin 2\theta}{g}, \end{aligned}$$

adică rezultatul așteptat. Menționăm că pentru ultima limită s-a folosit Corolarul 3.4.3.

Aplicația 3. A treia aplicație se referă la un aspect din teoria rețelelor aleatorii și, în particular, la o estimare a vitezei de propagare a unei pandemii. Fie o populație de N persoane. Presupunem că fiecare persoană se află în contact cu un număr de a persoane la întâmplare (a poate fi o expectanță, neconstantă pentru toate persoanele, și nu neapărat număr întreg). Dacă o persoană este infectată cu o boală transmisibilă și boala se transmite tuturor celor cu care este în contact, ne interesează numărul total de persoane care se vor infecta. Sau, mai bine spus, care este probabilitatea ca întreaga populație să fie infectată?

Pentru a transpune problema în limbajul inițial, N reprezintă un număr de neuroni și fiecare are a axoni ce pot forma sinapse cu alți 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 și B sunt conectați. Evident, se poate ca A și B să nu fie conectați, 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 aflați la cel mult t axoni de A . Conectivitatea slabă a rețelei se definește prin

$$\gamma(N, a) = \frac{x(N, a, N)}{N}.$$

Cum lucrăm cu N, a fixate, notăm $x(N, a, t)$ prin $x(t)$ și scriem formula de mai sus prin

$$\gamma = \frac{x(N)}{N}.$$

Vom descrie în continuare un argument euristic pentru deducerea valorii lui γ . Menționăm că, pe baza unui efort considerabil, se poate arăta că argumentul poate fi făcut riguros.

Căutăm să exprimăm numărul de neuroni aflați 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 așteptată a numărului de cutii ocupate de măcar o bilă este dată de

$$N \left(1 - \left(1 - \frac{1}{N} \right)^s \right). \quad (3.8)$$

La fiecare pas $a(x(t) - x(t-1))$ noi neuroni sunt conectați, deci la pasul $(t+1)$ vor fi conectați

$$N \left(1 - \left(1 - \frac{1}{N} \right)^{a(x(t) - x(t-1))} \right)$$

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 neconectați de A după t pași) și ecuația 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,

$$y(t+1) = K \left(1 - \frac{1}{N}\right)^{-ay(t)}. \quad (3.9)$$

Dar

$$y(0) = N - 1,$$

deci, conform (3.8),

$$y(1) = (N - 1)^{a+1} N^{-a}.$$

Obținem

$$K = N^{-aN} (N - 1)^{aN+1}.$$

Evident $y(t)$ este o cantitate descrescătoare în t , deci atunci când N crește și t crește, intuim din (3.9) că limita y a acestei cantități va satisface ecuația

$$y = (N - 1) \left(1 - \frac{1}{N}\right)^{a(N-y)}.$$

Pentru N mare această valoare poate fi aproximată prin

$$N e^{a(\frac{y}{N}-1)},$$

deci

$$\gamma = 1 - \frac{y(t)}{N}$$

poate fi aproximată prin soluția ecuației

$$\gamma = 1 - e^{-a\gamma}.$$

Astfel, conform Aplicației 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

$$N\gamma = N \left(1 + \frac{W(-ae^{-a})}{a} \right).$$

Folosind software-ul de calcul științific Octave, putem calcula procentajul de infecții în cadrul populației.

Astfel, funcția 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
```

obținem 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 infectați 79,68% din membrii populației;
- pentru $a = 3$, vor fi infectați 94,04% din membrii populației;
- pentru $a = 4$, vor fi infectați 98,01% din membrii populației;
- pentru $a = 5$, vor fi infectați 99,30% din membrii populației;
- pentru $a = 10$, vor fi infectați 99,99% din membrii populației.

Exercițiul 3.26 *Să se scrie un program propriu de determinare aproximativă a valorilor de mai sus.*

3.5 O metodă cvasi-Newton: metoda secantei

Discutăm în cazul funcțiilor de o variabilă o alternativă la metoda lui Newton despre care am văzut că are o viteză de convergență foarte bună, dar necesită la fiecare pas calculul derivatei într-un punct, calcul ce în unele situații poate fi costisitor din punct de vedere computațional. Ideea noii metode, intitulată metoda secantei (pentru motive pe care le vom înțelege în curând), este de a înlocui derivata din calculul iterației Newton

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

cu o aproximare a sa (deci este o metodă de tip cvasi-Newton). Astfel, obținem o iterație 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, iterația 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 recurența să fie definită, este nevoie de două date inițiale x_0, x_1 . O ilustrare grafică a semnificației iterației de mai sus justifică numele metodei. Este de asemenea ușor de intuit că limita $h_k \rightarrow 0$ nu este automat satisfăcută, ci depinde de unele detalii tehnice legate de alegerea datelor inițiale. Pe de altă parte, noua iterație nu face decât o singură evaluare a lui f la fiecare pas.

Prezentăm acum detaliile matematice ale convergenței acestei metode.

Lema 3.5.1 Fie $f : [a, b] \rightarrow \mathbb{R}$ o funcție derivabilă astfel încât

$$\mu = \inf_{x \in [a, b]} |f'(x)| > 0.$$

Atunci există $H > 0$ astfel încât pentru orice $x \in [a, b]$ și orice $h \in [-H, H]$ astfel încât $x+h \in [a, b]$ avem

$$\left| \frac{f(x+h) - f(x)}{h} \right| \geq \frac{\mu}{2}.$$

Demonstrație. Presupunem, prin reducere la absurd, că există $(x_k) \subset [a, b]$, $h_k \rightarrow 0$ astfel încât $x_k + h_k \in [a, b]$ cu

$$\left| \frac{f(x_k + h_k) - f(x_k)}{h_k} \right| < \frac{\mu}{2}, \quad \forall k.$$

Conform Teoremei lui Lagrange, pentru fiecare k , există ν_k între $x_k + h_k$ și 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 relație contrazice ipoteza. □

Următoarea leamnă este conținută în demonstrația Teoremei 3.3.3, dar o prezentăm explicit.

Lema 3.5.2 Fie $f : [a, b] \rightarrow \mathbb{R}$ o funcție derivabilă astfel încât f' este Lipschitz de constantă $L > 0$ pe $[a, b]$. Atunci pentru orice $x, y \in [a, b]$

$$|f(y) - f(x) - f'(x)(y-x)| \leq \frac{L}{2} (x-y)^2.$$

Demonstrație. Scrierea

$$|f(y) - f(x) - f'(x)(y - x)| = \left| \int_x^y (f'(t) - f'(x)) dt \right|$$

și ipoteza asigură concluzia. □

Teorema 3.5.3 Fie $f : [a, b] \rightarrow \mathbb{R}$ o funcție derivabilă astfel încât

- (i) f se anulează într-un punct $\bar{x} \in (a, b)$;
- (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 inițială $x_0 \in (\bar{x} - \delta, \bar{x} + \delta)$ și orice șir $\{h_k\} \subset (-\eta, \eta) \setminus \{0\}$, iterațiile

$$x_{k+1} = x_k - f(x_k) \frac{h_k}{f(x_k + h_k) - f(x_k)}$$

sunt bine definite și rămân în intervalul $(\bar{x} - \delta, \bar{x} + \delta)$ și, mai mult, $x_k \rightarrow \bar{x}$.

Demonstrație. Pentru buna definire este suficient ca $x_k + h_k, x_k \in (a, b)$ pentru orice k (cf. Teoremei lui Lagrange și ipotezei (ii)), lucru care se va realiza pentru alegeri potrivite ale constantelor δ și η , după cum se va vedea mai jos.

Definim $\varepsilon_k = \bar{x} - x_k$ pentru orice k . Din Lema 3.5.1,

$$d_k = \left| \frac{f(x_k + h_k) - f(x_k)}{h_k} \right| \geq \frac{\mu}{2}$$

dacă $|h_k| \leq H$ și $x_k, x_k + h_k \in (a, b)$. Alegem deci $\{h_k\} \subset (-H, H)$. Presupunem pentru moment că x_k și $x_k + h_k$ sunt în (a, b) . Atunci,

$$\begin{aligned} \varepsilon_{k+1} &= \bar{x} - x_{k+1} = \bar{x} - x_k + \frac{f(x_k)}{d_k} \\ &= \bar{x} - x_k + \frac{f(x_k)}{d_k} - \frac{f(\bar{x})}{d_k} \\ &= \varepsilon_k + \frac{f(x_k) - f(\bar{x})}{d_k} \\ &= \frac{1}{d_k} (f(x_k) - f(\bar{x}) + \varepsilon_k d_k) \\ &= \frac{1}{d_k} (f(x_k) - f(\bar{x}) + f'(x_k) \varepsilon_k + \varepsilon_k d_k - f'(x_k) \varepsilon_k), \end{aligned}$$

deci, folosind și Lema 3.5.2,

$$\begin{aligned} |\varepsilon_{k+1}| &\leq \frac{1}{|d_k|} (|f(\bar{x}) - f(x_k) - f'(x_k) \varepsilon_k| + |d_k - f'(x_k)| |\varepsilon_k|) \\ &\leq \frac{2}{\mu} \left(\frac{L |\varepsilon_k|^2}{2} + \frac{L |h_k| |\varepsilon_k|}{2} \right). \end{aligned}$$

Am găsit astfel estimarea,

$$|\varepsilon_{k+1}| \leq \frac{L}{\mu} (|\varepsilon_k| + |h_k|) |\varepsilon_k|$$

în ipotezele mai sus menționate.

Trecem acum la demonstrația efectivă a concluziei. Fie $\theta \in (0, 1)$, fie $\rho = \min \{|\bar{x} - a|, |\bar{x} - b|\}$ și $\delta = \min \left\{ \frac{\rho}{2}, \frac{\theta\mu}{2L} \right\}$, $\eta = \min \{H, \delta\}$. Fie (x_k) un șir de iterații generat ca în enunțul rezultatului. Pentru a obține concluziile este suficient să arătăm că

$$|\varepsilon_{k+1}| \leq \theta |\varepsilon_k|, \quad \forall k.$$

Pentru $k = 0$, $x_0 \in (a, b)$, $x_0 + h_0 \in (a, b)$ (din alegerea constantelor) și putem aplica estimarea anterioară

$$|\varepsilon_1| \leq \frac{L}{\mu} (|\varepsilon_0| + |h_0|) |\varepsilon_0| \leq \frac{L}{\mu} \cdot 2\delta \cdot |\varepsilon_0| \leq \frac{L}{\mu} \cdot 2 \frac{\theta\mu}{2L} \cdot |\varepsilon_0| = \theta |\varepsilon_0|.$$

Inductiv, se obține inegalitatea anunțată și demonstrația este încheiată. $\square$

Corolarul 3.5.4 În ipotezele Teoremei 3.5.3, dacă $h_k \rightarrow 0$, atunci convergența lui (x_k) este superliniară.

Demonstrație. Din demonstrația Teoremei 3.5.3 deducem

$$|\varepsilon_{k+1}| \leq \frac{2L}{\mu} \max \{|\varepsilon_k|, |h_k|\} |\varepsilon_k|, \quad \forall k \geq 0.$$

Cum $\varepsilon_k \rightarrow 0$ (conform Teoremei 3.5.3) și $h_k \rightarrow 0$ (conform ipotezei), avem și $\max \{|\varepsilon_k|, |h_k|\} \rightarrow 0$. Astfel, are loc convergența superliniară. $\square$

În continuare, precizăm ordinul de convergență al metodei secantei. Din nou, începem cu un rezultat ajutător.

Lema 3.5.5 Fie $f : [a, b] \rightarrow \mathbb{R}$ de două ori derivabilă și $x, y, z \in [a, b]$ puncte distincte. Atunci există $u \in (\min \{x, y, z\}, \max \{x, y, z\})$ astfel încât

$$e(x, y, z) := \frac{\frac{f(y)-f(z)}{y-z} - \frac{f(x)-f(y)}{x-y}}{z-x} = \frac{f''(u)}{2}.$$

Demonstrație Considerăm funcția polinomială

$$p(u) = f(x) + \frac{f(y) - f(x)}{y - x} (u - x) + e(x, y, z) (u - x) (u - y)$$

și constatăm că $f(x) = p(x)$, $f(y) = p(y)$, $f(z) = p(z)$. Aplicând Teorema lui Rolle funcției $f - p$, găsim două zerouri ale derivatei sale și tot din Teorema lui Rolle pentru $(f - p)'$ găsim u astfel încât $(f - p)''(u) = 0$, ceea ce conduce la concluzie. $\square$

Teorema 3.5.6 Fie $f \in C^2([a, b])$ astfel încât

(i) f se anulează într-un punct $\bar{x} \in (a, b)$;

(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 inițiale $x_0, x_1 \in [a, b]$ converge la $\bar{x}$ atunci există $C > 0$ astfel încât

$$|\varepsilon_{k+1}| \leq C |\varepsilon_k|^\varphi, \quad \forall k,$$

unde $\varphi = \frac{1+\sqrt{5}}{2}$ este numărul de aur.

Demonstrație. Pentru orice k avem

$$\begin{aligned} -\varepsilon_{k+1} &= x_{k+1} - \bar{x} = x_k - \bar{x} - f(x_k) \frac{x_{k-1} - x_k}{f(x_{k-1}) - f(x_k)} \\ &= -\varepsilon_k - f(x_k) \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \\ &= \frac{\frac{f(\bar{x}) - f(x_k)}{\bar{x} - x_k} - \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}}{\frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}} \cdot \varepsilon_k \varepsilon_{k-1} \\ &= \varepsilon_k \varepsilon_{k-1} \frac{f''(v_k)}{2f'(u_k)}, \end{aligned}$$

unde $u_k, v_k \in (a, b)$. Existența lui v_k rezultă din Teorema lui Lagrange, în timp ce existența lui (v_k) rezultă din Lema 3.5.5. De asemenea, este clar că $(u_k), (v_k) \rightarrow \bar{x}$.

Din ipoteză,

$$|\varepsilon_{k+1}| \leq \frac{2L}{\mu} |\varepsilon_k| |\varepsilon_{k-1}|, \quad \forall k \geq 0.$$

Notăm $M := \frac{2L}{\mu}$. Cum $|\varepsilon_{k-1}| \rightarrow 0$, deducem că ordinul de convergență ar putea fi mai mare decât 1.

Arătăm că ordinul de convergență este φ , soluția supraunitară a ecuației $p^2 = p + 1$. Pentru determina ordinul de convergență, căutăm două constante C și $p > 1$ astfel încât pentru orice k suficient de mare să avem

$$|\varepsilon_{k+1}| \leq C |\varepsilon_k|^p.$$

Atunci,

$$|\varepsilon_{k+1}| \leq C |\varepsilon_k|^p \leq C^{p+1} |\varepsilon_{k-1}|^{p^2}$$

și

$$|\varepsilon_{k+1}| \leq M |\varepsilon_k| |\varepsilon_{k-1}| \leq MC |\varepsilon_{k-1}|^p.$$

Este suficient ca

$$\begin{aligned} C^{p+1} &= MC \\ p^2 &= p + 1, \end{aligned}$$

de unde $p = \varphi$ și $C = M^{\frac{1}{\varphi}}$. □

Observația 3.5.7 Așadar, în anumite ipoteze metoda secantei are ordinul de convergență φ . Cum $\varphi^2 > 2$, iar uneori două iterații ale acestei metode sunt mai rapide decât o iterație a metodei lui Newton, în astfel de cazuri, metoda secantei este mai eficientă.

Exercițiul 3.27 Să se implementeze metoda secantei (eventual, pentru una dintre funcțiile studiate la metoda lui Newton).

Observația 3.5.8 Metoda secantei aplicată derivatei conduce la o metodă de ordin 1 pentru aproximarea punctelor critice (de minim, în unele cazuri).

Exercițiul 3.28 Să se folosească metoda secantei pentru determinarea minimului funcției de la Exemplul 1.1.1 (cu derivata calculată exact și/sau numeric).

Exercițiul 3.29 Folosiți metoda înjumătățirii intervalului, metoda lui Newton, metoda cvasi-Newton și metoda secantei pentru a aproxima rădăcini ale ecuațiilor de mai jos.

- (i) $\cos x - xe^x = 0$; (ii) $x^4 - x - 10 = 0$; (iii) $x - e^{-x} = 0$; (iv) $e^{-x}(x^2 - 5x + 2) + 1 = 0$;
(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 restricții

Așa cum am putut constata, multe dintre problemele de optimizare nu pot fi rezolvate exact. De asemenea, nu există o metodă universală de aproximare a soluțiilor, ci doar unele principii generale privind designul majorității algoritmilor pe care îi considerăm. Pe baza acestor principii vom încerca să construim clase de algoritmi ce sunt eficienți pentru clase specifice (particulare) de probleme. În acest capitol ne ocupăm de algoritmi pentru probleme fără restricții.

Până în acest moment, am discutat deja câțiva algoritmi pentru aproximarea punctelor de minim, dar cei mai eficienți dintre aceștia sunt bazați pe metode de determinare a zerourilor unei funcții aplicate diferențialei funcției de minimizat (metoda lui Newton și metoda secantei). În acest capitol vom studia algoritmi care sunt de la bun început concepuți pentru minimizare. Astfel, începem cu algoritmi de ordin zero pentru funcții de o variabilă, știut fiind faptul că astfel de algoritmi sunt greoi din punct de vedere computațional pentru dimensiuni mari (a se vedea Exemplul 3.1.4) și vom continua cu algoritmi de ordin superior în cazul general.

4.1 Algoritmi de ordin zero pentru funcții de o variabilă

Prima categorie de algoritmi de care ne ocupăm este cu oracol de ordin zero, adică fără diferențiale, baza reprezentând-o cunoașterea funcției obiectiv în punctele de control. Deci, în principiu, acești algoritmi funcționează și pentru funcții nediferențiabile.

Așa cum am văzut (Teorema 3.1.3), pentru mai funcții de mai multe variabile, complexitatea metodelor de ordin zero este descurajantă, motiv pentru care ne vom concentra, mai ales, pe cazul funcțiilor de o variabilă reală. Mai mult, funcțiile de care ne vom ocupa vor fi continue și definite pe intervale compacte. În plus, vom impune o proprietate foarte importantă, numită proprietatea de unimodalitate.

Fie deci $f : [a, b] \rightarrow \mathbb{R}$ o funcție continuă ($a, b \in \mathbb{R}, a < b$). În virtutea Teoremei lui Weierstrass, există $\bar{x} \in [a, b]$, un punct în care f își atinge minimul global pe $[a, b]$. Cum căutăm metode de a aproxima un astfel de punct, este rezonabil să considerăm că $\bar{x} \in (a, b)$, pentru că, în caz contrar, $\bar{x}$ este cunoscut cu exactitate. Spunem că f este unimodală dacă este strict descrescătoare pe $[a, \bar{x}]$ și strict crescătoare pe $[\bar{x}, b]$. Evident, în această ipoteză, $\bar{x}$ este unicul punct de minim global pentru f .

4.1.1 Delimitarea (bracketing) punctului de minim

Fie deci $f : [a, b] \rightarrow \mathbb{R}$ continuă și unimodală. Fie $\bar{x} \in (a, b)$ unicul său punct de minim. Ne interesează să determinăm un interval mai restrâns în care se găsește $\bar{x}$. Fie $x \in (a, b)$. Dacă $f(a) < f(x)$, cum f este unimodală, putem spune că $\bar{x} \in (a, x)$. Altfel, dacă $f(a) > f(x)$, nu se poate spune nimic cu privire la poziționarea lui $\bar{x}$ în raport cu cele două intervale delimitate de $\bar{x}$. Ideea este aceea de a crește suficient de încet punctul x plecând din a astfel încât să nu îl depășim pe $\bar{x}$.

Implementare 4.1 Ilustrăm această discuție 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: $a = 2.0616$; $\text{ans} = -1.2500$; $k = 40616$

Exercițiul 4.2 Să se deseneze graficul funcției $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^6 - 4x^5 + 3x^3 + 6x$ (a se vedea și Exercițiul 3.1) pe diferite intervale și să se deducă faptul că există două puncte de minim. Apoi, să se delimiteze în jurul fiecărui minim un interval pe care funcția este unimodală și să se aproximeze aceste puncte prin metoda de mai sus.

Chiar dacă este un procedeu simplu și oarecum euristic, această metodă ilustrează o idee ce se întâlnește des și la algoritmi mai sofisticăți, și anume ideea unei mișcări progresive în direcția potrivită cu un pas mic (dar totuși nu prea mic, pentru a evita un calcul prea costisitor).

Putem îmbunătăți metoda de mai sus după principiul rețelei 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 $i \in \overline{0, n}$ și alegem punctul în care se obține cea mai mică valoare. Obținem un punct $x_{\min}$ și 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 același procedeu: recalculăm noul δ și așa mai departe. Procesul de căutare se va termina când lungimea intervalului de căutare este mai mică decât o valoare $\varepsilon > 0$ prestabilită.

Exercițiul 4.3 Să se implementeze această metodă pe funcția de la exercițiul anterior.

Așa cum am văzut, acest algoritm este destul de costisitor din punct de vedere computațional pentru că presupune calcularea funcției (invocarea oracolului de ordin zero) în multe puncte.

În orice caz, pentru orice algoritm de acest tip apare problema determinării punctelor a și b între care facem căutarea. Aceasta poate fi rezolvată astfel: alegem u un punct arbitrar și $\delta > 0$ o constantă. Dacă

$$f(u - \delta) \geq f(u) \geq f(u + \delta),$$

atunci, cu siguranță, pe baza unimodalității, punctul de minim se află la dreapta lui u . În acest caz, ne deplasăm la dreapta cu constanta δ atât timp cât este respectată condiția $f(u) \geq f(u + \delta)$. Dacă această condiție nu este respectată, înseamnă că minimum se află în $(u - \delta, u + \delta)$. Un raționament similar se face dacă

$$f(u - \delta) \leq f(u) \leq f(u + \delta).$$

Dacă niciuna dintre condițiile de mai sus nu se întâmplă pentru alegerea u , cerem o altă dată inițială (unele raționamente ar permite să spunem că punctul de minim e relativ aproape de u , deci noua dată inițială nu trebuie să fie foarte departe de alegerea anterioară).

Exercițiul 4.4 Să se implementeze această idee pentru funcția $f: \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^2 - \sqrt{17}x + 3$.

O altă metodă de delimitare a punctului (punctelor) de minim, dar și a unor intervale pe care funcția este unimodulară, e bazată pe o vizualizare grafică. Pentru codul aferent reprezentărilor grafice ale funcțiilor de o variabilă, dar și o listă de funcții pe care se pot testa codurile din această secțiune prezentăm liniile Matlab/Octave de mai jos.

Implementare 4.5 Codul desen_o_var.m (fișier 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 funcțiile de mai sus provin din chestiuni cu caracter practic ce nu pot fi tratate cu ușurință prin metodele clasice de determinare cu exactitate a punctelor de extrem (a se vedea prima secțiune a lucrării).

4.1.2 Metoda secțiunii de aur

Prima metodă mai elaborată pe care o prezentăm este bazată pe o observație simplă și pe un detaliu tehnic, nu chiar simplu la rândul său.

Ideea este una similară celei de mai sus și constă în micșorarea intervalului ce conține 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 și un număr cât mai mic de invocări ale oracolului.

Fie $f : [a, b] \rightarrow \mathbb{R}$ o funcție 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 siguranță, $\bar{x} \in (a, x_2]$, iar altfel $\bar{x} \in (x_1, b)$. Deci, în orice caz putem micșora intervalul. Ideea este de a alege judicios punctele x_1 și x_2 în așa fel încât să avem o descreștere semnificativă a lungimii intervalului, indiferent de care parte a intervalului inițial este îndepărtată. O altă cerință este de a proceda suficient de eficient astfel încât acel punct dintre x_1 și x_2 care rămâne în interiorul noului interval ce delimitează minimul să poată servi ca punct intermediar pentru următorul pas al iterației. Dacă această a doua cerință este îndeplinită, atunci la noul pas este necesară evaluarea lui f doar într-un punct.

De exemplu, ne-am putea gândi la împărțirea intervalului inițial în trei părți egale, adică să luăm $x_1 = a + \frac{1}{3}(b - a)$ și $x_2 = a + \frac{2}{3}(b - a)$. Procedând astfel, dacă $f(x_1) < f(x_2)$ atunci trebuie să reținem 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 computațional) acest punct nu va oferi simetrie împărțirii intervalului. Deci, această împărțire nu este optimală din acest punct de vedere.

Fie $\rho \in (\frac{1}{2}, 1)$ și $x_1 = \rho a + (1 - \rho)b$, $x_2 = (1 - \rho)a + \rho b$. Această diviziune permite ca proporția din lungimea inițială care va fi eliminată să fie aceeași în oricare dintre cele două cazuri. Considerăm, la fiecare pas, această împărțire 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 reținem intervalul $(a, x_2]$. Pe baza cerințelor 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 reținem 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

$$(1 - \rho) a + \rho b = \rho (\rho a + (1 - \rho) b) + (1 - \rho) b,$$

și 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 și fiind tradițional notat cu φ .

Această metodă asigură faptul că intervalul ce delimitează minimumul se micșorează 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 funcției este necesară.

Din nou, căutarea se termină când lungimea intervalului este mai mică decât o valoare prescrisă. Aproximarea minimumului 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 funcția de două ori la fiecare pas.

Exercițiul 4.7 Să se implemeteze un cod care să remedieze această problemă.

Vom dori ca pe viitor să utilizăm această metodă pentru diverse funcții, motiv pentru care vom prefera unele variante care să implementeze o funcție Matlab/Octave.

Exercițiul 4.8 *Să se scrie o funcție 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 toți pașii, ci se bazează pe termenii șirului lui Fibonacci pentru construirea acestuia. Amintim principalele elemente despre acest șir necesare în continuare.

Propoziția 4.1.2 (*Șirul lui Fibonacci*) *Fie șirul $(F_k)_{k \geq 0}$, este definit cu termenii inițiali $F_0 = 0, F_1 = 1$ și prin relația de recurență:*

$$F_k = F_{k-1} + F_{k-2}, \quad \forall k \geq 2.$$

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)

$$\lim_{k \rightarrow \infty} \frac{F_k}{F_{k-1}} = \varphi.$$

(iii) *Seria*

$$\sum_{k=1}^{\infty} \frac{1}{F_k}$$

este convergentă.

Demonstrație. (i) O demonstrație se poate face prin inducție.

O altă demonstrație, prin care să se ajungă la forma anunțată a șirului este următoarea. Observăm că puterile rădăcinilor ecuației

$$x^2 - x - 1 = 0.$$

satisfac relația de recurență. Rădăcinile acestei ecuații sunt

$$\varphi = \frac{1 + \sqrt{5}}{2}$$

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

Deducem că pentru orice numere reale a, b , termenii șirului $(a\varphi^k + b\psi^k)_k$ satisfac relația de recurență. Cum datele inițiale împreună cu relația de recurență definesc în mod unic termenii șirului,

înseamnă că prin determinarea valorilor a, b pentru respectarea datelor inițiale, șirul obținut este șirul lui Fibonacci. Condițiile

$$\begin{aligned} a + b &= 0 \\ a\varphi + b\psi &= 1 \end{aligned}$$

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,

$$\lim \frac{F_k}{F_{k+1}} = \frac{1}{\varphi} < 1$$

și din criteriul raportului pentru serii cu termeni pozitivi obținem concluzia. $\square$

Prezentăm acum metoda anunțată. Cum am spus, procedăm ca la metoda secțiunii de aur, dar fără a avea un raport constant, dar ca mai înainte, la pasul cu numărul k funcția este testată în două puncte x_k^1 și x_k^2 din interiorul intervalului de delimitare $[a_k, b_k]$ și dorim să păstrăm unul dintre aceste puncte drept capăt al noului interval, iar pe celălalt ca punct interior. La iterația k intervalul are o lungime aflată în proporția $r_k \in [\frac{1}{2}, 1)$ 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 $f(x_k^1) \leq f(x_k^2)$, deci $a_{k+1} = a_k$ și $b_{k+1} = x_k^2$. Renotăm punctul x_k^1 prin x_{k+1}^2 (a se vedea cazul corespunzător de la metoda secțiunii de aur). Atunci, lungimea noului interval de căutare este

$$d_{k+1} = r_k (b_k - a_k).$$

Mai mult,

$$x_k^2 - x_k^1 = b_{k+1} - x_{k+1}^2$$

și avem

$$\begin{aligned} x_k^2 - x_k^1 &= (1 - r_k) a_k + r_k b_k - r_k a_k - (1 - r_k) b_k \\ &= (2r_k - 1) (b_k - a_k). \end{aligned}$$

În plus,

$$\begin{aligned} b_{k+1} - x_{k+1}^2 &= b_{k+1} - (1 - r_{k+1}) a_{k+1} - r_{k+1} b_{k+1} = (1 - r_{k+1}) (b_{k+1} - a_{k+1}) \\ &= (1 - r_{k+1}) r_k (b_k - a_k). \end{aligned}$$

Astfel, obținem

$$2r_k - 1 = (1 - r_{k+1}) r_k,$$

adică

$$r_{k+1} = \frac{1 - r_k}{r_k}.$$

În acest moment utilizăm termenii șirului lui Fibonacci.

Fixăm $n > 2$ și definim raportul r_k ca fiind

$$r_k = \frac{F_{n-k}}{F_{n-k+1}} \in \left[\frac{1}{2}, 1 \right), \quad \forall k \in \{1, 2, \dots, n-2\}.$$

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,

$$r_1 = \frac{F_{n-1}}{F_n}$$

și

$$r_2 = \frac{F_{n-2}}{F_{n-1}}, \dots, r_{n-2} = \frac{F_2}{F_3} = \frac{1}{2}.$$

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 $n - 2$ pași, unde n este numărul natural fixat mai sus. Avem

$$d_{n-1} = r_{n-2} d_{n-2} = \frac{F_2}{F_3} d_{n-2} = \frac{F_2}{F_3} \frac{F_3}{F_4} d_{n-3} = \dots = \frac{1}{F_n} d_1 = \frac{1}{F_n} (b - a),$$

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 secțiunii 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)

