Bor Plestenjak

Numerične metode za linearne kontrolne sisteme

skripta

verzija: 21. april 2006

Kazalo

1 Uvod **5**
- 1.1 Kontrolni sistemi . . . 5
- 1.2 Lastnosti kontrolnih sistemov . . . 7
- 1.3 Ponovitev Laplaceove transformacije . . . 9
- 1.4 Prenosna funkcija . . . 11
- 1.5 Računanje impulznega (stopničnega) odziva z nastavki . . . 11
- 1.6 Diskretni sistem . . . 12
- 1.7 Bločni diagrami . . . 12

2 Predstavitev v prostoru stanj **13**
- 2.1 Uvod . . . 13
- 2.2 Odziv sistema . . . 14
- 2.3 Povezava s klasično teorijo . . . 16
- 2.4 Diskretni sistemi . . . 18
- 2.5 Numerično računanje matrične eksponentne funkcije . . . 19
  - 2.5.1 Občutljivost matrične ekponentne funkcije . . . 19
- 2.6 Računanje frekvenčnega odziva . . . 21

3 Vodljivost in spoznavnost **22**
- 3.1 Uvod . . . 22
- 3.2 Vodljivost . . . 22
  - 3.2.1 Diskretni sistemi . . . 27
- 3.3 Spoznavnost . . . 28
- 3.4 Kanonične oblike . . . 30
  - 3.4.1 Vodljivostna normalna oblika . . . 30
  - 3.4.2 Luenbergerjeva vodljivostna kanonična oblika . . . 31
  - 3.4.3 Spoznavnostna normalna oblika . . . 33
- 3.5 Vodljivostna Hessenbergova oblika . . . 34
- 3.6 Razporejanje polov . . . 36
  - 3.6.1 Ackermanova formula . . . 37

4 Stabilnost **38**
- 4.1 Uvod . . . 38
- 4.2 Stabilnost po Ljapunovu . . . 39
- 4.3 Diskretni sistemi . . . 43
- 4.4 Klasična teorija . . . 44
- 4.5 Oddaljenost od nestabilnih sistemov . . . 45
  - 4.5.1 Robustna stabilnost . . . 47

5 Numerično reševanje Sylvestrove enačbe in enačbe Ljapunova **49**
5.1 Kroneckerjev produkt . . . 49
5.2 Občutljivost Sylvestrove enačbe . . . 51
5.3 Algoritmi za Sylvestrovo enačbo in enačbo Ljapunova . . . 57
5.3.1 Bartels-Stewartov algoritem za Sylvestrovo enačbo . . . 58
5.3.2 Bartels-Stewartov algoritem za enačbo Ljapunova . . . 59
5.3.3 Reševanje Sylvestrove enačbe preko Hessenberg-Schurove oblike . . . 60
5.3.4 Reševanje diskretne enačbe Ljapunova s simetričnim C . . . 62
5.3.5 Hammarlingov algoritem . . . 62
5.3.6 Hammarlingov algoritem za diskretno enačbo Ljapunova . . . 65

6 Realizacija in identifikacija **67**
6.1 Uvod . . . 67
6.2 Realizacija SISO sistema iz impulznega odziva . . . 67
6.3 Realizacija MIMO sistema v prostoru stanj . . . 71
6.3.1 Vodljiva realizacija . . . 72
6.3.2 Spoznavna realizacija . . . 73
6.3.3 Minimalna realizacija . . . 73
6.4 Identifikacija iz vhodno-izhodnih parov (SISO primer) . . . 76
6.5 Identifikacija iz vhodno-izhodnih parov (MIMO primer) . . . 79

7 Stabilizacija in razporejanje polov **84**
7.1 Uvod . . . 84
7.2 Stabilizacija s povratno zvezo iz stanja . . . 85
7.2.1 Stabilizacija preko vodljivostne Gramove matrike . . . 85
7.2.2 Stabilizacija preko enačbe Ljapunova . . . 87
7.3 Razporejanje polov . . . 88
7.3.1 Zveza med poli in prehodnim obnašanjem sistema . . . 89
7.4 Razporejanje polov enovhodnih sistemov . . . 89
7.4.1 Razporejanje polov preko Hessenbergove forme . . . 90
7.4.2 Metoda ortogonalnih transformacij na lastnih vektorjih . . . 90
7.4.3 Modifikacija QR algoritma . . . 92
7.5 Razporejanje polov večvhodnih sistemov . . . 93
7.5.1 Razporejanje polov preko Hessenbergove forme . . . 93
7.5.2 Metoda ortogonalnih transformacij na lastnih vektorjih . . . 94
7.5.3 Razporejanje polov preko Schurove forme . . . 94
7.6 Pogojenost polov zaprtozančnega sistema . . . 96
7.7 Robustno razporejanje polov . . . 99
7.8 Optimalno vodenje . . . 102
7.8.1 Diskretni sistemi . . . 109

8 Numerično reševanje Riccatijeve enačbe **112**
8.1 Uvod . . . 112
8.2 Občutljivost Riccatijeve enačbe . . . 112
8.3 Newtonova metoda . . . 114
8.4 Uporaba matričnega predznaka . . . 116
8.5 Metoda lastnih vektorjev . . . 119
8.6 Uporaba Schurove forme . . . 120
8.6.1 Analiza zaokrožitvenih napak . . . 121

8.6.2 Schurova metoda za DARE 122
8.7 Posplošeni problem lastnih vrednosti in DARE 122

Poglavje 1

Uvod

1.1 Kontrolni sistemi

Kontrolni sistemi nastopajo na najrazličnejših področjih. Skupno vsem je, da imamo dinamični sistem, sestavljen iz različnih komponent, ki vplivajo druga na drugo. Primeri takšnih sistemov so npr. električni motor, letalo, človeško telo, in podobno. Velja:

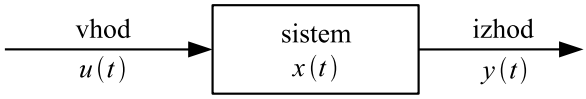
a) komponente sistema so povezane in medsebojno odvisne,

b) meje sistema ločijo notranje komponente od zunanjih.

Lastnost b) pomeni, da lahko sistem obravnavamo kot neko končno zaključeno celoto. Stanje takega sistema opisujejo notranje spremenljivke, ki jih imenujemo *spremenljivke stanja*. To še ne pomeni, da na sistem ne morejo vplivati zunanji dejavniki oz. *vhodi*. Ravno to, kako z vhodi od zunaj kontrolirati sistem, ki se sicer obnaša po nekih svojih zakonitostih, obravnavamo pri *teoriji kontrolnih sistemov*. Cilj je vplivati na sistem tako, da se bo njegovo obnašanje čim bolj ujemalo z zastavljenimi cilji. Npr.:

- Če si kot sistem predstavljamo sobo s klimatsko napravo, lahko s prižiganjem in ugašanjem naprave dosežemo, da bo temperatura v sobi čim bližje željeni.
- Kot sistem si lahko predstavljamo vse semaforje v mestu. Z ustreznim prižiganjem in ugašanjem luči lahko dosežemo, da bo promet čim bolj tekoč.
- Na ekonomsko situacijo v državi lahko vplivamo npr. z višino davkov in drugimi parametri.
- Na dlani držimo metlo in se trudimo, da bi stala pokonci. Tudi to je primer kontrolnega sistema.

Predstavljamo si lahko, da je dinamični sistem sestavljen iz prostora možnih stanj in pravil, ki na podlagi prejšnjih stanj in vhodov določajo trenutno stanje. V praksi ponavadi ne poznamo vrednosti vseh spremenljivk stanja, saj jih je ponavadi preveč, da bi lahko spremljali vse hkrati. Tako je npr. v ekonomiji inflacija odvisna od mnogih parametrov, nekatere poznamo, večino pa ne. Spremljamo le podmnožico oz. kombinacijo stanj, ki ji pravimo *izhod* oz. *odziv sistema*. Kar smo opisali, je sistem v t.i. *vhodno-izhodni obliki*. V tej obliki ga lahko predstavimo v obliki naslednjega diagrama:



Pri različnih aplikacijah je cilj preko vhoda regulirati sistem tako, da se obnaša po naših željah. To dosežemo s pomočjo *regulatorja* oz. *krmilnika*, ki vodi sistem tako, da generira ustrezni vhod. Povezava regulatorja in sistema je naš *kontrolni sistem*. Pomemben del teorije kontrolnih sistemov se ukvarja s konstrukcijo regulatorja, ki bo izpolnjeval te zahteve.

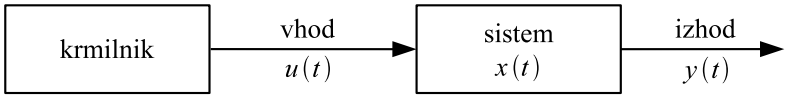
Označili bomo, da sistem reguliramo z vhodom $u(t)$, izhod iz sistema, ki je odvisen od vhoda, pa je $y(t)$. Notranje spremenljivke, ki opisujejo stanje sistema, naj bodo $x(t)$. Pravimo tudi, da sistem *vzbujamo* z vhodom $u(t)$, *odziv* sistema pa je izhod $y(t)$. Pri tem je cilj vhod določiti tako, da bo izhod čim bolj zadoščal izbranim kriterijem. Na nekaterih področjih (npr. v elektrotehniki) govorimo o signalih in sta tako $u(t)$ in $y(t)$ *vhodni* oz. *izhodni signal*.

Vodenje sistema ponavadi poteka avtomatično preko *regulatorja* ali *krmilnika*, ki proizvaja vhod $u(t)$. Pri tem ločimo sisteme na dve vrsti.

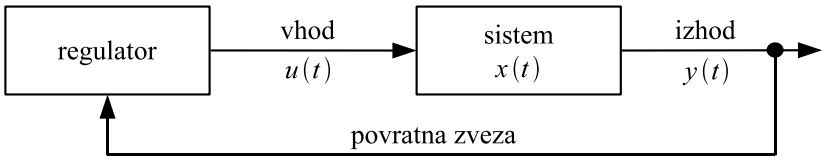
Preprostejša oblika so *odprtozančni sistemi*, kjer delovanje krmilnika ni odvisno od izhoda sistema. Npr.:

- Luči na semaforjih prižigamo in ugašamo v vnaprej predpisanih časovnih intervalih, neodvisno od prometne situacije.

Shemo odprtozančnega kontrolnega sistema predstavlja naslednja slika:



Kompleksnejša oblika so *zaprtozančni sistemi*, kjer imamo *povratno zanko* med izhodom in regulatorjem. Shema zaprtozančnega sistema je predstavljena na naslednji sliki:



Primeri zaprtozančnih sistemov so npr.:

- Sistem, ki odvisno od prometne situacije krmili semaforje v mestu s ciljem preprečevanja zastojev.
- Avtomatska klimatska naprava, kjer sistem glede na temperaturo sobe, ki je v bistvu izhod sistema, samodejno vklaplja in izklaplja napravo.

Za modeliranje dinamičnega sistema ponavadi uporabimo končni sistem diferencialnih enačb v obliki:

$$\begin{aligned} \dot{x}(t) &= f(t, x(t), u(t)), \qquad x(t_0) = x_0 \\ y(t) &= g(t, x(t), u(t)), \end{aligned}$$

kjer so

$$\begin{aligned} x(t) &= [x_1(t) \ \cdots \ x_n(t)]^T : \text{notranje spremenljivke stanja sistema,} \\ u(t) &= [u_1(t) \ \cdots \ u_m(t)]^T : \text{vhod,} \\ y(t) &= [y_1(t) \ \cdots \ y_r(t)]^T : \text{izhod.} \end{aligned}$$

Ponavadi je $r \leq n$ in $m \leq n$. Tako imamo preslikavi $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ in $g : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^r$. Ker je težko predvideti vse spremenljivke stanja sistema oziroma potem to tudi modelirati, na sistem lahko vplivajo še zunanji nepredvidljivi dejavniki. To so npr. močni sunki vetra med pristajanjem letala, zlom na borzi v ekonomskem sistemu, in podobno.

Opisali smo zvezni model. Drug pogost model so diferenčne enačbe v diskretnem primeru

$$\begin{aligned} x(k+1) &= f(k, x(k), u(k)), \qquad x(0) = x_0 \\ y(k) &= g(k, x(k), u(k)). \end{aligned}$$

Ponavadi diskretne vrednosti predstavljajo vzorce zveznega modela v izbranih trenutnih. Pri diskretnih modelih poznamo ponavadi še interval vzorčenja Δt, potem pa so $u(k)$, $y(k)$ in $x(t)$ približki za $u(k\Delta t)$, $y(k\Delta t)$ in $x(k\Delta t)$.

1.2 Lastnosti kontrolnih sistemov

Po Kalmanu lahko dinamični sistem formalno opišemo tako, da imamo dan prostor stanj $\mathcal{X}$ in končen časovni interval $\mathcal{T}$. Poleg tega imamo podan še prostor $\mathcal{U}$ vseh funkcij definiranih na $\mathcal{T}$, ki predstavljajo vse možne vhode v sistem.

Za vsak začetni trenutek $t_0 \in \mathcal{T}$, začetno stanje $x_0 \in \mathcal{X}$ in vhod $u \in \mathcal{U}$, definiran za $t \geq t_0$, $t \in \mathcal{T}$, so prihodnja stanja sistema določena s *prenosno preslikavo*

$$\Phi : \mathcal{T} \times \mathcal{X} \times \mathcal{U} \to \mathcal{X},$$

ki jo zapišemo kot

$$\Phi(t_1; t_0, x(t_0), u(t)) = x(t_1).$$

Za prenosno funkcijo veljajo naslednje lastnosti:

1. Prenosna preslikava je lahko identiteta: $\Phi(t_0; t_0, x(t_0), u(t)) = x(t_0)$.
2. Lastnost polgrupe:

   $$\Phi(t_2; t_0, x(t_0), u(t)) = \Phi(t_2; t_1, \Phi(t_1; t_0, x(t_0), u(t)), u(t))).$$

   To pomeni, da dobimo isto, če gremo iz t_0 direktno do t_2 ali pa če gremo najprej do t_1, potem pa od tam naprej do t_2.
3. *Vzorčnost*: če za za poljuben $t_0 \in \mathcal{T}$ za vse $t \geq t_0$, $t \in \mathcal{T}$, velja

   $$\Phi(t_1; t_0, x(t_0), u_1(t)) = \Phi(t_1; t_0, x(t_0), u_2(t)),$$

   potem velja $u_1(t) = u_2(t)$ za $t \geq t_0$, $t \in \mathcal{T}$.

   Vzorčnost v bistvu pomeni neke vrste injektivnost, saj za različne vhode dobimo različne izhode.

4. Izhod ima obliko preslikave
$$h : \mathcal{T} \times \mathcal{X} \times \mathcal{U} \to \mathcal{Y},$$
kjer je $\mathcal{Y}$ prostor izhodnih funkcij.

5. Φ in h sta zvezni funkciji.

Definicija 1.1 *Če za poljubna vhoda u_1, u_2 in skalarja c_1, c_2 velja*

$$\Phi(t_1; t_0, x(t_0), c_1u_1(t) + c_2u_2(t)) = c_1\Phi(t_1; t_0, x(t_0), u_1(t)) + c_2\Phi(t_1; t_0, x(t_0), u_2(t)),$$

za vse $t \geq t_0$ $t \in \mathcal{T}$, potem je sistem linearen.

Definicija 1.2 *Če za poljubna $t_0, t_1 \in \mathcal{T}$ in premik τ velja*

$$\Phi(t_1; t_0, x(t_0), u(t)) = \Phi(t_1 + \tau; t_0 + \tau, x(t_0), u(t)),$$

potem je sistem časovno nespremenljiv *oz.* časovno invarianten.

Zgled 1.1 Na treh preprostih zgledih lahko demonstriramo pojme vzročnosti, linearnosti in časovne nespremenljivosti.

a) $x(t) = u^2(t-1)$: vzročen, nelinearen, časovno nespremenljiv;

b) $x(t) = u(-t)$: ni vzročen, linearen, časovno nespremenljiv;

c) $x(t) = 3^t u(t-1)$: vzročen, linearen in časovno spremenljiv.

Mi se bomo ukvarjali z vzročnimi linearnimi časovno nespremenljivimi sistemi. Za takšne sisteme veljajo še dodatne uporabne lastnosti:

6. Če je sistem linearen, potem je odziv na ničelni vhod ničeln, oz.
$$\Phi(t; t_0, x(t_0), 0) = 0.$$
To sledi iz linearnosti $\Phi(t; t_0, x(t_0), \alpha u(t)) = \alpha\Phi(t; t_0, x(t_0), u(t))$, če vzamemo $\alpha = 0$.

7. Če je sistem vzročen, potem odziv ni odvisen od prihajajočih stanj in vhodov.
Če to ne bi bilo res, potem bi lahko poiskali dva vhoda $u_1(t)$ in $u_2(t)$, ki bi se na $\mathcal{T}$ ujemala, kasneje pa ne več, izhod pa bi bil na intervalu $\mathcal{T}$ različen.

8. Če je sistem linearen, potem je vzročnost ekvivalentna t.i. *pogoju začetnega mirovanja*:
$$\Phi(t; t_0, x(t_0), u(t)) = 0 \text{ za } t \leq t_1$$
velja natanko tedaj, ko velja $u(t) = 0$ za $t \leq t_1$.
Če Φ ne zadošča pogojem začetnega mirovanja, potem obstaja tak vhod $\widetilde{u}(t) \not\equiv 0$, da je $\Phi(t; t_0, x_0(t_0), \widetilde{u}(t)) = 0$. Zaradi tega nimamo vzročnosti, saj bo odziv enak za u in $u + \widetilde{u}$.
Za dokaz v drugo smer predpostavimo, da obstajata vhoda $u_1(t)$ in $u_2(t)$, ki imata enak odziv, ni pa $u_1(t) \equiv u_2(t)$. Zaradi tega je odziv na $u_1(t) - u_2(t)$ enak 0, to pa je v protislovju s pogojem začetnega mirovanja.

9. Če je sistem časovno nespremenljiv, potem je odziv na periodični vhod spet periodočen in to z isto periodo.

   Denimo, da je perioda τ oz. $u(t+\tau) = u(t)$. Potem zaradi časovne nespremenljivosti velja

   $$y(t) = \Phi(t; t_0, x_0, u(t)) = \Phi(t+\tau; t_0+\tau, x_0, u(t)) = \Phi(t+\tau; t_0+\tau, x_0, u(t+\tau)) = y(t+\tau).$$

Zgled 1.2 Dva opisa dinamičnega sistema, ki ju bomo uporabili v nadaljevanju za zvezne vzročne linearne časovno nespremenljive sisteme, sta

a) diferencialna enačba n-tega reda

$$y^{(n)}(t) + k_1 y^{(n-1)}(t) + \cdots + k_{n-1} y'(t) + k_n y_n(t) = \beta_0 u^{(m)}(t) + \beta_1 u^{(m-1)} + \cdots + \beta_m u(t),$$

b) opis z matrikami v prostoru stanj

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \qquad x(t_0) = x_0, \quad t \geq t_0, \\ y(t) &= Cx(t) + Du(t), \end{aligned}$$

kjer so $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{r\times n}$ $D \in \mathbb{R}^{r\times m}$, $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$ in $y(t) \in \mathbb{R}^r$. Ponavadi velja $m \leq n$ in $r \leq n$.

1.3 Ponovitev Laplaceove transformacije

Naj bo $f : [0, \infty) \to \mathbb{R}$ odsekoma zvezna in

$$\int_0^\infty |f(t)e^{-\sigma_0 t}|dt < \infty$$

za nek končen $\sigma_0 \in \mathbb{R}$. Potem za $\sigma_0 \leq \sigma$ velja $\int_0^\infty |f(t)e^{-\sigma t}|dt < \infty$ in lahko definiramo Laplaceovo transformacijo f kot

$$F(s) = \mathcal{L}(f(t)) = \int_0^\infty f(t)e^{-st}dt,$$

kjer je $s = \sigma + i\omega$ in $\sigma_0 \leq \sigma$.

Za meje moramo v bistvu privzeti, da so od $t = 0_-$ do $t = \infty$, da pokrijemo tudi primer impulzne funkcije, ki ima impulz pri $t = 0$.

Na kratko ponovimo glavne lastnosti Laplaceove transformacije:

- Linearnost: $\mathcal{L}(\alpha_1 f_1(t) + \alpha_2 f_2(t)) = \alpha_1 \mathcal{L}(f_1(t)) + \alpha_2 \mathcal{L}(f_2(t))$.
- Transformiranka odvoda:

$$\mathcal{L}\left[\frac{df(t)}{dt}\right] = sF(s) - \lim_{t\to 0} f(t) = sF(s) - f(0),$$

$$\begin{aligned} \mathcal{L}\left[\frac{d^n f(t)}{dt^n}\right] &= s^n F(s) - \lim_{t\to 0}\left(s^{n-1}f(t) + s^{n-2}f'(t) + \cdots + f^{n-1}(t)\right) \\ &= s^n F(s) - s^{n-1}f(0) - s^{n-2}f'(0) + \cdots - f^{(n-1)}(0). \end{aligned}$$

- Transformiranka integrala:
$$\mathcal{L}\left[\int_0^t f(\tau)d\tau)\right] = \frac{F(s)}{s},$$
$$\mathcal{L}\left[\int_0^{t_1}\int_0^{t_2}\cdots\int_0^{t_n} f(\tau)dt_1\ldots dt_{n-1}\right] = \frac{F(s)}{s},$$

- Časovni premik: $\mathcal{L}\left[f(t-T)u_s(t-T)\right] = e^{-Ts}F(s)$.

- Frekvenčni premik: $\mathcal{L}\left[f(t)e^{-\alpha t}\right] = F(s-\alpha)$.

- Izrek o začetni vrednosti: Če časovna limita obstaja, velja
$$\lim_{t\to 0} f(t) = \lim_{s\to\infty} sF(s).$$

- Izrek o končni vrednosti: Če je $sF(s)$ analitična na območju $\{s : \mathrm{Re}(s) \geq 0\}$, oziroma nima polov, katerih realni del je nenegativen, potem velja
$$\lim_{t\to\infty} f(t) = \lim_{s\to\infty} sF(s).$$

- Konvolucija: $f_1(t) = f_2(t) = 0$ za $t < 0$.
$$\begin{aligned} F_1(s)F_2(s) &= \mathcal{L}\left[\int_0^t f_1(\tau)f_2(t-\tau)d\tau)\right] = \mathcal{L}\left[\int_0^t f_2(\tau)f_1(t-\tau)d\tau)\right] \\ &= \mathcal{L}\left[f_1(t) * f_2(t)\right]. \end{aligned}$$
Opomba: $F_1(s)$ obstaja za $\mathrm{Re}(s) \geq \sigma_1$, $F_2(s)$ pa za $\mathrm{Re}(s)$. Zgornjo konvolucijo potem vzamemo na preseku obeh območij.
V drugo smer velja:
$$\mathcal{L}(f_1(t)f_2(t)) = F_1(s) * F_2(s) = \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} F(\rho)F_2(s-\rho)d\rho.$$

- Inverzna Laplaceova transformacija:
$$f(t) = \mathcal{L}^{-1}(F(s)) = \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} F(s)e^{st}ds,$$
kjer je σ večji od vseh realnih komponent polov $F(s)$.

V teoriji kontrolnih sistemov sta zelo pomembni naslednji vhodni funkciji:

a) *enotska stopnica* u_s, definirana z
$$u_s(t) = \begin{cases} 1 & \text{za } t \geq 0, \\ 0 & \text{za } t < 0. \end{cases}$$

b) *enotski impulz* δ, definiran z $\delta(t) = 0$ za $t \neq 0$ in $\int_{-\infty}^{\infty} \delta(t)dt = 1$. To je t.i. *Diracova delta funkcija*. Lahko si jo predstavljamo kot limito ustreznih funkcij z nepraznim nosilcem, možnosti je več, npr. $\delta = \lim_{\epsilon \to 0} \delta_\epsilon$ preko odsekoma konstantne funkcije

$$\delta_\epsilon(t) = \begin{cases} \frac{1}{2\epsilon}, & |x| \leq \epsilon \\ 0 & \text{sicer} \end{cases}$$

ali pa preko normalne distribucije

$$\delta_\epsilon(t) = \frac{1}{\epsilon\sqrt{\pi}} e^{-x^2/\epsilon^2}.$$

Enotski impulz si lahko predstavljamo tudi kot odvod enotske stopnice. Argument proti je, da strogo matematično gledano odvod enotske stopnice pri $t = 0$ ne obstaja, argument za pa je, da se Laplaceovi transformiranki enotske stopnice in enotskega impulza obnašata tako, kot da gre za transformiranki funkcije in njenega odvoda.

Laplaceova transformiranka enotske stopnice je $1/s$. Res,

$$\mathcal{L}(u_s(t)) = \int_0^\infty u_s(t)e^{-st}dt = \int_0^\infty e^{-st}dt = -\frac{1}{s}e^{-st}\Big|_0^\infty = \frac{1}{s}.$$

To velja za $\mathrm{Re}(s) > 0$, saj je v tem primeru

$$\int_0^\infty |u_s(t)e^{-\sigma t}|dt = \int_0^\infty |e^{-\sigma t}|dt < \infty.$$

Laplaceovo transformiranko za enotski impulz bomo dobili kot limito Laplaceovih transformacij za δ_ϵ, pri čemer bomo tudi spodnjo mejo integrala v limiti poslali proti 0 s spodnje strani.

$$\begin{aligned} \mathcal{L}(\delta(t) &= \lim_{\epsilon \to 0} \int_{-\epsilon}^{\epsilon} \delta_\epsilon(t)e^{-st}dt = \lim_{\epsilon \to 0} \int_{-\epsilon}^{\epsilon} \frac{1}{2\epsilon}e^{-st}st \\ &= \lim_{\epsilon \to 0} \frac{1}{2\epsilon}\left(\frac{-1}{s}\right)e^{-st}\Big|_{-\epsilon}^{\epsilon} = \lim_{\epsilon \to 0} \frac{1}{2\epsilon s}\left(e^{\epsilon s} - e^{-\epsilon s}\right) = 1. \end{aligned}$$

To velja za $s > 0$.

Tukaj manjka še: primeri uporabe Laplaceove transformacije.

1.4 Prenosna funkcija

Tukaj manjka še: poglavje o prenosni funkciji (8 strani priprav).

1.5 Računanje impulznega (stopničnega) odziva z nastavki

Tukaj manjka še: poglavje o računanju impulznega (stopničnega) odziva z nastavki (5 strani priprav).

1.6 Diskretni sistem

Tukaj manjka še: poglavje o diskretnem sistemu (Z-transformacija) (8 strani priprav).

1.7 Bločni diagrami

Tukaj manjka še: poglavje o bločnih diagramih (3 strani priprav).

Poglavje 2

Predstavitev v prostoru stanj

2.1 Uvod

Medtem, ko klasična teorija kontrolnih sistemov temelji na prenosni funkciji, je osnova moderne teorije obravnava v prostoru stanj. Prednosti so naslednje:

- na soroden način lahko obravnavamo probleme ene ali več spremenljivk, časovno nespremenljive in časovno spremenljive sisteme, linearne in nelinearne sisteme;
- prenosne funkcije so le za linearne časovno nespremenljive sisteme z enim vhodom in enim izhodom;
- pri povratni zvezi preko stanja imamo na voljo več parametrov s katerimi lahko nastavimo obnašanje sistema in ga stabiliziramo;
- pri prenosni zvezi imamo pri povratni zvezi na voljo le en parameter, to je ojačanje.

V prostoru stanj linearni zvezno nespremenljivi kontrolni sistem zapišemo v obliki

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad x(t_0) = x_0, \quad t \geq t_0, \tag{2.1}$$

$$y(t) = Cx(t) + Du(t), \tag{2.2}$$

kjer je $x(t) \in \mathbb{R}^n$ *vektor stanja*, $u(t) \in \mathbb{R}^m$ *vhodni signal* in $y(t) \in \mathbb{R}^r$ *izhodni signal*. Pri matrikah je $A \in \mathbb{R}^{n\times n}$ *matrika stanja*, $B \in \mathbb{R}^{n\times m}$ *vhodna matrika*, $C \in \mathbb{R}^{r\times n}$ *izhodna matrika* in $D \in \mathbb{R}^{r\times m}$ *matrika direktnega prenosa*, ki je ponavadi kar enaka 0. Matrike A, B, C in D lahko sestavimo v bločno matriko

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}.$$

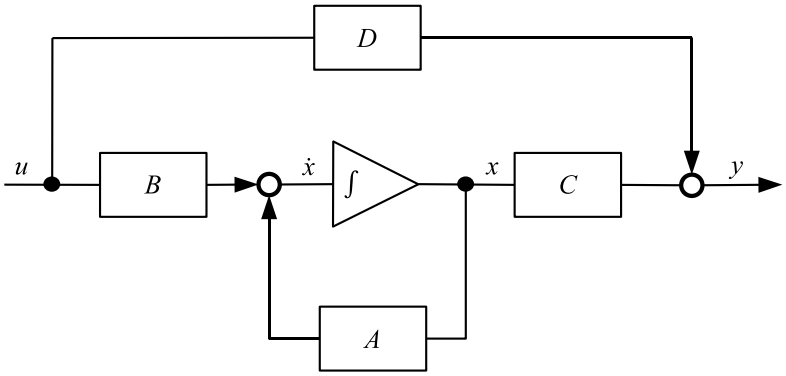
Običajno velja $m \leq n$ in $r \leq n$.

Enačba (2.1) je *enačba stanja*, (2.2) pa je *izhodna enačba*.

Z začetnim stanjem $x(t_0) = x_0$ in vhodom u na časovnem intervalu (t_0, t) je določen izhod za $t \geq t_0$. Če je $u(t) \equiv 0$, imamo *nevsiljen sistem*.

Če je $m = 1$ imamo *enovhodni sistem* in lahko pišemo kar $B = b \in \mathbb{R}^n$. Podobno imamo v primeru $r = 1$ *enoizhodni sistem* in lahko pišemo $C = c^T$ za $c \in \mathbb{R}^n$. Če je $m > 1$ imamo *večvhodni sistem*,

pri $r > 1$ pa *večizhodni sistem*. V primeru $m = 1$ in $r = 1$, ko je sistem enovhoden in enoizhoden, imamo *univariantni sistem* oz. *SISO sistem* (single-input single-output), v primeru $m > 1$ in $r > 1$ pa imamo *multivariantni sistem* oz. *MIMO sistem* (multiple-input multiple-output).



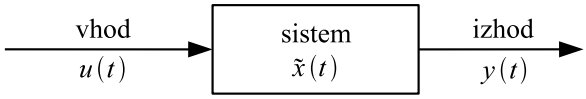
Opis sistema ni enoličen. Isti sistem lahko opišemo z različnimi modeli v prostoru stanj, odvisno od izbire vhodnih, izhodnih in notranjih spremenljivk. Če npr. z nesingularno transformacijo S spremenimo spremenljivke stanja v $x(t) = S\widetilde{x}(t)$, potem dobimo nov model

$$\begin{aligned} \dot{\widetilde{x}}(t) &= \widetilde{A}\widetilde{x}(t) + \widetilde{B}u(t), \qquad \widetilde{x}(t_0) = \widetilde{x}_0, \quad t \geq t_0, \\ y(t) &= \widetilde{C}\widetilde{x}(t) + Du(t), \end{aligned}$$

kjer je $\widetilde{A} = S^{-1}AS$, $\widetilde{B} = S^{-1}B$ in $\widetilde{C} = CS$. To lahko zapišemo kot

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \xrightarrow{S} \begin{bmatrix} S^{-1}AS & S^{-1}B \\ CS & D \end{bmatrix}.$$

Pri transformaciji se spremenijo le spremenljivke, ki predstavljajo stanje, vhod in izhod pa ostaneta nespremenjena.



Tukaj manjka še: zgled inverzno nihalo (3 strani priprav).

2.2 Odziv sistema

Brez škode za splošnost lahko predpostavimo, da je $t_0 = 0$. Če imamo nevsiljeni sistem, kjer je $u(t) \equiv 0$, imamo homogeno diferencialno enačbo

$$\dot{x}(t) = Ax(t), \qquad x(0) = x_0, \quad t \geq 0.$$

Vemo, da se rešitev izraža v obliki $x(t) = e^{At}x_0$, kjer je e^{At} matrična eksponentna funkcija, definirana s konvergentnim razvojem

$$e^{At} = \sum_{k=0}^{\infty} \frac{1}{k!}(At)^k = I_n + At + \frac{1}{2}(At)^2 + \cdots.$$

Osnovne lastnosti matrične eksponentne funkcije so:

1. $e^{A(t+s)} = e^{At}e^{As}$,
2. e^{At} je vedno nesingularna,
3. $(e^{At})^{-1} = e^{-At}$,
4. $\frac{d}{dt}(e^{At}) = Ae^{At} = e^{At}A$,
5. $e^{P^{-1}APt} = P^{-1}e^{At}P$ za vsako nesingularno matriko P,
6. $e^{(A+B)t} = e^{At} \cdot e^{Bt}$ natanko tedaj, ko A in B komutirata.

Tako dobimo t.i. *odziv na ničelni vhod* (zero-input response)

$$x(t) = e^{At}x_0 \Longrightarrow y(t) = Ce^{At}x_0 =: y_{\text{zi}}(t). \tag{2.3}$$

Matriko e^{At} imenujemo tudi *zvezna prehodna matrika stanja*, saj velja $x(t) = e^{A(t-s)}x(s)$. Z množenjem s prehodno matriko tako v primeru ničelnega vhoda pridemo iz enega stanja v drugega.

Za splošno rešitev nehomogene enačbe potrebujemo še *odziv z ničelnim stanjem* (zero-state response), kjer predpostavimo $x(0) = 0$. Na sistemu naredimo Laplaceovo transformacijo, ki nam sistem

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t) \end{aligned}$$

transformira v

$$\begin{aligned} s\widetilde{x}(s) &= A\widetilde{x}(s) + B\widetilde{u}(s), \\ \widetilde{y}_{\text{zs}}(s) &= C\widetilde{x}(s) + D\widetilde{u}(s). \end{aligned}$$

Rešitev transformiranega sistema je $y_{\text{zs}}(s) = G(s)\widetilde{u}(s)$, kjer je

$$G(s) = C(sI - A)^{-1}B + D \tag{2.4}$$

prenosna funkcija.

