# Tehnici de Optimizare

**Cristian OARA**

Facultatea de Automatica si Calculatoare
Universitatea Politehnica Bucuresti
Splaiul Independentei 313,
Bucuresti, Romania
Fax: + 40 21 3234 234
Email: oara@riccati.pub.ro

Cursul cuprinde trei mari parti: Programare Neliniara, Programare Liniara si Control Optimal si Robust. Structura cursului pe capitole si sectiuni este urmatoarea:

# PARTEA I: PROGRAMARE NELINIARA

## 3 METODE FUNDAMENTALE DE CAUTARE

2 Constructia Inversei
3 Metoda Davidon–Fletcher–Powell
4 Familii Broyden
5 Proprietati de Convergenta
6 Metode Qvasi–Newton fara Memorie

## 6 CONDITII DE MINIMIZARE CU CONSTRANGERI

1 Constrangeri
2 Plan Tangent
3 Conditii Necesare de Ordinul Intai (Constrangeri de Tip Egalitate)
4 Exemple
5 Conditii de Ordinul Doi
6 Valori Proprii in Subspatiul Tangent
7 Constrangeri de Tip Inegalitate

# 7 MINIMIZARE CU CONSTRANGERI – PRINCIPII GENERALE ALE ALGORITMILOR

# PARTEA a II–a: PROGRAMARE LINIARA

## 8 PROPRIETATI DE BAZA ALE PROGRAMARII LINIARE

1 Introducere
2 Exemple de Probleme de Programare Liniara
3 Teorema Fundamentala a Programarii Liniare
4 Relatii cu Convexitatea
5 Exercitii

## 9 METODA SIMPLEX

1 Pivoti
2 Puncte Extreme Adiacente
3 Determinarea unei Solutii Minime Fezabile
4 Algoritmul

# PARTEA a III–a: CONTROL OPTIMAL SI ROBUST

## 10 NOTIUNI DE BAZA

1 Sistem Liniar
2 Proprietati de Baza ale Sistemelor Liniare
3 Spatii de Semnale si Functii de Transfer
4 Evolutii pe Spatiul Starilor Generate de Intrari $L^2$
5 Operatorul $L^2$ Intrare–Iesire

## 11 OPTIMIZARE PATRATICA & TRIPLETE POPOV

1 Definitii si Echivalenta
2 Transformari sub Echivalenta:
   Indici Patratici
   Sistemul Riccati

Ecuatia Matriciala Algebrica Riccati
Sistemul Kalman–Yakubovich–Popov
Sistemul Hamiltonian
Functia Popov
Operatorul I/O al Sistemului Hamiltonian

## 12 TEORIE RICCATI: ABORDARE DINAMICA SI FRECVENTIALA

1 Operatorul I/O al Sistemului Hamiltonian
2 Principalul Rezultat in Domeniul Timp
3 Cazul Clasic de Pozitiviate: LQP
4 Ridicarea Ipotezei de Stabilitate
5 Conditia de Signatura
6 Optimizare Maxmin/Jocuri Dinamice
7 Conditii Frecventiale

8 Conditia de Signatura Frecventiala

9 Inegalitati matriciale Riccati

## 13 ECUATII RICCATI SI FASCICOLE MATRICIALE : CAZUL REGULAT

1 Valori proprii pentru fascicole matriciale: cazul regulat

2 Structura de valori proprii a unui fascicol Hamiltonian extins (EHP)

3 ARE si EHP

4 Ecuatia Bernoulli

5 Algoritmi numerici

## 14 APLICATII IN TEORIA SISTEMELOR

1 Lema de Real-Marginire

1 Formularea problemei
2 Presupuneri de baza
3 Problema $H^\infty$ si conditia de signatura
4 Solutia
5 Schita demonstratiei

# 18 PROBLEMA DE STABILIZARE ROBUSTA

# Capitolul 1: INTRODUCERE

Optimizare

Tipuri de Probleme

Dimensiunea Problemelor

Algoritmi Iterativi si Convergenta

# 1. Optimizare

- Principiu fundamental al unei multitudini de probleme complexe de alocare si decizie;

- Implica selectia de valori pentru o multime de variabile inter–relationate prin focalizarea atentiei asupra unei singur obiectiv proiectat (ales special) pentru a cuantifica performanta si a masura calitatea deciziei;

  Obiectivul:

- Este de obicei o functie sau functionala;

- Este maximizat sau minimizat cu posibile constrangeri care limiteaza alegerea valorilor pentru variabile;

- Izoleaza si caracterizeaza un aspect relevant al unei probleme in timp ce teoria optimizarii furnizeaza un cadru adecvat pentru analiza;

- Poate fi :
  - Profitul sau pierderea intr-un mediu de afaceri;
  - Viteza sau distanta intr–o problema de fizica;
  - Veniturile probabile intr-un mediu de investitii supuse riscului;
  - Bunastare sociala in contextul planificarii guvernamentale;

## Atentie !

- Este foarte rar posibil sa reprezentam toate complexitatile interactiunilor dintre variabile, constrangeri si obiective;
- Formularea unei probleme de optimizare trebuie sa fie privita doar ca o aproximatie;
- Intocmai ca si in alte domenii ale ingineriei abilitatile in modelare si o judecata corecta in interpretarea rezultatelor sunt absolut necesare pentru a obtine concluzii relevante;
- Experienta practica si intelegerea profunda a teoriei sunt cheia succesului;
- Trebuie stapanit **compromisul** in construirea unui model matematic

care descrie exact complexitatea problemei (suficient de complex) si care este fezabil din punct de vedere numeric (nu prea complex)

Accentul cursului: Stapanirea acestui compromis si obtinerea de abilitati pentru a putea construi modele in mod expert ceea ce inseamna:

- Identificarea si surprinderea aspectelor importante ale unei probleme;
- Abilitatea de a distinge modelele fezabile de cele nefezabile prin studierea tehnicilor existente si a teoriei asociate;
- Extinderea teoriei disponibile in cazul unor noi situatii;

# 2. Tipuri de Probleme

Vom considera in principal urmatoarele subiecte:

- Programare Liniara
- Probleme Neconstranse
- Probleme Constranse (cu constrangeri algebrice)
- Subiecte avansate in domeniul Controlului Automat al Sistemelor: (Control Optimal; Optimizare H-2 si H-infinit, optimizare Nehari si Aproximanti, Inegalitati Matriceale, Jocuri Diferentiale)

## Programare Liniara:

- Functiile obiectiv sunt combinatii liniare de necunoscute;
- Constrangerile sunt egalitati sau inegalitati matriciale in necunoscute;
- Aceste metode sunt relativ populare datorita fazei de modelare si nu atat datorita teoriei relativ simple;
- Se poate testa chiar si pentru probleme neliniare facandu–se in prealabil o liniarizare;

Exemplu: Problema cu o constrangere de buget. Avem o suma de

bani fixa pe care trebuie sa o impartim intre doua cheltuieli de tipul

$$x_1 + x_2 \leq B.$$

$x_i$: suma alocata activitatii $i$;
$B$: bugetul total. Sa presupunem ca functia obiectiv este greutatea maxima. Atunci avem

$$\begin{gathered} \text{maximizeaza} \quad w_1 x_1 + w_2 x_2 \\ \text{unde } x_1 + x_2 \leq B, \\ x_1 \geq 0, \quad x_2 \geq 0, \end{gathered}$$

unde $w_i$ este greutatea unitara (pe unitatea de pret) $i$. Aceasta este o problema tipica (si elementara) de programare liniara.

Problemele de programare liniara apar adesea in economie si finante

si mai rar in inginerie. De aceea ne vom concentra in special asupra problemelor de programare neliniara.

### Probleme fara Constrangeri:

- Au o importanta atat teoretica cat si practica;
- Din punct de vedere practic o problema constransa poate fi intotdeauna transformata printr-o simpla reformulare intr-o problema neconstransa (In problema noastra cu bugetul constrangerea nu este foarte relevanta intrucat intotdeauna putem imprumuta bani la un anumit cost (rata a dobanzii));
- Din punct de vedere teoretic problemele cu constrangeri pot fi in multe cazuri transformate in probleme fara constrangeri (de exemplu o egalitate de forma $x_1 + x_2 = B$);

- In majoritatea cazurilor problemele fara constrangeri sunt piatra de temelie pentru a intelege problemele mai sofisticate cu constrangeri;

## Probleme cu Constrangeri:

- In practica este mai uzual si mai simplu sa formulam direct o problema cu constrangerile ei naturale (de exemplu o problema complexa este reformulata sub forma catorva subprobleme cu anumite constrangeri inerente);

- Formularea matematica generala:

$$\begin{array}{ll} & \text{minimizeaza} \quad f(x) \\ \text{cu} & h_i(x) = 0, \quad i = 1, 2, \ldots, m, \\ & g_j(x) \leq 0, \quad j = 1, 2, \ldots, r, \\ & x \in \mathcal{S} \end{array}$$

unde $x$ este un vector $n$–dimensional de necunoscute, $f$, $h_i$ si $g_j$ sunt

functii de $x$ cu valori si $\mathcal{S}$ este o submultime a lui $\mathbb{R}^n$; $f$ este functia obiectiv si $h_i, g_j$ si $\mathcal{S}$ sunt constrangerile.

Nota: In mod uzual vom introduce anumite ipoteze suplimentare, in principal pentru a face problemele netede intr-un anumit sens, conducand la asa numita programare in variabile continue:

- $f$ (si $h_i$, $g_j$) sunt de obicei presupuse continue si uneori chiar cu derivate continue de un anumit ordin (mici variatii in $x$ conduc la mici variatii ale diverselor valori asociate cu problema);

- $\mathcal{S}$ este in mod uzual o submultime conexa pentru a asigura ca mici variatii ale lui $x$ raman in submultime;

# 3. Dimensiunea Problemelor

O masura a complexitatii unei probleme de programare este dimensiunea masurata in termenii numarului de variabile necunoscute si/sau numarul de constrangeri.

Avand in vedere performanta calculatoarelor moderne distingem urmatoarele clase de probleme:

- Dimensiune mica (avand pana la 5 variabile si/sau constrangeri);
- Dimensiune medie (avand intre 5 si 100 variabile);
- Dimensiune mare (avand mai mult de 100 poate 1000 sau mai multe variabile).

Aceasta clasificare nu este desigur rigida dar reflecta abordarile de calcul oarecum diferite pentru aceste clase de probleme. Teoria matematica clasica – incluzand mutiplicatori Lagrange, teoreme de tip Kuhn–Tucker si extensiile lor – se aplica problemelor indiferent de dimensiunea lor. Totusi teoria nu este potrivita atunci cand abordam o problema dpdv algoritmic intrucat nu tine seama de dificultatile asociate cu rezolvarea cu un calculator a ecuatiilor rezultand din conditiile necesare de ordinul 1. De aceea trebuie sa dezvoltam instrumente numerice eficiente pentru cautarea punctelor de optim in loc sa rezolvam direct respectivele ecuatii. In zilele noastre, tehnicile de cautare pot fi aplicate pentru rezolvarea unor probleme de programare neliniara de 500 variabile si probleme de programare liniara de 400 variabile si 1000 constrangeri (definim astfel de probleme ca fiind de dimensiune medie).

Probleme cu mai multe variabile si/sau constrangeri trebuie rezolvate cu proceduri speciale care exploateaza structura speciala a acestora.

# 4. Algoritmi Iterativi si Convergenta

Cei mai multi algoritmi dezvoltati pentru a rezolva probleme de optimizare sunt iterativi:

- Se selecteaza un vector intitial $x_0$ in multimea $\mathcal{S}$;
- Algoritmul genereaza un vector mai bun $x_1 \in \mathcal{S}$;
- Algoritmul este repetat si pe baza lui $x_0$ si/sau $x_1$ este selectat un vector si mai bun $x_2 \in \mathcal{S}$;
- Se gaseste un sir de puncte din ce in ce mai bune $x_0, x_1, x_2, x_3, \ldots, x_k, \ldots$ care tinde catre punctul solutie $x^*$;

- Procesul este terminat imediat ce se ajunge la $x^*$ sau cand este gasit un punct suficient de apropiat (din motive practice).

Teoria algoritmilor iterativi se impartite in trei aspecte ce se suprapun intr-o oarecare masura:

- Dezvoltarea algoritmului propriu–zis (bazata pe problema de programare particulara, structura sa inerenta si eficienta (performanta) calculatoarelor numerice);

- Verificarea ca respectivul algoritm genereaza un sir care converge la punctul solutie (convergenta globala);

- Viteza sau rata convergentei – se refera la cat de repede se apropie de solutie sirul de puncte (convergenta locala)

# "O teorie buna inlocuieste cu succes 1000 rulari pe computer"

Rezultate disponibile:

- Programare liniara: nu exista inca o teorie folositoare privind convergenta metodei simplex (probabil cea mai veche si importanta metoda de programare liniara); acest lucru se datoreaza convergentei intr–un numar finit de pasi; Sunt disponibile anumite estimari euristice si o vasta experienta practica ;

- Programare neliniara : Proprietatile de convergenta ale unui numar mare de algoritmi au fost deduse analitic prin metode relativ elementare si au fost verificate prin numeroase exemple numerice;

Teoria ratei de convergenta are anumite proprietati atractive:

- Este simpla pentru multi algoritmi sofisticati;
- O clasa larga de algoritmi relativ diferiti au acceasi rata de convergenta. Mai precis, in multe cazuri exista o rata canonica de convergenta asociata cu o problema de programare care guverneaza diversii algoritmi aplicati acelei probleme;
- Per global, teoria convergentei este simpla si puternica.

# Capitolul 2: PROPRIETATI DE BAZA ALE SOLUTIILOR SI ALGORITMILOR

In principal consideram urmatoarea problema:

$$\begin{array}{c} \text{minimizeaza} \quad f(\mathbf{x}) \\ \text{cu} \quad \mathbf{x} \in \Omega \end{array} \tag{1}$$

unde $f$ este o functie cu valori reale si (multimea fezabila) $\Omega \subset \mathbb{R}^n$.

Ipoteza de baza: $\Omega$ coincide cu $\mathbb{R}^n$ sau este o submultime simpla a lui $\mathbb{R}^n$.

# Cuprins

# 1. Conditii Necesare de Ordinul Intai

Prima problema: Exista vreo solutie ?

Principalul instrument teoretic: Teorema lui Weierstrass.

Principalul scop: Caracterizarea punctelor solutie si obtinerea unor algoritmi care sa le gaseasca.

Deosebim doua tipuri de puncte solutie:

- puncte de minim local

- puncte de minim global

**Definitia 1.** *Un punct $x^* \in \Omega$ este un punct de minim relativ (sau punct de minim local) a lui $f$ peste $\Omega$ daca exista $\epsilon > 0$ a.i. $f(x) \geq f(x^*)$ oricare $x \in \Omega$ aflat la o distanta de $x^*$ mai mica ca $\epsilon$. Daca $f(x) > f(x^*)$ oricare $x \in \Omega$, $x \neq x^*$, intr-o vecinatate $\epsilon$ a lui $x^*$, atunci $x^*$ este punct de minim relativ in sens strict al lui $f$ peste $\Omega$.*

**Definitia 2.** *Un punct $x^* \in \Omega$ este punct de minim global al lui $f$ peste $\Omega$ daca $f(x) \geq f(x^*)$ oricare $x \in \Omega$. Daca $f(x) > f(x^*)$ oricare $x \in \Omega$, $x \neq x^*$ atunci $x^*$ este un punct de minim global in sens strict al lui $f$ peste $\Omega$.*

Atentie !

- Cand abordam o problema presupunem implicit ca ne intereseaza minimele globale;

- Realitatea numerica cat si cea teoretica dicteaza ca trebuie sa ne multumim cu puncte de minim relativ;

- De regula, conditiile si solutiile globale pot fi obtinute doar daca functia poseda anumite proprietati de convexitate ce garanteaza ca orice minim relativ este global.

Pentru a obtine conditiile ce trebuie satisfacute intr-un punct de minim, incepem prin a studia variatiile de-alungul unor directii fezabile in jurul lui $x^*$, caz in care functia devine de o singura variabila.

**Definitia 3.** *Dandu-se $x \in \Omega$ vectorul $d$ este o directie fezabila in $x$ daca exista un $\overline{\alpha} > 0$ astfel incat $x + \alpha d \in \Omega$ oricare $\alpha$, $0 \leq \alpha \leq \overline{\alpha}$.*

**Propozitia 4. [Conditii necesare de ordinul 1]** *Fie $\Omega$ o submultime in $\mathbb{R}^n$ si fie $f \in \mathcal{C}^1$ o functie definita pe $\Omega$. Daca $x^*$ este un punct de*

*minim relativ al lui $f$ peste $\Omega$, atunci pentru orice $d \in \mathbb{R}^n$ care este o directie fezabila in $x^*$, avem*

$$\nabla f(x^*)d \geq 0.$$

Un caz particular important este cand $x^*$ este in interiorul lui $\Omega$ (ca de exemplu cand $\Omega$ este multime deschisa ) si fiecare directie din $x^*$ este fezabila !!!!

**Corolarul 5. [Cazul neconstrans]** *Fie $\Omega$ o submultime a lui $\mathbb{R}^n$ si fie $f \in \mathcal{C}^1$ o functie definita pe $\Omega$. Daca $x^*$ este un punct de minim relativ a lui $f$ pe $\Omega$ si daca $x^*$ este punct interior pentru $\Omega$, atunci*

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

Nota:

- In cazul neconstrans obtinem $n$ ecuatii cu $n$ necunoscute care pot fi rezolvate explicit in multe situatii (ca de exemplu cand functia obiectiv este patratica );

- Ecuatiile care rezulta sunt in general neliniare si foarte complicat de rezolvat;

- Abordarea noastra este nu inspre rezolvarea ecuatiilor ci inspre gasirea unui sir iterativ care converge catre $x^*$.

**Exemplul 6.** *Consideram problema*

$$\text{minimizeaza} \quad f(x_1, x_2) = x_1^2 - x_1 x_2 + x_2^2 - 3x_2$$

*peste* $\mathbb{R}^2$. *Punct de minim global la* $x_1 = 1$, $x_2 = 2$.

**Exemplul 7.**

$$\begin{aligned} &\text{minimizeaza} \quad f(x_1, x_2) = x_1^2 - x_1 + x_2 + x_1 x_2, \\ &\text{cand} \qquad x_1 \geq 0, \quad x_2 \geq 0. \end{aligned}$$

*Punct de minim global*

$$x_1 = 0.5, x_2 = 0$$

*in care derivatele partiale nu se anuleaza. Cu toate acestea deoarece orice directie fezabila trebuie sa satisfaca* $x_2 \geq 0$, *avem* $\nabla f(x^*)d \geq 0$ *pentru orice directie fezabila in* $x^*$.

# 2. Exemple de Probleme fara Constrangeri

**Exemplul 8. [Productie]** *O problema uzuala in economie este determinarea celui mai bun mod de a combina diverse intrari pentru a produce un oarecare produs. Se stie functia* $f(x_1, x_2, \cdots, x_n)$ *care da cantitatea de produse realizate ca functie de intrarile* $x_1, x_2, \cdots, x_n$. *Pretul unitar al produsului este* $q$ *si preturile unitare ale intrarilor sunt* $p_1, p_2, \cdots, p_n$. *Producatorul dorind sa maximizeze profitul trebuie sa rezolve problema*

$$\text{maximizeaza} \qquad qf(x_1, x_2, \cdots, x_n) - p_1x_1 - p_2x_2 - \cdots - p_nx_n.$$

*Conditiile necesare de ordinul intai spun ca derivatele partiale in raport*

*cu $x_i$ trebuie sa se anuleze, obtinand cele $n$ ecuatii*

$$q\frac{\partial f}{\partial x_i}(x_1, x_2, \cdots, x_n) = p_i, \qquad i = 1, 2, \cdots, n.$$

**Exemplul 9. [Aproximare]** *Optimizarea se foloseste curent in teoria aproximarii – de exemplu vezi Identificarea Sistemelor. Sa presupunem ca valorile functiei $g$ sunt determinate experimental in $m$ puncte, $x_1, x_2, \cdots, x_m$ ca avand valorile $g_1, g_2, \cdots, g_m$. Dorim sa aproximam functia $g$ printr-un polinom*

$$h(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0,$$

*unde $n < m$. Definim cea mai buna aproximare ca fiind polinomul ce face ca suma patratelor erorilor $\epsilon_k := g_k - h(x_k)$ sa fie cat mai mica,*

*i.e. minimizeaza*

$$\sum_{k=1}^{m} \epsilon_k^2 = \sum_{k=1}^{m} [g_k - \sum_{j=0}^{n} a_i x_k^i]^2 = f(a)$$

*in raport cu coeficientii necunoscuti* $a^T = \begin{bmatrix} a_0 & a_1 & \cdots & a_n \end{bmatrix}$ *(functie obiectiv patratica). Dupa prelucrarea algebrica a expresiei rezulta ca*

$$f(a) = a^T Q a - 2b^T a + c,$$

*in care* $Q = [q_{ij}]$ $i = \overline{1,n}, j = \overline{1,n}$ , $b = \begin{bmatrix} b_1, & b_2 & \cdots & b_{n+1} \end{bmatrix}$, *iar*

$$q_{ij} = \sum_{k=1}^{m} x_k^{i+j}, \qquad b_j = \sum_{k=1}^{m} g_k x_k^j, \quad c = \sum_{k=1}^{m} g_k^2.$$

*Conditiile necesare de ordinul intai cer ca gradientul lui $f$ sa se anuleze ceea ce conduce la sistemul de $n+1$ ecuatii*

$$Qa = b$$

*ce se rezolva si rezulta $a$.*

# 3. Conditii de Ordinul Doi

Prin folosirea derivatelor de ordinul 2 se pot obtine conditii suplimentare ce trebuie indeplinite de catre un punct de minim. Conditiile de ordinul 2 sunt definite in termenii matricii Hessiene $\nabla^2 f$ a derivatelor partiale de ordinul 2 ale lui $f$ si ele joaca un rol cheie.

**Propozitia 10. [Conditii de ordinul 2]** *Fie $\Omega$ o submultime a lui $\mathbb{R}^n$ si fie $f \in \mathcal{C}^2$ o functie definita pe $\Omega$. Daca $x^*$ este un punct de minim relativ pentru $f$ peste $\Omega$, atunci pentru orice $d \in \mathbb{R}^n$ care este o directie fezabila in $x^*$ avem*

i) $\nabla f(x^*)d \geq 0;$

ii) Daca $\nabla f(x^*)d \geq 0,$ atunci $d^T \nabla^2 f(x^*)d \geq 0.$

Un caz particular important este atunci cand $x^*$ este in interiorul lui $\Omega$ (ca in cazul complet neconstrans).

**Propozitia 11. [Conditii necesare de ordinul 2 – cazul neconstrans** *Fie $\Omega$ submultime a lui $\mathbb{R}^n$ si fie $f \in \mathcal{C}^2$ o functie pe $\Omega$. Daca $x^*$ este un punct de minim relativ pentru $f$ peste $\Omega$ si daca $x^*$ este un punct interior a lui $\Omega$, atunci*

$$\begin{aligned} &\text{i)} \quad \nabla f(x^*) = 0; \\ &\text{ii)} \quad \forall d, \quad d^T \nabla^2 f(x^*) d \geq 0. \end{aligned}$$

Notatie: Pentru simplitate notam matricea Hessiana $n \times n$ a lui $f$ alternativ cu

$$F(x) := \nabla^2 f(x).$$

**Exemplul 12.**

$$\begin{aligned} &\text{minimizeaza} \quad f(x_1, x_2) = x_1^2 - x_1 + x_2 + x_1 x_2, \\ &\text{cand} \quad x_1 \geq 0, \quad x_2 \geq 0. \end{aligned}$$

*Minimul global este*

$$x_1 = 0.5, x_2 = 0$$

*in care derivatele partiale nu se anuleaza. Pentru* $d = (d_1, d_2)$ *avem*

$$\nabla f(x^*)d = \frac{3}{2}d_2.$$

*Deci conditia (ii) din Propozitia 10 se aplica numai daca* $d_2 = 0$. *In acest caz avem*

$$d^T \nabla^2 f(x^*)d = 2d_1^2 \geq 0,$$

*astfel incat (ii) este satisfacuta.*

**Exemplul 13.** *Consideram problema*

$$\begin{aligned} &\text{minimizeaza} \quad f(x_1, x_2) = x_1^3 - x_1^2 x_2 + 2x_2^2, \\ &\text{cand} \quad x_1 \geq 0, \quad x_2 \geq 0. \end{aligned}$$

*Daca presupunem ca solutia este in interiorul multimii fezabile, adica daca* $x_1 > 0$, $x_2 > 0$, *atunci conditiile necesare de ordinul 1 sunt*

$$3x_1^2 - 2x_1 x_2 = 0, \qquad -x_1^2 + 4x_2 = 0.$$

*Acestea au o solutie* $x_1 = x_2 = 0$ *care este un punct pe frontiera, dar exista deasemenea o solutie* $x_1 = 6, x_2 = 9$. *Sa observam ca pentru* $x_1$ *fixat in* $x_1 = 6$, *functia obiectiv atinge un minim relativ in raport cu* $x_2$ *in* $x_2 = 9$. *Reciproc, cu* $x_2$ *fixat la* $x_2 = 9$, *functia obiectiv atinge un minim relativ in raport cu* $x_1$ *la* $x_1 = 6$. *In ciuda acestui fapt, punctul* $x_1 = 6$, $x_2 = 9$ *nu este un minim relativ pentru ca*

*Hessianul este*

$$F = \begin{bmatrix} 6x_1 - 2x_2 & -2x_1 \\ -2x_1 & 4 \end{bmatrix}$$

*care evaluat in* $x_1 = 6, x_2 = 9$ *este* $F = \begin{bmatrix} 18 & -12 \\ -12 & 4 \end{bmatrix}$ *ce nu este pozitiv semidefinit. Prin urmare solutia propusa nu este un punct de minim relativ.*

# Conditii Suficiente pentru un Minim Relativ

Intarind conditiile de ordinul 2 din Propozitia 11 obtinem un set de conditii ce implica ca $x^*$ este un minim relativ. Consideram in continuare numai probleme neconstranse (sau in care minimul este in interiorul regiunii fezabile ) deoarece cazul mai general este extrem de tehnic si are o importanta teoretica si practica marginala.

**Propozitia 14. [Conditii necesare de ordinul 2 – cazul neconstrans]** *Fie $f \in \mathcal{C}^2$ o functie definita intr-o regiune in care $x^*$ este interior. Presupunem in plus ca*

$$\begin{aligned} &\text{i)} \quad \nabla f(x^*) = 0; \\ &\text{ii)} \quad F(x) > 0. \end{aligned}$$

*Atunci $x^*$ este un punct de minim local strict al lui $f$.*

# 4. Functii Convexe si Concave

Pentru a obtine o teorie care este capabila sa caracterizeze puncte de minim global si nu numai local introducem anumite presupuneri de tip convexitate care fac rezultatele mai puternice dar in acelasi timp restrang aria de aplicatii.

**Definitia 15.** *O functie $f$ definita intr–o multime convexa $\Omega$ este numita convexa daca pentru orice $x_1, x_2 \in \Omega$ si orice $\alpha$, $0 \leq \alpha \leq 1$, are loc*

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

*Daca pentru orice* $\alpha$, $0 < \alpha < 1$, *si* $x_1 \neq x_2$, *are loc*

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

$f$ *se numeste strict convexa.*

**Definitia 16.** *O functie* $g$ *definita pe o multime convexa* $\Omega$ *se numeste concava daca functia* $f = -g$ *este convexa. Functia* $g$ *este strict concava daca* $-g$ *este strict convexa.*

# Combinatii de Functii Convexe

**Propozitia 17.** *Fie $f_1$ si $f_2$ functii convexe definite pe multimea convexa $\Omega$. Atunci functia $a_1 f_1 + a_2 f_2$ este convexa pe $\Omega$, oricare $a_1 \geq 0, a_2 \geq 0$.*

**Propozitia 18.** *Fie $f$ o functie convexa definita pe o multime convexa $\Omega$. Multimea*

$$\Gamma_c = \{x : x \in \Omega, f(x) \leq c\}$$

*este convexa oricare $c \in \mathbb{R}$.*

Nota:

- O combinatie liniara de functii convexe este o functie convexa;

- Deoarece intersectia unor multimi convexe este iar convexa, multimea punctelor ce satisfac simultan

$$f_1(x) \leq c_1, f_2(x) \leq c_2, ..., f_m(x) \leq c_m,$$

unde $f_i$ sunt convexe, defineste o multime convexa.

# Proprietatile Functiilor Diferentiabile Convexe

Pentru functii diferentiabile $f$ exista anumite rezultate mai sofisticate ce caracterizeaza convexitatea.

**Propozitia 19.** *Fie $f \in \mathcal{C}^1$. Atunci $f$ este convexa pe multimea convexa $\Omega$ daca si numai daca*

$$f(y) \geq f(x) + \nabla f(x)(y - x), \quad \forall x, y \in \Omega. \tag{2}$$

Nota: Proprietatea (2) poate fi folosita ca o definitie alternativa a convexitatii. Definitia originala afirma ca o interpolare liniara intre doua puncte **supraevalueaza** in timp ce propozitia de mai sus

afirma ca o combinatie liniara bazata pe derivata locala **subevalueaza** functia.

**Propozitia 20.** *Fie $f \in \mathcal{C}^2$. Atunci $f$ este convexa pe multimea convexa $\Omega$ ce contine un punct interior daca si numai daca matricea Hessiana $F$ a lui $f$ este pozitiv semi–definita pe $\Omega$.*

Nota:

- Hessianul este o extensie la $\mathbb{R}^n$ a notiunii de curbura a unei functii;
- Functiile convexe au curbura pozitiva in orice directie;
- Putem deasemenea introduce notiuni precum convexitate locala si convexitate locala stricta;
- Teoriile globala si locala sunt strans relationate;

# 5. Minimizari si Maximizari de Functii Convexe

**Teorema 21.** *Fie $f$ o functie convexa definita pe o multime convexa $\Omega$. Atunci multimea $\Gamma$ pe care $f$ is atinge minimumul este convexa, si orice minim relativ a lui $f$ este si minim global.*

"Toate punctele de minim sunt plasate impreuna intr-o multime convexa si toate minimele relative sunt si globale"

**Teorema 22.** *Fie $f \in \mathcal{C}^1$ o functie convexa definita pe o multime convexa $\Omega$. Daca exista un punct $x^* \in \Omega$ a.i.*

$$\forall y \in \Omega, \quad \nabla f(x^*)(y - x^*) \geq 0$$

*atunci $x^*$ este minim global pentru $f$ pe $\Omega$.*

"Pentru functii convexe continue si diferentiabile satisfacerea conditiilor de ordinul intai este si necesara si suficienta pentru ca un punct sa fie minim global"

**Teorema 23.** *Fie $f$ o functie convexa definita pe multimea marginita, inchisa si convexa $\Omega$. Daca $f$ are un maximum pe $\Omega$ atunci acesta este atins intr–un punct de extrem al lui $\Omega$.*

Nota: Exista o dualitate intre minimizarea functiilor convexe si maximizarea functiilor concave !

# 6. Convergenta Globala a Algoritmilor de Descrestere

Exista o serie de proprietati de baza pe care orice algoritm propus trebuie sa le indeplineasca:

- Iterativ: Algoritmul genereaza o serie de puncte, fiecare punct fiind calculat pe baza punctelor ce il preced;
- Descrestere: In fiecare nou punct generat de algoritm valoarea corespunzatoare a unei anumite functii descreste in raport cu valoarea in punctul precedent;
- Convergenta globala: Pentru puncte de plecare (initializari)

arbitrare algoritmul genereaza un sir care converge (intr-un numar finit sau infinit de pasi) la un punct solutie.

Atentie ! Cei mai importanti algoritmi ce vor fi prezentati in continuare **nu sunt (!!!!) global convergenti** in forma lor pura si cel mai adesea necesita anumite modificari speciale; anumiti algoritmi cu initializari arbitrare pot sa nu conveargа deloc iar altii pot converge la puncte care nu sunt solutii !!!

Convergenta globala a unui algoritm poate fi tratata intr-o maniera unitara prin intermediul teoriei generale a algoritmilor dezvoltate pentru prima data de catre Zangwill (cel mai important rezultat : Teorema de Convergenta Globala).

# Algoritmi

Un algoritm poate fi privit ca o aplicatie : dandu-se un punct $x$ intr-un spatiu $\mathcal{X}$, algoritmul $A$ aplicat lui $x$ genereaza un nou punct in acelasi spatiu; daca operam iterativ in aceeasi maniera se va genera un sir de puncte in $\mathcal{X}$ prin regula

$$x_{k+1} := A(x_k).$$

$A$ poate fi dat explicit de o formula sau poate fi un program lung de computer. In orice caz, pentru a putea avea o mai mare flexibilitate in analiza vom generaliza acest concept.

**Definitia 24.** *Un algoritm $A$ este o aplicatie definita pe un spatiu care atribuie fiecarui punct $x \in \mathcal{X}$ o submultime a lui $\mathcal{X}$.*

Nota:

- Termenul "spatiu" trebuie intepretat intr-un sens larg (o submultime in $\mathbb{R}^n$, sau un spatiu metric general);
- Un algoritm este o aplicatie punct–multime a lui $\mathcal{X}$ si nu o aplicatie de tip punct–punct;
- Dandu–se $x_k \in \mathcal{X}$, algoritmul genereaza multimea $A(x_k) \subset \mathcal{X}$ din care se selecteaza un element arbitrar;
- Incepand cu $x_0$ algoritmul genereaza siruri conform cu

$$x_{k+1} \in A(x_k);$$

- Incertitudinea (gradul de libertate) in aceasta descriere este gandit pentru a lua in calcul detaliile specifice ale diversilor algoritmi si nu inseamna ca algoritmii au un caracter aleator, i.e., algoritmii sunt in implementari aplicatii tip punct–punct (un anumit algoritm genereaza acelasi sir pornind din acelasi punct initial $x_0$);

**Exemplul 25.** *Pentru* $x \in \mathbb{R}$, *fie*

$$A(x) = \left[-\frac{|x|}{2}, \frac{|x|}{2}\right].$$

*Plecand din* $x_0 = 100$ *putem genera diverse siruri :*

$$100, 50, 25, 12, -6, -2, 1, \tfrac{1}{2};$$
$$100, -40, 20, -5, -2, 1, \tfrac{1}{4}, \tfrac{1}{8};$$
$$100, 10, -1, \tfrac{1}{16}, \tfrac{1}{100}, -\tfrac{1}{1000}, \tfrac{1}{10000}.$$

*Pentru a analiza anumite proprietati de convergenta nu este nevoie sa stim precis cum a fost selectat $x_{k+1}$ din $A(x_k)$ (sirul tinde oricum la zero).*

# Descrestere

**Definitia 26.** *Fie $\Gamma \subset \mathcal{X}$ o multime de solutii data si fie $A$ un algoritm pe $\mathcal{X}$. O functie continua cu valori reale $Z$ definita pe $\mathcal{X}$ se numeste functie de descrestere pentru $\Gamma$ si $A$ daca satisface*

*i) Daca $x \notin \Gamma$ si $y \in A(x)$, atunci $Z(y) < Z(x)$;*

*ii) Daca $x \in \Gamma$ si $y \in A(x)$, atunci $Z(y) \leq Z(x)$.*

Pentru problema

$$\begin{array}{c} \text{minimizeaza} \quad f(x) \\ \text{cand } x \in \Omega \end{array}$$

notiunea de multime de solutii, functie de descrestere si algoritm pot fi definite in moduri diferite:

- $\Gamma$: multimea punctelor de minim, si definim un algoritm a.i. la fiecare pas $f$ descreste, servind astfel ca functie de descrestere;

- $\Gamma$: multimea punctelor $x$ care satisfac $\nabla f(x) = 0$, si definim un algoritm a.i. la fiecare pas $|\nabla f(x)|$ (sau $f(x)$) descreste.

# Aplicatii Inchise

Un concept esential pentru a stabili proprietati de convergenta globala este acela de algoritm inchis (o extensie a notiunii de continuitate de la aplicatii de tip punct–punct la aplicatii de tip punct–submultime).

**Definitia 27.** *O aplicatie de tip punct–multime $A$ din $\mathcal{X}$ in $\mathcal{Y}$ este numita inchisa in $x \in \mathcal{X}$ daca ipotezele:*

- $x_k \longrightarrow x, \quad x_k \in \mathcal{X},$

- $y_k \longrightarrow y, \quad y_k \in A(x_k),$

*implica*

$$y \in A(x).$$

*In particular, $A$ se numeste inchisa daca este inchisa in fiecare punct in $\mathcal{X}$.*

**Exemplul 28.** *Aratati ca aplicatia definita in Exemplul 25 este inchisa.*

**Observatia 29.** *Ca un caz special, daca $A$ este de tip punct–punct si $A$ este continua atunci $A$ este inchisa. Reciproca nu este in general adevarata (demonstrati aceste afirmatii !!!)*

# Teorema de Convergenta Globala

Teorema de Convergenta Globala stabileste conditii tehnice generale in care convergenta globala poate fi garantata (globala = indiferent de punctul initial !!!).

**Teorema 30. [Teorema de Convergenta Globala]** *Fie $A$ un algoritm pe $\mathcal{X}$, si presupunem ca dandu-se $x_0$ sirul $\{x_k\}_{k=0}^{\infty}$ este generat prin*

$$x_{k+1} \in A(x_k).$$

*Fie o multime de solutii $\Gamma \subset \mathcal{X}$ si presupunem:*

*i) Toate punctele $x_k$ sunt continute intr-o multime compacta $\mathcal{S} \subset \mathcal{X}$;*

*ii) Exista o functie continua $Z$ pe $\mathcal{X}$ a.i.*

*a) Daca* $x \notin \Gamma$, *atunci* $Z(y) < Z(x)$, $\forall y \in A(x)$;
*b) Daca* $x \in \Gamma$, *atunci* $Z(y) \leq Z(x)$, $\forall y \in A(x)$;

*iii) Aplicatia* $A$ *este inchisa in punctele din afara lui* $\Gamma$.

*Atunci limita oricarui subsir convergent al lui* $\{x_k\}$ *este o solutie.*

**Corolarul 31.** *Daca in conditiile Teoremei de Convergenta Globala* $\Gamma$ *consta dintr-un singur punct* $\overline{x}$, *atunci sirul* $\{x_k\}$ *converge la* $\overline{x}$.

**Observatia 32.** *In multe privinte conditia (iii) a teoremei este cea mai importanta ; esecul multor algoritmi populari poate fi pusa pe seama nesatisfacerii acestei conditii.*

**Problema 33.** *Consideram algoritmul de tip punct–punct*

$$A(x) = \begin{cases} \frac{1}{2}(x-1)+1, & x > 1, \\ \frac{1}{2}x, & x \leq 1 \end{cases}$$

*si multimea solutiilor* $\Gamma = \{0\}$. *Aratati ca* $Z(x) = |x|$ *este o functie de descrestere pentru acest algoritm si aceasta multime de solutii. Cu toate acestea, plecand din* $x > 1$ *algoritmul genereaza un sir care converge la* $x = 1$ *si care nu este o solutie. Aratati ca acest lucru se intampla deoarece* $A$ *nu este inchis la* $x = 1$.

**Problema 34.** *Fie pe axa reala* $\mathcal{X} = \mathbb{R}$ *multimea solutiilor egala cu multimea vida, i.e.,* $\Gamma = \emptyset$, *functia de descrestere* $Z(x) = e^{-x}$, *si algoritmul* $A(x) = x + 1$. *Aratati ca toate conditiile teoremei sunt indeplinite cu exceptia lui i). Sirul generat plecand din orice punct initial diverge la infinit. Aceasta nu este o violare in sens strict a*

*teoremei (de ce ?).*

**Problema 35.** *Consideram algoritmul de tip punct–multime* $A$ *definit de*

$$A(x) = \begin{cases} [0, x), & 1 \geq x > 0 \\ 0, & x = 0. \end{cases}$$

*Fie* $\Gamma = \{0\}$. *Functia* $Z(x) = x$ *serveste drept functie de descrestere pentru ca pentru orice* $x \neq 0$ *toate punctele din* $A(x)$ *sunt mai mici decat* $x$. *Aratati ca sirul definit de*

$$x_0 = 1.$$
$$x_{k+1} = x_k - \frac{1}{2^{k+2}},$$

*satisface* $x_{k+1} \in A(x_k)$ *dar se poate vedea ca* $x_k \longrightarrow \frac{1}{2} \notin \Gamma$. *Aratati ca aceasta se intampla pentru ca algoritmul* $A$ *nu este inchis in afara multimii solutiilor.*

# 7. Viteza de convergenta

Viteza de convergenta a unui algoritm este un subiect extrem de important dar si complex. In mod esential da informatii despre cati pasi (iteratii) avem nevoie sa executam pentru a atinge un anumit grad de precizie. Introducem doua notiuni de baza:

- Ordinul de convergenta;
- Rata de convergenta

pentru un sir de numere reale $\{r_k\}_{k=0}^{k=\infty}$ ce converge la $r^*$.

**Definitia 36.** *Ordinul de convergenta a lui* $\{r_k\}$ *este definit ca*

*supremumul numerelor nenegative $p$ care satisfac*

$$0 \leq \overline{\lim}_{k \to \infty} \frac{|r_{k+1} - r^*|}{|r_k - r^*|^p} < \infty. \tag{3}$$

Nota: Pentru a fi siguri ca definitia se aplica oricarui sir trebuie sa facem niste precizari:

- limita superioara $\overline{\lim}$ se foloseste in locul limitei uzuale $\lim$;
- $0/0$ care apare atunci cand $r_k = r^*$ este considerata prin conventie finita.

Nota:

- Ordinul de convergenta este determinat doar de proprietatile "cozii" sirului (cand $k \to \infty$);

- Valori mai mari ale lui $p$ implica convergenta mai rapida deoarece distanta pana la $r^*$ este redusa (cel putin in coada) cu ordinul $p$ intr-un singur pas !!!;

- Daca sirul are ordinul $p$ si limita in (3) exista (in sens uzual), i.e.

$$\beta = \lim_{k \to \infty} \frac{|r_{k+1} - r^*|}{|r_k - r^*|^p}$$

atunci avem ca asimptotic

$$|r_{k+1} - r^*| = \beta |r_k - r^*|^p.$$

Multi algoritmi ce vor discutati mai departe au ordinul $1$ si de aceea consideram in continuare mai in detaliu acest caz.

**Definitia 37.** *Daca* $\{r_k\}$ *converge la* $r^*$ *astfel incat*

$$\lim_{k\to\infty} \frac{|r_{k+1} - r^*|}{|r_k - r^*|} = \beta < 1$$

*atunci sirul se numeste ca are convergenta liniara la* $r^*$ *cu rata de convergenta* $\beta$.

Nota:

- Un sir liniar convergent cu rata $\beta$ are o coada care converge cel putin la fel de repede ca seria geometrica $c\beta^k$ pentru o constanta $c$ (de aceea se mai numeste convergenta geometrica);

- Cand comparam doi algoritmi liniar convergenti, cu cat este mai mica rata cu atat este mai rapid algoritmul !

- Cazul extrem in care $\beta = 0$ este numit convergenta superliniara;

- Convergenta cu orice ordin mai mare decat $1$ este superliniara dar este posibil si sa avem convergenta superliniara cu ordinul $1$.

**Exemplul 38.** *Sirul* $r_k = a^k$ *unde* $0 < a < 1$ *converge la 0 cu ordinul* $1$, *deoarece*

$$\frac{r_{k+1}}{r_k} = a.$$

**Exemplul 39.** *Sirul* $r_k = a^{(2^k)}$ *unde* $0 < a < 1$ *converge la zero cu ordinul doi, deoarece*

$$\frac{r_{k+1}}{r_k^2} = 1.$$

**Exemplul 40.** *Sirul* $r_k = \frac{1}{k}$ *converge la 0. Convergenta este de ordinul 1 dar nu este liniara, deoarece*

$$\lim_{k \to \infty} \frac{r_{k+1}}{r_k} = 1$$

*adica* $\beta$ *nu este strict mai mic ca* $1$*.*

**Exemplul 41.** *Pentru sirul* $r_k = (\frac{1}{k})^k$ *ordinul de convergenta este* $1$ *pentru ca*

$$\lim_{k \to \infty} \frac{r_{k+1}}{r_k^p} = \infty$$

*pentru* $p > 1$*. Dar*

$$\lim_{k \to \infty} \frac{r_{k+1}}{r_k} = 0$$

*si prin urmare avem convergenta superliniara.*

**Observatia 42.** *Convergenta vectorilor este definita in mod uzual in raport cu o functie particulara $f$ care converteste sirul de vectori intr-un sir de numere. O astfel de functie se numeste functie de eroare. Prin urmare analizam convergenta sirului de vectori prin intermediul convergentei lui $f(x_k)$ la $f(x^*)$. In tehnicile de optimizare $f$ coincide in mod uzual cu functia obiectiv. Alternativ putem considera functia*

$$f(x) := |x - x^*|^2.$$

*In orice caz, alegerea functiei de eroare nu are o importanta capitala.*

# Capitolul 3: METODE FUNDAMENTALE DE CAUTARE

1. Cautare de tip Fibonacci si Sectiunea de Aur
2. Cautare Unidimensionala prin Metode de Interpolare
3. Convergenta Globala a Metodelor de Interpolare
4. Proprietati ale Algoritmilor de Cautare Unidimensionala
5. Cautari Unidimensionale Aproximative
6. Metoda Celei mai Abrupte Pante

7. Aplicatii ale Teoriei

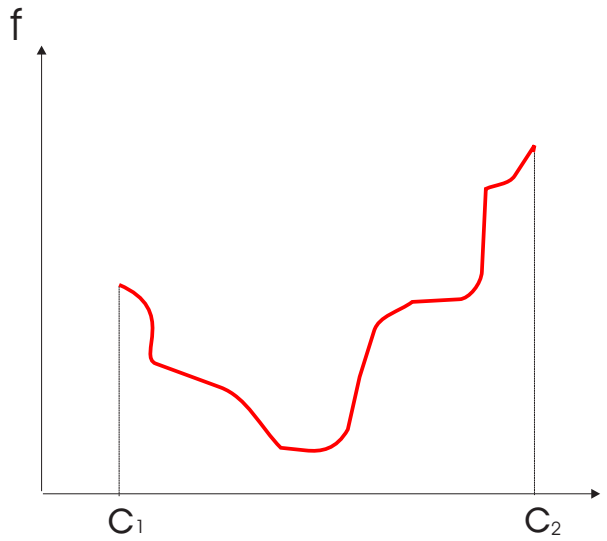
8. Metoda Newton

9. Metode de Cautare pe Coordonate

# 1. Metoda Sirului lui Fibonacci si Metoda Sectiunii de Aur

- Metode foarte populare pentru cautari unidimensionale.
- Relativ elegante dar exista alte metode care sunt in majoritatea cazurilor superioare.
- Metoda determina minimumul unei functii $f$ definita pe un interval inchis $[c_1, \quad c_2]$.
- $f$ poate fi specificata pe un interval mai larg dar in acest caz trebuie specificat un subinterval fixat pentru cautare.

- Ipoteze de baza: $f$ este unimodala.




Functie unimodala

- Nu se impun alte ipoteze !!!

- Minimul trebuie sa fie determinat evaluand functia $f$ intr-un numar

fixat de puncte (se presupune ca fiecare evaluare este relativ costisitoare dpdv al numarului de operatii/timp de calcul !!!

Problema: Gasiti o strategie pentru a alege succesiv $N$ puncte astfel incat sa putem determina cel mai mic interval de incertitudine in care se gaseste minimul !

Nota:

- Numarul $N$ este fixat *apriori*;
- Functia $f$ nu este cunoscuta apriori (i.e., valorile pe care le ia functia nu sunt cunsocute ci se calculeaza in fiecare punct).

# Sirul lui Fibonacci

Fie $N$ puncte ordonate

$$c_1 \leq x_1 < x_2 < \ldots < x_{N-1} < x_N \leq c_2$$

si fie $x_0 := c_1$ si $x_{N+1} := c_2$. Daca functia este evaluata in fiecare dintre aceste puncte stim sigur ca minimumul se gaseste in intervalul $[x_{k-1}, \quad x_{k+1}]$, unde $x_k$ este punctul in care functia ia valoarea minima dintre toate cele $N$ puncte selectate.

Problema:

Cum sa alegem $N$ puncte astfel incat intervalul $[x_{k-1}, \quad x_{k+1}]$ sa fie cat se poate de mic ?

## Raspuns:
## Sirul lui Fibonacci

Fie $d_1 := c_2 - c_1$, intervalul initial de incertitudine, $d_k$ : largimea intervalului de incertitudine dupa $k$ masuratori. Fie

$$d_k := \frac{F_{N-k+1}}{F_N} d_1 \tag{4}$$

unde $F_k$ sunt generate prin sirul lui Fibonacci definit de

$$F_0 = F_1 = 1, \qquad F_N := F_{N-1} + F_{N-2}.$$

Procedura de alegere a punctelor (se poate demonstra ca aceasta strategie este optimala !!!):

- Primele doua evaluari se fac simetric la o distanta de $\frac{F_{N-1}}{F_N} d_1$ de

capetele intervalului initial;

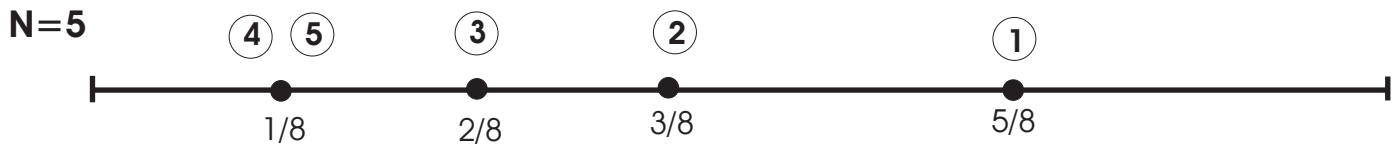
- In functie de care valore este mai mica, se alege un interval de marime $d_2 := \frac{F_{N-1}}{F_N} d_1$;

- Al treilea punct este dispus simetric in acest nou interval in raport cu pctele deja selectate;

- Folosind aceeasi regula se selecteaza un interval de lungime

$$d_3 := d_1 - \frac{F_{N-1}}{F_N} d_1 = \frac{F_{N-2}}{F_N} d_1;$$

- In general fiecare noua evaluare se face in intervalul curent de incertitudine in mod simetric cu punctul deja existent in acel interval;




## Metoda de cautare Fibonacci

Observati ca ultimele doua puncte coincid intotdeauna !!!

# Sectiunea de Aur

Principalul dezavantaj al cautarii Fibonacci este ca numarul de puncte $N$ trebuie precizat de la inceput. Daca dorim cresterea ulterioara a preciziei cautarii atunci intregul proces trebuie reluat de la capat.

O varianta alternativa este sa trecem $N \to \infty$ si sa facem o cautare corespunzatoare sirului rezultat definit de celebra Sectiune de Aur. Solutia recurentei $F_N = F_{N-1} + F_{N-2}$ este de forma

$$F_N = A\tau_1^N + B\tau_2^N \tag{5}$$

unde $\tau_1$ si $\tau_2$ sunt radacinile ecuatiei caracteristice

$$\tau^2 = \tau + 1$$

adica

$$\tau_1 = \frac{1+\sqrt{5}}{2}, \qquad \tau_2 = \frac{1-\sqrt{5}}{2}.$$

Numarul

$$\tau_1 = 1.618...$$

este numit Sectiunea de Aur fiind considerat de grecii antici ca valoarea cea mai estetica a raportului intre doua laturi adiacente ale unui dreptunghi.

Observati din (5) ca

$$\lim_{N\to\infty} \frac{F_{N-1}}{F_N} = \frac{1}{\tau_1} = 0.618...$$

Din (4) rezulta ca intervalul de incertitudine are marimea

$$d_k = \frac{1}{\tau_1^{k-1}} d_1$$

si ca

$$\frac{d_{k+1}}{d_k} = \frac{1}{\tau_1}.$$

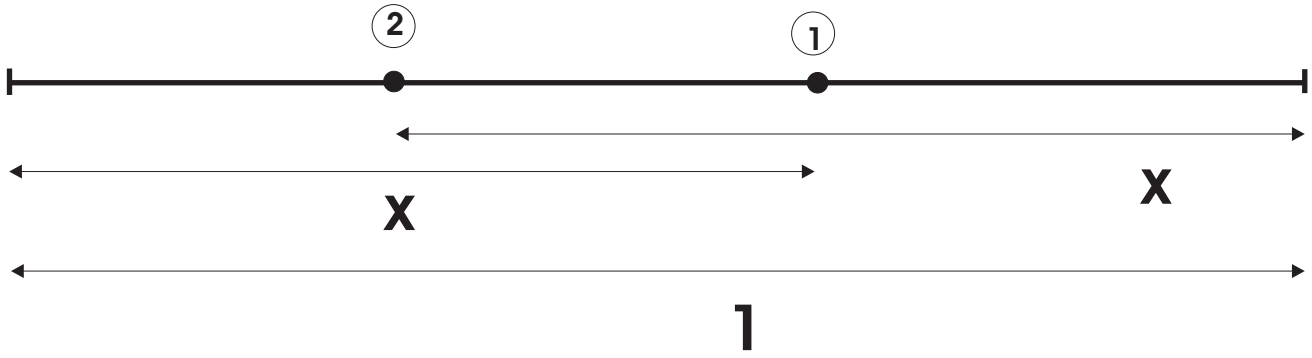
Aceasta arata ca in raport cu marimea intervalului metoda sectiunii de aur converge liniar (ordinul =1) cu rata de convergenta $\frac{1}{\tau_1}$.

Remarci:

- In practica functia se evalueaza tot intr-un numar finit de puncte chiar daca pentru gasirea acestora se foloseste sectiunea de aur. Atunci in raport cu acest numar finit de evaluari procesul generat de sectiunea de aur nu este optimal !

- Numarul dat de sectiunea de aur se poate obtine deasemenea punand conditia ca fiecare nou punct ales sa imparta noul interval in acelasi raport in care punctul precedent impartea intervalul

$$\frac{x}{1} = \frac{1-x}{x}.$$




Alegerea punctelor prin sectiunea de aur $\frac{x}{1} = \frac{1-x}{x}$

# 2. Cautare Unidimensionala prin Interpolare

Pentru functii cu un anumit grad de netezime se pot obtine algoritmi de cautare unidimensionala mai eficienti decat metoda lui Fibonacci.