```

Exercițiul 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).*

Observația 4.1.3 *Avantajul acestor algoritmi este acela că, oracolul fiind de ordin 0, se aplică și funcțiilor nederivabile, așa cum se vede din lista funcțiilor pentru care deja le-am implementat.*

Exercițiul 4.11 *Să se implementeze metoda secantei pentru derivată în scopul aproximării punctelor de minim pentru funcțiile studiate mai sus.*

Exercițiul 4.12 *Să se implementeze metoda înjumătățirii intervalului pentru derivată în scopul aproximării punctelor de minim pentru funcțiile studiate mai sus.*

Exercițiul 4.13 *Să se recapituleze și să se compare pe exemplele concrete de mai jos metodele de ordin 0, 1, 2 pentru optimizarea funcțiilor 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) $f(x) = 2x^2 + 16x^{-1}$, $x \in [1, 5]$.

Exercițiul 4.14 *Reprezentând grafic funcțiile de mai jos, să se observe că sunt unimodale pe intervalele indicate și 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 funcții de mai multe variabile, am văzut că algoritmii de ordin zero pot fi foarte costisitori din punct de vedere computațional. Pentru a obține algoritmi eficienți pentru astfel de funcții 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 secțiune vom studia câteva dintre aceste metode. Avantajul considerării diferențialei este că putem renunța la unimodalitate, întrucât semnul derivatei ne indică dacă în vecinătatea punctului funcția crește sau scade, iar dezavantajele sunt că nu putem trata funcții nediferențiabile și metodele ne indică, în general, că ne îndreptăm către puncte critice care, în absența convexității, ar putea să nu fie puncte de extrem.

4.2.1 Descrierea metodei direcției de descreștere (gradient descent, GD)

Pentru început, ne concentrăm pe modelul numit "căutarea direcției de descreștere".

Scopul acestui algoritm este acela de a realiza la fiecare pas o descreștere a funcției obiectiv $f : \mathbb{R}^p \rightarrow \mathbb{R}$, de clasă C^2 (i.e. $f(x_{k+1}) < f(x_k)$). Algoritmul calculează la fiecare pas k , o direcție p_k (un vector de pe sfera unitate) și un pas $\alpha_k > 0$ de mișcare pe direcția p_k . Astfel, plecând de la iterația x_k , noua iterație va fi

$$x_{k+1} = x_k + \alpha_k p_k. \quad (4.1)$$

Primul punct al iterației, adică x_0 , se alege arbitrar (vom vedea ulterior că, și de această dată, este indicat ca x_0 să satisfacă anumite cerințe). În cadrul acestei scheme sunt importante atât alegerea direcției cât și a pasului.

Alegerea direcției 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

$$f(x_k + \alpha p) - f(x_k) = \alpha \left[\nabla f(x_k)(p) + \frac{1}{2} \alpha \nabla^2 f(x_k + tp)(p, p) \right]. \quad (4.2)$$

Oprindu-ne cu evaluarea la diferențiala de ordinul întâi, direcția pe care f descrește cel mai mult este dată de soluția problemei de minimizare pe bila unitate închisă a funcției (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 și $\nabla f(x_k)$, este clar că minimum 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 numește metoda celei mai bune (sau celei mai abrupte) descreșteri (steepest descent method, în limba engleză). Pe de altă parte, orice direcție pentru care unghiul cu $\nabla f(x_k)$ este mai mare de $\frac{\pi}{2}$ (i.e. $\cos \theta < 0$) produce o descreștere a funcției f dacă α este suficient de mic (conform (4.2)), întrucât ultimul termen conține α , deci tinde mai rapid

către zero. O astfel de direcție (pentru care $\nabla f(x_k)(p_k) < 0$) se numește direcție de descreștere (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 și vom detalia consecințele celor mai populare dintre aceste posibilități.

Regula minimizării exacte Alegerea ideală este minimul, pentru $\alpha > 0$, al funcției $\alpha \mapsto f(x_k + \alpha p_k)$, dar, din punct de vedere practic, problema aceasta nu este neapărat una simplă de rezolvat. Totuși, aplicarea algoritmului secțiunii de aur conduce la o soluție aproximativă a acestei probleme.

Regula minimizării limitate Alegem α_k astfel încât

$$f(x_k + \alpha p_k) \leq f(x_k + t p_k), \quad \forall t \in [0, s],$$

unde $s > 0$ este o cantitate dinainte stabilită. Astfel, trebuie să minimizăm de fapt $\alpha \mapsto f(x_k + \alpha p_k)$ pe $[0, s]$. Din nou, această problemă intermediară poate fi una dificil de implementat.

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 variație a metodei celei mai bune direcții de descreștere, dar vom vedea că funcționează pentru funcții f având unele proprietăți suplimentare. În plus, în general, ea trebuie combinată și cu alte idei. Dacă $f(x_k + \beta \nabla f(x_k)) \geq f(x_k)$ atunci factorul β trebuie micșorat (procedură ce se numește backtracking).

Regula Armijo Fie $\alpha > 0$, $\beta \in (0, 1)$ și $c \in (0, 1)$. Se testează inegalitatea

$$f(x_k) + c(\beta^i \alpha) \nabla f(x_k)(p_k) > f(x_k + \beta^i \alpha p_k), \quad i = 0, 1, 2, \dots \quad (4.3)$$

iterativ pornind cu $i = 0$ (sau, mai practic, cu valoarea lui i obținută la precedenta iterație). Ideea este că dorim ca descreșterea funcției 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 micșorează la $\beta\alpha$ și 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 același fel, obținem o metodă foarte practică.