Za Laplaceovo transormacijo velja

$$\widetilde{y}_{\text{zs}}(s) = \widetilde{f}_1(s)\widetilde{f}_2(s) \quad \Longrightarrow \quad y_{\text{zs}}(t) = \int_0^\infty f_1(t-\tau)f_2(\tau)d\tau. \tag{2.5}$$

V našem primeru izberemo $\widetilde{f}_1(s) = C(sI - A)^{-1}$ in $\widetilde{f}_2(s) = B\widetilde{u}(s)$, torej $f_1(t) = Ce^{At}$ in $f_2(t) = Bu(t)$. Odtod iz (2.5) sledi

$$y_{\text{zs}}(t) = C\int_0^\infty e^{A(t-\tau)}Bu(\tau)d\tau + Du(t). \tag{2.6}$$

Splošna rešitev je potem vsota $y_{\text{zi}}(t)$ in $y_{\text{zs}}(t)$. Iz (2.3) in (2.6) sledi

$$y(t) = Ce^{At}x_0 + C\int_0^\infty e^{A(t-\tau)}Bu(\tau)d\tau + Du(t). \tag{2.7}$$

Trditev 2.1 *Prenosna funkcija $G(s)$ je neodvisna od izbire baze v prostoru stanj.*

Dokaz. Denimo, da z nesingularno transformacijo S spremenimo predstavitev sistema z matrikami (A, B, C, D) v $(\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D}) = (S^{-1}AS, S^{-1}B, CS, D)$. Potem za prenosno funkcijo v novi bazi velja

$$\widehat{G}(s) = \widehat{C}(sI - \widehat{A})\widehat{B} + \widehat{D} = CS(sI - S^{-1}AS)^{-1}S^{-1}B + D = C(sI - A)^{-1}B + D = G(s). \quad \blacksquare$$

Za numerično računanje odziva (npr. za potrebe simulacije sistema) potrebujemo numerični algoritem za računanje e^{At}. To bomo obravnavali kasneje v poglavju 2.5.

2.3 Povezava s klasično teorijo

Elementi prenosne funkcije (2.4), ki je matrika velikosti $r \times m$, so racionalne funkcije. Tako (i, j)-ti element $G(s)$ predstavlja prenosno funkcijo med j-to komponento vhoda in i-to komponento izhoda v smislu klasične teorije. Poli sistema so sedaj lastne vrednosti matrike A.

V klasični teoriji je prenosna funkcija Laplaceova transformiranka impulznega odziva. Podobno velja tudi sedaj. Če je $\delta_j(t) \in \mathbb{R}^m$ vhodna funkcija, katere j-ta komponenta je enaka enotskemu impulzu, ostale komponente pa so identično enake 0, potem je Laplaceova transformiranka odziva na $\delta_j(t)$ ravno j-ti stolpec v $G(s)$.

Zvezni sistem, ki je predstavljen v klasični teoriji z diferencialne enačbe n-tega reda

$$y^{(n)}(t) + k_1 y^{(n-1)}(t) + \cdots + k_{n-1}y'(t) + k_n y(t) = \beta_0 u^{(m)}(t) + \beta_1 u^{(m-1)}(t) + \cdots + \beta_m u(t), \quad (2.8)$$

lahko zapišemo tudi v prostoru stanj.

V najpreprostejšem primeru na desni strani enačbe (2.8) ni odvodov vhodne funkcije $u(t)$ in imamo enačbo oblike

$$y^{(n)}(t) + k_1 y^{(n-1)}(t) + \cdots + k_{n-1}y'(t) + k_n y(t) = u(t).$$

V tem primeru lahko diferencialno enačbo prevedemo na sistem diferencialnih enačb prvega reda, če npr. vzamemo za spremenljivke stanja

$$\begin{aligned} x_1(t) &= y(t), \\ x_2(t) &= y'(t), \\ &\vdots \\ x_n(t) &= y^{(n)}(t) = -k_1 y^{(n-1)}(t) - \cdots - k_n y(t) + u(t) \end{aligned}$$

Dobimo

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \vdots \\ \vdots \\ \dot{x}_n(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -k_n & -k_{n-1} & & \cdots & -k_1 \end{bmatrix} \cdot \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ \vdots \\ x_n(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ u(t) \end{bmatrix}. \quad (2.9)$$

Iz (2.9) lahko preberemo A in b, velja pa še $c = e_1$ in $d = 0$.

Če imamo na desni strani enačbe (2.8) še odvode, moramo ravnati drugače. Če uredimo člene po stopnji odvoda in predpostavimo $m = n$, dobimo

$$y^{(n)}(t) = \beta_0 u^{(n)}(t) - k_1 y^{(n-1)}(t) + \beta_1 u^{(n-1)}(t) + \cdots + [-k_n y(t) + \beta_n u(t)]. \tag{2.10}$$

Ko enačbo (2.10) n-krat integriramo dobimo (zaradi preglednosti je izpuščen argument t)

$$y = \beta_0 u + \int \left([-k_1 y + \beta_1 u] + \int \left([-k_2 y + \beta_2 u] + \cdots + \int [-k_n y + \beta_n u] dt_1) \cdots \right) dt_{n-1} \right) dt. \tag{2.11}$$

Sedaj lahko uvedemo nove spremenljivke

$$\begin{aligned} y &= \beta_0 u + x_1, \\ \dot{x}_1 &= -k_1 y + \beta_1 u + x_2, \\ &\vdots \\ \dot{x}_{n-1} &= -k_{n-1} y + \beta_{n-1} u + x_n, \\ \dot{x}_n &= -k_n y + \beta_n u. \end{aligned}$$

Od tod dobimo

$$\begin{aligned} \dot{x}_1 &= -k_1 x_1 + x_2 + (\beta_1 - k_1 \beta_0) u, \\ &\vdots \\ \dot{x}_{n-1} &= -k_{n-1} x_1 + x_n + (\beta_{n-1} - k_{n-1} \beta_0) u, \\ \dot{x}_n &= -k_n x_1 + (\beta_n - k_n \beta_0) u. \end{aligned}$$

V matrični obliki lahko sedaj sistem zapišemo kot

$$\dot{x} = \begin{bmatrix} -k_1 & 1 & & & \\ -k_2 & 0 & 1 & & \\ \vdots & & \ddots & \ddots & \\ \vdots & & & \ddots & 1 \\ -k_n & & & & 0 \end{bmatrix} x + \begin{bmatrix} \beta_1 - k_1 \beta_0 \\ \beta_2 - k_2 \beta_0 \\ \vdots \\ \vdots \\ \beta_n - k_n \beta_0 \end{bmatrix}$$

in

$$y = [\, 1 \quad 0 \cdots \quad 0 \,] \, x + \beta_0 u.$$

Tako smo sistem (2.8) v prostoru stanj predstavili v t.i. *spoznavnostni kanonični obliki*. Opazimo lahko, da se karakteristični polinom matrike A ujema z imenovalcem prenosne funkcije sistema (2.8). Poli so tako enaki lastnim vrednostim matrike A.

Podobno bi lahko sistem (2.8) zapisali v *vodljivostni kanonični obliki*

$$\dot{x} = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -k_n & -k_{n-1} & & \cdots & -k_1 \end{bmatrix} \cdot x + \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{bmatrix} u,$$

$$y = [\, \beta_n - k_n \beta_0 \quad \beta_{n-1} - k_{n-1} \beta_0 \quad \cdots \quad \beta_1 - k_1 \beta_0 \,] \, x + \beta_0 u.$$

Tu smo se srečali s pojmoma vodljivosti in spoznavnosti, ki ju bomo podrobneje spoznali v poglavju 3. V grobem vodljivost pomeni ali lahko sistem vzbudimo v poljubno stanje, spoznavnost pa ali lahko iz poznavanja vhoda in izhoda razberemo začetno stanje. V primeru SISO sistema ravno ti dve lastnosti odločata ali se da sistem zapisati v vodljivostni oz. spoznavnosti kanonični formi.

2.4 Diskretni sistemi

Vektorji stanja, vhoda in izhoda so lahko definirani le ob fiksnih trenutkih $t_k = k\Delta t$, kjer je Δt interval vzorčenja. V tem primeru dobimo linearni diskretni časovno nespremenljivi linearni sistem, ki je namesto z diferencialno enačbo predstavljen z diferenčno enačbo

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k, \\ y_{k+1} &= Cx_k + Du_k. \end{aligned}$$

Rešitev homogene enačbe $x_{k+1} = Ax_k$ je $x_k = A^k x_0$, rešitev nehomogene enačbe stanja pa

$$x_k = A^k x_0 + \sum_{i=0}^{k-1} A^{k-i-1} B u_i.$$

Sedaj za numerično računanje potrebujemo natančno in učinkovito računanje potenc matrike A.

Podobno kot pri zveznem sistemu lahko tu pridemo do prenosne funkcije z uporabo z-transformacije. Prenosna funkcija je tako kot pri zveznem sistemu enaka $G(s) = C(sI - A)^{-1}B + D$.

En način, kako pridemo do diskretnega sistema je aproksimacija zveznega sistema, ko predpostavimo, da ima $u(t)$ obliko kosoma konstantne funkcije oz. $u(t) = u(k\Delta t)$ za $k\Delta t \leq t < (k+1)\Delta t$. To velja npr. pri digitalnem vodenju. Pri teh predpostavkah za rešitev zveznega sistema

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t) \end{aligned}$$

za $t \geq k\Delta t$ velja

$$\begin{aligned} x(t) &= e^{A(t-k\Delta t)} x(k\Delta t) + \int_{k\Delta t}^{t} e^{A(t-s)} B u(s) ds, \\ y(t) &= Cx(t) + Du(t), \end{aligned}$$

torej pri $t = (k+1)\Delta t$ velja

$$x((k+1)\Delta t) = e^{A\Delta t} x(k\Delta t) + \left(\int_0^{\Delta t} e^{As} ds \right) B u(k\Delta t) ds.$$

Če označimo $x_k = x(k\Delta t)$, $u_k = u(k\Delta t)$ in $y_k = y(k\Delta t)$, dobimo diskretni sistem

$$\begin{aligned} x_{k+1} &= A_d x_k + B_d u_k, \\ y_{k+1} &= Cx_k + Du_k, \end{aligned}$$

kjer sta $A_d = e^{A\Delta t}$ in $B_d = \left(\int_0^{\Delta t} e^{As} ds \right) B$.

2.5 Numerično računanje matrične eksponentne funkcije

Za $n \times n$ matriko A je eksponentna funkcija definirana z razvojem

$$e^{At} = \sum_{k=0}^{\infty} \frac{(At)^k}{k!},$$

kjer je $t \geq 0$.

Pogledali bomo nekaj numeričnih metod za izračun e^{At} in obravnavali občutljivost eksponentne funkcije matrike. Lep pregled različnih metod za računanje matrične eksponentne funkcije je v [8].

2.5.1 Občutljivost matrične ekponentne funkcije

Pri občutljivosti nas zanima, kako velika je lahko relativna sprememba

$$\Phi(t) = \frac{\|e^{(A+E)t} - e^{At}\|}{\|e^{At}\|},$$

kjer je E motnja matrike A. Če matriki A in E komutirata, potem velja

$$e^{(A+E)t} - e^{At} = e^{At}(e^{Et} - I) = e^{At} Et \sum_{k=0}^{\infty} \frac{(Et)^k}{(k+1)!}.$$

Od tod lahko ocenimo

$$\Phi(t) \leq \|E\| t e^{\|E\| t}.$$

Če matriki A in E ne komutirata, potem analiza ni več tako enostavna. Če odvajamo $e^{A(t-s)}e^{(A+E)s}$ po s, dobimo

$$\begin{aligned} \frac{d}{ds}(e^{A(t-s)}e^{(A+E)s}) &= -Ae^{A(t-s)}e^{(A+E)s} + e^{A(t-s)}(A+E)e^{(A+E)s} \\ &= e^{A(t-s)}Ee^{(A+E)s}ds. \end{aligned}$$

Od tod sledi ocena

$$\Phi(t) \leq \frac{\|E\|}{\|e^{At}\|} \int_0^t \|e^{A(t-s)}\| \cdot \|e^{(A+E)s}\| ds.$$

Definiramo lahko tudi pogojenostno število za e^{At}.

$$\nu(A,t) := \max_{\|E\|=1} \left\| \int_0^t e^{A(t-s)} E e^{As} ds \right\| \cdot \frac{\|A\|}{\|e^{At}\|}.$$

Velja $\nu(A,t) \geq t\|A\|$. Enakost velja za vse $t \geq 0$, če je A normalna matrika. Pri matrikah, ki niso normalne, pa lahko $\nu(A,t)$ raste kot da gre za polinom visoke stopnje v t. Več o tem lahko najdemo v [10]

Poglejmo si nekaj možnosti za oceno $\|e^{At}\|$.

1. Iz Taylorjeve vrste sledi očitna ocena $\|e^{At}\| \leq e^{\|A\|t}$.
2. Dahlquistova ocena je $\|e^{At}\| \leq e^{\mu(A)t}$, kjer je
$$\mu(A) = \max\left\{\mu : \mu \text{ je lastna vrednost } \frac{1}{2}(A^* + A)\right\}.$$
3. Uporabimo Jordanovo formo. Če je $A = XJX^{-1}$, kjer je
$$J = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_m \end{bmatrix}, \qquad J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}$$
in je J_i matrika velikosti $n_i \times n_i$. Potem je
$$e^{Jt} = \begin{bmatrix} e^{J_1 t} & & \\ & \ddots & \\ & & e^{J_m t} \end{bmatrix},$$
kjer je
$$e^{J_i t} = e^{\lambda_i t} \begin{bmatrix} 1 & t & \frac{1}{2}t^2 & \cdots & \frac{1}{(n_i-1)!}t^{n_i-1} \\ & 1 & t & & \\ & & \ddots & \ddots & \vdots \\ & & & 1 & t \\ & & & & 1 \end{bmatrix}.$$
Tako dobimo oceno (uporabimo $\|A\|_2 \leq nN_\infty(A)$)
$$\|e^{J_i t}\|_2 \leq |e^{\lambda_i t}| n_i \max_{0 \leq j \leq n_i - 1} \frac{t^j}{j!}.$$
in
$$\|e^{At}\|_2 \leq \kappa(X) n_{\max} e^{\alpha(A)t} \max_{0 \leq j \leq n_{\max}-1} \frac{t^j}{j!},$$
kjer je
$$\alpha(A) = \max\{\mathrm{Re}(\lambda) : \ \lambda \text{ lastna vrednost } A\}$$
t.i. *spektralna abscisa*.
4. Uporabimo Schurovo formo. Če je $Q^*AQ = D + N$, kjer je $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, potem lahko ocenimo
$$\|e^{At}\|_2 \leq e^{\alpha(A)t} M_S(t),$$
kjer je
$$M_S(t) = \sum_{k=0}^{n-1} \frac{\|Nt\|_2^k}{k!}.$$

Oceni 1. in 2. sta lahko zelo nepraktični v primeru, ko je $\alpha(A) < 0$, saj z naraščajočim t rasteta in ne upoštevata tega, da je v tem primeru limita e^{At} enaka 0, ko gre $t \to \infty$.

Za oceno $\Phi(t)$ se da v primeru uporabe Schurove forme izpeljati (izpeljava je npr. v [10]), da je
$$\Phi(t) \leq t\|E\|_2 M_S(t)^2 e^{tM_S(t)\|E\|_2}.$$

Vemo, da je v primeru normalne matrike $M_S(t) \equiv 1$, torej lahko pri normalnih matrikah pričakujemo dobre rezultate, sicer pa je problem lahko zelo občutljiv.

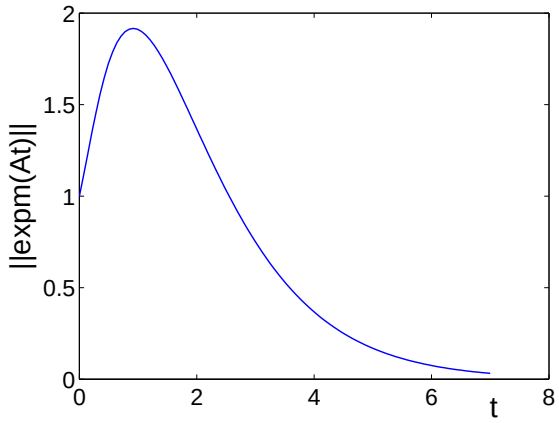
Zgled 2.1 *Za obnašanje e^{At} ni dovolj poznati le lastne vrednosti. Če vzamemo*

$$A = \begin{bmatrix} -1 & M \\ 0 & -1 \end{bmatrix},$$

potem je

$$e^{At} = e^{-t} \begin{bmatrix} 1 & tM \\ 0 & 1 \end{bmatrix}.$$

Preden e^{At} skonvergira proti 0, lahko vrednost $\|e^{At}\|$ nekaj časa narašča in graf $\|e^{At}\|$ ima grbo. Spodnja slika prikazuje grbo v primeru $M = 5$.



Tukaj manjka še: računanje matrične ekponentne funkcije.

Tukaj manjka še: računanje integralov z matrično ekponentno funkcijo.

2.6 Računanje frekvenčnega odziva

Tukaj manjka še: poglavje o ekonomičnem računanju frekvenčnega odziva (2 strani priprav).

Poglavje 3

Vodljivost in spoznavnost

3.1 Uvod

Imamo linearni zvezni kontrolni sistem

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \qquad x(t_0) = x_0, \quad t \geq t_0, \\ y(t) &= Cx(t) + Du(t). \end{aligned}$$

Grobo povedano nam vodljivost pove, v kolikšni meri lahko z vhodom $u(t)$ vplivamo na stanje $x(t)$. Spoznavnost pa nam pove, ali lahko iz poznavanja vhoda $u(t)$ in izhoda $y(t)$ razberemo stanje $x(t)$.

Z obema pojmoma se srečamo, ko želimo s povratno zvezo iz stanja stabilizirati sistem. Če poznamo stanje $x(t)$, lahko za povratno zvezo vzamemo

$$u(t) = v(t) - Kx(t),$$

kjer je $v(t)$ referenčna vhodna funkcija (signal). Dobimo

$$\begin{aligned} \dot{x}(t) &= (A - BK)x(t) + Bv(t), \qquad x(t_0) = x_0, \quad t \geq t_0, \\ y(t) &= (C - DK)x(t) + Dv(t). \end{aligned}$$

Pri stabilizaciji iščemo za dani A, B tako matriko K, da bo $A - BK$ stabilna. Izkaže se, da je obstoj take matrike povezan z vodljivostjo sistema.

Težava pri zgoraj opisani povratni zvezi je, da ponavadi ne poznamo stanja $x(t)$, temveč le izhod $y(t)$. Če želimo vseeno uporabiti povratno zvezo s stanjem, moramo z novim sistemom, t.i. *opazovalcem*, iz vhoda in izhoda generirati čim boljšo sproksimacijo za stanje. Obstoj takega opazovalca pa je povezan s spoznavnostjo sistema.

3.2 Vodljivost

Definicija 3.1 *Za sistem $\dot{x}(t) = Ax(t) + Bu(t)$, $y(t) = Cx(t) + Du(t)$ pravimo, da je vodljiv, če za poljubno začetno stanje x_0 in končno stanje x_1 obstaja končni t_1 in vhod $u(t)$, $0 \leq t \leq t_1$, da iz začetnega stanja $x(0) = x_0$ sistem pride v končno stanje $x(t_1) = x_1$.*

Ker se izkaže, da je vodljivost odvisna le od matrik A in B, govorimo tudi o tem, da je par (A, B) vodljiv.

Izrek 3.2 *Za $A \in \mathbb{R}^{n\times n}$ in $B \in \mathbb{R}^{n\times m}$, $m \le n$, je ekvivalentno:*

1. *sistem $\dot{x}(t) = Ax(t) + Bu(t)$ je vodljiv,*
2. vodljivostna matrika

$$C_M = [\, B \quad AB \quad A^2B \quad \cdots \quad A^{n-1}B \,] \in \mathbb{R}^{n\times(nm)}$$

   je polnega ranga,

3. *Matrika*

$$W_c = \int_0^{t_1} e^{At} BB^T e^{A^T t} dt$$

   je nesingularna za vsak $t_1 > 0$.

Dokaz. (1$\Longrightarrow$2): Denimo, da je $\operatorname{rang}(C_M) < n$. Potem obstaja neničelni vektor $w \in \mathbb{R}^n$, ki ni linearna kombinacija stolpcev matrike C_M. Splošna rešitev enačbe stanja je

$$x(t_1) = e^{At_1}x_0 + \int_0^{t_1} e^{A(t_1-t)}Bu(t)dt. \tag{3.1}$$

Od tod sledi

$$\begin{aligned} x(t_1) - e^{At_1}x_0 &= \int_0^{t_1} \left(I + A(t_1-t) + \frac{A^2(t_1-t)^2}{2} + \cdots \right) Bu(t)dt \\ &= B\int_0^{t_1} u(t)dt + AB\int_0^{t_1} (t_1-t)u(t)dt + A^2B\int_0^{t_1} \frac{(t_1-t)^2}{2}u(t)dt + \cdots. \end{aligned}$$

Po Cayley-Hamiltonovemu izreku je A^n linearna kombinacija $I, A, \ldots, A^{n-1}$, to pa pomeni, da je $x(t_1)$ linearna kombinacija stolpcev $B, AB, \ldots, A^{n-1}B$. Sedaj se iz začetnega stanja $x_0 = 0$ ne moremo premakniti v $x(t_1) = w$, saj $w \notin \operatorname{im}(C_M)$.

(2$\Longrightarrow$3): Denimo, da je matrika W_c singularna. Potem obstaja tak neničelni vektor v, da je $W_c v = 0$, torej tudi $v^T W_c v = 0$. Če definiramo $c(t) = B^T e^{A^T t} v$, lahko opazimo, da je

$$0 = \int_0^{t_1} v^T e^{At} BB^T e^{A^T t} v dt = \int_0^{t_1} c(t)^T c(t) dt = \int_0^{t_1} \|c(t)\|_2^2 \ge 0.$$

Zgornji izraz je lahko nič le v primeru, ko je $c(t) \equiv 0$, torej $v^T e^{At} B = 0$ ta $0 \le t \le t_1$. Potem so tudi vsi odvodi $c(t)$ enaki 0, z odvajanjem pa dobimo $v^T A^i B = 0$ za $i = 1, 2, \ldots$, torej je vektor v ortogonalen na vse stolpce C_M, ki potem ne more biti polnega ranga.

(3$\Longrightarrow$1): Za izbrani x_1 in iščemo $u(t)$, da bo $x(t_1) = x_1$. Pokažimo, da je dobra izbira

$$u(t) = B^T e^{A^T(t_1-t)} W_c^{-1}(-e^{At_1}x_0 + x_1). \tag{3.2}$$

Ko vstavimo (3.2) v (3.1), dobimo

$$\begin{aligned} x(t_1) &= e^{At_1}x_0 + \int_0^{t_1} e^{A(t_1-t)}BB^Te^{A^T(t_1-t)}W_c^{-1}(-e^{At_1}x_0+x_1)dt \\ &= e^{At_1}x_0 + \underbrace{\int_0^{t_1} e^{A(t_1-t)}BB^Te^{A^T(t_1-t)}dt}_{W_c}\cdot W_c^{-1}(-e^{At_1}x_0+x_1) \\ &= e^{At_1}x_0 - e^{At_1}x_0 + x_1 = x_1 \qquad \blacksquare \end{aligned}$$

Vodljivost ni odvisna od morebitnih nesingularnih transformacij vhoda in stanja. Če sta S in T nesingularni matriki in naredimo substituciji $x(t) = S\widetilde{x}(t)$ in $u(t) = T\widetilde{u}(t)$, dobimo

$$\dot{\widetilde{x}}(t) = \widetilde{A}\widetilde{x}(t) + \widetilde{B}\widetilde{u}(t),$$

kjer je $\widetilde{A} = S^{-1}AS$ in $\widetilde{B} = S^{-1}BT$. Ugotovimo lahko, da velja

$$\begin{aligned} \operatorname{rang}([\,\widetilde{B} \quad \widetilde{A}\widetilde{B} \quad \cdots \widetilde{A}^{n-1}\widetilde{B}\,]) &= \operatorname{rang}\left(S^{-1}\,[\,B \quad AB \quad \cdots \quad A^{n-1}B\,]\begin{bmatrix} T & & \\ & \ddots & \\ & & T \end{bmatrix}\right) \\ &= \operatorname{rang}([\,B \quad AB \quad \cdots \quad A^{n-1}B\,]). \end{aligned}$$

Denimo, da par (A, B) ni vodljiv in da je $\operatorname{rang}(C_M) = k < n$. Potem lahko sistem razdelimo na vodljiv in nevodljiv del.

Izrek 3.3 *Par (A, B) ni vodljiv natanko tedaj, ko obstaja taka nesingularna matrika T, da je*

$$\widetilde{A} = TAT^{-1} = \begin{matrix} \\ k \\ n-k \end{matrix}\begin{matrix} \begin{matrix} k & n-k \end{matrix} \\ \begin{pmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ 0 & \widetilde{A}_{22} \end{pmatrix} \end{matrix}, \qquad \widetilde{B} = TB = \begin{matrix} k \\ n-k \end{matrix}\begin{pmatrix} \widetilde{B}_1 \\ 0 \end{pmatrix}, \tag{3.3}$$

par $(\widetilde{A}_{11}, \widetilde{B}_1)$ je vodljiv in $\operatorname{rang}(C_M) = k < n$.

Dokaz. ($\Longrightarrow$): Naj bo $\operatorname{rang}(C_M) = k < n$ in naj bodo $v_1, \ldots, v_k$ linearno neodvisni stolpci iz C_M, ki jih dopolnimo do baze $\mathbb{R}^n$ z $v_{k+1}, \ldots, v_n$. Sedaj vzamemo $T^{-1} = [\,v_1 \quad \cdots \quad v_n\,]$. Dobimo

$$T\,[\,B \quad AB \quad \cdots \quad A^{n-1}B\,] = \begin{matrix} k \\ n-k \end{matrix}\begin{pmatrix} * \\ 0 \end{pmatrix} = [\,TB \quad TAT^{-1}TB \quad \cdots \quad TA^{n-1}T^{-1}TB\,]$$

od koder sledi

$$TB = \widetilde{B} = \begin{matrix} k \\ n-k \end{matrix}\begin{pmatrix} * \\ 0 \end{pmatrix}. \tag{3.4}$$

Iz

$$\widetilde{A} = TAT^{-1} = \begin{matrix} \\ k \\ n-k \end{matrix}\begin{matrix} \begin{matrix} k & n-k \end{matrix} \\ \begin{pmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ \widetilde{A}_{21} & \widetilde{A}_{22} \end{pmatrix} \end{matrix}, \qquad (TAT^{-1})(TB) = \begin{matrix} k \\ n-k \end{matrix}\begin{pmatrix} \widetilde{A}_{11}\widetilde{B}_1 \\ \widetilde{A}_{21}\widetilde{B}_1 \end{pmatrix}$$

in (3.4) sledi $\widetilde{A}_{21}\widetilde{B}_1 = 0$. Podobno iz

$$\widetilde{A}^2\widetilde{B} = \begin{matrix} k \\ n-k \end{matrix} \begin{pmatrix} \widetilde{A}_{11}^2\widetilde{B}_1 \\ \widetilde{A}_{21}\widetilde{A}_{11}\widetilde{B}_1 \end{pmatrix}$$

sledi $\widetilde{A}_{21}\widetilde{A}_{11}\widetilde{B}_1$. Če nadaljujemo na ta način, lahko pokažemo, da velja $\widetilde{A}_{21}\widetilde{A}_{11}^j\widetilde{B}_1 = 0$ za $j = 0, \ldots, n-1$. To pomeni

$$\widetilde{A}_{21}\,[\,\widetilde{B}_1 \quad \widetilde{A}_{11}\widetilde{B}_1 \quad \cdots \quad \widetilde{A}_{11}^{n-1}\widetilde{B}_1\,] = 0,$$

ker pa je zaradi (3.4) matrika

$$[\,\widetilde{B}_1 \quad \widetilde{A}_{11}\widetilde{B}_1 \quad \cdots \quad \widetilde{A}_{11}^{n-1}\widetilde{B}_1\,] \tag{3.5}$$

polnega ranga, mora biti potem $\widetilde{A}_{21} = 0$. Seveda poln rang (3.5) pomeni, da je par $(\widetilde{A}_{11}, \widetilde{B}_1)$ vodljiv.

($\Longleftarrow$): Če obstaja taka matrika T, potem ima vodljivostna matrika za par $(\widetilde{A}, \widetilde{B})$ obliko

$$\begin{bmatrix} \widetilde{B}_1 & \widetilde{A}_{11}\widetilde{B}_1 & \cdots & \widetilde{A}_{11}^{n-1}\widetilde{B}_1 \\ 0 & 0 & \cdots & 0 \end{bmatrix},$$

torej par $(\widetilde{A}, \widetilde{B})$ ni vodljiv, to pa je ekvivalentno temu, da par (A, B) ni vodljiv. ■

Opomba 3.1 *Transformacija T je lahko ortogonalna. Tako dobimo, če npr. uporabimo QR razcep s pivotiranjem na C_M.*

Tukaj manjka še: zgled za vodljivost.

Naslednja kriterija za vodljivost sta t.i. *PBH kriterija* (Popov-Belevitch-Hautus).

Izrek 3.4 *Za par (A, B) je ekvivalentno:*

1. *Par (A, B) je vodljiv.*
2. *Za vsak lastni par (λ, x) matrike A^T velja $x^T B \neq 0$.*
3. *Za vsako lastno vrednost λ matrike A velja* $\operatorname{rang}([\,A - \lambda I \quad B\,]) = n$.

Dokaz. ($1 \Longrightarrow 2$): Naj bo x tak lastni vektor za A^T, da je $A^T x = \lambda x$ in $x^T B = 0$. Potem je

$$x^T C_M = [\,x^T B \quad \lambda x^T B \quad \cdots \quad \lambda^{n-1} x^T B\,] = 0,$$

torej C_M ni polnega ranga in par (A, B) ni vodljiv.

($2 \Longleftrightarrow 3$): Naj bo λ taka lastna vrednost matrike A, da je $\operatorname{rang}([\,A - \lambda I \quad B\,]) < n$. To je ekvivalentno temu, da obstaja neničelni vektor x, da je $x^T\,[\,A - \lambda I \quad B\,] = 0$, to pa je res natanko tedaj, ko je $A^T x = \lambda x$ in $x^T B = 0$.

(2$\Longrightarrow$1): Denimo, da par (A,B) ni vodljiv, torej je $\text{rang}(C_M) = k < n$. Potem po izreku 3.3 obstaja nesingularna matrika T, da je

$$\widetilde{A} = TAT^{-1} = \begin{matrix} \\ k \\ n-k \end{matrix} \begin{matrix} \begin{matrix} k & n-k \end{matrix} \\ \begin{pmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ 0 & \widetilde{A}_{22} \end{pmatrix} \end{matrix}, \qquad \widetilde{B} = TB = \begin{matrix} k \\ n-k \end{matrix} \begin{pmatrix} \widetilde{B}_1 \\ 0 \end{pmatrix}.$$

Če je (λ, z) lastni par za $\widetilde{A}_{22}^T$, potem je $\begin{bmatrix} 0 \\ z \end{bmatrix}$ lastni vektor za $\widetilde{A}$, saj je $\widetilde{A}\begin{bmatrix} 0 \\ z \end{bmatrix} = \lambda \begin{bmatrix} 0 \\ z \end{bmatrix}$, velja pa tudi $[0 \; z^T]\widetilde{B} = 0$. ■

Omenili smo že, da vodljivost ni odvisna od nesingularnih transformacij vhoda in stanja. S pomočjo prave transformacije lahko A in B spravimo v obliko, iz katere se da direktno na zelo enostaven način preveriti vodljivost. Dve najpreprostejši možni transformaciji sta diagonalizacija matrike A, kadar je to možno, in pa transformacija matrike A v Jordanovo obliko. V tem primeru veljajo naslednje trditve.

Posledica 3.5 *Naj obstaja taka nesingularna matrika X, da je matrika $\widetilde{A} = X^{-1}AX$ diagonalna in naj bo $\widetilde{B} = X^{-1}B$. Potem je par (A,B) vodljiv natanko tedaj, ko so vse vrstice $\widetilde{B}$, ki pripadajo isti lastni vrednosti, linearno neodvisne.*

Dokaz. Vemo, da je par (A,B) vodljiv natanko tedaj, ko je vodljiv par $(\widetilde{A}, \widetilde{B})$. Zanj uporabimo točko 3. iz izreka 3.4. Očitno je, da vrstice $\widetilde{B}$, ki pripadajo lastni vrednosti λ, niso linearno neodvisne (kar v primeru ene same vrstice pomeni, da je ničelna), natanko tedaj, ko matrika $[\,\widetilde{A} - \lambda I \quad \widetilde{B}\,]$ ni polnega ranga. ■

Opomba 3.2 *Če ima matrika A v zgornji posledici same enostavne lastne vrednosti, potem je sistem vodljiv natanko tedaj, ko so vse vrstice matrike $\widetilde{B}$ neničelne.*

Posledica 3.6 *Če imamo enovhodni sistem in se da matriko A diagonalizirati, potem je sistem vodljiv natanko tedaj, ko so vse lastne vrednosti matrike A enostavne.*

Izrek 3.7 *Naj bo $\widetilde{A} = S^{-1}AS$ in $\widetilde{B} = S^{-1}B$, kjer je $\widetilde{A}$ Jordanova forma matrike A. Par $(\widetilde{A}, \widetilde{B})$ je vodljiv natanko tedaj, ko so vrstice $\widetilde{B}$, ki pripadajo zadnjim vrsticam Jordanovih kletk iste lastne vrednosti, linearno neodvisne.*

Posledica 3.8 *Če imamo enovhodni sistem, potem je potreben pogoj za vodljivost nederogatornost matrike A.*

Izrek 3.9 *Sistem $\dot{x}(t) = Ax(t) + Bu(t)$ je vodljiv natanko tedaj, ko lahko z ustrezno izbrano matriko K poljubno razporedimo lastne vrednosti $A - BK$ (s to omejitvijo, da morajo kompleksne lastne vrednosti nastopati v konjugiranih parih)*

Izrek bomo dokazali v nadaljevanju, ko bomo za poljubno razporeditev lastnih vrednosti eksplicitno skonstruirali ustrezno matriko K.

Definicija 3.10 *Lastna vrednost λ matrike A ni vodljiva, če za levi lastni vektor x velja $x^T B = 0$ (sicer pa je vodljiva).*

Če λ ni vodljiva lastna vrednosti, potem je λ tudi lastna vrednost matrike $A - BK$ nedovisno od izbire matrike K. Torej lahko sistem stabiliziramo s povratno zvezo preko stanja le, če so vse nevodljive lastne vrednosti stabilne. V tem primeru je sistem *stabilizabilen*. Če imamo razcep (3.3), potem je sistem stabilizabilen, če so vse lastne vrednosti iz nevodljivega dela $\widetilde{A}_{22}$ stabilne.