Ideea principala este interpolarea unui numar de cateva puncte (obtinute la ultimii pasi) printr–o curba neteda pe baza careia se obtine o estimare a punctului de minim. Exista o mare varietate de metode de interpolare depinzand de numarul de puncte de interpolare, ordinul derivatelor folosite, criteriul utilizat pentru evaluarea calitatii interpolarii :

- Metoda lui Newton

- Metoda Falsei Pozitii
- Interpolare Cubica
- Interpolare Patratica

Toate aceste metode au ordin de convergenta mai mare decat 1.

# Metoda lui Newton

Ipoteze: Functia $f : \mathbb{R} \to \mathbb{R}$, $f \in \mathcal{C}^2$, si primele doua derivate pot fi masurate (evaluate) in fiecare punct curent din domeniul de definitie.

Ideea: Construim un interpolant de gradul 2

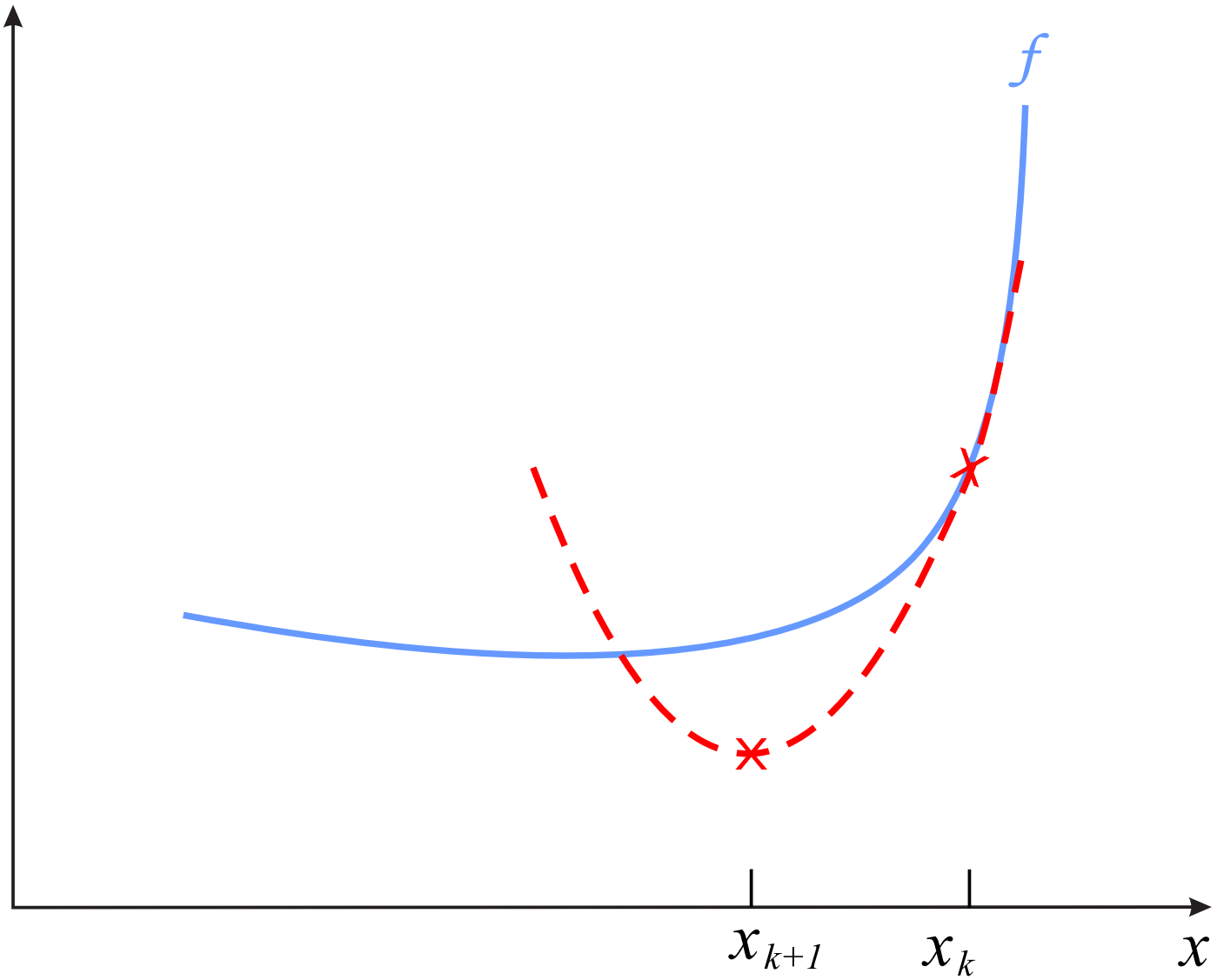
$$q(x) = f(x_k) + f^{'}(x_k)(x - x_k) + \frac{1}{2} f^{''}(x_k)(x - x_k)^2$$

in punctul curent $x_k$ si gasim un nou punct $x_{k+1}$ ca minim al lui $q(x)$:

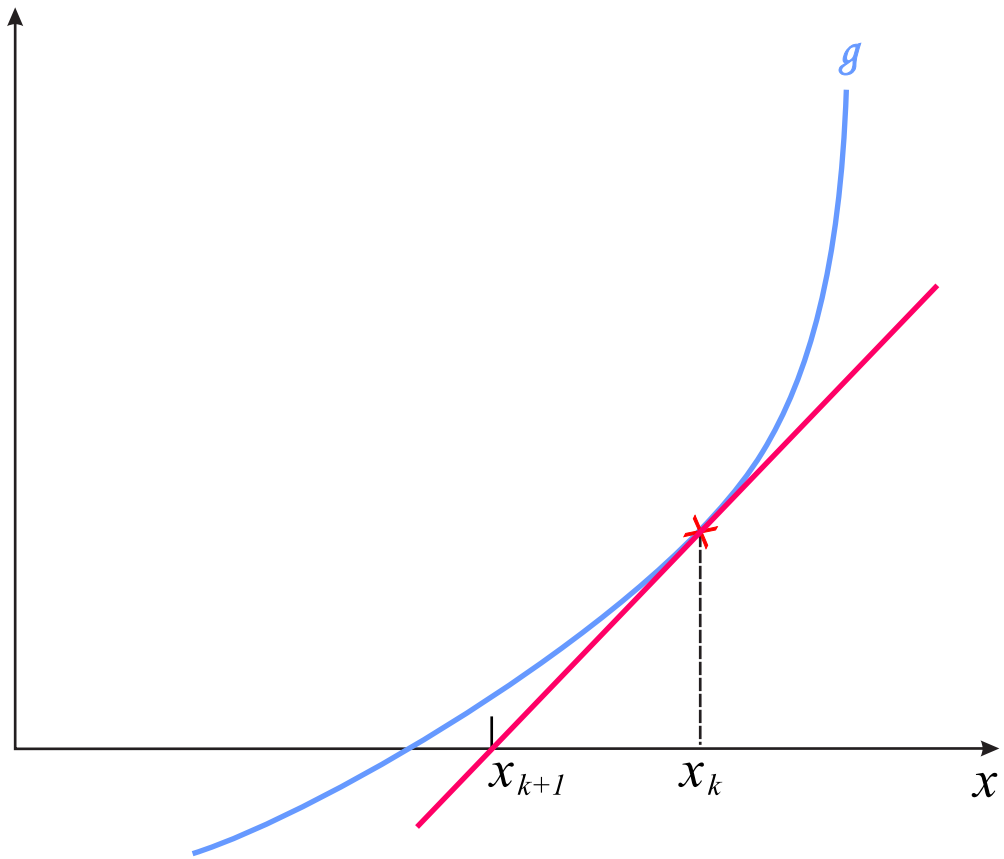
$$q^{'}(x_{k+1}) = 0 \quad \Rightarrow \quad x_{k+1} = x_k - \frac{f^{'}(x_k)}{f^{''}(x_k)}; \quad \left( x_{k+1} = x_k - \frac{g(x_k)}{g^{'}(x_k)} \right).$$

Nota:

- Procesul de cautare nu depinde de valoarea $f(x_k)$
- Metoda poate fi interpretata si ca o metoda de rezolvare iterativa a ecuatiilor $g(x) = 0$.




Metoda lui Newton




Metoda lui Newton - rezolvarea ecuatiilor

**Teorema 43.** *Fie* $g \in \mathcal{C}^2$, *si fie* $x^*$ *a.i.* $g(x^*) = 0$, $g^{'}(x^*) \neq 0$. *Daca*

*$x_0$ este suficient de aproape de $x^*$, sirul $\{x_k\}_{k=0}^{\infty}$ generat de metoda lui Newton converge la $x^*$ cu un ordin cel putin 2.*

# Metoda Falsei Pozitii

Metoda lui Newton se bazeaza pe informatie intr-un singur punct. Folosind mai mult decat un singur punct putem scadea ordinul derivatelor implicate in interpolare (folosim un numar mai mic de derivate).

Functia de interpolare este :

$$q(x) = f(x_k) + f^{'}(x_k)(x - x_k) + \frac{1}{2}\frac{f^{'}(x_{k-1}) - f^{'}(x_k)}{x_{k-1} - x_k}(x - x_k)^2$$

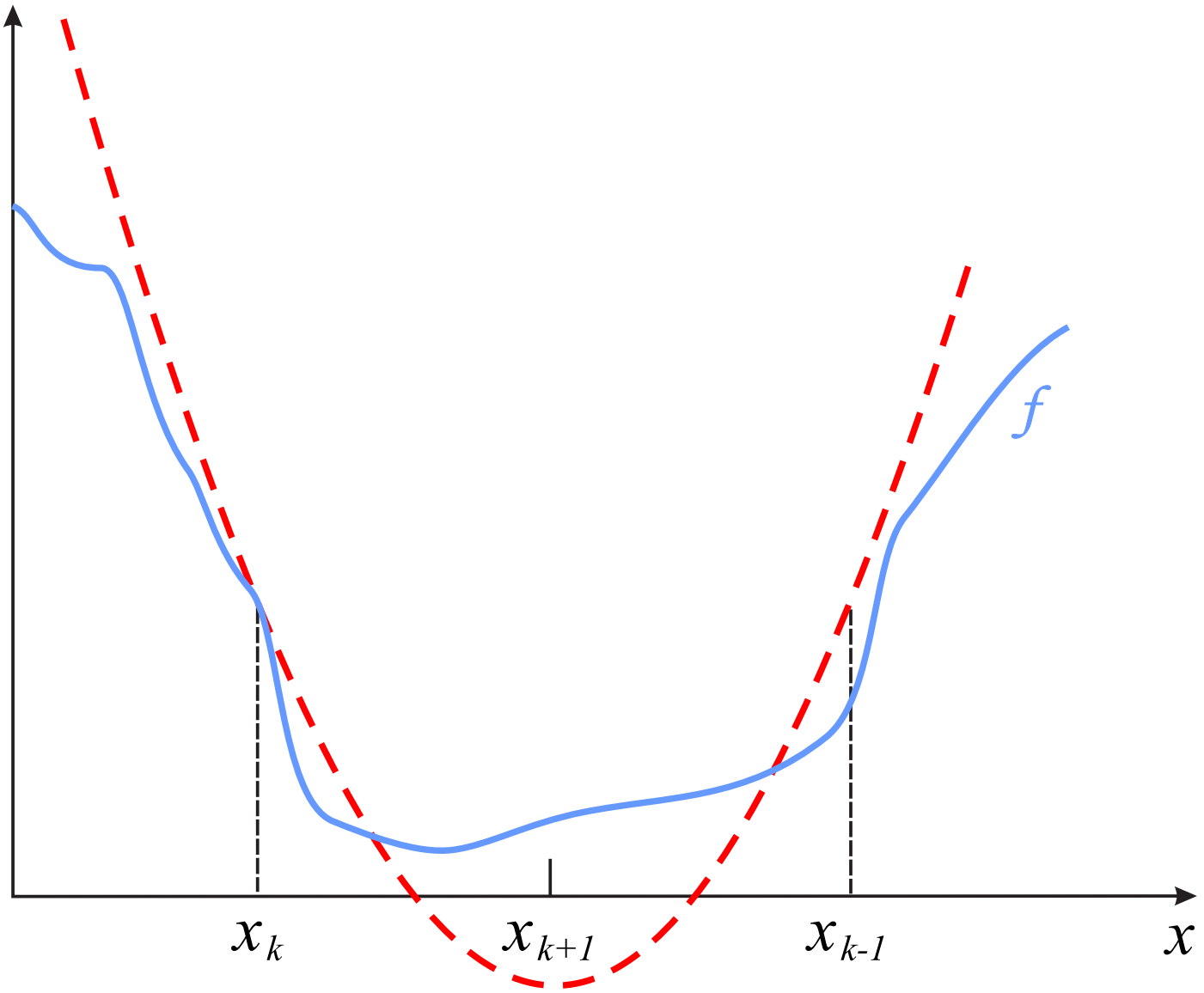
si noul punct $x_{k+1}$ este obtinut din

$$q^{'}(x_{k+1}) = 0 \Rightarrow x_{k+1} = x_k - f^{'}(x_k)\frac{x_{k-1} - x_k}{f^{'}(x_{k-1}) - f^{'}(x_k)}; \quad \left(x_{k+1} = x_k -\right.$$

Nota:

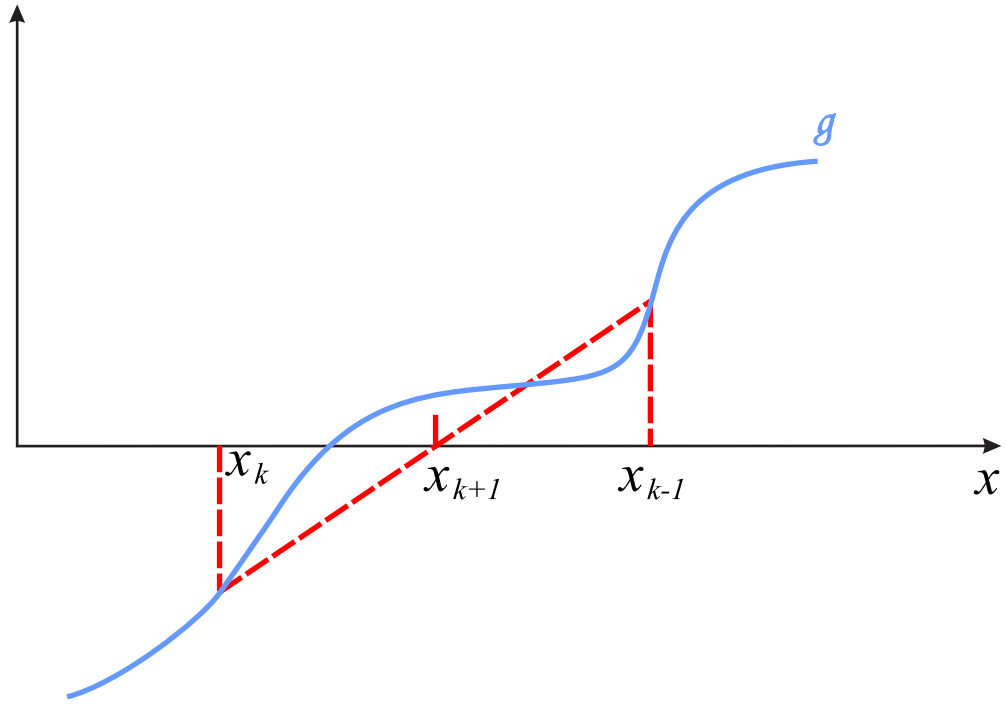
• Procesul de cautare nu depinde de $f(x_k)$ si interpolantul poate fi considerat la fel de bine in raport cu $x_{k-1}$ sau cu $x_k$;

• Metoda poate fi privita ca o aproximare a metodei lui Newton in care insa derivata a doua se inlocuieste cu o aproximare;




Metoda falsei pozitii

• Metoda poate din nou fi privita ca o metoda de rezolvare iterativa a ecuatiilor $g(x) = 0$, cu $g(x) = f^{'}(x)$.




Metoda falsei pozitii - rezolvarea ecuatiilor

**Teorema 44.** *Fie* $g \in \mathcal{C}^2$, *si fie* $x^*$ *a.i.* $g(x^*) = 0$, $g^{'}(x^*) \neq 0$. *Daca*

*$x_0$ este suficient de aproape de $x^*$, sirul $\{x_k\}_{k=0}^{\infty}$ generat de metoda falsei pozitii converge la $x^*$ cu ordinul $\tau_1 = 1.618...$ (sectiunea de aur).*

## Interpolare Cubica

Ideea: Dandu-se punctele $x_{k-1}$ si $x_k$ impreuna cu valorile $f(x_{k-1})$, $f'(x_{k-1})$, $f(x_k)$, $f'(x_k)$, se construieste un interpolant $q(x)$ de gradul 3, conducand la procesul iterativ

$$x_{k+1} = x_k - (x_k - x_{k-1})\frac{f'(x_k) + u_2 - u_1}{f'(x_k) - f'(x_{k-1}) + 2u_2}$$

unde

$$\begin{aligned} u_1 &= f^{'}(x_{k-1}) + f^{'}(x_k) - 3\frac{f(x_{k-1})-f(x_k)}{x_{k-1}-x_k}, \\ u_2 &= [u_1^2 - f^{'}(x_{k-1})f^{'}(x_k)]^{\frac{1}{2}}. \end{aligned}$$

Nota: Ordinul de convergenta al metodei de interpolare cubica este 2 in ciuda faptului ca metoda este exacta pentru functii cubice ceea ce ar fi putut sugera ca ordinul de convergenta este de 3 !!!

# Interpolare patratica

Aceasta metoda este cel mai adesea folosita in cautari unidimensionale avand avantajul esential ca nu necesita nici o informatie privitoare la derivate.

Ideea: Dandu-se punctele $x_1, x_2, x_3$, impreuna cu valorile $f(x_1) = f_1, f(x_2) = f_2, f(x_3) = f_3$, se construieste o functie de interpolare de gradul 2 $q(x)$ care trece prin aceste puncte:

$$q(x) = \sum_{i=1}^{3} f_i \frac{\Pi_{j \neq i}(x - x_j)}{\Pi_{j \neq i}(x_i - x_j)}.$$

Noul punct $x_4$ se determina ca punctul in care derivata lui $q(x)$ se

anuleaza. Se obtine

$$x_4 = \frac{1}{2}\frac{b_{23}f_1 + b_{31}f_2 + b_{12}f_3}{a_{23}f_1 + a_{31}f_2 + a_{12}f_3},$$

unde $a_{ij} = x_i - x_j$, $b_{ij} = x_i^2 - x_j^2$.

Ordinul de convergenta al metodei de interpolare cubica este $\lambda = 1.3...$, cea mai mare radacina a ecuatiei $\lambda^3 - \lambda - 1 = 0$.

# 3. Convergenta Globala a Metodelor de Interpolare

Pana in prezent am studiat convergenta algoritmilor de cautare unidimensionala in jurul punctului solutie. Mai departe trebuie insa sa ne asiguram ca acesti algoritmi converg cu adevarat indiferent de intializare, i.e., trebuie sa fim convinsi ca ei au convergenta globala.

Problema: Din pacate, algoritmii de cautare unidimensionala in forma lor pura nu au convergenta globala !!!. Din fericire, putem insa modifica acesti algoritmi intr-un mod simplu a.i. sa asiguram convergenta globala !!! In continuare vom descrie mai pe larg cum putem face aceste modificari in cazul interpolarii patratice.

Ipoteze: Functia $f$ care trebuie minimizata este strict unimodala si este de clasa $\mathcal{C}^2$.

Procedura:

Pasul 1. Gasim trei puncte $x_1$, $x_2$, si $x_3$, cu

$$x_1 < x_2 < x_3 \qquad \text{a.i.} \qquad f(x_1) \geq f(x_2) \leq f(x_3).$$

Acest lucru se poate face in mai multe feluri (Exercitiu !). Explicatia consta in aceea ca un interpolant patratic care trece prin aceste puncte va avea un minim (si nu un maxim) si acest minim se gaseste in intervalul $[x_1, x_3]$.

Pasul 2: Se gaseste $x_4$ in mod standard prin minimizarea functiei patratice, si $f(x_4)$ este evaluata (masurata). Presupunand

$$x_2 < x_4 < x_3$$

si datorita naturii unimodale a lui $f$ exista doar doua posibilitati:

$$1. \quad f(x_4) \leq f(x_2);$$

$$2. \quad f(x_2) < f(x_4) \leq f(x_3).$$

Pasul 3: Selectam un nou triplet in modul urmator:

- Cazul 1: $(\overline{x}_1, \overline{x}_2, \overline{x}_3) = (x_2, x_4, x_3)$.
- Cazul 2: $(\overline{x}_1, \overline{x}_2, \overline{x}_3) = (x_1, x_2, x_4)$.

Pasul 4: Continuam intr-un mod similar cu Pasul 2 gasind un nou interpolant patratic.

Cu aceasta modificare se poate demonstra ca procesul de interpolare patratica este convergent global !!!

Nota: • Modificarile de mai sus pentru interpolarea patratica pot fi adaptate pentru alti algoritmi de cautare unidimensionala (considerati ca exercitii cateva alte cazuri !!!)

• Ordinul de convergenta poate fi compromis (devine mai mic decat 1.3 cat este in forma pura a algoritmului) daca nici macar local nu se poate forma tripletul de interpolare din trei puncte precedente (in loc de patru) – acest lucru desigur nu se poate asigura in general.

Remarci:

• In forma pura a algoritmului de interpolare patratica punctul urmator se construieste pe baza punctului curent si a altor doua puncte precedente in timp ce in modificarea propusa mai sus punctul urmator se construieste pe baza punctului curent si a doua din trei puncte precedente ce impreuna formeaza un triplet de interpolare.

• Daca doua puncte sunt egale, de exemplu $x_1 = x_2$, atunci

$(x_1, x_2, x_3)$ formeaza un triplet de interpolare daca $f(x_2) \leq f(x_3)$ si $f^{'}(x_2) < 0$, si un interpolant patratic se gaseste folosind valorile functiei in cele doua puncte distincte si derivata in punctul care este duplicat

• Daca trei puncte sunt egale, $x_1 = x_2 = x_3$, atunci $(x_1, x_2, x_3)$ formeaza un triplet daca $f^{'}(x_2) = 0$ si $f^{''}(x_2) \geq 0$.

# 4. Proprietati ale Algoritmilor de Cautare Unidimensionala

Cautarile unidimensionale pentru gasirea minimului sunt o parte componenta a celor mai multi algoritmi de programarea neliniara. De aceea este esential sa studiem daca acesti algoritmi sunt inchisi.

Mai precis, dandu–se o functie $f$, pentru fiecare iteratie trebuie sa specificam doi vectori (punctul initial $x$ si directia de cautare $d$) iar rezultatul este un nou punct. Obtinem urmatoarea aplicatie de tip punct–multime

$$S(x,d) = \{y : y = x + \alpha d, \quad \alpha \geq 0, \quad f(y) = \min_{0 \leq \alpha \leq \infty} f(x + \alpha d)\}. \tag{6}$$

**Teorema 45.** *Daca $f$ este continua pe $\mathbb{R}^n$, atunci aplicatia (6) este inchisa in $(x, d)$ daca $d \neq 0$.*

# 5. Cautare Unidimensionala Aproximativa

In practica nu este posibil sa obtinem exact punctul de minim asa cum presupune algoritmul ideal de cautare unidimensionala $S$. Adesea se sacrifica precizia pentru a reduce timpul de executie.

Problema: Este teoria convergentei globale valabila daca folosim cautari aproximative in locul celor ideale ?

Raspuns: NU ! dar pentru a nu complica analiza vor presupune in continuare ca la fiecare pas s-a folosit un algoritm exact de cautare unidimensionala.

Criterii uzuale pentru terminarea cautarilor unidimensionale:

- Testul Procentual

- Regula Armijo
- Testul Goldstein
- Testul Wolfe.

## Testul Procentual

• Gaseste parametrul $\alpha$ cu o anumita precizie data de un procent fixat din valoarea sa reala $0 < c < 1$ (tipic $c = 0.1, 0.05, 0.02$ sau $0.01$).

• $\alpha$ este gasit astfel incat $|\alpha - \overline{\alpha}| \leq c\overline{\alpha}$, unde $\alpha$ este valoarea exacta a minimului.

• Desigur $\alpha$ nu este cunoscut dar pentru majoritatea algoritmilor de cautare unidimensionala se poate obtine relativ facil o margine buna a erorii fractionare (relative).

• Se poata arata ca acest algoritm este inchis.

## Regula Armijo

• Regula Armijo asigura ca $\alpha$ ales nu este prea mare si apoi ca nu

este prea mic !!!

- Fie

$$\phi(\alpha) = f(x_k + \alpha d_k).$$

Atunci regula este implementata considerand functia

$$\phi(0) + \epsilon\phi^{'}(0)\alpha, \quad \text{pentru} \quad \epsilon \in (0,1) \quad \text{fixat.}$$

- $\alpha$ nu este prea mare daca

$$\phi(\alpha) \leq \phi(0) + \epsilon\phi^{'}(0)\alpha \tag{7}$$

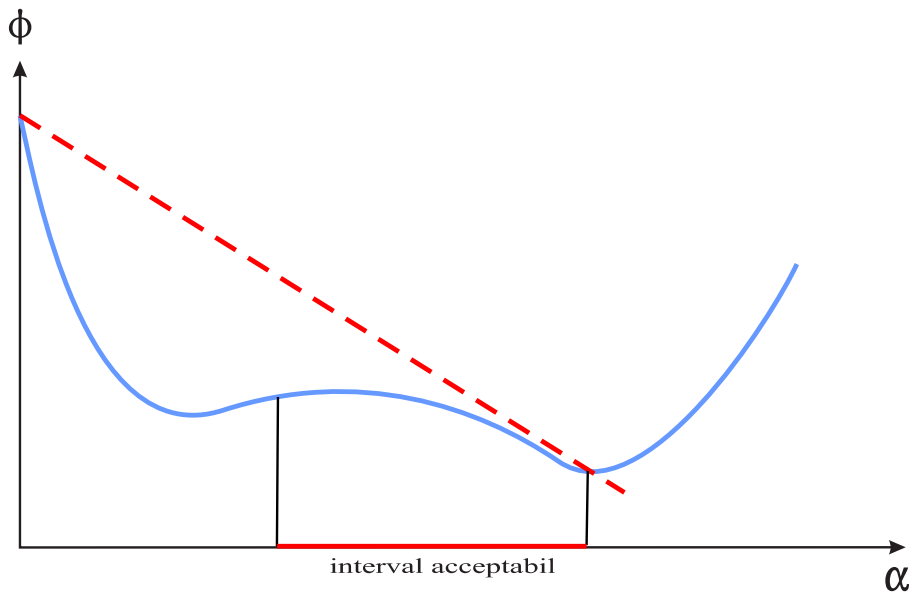
si nu este prea mic daca

$$\phi(\alpha) > \phi(0) + \epsilon\phi^{'}(0)\eta\alpha$$

unde $\eta > 1$ este ales (fixat) – aceasta este echivalent cu a spune ca

$\alpha$ este marit cu un factor de $\eta$ daca nu satisface testul (7).

• Uneori in practica regula Armijo este folosita pentru a defini un algoritm de cautare unidimensionala fara a folosi metodele de interpolare: se incepe cu un $\alpha$ arbitrar; daca satisface (7) este marit in mod repetat cu $\eta$ ( $\eta = 2$ sau $10$, si $\epsilon = 0.2$ sunt cel mai adesea folosite) pana cand (7) nu mai este satisfacut; penultimul $\alpha$ este atunci retinut. Daca $\alpha$ original nu satisface (7) este impartit la $\eta$ pana cand $\alpha$ rezultant satisface (7).



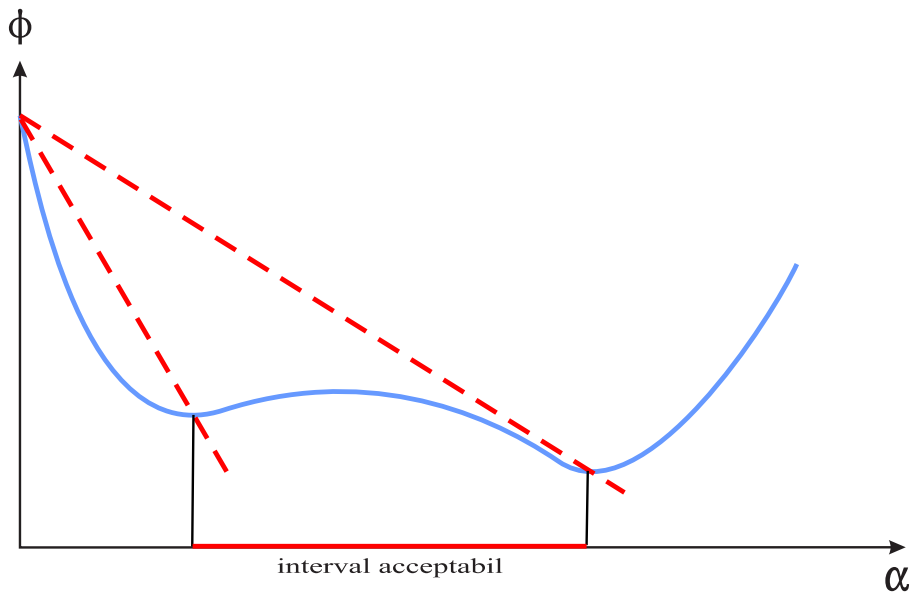

Regula lui Armijo

# Testul Goldstein

• Similar cu regula lui Armijo, $\alpha$ nu este prea mare daca satisface (7) pentru $0 < \epsilon < 0.5$, si este considerat a fi nu prea mic daca

$$\phi(\alpha) > \phi(0) + (1 - \epsilon)\phi^{'}(0)\alpha.$$

• Testul Goldstein conduce la un algoritm de cautare undimensionala inchis.




Testul Goldstein

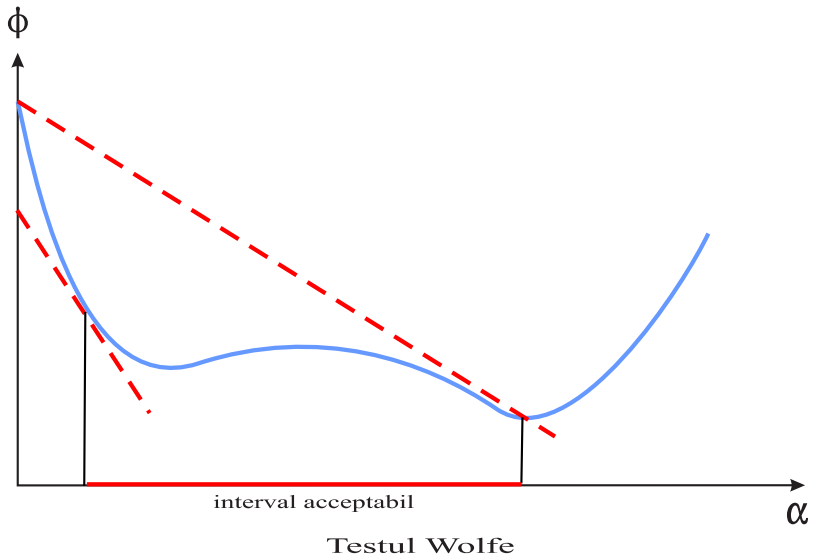
# Testul Wolfe

• Acesta este aplicat in locul testului Goldstein in cazul in care derivatele functiei obiectiv si valorile acestora pot fi obtinute relativ usor.

• $\epsilon$ este ales cu $0 < \epsilon < 0.5$, si $\alpha$ este ales sa satisfaca (7) si

$$\phi^{'}(\alpha) > (1 - \epsilon)\phi^{'}(0).$$

• Criteriul de mai sus este invariant la scalari.




Testul Wolfe

# 6. Metoda Celei Mai Abrupte Pante

• Una dintre cele mai vechi metode, cunoscuta si sub numele de metoda de gradient.

• Este extrem de importanta intrucat este cea mai simpla metoda pentru care exista o analiza satisfacatoare.

• Este simultan prima metoda care se incearca pe o problema noua si totdata referinta standard cu care se compara orice metoda noua.

## Metoda propriu–zisa

Fie $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1$. Pentru simplificarea notatiei fie

$$\begin{aligned} g(x) &:= \nabla f(x)^T \quad (\text{vector coloana}) \\ g_k &:= g(x_k). \end{aligned}$$

Metoda este definita de algoritmul iterativ

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

$\alpha_k$ este un scalar nenegativ care minimizeaza $f(x_k - \alpha_k g_k)$. Mai exact, cautarea minimului se face de–alungul directiei gradientului negativ $-g_k$ si punctul de minim este considerat $x_{k+1}$.

# Convergenta Globala

Algoritmul poate fi descris ca $A : \mathbb{R}^n \to \mathbb{R}^n$ si $x_{k+1} \in A(x_k)$. Acesta se poate descompune in $A = SG$, unde $G : \mathbb{R}^n \to \mathbb{R}^{2n}$, $G(x) = (x, -g(x))$, care da punctul initial si directia de cautare si $S : \mathbb{R}^{2n} \to \mathbb{R}^n$, dat in (6):

$$S(x,d) = \{y : y = x + \alpha d, \quad \alpha \geq 0, \quad f(y) = \min_{0 \leq \alpha \leq \infty} f(x + \alpha d)\}.$$

In Sectiunea 4 am vazut ca $S$ este inchis daca $\nabla f(x) \neq 0$, si este evident ca $G$ este continua. Prin urmare $A$ este inchisa.

Definim multimea solutiilor ca fiind punctele in care $\nabla f(x) = 0$. Atunci $Z(x) = f(x)$ este o functie de descrestere pentru $A$ deoarece

pentru $\nabla f(x) \neq 0$ avem

$$\min_{0 \leq \alpha < \infty} f(x - \alpha g(x)) < f(x).$$

Prin urmare concluzionam din Teorema de Convergenta Globala ca daca sirul $x_k$ este marginit va avea puncte limita si fiecare astfel de punct este o solutie.

# Convergenta Locala

Precum vom face in majoritatea cazurilor, incepem analiza cu cazul patratic ($f$ este o functie patratica

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

Presupunem ca $Q \in \mathbb{R}^{n\times n}$ este o matrice simetrica pozitiv definita $Q = Q^T > 0$, i.e., toate valorile ei proprii sunt pozitive, si ca toate aceste valori proprii sunt ordonate $0 < a = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n = A$. Atunci rezulta ca $f$ este strict convexa.

Punctul unic de minim $x^*$ poate fi gasit prin anularea gradientului

$g(x) = Qx - b$ obtinandu-se

$$Qx^* = b.$$

Fie functia eroare

$$E(x) := \frac{1}{2}(x - x^*)^T Q(x - x^*).$$

Atunci avem $E(x) = f(x) + \frac{1}{2}(x^*)^T Q x^*$, i.e., $E(x)$ si $f(x)$ difera doar printr-o constanta. In plus, ambele functii au acelasi punct de minim $x^*$. In multe situatii va fi mai convenabil sa consideram minimizarea lui $E$ in locul lui $f$ !

Cu aceste notatii putem exprima metoda celei mai abrupte pante

$$x_{k+1} = x_k - \alpha_k g_k, \quad g_k = Qx_k - b$$

unde $\alpha_k$ minimizeaza

$$f(x_k - \alpha g_k) = \frac{1}{2}(x_k - \alpha g_k)^T Q(x_k - \alpha g_k) - (x_k - \alpha g_k)^T b.$$

Derivand in aceasta expresie in raport cu $\alpha$ obtinem

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

In final, metoda celei mai abrupte pante ia forma explicita

$$x_{k+1} = x_k - \frac{g_k^T g_k}{g_k^T Q g_k} g_k. \tag{8}$$

**Teorema 46.** *Oricare ar fi $x_0 \in \mathbb{R}^n$, metoda celei mai abrupte pante (8) converge la punctul unic de minim $x^*$ al lui $f$. Mai mult, avem*

*la fiecare pas*

$$E(x_{k+1}) \leq \left(\frac{A-a}{A+a}\right)^2 E(x_k).$$

**Observatia 47.** • *Rata de convergenta este cu atat mai lenta cu cat contururile lui $f$ devin mai excentrice.*

- *Pentru "cercuri", convergenta este obtinuta intr-un singur pas ($A = a$).*

- *Chiar daca $n-1$ valori proprii sunt egale dar una singura este la mare distanta de ele, convergenta este lenta.*

- *In raport cu functia eroare (sau echivalent in raport cu $f$) metoda celei mai abrupte pante converge liniar cu o rata inferioara lui*

$\left(\frac{A-a}{A+a}\right)^2$.

- *Rata efectiva (reala) depinde puternic de $x_0$.*

- *Akaike a aratat ca aceasta margine se poate atinge pentru anumite puncte initiale $x_0$ si, mai mult, daca rata este defavorabila atunci este probabil ca procesul iterativ sa conveargа cu o rata apropiata de marginea superioara.*

- *Prin urmare putem spune in mare ca rata de convergenta a metodei celei mai abrupte pante este*

$$\left(\frac{A-a}{A+a}\right)^2 = \left(\frac{r-1}{r+1}\right)^2$$

*unde $r$ este numarul de conditionare al lui $Q$.*

- *Toate aceste observatii se pot aplica doar functiilor patratice. Cu toate acestea vom extinde aceste proprietati si la functii nepatratice folosind Hessianul functiei obiectiv evaluat intr-un punct solutie (in locul matricii $Q$). Aceasta tehnica este valabila pentru majoritatea metodelor avand ordinul 1 de convergenta.*

**Teorema 48.** *Fie $f$ definita pe $\mathbb{R}^n$, $f \in \mathcal{C}^2$ avand un minim local in $x^*$. Presupunem ca Hessianul $F(x^*)$ are cea mai mica valoare proprie $a > 0$ si cea mai mare valoare proprie $A > 0$. Daca $\{x_k\}$ este un sir generat de metoda celei mai abrupte pante care converge la $x^*$, atunci $f(x_k)$ converge la $f(x^*)$ liniar cu o rata de convergenta mai mica sau egala cu*

$$\left(\frac{A-a}{A+a}\right)^2.$$

# 7. Aplicatii ale Teoriei

– de discutat la Seminar –

# 8. Metoda lui Newton

Ideea: Functia obiectiv $f$ este aproximata local de o functie patratica, si functia patratica este minimizata exact:

$$f(x) \approx f(x_k) + \nabla f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T F(x_k)(x - x_k).$$

Membrul drept este minimizat exact la

$$x_{k+1} = x_k - [F(x_k)]^{-1} \nabla f(x_k)^T.$$

Presupuneri: Matricea Hesiana $F(x^*)$ este pozitiv definita (in lumina conditiilor de ordinul 2), si prin urmare metoda este bine definita in vecinatatea solutiei.

**Teorema 49.** *Fie $f$ definita pe $\mathbb{R}^n$, $f \in \mathcal{C}^3$ avand un minim local la $x^*$ si $F(x^*) > 0$. Atunci pentru o initializare suficient de aproape de $x^*$ sirul generat de metoda lui Newton converge la $x^*$ cu un ordin de convergenta de cel putin 2.*

# Modificari ale Metodei lui Newton

Metoda lui Newton necesita anumite modificari cand este startata relativ departe de punctul solutie.

Prima modificare: introducerea unui parametru de cautare $\alpha$ a.i.

$$x_{k+1} = x_k - \alpha_k [F(x_k)]^{-1} \nabla f(x_k)^T$$

unde $\alpha_k$ este ales sa minimizeze $f$. In apropierea solutiei ne asteptam desigur ca $\alpha_k \approx 1$. Introducerea acest parametru pentru puncte generale elimina posibilitatea ca functia obiectiv sa creasca cand $\alpha_k = 1$ ca urmare a termenilor nepatratici.

A doua modificare: Consideram clasa generala de algoritmi

$$x_{k+1} = x_k - \alpha M_k g_k \tag{9}$$

unde $M_k$ este o matrice $n \times n$, $\alpha$ este un parametru de cautare pozitiv, si $g_k := \nabla f(x_k)^T$. Atat metoda celei mai abrupte pante $(M_k = I)$ cat si metoda lui Newton $(M_k = F(x_k)^{-1})$ apartin acestei clase. Directia $d_k = -M_k g_k$ este o directie de descrestere daca pentru $\alpha$ mic valoarea lui $f$ descreste cand $\alpha$ creste de la zero in sus. Pentru $\alpha$ mic avem

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

sau folosind (9)

$$f(x_{k+1}) = f(x_k) - \alpha g_k^T M_k g_k + \mathcal{O}(\alpha^2).$$

Cand $\alpha \to 0$ al doilea termen il domina pe al treilea si trebuie sa avem $g_k^T M_k g_k > 0$ pentru a garanta descresterea. Modul cel mai simplu de a asigura acest lucru este sa cerem ca $M_k > 0$.

Pentru a asigura descresterea:

- Setam $M_k = I$ iar aceasta conduce la metoda celei mai abrupte pante care converge doar liniar.

- Setam $M_k = F(x_k)^{-1}$ care conduce la o descrestere foarte rapida in apropierea solutiei dar pentru un punct arbitrar poate sa nu genereze o directie de descrestere pentru ca este posibil ca $M_k$ sa nu fie pozitiv sau chiar este posibil sa nu existe.

- Setam $M_k = [\epsilon_k I + F(x_k)]^{-1}$ pentru un $\epsilon_k$ semipozitiv, ceea ce

constituie un compromis intre metoda celei mai abrupte pante ($\epsilon_k$ mare) si metoda Newton ($\epsilon_k = 0$).

Problema: Exista intotdeauana o modificare a.i. $M_k > 0$ ?

DA !!! Solutie:

Fie $F_k = F(x_k)$. Fixam $\delta > 0$. Calculam valorile proprii ale lui $F_k$ si fie $\epsilon_k$ cea mai mica constanta nenegativa pentru care matricea $\epsilon_k I + F_k$ are valoriile proprii mai mari sau egale cu $\delta$. Fie

$$x_{k+1} = x_k + \alpha_k d_k, \qquad d_k := -(\epsilon_k I + F_k)^{-1} g_k,$$

unde $\alpha_k$ minimizeaza $f(x_k + \alpha d_k)$, $\alpha \geq 0$.

Acest algoritm are proprietatile dorite de convergenta locala si globala: Presupunand ca $x_k$ este marginit, avem convergenta globala

si suficient de aproape de $x^*$ ordinul de convergenta este doi (alegand un $\delta > 0$ potrivit, adica mai mic decat cea mai mica valoare proprie a lui $F(x^*)$).

Intrebare esentiala: Cum se alege $\delta$ ?

Raspuns: Aceasta alegere implica o adevarata arta si este nevoie de mult exercitiu si experienta practica !!!! Valori mici pentru $\delta$ inseamna sa inversam matrici "aproape" singulare pe cand valori mari ale lui $\alpha$ inseamna ca ordinul doi de convergenta se poate pierde.

Dificultati suplimentare: La fiecare pas trebuie sa calculam valorile proprii ale lui $F(x_k)$ ceea ce se poate dovedi costisitor din punct de vedere numeric ! Prin urmare, in metodele de tip Levenberg–Marquardt pentru un $\epsilon_k$ dat se face o factorizare Cholesky pentru a

verifica pozitivitatea:

$$\epsilon_k I + F(x_k) = G_k^T G_k.$$

Daca algoritmul de factorizare Cholesky esueaza, $\epsilon_k$ este marit. Aceasta factorizare da deasemenea vectorul de deplasare rezolvand doua sisteme triangulare de ecuatii

$$G_k^T G_k d_k = g_k.$$

Se examineaza in continuare $f(x_k + d_k)$. Daca este destul de mai mic decat $f(x_k)$, atunci $x_{k+1}$ este acceptat si se determina un nou $\epsilon_{k+1}$. In principal $\epsilon_k$ este un parametru de cautare in aceste metode.

Atentie: Simplitatea aparenta a metodei lui Newton nu este in fapt transpusa in practica !

# 9. Metode de Cautare pe Coordonate

O clasa de algoritmi atractivi datorita simplitatii dar cu proprietati mai slabe de convergenta decat metoda celei mai abrupte pante.

Fie $f \in \mathcal{C}^1$. Descresterea in raport cu coordonata $x_i$ ($i$ fixat) este obtinuta prin rezolvarea

$$\min_{x_i} f(x_1, x_2, \ldots, x_n),$$

i.e., fiecare astfel de cautare poate fi privita ca o cautare pe directia $e_i$ (sau $-e_i$). Per ansamblu, se minimizeaza succesiv in raport cu toate coordonatele.

Moduri diverse de a implementa aceasta idee:

- Algoritmul de cautare ciclica: Se minimizeaza succesiv in raport cu $x_1, x_2, \ldots, x_n$, si apoi se repeta de la $x_1$ din nou (nu este necesara nici o informatie despre gradient).

- Metoda dublu ciclica a lui Aitken: Se minimizeaza succesiv in raport cu $x_1, x_2, \ldots, x_n$, si apoi inapoi $x_{n-1}, x_{n-2}, \ldots, x_1$ si apoi se repeta (nu este necesara nici o informatie asupra gradientului).

- Metoda Gauss–Southwell: La fiecare pas se minimizeaza coordonata corespunzatoare celei mai mari componente (in valoare absoluta) a gradientului.

# Convergenta Metodelor de Cautare pe Coordonate

Aceste metode au convergenta globala (sub anumite ipoteze rezonabile).

Convergenta locala este dificil de analizat. Pentru o problema patratica avem pentru metoda Gauss–Southwell

$$E(x_{k+1}) \leq (1 - \frac{a}{A(n-1)})E(x_k)$$

(care este o margine grosiera). Deoarece

$$\left(\frac{A-a}{A+a}\right)^2 \leq \left(1 - \frac{a}{A}\right) \leq \left(1 - \frac{a}{A(n-1)}\right)^{n-1}$$

concluzionam ca un pas din metoda celei mai abrupte pante este mai bun decat $n-1$ pasi din metoda Gauss–Southwell. Acest lucru este in acord cu multe experimente numerice. Oricum, daca variabilele sunt necuplate (corespunzand unei matrici Hessiene aproximativ diagonale, metodele de cautare pe coordonate sunt mult mai rapide decat se estimeaza.

# Capitolul 4: METODE DE DIRECTII CONJUGATE

Metodele de Directii Conjugate sunt intr-un anumit sens intermediare intre metoda celei mai abrupte pante si metoda lui Newton. Motivatia dezvoltarii acestor metode provine din dorinta de a accelera convergenta tipic lenta a metodei celei mai abrupte pante simultan cu evitarea evaluarii, memorarii si inversarii Hessianului necesare in metoda Newton. Aceste metode sunt printre cele mai eficiente metode de optimizare disponibile. Ca de obicei prezentarea metodelor de directii conjugate incepe cu problema patratica

$$\frac{1}{2}x^T Q x - b^T x, \qquad Q = Q^T > 0,$$

urmand ca apoi sa extindem, prin aproximare, la cazul problemelor

mai generale.

1. Directii Conjugate
2. Proprietati de Descrestere ale Metodei de Directii Conjugate
3. Metoda de Gradienti Conjugati
4. Metoda de Gradienti Conjugati ca Proces Optimal
5. Metoda Partiala de Gradienti Conjugati
6. Extensii la Probleme Nepatratice

# 1. Directii Conjugate

**Definitia 50.** *Dandu-se $Q = Q^T$, doi vectori $d_1$ si $d_2$ se numesc $Q$–ortogonali sau conjugati in raport cu $Q$ daca $d_1^T Q d_2 = 0$. O multime finita de vectori $d_0, d_1, \ldots, d_k$ se numeste $Q$-ortogonala daca*

$$d_i^T Q d_j = 0, \quad \forall i \neq j.$$

Cazuri particulare: $Q = 0$, $Q = I$ (in cazul nostru $Q > 0$ dar acest lucru nu este inerent in definitia de mai sus).

**Propozitia 51.** *Daca $Q = Q^T > 0$ si $d_0, d_1, \ldots, d_k$ sunt $Q$–ortogonali atunci sunt liniar independenti.*

De ce este folositoare $Q$–ortogonalitatea ? Sa presupunem ca avem problema

$$\min \frac{1}{2} x^T Q x - b^T x, \quad Q > 0,$$

care are o solutie unica $x^*$ ce satisface $Qx = b$. Fie $d_0, d_1, \ldots, d_{n-1}$ un numar de $n$ vectori nenuli si $Q$–ortogonali (care sunt automat liniar independenti). Prin urmare

$$x^* = \alpha_0 d_0 + \ldots + \alpha_{n-1} d_{n-1} \tag{10}$$

pentru anumiti $\alpha_i$ bine alesi. Inmultind cu $Q$ si luand produsele scalare cu $d_i$ obtinem

$$\alpha_i = \frac{d_i^T Q x^*}{d_i^T Q d_i} = \frac{d_i^T b}{d_i^T Q d_i} \tag{11}$$

ceea ce arata ca $\alpha_i$ si solutia $x^*$ pot fi evaluate prin intermediul unor produse scalare simple. In final avem

$$x^* = \sum_{i=0}^{n-1} \frac{d_i^T b}{d_i^T Q d_i} d_i.$$

Din relatia de mai sus se desprind cateva idei fundamentale:

• Multimea $d_i$–urilor s-a ales astfel incat toti termenii cu exceptia termenului $i$ in (10) sa se anuleze (aceasta se putea insa realiza si prin simpla ortogonalitate si nu neaparat prin $Q$–ortogonalitate).

• $\alpha_i$ s-au putut exprima in functie de vectorul cunoscut $b$ spre deosebire de situatia initiala in care erau exprimati in functie de vectorul necunoscut $x^*$.

• $x^*$ poate fi considerat rezultatul unui proces iterativ in care la pasul $i$ se adauga termenul $\alpha_i d_i$.

**Teorema 52.** *Fie $\{d_i\}_{i=0}^{n-1}$ o multime de vectori $Q$–ortogonal nenuli. Pentru oricare $x_0 \in \mathbb{R}^n$ sirul $x_k$ generat dupa formula*

$$x_{k+1} = x_k + \alpha_k d_k, \quad k \geq 0, \qquad \alpha_k = -\frac{g_k^T d_k}{d_k^T Q d_k}, \qquad g_k = Qx_k - b,$$

*converge la unica solutie $x^*$ a lui $Qx = b$ dupa $n$ pasi, i.e., $x_n = x^*$.*

# 2. Proprietati de Descrestere ale Metodei de Directii Conjugate

Fie $\mathcal{B}_k :=< d_0, d_1, \ldots, d_{k-1} >$ . Aratam ca pe masura ce metoda de directii conjugate avanseaza fiecare $x_k$ minimizeaza functia obiectiv pe varietatea liniara $k$–dimensionala $x_0 + \mathcal{B}_k$.

**Teorema 53. [Teorema de Expandare a Subspatiilor]** *Fie* $\{d_i\}_{i=0}^{n-1}$ *un sir de vectori nenuli* $Q$*–ortogonali in* $\mathbb{R}^n$*. Atunci pentru oricare* $x_0 \in \mathbb{R}^n$ *sirul* $\{x_k\}$ *generat de*

$$x_{k+1} = x_k + \alpha_k d_k, \qquad \alpha_k = -\frac{g_k^T d_k}{d_k^T Q d_k},$$

*are proprietatea ca* $x_k$ *minimizeaza* $f(x) = \frac{1}{2}x^T Q x - b^T x$ *pe dreapta*

$x = x_{k-1} + \alpha d_{k-1}$, $-\infty < \alpha < \infty$, *cat si pe varietatea liniara* $x_0 + \mathcal{B}_k$.

**Corolarul 54.** *In metoda de directii conjugate gradientii* $g_k$, $k = 0, 1, \ldots, n$ *satisfac*

$$g_k^T d_i = 0, \quad i < k.$$

Nota: • $\mathcal{B}_k \subset \mathcal{B}_{k+1}$.

• Deoarece $x_k$ minimizeaza $f$ pe $x_0 + \mathcal{B}_k$, este evident ca $x_n$ este minimumul global al lui $f$.

# 3. Metoda de Gradienti Conjugati

Directiile obtinute succesiv sunt versiuni conjugate ale gradientilor obtinuti pe masura ce metoda avanseaza (sunt generati succesiv la fiecare pas nefiind cunoscuti de la bun inceput).

La pasul $k$ se evalueaza gradientul negativ si se adauga unei combinatii liniare de directii precedente pentru a obtine o noua directie conjugata de-a lungul careia se face deplasarea.

**Avantajele metodei**:

- Cu exceptia situatiei in care solutia s-a obtinut in mai putin de $n$ pasi gradientul este intotdeauna nenul si liniar independent in raport

cu directiile precedente (gradientul este ortogonal pe subspatiul $\mathcal{B}_k$ generat de $d_0, d_1, \ldots, d_{k-1}$).

- Formula deosebit de simpla pentru generarea noii directii, facand metoda doar putin mai complicata decat metoda celei mai abrupte pante.

- Procesul de cautare progreseaza satisfactor si uniform inspre solutie la fiecare pas deoarece directiile sunt bazate pe gradienti (spre deosebire de multe metode de directii conjugate cu directii generate arbitrar si in care progresul poate fi lent pana la ultimii cativa pasi). Acest lucru nu este important pentru functii patratice ci pentru generalizarile la functii nepatratice.

# Algoritmul de Gradienti Conjugati

Incepand cu $x_0 \in \mathbb{R}^n$ definim $d_0 := -g_0 = b - Qx_0$ si

$$x_{k+1} = x_k + \alpha_k d_k, \quad \alpha_k = -\frac{g_k^T d_k}{d_k^T Q d_k}, \tag{12}$$

$$d_{k+1} = -g_{k+1} + \beta_k d_k, \qquad \beta_k = \frac{g_{k+1}^T Q d_k}{d_k^T Q d_k}, \tag{13}$$

unde $g_k = Qx_k - b$.

Nota:

- Primul pas este identic ca in metoda celei mai abrupte pante;
- La fiecare pas deplasarea se face intr-o directie ce este o combinatie

liniara a gradientului curent si a directiilor precedente;

• Remarcati formulele simple pentru actualizarea directiei de deplasare (13).

# Verificarea Algoritmului

Trebuie sa aratam ca vectorii $d_k$ sunt $Q$–ortogonali (acest lucru nu este deloc trivial !). Demonstram acest lucru impreuna cu alte proprietati in teorema ce urmeaza.

**Teorema 55. [Teorema de Gradienti Conjugati]** *Algoritmul de gradienti conjugati (12)–(13) este o metoda de directii conjugate.*

*Daca solutia nu este obtinuta la pasul $k$, i.e. in $x_k$, atunci:*

$$
\begin{aligned}
&a) \quad < g_0, g_1, \ldots, g_k > = < g_0, Qg_0, \ldots, Q^k g_0 >; \\
&b) \quad < d_0, d_1, \ldots, d_k > = < g_0, Qg_0, \ldots, Q^k g_0 >; \\
&c) \quad d_k^T Q d_i = 0, \quad \text{pentru} \quad i \leq k-1; \\
&d) \quad \alpha_k = \frac{g_k^T g_k}{d_k^T Q d_k}; \\
&e) \quad \beta_k = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k}.
\end{aligned}
$$

*Demonstratie:* Se demonstreaza simultan a), b) si c) prin inductie – (vezi notele de curs pentru demonstratia detaliata).