Observația 4.2.1 *Iterațiile prevăzute de Regula Armijo pentru determinarea pasului se termină la un moment dat. Demonstrăm această aserțiune. Evident, atunci când i este foarte mare, $t := \beta^i \alpha$ este foarte apropiat de 0. Presupunem că*

$$f(x_k + t p_k) \geq f(x_k) + ct \nabla f(x_k)(p_k), \quad \forall t.$$

Cum f este de clasă C^1 , există o funcție θ continuă în 0, cu $\lim_{t \rightarrow 0} \frac{\theta(t)}{t} = 0$ astfel încât

$$f(x_k + tp_k) = f(x_k) + t \nabla f(x_k)(p_k) + \theta(t).$$

Obținem

$$(c-1) \nabla f(x_k)(p_k) + \frac{\theta(t)}{t} \leq 0.$$

Pentru $t \rightarrow 0$ (adică pentru $i \rightarrow \infty$), 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 condiții

$$\begin{cases} f(x_k + \alpha p_k) < f(x_k) + c_1 \alpha \nabla f(x_k)(p_k), \\ c_2 \nabla f(x_k)(p_k) \leq \nabla f(x_k + \alpha p_k)(p_k). \end{cases} \quad (4.4)$$

În general, prima condiție (care este condiția Armijo de mai sus) poate fi satisfăcută de valori α foarte mici, lucru împiedicat de a doua condiție.

Consistența acestor condiții este arătată riguros mai jos.

Propoziția 4.2.2 Fie $f : \mathbb{R}^p \rightarrow \mathbb{R}$ de clasă C^1 , p_k o direcție de descreștere în x_k și $0 < c_1 < c_2 < 1$. Dacă f este mărginită inferior pe mulțimea $\{x_k + \lambda p_k \mid \lambda > 0\}$ atunci există $\alpha > 0$ care satisface condițiile din regula Wolfe.

Demonstrație Conform ipotezei, funcția $\alpha \mapsto f(x_k + \alpha p_k)$ este mărginită inferior pe $(0, \infty)$.

Pentru orice $\alpha > 0$ există $t \in (0, \alpha)$ astfel încât

$$\begin{aligned} f(x_k + \alpha p_k) &= f(x_k) + \alpha \nabla f(x_k)(p_k) + \frac{1}{2} \alpha^2 \nabla^2 f(x_k + tp_k)(p_k, p_k) \\ &= f(x_k) + c_1 \alpha \nabla f(x_k)(p_k) + \alpha \left[(1 - c_1) \nabla f(x_k)(p_k) + \frac{1}{2} \alpha \nabla^2 f(x_k + tp_k)(p_k, p_k) \right], \end{aligned}$$

Cum $c_1 \in (0, 1)$ și $\nabla f(x_k)(p_k) < 0$, 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, ținând cont de proprietatea de mărginire, ecuația (în α)

$$f(x_k + \alpha p_k) = f(x_k) + c_1 \alpha \nabla f(x_k)(p_k)$$

are cel puțin o soluție strict pozitivă. Din continuitatea (în α) a funcțiilor implicate, există o cea mai mică soluție pozitivă pe care o notăm cu α' . Evident, pentru orice $\alpha \in (0, \alpha')$ are loc prima condiție. 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

$$\nabla f(x_k + \alpha'' p_k)(p_k) = c_1 \nabla f(x_k)(p_k) > c_2 \nabla f(x_k)(p_k).$$

Așadar, pentru α'' are loc și a doua condiție. Cum pentru α'' inegalitățile din ambele condiții sunt stricte, există chiar un interval în jurul acestui punct pe care aceste condiții sunt îndeplinite. $\square$

Vom vedea ulterior unele criterii de oprire pentru algoritmi bazați pe Regula Wolfe.

4.2.2 Convergența metodei direcției de descreștere. Cazul general

Discutăm acum convergența metodei celei mai bune descreșteri pentru cele mai importante dintre regulile de mai sus: Regula Armijo și Regula Wolfe.

O primă cerință pentru a avea rezultate de convergență semnificative este legată de unghiul dintre $\nabla f(x_k)$ și direcția (vectorul de normă 1) p_k . Dacă notăm cu θ_k acest unghi, atunci impunem să existe $\varepsilon > 0$ astfel încât pentru orice k

$$\cos \theta_k = \frac{\nabla f(x_k)(p_k)}{\|\nabla f(x_k)\|} \in [-1, -\varepsilon]. \quad (4.5)$$

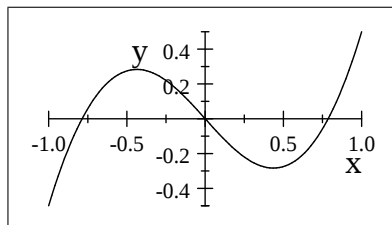
Astfel, chiar dacă p_k nu este cea mai bună direcție de descreștere, cerem totuși ca unghiul să nu se apropie oricât de mult de $\frac{\pi}{2}$. Evident, în cazul alegerii celei mai bune direcții de descreștere $\cos \theta_k = -1$ pentru orice k , deci condiția are loc.

Observația 4.2.3 Să considerăm funcția $f : \mathbb{R} \rightarrow \mathbb{R}$ 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 direcție de descreștere pentru o dată inițială $x_0 < -0.6$ (de exemplu) vom obține o descreștere infinită, deci metoda nu converge la punctul de minim. De aceea, și pentru această metodă, ideea de a aproxima (grafic sau prin altă tehnică) în primă instanță punctul de minim pentru o alegere potrivită a datei inițiale este importantă.



Teorema 4.2.4 Fie (x_k) șir de vectori generați după formula de recurență

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

unde direcțiile (p_k) satisfac (4.5) și pașii (α_k) satisfac Regula Armijo (4.3) cu parametrii α, β, c . Dacă $\bar{x}$ este un punct limită al șirului (x_k) , atunci $\bar{x}$ este punct critic al lui f (adică $\nabla f(\bar{x}) = 0$).

Demonstrație. Fără a restrânge generalitatea, putem presupune că (x_k) este convergent la $\bar{x}$, iar (p_k) este convergent la o direcție p . Dacă pasul α_k a fost validat la încercarea i în (4.3), atunci avem

$$f(x_k) - f(x_k + \alpha_k p_k) > -c\alpha_k \nabla f(x_k)(p_k) \geq \varepsilon c\alpha_k \|\nabla f(x_k)\|, \quad (4.6)$$

iar dacă încercarea $i - 1$ este invalidă

$$f(x_k) - f\left(x_k + \frac{\alpha_k}{\beta} p_k\right) \leq -c \frac{\alpha_k}{\beta} \nabla f(x_k)(p_k). \quad (4.7)$$

Presupunem, prin reducere la absurd, că $\nabla f(\bar{x}) \neq 0$. Cum $(f(x_k))$ este descrescător la $f(\bar{x})$ deducem că

$$f(x_k) - f(x_k + \alpha_k p_k) = f(x_k) - f(x_{k+1}) \rightarrow 0,$$

deci

$$\alpha_k \nabla f(x_k)(p_k) \rightarrow 0.$$

Dar, pe baza (4.5), prin trecere la limită, $\nabla f(\bar{x})(p) < 0$, deci $\alpha_k \rightarrow 0$. 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 funcții 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 $z_k \rightarrow \bar{x}$, prin simplificare și trecere la limită obținem contradicția $\nabla f(\bar{x})(p) \geq 0$. Deci $\nabla f(\bar{x}) = 0$. $\square$

Corolarul 4.2.5 *Presupunem că direcțiile de căutare satisfac (4.5). Dacă (α_k) sunt aleși după Regula minimizării exacte sau Regula minimizării limitate, atunci orice punct limită al lui (x_k) este punct critic pentru f .*

Demonstrație. Presupunem că un subșir (x_{k_r}) converge la un punct $\bar{x}$ care nu e critic. Fie $\bar{x}_{k_r} = x_{k_r} + \bar{\alpha}_{k_r} p_{k_r}$ punctul ales după Regula Armijo. Atunci, pentru orice r ,

$$f(x_{k_r}) - f(x_{k_r+1}) \geq f(x_{k_r}) - f(\bar{x}_{k_r+1}) \geq -c \bar{\alpha}_{k_r} \langle \nabla f(x_{k_r}), p_{k_r} \rangle.$$

Cum $f(x_k) \rightarrow f(\bar{x})$ avem $f(x_{k_r}) - f(x_{k_r+1}) \rightarrow 0$ și ca în demonstrația teoremei precedente, $\bar{\alpha}_{k_r} \rightarrow 0$.

Pe baza relației (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 raționamentul din rezultatul precedent, ajungem la aceeași contradicție. $\square$

Discutăm acum convergența algoritmului de căutare a direcției de descrescere pentru Regula Wolfe.

Teorema 4.2.6 (Zoutendijk) *Fie iterația (4.1) în care (p_k) sunt direcții de descrescere, iar (α_k) sunt pași ce satisfac condițiile Wolfe. Presupunem că f este de clasă C^1 și 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 condiția (4.5), atunci pentru orice punct limită $\bar{x}$ al lui (x_k) avem $\nabla f(\bar{x}) = 0$.

Demonstrație Din condițiile 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 condiția Lipschitz asupra diferențialei 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

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

Ținând 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 relațiile 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 funcției f , obținem convergența seriei din enunț.

Convergența demonstrată asigură faptul că

$$\cos^2 \theta_k \|\nabla f(x_k)\|^2 \rightarrow 0.$$

Dacă alegerea lui p_k se face astfel încât are loc condiția (4.5), atunci $\nabla f(x_k) \rightarrow 0$. Acum este evident că dacă $\bar{x}$ este punct limită pentru (x_k) , atunci $\bar{x}$ este punct critic. $\square$

Observația 4.2.7 Algoritmii de mai sus se aplică și funcțiilor convexe, determinându-se puncte de minim global (pentru că punctele critice au această proprietate) dar, cum vom vedea, forma particulară a convexității permite elaborarea unor algoritmi specifici puternici.

Implementare 4.15 Implementăm acum metoda celei mai bune descreșteri cu regula Armijo pentru o funcție de o variabilă reală. Așdar, direcția se alege

$$p_k = -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}$$

dacă $\|\nabla f(x_k)\| \neq 0$ iar, în ceea ce privește pasul α_k , vom utiliza doar condiția 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 îndeplinește condiția, iar în caz contrar, îl vom înmulți 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 funcția $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = 3e^x - x^3 + 5x$ care a fost discutată și 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*
        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 oprește atunci când valoarea absolută a derivatei în punctul curent este suficient de mică. Evoluția descreșterii pasului și a gradientului poate fi de asemenea elocventă.

Exercițiul 4.16 Să se implementeze metoda celei mai bune descreșteri cu regula Armijo pentru care derivata se va calcula numeric (a se vedea (3.7)).

Implementare 4.17 Scriem acum un cod pentru aplicarea metodei funcției Rosenbrock, funcție-test de două variabile reale pentru care știm 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)
        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 iterații, dar punctul de start poate fi îndepărtat de soluție.

Exercițiul 4.18 Să se implementeze algoritmul celei mai bune direcții de descreștere pentru funcția $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = e^x - 2x^2$.

Exercițiul 4.19 Fie $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = 2x^2 + 3e^{-2x}$. Să se arate că f este strict convexă și admite punct de minim. Să se determine acest punct cu aproximație folosind metoda lui Newton (aplicată lui $f'(x) = 0$) și metoda celei mai bune descreșteri.

Exercițiul 4.20 Să se arate că funcția $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^4 - x^2 + x - 1$ admite un singur punct de minim. Să se aproximeze acest punct folosind metoda celei mai bune descreșteri.

Exercițiul 4.21 Să se testeze metoda căutării direcției de descreștere cu data inițială $(-1, -1)$ pentru funcția $f : \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$f(x, y) = x^4 - 16x^2 + 5x + y^4 - 16y^2 + 5y.$$

Exercițiul 4.22 Să se implementeze metoda pentru $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, $f(x, y) = 4x^2 - 3xy + y^2$. Să se observe faptul că minimul (care este originea) găsindu-se într-o zonă mai accesibilă a graficului, numărul de iterații este semnificativ mai mic.

Exercițiul 4.23 Să se determine teoretic punctul de minim pentru $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ dată prin $f(x, y) = x^2(1 + y)^3 + y^2$ și să se aplice metoda de mai sus.

Exercițiul 4.24 Să se aplice metoda de mai sus pentru $f : \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$f(x, y) = x^6 + y^4 + 4x^2y^2 - 3xy + 2x + y.$$

Încheiem această secțiune folosind și Regula minimizării exacte pe care o implementăm cu ajutorul algoritmului raportului de aur. De asemenea, folosim un cod care calculează aproximativ gradientul funcției de minimizat.