Iz zgoraj omenjenega lahko zaključimo naslednji izrek.

Izrek 3.11 *Za par (A, B) je ekvivalentno:*

1. *par (A, B) je stabilizabilen,*
2. *obstaja taka matrika K, da je $A - BK$ stabilna,*
3. $\operatorname{rang}([A - \lambda I \ B]) = n$ *za vse* $\operatorname{Re}(\lambda) \geq 0$,
4. *za $x \neq 0$ in λ, da je $x^* A = \lambda x^*$ in $\operatorname{Re}(\lambda) \geq 0$, sledi $x^* B \neq 0$.*

Če je par (A, B) vodljiv, je očitno tudi stabilizabilen.

3.2.1 Diskretni sistemi

Pri diskretnih sistemih je situacija dokaj podobna. Za diskretni sistem

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k, \\ y_{k+1} &= Cx_k + Du_k. \end{aligned}$$

pravimo, da je vodljiv, če za poljubno začetno stanje x_0 in končno stanje $\widetilde{x}$ obstaja končno zaporedje vhodov $\{u_0, u_1, \ldots, u_{N-1}\}$, da je $x_N = \widetilde{x}$. Kadar je začetno stanje $x_0 = 0$, govorimo o *dosegljivem* (reachable) sistemu oz. o sistemu *vodljivem iz izhodišča*. Predpostavimo torej $x_0 = 0$. Ker je spet vse odvisno le od matrik A in B, spet govorimo o vodljivosti para (A, B).

Izrek 3.12 *Par (A, B) je vodljiv natanko tedaj, ko je rang vodljivostne matrike*

$$C_M = [\, B \quad AB \quad \cdots \quad A^{n-1}B \,]$$

enak n.

Dokaz. Vemo, da velja

$$x_N = \sum_{k=0}^{N-1} A^{N-k-1} B u_k = [\, B \quad AB \quad \cdots \quad A^{N-1}B \,] \begin{bmatrix} u_{N-1} \\ u_{N-2} \\ \vdots \\ u_0 \end{bmatrix}.$$

Očitno lahko poljubno rešitev $\widetilde{x}$ dosežemo le, če je matrika $[\, B \quad AB \quad \cdots \quad A^{N-1}B \,]$ polnega ranga. Ker za $N \geq n$ velja

$$\operatorname{rang}([\, B \quad AB \quad \cdots \quad A^{N-1}B \,]) = \operatorname{rang}([\, B \quad AB \quad \cdots \quad A^{n-1}B \,]),$$

mora biti vodljivostna matrika polnega ranga. ■

3.3 Spoznavnost

Definicija 3.13 *Linearni zvezni kontrolni sistem* $\dot{x}(t) = Ax(t) + Bu(t)$, $y(t) = Cx(t) + Du(t)$ *je spoznaven, če obstaja tak* $t - 1 > 0$, *da iz poznavanja* $u(t)$ *in* $y(t)$ *za* $0 \le t \le t_1$ *lahko določimo* $x(0)$.

Ker je to odvisno le od para (A, C), govorimo o spoznavnosti para (A, C).

Ko poznamo $x(0)$, lahko potem izračunamo $x(t)$ za poljuben t. Spoznavnost torej ni omejena le na možnost ugotavljanja $x(0)$, temveč lahko (če je sistem spoznaven) ugotovimo stanje v poljubnem času.

Vemo, da velja

$$y(t) = Ce^{At}x_0 + \int_0^t Ce^{A(t-\tau)}Bu(\tau)d\tau + Du(t).$$

Naj bo $g(t) = y(t) - \int_0^t Ce^{A(t-\tau)}Bu(\tau)d\tau - Du(t)$. Potem je

$$g(t) = Ce^{At}x_0, \tag{3.6}$$

kjer poznamo $g(t)$, saj poznamo $y(t)$ in $u(\tau)$ za $0 \le \tau \le t$. Če obe strani (3.6) pomnožimo z $e^{A^T t}C^T$ in integriramo, dobimo

$$\int_0^{t_1} e^{A^T t}C^T Ce^{At}dt\ x_0 = \int_0^{t_1} e^{A^T t}C^T g(t)dt.$$

Sedaj označimo $V(t_1) := \int_0^{t_1} e^{A^T t}C^T Ce^{At}dt$. Če je $V(t_1)$ obrnljiva, je

$$x_0 = V(t_1)^{-1} \int_0^{t_1} e^{A^T t}C^T g(t)dt.$$

To pomeni, da je v primeru, ko je $V(t_1)$ obrnljiva, sistem spoznaven.

Izrek 3.14 *Za linearni zvezni kontrolni sistem* $\dot{x}(t) = Ax(t) + Bu(t)$, $y(t) = Cx(t) + Du(t)$ *je ekvivalentno:*

1. *par* (A, C) *je spoznaven,*
2. *t.i. spoznavnostna matrika* $O_M = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}$ *ima poln rang* n,
3. *Matrika*
$$W_O = \int_0^{t_1} e^{A^T t}C^T Ce^{At}dt$$
je nesingularna za vse $t_1 > 0$,
4. *matrika* $\begin{bmatrix} \lambda I - A \\ C \end{bmatrix}$ *je ranga* n *za vse lastne vrednosti* λ *matrike* A,

5. *noben lastni vektor matrike A ni pravokoten na vrstice C: če je (λ, x) lastni par za A je $Cx \neq 0$,*
6. *obstaja taka matrika L, da so lastne vrednosti $A+LC$ poljubne (z omejitvijo, da kompleksne lastne vrednosti nastopajo v konjugiranih parih).*

Dokaz. Dokaz poteka podobno kot dokaz izreka 3.2. ■

Opazimo lahko, da je spoznavnostna matrika ravno transponirana vodljivostna matrika para (A^T, C^T). To pomeni, da sta vodljivost in spoznavnost dualni in zaradi tega lahko za ugotavljanje spoznavnosti uporabimo tudi kakšno izmed metod, ki jih imamo za ugotavljanje vodljivosti.

Izrek 3.15 *Par (A, C) ni spoznaven natanko tedaj, ko obstaja taka nesingularna matrika T, da je*

$$\widetilde{A} = TAT^{-1} = \begin{bmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ 0 & \widetilde{A}_{22} \end{bmatrix}, \qquad \widetilde{C} = CT^{-1} = [\,0 \quad \widetilde{C}_1\,],$$

kjer je par $(\widetilde{A}_{11}, \widetilde{C}_1)$ spoznaven, $\widetilde{A}_{11}$ pa je velikosti $k \times k$, kjer je k rang spoznavnostne matrike O_M.

Dokaz. Skličemo se na dualnost vodljivosti in spoznavnosti in uporabimo izrek 3.3. ■

Za linearni kontrolni sistem (2.1, 2.2) oz. za matrični par (A, C) pravimo, da je *zaznaven*, če se par (A^T, C^T) da stabilizirati. Zaznavnost je dualna stabilizabilnosti sistema.

Izrek 3.16 *Za par (A, C) je ekvivalentno:*

1. *par (A, C) je zaznaven,*
2. *par (A^T, C^T) je stabilizabilen,*
3. *obstaja taka matrika L, da je $A - LC$ stabilna matrika,*
4. $\operatorname{rang}(\begin{bmatrix} A - \lambda I \\ C \end{bmatrix}) = n$ *za vse* $\operatorname{Re}(\lambda) \geq 0$,
5. *za $x \neq 0$ in λ, da je $Ax = \lambda x$ in $\operatorname{Re}(\lambda) \geq 0$, sledi $Cx \neq 0$.*

Če je par (A, C) spoznaven, je očitno tudi zaznaven.

Tukaj manjka še: zgled za spoznavnost in rekonstrukcijo $x(0)$.

S kombiniranjem izrekov o razcepu na vodljiv in nevodljiv del (izrek 3.3) in na spoznaven in nespoznaven del (izrek 3.15), dobimo *Kalmanov kanonični razcep*.

Izrek 3.17 *Za linearni zvezni kontrolni sistem $\dot{x}(t) = Ax(t) + Bu(t)$, $y(t) = Cx(t) + Du(t)$ obstaja nesingularna transformacija T, da je*

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ & A_{22} & 0 & A_{24} \\ & & A_{33} & A_{34} \\ & & & A_{44} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \\ 0 \\ 0 \end{bmatrix} u(t)$$

in

$$y(t) = \begin{bmatrix} 0 & C_2 & 0 & C_4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix},$$

kjer za štiri podsisteme velja:

- x_1 *je vodljiv in ni spoznaven,*
- x_2 *je vodljiv in je spoznaven,*
- x_3 *ni vodljiv in ni spoznaven,*
- x_4 *ni vodljiv in je spoznaven.*

Prenosna funkcija je $G(s) = C_2(sI - A_{22})^{-1}B_2 + D$.

Dokaz. Ta vsebina je povezana s prvimi domačimi nalogami in bo dodana kasneje. ■

Tukaj manjka še: zgled za Kalmanov kanonični razcep.

3.4 Kanonične oblike

Vemo: če je T nesingularna matrika, potem je par (A, B) vodljiv natanko tedaj, ko je vodljiv par $(\widetilde{A}, \widetilde{B})$, kjer je $\widetilde{A} = TAT^{-1}$ in $\widetilde{B} = TB$. Pokazali bomo, da lahko transformacijo T izberemo tako, da imata $\widetilde{A}$ in $\widetilde{B}$ obliko, iz katere lahko na preprost način ugotovimo vodljivost.

3.4.1 Vodljivostna normalna oblika

Denimo, da imamo enovhodni linearni zvezni kontrolni sistem $\dot{x}(t) = Ax(t) + bu(t)$, kjer je par (A, b) vodljiv. To pomeni, da je vodljivostna matrika $C_M = \begin{bmatrix} b & Ab & \cdots & A^{n-1}b \end{bmatrix}$ ranga n.

Naj bo s_n zadnja vrstica C_M^{-1}. Za transformacijsko matriko vzamemo

$$T = \begin{bmatrix} s_n \\ s_nA \\ \vdots \\ s_nA^{n-1} \end{bmatrix}. \tag{3.7}$$

Matrika T je nesingularna, saj lahko hitro preverimo, da velja

$$T \cdot C_M = \begin{bmatrix} & & & 1 \\ & & 1 & \times \\ & \iddots & \iddots & \vdots \\ 1 & \times & \cdots & \times \end{bmatrix}.$$

Poglejmo, kakšno obliko imata $\widetilde{A}$ in $\widetilde{b}$. Ker je s_n zadnja vrstica C_M^{-1}, velja

$$s_nb = s_nAb = \cdots = s_nA^{n-2}b = 0 \quad \text{in} \quad s_nA^{n-1}b = 1.$$

Zaradi tega je

$$\widetilde{b} = Tb = \begin{bmatrix} s_n b \\ \vdots \\ s_n A^{n-2} b \\ s_n A^{n-1} b \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{3.8}$$

Za $\widetilde{A} = TAT^{-1}$ velja

$$TA = \widetilde{A}T = \begin{bmatrix} s_n A \\ \vdots \\ s_n A^n \end{bmatrix}.$$

Po Cayley-Hamiltonovemu izreku je $A^n = -a_0 I - a_1 A - \cdots - a_{n-1} A^{n-1}$, kjer je $a_0 + a_1 x + \cdots + a_{n-1} x^{n-1} + x^n$ karakteristični polinom matrike A. Torej je $s_n A^n = -a_0 s_n - a_1 s_n A - \cdots - a_{n-1} s_n A^{n-1}$. S primerjanjem leve in desne strani $TA = \widetilde{A}T$ dobimo

$$\widetilde{A} = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -a_0 & -a_1 & \cdots & \cdots & -a_{n-1} \end{bmatrix}. \tag{3.9}$$

Matrika $\widetilde{A}$ ni nič drugega kot pridružena matrika karakterističnega polinoma matrike A, katerega koeficiente lahko preberemo iz zadnje vrstice $\widetilde{A}$.

Če je sistem v taki obliki, kot sta $\widetilde{A}$ in $\widetilde{b}$ v (3.9) oz. (3.8), potem pravimo, da je v *vodljivostni normalni obliki*.

Tukaj manjka še: zgled za vodljivostno normalno obliko (Žak, str. 113).

3.4.2 Luenbergerjeva vodljivostna kanonična oblika

V prejšnjem razdelku smo videli, kako lahko transformiramo enovhodni sistem, če je vodljiv. Posplošitev obstaja tudi za primer, ko ima sistem več kot en vhod. Denimo torej, da imamo večvhodni sistem z vodljivim parom (A, B), kjer je A velikosti $n \times n$, B velikosti $n \times m$, $m \leq n$ in $\operatorname{rang}(B) = m$. Če matriko B zapišemo s stolpci kot $B = [\, b_1 \quad \cdots \quad b_m \,]$ potem lahko vodljivostno matriko C_M zapišemo kot

$$C_M = [\, b_1 \quad \cdots \quad b_m \quad Ab_1 \quad \cdots \quad Ab_m \quad A^2 b_1 \quad \cdots \quad A^2 b_m \quad \cdots \quad A^{n-1} b_1 \quad \cdots \quad A^{n-1} b_m \,].$$

Ker je C_M polnega ranga, lahko iz matrike izberemo n linearno neodvisnih stolpcev. Izbiramo jih po vrsti z leve proti desni, tako da so med njimi vsi stolpci $b_1, \ldots, b_m$, saj je matrika B polnega ranga. Hitro lahko tudi razmislimo, da v primeru, ko je $A^k b_j$ vsebovan v tem naboru linearno neodvisnih stolpcev C_M, to velja tudi za $A^{k-1} b_j$, saj jih izbiramo z leve proti desni. Od tod tudi sledi, da če $A^k b_j$ ni vsebovan v omenjenem naboru, potem tam tudi ni vektorja $A^{k+1} b_j$.

Ko preuiredimo vrstni red n linearno neodvisnih stolpcev, ki smo jih pobrali iz matrike C_M z leve proti desni, iz njih sestavimo $n \times n$ nesingularno matriko L, ki ima obliko

$$L = [\, b_1 \quad \cdots \quad A^{d_1 - 1} b_1 \quad \cdots \quad b_m \quad \cdots \quad A^{d_m - 1} b_m \,].$$

Števila $d_1, \ldots, d_m$ so *vodljivostni indeksi* para (A, B). Za njih velja $d_1 + \cdots + d_m = n$. Za *vodljivostni indeks* d para (A, B) vzamemo maksimalnega izmed vodljivostnih indeksov, oziroma

$d = \max\{d_i : \ i = 1, \ldots, m\}$. Definiramo še $\sigma_k = d_1 + \cdots + d_k$ za $k = 1, \ldots, m$, torej $\sigma_1 = d_1$, $\sigma_2 = d_1 + d_2$, ..., $\sigma_m = n$.

Sedaj iz matrike L^{-1} vzamemo m vrstic na mestih $\sigma_1, \ldots, \sigma_m$, označimo jih z $q_1, \ldots, q_m$. Iz njih sestavimo matriko velikosti $n \times n$ oblike

$$T = \begin{bmatrix} q_1 \\ q_1 A \\ \vdots \\ q_1 A^{d_1 - 1} \\ \vdots \\ q_m \\ q_m A \\ \vdots \\ q_m A^{d_m - 1} \end{bmatrix},$$

ki jo bomo uporabili za transformacijsko matriko. Preverimo lahko, da je matrika T nesingularna, saj je $\det(TL) = \pm 1$. Par $(\widetilde{A}, \widetilde{B}) = (TAT^{-1}, TB)$ ima sedaj bločno obliko

$$\widetilde{A} = \begin{bmatrix} \widetilde{A}_{11} & \cdots & \widetilde{A}_{1m} \\ \vdots & & \vdots \\ \widetilde{A}_{m1} & \cdots & \widetilde{A}_{mm} \end{bmatrix}, \quad \widetilde{B} = \begin{bmatrix} \widetilde{B}_1 \\ \vdots \\ \widetilde{B}_m \end{bmatrix},$$

kjer so bloki $\widetilde{A}_{ij}$ velikosti $d_i \times d_j$, bloki $\widetilde{B}_i$ pa velikosti $d_i \times m$. Diagonalni bloki matrike $\widetilde{A}$ imajo obliko pridružene matrike

$$\widetilde{A}_{ii} = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ \times & \times & \cdots & \cdots & \times \end{bmatrix},$$

izvendiagonalni pa imajo neničelno le zadnjo vrstico, torej

$$\widetilde{A}_{ij} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \\ \times & \times & \cdots & \times \end{bmatrix}.$$

Blok $\widetilde{B}_i$ ima oblko

$$\widetilde{B}_i = \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \times & \cdots & \times \end{bmatrix},$$

kjer ima zadnja vrstica na prvih $i - 1$ mestih vrednost 0.

Tukaj manjka še: zgled za Luenbergerjevo vodljivostno kanonično obliko (Žak, str. 116).

3.4.3 Spoznavnostna normalna oblika

Podobne rezultate kot pri vodljivosti dobimo tudi pri spoznavnosti. V primeru enoizhodnega sistema za spoznaven par (A, c) obstaja nesingularna transformacija P, da je

$$\widetilde{A} = PAP^{-1} = \begin{bmatrix} 0 & & & & -a_0 \\ 1 & 0 & & & -a_1 \\ & \ddots & \ddots & & \vdots \\ & & 1 & 0 & -a_{n-2} \\ & & & 1 & -a_{n-1} \end{bmatrix} \quad \text{in} \quad \widetilde{c} = cP^{-1} = [\,0 \quad \cdots \quad 0 \quad 1\,].$$

Zgornji obliki pravimo *spoznavnostna normalna oblika*.

Za večizhodne sisteme imamo *Luenbergerjevo spoznavnostno kanonično obliko*. Naj bo par (A, C) spoznaven, kjer je A velikosti $n \times n$, C velikosti $p \times n$, $p \leq n$ in $\operatorname{rang}(C) = p$. Potem lahko poiščemo tako nesingularno transformacijo P, da ima par $(\widetilde{A}, \widetilde{C}) = (PAP^{-1}, CP^{-1})$ bločno obliko

$$\widetilde{A} = \begin{bmatrix} \widetilde{A}_{11} & \cdots & \widetilde{A}_{1p} \\ \vdots & & \vdots \\ \widetilde{A}_{p1} & \cdots & \widetilde{A}_{pp} \end{bmatrix}, \quad \widetilde{C} = [\,\widetilde{C}_1 \quad \cdots \quad \widetilde{C}_p\,],$$

kjer so bloki $\widetilde{A}_{ij}$ velikosti $\widetilde{d}_i \times \widetilde{d}_j$, bloki $\widetilde{C}_i$ pa velikosti $n \times \widetilde{d}_i$. Diagonalni bloki matrike $\widetilde{A}$ imajo obliko pridružene matrike

$$\widetilde{A}_{ii} = \begin{bmatrix} 0 & & & & \times \\ 1 & 0 & & & \times \\ & \ddots & \ddots & & \vdots \\ & & 1 & 0 & \times \\ & & & 1 & \times \end{bmatrix},$$

izvendiagonalni pa imajo neničelen le zadnji stolpec, torej

$$\widetilde{A}_{ij} = \begin{bmatrix} 0 & \cdots & 0 & \times \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \times \end{bmatrix}.$$

Blok $\widetilde{C}_i$ ima oblko

$$\widetilde{C}_i = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & \times \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \times \end{bmatrix},$$

kjer ima zadnji stolpec na prvih $i-1$ mestih vrednost 0.

Opazimo lahko, da gre za transponirano obliko Luenbergerjeve vodljivostne kanonične oblike, kar se spet ujema z dualnostjo med vodljivostjo in spoznavnostjo. Do same transformacije P lahko v obeh zgornjih primerih pridemo tako, da vzamemo transponirano transformacijo v vodljivostno obliko za par (A^T, C^T).

Števila $\widetilde{d}_1, \ldots, \widetilde{d}_p$ so *spoznavnostni indeksi* para (A, C) in se ujemajo z vodljivostnimi indeki para (A^T, C^T). Maksimalni $\widetilde{d}_i$ je *spoznavnostni indeks* $\widetilde{d}$ para (A, C).

3.5 Vodljivostna Hessenbergova oblika

Težava z vodljivostno normalno oz. Luenbergerjevo obliko je, da je prehodna matrika T lahko zelo občutljiva. Za numerično stabilnost je bolje, če je matrika T ortogonalna.

Numerično stabilen test za preverjanje vodljivosti je redukcija para (A, B) v bločno Hessenbergovo obliko s pomočjo ortogonalnih transformacij. Dobimo ortogonalno matriko P, da je $\widetilde{A} = PAP^T = H$ bločna zgornja Hessenbergova matrika in $\widetilde{B} = PB = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}$. Tak par imenujemo *vodljivostna Hessenbergova oblika*, par $(H, \widetilde{B})$ pa je *vodljivostno Hessenbergov par* para (A, B). Redukcijo naredimo s t.i. *stopničastnim algoritmom*.

Denimo, da je A velikosti $n \times n$, B pa velikosti $n \times m$ in $m \leq n$. Algoritem za redukcijo para (A, B) je sestavljen iz naslednjih štirih korakov:

Korak 0 Matriko B spremenimo v zgornjo trikotno obliko. Uporabimo lahko QR razcep s pivotiranjem po stolpcih, da dobimo ortogonalno matriko P_1 in permutacijsko matriko E_1, da je

$$P_1 B E_1 = \begin{bmatrix} \widetilde{B}_1 \\ 0 \end{bmatrix},$$

kjer je $\widetilde{B}_1$ zgornja trikotna matrika velikosti $n_1 \times n$, kjer je $n_1 = \text{rang}(\widetilde{B}_1) = \text{rang}(B)$.

Korak 1 Posodobimo A in B. Izračunamo produkta

$$H_1 = P_1 A P_1^T = \begin{bmatrix} H_{11}^{(1)} & H_{12}^{(1)} \\ H_{21}^{(1)} & H_{22}^{(1)} \end{bmatrix} \quad \text{in} \quad \widetilde{B} = P_1 B = \begin{bmatrix} \widetilde{B}_1 \\ 0 \end{bmatrix} E^T \equiv \begin{bmatrix} B_1 \\ 0 \end{bmatrix},$$

kjer je $H_{11}^{(1)}$ velikosti $n_1 \times n_1$. Če je $H_{21}^{(1)} = 0$, končamo.

Korak 2 Matriko $H_{21}^{(1)} = 0$ spravimo s zgornjo trikotno obliko. Uporabimo lahko QR razcep s pivotiranjem po stolpcih, da dobimo ortogonalno matriko $\widehat{P}_2$ in permutacijsko matriko E_2, da je

$$\widehat{P}_2 H_{21}^{(1)} E_2 = \begin{bmatrix} H_{21}^{(2)} \\ 0 \end{bmatrix},$$

kjer je $H_{21}^{(2)}$ velikosti $n_2 \times n_1$, kjer je $n_2 = \text{rang}(H_{21}^{(2)}) = \text{rang}(H_{21}^{(1)})$. Če je $n_1 + n_2 = n$, končamo.

Sicer pa posodobimo

$$H_2 = P_2 H_1 P_2^T = \begin{bmatrix} H_{11}^{(1)} & H_{12}^{(2)} & H_{13}^{(2)} \\ H_{21}^{(2)} & H_{22}^{(2)} & H_{23}^{(2)} \\ 0 & H_{32}^{(2)} & H_{33}^{(2)} \end{bmatrix},$$

kjer je

$$P_2 = \begin{bmatrix} I_{n_1} & 0 \\ 0 & \widehat{P}_2 \end{bmatrix},$$

velikost $H_{22}^{(2)}$ je $n_2 \times n_2$, velikost $H_{32}^{(2)}$ je $(n - n_1 - n_2) \times n$ in $H_{21}^{(2)} = H_{21}^{(1)} E_2^T$.

Transformacijo P popravimo na $P_2 P_1$. Če je $H_{32}^{(2)} = 0$, potem končamo.

Korak 3 Zgornji postopek ponovimo za $H_{32}^{(2)}$ in ponavljamo toliko časa, dokler na nek $k \leq n$ ne dobimo

$$H = \begin{bmatrix} H_{11} & H_{12} & \cdots & H_{1k} \\ H_{21} & H_{22} & \cdots & H_{2k} \\ & \ddots & & \vdots \\ & & H_{k,k-1} & H_{kk} \end{bmatrix}, \qquad \widetilde{B} = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \tag{3.10}$$

kjer je

a) $H_{k,k-1}$ polnega ranga n_k, od koder sledi, da je par (A, B) vodljiv;

b) $H_{k,k-1} = 0$, od koder sledi, da je par (A, B) ni vodljiv.

Opomba 3.3 *V vsakem koraku algoritma določamo rang bloka matrike preko QR razcepa s pivotiranjem po stolpcih. Namesto tega lahko (kar je natančneje toda tudi dražje) uporabimo tudi singularni razcep.*

Opomba 3.4 *Kakor hitro v algoritmu naletimo na ničelni blok na poddiagonali H ali pa B_1 nima polnega ranga to pomeni, da par (A, B) ni vodljiv.*

Izrek 3.18 *Stopnični algoritem vrne tako ortogonalno matriko P, da je $PAP^T = H$ in $PB = \widetilde{B}$, kjer za H in $\widetilde{B}$ velja, da imata obliko (3.10). Če je par (A, B) vodljiv, je $H_{k,k-1}$ polnega ranga, če pa par (A, B) ni vodljiv, je $H_{k,k-1} = 0$.*

Dokaz. Par (A, B) je vodljiv natanko tedaj, ko je $\operatorname{rang}([\, B \quad A - \lambda I \,]) = n$ za vse λ. Velja $\operatorname{rang}([\, B \quad A - \lambda I \,]) = \operatorname{rang}([\, \widetilde{B} \quad H - \lambda I \,])$ in

$$[\, \widetilde{B} \quad H - \lambda I \,] = \begin{bmatrix} B_1 & H_{11} - \lambda I_1 & H_{12} & \cdots & H_{1k} \\ 0 & H_{21} & H_{22} - \lambda I_2 & \cdots & H_{2k} \\ \vdots & & \ddots & & \vdots \\ 0 & & & H_{k,k-1} & H_{kk} - \lambda I_k \end{bmatrix}.$$

Če je par (A, B) vodljiv, mora imeti $[\, \widetilde{B} \quad H - \lambda I \,]$ poln rang za vsak λ, torej mora biti $H_{k,k-1}$ polnega ranga. Če pa par (A, B) ni vodljiv, potem $[\, \widetilde{B} \quad H - \lambda I \,]$ ne more imeti polnega ranga in $H_{k,k-1}$ mora potem biti enak 0.

Algoritem je obratno stabilen. Za izračunani matriki $\widetilde{H}$ in $\widetilde{B}$ velja, da je $\widetilde{H} = H + \Delta H$ in $\widetilde{B} = B + \Delta B$, kjer je $\|\Delta H\| \leq c\|H\|_F u$, $\|\Delta B\| \leq c\|B\|_F u$ in je c majhna konstanta.

Zahtevnost algoritma je približno $6n^3 + 2n^m$ operacij.

Opomba 3.5 *Indeks k, kjer se konča stopničasti algoritem, je enak vodljivostnemu indeksu para (A, B).*

V primeru enovhodnega sistema se vodljivostna Hessenbergova oblika za par (A, b) spremeni v

$$PAP^T = H = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ & \ddots & & \vdots \\ & & h_{n,n-1} & h_{nn} \end{bmatrix}, \qquad \widetilde{b} = Pb = \begin{bmatrix} b_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Izrek 3.19 *Par (A,b) je vodljiv, če je pripadajoča vodljivostna Hessenbergova oblika $(H,\widetilde{b})$ taka, da je $b_1 \neq 0$, H pa je ireducilna zgornja Hessenbergova matrika, torej $h_{i,i-1} \neq 0$ za $i = 2,\ldots,n$. Če zgornji pogoj ni izpolnjen, potem par (A,b) ni vodljiv.*

Dokaz. Velja

$$\operatorname{rang}([\,b \quad Ab \quad \cdots A^{n-1}b\,]) = \operatorname{rang}([\,\widetilde{b} \quad H\widetilde{b} \quad \cdots H^{n-1}\widetilde{b}\,]).$$

Matrika $[\,\widetilde{b} \quad H\widetilde{b} \quad \cdots H^{n-1}\widetilde{b}\,]$ je zgornja trikotna matrika z diagonalnimi elementi b_1, $h_{21}b_1$, $h_{21}h_{32}b_1$, $\ldots$, $h_{21}\cdots h_{n,n-1}b_1$, ki je polnega ranga, če velja $b_1 \neq 0$ in $h_{i,i-1} \neq 0$ za $i = 2.\ldots,n$.

Če je $b_1 = 0$, potem sistem očitno ni vodljiv. Če velja $h_{i,i-1} = 0$, potem vodljivostna matrika C_M ni polnega ranga in (A,b) prav tako ni vodljiv. ∎

Opomba 3.6 *Podobno lahko naredimo pri spoznavnosti. Sedaj par (A,C) spremenimo v $H = QAQ^T$ in $\widetilde{C} = CQ^T = [\,0 \quad \cdots \quad 0 \quad C_1\,]$, kjer je H bločna zgornja Hessenbergova matrika. Če je par (A,C) spoznaven, je H bločno ireducibilna, C_1 pa je polnega ranga.*

Tukaj manjka še: kakšen zgled.

3.6 Razporejanje polov

Imamo linearni zvezni kontrolni sistem

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \qquad x(t_0) = x_0, \quad t \geq t_0, \\ y(t) &= Cx(t) + Du(t). \end{aligned}$$

S povratno zvezo iz stanja želimo stabilizirati sistem oziroma razporediti pole prenosne funkcije. Denimo da poznamo stanje $x(t)$. Potem za povratno zvezo lahko vzamemo $u(t) = v(t) - Kx(t)$, kjer je $v(t)$ referenčna vhodna funkcija in K povratnozančna matrika velikosti $m \times n$. Dobimo

$$\begin{aligned} \dot{x}(t) &= (A - BK)x(t) + Bv(t), \qquad x(t_0) = x_0, \quad t \geq t_0, \\ y(t) &= (C - DK)x(t) + Dv(t). \end{aligned}$$

Poli prenosne funkcije so sedaj ničle zaprtozančnega karakterističnega polinoma $\det(sI - A + BK)$. Denimo, da želimo, da so lastne vrednosti matrike $A - BK$ enake podanim vrednostim $s_1,\ldots,s_n$. Ker je matrika K realna, morajo morebitne kompleksne ničle nastopati v konjugiranih parih. Iz željenih lastnih vrednosti lahko sestavimo ciljni zaprtozančni karakteristični polinom

$$\alpha(s) = (s - s_1)\cdots(s - s_n) = s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_1 s + \alpha_0.$$

Naš cilj je poiskati tako povratnozančno matriko K, da bo

$$\det(sI - A + BK) = \alpha(s).$$

Tej nalogi pravimo tudi *problem razporejanja polov*.

Najprej si poglejmo, kako lahko problem rešimo kadar je sistem enovhoden.

3.6.1 Ackermanova formula

V tem primeru je $K = k \in \mathbb{R}^{1\times n}$. Denimo, da je sistem podan v vodljivostni normalni obliki. Potem je

$$A - bk = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -a_0 - k_1 & -a_1 - k_2 & \cdots & \cdots & -a_{n-1} - k_n \end{bmatrix}.$$

Od tod lahko preberemo, da moramo za elemente k vzeti $k_i = \alpha_{k-1} - a_{k-1}$ za $k = 1, \ldots, n$.

Če par (A, b) ni v vodljivostni normalni obliki, moramo najprej poiskati prehodno nesingularno matriko T, da bosta $\widetilde{A}$ in $\widetilde{b}$ prave oblike. Vemo, da taka matrika T obstaja natanko takrat, ko je par (A, b) vodljiv. Če je

$$\widetilde{k} = [\, \alpha_0 - a_0 \quad \alpha_1 - a_1 \quad \cdots \quad \alpha_{n-1} - a_{n-1} \,]$$

prava izbira povratne zveze za transformirani par $(\widetilde{A}, \widetilde{b}) = (TAT^{-1}, Tb)$, potem je za originalni par (A, b) prava izbira $k = \widetilde{k}T$.

Če upoštevamo formulo (3.7), potem velja

$$k = \widetilde{k}T = [\, \alpha_0 - a_0 \quad \alpha_1 - a_1 \quad \cdots \quad \alpha_{n-1} - a_{n-1} \,] \begin{bmatrix} s_n \\ s_n A \\ \vdots \\ s_n A^{n-1} \end{bmatrix}, \tag{3.11}$$

kjer je s_n zadnja vrstica inverza vodljivostne matrike C_M^{-1}. Ko zmnožimo (3.11), dobimo

$$\widetilde{k}T = s_n(\alpha_0 I + \alpha_1 A + \cdots + \alpha_{n-1}A^{n-1}) - s_n(a_0 I + a_1 A + \cdots + a_{n-1}A^{n-1}).$$

Sedaj upoštevamo, da je po Cayley-Hamiltonovemu izreku $A^n = -(a_0 I + a_1 A + \cdots + a_{n-1}A^{n-1})$. Od tod sledi

$$k = s_n \alpha(A). \tag{3.12}$$

Izraz (3.12) se imenuje *Ackermanova formula* za razporejanje polov.

Poglavje 4

Stabilnost

4.1 Uvod

Homogeni sistem $\dot{x}(t) = Ax(t)$, $x(0) = x_0$, je

- *asimptotično stabilen*, če za vsak x_0 velja, da gre $x(t)$ proti 0, ko gre t proti ∞;
- *stabilen*, če za vsak x_0 obstaja taka konstanta $c > 0$, da velja $\|x(t)\| < c$, ko gre t proti ∞;
- *nestabilen*, če obstaja tak x_0, pri katerem gre $\|x(t)\|$ proti ∞, ko gre t proti ∞.

Izrek 4.1 *Homogeni sistem $\dot{x}(t) = Ax(t)$, $x(0) = x_0$, je asimptotično stabilen natanko tedaj, ko za vse lastne vrednosti A velja* $\mathrm{Re}(\lambda) < 0$.

Sistem je stabilen, če velja $\mathrm{Re}(\lambda) \leq 0$, lastne vrednosti z $\mathrm{Re}(\lambda) = 0$ pa so polenostavne. Če vse lastne vrednosti matrike A zadoščajo pogoju $\mathrm{Re}(\lambda) \leq 0$, potem pravimo, da je A *stabilna matrika*.