# 4. Metoda de Gradienti Conjugati ca Proces Optimal

Vom da in continuare o descriere alternativa a metodei de G–C care va conduce la un rezultat profund de convergenta. Rezultatul este bazat in mod esential pe punctul b) al teoremei de G–C, mai precis pe faptul ca spatiile $\mathcal{B}_k$ in care se face succesiv minimizarea sunt determinate de gradientul intitial $g_0$ si multiplicari ale acestuia cu $Q$.

Consideram o metoda alternativa de rezolvare a problemei de minimizare patratica. Dandu-se $x_0$ arbitrar, fie

$$x_{k+1} = x_0 + P_k(Q)g_0, \tag{14}$$

unde $P_k$ este un polinom de grad $k$. Alegerea unei multimi de coeficienti pentru fiecare polinom $P_k$ determina un sir $x_k$. Obtinem

$$x_{k+1} - x^* = x_0 - x^* + P_k(Q)Q(x_0 - x^*) = [I + QP_k(Q)](x_0 - x^*)$$

de unde

$$E(x_{k+1}) = \frac{1}{2}(x_{k+1}-x^*)^T Q(x_{k+1}-x^*) = \frac{1}{2}(x_0-x^*)Q[I+QP_k(Q)]^2(x_0-$$

Ne punem problema alegerii polinomului $P_k$ astfel incat sa minimizam $E(x_{k+1})$ in raport cu clasa tuturor polinoamelor de grad $k$. Dezvoltand (14), obtinem

$$x_{k+1} = x_0 + \gamma_0 g_0 + \gamma_1 Q g_0 + \cdots + \gamma_k Q^k g_0, \qquad (15)$$

in care $\gamma_i$ sunt coeficientii lui $P_k$. Avand in vedere ca

$$\mathcal{B}_{k+1} =< d_0, d_1, \ldots, d_k >=< g_0, Qg_0, \ldots, Q^k g_0 >,$$

vectorul $x_{k+1} = x_0 + \alpha_0 d_0 + \alpha_1 d_1 + \ldots + \alpha_k d_k$ generat prin metoda de gradienti conjugati are exact aceasta forma. Mai mult, conform Teoremei de Expandare a Subspatiilor coeficientii $\gamma_i$ sunt determinati de metoda de gradienti conjugati astfel incat sa minimizeze $E(x_{k+1})$.

Nota: • Problema alegerii unui $P_k$ optimal este rezolvata de metoda de gradienti conjugati.

• Relatia explicita intre coeficientii optimali $\gamma_i$ ai lui $P_k$ si $\alpha_i$, $\beta_i$ este complicata precum este si relatia intre coeficientii lui $P_k$ si ai lui $P_{k+1}$.

• Forta metodei de gradienti conjugati consta in faptul ca pe masura ce progreseaza rezolva automat problemele optimale pentru $P_k$ prin

actualizarea unei cantitati mici de informatie la fiecare pas.

**Teorema 56.** *Punctul $x_{k+1}$ generat de metoda de gradienti conjugati satisface*

$$E(x_{k+1}) = \min_{P_k} \frac{1}{2}(x_0 - x^*)^T Q[I + QP_k(Q)]^2(x_0 - x^*),$$

*unde minimul este considerat in raport cu toate polinoamele $P_k$ de grad $k$.*

# Margini de Convergenta

**Teorema 57.** *In metoda gradientilor conjugati avem*

$$E(x_{k+1}) \leq \max_{\lambda_i}[1 + \lambda_i P_k(\lambda_i)]^2 E(x_0) \tag{16}$$

*pentru orice polinom $P_k$ de grad $k$, unde maximul este peste toate valorile proprii $\lambda_i$ ale lui $Q$.*

Fiecare pas al metodei de gradienti conjugati este cel putin la fel de bun ca un pas al metodei celei mai abrupte pante considerat in acelasi punct. Intr–adevar, presupunem ca $x_k$ s-a calculat prin metoda de gradienti conjugati. Din (15) avem $x_k = x_0 + \overline{\gamma}_0 g_0 + \overline{\gamma}_1 Q g_0 + \ldots \overline{\gamma}_{k-1} Q^{k-1} g_0$. Daca $x_{k+1}$ este calculat din $x_k$ prin metoda

celei mai abrupte pante atunci $x_{k+1} = x_k - \alpha_k g_k$ pentru anumiti $\alpha_k$. Din a) al Teoremei de Gradienti Conjugati $x_{k+1}$ va avea forma (15) si deoarece pentru metoda de gradienti conjugati $E(x_{k+1})$ este mai mic decat pentru orice alt $x_{k+1}$ de forma (15), concluzia rezulta imediat. Cand avem oarecare informatii privitoare la structura de valori proprii a lui $Q$, atunci putem exploata aceasta informatie pentru a construi $P_k$. De exemplu, daca stim ca matricea $Q$ are $m < n$ valori proprii distincte, atunci printr-o alegere potrivita a lui $P_{m-1}$ este posibil sa impunem conditia ca polinomul de grad $m$ $1 + \lambda P_{m-1}(\lambda)$ sa aibe cele $m$ zerouri egale cu cele $m$ valori proprii. Folosind acest polinom particular in (16) obtinem ca $E(x_m) = 0$ si deci solutia optimala se va obtine in $m$ pasi in loc de $n$ pasi.

# 5. Metoda Partiala de Gradienti Conjugati

Ideea: Executam $m+1 \leq n$ pasi din metoda G–C dupa care, in loc sa continuam, restartam procedura din punctul curent dupa care se executa alti $m+1$ pasi ai metodei G–C ! Aceasta clasa de metode are o importanta teoretica si practica de exceptie.

Cazuri particulare:

- $m = 0$: Metoda celei mai abrupte pante standard;
- $m = n - 1$: Metoda G–C standard.

Pentru problema $\min \frac{1}{2} x^T Q x - b^T x$, schema iterativa este :

$$x_{k+1} = x_k + P^{(k)}(Q) g_k, \qquad g_k = Q x_k - b^T x,$$

unde $P_k$ este un polinom de grad $m$. Coeficientii acestui polinom sunt alesi astfel incat sa se minimizeze

$$E(x_{k+1}) = \frac{1}{2}(x_{k+1} - x^*)Q(x_{k+1} - x^*),$$

unde $x^*$ este punctul de minim. Conform rezultatelor din sectiunea precedenta, $x_{k+1}$ poate fi gasit prin executarea a $m+1$ pasi din metoda de gradienti conjugati in locul gasirii directe a coeficientilor polinomului.

Rezultatele de convergenta din sectiunea precedenta sunt valabile si in acest caz. Totusi, atunci cand matricea $Q$ are o structura particulara de valori proprii ce este relativ des intalnita in problemele de optimizare, cu predilectie in optimizari cu constrangeri, obtinem un rezultat interesant.

Sa presupunem ca matricea $Q$ are $m$ valori proprii mari (nu neaparat grupate impreuna) si $n-m$ valori proprii mai mici care apartin intervalului $[a,b]$.

Exemplu: Problema cu constrangeri

$$\begin{aligned} &\min x^T Q x - b^T x \\ &\text{cu} \quad \text{constrangerea} \quad c^T x = 0, \end{aligned}$$

ce poate fi aproximata prin problema neconstransa

$$\min x^T Q x - b^T x + \frac{1}{2}\mu (c^T x)^2, \quad \mu >> 0,$$

a carei convergenta este dictata de valorile proprii ale termenului patratic $\frac{1}{2}(Q + \mu c c^T)$. Se poate arata ca atunci cand $\mu \to \infty$, o

valoare proprie a acestei matrici tinde la infinit in timp ce celelalte $n-1$ raman marginite in intervalul $[a, A]$.

Daca aplicam matoda celei mai abrupte pante rata de convergenta va fi defavorabila si se va deteriora pe masura ce il crestem pe $\mu$. In continuare vom arata ca executand $m+1$ pasi de gradienti conjugati efectul celor mai mari $m$ valori proprii este eliminat si rata de convergenta este determinata ca si cum aceste valori proprii nu ar fi prezente.

**Teorema 58. [Metoda partiala de gradienti conjugati]** *Presupunem ca $Q = Q^T > 0$ are $n-m$ valori proprii in intervalul $[a, b]$, $a > 0$, si cele ramase sunt mai mari decat $b$. Atunci pentru metoda partiala de*

*gradienti conjugati restartata dupa* $m+1$ *pasi avem*

$$E(x_{k+1}) \leq \left(\frac{b-a}{b+a}\right) E(x_k),$$

*unde punctul* $x_{k+1}$ *se determina din* $x_k$ *prin* $m+1$ *pasi de gradienti conjugati.*

Nota: Metoda partiala de gradienti conjugati poate fi vazuta ca o generalizare a metodei celei mai abrupte pante atat in filozofie cat si in implementare si comportare. Rata de convergenta este marginita de aceeasi formula insa cu cea mai mare valoare proprie eliminata.

# 6. Extensii la Probleme Nepatratice

Metoda G–C poarte fi extinsa la cazul general (nepatratic) intr-un numar de directii diferite in functie de ce proprietati ale lui $f$ sunt mai usor de calculat: Aproximare Patratica, Metoda Fletcher–Reeves, Metoda Polak–Ribiere, si PARTAN.

## Aproximare patratica

Facem urmatoarele asocieri

$$g_k \leftrightarrow \nabla f(x_k)^T, \qquad Q \leftrightarrow F(x_k).$$

Algoritmul ce rezulta este identic cu G–C pentru probleme patratice,

si implica o aproximare patratica ce trebuie sa fie satisfacatoare macar local.

Cand se aplica problemelor nepatratice, metodele G–C nu se vor termina dupa $n$ pasi si atunci avem urmatoarele posibilitati:

- Continuam sa gasim directii conform algoritmului G–C pana cand un anumit criteriu de oprire este indeplinit;

- Algoritmul se opreste dupa $n$ (sau $n+1$) pasi si se restarteaza cu un pas de metoda de gradient (cea mai abrupta panta).

Dintr-o serie de motive metoda a doua este de preferat conducand la urmatorul algoritm:

- Pasul 1. Incepand cu $x_0$ se calculeaza $g_0 = \nabla f(x_0)^T$ si se atribuie $d_0 = -g_0$;

- Pasul 2. Pentru $k = 0, 1, \ldots, n-1$:

  a) Setam $x_{k+1} = x_k + \alpha_k d_k$, unde $\alpha_k = \frac{-g_k^T d_k}{d_k^T F(x_k) d_k}$;

  b) Se calculeaza $g_{k+1} = \nabla f(x_{k+1})^T$;

  c) Cat timp $k < n-1$, setam $d_{k+1} = -g_{k+1} + \beta_k d_k$ unde

$$\beta_k = \frac{g_{k+1}^T F(x_k) d_k}{d_k^T F(x_k) d_k}$$

  si se repeta a);

  Pasul 3. Se inlocuieste $x_0$ cu $x_n$ si se reia cu Pasul 1.

Avantaje: • Nu sunt necesare cautari unidimensionale la nici un pas;
• Algoritmul converge intr-un numar finit de pasi pentru o problema patratica.
Dezavantaje: • $F(x_k)$ trebuie evaluata la fiecare pas;
• Algoritmul nu este global convergent.

## Cautari Unidimensionale

Este de preferat sa evitam asocierea $Q \leftrightarrow F(x_k)$.

Intai se gaseste $\alpha_k$ la Pasul 2 a) printr–o cautare unidimensionala care minimizeaza functia obiectiv (aceasta este in acord cu formula din cazul patratic).

In al doilea rand, formula pentru $\beta_k$ este inlocuita cu alta formula

(echivalenta in cazul patratic):

$$\beta_k = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k}$$

(Metoda Fletcher–Reeves ) sau

$$\beta_k = \frac{(g_{k+1} - g_k)^T g_{k+1}}{g_k^T g_k}$$

(Metoda Polak–Ribiere).

# Capitolul 5: METODE DE TIP QVASI–NEWTON

In acest capitol abordam intr-un nou mod problema gasirii unor metode intermediare intre metoda celei mai abrupte pante si metoda lui Newton. Acestea sunt din nou motivate de dorinta de a accelera convergenta tipic lenta a metodei celei mai abrupte pante in paralel cu evitarea evaluarii, memorarii si inversarii Hessianului cerute de metoda lui Newton. Ideea principala aici este sa folosim o aproximatie a inversei Hessianului in locul inversei exacte.

Aceasta metode sunt cele mai sofisticate metode disponibile Analiza convergentei este relativ complicata si ne vom rezuma doar la a prezenta principalele caracteristici.

1. Metoda Newton Modificata

2. Constructia Inversei

3. Metoda Davidon–Fletcher–Powell

4. Familii Broyden

5. Proprietati de Convergenta

6. Metode Qvasi–Newton fara Memorie

# 1. Metoda Newton Modificata

Consideram problema $\min f(x)$ si propunem o rezolvare bazata pe procesul iterativ general

$$x_{k+1} = x_k - \alpha_k S_k \nabla f(x_k)^T, \tag{17}$$

unde $S_k$ este o matrice $n \times n$ simetrica si $\alpha_k$ este ales astfel incat sa minimizeze $f(x_{k+1})$. (Pentru $S_k = I$ obtinem metoda celei mai abrupte pante iar pentru $S_k = F(x_k)^{-1}$ obtinem metoda lui Newton). Filosofia acestei modificari a metodei standard este sa alegem $S_k$ aproximativ egal cu $F(x_k)^{-1}$, cu conditia suplimentarea ca aceasta aproximatie sa fie usor de construit.

Nota: • Pentru a garanta descresterea pentru $\alpha$ mici trebuie sa

impunem $S_k > 0$.

• Ca de obicei prezentam metodele intai pentru problema patratica

$$\frac{1}{2}x^T Q x - b^T x, \qquad Q = Q^T > 0$$

si apoi extindem, prin aproximare, la probleme mai generale. Pentru cazul patratic avem

$$x_{k+1} = x_k - \alpha_k S_k g_k, \qquad g_k = Q x_k - b, \qquad \alpha_k = \frac{g_k^T S_k g_k}{g_k^T S_k Q S_k g_k}. \tag{18}$$

**Teorema 59. [Cazul patratic]** *Fie $x^*$ punctul unic de minim al lui $f$. Definim $E(x) := \frac{1}{2}(x-x^*)Q(x-x^*)$. Pentru algoritmul (18)–(18)*

*avem la fiecare pas* $k$

$$E(x_{k+1}) \leq \left(\frac{B_k - b_k}{B_k + b_k}\right)^2 E(x_k),$$

*unde* $b_k$ *si* $B_k$ *sunt cea mai mica si respectiv cea mai mare valoare proprie a matricii* $S_k Q$.

Nota: • Pentru cazul patratic trebuie sa asiguram ca $S_k$ este cat mai apropiat $Q^{-1}$ intrucat in acest caz atat $b_k$ cat si $B_k$ sunt aproape de 1 si convergenta va fi rapida.

• Pentru cazul nepatratic trebuie sa asiguram ca $S_k$ este cat mai apropiat de $F(x_k)^{-1}$.

• Algoritmul (17) nu contine idei noi ci este o extensie simpla a rezultatelor anterioare.

• Teorema precedenta constituie un instrument puternic ce poate

caracteriza proprietatile de convergenta ale unor algoritmi relativ complecsi de tipul qvasi–Newton prezentati in acest capitol.

## O Metoda Clasica

O metoda standard pentru a aproxima metoda lui Newton fara evaluarea lui $F(x_k)^{-1}$ pentru fiecare $k$ este

$$x_{k+1} = x_k - \alpha_k [F(x_0)]^{-1} \nabla f(x_k)^T,$$

adica Hessianul in punctul initial $x_0$ este folosit in intregul proces de cautare. Eficacitatea procedurii este guvernata de cat de rapid se schimba Hessianul, adica de marimea derivatei de ordinul 3.

# 2. Constructia Inversei

Ideile principale ale metodelor qvasi–Newton sunt:

• Constructia unei aproximatii ale inversei Hessianului folosind informatie obtinuta pe masura ce procesul de cautare evolueaza;

• Aproximarea curenta $H_k$ este folosita la fiecare pas pentru a defini noua directie de cautare punand $S_k = H_k$ in metoda Newton modificata;

• Aproximarea converge la inversa Hessianului in punctul solutie iar metoda per ansamblu se comporta relativ similar metodei lui Newton.

Fie $f : \mathbb{R}^n$ de clasa $\mathcal{C}^2$. Pentru doua puncte $x_{k+1}$ si $x_k$ definim

$g_{k+1} := \nabla f(x_{k+1})^T$, $g_k := \nabla f(x_k)^T$ si $p_k = x_{k+1} - x_k$. Atunci

$$g_{k+1} - g_k \approx F(x_k)p_k.$$

Daca Hessianul $F$ este constant atunci avem

$$q_k := g_{k+1} - g_k = Fp_k \tag{19}$$

si vedem ca evaluarea gradientului in doua puncte ne furnizeaza informatie despre $F$.

Presupunand ca $n$ directii liniar independente $p_0, p_1, \ldots, p_{n-1}$ si $q_k$ corespunzatori sunt cunoscuti atunci $F$ este determinat in mod unic. Intr–adevar, daca $P$ si $Q$ sunt matrici $n \times n$ cu coloane $p_k$ si $q_k$ respectiv, avem

$$F = QP^{-1}.$$

Construim aproximatii succesive $H_k$ pentru $F^{-1}$ bazandu-ne pe datele obtinute la primii $k$ pasi ai procesului de cautare astfel incat daca $F$ ar fi fost constant aproximarea sa fie consistenta cu (19) pentru acesti pasi adica

$$H_{k+1}q_i = p_i, \qquad 0 \leq i \leq k. \tag{20}$$

Dupa $n$ pasi liniar independenti vom avea $H_n = F^{-1}$.

Pentru orice $k < n$ fixat, problema construirii unui $H_k$ potrivit admite o infinitate de solutii deoarece exista mai multe grade de libertate decat constrangeri. In particular, se pot satisface proprietati suplimentare prin exploatarea gradelor de libertate.

## Corectii de Rang 1

Definim o recurenta de forma

$$H_{k+1} := H_k + a_k z_k z_k^T \quad (= H_{k+1}^T)$$

unde $z_k$ este un vector si $a_k$ un scalar ce definesc o matrice de rang (cel mult) unu. Alegem $z_k$ si $a_k$ a.i. (20) este indeplinita obtinand urmatorul rezultat.

**Teorema 60.** *Fie $F = F^T$ o matrice fixata si presupunem ca vectorii $p_0, p_1, \ldots p_k$ sunt dati. Definim $q_i = Fp_i$, $i = 0, 1, \ldots, k$. Initializand cu orice matrice simetrica $H_0$ fie*

$$H_{i+1} = H_i + \frac{(p_i - H_i q_i)(p_i - H_i q_i)^T}{q_i^T (p_i - H_i q_i)}. \tag{21}$$

*Atunci*

$$p_i = H_{k+1} q_i, \qquad i \leq k.$$

Procedura globala:

1. Calculam directia de cautare $d_k$ cu formula $d_k := -H_k g_k$;
2. Minimizam $f(x_k + \alpha_k d_k)$ in raport $\alpha \geq 0$;
3. Determinam $x_{k+1} = x_k + \alpha_k d_k$, $p_k = \alpha_k d_k$ si $g_{k+1}$;
4. Calculam $H_{k+1}$ din (21) pentru $i = k$.

Dificultati: • Formula de actualizare (21) pastreaza pozitivitatea doar daca

$$q_i^T (p_i - H_i q_i) > 0.$$

• Chiar daca aceasta cantitate este pozitiva, poate fi relativ mica, conducand deci la probleme numerice majore la inversare.

# 3. Metoda Davidon–Fletcher–Powell

- Prima metoda obtinuta original de Davidon si imbunatatita de Fletcher si Powell;
- Are proprietatea atractiva ca pentru o functie patratica genereaza directiile de gradienti conjugati construind simultan inversa Hessianului !
- La fiecare pas inversa Hessianului este actualizata printr-o suma de doua matrici simetrice de rang 1;
- Este o procedura de corectie de rang 2 numita si metoda de metrica variabila.

Procedura: Se incepe cu orice matrice simetrica si pozitiv definita $H_0$, orice punct $x_0$, si $k = 0$.

Pasul 1: $d_k := -H_k g_k$;

Pasul 2: Se minimizeaza $f(x_k + \alpha d_k)$ in raport cu $\alpha \geq 0$ pentru a obtine $x_{k+1}$, $p_k = \alpha_k d_k$, si $g_{k+1}$;

Pasul 3: Se seteaza $q_k = g_{k+1} - g_k$ si

$$H_{k+1} = H_k + \frac{p_k p_k^T}{p_k^T q_k} - \frac{H_k q_k q_k^T H_k}{q_k^T H_k q_k}.$$

Se actualizeaza $k$ si se trece la Pasul 1.

Se pot demonstra urmatoarele proprietati atractive:

- $H_k > 0 \rightarrow H_{k+1} > 0.$
- Daca $f$ este functie patratica cu Hessianul constant $F$ atunci metoda D–F–P produce directii $p_k$ care sunt $F$–conjugate. Daca metoda se ruleaza $n$ pasi atunci $H_n = F^{-1}$.

**Teorema 61.** *Daca $f$ este patratica avand Hessianul $F$ pozitiv definit, atunci pentru metoda D–F–P avem*

$$p_i^T F p_j = 0, \qquad 0 \leq i < j \leq k, \tag{22}$$

$$H_{k+1} F p_i = p_i, \qquad 0 \leq i \leq k. \tag{23}$$

Nota:

• Deoarece $p_k$ sunt $F$–ortogonali si deoarece minimizam $f$ succesiv de-alungul acestor directii observam ca procedura este de fapt o metoda de gradienti conjugati.

• Daca aproximarea initiala $H_0$ este luata egala cu matricea identitate $(H_0 = I)$ metoda coincide cu metoda de gradienti conjugati.

• In orice caz, procesul iterativ converge exact in $n$ pasi.

• $p_0, p_1, \ldots, p_k$ sunt vectori proprii corespunzatori unor valori proprii unitare pentru matricea $H_{k+1}F$ (asa cum rezulta din (23)). Acesti vectori proprii sunt liniar independenti deoarece sunt $F$–ortogonali si prin urmare $H_n = F^{-1}$.

# 4. Familii Broyden

Formulele de actualizare pentru inversul Hessianului sunt bazate pe satisfacerea relatiilor

$$H_{k+1}q_i = p_i, \qquad 0 \leq i \leq k, \tag{24}$$

care s-a obtinut din

$$q_i = Fp_i, \qquad 0 \leq i \leq k,$$

ce are loc in cazul pur patratic.

Exista posibilitatea sa actualizam aproximatii ale Hessianului propriu–zis in loc sa actualizam inversa acestuia caz in care trebuie

sa satisfacem relatiile

$$q_i = B_{k+1} p_i, \qquad 0 \leq i \leq k. \tag{25}$$

Ecuatia (25) are aceeasi forma ca (24) cu exceptia faptului ca $q_i$ si $p_i$ sunt interschimbate si $H$ este inlocuit cu $B$. Prin urmare orice formula pentru $H$ obtinuta pentru a satisface (24) poate fi transformata intr-o formula de actualizare pentru $B$, adica obtinem o formula complementara prin interschimbarea rolurilor lui $B$ si $H$ si respectiv ale lui $q$ si $p$. Viceversa, orice formula de actualizare care satisface (25) poate fi transformata intr-o formula complementara pentru actualizarea lui $H$. Este usor de verificat ca daca luam complementul unui complement obtinem formula originala.

De exemplu, actualizarea de rang unu este

$$H_{k+1} = H_k + \frac{(p_k - H_k q_k)(p_k - H_k q_k)^T}{q_k^T (p_k - H_k q_k)}$$

si formula complementara corespunzatoare este

$$B_{k+1} = B_k + \frac{(q_k - B_k p_k)(q_k - B_k p_k)^T}{p_k^T (q_k - B_k p_k)}.$$

Similar, formula D–F–P

$$H_{k+1}^{DFP} = H_k + \frac{p_k p_k^T}{p_k^T q_k} - \frac{H_k q_k q_k^T H_k}{q_k^T H_k q_k}$$

devine prin luarea complementului

$$B_{k+1} = B_k + \frac{q_k q_k^T}{q_k^T p_k} - \frac{B_k p_k p_k^T B_k}{p_k^T B_k p_k} \tag{26}$$

care mai este cunoscuta sub numele de actualizarea Broyden–Fletcher–Goldfarb–Shanno.

O modalitate alternativa de a transforma o formula de actualizare pentru $H$ intr–una pentru $B$ si viceversa este prin calcularea inversei. Incepand cu

$$H_{k+1}q_i = p_i, \qquad 0 \leq i \leq k,$$

avem

$$q_i = H_{k+1}^{-1}p_i, \qquad 0 \leq i \leq k,$$

ceea ce inseamna ca $H_{k+1}^{-1}$ satisface (25) care este criteriul pentru actualizarea lui $B$.

Deosebit de important, inversa unei formule de actualizare de rang 2 este tot o formula de rang 2. Acest lucru se poate verifica folosind celebra formula Sherman–Morrison

$$[A + ab^T]^{-1} = A^{-1} - \frac{A^{-1}ab^TA^{-1}}{1 + b^TA^{-1}a} \tag{27}$$

unde $A$ este o matrice $n \times n$ si $a$ si $b$ sunt vectori $n$–dimensionali (desigur ca formula este adevarata cu conditia ca inversele respective sa existe).

Actualizarea Broyden–Fletcher–Goldfarb–Shanno pentru $H$ produce, prin calcularea inversei, o actualizare corespunzatoare pentru $H$ de forma

$$H_{k+1}^{BFGS} = H_k + \left(1 + \frac{q_k^T H_k q_k}{q_k^T p_k}\right) \frac{p_k p_k^T}{p_k^T q_k} - \frac{p_k q_k^T H_k + H_k q_k p_k^T}{q_k^T p_k}. \quad (28)$$

Remarci:

• Aceasta formula poate fi folosita in mod identic cu formula D–F–P;

• Experimente numerice au indicat ca performantele ei sunt superioare formulei D–F–P;

• Ambele formule D–F–P si B–F–G–S contin corectii de rang 2

simetrice ce se construiesc din vectorii $p_k$ si $H_k q_k$. Prin urmare o combinatie liniara ponderata a acestor formule va fi de acelasi tip (simetrica, rang 2, construita pe baza lui $p_k$ si $H_k q_k$).

Obtinem astfel o intreaga colectie de actualizari cunoscute sub numele de familii Broyden definite de

$$H^{\phi} = (1 - \phi) H^{DFP} + \phi H^{BFGS} \tag{29}$$

unde $\phi$ este un parametru ce poate lua orice valoare reala.

O metoda Broyden este definita ca o metoda qvasi–Newton in care la fiecare iteratie un membru al familiei Broyden este folosit ca formula de actualizare. In general, parametrul $\phi$ variaza de la o iteratie la alta si prin urmare trebuie sa specificam sirul $\phi_1, \phi_2, \ldots$. O metoda Broyden pura foloseste un $\phi$ constant. Deoarece $H^{DFP}$ si $H^{BFGS}$

satisfac relatia fundamentala (24) aceasta relatie este deasemenea satisfacuta de toti membrii familiei Broyden.

**Teorema 62.** *Daca $f$ este patratica cu Hessian pozitiv definit $F$, atunci pentru o metoda Broyden avem*

$$p_i^T F p_j = 0, \qquad 0 \leq i < j \leq k, \tag{30}$$

$$H_{k+1} F p_i = p_i, \qquad 0 \leq i \leq k. \tag{31}$$

Observatii:

• Familia Broyden nu pastreaza cu necesitate pozitivitatea lui $H^\phi$ pentru toate valorile lui $\phi$;

• Alegerea sirului $\phi_k$ este irelevanta pentru functii patratice (vezi

teorema de mai sus). Se poate arata ca si pentru functii nepatratice si cautari unidimensionale exacte punctele generate de toate metodele Broyden coincid (daca singularitatile sunt evitate si minimele multiple sunt rezolvate in mod consistent). Deci anumite diferente apar numai in cazul cautarilor unidimensionale inexacte.

Proprietati ale metodelor Broyden:

• Necesita numai informatie despre gradient (derivate de ordinul 1);

• Directiile de deplasare sunt intotdeauna de descrestere daca asiguram $H_k > 0$;

• In general $H_k$ converge la inversa Hessianului (in cel mult $n$ pasi pentru functii patratice);

## Metode Partiale de tip Qvasi–Newton

Metodele de tip Broyden pot fi restartate la fiecare $m + 1 < n$

pasi, obtinandu-se metode partiale de tip qvasi–Newton. Pentru $m$ mic, acestea necesita memorie modesta intrucat aproximatia inversei Hessianului poate fi memorata implicit prin memorarea vectorilor $p_i$ si $q_i$, $i \leq m+1$. In cazul patratic aceasta corespunde exact cu metoda partiala de gradienti conjugati si are proprietati similare de convergenta.

# 5. Proprietati de Convergenta

Diversele metode qvasi–Newton sunt relativ dificil de analizat si prin urmare analiza lor este adesea facuta prin analogie.

## Convergenta Globala

Metodele qvasi–Newton sunt rulate intr-o maniera continua, incepand cu o aproximatie initiala care este imbunatatita de-alungul intregului proces iterativ.

Sub anumite ipoteze relativ stringente se poate demonstra ca procedurile sunt global convergente.

Alternativ, daca se restarteaza metoda qvasi–Newton la fiecare $n$

sau $n+1$ pasi atunci convergenta globala este garantata de prezenta primului pas iterativ la fiecare ciclu.

## Convergenta Locala

Metodele au in general convergenta superliniara. Trebuie remarcat ca metodele sunt relativ sensibile la cautari unidimensionale inexacte.

# 6. Metode Qvasi–Newton fara Memorie

Analiza precedenta a metodelor qvasi–Newton poate fi folosita ca baza pentru reconsiderarea metodelor de gradienti conjugati si obtinerea unei noi clase de proceduri atractive.

Sa consideram o simplificare a metodei qvasi–Newton BFGS in care $H_{k+1}$ este definit de o actualizare BFGS aplicata lui $H = I$ in locul lui $H_k$. Astfel $H_{k+1}$ este determinat fara referinta la precedentul $H_k$ si prin urmare actualizarea se numeste fara memorie.

Procedura: Startam in orice punct $x_0$ cu $k = 0$.

Pasul 1: Setam $H_k = I$;

Pasul 2: Setam

$$d_k = -H_k g_k; \tag{32}$$

Pasul 3: Minimizam $f(x_k + \alpha d_k)$ in raport cu $\alpha \geq 0$ pentru a obtine $\alpha_k$, $x_{k+1}$, $p_k = \alpha_k d_k$, $g_{k+1}$, si $q_k = g_{k+1} - g_k$ (trebuie ales $\alpha_k$ suficient de precis pentru a asigura $p_k^T q_k > 0$).

Pasul 4: Daca $k$ nu este multiplu de $n$ setam

$$H_{k+1} = I + \left(1 + \frac{q_k^T q_k}{q_k^T p_k}\right) \frac{p_k p_k^T}{p_k^T q_k} - \frac{p_k q_k^T + q_k p_k^T}{q_k^T p_k}. \tag{33}$$

Fie $k := k + 1$ si ne intoarcem la Pasul 2. Daca $k$ este multiplu de $n$ ne intoarcem la Pasul 1.

Combinand (32) si (33) este usor de vazut ca

$$d_{k+1} = -g_{k+1} - \left(1 + \frac{q_k^T q_k}{q_k^T p_k}\right) \frac{p_k p_k^T g_{k+1}}{p_k^T q_k} + \frac{p_k q_k^T g_{k+1} + q_k p_k^T g_{k+1}}{q_k^T p_k}. \tag{34}$$

Daca fiecare cautare unidimensionala este exacta, atunci $p_k^T g_{k+1} = 0$ si prin urmare $p_k^T q_k = -p_k^T g_k$. In acest caz (34) este echivalenta cu

$$d_{k+1} = -g_{k+1} + \frac{q_k^T g_{k+1}}{p_k^T q_k} = -g_{k+1} + \beta_k d_k,$$

unde $\beta_k = \frac{q_k q_{k+1}^T}{g_k^T q_k}$. Aceasta coincide cu forma Polak–Ribiere a gradientilor conjugati. Prin urmare folosirea actualizarii BFGS in aceasta maniera conduce la un algoritm de tip Newton modificat cu coeficient matriceal pozitiv definit si care este echivalent cu o

implementare standard a metodei de gradient conjugat atunci cand cautarea unidimensionala este exacta.

Cu toate acestea acest algoritm este folosit fara cautare unidimensionala exacta intr-o forma similara cu metoda de gradienti conjugati folosind (34).

Mai general, se poate extinde aceasta idee si se poate obtine o actualizare Broyden fara memorie de forma

$$d_{k+1} = -g_{k+1} + (1-\phi)\frac{q_k^T g_{k+1}}{q_k^T q_k} q_k + \phi \frac{q_k^T g_{k+1}}{q_k^T p_k} p_k.$$

Aceasta este echivalenta cu metoda de gradienti conjugati doar pentru $\phi = 1$, corespunzand actualizarii BFGS. Prin urmare, in multe situatii, se prefera $\phi = 1$ pentru aceasta metoda.

# Capitolul 6: CONDITII DE MINIMIZARE CU CONSTRANGERI

Constrangeri

Plan Tangent

Conditii Necesare de Ordinul Intai (Constrangeri de Tip Egalitate)

Exemple

Conditii de Ordinul Doi

Valori Proprii in Subspatiul Tangent

## Constrangeri de Tip Inegalitate

Principalele idei in studiul problemelor cu constrangeri de tip egalitate sunt:

- Caracterizarea suprafetei determinate de curba fezabila definita de aceste egalitati (in $\mathbb{R}$);

- Considerarea valorii functiei obiectiv de–alungul unor curbe pe aceasta suprafata;

Incepem prin studiul conditiilor necesare si suficiente satisfacute in punctele solutie. Principalele instrumente tehnice sunt multiplicatorii Lagrange si o matrice Hessiana speciala care, folosite impreuna,

formeaza fundatia pentru dezvoltarea si analiza algoritmilor folositi in cazul constrans.

# 1. Constrangeri

Problema generala neliniara de care ne ocupam este

$$\begin{array}{rc} & \text{minimizeaza} \quad f(x) \\ \text{cand} & h_i(x) = 0, \quad i = 1, 2, \ldots, m, \\ & g_j(x) \leq 0, \quad j = 1, 2, \ldots, p, \\ & x \in \Omega \subset \mathbb{R}^n, \end{array}$$

unde $m \leq n$ iar functiile $f$, $h_i$ si $g_j$ $(i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, p)$ sunt continue (de obicei chiar de clasa $\mathcal{C}^2$). Pentru simplificarea notatiei introducem functiile vectoriale $\mathbf{h} = (h_1, h_2, \ldots, h_m)$ si $\mathbf{g} = (g_1, g_2, \ldots, g_p)$.

Constrangerile $\mathbf{h}(x) = 0$ si $\mathbf{g}(x) \leq 0$ sunt numite constrangeri

functionale in timp ce $x \in \Omega$ este o constrangere de tip multime. Intocmai ca in cazul neconstrans, consideram ca $\Omega$ este $\mathbb{R}^n$ sau presupunem ca punctele solutie sunt in interiorul lui $\Omega$.

Punct fezabil: Un punct satisfacand toate constrangerile functionale.

Constrangeri active: O constrangere de tip inegalitate $g_i(x) \leq 0$ se numeste activa intr-un punct fezabil daca $g_i(x) = 0$ si inactiva daca $g_i(x) < 0$. Prin conventie toate constrangerile de tip egalitate sunt active.

Observatii:

- Constrangerile active intr–un punct fezabil restrang domeniul de fezabilitate intr-o vecinatate a lui $x$ in timp ce cele inactive nu au

nici o influenta (in vecinatatea lui $x$);

- Presupunand ca stim a priori care constrangeri sunt active in punctele solutie, punctul solutie va fi un minim local al problemei definita prin ignorarea constrangerilor inactive si tratarea tuturor constrangerilor active drept constrangeri de tip egalitate. In acest caz se poate considera ca problema are numai constrangeri de tip egalitate;

- Vom considera in continuare probleme avand numai constrangeri de tip egalitate pentru a economisi notatie si a izola principalele idei ce guverneaza problemele constranse. Ulterior, vom proceda la anumite adaugiri privitoare la modul de alegere al constrangerilor active;

- Principala noastra preocupare o constituie solutiile locale si vom da doar anumite scurte remarci privind solutiile globale;

# 2. Plan Tangent

Multimea de egalitati

$$h_i(x) = 0, \quad i = 1, 2, \ldots, m, \tag{35}$$

defineste o submultime a lui $\mathbb{R}^n$ numita o (hiper)suprafata. Daca egalitatile sunt regulate, intr-un sens ce va fi descris mai jos, atunci hipersuprafata este regulata si are dimensiune $n - m$.

Suprafata neteda: Daca functiile $h_i$ apartin lui $\mathcal{C}^1$, $(i = 1, 2, \ldots, m)$.

Curba pe suprafata $S$: O familie de puncte $x(t) \in S$ parametrizate continuu de $t$, $a \leq t \leq b$.

Curba diferentiabila $x(t)$: Daca $x(t)$ este diferentiabila in raport cu $t$ (acesta este o functie vectoriala).

O curba $x(t)$ trece printr–un punct $x^*$ daca $x^* = x(t^*)$ pentru un $t^*$ a.i. $a \leq t^* \leq b$.

Plan tangent in punctul $x^*$: Multimea tuturor derivatelor in $x^*$ ale tuturor curbelor diferentiabile pe $S$ care trec prin $x^*$ (acesta este un subspatiu a lui $\mathbb{R}^n$).

Pentru caracterizarea planului tangent la o curba (35) introducem urmatoarele definitii.

**Definitia 63.** *Un punct $x^*$ care satisface constrangerea $\mathbf{h}(x^*) = 0$ este numit regulat pentru respectiva constrangere daca vectorii gradient $\nabla h_1(x^*), \ldots, \nabla h_m(x^*)$ sunt liniar independenti.*

**Teorema 64.** *Intr-un punct regulat* $x^*$ *pe suprafata definita de* $\mathbf{h}(x) = 0$ *planul tangent este*

$$M = \{y : \nabla \mathbf{h}(x^*)y = 0\}.$$

Observatie: Conditia de regularitate a unui punct nu este intrinseca suprafetei definita de constrangeri ci este specifica reprezentarii particulare in termenii lui $\mathbf{h}$. Planul tangent este independent de reprezentare in timp ce $M$ nu este.

**Exemplul 65.** *Fie* $\mathbf{h}(x_1, x_2) = 0$. *Atunci* $\mathbf{h} = 0$ *genereaza axa* $x_2$, *si fiecare punct pe aceasta axa este regulat. Alternativ, daca* $\mathbf{h}(x_1, x_2) = x_1^2$ *din nou* $S$ *este axa* $x_2$ *dar acum nici un punct de pe axa nu este regulat. Intr-adevar, avem* $M = \mathbb{R}^2$, *in timp ce planul*

*tangent este axa* $x_2$.

# 3. Conditii Necesare de Ordinul Intai

**Lema 66.** *Fie $x^*$ punct regulat al constrangerilor $\mathbf{h}(x) = 0$ si punct de extrem local al lui $f$ cu constrangerile $\mathbf{h}$. Atunci oricare ar fi $y \in \mathbb{R}^n$ care satisface*

$$\nabla \mathbf{h}(x^*)y = 0$$

*satisface deasemenea*

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

Aceasta lema arata ca $\nabla f(x^*)$ este ortogonal la planul tangent ceea ce implica ca este o combinatie liniara a gradientilor lui $\mathbf{h}$ evaluati in $x^*$ (coeficientii acestei combinatii liniare sunt exact multiplicatorii Lagrange).

**Teorema 67. [Conditii necesare de ordinul intai]** *Fie* $x^*$ *un punct de extrem local al lui* $f$ *cu constrangerile* $h(x) = 0$. *Presupunem ca* $x^*$ *este un punct regulat pentru aceste constrangeri. Atunci exista* $\lambda \in \mathbb{R}^m$ *astfel incat*

$$\nabla f(x^*) + \lambda^T \nabla \mathbf{h}(x^*) = 0.$$

Observatie: Conditiile necesare de ordinul intai

$$\nabla f(x^*) + \lambda^T \nabla \mathbf{h}(x^*) = 0$$

impreuna cu constrangerile

$$\mathbf{h}(x^*) = 0$$

dau $n + m$ ecuatii (in general neliniare) cu $n + m$ variabile ce sunt componentele lui $x^*$ si $\lambda$. Deci in cazuri "nepatologice" conditiile necesare pot furniza o solutie unica !

In termenii Lagrangianului asociat cu problema constransa

$$\ell(x, \lambda) := f(x) + \lambda^T h(x)$$

conditiile necesare pot fi exprimate in forma

$$\begin{aligned} \nabla_x \ell(x, \lambda) &= 0, \\ \nabla_\lambda \ell(x, \lambda) &= 0. \end{aligned}$$

## 4. Exemple

– de discutat la Seminar –

# 5. Conditii de Ordinul Doi

In aceasta sectiune presupunem ca $f, \mathbf{h} \in \mathcal{C}^2$.

**Teorema 68. [Conditii Necesare de Ordinul Doi]** *Presupunem ca $x^*$ este un minim local al lui $f$ cu constrangerile $\mathbf{h} = 0$ si este punct regulat al acestor constrangeri. Atunci exista $\lambda \in \mathbb{R}^m$ astfel incat*

$$\nabla f(x^*) + \lambda^T \nabla \mathbf{h}(x^*) = 0.$$

*Daca $M = \{y : \nabla \mathbf{h}(x^*)y = 0\}$ este planul tangent atunci matricea*

$$L(x^*) = F(x^*) + \lambda^T H(x^*)$$

*este pozitiv semidefinita pe $M$ adica*

$$y^T L(x^*) y \geq 0, \qquad \forall y \in M.$$

Observatii: Pentru o functie vectoriala $\mathbf{h} = (h_1, h_2, \ldots, h_m)$ prima derivata este o matrice $m \times n$

$$\nabla \mathbf{h}(x) := \left[ \frac{\partial h_i(x)}{\partial x_j} \right]$$

iar a doua derivata este un tensor tridimensional definit in termenii celor $m$ Hessieni $H_1(x), \ldots, H_m(x)$ corespunzatori celor $m$ componente ale functiei. Avem deasemenea ca gradientul lui $\lambda^T \mathbf{h}(x)$

este $\lambda^T \nabla \mathbf{h}(x)$ iar Hessianul este

$$\lambda^T H(x) := \sum_{i=1}^{m} \lambda_i H_i(x).$$

**Teorema 69. [Conditii Suficiente de Ordinul Doi]** *Presupunem ca exista un punct $x^*$ care satisface $\mathbf{h}(x^*) = 0$ si un $\lambda \in \mathbb{R}^n$ astfel incat*

$$\nabla f(x^*) + \lambda^T \nabla \mathbf{h}(x^*) = 0.$$

*Sa presupunem deasemenea ca matricea*

$$L(x^*) = F(x^*) + \lambda^T H(x^*)$$

*este pozitiv definita pe $M = \{y : \nabla \mathbf{h}(x^*)y = 0\}$, adica pentru*

$y \in M$, $y \neq 0$, *avem* $y^T L(x^*)y > 0$. *Atunci* $x^*$ *este un minim local strict al lui* $f$ *cu constrangerile* $h(x) = 0$.

# 6. Valori Proprii in Subspatiul Tangent

Matricea $L$ restrictionata la subspatiul $M$ care este tangent la suprafata definita de constrangeri joaca in cadrul conditiilor de ordin doi un rol similar cu cel jucat de Hessianul functiei obiectiv in cazul fara constrangeri.

Mai mult, valorile proprii corespunzatoare ale lui $L$ restrictionate la $M$ (notam acest operator cu $L_M$) determina rata de convergenta a algoritmilor corespunzatori.

Definim $L_M : M \to M$ prin $L_M(y) :=$ proiectia lui $Ly$ pe $M$. Aceasta este o aplicatie liniara din $M$ in $M$.

Un vector $y \in M$ este un vector propriu al lui $L_M$ daca exista

un numar real $\lambda$ astfel incat $L_M y = \lambda y$, si $\lambda$ este atunci o valoare proprie. In particular, $y$ este un vector propriu daca $Ly$ poate fi scris ca suma intre $\lambda y$ si un vector ortogonal pe $M$.

Reprezentare matriciala pentru $L_M$: Fie $e_1, e_2, \ldots, e_{n-m}$ o baza ortonormala pentru $M$ si $E$ matricea baza corespunzatoare de dimensiune $n \times (n-m)$,

$$\forall y \in M, \exists z \in \mathbb{R}^{n-m}, \quad \text{a. i.} \quad y = Ez.$$

Prin urmare $E^T L E z$ este vectorul ale carui componente dau reprezentarea in termenii bazei si, in mod corespunzator, matricea de dimensiune $(n-m) \times (n-m)$ $E^T L E$ este reprezentarea matriceala a lui $L$ restrictionat la $M$.

Valorile proprii ale lui $L$ restrictionat la $M$ sunt exact valorile

proprii ale lui $E^T LE$ si acestea sunt independente de matricea baza ortogonala $E$.

## Hessieni Bordati

Abordarea de mai sus pentru determinarea valorilor proprii ale lui $L_M$ este simpla si directa. Putem insa considera si o alta abordare utila bazata pe constructia unor matrici si determinanti de ordin $n+m$ in loc de $n-m$. Vom da intai cateva caracterizari :

$$M = \{x \in \mathbb{R}^n : \nabla \mathbf{h} x = 0\};$$

$$z \perp M \Leftrightarrow z = \nabla \mathbf{h}^T w, \quad w \in \mathbb{R}^m$$

(necesitatea este mai greu de aratat). In plus, $x$ este un vector propriu al lui $L_M$ daca satisface urmatoarele doua conditii:

1. $x \in M$;

2. $Lx = \lambda x + z$, unde $z \perp M$.

Acestea sunt echivalente in termenii lui $z$ cu

$$\begin{aligned} \nabla \mathbf{h} x &= 0 \\ Lx &= \lambda x + \nabla \mathbf{h}^T w \end{aligned}$$

care este un sistem de $n + m$ ecuatii liniare in necunoscutele $w$, $x$. Acesta are o solutie nenula daca si numai daca determinantul este zero, i.e.,

$$\det \begin{bmatrix} O & \nabla \mathbf{h} \\ -\nabla \mathbf{h}^T & L - \lambda I \end{bmatrix} \equiv p(\lambda) = 0.$$

Observati ca $p(\lambda)$ este un polinom in $\lambda$ de grad $n - m$. Prin modul de obtinere el este si polinomul caracteristic al lui $L_M$. Obtinem

urmatorul criteriu:

**Teorema 70. [Hessian Bordat]** *Matricea $L$ este pozitiv definita pe subspatiul $M = \{x : \nabla h x = 0\}$ daca si numai daca ultimii $n - m$ minori principali ai lui*

$$B = \begin{bmatrix} O & \nabla \mathbf{h} \\ \nabla \mathbf{h}^T & L \end{bmatrix}$$

*au toti semnul $(-1)^m$.*

# 7. Constrangeri de Tip Inegalitate

Consideram acum probleme de forma

$$\begin{array}{ll} & \text{minimizeaza} \quad f(x) \\ \text{cu} & \mathbf{h}(x) = 0, \\ & \mathbf{g}(x) \leq 0, \end{array} \tag{36}$$

unde $f$, $\mathbf{h}$, $\mathbf{g}$ sunt de clasa $\mathcal{C}^1$ iar $\mathbf{g}$ este o functie $p$–dimensionala.

**Definitia 71.** *Fie $x^*$ un punct ce satisface constrangerile*

$$h(x^*) = 0, \qquad g(x^*) \leq 0$$

*si fie $J$ multimea indicilor $j$ pentru care $g_j(x^*) = 0$. Atunci $x^*$*

*este numit un punct regulat al constrangerilor daca vectorii gradient* $\nabla h_i(x^*)$, $\nabla g_j(x^*)$, $1 \leq i \leq m, j \in J$ *sunt liniar independenti.*

**Teorema 72. [Kuhn–Tucker]** *Fie* $x^*$ *un punct de minim relativ pentru problema (36) si presupunem ca* $x^*$ *este un punct regulat pentru constrangeri. Atunci exista un vector* $\lambda \in \mathbb{R}^n$ *si un vector* $\mu \in \mathbb{R}^p$, *cu* $\mu \geq 0$, *astfel incat*

$$\begin{aligned} \nabla f(x^*) + \lambda^T \nabla h(x^*) + \mu^T \nabla g(x^*) &= 0, \\ \mu^T g(x^*) &= 0. \end{aligned} \tag{37}$$

**Teorema 73. [Conditii Necesare de Ordinul Doi]** *Presupunem ca functiile* $f, \mathbf{h}, \mathbf{g} \in \mathcal{C}^2$, $x^*$ *este un minim local pentru (36) si este punct regulat al respectivelor constrangeri. Atunci exista* $\lambda \in \mathbb{R}^m$, $\mu \in \mathbb{R}^p$,

*$\mu \geq 0$ a. i. (37) este adevarata si astfel incat*

$$L(x^*) = F(x^*) + \lambda^T H(x^*) + \mu^T G(x^*)$$

*este pozitiv semidefinita pe subspatiul tangent al constrangerilor active in $x^*$.*

**Teorema 74. [Conditii Suficiente de Ordinul Doi]** *Fie $f, \mathbf{h}, \mathbf{g} \in \mathcal{C}^2$. $x^*$ satisfacand constrangerile (36) este un punct de minim relativ in sens strict daca au loc conditiile :*

- *Exista $\lambda \in \mathbb{R}^m$ si $\mu \in \mathbb{R}^p$ a.i.*

$$\begin{aligned} \mu &\geq 0, \\ \mu^T g(x^*) &= 0, \\ \nabla f(x^*) + \lambda^T \nabla h(x^*) + \mu^T \nabla g(x^*) &= 0. \end{aligned}$$

- *Hessianul*

$$L(x^*) = F(x^*) + \lambda^T H(x^*) + \mu^T G(x^*)$$

*este pozitiv definit pe subspatiul*

$$M^{'} = \{y : \nabla \mathbf{h}(x^*)y = 0, \quad \nabla g_j(x^*)y = 0, \forall j \in J\}$$

*unde*

$$J = \{j : g_j(x^*) = 0, \mu_j > 0\}.$$

# Capitolul 7: MINIMIZARE CU CONSTRANGERI – PRINCIPII GENERALE ALE ALGORITMILOR

Introducere

Metode Primale

Metoda de Penalizare si Bariera

Metode Duale si de Plan Secant

Metode de Tip Lagrange

# 1. Introducere

In acest capitol vom da o descriere foarte sumara a principiilor ce guverneaza algoritmii in cazul optimizarilor cu constrangeri.

In general, problema de optimizare cu constrangeri este intotdeauna redusa la una fara constrangeri, cea din urma rezolvandu-se printr-o modificare oarecare a unuia dintre algoritmii cunoscuti. Cele patru clase de metode amintite mai sus corespund unei scheme de clasificare ce are in vedere multimea pe care se face efectiv cautarea minimului.

Consideram o problema de minimizare a unei functii de $n$ variabile si avand $m$ constrangeri. Exista diverse metode de rezolvare a acestei probleme ce lucreaza in spatii de dimensiune $n-m$, $n$, $m$ sau $n+m$.

Aceste patru clase de metode isi au fundamentul in diverse parti ale teoriei prezentate in capitolul anterior.

In orice caz, exista puternice interconexiuni intre diversele metode atat in forma finala in care se implementeaza algoritmul cat si in performantele specifice. De exemplu, rata de convergenta a algoritmilor cei mai mai buni dpdv practic este determinata de structura Hessianului Lagrangianului intocmai precum structura Hessianului functiei obiectiv determina rata de convergenta in multe metode pentru probleme fara constrangeri. Pe scurt, cu toate ca diversii algoritmi difera substantial in motivatie, in final sunt guvernati de un set comun de principii.

# 2. Metode Primale

Consideram problema

$$\begin{array}{ll} & \min \ \ f(x), \\ \text{cu} & h(x) = 0, \\ & g(x) \leq 0, \end{array}$$

unde $x$ este de dimensiune $n$, in timp ce $f$, $g$ si $h$ au dimensiunile egale cu $1$, $p$ si $m$.

**Metodele primale** rezolva problema prin minimizarea lui $f$ in regiunea din $\mathbb{R}^n$ definita de constrangeri. O metoda primala este o metoda de cautare care actioneaza direct pe functia originala cautand solutia optimala intr–o regiune fezabila. Fiecare punct curent este

fezabil iar valoarea functiei obiectiv descreste continuu. Pentru o problema cu $n$ variabile si avand $m$ constrangeri de tip egalitate, metodele primale actioneaza in spatiul fezabil care are dimensiune $n - m$.

Avantajele principale:

• Fiecare punct generat de metoda este fezabil. Prin urmare, punctul final (la care se termina cautarea) este fezabil si probabil aproape optimal reprezentand prin urmare o solutie acceptabila a problemei practice ce a motivat programarea neliniara;

• Cel mai adesea, se poate garanta ca daca metoda genereaza un sir convergent atunci punctul limita al sirului este cel putin un minim local cu constrangeri. Mai precis, convergenta globala este

cel mai adesea satisfacatoare pentru aceste metode;

• Cele mai multe dintre metodele din aceasta clasa nu se bazeaza pe structura particulara a problemei, precum ar fi convexitatea, si prin urmare sunt aplicabile problemelor generale de programare neliniara.

Dezavantaje majore:

• Aceste metode necesita o procedura preliminara pentru a obtine un punct fezabil initial;

• Apar dificultati majore de calcul intrucat trebuie sa ramanem permanent in regiunea fezabila;

• Anumite metode pot sa nu convearga daca nu sunt luate anumite precautii speciale.

Concluzii: Ratele de convergenta sunt relativ bune si pentru constrangeri liniare metodele primale sunt dintre cele mai eficiente. Mai mult, sunt simple si general aplicabile.