Implementare 4.25 fișier funcție

```
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);
fișier funcție
function deriv = grad_vec(x,dex)
xvec = x;
xvec1 = x;
for i = 1:length(x)
    xvec = x;
    xvec1 = x;
    xvec(i) = x(i) + dex;
    xvec1(i) = x(i) - dex;
    deriv(i) = (func_multivar(xvec) - func_multivar(xvec1))/(2*dex);
end
fișier funcție
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);
fișier 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 funcționează foarte greu pe funcția Rosenbrock din cauza specificității acestei funcții și e de recomandat înlocuirea valorii 10^{-4} cu 10^{-2} în testul de stop.

Exercițiul 4.26 Fie $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^7 + x^3 + 1$. Să se arate că ecuația $f(x) = 0$ admite o singură rădăcină reală, $\bar{x}$. Să se justifice faptul că $\bar{x}$ este punct de minim global pentru $F : \mathbb{R} \rightarrow \mathbb{R}$, $F(x) = \int_0^x f(s)ds$. Să se aproximeze $\bar{x}$ folosind metoda celei mai bune descreșteri. Să se compare cu metoda lui Newton aplicată direct ecuației. Similar pentru funcția $f : (0, \infty) \rightarrow \mathbb{R}$, $f(x) = \ln x + 6x$.

Exercițiul 4.27 Fie $f : \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$f(x, y) = 0.6y^4 + 5x^2 - 7y^2 + \sin(xy) - 5y.$$

(i) Să se reprezinte grafic funcția f pe pătratul $[-5, 5]^2$ și să se deducă existența a două puncte de minim local.

(ii) Să se aproximeze punctele de minim local folosind metodele de mai sus.

Exercițiul 4.28 Fie $f : \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$f(x, y) = 0.7x^4 + 6y^2 - 8x^2 + \cos(xy) - 8x.$$

(i) Să se reprezinte grafic funcția f pe pătratul $[-5, 5]^2$.

(ii) Să se aproximeze punctele de minim local folosind metodele de mai sus.

Exercițiul 4.29 Fie $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ dată prin

$$f(x) = x_1^2 + 2x_2^2 + 4x_1 + 4x_2.$$

Considerăm metoda celei mai bune descreșteri cu regula minimizării exacte și data inițială $(0, 0)$. Să se arate că

$$x_k = \left(\frac{2}{3^k} - 2, \left(-\frac{1}{3} \right)^k - 1 \right), \quad \forall k$$

și să se deducă punctul de minim.

Exercițiul 4.30 Fie $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ 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 direcției de descreștere (cu diverse reguli de determinare a pasului).

(ii) Să se explice de ce metoda menționată necesită un număr relativ mare de iterații pentru aproximarea minimului.

Exercițiul 4.31 Să se rezolve problema minimizării funcției $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ 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 inițiale $(1, 1, 1)$, $(-2.3, 0, 0)$, $(0, 2, -12)$.

Exercițiul 4.32 Fie $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ dată prin

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

Să se reprezinte grafic, să se găsească punctele de minim global și să se arate că $(\sqrt[3]{4^{-1}}, \sqrt[3]{4^{-1}})$ este un punct critic care nu este extrem local. Să se utilizeze metoda căutării direcției de descreștere (cu diverse reguli de determinare a pasului) pentru datele inițiale $(4, 4)$, $(4, -4)$, $(-4, 4)$, $(-4, -4)$ și să se justifice rezultatele obținute.

Exercițiul 4.33 Fie $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ dată prin

$$f(x) = x_1^3 e^{x_2 - x_1^2 - 10(x_1 - x_2)^2}.$$

Să se reprezinte grafic pe $[-1.5, -0.5]^2$ și să se utilizeze metoda căutării direcției de descreștere (cu diverse reguli de determinare a pasului) pentru datele inițiale $(-3, -3)$, $(-3, 3)$, $(3, -3)$, $(3, 3)$.

Exercițiul 4.34 Fie $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ dată prin

$$f(x) = -x_2^2 e^{1 - x_1^2 - 20(x_1 - x_2)^2}.$$

Să se reprezinte grafic și să se utilizeze metoda căutării direcției de descreștere (cu diverse reguli de determinare a pasului) pentru datele inițiale $(0.1, 0.1)$, $(0.8, 0.1)$, $(1.1, 0.1)$.

4.2.3 Convergența metodei direcției de descreștere. Cazul convex

Atunci când adăugăm proprietăți suplimentare funcțiilor pe care le studiem, putem demonstra că algoritmi sunt mai eficienți decât în cazul general. Exemplificăm cu algoritmul direcției de descreștere pentru funcții convexe.

Astfel, în această secțiune, în cadrul rezultatelor principale considerăm $f : \mathbb{R}^p \rightarrow \mathbb{R}$ o funcție convexă, diferențiabilă cu diferențiala Lipschitz. Presupunem deci, pe lângă convexitate, că există $L > 0$ astfel încât

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \quad \forall x, y \in \mathbb{R}^p.$$

Această din urmă ipoteză a fost întâlnită și la Teorema lui Zoutendijk (Teorema 4.2.6). Începem cu un rezultat tehnic care pune în evidență rolul proprietății Lipschitz a diferențialei în obținerea unei majorări a funcției date (nu neapărat convexă în acest caz) printr-o funcție pătratică.

Lema 4.2.8 *Dacă $f : \mathbb{R}^p \rightarrow \mathbb{R}$ este diferențiabilă, cu diferențiala L -Lipschitz, atunci*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2, \quad \forall x, y \in \mathbb{R}^p. \quad (4.8)$$

Demonstrație. Fie $x, y \in \mathbb{R}^p$ și $\varphi_{y,x} : \mathbb{R} \rightarrow \mathbb{R}$,

$$\varphi_{y,x}(\lambda) = f(\lambda y + (1 - \lambda)x).$$

Avem

$$\varphi_{y,x}(1) - \varphi_{y,x}(0) = \int_0^1 \varphi'_{y,x}(\lambda) d\lambda = \varphi'_{y,x}(0) + \int_0^1 (\varphi'_{y,x}(\lambda) - \varphi'_{y,x}(0)) d\lambda.$$

Înlocuind și folosind ipoteza, obținem

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= \int_0^1 \langle \nabla f(x + \lambda(y - x)) - \nabla f(x), y - x \rangle d\lambda \\ &\leq \|y - x\| \int_0^1 \|\nabla f(x + \lambda(y - x)) - \nabla f(x)\| d\lambda \\ &\leq \|y - x\|^2 \int_0^1 \lambda d\lambda = \frac{L}{2} \|y - x\|^2. \end{aligned}$$

Aceasta este concluzia. □

Exercițiul 4.35 *Să se arate că condiția (4.8) este echivalentă cu convexitatea funcției $\mathbb{R}^p \ni x \mapsto \frac{L}{2} \|x\|^2 - f(x)$.*

Exercițiul 4.36 *Folosind funcția $\mathbb{R} \ni x \mapsto -e^x$ să se arate că condiția (4.8) nu implică faptul că diferențiala lui f este Lipschitz.*

Exercițiul 4.37 *Considerăm $f : \mathbb{R} \rightarrow \mathbb{R}$ dată prin*

$$f(x) = \begin{cases} \frac{3}{4}x^2, & \text{dacă } |x| < \frac{2}{3} \\ -\frac{3}{2}x^2 + 3x - 1, & \text{dacă } x \in [\frac{2}{3}, 1] \\ -\frac{3}{2}x^2 - 3x - 1, & \text{dacă } x \in [-1, -\frac{2}{3}] \\ |x| - \frac{x^2}{1}, & \text{dacă } |x| > 1. \end{cases}$$

Să se arate că f este diferențiabilă (derivabilă), iar cea mai mică constantă Lipschitz pentru derivată este 3. Să se arate că f satisface condiția (4.8) pentru $L = \frac{3}{2}$ (folosind, eventual, un grafic).

Observația 4.2.9 Așadar, pentru o funcție $f : \mathbb{R}^p \rightarrow \mathbb{R}$ o funcție convexă, diferențiabilă cu diferențiala L -Lipschitz, pentru orice $x, y \in \mathbb{R}^p$

$$f(y) + \langle \nabla f(y), x - y \rangle \leq f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|y - x\|^2. \quad (4.9)$$

Privită în raport cu x , adică pentru y fixat, estimarea inferioară (din partea stângă) este o funcție afină, în timp ce estimarea superioară (din partea dreaptă) este o funcție pătratică. Această funcție se scrie ca

$$\begin{aligned} h(x) &= f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|y - x\|^2 \\ &= \frac{L}{2} \left\| x - \left(y - \frac{1}{L} \nabla f(y) \right) \right\|^2 - \frac{1}{2L} \|\nabla f(y)\|^2 + f(y) \end{aligned}$$

și are minimul global în

$$\bar{x} = y - \frac{1}{L} \nabla f(y).$$

Aceasta sugerează următoare abordare algoritmică: dată fiind iterația 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 descreșteri cu pas

$$\alpha_k = \frac{\|\nabla f(x_k)\|}{L}.$$

Această observație este ilustrată prin exercițiul de mai jos, iar consistența acestei metode este demonstrată riguros după aceea.

Exercițiul 4.38 Fie $f : \mathbb{R} \rightarrow \mathbb{R}$ dată prin

$$f(x) = \frac{1}{2} (x - 1)^2 - \frac{1}{2} \ln(x^2 + 1) + 3.$$

Avem

$$\begin{aligned} f'(x) &= x - 1 - \frac{x}{x^2 + 1}, \\ f''(x) &= \frac{x^2(x^2 + 3)}{(x^2 + 1)^2}, \quad \forall x \in \mathbb{R}. \end{aligned}$$

Deci f este strict convexă pe $\mathbb{R}$. Se poate cu ușurință 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 același grafic cele trei funcții din estimarea (4.9).

(ii) Rezolvând ecuația $f'(x) = 0$ să se aproximeze punctul de minim.

(iii) Folosind iterația

$$x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$$

cu data inițială $x_0 = 3$ să se determine x_7 și să se compare cu punctul de minim determinat anterior.

Cum spuneam, demonstrăm acum faptul că iterația intuită mai sus și apoi ilustrată prin exemplul anterior conduce la un punct de minim. Începem cu un rezultat ajutător care pune în evidență caracterizări ale proprietății Lipschitz pentru diferențială în cazul funcțiilor convexe.

Lema 4.2.10 Dacă $f : \mathbb{R}^p \rightarrow \mathbb{R}$ este convexă și diferențiabilă pe C , atunci următoarele afirmații sunt echivalente:

(i) diferențiala lui f este L -Lipschitz, adică

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \quad \forall x, y \in \mathbb{R}^p;$$

(ii)

$$\frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2} \|y - x\|^2, \quad \forall x, y \in \mathbb{R}^p;$$

(iii)

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

Demonstrație. (i) $\implies$ (ii) Ca în lema de mai sus se obține a doua inegalitate de la (ii).

Pentru a proba prima inegalitate de la (ii), fixăm x și definim funcția convexă diferențiabilă $g : \mathbb{R}^p \rightarrow \mathbb{R}$ prin

$$g(y) = f(y) - \langle \nabla f(x), y \rangle.$$

Evident, pentru orice $y \in \mathbb{R}^p$

$$\nabla g(y) = \nabla f(y) - \nabla f(x)$$

și cum $\nabla g(x) = 0$ și g este convexă, deducem că x este punct de minim global pentru g . În plus, este simplu de observat că diferențiala lui g este L -Lipschitz. Astfel, folosind și 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 și avem

$$f(x) - \langle \nabla f(x), x \rangle \leq f(y) - \langle \nabla f(x), y \rangle - \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2,$$

și (ii) este complet demonstrată.

(ii) $\implies$ (iii) Scriem relațiile de la (ii) pentru cuplurile (x, y) și (y, x) și le adunăm.

(iii) $\implies$ (i) Evident. □

Teorema 4.2.11 Fie $f : \mathbb{R}^p \rightarrow \mathbb{R}$ o funcție convexă, diferențiabilă cu diferențiala L -Lipschitz. Presupunem că f admite punct de minim și considerăm următoarea metodă a căutării direcției de descreștere

$$x_0 \in \mathbb{R}^p,$$

$$x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k), \quad \forall k \geq 0.$$

Atunci, dacă $\nabla f(x_k) \neq 0$, pentru orice k , (x_k) converge la un punct de minim $\bar{x}$ al lui f , $(\|x_k - \bar{x}\|)$ este strict descrescător la 0 și au loc următoarele estimări

$$f(x_k) - f(\bar{x}) \leq \frac{5L}{2 \sum_{i=0}^{k-1} \|x_i - \bar{x}\|^{-2}}, \quad \forall k$$

$$\lim_k (f(x_k) - f(\bar{x})) = 0.$$

Demonstrație. Conform Lemei 4.2.8,

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

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

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, șirul este staționar și 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 și din Lema 4.2.8 avem

$$w_{k+1} - w_k \leq -\frac{L}{2} \|x_{k+1} - x_k\|^2, \quad \forall k,$$

și

$$w_{k+1} < w_k \leq \langle \nabla f(x^*), x_k - x^* \rangle + \frac{L}{2} \|x_k - x^*\|^2 = \frac{L}{2} \|x_k - x^*\|^2, \quad \forall k.$$

Avem, tot pe baza convexității,

$$f(x^*) \geq f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle, \quad \forall k,$$

deci

$$w_k \leq -\langle \nabla f(x_k), x^* - x_k \rangle \leq \|\nabla f(x_k)\| \|x_k - x^*\| = L \|x_{k+1} - x_k\| \|x_k - x^*\|, \quad \forall k.$$

Deducem

$$\begin{aligned} f(x_k) - f(x_{k+1}) &= w_k - w_{k+1} \geq \frac{L}{2} \|x_{k+1} - x_k\|^2 \\ &\geq \frac{L}{2} \left(\frac{w_k}{L \|x_k - x^*\|} \right)^2 = \frac{w_k^2}{2L \|x_k - x^*\|^2} \geq \frac{w_{k+1}^2}{2L \|x_k - x^*\|^2}, \quad \forall k. \end{aligned}$$

Rescriem inegalitatea de mai sus astfel:

$$\frac{1}{w_k} \leq \frac{1}{w_{k+1}} \left(1 + \frac{w_{k+1}}{2L \|x_k - x^*\|^2} \right)^{-1}, \quad \forall k.$$

Cum

$$w_{k+1} \leq \frac{L}{2} \|x_k - x^*\|^2,$$

obținem că

$$\frac{w_{k+1}}{2L \|x_k - x^*\|^2} \leq \frac{1}{4}.$$

Pe de altă parte, este simplu de observat că

$$\forall t \in \left[0, \frac{1}{4}\right], \quad \frac{1}{1+t} \leq 1 - \frac{4t}{5},$$

deci

$$\frac{1}{w_k} \leq \frac{1}{w_{k+1}} - \frac{2}{5L \|x_k - x^*\|^2}, \quad \forall k.$$

Sumăm aceste relații de la 0 la $k-1$ și obținem

$$0 \leq \frac{1}{w_0} \leq \frac{1}{w_k} - \frac{2}{5L} \sum_{i=0}^{k-1} \|x_i - x^*\|^{-2},$$

ceea ce ne conduce la relația

$$f(x_k) - \inf_{x \in \mathbb{R}^p} f(x) \leq \frac{5L}{2 \sum_{i=0}^{k-1} \|x_i - x^*\|^{-2}}, \quad \forall k.$$

Arătăm în continuare că $(\|x_k - x^*\|)$ este strict descrescător. Avem, folosind faptul că $\nabla f(x^*) = 0$ și Lema 4.2.10,

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \left\| x_k - x^* - \frac{1}{L} \nabla f(x_k) \right\|^2 \\ &= \|x_k - x^*\|^2 + \left\| \frac{1}{L} \nabla f(x_k) \right\|^2 - \frac{2}{L} \langle x_k - x^*, \nabla f(x_k) - \nabla f(x^*) \rangle \\ &\leq \|x_k - x^*\|^2 + \frac{\|\nabla f(x_k)\|^2}{L^2} - 2 \frac{\|\nabla f(x_k)\|^2}{L^2} < \|x_k - x^*\|^2, \quad \forall k, \end{aligned}$$

adică ceea ce dorim.

În particular, (x_k) este mărginit, deci are un subșir (x_{k_r}) convergent la un punct pe care îl notăm $\bar{x}$. Dar, am văzut mai sus,

$$f(x_{k_{r+1}}) - f(x_{k_r}) \leq f(x_{k_r+1}) - f(x_{k_r}) \leq -\frac{1}{2L} \|\nabla f(x_{k_r})\|^2, \quad \forall r$$

și trecând la limită deducem

$$0 \leq -\frac{1}{2L} \|\nabla f(\bar{x})\|^2,$$

adică $\nabla f(\bar{x}) = 0$ și $\bar{x}$ 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 și pentru $\bar{x}$ în loc de x^* . În particular, $(\|x_k - \bar{x}\|)$ este strict descrescător și are loc inegalitatea din concluzie. Aceasta înseamnă că șirul (x_k) este convergent la $\bar{x}$.

Demonstrăm acum ultima concluzie. Definim $(\alpha_k) = (\|x_k - \bar{x}\|^{-2})$, șir care este crescător cu limita $+\infty$. Relația

$$f(x_k) - f(\bar{x}) \leq \frac{5L}{2 \sum_{i=0}^{k-1} \|x_i - \bar{x}\|^{-2}}, \quad \forall k$$

se rescrie în forma

$$k(f(x_k) - f(\bar{x})) \leq \frac{5L}{2} \left(\frac{1}{k} \sum_{i=0}^{k-1} \alpha_i \right)^{-1}, \quad \forall k.$$

Pentru a obține concluzia, este suficient să arătăm că

$$\frac{1}{k} \sum_{i=0}^{k-1} \alpha_i \rightarrow +\infty.$$

Fie $M > 0$ și $n \in \mathbb{N}$ 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 ușor de verificat, pe baza monotoniei lui $(\|x_k - \bar{x}\|)$ că șirul $\left(\frac{1}{k} \sum_{i=0}^{k-1} \alpha_i\right)$ este crescător. Aceste observații încheie demonstrația. $\square$

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);

```

Exercițiul 4.40 Să se implementeze metoda de mai sus pentru $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ dată prin

$$f(x, y) = 20x^2 + y^4 + y.$$

4.2.4 Posibilități de accelerare a metodei gradient descent

Algoritmul gradient descent poate avansa foarte lent în regiuni plate ale graficului funcției obiectiv, după cum am văzut deja la unele funcții considerate anterior (de exemplu, funcția Rosenbrock). Câteva soluții pentru evitarea acestei dificultăți 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 direcții care s-au dovedit eficiente anterior și a încetini în direcții unde gradientul s-a schimbat brusc, similar cu o bilă care se rostogolește la vale. Aceasta poate fi implementată astfel:

$$\begin{aligned}
m_{k+1} &= \gamma m_k - \alpha \nabla f(x_k), \\
x_{k+1} &= x_k + m_{k+1},
\end{aligned}$$

unde m_k se numește impuls, $m_0 = 0$ și $0 < \gamma < 1$. O valoare tipică pentru γ este 0.9. Pentru $\gamma = 0$, metoda devine gradient descent obișnuită.

Observăm ca m_k este ca o medie exponențial ponderată a gradientilor anteriori:

$$m_{k+1} = \gamma m_k - \alpha \nabla f(x_k) = \gamma^2 m_{k-1} - \alpha \gamma \nabla f(x_{k-1}) - \alpha \nabla f(x_k) = \dots = -\alpha \sum_{i=0}^k \gamma^i \nabla f(x_{k-i}).$$

Deoarece actualizăm parametrii folosind media gradientului m_{k+1} , și nu ultimul gradient, $\nabla f(x_k)$, observăm ca gradientii anteriori pot influența iterația 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 computaționale suplimentare.

O problemă care poate apărea în momentum method este că poate să nu încetinească suficient în partea de jos a unei "văi", cauzând oscilații. Metoda *gradientului accelerat Nesterov* (sau *Nesterov accelerated gradient*) modifică metoda de tip gradient descent pentru a include un pas de extrapolare, astfel:

$$\begin{aligned}
\tilde{x}_{k+1} &= x_k + \beta(x_k - x_{k-1}), \\
x_{k+1} &= \tilde{x}_{k+1} - \alpha_k \nabla f(\tilde{x}_{k+1}),
\end{aligned}$$

sau, altfel zis,

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

Aceasta este, în mod esențial, o formă de "a privi înainte" cu un pas, care poate reduce nivelul oscilației.

Gradientul accelerat Nesterov poate fi rescris în aceeași formă ca momentum method. În acest caz, impulsul este actualizat folosind gradientul la noua locație prevăzută:

$$\begin{aligned} m_{k+1} &= \beta m_k - \alpha \nabla f(x_k + \beta m_k), \\ x_{k+1} &= x_k + m_{k+1}. \end{aligned}$$

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 direcția corectă, astfel încât măsurarea gradientului la noua locație, $x_k + \beta m_k$, și nu la locația curentă, x_k , poate fi mai precisă.

Metoda gradientului accelerat Nesterov este demonstrabil mai rapidă decât gradientul descendent pentru funcții convexe, pentru β și α_k sunt alese corespunzător. Se numește "accelerat" datorită acestei îmbunătățiri, care este optimă pentru metodele de tip gradient descent folosind doar informații de ordinul întâi, atunci când funcția obiectiv este convexă și are gradienti Lipschitz continui. În practică, totuși, această metodă poate fi mai lentă decât steepest descent, și poate deveni instabilă pentru alegeri nepotrivite ale lui β și α .