Sistem

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \qquad x(t_0) = x_0, \quad t \geq t_0, \\ y(t) &= Cx(t) + Du(t). \end{aligned} \tag{4.1}$$

je *BIBO stabilen* (bounded input-bounded output), če je za vsak omejen vhod tudi izhod omejen. BIBO stabilnost je odvisna od polov prenosne funkcije $G(s) = C(sI - A)^{-1}B + D$.

Izrek 4.2 *Sistem (4.1) je BIBO stabilen natanko tedaj, ko imajo vsi poli prenosne funkcije $G(s)$ negativni realni del.*

Vsak pol $G(s)$ je lastna vrednost matrike A, torej je vsak asimptotično stabilen sistem tudi BIBO stabilen. Obratno pa ni nujno res.

Zgled 4.1 *Če vzamemo*

$$\dot{x} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u, \qquad y = \begin{bmatrix} 1 & 1 \end{bmatrix} x,$$

potem je

$$G(s) = C(sI - A)^{-1}B = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} s-1 & 0 \\ 0 & s+1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} u = \frac{1}{s+1}.$$

Od tod sledi, da je sistem BIBO stabilen, ni pa asimptotično stabilen. □

Izrek 4.3 *Sistem (4.1) je BIBO stabilen natanko tedaj, ko*

1. *vse lastne vrednosti A imajo nepozitivne realne dele,*
2. *če je $\mathrm{Re}(\lambda_i) = 0$, je λ_i polenostavna,*
3. *če je $\mathrm{Re}(\lambda_i) = 0$, potem λ_i ni vodljiva (za levi lastni vektor x velja $x^T B = 0$).*

4.2 Stabilnost po Ljapunovu

S pomočjo kriterija Ljapunova lahko teoretično brez računanja lastnih vrednosti matrike A ugotovimo, ali je matrika A stabilna ali ne.

Izrek 4.4 *Sistem $\dot{x}(t) = Ax(t)$, $x(0) = x_0$, je asimptotično stabilen natanko tedaj, ko je za poljubno simetrično pozitivno definitno matriko Q rešitev* zvezne enačbe Ljapunova

$$A^T P + PA = -Q \tag{4.2}$$

enolična, simetrična in pozitivno definitna.

Dokaz. ($\Longleftarrow$:) Naj bo (λ, x) lastni par matrike A. Potem iz

$$A^T P + PA = -Q$$

sledi

$$x^* A^T P x + x^* P A x = -x^* Q x$$

in

$$\overline{\lambda} x^* P x + \overline{\lambda} x^* P x = (\lambda + \overline{\lambda}) \underbrace{x^* P x}_{>0} = -\underbrace{x^* Q x}_{>0}$$

Od tod sledi, da je $\lambda + \overline{\lambda} < 0$, torej $\mathrm{Re}(\lambda) < 0$.

($\Longrightarrow$:) Pokažimo, da za rešitev P lahko vzamemo kar

$$P = \int_0^\infty e^{A^T t} Q e^{At} dt. \tag{4.3}$$

$$\begin{aligned} AP + P^T A &= \int_0^\infty e^{A^T t} Q e^{At} A dt + \int_0^\infty A^T e^{A^T t} Q e^{At} dt \\ &= \int_0^\infty \frac{d}{dt} \left(e^{A^T t} Q e^{At} \right) dt = e^{A^T t} Q e^{At} \Big|_0^\infty . \end{aligned}$$

Ker je A stabilna matrika, gre $e^{A^T t}$ proti 0, ko gre $t \to \infty$. To pa pomeni, da je $AP + P^T A = Q$ in P je res rešitev (4.2).

Simetrija P sledi iz same konstrukcije (4.3), pokazati pa moramo še pozitivno definitnost. Za poljuben vektor u velja

$$u^T P u = \int_0^\infty u^T e^{A^T t} Q e^{At} u \, dt = \int_0^\infty \left(e^{AT t} u\right)^T Q e^{At} u \, dt.$$

Ker je matrika e^{At} nesingularna, Q pa pozitivno definitna, je $u^T P u > 0$.

Pokažimo še enoličnost. Denimo, da sta P_1 in P_2 različni rešitvi (4.2). Potem je

$$A^T(P_1 - P_2) + (P_1 - P_2)A = 0,$$

od tod pa sledi

$$e^{A^T t}\left(A^T(P_1 - P_2) + (P_1 - P_2)A\right) e^{At} = 0,$$

kar je ekvivalentno

$$\frac{d}{dt}\left(e^{A^T t}(P_1 - P_2)e^{At}\right) = 0.$$

Od tod sledi, da je $e^{A^T t}(P_1 - P_2)e^{At}$ konstantna matrika za vsak t. Ko vstavimo $t = 0$ in $t = \infty$, ugotovimo, da mora biti potem $P_1 - P_2 = 0$. ∎

Opomba 4.1 *Iz samega dokaza je razvidno, da matrika P, definirana z (4.3), zadošča enačbi Ljapunova tudi v primeru, ko matrika Q ni pozitivno definitna.*

Zvezna matrična enačba Ljapunova

$$PA + A^T P = -Q$$

nastopa tudi v dualni obliki

$$AP + P^T A = -Q.$$

Če je A stabilna matrika in Q simetrična pozitivno definitna, potem velja:

1. enolična rešitev $PA + A^T P = -Q$ je $P = \int_0^\infty e^{A^T t} Q e^{At} dt$,
2. enolična rešitev $AP + P^T A = -Q$ je $P = \int_0^\infty e^{At} Q e^{A^T t} dt$.

Enačba Ljapunova nastopa pogoste tudi v obliki, ko Q ni pozitivno definitna, temveč le nenegativno definitna, npr. ko je $Q = BB^T$ ali $Q = C^T C$, kjer sta B in C vhodna oz. izhodna matrika.

Izrek 4.5 *Naj bo P rešitev enačbe Ljapunova $PA + A^T P = -C^T C$. Velja:*

1. *Če je P s.p.d. in je par (A, C) spoznaven, potem je A stabilna;*
2. *Če A stabilna in je par (A, C) spoznaven, potem je je P s.p.d.;*
3. *Če A stabilna in je P s.p.d., potem je par (A, C) spoznaven.*

Dokaz.

1. Naj bo (λ, x) lastni par matrike A. Potem iz enačbe Ljapunova sledi

$$x^*PAx + x^*A^TPx = -x^*C^TCx,$$

od tod pa

$$(\lambda + \overline{\lambda})x^*Px = -\|Cx\|^2.$$

Ker je par (A, c) spoznaven, je $Cx \neq 0$, zato mora veljati $\lambda + \overline{\lambda} < 0$ in Aej stabilna.

2. Ker je A stabilna, je rešitev podana z

$$P = \int_0^\infty e^{A^Tt}C^TCe^{At}dt.$$

Če P ni simetrična pozitivno definitna, potem obstaja $x \neq 0$, da je $Px = 0$, torej $x^TPx = 0$, od tod pa iz

$$x^TPx = \int_0 \infty x^Te^{A^Tt}C^TCe^{At}t\,dt = \int_0^\infty \left\|Ce^{At}x\right\| dt$$

sledi $Ce^{At} = 0$, to pa pomeni, da par (A, C) ni spoznaven.

3. Denimo, da par (A, C) ni spoznaven. Potem obstaja tak lastni par (λ, x) matrike A, da je $Cx = 0$. Od tod podobno kot v točki 1. sledi

$$(\lambda + \overline{\lambda})x^*Px = -\|Cx\|^2 = 0.$$

Ker je $\lambda + \overline{\lambda} < 0$, mora biti $x^*Px = 0$, torej P ni s.p.d..

■

Soroden izrek lahko pokažemo tudi za vodljivost.

Izrek 4.6 *Naj bo P rešitev enačbe Ljapunova $AP + PA^T = -BB^T$. Velja:*

1. *Če je P s.p.d. in je par (A, B) vodljiv, potem je A stabilna;*
2. *Če A stabilna in je par (A, B) vodljiv, potem je je P s.p.d.;*
3. *Če A stabilna in je P s.p.d., potem je par (A, B) vodljiv.*

Definicija 4.7 *Matriko*

$$O_G = \int_0^\infty e^{A^Tt}BB^Te^{At}dt$$

imenujemo vodljivostna Gramova matrika*,*

$$C_G = \int_0^\infty e^{A^Tt}C^TCe^{At}dt$$

pa spoznavnostna Gramova matrika*.*

Izrek 4.8 *Naj bo matrika A stabilna.*

1. *Vodljivostna Gramova matrika C_G reši enačbo Ljapunova*

$$AC_G + C_G A^T = -BB^T$$

in je s.p.d. natanko tedaj, ko je par (A, B) vodljiv.

2. *Spoznavnostna Gramova matrika O_G reši enačbo Ljapunova*

$$O_G A + A^T O_G = -C^T C$$

in je s.p.d. natanko tedaj, ko je par (A, C) spoznaven.

V nekaterih primerih se da stabilnost matrike preveriti tudi na lažji način.

Zgled 4.2 *Za par*

$$A = \begin{bmatrix} -1 & -2 & -3 \\ 0 & -2 & -1 \\ 0 & 0 & -3 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

dobimo vodljivostno Gramovo matriko

$$C_G = \frac{1}{24} \begin{bmatrix} 7 & 1 & 1 \\ 1 & 4 & 4 \\ 1 & 4 & 4 \end{bmatrix},$$

ki je očitno singularna. Ker C_G ni pozitivno definitna, par (A, b) ni vodljiv. To je razvidno tudi iz vodljivostne matrike, ki je enaka

$$C_M = \begin{bmatrix} 1 & -6 & 21 \\ 1 & -3 & 9 \\ 1 & -3 & 9 \end{bmatrix}.$$

□

Izrek 4.9 *Če je matrika A strogo diagonalno dominantna, torej*

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$$

za $i = 1, \ldots, n$ in so vsi diagonalni elementi a_{ii} negativni, potem je A stabilna matrika.

Dokaz. Skličemo se na Gerschgorinov izrek, ki pravi, da vse lastne vrednosti matrike A ležijo v uniji krogov

$$K_i = \left\{ z \in \mathbb{C} : \ |z - a_{ii}| < \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \right\}$$

za $i = 1, \ldots, n$. Ker v našem primeru vsi krogi K_i ležijo v levi polravnini kompleksne ravnine, morajo imeti vse lastne vrednosti negativne realne dele. ■

4.3 Diskretni sistemi

Diskretni sistem $x_{k+1} = Ax_k$ je *asimptotično stabilen*, če so absolutne vrednosti vseh lastnih vrednosti matrike A strogo manjše od 1. Podobno kot pri zveznem primeru je sistem *stabilen*, če za vse lastne vrednosti λ matrike A velja $|\lambda| \leq 1$, tiste z absolutno vrednostjo 1 pa so polenostavne.

Če za vse lastne vrednosti λ matrike A velja $|\lambda| < 1$ pravimo, da je A *konvergentna matrika*.

Podobno kot pri zveznem sistemu je tudi tokrat stabilnost povezana z rešitvijo enačbe Ljapunova. *Diskretna enačba Ljapunova* nastopa v dualnih oblikah

$$P - A^T P A = Q$$

in

$$P - APA^T = Q.$$

Izrek 4.10 *Diskretni sistem $x_{k+1} = Ax_k$ je asimptotično stabilen natanko tedaj, ko je za poljubno simetrično pozitivno definitno matriko Q rešitev diskretne enačbe Ljapunova*

$$P - A^T P A = Q \tag{4.4}$$

enolična, simetrična in pozitivno definitna.

Dokaz. ($\Longleftarrow$:) Naj bo (λ, x) lastni par matrike A. Potem iz

$$P - A^T P A = Q$$

sledi

$$x^* P x + x^* A^T P A x = x^* Q x$$

in

$$x^* P x + \lambda\overline{\lambda} x^* P x = (1 - |\lambda|^2) \underbrace{x^* P x}_{>0} = \underbrace{x^* Q x}_{>0}.$$

Od tod sledi, da je $1 - |\lambda|^2 > 0$, torej $|\lambda| < 1$.

($\Longrightarrow$:) Pokažimo, da za rešitev P lahko vzamemo kar

$$P = \sum_{k=0}^{\infty} (A^T)^k Q A^k. \tag{4.5}$$

Ker je matrika A konvergentna, je vsota v (4.5) konvergentna in P je v redu definirana. Očitno je, da je P simetrična pozitivno definitna matrika. Zaradi

$$P - A^T P A = \sum_{k=0}^{\infty} (A^T)^k Q A^k - \sum_{k=1}^{\infty} (A^T)^k Q A^k = Q$$

je P tudi rešitev diskretne Ljapunove enačbe (4.4).

Pokazati moramo še enoličnost. Denimo, da poleg (4.5) enačbo (4.4) reši še matrika $\widetilde{P}$. Potem velja

$$P = \sum_{k=0}^{\infty}(A^T)^k Q A^k = \sum_{k=0}^{\infty}(A^T)^k(\widetilde{P} - A^T\widetilde{P})A^k = \sum_{k=0}^{\infty}(A^T)^k\widetilde{P}A^k - \sum_{k=1}^{\infty}(A^T)^k\widetilde{P}A^k = \widetilde{P},$$

torej je rešitev enolična. ■

Podobno kot pri zveznem problemu nas tudi sedaj zanimajo rešitve enačbe Ljapunova, kadar je Q le nenegativno definitna, torej npr. $Q = BB^T$ ali $Q = C^TC$. Naslednja dva izreka sta diskretni obliki izrekov 4.5 in 4.6.

Izrek 4.11 *Naj bo P rešitev diskretne enačbe Ljapunova $P - A^TPA = C^TC$. Velja:*

1. *Če je P s.p.d. in je par (A, C) spoznaven, potem je A konvergentna;*
2. *Če A konvergentna in je par (A, C) spoznaven, potem je je P s.p.d.;*
3. *Če A konvergentna in je P s.p.d., potem je par (A, C) spoznaven.*

Izrek 4.12 *Naj bo P rešitev diskretne enačbe Ljapunova $P - A^TPA = BB^T$. Velja:*

1. *Če je P s.p.d. in je par (A, B) vodljiv, potem je A konvergentna;*
2. *Če A konvergentna in je par (A, B) vodljiv, potem je je P s.p.d.;*
3. *Če A konvergentna in je P s.p.d., potem je par (A, B) vodljiv.*

Definicija 4.13 *Matriko*

$$O_G^D = \sum_{k=0}^{\infty}(A^T)^k BB^T A^k$$

imenujemo diskretna vodljivostna Gramova matrika,

$$C_G^D = \sum_{k=0}^{\infty}(A^T)^k C^T C A^k$$

pa diskretna spoznavnostna Gramova matrika.

4.4 Klasična teorija

Tukaj manjka še: poglavlje o stabilnosti klasičnih sistemov (5 strani priprav).

4.5 Oddaljenost od nestabilnih sistemov

Naj bo A kompleksna stabilna matrika. Zanima nas, kako močno je matrika stabilna oz. koliko je blizu nestabilnosti.

Eno merilo naj bi bila t.i. *stabilnostna abscisa*

$$\alpha(A) := \max\{\mathrm{Re}(\lambda) : \ \lambda \text{ lastna vrednost } A\}.$$

Če je $\alpha(A) < 0$ je matrika stabilna in pričakujemo da manjša ko je vrednost $\alpha(A)$, dalje je A od nestabilne matrike.

Zgled 4.3 *Za matriko*

$$A = \begin{bmatrix} -0.5 & 1 & 1 & 1 & 1 & 1 \\ & -0.5 & 1 & 1 & 1 & 1 \\ & & -0.5 & 1 & 1 & 1 \\ & & & -0.5 & 1 & 1 \\ & & & & -0.5 & 1 \\ & & & & & -0.5 \end{bmatrix},$$

je $\alpha(A) = -0.5$. *Če pa* A *zmotimo na mestu* $(6,1)$ *v* $1/324$, *ima nova matrika lastne vrednosti* -0.8006, $-0.7222 \pm 0.2485i$, $-0.3775 \pm 0.4120i$, 0 *in je nestabilna. V tem primeru je stabilna matrika* A *dosti bližje nestabilni matriki kot kaže* $\alpha(A)$.

Boljša mera za stabilnost je razdalja do množice nestabilnih matrik. Definiramo jo kot

$$\beta(A) = \min\{\|E\| : \ A + E \text{ ni stabilna matrika}\}.$$

Ekvivalentno bi lahko zapisali

$$\beta(A) = \min\{\|E\| : \ A + E \text{ ima lastno vrednost } \lambda, \text{ kjer je } \mathrm{Re}(\lambda) \geq 0\}.$$

Lema 4.14 *Naj bo* A *stabilna kompleksna matrika. Potem je*

$$\beta(A) = \min_{\omega \in \mathbb{R}} \sigma_{\min}(A - i\omega I), \tag{4.6}$$

kjer je $\sigma_{\min}$ *najmanjša singularna vrednost.*

Preko formule (4.6) lahko izračunamo $\beta(A)$ z uporabo kakšne metode za nelinearno optimizacijo.

Za poljuben $\omega \in \mathbb{R}$ dobimo zgornjo mejo $\beta(A) \leq \sigma_{\min}(A - i\omega I)$.

Van Loan (1985) je predlagal, da za približek za $\beta(A)$ vzamemo kar

$$\min\left\{\sigma_{\min}(A - i \cdot \mathrm{im}(\lambda) I) : \ \lambda \text{ lastna vrednost } A\right\}.$$

Ko se A spremeni v najbližjo matriko, ki ni stabilna, ima $A + E$ neko strogo imaginarno lastno vrednost. Za pričakovati je, da bo imaginarni del te lastne vrednosti ostal nespremenjen ali vsaj blizu imaginarnemu delu lastne vrednosti matrike A.

Zgled 4.4 *Demmel (1987) je našel protiprimer za zgornjo domnevo. Če vzamemo*

$$A = \begin{bmatrix} -1 & -b & -b^2 \\ & -1 & -b \\ & & -1 \end{bmatrix}, \qquad b \gg 1,$$

potem je

$$\min\{\sigma_{\min}(A - i \cdot \mathrm{im}(\lambda)I) : \ \lambda \in \lambda(A)\} = \sigma_{\min}(A) = \mathcal{O}(b^{-1}),$$

vendar je $\beta(A) = \mathcal{O}(b^{-2})$.

Je pa res, da v praksi ocena vrača zadovoljive vrednosti, odpove le v primerih, ko je matrika A hkrati defektna in derogatorna.

Izrek 4.15 *Naj bo A stabilna kompleksna matrika. Za $\sigma \geq 0$ velja, da je $\sigma \geq \beta(A)$ natanko tedaj, ko ima matrika*

$$H(\sigma) = \begin{bmatrix} A & -\sigma I \\ \sigma I & -A^* \end{bmatrix}$$

čisto imaginarno lastno vrednost.

Matrika $H(\sigma)$ je Hamiltonska matrika. V realnem primeru je matrika Hamiltonska, če ima strukturo

$$H = \begin{bmatrix} A & G \\ Q & -A^T \end{bmatrix},$$

kjer sta G in Q simetrični. Lastne vrednosti realne Hamiltonske matrike nastopajo v

- parih $\{\lambda, -\lambda\}$, kadar je $\lambda \in \mathbb{R}$ ali $\lambda \in i\mathbb{R}$,
- četvorkah $\{\lambda, -\lambda, \overline{\lambda}, -\overline{\lambda}\}$, kadar je $\lambda \in \mathbb{C}\backslash(\mathbb{R} \cup i\mathbb{R})$.

$H \in \mathbb{R}^{2n \times 2n}$ je Hamiltonska $\Leftrightarrow (HJ)^T = HJ$, kjer je $J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}$.

Byers (1988) je razvil naslednji algoritem, ki oceni $\beta(A)$ do faktorja 10 natančno ali pa ugotovi, da je $\beta(A)$ pod predpisano toleranco.

Dana je matrika A in toleranca $\tau > 0$

$\alpha = 0$
$\nu = \frac{1}{2}\|A + A^*\|_2$
Dokler je $\nu > 10\max(\tau, \alpha)$
  $\sigma = \sqrt{\nu \max(\tau, \alpha)}$
  Če ima $H(\sigma)$ čisto imaginarno lastno vrednost vzemi $\nu = \sigma$ sicer pa $\alpha = \sigma$.

Glavni del algoritma je ugotavljanje, ali ima $H(\sigma)$ imaginarno lastno vrednosti ali ne, saj to preverjamo numerično. Tu pomagajo algoritmi, ki ohranjajo Hamiltonsko strukturo.

Sicer pa velja, da lahko pri izračunani lastni vrednosti postavimo realni del na 0, če je po absolutni vrednosti reda $\mathcal{O}(u^{1/2}\|A\|_F)$, kjer je u osnovna zaokrožitvena napaka.

Vemo, da je stabilnost povezan z enačbo Ljapunova in ta povezava nam da naslednjo oceno za $\beta(A)$.

Izrek 4.16 *Naj bo A kompleksna stabilna matrika in X enolična hermitska pozitivno definitna rešitev enačbe Ljapunova*

$$XA + A^*X = -M,$$

kjer je M hermitska pozitivno definitna matrika. Potem je

$$\beta(A) \geq \frac{\lambda_{\min}(M)}{2\|X\|},$$

kjer je $\lambda_{\min}(M)$ najmanjša lastna vrednost M.

Razdalja do nestabilnosti v diskretnem primeru

V primeru konvergentne matrike je ustrezna razdalja do nestabilnosti definirana z

$$\gamma(A) = \min\{\|E\| : \text{ za nek } \theta \in \mathbb{R} \text{ je } e^{i\theta} \text{ lastna vrednost } A + E\}.$$

Izrek 4.17 *Za $n \times n$ matriko A obstaja taka vrednost $\Gamma(A) \in \mathbb{R}$, da je $\Gamma(A) \geq \gamma(A)$ in da ima v primeru $\Gamma(A) \geq \sigma \geq \gamma(A)$ Hamiltonski matrični šop*

$$H_D(\sigma) = F(\sigma) - \lambda G(\sigma) = \begin{bmatrix} -\sigma I_n & A \\ I_n & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & I_n \\ A^T & -\sigma I_n \end{bmatrix}$$

posplošeno lastno vrednost z absolutno vrednostjo 1.

Na podlagi zgornjega izreka lahko tudi v diskretnem primeru ocenimo razdaljo do nestabilnosti s pomočjo bisekcije.

4.5.1 Robustna stabilnost

V praksi je sistem vedno podvržen določenim motnjam. Tako si lahko npr. mislimo, da matrika A ni znana eksaktno, temveč je znana matrika $A + E$, kjer je E neka motnja.

Zato bi radi vedeli, ali bo sistem, kjer je namesto A v resnici $A + E$, še vedno stabilen.

Izrek 4.18 *Naj bo A stabilna matrika in naj ima motnja E obliko*

$$E = \sum_{i=1}^{r} p_i E_i,$$

kjer matrike $E_1, \ldots, E_r$ določajo strukturo same motnje. Če je X enolična simetrična pozitivno definitna rešitev enačbe Ljapunova

$$XA + A^T X = -Q,$$

kjer je Q simetrična pozitivno definitna matrika, potem je matrika $A+E$ stabilna za

$$\sum_{i=1}^{r} |p_i|^2 \leq \frac{\sigma_{\min}^2(Q)}{\sum_{i=1}^{r} \|E_i^T X + X E_i\|^2}.$$

Denimo, da je matrika A stabilna. Količina $\beta(A)$ meri oddaljenost od najbližje matrike, ki ni stabilna. V praksi pa imajo motnje E lahko posebno strukturo in potem je bolj smiselno vprašanje, koliko je oddaljena najbližja nestabilna matrika z dano strukturo.

Iščemo motnje oblike BEC, kjer sta matriki B in C fiksni (in sta v praksi ravno matriki iz predstavitve sistema v prostoru stanj), matrika E pa je spremenljiva.

Za stabilno matriko A velikosti $n \times n$ in matriki B in C velikosti $n \times m$ in $r \times n$ definiramo *radij stabilnosti* ($\mathbb{F} = \mathbb{R}$ ali $\mathbb{F} = \mathbb{C}$)

$$r_{\mathbb{F}}(A,B,C) = \inf\{\|E\|_2 : \ E \in \mathbb{F}^{m\times r} \text{ in } A+BEC \text{ ni stabilna}\}.$$

Lema 4.19 *Velja*

$$r_{\mathbb{F}}(A,B,C) = \inf_{\omega\in\mathbb{R}} \{\|E\|_2 : \ E \in \mathbb{F}^{m\times r} \text{ in } \det(I - EG(i\cdot\omega)) = 0\},$$

kjer je $G(s) = C(sI-A)^{-1}B$.

Izrek 4.20

$$r_{\mathbb{C}}(A,B,C) = \left(\sup_{\omega\in\mathbb{R}} \|G(i\cdot\omega)\|\right)^{-1}.$$

Če imamo realne matrike in dopuščamo le realne motnje, dobimo večje radije stabilnosti, saj očitno vedno velja $r_{\mathbb{R}}(A,B,C) \geq r_{\mathbb{C}}(A,B,C)$. Razmerje je lahko poljubno veliko.

Izrek 4.21

$$r_{\mathbb{R}}(A,B,C) = \left(\sup_{\omega\in\mathbb{R}} \mu_{\mathbb{R}}(G(i\cdot\omega))\right)^{-1},$$

kjer je $\mu_{\mathbb{R}}(M) = \left(\inf\{\|E\| : E \in \mathbb{R}^{m\times r} \text{ in } \det(I-EM) = 0\}\right)^{-1}$.

Lema 4.22

$$\mu_{\mathbb{R}}(M) = \inf_{\gamma\in(0,1]} \sigma_2\left(\begin{bmatrix} \mathrm{Re}(M) & -\gamma\,\mathrm{im}(M) \\ \gamma^{-1}\,\mathrm{im}(M) & \mathrm{Re}(M) \end{bmatrix}\right).$$

Če je A kompleksna matrika, velja $\beta(A) = r_{\mathbb{C}}(A,I,I) = \min_{\omega\in\mathbb{R}} \sigma_{\min}(A - i\cdot\omega I)$, za realno matriko pa velja

$$\beta(A) = r_{\mathbb{R}}(A,I,I) = \min_{\omega\in\mathbb{R}} \max_{\gamma\in(0,1]} \sigma_{2n-1}\left(\begin{bmatrix} A & -\gamma\omega I \\ \gamma^{-1}\omega I & A \end{bmatrix}\right),$$

kjer je σ_{2n-1} druga najmanjša singularna vrednost.

Poglavje 5

Numerično reševanje Sylvestrove enačbe in enačbe Ljapunova

5.1 Kroneckerjev produkt

V nadaljevanju bomo potrebovali naslednji dve definiciji. *Kroneckerjev produkt* matrik $W \in \mathbb{R}^{p\times q}$ in $Z \in \mathbb{R}^{s\times t}$ je matrika velikosti $ps \times qt$ bločne oblike

$$W \otimes Z = \begin{bmatrix} w_{11}Z & \cdots & w_{1p}Z \\ \vdots & & \vdots \\ w_{p1}Z & \cdots & w_{pq}Z \end{bmatrix}.$$

Vektorizacija matrike $W = [\, w_1 \quad \cdots \quad w_q \,] \in \mathbb{R}^{p\times q}$ je vektor velikosti pq bločne oblike

$$\mathrm{vec}(W) = \begin{bmatrix} w_1 \\ \vdots \\ w_q \end{bmatrix}.$$

Enačba Ljapunova $A^TX + XA = -Q$ je poseben primer Sylvestrove enačbe

$$AX + XB = C, \tag{5.1}$$

kjer so dane matrike $A \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{n\times n}$ in $C \in \mathbb{R}^{m\times n}$, iščemo pa $X \in \mathbb{R}^{m\times n}$. Hitro lahko preverimo, da veljajo naslednje enakosti

$$\begin{aligned} \mathrm{vec}(AX) &= (I_n \otimes A)\mathrm{vec}(X), && (5.2)\\ \mathrm{vec}(XB) &= (B^T \otimes I_m)\mathrm{vec}(X), && (5.3)\\ \mathrm{vec}(AXB) &= (B^T \otimes A)\mathrm{vec}(X), && (5.4)\\ \|\mathrm{vec}(X)\|_2 &= \|X\|_F. && (5.5) \end{aligned}$$

Iz enakosti (5.2) in (5.3) sledi, da je Sylvestrova enačba (5.1) ekvivalentna linearnemu sistemu

$$(I_n \otimes A + B^T \otimes I_m)\mathrm{vec}(X) = \mathrm{vec}(C) \tag{5.6}$$

Lema 5.1 *Za matriki $A \in \mathbb{R}^{m\times m}$ in $B \in \mathbb{R}^{n\times n}$ velja*

1. *lastne vrednosti $A \otimes B$ so $\lambda_i \mu_j$,*
2. *lastne vrednosti $I_n \otimes A + B^T \otimes I_m$ so $\lambda_i + \mu_j$,*

kjer so λ_i za $i = 1, \ldots, m$ lastne vrednosti matrike A, μ_j za $j = 1, \ldots, n$ pa lastne vrednosti matrike B.

Dokaz. Za matriki A in B obstajata unitarni matriki U, V in zgornji trikotni matriki R, S, da je $R = U^*AU$ in $S = V^*BV$. Potem velja

$$A \otimes B = (U \otimes V)(R \otimes S)(U \otimes V)^*,$$

$R \otimes S$ pa je zgornja trikotna matrika, ki ima na diagonali vse možne produkte $r_{ii}s_{jj}$. To pa so ravno vsi možni produkti parov lastnih vrednosti matrik A in B. Podobno je

$$I_n \otimes A + B^T \otimes I_m = (U \otimes V)(I_n \otimes R + S^T \otimes I_m)(U \otimes V)^*,$$

$I_n \otimes R + S^T \otimes I_m$ pa je spet zgornja trikotna matrika, ki ima na diagonali vse možne vsote $r_{ii} + s_{jj}$. ■

Posledica (5.6) in leme 5.1 je naslednji izrek.

Izrek 5.2 *Sylvestrova enačba (5.1) ima enolično rešitev natanko tedaj, ko velja*

$$\sigma(A) \cap \sigma(-B) = \emptyset,$$

to je takrat, ko matriki A in $-B$ nimata nobene skupne lastne vrednosti.

Dokaz. Po lemi 5.1 so lastne vrednosti matrike $I_n \otimes A + B^T \otimes I_m$ enake $\lambda_i + \mu_j$, kjer sta λ_i in μ_j po vrsti lastni vrednosti matrik A in B za $i = 1, \ldots, m$ in $j = 1, \ldots, n$. Zaradi tega je sistem (5.6) enolično rešljiv natanko tedaj, ko matriki A in $-B$ nimata nobene skupne lastne vrednosti. ■

Posledica 5.3 *Če je matrika A stabilna, je enačba Ljapunova enolično rešljiva.*

Dokaz. Po izreku 5.2 je enačba Ljapunova enolično rešljiva takrat, ko matriki A in $-A^T$ nimata nobene skupne lastne vrednosti. Ker so lastne vrednosti A^T enake lastnim vrednostim A, to pomeni, da mora za poljubni lastni vrednosti λ in μ matrike A veljati $\lambda + \mu \neq 0$. To pa je res, ko je A stabilna matrika, saj takrat velja $\mathrm{Re}(\lambda + \mu) = \mathrm{Re}(\lambda) + \mathrm{Re}(\mu) < 0$. ■

Podobne ugotovitve lahko naredimo tudi za diskretne sisteme. Diskretna enačba Ljapunova $A^TXA - X = -Q$ je poseben primer *diskretne Sylvestrove enačbe*

$$AXB - X = C,$$

kjer so dane matrike $A \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{n\times n}$ in $C \in \mathbb{R}^{m\times n}$, iščemo pa $X \in \mathbb{R}^{m\times n}$. Diskretna Sylvestrova enačba $AXB - X = C$ je ekvivalentna linearnemu sistemu

$$(B^T \otimes A - I_n \otimes I_m)\mathrm{vec}(X) = \mathrm{vec}(C).$$

Izrek 5.4 *Diskretna Sylvestrova enačba $AXB - X = C$ ima enolično rešitev natanko tedaj, ko za vse lastne vrednosti λ matrike A in μ matrike B velja $\lambda\mu \neq 1$.*

Dokaz. Po lemi 5.1 so lastne vrednosti matrike $B^T \otimes A - I_n \otimes I_m$ enake $\lambda_i\mu_j - 1$, kjer sta λ_i in μ_j po vrsti lastni vrednosti matrik A in B za $i = 1, \ldots, m$ in $j = 1, \ldots, n$. ■

Posledica 5.5 *Če je A konvergentna, je diskretna enačba Ljapunova enolično rešljiva.*

Dokaz. Ker so lastne vrednosti A^T in A enake, mora za poljubni lastni vrednosti λ in μ matrike A veljati $\lambda\mu \neq 1$. Če je A konvergentna matrika, velja $|\lambda\mu| = |\lambda| \cdot |\mu| < 1$. ■

5.2 Občutljivost Sylvestrove enačbe

Za kvadratni matriki $A \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{n\times n}$ lahko definiramo *ločenost matrik* kot

$$\text{sep}(A, B) = \min_{X\neq 0} \frac{\|AX - XB\|_F}{\|X\|_F}.$$

V nadaljevanju bomo pokazali, da je ločenost matrik A in B povezana z občutljivostjo Sylvestrove enačbe $AX + BX = C$.

Kot nam pove naslednji izrek, je ločenost $\text{sep}(A, B)$ tudi merilo za občutljivost invariantnega podprostora.

Izrek 5.6 ([5, Izrek 7.2.4]) *Naj bo $U^*AU = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}$ Schurov razcep A in $U = [\,U_1 \quad U_2\,]$, torej stolpci U_1 razpenjajo invariantni podprostor matrike A. Matriko A zmotimo za E, kjer je $U^*EU = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}$. Če je*

$$\text{sep}(T_{11}, T_{22}) > 0$$

in

$$\|E\|_2 \left(1 + \frac{5\|T_{12}\|_2}{\text{sep}(T_{11}, T_{22})}\right) \leq \frac{\text{sep}(T_{11}, T_{22})}{5},$$

potem obstaja taka matrika P, ki zadošča

$$\|P\|_2 \leq \frac{4\|E_{21}\|_2}{\text{sep}(T_{11}, T_{22})}\|E_{21}\|_2$$

in so stolpci $\widetilde{U}_1 = (U_1 + U_2P)(I + P^HP)^{-1/2}$ ortogonalna baza za invariantni podprostor $A + E$.