Exista doua clase de metode primale:

- Metode de directii fezabile
- Metode de constrangeri active

# Metode de directii fezabile

Ideea este ca respectiva cautare sa fie facuta intr-o regiune fezabila de forma

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

unde $d_k$ este o directie si $\alpha_k$ un scalar nenegativ. Scalarul se alege a.i. sa minimizeze functia obiectiv $f$ cu restrictia ca punctul $x_{k+1}$ si segmentul de dreapta ce uneste $x_k$ si $x_{k+1}$ sa fie fezabile. Prin urmare, un intreg segment $x_k + \alpha d_k$, $\alpha > 0$, trebuie sa fie continut in regiunea fezabila. Deci fiecare pas este o compunere de doi subpasi:

- Se alege o directie fezabila
- Se executa o cautare unidimensionala cu constrangeri

Dezavantaje:

• Pentru probleme generale este posibil sa nu existe nici o directie fezabila. Prin urmare trebuie sau sa relaxam cerinta de fezabilitate permitand ca punctele de cautare sa poata devia usor de la suprafata determinata de constrangeri sau sa introducem deplasari de–alungul unor curbe pe suprafata determinata de constrangeri;
• In forma lor pura majoritatea metodelor de directii fezabile nu sunt global convergente;

# Metode de Constrangeri Active

Ideea centrala a acestor metode este sa impartim constrangerile de tip inegalitate in doua grupe: unele tratate drept constrangeri active iar altele care sunt tratate drept inactive. Cele inactive sunt ignorate !

La fiecare pas al unui algoritm se defineste o multime de constrangeri numite multimea curenta (sau de lucru) si care sunt tratate drept active. Multimea de lucru este intotdeauna o submultime a constrangerilor active in punctul curent. Prin urmare punctul curent este fezabil pentru multimea de lucru. Algoritmul se deplaseaza pe suprafata definita de multimea de lucru spre un punct mai apropiat de solutie. In noul punct multimea de lucru se poate schimba. Deci metoda constrangerilor active consta in urmatoarele etape:

• Determinarea unei multimi de lucru curente care este o submultime a constrangerilor active ;
• Deplasarea pe suprafata definita de multimea de lucru curenta spre un punct superior (calitativ);

Directia de deplasare este in general determinata de aproximarile de odinul intai si ordinul doi ale functiilor in punctul curent intr-un mod asemanator ca pentru cazul neconstrans.

# 3. Metode de Penalizare si Bariera

Acestea sunt proceduri care aproximeaza problemele de optimizare cu constrangeri prin probleme fara constrangeri. Pentru metodele de penalizare aproximarea este realizata prin adaugarea la functia obiectiv a unui termen ce are o valoare mare atunci cand constrangerile sunt violate. Pentru metodele de bariera se adauga un termen ce favorizeaza punctele interioare regiunii fezabile in raport cu cele de pe frontiera.

In aceste metode intervine un parametru $c$ care determina severitatea penalizarii sau barierei si indica gradul in care problema neconstransa aproximeaza problema originala constransa. Cand $c \rightarrow \infty$ aproximarea devine din ce in ce mai exacta si exista si

anumite functii de penalizare care dau solutii exacte pentru valori finite ale parametrului.

Pentru o problema cu $n$ variabile si $m$ constrangeri, metodele de penalizare si bariera actioneaza direct in spatiul $n$–dimensional al variabilelor.

Exista doua chestiuni fundamentale care trebuie considerate:

- Cat de bine aproximeaza problema neconstransa problema originala. Mai precis, trebuie vazut daca pe masura ce $c$ este crescut spre infinit solutia problemei neconstranse converge la solutia problemei originale constranse.

- Cum poate fi rezolvata o problema neconstransa atunci cand functia obiectiv contine un termen de penalizare sau bariera;

analizand acest fenomen rezulta ca pe masura ce $c$ este crescut (pentru a obtine o buna aproximare), structura corespunzatoare a problemei neconstranse devine progresiv mai nefavorabila incetinind prin urmare rata de convergenta. Prin urmare, trebuie sa deducem proceduri de accelerare pentru a evita aceasta convergenta lenta.

## Metode de Penalizare

Consideram problema

$$\begin{array}{ll} & \text{minimizeaza} \quad f(x) \\ \text{unde} & x \in S \end{array} \tag{38}$$

unde $f$ este continua in $\mathbb{R}^n$ si $S$ este o multime in $\mathbb{R}^n$. In majoritatea aplicatiilor $S$ este definita implicit printr-un numar de constrangeri functionale dar aici putem considera direct cazul general.

Ideea metodei de penalizare este sa inlocuim problema (38) cu o problema neconstransa de forma

$$\text{minimizeaza} \quad f(x) + cP(x) \tag{39}$$

unde $c$ este o constanta pozitiva si $P$ este o functie definita pe $\mathbb{R}^n$ care satisface: (i) $P$ este continua; (ii) $P(x) \geq 0, \forall x \in \mathbb{R}^n$; (iii) $P(x) = 0$ daca si numai daca $x \in S$.

Procedura pentru rezolvarea problemei (38) prin metoda de penalizare este urmatoarea: Fie $\{c_k\}, k = 1, 2, \ldots$ un sir care tinde la infinit astfel incat pentru fiecare $k$, $c_k \geq 0$, $c_{k+1} > c_k$. Definim functia

$$q(c, x) := f(x) + cP(x).$$

Pentru fiecare $k$ rezolvam problema

$$\text{minimizeaza} \quad q(c_k, x) \tag{40}$$

obtinand o solutie $x_k$. In general, presupunem ca pentru fiecare $k$ problema (40) are o solutie. Aceasta este adevarat in particular atunci cand $q(c, x)$ creste nemarginit pentru $|x| \longrightarrow \infty$.

## Metode de Bariera

Metodele de bariera sunt aplicabile problemelor de forma

$$\begin{array}{ll} & \text{minimizeaza} \quad f(x) \\ \text{unde} & x \in S \end{array} \tag{41}$$

unde $S$ are un interior nevid care este oricat de aproape de orice punct al lui $S$. Intuitiv, aceasta inseamna ca multimea are un interior

si este posibil sa ajungem la orice punct de frontiera cu puncte din interior. O astfel de multime se numeste robusta.

Metodele de bariera stabilesc o bariera pe frontiera multimii fezabile care impiedica procedura de cautare sa paraseasca multimea. O functie bariera este o functie $B$ definita pe interiorul lui $S$ astfel incat (i) $B$ este continua; (ii) $B(x) \geq 0$; (iii) $B(x) \to \infty$ pe masura ce $x$ se apropie de frontiera lui $S$.

Corespunzator problemei (41) consideram problema aproximativa

$$\begin{array}{rc} & \text{minimizeaza} \quad f(x) + \frac{1}{c}B(x) \\ \text{unde} & x \in \text{ interiorul lui } S \end{array} \tag{42}$$

unde $c$ este o constanta pozitiva. Aceasta problema este oarecum mai complicata decat problema originala dar poate fi rezolvata printr-o

metoda de cautare fara constrangeri. Pentru a gasi solutia pornim cu un punct initial interior si folosim ulterior metoda celei mai abrupte pante sau alta procedura de cautare iterativa pentru probleme neconstranse. Deorece valoarea functiei tinde la infinit in apropierea frontierei lui $S$, procedura de cautare va ramane automat in interiorul lui $S$ si nu mai trebuie sa tinem seama explicit de constrangere. Astfel, cu toate ca problema (42) este din punct de vedere formal o problema constransa, din punct de vedere procedural este neconstransa.

Metoda lucreaza in modul urmator: Fie $c_k$ un sir care tinde la infinit astfel incat pentru orice $k$, $k = 1, 2, \ldots$, $c_k \geq 0$, $c_{k+1} > c_k$. Definim functia

$$r(c, x) = f(x) + \frac{1}{c} B(x).$$

Pentru fiecare $k$ rezolvam problema

$$\begin{aligned} &\text{minimizeaza} \;\; r(c_k, x) \\ \text{unde} \quad &x \in \text{ interiorul lui } S \end{aligned}$$

si obtinem punctul $x_k$. Pentru metodele de penalizare si bariera avem urmatorul rezultat:

**Teorema 75.** *Orice punct limita al unui sir generat de metoda de penalizare sau bariera este o solutie a problemei originale (38).*

# 4. Metode Duale si de Plan Secant

Metodele Duale sunt bazate pe faptul ca multiplicatorii Lagrange sunt necunoscutele fundamentale asociate cu problema constransa; indata ce acesti multiplicatori sunt determinati, gasirea punctelor solutie este simpla (cel putin in anumite cazuri particulare). Tehnicile duale nu abordeaza problema originala constransa in mod direct ci abordeaza o asa numita problema duala ale carei necunoscute sunt multiplicatorii Lagrange ai primei probleme. Pentru o functie obiectiv cu $n$ variabile si $m$ constrangeri de tip egalitate metodele duale fac cautarea in spatiul $m$ dimensional al multiplicatorilor Lagrange.

Metodele de plan secant determina o serie de programe liniare din ce in ce mai bune a caror solutie converge la solutia problemei

originale. Aceste metode sunt adesea foarte usor de implementat dar teoria asociata lor nu este foarte bine dezvoltata si proprietatile lor de convergenta nu sunt foarte atractive.

# 5. Metode Lagrange

Aceste metode sunt bazate pe rezolvarea directa a conditiilor necesare de ordinul intai de tip Lagrange. Pentru probleme cu constrangeri de tip egalitate

$$\begin{array}{cc} & \text{minimizeaza} \quad f(x) \\ \text{cu} & h(x) = 0 \end{array}$$

unde $x$ este $n$–dimensional si $h(x)$ este $m$–dimensional, aceasta abordare conduce la rezolvarea unui sistem de ecuatii

$$\begin{array}{rcl} \nabla f(x) + \lambda^T \nabla h(x) & = & 0, \\ h(x) & = & 0, \end{array}$$

in necunoscutele $x$ si $\lambda$. Multimea conditiilor necesare este un sistem de $n + m$ ecuatii in $n + m$ necunoscute (componentele lui $x$ si $\lambda$). Aceste metode lucreaza intr-un spatiu $(n + m)$–dimensional.

# Partea a–II–a: Programare Liniara

## Capitolul 8: PROPRIETATI DE BAZA ALE PROGRAMARII LINIARE

Introducere

Exemple de Probleme de Programare Liniara

Teorema Fundamentala a Programarii Liniare

Relatii cu Convexitatea

Exercitii

# 1. Introducere

In acest capitol dam o descriere a principiilor ce guverneaza problemele de programare (optimizare) liniara cu constrangeri.

**Problema de programare liniara** : problema matematica in care functia obiectiv este liniara in necunoscute si constrangerile constau din egalitati liniare si inegalitati liniare (Totul este liniar !).

Forma concreta a inegalitatilor poate in general diferi dar, asa cum vom arata putin mai tarziu, orice program liniar se poate aduce in

urmatoarea forma standard:

$$\begin{array}{ll} \text{minimizati} & c_1x_1 + c_2x_2 + \cdots + c_nx_n \\ \text{atunci cand} & a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ & a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ & \vdots \qquad\qquad\qquad\qquad \vdots \\ & a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \\ \text{si} & x_1 \geq 0, x_2 \geq 0, \ldots, x_n \geq 0, \end{array} \tag{43}$$

unde $b_i$, $c_i$ si $a_{ij}$ sunt constante reale fixate si $x_j$ sunt numere reale ce trebuie determinate.

Fara a restrange generalitatea, **presupunem intotdeauna ca** $b_i \geq 0$ (printr-o eventuala multiplicare a fiecarei ecuatii cu -1 !).

Relatiile (43) se pot rescrie compact in forma

$$\begin{array}{lll} \text{minimizati} & c^T x & \\ \text{atunci cand} & Ax = b \quad \text{si} & x \geq 0 \end{array} \tag{44}$$

in care $x$ este un vector $n$–dimensional (coloana), $c^T$ este un vector $n$–dimensional (linie), $A$ este o matrice de dimensiune $m \times n$, $b$ este un vector $m$–dimensional (coloana) iar inegalitatea vectoriala $x \geq 0$ inseamna ca fiecare componenta in parte a lui $x$ este semipozitiva (mai mare sau egala cu zero).

Multe alte forme aparent diverse de programe liniare pot fi convertite in forma standard:

**Exemplul 76.** *Consideram problema*

$$
\begin{array}{ll}
\text{minimizati} & c_1x_1 + c_2x_2 + \cdots + c_nx_n \\
\text{atunci cand} & a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \leq b_1 \\
& a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \leq b_2 \\
& \vdots \qquad\qquad\qquad\qquad \vdots \\
& a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \leq b_m \\
\text{si} & x_1 \geq 0, x_2 \geq 0, \ldots, x_n \geq 0
\end{array}
$$

*in care multimea constrangerilor este determinata in totalitate de*

*inegalitati liniare. Problema se poate exprima alternativ sub forma*

$$
\begin{array}{ll}
\text{minimizati} & c_1x_1 + c_2x_2 + \cdots + c_nx_n \\
\text{atunci cand} & a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + y_1 = b_1 \\
& a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n + y_2 = b_2 \\
& \vdots \qquad\qquad\qquad\qquad \vdots \\
& a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n + y_m = b_m \\
\text{si} & x_1 \geq 0, x_2 \geq 0, \ldots, x_n \geq 0, \\
\text{si} & y_1 \geq 0, y_2 \geq 0, \ldots, y_m \geq 0.
\end{array}
$$

*Aceasta noua problema considerata in $n+m$ necunoscute $x_1, \ldots x_n$, $y_1, \ldots, y_m$ este in forma standard iar matricea de tip "$A$" ce descrie acum multimea de constrangeri de tip egalitate are forma speciala $\begin{bmatrix} A & I \end{bmatrix}$.*

**Exemplul 77.** *Daca inegalitatile liniare de la Exemplul 1 sunt*

*reversate (i.e. sunt de tipul*

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n \geq b_i),$$

*atunci aceasta este echivalenta cu*

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n - y_i = b_i, \quad \text{si} \quad y_i \geq 0.$$

*Din Exemplele 1 si 2 este clar ca prin multiplicare cu -1 si introducerea unor variabile suplimentare orice set de inegalitati liniare se poate converti la o forma standard constrangand corespunzator variabilele necunoscute (de tip $y_i$) sa fie semipozitive !*

**Exemplul 78.** *Fie un program liniar dat in forma standard cu exceptia faptului ca una sau mai multe variabile necunoscute nu trebuie sa fie semipozitive (ci arbitrare).*

*Problema se transforma intr–una standard prin metoda urmatoare. Sa presupunem ca in (43)* $x_1 \geq 0$ *nu apare si atunci* $x_1$ *este libera sa ia valori arbitrare (pozitive sau negative). Fie*

$$x_1 = u_1 - v_1 \tag{45}$$

*in care cerem ca* $u_1 \geq 0$ *si* $v_1 \geq 0$. *Substituind* $u_1 - v_1$ *in locul lui* $x_1$ *peste tot in (43) obtinem ca liniaritatea constrangerilor se pastreaza si acum toate cele* $n + 1$ *variabile* $u_1, v_1, x_2, \ldots, x_n$ *trebuie sa fie semipozitive (deci din nou un program standard).*

*Este usor de observat aparitia unei redundante introduse de aceasta tehnica (o constanta aditiva !).*

**Exemplul 79.** *Aceeasi problema – o alta metoda ! O alta abordare in rezolvarea problemei 3 cand* $x_1$ *nu este constrans ca semn este sa–l*

*eliminam pe $x_1$ impreuna cu una dintre ecuatiile de constrangere, de exemplu*

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n = b_i \tag{46}$$

*(singura cerinta este ca aceasta ecuatie sa aibe $a_{i1}$ – coeficientul lui $x_1$ – nenul).*

*Din aceasta ecuatie $x_1$ se poate exprima ca o combinatie liniara de celelalte variabile + o constanta, expresie ce se inlocuieste peste tot in (43). Se obtine o noua problema de acelasi tip cu cea originala in necunoscutele $x_2, x_3, \ldots, x_n$ iar ecuatia $i$ devine identic zero si deci se poate elimina.*

*Schema de lucru prevede rezolvarea acestui din urma program si obtinerea in final a lui $x_1$ pe baza ecuatiei eliminate (46).*

**Exemplul 80.** *Dam un exemplu de aplicare a tehnicii de mai sus.*

*Fie programul*

$$\begin{array}{ll} \text{minimizati} & x_1 + 3x_2 + 4x_3 \\ \\ \text{atunci cand} & x_1 + 2x_2 + x_3 = 5 \\ & 2x_1 + 3x_2 + x_3 = 6 \\ \text{si} & x_2 \geq 0, x_3 \geq 0. \end{array}$$

*Deoarece $x_1$ este liber, rezolvam prima constrangere obtinand*

$$x_1 = 5 - 2x_2 - x_3 \tag{47}$$

*care inlocuit in functia obiectiv si intr–a doua constrangere genereaza*

*problema echivalenta* **in forma standard**

$$\begin{array}{ll} \text{minimizati} & x_2 + 3x_3 + 5 \\ \\ \text{atunci cand} & x_2 + x_3 = 4 \\ \text{si} & x_2 \geq 0, x_3 \geq 0. \end{array}$$

*Dupa ce se rezolva aceasta problema (avand solutia* $x_2 = 4$, $x_3 = 0$*), valoarea lui* $x_1 = -3$ *se obtine din ecuatia (47).*

# 2. Exemple de Probleme de Programare Liniara

**Exemplul 81. [Problema dietei economice]** *Se pune problema sa determinam o dieta cat mai economica care sa acopere insa in totalitate substantele nutritive necesare organismului uman – problema tipica de alocat resurse pentru dieteticianul unei mari armate ! Ipotezele sunt: exista pe piata $n$ alimente care se vand la pretul $c_i$ per bucata, si exista $m$ ingrediente nutritionale de baza pe care fiecare om trebuie sa le consume intr-o cantitate de minim $b_j$ unitati (din elementul $j$). In plus, fiecare aliment contine $a_{ji}$ unitati din elementul nutritional $j$.*
*Fie $x_i$ numarul de unitati din alimentul $i$ din dieta. Problema este atunci sa selectam $x_i$ care minimizeaza costul total $\sum_{i=1}^{n} c_i x_i$ atunci*

*cand avem constrangerile nutritionale*

$$
\begin{aligned}
& a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \geq b_1 \\
& a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \geq b_2 \\
& \vdots \qquad\qquad\qquad\qquad \vdots \\
& a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \geq b_m \\
& \text{si} \quad \text{constrangerile } x_1 \geq 0, x_2 \geq 0, \ldots, x_n \geq 0,
\end{aligned}
$$

*asupra cantitatilor de alimente.*

**Exemplul 82. [Problema transportului]** *Trebuie livrate cantitatile* $a_1, a_2, \ldots, a_m$ *dintr-un produs depozitat in* $m$ *locatii astfel incat sa ajunga in final cantitatile* $b_1, b_2, \ldots, b_n$ *la fiecare dintre cele* $n$ *destinatii finale. Costul de a transporta o unitate de produs de la locatia initiala* $i$ *la cea finala* $j$ *este de* $c_{ij}$*. Se cere obtinerea cantitatilor* $x_{ij}$ *ce trebuie transportate de* $i$ *la* $j$ *astfel incat sa se minimizeze costul*

*transportului. Problema de minimizat se formuleaza precum urmeaza*

$$\begin{array}{ll} \text{minimizati} & \sum_{ij} c_{ij}x_{ij} \\ \text{atunci cand} & \sum_j x_{ij} = a_i, \quad pentru \quad i = 1, 2, \ldots, m, \\ & \sum_i x_{ij} = b_j, \quad pentru \quad j = 1, 2, \ldots, n, \\ & x_{ij} \geq 0, \quad \text{pentru} \quad i = 1, 2, \ldots, m, \qquad j = 1, 2, \ldots, \end{array}$$

*Pentru consistenta problemei trebuie ca* $\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j$ *(cantitatea trimisa este egala cu cantitatea receptionata). Problema de transport este deci o problema de programare liniara in* $mn$ *variabile. Ecuatille ce descriu constrangerile se pot pune in forma uzuala rezultand o matrice de dimensiune* $(m+n) \times mn$ *avand drept elemente 0 si 1 !*

**Exemplul 83. [Problema de productie]** *Presupunem ca avem o unitate industriala ce efectueaza* $n$ *activitati productive si fiecare*

*activitate productiva consta in obtinerea a diverse cantitati din $m$ produse distincte. Fiecare activitate productiva se poate efectua la orice nivel $x_i \geq 0$ dar cand este operata la nivel unitar activitatea $i$ costa $c_i$ si produce $a_{ji}$ unitati din produsul $j$. Presupunem liniaritatea unitatii industriale. Nr. de produse finite ce trebuie obtinut este specificat prin $b_1, b_2, \ldots, b_m$ si le dorim produse la un cost minim. Programul liniar se sintetizeaza prin (43) !*

**Exemplul 84. [Problema de depozitare]** *– De Discutat la Seminar –*

# 3. Teorema Fundamentala a Programarii Liniare

**Ipoteza fundamentala (implicita)**: Constrangerile de tip egalitate

$$Ax = b, \quad A \in \mathbb{R}^{m \times n}, \quad b \in \mathbb{R}^m \tag{48}$$

au **matricea $A$ surjectiva ( $m$ linii liniar independente $(\leq n)$)**.

Ratiunea introducerii acestei ipoteze sta in alte doua presupuneri naturale:

- Sunt mai multe variabile decat ecuatii (avem unde optimiza !!!)
- Ecuatiile sunt liniar independente (problema are solutie si nu are ecuatii redundante !!! )

**Definitia 85.** *Fie $B$ orice $m \times m$ matrice nesingulara (numita si matrice baza) formata din coloane ale lui $A$, fie $x_b$ unica solutie a ecuatiei $Bx_b = b$ si $x \in \mathbb{R}^n$ obtinuta extinzand $x_b$ cu zerouri corespunzatoare componentelor ce nu sunt asociate coloanelor lui $B$. Atunci $x$ se numeste* **solutie fundamentala a ecuatiei (48)** *in raport cu baza $B$ iar componentele sale asociate cu coloanele lui $B$ sunt numite* **variabile fundamentale**.

**Observatia 86.** *Sub ipoteza fundamentala facuta* **sistemul (48) are intotdeauna o solutie si cel putin o solutie fundamentala** *! Solutiile fundamentale nu sunt neaparat nenule !*

**Definitia 87.** *Daca o solutie fundamentala are macar una dintre variabilele fundamentale nule atunci solutia se numeste degenerata.*

**Observatia 88.** *Intr-o solutie fundamentala nedegenerata se pot*

*deosebi imediat variabilele fundamentale de celelalte (pozitiv vs. zero) pe cand in cele degenerate variabilele fundamentale nule se pot interschimba cu cele nefundamentale !*

Consideram in continuare sistemul complet de constrangeri al unui program liniar

$$\begin{aligned} Ax &= b, \\ x &\geq 0. \end{aligned} \tag{49}$$

**Definitia 89.** *Vectorul $x$ satisfacand (49) se numeste fezabil.*

Deci putem avea **solutii fundamentale fezabile si nefezabile**, si **o solutie fundamentala fezabila poate fi degenerata sau nu**.

Teorema centrala a programarii liniare pune in evidenta **caracterul primordial jucat de solutiile fundamentale fezabile**.

Esentialmente, teorema afirma ca este necesar sa consideram doar solutii fundamentale fezabile atunci cand cautam optimul unui program liniar pentru ca valoarea optima este intotdeauna obtinuta intr-o astfel de solutie !

Corespunzator unui program in forma standard

$$\begin{array}{ll} \text{minimizati} & c^T x \\ \text{atunci cand} & Ax = b \quad \text{si} \quad x \geq 0 \end{array} \tag{50}$$

o solutie fezabila a constrangerilor ce atinge minimul functiei obiectiv cu aceste constrangeri se numeste **solutie fezabila optimala**. Daca aceasta solutie este in plus fundamentala atunci atunci se numeste **solutie fundamentala fezabila optimala !**

**Teorema 90. [Teorema fundamentala a programarii liniare]** *Fie*

*$A$ o $m \times n$ matrice surjectiva si fie (50) programul liniar asociat.*

- *Daca exista o solutie fezabila atunci exista si o solutie fundamentala fezabila.*
- *Daca exista o solutie fezabila optimala atunci exista si o solutie fundamentala fezabila optimala.*

**Observatia 91.** Teorema reduce sarcina complexa de a rezolva o problema de programare linara la aceea de a cauta in submultimea solutiilor fundamentale fezabile care este considerabil mai ieftina decat sarcina originala intrucat avem cel mult

$$C_n^m = \frac{n!}{m!(n-m)!}$$

solutii fundamentale (corespunzatoare modalitatilor diverse de a alege $m$ coloane din $n$ coloane). Prin urmare sarcina calculatorie se termina intr-un numar finit de pasi (chiar daca mare) prin intermediul unui algoritm foarte simplu dar extrem de ineficient !

O metoda eficienta de calcul – numita simplex – este prezenta in capitolul urmator.

# 4. Relatii cu Convexitatea

Principalele rezultate prezentate anterior au interpretari deosebit de interesante in termenii teoriei multimilor convexe si proprietatilor asociate. Principala legatura intre proprietatile algebrice si cele geometrice consta in relatia formala dintre solutiile fundamentale fezabile ale inegalitatilor liniare si punctele extreme ale varietatilor liniare (politop).

**Definitia 92.** *Un punct $x$ apartinand unei multimi convexe $C$ se numeste* **punct extrem** *al lui $C$ daca nu exista doua puncte distincte $x_1$ si $x_2$ in $C$ a.i.*

$$x = \alpha x_1 + (1 - \alpha)x_2$$

*pentru un $\alpha$, $0 < \alpha < 1$.*

**Teorema 93. [Echivalenta punctelor extreme si a solutiilor fundam**
*Fie $A$ o $m \times n$ matrice surjectiva si $b \in \mathbb{R}^m$. Fie $K$ politopul convex constand din toti vectorii $x \in \mathbb{R}^n$ satisfacand*

$$\begin{aligned} Ax &= b, \\ x &\geq 0. \end{aligned} \tag{51}$$

*Atunci $x$* **este un punct extrem al lui $K$ daca si numai daca $x$ este solutie fundamentala fezabila a lui (51)**.

Aceasta corespondenta ne permite sa demonstram cateva proprietati geometrice ale unui politop convex definit de constrangerile unui program liniar.

**Corolarul 94.** *Daca politopul corespunzator lui (51) este nevid atunci are cel putin un punct extrem.*

**Corolarul 95.** *Daca o problema de programare liniara are o solutie finita optimala atunci are si o solutie finita optimala care este un punct extrem al multimii constrangerilor.*

**Corolarul 96.** *Multimea constransa $K$ definita de (51) are cel mult un numar finit de puncte extreme.*

In final, prezentam un caz special ce este caracteristic problemelor de programare liniara bine formulate: $K$ este nevid si marginit.

**Corolarul 97.** *Daca politopul $K$ definit de (51) este marginit, atunci $K$ este un poliedru convex, i.e., $K$ consta din puncte ce sunt combinatii convexe ale unui numar finit de puncte.*

**Exemplul 98.** *Consideram multimea constransa in* $\mathbb{R}^3$ *definita de*

$$x_1 + x_2 + x_3 = 1,$$
$$x_1 \geq 0, x_2 \geq 0, x_3 \geq 0$$

*si ilustrata in Figura 1.*

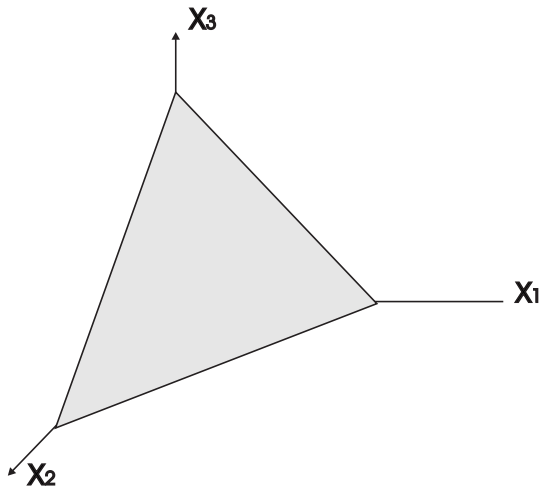



*Figura 1: Multimea Fezabila ptr. Exemplul 98*

*Aceasta multime are trei puncte extreme, corespunzand celor*

$C_3^2 = 3$ *solutii fundamentale ale lui* $x_1 + x_2 + x_3 = 1$.

**Exemplul 99.** *Consideram multimea constransa in* $\mathbb{R}^3$ *definita de*

$$\begin{aligned} x_1 + x_2 + x_3 &= 1, \\ 2x_1 + 3x_2 &= 1, \\ x_1 \geq 0, x_2 \geq 0, x_3 &\geq 0 \end{aligned}$$

*si ilustrata in Figura 2.*

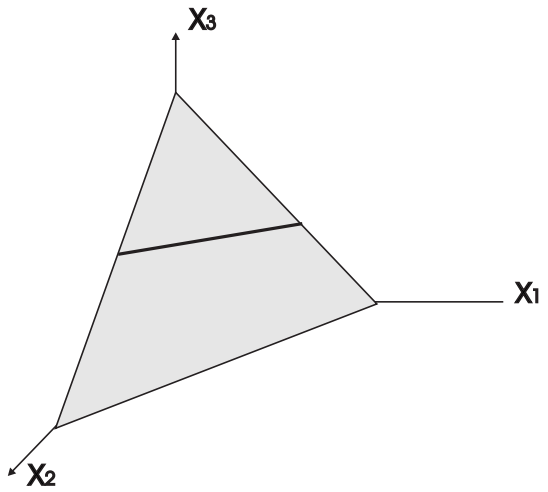



*Figura 2: Multimea Fezabila ptr. Exemplul 99*

*Aceasta multime are doua puncte extreme corespunzand celor doua solutii fundamentale fezabile. Observati deasemenea ca sistemul de ecuatii are trei solutii fundamentale*

$$(2, -1, 0), \qquad (\frac{1}{2}, 0, \frac{1}{2}), \qquad (0, \frac{1}{3}, \frac{2}{3}),$$

*din care insa prima nu este fezabila !*

**Exemplul 100.** *Consideram multimea constransa in* $\mathbb{R}^2$ *definita de*

$$x_1 + \frac{8}{3}x_2 \leq 4,$$
$$x_1 + x_2 \leq 2,$$
$$2x_1 \leq 3,$$
$$x_1 \geq 0, x_2 \geq 0$$

*si ilustrata in Figura 3.*

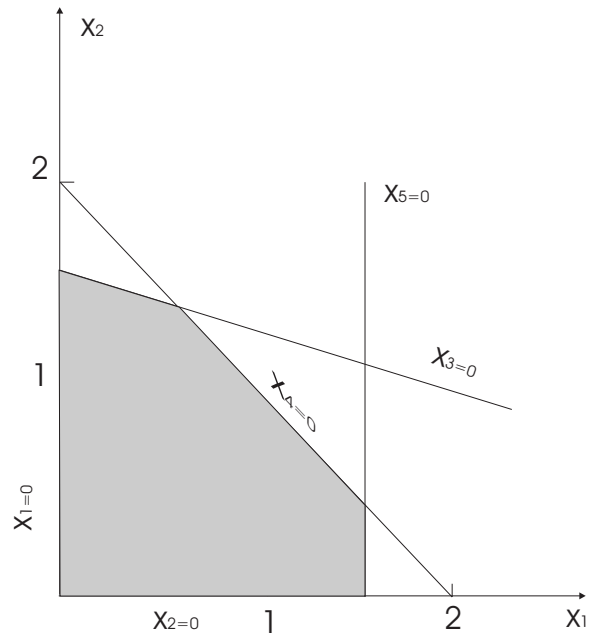



*Figura 3: Multimea Fezabila ptr Exemplul 100*

*Prin inspectie vizuala observam ca aceasta multime are 5 puncte extreme. Pentru a compara acest exemplu cu rezultatele generale obtinute pana acum aplicam procedura de transformare introdusa in Exemplul 1 din Sectiunea 1 obtinand in final multimea echivalenta in*

$\mathbb{R}^5$

$$x_1 + \frac{8}{3}x_2 + x_3 = 4,$$
$$x_1 + x_2 + x_4 = 2,$$
$$2x_1 + x_5 = 3,$$
$$x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0, x_5 \geq 0.$$

*Solutiile fundamentale pentru acest sistem se obtin punand oricare doua variabile egale cu zero si rezolvand pentru obtinerea celorlalte trei, in total un numar de $C_5^3 = 10$ combinatii. Dintre acestea doar 5 sunt fezabile corespunzand celor 5 colturi ale poligonului din Figura 3. Alternativ, laturile poligonului corespund unei variabile egale cu zero iar colturile (punctele extreme) sunt punctele in care doua variabile sunt egale cu zero.*

**Exemplul 101.** *Ultimul exemplu indica ca si atunci cand nu sunt exprimate in forma standard, punctele extreme ale multimii definite*

*de constrangerile unui program liniar corespunde posibilelor puncte solutie. Ilustram acest fapt pe constrangerile din exemplul precedent avand functia de minimizat*

$$-2x_1 - x_2.$$

*Multimea punctelor satisfacand* $-2x_1 - x_2 = z$ *pentru* $z$ *fixat este o dreapta. Cand* $z$ *variaza se obtin diverse drepte paralele asa cum este prezentat in Figura 4.*

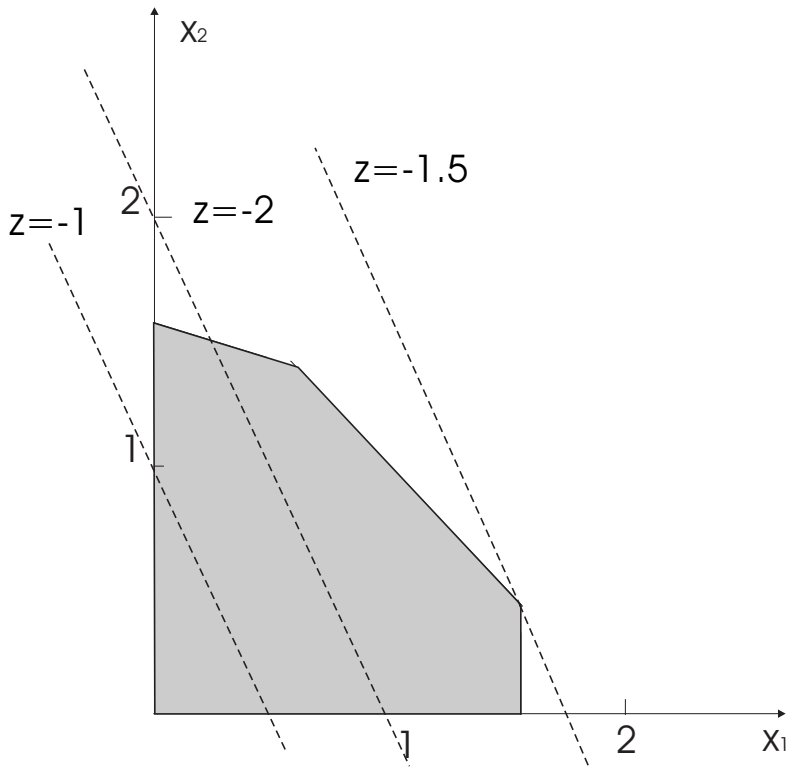



*Figura 4: Solutia extrema*

*Valoarea optimala a problemei de programare liniara este cea mai mica valoare a lui $z$ pentru care dreapta respectiva are un punct in comun cu multimea fezabila. Este clar, macar in doua dimensiuni, ca*

*punctele solutie vor include intotdeauna un punct extrem ! Observati din figura ca aceasta se intampla in punctul* $(\frac{3}{2}, \frac{1}{2})$ *cu* $z = -3\frac{1}{2}$ *!*

# Capitolul 9: METODA SIMPLEX

Ideea de baza a metodei simplex este ca pornind de la o solutie fundamentala fezabila sa gasim o noua solutie fundamentala fezabila in care functia obiectiv sa descreasca. Procesul se continua pana cand se obtine minimul.

Teorema programarii liniare din capitolul precedent ne asigura ca parcurgand numai aceste puncte in $\mathbb{R}^n$ (solutii fundamentale fezabile) putem obtine in final solutia optimala dorita. Metoda simplex arata ca in plus aceasta cautare se poate face intr-o maniera eficienta dpdv numeric.

In primele cinci sectiuni ale capitolului dam o descriere detaliata a metodei simplex pornind de la o examinare atenta a sistemului

de ecuatii liniare ce defineste constrangerile. Aceasta abordare desi necesita matematica extrem de simpla nu se poate exprima prea usor intr-o forma matriceala compacta. In ultimele sectiuni ale capitolului, schimbam punctul de vedere si ne concentram asupra unei viziuni matriceale ce conduce la o reprezentare compacta a metodei simplex cat si la metode alternative de implementare.

Pivoti

Puncte Extreme Adiacente

Determinarea unei Solutii Minime Fezabile

Algoritmul

Variabile Artificiale

Forma Matriceala a Metodei Simplex

Metoda Simplex Revizuita

Metoda Simplex via Descompunerea LU

Concluzii Finale

# 1. Pivoti

Reamintim in contextul programarii liniare operatiile de pivotare asupra unui sistem de ecuatii liniare. Avem doua tipuri de pivotari cu semnificatii duale:

- Pivotare pe linii;
- Pivotare pe coloane.

Fie sistemul $Ax = b$, cu $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$, $m \leq n$, avand forma explicita

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots \qquad\qquad & \quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned} \tag{52}$$

Sub ipotezele uzuale (i.e. $A$ surjectiva) si cu $m < n$ sistemul (52) nu are o solutie unica ci o varietate liniara de solutii. Presupunand pentru

comoditate ca primele $m$ coloane ale lui $A$ sunt liniar independente, sistemul (52) poate fi redus prin scheme clasice de eliminare Gaussiana pe linii (multiplicarea unei linii cu o constanta nenula si adunarea de combinatii liniare de linii) la urmatoarea **forma canonica**:

$$\begin{aligned}
x_1 \quad & + \widetilde{a}_{1,m+1}x_{m+1} + \widetilde{a}_{1,m+2}x_{m+2} + \cdots + \widetilde{a}_{1,n}x_n = \widetilde{b}_1, \\
x_2 \quad & + \widetilde{a}_{2,m+1}x_{m+1} + \widetilde{a}_{2,m+2}x_{m+2} + \cdots + \widetilde{a}_{2,n}x_n = \widetilde{b}_2, \\
& \vdots \qquad\qquad\qquad\qquad \vdots \\
x_m \quad & + \widetilde{a}_{m,m+1}x_{m+1} + \widetilde{a}_{m,m+2}x_{m+2} + \cdots + \widetilde{a}_{m,n}x_n = \widetilde{b}_m.
\end{aligned} \tag{53}$$

In aceasta reprezentare echivalenta **variabilele** $x_1, \ldots, x_m$ sunt numite **fundamentale** iar celelalte **nefundamentale**. Solutia fundamentala corespunzatoare

$$x_1 = \widetilde{b}_1, \quad x_2 = \widetilde{b}_2, \quad \ldots, \quad x_m = \widetilde{b}_m, \quad x_{m+1} = 0, \ldots, \quad x_n = 0.$$

Extindem definitia formei canonice a unui sistem astfel: un sistem este in f.c. daca dintre cele $n$ variabile exact $m$ sunt fundamentale si au proprietatile: • fiecare variabila fundamentala apare intr-o singura ecuatie cu coeficientul 1; • nu apar doua variabile fundamentale in aceeasi ecuatie (echivalent cu a spune ca sistemul dupa o anumita reordonare a variabilelor si ecuatiilor are forma (53)).

Din economie de notatie, in programarea liniara se foloseste urmatoarea reprezentare a coeficientilor sistemului (53) sub forma

de tabel:

$$\begin{array}{cccccccc}
1 & 0 & \cdots & 0 & \widetilde{a}_{1,m+1} & \widetilde{a}_{1,m+2} & \cdots & \widetilde{a}_{1,n} & \widetilde{b}_1 \\
0 & 1 & \cdots & 0 & \widetilde{a}_{2,m+1} & \widetilde{a}_{2,m+2} & \cdots & \widetilde{a}_{2,n} & \widetilde{b}_2 \\
0 & 0 & \cdots & 0 & \cdot & \cdot & \cdot & \cdot & \cdot \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & 1 & \widetilde{a}_{m,m+1} & \widetilde{a}_{m,m+2} & \cdots & \widetilde{a}_{mn} & \widetilde{b}_m
\end{array} \tag{54}$$

Prin pivotare se poate inlocui o variabila fundamentala cu una nefundamentala si reciproc. Procedura se poate realiza pe linii sau coloane.

**Pivotare pe linii:** Foloseste reprezentarea sistemului $Ax = b$ sub

forma ecuatiilor

$$\begin{aligned} a^1 x &= b_1 \\ a^2 x &= b_2 \\ &\vdots \\ a^m x &= b_m \end{aligned}$$

unde $a^i$ este linia $i$ a matricii $A$.

Sa presupunem ca in sistemul canonic (53) vrem sa inlocuim variabila fundamentala $x_p$, $1 \leq p \leq m$ cu variabila nefundamantala $x_q$ (este posibil numai daca $\widetilde{a}_{pq} \neq 0$). Acest lucru se poate realiza prin impartirea liniei $p$ la $\widetilde{a}_{pq}$ pentru a obtine un coeficient unitar pentru $x_q$ in aceasta ecuatie, si prin scaderea liniei $p$ multiplicate cu un coeficient potrivit din toate celelalte linii pentru a obtine un coeficient zero pentru $x_q$ (in respectivele ecuatii). Fie $\widehat{a}_{ij}$ coeficientii

noului sistem obtinut. Avem:

$$\begin{cases} \widehat{a}_{ij} = \widetilde{a}_{ij} - \frac{\widetilde{a}_{pi}}{\widetilde{a}_{pq}}\widetilde{a}_{iq}, & i \neq p, \\ \widehat{a}_{pj} = \frac{\widetilde{a}_{pj}}{\widetilde{a}_{pq}}. \end{cases} \tag{55}$$

**Ecuatiile** de mai sus se numesc **de pivotare** iar elementul $\widetilde{a}_{pq}$ este **pivotul**.

**Pivotare pe coloane:** Foloseste reprezentarea sistemului $Ax = b$ in ideea ca $b$ este o combinatie liniara de coloane ale lui $A$ de tipul

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b,$$

in care $a_i$ este coloana $i$ a matricii $A$.

Fie sistemul canonic (54) cu primele $m$ coloane formand o baza.

Coeficientii unei coloane din tabel reprezinta coeficientii combinatiei liniare a acestor vectori baza care genereaza respectivul vector coloana, i.e.,

$$a_j = \sum_{k=1}^{m} \widetilde{a_{kj}} a_k \tag{56}$$

Tabelul se mai poate reinterpreta ca dand expresia vectorului $a_j$ in termenii vectorilor baza $a_1, a_2, \cdots, a_m$ (similar pentru $\widetilde{b}$).

Sa presupunem ca vrem sa inlocuim unul dintre vectorii din baza $a_p$, $1 \leq p \leq m$ cu alt vector $a_q$. Daca noii vectori $a_1, \ldots, a_m$ sunt liniar independenti, ei constituie o baza in functie de care orice alt vector poate fi reprezentat (este posibil numai daca $\widetilde{a}_{pq} \neq 0$).

Operatiile de actualizare a tabloului pleaca de la reprezentarea

$$a_q = \sum_{i=1, i\neq p}^{m} \widetilde{a}_{iq} a_i + \widetilde{a}_{pq} a_p$$

din care rezulta $a_p$,

$$a_p = \frac{1}{\widetilde{a}_{pq}} a_q - \sum_{i=1, i\neq p} \frac{\widetilde{a}_{iq}}{\widetilde{a}_{pq}} a_i. \tag{57}$$

Introducand (57) in (56) obtinem

$$a_j = \sum_{i=1, i\neq p}^{m} \left( \widetilde{a}_{ij} - \frac{\widetilde{a}_{iq}}{\widetilde{a}_{pq}} \widetilde{a}_{pj} \right) a_i + \frac{\widetilde{a}_{pj}}{\widetilde{a}_{pq}} a_q. \tag{58}$$

Daca $\widehat{a}_{ij}$ sunt coeficientii noului sistem atunci obtinem imediat expresii identice ca in cazul pivotarii pe linii (55).

**Observatia 102.** *Daca un sistem de ecuatii nu este de la inceput in forma canonica el se poate aduce in forma canonica concatenand $m$ vectori unitari tabloului si, plecand de la acesti vectori ca baza, inlocuind fiecare dintre ei prin pivotare cu coloane de-ale lui $A$.*

# 2. Puncte Extreme Adiacente

Solutia unui program liniar se gaseste in multimea solutiilor fundamentale fezabile ale sistemului $Ax = b,\ x \geq 0$ (Teorema programarii liniare) iar prin pivotare putem sa ne deplasam dintr-o solutie fundamentala in alta. In aceasta sectiune specializam cautarea a.i. orice noua solutie fundamentala gasita prin pivotare sa fie fezabila !

**Concluzia de baza**: Cu toate ca nu este in general posibil sa specificam perechile de variabile ce sunt interschimbate (pastrand evident pozitivitatea lui $x$), este posibil sa **specificam care variabila nefundamentala va deveni fundamentala** si apoi **sa determinam** cu care variabila fundamentala o inlocuim !

**Ipoteza de Nedegenerare:** Orice solutie fundamentala fezabila a lui $Ax = b$, $x \geq 0$, este o solutie nedegenerata!

Ipoteza simplifica mult o suma de argumentatii ale metodei simplex. In cazul in care nu este indeplinita poate conduce la esuarea metodei. Ipoteza este facuta pentru a simplifica expunerea si intotdeauna se poate asigura extensia la cazul general prin anumite modificari simple ale metodei simplex.

## Determinarea vectorului care paraseste baza

Presupunem ca avem solutia fundamentala fezabila $x^T = \begin{bmatrix} x_1 & x_2 & \dots & x_m & 0 & \dots & 0 \end{bmatrix}$ sau, echivalent, reprezentarea

$$b = \sum_{k=1}^{m} x_k a_k. \tag{59}$$

Din ipoteza de nedegenerare avem $x_i > 0$, $i = 1, \dots, m$. Presupunem ca vrem sa introducem in baza vectorul $a_q$, $q > m$, care are reprezentarea in vechea baza

$$a_q = \sum_{k=1}^{m} \widetilde{a}_{kq} a_k. \tag{60}$$

Multiplicand (60) cu $\epsilon \geq 0$ si extragand-o din (59) obtinem

$$\sum_{k=1}^{m}(x_k - \epsilon\widetilde{a}_{kq})a_k + \epsilon a_q = b. \tag{61}$$

Aceasta relatie arata ca pentru orice $\epsilon \geq 0$ $b$ este o combinatie de cel mult $m+1$ vectori. Pentru $\epsilon = 0$ obtinem vechea solutie. Cand $\epsilon$ creste de la zero in sus rezulta ca respectivul coeficient al lui $a_q$ creste, si este clar ca pentru $\epsilon$ nu prea mare **(61) da o solutie fezabila dar nefundamentala**. Coeficientii celorlalti vectori vor creste sau vor scadea odata cu cresterea lui $\epsilon$ pana cand unul (sau mai multi) se vor anula. Acest lucru se intampla ori de cate ori cel putin un coeficient $\widetilde{a}_{kq} > 0$.

In acest caz punem

$$\epsilon = \min_k \{\frac{x_k}{\widetilde{a}_{kq}} : \widetilde{a}_{kq} > 0\} \tag{62}$$

si am obtinut o noua solutie fezabila cu vectorul $a_q$ inlocuit de $a_p$, unde $p$ este indexul ce minimizeaza (62). Daca minimul (62) este obtinut simultan pentru mai multi indici atunci noua solutie este degenerata si oricare vector avand componenta corespunzatoare nula se poate considera ca parasind baza !

Daca nici unul dintre $\widetilde{a}_{kq}$ nu este pozitiv, atunci toti coeficientii lui (61) cresc (sau raman constanti) cand il crestem pe $\epsilon$ si nu se poate obtine nici o noua solutie fundamentala ! In aceasta situatie observam ca sistemul $Ax = b$, $x \geq 0$ are solutii fezabile cu coeficienti oricat de mari de unde rezulta ca multimea $K$ a solutiilor fezabile

este nemarginita !

**Pe scurt:** Dandu-se o solutie fundamentala fezabila si un vector arbitrar $a_q$ sunt posibile doua situatii:

• Exista o noua solutie fundamentala fezabila avandu-l pe $a_q$ in baza (in locul unuia dintre vectorii originali);

• Multimea solutiilor fezabile este nemarginita !

## Ilustrarea operatiilor in tablou:

$$\begin{array}{ccccccccc|c} a_1 & a_2 & a_3 & \cdots & a_m & a_{m+1} & a_{m+2} & \cdots & a_n & b \\ 1 & 0 & 0 & \cdots & 0 & \widetilde{a}_{1,m+1} & \widetilde{a}_{1,m+2} & \cdots & \widetilde{a}_{1,n} & \widetilde{b}_1 \\ 0 & 1 & 0 & \cdots & 0 & \widetilde{a}_{2,m+1} & \widetilde{a}_{2,m+2} & \cdots & \widetilde{a}_{2,n} & \widetilde{b}_2 \\ 0 & 0 & 1 & \cdots & 0 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & \cdots & & 1 & \widetilde{a}_{m,m+1} & \widetilde{a}_{m,m+2} & \cdots & \widetilde{a}_{mn} & \widetilde{b}_m \end{array} \qquad (63)$$

Acest tablou are o solutie cu baza $a_1, a_2, \cdots, a_m$. Presupunem ca $\widetilde{b}_i \geq 0$ a.i. solutia fundamentala corespunzatoare este fezabila. Dorim introducerea coloanei $q > m$ si mentinerea fezabilitatii. Pentru a determina care element din coloana $q$ sa–l folosim ca pivot (si deci ce vector al bazei va fi eliminat) folosim (62) ptr. calcularea rapoartelor $\frac{x_k}{\widetilde{a}_{kq}} = \frac{\widetilde{b}_k}{\widetilde{a}_{kq}}$, $k = 1, 2, \cdots, m$, din care il alegem pe cel mai mic nenegativ, si pivotam pe respectivul $\widetilde{a}_{kq}$.

## Interpretare Geometrica

Exista doua interpretari geometrice ale conceptelor introduse:

• In spatiul in care este reprezentat $x$: Regiunea fezabila este o multime convexa iar solutiile fundamentale fezabile sunt punctele extreme (vezi cursul precedent). Punctele extreme adiacente sunt puncte care stau pe o latura comuna.

• In spatiul in care sunt reprezentate coloanele lui $A$ si $b$; Relatia fundamentala este

$$b = \sum_{k=1}^{n} a_i x_i.$$

Un exemplu cu $m = 2$ si $n = 4$ este reprezentat in Figura 9.1. O solutie fezabila este o reprezentare a lui $b$ folosind combinatii semipozitive de $a_i$. O solutie fundamentala fezabila este o reprezentare a lui $b$ folosind numai $m$ combinatii pozitive.

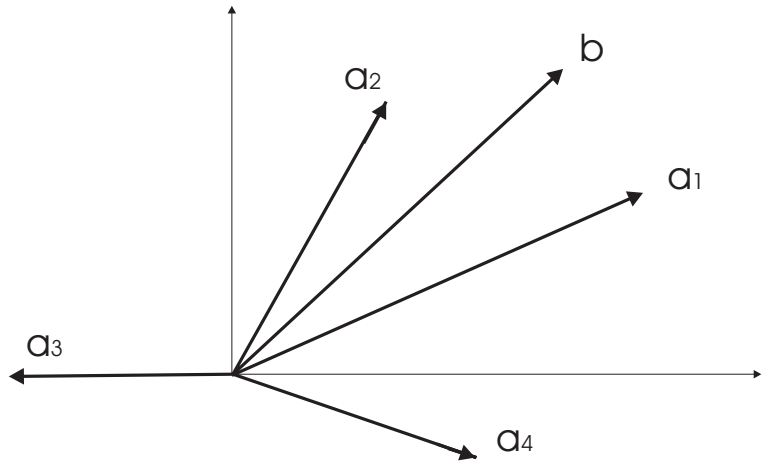



Figura 9.1

Observati ca putem pleca cu o baza formata din $a_1$ si $a_2$ dar nu din $a_1$ si $a_4$! Daca plecam cu $a_1$ si $a_2$ si pentru a obtine o baza adiacenta il adaugam pe $a_3$, atunci automat iese $a_2$. Daca il adaugam pe $a_4$ atunci iese automat $a_1$ !

# 3. Determinarea unei Solutii Minime Fezabile

**Ce stim deja ?** Cum sa pivotam si indata ce am selectat o noua coloana arbitrara care va intra in baza stim sa determinam care coloana iese din baza a.i. solutia sa ramana fezabila (sau respectiv sa determinam ca solutia este nemarginita).

**Ce dorim in continuare ?** Cum selectam coloana pe care o vom introduce in baza a.i. sa reducem valoarea functiei !

**Cuplam cele doua idei si obtinem metoda simplex !**

Presupunem ca avem solutia fundamentala fezabila $x^T =$

$\left[\begin{array}{cc} x_B^T & 0 \end{array}\right] = \left[\begin{array}{ccccccc} x_1 & x_2 & \dots & x_m & 0 & \dots & 0 \end{array}\right]$ impreuna cu tabloul

$$\begin{array}{ccccccccc} a_1 & a_2 & a_3 & \cdots & a_m & a_{m+1} & a_{m+2} & \cdots & a_n & b \\ 1 & 0 & 0 & \cdots & 0 & \widetilde{a}_{1,m+1} & \widetilde{a}_{1,m+2} & \cdots & \widetilde{a}_{1,n} & \widetilde{b}_1 \\ 0 & 1 & 0 & \cdots & 0 & \widetilde{a}_{2,m+1} & \widetilde{a}_{2,m+2} & \cdots & \widetilde{a}_{2,n} & \widetilde{b}_2 \\ 0 & 0 & 1 & \cdots & 0 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & \cdots & & 1 & \widetilde{a}_{m,m+1} & \widetilde{a}_{m,m+2} & \cdots & \widetilde{a}_{mn} & \widetilde{b}_m. \end{array} \tag{64}$$

Valoarea functiei corespunzatoare oricarei solutii este

$$z = \sum_{i=1}^{n} c_i x_i = c_B^T x_B =: z_0, \qquad c_B^T = \left[\begin{array}{ccc} c_1 & \dots & c_m \end{array}\right]. \tag{65}$$

Daca atribuim valori arbitrare lui $x_{m+1}, \dots, x_n$ variabilele

fundamentale se obtin din (64) prin formulele

$$\begin{array}{rcl} x_1 & = & \widetilde{b}_1 - \sum_{j=m+1}^{n} \widetilde{a}_{1j} x_j \\ x_2 & = & \widetilde{b}_2 - \sum_{j=m+1}^{n} \widetilde{a}_{2j} x_j \\ \vdots & & \vdots \\ x_m & = & \widetilde{b}_m - \sum_{j=m+1}^{n} \widetilde{a}_{mj} x_j. \end{array} \tag{66}$$

Folosind (66) se pot elimina variabilele $x_1, \ldots, x_m$ din formula generala (65) si se obtine

$$z = c^T x = z_0 + (c_{m+1} - z_{m+1}) x_{m+1} + (c_{m+2} - z_{m+2}) x_{m+2} + \cdots + (c_n - z_n) x \tag{67}$$

unde

$$z_j := \sum_{i=1}^{m} \widetilde{a}_{ij} c_j, \quad m+1 \leq j \leq n. \tag{68}$$

Relatia (67) este fundamentala pentru determinarea coloanei pivot ce va fi introdusa in baza intrucat da valoarea functiei in orice punct al solutiei $Ax = b$ in termenii variabilelor "libere" $x_{m+1}, \cdots, x_n$.