Implementare 4.41 Considerăm următoarea implementare a GD pentru funcția 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 și GD cu impuls Nesterov și să se compare
rezultatul cu metoda GD originală de mai sus.

```

4.2.5 Algoritmi specifici *machine learning* bazați pe metoda direcției de descreștere

Descriem acum, fără demonstrații privind convergența, câțiva algoritmi derivați din metoda direcției de descreștere care sunt adaptați specificității unor probleme ce apar recurent în machine learning. Încercăm să prezentăm la nivel euristic, intuitiv motivele pentru care aceste variații sunt mai potrivite anumitor probleme de mari dimensiuni decât algoritmul original.

Stochastic Gradient Descent

Se consideră un set de date de antrenament, $\{(a_i, y_i)\}_{i=1, \overline{n}}$, cu $a_i \in \mathbb{R}$ și $y_i \in \mathbb{R}^p$, ce urmează o distribuție necunoscută. Funcția obiectiv este:

$$f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x) = \frac{1}{n} \sum_{i=1}^n \ell(h(x, a_i), y_i).$$

Exemplu de funcție de pierdere ("loss functions"): "square loss":

$$\ell(h(x, a_i), y_i) = \sum_{i \in \overline{1, n}} \|y_i - h(x, a_i)\|^2.$$

Vom considera algoritmul de tipul steepest descent method (or gradient descent, GD):

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

(adică $\alpha_k = \beta \|\nabla f(x_k)\|$, pentru orice k), unde gradientul funcției obiectiv este

$$\nabla f(x) = \frac{1}{n} \sum_{i=1}^n \nabla f_i(x).$$

Factorul $\beta > 0$ se numește rată de învățare (*learning rate*). Dacă folosim acest algoritm, costul computațional crește liniar cu n . Așadar, cu cât setul de date de antrenament este mai mare, cu atât costul computațional pentru fiecare iterație este mai mare.

Stochastic Gradient Descent (SGD) reduce costul computațional la fiecare iterație. Fie $m < n$. La fiecare pas se alege o mulțime $\mathcal{I}$ de indici (adică un lot, *minibatch*) având cardinalul m și actualizăm iterația 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 observația (ce poate fi riguros demonstrată) conform căreia, dacă mulțimile $\mathcal{I}$ sunt aleator alese, atunci gradientul stochastic este un estimator nedeplasat al gradientului complet, i.e.,

$$\mathbb{E} \left(\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \nabla f_i(x) \right) = \nabla f(x), \quad \forall x \in \mathbb{R}^p.$$

Implementare 4.42 Considerăm exemplul $f : \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$f(x, y) = (x - 2)^2 + (y - 3)^2 + 2(x + 1)^2 + (y - 1)^2 + (x - 1)^2 + 2(y + 2)^2,$$

adică suma funcțiilor $f_i : \mathbb{R}^2 \rightarrow \mathbb{R}$ ($i \in \overline{1, 3}$),

$$f_1(x, y) = (x - 2)^2 + (y - 3)^2, \quad f_2(x, y) = 2(x + 1)^2 + (y - 1)^2, \quad f_3(x, y) = (x - 1)^2 + 2(y + 2)^2.$$

Să se scrie un program care să implementeze SGD cu rata de învățare este constantă 0.05, rulează 100 de iterații și alege câte unul din cei trei gradienti cu egală probabilitate (minibatch de cardinal 1). Pentru selecția gradientilor 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 direcției de descreștere corespunzător.

Implementare 4.43 Să se implementeze SGD cu minibatch de cardinal 2 pentru minimizarea funcției de mai sus. Să se compare cu metoda GD originală.

Deși rata teoretică de convergență a SGD este mai lentă decât cea a GD pe întregul set de date de antrenament $\{1, 2, \dots, n\}$, în practică SGD este adesea mai rapid, deoarece timpul pentru fiecare pas este mult mai scăzut (într-un program Matlab/Octave timpul de execuție este afișat prin includerea instrucțiunilor `tic...toc` la începutul, respectiv sfârșitul secvenței avute în vedere). Pe de altă parte, se pot considera rate de învățare dinamice. Pentru a înțelege 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 puțin) de K ori mai lentă decât SGD, deoarece va pierde timp calculând gradientul pentru exemplele repetate. Chiar și în absența duplicatelor, antrenarea pe întregul set de date poate fi ineficientă, deoarece, la începutul antrenării, parametrii nu sunt bine estimați, ceea ce face ca evaluarea în detaliu a gradientului poate să nu merite efortul.

SGD preconditionat

Se poate considera de asemenea un SGD preconditionat, 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 preconditionare, 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 direcția de descreștere folosind o matrice de preconditionare completă. Prin urmare, în practică vom folosi de obicei o matrice diagonală M_k . Aceste preconditionări nu folosesc neapărat informații de ordinul doi, dar adesea duc la accelerări în comparație cu SGD simplu.

AdaGrad

Metoda AdaGrad (de la "gradient adaptiv") a fost concepută inițial pentru optimizarea funcțiilor obiectiv convexe unde multe componente ale gradientului sunt zero; aceasta corespunde situației

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 $g_k = \nabla f(x_k)$,

$$s_{k,d} = \sum_{i=1}^k g_{i,d}^2$$

reprezintă suma componentelor gradientilor la pătrat, iar $\varepsilon > 0$ este un termen pozitiv mic introdus pentru a evita împărțirea la zero.

Văzută ca SGD preconditionat, aceasta este echivalentă cu

$$M_k = \text{diag}(s^{(k)} + \varepsilon)^{\frac{1}{2}}$$

și este un exemplu de *rată de învățare adaptativă*.

RMSProp și AdaDelta

O caracteristică definitorie a algoritmului AdaGrad este că termenul de la numitor devine mare în timp, deci rata efectivă de învățare scade. Deși necesară pentru a asigura convergența, poate influența în mod negativ performanța dacă numitorul devine mare prea repede.

O metodă alternativă este de a utiliza o medie ponderată exponențial a pătratelor gradientilor precedenți, în locul sumei acestora:

$$s_{k+1,d} = \beta s_{k,d} + (1 - \beta) g_{k,d}^2.$$

În practică, se folosește de obicei $\beta \sim 0.9$, care dă mai multă importanță iterațiilor recente. În acest caz,

$$\sqrt{s_{k,d}} \approx \text{RMS}(g_{1:k,d}) = \sqrt{\frac{1}{k} \sum_{i=1}^k g_{i,d}^2},$$

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 exponențial 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 aceeași expresie ca la metoda RMSProp de mai sus. Metoda are avantajul că numărătorul și 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 învățare α_k , pe care o putem alege 1, însă deoarece aceste rate adaptative nu descresc, aceste metode nu converg în mod sigur la o soluție.

Adam

Este posibil să combinăm RMSProp cu metoda impulsului (momentum method). În particular, calculăm EWMA, o medie ponderată exponențială a gradientilor (ca în metoda impulsului) și a gradientilor la pătrat (ca în RMSProp):

$$\begin{aligned} m_k &= \beta_1 m_{k-1} + (1 - \beta_1) g_k, \\ s_k &= \beta_2 s_{k-1} + (1 - \beta_2) g_k^2, \end{aligned}$$

unde în a doua ecuație egalitatea se realizează pe componente. Facem apoi următoarea actualizare (unde egalitatea este înțeleasă 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 $\beta_1 = 0.9$, $\beta_2 = 0.999$ și $\varepsilon = 10^{-6}$. (Dacă setăm $\beta_1 = 0$ și nu folosim corecția de bias, recuperăm RMSProp, care nu utilizează momentum). Pentru rata de învățare generală, se obișnuiește să folosim o valoare fixă, cum ar fi $\alpha_k = 0.001$. Din nou, deoarece rata de învățare adaptivă poate să nu scadă în timp, convergența nu este garantată.

Dacă inițializăm $m_0 = s_0 = 0$, atunci estimările inițiale 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 învățare este determinată de $\alpha_0 M_k^{-1}$, care se schimbă în timp. Prin urmare, aceste metode sunt adesea numite metode cu *rată de învățare adaptivă*. Totuși, ele necesită în continuare setarea unei rate de învățare inițiale α_0 .

Deoarece metodele EWMA sunt utilizate de obicei în cadru stochastic, unde estimările gradientului au zgomot, adaptarea ratei de învățare poate duce la pierderea convergenței chiar și pentru probleme convexe. Au fost propuse diverse soluții 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

$$s_k = s_{k-1} + (1 - \beta_2)(g_k^2 - s_{k-1}) \odot \text{sgn}(g_k^2 - s_{k-1}),$$

unde sgn reprezintă funcția signum, având valoarea 1 pentru numere strict pozitive, -1 pentru numere strict negative și 0 pentru 0.

Așadar, am obținut 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 inițial $x^{(0)}$

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: $s_k = s_{k-1} + (1 - \beta_2)(g_k^2 - s_{k-1}) \odot \text{sgn}(g_k^2 - s_{k-1})$

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 afișează x_{k+1} și $f(x_{k+1})$.

Comparativ cu celelalte metode prezentate, algoritmiul Adam se comportă mai bine, în special aplicat pentru rețele neuronale de tip multilayer.

4.2.6 Convergența metodei direcției de descreștere. Cazul funcțiilor pătratice

Să analizăm cazul funcției $f : \mathbb{R}^p \rightarrow \mathbb{R}$ 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, funcția este strict convexă și are mulțimile de nivel mărginite, deci există un unic punct de minim dat de ecuația $\nabla f(x) = 0$, adică $Qx + b = 0$. Deci minimul este $\bar{x} = -Q^{-1}b$.

Pentru probleme de mari dimensiuni calculul lui Q^{-1} poate fi costisitor din punct de vedere computațional.

Să studiem acum comportarea algoritmului celei mai bune direcții de descreștere cu pasul dat de Regula minimizării exacte.

Astfel, $x_0 \in \mathbb{R}^p$ și pentru orice $k \geq 0$,

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

unde

$$p_k = -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|} = -\frac{1}{\|Qx_k + b\|} (Qx_k + b),$$

iar α_k este minimul funcției scalare $\alpha \mapsto f(x_k + \alpha p_k)$.

Presupunem că gradientul nu se anulează în punctele iterației (adică $Qx_k + b \neq 0$), ceea ce, ținând cont de convexitatea problemei, este echivalent cu faptul că $f(x_k) > \bar{f}$, unde $\bar{f}$ este valoarea minimă a funcției.

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 funcție are punctul de minim în

$$\alpha_k = -\frac{\langle Qx_k + b, p_k \rangle}{\langle Qp_k, p_k \rangle} = \frac{\|Qx_k + b\|^3}{\langle Q(Qx_k + b), Qx_k + b \rangle}.$$

Deci iterația este

$$\begin{aligned} x_{k+1} &= x_k + \frac{\|Qx_k + b\|^3}{\langle Q(Qx_k + b), Qx_k + b \rangle} p_k \\ &= x_k + \frac{\|Qx_k + b\|^2}{\langle Q(Qx_k + b), Qx_k + b \rangle} (-Qx_k - b) \\ &= x_k - \frac{\|Qx_k + b\|^2}{\langle Q(Qx_k + b), Qx_k + b \rangle} (Qx_k + b) \end{aligned}$$

Observația 4.2.12 Dacă $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ 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

$$Q = \begin{pmatrix} 2\alpha & \beta \\ \beta & 2\gamma \end{pmatrix} \text{ și } b = (\delta, \varepsilon).$$

Exercițiul 4.44 Să se implementeze algoritmul descris mai sus.

Este clar că rezultatele generale de convergență pentru metoda celei mai bune descreșteri se aplică și acestei iterației de mai sus. Particularitățile problemei conduc la un rezultat de convergență specific acestui caz. Pentru o discuție 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 numește vector propriu pentru A dacă există $\lambda \in \mathbb{C}$ astfel încât $Ax = \lambda x$. Scalarul λ se numește 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 $U^t U = U U^t = I$ și o matrice diagonală $D = \text{diag}(d_1, \dots, d_p)$ astfel încât $U^t A U = D$.

Mai mult, coloanele matricii U reprezintă o bază ortonormală a spațiului $\mathbb{R}^p$ formată din vectori proprii și elementele lui D sunt valorile proprii corespunzătoare.

O consecință a Teoremei 4.2.13 este că urma matricii A este suma valorilor sale proprii, iar determinantul 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

$$\|x\|^4 \leq \langle Qx, x \rangle \cdot \langle Q^{-1}x, x \rangle \leq \frac{1}{4} \left(\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} + \sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}} \right)^2 \|x\|^4,$$

unde $\lambda_{\max}$ și $\lambda_{\min}$ reprezintă cea mai mare și respectiv cea mai mică valoare proprie a lui Q .

Demonstrație Este suficient să demonstrăm inegalitatea pentru orice x de normă 1. Așa cum am amintit în introducere, valorile proprii ale lui Q sunt reale strict pozitive și, fără a restrânge generalitatea, le ordonăm descrescător: $\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p = \lambda_{\min}$. Notăm matricea diagonală având valorile proprii pe diagonala principală, în ordinea precizată, cu $D := \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$. În plus, conform Teoremei 4.2.13, există o matrice ortogonală B cu proprietatea că $Q = B^t D B$. Atunci $Q^{-1} = (B^t D B)^{-1} = B^t D^{-1} B$, iar $D^{-1} = \text{diag}(\lambda_1^{-1}, \lambda_2^{-1}, \dots, \lambda_p^{-1})$. Deci

$$\langle Qx, x \rangle = \langle B^t D B x, x \rangle = \langle D B x, B x \rangle$$

și

$$\langle Q^{-1}x, x \rangle = \langle B^t D^{-1} B x, x \rangle = \langle D^{-1} B x, B x \rangle.$$

Pe de altă parte, aplicația $x \mapsto (Bx)$ este o bijecție de la sfera unitate a lui $\mathbb{R}^p$ în ea însăși deci, pentru a obține relația din enunț, este suficient să demonstrăm că pentru orice $u \in \mathbb{R}^p$ de normă 1 are loc

$$1 \leq \langle Du, u \rangle \cdot \langle D^{-1}u, u \rangle \leq \frac{1}{4} \left(\sqrt{\frac{\lambda_1}{\lambda_p}} + \sqrt{\frac{\lambda_p}{\lambda_1}} \right)^2 \|u\|^4.$$

Dacă $\lambda_1 = \lambda_p$ atunci are loc egalitatea. Presupunem că $\lambda_p < \lambda_1$. Avem

$$\langle Du, u \rangle = \sum_{i=1}^p u_i^2 \lambda_i; \quad \langle D^{-1}u, u \rangle = \sum_{i=1}^p u_i^2 \frac{1}{\lambda_i}.$$

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

$$\frac{1}{(\sum_{i=1}^p u_i^2 \lambda_i)} \leq \sum_{i=1}^p u_i^2 \frac{1}{\lambda_i}.$$

Cum $\sum_{i=1}^p u_i^2 = 1$, inegalitatea anterioară rezultă din inegalitatea lui Jensen aplicată funcției 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

$$\sum_{i=1}^p u_i^2 \frac{1}{\lambda_i} \leq \frac{1}{\lambda_1} + \frac{1}{\lambda_p} - \frac{\sum_{i=1}^p u_i^2 \lambda_i}{\lambda_1 \lambda_p}.$$

Obținem

$$\begin{aligned} \left(\sum_{i=1}^p u_i^2 \frac{1}{\lambda_i} \right) \left(\sum_{i=1}^p u_i^2 \lambda_i \right) &\leq \left(\sum_{i=1}^p u_i^2 \lambda_i \right) \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_p} - \frac{\sum_{i=1}^p u_i^2 \lambda_i}{\lambda_1 \lambda_p} \right) \\ &= \frac{(\sum_{i=1}^p u_i^2 \lambda_i) (\lambda_1 + \lambda_p - \sum_{i=1}^p u_i^2 \lambda_i)}{\lambda_1 \lambda_p}. \end{aligned}$$