Vemo, da je Sylvestrova enačba $AX + XB = C$ ekvivalentna linearnemu sistemu

$$(I_n \otimes A + B^T \otimes I_m)\text{vec}(X) = \text{vec}(C).$$

Označimo $P := I_n \otimes A + B^T \otimes I_m$. Za matriko P velja

$$\|P^{-1}\|_2^{-1} = \min_{z\neq 0} \frac{\|Pz\|_2}{\|z\|_2} = \min_{X\neq 0} \frac{\|AX + XB\|_F}{\|X\|_F} = \text{sep}(A, -B),$$

torej

$$\|P^{-1}\|_2 = \frac{1}{\mathrm{sep}(A,-B)}.$$

Poglejmo, kako je Sylvestrova enačba $AX + BX = C$ občutljiva na motnje. Če zmotimo A, B in C, se zmoti tudi X in dobimo

$$(A + \Delta A)(X + \Delta X) + (X + \Delta X)(B + \Delta B) = C + \Delta C.$$

V tenzorskem produktu to ustreza

$$(P + \Delta P)\mathrm{vec}(X + \Delta X) = \mathrm{vec}(C + \Delta C),$$

kjer je $\Delta P = I_n \otimes \Delta A + \Delta B^T \otimes I_m$.

Iz teorije zaokrožitvenih napak za nesingularni linearni sistem $Tz = y$ poznamo oceno

$$\frac{\|\delta z\|}{\|z\|} \le \frac{\|T^{-1}\|}{1 - \|T^{-1}\|\|\delta T\|}\left(\|\delta T\| + \frac{\|\delta y\|}{\|z\|}\right) \tag{5.7}$$

za rešitev $(T + \delta T)(z + \delta z) = (y + \delta y)$, ki velja v primeru $\|T^{-1}\|\|\delta T\| < 1$.

Predpostavimo, da v našem primeru velja $\|P^{-1}\|_2\|\Delta P\|_2 < 1$. Potem lahko uporabimo (5.7) in dobimo

$$\frac{\|\mathrm{vec}(\Delta X)\|_2}{\|\mathrm{vec}(X)\|_2} \le \frac{\|P^{-1}\|_2}{1 - \|P^{-1}\|_2\|\Delta P\|_2}\left(\|\Delta P\|_2 + \frac{\|\mathrm{vec}(\Delta C)\|_2}{\|\mathrm{vec}(X)\|_2}\right). \tag{5.8}$$

Če velja $\|\Delta A\|_F \le \|A\|_F\,\varepsilon$, $\|\Delta B\|_F \le \|B\|_F\,\varepsilon$, $\|\Delta C\|_F \le \|C\|_F\,\varepsilon$ in

$$\delta := \frac{\|A\|_F + \|B\|_F}{\mathrm{sep}(A,-B)}\,\varepsilon < 1,$$

potem iz ocen

$$\begin{aligned}
\|P\|_2 &\le \|A\|_F + \|B\|_F,\\
\|\Delta P\|_2 &\le \|\Delta A\|_F + \|\Delta B\|_F,\\
\|\mathrm{vec}(C)\|_2 &\le \|C\|_F \le (\|A\|_F + \|B\|_F)\|X\|_F,\\
\|\mathrm{vec}(\Delta C)\|_2 &= \|\Delta C\|_F.
\end{aligned}$$

dobimo

$$\begin{aligned}
\frac{\|\Delta X\|_F}{\|X\|_F} &\le \frac{1}{(1-\delta)\mathrm{sep}(A,-B)}\left(\varepsilon\|A\|_F + \varepsilon\|B\|_F + \varepsilon\frac{\|C\|_F}{\|X\|_F}\right)\\
&\le \frac{1}{(1-\delta)\mathrm{sep}(A,-B)}\left(\|A\|_F + \|B\|_F\right)\varepsilon.
\end{aligned}$$

Če vzamemo $\delta < \frac{1}{2}$, smo dokazali naslednji izrek.

Izrek 5.7 *Sylvestrova enačba $AX + XB = C$ naj ima enolično rešitev X za $C \neq 0$. Če velja*

$$\frac{\|A\|_F + \|B\|_F}{\mathrm{sep}(A,-B)}\,\varepsilon < \frac{1}{2},$$

potem za rešitev zmotene Sylvestrove enačbe $(A+\Delta A)(X+\Delta X)+(X+\Delta X)(B+\Delta B) = C+\Delta C$, *kjer je*

$$\varepsilon = \max\left\{\frac{\|\Delta A\|_F}{\|A\|_F}, \frac{\|\Delta B\|_F}{\|B\|_F}, \frac{\|\Delta C\|_F}{\|C\|_F}\right\},$$

velja

$$\frac{\|\Delta X\|_F}{\|X\|_F} \leq 4\,\frac{\|A\|_F + \|B\|_F}{\mathrm{sep}(A,-B)}\,\varepsilon.$$

Občutljivost Sylvestrove enačbe je torej enaka

$$\frac{\|A\|_F + \|B\|_F}{\mathrm{sep}(A,-B)}.$$

Če izrek 5.7 uporabimo za enačbo Ljapunova $A^TX + XA = -Q$, ugotovimo, da je občutljivost odvisna od

$$\frac{\|A\|_F}{\mathrm{sep}(A^T,-A)}.$$

Če velja

$$\varepsilon = \max\left\{\frac{\|\Delta A\|_F}{\|A\|_F}, \frac{\|\Delta Q\|_F}{\|Q\|_F}\right\}$$

in je izpolnjen pogoj

$$\frac{\|A\|_F}{\mathrm{sep}(A^T,-A)}\,\varepsilon < \frac{1}{4},$$

potem za rešitev $(A+\Delta A)^T(X+\Delta X)+(X+\Delta X)(A+\Delta A) = -(Q+\Delta Q)$ velja

$$\frac{\|\Delta X\|_F}{\|X\|_F} \leq 8\,\frac{\|A\|_F}{\mathrm{sep}(A^T,-A)}\,\varepsilon.$$

V oceni iz izreka 5.7 nismo upoštevali, da ima P posebno strukturo, temveč smo se sklicali na teorijo, ki velja za poljubno polno matriko. Lahko pa dobimo še boljšo oceno. Če vzamemo $(A+\Delta A)(X+\Delta X)+(X+\Delta X)(B+\Delta B) = C+\Delta C$ in zanemarimo kvadratne popravke, dobimo novo Sylvestrovo enačbo

$$A\Delta X + \Delta X B = \Delta C - \Delta A X - X\Delta B,$$

ki jo lahko zapišemo v obliki

$$P\,\mathrm{vec}(\Delta X) = -\begin{bmatrix} X^T\otimes I_m & I_n\otimes X & -I_{mn}\end{bmatrix}\begin{bmatrix}\mathrm{vec}(\Delta A)\\ \mathrm{vec}(\Delta B)\\ \mathrm{vec}(\Delta C)\end{bmatrix}.$$

Naj bo

$$\varepsilon = \max\left(\frac{\|\Delta A\|_F}{\alpha}, \frac{\|\Delta B\|_F}{\beta}, \frac{\|\Delta C\|_F}{\gamma}\right),$$

kjer so α, β in γ tolerance (ponavadi $\alpha = \|A\|_F$, $\beta = \|B\|_F$, $\gamma = \|C\|_F$). Dobimo

$$\frac{\|\Delta X\|_F}{\|X\|_F} \leq \sqrt{3}\psi\varepsilon,$$

kjer je

$$\psi = \frac{\|P^{-1}\,[\,\alpha(X^T \otimes I_m) \quad \beta(I_n \otimes X) \quad -\gamma I_{mn}\,]\,\|_2}{\|X\|_F} \tag{5.9}$$

natančnejše pogojenostno število Sylvestrove enačbe.

Zgled 5.1 *Če vzamemo Sylvestrovo enačbo z matrikami*

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} -0.9888 & 0 & 0 \\ 0 & -0.9777 & 0 \\ 0 & 0 & -0.9666 \end{bmatrix}, \quad C = \begin{bmatrix} 0.0112 & 1.0112 & 2.0112 \\ 0.0223 & 1.0223 & 2.0223 \\ 0.0334 & 1.0334 & 2.0334 \end{bmatrix},$$

potem je točna rešitev

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Če spremenimo a_{11} *v* $1 - 10^{-6}$*, se rešitev spremeni v*

$$\widehat{X} = \begin{bmatrix} 1.0001 & 0.9920 & 1.7039 \\ 1.0000 & 0.9980 & 1.0882 \\ 1.0000 & 0.9991 & 1.0259 \end{bmatrix}.$$

Vidimo, da je majhna relativna motnja v matriki A *povzročila veliko motnjo v rešitvi* X*. Izračun nam da*

$$\|\Delta A\|_F / \|A\|_F = 2.4 \cdot 10^{-1}$$

in

$$\|\Delta X\|_F / \|X\|_F = 4.1 \cdot 10^{-7}.$$

Velika sprememba je v skladu z izrekom 5.7 posledica majhne ločenosti matrik A *in* B*, saj velja*

$$\mathrm{sep}(A, -B) = 1.4 \cdot 10^{-6}.$$

□

Ločenost matrik $\mathrm{sep}(A, -B)$ ni direktno povezana z razdaljami med spektrom matrik A in B. V zgledu 5.1 je tako $\mathrm{sep}(A, -B)$ močno manjša od minimalne razdalje med lastnimi vrednosti A in $-B$.

V nekaterih primerih lahko že vnaprej iz lastnosti matrik A in B sklepamo, da bo Sylvestrova enačba slabo pogojena.

Izrek 5.8 *Sylvestrova enačba* $XA + XB = C$ *je zelo občutljiva, če sta obe matriki* A *in* B *zelo občutljivi.*

Obratno ni nujno res in lahko imamo zelo občutljivo Sylvestrovo enačbo, a zelo dobro pogojeni matriki A in B. To se zgodi npr. tudi v zgledu 5.1.

Posledica 5.9 *Če je matrika* A *skoraj singularna, je enačba Ljapunova zelo občutljiva.*

Računanje $\|P^{-1}\|_2 = \mathrm{sep}(A,-B)^{-1}$ je zahtevno, saj je dimenzija matrike P lahko zelo velika.

Na voljo so cenilke, s katerimi lahko ocenimo $\mathrm{sep}(A,-B)$ ceneje kot pa če bi računali najmanjšo singularno vrednost matrike P. Uporabljajo se podobne ocene kot jih imamo za ocenjevanje $\|A^{-1}\|$. Poskušamo poiskati tak z, da bo $\|y\|/\|z\|$, kjer je $Py = z$, čim boljša ocena za $\sigma_{\min}(P)$. V vsakem koraku cenilke moramo rešiti eno Sylvestrovo enačbo. Več podrobnosti je na voljo v [2].

Natančnejšo oceno (5.9) lahko izpeljemo tudi za enačbo Ljapunova. Naj bo $\|\Delta A\|_F \le \alpha\epsilon$, $\Delta Q = \Delta Q^T$ in $\|\Delta Q\|_F \le \gamma\epsilon$. Za enačbo Ljapunova $A^TX + XA = -Q$ potem dobimo

$$\frac{\|\Delta X\|_F}{\|X\|_F} \le \sqrt{3}\psi\varepsilon, \quad \text{kjer je} \quad \psi = \frac{\|P^{-1}\,[\,\alpha((X^T\otimes I_m) + (I_n\otimes X)\Pi^T) \quad -\gamma I_{n^2}\,]\,\|_2}{\|X\|_F}$$

za

$$P = I_n \otimes A^T + A^T \otimes I_n$$

in permutacijo

$$\Pi = \sum_{i,j=1}^{n} (e_i e_j^T) \otimes (e_j e_i^T),$$

za katero velja

$$\mathrm{vec}(A^T) = \Pi\,\mathrm{vec}(A).$$

Občutljivost enačbe Ljapunova, kjer je A stabilna matrika, je povezana z 2-normo simetrične pozitivno definitne rešitve enačbe $A^TX + XA = -I$.

Izrek 5.10 *Naj bo A stabilna matrika in X rešitev $A^TX + XA = -Q$. Če je*

$$(A+\Delta A)^T(X+\Delta X) + (X+\Delta X)(A+\Delta A) = -(Q+\Delta Q), \tag{5.10}$$

potem velja

$$\frac{\|\Delta X\|}{\|X+\Delta X\|} \le 2\|A+\Delta A\|\cdot\|H\|\cdot\left(\frac{\|\Delta A\|}{\|A+\Delta A\|} + \frac{\|\Delta Q\|}{\|Q+\Delta Q\|}\right), \tag{5.11}$$

kjer je H rešitev enačbe Ljapunova $A^TH + HA = -I$.

Dokaz. Ker je matrika A stabilna, iz enačbe (4.3) sledi, da lahko zapišemo

$$H = \int_0^\infty e^{A^Tt} e^{At}\,dt.$$

Iz zveze (5.10) lahko izrazimo

$$A^T\Delta X + \Delta X A = -(\Delta Q + \Delta A^T(X+\Delta X) + (X+\Delta X)\Delta A).$$

To je enačba Ljapunova za ΔX in spet, ker je A stabilna, velja

$$\Delta X = \int_0^\infty e^{A^Tt}(\Delta Q + \Delta A^T(X+\Delta X) + (X+\Delta X)\Delta A)e^{At}\,dt.$$

Naj bosta u in v levi in desni singularni vektor matrike ΔX, ki pripadata njeni največji singularni vrednosti. Potem velja

$$\begin{aligned}
\|\Delta X\| &= |u^*\Delta X v| = \int_0^\infty |u^* e^{A^T t}(\Delta Q + \Delta A^T(X+\Delta X) + (X+\Delta X)\Delta A)e^{At}v|dt \\
&\le \|\Delta Q + \Delta A^T(X+\Delta X) + (X+\Delta X)\Delta A\| \int_0^\infty \|e^{At}u\|\|e^{At}v\|dt \\
&\le (\|\Delta Q\| + 2\|\Delta A\|\|X+\Delta X\|) \int_0^\infty \|e^{At}u\|\|e^{At}v\|dt \\
&\le (\|\Delta Q\| + 2\|\Delta A\|\|X+\Delta X\|) \left(\int_0^\infty \|e^{At}u\|^2 dt\right)^{1/2} \left(\int_0^\infty \|e^{At}v\|^2 dt\right)^{1/2},
\end{aligned}$$

kjer smo za zadnjo neenakost uporabili Cauchy-Schwarzevo neenakost. Zaradi

$$\begin{aligned}
\int_0^\infty \|e^{At}u\|^2 dt &= \int_0^\infty u^* e^{A^T t} e^{At} u \, dt \\
&= u^* \left(\int_0^\infty e^{A^T t} e^{At}\, dt\right) u = u^* H u \le \|H\|
\end{aligned}$$

lahko $\|\Delta X\|$ končno ocenimo z

$$\|\Delta X\| \le (\|\Delta Q\| + 2\|\Delta A\|\|X+\Delta X\|)\|H\|.$$

Sedaj upoštevamo še oceno

$$\|Q+\Delta Q\| \le 2\|A+\Delta A\|\|X+\Delta X\|,$$

ki sledi iz (5.10), in dobimo (5.11). ∎

Posledica 5.11 *Naj bo* $\|\Delta A\| \le \|A\|\,\varepsilon$, $\|\Delta Q\| \le \|Q\|\,\varepsilon$ *in*

$$8\,\varepsilon\|A\| \cdot \|H\| \le \frac{1-\varepsilon}{1+\varepsilon}.$$

Potem je

$$\frac{\|\Delta X\|}{\|X\|} \le 8\,\varepsilon\|A\| \cdot \|H\| + \mathcal{O}(\varepsilon^2).$$

Zgled 5.2

$$A = \begin{bmatrix} -1 & 2 & 3 \\ 0 & -0.0001 & 3 \\ 0 & 0 & -3 \end{bmatrix}, \quad C = \begin{bmatrix} -2 & 0.9999 & 2 \\ 0.9999 & 3.9998 & 4.9999 \\ 2 & 4.9999 & 6 \end{bmatrix}.$$

Točna rešitev je

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Rešitev enačbe Ljapunova $A^T H + HA = -I$ *je*

$$H = 10^4 \cdot \begin{bmatrix} 0.0001 & 0.0001 & 0.0001 \\ 0.0001 & 2.4998 & 2.4999 \\ 0.0001 & 2.4999 & 2.5000 \end{bmatrix}.$$

Če spremenimo a_{11} *v* $-1+10^{-7}$*, dobimo*

$$\widehat{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1.006 & 1.006 \\ 1 & 1.006 & 1.006 \end{bmatrix},$$

$\|\Delta X\|/\|X\| = 4 \cdot 10^{-3}$, $\|\Delta A\|/\|A\| = 1.9 \cdot 10^{-8}$, $\|H\| = 5.00 \cdot 10^{4}$, $\mathrm{sep}(A^T, -A) = 2.00 \cdot 10^{-5}$.

Pri diskretni enačbi Ljapunova $A^T X A - X = -Q$ nastopa količina

$$\mathrm{sep}_d(A^T, A) = \min_{X \neq 0} \frac{\|A^T X A - X\|_F}{\|X\|_F} = \sigma_{\mathsf{min}}(A^T \otimes A^T - I_{n^2}).$$

Če je A konvergentna imajo vse lastne vrednosti matrike A absolutne vrednosti strogo pod 1 in velja $0 < \mathrm{sep}_d(A^T, A) \leq 1$.

Izrek 5.12 *Naj bo* $X + \Delta X$ *rešitev zmotene diskretne enačbe Ljapunova z* $A + \Delta A$ *in* $Q + \Delta Q$*, kjer je* $\|\Delta A\|_F \leq \|A\|_F \, \varepsilon$*,* $\|\Delta Q\|_F \leq \|Q\|_F \, \varepsilon$ *in*

$$\varepsilon \frac{\|A\|_F^2}{\mathrm{sep}_d(A^T, A)} < \frac{1}{4}.$$

Potem je

$$\frac{\|\Delta X\|_F}{\|X\|_F} \leq 2\,\varepsilon \frac{3\|A\|_F^2 + 1}{\mathrm{sep}_d(A^T, A)}.$$

Izrek 5.13 *Če je* A *konvergentna matrika in je* H *rešitev diskretne enačbe Ljapunova* $A^T H A - H = -I$*, potem je*

$$\mathrm{sep}_d(A^T, A) \geq \frac{\sqrt{n}}{\|H\|_2}.$$

5.3 Algoritmi za Sylvestrovo enačbo in enačbo Ljapunova

Prvi možni način je uporaba sistema $(I_n \otimes A + B^T \otimes I_m)\mathrm{vec}(X) = \mathrm{vec}(C)$. V tem primeru (če ne upoštevamo posebne strukture) potrebujemo $\mathcal{O}(m^3 n^3)$ operacij in $\mathcal{O}(m^2 n^2)$ prostora.

A in B lahko transformiramo v lepšo obliko. Če sta U in V nesingularni matriki in

$$R = U^{-1} A U, \qquad S = V^{-1} B V, \qquad D = V^{-1} C U,$$

se $XA + BX = C$ spremeni v $YR + SY = D$, kjer je $Y = V^{-1} X U$.

Denimo, da lahko A in B diagonaliziramo. Potem sta R in S diagonalni in velja

$$y_{ij} = \frac{d_{ij}}{r_{jj} + s_{ii}} \quad \text{za} \quad i = 1, \ldots, m, \; j = 1, \ldots, n.$$

To pride v poštev le, ko sta npr. A in B simetrični matriki, sicer pa imamo lahko velike težave, če U in V nista ortogonalni.

Da ohranimo stabilnost se omejimo na ortogonalne transformacije, saj so matrike realne.

5.3.1 Bartels-Stewartov algoritem za Sylvestrovo enačbo

Rešujemo Sylvestrovo enačbo $AX + BX = C$. Pri tem algoritmu A reduciramo na spodnjo realno Schurovo formo, B pa na zgornjo realno Schurovo formo:

$$R = U^T AU, \qquad S = V^T BV,$$

kjer sta U in V ortogonalni, R kvazi spodnja trikotna in S kvazi zgornja trikotna. Transformirana enačba je $RY + YS = D$, kjer sta $D = U^T CV$ in $Y = U^T XV$.

V bločni obliki lahko zapišemo R in S kot

$$R = \begin{bmatrix} R_{11} & & & \\ R_{21} & R_{22} & & \\ \vdots & & \ddots & \\ R_{p1} & \cdots & \cdots & R_{pp} \end{bmatrix}, \qquad S = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1q} \\ & S_{22} & & \vdots \\ & & \ddots & \vdots \\ & & & S_{qq} \end{bmatrix},$$

kjer so diagonalni bloki velikosti 1×1 in 2×2. Če skladno bločno zapišemo D in Y v obliki

$$D = \begin{bmatrix} D_{11} & \cdots & D_{1q} \\ \vdots & & \vdots \\ D_{p1} & \cdots & Dpq \end{bmatrix}, \qquad Y = \begin{bmatrix} Y_{11} & \cdots & Y_{1q} \\ \vdots & & \vdots \\ Y_{p1} & \cdots & Ypq \end{bmatrix},$$

dobimo enačbe

$$R_{kk}Y_{kl} + Y_{kl}S_{ll} = D_{kl} - \sum_{j=1}^{k-1} R_{kj}Y_{jl} - \sum_{i=1}^{l-1} Y_{kl}S_{il} \tag{5.12}$$

za $k = 1, \ldots, p$ in $l = 1, \ldots, q$.

Enačba (5.12) je Sylvestrova enačba z matrikami velikosti 1×1 ali 2×2. Če želimo iz (5.12) izračunati Y_{kl}, potem moramo poznati bloke $Y_{1l}, \ldots, Y_{k-1,l}$ in $Y_{k1}, \ldots, Y_{k,l-1}$. Za to lahko poskrbimo, če bloke računamo v pravilnem vrstnem redu, npr. po vrsticah ali po stolpcih.

Algoritem 5.1 Bartels-Stewartov algoritem za reševanje Sylvestrove enačbe $AX + BX = C$

$R = U^T AU$, R spodnja realna Schurova forma, $U^T U = I$
$S = V^T BV$, S zgornja realna Schurova forma, $V^T V = I$
$D = U^T CV$
$l = 1, \ldots, q$
  $k = 1, \ldots, p$
    $\widetilde{D}_{kl} = D_{kl} - \sum_{j=1}^{k-1} R_{kj}Y_{jl} - \sum_{i=1}^{l-1} Y_{kl}S_{il}$
    reši $R_{kk}Y_{kl} + Y_{kl}S_{ll} = \widetilde{D}_{kl}$ za Y_{kl}
$X = UYV^T$

Znotraj algoritma rešujemo sisteme $R_{kk}Y_{kl} + Y_{kl}S_{ll} = \widetilde{D}_{kl}$, kjer sta R_{kk} in S_{ll} velikosti 1×1 ali 2×2. Maksimalno imamo tako opravka s sistemom velikosti 4×4, ki ima obliko

$$\begin{bmatrix} r_{11} + s_{11} & r_{12} & s_{21} & 0 \\ r_{21} & r_{22} + s_{11} & 0 & s_{21} \\ s_{12} & 0 & r_{11} + s_{22} & r_{12} \\ 0 & s_{12} & r_{21} & r_{22} + s_{22} \end{bmatrix} \begin{bmatrix} y_{11} \\ y_{21} \\ y_{12} \\ y_{22} \end{bmatrix} = \begin{bmatrix} \widetilde{d}_{11} \\ \widetilde{d}_{21} \\ \widetilde{d}_{12} \\ \widetilde{d}_{22} \end{bmatrix}.$$

Število operacij:

1. redukcija na Schurovi obliki: $26m^3 + 26n^3$,
2. izračun $D = U^T CV$: $2m^2n + 2mn^2$,
3. izračun Y: $m^2n + mn^2$,
4. izračun $X = UYV^T$: $2m^2n + 2mn^2$.

Skupaj: $26(m^3 + n^3) + 5(mn^2 + mn^2)$.

Poraba pomnilnika: $2n^2 + 2m^2 + mn$.

Izrek 5.14 *Če je*

$$\frac{\|A\|_F + \|B\|_F}{\mathrm{sep}(A, -B)}\, u \leq \frac{1}{4},$$

potem za izračunani $\widehat{X}$ *velja*

$$\frac{\|\widehat{X} - X\|_F}{\|X\|_F} \leq 8\frac{\|A\|_F + \|B\|_F}{\mathrm{sep}(A, -B)}\, u.$$

5.3.2 Bartels-Stewartov algoritem za enačbo Ljapunova

Prejšnji algoritem lahko uporabimo tudi za enačbo Ljapunova $A^T X + XA = C$. V primeru $C = C^T$ pa lahko dodatno izkoristimo dejstvo, da je tudi rešitev X simetrična. Tako dobimo še ekonomičnejši algoritem.

Matriko A najprej reduciramo na realno Schurovo formo $R = U^T AU$, kjer je U ortogonalna in R kvazi zgornja trikotna. Potem izračunamo $D = U^T CU$, kjer z upoštevanjem simetrije spet lahko prihranimo pri številu operacij. Če označimo še $Y = U^T XU$, rešujemo sedaj enačbo $R^T Y + YR = D$.

Naj bo

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}, \quad D = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}, \quad Y = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix},$$

kjer je R_{11} matrika 1×1 ali 2×2, velikosti blokov v matrikah D in Y pa so skladne z velikostmi blokov v R. Dobimo:

$$R_{11}^T Y_{11} + Y_{11} R_{11} = D_{11},$$

in

$$R_{22}^T Y_{22} + Y_{22} R_{22}^T = D_{22} - R_{12}^T Y_{21} - Y_{21} R_{12},$$

kar pomeni, da se problem, ko izračunamo Y_{11} in $Y_{12} = Y_{21}^T$, reducira na manjši problem. Pri tem omenimo, da Y_{12} izračunamo s podobnim postopkom kot računamo izvendiagonalne bloke v algoritmu 5.1.

Algoritem 5.2 porabi $26n^3 + 7n^3$ operacij in $3n^2$ pomnilnika.

Algoritem 5.2 Bartels-Stewartov algoritem za reševanje enačbe Ljapunova $A^TX + XA = C$

$R = U^TAU$, R zgornja realna Schurova forma, $U^TU = I$
$D = U^TCU$
$l = 1, \ldots, p$
  $k = l, \ldots, p$
    $D_{kl} = D_{kl} - \sum_{i=l}^{k-1} R_{ik}^T Y_{il}$
    reši $R_{kk}^T Y_{kl} + Y_{kl} R_{ll} = \widetilde{D}_{kl}$ za Y_{kl}
  $j = l+1, \ldots, p$
    $Y_{lj} = Y_{jl}^T$
    $i = j, \ldots, p$
      $D_{ij} = D_{ij} - Y_{il}^T R_{lj} - R_{li}^T Y_{lj}$
      $D_{ji} = D_{ij}^T$
$X = UYU^T$

5.3.3 Reševanje Sylvestrove enačbe preko Hessenberg-Schurove oblike

Pri reševanju Sylvestrove enačbe $AX + XB = C$ lahko prihranimo, če namesto v Schurovo obliko večjo izmed matrik reduciramo le v Hessenbergovo obliko. Denimo, da je $m > n$. Potem A reduciramo v zgornjo Hessenbergovo obliko, B pa v Schurovo:

$$H = U^TAU, \qquad S = V^TBV.$$

V primeru $m < n$ delamo na transponiranem sistemu $B^TX^T + X^TA^T = C^T$.

Dobljeno enačbo $HY + YS = D$ rešujemo po stolpcih. Naj bo

$$D = U^TCV = [\, d_1 \quad d_2 \quad \cdots \quad d_n \,], \qquad Y = U^TXV = [\, y_1 \quad y_2 \quad \cdots \quad y_n \,].$$

Denimo, da smo že izračunali stolpce $y_{k+1}, \ldots, y_n$.

a) če je $s_{k,k-1} = 0$, rešimo zgornji Hessenbergov sistem

$$(H + s_{kk}I)y_k = d_k - \sum_{j=k+1}^{n} s_{kj} y_j,$$

b) če je $s_{k,k-1} \neq 0$, rešimo sistem za stolpca y_{k-1} in y_k hkrati:

$$\begin{bmatrix} H + s_{k-1,k-1}I & s_{k-1,k}I \\ s_{k,k-1}I & H + s_{kk}I \end{bmatrix} \begin{bmatrix} y_{k-1} \\ y_k \end{bmatrix} = \begin{bmatrix} d_{k-1} \\ d_k \end{bmatrix} - \sum_{j=k+1}^{n} \begin{bmatrix} s_{k-1,j} y_j \\ s_{kj} y_j \end{bmatrix}.$$

To je sedaj sistem velikosti $2m \times 2m$, a se ga da preurediti tako, da bo imel v spodnjem trikotniku neničelni le dve glavni poddiagonali, kar pomeni, da ga lahko z Gaussovo eliminacijo z delnim pivotiranjem rešimo v $\mathcal{O}(m^2)$ operacijah.

Število operacij:

1. redukcija na Hessenbergovo in Schurovo obliko: $\frac{10}{3}m^3 + 26n^3$,

2. izračun $D = U^T C V$: $2m^2n + 2mn^2$,
3. izračun Y: $6m^2n + mn^2$,
4. izračun $X = UYV^T$: $2m^2n + 2mn^2$.

Skupaj: $\frac{10}{3}m^3 + 26n^3 + 10m^2n + 5mn^2$. To je dosti ceneje od algoritma 5.1 če je $m \gg n$. Poraba pomnilnika je $2n^2 + 3m^2 + mn$, kar je za m^2 več kot pri algoritmu 5.1, a imamo zato manj operacij.

Numerična stabilnost Hessenberg-Schurovega algoritma je skoraj identična stabilnosti Bartels-Stewartovega algoritma. Velja

Izrek 5.15 *Če je*

$$\frac{\|A\|_F + \|B\|_F}{\mathrm{sep}(A, -B)}\, u \leq \frac{1}{4},$$

potem za izračunani $\widehat{X}$ *velja*

$$\frac{\|\widehat{X} - X\|_F}{\|X\|_F} \leq 9\frac{\|A\|_F + \|B\|_F}{\mathrm{sep}(A, -B)}\, u.$$

Pri enačbi Ljapunova tega algoritma ne moremo uporabiti, saj imamo le eno matriko, to pa reduciramo na Schurovo obliko.

Obstaja tudi Hessenberg-Hessenbergov algoritem, kjer A in B reduciramo na Hessenbergovo obliko, a njegova stabilnost še ni raziskana [2]. Iz tega algoritma lahko potem izpeljemo tudi algoritem za reševanje enačbe Ljapunova, kjer matriko A reduciramo le v Hessenbergovo obliko.

Reševanje diskretne enačbe Ljapunova preko Schurove oblike

Rešujemo $A^T X A - X = C$. Matriko A reduciramo v realno Schurovo formo. Tako se enačba spremeni v $RYR^T - Y = D$, kjer je $D = U^T C U$ in $Y = U^T X U$.

Podobno kot prej enačbo rešujemo po stolpcih. Naj bo

$$D = [\, d_1 \quad d_2 \quad \cdots \quad d_n \,], \qquad Y = [\, y_1 \quad y_2 \quad \cdots \quad y_n \,].$$

Denimo, da smo že izračunali stolpce $y_{k+1}, \ldots, y_n$.

a) če je $r_{k,k-1} = 0$, rešimo zgornji kvazi trikotni sistem

$$(r_{kk}R - I)y_k = d_k - R\sum_{j=k+1}^{n} r_{kj}y_j,$$

b) če je $r_{k,k-1} \neq 0$, rešimo sistem za stolpca y_{k-1} in y_k hkrati:

$$\begin{bmatrix} r_{11}R - I & r_{12}R \\ r_{21}R & r_{22}R - I \end{bmatrix} \begin{bmatrix} y_{k-1} \\ y_k \end{bmatrix} = \begin{bmatrix} d_{k-1} \\ d_k \end{bmatrix} - R\sum_{j=k+1}^{n} \begin{bmatrix} r_{k-1,j}y_j \\ r_{kj}y_j \end{bmatrix}.$$

5.3.4 Reševanje diskretne enačbe Ljapunova s simetričnim C

Rešujemo $A^TXA - X = C$, kjer je $C = C^T$. Matriko A reduciramo v realno Schurovo formo in zapišemo v bločni obliki

$$R = U^TAU = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1p} \\ & R_{22} & & \vdots \\ & & \ddots & \vdots \\ & & & R_{pp} \end{bmatrix},$$

kjer so diagonalni bloki velikosti 1×1 ali 2×2. Skladno bločno zapišemo tudi

$$D = U^TCU = \begin{bmatrix} D_{11} & \cdots & D_{1p} \\ \vdots & & \vdots \\ D_{p1} & \cdots & D_{pp} \end{bmatrix}, \qquad Y = U^TXU = \begin{bmatrix} Y_{11} & \cdots & Y_{1p} \\ \vdots & & \vdots \\ Y_{p1} & \cdots & Y_{pp} \end{bmatrix}.$$

Za Y_{kl} velja enačba $R_{kk}^TY_{kl}R_{ll} - Y_{kl} = \widetilde{D}_{kl}$, kjer je

$$\widetilde{D}_{kl} = D_{kl} - R_{kk}^T \sum_{j=1}^{l-1} Y_{kj}R_{jl} - \sum_{i=1}^{k-1} R_{ik}^T \sum_{j=1}^{l} Y_{ij}R_{jl}.$$

Bloke lahko po vrsti izračunamo z algoritmom 5.3.

Algoritem 5.3 Reševanje diskretne enačbe Ljapunova s simetričnim C

$R = U^TAU$, R zgornja realna Schurova forma, $U^TU = I$
$D = U^TCU$
$l = 1, \ldots, p$
  $k = l, \ldots, p$
    $D_{kl} = D_{kl} - R_{kk}^T \sum_{J=1}^{l-1} Y_{kj}R_{jl} - \sum_{i=1}^{k-1} R_{ik}^T \sum_{j=1}^{l} Y_{ij}R_{jl}$
    reši $R_{kk}^TY_{kl}R_{ll} - Y_{kl} = D_{kl}$ za Y_{kl}
    $Y_{lk} = Y_{kl}^T$

Algoritem 5.3 porabi $26n^3 + 7n^3$ operacij in $3n^2$ pomnilnika.

5.3.5 Hammarlingov algoritem

Imamo enačbo Ljapunova $XA + A^TX = -C^TC$, kjer je $A \in \mathbb{R}^{n\times n}$ stabilna matrika in $C \in \mathbb{R}^{r\times n}$.

Takšna rešitev ima simetrično pozitivno semidefinitno rešitev X, za katero obstaja razcep Choleskega $X = Y^TY$, kjer je Y zgornja trikotna matrika.