Din aceste formule se poate deduce daca exista vreun avantaj in inlocurirea unei varibile fundamentale cu una noua. De exemplu, daca unul dintre coeficientii $c_j - z_j$ este negativ pentru un $j$, $m+1 \leq j \leq n$, atunci cresterea lui $x_j$ de la zero la o valoare pozitiva va descreste valoarea functiei si deci va genera o solutie mai buna !

Deducem acum relatiile dintr-o alta perspectiva. Fie $a_i$ coloana $i$ a tabloului. Orice solutie satisface

$$x_1 e_1 + x_2 e_2 + \cdots + x_m e_m = b - \sum_{j=m+1}^{n} x_j a_j$$

Luand produsul scalar al acestei ecuatii vectoriale cu $c_B^T$ obtinem

$$\sum_{i=1}^{m} c_i x_i = c_B^T b - \sum_{j=m+1}^{n} x_j z_j$$

in care $z_j := c_B^T a_j$. Adunand $\sum_{j=m+1}^{n} c_j x_j$ ambilor membri se obtine relatia anterioara

$$c^T x = z_0 + \sum_{j=m+1}^{n} (c_j - z_j) x_j. \tag{69}$$

**Teorema 103. [Imbunatatirea solutiei fundamentale fezabile]** *Fie o solutie fundamentala fezabila nedegenerata avand valoarea corespunzatoare a functiei obiectiv $z_0$ si presupunem ca exista $j$ a.i. $c_j - z_j < 0$. Atunci exista o solutie fundamentala fezabila cu*

*$z < z_0$. Daca coloana $a_j$ poate inlocui un vector in baza fundamentala originala pentru a obtine o noua solutie fundamentala fezabila atunci aceasta noua solutie va avea $z < z_0$. Daca $a_j$ nu poate inlocui un vector din baza originala atunci $K$ este nemarginit si functia obiectiv poate fi facuta oricat de mica (spre $-\infty$).*

Este clar ca daca la orice pas avem $c_j - z_j < 0$ pentru un $j$ oarecare, este posbil sa-l facem pe $x_j > 0$ si sa descrestem functia obiectiv. Ramane de determinat daca $c_j - z_j \geq 0$ pentru toti $j$ implica optimalitatea !

**Teorema 104. [Conditia de optimalitate]** *Daca pentru o solutie fundamentala fezabila avem $r_j := c_j - z_j \geq 0$, pentru toti $j$, atunci solutia este optimala.*

**Coeficienti de cost relativ sau redus:** $r_j := c_j - z_j$ (se introduc

chiar si pentru variabilele fundamentale fiind 0) joaca un rol primordial in metoda simplex !

# 4. Algoritmul Metodei Simplex

**Presupuneri initiale:** Avem o solutie fundamentala fezabila si tabloul corespunzator lui $Ax = b$ este in forma canonica (metode de obtinere a unei solutii fundamentale fezabile initiale vor fi discutate in sectiunea urmatoare).

Folosim tabelul simplex (are o linie in plus):

$$
\begin{array}{ccccccccccc}
a_1 & a_2 & a_3 & \cdots & a_m & a_{m+1} & a_{m+2} & \cdots & a_n & b \\
1 & 0 & 0 & \cdots & 0 & \widetilde{a}_{1,m+1} & \widetilde{a}_{1,m+2} & \cdots & \widetilde{a}_{1,n} & \widetilde{b}_1 \\
0 & 1 & 0 & \cdots & 0 & \widetilde{a}_{2,m+1} & \widetilde{a}_{2,m+2} & \cdots & \widetilde{a}_{2,n} & \widetilde{b}_2 \\
0 & 0 & 1 & \cdots & 0 & \cdot & \cdot & \cdot & \cdot & \cdot \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
0 & 0 & \cdots & & 1 & \widetilde{a}_{m,m+1} & \widetilde{a}_{m,m+2} & \cdots & \widetilde{a}_{mn} & \widetilde{b}_m \\
\hline
0 & 0 & 0 & \cdots & 0 & r_{m+1} & r_{m+2} & \cdots & r_n & -z_0.
\end{array}
\tag{70}
$$

Solutia fundamentala corespunzatoare acestui tabel este

$$
x_i = \begin{cases} \widetilde{b}_i, & 1 \leq i \leq m, \\ 0, & m+1 \leq i \leq n \end{cases}
$$

si fiind fezabila avem $\widetilde{b}_i \geq 0,\ i = 1, 2, \cdots, m$. Valoarea functiei obiectiv este $z_0$.

Coeficientii de cost relativ $r_j$ indica daca valoarea functiei obiectiv va descreste sau va creste daca $x_j$ este introdus in solutie. Daca toti coeficientii sunt semipozitivi atunci solutia este optimala. Daca unii sunt negativi atunci se poate imbunatati solutia. Cand sunt mai multi negativi oricare se poate selecta pentru a determina in ce coloana se va face pivotarea dar de obicei se ia cel mai puternic negativ.

Ultima linie a tabelului poate fi tratata similar cu celelalte cum este explicat in continuare. Intr–adevar, putem vedea $z$ ca o variabila suplimentara si

$$\sum_{i=1}^{n} c_i x_i - z = 0$$

ca o ecuatie suplimentara. Atunci tabelul simplex va avea $m+1$ variabile fundamentale si cerand ca permanent $z$ sa fie una dintre ele **nu mai este nevoie de adaugarea unei coloane in tabel !**

Deci putem intotdeauna incepe un tabel avand ca ultima linie coeficientii $c_i,\ i = 1, 2, \ldots, n$ si un membru drept nul. Folosind operatii standard de pivotare anulam (prin pivotare pe linii) elementele din aceasta linie ce corespund variabilelor fundamentale. Aceasta este echivalent cu a tranforma ecuatia suplimentara in

$$\sum_{j=m+1}^{n} r_j x_j - z = -z_0 \tag{71}$$

care este echivalenta cu (69) si deci $r_j$ obtinuti sunt exact coeficientii de cost relativ.

Dupa ce a fost selectata o coloana $q$ in care se pivoteaza, selectia finala a pivotului se face pe baza calcularii rapoartelor $\frac{b_i}{\widetilde{a}_{iq}}$ pentru elementele $\widetilde{a}_{iq} > 0,\ i = 1, 2, \cdots, m$ si selectand acel indice $p$ corespunzator celui mai mic raport. Pivotarea pe acest element pastreaza fezabilitatea si descreste valoarea functiei obiectiv (in ipoteza de nedegenerare). Daca exista mai multi indici care ating minimul se poate folosi oricare dintre acestia. Daca nu sunt elemente semipozitive in coloana problema este nemarginita. Dupa ce actualizam in final tabelul cu pivotul $\widetilde{a}_{pq}$ (inclusiv ultima linie) se obtine un nou tabel in forma canonica iar noua valoare a functiei obiectiv va apare in coltul dreapta–jos.

**Algoritmul simplex:**

**Pasul 0:** Formeaza un tablou simplex corespunzantor unei solutii

fundamentale fezabile. Coeficientii de cost relativ se obtin prin reducere pe linii;

**Pasul 1:** Daca toti $r_j \geq 0$, **Stop**. Solutia curenta este optimala;

**Pasul 2:** Se alege $q$ a.i. $r_q < 0$ ptr. a afla care variabila nefundamentala va deveni fundamentala;

**Pasul 3:** Se calculeaza $\frac{\widetilde{b}_i}{\widetilde{a}_{iq}}$ ptr. $\widetilde{a}_{iq} > 0$, $i = 1, 2, \cdots, m$. Daca nu exista $\widetilde{a}_{iq} > 0$, **Stop**. Problema este nemarginita. Altfel se alege un $i = p$ pentru care se atinge minimul rapoartelor.

**Pasul 4:** Pivoteaza elementul $\widetilde{a}_{pq}$ actualizand toate liniile incluzand-o pe ultima. **Returnare la Pasul 1.**

In ipoteza de nedegenerare **algoritmul se opreste cand este atinsa optimalitatea sau cand se descopera nemarginirea**.

Daca nici una dintre conditii nu este descoperita la o anumita solutie fundamentala atunci functia obiectiv este strict descrescuta.

Deoarece exista numai un numar finit de solutii si nici o baza nu se repeta intrucat functia obiectiv ia o valoare strict mai mica, **algoritmul va gasi in final o baza fundamentala ce satisface una dintre cele doua conditii de terminare** !

## Degenerare

In timpul aplicarii simplex este posibil sa gasim baze degenerate ! Cum procedam ?

• In multe cazuri putem sa le tratam ca baze fundamentale fezabile nedegenerate;

• Exista posbilitatea teoretica ca minimul lui $\frac{\widetilde{b}_i}{\widetilde{a}_{iq}}$ sa fie zero implicand ca aceasta variabila fundamentala egala cu zero trebuie sa paraseasca baza $\Rightarrow$

• Noua variabila $x_q$ intra in baza la valoarea zero, functia obiectiv nu descreste, si noua baza este deasemenea degenerata ! Mai mult, exista posbilitatea (mai mult teoretica) ca acest proces sa continue pentru cativa pasi dupa care se reapara aceeasi baza degenerata

originala rezultand **un ciclu care se repeta indefinit de mult !**

• Exista metode care pot evita astfel de cicluri si care sunt bazate pe perturbarea usoara a datelor problemei a.i. valorile nule sa fie inlocuite cu mici valori pozitive (vezi Exercitiile).

• In practica nu sunt necesare astfel de proceduri intrucat simplex nu intra in general niciodata in astfel de cicluri.

• Deoarece procedurile anticiclare sunt simple ele sunt intotdeauna incluse in software-ul comercial pentru a evita orice posibila evolutie neplacuta.

# 5. Variabile Artificiale

Cum gasim o solutie initiala fundamentala si fezabila ?

Caz simplu: Consideram ca avem o problema de optimizare cu constrangeri de tipul $Ax \leq b$ cu $b \geq 0$. Aceasta problema se trateaza prin introducerea de variabile suplimentare care automat furnizeaza un punct de plecare pentru metoda simplex.

Caz general: Rareori este posibil sa obtinem sau sa ghicim o solutie initiala fundamentala si fezabila !! Trebuie dezvoltat un mecanism care sa functioneze pentru orice problema !! Interesant (si din fericire adevarat), in orice problema solutia intitiala se poate obtine **rezolvand TOT o problema de programare liniara** avand insa

initializarea evidenta (pentru care putem deci utiliza metoda simplex )!

**Solutie:** Prin operatii elementare constrangerile unui program liniar se pot aduce intotdeauna la forma

$$\begin{aligned} Ax &= b, \\ x &\geq 0, \end{aligned} \tag{72}$$

cu $b \geq 0$. Pentru a gasi o solutie intitiala pentru aceasta problema consideram problema (artificala) de minimizare

$$\begin{aligned} \text{minimizati} \quad & \textstyle\sum_{i=1}^{m} y_i \\ \text{atunci} \quad \text{cand} \quad & Ax + y = b, \\ & x \geq 0, \\ & y \geq 0, \end{aligned} \tag{73}$$

unde $y^T = \left[ \begin{array}{cccc} y_1 & y_2 & \cdots & y_m \end{array} \right]$ este un vector de variabile artificiale.

Daca (72) are o solutie fezabila, atunci este clar ca (73) va avea o solutie (considerata in variabilele $x$ **si** $y$) avand minimul egal cu zero si $y = 0$. Daca (72) nu are solutie fezabila atunci valoarea minima a lui (73) este mai mare decat zero.

Problema (73) este o problema de programare liniara in variabilele $x$ si $y$ si sistemul este deja in forma canonica cu solutia fundamentala fezabila $y = b$. Daca rezolvam sistemul (73) cu metoda simplex obtinem o solutie fundamentala fezabila la fiecare pas! Daca valoarea minima a functiei obiectiv este zero, atunci solutia finala va avea toti $y_i = 0$ si atunci solutia finala nu va contine niciunul dintre $y_i$ ca variabila fundamentala (desigur sub ipoteza de nedegenerare). Daca in solutia finala unii $y_i$ sunt si zero si variabile fundamentale

atunci intotdeauna se pot intershimba cu $x_i$ pastrand niveleul zero al functiei a.i. solutia finala sa aibe numai variabile fundamentale de tip $x_i$ (situatia este mai complicata in acest caz, vezi Exercitiile).

**Concluzii generale** : Metoda generala este in **doua faze**:

**Faza I:** Se introduc variabile artificiale, se obtine o problema de programare liniara care se rezolva (prin simplex) obtinandu-se o solutie fundamentala fezabila pentru problema originala ! In aceasta faza este posibil sa determinam ca problema originala nu are nici o solutie fundamentala fezabila ! In aceasta faza se introduc variabile artificiale doar in acele ecuatii care nu contin variabile suplimentare !

**Faza II:** Folosind solutia fundamentala fezabila determinata la Faza I se minimizeaza functia obiectiv originala. In faza II se omit functia si variabilele artificiale de la Faza I !

# 6. Forma Matriceala a Metodei Simplex

Permite o scriere condensata a metodei simplex ce conduce in final la *metoda simplex revizuita* care prezenta numeroase avantaje numerice.

noindent Fie $B$ matricea coeficientilor variabilelor fundamentale (si care este submatrice a lui $A$) – este formata din (primele) $m$ coloane liniar independente si deci este matrice baza !

Partitionand

$$A = \left[ \begin{array}{cc} B & D \end{array} \right], \quad x = \left[ \begin{array}{c} x_B \\ x_D \end{array} \right], \quad c = \left[ \begin{array}{c} c_B \\ c_D \end{array} \right]$$

obtinem problema de programare liniara

$$\begin{array}{l} \text{minimizeaza} \quad c_B^T x_B + c_D^T x_D \\ \text{atunci} \quad \text{cand} \quad Bx_B + Dx_D = b \\ x_B \geq 0, \quad x_D \geq 0. \end{array} \tag{74}$$

Solutia fundamentala ce este presupusa fezabila corespunzatoare bazei $B$ este $x^T = \left[ \begin{array}{cc} x_B^T & O \end{array} \right]$ unde $x_B = B^{-1}b$ (obtinuta punand automat $x_D = 0$). Pentru un $x_D$ arbitrar valoarea lui $x_B$ este

$$x_B = B^{-1}x_B - B^{-1}Dx_D \tag{75}$$

care substituita in expresia functiei obiectiv conduce la

$$z = c_B^T(B^{-1}b - B^{-1}Dx_D) + c_D^T x_D = c_B^T B^{-1}b + (c_D^T - c_B^T B^{-1}D)x_D \tag{76}$$

care exprima costul oricarei solutii in termenii lui $x_D$. Prin urmare

$$r_D^T = c_D^T - c_B^T B^{-1} D \tag{77}$$

este vectorul de cost relativ (pentru variabilele nefundamentale). Componentele acestui vector se folosesc pentru a hotari care vector va fi adus in baza.

Tabloul initial este

$$\begin{bmatrix} A & b \\ c^T & 0 \end{bmatrix} = \begin{bmatrix} B & D & b \\ c_B^T & c_D^T & 0 \end{bmatrix} \tag{78}$$

care nu este in general in forma canonica si nu corespunde unui punct in procedura simplex. Daca folosim $B$ drept matrice baza atunci

tabloul corespunzator devine

$$\begin{bmatrix} I & B^{-1}D & B^{-1}b \\ 0 & c_D^T - c_B^T B^{-1} D & -c_B^T B^{-1} b \end{bmatrix} \tag{79}$$

care este forma matriceala cautata !

# 7. Metoda Simplex Revizuita

Experienta practica extensiva (cu probleme din domenii extrem de diverse si avand diferite valori pentru $m$ si $n$) a aratat ca metoda converge la o solutie optimala in aproximativ $m$ pivotari (posibil si $\frac{3m}{2}$).

**Concluzie interesanta:** Daca $m << n$ (matricea $A$ are mult mai putine linii decat coloane) pivotii vor aparea in timpul optimizarii doar intr-o mica parte dintre coloane ! Deci efortul facut pentru actualizarea celorlalte coloane dupa pivotare este complet inutil !

Metoda simplex revizuita rezolva aceste probleme:

- Ordoneaza diferit calculele metodei simplex a.i. sa evite calculele

inutile;

- Chiar in situatia in care pivotarea este necesara in toate coloanele dar $m$ este mic in comparatie cu $n$ metoda economiseste uzual un numar important de operatii;

**Algoritmul simplex revizuit:**

**Pasul 0:** Se dau $B^{-1}$ (inversa unei baze curente), si solutia curenta $x_B = \widetilde{b} = B^{-1}b$;

**Pasul 1:** Se calculeaza coeficientul de cost relativ $r_D^T := c_D^T - c_B^T B^{-1} D$. Operatia se efectueaza **calculand intai** $\lambda^T = c_B^T B^{-1}$ **si apoi vectorul de cost** $r_D^T := c_D^T - \lambda^T D$. Daca $r_D \geq 0$, **STOP**; valoarea curenta este optimala;

**Pasul 2:** Alegeti $q$ a.i. $r_q < 0$ ptr. a afla care variabila nefundamentala va deveni fundamentala (se alege cel mai negativ coeficient de cost); se calculeaza $\widetilde{a}_q := B^{-1}a_q$ care da vectorul $a_q$ in baza curenta;

**Pasul 3:** Calculati $\frac{\widetilde{b}_i}{\widetilde{a}_{iq}}$ ptr. $\widetilde{a}_{iq} > 0$, $i = 1, 2, \cdots, m$. Daca nu exista $\widetilde{a}_{iq} > 0$, **STOP**. Problema este nemarginita. Altfel alege un $i = p$ pentru care se atinge minimul si deci vectorul $p$ va parasi baza;

**Pasul 4:** Actualizeaza $B^{-1}$ si solutia curenta $B^{-1}\widetilde{b}$. Returnare la Pasul 1.

**Observatia 105.** *Actualizarea lui $B^{-1}$ se obtine prin operatiile uzuale de pivotare aplicate unei matrici formate din $B^{-1}$ si din $\widetilde{a}_q$,*

*unde pivotul este dat de elementul corespunzator din $\widetilde{a}_q$. Simultan se poate desigur actualiza si $B^{-1}\widetilde{b}$ adaugand o singura coloana.*

**Observatia 106.** *Pentru inceperea procedurii avem nevoie de o solutie initiala fezabila si de inversa bazei initiale. In multe probleme baza initiala este matricea identitate $I_m$ (si evident si inversa) – rezultand din variabile suplimentare sau artificiale !*

## Forma Produs a Inversei

Aceasta varianta de metoda simplex revizuita se bazeaza pe o reprezentare de tip produs a inversei bazei. Avantajele variantei:

- Are nevoie de mai putine calcule decat alte metode;
- Avantajul principal consta in necesarul foarte mic de memorie "rapida".

Sa presupunem ca tabloul $T$ se modifica prin pivotarea cu un pivot $\widetilde{a}_{pq}$ din coloana $a_q = \begin{bmatrix} \widetilde{a}_{1q} \\ \widetilde{a}_{2q} \\ \dots \\ \widetilde{a}_{mq} \end{bmatrix}$. Rezultatul pivotarii este matricea

$ET$ in care $E$ este matricea elementara

$$E = \begin{bmatrix} 1 & 0 & \cdots & 0 & v_1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 & v_2 & 0 & & 0 \\ & & & 1 & \vdots & 0 & & \\ & & & & v_p & 0 & & \vdots \\ & & & & & 1 & & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & v_m & 0 & \cdots & 1 \end{bmatrix}, \tag{80}$$

cu elementele $v_i$ din coloana $p$ avand valorile

$$\begin{aligned} v_i &= -\frac{\widetilde{a}_{iq}}{\widetilde{a}_{pq}}, \quad i \neq p, \\ v_p &= \frac{1}{\widetilde{a}_{pq}}. \end{aligned} \tag{81}$$

Matricea $E$ se determina pe baza coloanei pivot. Multiplicarea

oricarei coloane la stanga cu matricea $E$ adauga unei componente arbitrare $i \neq p$ componenta $p$ multiplicata cu $v_i$ si multiplica componenta $p$ cu $v_p$ – aceasta corespunde exact operatiilor de pivotare facute intr-o coloana a tabelului!

Initializand algoritmul simplex cu o matrice baza identitate obtinem dupa $k$ pivotari succesive inversa

$$B^{-1} := E_k E_{k-1} \dots E_2 E_1 \tag{82}$$

in care $E_i$ este matricea elementara corespunzatoare pivotarii $i$. Aplicand aceasta reprezentare a inversei se obtine o noua varianta a metodei simplex revizuite.

## Metoda Simplex Revizuita de Tip Produs

**Pasul 0:** Presupunem ca dupa pivotarea $k$ am avem memorate $E_1, \ldots E_k$.

**Pasul 1:** Se calculeaza solutia de baza prin formula recursiva

$$x_B = (E_k(E_{k-1} \ldots (E_1 b)))$$

**Pasul 2:** Se calculeaza coeficientii de cost relativ $r_D^T = c_D^T - c_B^T B^{-1} D$ (se foloseste aceeasi metoda de calcul cu $\lambda$ ca anterior). Vectorul $\lambda$ se obtine prin formula recursiva

$$\lambda^T = (((c_B^T E_k) E_{k-1}) \ldots E_1).$$

Daca $r_D \geq 0$, **STOP**; solutia curenta este optimala.

**Pasul 3:** Se selecteaza cel mai negativ coeficient de cost relativ si se calculeaza vectorul corespunzator

$$\widetilde{a}_q = (((E_k)E_{k-1}\dots(E_1\widetilde{a}_q))).$$

**Pasul 4:** Daca nu exista $\widetilde{a}_{iq} > 0$, **STOP**. Problema este nemarginita. Altfel alege un $i = p$ pentru care se atinge minimul $\frac{\widetilde{b}_i}{\widetilde{a}_{iq}}$ pentru care $\widetilde{a}_{iq} > 0$ stabilind astfel care vector paraseste baza. In felul aceasta s-a determinat care componenta este noul pivot si deci $E_{k+1}$. Returnare la Pasul 1.

**Observatia 107.** *Pentru stocarea matricilor $E$ este necesara memorarea a numai $m + 1$ numere: $(p, v_1, \dots, v_m)$. Mai mult, nici nu este nevoie sa reconstituim matricile $E$ in mod explicit ci doar formulele de multiplicat $E$ la dreapta cu un vector coloana (Pasii 1 si*

*3) si la stanga cu un vector linie (Pasul 2) !*

**Observatia 108.** *In cazul rezolvarii unor probleme de dimensiuni mari prin aceasta metoda majoritatea informatiilor se pot pastra in memoriile mai lente ale masinii de calcul (HDD, DVD, etc). Coloanele lui $A$ se pot aduce in memoria RAM in mod individual, si se pot inmulti cu $B^{-1}$ prin aducerea* **succesiva** *in memoria RAM a vectorilor de dimensiune $m+1$ ce definesc matricile $E$.*

**Observatia 109.** *Problemele de dimensiuni mari care apar uzual in programarea liniara au multa structura, implicand adesea ca matricea $A$ (si desemenea matricea $B$) sunt matrici rare (cu putine elemente nenule). In orice caz, $B^{-1}$ nu este matrice rara in general si acest fapt evidentiaza avantajul formei produs. Un avantaj suplimentar al formei produs este ca vectorii $(p, v_1, v_2, \ldots, v_m)$ definind matricile $E$*

*sunt probabil ei insasi rari, generand o reducere masiva in memoria si timpul de executie necesare.*

## Reinversare

Pentru probleme de dimensiuni mari sau prost conditionate numeric erorile de calcul se acumuleaza pana la un nivel la care $B^{-1}$ **este puternic imprecisa** ! Evaluarea erorii la pasul curent se face evaluand

$$Bx_B - b$$

in care $x_B$ este valoarea curenta gasita prin formula $x_B = B^{-1}b$ in care pentru calculul lui $B^{-1}$ se foloseste valoarea curenta a lui $B^{-1}$. Daca eroarea este semnificativa este recomandata **reinversarea cat mai exact posibil a lui $B$ si restartatea procedurii**. Aceasta operatie se numeste **reinversare** si este o componenta esentiala a

oricarei biblioteci de programe. Pentru inversarea lui $B$ se poate folosi orice procedura numeric stabila, cel mai adesea folosindu-se proceduri de pivotare in care pivotii se aleg la fiecare pas a.i. sa minimizeze erorile de calcul.

In cazul formei produs reinversarea serveste si pentru **controlul necesarului de memorie**: sirul lung de matrici elementare ce descriu inversa lui $B$ este inlocuit cu un sir de $m$ matrici ! Mai mult, in acest caz alegerea pivotilor se face a.i. sa mentinem structura rara (in afara de acuratetea calculelor).

# 8. Metoda Simplex via Descompunerea LU

In procedura propusa in sectiunea precedenta am vazut ca parcurgerea unui pas din metoda simplex nu este dependenta de cunoasterea explicita a inversei $B^{-1}$ ci de rezolvarea unor sisteme de ecuatii avand matricea coeficient $B$. Cu aceasta observatie procedura simplex din sectiunea precedenta se poate da in forma:

**Algoritmul simplex revizuit:**

**Pasul 0:** Se da baza curenta $B$.

**Pasul 1:** Se calculeaza solutia curenta $x_B$ ce satisface $Bx_B = \widetilde{b}$;

**Pasul 2:** Se rezolva $\lambda^T B = c_B^T$ si se calculeaza coeficientul de cost relativ $r_D^T := c_D^T - \lambda^T D$. Daca $r_D \geq 0$, **STOP**; valoarea curenta este optimala;

**Pasul 3:** Se alege $q$ a.i. $r_q < 0$ ptr. a afla care variabila nefundamentala va deveni fundamentala (se alege cel mai negativ coeficient de cost); se rezolva $B\widetilde{a}_q := a_q$ care da vectorul $a_q$ in baza curenta;

**Pasul 4:** Se calculeaza $\frac{\widetilde{b}_i}{\widetilde{a}_{iq}}$ ptr. $\widetilde{a}_{iq} > 0$, $i = 1, 2, \cdots, m$. Daca nu exista $\widetilde{a}_{iq} > 0$, **STOP**. Problema este nemarginita. Altfel se alege un $i = p$ pentru care se atinge minimul si deci vectorul $p$ va parasi baza;

**Pasul 5:** Actualizeaza $B$. Returnare la Pasul 1.

Din procedura de mai sus se observa ca intr-adevar nu este nevoie de $B^{-1}$ ci doar de rezolvarea a trei sisteme de ecuatii liniare, doua avandu-l pe $B$ drept coeficient matriceal si unul pe $B^T$. In procedurile anterioare sistemele erau implicit rezolvate prin operatiile corespunzatoare de pivotare pe masura ca metoda progresa. Dpdv al eficientei si preciziei numerice metoda de pivotare prezentata nu este la fel de eficienta precum eliminarea Gaussiana pentru sisteme generale de ecuatii liniare. In consecinta are rost sa adaptam eliminarea Gaussiana pentru metoda simplex, rezultatul fiind o versiune a metodei simplex revizuite care are o stabilitate numerica mai buna si pentru probleme de dimensiuni mari ofera economii importante de memorie.

Ne concentram asupra celor trei sisteme ce trebuie rezolvate la

fiecare pas:

$$Bx_B = b, \qquad \lambda^T B = c_B^T, \qquad B\widetilde{a}_q = a_q. \tag{83}$$

Presupunem ca $B$ a fost descompus sub forma $B = LU$ unde $L$ si $U$ sunt matrici inferior respectiv superior triangulare (presupunem pentru simplitate ca nu avem nevoie de reordonare pe linii – se poate relaxa dar complexitatea expunerii creste considerabil). Deci fiecare sistem (83) se poate rezolva prin rezolvarea a doua sisteme triangulare.

Cheia cresterii eficientei metodei simplex consta in actualizarea eficienta a acestor doua matrici triangulare cand se schimba un singur

vector baza ! Presupunem ca la inceputul metodei simplex avem

$$B = \left[ \begin{array}{cccc} a_1 & a_2 & \ldots & a_m \end{array} \right]$$

si la sfarsitul ciclului vom avea

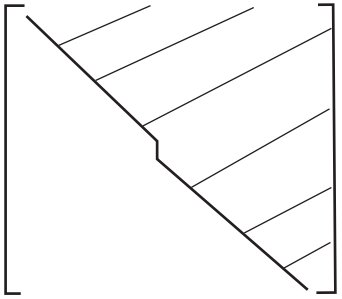
$$\overline{B} = \left[ \begin{array}{ccccccc} a_1 & a_2 & \ldots & a_{p-1} & a_{p+1} & \ldots & a_m & a_q \end{array} \right]$$

(vectorul $a_p$ s-a inlocuit cu $a_q$). Sa observam ca $\overline{B}$ se obtine din $B$ prin eliminarea lui $a_p$, siftarea vectorilor cu indici mai mari spre stanga si adaugarea lui $a_q$ in extremitatea dreapta. Avem

$$L^{-1}\overline{B} = \left[ \begin{array}{ccccccc} L^{-1}a_1 & L^{-1}a_2 & \ldots & L^{-1}a_{p-1} & L^{-1}a_{p+1} & \ldots & L^{-1}a_m & L^{-1}a \end{array} \right.$$

$$= \left[ \begin{array}{ccccccc} u_1 & u_2 & \ldots & u_{p-1} & u_{p+1} & \ldots & u_m & L^{-1}a_q \end{array} \right] = \overline{H}$$

in care $u_i$ sunt exact coloanele lui $U$ iar matricea $\overline{H}$ are forma

$$\overline{\mathbf{H}} = \left[ \right.$$

$$\left. \right]$$

cu zerouri sub diagonala principala in primele $k-1$ coloane si zerouri sub elementul imediat de sub diagonala principala in restul de coloane. Matricea $\overline{H}$ se poate construi fara calcule suplimentare deoarece $u_i$ sunt cunoscute si $L^{-1}a_q$ este deja calculat atunci cand se rezolva al treilea sistem (83). Tot ce avem de facut este sa anulam elementele nenule subdiagonale folosind de exemplu eliminarea Gaussiana. In final se obtine matricea superior triangulara $\overline{U}$ ca urmare a aplicarii

unor transformari de tipul

$$M_i = \begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & 1 & & & \\ & & & m_i & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix}, \qquad i = k, k+1, \ldots, m-1 \tag{84}$$

sub forma

$$\overline{U} = M_{m-1} M_{m-2} \cdots M_k \overline{H}. \tag{85}$$

Obtinem in final

$$\overline{B} = L\overline{H} = L M_k^{-1} M_{k+1}^{-1} \cdots M_{m-1}^{-1} \overline{U}, \tag{86}$$

$$\overline{L} = LM_k^{-1} \dots M_{m-1}^{-1} \tag{87}$$

si deci descompunerea

$$\overline{B} = \overline{LU}. \tag{88}$$

Se observa ca $\overline{L}$ se calculeaza simplu si este inferior triunghiulara deoarece

$$M_i^{-1} = \begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & 1 & & & \\ & & & -m_i & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix}$$

**Observatia 110.** *Exista multe variante concrete de implementare:*

*transformarile $M_i$ se pot memora in loc sa evaluam $\overline{L}$; descompunerea LU se poate reevalua periodic (similar ca la reinversare) si se pot folosi intreschimbari de linii si coloane pentru a mximiza stabilitatea sau pentru a minimiza "densitatea" descompunerii (vezi Exercitiile)*

# 9. Concluzii

Metoda simplex se bazeaza pe faptul ca daca valoarea optimala a unui program liniar este finita atunci **se atinge la o solutie fundamentala fezabila**. Exista **doua interpretari posibile** ale metodei simplex:

- **(Prin variatii continue)**: Se pleaca cu o solutie fezabila fundamentala si una dintre variabilele nefundamentale este crescuta incet de la zero. In timp ce aceasta variabila este crescuta, restul de variabile fundamentale sunt ajustate a.i. sa se pastreze fezabiliatea. Schimbarea functiei obiectiv la o schimbare unitara a variabilei nefundamentale este coeficientul de cost relativ asociat cu variabila nefundamentala. Daca coeficientul este negativ, valoarea functiei

obiectiv este imbunatatita pe masura ce variabila nefundamentala este crescuta pana in punctul in care o crestere mai mare violeaza fezabilitatea. In acest punct una dintre variabilele fundamentale este zero si se face intreschimbarea cu varibila nefundamentala.

- **(Prin variatii discrete)**: Deoarece este necesar sa consideram doar solutii fundamentale fezabile, se selecteaza diverse baze si se calculeaza solutiile fundamentale corespunzatoare prin rezolvarea unor sisteme de ecuatii liniare. Logica alegerii unor noi baze este generata de coeficientii de cost relativ.

# Partea a–III–a: CONTROL OPTIMAL SI ROBUST

## Capitolul 10 : NOTIUNI DE BAZA

Sistem Liniar

Proprietati de Baza ale Sistemelor Liniare

Spatii de Semnale si Functii de Transfer

Evolutii pe Spatiul Starilor Generate de Intrari $L^2$

Operatorul $L^2$ Intrare–Iesire

Descriere Dinamica:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{89}$$

in care $t \in \mathbb{R}$, $x(t) \in \mathbb{R}^n$ este starea, $u(t) \in \mathbb{R}^m$ este intrarea, $y(t) \in \mathbb{R}^p$ este iesirea sistemului, si $A$, $B$, $C$, $D$ sunt matrici constante de dimensiuni adecvate

$$A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{p\times n}, D \in \mathbb{R}^{p\times m}.$$

Comportarea intrare–iesire a sistemului (89) este descrisa convenabil de matricea de transfer care este matricea rationala proprie

de dimensiuni $p \times m$

$$\mathbf{G}(s) := C(sI - A)^{-1}B + D = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right](s). \tag{90}$$

Facand o schimbare de variabile de forma $\widetilde{x} := Tx$ ($T$ matrice inversabila), vectorul transformat $\widetilde{x}$ satisface

$$\begin{array}{rcl} \dot{\widetilde{x}}(t) & = & \widetilde{A}\widetilde{x}(t) + \widetilde{B}u(t), \\ y(t) & = & \widetilde{C}\widetilde{x}(t) + \widetilde{D}u(t), \end{array} \tag{91}$$

in care

$$\widetilde{A} := TAT^{-1}, \quad \widetilde{B} := TB, \quad \widetilde{C} := CT^{-1}, \quad \widetilde{D} := D. \tag{92}$$

NOTA:

- $\mathbf{G}(s)$ se obtine luand transformata Laplace in (89) si explicitand $y(s)$ ca functie de $u(s)$ in forma

$$y(s) = \mathbf{G}(s)u(s) \tag{93}$$

- Incepand cu o reprezentare a lui $\mathbf{G}$ ca matrice rationala proprie putem intotdeauna obtine reprezentare de stare (89) scriind realizarea standard $(A, B, C, D)$ a lui $\mathbf{G}$.

- Se poate intotdeauna asigura minimalitatea realizarii.

- O schimbare de variabila (transformare de similaritate sau transformare de echivalenta) nu afecteaza spectrul matricii $A$ nici matricea de transfer intrare–iesire.

# Proprietati de Baza ale Sistemelor Liniare

Un sistem se numeste

- stabil: daca toate valorile proprii ale lui $A$ sunt in $\mathbb{C}_-$
- antistabil: daca toate valorile proprii ale lui $A$ sunt in $\mathbb{C}_+$
- dihotomic: daca $A$ nu are valori proprii pe axa imaginara

NOTA:

- Stabilitatea, antistabilitatea, dihotomia si proprietatile I/O sunt invariante sub transformari de coordonate

- Facand o transformare de coordonate $T$, un sistem dihotomic se poate descompune intr–o parte stabila si una antistabila:

$$\widetilde{A} = TAT^{-1} = \left[ \begin{array}{cc} A_- & O \\ O & A_+ \end{array} \right] \begin{array}{l} \}n_- \\ \}n_+ \end{array}, \widetilde{B} = TB = \left[ \begin{array}{c} B_- \\ B_+ \end{array} \right], \widetilde{C} = CT^{-1} =$$

(94)

in care $A_-$ este stabila si $A_+$ este antistabila.

Un sistem se numeste:

- Controlabil: daca $(A, B)$ satisface

$$\text{rank} \left[ \begin{array}{cc} sI - A & B \end{array} \right] = n, \forall s$$

- Stabilizabil: daca $(A, B)$ satisface

$$\text{rank} \begin{bmatrix} sI - A & B \end{bmatrix} = n, \quad \forall s, \quad \text{Re}\, s \geq 0$$

$\Leftrightarrow$ Exista $F$ a.i. $A + BF$ este stabila

- Antistabilizabil: daca $(A, B)$ satisface

$$\text{rank} \begin{bmatrix} sI - A & B \end{bmatrix} = n, \quad \forall s, \quad \text{Re}\, s \leq 0.$$

Notiuni duale: Observabilitate, Detectabilitate, Antidetectabilitate

# Spatii de Semnale si Matrici de Transfer

DEFINIM:

- $L^\infty(\mathbb{C}_0)$: spatiul Banach al functiilot matriciale complexe $G(s)$ care sunt marginite pe $\mathbb{C}_0$, cu *norma–*$L^\infty(\mathbb{C}_0)$

$$\|G\|_\infty := \sup_{s\in\mathbb{C}_0} \overline{\sigma}(G(s)) \tag{95}$$

  in care $\overline{\sigma}$: valoarea singulara maxima.

- $RH^\infty$: subspatiul rational constand din matrici proprii $p \times m$ cu elemente marginite pe axa imaginara.

- $RH^{\infty}_{+,p\times m}$ spatiul matricilor rationale proprii cu elemente analitice in $\mathbb{C}_+ \cup \mathbb{C}_0$

- $RH^{\infty}_{-,p\times m}$ spatiul matricilor rationale proprii cu elemente analitice in $\mathbb{C}_- \cup \mathbb{C}_0$.

NOTA:

- Daca $A$ este dicotomic $\Rightarrow G \in RL^{\infty}$.

- Daca $A$ este stabila $\Rightarrow G \in RH^{\infty}_{+}$

- Daca $A$ este antistabila $\Rightarrow G \in RH^{\infty}_{-}$.

- Un sistem cu matricea de transfer $\in RH^{\infty}_{+}$ este stabil intrare–iesire (coincide cu $RH^{\infty}_{+}$)

- Un sistem stabil (numit si intern stabil – cu $A$ stabila) este intotdeauna stabil in sens intrare–iesire.

- Reciproca nu este adevarata decat daca realizarea este stabilizabila si detectabila.

Problema centrala in Teoria sistemelor: stabilitatea interna pentru ca garanteaza ca orice semnal de energie finita (norma $L^2$ finita) injectat oriunde in sistem genereaza semnale de energie marginita (norma $L^2$ finita).

Pentru functii de transfer avem

$$\|\mathbf{G}\|_\infty = \sup_{\omega\in\mathbb{R}} \overline{\sigma}(\mathbf{G}(j\omega)) = \sup_{\omega\in\mathbb{R}} \|\mathbf{G}(j\omega)\|$$

in care $\|\cdot\|$ este norma operatoriala indusa a matricii (= norma indusa a operatorului intrare–iesire a sistemului).

NOTA:

- Norma $L_\infty$ a unui sistem este finita daca si numai daca matricea de transfer este analitica pe axa imaginara incluzand la infinit !

- Norma $L^\infty$ a unui sistem da energia maxima a semnalului de iesire cand intrarea este un semnal de energie marginita de $1$ ! — Cel mai prost caz!

# Rezultate asupra normei $L^\infty$

**Teorema 111. [Teorema lui Rellich]** *Fie matricea $n \times n$ $M(x) = M^*(x)$ depinzand continuu de parametrul real $x$ apartinand intervalului $I$. Atunci exista functii reale continue $\mu_1, \ldots, \mu_n$ definite pe $I$, a.i. $\mu_n(x) \leq \ldots \leq \mu_1(x)$, pentru orice $x$ in $I$, si $\Lambda(M(x)) = \{\mu_i(x)\}$.*

**Corolarul 112.** *Pentru $\mathbf{G} \in RL^\infty_{p\times m}$, $\overline{\sigma}(\mathbf{G}(j\omega))$ depinde continuu de $\omega \in \mathbb{R}$, este marginita si isi atinge maximum pentru $\omega \in [-\infty, \infty]$, unde prin definitie $\mathbf{G}(j\infty) = \mathbf{G}(-j\infty) = D$.*

**Teorema 113.** *Presupunem ca sirul de sisteme*

$$\mathbf{G}_k = \left[\begin{array}{c|c} A_k & B_k \\ \hline C_k & D_k \end{array}\right] \quad , \quad k \in \mathbf{N}, \tag{96}$$

*este a.i.* $A_k \to A, B_k \to B, C_k \to C$ *si* $D_k \to D$ *cand* $k \to \infty$, $A$ *si* $A_k$ *sunt stabile, si* $\mathbf{G}(s) = C(sI - A)^{-1}B + D \equiv 0$. *Atunci*

$$\lim_{k \to \infty} \|\mathbf{G}_k\|_\infty = 0.$$

Demonstratie: Rezultatul poate suna evident dar nu este deloc trivial !!!!

# Rezultate privind norma $L^2$

DEFINIM: Spatiul $L^2(\mathbf{M})$: Spatiul Hilbert al functiilor complexe matriciale (vectoriale sau scalare ) $G(t)$ pentru care

$$\int_{\mathbf{M}} \mathrm{Trace}\,[G^*(t)G(t)]\mathrm{d}t < \infty, \tag{97}$$

unde $\mathbf{M}$ este sau $\mathbf{Z}$, $(-\infty, \infty)$, $\mathbb{C}_0$, sau $\partial\mathbf{D}$.

In particular, $L^2(\mathbb{C}_0)$ este spatiul Hilbert al functiilor matriciale (vectoriale sau scalare) $G(s)$ cu norma $L^2(\mathbb{C}_0)$ finita

$$\|G\|_2 := \{\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{Trace}\,[G^*(j\omega)G(j\omega)]\mathrm{d}\omega\}^{\frac{1}{2}}. \tag{98}$$

## NOTA:

- Daca $G \equiv \mathbf{G}$ este o functie rationala atunci norma $L^2(\mathbb{C}_0)$ este finita daca si numai daca $\mathbf{G}$ nu are poli pe axa imaginara si este strict proprie.

- Norma $L^2$ a unui sistem da energia semnalului de iesire cand la intrare se aplica un impuls Dirac!!!

# Calculul alternativ al normelor

Presupunem $D = 0$ si $A$ stabila. Norma $L^2(\mathbb{C}_0)$ a lui $\mathbf{G}$ se poate calcula folosind

$$\|\mathbf{G}\|_2 = [\mathrm{Trace}\,(B^T Q B)]^{\frac{1}{2}} = [\mathrm{Trace}\,(C P C^T)]^{\frac{1}{2}} \qquad (99)$$

in care $P$ si $Q$ sunt unicele solutii ale urmatoarelor ecuatii

$$AP + PA^T + BB^T = 0, \qquad A^T Q + QA + C^T C = 0. \qquad (100)$$

Pp $D = 0$ si $A$ antistabila. Norma $L^2(\mathbb{C}_0)$ a lui $\mathbf{G}$ se poate calcula folosind

$$\|\mathbf{G}\|_2 = [\mathrm{Trace}\,(B^T \widehat{Q} B)]^{\frac{1}{2}} = [\mathrm{Trace}\,(C \widehat{P} C^T)]^{\frac{1}{2}} \qquad (101)$$

unde $\widehat{P}$ si $\widehat{Q}$ sunt unicele solutii ale ecuatiilor

$$A\widehat{P} + \widehat{P}A^T - BB^T = 0, \qquad A^T\widehat{Q} + \widehat{Q}A - C^TC = 0. \qquad (102)$$

Calculul normei $L^\infty$ necesita intotdeauna o cautare iterativa !!!

# Evolutii in Spatiul Starilor Corespunzatoare Intrarilor $L^2$

Investigam solutiile particulare ale sistemului diferential

$$\dot{x} = Ax + Bu, \tag{103}$$

unde $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}$, in cazul in care $u$ este o functie $L^{2,m}(-\infty, \infty)$: spatiul Hilbert al functiilor de patrat integrabil in sens Lebesgue pe $(-\infty,\ \infty)$, luand valori in $\mathbb{R}^m$.

Daca $u \in L^{2,m}$ definim norma $L^2$ a lui $u$:

$$\|u\|_2 := (\int_{-\infty}^{\infty} \|u(t)\|^2 \mathrm{d}t)^{\frac{1}{2}} < \infty.$$

Fie deasemenea

- $L_+^{2,m}$: subspatiile inchise a lui $L^{2,m}$ de functii cu suport $[0,\ \infty)$
- $L_-^{2,m}$: subspatiile inchise a lui $L^{2,m}$ al functiilor cu suport $(-\infty, 0]$

Nota:

$$L^{2,m} = L_+^{2,m} \oplus L_-^{2,m}.$$

# Solutii $L^2$ ale lui $\dot{x} = Ax + Bu$

Cautam solutii $L^2$ in cazurile urmatoare:

1. Sub conditii initiale

2. Pe intreaga axa de timp:

   - $A$ stabila
   - $A$ antistabila
   - $A$ dihotomica
   - $A$ arbitrara

## 1. Solutii sub conditii intitiale

Fie $\xi \in \mathbb{R}^n$ arbitrar, si $u \in L_+^{2,m}$. Fie $x^{\xi,u}$ solutia care satisface $x(0) = \xi$ data de formula de variatie a constantelor

$$x^{\xi,u}(t) = e^{At}\xi + \int_0^t e^{A(t-\tau)}Bu(\tau)\mathrm{d}\tau, \quad t \geq 0 \tag{104}$$

sau in forma operatoriala

$$x^{\xi,u} = \Phi\xi + \mathcal{L}u \tag{105}$$

cu

$$\Phi : \mathbb{R}^n \to L_+^{2,n}\ , \quad (\Phi\xi)(t) := e^{At}\xi, \quad t \geq 0, \tag{106}$$

si

$$\mathcal{L} : L_+^{2,m} \to L_+^{2,n} \ , \quad (\mathcal{L}u)(t) := \int_0^t e^{A(t-\tau)} B u(\tau) \mathrm{d}\tau, \quad t \geq 0. \ (107)$$

NOTA:

- Acesti doi operatori nu sunt cu necesitate marginiti
- Daca $A$ este stabila, atunci $x^{\xi,u} \in L_+^{2,n}$ si ambii operatori $\Phi$ si $\mathcal{L}$ sunt marginiti.

Ne intereseaza si alte solutii ale $\dot{x} = Ax + Bu$ (pe intreaga axa de timp). In orice caz, proprietatile lor depind puternic de spectrul lui $A$ si trebuie deci considerate separat cazurile in care $A$ este stabila, antistabila, dihotomica, si arbitrara.

## 2. Solutii pe intreaga axa

**Teorema 114.** *Fie $A$ stabila. Atunci pentru fiecare $u \in L^{2,m}$, ecuatia $\dot{x} = Ax + Bu$ are o solutie unica $x_e^u$ (absolut continua) in $L^{2,n}$, data de*

$$x_e^u(t) = \int_{-\infty}^{t} e^{A(t-\tau)} B u(\tau) \mathrm{d}\tau \ , \quad t \in \mathbb{R}. \tag{108}$$

Putem defini operatorul marginit $\mathcal{L}_e : L^{2,m} \to L^{2,n}$ prin

$$(\mathcal{L}_e u)(t) := \int_{-\infty}^{t} e^{A(t-\tau)} B u(\tau) \mathrm{d}\tau \ , \quad t \in \mathbb{R} \tag{109}$$

si obtinem forma operatoriala a solutiei

$$x_e^u = \mathcal{L}_e u. \tag{110}$$

Mai mult, avem urmatoarele conexiuni intre cei doi operatori $\mathcal{L}$ introdusi pana acum

$$\mathcal{L} = P_+^n \mathcal{L}_e P_+^m = \mathcal{L}_e P_+^m, \tag{111}$$

unde $P_\pm^r$ sunt proiectiile ortogonale ale lui $L^{2,r}$ pe $L_\pm^{2,r}$.

**Teorema 115.** *Fie $A$ antistabila. Atunci pentru oricare $u \in L^{2,m}$, ecuatia $\dot{x} = Ax + Bu$ are o solutie unica $x_e^u$ in $L^{2,n}$ (absolut continua) data de*

$$x_e^u(t) = -\int_t^\infty e^{A(t-\tau)} B u(\tau) \mathrm{d}\tau \ , \quad t \in \mathbb{R}. \tag{112}$$

Analog, putem defini in acest caz operatorul liniar marginit $\mathcal{L}_e$ :

$L^{2,m} \to L^{2,n}$ dat de

$$(\mathcal{L}_e u)(t) := -\int_t^\infty e^{A(t-\tau)} B u(\tau) \mathrm{d}\tau \ , \quad t \in \mathbb{R}, \tag{113}$$

a.i. $x_e^u = \mathcal{L}_e u$ continua sa fie valida.

**Teorema 116.** *Fie $A$ dihotomica. Atunci, oricare $u$ in $L^{2,m}$, ecuatia $\dot{x} = Ax + Bu$ are o unica solutie in $L^{2,n}$ data de*

$$x_e^u(t) = \int_{-\infty}^t e^{A(t-\tau)} \Pi_- B u(\tau) \mathrm{d}\tau - \int_t^\infty e^{A(t-\tau)} \Pi_+ B u(\tau) \mathrm{d}\tau \ , \quad t \in \mathbb{R} \tag{114}$$

*unde $\Pi_-$ and $\Pi_+$ sunt proiectiile lui $\mathbb{R}^n$ pe subspatiile stabile si antistabile ale lui $A$.*

Similar, definim operatorul liniar marginit $\mathcal{L}_e : L^{2,m} \to L^{2,n}$, $x_e^u :=$

$\mathcal{L}_e u$ si $\mathcal{L} : L_+^{2,m} \to L_+^{2,n}$, $\mathcal{L} := P_+^n \mathcal{L}_e P_+^m, x^u := \mathcal{L} u$.

# Cazul general: $A$ arbitrar

Asa cum am vazut, in acest caz definim solutiile (cu conditii initiale) $x^{\xi,u}$ prin

$$x^{\xi,u}(t) = e^{At}\xi + \int_0^t e^{A(t-\tau)}Bu(\tau)\mathrm{d}\tau, \quad t \geq 0. \tag{115}$$

In general, $x^{\xi,u}$ nu este in $L_+^{2,n}$. Este interesant sa consideram multimea acelor functii de intrare $u$ care fac $x^{\xi,u}$ integrabila (in patratul normei). Aceste intrari pot fi vazute ca stabilizand sistemul in bucla deschisa. Introducem

$$\mathcal{U}_\xi := \{u : u \in L_+^{2,m} \text{ a.}i.\ x^{\xi,u} \in L_+^{2,n}\}. \tag{116}$$

Daca $A$ este stabila atunci $\mathcal{U}_\xi = L_+^{2,m}$. In general, avem:

**Lema 117.** *Presupunem $(A, B)$ stabilizabila si fie $F$ o reactie stabilizatoare. Fie $\xi \in \mathbb{R}^n$ si $u \in L_+^{2,m}$. Atunci $u \in \mathcal{U}_\xi$ daca si numai daca $u = Fx_F^{\xi,v} + v$ , in care $v$ este arbitrara in $L_+^{2,m}$ si $x_F^{\xi,v}$ este unica solutie in $L_+^{2,n}$ a lui*

$$\dot{x}_F = (A + BF)x_F + Bv \ , \quad x_F(0) = \xi.$$

# Operatorul $L^2$ Intrare–Iesire

Consideram un sistem liniar cu $A$ dihotomica si un $u \in L^{2,m}$ arbitrar. Atunci

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{117}$$

se poate rescrie ca

$$\begin{aligned} x &= x_e^u = \mathcal{L}_e u \\ y &= (C\mathcal{L}_e + D)u \qquad (\in L^{2,p}). \end{aligned}$$

Sistemul (117) defineste astfel un operator liniar marginit din $L^{2,m}$ in

$L^{2,p}$:

$$(\mathcal{G}u)(t) = \int_{-\infty}^{t} Ce^{A(t-\tau)}\Pi_- Bu(\tau)\mathrm{d}\tau - \int_{t}^{\infty} Ce^{A(t-\tau)}\Pi_+ Bu(\tau)\mathrm{d}\tau + Du($$

(118)

Printr-o transformare $T$ care descompune sistemul in partile sale stabile si antistabile obtinem

$$(\mathcal{G}u)(t) = \int_{-\infty}^{t} C_- e^{A_-(t-\tau)} B_- u(\tau)\mathrm{d}\tau - \int_{t}^{\infty} C_+ e^{A_+(t-\tau)} B_+ u(\tau)\mathrm{d}\tau + D$$

(119)

Deoarece $A$ este dihotomica,

$$\mathbf{G}(s) = C(sI - A)^{-1}B + D \in RL^{\infty}_{p\times m}.$$

Daca $y \in L^{2,p}, u \in L^{2,m}$ sunt a.i.

$$y = \mathcal{G}u,$$

atunci transformatele Fourier $\widehat{y}, \widehat{u} \in L^2(\mathbb{C}_0)$,

$$\widehat{y}(j\omega) = \mathbf{G}(j\omega)\widehat{u}(j\omega), a.p.t.\omega$$

**Teorema 118.** *Daca sistemul este dihotomic, atunci norma operatorului intrare–iesire este egala cu norma* $L^\infty(\mathbb{C}_0)$ *matricii sale de transfer, i.e.,*

$$\|\mathcal{G}\| = \|\mathbf{G}\|_\infty. \tag{120}$$

Avem urmatorul rezultat in cazul unui sistem inversabil (avand $D$ inversabila).