Funcția de gradul al doilea

$$\lambda \mapsto \frac{\lambda(\lambda_1 + \lambda_p - \lambda)}{\lambda_1 \lambda_p}$$

își atinge maximul pentru $\lambda = \frac{\lambda_1 + \lambda_p}{2}$, deci

$$\begin{aligned} \left(\sum_{i=1}^p u_i^2 \frac{1}{\lambda_i} \right) \left(\sum_{i=1}^p u_i^2 \lambda_i \right) &\leq \left(\frac{\lambda_1 + \lambda_p}{2} \right)^2 \frac{1}{\lambda_1 \lambda_p} \\ &= \frac{1}{4} \left(\sqrt{\frac{\lambda_1}{\lambda_p}} + \sqrt{\frac{\lambda_p}{\lambda_1}} \right)^2. \end{aligned}$$

Demonstrația este completă. □

Teorema 4.2.15 *Cu notațiile de mai sus are loc inegalitatea*

$$f(x_k) - \min f \leq (f(x_0) - \min f) \left(\frac{c-1}{c+1} \right)^{2k}, \quad \forall k \geq 0,$$

unde $c = \lambda_{\max}/\lambda_{\min}$, unde $\lambda_{\max}$ și $\lambda_{\min}$ reprezintă cea mai mare și respectiv cea mai mică valoare proprie a lui Q .

Demonstrație. Notăm, pentru orice k ,

$$\begin{aligned} e_k &= f(x_k) - \min f, \\ r_k &= Qx_k + b. \end{aligned}$$

Folosind calculele de mai sus, avem

$$\begin{aligned} p_k &= -\frac{r_k}{\|r_k\|}, \\ \alpha_k &= \frac{\|r_k\|^3}{\langle Qr_k, r_k \rangle}, \end{aligned}$$

$$\begin{aligned}
e_{k+1} &= e_k + \frac{1}{2} \alpha_k^2 \langle Q p_k, p_k \rangle + \alpha_k \langle Q x_k + b, p_k \rangle \\
&= e_k + \frac{1}{2} \frac{\|r_k\|^6}{\langle Q r_k, r_k \rangle^2} \frac{\langle Q r_k, r_k \rangle}{\|r_k\|^2} + \frac{\|r_k\|^3}{\langle Q r_k, r_k \rangle} \left\langle r_k, -\frac{r_k}{\|r_k\|} \right\rangle \\
&= e_k - \frac{1}{2} \frac{\|r_k\|^4}{\langle Q r_k, r_k \rangle}, \quad \forall k.
\end{aligned}$$

Similar,

$$\begin{aligned}
e_k &= f(x_k) - f(\bar{x}) = \langle \nabla f(\bar{x}), x_k - \bar{x} \rangle + \frac{1}{2} \langle Q(x_k - \bar{x}), x_k - \bar{x} \rangle \\
&= \frac{1}{2} \langle Q(x_k - \bar{x}), x_k - \bar{x} \rangle = \frac{1}{2} \langle Q x_k + b, x_k + Q^{-1} b \rangle \\
&= \frac{1}{2} \langle Q^{-1} r_k, r_k \rangle, \quad \forall k.
\end{aligned}$$

Astfel, prin aplicarea Inegalității lui Kantorovici, obținem

$$\begin{aligned}
\frac{e_{k+1}}{e_k} &= 1 - \frac{1}{2} \frac{1}{e_k} \frac{\|r_k\|^4}{\langle Q r_k, r_k \rangle} = 1 - \frac{1}{2} \frac{2}{\langle Q^{-1} r_k, r_k \rangle} \frac{\|r_k\|^4}{\langle Q r_k, r_k \rangle} \\
&= 1 - \frac{\|r_k\|^4}{\langle Q^{-1} r_k, r_k \rangle \langle Q r_k, r_k \rangle} \leq 1 - 4 \left(\sqrt{\frac{\lambda_1}{\lambda_p}} + \sqrt{\frac{\lambda_p}{\lambda_1}} \right)^{-2} \\
&= 1 - 4 \left(\frac{\lambda_{\max} + \lambda_{\min}}{\sqrt{\lambda_{\max} \cdot \lambda_{\min}}} \right)^{-2} \\
&= 1 - 4 \frac{\lambda_{\max} \cdot \lambda_{\min}}{(\lambda_{\max} + \lambda_{\min})^2} = \frac{(\lambda_{\max} - \lambda_{\min})^2}{(\lambda_{\max} + \lambda_{\min})^2} = \left(\frac{c-1}{c+1} \right)^2, \quad \forall k.
\end{aligned}$$

Aceasta asigură validitatea concluziei. □

Observația 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 convergență a metodei să fie lentă, în general.

Exercițiul 4.45 Fie $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ 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 descreșteri pentru $x = (2, 3)$ să se determine teoretic și practic câte iterații sunt necesare pentru ca $f(x_k) - \min f \leq 10^{-10}$.

4.2.7 Metoda gradientilor conjugați

Metoda pe care o prezentăm acum se referă la rezolvarea unui sistem de ecuații liniare de forma

$$Qx = b,$$

unde Q este o matrice pătratică de dimensiune p , simetrică și pozitiv definită. Este de observat că rezolvarea sistemului de mai sus este echivalentă cu minimizarea funcției pătratice strict convexe $q : \mathbb{R}^p \rightarrow \mathbb{R}$,

$$q(x) = \frac{1}{2} \langle Qx, x \rangle - \langle b, x \rangle$$

pentru că

$$\nabla q(x) = Qx - b.$$

Notăm

$$r(x) := \nabla q(x) = Qx - b.$$

Definiția 4.2.17 Fie Q o matrice pătratică de dimensiune p , simetrică și pozitiv definită. Definim pe $\mathbb{R}^p \times \mathbb{R}^p$ aplicația cu valori reale

$$\langle x, y \rangle_Q = \langle Qx, y \rangle,$$

numită Q -produs scalar.

Observația 4.2.18 Se verifică ușor că aplicația de mai sus este într-adevăr un produs scalar, iar norma corespunzătoare, numită Q -normă, este

$$\|x\|_Q = \sqrt{\langle Qx, x \rangle}.$$

Definiția 4.2.19 O mulțime de direcții (i.e., vectori nenuli) $(d_i)_{i \in \overline{1, k}}$ (unde $k \in \mathbb{N} \setminus \{0\}$) se numește Q -ortogonală, iar elementele sale Q -conjugate dacă

$$\|d_i\|_Q \neq 0, \forall i \in \overline{1, k} \text{ și } \langle d_i, d_j \rangle_Q = 0, \forall i, j \in \overline{1, k}, i \neq j.$$

Lema 4.2.20 O mulțime de direcții Q -conjugate este liniar independentă.

Demonstrație. Fie $(d_i)_{i \in \overline{1, k}}$ (unde $k \in \mathbb{N} \setminus \{0\}$) o mulțime de direcții Q -conjugate și fie $(\alpha_i)_{i \in \overline{1, n}}$ numere reale. Dacă $\sum_{i=1}^k \alpha_i d_i = 0$, atunci, pentru orice $j \in \overline{1, k}$,

$$0 = \left\langle \sum_{i=1}^k \alpha_i d_i, Qd_j \right\rangle = \sum_{i=1}^k \alpha_i \langle d_i, Qd_j \rangle = \alpha_j \|d_j\|_Q^2.$$

Obținem astfel că toți scalarii sunt nuli. □

Observația 4.2.21 Dacă $(d_i)_{i \in \overline{1, k}}$ ($k \in \mathbb{N} \setminus \{0\}$) este o mulțime de direcții Q -conjugate și $x_0 \in \mathbb{R}^p$, atunci a minimiza funcția q pe subspațiul afin

$$A = x_0 + \text{span} \left((d_i)_{i \in \overline{1, k}} \right)$$

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 obține minimizând separat pentru fiecare a_i ; avem minimul

$$\bar{x} = x_0 + \sum_{i=1}^k a_i d_i, \text{ unde } a_i = -\frac{\langle r_0, d_i \rangle}{\langle Qd_i, d_i \rangle}.$$

Observația 4.2.22 Având o mulțime de vectori liniar independenți $(v_i)_{i \in \overline{1, k}}$ putem genera, prin procedeul Gram-Schmidt, o mulțime de direcții Q -conjugate $(d_i)_{i \in \overline{1, k}}$ astfel încât $\text{span} \left((v_i)_{i \in \overline{1, k}} \right) = \text{span} \left((d_i)_{i \in \overline{1, k}} \right)$. Procedeul este descris mai jos.

Alegem $d_1 = v_1$. Dacă am construit $(d_j)_{j \in \overline{1, i-1}}$, definim

$$d_i = v_i + \sum_{j=1}^{i-1} \beta_j d_j,$$

unde coeficienții β_j sunt determinați astfel: cum dorim ca $\langle d_i, d_l \rangle_Q = 0$ pentru orice $l < i$, trebuie ca

$$\begin{aligned} 0 &= \langle d_i, d_l \rangle_Q = \left\langle v_i + \sum_{j=1}^{i-1} \beta_j d_j, d_l \right\rangle_Q \\ &= \langle v_i, d_l \rangle_Q + \sum_{j=1}^{i-1} \beta_j \langle d_j, d_l \rangle_Q = \langle v_i, d_l \rangle_Q + \beta_l \langle d_l, d_l \rangle_Q, \quad \forall l < i \end{aligned}$$

deci

$$\beta_l = -\frac{\langle v_i, d_l \rangle_Q}{\langle d_l, d_l \rangle_Q}, \quad \forall l < i$$

și 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 direcțiilor conjugate.

Fie $x_0 \in \mathbb{R}^p$ și $(d_i)_{i \in \overline{0, n-1}}$ ($n \in \mathbb{N}^*$) este o mulțime de direcții Q -conjugate. Pentru $k \in \overline{0, n-1}$, definim

$$x_{k+1} = x_k - \frac{\langle r_k, d_k \rangle}{\langle Qd_k, d_k \rangle} d_k, \text{ unde } r_k = \nabla q(x_k) = Qx_k - b. \quad (4.10)$$

Astfel, conform Observației 4.2.21, x_{k+1} este minimul lui q pe subspațiul 1-dimensional

$$\{x_k + \alpha d_k \mid \alpha \in \mathbb{R}\}.$$

Are loc rezultatul următor care deschide calea către metoda gradientilor conjugați.

Teorema 4.2.23 Fie $x_0 \in \mathbb{R}^p$ și $(d_i)_{i \in \overline{0, p-1}}$ este o mulțime de direcții Q -conjugate. Fie $(x_i)_{i \in \overline{1, p}}$ punctele generate pe baza metodei direcțiilor conjugate de mai sus. Atunci pentru orice $k \in \overline{1, p}$, x_k este minimum global al funcției q pe subspațiul afin k -dimensional $x_0 + \text{span} \left((d_i)_{i \in \overline{0, k-1}} \right)$. În particular, soluția ecuației $Qx = b$ poate fi găsită în cel mult p pași. Mai mult, pentru orice $k \in \overline{1, p}$ și $i \in \overline{0, k-1}$, $\langle r_k, d_i \rangle_Q = 0$.

Demonstrație. Fie $k \in \overline{1, p}$. Conform Observației 4.2.21,

$$x_k := x_0 + \sum_{i=0}^{k-1} -\frac{\langle r_0, d_i \rangle}{\langle Qd_i, d_i \rangle} d_i,$$

este minimul lui q pe $x_0 + \text{span} \left((d_i)_{i \in \overline{0, k-1}} \right)$. Arătăm că aceste elemente coincid cu cele definite prin relația (4.10). Dacă, pentru $i \in \overline{0, k-1}$, $x \in x_0 + \text{span} \left((d_j)_{j \in \overline{0, i-1}} \right)$, atunci există $(\gamma_j)_{j \in \overline{0, i-1}}$ astfel încât

$$x = x_0 + \sum_{j=0}^{i-1} \gamma_j d_j$$

și atunci

$$\nabla q(x) = r_0 + \sum_{j=0}^{i-1} \gamma_j Qd_j$$

și

$$\langle \nabla q(x), d_i \rangle = \langle r_0, d_i \rangle + \sum_{j=0}^{i-1} \gamma_j \langle Qd_j, d_i \rangle = \langle r_0, d_i \rangle.$$

Alegem $x = x_i \in x_0 + \text{span} \left((d_j)_{j \in \overline{0, i-1}} \right)$ și deducem

$$\langle r_i, d_i \rangle = \langle r_0, d_i \rangle.$$

Astfel deducem că pentru orice $i \in \overline{0, p-1}$

$$\frac{\langle r_0, d_i \rangle}{\langle Qd_i, d_i \rangle} = \frac{\langle r_i, d_i \rangle}{\langle Qd_i, d_i \rangle},$$

de unde obținem prima concluzie.

Acum, a doua concluzie este evidentă.

Pentru ultima concluzie, observăm că pentru orice $k \in \overline{1, p}$, x_k minimizează (fără restricții) funcția $h : \mathbb{R}^k \rightarrow \mathbb{R}$

$$h(\gamma_0, \gamma_1, \dots, \gamma_{k-1}) = q \left(x_0 + \sum_{i=0}^{k-1} \gamma_i d_i \right).$$

ceea ce înseamnă că pentru orice $i \in \overline{0, k-1}$

$$\frac{\partial h}{\partial \gamma_i}(\gamma_0, \gamma_1, \dots, \gamma_{k-1}) = \langle \nabla q(x_k), d_i \rangle = 0.$$

Demonstrația este completă. □

Metoda gradientilor conjugați este o metodă a direcțiilor conjugate în care se construiesc la fiecare pas direcții 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 independență necesară aplicării procedurii menționat. În versiunea sa preliminară, metoda este următoarea.

Se alege $x_0 \in \mathbb{R}^p$ și se definește $d_0 = -\nabla q(x_0) = -r_0$.

Atâta timp cât $r_k \neq 0$ definim

$$\begin{aligned} d_k &= -r_k + \sum_{j=0}^{k-1} \frac{\langle Qr_k, d_j \rangle}{\langle Qd_j, d_j \rangle} d_j, \\ \alpha_k &= \frac{\|r_k\|^2}{\langle Qd_k, d_k \rangle}, \\ x_{k+1} &= x_k + \alpha_k d_k. \end{aligned}$$

Alegerea lui α_k provine din Metoda direcțiilor conjugate și din faptul că, pe baza Teoremei 4.2.23,

$$\langle r_k, d_k \rangle = \left\langle r_k, -r_k + \sum_{j=0}^{k-1} \frac{\langle Qr_k, d_j \rangle}{\langle Qd_j, d_j \rangle} d_j \right\rangle = -\|r_k\|^2, \quad \forall k.$$

Așa cum e descrisă mai sus, metoda este costisitoare din punct de vedere computațional din cauza formulei direcțiilor d_k . Următorul rezultat permite simplificarea semnificativă a acestei formule.

Teorema 4.2.24 *Dacă elementele $(r_i)_{i \in \overline{0, k}}$ nu sunt zero, atunci ele sunt mutual Q -ortogonale și*

$$d_k = -r_k + \beta_k d_{k-1}, \quad \text{unde } \beta_k = \frac{\|r_k\|^2}{\|r_{k-1}\|^2}.$$

Demonstrație. Cum $(d_i)_{i \in \overline{0, k-1}}$ sunt generați prin procedeul Gram-Schmidt din $(-r_i)$ avem

$$\text{span} \left((d_i)_{i \in \overline{0, k-1}} \right) = \text{span} \left((r_i)_{i \in \overline{0, k-1}} \right).$$

Cum $(d_i)_{i \in \overline{0, k-1}}$ sunt Q -conjugate, din Teorema 4.2.23, r_k este Q -ortogonal pe toți vectorii $(d_i)_{i \in \overline{0, k-1}}$, deci și 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

$$Qd_i = Q \left(\frac{x_{i+1} - x_i}{\alpha_i} \right) = \frac{(Qx_{i+1} - b) - (Qx_i - b)}{\alpha_i} = \frac{r_{i+1} - r_i}{\alpha_i}, \quad \forall i \in \overline{0, k-1}.$$

Deci,

$$\langle Qr_k, d_i \rangle = \langle r_k, Qd_i \rangle = \frac{1}{\alpha_i} \langle r_k, r_{i+1} - r_i \rangle, \quad \forall i \in \overline{0, k-1}.$$

Astfel,

$$\begin{aligned}\langle Qr_k, d_i \rangle &= 0, \quad \forall i \in \overline{0, k-2} \\ \langle Qr_k, d_{k-1} \rangle &= \frac{\|r_k\|^2}{\alpha_{k-1}}.\end{aligned}$$

Substituim în expresia lui d_k și avem

$$d_k = -r_k + \frac{\|r_k\|^2}{\alpha_{k-1} \langle Qd_{k-1}, d_{k-1} \rangle} = -r_k + \frac{\|r_k\|^2}{\|r_{k-1}\|^2} d_{k-1},$$

ceea ce încheie demonstrația. □

Pe baza teoremei tocmai demonstrate, putem formula Metoda gradientilor conjugați astfel:

Se alege $x_0 \in \mathbb{R}^p$ și se definește $d_0 = -\nabla q(x_0) = -r_0$.

Atâta timp cât $r_k \neq 0$ definim

$$\begin{aligned}\alpha_k &= \frac{\|r_k\|^2}{\langle Qd_k, d_k \rangle}, \\ x_{k+1} &= x_k + \alpha_k d_k, \\ r_{k+1} &= r_k + \alpha_k Qd_k, \\ \beta_{k+1} &= \frac{\|r_{k+1}\|^2}{\|r_k\|^2}, \\ d_{k+1} &= -r_{k+1} + \beta_{k+1} d_k.\end{aligned}$$

Teoretic, Metoda gradientilor conjugați determină exact soluția în cel mult p pași. Totuși, din cauza rotunjirilor din cadrul calculelor, aceasta este o situație rară în practică. În schimb, este o metodă eficientă pentru sisteme cu dimensiune mare.

Implementare 4.46 Să se scrie o funcție `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 comparație cu metodele de ordin zero sau unu, mai costisitoare din punct de vedere computațional, dar au, în

general, o convergență 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

Exercițiul 4.47 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 inițiale $(1, 1)$, $(-1, -1)$, $(-1, 1)$.

Așa cum am arătat pe scurt în Observația 3.3.5, metoda lui Newton poate fi folosită și la probleme de optimizare, fiind în acest context o metodă cu oracol de ordinul al doilea. Aplicată ecuației $\nabla f(x) = 0$, metoda lui Newton returnează o aproximare a unui punct critic. Ca de obicei, în condiții de convexitate obținem convergența către un punct de minim global. Pentru o funcție $f: \mathbb{R}^p \rightarrow \mathbb{R}$ iterația 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), $\varepsilon_1, \varepsilon_2 > 0$ (toleranțe), $\Delta x > 0$ (pentru calculul gradientului);

Pasul al 2-lea: fiind calculat x_k , se calculează $\nabla f(x_k)$, $H_f(x_k)$, gradientul și Hessiana în x_k ;

$p_k = -H_f^{-1}(x_k) \nabla f(x_k)$ (direcția 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 afișează x_{k+1} și $f(x_{k+1})$.

Prezentăm două modele de aplicare a acestei metode.

Exercițiul 4.48 Să se implementeze această metodă pentru funcțiile test deja utilizate.

Implementare 4.49 Putem folosi de asemenea un program care să calculeze (cu aproximație) hessiana, așa încât să nu fim nevoiți să calculăm prin derivare directă această matrice.