Želimo izračunati Y direktno iz C brez ekplicitnega računanja C^TC ali X:

- v številnih aplikacijah v resnici potrebujemo Y in ne X,
- pri ekplicitnem izračunu C^TC lahko pride do izgube natančnosti,
- $\kappa_2(X) = \kappa_2(Y)^2$.

Zgled 5.3 *Pri redukciji modela potrebujemo faktorja Choleskega za vodljivostno in spoznavnostno Gramovo matriko. Zanju vemo, da sta rešitvi enačb Ljapunova*

$$\begin{aligned} AC_G + C_G A^T &= -BB^T, \\ O_G A + A^T O_G &= -C^T C. \end{aligned}$$

Zgled 5.4 *Če vzamemo*

$$A = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \\ 1 & \mu \end{bmatrix},$$

potem za rešitev enačbe Ljapunova $XA + A^T X = -C^T C$ dobimo

$$X = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1+\mu^2 \end{bmatrix}.$$

Če je $|\mu| < u^{1/2}$, kjer je u osnovna zaokrožitvena napaka, dobimo

$$\mathrm{fl}(X) = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Točen faktor Choleskega je

$$Y = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 \\ 1 & \mu \end{bmatrix}.$$

Če je $\mu < 1$, velja $\kappa(X) = 4/\mu^2$ in $\kappa_2(Y) = 2/\mu$.

Če pišemo $X = Y^T Y$, potem želimo iz

$$(Y^T Y)A + A^T (Y^T Y) = -C^T C$$

izračunati Y.

Najprej reduciramo A na realno Schurovo formo $A = USU^T$, kjer je U ortogonalna, S pa kvazi zgornja trikotna. Tako dobimo

$$U^T (Y^T Y) U S + S^T U^T (Y^T Y) U = -U^T C^T C U.$$

Sedaj izračunamo QR razcep matrike CU. Če je $CU = QR$, potem je

$$(CU)^T (CU) = R^T R$$

in imamo razcep Choleskega za $U^T C^T C U$. Če označimo še $Z = YU$, smo problem prevedli na

$$S^T (Z^T Z) + (Z^T Z) S = -R^T R.$$

Matrika Z ni zgornja trikotna, ker pa ni enolična, lahko v nadaljevanju predpostavimo, da ima enako obliko kot S.

Matrike Z, R, S bločno zapišemo kot

$$S = \begin{bmatrix} S_1 & s \\ & \sigma \end{bmatrix}, \qquad R = \begin{bmatrix} R_1 & r \\ & \rho \end{bmatrix}, \qquad Z = \begin{bmatrix} Z_1 & z \\ & \zeta \end{bmatrix},$$

kjer je σ blok velikosti 1×1 ali 2×2. Dobimo naslednje enačbe

$$S_1^T(Z_1^T Z_1) + (Z_1^T Z_1)S_1 = -R_1^T R_1 \tag{5.13}$$

$$S_1^T Z_1^T z + (Z_1^T Z_1)s + Z_1^T z\sigma = -R_1^T r \tag{5.14}$$

$$s^T Z_1^T z + \sigma^T(z^T z + \zeta^T\zeta) + z^T Z_1 s + (z^T z + \zeta^T\zeta)\sigma = -\left(r^T r + \rho^T\rho\right) \tag{5.15}$$

Enačba (5.13) je enačba Ljapunova za vodilne podmatrike S, R, Z. Denimo, da že poznamo rešitev Z_1 iz enačbe (5.13). Potem je Z_1 obrnljiva matrika in (5.14) lahko pomnožimo z Z_1^{-T} z leve. Tako dobimo enačbo

$$(Z_1^{-T} S_1 Z_1^T)z + z\sigma = -Z_1 s - Z_1^{-T} R_1^T r$$

iz katere lahko izračunamo z.

a) Če je σ velikosti 1×1, je z vektor velikosti $n-1$, za katerega rešimo sistem

$$(Z_1^{-T} S_1 Z_1^T + \sigma I)z = -Z_1 s - Z_1^{-T} R_1^T r.$$

b) Če je σ velikosti 2×2, potem ima z dva stolpca velikosti $n-2$. Če označimo

$$\sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}, \qquad z = [\, z_1 \quad z_2 \,], \quad -Z_1 s - Z_1^{-T} R_1^T r = [\, g_1 \quad g_2 \,],$$

potem rešimo sistem

$$\begin{bmatrix} Z_1^{-T} S_1 Z_1^T + \sigma_{11} I & \sigma_{11} I \\ \sigma_{21} I & Z_1^{-T} S_1 Z_1^T + \sigma_{22} I \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}.$$

Iz enačbe (5.15) lahko sedaj izračunamo ζ. Enačbo lahko preuredimo tako, da je razvidno, da gre za enačbo Ljapunova

$$\sigma(\zeta^T\zeta) + (\zeta^T\zeta)\sigma = -\left(r^T r + \rho^T\rho + s^T Z_1^T z + \sigma^T(z^T z) + z^T Z_1 s + (z^T z)\sigma \right)$$

za ζ, ki je velikosti 1×1 ali 2×2.

Sedaj se ne moremo izogniti temu, da najprej izračunamo $\zeta^T\zeta$ in nato iz tega ζ.

S prej navedenim postopkom lahko izračunamo matriko Z, da je

$$S^T(Z^T Z) + (Z^T Z)S = -R^T R.$$

Za $Y = ZU^T$ velja

$$(Y^T Y)A + A^T(Y^T Y) = -C^T C,$$

a ker Y ni zgornja trikotna, moramo rešitev še popraviti.

Za matriko ZU^T izračunamo QR razcep

$$ZU^T = \widetilde{Q}\widetilde{Y},$$

kjer je $\widetilde{Q}$ ortogonalna, $\widetilde{Y}$ pa zgornja trikotna. Ker je $\widetilde{Y} = \widetilde{Q}^T Z U^T$, je

$$\widetilde{Y}^T\widetilde{Y} = UZ^T\widetilde{Q}\widetilde{Q}^T ZU^T = UZ^T ZU = Y^T Y,$$

in $\widetilde{Y}$ je rešitev, ki jo iščemo.

Pri QR razcepu moramo paziti še na to, da so diagonalni elementi $\widetilde{Y}$ pozitivni. Če to ni res, na koncu ustrezne stolpce $\widetilde{Y}$ pomnožimo z -1.

Vse skupaj lahko združimo v algoritmu 5.4. za reševanje $(Y^TY)A + A^T(Y^TY) = -C^TC$.

Algoritem 5.4 Računanje faktorja Choleskega iz enačbe $(Y^TY)A + A^T(Y^TY) = -C^TC$.

---

1) $A = USU^T$, $U^TU = I$, S kvazi zgornja trikotna
2) $CU = QR$, Q z ON stolpci, R zgornja trikotna
3) Vzemi 1. diagonalni blok S_1 (1×1 ali 2×2) in izračunaj Z_1 iz

$$S_1^T(Z_1^TZ_1) + (Z_1^TZ_1)S_1 = -R_1^TR_1.$$

4) Ponavljaj:

4a) $S = \begin{bmatrix} S_1 & s & \times \\ 0 & \sigma & \times \\ 0 & 0 & \times \end{bmatrix}$, $Z = \begin{bmatrix} Z_1 & z & \times \\ 0 & \zeta & \times \\ 0 & 0 & \times \end{bmatrix}$, $R = \begin{bmatrix} R_1 & r & \times \\ 0 & \rho & \times \\ 0 & 0 & \times \end{bmatrix}$,
σ je 1×1 ali 2×2, Z_1 že poznamo.

4b) Izračunaj z iz

$$(Z_1^{-T}S_1Z_1^T)z + z\sigma = -Z_1s - Z_1^{-T}R_1^Tr.$$

4c) Izračunaj ζ iz

$$\sigma(\zeta^T\zeta) + (\zeta^T\zeta)\sigma = -\Big(r^Tr + \rho^T\rho + s^TZ_1^Tz + \sigma^T(z^Tz) + z^TZ_1s + (z^Tz)\sigma\Big).$$

5) $ZU^T = \widetilde{Q}Y$, $\widetilde{Q}$ z ON stolpci, Y zgornja trikotna.

5.3.6 Hammarlingov algoritem za diskretno enačbo Ljapunova

Podobno lahko poiščemo faktor Choleskega tudi v diskretnem primeru, če je na desni simetrična semidefinitna matrika.

Rešujemo $A^TXA - X = -C^TC$. Podobno kot prej

1. pišemo $X = Y^TY$,
2. A reduciramo v realno Schurovo formo $A = USU^T$,
3. QR razcep $CU = QR$.

Dobimo sistem $S^TZ^TZS - Z = -R^TR$.

Bločno zapišemo $S = \begin{bmatrix} S_1 & s \\ & \sigma \end{bmatrix}$, $R = \begin{bmatrix} R_1 & r \\ & \rho \end{bmatrix}$, $Z = \begin{bmatrix} Z_1 & z \\ & \zeta \end{bmatrix}$, kjer je σ blok velikosti 1×1 ali 2×2. Dobimo enačbe

$$S_1^T(Z_1^TZ_1)S_1 - Z_1^TZ_1 \;=\; -R_1^TR_1,$$

$$\begin{aligned} Z_1^T z - S_1^T Z_1^T z\sigma &= S_1^T Z_1^T Z_1 s + R_1^T r, \\ \sigma^T \zeta^T \zeta \sigma - \zeta^T \zeta &= -\left(\rho^T \rho + r^T r + (Z_1 s + z\sigma)^T (Z_1 s + z\sigma) - z^T z\right) \end{aligned}$$

Na koncu spet naredimo QR razcep ZU^T in za Y vzamemo zgornji trikotni faktor.

Poglavje 6

Realizacija in identifikacija

6.1 Uvod

Če poznamo prenosno funkcijo $G(s)$ ali pa matrike (A, B, C, D) iz predstavitve v prostoru stanj, potem lahko iz danega vhoda $u(t)$ izračunamo izhod $y(t)$ in vrednosti spremenljivk stanja $x(t)$.

Obratni problem, s katerim se bomo ukvarjali v tem poglavju je, kako iz danih podatkov, ki jih dobimo iz sistema, razbrati prenosno funkcijo oz. predstavitev v prostoru stanj.

Tokrat se bomo v glavnem ukvarjali z diskretnimi sistemi, tudi zvezne se ponavadi rešuje z vzorčenjem.

Poznamo torej:

- impulzni odziv sistema,
- množico vhodno-izhodnih parov u_i, y_i,
- kovariance izhodnih vektorjev sistema, ki je vzbujen z belim šumom,

iščemo pa prenosno funkcijo oz. matrike (A, B, C, D).

6.2 Realizacija SISO sistema iz impulznega odziva

Denimo, da se da SISO sistem predstaviti s prenosno funkcijo

$$H(z) = \frac{p(z)}{q(z)} = \frac{p_m z^m + p_{m-1} z^{m-1} + \cdots + p_0}{q_n z^n + q_{n-1} z^{n-1} + \cdots + q_0},$$

kjer sta polinoma p in q tuja, $m \leq n$ (to je pogoj za vzročnost) in $q_n \neq 0$.

V tem primeru je n *red sistema*. To ustreza diskretnemu sistemu

$$q_n y_{k+1} + q_{n-1} y_k + \cdots + q_0 y_{k-n} = p_m u_{k+1} + p_{m-1} u_k + \cdots + p_0 u_{k-m}.$$

Po Z-transformaciji dobimo $Y(z) = H(z)U(z)$. Če je $(u_i)_{i=0}^{\infty} = (1, 0, 0, \ldots)$ enotski impulz in je impulzni odziv enak $(y_i)_{i=0}^{\infty} = (h_0, h_1, h_2, \ldots)$, potem so h_i koeficienti razvoja

$$H(z) = h_0 + \frac{h_1}{z} + \frac{h_2}{z^2} + \frac{h_3}{z^3} + \cdots .$$

Koeficiente h_i imenujemo *Markovski koeficienti*.

Pri identifikaciji je naš cilj iz neskončno (v praksi jih seveda poznamo le končno) Markovskih koeficientov h_i določiti koeficiente $p_0, \ldots, p_m$ in $q_0, \ldots, q_n$, ki določajo prenosno funkcijo.

Iz neskončno (v praksi seveda končno) Markovskih koeficientov h_i želimo določiti koeficiente $p_0, \ldots, p_m$ in $q_0, \ldots, q_n$, ki določajo prenosno funkcijo. Iz zveze $H(z)q(z) = p(z)$ s primerjanjem koeficientov pri $z^m, z^{m-1}, \ldots$ dobimo:

$$\left[\begin{array}{cccc} & & & h_0 \\ & & h_0 & h_1 \\ & \iddots & \iddots & \vdots \\ h_0 & h_1 & \cdots & h_n \\ \hline h_1 & h_2 & \cdots & h_{n+1} \\ \vdots & \vdots & \vdots & \vdots \end{array}\right] \left[\begin{array}{c} q_0 \\ q_1 \\ \vdots \\ q_{n-1} \\ q_n \end{array}\right] = \left[\begin{array}{c} p_n \\ p_{n-1} \\ \vdots \\ p_0 \\ \hline 0 \\ \vdots \end{array}\right] . \tag{6.1}$$

Pri tem je $p_n = \cdots = p_{m+1} = 0$, če je $m < n$. Če upoštevamo še normalizacijo $q_n = 1$, potem se nam spodnje enačbe iz (6.1) spremenijo v

$$\left[\begin{array}{cccc} h_0 & h_1 & \cdots & h_n \\ h_1 & h_2 & \cdots & h_{n+1} \\ \vdots & \vdots & \iddots & \vdots \\ h_n & h_{n+1} & \cdots & h_{2n-1} \\ \hline h_{n+1} & h_{n+2} & \cdots & h_{2n} \\ \vdots & \vdots & \vdots & \vdots \end{array}\right] \left[\begin{array}{c} q_0 \\ q_1 \\ \vdots \\ q_{n-1} \end{array}\right] = - \left[\begin{array}{c} h_{n+1} \\ h_{n+2} \\ \vdots \\ h_{2n} \\ \hline h_{2n+1} \\ \vdots \end{array}\right] . \tag{6.2}$$

Sistem je rešljiv, če ima matrika poln rang. Ko poznamo koeficiente $q_0, \ldots, q_{n-1}$, lahko iz prvih $n+1$ enačb (6.1) izračunamo še koeficiente $p_0, \ldots, p_n$.

Pri tem se lahko zgodi, da je v primeru $h_0 = \cdots = h_k = 0$ tudi $p_0 = \cdots = p_k = 0$ (zamik odziva).

Matriko z obliko

$$\left[\begin{array}{cccc} a_1 & a_2 & \cdots & a_n \\ a_2 & a_3 & \cdots & a_{n+1} \\ \vdots & \vdots & & \vdots \\ a_n & a_{n+1} & \cdots & a_{2n-1} \end{array}\right]$$

imenujemo *Hanklova matrika*.

Strukturo Hanklove matrike se da uporabiti za učinkovitejše algoritme za reševanje linearnih sistemov, množenje z matriko, itd.

V nadaljevanju si bomo pomagali z oznako

$$M_{i,j} = \begin{bmatrix} h_1 & h_2 & \cdots & h_j \\ h_2 & h_3 & \cdots & h_{j+1} \\ \vdots & \vdots & & \vdots \\ h_i & h_{i+1} & \cdots & h_{i+j-1} \end{bmatrix}.$$

Izrek 6.1 *Če je red sistema z impulznim odzivom $(h_0, h_1, h_2, \ldots)$ enak n, potem ima matrika*

$$M_{\infty,i} = \left[\begin{array}{cccc} h_1 & h_2 & \cdots & h_n \\ h_2 & h_3 & \cdots & h_{n+1} \\ \vdots & \vdots & \iddots & \vdots \\ h_n & h_{n+1} & \cdots & h_{2n-1} \\ \hline h_{n+1} & h_{n+2} & \cdots & h_{2n} \\ \vdots & \vdots & \vdots & \vdots \end{array}\right]$$

z neskončno vrsticami in i stolpci v primeru $i \geq n$ rang n.

Dokaz. Ker je red sistema n, obstaja neničelna rešitev (6.2). Iz nje sledi, da je vsak j-ti stolpcec $M_{\infty,i}$ za $j > n$ linearna kombinacija prvih n stolpcev, torej rang ne more biti večji kot n. V primeru $i = n$ mora biti rang enak n, saj bi sicer iz (6.2) lahko dobili model nižjega reda. ■

V primeru, ko je red sistema enak n, je Hanklova matrika

$$M_{n,n} = \begin{bmatrix} h_1 & h_2 & \cdots & h_n \\ h_2 & h_3 & \cdots & h_{n+1} \\ \vdots & \vdots & \iddots & \vdots \\ h_n & h_{n+1} & \cdots & h_{2n-1} \end{bmatrix}$$

nesingularna. Od tod iz (6.1) in (6.2) dobimo prvi algoritem

1. Reši $M_{n,n}q = b$, kjer je $b = -\begin{bmatrix} h_{n+1} & \cdots & h_{2n} \end{bmatrix}^T$ in $q = \begin{bmatrix} q_0 & \cdots & q_{n-1} \end{bmatrix}^T$.
2. Izračunaj $p = S \begin{bmatrix} q \\ 1 \end{bmatrix}$, kjer je $p = \begin{bmatrix} p_n & \cdots & p_0 \end{bmatrix}^T$ in

$$S = \begin{bmatrix} & & & h_0 \\ & & h_0 & h_1 \\ & \iddots & \iddots & \vdots \\ h_0 & h_1 & \cdots & h_n \end{bmatrix}.$$

Če imamo na voljo več kot n merjenj impulznega odziva, lahko reševanje linearnega sistema

$$M_{n,n}q = b \tag{6.3}$$

nadomestimo z reševanjem predoločenega sistema

$$M_{N,n}q = b_N, \tag{6.4}$$

kjer je $N > n$, $b_N = -\left[\, h_{n+1} \quad \cdots \quad h_{n+N} \,\right]^T$ in

$$M_{N,n} = \begin{bmatrix} h_1 & h_2 & \cdots & h_n \\ h_2 & h_3 & \cdots & h_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ h_N & h_{N+1} & \cdots & h_{N+n-1} \end{bmatrix}.$$

Eksaktno gledano sta metodi enaki, a ker so pri meritvah vedno prisotne napake, je stabilneje reševati predoločeni sistem. Velikokrat ima predoločeni sistem celo manjše pogojenostno število kot pa linearni sistem.

Analiza zaokrožitvenih napak za linearni sistem (6.3) pravi, da v primeru, ko malo zmotimo matriko $M_{n,n}$ in desno stran b, lahko pričakujemo, da za spremembo rešitve velja

$$\frac{\|\delta Q\|}{\|q\|} \leq \kappa(M_{n,n}) \left(\frac{\|\delta M_{n,n}\|}{\|M_{n,n}\|} + \frac{\|\delta b\|}{\|b\|} \right) c_1,$$

kjer je c_1 konstanta blizu 1. Za predoločeni sistem (6.4) velja

$$\frac{\|\delta Q\|}{\|q\|} \leq \kappa(M_{N,n}) \left(\frac{\|\delta M_{n,n}\|}{\|M_{n,n}\|} + \frac{\|\delta b\|}{\|b\|} \right) \left(c_2 + c_3 \kappa(M_{N,n}) \frac{\|r\|}{\|b\|} \right),$$

kjer sta c_2 in c_3 konstanti blizu 1 in je r ostanek predoločenega sistema $H_{N,n}q - b_N$. Ker pa je desna stran (6.4) v sliki matrike, je v našem primeru ostanek r enak 0 in dobimo oceno, ki je zelo podoba oceni za občutljivost linearnega sistema.

Velja še več, občutljivost matrike je enaka razmerju med največjo in najmanjšo singularno vrednostjo. Ker je $M_{n,n}$ podmatrika $M_{N,n}$, od tod sledi $\sigma_n(M_{n,n}) \leq \sigma_n(M_{N,n})$. Ker se po drugi strani največji singularni vrednosti $M_{n,n}$ in $M_{N,n}$ običajno ne razlikujeta dosti, najmanjši pa se, je lahko linearni sistem (6.3) veliko bolj numerično občutljiv kot pa predoločeni sistem (6.4).

Ponavadi red sistema ni vnaprej znan in ga moramo prav tako določiti iz impulznega odziva. To moramo narediti še preden začnemo računati koeficiente polinomov p in q.

Kot bomo pokazali v nadaljevanju pri MIMO sistemih, je red sistema enak maksimalnemu rangu $M_{k,k}$ za $k = 1, 2, \ldots$. Če je red sistema n, potem je $n = \operatorname{rang}(M_{n,n})$.

Rang sistema torej razberemo iz ranga matrike $M_{N,i}$, kjer sta $N, i \geq n$. Rang lahko ugotovimo npr. z uporabo QR razcepa s pivotiranjem ali pa s pomočjo singularnega razcepa.

Če uporabimo SVD in je $N \gg i$, potem se spača najprej narediti QR razcep

$$M_{N,i} = Q \begin{bmatrix} R \\ 0 \end{bmatrix},$$

potem pa uporabimo SVD na R. Dobimo $R = U\Sigma V^T$, torej je

$$M_{N,i} = Q \begin{bmatrix} U & \\ & I \end{bmatrix} \cdot \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} \cdot V^H.$$

Tako delamo singularni razcep na manjši matriki in prihranimo veliko operacij.

Končni algoritem ima obliko Algoritma 6.1.

Opombe:

Algoritem 6.1 Identifikacija prenosne funkcije iz Markovskih koeficientov

1. Oceni rang n Hanklove matrike $M_{N,i}$ za naraščajoče i iz
$$n = \max\{i : \text{rang}(M_{N,i}) = i\}.$$
2. Reši predoločeni sistem
$$M_{N,n} \cdot q = b_N,$$
kjer je $b_N = -[\, h_{n+1} \quad \cdots \quad h_{n+N} \,]^T$ in $q = [\, q_0 \quad \cdots \quad q_{n-1} \,]^T$.
3. Izračunaj $p = S \begin{bmatrix} q \\ 1 \end{bmatrix}$, kjer je $p = [\, p_n \quad \cdots \quad p_0 \,]^T$ in
$$S = \begin{bmatrix} & & & h_0 \\ & & h_0 & h_1 \\ & \iddots & \iddots & \vdots \\ h_0 & h_1 & \cdots & h_n \end{bmatrix}.$$

- QR razcep v točki 1. lahko učinkovito izračunamo s posodabljanjem razcepa za matriko brez zadnjega stolpca.
- V točki 2. se da izkoristiti Hanklovo strukturo in zmanjšati zahtevnost iz $\mathcal{O}(Nn^2)$ v $\mathcal{O}(Nn)$.

6.3 Realizacija MIMO sistema v prostoru stanj

Imamo MIMO sistem, kjer je na vhodu m parametrov, na izhodu pa r, torej $u_i \in \mathbb{R}^m$ in $y_i \in \mathbb{R}^r$ za vsak i. Prenosna funkcija $G(s)$ je potem matrika velikosti $r \times m$, kjer je vsak element $g_{ij}(z)$ prenosna funkcija, ki predstavlja impulzni odziv i-te komponente izhoda $y(i)$ na enotski impulz v j-ti komponenti vhoda $y(j)$.

Po formulah iz prejšnjega razdelka bi lahko izračunali vsako funkcijo g_{ij} posebej, a to ni ne praktično ne stabilno. Zaradi numeričnega računanja bi se npr. hitro zgodilo, da bi bil največji skupni delitelj imenovalcev g_{ij} enak 1.

Zaradi tega raje direktno računamo matrike A, B, C, D.

Naj bo $G(z)$ prenosna funkcija dimenzije $r \times n$, kjer za vsak element $g_{ij}(z)$ velja, da sta si števec in imenovalec tuja in stopnja števca ni vecja od stopnje imenovalca. Če za matrike (A, B, C, D) velja
$$G(s) = C(zI - A)^{-1}B + D,$$
potem pravimo, da matrike (A, B, C, D) predstavljajo *realizacijo $G(z)$ v prostoru stanj*.

Razvoj prenosne funkcije ima sedaj obliko
$$G(z) = H_0 + H_1 z^{-1} + H_2 z^{-2} + H_3 z^{-3} + \cdots,$$

kjer za Markovske koeficiente prenosne funkcije $G(z)$ velja

$$\begin{aligned} H_0 &= \lim_{s\to\infty} G(s) \\ H_1 &= \lim_{s\to\infty} s(G(s)-H_0) \\ H_2 &= \lim_{s\to\infty} s^2(G(s)-H_0-H_1 s) \\ &\vdots \end{aligned}$$

Če je (A,B,C,D) realizacija $G(z)$, potem velja

$$G(z) = C(zI-A)^{-1}B + D = C(z^{-1}I + z^{-2}A + z^{-3}A + \cdots)B + D,$$

torej $H_0 = D$ in $H_k = CA^{k-1}B$ za $k = 1, 2, \ldots$.

Možnih realizacij je več, zato iščemo še kakšne uporabne lastnosti.

6.3.1 Vodljiva realizacija

Prenosno funkcijo $G(z)$ zapišemo v obliki

$$G(z) = H_0 + \frac{P(z)}{q(z)}, \tag{6.5}$$

kjer je $q(z) = z^n + q_{n-1}z^{n-1} + \cdots + q_0$ polinom stopnje n (najmanjši skupni večkratnik imenovalcev $q_{ij}(z)$) in $P(z) = P_0 + P_1 z + \cdots + P_{n-1}z^{n-1}$.

Če vzamemo A, B, C, D v bločnem zapisu

$$A = \begin{bmatrix} 0 & I_m & & & \\ & 0 & I_m & & \\ & & \ddots & \ddots & \\ & & & 0 & I_m \\ -q_0 I_m & -q_1 I_m & \cdots & -q_{n-2}I_m & -q_{n-1}I_m \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ I_m \end{bmatrix},$$

$$C = [\, P_0 \quad P_1 \quad \cdots \quad P_{n-1} \,], \qquad D = H_0,$$

lahko preverimo, da velja $G(z) = C(zI-A)^{-1}B + D$. Dobljena realizacija je reda mn in je očitno vodljiva.

Pokažimo, da zgornje matrike (A,B,C,D) res predstavljajo realizacijo prenosne funkcije $G(z)$. Ker je $D = H_0$, nam ostane še pokazati, da velja $C(zI-A)^{-1}B = \frac{P(z)}{q(z)}$. Označimo $X = (zI-A)^{-1}$. Z množenjem $(zI-A)X$ lahko hitro preverimo, da ima X obliko

$$X = \frac{1}{q(z)} \begin{bmatrix} I \\ zI \\ \vdots \\ z^{n-1}I \end{bmatrix}.$$

Od tod sledi

$$C(zI-A)^{-1}B = CX = \frac{1}{q(x)}(P_0 + zP_1 + \cdots + z^{n-1}P_{n-1}) = \frac{P(z)}{q(z)}$$

in pokazali smo, da je realizacija pravilna.

6.3.2 Spoznavna realizacija

Prenosno funkcijo $G(z)$ zapišemo v obliki

$$G(z) = H_0 + H_1 z^{-1} + H_2 z^{-2} + H_3 z^{-3} + \cdots,$$

upoštevamo še (6.5) in vzamemo

$$\widetilde{A} = \begin{bmatrix} 0 & I_r & & & \\ & 0 & I_r & & \\ & & \ddots & \ddots & \\ & & & 0 & I_r \\ -q_0 I_r & -q_1 I_r & \cdots & -q_{n-2} I_r & -q_{n-1} I_r \end{bmatrix}, \qquad \widetilde{B} = \begin{bmatrix} H_1 \\ H_2 \\ \vdots \\ H_n \end{bmatrix},$$

$$\widetilde{C} = [\, I_r \quad 0 \quad \cdots \quad 0 \,], \qquad \widetilde{D} = H_0.$$

Spet lahko preverimo, da velja $G(z) = \widetilde{C}(zI - \widetilde{A})^{-1}\widetilde{B} + \widetilde{D}$. Dobljena realizacija je reda rn in je očitno spoznavna.

Pokažimo, da matrike $(\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D})$ res realizirajo prenosno funkcijo $G(z)$. Pokazati moramo, da dobimo Markovske parametre $H_0, H_1, \ldots$. Ker velja $\widetilde{D} = H_0$, moramo preveriti še, da je $\widetilde{C}\widetilde{A}^{i-1}\widetilde{B} = H_i$ za $i = 1, 2, \ldots$. Poglejmo, kako zgleda produkt $\widetilde{A}\widetilde{B}$. Dobimo

$$\widetilde{A}\widetilde{B} = \begin{bmatrix} H_2 \\ \vdots \\ H_n \\ -(q_0 H_1 + q_1 H_2 + \cdots + q_{n-1} H_n) \end{bmatrix}.$$

V zadnjem bloku v zgornjem izrazu lahko matrike H_i nadomestimo z $H_i = CA^{i-1}B$, kjer za matrike A, B, C vzamemo matrike iz vodljive realizacije, za katero smo že prej dokazali, da je dobra. Potem dobimo

$$-(q_0 H_1 + q_1 H_2 + \cdots + q_{n-1} H_n) = -C(q_0 I + q_1 A + \cdots + q_{n-1} A^{n-1})B = CA^n B = H_{n+1},$$

saj je $q(A) = 0$. Od tod z indukcijo lahko pokažemo, da je

$$\widetilde{A}^{i-1}\widetilde{B} = \begin{bmatrix} H_i \\ \vdots \\ H_{i+n-1} \end{bmatrix}$$

in $\widetilde{C}\widetilde{A}^{i-1}\widetilde{B} = H_i$ za $i = 1, 2, \ldots$.

6.3.3 Minimalna realizacija

Realizacij je neomejeno, saj iz vsake lahko z nesingularno transformacijo T dobimo novo.

$$\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right) \longrightarrow \left(\begin{array}{c|c} T^{-1}AT & T^{-1}B \\ \hline CT & D \end{array}\right).$$

Videli smo tudi, da imamo lahko realizacije različnih redov.

Pravimo, da je (A, B, C, D) *minimalna realizacija* prenosne funkcije $G(z)$, če za vsako realizacijo $(\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D})$ iste prenosne funkcije velja, da je velikost matrike $\widetilde{A}$ večja ali enaka velikosti matrike A. Velikost matrike A v minimalni realizaciji je *McMillanova stopnja*.

Izrek 6.2 *Realizacija (A, B, C, D) prenosne funkcije $G(z)$ je minimalna natanko takrat, ko je par (A, B) vodljiv in par (A, C) spoznaven.*

Dokaz. ($\Rightarrow$): Denimo, da par (A, B) ni vodljiv ali pa par (A, C) ni spoznaven. Potem iz Kalmanovega razcepa sledi, da obstaja realizacija še manjše stopnje, ki je vodljiva in spoznavna.

($\Leftarrow$): Denimo, da imamo poleg (A, B, C, D) še minimalno realizacijo $(\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D})$, ki ima stopnjo $\widetilde{n} < n$. Ker imata obe realizaciji enako prenosno funkcijo se morajo ujemati tudi Markovski koeficienti, torej

$$CA^{i-1}B = \widetilde{C}\widetilde{A}^{i-1}\widetilde{B}$$

za $i = 1, 2, \ldots$. Od tod sledi

$$O_M C_M = \widetilde{O}_M \widetilde{C}_M,$$

kjer sta O_M in C_M spoznavnostna in vodljivostna matrika. Ker je sistem vodljiv in spoznaven, je $\operatorname{rang}(O_M) = \operatorname{rang}(C_M) = n$, torej je tudi $\operatorname{rang}(O_M C_M) = n$, po drugi strani pa je $\operatorname{rang}(\widetilde{O}_M \widetilde{C}_M) = \widetilde{n} < n$, kar je protislovje. ∎

Pravimo, da je $A = LR$ *razcep polnega ranga*, če za matriki L in R velja, da imata enak rang in da sta obe polnega ranga. Od tod sledi, da je rang matrike A enak rangu matrike L oziroma R.

Zgled za takšen razcep dobimo npr. iz singularnega razcepa. Če je

$$A = U\Sigma V^T = U\Sigma^{1/2}\Sigma^{1/2}V^T$$

in vzamemo $L = U\Sigma^{1/2}$ in $R = \Sigma^{1/2}V^T$, je LR razcep polnega ranga za matriko A.

Razcep, ki ga dobimo iz singularnega razcepa, je stabilno izračunan.

Lema 6.3 *Naj bosta $A = LR$ in $A = \widetilde{L}\widetilde{R}$ dva razcepa matrike $A \in \mathbb{R}^{m\times r}$, kjer je $m \geq r$ in so $L, R, \widetilde{L}$ in $\widetilde{R}$ polnega ranga. Potem obstaja taka nesingularna matrika T, da je $\widetilde{L} = LT$ in $\widetilde{R} = T^{-1}R$.*

Dokaz. Ker je L polnega ranga je njegov pseudoinverz enak

$$L^+ = (L^T L)^{-1} L^T.$$

Podobno pri $\widetilde{R}$ dobimo

$$\widetilde{R}^+ = \widetilde{R}(\widetilde{R}\widetilde{R}^T)^{-1}.$$

Iz $LR = \widetilde{L}\widetilde{R}$ sledi $L^+\widetilde{L} = R\widetilde{R}^+$ in to vzamemo za T. Potem dobimo

$$LR = \widetilde{L}\widetilde{R}, \quad \Rightarrow \quad L^+LR = L^+\widetilde{L}\widetilde{R}, \quad \Rightarrow R = T\widetilde{R}$$

in

$$LR = \widetilde{L}\widetilde{R}, \quad \Rightarrow \quad LR\widetilde{R}^+ = \widetilde{L}\widetilde{R}\widetilde{R}^+, \quad \Rightarrow \quad LT = \widetilde{L}.$$

∎

Posledica 6.4 *Če sta (A, B, C, D) in $(\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D})$ minimalni realizaciji iste prenosne funkcije, potem obstaja enolična obrnljiva matrika T, da je $\widetilde{A} = T^{-1}AT$, $\widetilde{B} = T^{-1}B$, $\widetilde{C} = CT$ in $\widetilde{D} = D$.*

Velja še $T = O_M^+ \widetilde{O}_M = C_M \widetilde{C}_M^+$.

Dokaz. $M_n = O_M C_M = \widetilde{O}_M \widetilde{C}_M$ je razcep polnega ranga, po lemi 6.3 je potem $T = O_M^+ \widetilde{O}_M = C_M \widetilde{C}_M^+$ nesingularna in $C_M = T\widetilde{C}_M$ in $O_M T = \widetilde{O}_M$.