**Propozitia 119.** *Fie $D$ inversabila si $A - BD^{-1}C$ este dihotomica. Atunci operatorul intrare–iesire $\mathcal{G}$ este inversabil si inversa sa este operatorul intrare–iesire al sistemului*

$$\begin{array}{rcrcl} \dot{x}(t) & = & (A - BD^{-1}C)x(t) & + & BD^{-1}y(t), \\ u(t) & = & -D^{-1}Cx(t) & + & D^{-1}y(t). \end{array} \tag{121}$$

# Capitolul 11: OPTIMIZARE PATRATICA & TRIPLETE POPOV

Definitii si Echivalenta

Transformari sub Echivalenta

- Indici Patratici
- Sistemul Riccati
- Ecuatia Matriciala Algebrica Riccati
- Sistemul Kalman–Yakubovich–Popov
- Sistemul Hamiltonian
- Functia Popov
- Operatorul I/O al Sistemului Hamiltonian

# Definitii si Echivalenta

**Definitia 120. [Triplet Popov]** *Un triplet de matrici*

$$\Sigma := (A, B; P), \quad P := \begin{bmatrix} Q & L \\ L^T & R \end{bmatrix} = P^T,$$

*unde* $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $Q = Q^T \in \mathbb{R}^{n\times n}$, $L \in \mathbb{R}^{n\times m}$, $R = R^T \in \mathbb{R}^{m\times m}$, *si* $P \in \mathbb{R}^{(n+m)\times(n+m)}$, *se numeste* triplet Popov.

- Semnificatie: Reprezentare pe scurt a unei perechi constand din:
  - Un sistem dinamic controlat

$$\dot{x} = Ax + Bu \tag{122}$$

– Un indice patratic de performanta

$$J = \int_{t_0}^{t_1} w^T(t) P w(t) \mathrm{d}t, \quad w(t) := \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}, \tag{123}$$

unde $x(t)$ si $u(t)$ satisfac (122).

- Reprezentare alternativa explicita a unui triplet Popov ca $\Sigma = (A, B; Q, L, R)$

**Definitia 121. [Echivalenta si triplete Popov]** *Doua triplete Popov*

$$\begin{aligned} \Sigma &= (A, B; Q, L, R), \\ \widetilde{\Sigma} &= (\widetilde{A}, \widetilde{B}; \widetilde{Q}, \widetilde{L}, \widetilde{R}), \end{aligned}$$

*se numesc* $(X, F)$*–echivalente daca exista* $F \in \mathbb{R}^{m \times n}$ *si* $X = X^T$

$\in \mathbb{R}^{n\times n}$ *a.i.*

$$\begin{aligned}
\widetilde{A} &= A + BF, \\
\widetilde{B} &= B, \\
\widetilde{Q} &= Q + LF + F^T L^T + F^T RF + \widetilde{A}^T X + X\widetilde{A}, \\
\widetilde{L} &= L + F^T R + XB, \\
\widetilde{R} &= R.
\end{aligned} \tag{124}$$

- *Echivalent Feedback* (*sau* $F$*–echivalent*): $X = 0$
- *Echivalent Redus* $F = 0$ *si* $X$ *a.i.* $\widetilde{Q} = 0$ *( ecuatie Lyapunov:* $Q + A^T X + XA = 0$*)*
- Relatia de mai sus este intr-adevar o relatie de echivalenta ce satisface axiomele corespunzatoare.

# Transformari sub Echivalenta

Vedem in continuare cum se transforma diversele obiecte asociate cu un indice patratic + sistem dinamic controlat (sau cu un triplet Popov) sub relatia de echivalenta :

- Indice patratic
- Sistemul algebric Riccati (ARS)
- Ecuatia algebrica Riccati (ARE)
- Sistemul Kalman-Popov-Yakubovich (KPYS)
- Sistemul Hamiltonian

- Fascicolul (matricial) Hamiltonian extins (EHP)
- Functia Popov
- Operatorul I/O al sistemului Hamiltonian

# Indice Patratic

$$J_\Sigma(\xi,u):=\int_0^\infty \left[\begin{array}{c} x \\ u \end{array}\right]^T \left[\begin{array}{cc} Q & L \\ L^T & R \end{array}\right] \left[\begin{array}{c} x \\ u \end{array}\right] \mathrm{d}t = \left\langle \left[\begin{array}{c} x^{\xi,u} \\ u \end{array}\right], \left[\begin{array}{cc} Q & L \\ L^T & R \end{array}\right] \left[ \right.$$

(125)

unde

$$x^{\xi,u}(t)=e^{At}\xi+\int_0^t e^{A(t-\tau)}Bu(\tau)\mathrm{d}\tau, \quad t\geq 0 \tag{126}$$

este solutia

$$\dot{x}=Ax+Bu, \quad x(0)=\xi \tag{127}$$

$$u\in\mathcal{U}_\xi:=\{u: u\in L_+^{2,m} \text{ a.i. } x^{\xi,u}\in L_+^{2,n}\} \tag{128}$$

- Reamintire: Multimea $\mathcal{U}_\xi$ contine acele intrari care fac solutia $x^{\xi,u}$ de patrat integrabil. Aceste intrari stabilizeaza sistemul in bucla deschisa !!!

- Daca $A$ este stabila, $J_\Sigma(\xi, u)$ este definita pentru toti $\xi \in \mathbb{R}^n$ si $u \in L_+^{2,m}$ i.e. $\mathcal{U}_\xi = L_+^{2,m}$.

**Propozitia 122.** *Fie $\Sigma = (A, B; P)$ si fie $\widetilde{\Sigma} = (\widetilde{A}, \widetilde{B}; \widetilde{P})$ un $(X, F)$ echivalent. Atunci:*

$J_\Sigma(\xi, u) = J_{\widetilde{\Sigma}}(\xi, \widetilde{u}) + \xi^T X \xi, \quad (u \in \mathcal{U}_\xi)$ *unde* $\widetilde{u} = -Fx^{\xi,u} + u$.

# Sistemul Algebric Riccati

Sistemul de ecuatii in necunoscutele $F \in \mathbb{R}^{m\times n}, X = X^T \in \mathbb{R}^{n\times n}$,

$$\underbrace{\begin{bmatrix} A^TX + XA + Q & L + XB \\ B^TX + L^T & R \end{bmatrix}}_{\text{Matricea de Disipativitate}} \begin{bmatrix} I \\ F \end{bmatrix} = 0 \qquad (129)$$

se numeste sistemul algebric Riccati (ARS) asociat cu $\Sigma$ : ARS($\Sigma$).

**Definitia 123.** *O pereche $(X, F)$ satisfacand (129), cu $X = X^T$ si $A + BF$ stabila, se numeste solutie stabilizanta a ARS($\Sigma$).*

**Propozitia 124.** *Fie $(X, F)$, $(\widehat{X}, \widehat{F})$ solutii stabilizante ale ARS($\Sigma$). Atunci $X = \widehat{X}$.*

Observatie: $F$ este in general ne–unica daca $R$ nu este inversabila!!!

**Propozitia 125.** *Fie* $\Sigma = (A, B; P)$ *si* $\widetilde{\Sigma} = (\widetilde{A}, \widetilde{B}; \widetilde{P})$ *un* $(X, F)$*–equivalent.*

$(X_s, F_s)$ *este o solutie (stabilizanta) a ARS(*$\Sigma$*)* $\Leftrightarrow$

$(X_s - X, F_s - F)$ *este o solutie (stabilizanta) a ARS(*$\widetilde{\Sigma}$*).*

# Ecuatia Algebrica Riccati (ARE)

Ecuatia algebrica Riccati este un caz particular al sistemului Riccati atunci cand $R$ este nensingulara. Intr-adevar, avem pentru $X = X^T$,

$$\begin{gathered} F = -R^{-1}(B^T X + L^T), \\ A^T X + XA - (XB + L)R^{-1}(B^T X + L^T) + Q = 0 \end{gathered} \tag{130}$$

care este ecuatia algebrica Riccati asociata cu $\Sigma$.

**Definitia 126.** *O solutie simetrica $X$ se numeste stabilizanta daca $A + BF$ este stabila. $F$ se numeste reactie stabilizanta.*

**Corolarul 127.** *Daca ARE($\Sigma$) are o solutie stabilizanta atunci aceasta este unica.*

**Corolarul 128.** *Fie* $\Sigma = (A, B; P)$ *si* $\widetilde{\Sigma} = (\widetilde{A}, \widetilde{B}; \widetilde{P})$ *doua triplete Popov* $(X, F)$ *echivalente.* $X_s$ *este solutia stabilizanta a ARE(*$\Sigma$*)*

$$\Leftrightarrow$$

$X_s - X$ *este solutia stabilizanta a ARE(*$\widetilde{\Sigma}$*). Mai mult,*

$$\widetilde{F}_{Ricc} = F_{Ricc} - F. \tag{131}$$

# Sistemul Kalman–Yakubovich–Popov

O inlocuire comoda a ecuatiei Riccati folosita pentru a completa forma patrata.

Fie $\Sigma = (A, B; Q, L, R)$, $R$ inversabila, $J$ o $m \times m$ matrice de semn. Sistemul neliniar

$$\begin{aligned} R &= V^T J V, \\ L + XB &= W^T J V, \\ Q + A^T X + XA &= W^T J W, \end{aligned} \tag{132}$$

se numeste sistemul Kalman–Popov–Yakubovich (KPYS) in forma $J$: KPYS$(\Sigma, J)$.

NOTA: $R$ inversabila $\Rightarrow$ $J$ inversabila

$$J = \begin{bmatrix} -I_{m_1} & \\ & I_{m_2} \end{bmatrix}, \quad \operatorname{sgn}(R) = J. \tag{133}$$

**Definitia 129.** *Un triplet de matrici* $(X = X^T, V, W)$ *ce satisface (132) a.i.* $A + BF$ *este stabila, pentru*

$$F := -V^{-1}W, \tag{134}$$

*se numeste solutie stabilizanta a sistemului KPYS*$(\Sigma, J)$ *si* $F$ *se numeste reactia stabilizanta.*

NOTA:

- Daca $(X = X^T, V, W)$ este o solutie a KPYS$(\Sigma, J) \Rightarrow X$ este o solutie simetrica a ARE$(\Sigma)$.

- Daca $X$ este o solutie simetrica a ARE($\Sigma$), atunci putem alege $V$ a.i. $R = V^T J V$ (cu conditia $\operatorname{sgn}(R) = J$). Atunci luam $W := (L + XB)V^{-1}J^{-1}$ care satisface deasemenea ultima ecuatie. Prin urmare $(X, V, W)$ este o solutie a KPYS($\Sigma, J$).

- Reactia Riccati coincide cu reactia sistemului KPYS $F := -V^{-1}W$.

CONCLUZII:

- Sistemul KPYS($\Sigma, J$) are o solutie (stabilizanta) $(X = X^T, V, W)$ daca si numai daca ARE($\Sigma$) are o solutie (stabilizanta) $X$ si $\operatorname{sgn}(R) = J$.

- Daca $(X, V, W)$, $(\widehat{X}, \widehat{V}, \widehat{W})$ sunt solutii stabilizante ale KPYS($\Sigma, J$), atunci $X = \widehat{X}$.

- Daca $J = I_m$ sau $J = -I_m$, atunci KPYS devine KPYS in forma de pozitivitate

$$\begin{aligned} R &= V^T V, \\ L + XB &= W^T V, \\ Q + A^T X + XA &= W^T W, \end{aligned} \tag{135}$$

sau in forma de negativitate

$$\begin{aligned} R &= -V^T V, \\ L + XB &= -W^T V, \\ Q + A^T X + XA &= -W^T W. \end{aligned} \tag{136}$$

**Propozitia 130.** *Fie $\Sigma$ un triplet Popov si $(X, V, W)$ o solutie a*

*KPYS($\Sigma, J$). Pentru oricare $u \in \mathcal{U}_\xi$ avem*

$$J_\Sigma(\xi, u) = \left\langle Vu + Wx^{\xi,u}, J(Vu + Wx^{\xi,u}) \right\rangle_{L_+^{2,m}} + \xi^T X \xi. \quad (137)$$

**Propozitia 131.** *Fie $\Sigma = (A, B; P)$ si $\widetilde{\Sigma} = (\widetilde{A}, \widetilde{B}; \widetilde{P})$ doua triplete Popov $(X, F)$ echivalente.*

*$(X_s, V_s, W_s)$ este o solutie (stabilizanta) pentru KPYS($\Sigma, J$)*

$\Leftrightarrow$

*$(X_s - X, V_s, W_s + V_s F)$ este o solutie (stabilizanta) pentru KPYS($\widetilde{\Sigma}, J$).*

# Sistemul Hamiltonian

Fie $\Sigma = (A, B; Q, L, R)$. Sistemul diferential liniar

$$\begin{array}{rcrrr} \dot{x} & = & Ax & & + Bu, \\ \dot{\lambda} & = & -Qx & - A^T\lambda & - Lu, \\ v & = & L^T x & + B^T\lambda & + Ru, \end{array} \tag{138}$$

se numeste sistemul Hamiltonian cu timp continuu: HS($\Sigma$).

Ne intereseaza acele functii $(u, x, \lambda)$ care fac $v = 0$ pentru sistemul Hamiltonian (determinam in acest fel punctele critice).

**Propozitia 132.** *Fie* $\Sigma = (A, B; P)$ *si* $\widetilde{\Sigma} = (\widetilde{A}, \widetilde{B}; \widetilde{P})$ $(X, F)$ *echivalente.*

$(u, x, \lambda) \Rightarrow v = 0$ *pentru HS($\Sigma$) (pentru un anumit interval de timp)*

$\Leftrightarrow$

$(u - Fx, x, \lambda - Xx) \Rightarrow v = 0$ *pentru HS($\widetilde{\Sigma}$) (pe acelasi interval de timp).*

# Fascicolul (matricial) Hamiltonian extins (EHP)

Fie $\Sigma = (A, B; Q, L, R)$. Fascicolul matricial $sM_\Sigma - N_\Sigma$, cu

$$M_\Sigma := \begin{bmatrix} I & O & O \\ O & I & O \\ O & O & O \end{bmatrix}, \quad N_\Sigma := \begin{bmatrix} A & O & B \\ -Q & -A^T & -L \\ L^T & B^T & R \end{bmatrix}, \qquad (139)$$

se numeste fascicolul Hamiltonian extins: EHP($\Sigma$).

Observatii:

- Matricile $M_\Sigma$ si $N_\Sigma$ sunt patrate $(2n + m) \times (2n + m)$, dar $sM_\Sigma - N_\Sigma$ nu este o problema de valori proprii standard intrucat $M_\Sigma$ este singular !!!

- Cu toate acestea, daca $R$ este inversabil, avem

$$Q(\lambda M_\Sigma - N_\Sigma)Z = \lambda \begin{bmatrix} I_{2n} & O \\ O & O \end{bmatrix} - \begin{bmatrix} H_\Sigma & O \\ O & I_m \end{bmatrix}, \qquad (140)$$

$$Q = \begin{bmatrix} I_n & O & -BR^{-1} \\ O & I_n & LR^{-1} \\ O & O & I_m \end{bmatrix}, \qquad Z = \begin{bmatrix} I_n & O & O \\ O & I_n & O \\ -R^{-1}L^T & -R^{-1}B^T & R^{-1} \end{bmatrix}$$

$$H_\Sigma := \begin{bmatrix} A - BR^{-1}L^T & -BR^{-1}B^T \\ -Q + LR^{-1}L^T & -A^T + LR^{-1}B^T \end{bmatrix}. \qquad (141)$$

Deci avem o problema de valori proprii clasica $sI_{2n} - H_\Sigma$ pentru matricea Hamiltoniana $H_\Sigma$.

- EHP($\Sigma$) este exact fascicolul de transmisie (sistem) al HS($\Sigma$). Cum suntem interesati in solutii ce anuleaza iesirea sistemului HS este

natural ca fascicolul sistem sa joace un rol central in caracterizarea acestora.

- Daca luam $v = 0$ in HS($\Sigma$) obtinem un sistem descriptor diferential

$$z(t) := \begin{bmatrix} x(t) \\ \lambda(t) \\ u(t) \end{bmatrix} \Rightarrow M_\Sigma \dot{z} = N_\Sigma z.$$

**Propozitia 133.** *Fie* $\Sigma = (A, B; Q, L, R)$, $\widetilde{\Sigma} = (\widetilde{A}, \widetilde{B}; \widetilde{Q}, \widetilde{L}, \widetilde{R})$ *doua triplete Popov* $(X, F)$ *echivalente. Atunci EHP($\Sigma$) si EHP($\widetilde{\Sigma}$) sunt legate prin*

$$sM_\Sigma - N_\Sigma = U(sM_{\widetilde{\Sigma}} - N_{\widetilde{\Sigma}})V, \tag{142}$$

*unde*

$$U := \begin{bmatrix} I & O & O \\ X & I & F^T \\ O & O & I \end{bmatrix}, \quad V := \begin{bmatrix} I & O & O \\ -X & I & O \\ -F & O & I \end{bmatrix}. \tag{143}$$

# Functia Popov (in timp continuu)

Fie $\Sigma = (A, B; Q, L, R)$. Pentru $s \in \mathbb{C} \backslash \{\Lambda(A) \cup \Lambda(-A^T)\}$ definim functia matriciala rationala

$$\mathbf{\Pi}_\Sigma(s) := [B^T(-sI - A^T)^{-1} \; I] \begin{bmatrix} Q & L \\ L^T & R \end{bmatrix} \begin{bmatrix} (sI - A)^{-1}B \\ I \end{bmatrix}, \tag{144}$$

ce se numeste functia lui Popov.

Nota:

- $\mathbf{\Pi}_\Sigma = \left[\begin{array}{cc|c} A & O & B \\ -Q & -A^T & -L \\ \hline L^T & B^T & R \end{array}\right]$, i.e., $\mathbf{\Pi}_\Sigma$ este exact matricea de transfer de la $u$ la $v$ a HS($\Sigma$).

- EHP($\Sigma$) este exact fascicolul de transmisie asociat cu aceasta realizarea a functiei Popov.

- $\Pi_\Sigma$ este auto–adjuncta, i.e., are urmatoarea simetrie : $\Pi_\Sigma(s) = \Pi_\Sigma^T(-s) \quad (=: \Pi_\Sigma^*(s))$.

- $\Pi_\Sigma$ se numeste si densitate spectrala – introdusa de Popov in celebra monografie: "Hiperstabilitatea Sistemelor Automate" (1960); joaca un rol major in identificari, prelucrarea semnalelor, sisteme automate, teoria circuitelor, etc.

# Factorizare Spectrala

**Propozitia 134.** *Fie* $\Sigma = (A, B; P)$, $\widetilde{\Sigma} = (\widetilde{A}, \widetilde{B}; \widetilde{P})$ *doua triplete* $(X, F)$ *echivalente.*

1. *Pentru* $s \in \mathbb{C} \setminus \{\Lambda(A) \cup \Lambda(-A^T) \cup \Lambda(\widetilde{A}) \cup \Lambda(-\widetilde{A}^T)\}$

$$\mathbf{\Pi}_{\Sigma}(s) = \mathbf{S}_F^*(s)\mathbf{\Pi}_{\widetilde{\Sigma}}(s)\mathbf{S}_F(s), \tag{145}$$

*unde*

$$\mathbf{S}_F := \left[\begin{array}{c|c} A & B \\ \hline -F & I \end{array}\right]. \tag{146}$$

2. *Mai mult, daca* $(X_s = X_s^T, F_s)$ *este o solutie a ARS(*$\Sigma$*) (sau ARE),*

*atunci*

$$\mathbf{\Pi}_{\Sigma}(s) = \mathbf{S}^{*}_{F_s}(s) R \mathbf{S}_{F_s}(s) \tag{147}$$

*unde* $\mathbf{S}_{F_s}$ *este (146) scrisa pentru* $F = F_s$. *(Identitatea de factorizare spectrala )*

# Capitolul 12: TEORIE RICCATI: ABORDARE DINAMICA SI FRECVENTIALA

Operatorul I/O al Sistemului Hamiltonian

Principalul Rezultat in Domeniul Timp

Cazul Clasic de Pozitiviate: LQP

Ridicarea Ipotezei de Stabilitate

Conditia de Signatura

Optimizare Maxmin/Jocuri Dinamice

Conditii Frecventiale

## Conditia de Signatura Frecventiala

## Inegalitati matriciale Riccati

# Operatorul I/O al Sistemului Hamiltonian

Fixam un triplet Popov $\Sigma = (A, B; Q, L, R)$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, si investigam expresiile indicilor patratici pentru diverse solutii ale ecuatiilor diferentiale. Fie $A$ dihotomica. Investigam intai criteriul patratic extins

$$J_{e,\Sigma}(u) := \int_{-\infty}^{\infty} \begin{bmatrix} x \\ u \end{bmatrix}^T \begin{bmatrix} Q & L \\ L^T & R \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} \mathrm{d}t$$

unde $u \in L^{2,m}$ si $x$ este solutia pe intreaga axa a ecuatiei diferentiale liniare $\dot{x} = Ax + Bu$. Substituind solutia $L^{2,n}$ data de $x_e^u = \mathcal{L}_e u$,

obtinem

$$\begin{aligned} J_{e,\Sigma}(u) &= \left\langle \begin{bmatrix} x_e^u \\ u \end{bmatrix}, \begin{bmatrix} Q & L \\ L^T & R \end{bmatrix} \begin{bmatrix} x_e^u \\ u \end{bmatrix} \right\rangle \\ &= \left\langle u, \begin{bmatrix} \mathcal{L}_e^* & I \end{bmatrix} \begin{bmatrix} Q & L \\ L^T & R \end{bmatrix} \begin{bmatrix} \mathcal{L}_e \\ I \end{bmatrix} u \right\rangle = \langle u, \mathcal{R}_{e,\Sigma} u \rangle \end{aligned} \tag{148}$$

unde $\mathcal{R}_{e,\Sigma}$ este un operator liniar marginit din $L^{2,m}$ in $L^{2,m}$ dat de

$$\mathcal{R}_{e,\Sigma} := R + L^T \mathcal{L}_e + \mathcal{L}_e^* L + \mathcal{L}_e^* Q \mathcal{L}_e = \mathcal{R}_{e,\Sigma}^*. \tag{149}$$

**Teorema 135.** *Fie $\Sigma = (A, B; Q, L, R)$ un triplet Popov. Daca $A$ este dihotomica, atunci operatorul $\mathcal{R}_{e,\Sigma}$ este operatorul $L^2$ intrare–iesire al sistemului Hamiltonian.*

Fie $A$ stabila. Pentru fiecare $u \in L_+^{2,m}$ indicele patratic $J_\Sigma(\xi, u)$

este bine definit. Inlocuim solutia $x^{\xi,u} = \Phi\xi + \mathcal{L}u$, pentru $x(0) = \xi$ :

$$\begin{aligned} J_\Sigma(\xi,u) &= \left\langle \begin{bmatrix} x^{\xi,u} \\ u \end{bmatrix}, \begin{bmatrix} Q & L \\ L^T & R \end{bmatrix} \begin{bmatrix} x^{\xi,u} \\ u \end{bmatrix} \right\rangle \\ &= \left\langle \begin{bmatrix} \Phi & \mathcal{L} \\ O & I \end{bmatrix} \begin{bmatrix} \xi \\ u \end{bmatrix}, \begin{bmatrix} Q & L \\ L^T & R \end{bmatrix} \begin{bmatrix} \Phi & \mathcal{L} \\ O & I \end{bmatrix} \begin{bmatrix} \xi \\ u \end{bmatrix} \right\rangle \\ &= \left\langle \begin{bmatrix} \xi \\ u \end{bmatrix}, \begin{bmatrix} \Phi^* & O \\ \mathcal{L}^* & I \end{bmatrix} \begin{bmatrix} Q & L \\ L^T & R \end{bmatrix} \begin{bmatrix} \Phi & \mathcal{L} \\ O & I \end{bmatrix} \begin{bmatrix} \xi \\ u \end{bmatrix} \right\rangle \\ &= \left\langle \begin{bmatrix} \xi \\ u \end{bmatrix}, \begin{bmatrix} \mathcal{P}_{o,\Sigma} & \mathcal{P}_\Sigma \\ \mathcal{P}_\Sigma^* & \mathcal{R}_\Sigma \end{bmatrix} \begin{bmatrix} \xi \\ u \end{bmatrix} \right\rangle, \end{aligned} \tag{150}$$

in care toti operatorii ce intervin sunt liniar marginiti:

$$\mathcal{P}_{o,\Sigma} : \mathbb{R}^n \to \mathbb{R}^n, \quad \mathcal{P}_{o,\Sigma} := \Phi^* Q \Phi, \tag{151}$$

$$\mathcal{P}_\Sigma : L_+^{2,m} \to \mathbb{R}^n, \quad \mathcal{P}_\Sigma := \Phi^*(Q\mathcal{L} + L), \tag{152}$$

$$\mathcal{R}_\Sigma : L_+^{2,m} \to L_+^{2,m}, \quad \mathcal{R}_\Sigma := R + L^T\mathcal{L} + \mathcal{L}^*L + \mathcal{L}^*Q\mathcal{L} = \mathcal{R}_\Sigma^*. \tag{153}$$

NOTA: Daca $A$ este stabila atunci $\mathcal{R}_\Sigma$ este operatorul Toeplitz avand drept simbol functia Popov $\mathbf{\Pi}_\Sigma$. Intr–adevar, functia Popov este matricea de transfer a sistemului Hamiltonian si $\mathcal{R}_\Sigma$ este operatorul Toeplitz asociat cu operatorul intrare–iesire al sistemului Hamiltonian $\mathcal{R}_{e,\Sigma}$.

**Definitia 136.** *Fie $\mathcal{G}$ un operator liniar marginit din $L^{2,m}$ in $L^{2,p}$. Definim*

1. Operatorul cauzal Hankel *asociat cu $\mathcal{G}$:* $\mathbf{H}_\mathcal{G} := P_+^p\mathcal{G}|_{L_-^{2,m}}$

2. Operatorul Hankel anticauzal *asociat cu $\mathcal{G}$:* $\widehat{\mathbf{H}}_\mathcal{G} := P_-^p\mathcal{G}|_{L_+^{2,m}}$.

*3.* Operatorul Toeplitz cauzal *asociat cu* $\mathcal{G}$, $\mathbf{T}_{\mathcal{G}} := P_+^p \mathcal{G}|_{L_+^{2,m}}$.

*4.* Operatorul Toeplitz anticauzal *asociat cu* $\mathcal{G}$: $\widehat{\mathbf{T}}_{\mathcal{G}} := P_-^p \mathcal{G}|_{L_-^{2,m}}$

*unde* $P_\pm^r$ *sunt proiectiile ortogonale uzuale ale lui* $L^{2,r}$ *pe* $L_\pm^{2,r}$.

# Rezultatul Central

Dat $\Sigma = (A, B; Q, L, R)$ cu $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, scopul principal este obtinerea conditiilor necesare si suficiente de existenta a solutiei stabilizatoare a ARE($\Sigma$)

$$A^T X + XA - (XB + L)R^{-1}(B^T X + L^T) + Q = 0, \qquad (154)$$

in termenii operatorului Toeplitz $\mathcal{R}_\Sigma$ introdus mai sus.

$$\mathcal{R}_\Sigma : L_+^{2,m} \to L_+^{2,m}, \quad \mathcal{R}_\Sigma := R + L^T\mathcal{L} + \mathcal{L}^* L + \mathcal{L}^* Q \mathcal{L} = \mathcal{R}_\Sigma^*. \quad (155)$$

**Teorema 137.** *Fie $\Sigma = (A, B; Q, L, R)$ un triplet Popov cu A stabila. Urmatoarele afirmatii sunt echivalente:*

*1. $R$ este nesingulara si ARE($\Sigma$) are o (unica) solutie stabilizanta*

$$X := \mathcal{P}_{o,\Sigma} - \mathcal{P}_\Sigma \mathcal{R}_\Sigma^{-1} \mathcal{P}_\Sigma^*$$

*2. Operatorul Toeplitz $\mathcal{R}_\Sigma$ (avand drept simbol functia Popov) are o inversa marginita*

**Corolarul 138.** *Fie $\Sigma = (A, B; Q, L, R)$ un triplet Popov cu $A$ stabila. Presupunem ca operatorul Toeplitz $\mathcal{R}_\Sigma$ are o inversa marginita si fie $X$ solutia stabilizanta a ARE($\Sigma$). Atunci avem*

$$J_\Sigma(\xi, u^\xi) = \left\langle \begin{bmatrix} x^{\xi,u^\xi} \\ u \end{bmatrix}, \begin{bmatrix} Q & L \\ L^T & R \end{bmatrix} \begin{bmatrix} x^{\xi,u^\xi} \\ u \end{bmatrix} \right\rangle_{L_+^{2,n} \times L_+^{2,m}} = \xi^T X \xi \tag{156}$$

*pentru orice $\xi \in \mathbb{R}^n$ si $u^\xi = -\mathcal{R}_\Sigma^{-1} \mathcal{P}_\Sigma^* \xi$.*

Nota:

- Nu se face nici o presupunere privind pozitivitatea lui $R$, $Q$, sau a altor matrici. Putem deci formula in acest cadru jocuri necooperative, teoriile de tip Nash, puncte sa, $H^\infty$, etc.

- In plus avem formule operatoriale pentru solutia stabilizanta a ecuatiei Riccati si pentru valoarea criteriului patratic in punctul critic.

# Cazul Clasic de Pozitivitate: LQP

Presupunerea de baza aici este ca matricea $R$ este pozitiv definita (coeficientul termenului patratic).

**Corolarul 139.** *Fie $\Sigma = (A, B; Q, L, R)$ un triplet Popov cu $A$ stabila.*

1. • *$R > 0$ si ARE($\Sigma$) are o solutie stabilizanta $X$*
   $\Leftrightarrow$
   • *Operatorul $\mathcal{R}_{e,\Sigma}$ este coerciv*

2. *Daca operatorul $\mathcal{R}_{e,\Sigma}$ este coerciv atunci minimul indicelui patratic*

*este*

$$\min_{u\in L_+^{2,m}} J_\Sigma(\xi,u)=\xi^T X\xi,\quad \forall \xi\in\mathbb{R}^n, \tag{157}$$

*si minimul se atinge pentru* $u=u^\xi=-\mathcal{R}_\Sigma^{-1}\mathcal{P}_\Sigma^*\xi$ *si satisface deasemenea* $u^\xi=Fx^\xi$, *unde* $F:=-R^{-1}(B^TX+L^T)$ *este reactia Riccati.*

3. *Daca in plus*

$$\left[\begin{array}{cc} Q & L \\ L^T & R \end{array}\right]\geq 0, \tag{158}$$

*si operatorul* $\mathcal{R}_{e,\Sigma}$ *este coerciv, atunci* $X\geq 0$.

Putem relaxa conditia de coercivitate precum urmeaza.

**Corolarul 140.** *Fie* $\Sigma=(A,B;Q,L,R)$ *un triplet Popov cu* $A$ *stabila. Presupunem ca* $R>0$ *si ca perechea* $(A,B)$ *este controlabila.*

*Daca*

$$\mathcal{R}_{e,\Sigma} \geq 0, \tag{159}$$

*atunci exista o solutie simetrica $X$ a ARE($\Sigma$), a.i. pentru $F := -R^{-1}(B^T X + L^T)$ sa avem $\Lambda(A+BF) \subset \mathbb{C}_- \cup \mathbb{C}_0$. (Numim $X$ o solutie* semi–stabilizanta *a ARE).*

**Corolarul 141. [Ecuatia Riccati clasica de control]** *Fie $\Sigma_s = (A, B; C^T C, 0, I)$ tripletul Popov standard de control si presupunem ca $(A, B)$ si $(C, A)$ sunt stabilizabila si respectiv detectabila. Atunci ARE($\Sigma_s$)*

$$A^T X + XA - XBB^T X + C^T C = 0 \tag{160}$$

*are o solutie stabilizanta pozitiv semidefinita.*

# Ridicarea Ipotezei de Stabilitate

Folosind o reactie este trivial sa relaxam ipoteza de stabilitate in cea de stabilizabilitate asupra perechii $(A, B)$.

NOTA: Daca ARE($\Sigma$) are o solutie stabilizanta atunci automat perechea $(A, B)$ trebuie sa fie stabilizabila.

**Teorema 142.** *Fie $\Sigma = (A, B; Q, L, R)$ un triplet Popov cu $(A, B)$ stabilizabila. Atunci urmatoarele doua afirmatii sunt echivalente.*

1. *$R$ este nensingulara si ARE($\Sigma$) are o (unica) solutie stabilizanta.*

2. *Exista $F \in \mathbb{R}^{m \times n}$ a.i. $A + BF$ este stabila si $F$–echivalentul lui $\Sigma$,*

$\widetilde{\Sigma} = (\widetilde{A}, \widetilde{B}; \widetilde{Q}, \widetilde{L}, \widetilde{R})$, *are un operator Toeplitz asociat* $\mathcal{R}_{\widetilde{\Sigma}}$ *ce este inversabil.*

NOTA: Daca punctul 2 al Teoremei este adevarat pentru un $F$ stabilizant, atunci are loc pentru orice $F$ stabilizant.

# Conditia de Signatura

Am vazut ca existenta unei solutii stabilizante a ecuatiei ARE este intim legata de existenta unei inverse marginite a operatorului Toeplitz $\mathcal{R}_\Sigma$. Un caz simplu in care inversabilitatea acestui operator este garantata este cazul de pozitivitate.

In continuare elaboram o conditie mai sofisticata pentru inversabilitatea operatorului Toeplitz $\mathcal{R}_\Sigma$. Aceasta conditie numita conditia de signatura, este relevanta pentru jocurile dinamice (Nash) si va juca un rol central in rezolvarea celor mai importante probleme din teoria sistemelor optimale si robuste.

Fie $\Sigma = (A, B; Q, L, R)$, cu $A \in \mathbb{R}^{n\times n}$ stabila, $B \in \mathbb{R}^{n\times m}$.

Partitionam

$$B = [B_1 \quad B_2],\ B_i \in \mathbb{R}^{n\times m_i},\ i = 1, 2,\ m_1 + m_2 = m, \tag{161}$$

si matricile $L$ si $R$ le partitionam corespunzator in acord cu $B$, i.e.,

$$L = [L_1 \quad L_2], \quad R = \begin{bmatrix} R_{11} & R_{12} \\ R_{12}^T & R_{22} \end{bmatrix}. \tag{162}$$

Solutia problemei cu conditii initiale

$$\dot{x} = Ax + B_1u_1 + B_2u_2\ , \quad x(0) = \xi, \tag{163}$$

se scrie ca

$$x^{\xi,u_1,u_2} = \Phi\xi + \mathcal{L}_1u_1 + \mathcal{L}_2u_2.$$

Avem deasemenea

$$\mathcal{R}_\Sigma = \left[ \begin{array}{cc} \mathcal{R}_{11} & \mathcal{R}_{12} \\ \mathcal{R}_{12}^* & \mathcal{R}_{22} \end{array} \right]. \tag{164}$$

NOTA: $\mathcal{R}_{22}$ coincide cu operatorul Toeplitz $\mathcal{R}_{\Sigma_2}$ asociat cu tripletul Popov

$$\Sigma_2 = (A, B_2; Q, L_2, R_{22}). \tag{165}$$

**Teorema 143.** *Fie $\Sigma = (A, B; Q, L, R)$ un triplet Popov avand matricile partitonate ca mai sus si $A$ stabila. Daca $\mathcal{R}_{22} >> 0$ si*

$$\mathcal{R}_{11}^\times = \mathcal{R}_{11} - \mathcal{R}_{12}\mathcal{R}_{22}^{-1}\mathcal{R}_{12}^* << 0, \tag{166}$$

*atunci*

*1. ARE($\Sigma_2$) are o solutie stabilizanta $X_2$.*

*2. ARE($\Sigma$) are o solutie stabilizanta $X$ a.i.*

$$X \geq X_2. \tag{167}$$

# Probleme Maxmin /Optimizarea Dinamica a Jocurilor

**Teorema 144.** *Aceleasi presupuneri ca in teorema precedenta. Fie*

$$J_\Sigma(\xi,u_1,u_2)=\left\langle \begin{bmatrix} x^{\xi,u_1,u_2} \\ u_1 \\ u_2 \end{bmatrix}, \begin{bmatrix} Q & L_1 & L_2 \\ L_1^T & R_{11} & R_{12} \\ L_2^T & R_{12}^T & R_{22} \end{bmatrix} \begin{bmatrix} x^{\xi,u_1,u_2} \\ u_1 \\ u_2 \end{bmatrix} \right\rangle$$

*indicele patratic definit pentru toti* $(\xi,u_1,u_2)\in\mathbb{R}^n\times L_+^{2,m_1}\times L_+^{2,m_2}$. *Avem:*

*1. Pentru fiecare* $\xi\in\mathbb{R}^n$ *si fiecare* $u_1\in L_+^{2,m_1}$, *exista un unic* $u_2$ *in*

$L_+^{2,m_2}$, *notat* $u_2^{\xi,u_1}$, *ce minimizeaza* $J_\Sigma(\xi, u_1, u_2)$, *i.e.,*

$$J_\Sigma(\xi, u_1, u_2) \geq J_\Sigma(\xi, u_1, u_2^{\xi,u_1}) \tag{168}$$

*pentru toti* $u_2 \in L_+^{2,m_2}$.

2. *Pentru fiecare* $\xi$ *in* $\mathbb{R}^n$, *exista un unic* $u_1$, *notat* $u_1^\xi$, *care maximizeaza* $J_\Sigma(\xi, u_1, u_2^{\xi,u_1})$.

3. *Daca notam* $u_2^\xi = u_2^{\xi,u_1^\xi}$, *atunci*

$$J_\Sigma(\xi, u_1^\xi, u_2^\xi) = \xi^T X \xi \tag{169}$$

*pentru fiecare* $\xi \in \mathbb{R}^n$, *unde* $X$ *este solutia stabilizanta a ARE(*$\Sigma$*).*

*Mai mult, daca $x^\xi$ este solutia corespunzand lui $u_1 = u_1^\xi$ si $u_2 = u_2^\xi$, atunci*

$$u^\xi = \left[ \begin{array}{c} u_1^\xi \\ u_2^\xi \end{array} \right] = F_{Ricc}\, x^\xi, \tag{170}$$

*unde $F_{Ricc}$ este reactia stabilizanta Riccati.*

NOTA: Concluzia Teoremei 144 se poate exprima sintetic in felul urmator:

$$\max_{u_1} \min_{u_2} J_\Sigma(\xi, u_1, u_2) = \xi^T X \xi. \tag{171}$$

Si solutia maxmin este disponibila in forma "feedback" (170).

# Conditii Frecventiale

Pana in prezent am dat un numar de conditii necesare si suficiente pentru existenta solutiei stabilizante a ARE. Aceste conditii au fost exprimate in termenii diversilor operatori ce intervin in descrierea dinamica (in domeniul timp) a sistemului Hamiltonian HS. In aceasta sectiune exprimam aceste conditii in domeniul frecvential.

Traditional, acesta a constituit un subiect major de cercetare si a fost unul dintre punctele centrale ce au facut atat de atractiva Teoria Pozitivitatii. Conditiile frecventiale erau "usor" de verificat de catre ingineri.

Era extrem de dezirabil atunci sa avem o legatura intre aceste conditii si proprietati de spatiul starilor (realizari). Odata cu

dezvoltarea tehnicilor moderne de control automat accentul s-a mutat oarecum spre tehnicile de spatiul starilor.

Existenta solutiei stabilizante a ARE($\Sigma$) a fost legata de existenta inversei marginite a operatorului Toeplitz $\mathcal{R}_\Sigma$. Exista rezultate puternice de teoria operatorilor ce leaga inversabilitatea acestui operator de proprietati de factorizare ale simbolului rational asociat. Vom folosi astfel de rezultate pentru a da complementul frecvential al rezultatelor operatoriale.

**Teorema 145. [Rezultatul central]** *Fie $\Sigma = (A, B; P)$ un triplet Popov cu $A$ stabila. Urmatoarele afirmatii sunt echivalente.*

1. *$R$ este nesingulara si ARE($\Sigma$) are o solutie stabilizanta.*

2. *Functia Popov $\mathbf{\Pi}_\Sigma$ este antianalitic factorizabila, i.e., exista $\mathbf{\Xi}$ si $\mathbf{\Omega}$, doua unitati in $RH^\infty_{+,m\times m}$ a.i.*

$$\mathbf{\Pi}_\Sigma(s) = [\mathbf{\Xi}(s)]^* \mathbf{\Omega}(s). \tag{172}$$

*In particular,*

$$\mathbf{\Omega}(s) := R(I - F_{Ricc}(sI - A)^{-1}B), \quad \mathbf{\Xi}(s) := I - F_{Ricc}(sI - A)^{-1}B$$

**Teorema 146. [Indepartarea ipotezei de stabilitate]** *Fie* $\Sigma =$

$(A, B; P)$ *un triplet Popov cu* $(A, B)$ *stabilizabila. Atunci urmatoarele doua afirmatii sunt echivalente.*

1. $R$ *este nensingulara si ARE(*$\Sigma$*) are o solutie stabilizanta.*

2. *Functia Popov* $\mathbf{\Pi}_\Sigma$ *este antianalitic prefactorizabila, i.e., exista un* $F$ *a.i. pentru tripletul Popov* $F$*–echivalent* $\widetilde{\Sigma}$ *sa avem* $\mathbf{\Pi}_{\widetilde{\Sigma}}$ *este analitic factorizabila.*

# Cazul de Pozitivitate: $A$ Stabila

**Corolarul 147.** *Fie* $\Sigma = (A, B; Q, L, R)$ *un triplet Popov cu* $A$ *stabila.*

1. $R > 0$ *si ARE(*$\Sigma$*) are o solutie stabilizanta* $X$
   $\Leftrightarrow$
   $\mathbf{\Pi}_\Sigma > 0$ *(pe axa imaginara)*

2. *Daca oricare afirmatie de la 1. este indeplinita, minimul indicelui patratic este* $\min_{u \in L_+^{2,m}} J_\Sigma(\xi, u) = \xi^T X \xi \ \forall \xi \in \mathbb{R}^n$ *si se atinge pentru* $u^\xi = F x^\xi$, *unde* $F := -R^{-1}(B^T X + L^T)$ *este reactia Riccati stabilizanta.*

*3. Daca*

$$\begin{bmatrix} Q & L \\ L^T & R \end{bmatrix} \geq 0, \tag{173}$$

*si oricare afirmatie 1. este indeplinita atunci* $X \geq 0$.

**Corolarul 148. [Relaxarea ipotezei de pozitivitate]** *Fie* $\Sigma = (A, B; Q, L, R)$ *un triplet Popov cu* $A$ *stabila. Presupunem* $R > 0$ *si* $(A, B)$ *este controlabila. Daca* $\mathbf{\Pi}_\Sigma \geq 0$, *atunci exista o solutie simetrica* $X$ *a ARE(*$\Sigma$*), a.i. pentru* $F := -R^{-1}(B^T X + L^T)$ *sa avem* $\Lambda(A + BF) \subset \mathbb{C}_- \cup \mathbb{C}_0$. *(Numim* $X$ *o solutie semistabilizanta a ARE).*

# Cazul de Pozitivitate: $A$ arbitrara

**Corolarul 149.** *Fie* $\Sigma = (A, B; Q, L, R)$ *un triplet Popov. Presupunem ca :*

**a.** *Perechea* $(A, B)$ *este stabilizabila.*

**b.** *Realizarea lui* $\mathbf{\Pi}_\Sigma$,

$$\mathbf{\Pi}_\Sigma = \left[\begin{array}{cc|c} A & O & B \\ -Q & -A^T & -L \\ \hline L^T & B^T & R \end{array}\right], \tag{174}$$

*este controlabila si observabila pe axa imaginara.*

**c.** $\mathbf{\Pi}_\Sigma(j\omega) > 0, \quad \forall \omega \in \overline{\mathbb{R}}.$

*Atunci ARE($\Sigma$) are o solutie stabilizanta* $X$ *si* $\min_{u \in \mathcal{U}^\xi} J_\Sigma(\xi, u) = \xi^T X \xi$.

*Daca* $\begin{bmatrix} Q & L \\ L^T & R \end{bmatrix} \geq 0$, *atunci* $X \geq 0$.

# Conditia de Signatura in Domeniul Frecvential

Prezentam in continuare o versiune in domeniul frecvential a conditii de signatura pentru operatori. De fapt vom da un rezultat mai puternic pentru sistemul KYPS($\Sigma$).

Consideram din nou tripletul Popov $\Sigma = (A, B; Q, L, R)$ cu $B$, $L$ si $R$ partitionate ca

$$B = \left[ \begin{array}{cc} B_1 & B_2 \end{array} \right], L = \left[ \begin{array}{cc} L_1 & L_2 \end{array} \right], R = \left[ \begin{array}{cc} R_{11} & R_{12} \\ R_{12}^T & R_{22} \end{array} \right], \quad (175)$$

($B_i \in \mathbb{R}^{n \times m_i}, i = 1, 2, m = m_1 + m_2$). Fie functia Popov $\mathbf{\Pi}_\Sigma$

partitionata in concordanta cu $R$,

$$\mathbf{\Pi}_\Sigma = \begin{bmatrix} \mathbf{\Pi}_{11} & \mathbf{\Pi}_{12} \\ \mathbf{\Pi}_{12}^* & \mathbf{\Pi}_{22} \end{bmatrix}. \tag{176}$$

Introducem deasemenea matricea de semn

$$J := \begin{bmatrix} -I_{m_1} & \\ & I_{m_2} \end{bmatrix}. \tag{177}$$

Notam tripletul Popov $\Sigma_2 = (A, B_2; Q, L_2, R_{22})$. Este usor de verificat ca $\mathbf{\Pi}_{22} = \mathbf{\Pi}_{\Sigma_2}$.

# O conditie suficienta

**Teorema 150.** *Fie* $\Sigma = (A, B; Q, L, R)$ *un triplet Popov cu* $A$ *stabila si intrarile partitionate ca mai sus. Presupunem* $\mathbf{\Pi}_{22} > 0$ *pe* $\overline{\mathbb{C}}_0$. *Daca exista un* $\mathbf{S} \in RH^{\infty}_{+,m_2 \times m_1}$ *a.i.*

$$\left[\begin{array}{cc} I & \mathbf{S}^* \end{array}\right] \mathbf{\Pi}_{\Sigma} \left[\begin{array}{c} I \\ \mathbf{S} \end{array}\right] < 0 \ pe \ \overline{\mathbb{C}}_0, \tag{178}$$

*atunci KPYS*$(\Sigma, J)$ *are o solutie stabilizanta*

$$(X, V, W) = (X, \left[\begin{array}{cc} V_{11} & O \\ V_{21} & V_{22} \end{array}\right], \left[\begin{array}{c} W_1 \\ W_2 \end{array}\right])$$

*in care partitiile lui* $V$ *si* $W$ *sunt in concordanta cu* $J$.

Conditia (178) este doar suficienta pentru satisfacerea (166) care, in schimb, este o conditie suficienta pentru inversabilitatea lui $\mathcal{R}_\Sigma$ (ptr. ca $\mathcal{R}_{22} >> 0$). Din moment ce stim ca existenta unei solutii stabilizante este echivalenta cu inversabilitatea lui $\mathcal{R}_\Sigma$, este extrem de interesant sa gasim o situatie in care existenta solutiei stabilizante implica conditia (178). Acesta este scopul urmatorului rezultat care este piatra de temelie a solutiilor problemelor de control optimal si robust: problema Nehari, problema de control $H^\infty$ si robustetea stabilizarii.

# O Conditie Necesara si Suficienta

**Teorema 151.** *Fie $\Sigma = (A, B; Q, L, R)$ un triplet Popov cu intrarile partitionate ca mai sus. Presupunem ca :*

**a.** *$A$ este stabila.*

**b.** $\mathbf{\Pi}_{22} > 0$ *pe* $\overline{\mathbb{C}}_0$.

**c.** $\begin{bmatrix} Q & L_2 \\ L_2^T & R_{22} \end{bmatrix} \geq 0.$

**I.** *Urmatoarele afirmatii sunt echivalente:*

1. *Exista* $\mathbf{S} \in RH^{\infty}_{+,m_2 \times m_1}$ *care satisface inegalitatea de signatura.*

$$\begin{bmatrix} I & \mathbf{S}^* \end{bmatrix} \mathbf{\Pi}_{\Sigma} \begin{bmatrix} I \\ \mathbf{S} \end{bmatrix} < 0 \ on \ \overline{\mathbb{C}}_0. \tag{179}$$

2. *KPYS($\Sigma, J$) are o solutie stabilizanta*

$$(X, V, W) = (X, \begin{bmatrix} V_{11} & O \\ V_{21} & V_{22} \end{bmatrix}, \begin{bmatrix} W_1 \\ W_2 \end{bmatrix})$$

*cu $X \geq 0$, unde partitiile lui $V$ si $W$ sunt in acord cu $J$.*

**II.** *Daca 1 sau 2 are loc, atunci exista o matrice rationala*

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_1 & \mathbf{G}_2 \\ \mathbf{G}_3 & \mathbf{G}_4 \end{bmatrix}$$

*partitionata in acord cu $\mathbf{\Pi}_\Sigma$ a.i. $\mathbf{G}$ si $\mathbf{G}_4$ sunt unitati in $RH_+^\infty$ si*

$$\mathbf{\Pi}_\Sigma = \mathbf{G}^* J \mathbf{G}, \tag{180}$$

*(i.e., $\mathbf{G}$ este un factor $J$–spectral al lui $\mathbf{\Pi}_\Sigma$). Clasa tuturor $\mathbf{S}$ care rezolva inegalitatea de signatura este data de*

$$\mathbf{S} = \mathbf{S}_2\mathbf{S}_1^{-1}, \quad \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \end{bmatrix} = \mathbf{G}^{-1} \begin{bmatrix} I \\ \boldsymbol{\theta} \end{bmatrix}, \tag{181}$$

*unde $\boldsymbol{\theta}$ este un parametru liber in $RH^\infty_{+,m_2\times m_1}$ , cu $\|\boldsymbol{\theta}\|_\infty < 1$.*

# Inegalitati Algebrice Riccati

Dam in continuare un rezultat remarcabil privind inegalitatile algebrice Riccati. In acest scop introducem urmatoarele notatii: pentru un triplet Popov fixat $\Sigma = (A, B; Q, L, R)$ si pentru $X = X^T$,

$$\mathrm{CRICOP}(\Sigma, X) := A^T X + XA - (XB + L)R^{-1}(B^T X + L^T) + Q.$$

**Teorema 152.** *Fie $\Sigma = (A, B; Q, L, R)$ un triplet Popov cu $Q \geq 0$. Urmatoarele afirmatii sunt echivalente.*

1. *$A$ este stabila si $\mathbf{\Pi}_\Sigma < 0$ pe $\overline{\mathbb{C}}_0$.*

2. *$R < 0$ si exista $X > 0$ a.i. $\mathrm{CRICOP}(\Sigma, X) < 0$.*

NOTA: Afirmatia 2 este echivalenta cu

$$D_\Sigma(X) = \begin{bmatrix} A^TX + XA + Q & XB + L \\ L^T + B^TX & R \end{bmatrix} < 0$$

pentru $X > 0$.

# Capitolul 13: Ecuatii Riccati si Fascicole Matriciale: Cazul Regulat

In acest capitol prezentam o noua abordare a ecuatiei matriciale algebrice Riccati prin intermediul fasciolelor matriciale. Ca si in capitolele precedente consideram ecuatii Riccati sub ipoteze foarte generale asupra coeficientilor matriciali ce permit includerea situatiilor de joc (de tip $H^\infty$) in care coeficientul patratic are semn nedefinit. Singura restrictie majora a capitolului (pentru pastrarea dezvoltarilor teoretice la un nivel acceptabil de complexitate) consta in **ipoteza de regularitate** sub care functia Popov corespunzatoare ia valori de rang intreg, i.e. este inversabila ca matrice rationala.

## Valori proprii pentru fascicole matriciale: cazul regulat

Structura de valori proprii a unui fascicol Hamiltonian extins (EHP)

ARE si EHP

Ecuatia Bernoulli

Algoritmi numerici

# Valori proprii pentru fascicole matriciale: cazul regulat

Fie $M, N$ matrici $m \times n$ cu elemente in $\mathbb{R}$.

- Fascicol Matricial: Matricea polinomiala de ordinul 1 $\lambda M - N$ in variabila $\lambda$
- Fascicol matricial regulat: Un fascicol matricial patrat $(n \times n)$ cu un determinant neidentic nul $\det(\lambda M - N) \not\equiv 0$.
- Problema de valori proprii generalizate: Problema rezolvarii pentru $\lambda \in \mathbb{C}$ a ecuatiei polinomiale

$$\chi(\lambda) := \det(\lambda M - N) = 0 \tag{182}$$

## Observatii:

- Daca $M$ sau $N$ sunt inversabile atunci fascicolul $\lambda M - N$ este regulat
- Un fascicol matricial regulat poate insa avea ambele matrici $N$ si $M$ singulare
- Deoarece $M$ poate fi singulara, polinomul caracteristic $\chi(\lambda)$ are gradul $n_f \leq n$
- Valori proprii (generalizate) finite: Cele $n_f$ radacini ale $\chi(\lambda)$
- Valori proprii infinite: Spunem ca $\lambda = \infty$ este o valoare proprie (generalizata) a lui $\lambda M - N$ daca $\lambda = 0$ este o valoare proprie

generalizata a fasciolului reciproc $\lambda N - M$; valabil si pentru mutiplicitati algebrice si geometrice

Observatie:

- $n_\infty = n - n_f$
- Un fascicol regulat $n \times n$ $\lambda M - N$ intotdeauna are $n$ valori proprii (finite si infinite) care impreuna formeaza spectrul $\Lambda(M, N)$
- Pentru o valoare proprie generalizata $\lambda_0$ exista un vector propriu (generalizat) $x(\neq 0) \in \mathbb{C}^n$ a. i.

$$\begin{cases} Nx = \lambda_0 Mx & \text{daca } \lambda_0 \text{ este finita,} \\ Mx = 0 & \text{daca } \lambda_0 \text{ este infinita} \end{cases}$$

**Definitia 153.** *Doua fascicole* $\lambda M - N$ *si* $\lambda\widetilde{M} - \widetilde{N}$, *cu* $M, N, \widetilde{M}, \widetilde{N} \in \mathbb{C}^{m\times n}$ *se numesc (strict)* echivalente *daca exista doua matrici constante inversabile* $Q \in \mathbb{C}^{m\times m}$, $Z \in \mathbb{C}^{n\times n}$, *a.i.*

$$Q(\lambda M - N)Z = \lambda\widetilde{M} - \widetilde{N}. \tag{183}$$

# Forma canonica Weierstrass a unui fascicol matricial

Relatia de ecchivalenta stricta induce o forma canonica in multiumea fascicolelor regulate $n \times n$ numita forma canonica Weierstrass (extinde forma canonica Jordan).