```
function sec_deriv = hessian(x,dx)
for i = 1:length(x)
    for j = 1:length(x)
        if i == j
            temp = x;
            temp(i) = x(i) + dx;
            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

```

Exercițiul 4.50 Folosind această funcție, să se implementeze din nou metoda lui Newton (a se vedea funcț `func_multivar` și `grad_vec` de mai sus).

4.3.2 Metode derivate din metodele fundamentale

Metoda căutării direcției de descreștere, metoda gradientilor conjugați și metoda lui Newton sunt metode fundamentale în optimizarea fără restricții. Pe baza lor au fost derivate și alte metode care fie sunt hibride, în sensul că prin construcția lor combină caracteristici ale celor trei metode de bază, fie sunt aproximative, în sensul că oferă posibilități de evitare a calculului elementelor celor mai costisitoare din punct de vedere computațional ale acestora (ca de exemplu, calculul Hessiane în cazul metodei lui Newton).

Prezentăm mai jos, fără demonstrații privind convergența, principalele metode derivate. În orice caz, fiind vorba de variații ale metodelor de bază, considerăm că, prin însăși natura lor, acești algoritmi pot fi înțeleși, acceptați și folosiți cu ușurință.

1. Metoda lui Newton modificată

Pasul 1: se dau x_0 (punctul de start), $\varepsilon_1, \varepsilon_2 > 0$ (toleranțe), $\Delta x > 0$ (pentru calculul gradientului);

Pasul al 2-lea: fiind calculat x_k , se calculează $\nabla f(x_k), H_f(x_k)$, gradientul și Hessiana în x_k ;

$p_k = -H_f^{-1}(x_k) \nabla f(x_k)$ (direcția de căutare);

se minimizează $\alpha \rightarrow f(x_k + \alpha p_k)$ (utilizând, de exemplu, metoda secțiunii de aur) și se obține 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 afișează x_{k+1} și $f(x_{k+1})$.

În unele situații această metodă este mai rapidă decât metoda lui Newton. Spre comparație, această metodă mai caută un pas α suplimentar față de metoda lui Newton, deci combină metoda clasică cu metoda căutării direcției de descreștere.

Exercițiul 4.51 *Să se exemplifice această metodă folosindu-se funcțiile Matlab/Octave deja prezentate.*

2. Metoda Levenberg–Marquardt

Pasul 1: se dau x_0 (punctul de start), $\varepsilon_1, \varepsilon_2 > 0$ (toleranțe), $\Delta x > 0$ (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 și Hessiana în x_k ;

$p_k = -[H_f(x_k) + \lambda I]^{-1} \nabla f(x_k)$ (direcția 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 afișează x_{k+1} și $f(x_{k+1})$.

Se poate observa că și această metodă este una hibrid între metoda lui Newton și metoda căutării direcției de descreștere.

Exercițiul 4.52 *Să se exemplifice această metodă folosindu-se funcțiile Matlab/Octave deja prezentate.*

3. Metoda Fletcher–Reeves a gradientilor conjugați

Pasul 1: se dau x_0 (punctul de start), $\varepsilon_1, \varepsilon_2 > 0$ (toleranțe), $\Delta x > 0$ (pentru calculul gradientului);

se minimizează $\alpha \rightarrow f(x_0 - \alpha \nabla f(x_0))$ (utilizând, de exemplu, metoda secțiunii de aur) și se obține 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 $\alpha \rightarrow f(x_k + \alpha p_k)$ (utilizând, de exemplu, metoda secțiunii de aur) și se obține 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 afișează x_{k+1} și $f(x_{k+1})$.

Această metodă este o extindere la funcții nu neapărat pătratice a metodei gradientilor conjugăți. În particular, este o metodă de ordinul întâi.

Exercițiul 4.53 *Să se exemplifice această metodă folosindu-se funcțiile Matlab/Octave deja prezentate.*

4. Metoda Davidon–Fletcher–Powell (Metoda DFP)

Ideea acestei metode este de a evita calculul Hessienei în metoda lui Newton, folosind în schimb o aproximare a inversei acesteia. Este o metodă de ordinul întâi cu convergență pătratică, mai fiind numită și metoda quasi-Newton.

Aproximarea inversei Hessienei în punctul curent se face printr-un șir de matrici (A_k) , cu primul termen identitatea care respectă recurența

$$A_{k+1} = A_k + \frac{(x_{k+1} - x_k)^t (x_{k+1} - x_k)}{(x_{k+1} - x_k) (\nabla f(x_{k+1}) - \nabla f(x_k))^t} - \frac{A_k (\nabla f(x_{k+1}) - \nabla f(x_k))^t (\nabla f(x_{k+1}) - \nabla f(x_k)) A_k}{(\nabla f(x_{k+1}) - \nabla f(x_k)) A_k (\nabla f(x_{k+1}) - \nabla f(x_k))^t}.$$

Metoda poate fi descrisă astfel:

Pasul 1: se dau x_0 (punctul de start), $\varepsilon_1, \varepsilon_2 > 0$ (toleranțe), $\Delta x > 0$ (pentru calculul gradientului), matricea A inițializată cu identitatea;

se minimizează $\alpha \rightarrow f(x_0 - \alpha \nabla f(x_0))$ (utilizând, de exemplu, metoda secțiunii de aur) și se obține 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 și 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 $\alpha \rightarrow f(x_k + \alpha p_k)$ (utilizând, de exemplu, metoda secțiunii de aur) și se obține 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 afișează x_{k+1} și $f(x_{k+1})$.

Exercițiul 4.54 Să se exemplifice această metodă folosindu-se funcțiile Matlab/Octave deja prezentate.

5. Metoda Broyden–Fletcher–Goldfarb–Shanno (Metoda BFGS)

Din nou, ideea este de a evita calculul Hessienei 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 șirul de matrici (A_k) cu primul termen identitatea care respectă recurența

$$A_{k+1} = A_k + \frac{(\nabla f(x_{k+1}) - \nabla f(x_k))^t (\nabla f(x_{k+1}) - \nabla f(x_k))}{(\nabla f(x_{k+1}) - \nabla f(x_k))(x_{k+1} - x_k)^t} - \frac{\nabla f(x_k)^t \nabla f(x_k)}{\nabla f(x_k) p_k^t},$$

unde p_k este definit în descrierea de mai jos a metodei:

Pasul 1: se dau x_0 (punctul de start), $\varepsilon_1, \varepsilon_2 > 0$ (toleranțe), $\Delta x > 0$ (pentru calculul gradientului), matricea A inițializată cu identitatea;

se minimizează $\alpha \rightarrow f(x_0 - \alpha \nabla f(x_0))$ (utilizând, de exemplu, metoda secțiunii de aur) și se obține 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 și 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 $\alpha \rightarrow f(x_k + \alpha p_k)$ (utilizând, de exemplu, metoda secțiunii de aur) și se obține 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 afișează x_{k+1} și $f(x_{k+1})$.

Exercițiul 4.55 Să se exemplifice această metodă folosindu-se funcțiile Matlab/Octave deja prezentate.

Exercițiul 4.56 Să se testeze și să se compare metodele pe funcția Rosenbrock.

Exercițiul 4.57 Fie $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ dată prin

$$f(x, y) = e^{x^2+y^2} + x + y - 3 - \sin(3(x+y)).$$

Să se arate deseneze graficul și liniile de contur pe $[-1, 1]^2$ și 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.

Exercițiul 4.58 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 soluție a acestui sistem folosind metodele anterioare pentru funcția $f : \mathbb{R}^3 \rightarrow \mathbb{R}$,

$$f(x, y, z) = \left(3x - \cos yz - \frac{3}{2}\right)^2 + (3x^2 - 625y^2 + 2y - 1)^2 + (e^{-xy} + 20z + 9.1)^2.$$

Exercițiul 4.59 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 soluții a acestui sistem folosind metodele anterioare pentru funcția construită ca la exercițiul anterior pentru datele inițiale $(1.5, -1.75)$, $(1.5, 1.5)$.

Exercițiul 4.60 Să se compare metodele studiate pe cazul minimizării funcției lui Himmelblau, $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ date prin

$$f(x) = (x_1^2 + x_2 - 11)^2 + (x_2^2 + x_1 - 7)^2$$

luând diverse date inițiale (se va reprezenta grafic și se va constata că funcția are patru puncte de minim local).

Exercițiul 4.61 Să se verifice pe cazul funcției $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ date prin

$$f(x) = x_1^2 + 3x_1x_2 + 5x_2^2$$

luând data inițială $x = (1, 1)$ că

(i) iterația pentru generarea matricilor $(A_k)_k$ din metoda DFP conduce la o aproximare a inversei hessianei funcției;

(ii) iterația pentru generarea matricilor $(A_k)_k$ din metoda BFGS conduce la o aproximare a hessianei funcției.

Exercițiul 4.62 Fie $f : \mathbb{R}^4 \rightarrow \mathbb{R}$,

$$f(x) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^2 + 100(x_1 - x_4)^4.$$

Verificați că $0 \in \mathbb{R}^4$ este singurul punct de minim global și că nu există alte puncte de minim local. Testați pe funcția f

(i) metoda celei mai bune descreșteri cu datele inițiale $(-2, -1, 1, 2)$, $(200, -200, 100, -100)$;

(ii) metodele derivate din metoda lui Newton.

Exercițiul 4.63 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 și 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 și este modelată de funcția $f : \mathbb{R}^3 \rightarrow \mathbb{R}$,

$$f(x) = x_3 t^{x_1} \frac{1}{(1-t)^{x_2}}.$$

Funcția de minimizat pentru a găsi modelul cel mai apropiat de aceste date este, conform metodei celor mai mici pătrate (a se vedea și Problema 2.22),

$$x \mapsto \sum_{i=1}^6 \left[m_i - x_3 t_i^{x_1} \frac{1}{(1-t_i)^{x_2}} \right]^2.$$

Determinați parametrii modelului și valoarea totală a pătratelor abaterilor minimizând funcția de mai sus prin diverse metode.

Exercițiul 4.64 Să se testeze metodele studiate pe următoarele funcții test:

(i) funcția Beale, $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ dată prin

$$f(x) = (1.5 - x_1 + x_1 x_2)^2 + (2.25 - x_1 + x_1 x_2^2)^2 + (2.625 - x_1 + x_1 x_2^3)^2,$$

pentru care punctul de minim global este $(3, 0.5)$, iar valoarea minimă globală este 0;

(ii) funcția „câmilă cu trei cocoșe”, $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ dată prin

$$f(x) = 2x_1^2 - 1.05x_1^4 + \frac{x_1^6}{6} + x_1 x_2 + x_2^2,$$

pentru care punctul de minim global este $(0, 0)$, iar valoarea minimă globală este 0;

(iii) funcția Branin, $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ 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 funcțiilor cu valori vectoriale: optimizare Pareto

Optimizarea Pareto (după numele economistului italian W. F. Pareto) se ocupă de situația în care minimizarea (sau maximizarea) se face după mai multe criterii, unele fiind în concurență cu altele. Formalizarea matematică a acestei idei este prezentată succint mai jos.

Ca de obicei, considerăm $p \in \mathbb{N} \setminus \{0\}$.

Definiția 4.4.1 O submulțime nevidă $C \subset \mathbb{R}^p$ se numește con dacă are loc proprietatea

$$\forall x \in C, \forall \lambda \in \mathbb{R}_+ := [0, \infty) : \lambda x \in C.$$

Conform acestei definiții, orice con conține elementul 0. Dacă $A \subset \mathbb{R}^p$ este o mulțime nevidă, atunci mulțimea

$$\mathbb{R}_+ A := \{\lambda a \mid \lambda \in \mathbb{R}_+, a \in C\}$$

este con, numit conul generat de A .

Propoziția 4.4.2 *Un con C este mulțime convexă dacă și numai dacă $C + C \subset C$.*

Prin urmare, un con convex este o mulțime nevidă închisă la adunare și la înmulțirea cu scalari pozitivi.

Definiția 4.4.3 (i) *Un con C se numește propriu dacă $C \neq \{0\}$ și $C \neq \mathbb{R}^p$.*

(ii) *Un con C se numește punctat dacă $C \cap (-C) = \{0\}$.*

Următorul rezultat arată că un con introduce pe $\mathbb{R}^p$ o relație de ordine (parțială) compatibilă cu structura de spațiu liniar.

Teorema 4.4.4 *Fie $\mathbb{R}^p$ un spațiu liniar și $C \subset \mathbb{R}^p$ un con. Atunci relația*

$$\mathcal{R}_C := \{(x_1, x_2) \in \mathbb{R}^p \times \mathbb{R}^p \mid x_2 - x_1 \in C\}$$

este reflexivă și

$$\forall x_1, x_2 \in \mathbb{R}^p, \forall \lambda \in \mathbb{R}_+, x_1 \mathcal{R}_C x_2 \Rightarrow \lambda x_1 \mathcal{R}_C \lambda x_2 \quad (4.11)$$

$$\forall x_1, x_2, x \in \mathbb{R}^p, x_1 \mathcal{R}_C x_2 \Rightarrow (x_1 + x) \mathcal{R}_C (x_2 + x). \quad (4.12)$$

Mai mult, C este convex dacă și numai dacă $\mathcal{R}_C$ este tranzitivă și, respectiv, C este punctat dacă și numai dacă $\mathcal{R}_C$ este antisimetrică.

Invers, dacă $\mathcal{R}$ este o relație reflexivă pe $\mathbb{R}^p$ care satisface (4.11) și (4.12), atunci $C := \{x \in \mathbb{R}^p \mid 0 \mathcal{R} x\}$ este con și $\mathcal{R} = \mathcal{R}_C$.

Demonstrație. Toate afirmațiile se verifică direct, pe baza definițiilor. □

Observația 4.4.5 *Pentru $p = 1$ și $C = \mathbb{R}_+$, $\mathcal{R}_C$ 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 $x_1 \mathcal{R}_C x_2$.

Exercițiul 4.65 *Fie $C \subset \mathbb{R}^p$ un con convex închis punctat cu interior nevid. Să se arate că:*

- (i) $C + \text{int } C = \text{int } C$;
- (ii) $\mathbb{R}^p \setminus -\text{int } C + C = \mathbb{R}^p \setminus -\text{int } C$ și $(\mathbb{R}^p \setminus -C) \cup \{0\} + C = (\mathbb{R}^p \setminus -C) \cup \{0\}$;
- (iii) dacă $x \in \text{int } C$ și $x^* \in C^+ \setminus \{0\}$, atunci $x^*(x) > 0$.

Exercițiul 4.66 *Fie $C \subset \mathbb{R}^p$ un con închis convex cu interior nevid și $\emptyset \neq A \subset \mathbb{R}^p$. Să se arate că următoarele afirmații sunt echivalente:*

- (i) $A \cap -\text{int } C = \emptyset$;
- (ii) $\text{cl } A \cap -\text{int } C = \emptyset$;
- (iii) $(A + C) \cap -\text{int } C = \emptyset$;
- (iv) $\text{cl}(\text{con}(A + C)) \cap -\text{int } C = \emptyset$.

Exercițiul 4.67 Fie $C \subset \mathbb{R}^p$ un con închis convex cu interior nevid care nu coincide cu $\mathbb{R}^p$ și $e \in \text{int } C$. Să se arate că:

- (i) $C + [0, \infty)e \subset C$;
- (ii) $C + (0, \infty)e = \text{int } C$;
- (iii) $\mathbb{R}e - C = \mathbb{R}^p$;
- (iv) pentru orice $x \in \mathbb{R}^p$, $x + \mathbb{R}e \not\subset C$.

Definiția 4.4.6 Considerăm $\mathbb{R}^p$ ca fiind ordonat cu ordinea parțială indusă de un con convex, punctat, propriu C și $A \subset \mathbb{R}^p$ o mulțime nevidă.

(i) Un punct $\bar{x} \in A$ se numește minim tare al mulțimii A (în raport cu conul C) dacă $\bar{x} \leq_C x$ pentru orice $x \in A$, sau, echivalent, $A \subset \bar{x} + C$.

(ii) Un punct $\bar{x} \in A$ se numește minim Pareto al mulțimii A (în raport cu conul C) dacă $x \in A$, $x \leq_C \bar{x}$ implică $x = \bar{x}$, sau, echivalent, $A \cap (\bar{x} - C) = \{\bar{x}\}$.

(iii) Presupunem că C are interior nevid. Un punct $\bar{x} \in A$ se numește minim slab al mulțimii A (în raport cu conul C) dacă $A \cap (\bar{x} - \text{int } C) = \emptyset$.

Observația 4.4.7 Este evident că definiția minimului Pareto este echivalentă cu oricare dintre relațiile

$$(A - \bar{x}) \cap (-C) = \{0\},$$

$$A \cap (\bar{x} - C) = \{\bar{x}\}$$

și

$$(A + C) \cap (\bar{x} - C) = \{\bar{x}\}.$$

Având în vedere incluziunea $C + \text{int } C \subset \text{int } C$, definiția minimului slab este echivalentă cu

$$(A - \bar{x}) \cap (-\text{int } C) = \emptyset,$$

$$A \cap (\bar{x} - \text{int } C) = \emptyset,$$

$$(A + C) \cap (\bar{x} - \text{int } C) = \emptyset$$

și cu

$$\bar{x} \notin A + \text{int } C.$$

Mulțimea punctelor de minim tare (minim Pareto, minim slab) ale mulțimii A va fi notată cu $IMin(A, C)$ (respectiv, $Min(A, C)$, $WMin(A, C)$). Evident, dacă $\text{int } C \neq \emptyset$, $IMin(A, C) \subset Min(A, C) \subset WMin(A, C)$.

Observația 4.4.8 Pentru $p = 1$ și $C = \mathbb{R}_+$, toate noțiunile de mai sus coincid cu minimalitatea scalară standard.

Exercițiul 4.68 Să se illustreze minimalitatea Pareto, căutând mulțimea punctelor optime Pareto în raport cu $\mathbb{R}_+^2$ pentru un grup de puncte generate aleator în primul cadran.

Exercițiul 4.69 Să se genereze 50 puncte echidistante pe un cerc și să se identifice dintre aceste punctele optime în sens Pareto (din nou, în raport cu $\mathbb{R}_+^2$).

Definiția 4.4.9 Fie $f : \mathbb{R}^n \rightarrow \mathbb{R}^p$ unde $n, p \in \mathbb{N} \setminus \{0\}$, $C \subset \mathbb{R}^p$ un con convex închis punctat propriu și $A \subset \mathbb{R}^n$ o mulțime închisă. Spunem că $\bar{x} \in A$ este minim (slab) Pareto pentru f pe A dacă $f(\bar{x})$ este minim (slab) Pareto al mulțimii $f(A)$.

Următorul rezultat prezintă o situație în care găsirea unui minim Pareto pentru o funcție vectorială se reduce la găsirea unui minim pentru o funcție scalară.

Propoziția 4.4.10 Cu notațiile din Definiția 4.4.9, considerăm $C = \mathbb{R}_+^p$, $w \in \text{int } \mathbb{R}_+^p$ cu $\sum_{i=1}^p w_i = 1$. Fie $g : \mathbb{R}^n \rightarrow \mathbb{R}$,

$$g(x) = \sum_{i=1}^p w_i f_i(x).$$

Dacă $\bar{x} \in A$ este punct de minim (global) pe A pentru g , atunci $\bar{x}$ este minim Pareto pentru f pe A .

Demonstrație. Presupunem, prin reducere la absurd, că $\bar{x}$ nu este minim Pareto pentru f pe A . Atunci există $x \in A$ astfel încât pentru orice $i \in \overline{1, n}$, $f_i(x) \leq f_i(\bar{x})$ și există $j \in \overline{1, n}$ astfel încât $f_j(x) < f_j(\bar{x})$. Înmulțind aceste inegalități cu scalarii $w_i > 0$ corespunzători și adunându-le, obținem $g(x) < g(\bar{x})$, ceea ce reprezintă o contradicție. $\square$

Exercițiul 4.70 Fie $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$,

$$f(x) = (x_1^2 + x_2^2, (x_1 - 1)^2 + (x_2 - 3)^2).$$

Să se scrie un program care să determine puncte de minim Pareto pentru f (pe $\mathbb{R}^2$) prin metoda descrisă în Propoziția 4.4.10 (de exemplu, prin modificarea corespunzătoare a metodei DFP).

4.5 Obținerea unor algoritmi pentru probleme cu restricții - o metodă de penalizare

În cazul problemelor cu restricții, cei mai importanți algoritmi derivă din metodele studiate pentru probleme fără restricții 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 $f, g_1, g_2 : \mathbb{R}^2 \rightarrow \mathbb{R}$ date respectiv prin

$$f(x) = (x_1 - 1)^2 + (x_2 - 5)^2$$

$$g_1(x) = -x_1^2 + x_2 - 4$$

$$g_2(x) = -(x_1 - 2)^2 + x_2 - 3.$$

Studiem problema

$$(P_1) \min f(x) \text{ cu } g_1(x) \leq 0, g_2(x) \leq 0.$$

Este evident că problema admite soluție. O analiză geometrică, combinată cu una analitică, arată că soluția 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 $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^2 - 8x + 3$ cu restricția $g(x) \leq 0$, unde $g : \mathbb{R} \rightarrow \mathbb{R}$, $g(x) = x - 1$. Evident, soluția este $x = 1$.

Evident, există și algoritmi direcți cum ar fi algoritmul grilei uniforme care pot fi adaptați foarte ușor.

Exercițiul 4.71 *Să se implementeze un algoritm de tipul grilei uniforme adaptat pentru problema (P_1) .*

Vom constata că metodele de aproximare a soluțiilor problemelor cu restricții se reduc, prin diverse procedee, la minimizarea unor familii de funcții fără restricții pentru care se pot aplica algoritmii studiați anterior.

O metodă mai elaborată pe care o descriem este bazată pe o tehnică numită penalizare. Ideea este de a îngloba restricțiile problemei într-o funcție, numită termen de penalizare, astfel încât rezolvarea problemei fără restricții care constă din minimizarea funcției obiectiv adunată cu acest termen să conducă la aproximarea soluției problemei inițiale. De regulă, trebuie considerată o familie parametrică de termeni de penalizare pentru a obține o aproximare bună a soluției pe care o căutăm. Prezentăm în continuare aspectele matematice cele mai relevante ale acestei abordări.

Considerăm problema generală cu restricții

$$\min f(x) \text{ cu } x \in M,$$

unde $f : \mathbb{R}^p \rightarrow \mathbb{R}$ este o funcție netedă și $M \subset \mathbb{R}^p$ este o mulțime închisă. Dorim să aproximăm soluțiile acestei probleme aceasta prin intermediul soluțiilor problemelor fără restricții de forma

$$\min (f(x) + \gamma P(x)),$$

unde $\gamma > 0$ și $P : \mathbb{R}^p \rightarrow \mathbb{R}$ este o funcție. Constanta γ se numește parametru de penalizare, iar funcția P se numește funcție de penalizare, adică o funcție care satisface definiția de mai jos.

Definiția 4.5.1 *O funcție $P : \mathbb{R}^p \rightarrow \mathbb{R}$ se numește funcție de penalizare pentru problema generală cu restricții dacă este continuă, cu valori nenegative și*

$$P(x) = 0 \iff x \in M.$$

Ideea este că o astfel de funcție "penalizează" punctele din afara mulțimii M astfel încât acestea nu pot fi soluții pentru problema fără restricții asociată mai sus.

Exemplul 4.5.2 Fie

$$M = \{x \in \mathbb{R}^p \mid g(x) \leq 0\}$$

unde $g : \mathbb{R}^p \rightarrow \mathbb{R}^n$ este o funcție diferențiabilă, iar inegalitatea este înțeleasă pe coordonate. Atunci funcția $P : \mathbb{R}^p \rightarrow \mathbb{R}$ dată prin

$$P(x) = \sum_{i=1}^n \max \{0, g_i(x)\}$$

este o funcție de penalizare. Evident, această funcție este una nediferențiabilă, în general.

O altă posibilitate este să alegem $P : \mathbb{R}^p \rightarrow \mathbb{R}$ dată prin

$$P(x) = \sum_{i=1}^n (\max\{0, g_i(x)\})^2,$$

iar aceasta este o funcție de penalizare diferențiabilă.

Exemplul 4.5.3 Fie

$$M = \{x \in \mathbb{R}^p \mid h(x) = 0\}$$

unde $h : \mathbb{R}^p \rightarrow \mathbb{R}^m$ este o funcție diferențiabilă. Atunci funcția $P : \mathbb{R}^p \rightarrow \mathbb{R}$ dată prin

$$P(x) = \sum_{i=1}^n h_i^2(x)$$

este o funcție de penalizare.

Exercițiul 4.72 Să se scrie prima variantă a funcției de penalizare din exemplul anterior pentru $g : \mathbb{R} \rightarrow \mathbb{R}^2$,

$$g(x) = (x - 2, -(x + 1)^3).$$

Studiați diferențiabilitatea și graficul funcției de penalizare.

Exemplul 4.5.4 Fie Q o matrice pătratică simetrică de dimensiune p . Considerăm aplicația $f : \mathbb{R}^p \rightarrow \mathbb{R}$,

$$f(x) = \langle Qx, x \rangle$$

și problema minimizării și maximizării acesteia pe sfera unitate din $\mathbb{R}^p$, adică pe mulțimea

$$\{x \in \mathbb{R}^p \mid \|x\| = 1\}.$$

Evident, în ambele cazuri există soluție, sfera fiind compactă.

Să remarcăm că într-un punct x de pe sferă conul tangent Bouligand la acesta este mulțimea

$$\{u \in \mathbb{R}^p \mid \langle x, u \rangle = 0\}$$

(Teorema lui Lyusternik), deci conul normal corespunzător este

$$\{\alpha x \mid x \in \mathbb{R}\}.$$

Atunci, conform teoremei care dă condiții necesare de optimalitate de ordinul I, un punct $\bar{x}$ de pe sferă este soluție a problemei de minimizare sau de maximizare dacă există $\mu \in \mathbb{R}$ cu

$$Q\bar{x} = \mu\bar{x},$$

adică

$$\mu = \langle Q\bar{x}, \bar{x} \rangle.$$

(Aceeași concluzie rezultă prin aplicarea Teoremei Karush-Kuhn-Tucker.) Deducem de aici că punctele ce satisfac condițiile de optimalitate pentru problemele noastre corespund vectorilor proprii de normă 1 ai lui Q , iar valorile funcției 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 obține

$$\lambda_{\max} = \max_{x \in \mathbb{R}^p \setminus \{0\}} \frac{\langle Qx, x \rangle}{\|x\|^2} \text{ și } \lambda_{\min} = \min_{x \in \mathbb{R}^p \setminus \{0\}} \frac{\langle Qx, x \rangle}{\|x\|^2}.$$

Să considerăm acum pentru problema de mai sus funcția de penalizare $P(x) = (\|x\|^2 - 1)^2$ și un parametru $\gamma > 0$. Conform ideii generale, problema fără restricții asociată este

$$\min \left(\langle Qx, x \rangle + \gamma (\|x\|^2 - 1)^2 \right).$$

O soluție $\bar{x}_\gamma$ a acestei probleme trebuie să satisfacă, conform Teoremei lui Fermat, ecuația

$$2Q\bar{x}_\gamma + 4\gamma (\|\bar{x}_\gamma\|^2 - 1) \bar{x}_\gamma = 0,$$

adică

$$Q\bar{x}_\gamma = 2\gamma (1 - \|\bar{x}_\gamma\|^2) \bar{x}_\gamma,$$

ceea ce înseamnă în particular că $\bar{x}_\gamma$ este vector propriu al lui Q .

În plus, valoarea proprie asociată este $2\gamma (1 - \|\bar{x}_\gamma\|^2)$, deci

$$\lambda_{\min} \leq 2\gamma (1 - \|\bar{x}_\gamma\|^2) \leq \lambda_{\max}.$$

Astfel,

$$\frac{\lambda_{\min}}{2\gamma} \leq 1 - \|\bar{x}_\gamma\|^2 \leq \frac{\lambda_{\max}}{2\gamma}$$

și pentru $\gamma \rightarrow \infty$, norma lui $\bar{x}_\gamma$ tinde la 1, deci $\bar{x}_\gamma$ tinde (măcar pe un subsir) la un punct fezabil al problemei inițiale.

Lema 4.5.5 *Fie P o funcție de penalizare pentru problema minimizării lui $f(x)$ cu $x \in M$. Fie $\bar{x}$ o soluție globală a acestei probleme și (γ_k) un șir strict crescător de numere reale pozitive. Pentru fiecare $k \in \mathbb{N}$ considerăm funcția $q_k : \mathbb{R}^p \rightarrow \mathbb{R}$ dată prin*

$$q_k(x) = f(x) + \gamma_k P(x).$$

Presupunem că fiecare dintre aceste funcții 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) $f(\bar{x}) \geq q_k(x_k) \geq f(x_k)$.

Demonstrație. (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 obținem prima concluzie.

(ii) Avem, ca mai sus,

$$q_k(x_k) = f(x_k) + \gamma_k P(x_k) \leq f(x_{k+1}) + \gamma_k P(x_{k+1})$$

și

$$q_{k+1}(x_{k+1}) = f(x_{k+1}) + \gamma_{k+1}P(x_{k+1}) \leq f(x_k) + \gamma_{k+1}P(x_k),$$

inegalități 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 $\gamma_{k+1} > \gamma_k$ obținem 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

$$f(\bar{x}) + \gamma_k P(\bar{x}) \geq q_k(x_k) = f(x_k) + \gamma_k P(x_k).$$

Dar $\bar{x}$ este punct fezabil pentru problema inițială, deci $P(\bar{x}) = 0$. Astfel,

$$f(\bar{x}) \geq q_k(x_k) = f(x_k) + \gamma_k P(x_k) \geq f(x_k).$$

Demonstrația este completă. □

Rezultatul următor este esențial pentru fundamentarea metodei de penalizare.

Teorema 4.5.6 *Cu notațiile de mai sus, în ipotezele Lemei 4.5.5, dacă $\gamma_k \rightarrow \infty$, atunci orice punct limită al șirului (x_k) este soluție a problemei cu restricții.*

Demonstrație. Fie (x_{k_l}) un subșir al lui (x_k) convergent la un punct x^* . Conform Lemei 4.5.5, șirul $(q_k(x_k))$ este crescător și mărginit superior de $f(\bar{x})$. Deci există limita reală $\bar{q}$ al lui $(q_k(x_k))$ și $\bar{q} \leq f(\bar{x})$. Cum f este continuă, avem

$$\lim_l f(x_{k_l}) = f(x^*) \leq f(\bar{x}).$$

Dar şirul cu termenii

$$\gamma_{k_l} P(x_{k_l}) = q_{k_l}(x_{k_l}) - f(x_{k_l})$$

este convergent la $\bar{q} - f(x^*)$. Tot din lema precedentă, şirul $(P(x_k))$ este descrescător şi mărginit inferior de 0, deci este convergent. Cum $\gamma_{k_l} \rightarrow \infty$, deducem că $P(x_{k_l}) \rightarrow 0$, 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 $f(x^*) \leq f(\bar{x})$, deducem că x^* este la rândul său o soluţie globală a acestei probleme. $\square$

Observaţia 4.5.7 Conform rezultatului precedent, pentru convergenţă trebuie să considerăm un şir de parametri $\gamma_k \rightarrow \infty$.

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 $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = 5 - 3x$ pentru $x \in [0, 1]$. Evident, $\bar{x} = 1$ 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 $\bar{x}$ 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

$$\min f(x) \text{ cu } g(x) \leq 0, \quad h(x) = 0,$$

unde $f : \mathbb{R}^p \rightarrow \mathbb{R}$, $g : \mathbb{R}^p \rightarrow \mathbb{R}^n$, $h : \mathbb{R}^p \rightarrow \mathbb{R}^m$, cu notaţiile standard.

Astfel, în notaţiile de mai sus, considerăm funcţia de penalizare

$$P(x) = \sum_{i=1}^n (\max\{0, g_i(x)\})^2 + \sum_{j=1}^m h_j^2(x)$$

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 4.73 Mai întâi, evidențiem pe baza unui desen comportamentul funcției de penalizare în cazul problemei (P_2) la varierea (creșterea) 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]);
```