Iz prve bločne vrstice $C_M = T\widetilde{C}_M$ sledi $\widetilde{C} = CT$, iz prvega bločnega stolpca $O_M T = \widetilde{O}_M$ pa $T\widetilde{B} = B$.

Pokazati moramo še $\widetilde{A} = T^{-1}AT$. Ker gre za realizaciji iste prenosne funkcije, se morajo Markovski koeficienti ujemati, torej $CA^{i-1}B = \widetilde{C}\widetilde{A}^{i-1}\widetilde{B}$ za $i = 1, 2, \ldots$. Potem pa velja tudi

$$O_M A C_M = \widetilde{O}_M \widetilde{A} \widetilde{C}_M,$$

od koder sledi

$$AC_M = O_M^+ \widetilde{O}_M \widetilde{A} \widetilde{C}_M, \quad \Rightarrow \quad AC_M \widetilde{C}_M^+ = O_M^+ \widetilde{O}_M \widetilde{A}, \quad \Rightarrow \quad \widetilde{A} = T^{-1}AT.$$

■

Izrek 6.5 *Za Hanklovo matriko Markovskih koeficientov velja*

1. $\operatorname{rang}(M_k) \leq \operatorname{rang}(M_{k+1})$ *za vsak* k,
2. *če je* (A, B, C, D) *realizacija reda* n, *je* $\operatorname{rang}(M_k) = \operatorname{rang}(M_n)$ *za* $k \geq n$,
3. *če sta* (A, B, C, D) *in* $(\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D})$ *realizaciji reda* n *in* $\widetilde{n}$, *je* $\operatorname{rang}(M_n) = \operatorname{rang}(M_{\widetilde{n}})$,
4. *če je* d *McMillanova stopnja, je* $d = \max_k(\operatorname{rang}(M_k))$
5. *če je* (A, B, C, D) *realizacija reda* n *je*

$$d = \operatorname{rang}(M_n) = \operatorname{rang}(O_K C_M).$$

Dokaz.

1. Očitno.
2. Hanklovo matriko M_k lahko zapišemo kot

$$M_k = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{k-1} \end{bmatrix} \begin{bmatrix} B & AB & \cdots & A^{k-1}B \end{bmatrix}.$$

   V primeru $k \geq n$ za faktorja na desni velja, da sta njuna ranga enaka rangu spoznavnostne matrike O_M oziroma vodljivostne matrike C_M, zato je $\operatorname{rang}(M_k) = \operatorname{rang}(M_n) = \operatorname{rang}(O_M C_M)$.

3. Naj bo $r = \max(n, \widetilde{n})$. Ker imata obe realizaciji enake Markovske koeficiente, mora veljati $\widetilde{M}_r = M_r$. Po točki 2. je potem

$$\operatorname{rang}(M_n) = \operatorname{rang}(M_r) = \operatorname{rang}(\widetilde{M}_r) = \operatorname{rang}(\widetilde{M}_{\widetilde{n}}).$$

4. Denimo, da obstaja minimalna realizacija (A, B, C, D) reda $\widetilde{d} < d$. Potem bi moralo biti $\max_k(\operatorname{rang}(M_k)) = \widetilde{d} < d$, kar je protislovje.

5. Uporabimo točki 3. in 4.

■

Iz vsakega razcepa polnega ranga Hanklove matrike lahko dobimo realizacijo.

Naj bo $M_{i,j} = O_i C_j$ tak razcep, da sta O_i in C_j polnega ranga, pri čemer sta $i, j \geq n$. Sedaj bločno zapišemo

$$O_i = \begin{bmatrix} C \\ O_+ \end{bmatrix} = \begin{bmatrix} O_- \\ CA^{i-1} \end{bmatrix}, \qquad C_j = [\, B \quad C_+ \,] = [\, C_- \quad A^{j-1}B \,].$$

Potem A dobimo iz enačbe

$$O_- A = O_+$$

ali

$$C_+ = AC_-,$$

torej

$$A = O_-^+ O_+ = C_+ C_-^+.$$

Najbolj stabilno lahko razcep polnega ranga in s tem tudi realizacijo dobimo preko singularnega razcepa. Če za razcep uporabimo ekonomični singularni razcep, dobimo

$$M_{i,j} = U \Sigma V^T.$$

Potem vzamemo $O_i = U\Sigma^{1/2}$ in $C_j = \Sigma^{1/2} V^T$. Dobimo

$$O_i^T O_i = C_j^T C_j = \Sigma,$$

kar pomeni, da je ta realizacija *uravnotežena*. To pravimo takrat, kadar za končni Gramovi matriki, vodljivostno in spoznavnostno, velja, da sta enaki in diagonalni.

$$C_G^D(0, j-1) = C_j C_j^T = \sum_{k=1}^{i-1} A^K B B^T (A^T)^k = \Sigma,$$

$$O_G^D(0, j-1) = O_i^T O_j = \sum_{k=1}^{i-1} (A^T)^k C^T C (A^k) = \Sigma.$$

Vidimo, da ima v primeru, ko do minimalne realizacije pridemo preko singularnega razcepa, dobljena realizacija lepe lastnosti.

6.4 Identifikacija iz vhodno-izhodnih parov (SISO primer)

Pogosto nimamo na voljo impulznega odziva, npr. , možno je, da sistem že deluje in ga ni moč zaustaviti in pogledati impulzni odziv. V tem primeru moramo uporabiti tiste podatke, ki jih lahko dobimo, to pa pomeni, poznavanje vzhodno-izhodnih parov $(u_i)_{i=0}^{\infty}$ in $(y_i)_{i=0}^{\infty}$. .

Iz teh parov želimo poiskati prenosno funkcijo ali pa matrike (A, B, C, D) za matriko sistema, s katerim so bili podatki generirani.

Denimo, da ima SISO sistem red n in se da predstaviti s prenosno funkcijo

$$H(z) = \frac{p(z)}{q(z)} = \frac{p_m z^m + p_{m-1} z^{m-1} + \cdots + p_0}{q_n z^n + q_{n-1} z^{n-1} + \cdots + q_0},$$

kjer sta polinoma p in q tuja, $m \leq n$ in $q_n \neq 0$.

To ustreza diskretnemu sistemu

$$q_n y_{k+1} + q_{n-1} y_k + \cdots + q_0 y_{k-n} = p_m u_{k+1} + p_{m-1} u_k + \cdots + p_0 u_{k-m}.$$

Iz vhodno-izhodnih parov lahko sestavimo sistem za koeficiente p in q:

$$\begin{bmatrix} & & & y_0 \\ & & \iddots & y_1 \\ & y_0 & \iddots & y_2 \\ y_0 & y_1 & \iddots & \vdots \\ y_1 & y_2 & & \\ y_2 & \vdots & & \\ \vdots & & & \end{bmatrix} \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ q_n \end{bmatrix} = \begin{bmatrix} & & & u_0 \\ & & \iddots & u_1 \\ & u_0 & \iddots & u_2 \\ u_0 & u_1 & \iddots & \vdots \\ u_1 & u_2 & & \\ u_2 & \vdots & & \\ \vdots & & & \end{bmatrix} \begin{bmatrix} p_0 \\ p_1 \\ \vdots \\ p_n \end{bmatrix}. \tag{6.6}$$

Če upoštevamo še normalizacijo $q_n = 1$, se sistem spremeni v

$$\left[\begin{array}{cccc|ccc} & & & u_0 & & & 0 \\ & & \iddots & u_1 & & & y_0 \\ & & \iddots & u_2 & & \iddots & y_1 \\ & u_0 & \iddots & \vdots & & y_0 & y_2 \\ u_0 & u_1 & \iddots & & y_0 & y_1 & \vdots \\ u_1 & u_2 & & & y_1 & y_2 & \\ u_2 & \vdots & & & y_1 & \vdots & \\ \vdots & & & & \vdots & & \end{array}\right] \left[\begin{array}{c} p_0 \\ \vdots \\ p_n \\ \hline -q_1 \\ \vdots \\ -q_{n-1} \end{array}\right] = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \end{bmatrix}. \tag{6.7}$$

$\underbrace{\qquad\qquad}_{n+1}$ $\underbrace{\qquad\qquad}_{n}$

Izrek 6.6 *Naj bo red sistema, ki je zgeneriral vhodno izhodne podatke u_i in y_i, enak n. Potem*

matrika

$$\left[\begin{array}{c|c} U_{\infty,n+1} & Y_{\infty,n+1} \end{array}\right] = \left[\begin{array}{ccccc|ccccc} & & & & u_0 & & & & & y_0 \\ & & & \iddots & u_1 & & & \iddots & & y_1 \\ & & u_0 & \iddots & u_2 & & y_0 & \iddots & & y_2 \\ u_0 & u_1 & \iddots & & \vdots & y_0 & y_1 & \iddots & & \vdots \\ u_1 & u_2 & & & & y_1 & y_2 & & & \\ u_2 & \vdots & & & & y_2 & \vdots & & & \\ \vdots & & & & & \vdots & & & & \end{array}\right]$$

$$\underbrace{\qquad\qquad}_{n+1} \qquad\qquad \underbrace{\qquad\qquad}_{n+1}$$

ni polnega ranga, matrika

$$\left[\begin{array}{c|c} U_{\infty,n+1} & \widetilde{Y}_{\infty,n+1} \end{array}\right] = \left[\begin{array}{ccccc|ccccc} & & & & u_0 & & & & & 0 \\ & & & \iddots & u_1 & & & & & y_0 \\ & & \iddots & \iddots & u_2 & & & & \iddots & y_1 \\ & u_0 & \iddots & \iddots & \vdots & & y_0 & \iddots & \iddots & y_2 \\ u_0 & u_1 & \iddots & & & y_0 & y_1 & \iddots & \iddots & \vdots \\ u_1 & u_2 & & & & y_1 & y_2 & & & \\ u_2 & \vdots & & & & y_1 & \vdots & & & \\ \vdots & & & & & \vdots & & & & \end{array}\right]$$

$$\underbrace{\qquad\qquad}_{n+1} \qquad\qquad \underbrace{\qquad\qquad}_{n}$$

pa je.

Dokaz. Po predpostavkah izreka velja (6.6). Potem iz (6.7) vidimo, da $\left[\begin{array}{c|c} U_{\infty,n+1} & Y_{\infty,n+1} \end{array}\right]$ ni polnega ranga, saj ima v jedru neničelni vektor $[\,p_0 \;\; \cdots \;\; p_n \;\; -q_0 \;\; \cdots \;\; -q_n\,]^T$.

Za drugi del izreka predpostavimo, da matrika $\left[\begin{array}{c|c} U_{\infty,n+1} & Y_{\infty,n} \end{array}\right]$ ni polnega ranga. Potem ima v jedru neničelni vektor $[\,p_0 \;\; \cdots \;\; p_n \;\; -q_0 \;\; \cdots \;\; -q_{n-1}\,]^T$. Zaradi vzročnosti lahko predpostavimo, da je $u_0 \neq 0$, saj sicer iz $u_0 = \cdots = u_i = 0$ sledi $y_0 = \cdots = y_i = 0$ in lahko vse skupaj premaknemo za indeks i. Prva enačba v (6.7) nam potem da $u_0 p_n = 0$, od tod pa sledi $p_n = 0$. Potem pa velja

$$Y_{\infty,n} \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ q_{n-1} \end{bmatrix} = U_{\infty,n} \begin{bmatrix} p_0 \\ p_1 \\ \vdots \\ p_{n-1} \end{bmatrix}.$$

in red sistema je manjši od n, kar je protislovje. ■

Koeficiente p in q lahko izračunamo iz prvih $2n+1$ vrstic $\left[\begin{array}{c|c} U_{\infty,n+1} & Y_{\infty,n} \end{array}\right]$, stabilnejše pa je, če vzamemo več podatkov in rešimo sistem po metodi najmanjših kvadratov.

6.5 Identifikacija iz vhodno-izhodnih parov (MIMO primer)

Sedaj imamo podane

- vhode $u_0, u_1, \ldots$, kjer je $u_i \in \mathbb{R}^m$ in
- izhode $y_0, y_1, \ldots$, kjer je $y_i \in \mathbb{R}^r$.

Iščemo realizacijo sistema, ki bo imel ustrezni odziv.

Če bi imeli na voljo še stanja $x_i \in \mathbb{R}^n$, bi se zadeva močno poenostavila, saj iz

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k \\ y_k &= Cx_k + Du_k \end{aligned}$$

sledi, da iščemo rešitev sistema

$$\begin{bmatrix} x_2 & x_3 & \cdots & x_N \\ y_1 & y_2 & \cdots & y_{N-1} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \begin{bmatrix} x_1 & x_2 & \cdots & x_{N-1} \\ u_1 & u_2 & \cdots & u_{N-1} \end{bmatrix}.$$

Iz tega sistema lahko izračunamo A, B, C, D ali direktno ali pa po metodi najmanjših kvadratov, če seveda poznamo vektorje stanja x_i.

Za razliko od vhoda in izhoda vektorji stanja niso enolični. S poljubno nesingularno transformacijo T dobimo $\widetilde{A} = T^{-1}AT$, $\widetilde{B} = T^{-1}B$, $\widetilde{C} = CT$, $\widetilde{D} = D$ in $\widetilde{x}_i = T^{-1}X_i$. V nadaljevanju bomo videli, kako lahko kljub temu izračunamo vektorje stanja v neki bazi in potem izračunamo matrike A, B, C, D.

Iz

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k \\ y_k &= Cx_k + Du_k \end{aligned} \quad \text{in} \quad \begin{aligned} x_{k+2} &= Ax_{k+1} + Bu_{k+1} \\ y_{k+1} &= Cx_{k+1} + Du_{k+1} \end{aligned}$$

dobimo

$$\begin{aligned} x_{k+2} &= A^2x_k + \begin{bmatrix} AB & B \end{bmatrix} \begin{bmatrix} u_k \\ u_{k+1} \end{bmatrix} \\ \begin{bmatrix} y_k \\ y_{k+1} \end{bmatrix} &= \begin{bmatrix} C \\ CA \end{bmatrix} x_k + \begin{bmatrix} D & 0 \\ CB & D \end{bmatrix} \begin{bmatrix} u_k \\ u_{k+1} \end{bmatrix}. \end{aligned}$$

Če nadaljujemo, dobimo

$$\left[\begin{array}{c} x_{k+i} \\ \hline y_k \\ y_{k+1} \\ y_{k+2} \\ \vdots \\ y_{k+i-1} \end{array}\right] = \left[\begin{array}{c|cccccc} A^i & A^{i-1}B & A^{i-2}B & A^{i-3}B & \cdots & B \\ \hline C & D & & & & \\ CA & CB & D & & & \\ CA^2 & CAB & CB & D & & \\ \vdots & \vdots & & & \ddots & \\ CA^{i-1} & CA^{i-2}B & \cdots & \cdots & \cdots & D \end{array}\right] \left[\begin{array}{c} x_k \\ \hline u_k \\ u_{k+1} \\ u_{k+2} \\ \vdots \\ u_{k+i-1} \end{array}\right].$$

Zgornji sistem lahko zapišemo v kompaktnejši obliki

$$\begin{aligned} x_{k+i} &= A^i x_k + \mathbf{B}_i U_{k,i} \\ Y_{k,i} &= \mathbf{C}_i x_k + \mathbf{D}_i U_{k,i}, \end{aligned}$$

kjer je

$$\mathbf{B}_i = \begin{bmatrix} A^{i-1}B & \cdots & B \end{bmatrix}, \quad \mathbf{D}_i = \begin{bmatrix} D & & & \\ CB & D & & \\ \vdots & & \ddots & \\ CA^{i-2}B & \cdots & \cdots & D \end{bmatrix},$$

$$\mathbf{C}_i = \begin{bmatrix} C \\ \vdots \\ CA^{i-1} \end{bmatrix}, \qquad Y_{k,i} = \begin{bmatrix} y_k \\ \vdots \\ y_{k+i-1} \end{bmatrix}, \qquad U_{k,i} = \begin{bmatrix} u_k \\ \vdots \\ u_{k+i-1} \end{bmatrix}.$$

Če sedaj definiramo

$$\begin{aligned} Y_{k,i,j} &= [\,Y_{k,i} \quad Y_{k+1,i} \quad \cdots \quad Y_{k+j-1,i}\,] = \begin{bmatrix} y_k & y_{k+1} & \cdots & y_{k+j-1} \\ \vdots & & & \vdots \\ y_{k+i-1} & y_{k+i} & \cdots & y_{k+i+j-2} \end{bmatrix}, \\ U_{k,i,j} &= [\,U_{k,i} \quad U_{k+1,i} \quad \cdots \quad U_{k+j-1,i}\,] = \begin{bmatrix} u_k & u_{k+1} & \cdots & u_{k+j-1} \\ \vdots & & & \vdots \\ u_{k+i-1} & u_{k+i} & \cdots & u_{k+i+j-2} \end{bmatrix}, \\ X_{k,j} &= [\,x_k \quad x_{k+1} \quad \cdots \quad x_{k+j-1}\,], \end{aligned}$$

sledi

$$X_{k+i,j} = A^i X_{k,j} + \mathbf{B}_i U_{k,i,j} \tag{6.8}$$

$$Y_{k,i,j} = \mathbf{C}_i X_{k,j} + \mathbf{D}_i U_{k,i,j}. \tag{6.9}$$

Iz $Y_{k,i,j}$ in $U_{k,i,j}$ sestavimo matriko $H_{k,i,j} = \begin{bmatrix} Y_{k,i,j} \\ U_{k,i,j} \end{bmatrix}$ velikosti $(m+r)i \times j$.

Izrek 6.7 *Naj bo odziv nenehno vzbujen in naj velja $i \geq n$ in $j \geq (m+r)i$, kjer je n red sistema. Potem je*

$$\operatorname{rang}(H_{k,i,j}) = \operatorname{rang}(U_{k,i,j}) + \operatorname{rang}(X_{k,j}) = mi + n.$$

Dokaz. Predpostavimo lahko, da je realizacija spoznavna. Zaradi tega je za $i \geq n$, matrika $\mathbf{C}_I$ polnega ranga.

V primeru $j \geq (m+r)i$ imajo matrike $(U_{k,i,j})$, $Y_{k,i,j}$ in $X_{k,j}$ več stolpcev kot vrstic. Iz (6.9) sledi, da so vrstice $Y_{k,i,j}$ linearne kombinacije vrstic $(U_{k,i,j})$ in $X_{k,j}$x. Zaradi tega lahko pišemo

$$\operatorname{rang}(H_{k,i,j}) \leq \operatorname{rang}(U_{k,i,j}) + \operatorname{rang}(X_{k,j}).$$

Ker pa so vrstice $U_{k,i,j}$ in $X_{k,j}$ linearno neodvisne (to sledi iz nenehne vzujenosti), mora biti

$$\operatorname{rang}(H_{k,i,j}) = \operatorname{rang}(U_{k,i,j}) + \operatorname{rang}(X_{k,j}).$$

■

Vse skupaj lahko premaknemo za i in zapišemo

$$\operatorname{rang}(H_{k,i,j}) = \operatorname{rang}(U_{k,i,j}) + \operatorname{rang}(X_{k,j}). \tag{6.10}$$

Izrek 6.8

$$\mathrm{Lin}(X_{k+i,j}^T) = \mathrm{Lin}(H_{k,i,j}^T) \cap \mathrm{Lin}(H_{k+i,i,j}^T).$$

Dokaz. Matrika

$$H_{k,2i,j} = \begin{bmatrix} Y_{k,2i,j} \\ U_{k,2i,j} \end{bmatrix}$$

ima po izreku 6.7 rang $2mi + n$. Hitro lahko vidimo, da obstaja taka permutacijska matrika P, da je

$$H_{k,2i,j} = P \begin{bmatrix} H_{k,i,j} \\ H_{k+i,i,j} \end{bmatrix}.$$

Od tod sledi, da je

$$\dim\left(\mathrm{Lin}(H_{k,i,j}^T) \cap \mathrm{Lin}(H_{k+i,i,j}^T)\right) = n,$$

saj je $\mathrm{rang}(H_{k,i,j}^T) + \mathrm{rang}(H_{k+i,i,j}^T) = 2mi + 2n$.

Sedaj moramo pokazati še, da je $\mathrm{Lin}(X_{k,+i,j}^T) \subset \mathrm{Lin}(H_{k,i,j}^T) \cap \mathrm{Lin}(H_{k+i,i,j}^T)$. To bi že pomenilo enakost, saj vemo, da je $\mathrm{rang}(X_{k,+i,j}^T) = n$.

Iz zveze (6.10) sledi, da je $\mathrm{Lin}(X_{k,+i,j}^T) \subset \mathrm{Lin}(H_{k+i,i,j}^T)$. Ostalo nam le še dokazati, da je $\mathrm{Lin}(X_{k,+i,j}^T) \subset \mathrm{Lin}(H_{k,i,j}^T)$. Ker je matrika $\mathbf{C}_i$ polnega ranga, iz enačbe (6.9) sledi

$$X_{k,j} = \mathbf{C}_i^+ U_{k,i,j} - \mathbf{C}_i^+ Y_{k,i,j},$$

kar vstavimo v (6.8) in dobimo

$$X_{k,i,j} = (A^i \mathbf{C}_i^+ \mathbf{B}_i) U_{k,i,j} - A^i \mathbf{C}_i^+ Y_{k,i,j}.$$

Torej so vrstice $X_{k,+i,j}^T$ linearne kombinacije vrstic $Y_{k,i,j}$ in $U_{k,i,j}$, ki skupaj sestavljajo vrstice $H_{k,i,j}$ in dokaz je končan. ■

S pomočjo zveze

$$\mathrm{Lin}(X_{k+i,j}^T) = \mathrm{Lin}(H_{k,i,j}^T) \cap \mathrm{Lin}(H_{k+i,i,j}^T)$$

lahko $X_{k+i,j}$ izračunamo iz $H_{k,i,j}$ in $H_{k+i,i,j}$, ko pa enkrat poznamo vektorje stanja, lahko izračunamo matrike A, B, C, D iz sistema

$$\begin{bmatrix} x_{k+i+1} & x_{k+i+2} & \cdots & x_{k+i+j-1} \\ y_{i+k} & y_{i+k+1} & \cdots & y_{i+k+j-2} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \begin{bmatrix} x_{k+i} & x_{k+i+1} & \cdots & x_{k+i+j-2} \\ u_{k+i} & u_{k+i+1} & \cdots & u_{k+i+j-2} \end{bmatrix}.$$

Najboljši približek za $\mathrm{Lin}(H_{k,i,j}^T) \cap \mathrm{Lin}(H_{k+i,i,j}^T)$ dobimo preko singularnega razcepa. Poiščemo lahko ortogonalni matriki Q in V, da je

$$Q^T \begin{bmatrix} H_{k,i,j} \\ H_{k+i,i,j} \end{bmatrix} V = \begin{bmatrix} M_{11} & M_{12} & 0 \\ \times & \times & M_{23} \\ M_{31} & 0 & 0 \end{bmatrix} \begin{matrix} (m+r)i \\ mi \\ ri \end{matrix} \quad . \qquad (6.11)$$

$$\begin{matrix} n & mi & j-mi-n \end{matrix}$$

Ker imata matriki M_{11} in M_{31} linearno neodvisne stolpce, presek

$$\mathrm{Lin}(\begin{bmatrix} M_{11} & M_{12} & 0^T \end{bmatrix} \cap \mathrm{Lin}\left(\begin{bmatrix} \times & \times & M_{23} \\ M_{31} & 0 & 0 \end{bmatrix}^T\right) \qquad (6.12)$$

očitno vsebuje vse vrstice oblike $w = [\, z^T \quad 0\,]$, kjer je $w \in \mathbb{R}^j$ in $z \in \mathbb{R}^n$. Ker pa je dimenzija podprostora takih vektorjev n, to pa je tudi dimenzija preseka (6.12), so to vsi vektorji iz preseka. Iz (6.11) potem sledi, da prvih n vrstic matrike V^T razpenja podprostor $\mathrm{Lin}(X_{k+i,j}^T)$.

Poglejmo, kako lahko s tremi singularnimi razcepi pridemo do ortogonalnih matrik Q in V, ki zadoščata (6.11).

Naj bo

$$\mathbf{H} = \begin{bmatrix} H_{k,i,j} \\ H_{k+i,i,j} \end{bmatrix}.$$

Najprej izračunamo singularni razcep $H_{k,i,j} = U_1 S_1 V_1^T$, kjer sta U_1 in V_1 ortogonalni matriki velikosti $(m+r)i \times (m+r)i$ oziroma $j \times j$, S_1 pa je diagonalna matrika reda $(m+r)i \times j$, ki ima le prvih $p = \mathrm{rang}(H_{k,i,j})$ diagonalnih elementov neničelnih. Ker vemo, da je $\mathrm{rang}(H_{k,i,j}) = mi + n$, od tod lahko izračunamo red kontrolnega sistema $n = p - mi$. Od tod sledi

$$\mathbf{H}V_1 = \begin{matrix} (m+r)i \\ (m+r)i \end{matrix} \begin{bmatrix} \times & 0 \\ \times & H_2 \\ p & j-p \end{bmatrix} \begin{matrix} (m+r)i \\ (m+r)i \end{matrix}.$$

Sedaj izračunamo singularni razcep $H_2 = U_2 S_2 V_2^T$, kjer je U_2 ortogonalna matrika dimenzije $(m+r) \times (m+r)i$, S_2 je matrika velikosti $(m+r)i \times (j-p)$, ki ima prvih mi singularnih vrednosti neničelnih, V_2 pa je ortogonalna matrika velikosti $(j-p) \times (j-p)$. Tako dobimo

$$\begin{matrix} (m+r)i \\ (m+r)i \end{matrix} \begin{bmatrix} I & 0 \\ 0 & U_2^T \\ (m+r)i & (m+r)i \end{bmatrix} \mathbf{H}V_1 = \begin{bmatrix} \times & 0 \\ \times & \times \\ H_3 & 0 \\ p & j-p \end{bmatrix} \begin{matrix} (m+r)i \\ mi \\ ri \end{matrix}.$$

Za konec izračunamo še singularni razcep $H_3 = U_3 S_3 V_3^T$, kjer sta U_3 in V_3 ortogonalni matriki velikosti $ri \times ri$ oziroma $p \times p$, S_1 pa je diagonalna matrika reda $ri \times p$, ki ima le prvih n diagonalnih elementov neničelnih. Potem je

$$\begin{matrix} (m+r)i \\ (m+r)i \end{matrix} \begin{bmatrix} I & 0 \\ 0 & U_2^T \\ (m+r)i & (m+r)i \end{bmatrix} \mathbf{H}V_1 \begin{matrix} j-p \\ p \end{matrix} \begin{bmatrix} I & 0 \\ 0 & V_3 \\ j-p & p \end{bmatrix} = \begin{bmatrix} M_{11} & M_{12} & 0 \\ \times & \times & M_{23} \\ M_{31} & 0 & 0 \\ n & mi & j-mi-n \end{bmatrix} \begin{matrix} (m+r)i \\ mi \\ ri \end{matrix}$$

razcep, ki smo ga iskali in iz katerega lahko preberemo matriki Q in V.

Celotni algoritem za izračun realizacije je podan v algoritmu 6.2.

Algoritem 6.2 Identifikacija MIMO sistema iz vhodno-izhodnih parov

Začetni podatek je $H = \begin{bmatrix} H_{k,i,j} \\ H_{k+i,i,j} \end{bmatrix}$, $i \geq n$, $j \geq (m+r)i$.

1. $H_{k,i,j} = U_1 S_1 V_1^T$, S_1 velikosti $p \times p$, $n = p - mi$, $H = HV_1$.

2. $H = \begin{matrix} \begin{bmatrix} \times & 0 \\ \times & H_2 \end{bmatrix} \\ \begin{matrix} p & j-p \end{matrix} \end{matrix} \begin{matrix} (m+r)i \\ (m+r)i \end{matrix}$, $H_2 = U_2 S_2 V_2^T$, S_2 velikosti $mi \times mi$, $H = \begin{bmatrix} I & 0 \\ 0 & U_2^T \end{bmatrix} H$.

3. $H = \begin{matrix} \begin{bmatrix} \times & 0 \\ \times & \times \\ H_3 & 0 \end{bmatrix} \\ \begin{matrix} p & j-p \end{matrix} \end{matrix} \begin{matrix} (m+r)i \\ mi \\ ri \end{matrix}$, $H_3 = U_3 S_3 V_3^T$, S_3 velikosti $n \times n$,

4. $V_1 = [\, V_{1a} \quad V_{1b} \,]$, $X_{k+i,j} = V_3^T V_{1a}^T$.

5. Reši predoločeni sitem

$$\begin{bmatrix} x_{k+i+1} & \cdots & x_{k+i+j-1} \\ y_{i+k} & \cdots & y_{i+k+j-2} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \begin{bmatrix} x_{k+i} & \cdots & x_{k+i+j-2} \\ u_{k+i} & \cdots & u_{k+i+j-2} \end{bmatrix}.$$

Poglavje 7

Stabilizacija in razporejanje polov

7.1 Uvod

Imamo linearni zvezni kontrolni sistem

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \qquad x(t_0) = x_0, \quad t \geq t_0, \\ y(t) &= Cx(t) + Du(t). \end{aligned}$$

Pri vodenju sistema ponavadi uporabljamo povratno zvezo. Če poznamo stanje $x(t)$, lahko za povratno zvezo vzamemo

$$u(t) = v(t) - Kx(t),$$

kjer je $v(t)$ nova vhodna funkcija. Zaprtozančni sistem je

$$\begin{aligned} \dot{x}(t) &= (A - BK)x(t) + Bv(t), \qquad x(t_0) = x_0, \quad t \geq t_0, \\ y(t) &= (C - DK)x(t) + Dv(t). \end{aligned}$$

Ukvarjali se bomo z dvema nalogama:

- Pri *stabilizaciji sistema* iščemo tako matriko povratne zveze K, da bo zaprtozančni sistem stabilen.
- Pri *razporejanju polov* oz. *EVA problemu* (eigenvalue assignment problem) iščemo tako matriko povratne zveze K, da bo zaprtozančni sistem imel točno določene pole.

Očitno je, da če lahko poljubno razporedimo lastne vrednosti, potem lahko vse pošljemo v levo polravnino $\mathbb{C}$ in tako stabiliziramo sistem. Moramo pa že vnaprej določiti, kje naj bodo nove lastne vrednosti. Pri problemu stabilizacije ne moremo vplivati na to, kakšni bodo novi poli, bo pa novi sistem stabilen.

Izrek 7.1 *Za realni matriki A in B iz linearnega sistema $\dot{x}(t) = Ax(t) + Bu(t)$ obstaja realna matrika K, da ima $A - BK$ predpisan spekter (zaprt za konjugiranje), natanko tedaj, ko je par (A, B) vodljiv.*

7.2 Stabilizacija s povratno zvezo iz stanja

Iščemo tako matriko K, da bo zaprtozančni sistem stabilen, ne moremo pa podati, kje naj ležijo novi poli.

Predpostavka je, da je sistem *stabilizabilen* oz. ga je možno stabilizirati. Zadosten pogoj za stabilizabilnost je, da je par (A, B) vodljiv, ni pa potreben.

Če par (A, B) ni vodljiv, potem obstaja taka nesingularna matrika T, da je

$$TAT^{-1} = \begin{bmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ 0 & \widetilde{A}_{22} \end{bmatrix}, \qquad TB = \begin{bmatrix} \widetilde{B}_1 \\ 0 \end{bmatrix}$$

in je par $(\widetilde{A}_{11}, \widetilde{B}_1)$ vodljiv. Potreben pogoj za stabilizabilnost je potem stabilnost $\widetilde{A}_{22}$. Matrika povratne zveze ima potem bločno obliko

$$K = \begin{bmatrix} K_1 & 0 \\ 0 & K_2 \end{bmatrix},$$

kjer je K_1 taka matrika, da je $\widetilde{A}_{11} - \widetilde{B}_1 K_1$ stabilna, matrika K_2 pa je poljubna.

7.2.1 Stabilizacija preko vodljivostne Gramove matrike

Če je par (A, B) vodljiv, potem je za poljuben $t > 0$ Gramova vodljivostna matrika

$$W_C(t) = \int_0^t e^{-As} B B^T e^{-A^T s} ds \tag{7.1}$$

pozitivno definitna. Iz (7.1) dobimo

$$AW_C(t) + W_C(t)A^T = -\int_0^t \frac{d}{ds}\left(e^{-As} B B^T e^{-A^T s}\right) ds = e^{-At} B B^T e^{-A^T t} + B B^T.$$

Na obeh straneh prištejemo $-2BB^T = W_C(t) W_C^{-1}(t) B B^T - B B^T W_C^{-1}(t) W_C(t)$, da dobimo

$$F W_C(t) + W_C(t) F^T = -Q, \tag{7.2}$$

kjer je $F = A - B B^T W_C^{-1}(t)$ in $Q = e^{-At} B B^T e^{-A^T t} + B B^T$.

F je rešitev enačbe Ljapunova (7.2), kjer sta $W_C(t)$ in Q simetrični pozitivno definitni matriki, zato je F stabilna matrika.

Za izračun $W_C(t)$ lahko uporabimo eno izmed metod za numerično računanje eksponentne funkcije matrike. Če je

$$\exp\left(\begin{bmatrix} A & B B^T \\ & -A^T \end{bmatrix} t\right) = \begin{bmatrix} E_1(t) & G(t) \\ & E_2(t) \end{bmatrix},$$

potem je očitno $E_1(t) = e^{At}$, $E_2(t) = e^{-A^T t}$, za $G(t)$ pa preverimo, da velja $G(t) = e^{At} W_C(t)$.

Res, če upoštevamo $\frac{d}{dt}(e^{Ct}) = C e^{Ct}$, potem bi moralo za $G(t)$ veljati

$$G'(t) = A G(t) + B B^T E_2(t),$$

to pa je točno to, kar dobimo z odvajanjem izraza $G(t) = \int_0^t e^{A(t-s)} BB^T e^{-A^T s} ds$.

Sledi, da je $W_C(t) = E_2(t)^T G(t)$, potem pa za K vzamemo $B^T W_C(t)^{-1}$.

Denimo, da sistem ni vodljiv, je pa stabilizabilen. Obstaja ortogonalna matrika U, da je

$$U^T A U = \begin{bmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ 0 & \widetilde{A}_{22} \end{bmatrix}, \qquad U^T B = \begin{bmatrix} \widetilde{B}_1 \\ 0 \end{bmatrix}$$

in je par $(\widetilde{A}_{11}, \widetilde{B}_1)$ vodljiv, matrika $\widetilde{A}_{22}$ pa stabilna.