Pentru orice fascicol regulat $\lambda M - N$, cu $M, N \in \mathbb{C}^{n\times n}$, exista doua matrici inversabile $Q, Z \in \mathbb{C}^{n\times n}$ a. i.

$$Q(\lambda M - N)Z = \lambda M_W - N_W$$

unde

$$\lambda M_W - N_W := \begin{bmatrix} \lambda M_\infty - I_{n_\infty} & O \\ O & \lambda I_{n_f} - N_f \end{bmatrix}, \tag{184}$$

iar $N_f$ este in forma canonica Jordan, $M_\infty$ este nilpotenta si in forma canonica Jordan

$$M_\infty := \mathrm{diag}\left(J_{s_1^\infty}(0), J_{s_2^\infty}(0), \ldots, J_{s_{h_\infty}^\infty}(0)\right) \tag{185}$$

si $J_s(0)$ este un bloc elementar Jordan $s \times s$ (nilpotent) (avand o singura valoare proprie in $\lambda = 0$).

Observatie:

- Valorile proprii generalizate (si multiplicitatile lor) ale fascicolului $\lambda M - N$ coincid cu valorile proprii (si multiplicitatile lor) ale matricii $N_f$

- Multiplicitatile partiale, algebrice si geometrice ale valorii proprii infinite a lui $\lambda M - N$ coincid cu multiplicitatile valorii proprii din

zero a lui $M_\infty$

- Multimea valorilor proprii ($+$ multiplicitatile) determina complet forma canonica Weierstrass a unui fascicol regulat (pana la o permutare a blocurilor diagonale pe diagonala principala).

- Daca $M$ este inversabil nu exista valori proprii infinite si forma canonica Weierstrass se reduce la forma canonica Jordan a lui $M^{-1}N$

- Exista o varianta reala a acestor concepte (analog cu Jordan)!

# Subspatii de deflatie ale unui fascicol

Extindem in continuare notiunea de spatiu invariant al unei matrici la notiunea de spatiu de deflatie al unui fascicol (regulat).

**Definitia 154.** *Fie $\lambda M - N$ un fascicol regulat, cu $M, N \in \mathbb{C}^{n\times n}$. Subspatiul liniar $\mathcal{V} \subset \mathbb{C}^n$ se numeste spatiu de deflatie a lui $\lambda M - N$ daca*

$$\dim(M\mathcal{V} + N\mathcal{V}) = \dim\mathcal{V}. \tag{186}$$

Observatii:

- Cele mai simple spatii de deflatie sunt vectorii proprii (generalizati) !

- Pentru $M = I_n$, definitia spatiului de deflatie se reduce la ce de spatiu invariant a lui $N$:

$$\dim(\mathcal{V} + N\mathcal{V}) = \dim\mathcal{V} \Leftrightarrow N\mathcal{V} \subset \mathcal{V}$$

- Fie $\mathcal{V} \subset \mathbb{C}^n$ un spatiu arbitrar (nu neaparat de deflatie) avand dimensiunea $\ell$ si
$\mathcal{W} := M\mathcal{V} + N\mathcal{V}$. Avem:

$$k := \dim\mathcal{W} \geq \dim\mathcal{V} = \ell,$$

(egalitatea avand loc doar daca $\mathcal{V}$ este spatiu de deflatie).

Fie $\mathcal{V} = \operatorname{Im} Z_1, \mathcal{W} = \operatorname{Im} Q_1$. Construim matricile inversabile $Q$ si $Z$,

partitionate

$$Z = [\ \underbrace{Z_1}_{\ell}\ \underbrace{Z_2}_{n-\ell}\ ]\,, \qquad Q = [\ \underbrace{Q_1}_{k}\ \underbrace{Q_2}_{n-k}\ ]\,. \tag{187}$$

Din moment ce $M\mathcal{V} \subset \mathcal{W}$ si $N\mathcal{V} \subset \mathcal{W}$:

$$Q^{-1}(\lambda M - N)Z = \begin{bmatrix} \lambda M_{11} - N_{11} & \lambda M_{12} - N_{12} \\ O_{\cdot} & \lambda M_{22} - N_{22} \end{bmatrix} \begin{matrix} \} k \\ \} n-k, \end{matrix} \tag{188}$$

$$\underbrace{\qquad\qquad}_{\ell} \qquad \underbrace{\qquad\qquad\qquad}_{n-\ell}$$

unde $k \geq \ell$. In acest nou sistem de coordonate $\mathcal{V}$ si $\mathcal{W}$ sunt

reprezentate de

$$\mathcal{V} = \mathrm{Im} \begin{bmatrix} I_\ell \\ O \end{bmatrix}, \qquad \mathcal{W} = \mathrm{Im} \begin{bmatrix} I_k \\ O \end{bmatrix}. \tag{189}$$

Daca $\mathcal{V}$ este acum un spatiu de deflatie atunci $\dim \mathcal{W} = k = \ell = \dim \mathcal{V}$, $M_{ii}$ si $N_{ii}$ sunt patrati ptr. $i = 1, 2$, si definim fascicolul restrictionat si spectrul restrictionat al fascicolului la $\mathcal{V}$ ca fiind

$$(\lambda M - N)|_{\mathcal{V}} := \lambda M_{11} - N_{11}, \quad \Lambda(M, N)|_{\mathcal{V}} := \Lambda(M_{11}, N_{11}). \tag{190}$$

Consideram o partitie a planului complex in doua multimi disjuncte

$$\mathbb{C} = \mathbb{C}_b \cup \mathbb{C}_r \tag{191}$$

(cel bun si cel rau ). Un spatiu de deflatie $\mathcal{V}$ este de deflatie $\mathbb{C}_b$ daca $\Lambda(M,N)|_{\mathcal{V}} \subset \mathbb{C}_b$.

**Propozitia 155.** *Fie $\lambda M - N$ un fascicol regulat cu $M, N \in \mathbb{F}^{n\times n}$.*

1. *Daca $\mathcal{V} \subset \mathbb{C}^n$ este un spatiu de deflatie de dimensiune $\ell$ a lui $\lambda M - N$ si $\mathcal{W} := M\mathcal{V} + N\mathcal{V}$ atunci exista matrici complexe inversabile $Z = \left[\begin{array}{cc} Z_1 & Z_2 \end{array}\right]$, $Q = \left[\begin{array}{cc} Q_1 & Q_2 \end{array}\right]$, cu $\mathcal{V} = \operatorname{Im} Z_1$ si $\mathcal{W} = \operatorname{Im} Q_1$, a.i.*

$$Q^{-1}(\lambda M - N)Z = \begin{array}{c} \left[\begin{array}{cc} \lambda M_{11} - N_{11} & \lambda M_{12} - N_{12} \\ O & \lambda M_{22} - N_{22} \end{array}\right] \begin{array}{l} \} \ell \\ \} n-\ell, \end{array} \\ \underbrace{\qquad\qquad}_{\ell} \quad \underbrace{\qquad\qquad}_{n-\ell} \end{array} \tag{192}$$

*unde $\lambda M_{ii} - N_{ii}$ $(i = 1, 2)$ sunt regulate, $\Lambda(M_{11}, N_{11}) \cup$*

$\Lambda(M_{22}, N_{22}) = \Lambda(M, N)$ *si* $\Lambda(M, N)|_{\mathcal{V}} = \Lambda(M_{11}, N_{11})$.

2. *Reciproc, daca are loc (192) pentru matricile inversabile $Q$ si $Z$ partitionate ca mai sus atunci $\mathcal{V} = \operatorname{Im} Z_1$ este un spatiu de deflatie a lui $\lambda M - N$, $\mathcal{W} = \operatorname{Im} Q_1$ unde $\mathcal{W} := M\mathcal{V} + N\mathcal{V}$, si $\Lambda(M, N)|_{\mathcal{V}} = \Lambda(M_{11}, N_{11})$.*

**Corolarul 156.** *Fie $\lambda M - N$ un fascicol regulat cu coeficienti in $\mathbb{C}$ si fie $\Lambda_1 \cup \Lambda_2 = \Lambda(M, N)$ o partitie a spectrului. Atunci exista un spatiu de deflatie $\mathcal{V} \subset \mathbb{C}^n$ a.i.*

$$\Lambda(M, N)|_{\mathcal{V}} = \Lambda_1.$$

*Daca $M$ si $N$ au elemente reale si $\Lambda_1$ este o multime simetrica, atunci putem alege $\mathcal{V} \subset \mathbb{R}^n$. Daca $\Lambda_1 \cap \Lambda_2 = \emptyset$, atunci $\mathcal{V}$ este unic.*

Dam in continuare o caracterizare utila a unui spatiu de deflatie in termenii unei matrici baza asociate.

**Teorema 157.** *Fie $\lambda M - N$ un fascicol regulat, cu $M, N \in \mathbb{C}^{n\times n}$.*

1. *Daca $V \in \mathbb{C}^{n\times \ell}$ este o baza pentru spatiul de deflatie $\mathcal{V}$ cu spectru finit atunci exista o matrice patrata $S \in \mathbb{C}^{\ell\times\ell}$ a.i.*

$$MVS = NV \tag{193}$$

   *si $\lambda I - S$ este echivalent cu $(\lambda M - N)|_{\mathcal{V}}$.*

2. *Reciproc, daca are loc $MVS = NV$ pentru o matrice baza $V \in \mathbb{C}^{n\times\ell}$ si o matrice $S \in \mathbb{C}^{\ell\times\ell}$ atunci $\mathcal{V} = \operatorname{Im} V$ este un spatiu de deflatie a lui $\lambda M - N$ si $\lambda I - S$ este echivalent cu $(\lambda M - N)|_{\mathcal{V}}$.*

Relatia (193) generalizeaza caracterizarea unui spatiu invariant al unei matrici la acea de spatiu de deflatie al unui fascicol regulat.

# Structura de valori proprii a unui fascicol Hamiltonian extins (regulat)

Fie $\Sigma = (A, B; Q, L, R)$ un triplet Popov, cu $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$. Asociem:

$$\text{EHP}(\Sigma): \quad \lambda M_\Sigma - N_\Sigma := \lambda \begin{bmatrix} I & O & O \\ O & I & O \\ O & O & O \end{bmatrix} - \begin{bmatrix} A & O & B \\ -Q & -A^T & -L \\ L^T & B^T & R \end{bmatrix} \tag{194}$$

si functia Popov

$$\mathbf{\Pi}_\Sigma(\lambda) = \left[\begin{array}{cc|c} A & O & B \\ -Q & -A^T & -L \\ \hline L^T & B^T & R \end{array}\right]. \tag{195}$$

Reamintire:

EHP($\Sigma$) este fascicolul de transmisie (sistem) asociat cu realizarea de mai sus a $\mathbf{\Pi}_\Sigma$.

EHP (si in general un fascicol matricial oarecare) se numeste regulat daca

$$\det(sM_\Sigma - N_\Sigma) \not\equiv 0.$$

Notam pentru $\lambda M_\Sigma - N_\Sigma$:

$n_f^-$: numarul de v.p. generalizate finite in $\mathbb{C}_-$

$n_f^0$: numarul de v.p. generalizate finite in $\mathbb{C}_0$

$n_f^+$: numarul de v.p. generalizate finite in $\mathbb{C}_+$

$n_\infty$: numarul de v.p. generalizate infinite

Fie $\pi_0$ numarul de zerouri ale $\mathbf{\Pi}_\Sigma(\lambda^{-1})$ din zero ($\equiv$ numarul de zerouri ale lui $\mathbf{\Pi}_\Sigma(\lambda)$ la $\infty$).

**Teorema 158.** *Pentru un EHP regulat avem:*

*1.* $n_f^- = n_f^+$.

*2.* $\operatorname{rank}_{\mathbb{R}(\lambda)}(\lambda M_\Sigma - N_\Sigma) = 2n + m, \quad \operatorname{rank}_{\mathbb{R}(\lambda)} \mathbf{\Pi}_\Sigma(\lambda) = m.$

*3.* $n_\infty = m + \pi_0.$

**Demonstratie.** 1. Se poate verifica direct

$$P(\lambda M_\Sigma - N_\Sigma)^T P = -\lambda M_\Sigma - N_\Sigma,$$

unde

$$P := \begin{bmatrix} O & -I_n & O \\ I_n & O & O \\ O & O & I_m \end{bmatrix}.$$

Deci

$$\det(\lambda M_\Sigma - N_\Sigma) = \det(-\lambda M_\Sigma - N_\Sigma).$$

EHP-ul fiind regulat rezulta ca $\lambda$ este o v.p. generalizata a lui $\lambda M_\Sigma - N_\Sigma$ daca si numai daca $-\lambda$ este o v.p. generalizata a lui $\lambda M_\Sigma - N_\Sigma$ cu aceeasi multiplicitate.

2. Deoarece $\lambda M_\Sigma - N_\Sigma$ este fascicolul de transmisie al functiei Popov $\mathbf{\Pi}_\Sigma(\lambda)$ avem

$$\mathrm{rank}_{\mathbb{R}(\lambda)}(\lambda M_\Sigma - N_\Sigma) = 2n + \mathrm{rank}_{\mathbb{R}(\lambda)}\mathbf{\Pi}_\Sigma(\lambda). \qquad (196)$$

Din regularitatea fascicolului rezulta ca $\mathrm{rank}_{\mathbb{R}(\lambda)}(\lambda M_\Sigma - N_\Sigma) = 2n + m$ care combinat cu (196) conduce la $\mathrm{rank}_{\mathbb{R}(\lambda)}\mathbf{\Pi}_\Sigma(\lambda) = m$.

3. Pentru orice $\lambda$ pentru care $I - \lambda A$ si $I + \lambda A^T$ sunt inversabile

avem identitatea

$$M_\Sigma - \lambda N_\Sigma = \begin{bmatrix} I_n & O & O \\ O & I_n & O \\ \mathbf{X}(\lambda) & \mathbf{Y}(\lambda) & -I_m \end{bmatrix} \begin{bmatrix} I_n - \lambda A & O & \lambda B \\ \lambda Q & I_n + \lambda A^T & -\lambda L \\ O & O & \lambda \mathbf{\Pi}_\Sigma(\lambda) \end{bmatrix}$$

unde

$$\begin{aligned} \mathbf{X}(\lambda) &:= -\lambda[L^T + \lambda B^T (I_n + \lambda A^T)^{-1} Q](I_n - \lambda A)^{-1}, \\ \mathbf{Y}(\lambda) &:= -\lambda B^T (I_n + \lambda A^T)^{-1}, \end{aligned}$$

ambele nu au poli in $\lambda = 0$. Deoarece fascicolul $\lambda M_\Sigma - N_\Sigma$ este regulat,

$$\det(M_\Sigma - \lambda N_\Sigma) = (-1)^m \lambda^m \det(I_n - \lambda A) \det(I_n + \lambda A^T) \det(\mathbf{\Pi}_\Sigma(\lambda^{-1})).$$

Folosind forma Smith–McMillan a lui $\mathbf{\Pi}_\Sigma$, obtinem ca $\lambda = 0$ este o

valoare proprie generalizata a lui $M_\Sigma - \lambda N_\Sigma$ cu multiplicitatea $m+\pi_0$ ceea ce inseamna ca $\lambda = \infty$ este o v.p. a lui $\lambda M_\Sigma - N_\Sigma$ avand aceeasi multiplicitate $n_\infty = m + \pi_0$. ∎

**Definitia 159.** *Un EHP regulat se numeste dihotomic daca*

$$n_f^0 = 0, \quad n_\infty = m$$

*(sau, echivalent, daca $n_f^0 = 0$ si $\pi_0 = 0$).*

Observatii

Zerourile $\mathbf{\Pi}_\Sigma(\lambda)$ sunt printre valorile proprii generalizate ale $\lambda M_\Sigma - N_\Sigma$.

Dihotomia implica ca $\mathbf{\Pi}_\Sigma(\lambda)$ nu are zerouri pe axa imaginara inclusiv la infinit

Dihotomia implica $R = \mathbf{\Pi}_\Sigma(\infty)$ nesingulara.

Dihotomia implica ca toate multiplicitatile valorii proprii de la infinit sunt egale cu $1$.

**Propozitia 160. [Dihotomie si Spatii de Deflatie ]** *Un EHP regulat este dihotomic daca si numai daca are un spatiu de deflatie $\mathbb{C}_-$ de dimensiune $n$.*

In particular, rezultatul anterior implica ca EHP este dihotomic daca

si numai daca

$$M_\Sigma V S = N_\Sigma V, \qquad V = \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix} \begin{matrix} \}n \\ \}n \\ \}m \end{matrix} \tag{197}$$

unde $V$ este o matrice baza pentru subspatiul de deflatie $\mathbb{C}_-$ de dimensiune $n$ si $S$ este o matrice $n \times n$ stabila.

**Definitia 161.** *Un subspatiu de deflatie $\mathcal{V}$ se numeste disconjugat daca admite o matrice baza cu $V_1$ avand rang intreg pe coloane (injectiva).*

Observati ca disconjugarea este independenta de alegerea matricii baza.

**Definitia 162.** *Un EHP regulat se numeste stabil (antistabil)*

*disconjugat daca este dihotomic si are un subspatiu de deflatie disconjugat $\mathbb{C}_-$ ($\mathbb{C}_+$) de dimensiune $n$.*

Observatie:

Un EHP dihotomic are un unic subspatiu de deflatie (maximal) $n$–dimensional $\mathbb{C}_-$ (si $\mathbb{C}_+$).

Disconjugarea unui EHP este bine definita

Un EHP dihotomic are intotdeauna un subspatiu de deflatie $n$–dimensional $\mathbb{C}_-$ si un subspatiu de deflatie $n$–dimensional $\mathbb{C}_+$, fiecare in parte putand fi disconjugat sau nu.

**Propozitia 163.** *Daca* $V = \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix} \begin{matrix} \}n \\ \}n \\ \}m \end{matrix}$ *este o matrice baza a unui spatiu de deflatie* $\mathbb{C}_-$ *(sau* $\mathbb{C}_+$*) a unui EHP regulat atunci*

$$V_1^T V_2 = V_2^T V_1. \tag{198}$$

**Demonstratie.** Scriind (197) explicit obtinem

$$\begin{array}{rcrcrcl} V_1 S & = & AV_1 & & & + & BV_3, \\ V_2 S & = & -QV_1 & - & A^T V_2 & - & LV_3, \\ 0 & = & L^T V_1 & + & B^T V_2 & + & RV_3, \end{array} \tag{199}$$

unde $S$ este stabila (sau antistabila). Adunand prima ecuatie

premultiplicata cu $V_2^T$ celei de-a doua transpusa si postmultiplicata cu $V_1$ si folosind a treia ecuatie (199) obtinem

$$V_2^T V_1 S + S^T V_2^T V_1 + V_1^T Q V_1 - V_2^T B V_3 - V_3^T B^T V_2 - V_3^T R V_3 = 0. \tag{200}$$

Aceasta ecautie este de tip Lyapunov in $V_2^T V_1$, avand un termen liber simetric. Deoarece $S$ este stabila (sau antistabila) $V_2^T V_1$ este unica solutie simetrica. ∎

# ARE si EHP

Dam in continuare conditii necesare si suficiente de existenta a solutiei stabilizante (si antistabilizante) a ARE, impreuna cu formule de calcul si un algoritm numeric stabil.

**Teorema 164.** *Urmatoarele afirmatii sunt echivalente:*

1. *Matricea $R$ este inversabila si*

$$A^T X + XA - (XB + L)R^{-1}(L^T + B^T X) + Q = 0 \qquad (201)$$

*are o solutie stabilizanta (antistabilizanta).*

*2. EHP*

$$\lambda M_\Sigma - N_\Sigma = \lambda \begin{bmatrix} I & O & O \\ O & I & O \\ O & O & O \end{bmatrix} - \begin{bmatrix} A & O & B \\ -Q & -A^T & -L \\ L^T & B^T & R \end{bmatrix} \tag{202}$$

*este regulat si stabil (antistabil) disconjugat.*

*Mai mult, solutia si reactia se pot calcula cu formulele*

$$X = V_2 V_1^{-1}, \qquad F = V_3 V_1^{-1} \tag{203}$$

*in care* $V = \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix} \begin{matrix} \}n \\ \}n \\ \}m \end{matrix}$ *este o matrice baza a subspatiului de deflatie* $n$*–dimensional* $\mathbb{C}_-$ *(*$\mathbb{C}_+$*) a EHP.*

Pentru $R$ inversabil, obtinem identitatea

$$Q(\lambda M_\Sigma - N_\Sigma)Z = \lambda \begin{bmatrix} I_{2n} & O \\ O & O \end{bmatrix} - \begin{bmatrix} H_\Sigma & O \\ O & I_m \end{bmatrix}, \qquad (204)$$

in care

$$H_\Sigma := \begin{bmatrix} A - BR^{-1}L^T & -BR^{-1}B^T \\ -Q + LR^{-1}L^T & -A^T + LR^{-1}B^T \end{bmatrix} \qquad (205)$$

si

$$Q = \begin{bmatrix} I_n & O & -BR^{-1} \\ O & I_n & LR^{-1} \\ O & O & I_m \end{bmatrix}, \qquad Z = \begin{bmatrix} I_n & O & O \\ O & I_n & O \\ -R^{-1}L^T & -R^{-1}B^T & R^{-1} \end{bmatrix}.$$

Matricea $H_\Sigma$ este o matrice Hamiltoniana de dimensiune $2n \times 2n$ ce

satisface prin definitie

$$H_\Sigma^T \widehat{J} + \widehat{J} H_\Sigma = 0, \quad \text{pentru} \quad \widehat{J} := \begin{bmatrix} O & -I_n \\ I_n & O \end{bmatrix}.$$

**Corolarul 165.** *Presupunem $R$ inversabila. Atunci urmatoarele afirmatii sunt echivalente.*

1. *ARE are o solutie stabilizanta (antistabilizanta).*

2. *Matricea Hamiltoniana $H_\Sigma$ are un subspatiu invariant $n$–dimensional $\mathbb{C}_-$ ($\mathbb{C}_+$) care admite o baza cu $V_1$ inversabila:*

$$V = \underbrace{\begin{bmatrix} V_1 \\ V_2 \end{bmatrix}}_{n} \begin{matrix} \}n \\ \}n \end{matrix}. \tag{206}$$

*Solutia stabilizanta si reactia corespunzatoare se pot calcula din*

$$X = V_2 V_1^{-1}, \qquad F = -R^{-1}(B^T X + L^T).$$

Observatii:

- Existenta solutiei stabilizante pentru ARE se poate verifica alternativ folosind matricea Hamiltoniana sau EHP.

- Solutiile se pot calcula rezolvand problema de valori proprii pentru matricea Hamiltoniana de dimensiune $2n$ in loc sa rezolvam problema mai complicata de valori proprii generalizate pentru EHP avand dimensiune $2n + m$.

- Folosirea matricii Hamiltoniene poate conduce la solutii extrem de inexacte daca $R$ este prost conditionata in raport cu inversarea

numerica. In acest caz se foloseste intotdeauna EHP in loc de matricea Hamiltoniana.

# Ecuatia Bernoulli

In cazul particular in care $Q = 0, L = 0,$ si $R$ este nensingulara, ecuatia algebrica Riccati se numeste ecuatie Bernoulli:

$$A^T X + XA - XBR^{-1}B^T X = 0. \tag{207}$$

Din cauza formei sale particulare, ecuatia Bernoulli are cateva proprietati suplimentare.

**Propozitia 166.** *Fie $R$ inversabila. Daca ecuatia Bernoulli are o solutie stabilizanta (antistabilizanta) atunci $A$ este dihotomica.*

**Propozitia 167.** *Fie $R < 0$. Daca ecuatia Bernoulli are o solutie stabilizanta pozitiv semidefinita $X$ atunci $A$ este stabila.*

**Teorema 168.** *Fie $A$ dihotomica, perechea $(A, B)$ stabilizabila, si $R > 0$. Atunci ecuatia Bernoulli are o solutie stabilizanta pozitiv semidefinita.*

# Capitolul 14: Aplicatii in Teoria Sistemelor

In acest capitol prezentam cateva aplicatii generale ale rezultatelor Riccati in teoria generala a sistemelor, rezultate ce vor fi ulterior folosite in teoria controlului optimal.

Lema de Real-Marginire

Factorizari Coprime Normalizate

Teorema Micii Amplificari

Factorizare Spectrala si Inner-outer (interioara-exterioara)

Lema urmatoare contine un rezultat celebru in Teoria Sistemelor dand conditii necesare si suficiente pentru ca un sistem sa fie marginit in norma $H^\infty$ de un $\gamma > 0$.

**Lema 169.** *Fie $\gamma > 0$ si*

$$\mathbf{G} = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right].$$

*Urmatoarele afirmatii sunt echivalente:*

1. *$A$ este dihotomica, $(A, B)$ este stabilizabila si $\|\mathbf{G}\|_\infty < \gamma$.*

2. *ARE*

$$A^TX+XA-(XB+C^TD)(-\gamma^2I+D^TD)(D^TC+B^TX)+C^TC=0$$

*are o solutie stabilizanta $X$ care satisface urmatoarele proprietati:*

$$\delta_c(X)+\pi_c(X)=\nu_c(A), \qquad \nu_c(X)=\pi_c(A)$$

*in care $\nu_c(X), \pi_cX, \delta_c(X)$ reprezinta numarul de valori proprii cu parte reala strict negativa, strict pozitiva si respectiv egala cu zero.*

**Lema 170. [Varianta stricta]** *Fie $\gamma > 0$ si*

$$\mathbf{G}=\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right].$$

*Urmatoarele doua afirmatii sunt echivalente:*

1. $A$ *este stabila si* $\|\mathbf{G}\|_\infty < \gamma$

2. $-\gamma^2 I + D^T D < 0$ *si exista* $X > 0$ *a.i.*

$$\begin{bmatrix} A^T X + X A + C^T C & X B + C^T D \\ D^T C + B^T X & -\gamma^2 I + D^T D \end{bmatrix} < 0$$

# Factorizari Coprime Normalizate

**Teorema 171.** *Fie* $\mathbf{G} = \left[\begin{array}{c|c} A & B \\ \hline C & O \end{array}\right]$, *cu* $(A, B)$ *stabilizabila si* $(C, A)$ *detectabila. Fie*

$$\Sigma_X = (A, B; C^T C, 0, I), \quad \Sigma_Y = (A^T, C^T; BB^T, 0, I)$$

*si fie* $X$ *si* $Y$ *solutii stabilizante a ARE(*$\Sigma_X$*) si ARE(*$\Sigma_Y$*). Definim* $F := -B^T X$, $K := -YC^T$, $A_F := A + BF$, $A_K := A + KC$., *si*

*matricile de transfer* $RH_+^\infty$ *:*

$$
\begin{aligned}
\mathbf{M} &= \left[\begin{array}{c|c} A_F & B \\ \hline F & I \end{array}\right], \quad & \mathbf{N} &= \left[\begin{array}{c|c} A_F & B \\ \hline C & O \end{array}\right], \\
\mathbf{U} &= \left[\begin{array}{c|c} A_K & B \\ \hline -F & I \end{array}\right], \quad & \mathbf{W} &= \left[\begin{array}{c|c} A_K & K \\ \hline F & O \end{array}\right], \\
\widetilde{\mathbf{M}} &= \left[\begin{array}{c|c} A_K & K \\ \hline C & I \end{array}\right], \quad & \widetilde{\mathbf{N}} &= \left[\begin{array}{c|c} A_K & B \\ \hline C & O \end{array}\right], \\
\widetilde{\mathbf{U}} &= \left[\begin{array}{c|c} A_F & -K \\ \hline C & I \end{array}\right], \quad & \widetilde{\mathbf{W}} &= \left[\begin{array}{c|c} A_F & K \\ \hline F & O \end{array}\right].
\end{aligned}
\tag{208}
$$

*Atunci avem:*

1.

$$\begin{bmatrix} -\mathbf{W} & \mathbf{U} \\ \widetilde{\mathbf{M}} & \widetilde{\mathbf{N}} \end{bmatrix} \begin{bmatrix} -\mathbf{N} & \widetilde{\mathbf{U}} \\ \mathbf{M} & \widetilde{\mathbf{W}} \end{bmatrix} = \begin{bmatrix} I & O \\ O & I \end{bmatrix} \tag{209}$$

*( identitatea de factorizare coprima (dubla))*

2.

$$\mathbf{G} = \mathbf{N}\mathbf{M}^{-1}, \qquad \mathbf{N}^*\mathbf{N} + \mathbf{M}^*\mathbf{M} = I, \tag{210}$$

*i.e., perechea* $(\mathbf{N}, \mathbf{M})$ *este o factorizare coprima normalizata la dreapta a lui* $\mathbf{G}$ *peste* $RH_+^\infty$.

3.

$$\mathbf{G} = \widetilde{\mathbf{M}}^{-1}\widetilde{\mathbf{N}}, \qquad \widetilde{\mathbf{N}}\widetilde{\mathbf{N}}^* + \widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^* = I, \tag{211}$$

*i.e. perechea* $(\widetilde{\mathbf{N}}, \widetilde{\mathbf{M}})$ *defineste o factorizare coprima normalizata la stanga a lui* $\mathbf{G}$ *peste* $RH_+^\infty$.

# Teorema Micii Amplificari

Aceasta teorema da un instrument puternic pentru evaluarea stabilitatii interne a unui sistem in bucla inchisa.

**Teorema 172. [Teorema Micii Amplificari]** *Fie*

$$\mathbf{G}_i = \left[\begin{array}{c|c} A_i & B_i \\ \hline C_i & D_i \end{array}\right], \quad y_i = \mathbf{G}_i u_i, \quad i = 1, 2,$$

*doua sisteme cu* $A_i$ *stabila*, $i = 1, 2$, *avand functiile de transfer* $\mathbf{G}_1$ *si* $\mathbf{G}_2$ *de dimensiune* $p \times m$ *si* $m \times p$. *Presupunem ca* $S_c := I_m - D_2 D_1$ *(sau, echivalent* $\widehat{S}_c := I_p - D_1 D_2$*) este nensingulara.*

*Daca*

$$\|\mathbf{G}_1\|_\infty < \frac{1}{\gamma} \qquad \|\mathbf{G}_2\|_\infty \leq \gamma$$

*pentru* $\gamma > 0$, *atunci sistemul in bucla inchisa a lui* $\mathbf{G}_1$ *cu* $\mathbf{G}_2$ *este (intern) stabil.*

# Factorizari spectrale si interioare–exterioare

**Definitia 173.** *Fie* $\mathbf{G}(\lambda) \in RH^{\infty}_{+,p\times m}$ *avand rang intreg pe coloane pe intreaga axa imaginara, i.e.,*

$$\text{rank } \mathbf{G}(j\omega) = m \ (= \text{rank}_{\mathbb{R}(\lambda)}\mathbf{G}(\lambda)), \ \forall \omega \in \overline{\mathbb{R}}. \tag{212}$$

1. *Un* $\mathbf{G}_o(\lambda) \in RH^{\infty}_{+,m\times m}$ *inversabil cu inversa in* $RH^{\infty}_{+,m\times m}$ *a.i.*

$$\mathbf{G}^*(\lambda)\mathbf{G}(\lambda) = \mathbf{G}^*_o(\lambda)\mathbf{G}_o(\lambda) \tag{213}$$

*se numeste un factor spectral al lui* $\mathbf{G}$*.*

2. *Fie* $\mathbf{G}_i := \mathbf{G}\mathbf{G}_o^{-1} \in RH^{\infty}_{+,p\times m}$*. Atunci* $\mathbf{G}_i$ *este interioara, i.e., este*

*in* $RH_+^\infty$ *si verifica*

$$\mathbf{G}_i^*(\lambda)\mathbf{G}_i(\lambda) = I \tag{214}$$

*Mai mult,*

$$\mathbf{G}(\lambda) = \mathbf{G}_i(\lambda)\mathbf{G}_o(\lambda) \tag{215}$$

*defineste o factorizare interioara–exterioara a lui* $\mathbf{G}$.

**Teorema 174.** *Fie*

$$\mathbf{G} = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]. \tag{216}$$

*Presupunem ca* $\mathbf{G}$ *este injectiva (rang intreg pe coloane) pe axa imaginara extinsa si realizarea ei este stabilizabila si* $\mathbb{C}_0$*–observabila. Definim tripletul Popov de pozitivitate*

$$\Sigma = (A, B; C^T C, C^T D, D^T D).$$

1. *ARE*($\Sigma$) *are o solutie stabilizatoare* $X$*, cu reactia Riccati corespunzatoare* $F$*.*

2. *Fie $H$ o matrice inversabila a.i. $H^T H = D^T D$. Definim*

$$\begin{aligned} \mathbf{G}_o &:= \left[\begin{array}{c|c} A & B \\ \hline -HF & H \end{array}\right], \\ \mathbf{G}_i &:= \left[\begin{array}{c|c} A+BF & BH^{-1} \\ \hline C+DF & DH^{-1} \end{array}\right]. \end{aligned} \tag{217}$$

*Atunci $\mathbf{G}_o$ este un factor spectral si $\mathbf{G}_i$ este factorul interior a lui $\mathbf{G}$, si impreuna ele definesc o factorizare interioara-exterioara.*

# Capitolul 15: Problema Nehari patru bloc

In acest capitol dam solutia problemei Nehari care consta in a aproxima optimal (!!!!) in norma $L^\infty$ un sistem stabil cu unul antistabil. Consideram in acest capitol cazul general in care aproximantul actioneaza doar pe un colt al sistemului aproximat, problema ce mai este cunoscuta si sub numele de tip patru bloc.

Abordaarea pe care o dam consta in reducerea problemei Nehari la o conditie de signatura care poate fi rezolvata pe baza rezultatelor ce le-am obtinut anterior.

Problema Nehari si conditia de signatura

Problema Parrott

# Solutia problemei Nehari

# Problema Nehari si conditia de signatura

Consideram un sistem antistabil (avand matricea $A$ antistabila)

$$\mathbf{T} = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{array}\right] = \left[\begin{array}{cc} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{array}\right] \in RH_-^\infty. \tag{218}$$

Definim un aproximant pentru sistemul $\mathbf{T}$ ca fiind un sistem stabil (cu $A_S$ stabil)

$$\mathbf{S} = \left[\begin{array}{c|c} A_S & B_S \\ \hline C_S & D_S \end{array}\right] \in RH_+^\infty. \tag{219}$$

Dandu-se sistemul antistabil $\mathbf{T}$ si un $\gamma > 0$, problema Nehari suboptimala (4–NP) consta in gasirea tuturor aproximantilor (stabili) $\mathbf{S}$ (atunci cand exista !), a.i.

$$\left\| \begin{bmatrix} \mathbf{T}_{11} + \mathbf{S} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix} \right\|_\infty < \gamma. \tag{220}$$

Problema Nehari optimala consta in gasirea aproximantilor (stabili) $\mathbf{S}$ a.i.

$$\min_{\widehat{\mathbf{S}} \in RH^\infty_{+,p_1 \times m_1}} \left\| \begin{bmatrix} \mathbf{T}_{11} + \widehat{\mathbf{S}} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix} \right\|_\infty = \left\| \begin{bmatrix} \mathbf{T}_{11} + \mathbf{S} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix} \right\|_\infty = \gamma_{\min}. \tag{221}$$

Observatie:

- Cazul optimal este echivalent cu gasirea lui $\gamma_{\min}$ si toti $\mathbf{S}$ stabili care satisfac (220) cu inegalitatea stricta relaxata la una nestricta, si $\gamma = \gamma_{\min}$. Din motive de timp vom trata numai cazul suboptimal care este esentialmente echivalent cu cel optimal dpdv numeric.

- Nu restrangem cu nimic generalitatea presupunand $D_{11} = 0$ si $\gamma = 1$. Presupunem indeplinite aceste ipoteze.

Inegalitatea (220) este echivalenta cu o problema doi bloc

$$\left\| \begin{bmatrix} \mathbf{T}_1 + \overline{\mathbf{S}} \\ \mathbf{T}_2 \end{bmatrix} \right\|_\infty < 1 \tag{222}$$

cu constrangerea structurala $\overline{\mathbf{S}} = \begin{bmatrix} \mathbf{S} & O \end{bmatrix}$, unde $\mathbf{T}_1 := \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \end{bmatrix}$, $\mathbf{T}_2 := \begin{bmatrix} \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix}$. Mai departe, (222) este

echivalenta cu

$$(\mathbf{T}_1+\overline{\mathbf{S}})^*(\mathbf{T}_1+\overline{\mathbf{S}})+\mathbf{T}_2^*\mathbf{T}_2-I=\left[\begin{array}{c}\mathbf{T}_1\\ \mathbf{T}_2\end{array}\right]^*\left[\begin{array}{c}\mathbf{T}_1\\ \mathbf{T}_2\end{array}\right]+\overline{\mathbf{S}}^*\mathbf{T}_1+\mathbf{T}_1^*\overline{\mathbf{S}}-I$$

$$=\mathbf{T}^*\mathbf{T}+\overline{\mathbf{S}}^*\mathbf{T}_1+\mathbf{T}_1^*\overline{\mathbf{S}}-I<0,\ \text{pe}\ \overline{\mathbb{C}}_0. \qquad (223)$$

Rescriind (223) in mod compact rezulta conditia de signatura

$$\left[\begin{array}{cc} I & \overline{\mathbf{S}}^*\end{array}\right]\left[\begin{array}{cc}\mathbf{T}^*\mathbf{T}-I_m & \mathbf{T}_1^*\\ \mathbf{T}_1 & I_{p_1}\end{array}\right]\left[\begin{array}{c} I\\ \overline{\mathbf{S}}\end{array}\right]<0,\ \text{pe}\ \overline{\mathbb{C}}_0 \qquad (224)$$

asupra functiei Popov $\mathbf{\Pi}_{\overline{\Sigma}}:=\left[\begin{array}{cc}\mathbf{T}^*\mathbf{T}-I_m & \mathbf{T}_1^*\\ \mathbf{T}_1 & I_{p_1}\end{array}\right]$. Aceasta este in schimb asociata tripletului Popov

$$\overline{\Sigma}\ =\ (\overline{A},\overline{B};\overline{Q},\overline{L},\overline{R})$$

$$:= \quad (A, \begin{bmatrix} B & O_{p_1} \end{bmatrix}; C^T C, \begin{bmatrix} C^T D & C_1^T \end{bmatrix}, \begin{bmatrix} D^T D - I & D_1^T \\ D_1 & I_{p_1} \end{bmatrix}). \tag{225}$$

Astfel problema Nehari 4–bloc are o solutie $\mathbf{S}$ daca si numai daca conditia de signatura (224) are loc pentru $\overline{\mathbf{S}}$.

Oricum, nu putem aplica direct rezultatele date in Capitolul 13 din doua motive:

- $\overline{A}$ este antistabila pe cand noi avem nevoie sa fie stabila
- Solutia $\overline{\mathbf{S}}$ a inegalitatii (224) trebuie sa aibe a un patern de zerouri care nu este dat automat de conditia generala de signatura.

Pentru a depasi prima dificultate vom transforma intai tripletul antistabil $\overline{\Sigma}$ intr-unul stabil prin luarea unuor echivalenti potriviti.

Pentru a defini echivalentii Popov potriviti $\widehat{\Sigma}$ folosim solutia ecuatiei Riccati care exprima proprietatea de contractivitate

$$\left\| \begin{bmatrix} \mathbf{T}_{12} \\ \mathbf{T}_{22} \end{bmatrix} \right\|_{\infty} < 1. \tag{226}$$

Observati ca (226) este o conditie necesara de rezolvare a problemei Nehari 4 bloc.

Pentru a depasi a doua dificultate vom actualiza solutia conditiei de signatura printr-o structura potrivita. Cheia consta intr-o factorizare $J$ spectrala a unei functii Popov avand o structura prescrisa de "zerouri".

**Teorema 175.** *Fie* $\mathbf{\Pi} := \begin{bmatrix} \mathbf{\Pi}_{11} & \mathbf{\Pi}_{12} \\ \mathbf{\Pi}_{12}^* & \mathbf{\Pi}_{22} \end{bmatrix} = \mathbf{\Pi}^*$, *cu* $\mathbf{\Pi}_{22} > 0$ *pe* $\overline{\mathbb{C}}_0$.

*Presupunem ca* $\mathbf{\Pi}$ *admite o factorizare* $J$ *spectrala*

$$\mathbf{\Pi} = \mathbf{G}^* J \mathbf{G}, \tag{227}$$

*unde* $J := \operatorname{diag}(-I_m, I_{p_1}) = \operatorname{diag}(-I_{m_1}, -I_{m_2}, I_{p_1})$, *factorul spectral are structura particulara*

$$\mathbf{G} = \left[\begin{array}{c:c} \mathbf{G}_1 & \mathbf{G}_2 \\ \hdashline \mathbf{G}_3 & \mathbf{G}_4 \end{array}\right] = \left[\begin{array}{cc:c} \mathbf{G}_{11} & O & \mathbf{G}_{13} \\ \mathbf{G}_{21} & \mathbf{G}_{22} & \mathbf{G}_{23} \\ \hdashline \mathbf{G}_{31} & O & \mathbf{G}_{33} \end{array}\right], \tag{228}$$

$\mathbf{G}$ *este o unitate in* $\in RH^\infty_{+,(m+p_1)\times(m+p_1)}$ *si* $\mathbf{G}_4 (= \mathbf{G}_{33})$ *este o unitate in* $RH^\infty_{+,p_1\times p_1}$.

*Atunci clasa tuturor* $\overline{\mathbf{S}} \in RH^\infty_+$, ( $\overline{\mathbf{S}} = \left[\begin{array}{cc} \mathbf{S} & O_{p_1\times m_2} \end{array}\right]$, $\mathbf{S} \in$

$RH^{\infty}_{+,p_1\times m_1}$), *care satisfac conditia de signatura*

$$\left[\begin{array}{cc} I & \overline{\mathbf{S}}^* \end{array}\right] \mathbf{\Pi} \left[\begin{array}{c} I \\ \overline{\mathbf{S}} \end{array}\right] < 0, \quad \text{pe } \overline{\mathbb{C}}_0, \tag{229}$$

*este data de*

$$\overline{\mathbf{S}} := \overline{\mathbf{S}}_2 \overline{\mathbf{S}}_1^{-1}, \ \left[\begin{array}{c} \overline{\mathbf{S}}_1 \\ \overline{\mathbf{S}}_2 \end{array}\right] = \mathbf{G}^{-1} \left[\begin{array}{c} I_m \\ \overline{\boldsymbol{\theta}} \end{array}\right], \ \overline{\boldsymbol{\theta}} = \left[\begin{array}{cc} \boldsymbol{\theta} & O \end{array}\right], \tag{230}$$

*unde* $\boldsymbol{\theta}$ *este un element arbitrar a lui* $RH^{\infty}_{+,p_1\times m_1}$ *cu* $\|\boldsymbol{\theta}\|_{\infty} < 1$.

# Problema Parrott

Dam intai solutia problemei Nehari patru bloc in cazul matricial constant – in aceasta forma este cunoscuta si sub numele de problema lui Parrott (elementara dar netriviala !!!).

**Teorema 176. [Problema Parrott]** *Fie* $D = \begin{bmatrix} O & D_{12} \\ D_{21} & D_{22} \end{bmatrix}$. *Exista* $S$ *a. i.*

$$\| \begin{bmatrix} S. & D_{12} \\ D_{21} & D_{22} \end{bmatrix} \| < 1 \tag{231}$$

*daca si numai daca:* $\|D_2\| < 1$ *si* $\|\overline{D}_2\| < 1$, *unde* $D_2 := \begin{bmatrix} D_{21} & D_{22} \end{bmatrix}$ *si* $\overline{D}_2 := \begin{bmatrix} D_{12} \\ D_{22} \end{bmatrix}$.

*Daca aceste conditii sunt indeplinite atunci*

$$R := \left[\begin{array}{c|c} D^T D - I & D_1^T \\ \hline D_1 & I \end{array}\right] = \left[\begin{array}{cc|c} D_{21}^T D_{21} - I & D_{21}^T D_{22} & O \\ D_{22}^T D_{21} & \overline{D}_2^T \overline{D}_2 - I & D_{12}^T \\ \hline O & D_{12} & I \end{array}\right], \tag{232}$$

*are o* $J$ *factorizare*

$$R = V^T J V, \qquad V = \left[\begin{array}{ccc} V_{11} & O & O \\ V_{21} & V_{22} & V_{23} \\ V_{31} & O & V_{33} \end{array}\right]. \tag{233}$$

*Clasa tutoror matricilor* $S$ *este data de*

$$S = -V_{33}^{-1}(V_{31} + \theta V_{11}),\ \forall \theta \in \mathbb{R}^{p_1 \times m_1},\ \text{cu}\ \|\theta\| < 1. \tag{234}$$

# Solutia problemei Nehari 4 bloc

Introducem intai niste notatii. Definim tripletele Popov

$$\begin{aligned}
\Sigma_1 &= (-A, -B_2; C^T C, C^T \overline{D}_2, \overline{D}_2^T \overline{D}_2 - I), \quad \overline{D}_2 := \begin{bmatrix} D_{12} \\ D_{22} \end{bmatrix} \\
\Sigma_2 &= (-A^T, -C_2^T; BB^T, BD_2^T, D_2 D_2^T - I), \quad D_2 := \begin{bmatrix} D_{21} & D_{22} \end{bmatrix} \\
\widetilde{\Sigma} &= (\widetilde{A}, \widetilde{B}; \widetilde{Q}, \widetilde{L}, \widetilde{R}),
\end{aligned} \tag{235}$$

$$
\begin{aligned}
\widetilde{A} &:= -(A + B_2 F_1)^T, \\
\widetilde{B} &:= -\left[ \begin{array}{ccc} (C_2^T + F_1 D_{22}^T) D_{21} - X_1 B_1 & O & C_1^T + F_1^T D_{12}^T \end{array} \right] \\
\widetilde{Q} &:= 0, \\
\widetilde{L} &= \left[ \begin{array}{ccc} B_1 & B_2 & O \end{array} \right], \\
\widetilde{R} &:= \left[ \begin{array}{cc} D^T D - I & D_1^T \\ D_1 & I \end{array} \right]
\end{aligned}
\tag{236}
$$

unde $X_1$ si $F_1$ sunt solutia stabilizanta si reactia stabilizanta a ARE($\Sigma_1$).

**Teorema 177.** *Problema Nehari suboptimala are o solutie* $\mathbf{S}$ *daca si numai daca:*

1. *KYPS*$(\Sigma_1, -I)$ *are o solutie stabilizanta* $(X_1, V_1, W_1)$, *cu* $X_1 \geq 0$.

2. *KYPS*$(\Sigma_2, -I)$ *are o solutie stabilizanta* $(X_2, V_2, W_2)$, *cu* $X_2 \geq 0$.

3. $\rho(X_1 X_2) < 1$.

*Daca aceste conditii au loc atunci KYPS*$(\widetilde{\Sigma}, J)$ *are o solutie stabilizanta* $(\widetilde{X}, V, W)$, *cu* $\widetilde{X} \geq 0$ *satisfacand* $\widetilde{X} = X_2(I - X_1 X_2)^{-1}$. *Clasa solutiilor problemei Nehari patru bloc este*

$$\mathbf{S} = (\overline{\mathbf{G}}_{31} + \overline{\mathbf{G}}_{33}\boldsymbol{\theta})(\overline{\mathbf{G}}_{11} + \overline{\mathbf{G}}_{13}\boldsymbol{\theta})^{-1},$$

*unde* $\boldsymbol{\theta}$ *este stabil,* $\|\boldsymbol{\theta}\|_\infty < 1$,

$$\begin{bmatrix} \overline{\mathbf{G}}_{11} & \overline{\mathbf{G}}_{13} \\ \overline{\mathbf{G}}_{31} & \overline{\mathbf{G}}_{33} \end{bmatrix} = \left[\begin{array}{c|cc} \widetilde{A}+\widetilde{B}\widetilde{F} & \widetilde{B}_1\overline{V}_{11}+\widetilde{B}_3\overline{V}_{31} & \widetilde{B}_3\overline{V}_{33} \\ \hline \widetilde{F}_1 & \overline{V}_{11} & O \\ \widetilde{F}_3 & \overline{V}_{31} & \overline{V}_{33} \end{array}\right],$$

*si* $\widetilde{F} := -V^{-1}W = \begin{bmatrix} \widetilde{F}_1^T & \widetilde{F}_2^T & \widetilde{F}_3^T \end{bmatrix}^T$.

# Capitolul 16: Problema de Reglare $H^2$ Optimala

In acest capitol consideram problema de control (reglare) $H^2$ optimala care consta in gasirea unui regulator care **stabilizeaza intern si minimizeaza norma $H^2$ a sistemului rezulta in bucla inchisa**.

Fromularea problemei

Evaluarea normei $H^2$

Rezultatul central

# Formularea problemei

Consideram un sistem dinamic liniar

$$\begin{aligned} \dot{x} &= Ax + B_1 u_1 + B_2 u_2, \\ y_1 &= C_1 x + D_{12} u_2, \\ y_2 &= C_2 x + D_{21} u_1, \end{aligned} \tag{237}$$

avand matricea de transfer data de : $\mathbf{T} = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] =$

$$\left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & O & D_{12} \\ C_2 & D_{21} & O \end{array}\right] = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix}$$ unde $u$ si $y$ sunt intrarea si

iesirea lui $\mathbf{T}$ partitionate in acord cu $\mathbf{T}$ a.i.

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{T}u = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}. \tag{238}$$

- $u_1$: vectorul intrarilor exogene (perturbatii, zgomote si semnale referinta)
- $y_1$: vectorul iesirilor reglate (semnale de tip eroare)
- $u_2$: vectorul intrarilor reglate
- $y_2$: vectorul iesirilor masurate

Controlul (Reglarea) sistemului: folosind informatia masurata $y_2$ gasiti o marime de comanda $u_2$ a.i. raspunsul rezultant $y_1$ la intrarea

$u_1$ sa indeplineasca anumite cerinte fixate prin proiectare.

Problema: Gasiti un $u_2$ potrivit, ca iesire a unui sistem numit regulator si avand intrarea $y_2$, a.i. raspunsul sistemului rezultant **sa aibe energie cat mai mica $y_1$ cand intrarea in sistem $u_1$ este un impuls Dirac**.

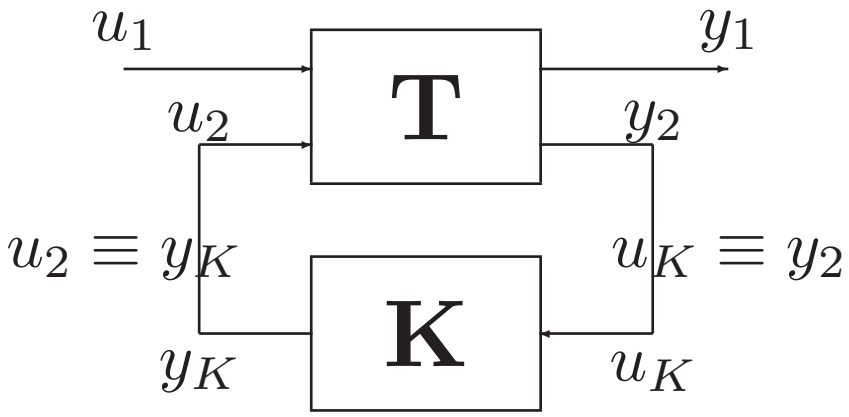
Definim un regulator pentru sistemul dat ca fiind alt sistem dinamic liniar

$$\begin{aligned} \dot{x}_K &= A_K x_K + B_K u_K, \\ y_K &= C_K x_K, \end{aligned} \tag{239}$$

avand matricea de transfer $\mathbf{K} = \left[\begin{array}{c|c} A_K & B_K \\ \hline C_K & O \end{array}\right]$, unde $u_K$ si $y_K$ sunt intrarea si iesirea lui $\mathbf{K}$.

Sistemul in bucla inchisa se obtine conectand regulatorul la sistemul

dat a.i. $u_2 \equiv y_K$ si $u_K \equiv y_2$.




Sistemul in bucla inchisa rezultat avand intrarea $u_1$ si iesirea $y_1$ este bine definit si dat de

$$y_1 = \mathbf{T}_{y_1u_1}u_1, \quad \mathbf{T}_{y_1u_1} = \mathrm{LFT}(\mathbf{T},\mathbf{K}) = \mathbf{T}_{11} + \mathbf{T}_{12}\mathbf{K}(I - \mathbf{T}_{22}\mathbf{K})^{-1}\mathbf{T} \tag{240}$$

$$\mathbf{T}_{y_1u_1} \;=\; \left[\begin{array}{c|c} A_{_R} & B_{_R} \\ \hline C_{_R} & O \end{array}\right] = \left[\begin{array}{cc|c} A & B_2C_{_K} & B_1 \\ B_{_K}C_2 & A_{_K} & B_{_K}D_{21} \\ \hline C_1 & D_{12}C_{_K} & O \end{array}\right].$$

Problema de Control Optimal $H^2$: Dandu-se $\mathbf{T}$, gasiti un regulator strict propriu $\mathbf{K}$ pentru care sistemul in bucla inchisa $\mathbf{T}_{y_1u_1}$:

1. Este intern stabil ($\Lambda(A_{_R}) \subset \mathbb{C}_-$)

2. Are norma $H^2$ minima ( $\|\mathbf{T}_{y_1u_1}\|_2$ isi atinge minimul peste clasa tuturor sistemelor
   $\mathcal{K}$= { $\mathbf{K}$ : $\mathbf{K}$ strict proprii si care stabilizeaza intern $\mathbf{T}_{y_1u_1}$}).

- Regulator Stabilizant : Un regulator care satisface conditia de stabilitate interna

- Regulator Optimal: Un regulator care atinge minimul normei $L^2$

- Am presupus implicit: (i) $\mathbf{T}_{11}(\infty) = D_{11} = 0$; (ii) $\mathbf{T}_{22}(\infty) =$

$D_{22} = 0$; (iii) $\mathbf{K}(\infty) = D_K = 0$;

- Ipotezele (i) si (iii) asigura $\mathbf{T}_{y_1 u_1}(\infty) = D_R = 0$ care este o conditie necesara pentru finitudinea normei $H^2$ a sistemului in bucla inchisa;

- Ipoteza (iii) asigura automat ca sistemul in bucla inchisa este bine definit;

- Presupunerea (ii) este facuta pentru simplificarea formulelor (nu se pierde din generalitate); intr-adevar, daca $\mathbf{K}$ rezolva problema cu $D_{22} = 0$, atunci $\mathbf{K}(I + D_{22}\mathbf{K})^{-1}$ rezolva problema originala cu $D_{22}$ arbitrar;

- Nici (i) nici (iii) nu sunt necesare pentru rezolvarea problemei de reglare $H^2$ si pot fii relaxate (s-au facut pentru simplificarea expunerii).