Exercițiul 4.74 Să se scrie fișierele care determină soluția problemei (P_2) folosind algoritmul secțiunii de aur pentru funcția penalizată (prin creșterea progresivă a parametrului γ).

Exercițiul 4.75 Să se scrie fișierele care adaptează metoda DFP pentru determinarea soluției problemei (P_1).

Observația 4.5.9 O metodă foarte simplă, dar eficientă în unele situații, bazată de asemenea pe ideea penalizării, este prezentată în continuare. Considerăm problema generală cu restricții funcționale și date netede (în notația standard):

$$\min f(x) \text{ cu } g(x) \leq 0, \quad h(x) = 0,$$

unde $f: \mathbb{R}^p \rightarrow \mathbb{R}$, $g: \mathbb{R}^p \rightarrow \mathbb{R}^n$, $h: \mathbb{R}^p \rightarrow \mathbb{R}^m$. Pentru metoda pe care o vom prezenta, o inegalitate de tipul $h_j(x) = 0$ se scrie echivalent în forma ($h_j(x) \leq 0$ și $-h_j(x) \leq 0$), de aceea ne vom ocupa de problema cu inegalități:

$$\min f(x) \text{ cu } g(x) \leq 0,$$

unde $g = (g_1, \dots, g_n)$. Algoritmul propus este următorul:

Pasul 1. Se alege $y_0 = (1, \dots, 1) \in \mathbb{R}^n$ (un parametru) și $x_{-1} = 0 \in \mathbb{R}^p$ (o dată inițială), $\varepsilon > 0$ (o toleranță)

Pasul 2. Date fiind $y_k = (y_k^i)_{i \in \overline{1, n}} \in \mathbb{R}^n$ și $x_{k-1} \in \mathbb{R}^p$ minimizăm funcția (fără restricții)

$$f(x) + \sum_{i=1}^n e^{y_k^i g_i(x)}$$

cu data inițială x_{k-1} . Fie x_k punct de minim pentru această funcție.

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 convergență a acestui algoritm este următorul: dacă șirurile obținute pe baza pașilor anteriori satisfac

$$y_k^i g_i(x_k) \rightarrow 0, \forall i \in \overline{1, n} \text{ și } x_k \rightarrow \bar{x},$$

atunci $\bar{x}$ este soluție a problemei cu restricții.

Pentru a demonstra, este suficient să avem în vedere cazul $n = 1$. Astfel, considerăm $g : \mathbb{R}^p \rightarrow \mathbb{R}$.

Mai întâi arătăm că $\bar{x}$ este punct fezabil. Presupunând contrariul, $g(\bar{x}) > 0$, deci $g(x_k) > 0$ pentru k suficient de mare. Obținem:

$$y_{k+1} = e^{y_k g(x_k)} y_k \geq y_k.$$

Deci (y_k) este un șir crescător și cum $y_0 > 0$, acesta nu poate avea limita 0. Această observație este în contradicție cu ipoteza $y_k g(x_k) \rightarrow 0$. Deducem că $g(\bar{x}) \leq 0$, deci $\bar{x}$ este fezabil.

Presupunem, tot prin reducere la absurd, că pentru orice $\rho > 0$ există $x_\rho \in B(\bar{x}, \rho) \cap M$ astfel încât $f(x_\rho) < f(\bar{x})$. Atunci pentru orice k .

$$f(x_\rho) + e^{y_k g(x_\rho)} \leq f(x_\rho) + 1 < f(\bar{x}) + 1.$$

Cum $y_k g(x_k) \rightarrow 0$ și $x_k \rightarrow \bar{x}$, 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 .

Așadar, convergența menționată are loc.

Modele de examene scrise

Model examen parțial

Subiectul 1. Fie mulțimea $M = \{x \in \mathbb{R}^2 \mid x + y \leq 1\}$. Să se scrie expresia conului tangent la M în $(2, -1)$. Fie $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ o funcție de clasă C^1 . Să se scrie condiția necesară de optimalitate de ordinul I astfel încât $(2, -1)$ să fie soluție locală a problemei de optimizare $\min f(x, y)$ cu $(x, y) \in M$.

Subiectul 2. Se dorește construirea unei cutii paralelipipedice cu baza pătrat și fără capac cu volumul de 100. Determinați dimensiunile unei astfel de cutii astfel încât materialul folosit să fie minim.

Subiectul 3. Fie $f, g : \mathbb{R}^2 \rightarrow \mathbb{R}$ date prin $f(x, y) = (x + 1)^2 + y^2$ și $g(x, y) = -x^3 + y^2$. Considerăm problema minimizării lui $f(x, y)$ cu restricția $g(x, y) \leq 0$.

- (i) Să se verifice dacă are loc condiția Mangasarian-Fromovitz în $(0, 0)$.
- (ii) Să se determine punctele care satisfac condițiile de optimalitate Fritz John.
- (iii) Să se determine soluțiile problemei.

Subiectul 4. Fie $f, g, h : \mathbb{R}^2 \rightarrow \mathbb{R}$ date prin $f(x, y) = (x - 1)^2 + y - 2$, $g(x, y) = x + y - 2$ și $h(x, y) = y - x - 1$. Considerăm problema minimizării lui $f(x, y)$ cu restricțiile $g(x, y) \leq 0$, $h(x, y) = 0$.

- (i) Să se găsească punctele care satisfac condițiile de optimalitate Karush-Kuhn-Tucker.
- (ii) Să se verifice dacă condițiile suficiente de ordinul al doilea au loc în aceste puncte și să se concluzioneze.

Barem de notare:

- 1. 1p – baza; conul tangent 4p; condiția necesară de optimalitate 5p.
- 2. 1p – baza; modelarea 2p; rezolvarea problemei 5p; concluzia 2p.
- 3. 1p – baza; condiția MF: 2p; punctele Fritz John 5; soluțiile 2p.
- 4. 1p – baza; punctele Karush-Kuhn-Tucker 4p; condiția suficientă de ordinul al doilea 4p; concluzia 1p.

Timp de lucru: 100 minute

Model examen final

Partea I (30') (scris) Metoda celei mai bune descreșteri cu regula minimizării exacte. Descriere, convergență

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 soluții ale acestui sistem folosind pentru funcția $f : \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$f(x, y) = (x^2 - y^2 + x - 3y - 2)^2 + (x^3 - y^4 + 2)^2 + (x^2 + y^3 + 2x - y + 1.1)^2,$$

cu data inițială $(1.5, -1.75)$, una dintre metodele de mai jos:

- (1) metoda lui Newton;
- (2) metoda celei mai bune direcții de descreștere 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: enunț 1 p, demonstrație 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 variationnelle et optimization*, Editions Cépenduès, Toulouse, 2010.
- [6] A. Beck, *First-order methods in optimization*, SIAM, 2017.
- [7] M. Durea, *O introducecere în teoria optimizării 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] P.T. Leeson, E.E. Piano, *The golden age of mercenaries*, European Review of Economic History, 25 (2020), 429–446.
- [11] Y. Nesterov, *Introductory lectures on convex optimization*, Springer, 2004.
- [12] J. Nocedal, S.J. Wright, *Numerical optimization*, Springer, 2006.
- [13] P. Pedregal, *Introduction to optimization*, Springer, 2004.
- [14] P. Pedregal, *Optimization and approximation*, Springer, 2017.