# O evaluare a normei $L^2$

Dam in continuare o evaluare utila a normei $L^2$ a unui sistem stabil in termenii iesirii sistemului atunci cand semnalul de intrare este un impuls Dirac.

Consideram sistemul stabil ($A$ stabila)

$$\mathbf{T} = \left[\begin{array}{c|c} A & B \\ \hline C & O \end{array}\right], \qquad y = \mathbf{T}u. \tag{241}$$

Norma $L^2$ a lui $\mathbf{T}$ este definita ca

$$\|\mathbf{T}\|_2 := \left\{ \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{Trace}\,[\mathbf{T}^*(j\omega)\mathbf{T}(j\omega)]\mathrm{d}\omega \right\}^{\frac{1}{2}}. \tag{242}$$

Fie $u^i := \delta e_i$, $(i = 1, \ldots, m)$, unde $\delta$ este distributia Dirac si $e_1, \ldots, e_m$ este baza Euclidiana standard in $\mathbb{R}^m$.

Fie $y^i$ iesirea pentru $u = u^i$, $(i = 1, \ldots, m)$.

Din stabilitatea lui $A$ rezulta ca $y^i \in L_+^{2,p}$ si are forma explicita:

$$y^i(t) = 0, \text{ ptr. } t < 0,$$

$$y^i(t) = Ce^{At}Be_i, \text{ ptr. } t \geq 0.$$

Fie deasemenea matricea de raspuns cauzal la impuls:

$$\mathcal{G}(t) := 0, \text{ ptr. } t < 0$$

$$\mathcal{G}(t) = Ce^{At}B, \text{ ptr. } t \geq 0.$$

Avem succesiv

$$\begin{aligned}\sum_{i=1}^{m} \|y^i\|_2^2 &= \sum_{i=1}^{m} \int_0^\infty e_i^T \mathcal{G}^T(t)\mathcal{G}(t)e_i \mathrm{d}t = \int_0^\infty [\sum_{i=1}^{m} e_i^T \mathcal{G}^T(t)\mathcal{G}(t)e_i]\mathrm{d}t \\ &= \int_0^\infty \mathrm{Trace}\,[\mathcal{G}^T(t)\mathcal{G}(t)]\mathrm{d}t = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{Trace}\,[\mathbf{T}^*(j\omega)\mathbf{T}(j\omega)]\end{aligned}$$

in care ultima ecuatie este o consecinta a formulei lui Parseval. Deci

$$\|\mathbf{T}\|_2^2 = \sum_{i=1}^{m} \|y^i\|_2^2. \tag{244}$$

# Principalul rezultat

**Teorema 178.** *Fie sistemul* $\mathbf{T}$. *Asociem doua triplete Popov*

$$\Sigma_{12} = (A, B_2; C_1^T C_1, C_1^T D_{12}, D_{12}^T D_{12}), \tag{245}$$

$$\Sigma_{21} = (A^T, C_2^T; B_1 B_1^T, B_1 D_{21}^T, D_{21} D_{21}^T). \tag{246}$$

*Presupunem:*

**(H1)** *Perechea* $(A, B_2)$ *este stabilizabila,* $D_{12}$ *are rang intreg pe coloane, si*

$$\operatorname{rank} \begin{bmatrix} A - j\omega I & B_2 \\ C_1 & D_{12} \end{bmatrix} = n + m_2, \quad \forall\, \omega \in \mathbb{R}.$$

**(H2)** *Perechea* $(C_2, A)$ *este detectabila,* $D_{21}$ *are rang intreg pe linii, si*

$$\operatorname{rank} \begin{bmatrix} A - j\omega I & B_1 \\ C_2 & D_{21} \end{bmatrix} = n + p_2, \quad \forall\, \omega \in \mathbb{R}.$$

*Atunci ARE(*$\Sigma_{12}$*) data de*

$$A^T X + X A - (X B_2 + C_1^T D_{12})(D_{12}^T D_{12})^{-1}(B_2^T X + D_{12}^T C_1) + C_1^T C_1 = 0 \tag{247}$$

*ate o solutie stabilizanta* $X_s \geq 0$*, si ARE(*$\Sigma_{21}$*) data de*

$$A Y + Y A^T - (Y C_2^T + B_1 D_{21}^T)(D_{21} D_{21}^T)^{-1}(C_2 Y + D_{21} B_1^T) + B_1 B_1^T = 0 \tag{248}$$

*are o solutie stabilizanta* $Y_s \geq 0$.

*Mai mult, exista un regulator strict propriu*

$$\mathbf{K} = \left[\begin{array}{c|c} A + B_2F_s + K_s^TC_2 & K_s^T \\ \hline -F_s & O \end{array}\right] \tag{249}$$

*care rezolva problema de reglare $H^2$ optimala, unde*

$$F_s := -(D_{12}^TD_{12})^{-1}(B_2^TX_s + D_{12}^TC_1),$$

$$K_s := -(D_{21}D_{21}^T)^{-1}(C_2Y_s + D_{21}B_1^T),$$

*sunt reactiile Riccati stabilizante. Valoarea optimala (minimala) a normei $H^2$ este*

$$\min_{\mathbf{K}\in\mathcal{K}} \|\mathbf{T}_{y_1u_1}\|_2 = [\mathrm{Trace}\,(B_1^TX_sB_1) + \mathrm{Trace}\,(C_1Y_sC_1^T)]^{\frac{1}{2}}. \tag{250}$$

Observatii:

- $(A, B_2)$ stabilizabila si $(C_2, A)$ detectabila sunt necesare pentru indeplinirea cerintei de stabilitate interna.

- Celelalte ipoteze sunt necesare pentru existenta solutiei stabilizante a celor doua ecuatii Riccati dar NU sunt necesare pentru existenta unei solutii a problemei $H^2$ optimale. In caz ca una dintre cele doua ipoteze nu este indeplinita obtinem o asa numita problema singulara.

Demonstratia se bazeaza pe solutia a trei probleme particulare $H^2$ optimale, fiecare in parte fiind data sub forma unei propozitii separate: 1 bloc, 2 bloc si 2 bloc duala.

Schema de demonstratie este:

1. Rezolvam problema 1 bloc

2. Reducem problema 2 bloc la cea 1 bloc

3. Scriem solutia problemei 2 bloc duale prin argumente de dualitate

4. Reducem cazul general la o problema 2 bloc duala

**Propozitia 179. [1 bloc]** *Presupunem ca* $\mathbf{T}$ *satisface ipotezele:*

**(1 bloc–H1)** $D_{12}$ *este patrata si inversabila si* $A - B_2 D_{12}^{-1} C_1$ *este stabila.*

**(1 bloc–H2)** $D_{21}$ *este patrata si inversabila si* $A - B_1 D_{21}^{-1} C_2$ *este stabila.*

*Atunci o solutie si valoare optimala a normei $H^2$ sunt date de*

$$\mathbf{K} = \left[\begin{array}{c|c} A - B_1 D_{21}^{-1} C_2 - B_2 D_{12}^{-1} C_1 & B_1 D_{21}^{-1} \\ \hline -D_{12}^{-1} C_1 & O \end{array}\right], \qquad (251)$$

$$\min_{\mathbf{K} \in \mathcal{K}} \|\mathbf{T}_{y_1 u_1}\|_2 = 0.$$

**Demonstratie.** Inlocuim expresia regulatorului si obtinem

$$\mathbf{T}_{y_1 u_1} = \left[\begin{array}{cc|c} A & -B_2 D_{12}^{-1} C_1 & B_1 \\ B_1 D_{21}^{-1} C_2 & A - B_1 D_{21}^{-1} C_2 - B_2 D_{12}^{-1} C_1 & B_1 \\ \hline C_1 & -C_1 & O \end{array}\right]$$

$$= \left[\begin{array}{cc|c} A - B_2 D_{12}^{-1} C_1 & -B_2 D_{12}^{-1} C_1 & B_1 \\ O & A - B_1 D_{21}^{-1} C_2 & O \\ \hline O & -C_1 & O \end{array}\right]. \qquad (252)$$

■

**Propozitia 180. [2 bloc]** *Presupunem ca* $\mathbf{T}$ *satisface urmatoarele ipoteze:*

**(2 bloc–H1)** *Perechea* $(A, B_2)$ *este stabilizabila,* $D_{12}$ *are rang intreg pe coloane, si*

$$\operatorname{rank} \left[\begin{array}{cc} A - j\omega I & B_2 \\ C_1 & D_{12} \end{array}\right] = n + m_2, \quad \forall\, \omega \in \mathbb{R}.$$

**(2 bloc–H2)** *$D_{21}$ este patrata si inversabila si $A - B_1 D_{21}^{-1} C_2$ este stabila.*

*Atunci ARE*

$$A^T X + X A - (X B_2 + C_1^T D_{12})(D_{12}^T D_{12})^{-1}(B_2^T X + D_{12}^T C_1) + C_1^T C_1 = 0 \tag{253}$$

*are o solutie stabilizanta $X_s \geq 0$. O solutie si valoarea optimala a normei $H^2$ sunt date de*

$$\mathbf{K} = \left[\begin{array}{c|c} A - B_1 D_{21}^{-1} C_2 + B_2 F_s & B_1 D_{21}^{-1} \\ \hline F_s & O \end{array}\right], \tag{254}$$

$$\min_{\mathbf{K} \in \mathcal{K}} \|\mathbf{T}_{y_1 u_1}\|_2 = [\mathrm{Trace}\,(B_1^T X_s B_1)]^{\frac{1}{2}}. \tag{255}$$

**Demonstratie.** Din (2 bloc–H1) rezulta ca ARE are o solutie stabilizanta $X_s \geq 0$. Mai mult, existenta solutiei stabilizante $X_s$ a ARE($\Sigma_{12}$) este echivalenta cu existenta solutiei stabilizante $(X_s, V_s, W_s)$, $X_s \geq 0$, a KPYS

$$\begin{aligned} D_{12}^T D_{12} &= V^T V, \\ C_1^T D_{12} + X B_2 &= W^T V, \\ C_1^T C_1 + A^T X + X A &= W^T W, \end{aligned} \tag{256}$$

unde $F_s = -V_s^{-1}W_s$ este reactia stabilizanta. Consideram sistemul

$$\widetilde{\mathbf{T}} = \begin{bmatrix} \widetilde{\mathbf{T}}_{11} & \widetilde{\mathbf{T}}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix} = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline W_s & O & V_s \\ C_2 & D_{21} & O \end{array}\right], \tag{257}$$

$$\begin{bmatrix} \widetilde{y}_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{T}}_{11} & \widetilde{\mathbf{T}}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

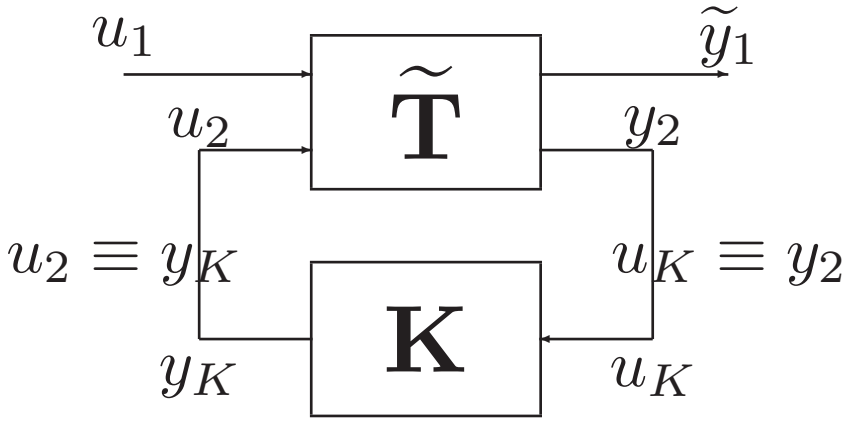
care se obtine inlocuind iesirea $y_1 = C_1 x + D_{12} u_2$ cu iesirea

$$\widetilde{y}_1 = W_s x + V_s u_2.$$

Cum $V_s$ este inversabil, se verifica imediat ca $\widetilde{\mathbf{T}}$ satisface ipotezele (1 bloc–H1) si (1 bloc–H2).

Daca scriem regulatorul pentru cazul 1–bloc cu date obtinute din

$\widetilde{\mathbf{T}}$ obtinem (254). Cu alte cuvinte, regulatorul $\mathbf{K}$ din (254) rezolva problema 1–bloc formulata pentru sistemul (257). Prin urmare sistemul in bucla inchisa din figura




dat de

$$\widetilde{\mathbf{T}}_{y_1u_1} = \mathrm{LFT}(\widetilde{\mathbf{T}}, \mathbf{K}) = \widetilde{\mathbf{T}}_{11} + \widetilde{\mathbf{T}}_{12}\mathbf{K}(I - \mathbf{T}_{22}\mathbf{K})^{-1}\mathbf{T}_{21}, \qquad (258)$$

este intern stabil si

$$\|\widetilde{\mathbf{T}}_{y_1u_1}\|_2^2 = \sum_{i=1}^{m} \|\widetilde{y}_1^i\|_2^2 = 0 \tag{259}$$

unde $\widetilde{y}_1^i$ este iesirea lui $\widetilde{\mathbf{T}}_{y_1u_1}$ pentru intrarea $u_1 = u_1^i = \delta e_i$. Stabilitatea interna a (258) este echivalenta cu faptul ca matricea de stare

$$\widetilde{A}_R = \begin{bmatrix} A & B_2F_s \\ B_1D_{21}^{-1}C_2 & A - B_1D_{21}^{-1}C_2 + B_2F_s \end{bmatrix}$$

este stabila. In orice caz, este usor de sa vedem ca $\widetilde{A}_R$ este matricea de stare a lui

$$\mathbf{T}_{y_1u_1} = \mathrm{LFT}(\mathbf{T}, \mathbf{K}) = \mathbf{T}_{11} + \mathbf{T}_{12}\mathbf{K}(I - \mathbf{T}_{22}\mathbf{K})^{-1}\mathbf{T}_{21}, \tag{260}$$

si prin urmare $\mathbf{T}_{y_1u_1}$ este deasemenea intern stabila. Rezulta ca $\mathbf{K}$

satisface cerinta de stabilitate a problemei $H^2$ formulate pentru $\mathbf{T}$ si ramane sa aratam ca cerinta de optimalitate este deasemenea indeplinita. Evaluam succesiv:

$$\begin{aligned}
\|y_1(t)\|^2 &= \|C_1x(t)+D_{12}u_2(t)\|^2 \\
&= x^T(t)C_1^TC_1x(t)+2x^T(t)C_1^TD_{12}u_2(t)+u_2^T(t)D_{12}^TD_{12}u_2(t) \\
&\overset{(256)}{=} x^T(t)W_s^TW_sx(t)-x^T(t)(A^TX_s+X_sA)x(t)+2x^T(t)W_s^TV_su_2( \\
&\quad -2x^T(t)X_sB_2u_2(t)+u_2^T(t)V_s^TV_su_2(t) \\
&= \|W_sx(t)+V_su_2(t)\|^2-x^T(t)X_s(Ax(t)+B_2u_2(t)) \\
&\quad -(Ax(t)+B_2u_2(t))^TX_sx(t) \\
&= \|\widetilde{y}_1(t)\|^2-x^T(t)X_s\dot{x}(t)-\dot{x}^T(t)X_sx(t) \\
&\quad +x^T(t)X_sB_1u_1(t)+u_1^T(t)B_1^TX_sx(t)
\end{aligned}$$

$$= \quad \|\widetilde{y}_1(t)\|^2 - \frac{\mathrm{d}}{\mathrm{d}t}(x^T(t)X_s x(t)) + x^T(t)X_s B_1 u_1(t) + u_1^T(t)B_1^T X_s x($$

Fie $\mathbf{K}$ orice regulator a.i. $\mathbf{T}_{y_1 u_1} = \mathrm{LFT}(\mathbf{T}, \mathbf{K})$ este intern stabila. Asa cum a rezultat mai sus, $\widetilde{\mathbf{T}}_{y_1 u_1} = \mathrm{LFT}(\widetilde{\mathbf{T}}, \mathbf{K})$ este desemenea intern stabila.

Fie $x_{_K}(t)$ si $x_{_R}(t)$ vectorul de stare al lui $\mathbf{K}$ si respectiv al lui $\mathbf{T}_{y_1 u_1}$. Consideram intrarea impulsiva $u_1 = u_1^i := \delta e_i$ si conditia initiala $x(0_-) = 0, x_{_K}(0_-) = 0$. Atunci obtinem cu (240),

$$x_{_R}(0_+) = \left[ \begin{array}{c} x(0_+) \\ x_{_K}(0_+) \end{array} \right] = B_{_R} e_i = \left[ \begin{array}{c} B_1 \\ B_{_K} D_{21} \end{array} \right] e_i.$$

Prin urmare

$$x(0_+) = B_1 e_i, \tag{261}$$

$x_{_K}(0_+) = B_{_K}D_{21}e_i$, $u_1^i(t) = 0$, $\forall t > 0$, si toate semnalele din bucla inchisa ale $\mathbf{T}_{y_1u_1}$ si $\widetilde{\mathbf{T}}_{y_1u_1}$ sunt **functii exponential descrescatoare in** $L_+^2$. Notam $y_1^i$ si $\widetilde{y}_1^i$ iesirile lui $\mathbf{T}_{y_1u_1}$ si $\widetilde{\mathbf{T}}_{y_1u_1}$ la intrarea $u_1^i$.

Integram pe $(0,\infty)$ ambii membrii pentru $u_1(t) = u_1^i := \delta e_i$ si cu (261) obtinem succesiv

$$\|y_1^i\|_2^2 = \|\widetilde{y}_1^i\|_2^2 + e_i^T B_1^T X B_1 e_i, \quad \sum_{i=1}^{m_1} \|y_1^i\|_2^2 = \sum_{i=1}^{m_1} \|\widetilde{y}_1^i\|_2^2 + \mathrm{Trace}\,(B_1^T X_s B_1$$

de unde concluzionam

$$\|\mathbf{T}_{y_1u_1}\|_2^2 = \|\widetilde{\mathbf{T}}_{y_1u_1}\|_2^2 + \mathrm{Trace}\,(B_1^T X_s B_1).$$

Aceasta egalitate are loc pentru toate regulatoarele $\mathbf{K}$ pentru care $\mathbf{T}_{y_1u_1}$ (si automat $\widetilde{\mathbf{T}}_{y_1u_1}$) este intern stabila. Dar $\|\mathbf{T}_{y_1u_1}\|_2^2 \geq$

$\mathrm{Trace}\,(B_1^T X B_1)$ si egalitatea are loc pentru $\mathbf{K}$ dat de (254) deoarece are loc (259). Din aceasta concluzionam ca (255) are loc si demonstratia este incheiata. ∎

**Propozitia 181. [2 bloc duala]** *Prespunem ca* $\mathbf{T}$ *satisface urmatoarele ipoteze:*

**(2 bloc dual–H1)** $D_{12}$ *este patrata si inversabila si* $A - B_2 D_{12}^{-1} C_1$ *este stabila.*

**(2 bloc dual–H2)** *Perechea* $(C_2, A)$ *este detectabila,* $D_{21}$ *are rang intreg pe linii si*

$$\mathrm{rank} \begin{bmatrix} A - j\omega I & B_1 \\ C_2 & D_{21} \end{bmatrix} = n + p_2, \quad \forall\, \omega \in \mathbb{R}.$$

*Atunci ARE($\Sigma_{21}$)*

$$AY+YA^T-(YC_2^T+B_1D_{21}^T)(D_{21}D_{21}^T)^{-1}(C_2Y+D_{21}B_1^T)+B_1B_1^T=0 \tag{262}$$

*are o solutie stabilizanta* $Y_s \geq 0$. *O solutie si valoarea optimala (minima) a normei* $H^2$ *sunt date de*

$$\mathbf{K}=\left[\begin{array}{c|c} A-B_2D_{12}^{-1}C_1+K_s^TC_2 & K_s^T \\ \hline D_{12}^{-1}C_1 & O \end{array}\right] \tag{263}$$

$$\min_{\mathbf{K}\in\mathcal{K}}\|\mathbf{T}_{y_1u_1}\|_2=[\mathrm{Trace}\,(C_1Y_sC_1^T)]^{\frac{1}{2}}. \tag{264}$$

**Demonstratie.** Demonstratia este o consecinta a dualitatii cu problema 2 bloc. Mai precis, daca $\mathbf{T}$ satisface ipotezele problemei 2

bloc duale, atunci $\mathbf{T}^T$ satisface ipotezele problemei 2 bloc. Scriem solutia $\mathbf{K}$ a problemei 2 bloc formulate pentru $\mathbf{T}^T$ si obtinem ca $\mathbf{K}^T$ este o solutie a problemei 2 bloc duale formulate pentru $\mathbf{T}$. Restul demonstratiei este o simpla chestiune de substitutii in formule. ∎

Demonstratia Teoremei 178. Observam intai ca cele doua ARE sunt ecuatiile Riccati de pozitivitate asociate cu $\mathbf{T}_{12}$ si respectiv $\mathbf{T}_{21}$. Atunci cu (H1) si (H2) rezulta ca Riccati-urile corespunzatoare au solutii stabilizante $X_s \geq 0$ si $Y_s \geq 0$.

Mai departe urmam acceasi linie de rationament ca la cazul 2 bloc. Existenta solutiei stabilizante $X_s$ a ARE este echivalenta cu existenta unei solutii stabilizante $(X_s, V_s, W_s)$, $X_s \geq 0$, a KPYS dat in (256). Consideram din nou sistemul (257) care se obtine inlocuind iesirea

$y_1$ cu iesirea $\widetilde{y}_1$. Cum $V_s$ este inversabil, este usor de verificat ca sistemul $\widetilde{\mathbf{T}}$ satisface acum ipotezele problemei 2 bloc duale. Daca scriem regulatorul corespunzator pentru $\widetilde{\mathbf{T}}$ dat de (257) ajungem exact la (249). Cu alte cuvinte, regulatorul $\mathbf{K}$ in (249) rezolva problema 2 bloc duala formulata pentru sistemul (257).

Deci $\widetilde{\mathbf{T}}_{y_1u_1}$ este desemenea intern stabil. Din stabilitatea interna a $\widetilde{\mathbf{T}}_{y_1u_1}$ rezulta in mod analog ca in demonstratia Propozitiei 180 ca $\mathbf{T}_{y_1u_1}$ este deasemenea intern stabil.

A ramas de aratat ca $\mathbf{K}$ este intr–adevar optimal si sa obtinem evaluarea (250). In acest scop fie $\mathbf{K}$ orice regulator a.i. $\mathbf{T}_{y_1u_1}$ este intern stabil. Din Propozitia 181 aplicata lui (257) obtinem

$$\|\widetilde{\mathbf{T}}_{y_1u_1}\|_2^2 \geq \mathrm{Trace}\,(C_1 Y_s C_1^T) \tag{265}$$

si mai departe

$$\|\mathbf{T}_{y_1u_1}\|_2^2 \geq \mathrm{Trace}\,(B_1^T X_s B_1) + \mathrm{Trace}\,(C_1 Y_s C_1^T). \tag{266}$$

Egalitatea din (266) are loc daca si numai daca egalitatea din (265) are loc. Dar, asa cum am explicat mai sus, pentru $\mathbf{K}$ dat de (249) egalitatea (265) are loc si prin urmare rezulta (250) incheind astfel demonstratia Teoremei 178. ∎

# Capitolul 17 Problema de reglare $H^\infty$ (sub)optimala

In acest capitol consideram problema de reglare $H^\infty$ (sub)optimala (numita si problema atenuarii perturbatiei) ce consta in gasirea clasei tuturor regulatoarelor care stabilizeaza intern sistemul in bucla inchisa si ii minimizeaza norma $H^\infty$ (o fac mai mica decat un $\gamma > 0$).

Solutia propusa se bazeaza din nou pe conditia de signatura pe care deja stim s-o rezolvam.

Formularea problemei

Presupuneri de baza

Problema $H^\infty$ si conditia de signatura

## Solutia

## Schita demonstratiei

# Formularea problemei

Cadrul este complet similar celui din capitolul anterior. Consideram un sistem dinamic liniar:

$$\begin{aligned} \dot{x} &= Ax + B_1 u_1 + B_2 u_2, \\ y_1 &= C_1 x + D_{11} u_1 + D_{12} u_2, \\ y_2 &= C_2 x + D_{21} u_1 + D_{22} u_2, \end{aligned} \tag{267}$$

avand matricea de transfer data de : $\mathbf{T} = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{array}\right] = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix}$ unde $u$ si $y$ semnifica intrarea

si iesirea lui $\mathbf{T}$ partitionate conform a.i.

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{T}u = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}. \tag{268}$$

- $u_1$: vectorul intrarilor exogene
- $y_1$: vectorul iesirilor reglate
- $u_2$: vectorul intrarilor reglate
- $y_2$: vectorul iesirilor masurate

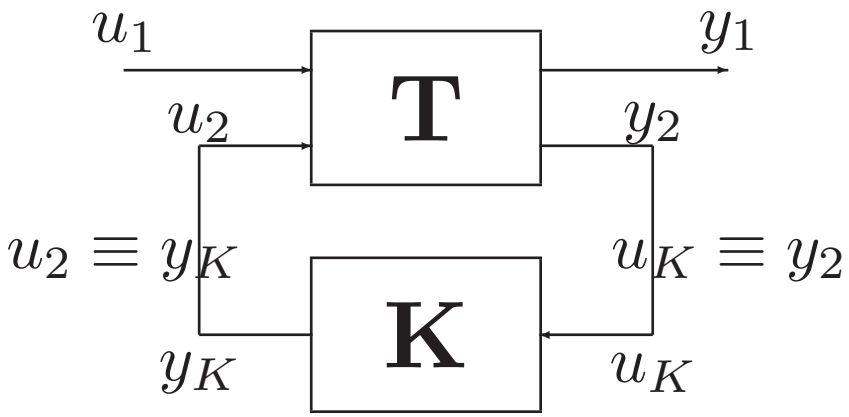
Problema: Gasiti o intrare potrivita $u_2$ (ce este iesire a unui sistem numit regulator avand intrarea $y_2$) astfel incat raspunsul rezultat $y_1$ **sa aibe energie cat mai mica** atunci cand intrarea $u_1$ este cel mai

prost semnal posibil de energie unitara (i.e., o functie in $L^2$, de norma 1, aleasa astfel incat sa maximizeze $y_1$). Cum $u_1$ este vazut ca un semnal perturbator conditia poate fi privita ca o cerinta de atenuare a perturbatiei.

Definim regulatorul prin formulele generale

$$\begin{aligned} \dot{x}_K &= A_K x_K + B_K u_K, \\ y_K &= C_K x_K + D_K u_K, \end{aligned} \tag{269}$$

avand matricea de transfer $\mathbf{K} = \left[\begin{array}{c|c} A_K & B_K \\ \hline C_K & D_K \end{array}\right]$, unde $u_K$ si $y_K$ sunt intrarea si iesirea lui $\mathbf{K}$. Sistemul in bucla inchisa se obtine conectand regulatorul la sistemul nominal a.i. $u_2 \equiv y_K$ si $u_K \equiv y_2$.




Sistemul in bucla inchisa este bine definit daca matricea $\begin{bmatrix} I_{p_2} & D_{22} \\ D_K & I_{m_2} \end{bmatrix}$ este nensingulara. Buna definire poate fi exprimata echivalent prin conditia $S = I_{p_2} - D_{22}D_K$ nensingulara, sau ca $\widetilde{S} := I_{m_2} - D_K D_{22}$ nensingulara. Daca sistemul in bucla inchisa este bine definit atunci sistemul rezultat avand intrarea $u_1$ si iesirea $y_1$

este dat de

$$\mathbf{T}_{y_1u_1} = \mathrm{LFT}(\mathbf{T}, \mathbf{K}) = \mathbf{T}_{11} + \mathbf{T}_{12}\mathbf{K}(I - \mathbf{T}_{22}\mathbf{K})^{-1}\mathbf{T}_{21},$$

$$\mathbf{T}_{y_1u_1} = \left[\begin{array}{c|c} A_R & B_R \\ \hline C_R & D_R \end{array}\right]$$

$$= \left[\begin{array}{cc|c} A + B_2\widetilde{S}^{-1}D_KC_2 & B_2\widetilde{S}^{-1}C_K & B_1 + B_2\widetilde{S}^{-1}D_KD \\ B_KS^{-1}C_2 & A_K + B_KS^{-1}D_{22}C_K & B_KS^{-1}D_{21} \\ \hline C_1 + D_{12}D_KS^{-1}C_2 & D_{12}\widetilde{S}^{-1}C_K & D_{11} + D_{12}D_KS^{-1} \end{array}\right] \quad (270)$$

Problema $H^\infty$ (sub)optimala: Dandu-se un sistem $\mathbf{T}$ si $\gamma > 0$, gasiti toate regulatoarele $\mathbf{K}$ pentru care sistemul in bucla inchisa $\mathbf{T}_{y_1u_1}$ este bine definit ($I - D_{22}D_K$ este inversabila), intern stabil, i.e.,

$$\Lambda(A_{_R}) \subset \mathbb{C}_-, \tag{271}$$

si marginit in norma $H^\infty$ de $\gamma$, i.e.,

$$\|\mathbf{T}_{y_1u_1}\|_\infty < \gamma. \tag{272}$$

Deoarece $\|\mathbf{T}_{y_1u_1}\|_\infty = \sup_{\|u_1\|_2=1} \|y_1\|_2$, rezulta ca (272) impune de fapt o margine asupra normei $L_2$ a raspunsului sistemului atunci cand intrarea este arbitrara de norma $L_2$ egala cu unitatea.

Observatie: Calcule simple arata ca realizarea lui $(I - \mathbf{T}_{22}\mathbf{K})^{-1}$, $\mathbf{T}_{y_2u_1} := (I - \mathbf{T}_{22}\mathbf{K})^{-1}\mathbf{T}_{21}$, si $\mathbf{T}_{u_2u_1} = \mathbf{K}(I - \mathbf{T}_{22}\mathbf{K})^{-1}\mathbf{T}_{21}$ impart

aceeasi matrice de stare care este exact $A_R$:

$$(I-\mathbf{T}_{22}\mathbf{K})^{-1} = \left[\begin{array}{c|c} A_R & \star \\ \hline \star & \star \end{array}\right], \quad \mathbf{T}_{y_2u_1} = \left[\begin{array}{c|c} A_R & \star \\ \hline \star & \star \end{array}\right], \quad \mathbf{T}_{u_2u_1} = \left[\begin{array}{c|c} A_R & \star \\ \hline \star & \star \end{array}\right],$$

Daca regulatorul $\mathbf{K}$ stabilizeaza $\mathbf{T}$, atunci sistemul rezultant este automat stabil, i.e., $\mathbf{T}_{y_1u_1} \in RH^\infty_{+,p_1\times m_1}$, si deasemenea $(I - \mathbf{T}_{22}\mathbf{K})^{-1} \in RH^\infty_{+,p_2\times p_2}$, $\mathbf{T}_{y_2u_1} \in RH^\infty_{+,p_2\times m_1}$, si $\mathbf{T}_{u_2u_1} \in RH^\infty_{+,m_2\times m_1}$. O problema naturala in acest context este determinarea unui $\gamma_{\min}$ peste clasa tuturor regulatoarelor stabilizante, si scrierea unei formule pentru regulator ce asigura optimul (minimul). Problema se numeste optimala si este considerabil mai dificila.

# Presupuneri de baza

Facem o multime de presupuneri aditionale asupra sistemului dat care sau simplifica considerabil expunerea (fara restrangerea generalitatii) sau sunt de natura tehnica intrinseca.

(H1) $\gamma = 1$: Este o simpla scalare

(H2) $D_{22} = 0$: Daca $\mathbf{K}$ este o solutie a problemei in cazul $D_{22} = 0$ atunci $\mathbf{K}(I + D_{22}\mathbf{K})^{-1}$ este o solutie a problemei originale. Aceasta presupunere simplifica considerabil restrictia impusa de buna definire a buclei (implica automat buna definire) si conduce la formule mai simple pentru solutie.

(H3) Perechea $(A, B_2)$ este stabilizabila si perechea $(C_2, A)$ este detectabila: Aceste presupuneri sunt necesare pentru existenta clasei compensatoarelor stabilizante. Intr-adevar:

$$A_R = \begin{bmatrix} A + B_2\widetilde{S}^{-1}D_K C_2 & B_2\widetilde{S}^{-1}C_K \\ B_K S^{-1}C_2 & A_K + B_K S^{-1}D_{22}C_K \end{bmatrix} \quad (2$$

$$= \begin{bmatrix} A & O \\ O & A_K \end{bmatrix} + \begin{bmatrix} B_2 & O \\ O & B_K \end{bmatrix} \begin{bmatrix} \widetilde{S}^{-1}D_K & \widetilde{S}^{-1} \\ S^{-1} & S^{-1}D_{22} \end{bmatrix} \begin{bmatrix} C_2 & O \\ O & C_K \end{bmatrix}$$

Daca $A_R$ este stabila deducem ca perechea matriciala

$$\left( \begin{bmatrix} A & O \\ O & A_K \end{bmatrix}, \begin{bmatrix} B_2 & O \\ O & B_K \end{bmatrix} \right)$$

este stabilizabila si perechea matriciala

$$\left(\begin{bmatrix} C_2 & O \\ O & C_K \end{bmatrix}, \begin{bmatrix} A & O \\ O & A_K \end{bmatrix}\right)$$

este detectabila de unde rezulta ca perechea $(A, B_2)$ este stabilizabila si perechea $(C_2, A)$ este detectabila. Rezulta in particular si ca $(A_K, B_K)$ si $(C_K, A_K)$ trebuie sa fie stabilizabile si respectiv detectabile.

(H4) Ipoteze de regularitate.

(R1) Fascicolul de transmisie $\mathbf{T}_{12}$ dat de

$$\begin{bmatrix} sI - A & -B_2 \\ -C_1 & -D_{12} \end{bmatrix}$$

are rang intreg pe coloane ptr. $s \in \overline{\mathbb{C}}_0$.
(R2) Fascicolul de transmisie $\mathbf{T}_{21}$ dat de

$$\begin{bmatrix} sI - A & -B_1 \\ -C_2 & -D_{21} \end{bmatrix}$$

are rang intreg pe linii ptr. $s \in \overline{\mathbb{C}}_0$.

O problema ce satisface aceste doua presupuneri se numeste regulata, altfel se numeste singulara (mult mai dificila !!!!). Desigur aceste doua presupuneri nu sunt necesare ci sunt datorate complicatiilor tehnice ce apar in absenta lor.

# DAP si conditia de signatura

Este usor de verificat ca functia Popov asociata cu $\Sigma_c$ este data de

$$\mathbf{\Pi}_{\Sigma_c} = \begin{bmatrix} \mathbf{T}_{11}^* \mathbf{T}_{11} - I & \mathbf{T}_{11}^* \mathbf{T}_{12} \\ \mathbf{T}_{12}^* \mathbf{T}_{11} & \mathbf{T}_{12}^* \mathbf{T}_{12} \end{bmatrix}, \qquad \left[\begin{array}{c|cc} A & B_1 & B \\ \hline C_1 & D_{11} & D \end{array}\right] := \left[\begin{array}{c|cc} A & B_1 & B \\ \hline \end{array}\right.$$

$$\left.\begin{array}{c|cc} C_1 & D_{11} & D \end{array}\right]$$

Din presupunerile de regularitate (R1) avem

$$\mathbf{\Pi}_{\Sigma_c,22} = \mathbf{T}_{12}^* \mathbf{T}_{12} > 0 \tag{274}$$

pe $\overline{\mathbb{C}}_0$ si este clar ca $\begin{bmatrix} C_1^T \\ D_{12}^T \end{bmatrix} \begin{bmatrix} C_1 & D_{12} \end{bmatrix} \geq 0.$

Daca luam $\mathbf{S} := \mathbf{T}_{u_2u_1} = \mathbf{K}(I - \mathbf{T}_{22}\mathbf{K})^{-1}\mathbf{T}_{21}$ (functia de transfer in bucla inchisa de la $u_1$ la $u_2$), atunci este clar ca $\mathbf{S} \in RH_+^\infty$ si

$$\left[\begin{array}{cc} I & \mathbf{S}^* \end{array}\right] \mathbf{\Pi}_{\Sigma_c} \left[\begin{array}{c} I \\ \mathbf{S} \end{array}\right] = -I + (\mathbf{T}_{11} + \mathbf{T}_{12}\mathbf{S})^*(\mathbf{T}_{11} + \mathbf{T}_{12}\mathbf{S}).$$

Dar $\mathbf{T}_{11} + \mathbf{T}_{12}\mathbf{S} = \mathbf{T}_{11} + \mathbf{T}_{12}\mathbf{T}_{u_2u_1} = \mathbf{T}_{y_1u_1}$.

Deoarece regulatorul este contractiv, avem $-I + \mathbf{T}_{y_1u_1}^*\mathbf{T}_{y_1u_1} < 0$ pe $\overline{\mathbb{C}}_0$, si rezulta ca

$$\left[\begin{array}{cc} I & \mathbf{S}^* \end{array}\right] \mathbf{\Pi}_{\Sigma_c} \left[\begin{array}{c} I \\ \mathbf{S} \end{array}\right] < 0 \tag{275}$$

pe $\overline{\mathbb{C}}_0$. De aici automat avem ca ARE($\Sigma_c$) - sau KPYS($\Sigma_c$) - are o solutie stabilizanta.

Reciproca este deasemenea valabila si da clasa solutiilor problemei $H^\infty$ atunci cand stim clasa solutiilor inegalitatii de signatura.

# Solutia

Rezultatul central contine doua aspecte. Pe de-o parte sunt conditiile de existenta ale solutiei, care sunt in forma necesara si suficienta si care se pot verifica usor dpdv numeric. Pe de-alta parte este formula explicita ce genereaza clasa tutoror regulatoarelor (atunci cand exista). Conditiile de existenta si clasa regulatoarelor sunt exprimate in termenii solutiei stabilizante a sistemului Kalman–Popov–Yakubovich (sau, echivalent, a ARE).

**Teorema 182.** *Presupunem ca* $\mathbf{T}$ *satisface ipotezele (H2)-(H4). Asociem cu* $\mathbf{T}$ *tripletele Popov :*

$$\Sigma_c = (A, \left[\begin{array}{cc} B_1 & B_2 \end{array}\right]; Q_c, L_c, R_c); \quad \Sigma_o = (A^T, \left[\begin{array}{cc} C_1^T & C_2^T \end{array}\right]; Q_o, L_o, R_o$$

*unde*

$$Q_c = C_1^T C_1; L_c = C_1^T \left[ \begin{array}{cc} D_{11} & D_{12} \end{array} \right]; R_c = \left[ \begin{array}{c} D_{11}^T \\ D_{12}^T \end{array} \right] \left[ \begin{array}{cc} D_{11} & D_{12} \end{array} \right] - \left[ \begin{array}{c} I_m \\ O \end{array} \right.$$

$$Q_o = B_1 B_1^T; L_o = B_1 \left[ \begin{array}{cc} D_{11}^T & D_{21}^T \end{array} \right]; R_o = \left[ \begin{array}{c} D_{11} \\ D_{21} \end{array} \right] \left[ \begin{array}{cc} D_{11}^T & D_{21}^T \end{array} \right] - \left[ \begin{array}{c} I_{p_1} \\ O \end{array} \right.$$

*Fie*

$$J_c = \left[ \begin{array}{cc} -I_{m_1} & O \\ O & I_{m_2} \end{array} \right], \quad J_o = \left[ \begin{array}{cc} -I_{p_1} & O \\ O & I_{p_2} \end{array} \right].$$

*Atunci avem :*

**(I)** *Problema $H^\infty$ suboptimala are solutie daca si numai daca urmatoarele conditii sunt indeplinite:*

*(C1) Sistemul Kalman–Popov–Yakubovich KPYS($\Sigma_c, J_c$) are*

*o solutie stabilizanta* $(X, V_c, W_c) = (X, \begin{bmatrix} V_{c11} & O \\ V_{c21} & V_{c22} \end{bmatrix}, \begin{bmatrix} W_{c1} \\ W_{c2} \end{bmatrix})$, *cu* $X \geq 0$.

*(C2) Sistemul Kalman–Popov–Yakubovich KPYS*$(\Sigma_o, J_o)$ *are o solutie stabilizanta* $(Y, V_o, W_o) = (Y, \left[ \begin{array}{cc} V_{o11} & O \\ V_{o21} & V_{o22} \end{array} \right], \left[ \begin{array}{c} W_{o1} \\ W_{o2} \end{array} \right])$, *cu* $Y \geq 0$.

*(C3)* $\rho(XY) < 1$.

**(II)** *Presupunem (C1). Fie* $J_\times = \left[ \begin{array}{cc} -I_{m_2} & O \\ O & I_{p_2} \end{array} \right]$, *si fie:*

$$\Sigma_\times = (A^T + F_1^T B_1^T, \left[ \begin{array}{cc} -F_2^T V_{c22}^T & C_2^T + F_1^T D_{21}^T \end{array} \right]; Q_\times, L_\times, R_\times),$$

*unde* $\begin{bmatrix} F_1 \\ F_2 \end{bmatrix}$ *este reactia stabilizanta a KPYS*$(\Sigma_c,\ J_c)$ *si* $Q_\times := B_1(V_{c11}^T V_{c11})^{-1} B_1^T$;

$$L_\times := B_1(V_{c11}^T V_{c11})^{-1} \begin{bmatrix} V_{c21}^T & D_{21}^T \end{bmatrix}; R_\times := \begin{bmatrix} V_{c21} \\ D_{21} \end{bmatrix} (V_{c11}^T V_{c11})^{-1} \begin{bmatrix} V_{c2}^T$$

*Problema* $H^\infty$ *are solutie daca si numai daca (C1) este adevarata si urmatoarea conditie (C4) este adevarata :*

*(C4) Sistemul KPYS*$(\Sigma_\times, J_\times)$ *are o solutie stabilizanta* $(Z, V_\times, W_\times) = (Z, \begin{bmatrix} V_{\times 11} & O \\ V_{\times 21} & V_{\times 22} \end{bmatrix}, \begin{bmatrix} W_{\times 1} \\ W_{\times 2} \end{bmatrix})$, *cu* $Z \geq 0$.

*Mai mult, avem relatia intre solutiile KPYS:*

$$Z := Y(I - XY)^{-1}. \tag{276}$$

**(III)** *Presupunem ca (C1) si (C4) sunt adevarate (sau, echivalent, (C1)-(C3)). Atunci clasa tuturor solutiilor problemei $H^\infty$ suboptimale este $\mathbf{K} = \mathrm{LFT}(\mathbf{K}_g, \mathbf{Q})$, unde $\mathbf{Q} \in RH^\infty_{+,m_2\times p_2}$ este un parametru arbitrar cu $\|\mathbf{Q}\|_\infty < 1$ si*

$$\mathbf{K}_g = \left[\begin{array}{c|cc} A_g & B_{g1} & B_{g2} \\ \hline C_{g1} & D_{g11} & D_{g12} \\ C_{g2} & D_{g21} & D_{g22} \end{array}\right] = \left[\begin{array}{cc} \mathbf{K}_{g11} & \mathbf{K}_{g12} \\ \mathbf{K}_{g21} & \mathbf{K}_{g22} \end{array}\right], \qquad (277)$$

*are coeficientii scrisi in termenii sistemului original si ale solutiilor ARE.*

# Schita demonstratiei

- Necesitatea lui (C1) – se foloseste direct conditia de signatura
- Necesitatea lui (C2) – prin dualitate
- Daca (C1) are loc atunci (C4) este necesara pentru existenta solutiei $H^\infty$ – din nou din conditia de signatura.
- Daca (C1) are loc atunci (C4) este echivalenta cu (C2) impreuna cu (C3) – se foloseste o transformare de echivalenta pe fascicole matriciale – acest fapt demonstreaza si necesitatea lui (C3) si celelalte afirmatii cu exceptia clasei regulatoarelor

- Pentru clasa regulatoarelor consideram intai o problema mai simpla de tip 2 bloc – solutia ei se poate scrie direct pe baza solutiilor inegalitatii de signatura

- In final, in cazul general se foloseste ARE($\Sigma_c, J_c$) pentru a reduce problema la una mai simpla care este la randul ei o problema de tip 2 bloc duala.

# Capitolul 18: Stabilizare Robusta

In acest capitol aratam ca problema stabilizarii robuste pentru diverse tipuri de incertitudini (aditive, multiplicative, pe factori coprimi) este echivalenta cu o problema de tip $H^\infty$ optimala formulata pentru un sistem generalizat particular a carei solutie se poate obtine pe baza rezultatelor din capitolul precedent.

Fie sistemul $p \times m$ dat de $\mathbf{G}$, $y = \mathbf{G}u$, si fie $(\widetilde{\mathbf{N}}, \widetilde{\mathbf{M}})$ o factorizare corpima peste $RH_+^\infty$ astfel incat $\mathbf{G} = \widetilde{\mathbf{M}}^{-1}\widetilde{\mathbf{N}}$. Definim un regulator stabilizant pentru $\mathbf{G}$ ca fiind un sistem $\mathbf{K}$ pentru care sistemul in bucla inchisa format din $\mathbf{G}$ si $\mathbf{K}$ este intern stabil.

Pentru $\delta > 0$ introducem clasele de sisteme

$$\mathcal{D}_\delta^a := \{\mathbf{G}_\Delta^a : \mathbf{G}_\Delta^a = \mathbf{G} + \mathbf{\Delta}, \mathbf{\Delta} \in RH_{+,p\times m}^\infty \text{ si } \|\mathbf{\Delta}\|_\infty < \delta\}, \quad (278)$$

$$\mathcal{D}_\delta^m := \{\mathbf{G}_\Delta^m : \mathbf{G}_\Delta^m = (I + \mathbf{\Delta})\mathbf{G}, \mathbf{\Delta} \in RH_{+,p\times p}^\infty \text{ si } \|\mathbf{\Delta}\|_\infty < \delta\}, \quad (279)$$

$$\mathcal{D}_\delta^{cf} := \{\mathbf{G}_\Delta^{cf} : \mathbf{G}_\Delta^{cf} = (\widetilde{\mathbf{M}} + \mathbf{\Delta}_M)^{-1}(\widetilde{\mathbf{N}} + \mathbf{\Delta}_N),$$

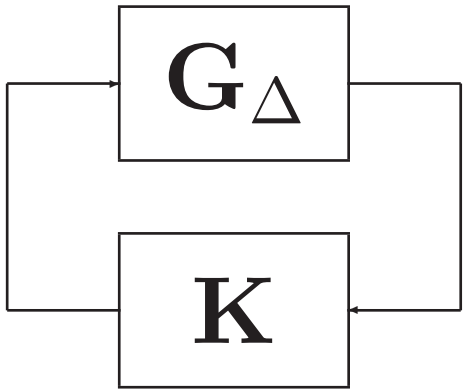
$$\mathbf{\Delta} = \left[\begin{array}{c|c} \mathbf{\Delta}_N & -\mathbf{\Delta}_M \end{array}\right] \in RH_{+,p\times(p+m)}^\infty \text{ si } \|\mathbf{\Delta}\|_\infty < \delta\} \quad (280)$$

Fiecare clasa $\mathcal{D}_\delta^a$, $\mathcal{D}_\delta^m$ si $\mathcal{D}_\delta^{cf}$ reprezinta o clasa de sisteme perturbate obtinute din $\mathbf{G}$ prin considerarea unor incertitudini aditive, multiplicative sau pe factori coprimi. Notam cu $\mathcal{D}_\delta$ oricare dintre aceste trei clase, si prin $\mathbf{G}_\Delta$ un element generic a lui $\mathcal{D}_\delta$.

Cel mai mare $\delta = \delta_{\max}$ pentru care exista un singur regulator stabilizator pentru toate sistemele $\mathcal{D}_{\delta_{\max}}$ se numeste marginea maxima

de stabilitate pentru clasa respectiva de incertitudini.

Problema de Stabilizare Robusta (suboptimala): Dandu-se un sistem nominal $\mathbf{G}$, o margine $\delta$ si clasa sistemelor $\mathcal{D}_\delta$, unde $\delta < \delta_{\max}$, gasiti un regulator (sau toata clasa) $\mathbf{K}$, $u = \mathbf{K}y$, care stabilizeaza toate sistemele $\mathbf{G}_\Delta \in \mathcal{D}_\delta$.




Regulatorul $\mathbf{K}$ se numeste robust. Problema de stabilizare robusta optimala consta in constructia unui regulator $\mathbf{K}$ (sau toata clasa) pentru toate sistemele in $\mathcal{D}_{\delta_{\max}}$.

Aratam in continuare cum o problema de stabilizare robusta se poate converti intr-o problema echivalenta de tip $H^\infty$ de tipul celor studiate in capitolul precedent.

Fiecare clasa de sisteme incerte $\mathcal{D}_\delta$ se poate reprezenta ca o transformare liniar fractionara superioara (ULFT) a unui sistem generalizat si a unui sistem perturbator $\mathbf{\Delta}$:

$$\mathbf{T}^a := \left[\begin{array}{c|c} \mathbf{T}^a_{11} & \mathbf{T}^a_{12} \\ \hline \mathbf{T}^a_{21} & \mathbf{T}^a_{22} \end{array}\right] = \left[\begin{array}{c|c} O & I_m \\ \hline I_p & \mathbf{G} \end{array}\right], \tag{281}$$

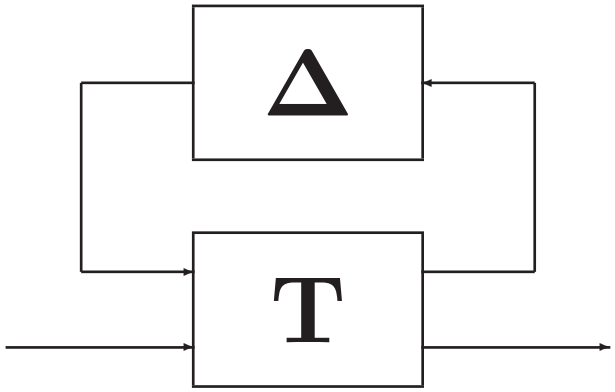
$$\mathbf{T}^m := \left[\begin{array}{c|c} \mathbf{T}^m_{11} & \mathbf{T}^m_{12} \\ \hline \mathbf{T}^m_{21} & \mathbf{T}^m_{22} \end{array}\right] = \left[\begin{array}{c|c} O & \mathbf{G} \\ \hline I_p & \mathbf{G} \end{array}\right], \tag{282}$$

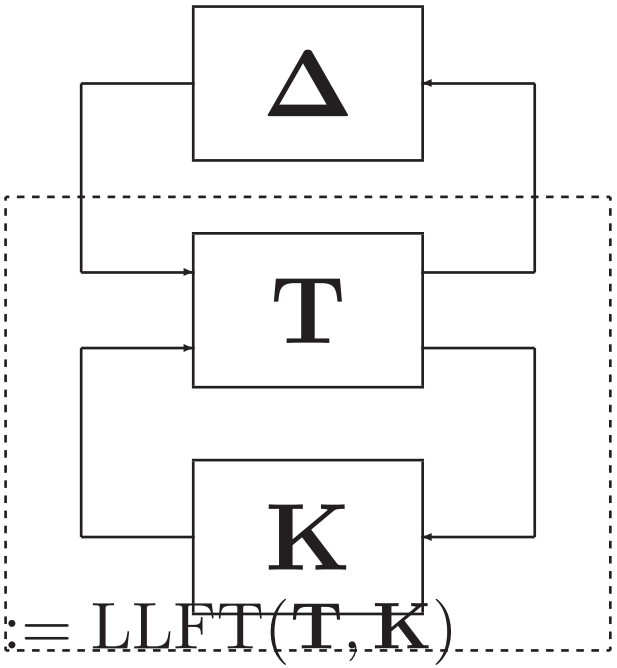
$$\mathbf{T}^{cf} = \left[\begin{array}{c|c} \mathbf{T}^{cf}_{11} & \mathbf{T}^{cf}_{12} \\ \hline \mathbf{T}^{cf}_{21} & \mathbf{T}^{cf}_{22} \end{array}\right] = \left[\begin{array}{c|c} O & I_m \\ \widetilde{\mathbf{M}}^{-1} & \mathbf{G} \\ \hline \widetilde{\mathbf{M}}^{-1} & \mathbf{G} \end{array}\right]. \tag{283}$$

Un calcul de rutina arata ca

$$\begin{aligned}
\mathrm{ULFT}(\mathbf{T}^a, \mathbf{\Delta}) &= \mathbf{G} + \mathbf{\Delta} = \mathbf{G}^a_\Delta, \\
\mathrm{ULFT}(\mathbf{T}^m, \mathbf{\Delta}) &= (I + \mathbf{\Delta})\mathbf{G} = \mathbf{G}^m_\Delta, \\
\mathrm{ULFT}(\mathbf{T}^{cf}, \mathbf{\Delta}) &= (\widetilde{\mathbf{M}} + \mathbf{\Delta}_M)^{-1}(\mathbf{N} + \mathbf{\Delta}_N) = \mathbf{G}^{cf}_\Delta, \quad \mathbf{\Delta} = \left[ \begin{array}{c|c} \mathbf{\Delta}_N & - \end{array} \right.
\end{aligned}$$

Acesta ne arata ca in oricare caz clasa sistemelor incerte se poate reprezenta sub forma unei $\mathrm{ULFT}(\mathbf{T}, \mathbf{\Delta})$ asa cum este descris in figurile :






Deci obtinem configuratiile echivalente de mai sus in care atat $\mathbf{T}_{y_1u_1}$ cat si $\boldsymbol{\Delta}$ sunt stabile si $\mathbf{T}_{y_1u_1} = \mathrm{LLFT}(\mathbf{T}, \mathbf{K})$.

Intorcandu-ne la problema originala de convertire a problemei de stabilizare robusta intr-una de tip $H^\infty$ obtinem urmatorul rezultat.

**Teorema 183.** *Un regulator* $\mathbf{K}$ *este o solutie a problemei de stabilizare robusta in raport cu orice clasa de sisteme* $\mathcal{D}_\delta$, $\delta \leq \delta_{\max}$,

*daca si numai daca* $\mathbf{K}$ *este o solutie a problemei* $H^\infty$ *suboptimale cu* $\gamma = \frac{1}{\delta}$ *pentru* $\mathbf{T}$.

Prin urmare putem obtine o expresie a marginii de stabilitate si o solutie pentru problema stabilizarii robuste evaluand intai $\gamma_{\min} = \frac{1}{\delta_{\max}}$ si construind ulterior regulatorul optimal pentru problema $H^\infty$ optimala corespunzatoare datelor respective.

Observatie: Problema optimala se poate rezolva in acest caz pornind de la una suboptimala si facand o analiza de perturbatii singulare.