Potem je

$$W_C^+(t) = U^T \begin{bmatrix} W_{11}^{-1}(t) & 0 \\ 0 & 0 \end{bmatrix} U,$$

kjer je

$$W_{11}(t) = \int_0^t e^{-A_{11} s} BB^T e^{-A_{11}^T s} ds.$$

Zato je stabilen zaprtozančni model

$$A - BB^T W_C^+ = U \begin{bmatrix} A_{11} - B_1 B_1^T W_{11}^T & A_{12} \\ 0 & A_{22} \end{bmatrix} U^T.$$

Tako smo izpeljali algoritem 7.1 za stabilizacijo z Gramovo matriko.

Algoritem 7.1 Stabilizacija sistema preko Gramove spoznavnostne matrike

1. $H = \begin{bmatrix} A & BB^T \\ 0 & -A^T \end{bmatrix}$,
2. za $t > 0$ izračunaj $e^{Ht} = \begin{bmatrix} E_1(t) & G(t) \\ & E_2(t) \end{bmatrix}$,
3. $W_C(t) = E_2^T(t) G(t)$,
4. izračunaj lastni razcep
$$W_C(t) = Q \begin{bmatrix} \Lambda & \\ & 0 \end{bmatrix} Q^T,$$
5. matrika povratne zveze je
$$K = B^T Q \begin{bmatrix} \Lambda^{-1} & \\ & 0 \end{bmatrix} Q^T.$$

V primeru diskretnega sistema $x_{k+1} = Ax_k + Bu_k$ je analogna formula

$$K = B^T \left[(A^T)^N W_N^+ A^N \right] A = (I + B^T V_N B)^{-1} B^T V_N A,$$

kjer vzamemo $N \geq n$ in

$$W_N = \sum_{i=0}^{N} A^i BB^T (A^T)^i,$$

$$V_N = (A^T)^N W_N^+ + A^N.$$

V primeru, ko je par (A, B) vodljiv, namesto W_N^+ lahko vzamemo W_N^{-1}, saj je v tem primetu W_N nesingularna matrika.

7.2.2 Stabilizacija preko enačbe Ljapunova

Izrek 7.2 *Naj bo (A, B) vodljiv par in $\beta > |\lambda_{\max}|$, kjer je $\lambda_{\max}$ lastna vrednost matrike A z največjim absolutnim realnim delom. Če vzamemo za matriko povratne zveze $K = B^T Z^{-1}$, kjer je Z s.p.d. rešitev enačbe Ljapunova*

$$-(A + \beta I)Z + Z(-(A + \beta I))^T = -2BB^T, \tag{7.3}$$

potem je matrika $A - BK$ stabilna.

Dokaz. Matrika $-(A+\beta I)$ je stabilna, saj imajo njene lastne vrednosti očitno negativne realne dele. Iz vodljivosti para (A, B) sledi vodljivost para $(-(A + \beta I), B)$. Zdaj iz izreka 4.5 sledi, da je Z, ki reši enačbo Ljapunova (7.3), simetrična pozitivno definitna matrika.

Enačbo (7.3) lahko prepišemo v obliki

$$(A - BB^T Z^{-1})Z + Z(A - BB^T Z^{-1})^T = -2\beta Z$$

oziroma

$$(A - BK)Z + Z(A - BK)^T = -2\beta Z, \tag{7.4}$$

kjer je $K = B^T Z^{-1}$. Iz simetrične pozitivne definitnosti Z sedaj sledi, da je matrika $A - BK$ stabilna, saj če je npr. $(A - BK)x = \lambda x$ lastni par matrike $A - BK$, potem iz (7.4) dobimo

$$(\overline{\lambda} + \lambda)x^H Z x = -2\beta x^H Z x,$$

od tod pa sledi $\mathrm{Re}(\lambda) > 0$. ■

Če nimamo na voljo boljše ocene, lahko vzamemo $\beta > \|A\|$. Stabilizacija preko enačbe Ljapunova je predstavljena v algoritmu 7.2.

Podobno kot prej, če par (A, B) ni vodljiv, se pa da stabilizirati, potem se matrika povratne zveze izraža s pseudoinverzom $K = B^T Z^+$, kjer je Z simetrična nenegativno definitna matrika, ki zadošča (7.3).

V primeru diskretnega sistema $x_{k+1} = Ax_k + Bu_k$ imamo analogni izrek.

Izrek 7.3 *Naj bo $x_{k+1} = Ax_k + Bu_k$ vodljiv diskretni sistem. Naj bo $0 < \beta \leq 1$, da velja $|\lambda| \geq \beta$ za vse lastne vrednosti λ matrike A in naj Z zadošča diskretni enačbi Ljapunova*

$$AZA^T - \beta^2 z = 2BB^T.$$

Če vzamemo

$$K = B^T(Z + BB^T)^{-1}A,$$

potem je $A - BK$ konvergentna matrika.

Algoritem 7.2 Stabilizacija sistema preko enačbe Ljapunova

1. izberi $\beta > \|A\|$,
2. izračunaj $Q = 2BB^T$,
3. reši enačbo Ljapunova
$$-(A+\beta I)Z + Z(-(A+\beta I))^T = -Q,$$
4. izračunaj lastni razcep
$$Z = V \begin{bmatrix} \Lambda & \\ & 0 \end{bmatrix} V^T,$$
5. matrika povratne zveze je
$$K = B^T V \begin{bmatrix} \Lambda^{-1} & \\ & 0 \end{bmatrix} V^T.$$

7.3 Razporejanje polov

Sistem lahko stabiliziramo tudi tako, da izberemo tako matriko povratne zveze, da ima $A - BK$ vnaprej izbrane stabilne lastne vrednosti. To lahko naredimo, če je par (A, B) vodljiv. Na voljo imamo več algoritmov, osnovni je Ackermanov, ki smo ga že opisali v razdelku 3.6.1.

Naj bo
$$\alpha(s) = s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_1 s + \alpha_0$$
ciljni karakteristični polinom. Če je sistem podan v vodljivostni normalni obliki
$$A = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -a_0 & -a_1 & \cdots & \cdots & -a_{n-1} \end{bmatrix}, \qquad b = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ b_n \end{bmatrix},$$
potem moramo vzeti
$$\widetilde{k} = [\,\alpha_0 - a_0 \quad \alpha_1 - a_1 \quad \cdots \quad \alpha_{n-1} - a_{n-1}\,].$$

Če par (A, b) ni v vodljivostni normalni obliki, moramo najprej poiskati prehodno nesingularno matriko T, da bosta $\widetilde{A}$ in $\widetilde{b}$ prave oblike. Če je
$$\widetilde{k} = [\,\alpha_0 - a_0 \quad \alpha_1 - a_1 \quad \cdots \quad \alpha_{n-1} - a_{n-1}\,]$$
prava izbira povratne zveze za transformirani par $(\widetilde{A}, \widetilde{b}) = (TAT^{-1}, Tb)$, potem je za originalni par (A, b) prava izbira $k = \widetilde{k}T$. Vemo
$$k = \widetilde{k}T = [\,\alpha_0 - a_0 \quad \alpha_1 - a_1 \quad \cdots \quad \alpha_{n-1} - a_{n-1}\,] \begin{bmatrix} s_n \\ s_n A \\ \vdots \\ s_n A^{n-1} \end{bmatrix},$$

kjer je s_n zadnja vrstica inverza vodljivostne matrike C_M^{-1}. Sledi

$$\widetilde{k}T = s_n(\alpha_0 I + \alpha_1 A + \cdots + \alpha_{n-1}A^{n-1}) - s_n(a_0 I + a_1 A + \cdots + a_{n-1}A^{n-1})$$

in preko Cayley-Hamiltonovega izreka dobimo *Ackermanovo formulo*

$$k = e_n^T C_M^{-1} \alpha(A).$$

7.3.1 Zveza med poli in prehodnim obnašanjem sistema

Ponavadi ni dovoj, da sistem le stabiliziramo, temveč bi radi tudi kontrolirali, kako se odziva na spremembe, torej kakšno je njegovo prehodno obnašanje.

Za zgled si poglejmo sistem druge stopnje

$$\ddot{x}(t) + 2\zeta\omega\dot{x}(t) + \omega^2 x(t) = u(t),$$

kjer je ζ faktor dušenja, ω pa naravna frekvenca sistema.

Pola sistema sta $\lambda_{1,2} = \zeta\omega \pm \omega\sqrt{1-\zeta^2}$.

Prehodno obnašanje sistema je odvisno od ζ in ω.

Pri fiksnem ω imamo

- za $\zeta \geq 1$ dobimo počasen in gladek odziv,
- za $0 \leq \zeta < 1$ imamo hiter a oscilirajoč odziv.

Pri večjih sistemih ponavadi odziv določata dva dominantna pola, ki sta najbližja imaginarni osi.

V primeru, ko pole dosti premaknemo, lahko pričakujemom da bomo potrebovali velike vrednosti v vhodnem vektorju. Velja namreč

$$u(t) = v(t) - Kx(t)$$

in če je norma K velika, lahko pričakujemo, da bo zaprtozančni sistem potreboval velike vrednosti na vhodu. Ker so ponavadi maksimalne vrednosti vhoda omejene, to pomeni, da si ne moremo privoščiti povratnih zvez, kjer pride do velikih premikov polov.

7.4 Razporejanje polov enovhodnih sistemov

Imamo linearni zvezni kontrolni sistem

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \qquad x(t_0) = x_0, \quad t \geq t_0, \\ y(t) &= Cx(t) + Du(t). \end{aligned}$$

Predpostavimo lahko, da je sistem vodljiv. Potem obstaja ortogonalna matrika U, da je

$$\widetilde{A} = U^T A U$$

ireducibilna zgornja Hessenbergova matrika in

$$\widetilde{b} = U^T b = [\, b_1 \quad 0 \quad \cdots \quad 0 \,]^T .$$

Zato lahko predpostavimo, da je že začetni sistem v t.i. Hessenbergovi vodljivostni obliki.

Iščemo povratno zvezo k, da bo $A - bk$ imela predpisane lastne vrednosti $\lambda_1, \ldots, \lambda_n$.

7.4.1 Razporejanje polov preko Hessenbergove forme

Predpostavimo, da so predpisane lastne vrednosti vse enostavne. Potem imamo n lastnih vektorjev $x_1, \ldots, x_n$ in $A - bk$ se da diagonalizirati kot

$$A - bk = X \Lambda X^{-1}.$$

Če bi poznali X, bi lahko izračunali povratno zvezo k iz

$$k = \frac{1}{b_1} \Big(e_1^T (A - X \Lambda X^{-1}) \Big).$$

Izkaže se, da je to izvedljivo, saj lahko lastne vektorje $x_1, \ldots, x_n$ izračunamo, ne da bi poznali k.

Če je z lastni vektor za lastno vrednost λ, je $(A - bk)z = \lambda z$, kjer je

$$A - bk = \begin{bmatrix} a_{11} - b_1 k_1 & a_{12} - b_1 k_2 & \cdots & a_{1n} - b_1 k_n \\ a_{21} & a_{22} & \cdots & a_{2n} \\ & \ddots & & \vdots \\ & & a_{n,n-1} & a_{nn} \end{bmatrix}.$$

Če izberemo $z_n \neq 0$, lahko preostale komponente z izračunamo neodvisno od k kot

$$z_k = \frac{1}{a_{k+1,k}} \left(\lambda z_{k+1} - \sum_{j=k+1}^{n} a_{k+1,j} z_j \right), \quad k = n-1, \ldots, 2, 1.$$

Ker je A ireducibilna, so vsi poddiagonalni elementi $a_{i+1,i}$ neničelni.

Tako lahko za vsako lastno vrednost λ_k izračunamo lastni vektor x_k, potem pa iz formule

$$k = \Big(e_1^T (A - X \Lambda X^{-1}) \Big)$$

dobimo k.

7.4.2 Metoda ortogonalnih transformacij na lastnih vektorjih

Prejšnja metoda ni popolna. Pri računanju $X \Lambda X^{-1}$ lahko pride do velikih napak, če je matrika lastnih vektorjev zelo občutljiva, težava pa je tudi, da smo omejeni na enostavne lastne vrednosti. Težavam se bomo poskusili izogniti z računanjem z ortogonalnimi transformacijami.

Če je $\|z\|_2 = 1$ lastni vektor za lastno vrednost λ, je $(A-bk)z = \lambda z$. Naj bo Q taka ortogonalna matrika, da je $z = Qe_1$. Potem ima $Q^T(A-bk)Q$ obliko

$$Q^T(A-bk)Q = \begin{bmatrix} \lambda & \times & \cdots & \times \\ 0 & \times & \cdots & \times \\ \vdots & \vdots & & \vdots \\ 0 & \times & \cdots & \times \end{bmatrix}.$$

Sedaj bi lahko nadaljevali na preostanku matrike, a bomo pokazali, da lahko to redukcijo izvedemo tako, da bo reducirani problem spet v Hessenbergovi vodljivostni obliki.

Predpostavimo, da so izbrane lastne vrednosti realne. Naj bo λ lastna vrednost in z lastni vektor za $A-bk$. Izberemo $z_n \neq 0$ in iz $(A-bk)z = \lambda z$ izračunamo

$$\begin{aligned} z_{n-1} &= ((\lambda - a_{nn})z_n)/a_{n,n-1} \\ z_{n-2} &= ((\lambda - a_{n-1,n-1})z_{n-1} - a_{n-1,n}z_n)/a_{n-1,n-2} \end{aligned}$$

Z rotacijo $z^{(1)} = R_{n-1,n}^T z = [\,\times \quad \cdots \quad \times \quad z_{n-2} \quad \widetilde{z}_{n-1} \quad 0\,]^T$ uničimo zadnji element v z. Zaradi $z_n \neq 0$ je tudi $\widetilde{z}_{n-1} \neq 0$.

$(A-bk)z = \lambda z$ se spremeni v $R_{n-1,n}^T(A-bk)R_{n-1,n}z^{(1)} = \lambda z^{(1)}$, kjer je

$$R_{n-1,n}^T(A-bk)R_{n-1,n} = \begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & + & \times & \times \end{bmatrix},$$

na mestu $(n, n-2)$ je novi neničelni element.

Sedaj izračunamo

$$z_{n-3} = ((\lambda - a_{n-2,n-1})z_{n-2} - a_{n-2,n-1}\widetilde{z}_{n-1})/a_{n-2,n-3},$$

kjer so a_{ij} elementi matrike $R_{n-1,n}^T A R_{n-1,n}$.

Z rotacijo $R_{n-2,n-1}$ uničimo naslednji element v z:

$$z^{(2)} = R_{n-2,n-1}^T z^{(1)} = [\,\times \quad \cdots \quad \times \quad z_{n-3} \quad \widetilde{z}_{n-2} \quad 0 \quad 0\,]^T.$$

Dobimo

$$R_{n-2,n-1}^T R_{n-1,n}^T(A-bk)R_{n-1,n}R_{n-2,n-1} = \begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & + & \times & \times & \times \\ & & 0 & \times & \times \end{bmatrix}.$$

Na mestu $(n-1, n-3)$ je nov neničelni element, neničelni element na mestu $(n, n-2)$ iz prejšnjega koraka pa je spet nič. To pa zato, ker iz zadnje vrstice sledi $a_{n,n-2}\widetilde{z}_{n-2} = 0$, za $\widetilde{z}_{n-2}$ pa vemo, da ni 0.

Tako imamo nekaj podobnega preganjanju grbe pri QR algoritmu. Po $n-1$ rotacijah ostane $Q^T(A-bk)Qz^{(n-1)} = \lambda z^{(n-1)}$, kjer je $Q = R_{n-1,n}\cdots R_{12}$, $z^{(n-1)} = [\times \quad 0 \quad \cdots \quad 0]^T$ in

$$Q^TAQ = \begin{bmatrix} a_{11} & \times & \cdots & \times \\ a_{21} & & & \\ 0 & & A^{(2)} & \\ \vdots & & & \\ 0 & & & \end{bmatrix}, \qquad Q^Tb = \begin{bmatrix} \widetilde{b}_1 \\ \widetilde{b}_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

kjer je $\widetilde{b}_2 \neq 0$ in je $A^{(2)}$ ireducibilna Hessenbergova matrika. Iz zveze $\widetilde{b}k_1 = a_{11} - \lambda$ in $\widetilde{b}_2k_1 = a_{21}$ dobimo $k_1 = a_{21}/\widetilde{b}_1$. Postopek rekurzivno nadaljujemo na $A^{(2)}$ in $b^{(2)} = [\widetilde{b}_2 \quad 0 \quad \cdots \quad 0]^T$.

7.4.3 Modifikacija QR algoritma

Tu bomo uporabili znano dejstvo, da v primeru, ko je matrika H ireducibilna Hessenbergova matrika in je λ njena lastna vrednosti, pri QR algoritmu s premikom λ v enem koraku izločimo to lastno vrednost.

En korak QR algoritma s premikom λ je

$$\begin{aligned} H - \lambda I &= QR \\ \widetilde{H} &= RQ + \lambda I \end{aligned}$$

Ker je λ lastna vrednost, ima zgornja trikotna matrika R prvih $n-1$ stolpcev linearno neodvisnih, v zadnjem stolpcu pa je $r_{nn} = 0$. Zaradi tega ima $\widetilde{H}$ zadnjo vrstico $[0 \quad \cdots \quad 0 \quad \lambda]$.

Sedaj bomo to obrnili tako, da bomo nastavili vrednost koeficientov povratne zveze k tako, da bomo po enem koraku QR s premikom λ, kjer bo λ ena izmed predpisanih lastnih vrednosti, izločili λ.

Na začetku je A ireducibilna zgornja Hessenbergova in $b = [b_1 \quad 0 \quad \cdots \quad 0]^T$, iščemo pa $k = [k_1 \quad \cdots \quad k_n]^T$, da bo $A - bk$ imela lastne vrednosti $\lambda_1, \ldots, \lambda_n$.

Označimo $A^{(1)} = A$, $b^{(1)} = b$ in $k^{(1)} = k$. Če naj bo λ_1 lastna vrednost je $A^{(1)} - b^{(1)}k^{(1)} - \lambda_1 I$ singularna matrika. Če sedaj na matriki $A^{(1)} - b^{(1)}k^{(1)} - \lambda_1 I$ z desne uporabimo Givensove rotacije $R_{n-1,n}, \ldots, R_{12}$, da jo spravimo v zgornjo trikotno obliko, dobimo

$$(A^{(1)} - b^{(1)}k^{(1)} - \lambda_1 I)R_{n-1,n}\cdots R_{12} = \begin{bmatrix} 0 & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \\ & & & & \times \end{bmatrix}.$$

To pomeni

$$(A^{(1)} - \lambda_1 I)Q_1 = \begin{bmatrix} t_{11} & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \\ & & & & \times \end{bmatrix}, \qquad b^{(1)}k^{(1)}Q_1 = \begin{bmatrix} b_1k_1 & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \\ & & & & \times \end{bmatrix},$$

kjer je $Q_1 = R_{n-1,n}\cdots R_{12}$ in $k^{(1)}Q_1 = [k_1 \quad k^{(2)}]$. Izbrati moramo $k_1 = t_{11}/b_1$.

Po QR koraku dobimo

$$Q_1^H(A^{(1)} - b^{(1)}k^{(1)})Q_1 = \begin{bmatrix} \lambda_1 & \times & \cdots & \times \\ 0 & & & \\ \vdots & & A^{(2)} - b^{(2)}k^{(2)} & \\ 0 & & & \end{bmatrix}, \qquad Q_1^T b = \begin{bmatrix} \widetilde{b}_1 \\ b_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

kjer je $A^{(2)}$ ireducibilna Hessenbergova in $b^{(2)} = [\, b_2 \quad 0 \quad \cdots \quad 0 \,]^T$. Ponovimo postopek in določimo prvi element $k^{(2)}$. Na koncu ostane 1×1 matrika $A^{(n)}$ iz katere dobimo $k^{(n)} = k_n$ iz zveze

$$A^{(n)} - b^{(n)}k^{(n)} = \lambda_n.$$

Potem moramo še računati nazaj. Iz $k^{(j)}Q_j = [\, k_j \quad k^{(j+1)} \,]$ sledi

$$k^{(j)} = [\, k_j \quad k^{(j+1)} \,]\, Q_j^H$$

za $j = n-1, \ldots, 2, 1$, vse pa se začne s $k^{(n)} = k_n$.

Celotni algoritem je zapisan v algoritmu 7.3.

Algoritem 7.3 Razporejanje polov z modificiranim QR algoritmom

$A^{(1)} = A$
$j = 1, 2, \ldots, n-1$
  izračunaj unitarno matriko $Q_j = R_{12} \cdots R_{n-j,n-j+1}$ velikosti $(n-j+1) \times (n-j+1)$,
  da je $(A^{(j)} - \lambda_j I)Q_j = T^{(j)}$ zgornja trikotna
  $t_{jj} = (T^{(j)})_{jj}$.
  $k_j = t_{jj}/b_j$
  $\begin{bmatrix} \times \\ b^{(j+1)} \end{bmatrix} = R_{12}^H \begin{bmatrix} b_j \\ 0 \end{bmatrix}$

  $\begin{bmatrix} \times & \times & \cdots & \times \\ 0 & & & \\ \vdots & & A^{(j+1)} & \\ 0 & & & \end{bmatrix} = Q_j^H T^{(j)} + \lambda_j I$

$k^{(n)} = (A^{(n)} - \lambda_n)/b_n$
$j = n-1, \ldots, 2, 1$
  $k^{(j)} = [\, k_j \quad k^{(j+1)} \,]\, Q_j^H$
$k = k^{(1)}$.

7.5 Razporejanje polov večvhodnih sistemov

7.5.1 Razporejanje polov preko Hessenbergove forme

Tukaj manjka še: Razporejanje polov preko Hessenbergove forme.

7.5.2 Metoda ortogonalnih transformacij na lastnih vektorjih

Tukaj manjka še: Metoda ortogonalnih transformacij na lastnih vektorjih.

7.5.3 Razporejanje polov preko Schurove forme

Pri tem algoritmu namesto vodljivostne Hessenbergove forme uporabimo Schurovo formo matrike A. Naj bo par (A, B) vodljiv in naj bo

$$\widetilde{A} = QAQ^T = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \quad \widetilde{B} = QB = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix},$$

kjer sta A_{11} in A_{22} kvazi zgornje trikotni matriki.

Denimo, da smo diagonalne elemente (kjer dopuščamo tudi bloke 2×2) v Schurovi formi razporedili tako, da imamo v A_{11} že dobre lastne vrednosti, ki jih ne bi radi več spreminjali, v A_{22} pa so slabe lastne vrednosti, ki bi jih radi premaknili.

Če je $\widetilde{K} = [\,0 \quad K_2\,]$, potem dobimo

$$\widetilde{A} - \widetilde{B}\widetilde{K} = \begin{bmatrix} A_{11} & A_{12} - B_1K_2 \\ 0 & A_{22} - B_2K_2 \end{bmatrix}.$$

Vidimo, da se je A_{22} premaknila v $A_{22} - B_2K_2$, s tem pa so se spremenile tudi tiste lastne vrednosti A, ki so pripadale bloku A_{22}. Iz vodljivosti para $(\widetilde{A}, \widetilde{B})$ sledi tudi vodljivost para (A_{22}, B_2). Dovolj je torej rešiti le problem razporeditve polov za par (A_{22}, B_2). Od tod dobimo K_2, potem pa je $K = [\,0 \quad K_2\,]\,Q$.

Na začetku moramo lastne vrednosti A razdeliti na dva dela, prvi del bo v A_{11}, drugi del pa v A_{22}. Ker pri računanju Schurove forme matrike A dobimo razporeditev, ki ni pravilno razdeljena na bloka, moramo najprej ustrezno preurediti lastne vrednosti na diagonali $\widetilde{A}$. Poiščemo ortogonalno matriko U, da ima matrika $U\widetilde{A}U^T$ pravilno razporeditev.

Preurejanje izvedemo z zamenjavami sosednjih lastnih vrednosti. Če imamo matriko $\begin{bmatrix} \lambda_1 & r \\ 0 & \lambda_2 \end{bmatrix}$ in poiščemo rotacijo R_{12}, da je $R_{12}^T \begin{bmatrix} r \\ \lambda_2 - \lambda_1 \end{bmatrix} = \begin{bmatrix} \times \\ 0 \end{bmatrix}$, potem je

$$R_{12}^T A R_{12} = \begin{bmatrix} \lambda_2 & r \\ 0 & \lambda_1 \end{bmatrix}.$$

V primeru $n \times n$ matrike lahko zamenjamo sosednja 1×1 diagonalna bloka:

$$R_{p,p+1}^T \begin{bmatrix} \lambda_1 & \cdots & \cdots & \cdots & \cdots & \times \\ & \ddots & & & & \vdots \\ & & \lambda_p & & & \vdots \\ & & & \lambda_{p+1} & & \vdots \\ & & & & \ddots & \vdots \\ & & & & & \lambda_n \end{bmatrix} R_{p,p+1} = \begin{bmatrix} \lambda_1 & \cdots & \cdots & \cdots & \cdots & \times \\ & \ddots & & & & \vdots \\ & & \lambda_{p+1} & & & \vdots \\ & & & \lambda_p & & \vdots \\ & & & & \ddots & \vdots \\ & & & & & \lambda_n \end{bmatrix}.$$

Poljubno permutacijo lahko sestavimo iz transpozicij, torej lahko lastne vrednosti poljubno preuredimo. V primeru kompleksnih lastnih vrednosti imamo še dodatne algoritme za izmenjavo diagonalnih blokov velikosti 2×1 ali 2×2.

Poglejmo si, kako lahko razporedimo lastne vrednosti v primeru matrik 1×1 ali 2×2.

Naj bo M matrika $p \times p$, kjer je $p = 1$ ali $p = 2$. Naj bo G matrika $p \times m$ in $\Lambda = \{\lambda_1, \ldots, \lambda_p\}$, pri čemer velja $\Lambda = \overline{\Lambda}$. Algoritem za matriko povratne zveze F je:

$G = U\,[\,\widetilde{G} \quad 0\,]\,V^T$, kjer je $\widetilde{G}$ matrika $r \times r$ in $r = \mathrm{rang}(G)$
$\widetilde{M} = U^T M U$
Če je $r = p$, potem
    izberi poljubno $p \times p$ matriko T z lastnimi vrednostmi v Λ.
    $\widetilde{F} = \widetilde{G}^{-1}(\widetilde{M} - T)$
sicer (edina možnost je $r = 1$, $p = 2$) izračunaj $\widetilde{F} = [\,f_1 \quad f_2\,]$
    $\widetilde{M} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}$, $\widetilde{G} = [\beta]$
    $f_1 = (m_{11} + m_{22} - \lambda_1 - \lambda_2)/\beta$
    $f_2 = \left(\frac{m_{22}}{m_{21}}\right) f_1 - (m_{11}m_{22} - m_{12}m_{21} - \lambda_1\lambda_2)/(m_{21}\beta)$
$F = V \begin{bmatrix} \widetilde{F} \\ 0 \end{bmatrix} U^T$

Sedaj si lahko pogledamo, kako lahko preuredimo lastne vrednosti na matriki ??

Pretvori A v urejeno Schurovo formo: $A = QAQ^T = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$,
    da so v A_{11} dobre lastne vrednosti $\lambda_1, \ldots, \lambda_r$, v A_{22} pa $\lambda_{r+1}, \ldots, \lambda_n$
$B = QB, \quad \widehat{Q} = Q$
$K = 0, \quad i = r + 1$
Dokler je $i \le n$ ponavljaj:
    Za M vzemi zadnji blok A velikosti $p \times p$, kjer je $p = 1$ ali $p = 2$
    Za G vzemi zadnjih p vrstic B
    Izračunaj matriko F, ki premakne p lastnih vrednosti
    Posodobi K in A: $K = K - [\,0 \quad F\,]\,\widehat{Q}, \quad A = A - B\,[\,0 \quad F\,]$
    Premakni zadnji blok A na mesto (i, i), transformacije shrani v Q
    Posodobi B in $\widehat{Q}$: $B = QB, \quad \widehat{Q} = Q\widehat{Q}$
    $i = i + p$

Pri problemu *PEVA* (Partial Eigenvalue Assignment) imamo dano matriko $A \in \mathbb{R}^{n \times n}$, matriko $B \in \mathbb{R}^{n \times m}$, del spektra $\{\lambda_1, \ldots, \lambda_p\}$ matrike A, kjer je $p < n$ in pa želene lastne vrednosti $\{\mu_1, \ldots, \mu_p\}$. Cilj je poiskati matriko povratne zveze K, da bo matrika $A - BK$ imela lastne vrednosti $\{\mu_1, \ldots, \mu_p, \lambda_{p+1}, \ldots, \lambda_n\}$.

Problem lahko rešimo s prejšnjim algoritmom, kjer smo pole razporejali s pomočjo Schurove forme. To je lepa lastnost tega algoritma, kjer lahko premaknemo le izbrani del spektra, drugi del pa ostane nespremenjen.

Vseeno pa algoritem ni najbolj ekonomičen, saj moramo najprej pretvoriti A v Schurovo formo, kar v praksi pomeni, da izračunamo vse lastne vrednosti matrike A. V primeru, ko je matrika A velikih dimenzij, so primernejši algoritmi za reševanje problema PEVA, kjer izračunamo samo toliko lastnih vrednosti kot jih je potrebno zamenjati.

Primer uporabe bi npr. bil, da za sistem poiščemo samo lastne vrednosti s pozitivnim realnim delom in jih premaknemo v levo polravnino.

7.6 Pogojenost polov zaprtozančnega sistema

Zanima nas, za koliko se bodo poli zaprtozančnega sistema z izračunano matriko povratne zveze $\widehat{K}$ razlikovali od želenih polov Λ, ki so bili vhodni podatek za izračun K.

Pri računanju K lahko v najboljšem primeru pričakujemo, da je algoritem obratno stabilen in da smo torej izračunali točen $\widehat{K}$ za bližnji par $\widehat{A}$ in $\widehat{B}$. Imamo $\widehat{K} = K + \Delta K$, $\widehat{A} = A + \Delta A$ in $\widehat{B} = B + \Delta B$, kjer sta $\|\Delta A\|$ in $\|\Delta B\|$ majhni glede na $\|A\|$ oziroma $\|B\|$, matrika $\widehat{F} = \widehat{A} - \widehat{B}\widehat{K}$ pa ima predpisane lastne vrednosti $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$.

V zaprtozančnem sistemu imamo potem matriko $\widetilde{F} = A - B\widehat{K}$ in zanima nas, za koliko se njene lastne vrednosti razlikujejo od predpisanih Λ.

Če označimo $F = A - BK$, potem velja $F - \widetilde{F} = B\Delta K$. Če so vse lastne vrednosti F enostavne in je X matrika lastnih vektorjev za F, potem nam Bauer-Fikeov izrek pove, da za lastne vrednosti μ matrike $\widetilde{F}$ velja

$$\min_i |\mu - \lambda_i| \leq \kappa_2(X)\|B\Delta K\|_2.$$

V primeru velike občutljivosti lastnih vektorjev matrike F ali velike norme $B\Delta K$ lahko pričakujemo velika odstopanja lastnih vrednosti $\widetilde{F}$ od predpisanih $\lambda_1, \ldots, \lambda_n$.

Na odstopanja lastnih vrednosti $\widetilde{F}$ od predpisanih $\lambda_1, \ldots, \lambda_n$ vplivajo številni faktorji, ki so tudi medsebojno povezani. Prvi faktor je občutljivost samega problema izračuna matrike povratne zveze K.

Samo računanje K je lahko zelo občutljivo, kar pomeni, da pri danih matrikah A, B in predpisanem spektru Λ majhne motnje ΔA, ΔB ali $\Delta\Lambda$ lahko povzročijo velike motnje ΔK matrike K.

Sun je razvil naslednjo teorijo za oceno spremembe K.

Naj bo $A \in \mathbb{R}^{n\times n}$, $B = [\,b_1 \;\; \cdots \;\; b_m\,] \in \mathbb{R}^{m\times n}$, $K = [\,k_1 \;\; \cdots \;\; k_m\,]^T \in \mathbb{R}^{n\times m}$. Vse lastne vrednosti $A - BK$ naj bodo enostavne, $X = [\,x_1 \;\; \cdots \;\; x_n\,]$ naj bo matrika normiranih lastnih vektorjev, da je

$$A - BK = X\Lambda X^{-1},$$

kjer je $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Naj bodo $Y = [\,y_1 \;\; \cdots \;\; y_n\,]$ levi lastni vektorji, normirani tako, da je $y_i^T x_j = \delta_{ij}$ za vse i, j. Motnje ΔA, ΔB in $\Delta\Lambda$ naj bodo tako majhne, da ima zmoteni zaprotzančni sistem še vedno enostavne lastne vrednosti.

$$\begin{aligned} W_\lambda &= [\,S_1X^T \quad \cdots \quad S_mX^T\,] \in \mathbb{R}^{n\times mn}, & S_j &= \operatorname{diag}(y_1^Tb_j,\ldots,y_n^Tb_j),\\ W_a &= [\,D_1(X)X^{-1} \quad \cdots \quad D_n(X)X^{-1}\,] \in \mathbb{R}^{n\times n^2}, & D_i(X) &= \operatorname{diag}(x_{i1},\ldots,x_{in}),\\ W_b &= \operatorname{diag}(T_1X^{-1},\ldots,T_mX^{-1}) \in \mathbb{R}^{n\times mn}, & T_j &= \operatorname{diag}(k_j^Tx_1,\ldots,k_j^Tx_n), \end{aligned}$$

$$Z = W_\lambda^+, \qquad \Phi = -ZW_a, \qquad \Psi = -ZW_b,$$

predpostavimo, da sta para (A,B) in $(\widetilde{A},\widetilde{B})$ vodljiva, množici $S=\{\lambda_1,\ldots,\lambda_n\}$ in $\widetilde{S}=\{\widetilde{\lambda}_1,\ldots,\widetilde{\lambda}_n\}$ pa zaprti za konjugiranje.

Če je K matrika povratne zveze za A,B,S, potem obstaja matrika povratne zveze $\widetilde{K}$ za $\widetilde{A},\widetilde{B},\widetilde{S}$, da je

$$\|\widetilde{K}-K\| \le \delta_K + \mathcal{O}\left(\left\|\begin{bmatrix}\widetilde{a}\\ \widetilde{b}\\ \widetilde{\lambda}\end{bmatrix} - \begin{bmatrix}a\\ b\\ \lambda\end{bmatrix}\right\|^2\right) \le \Delta_K + \mathcal{O}\left(\left\|\begin{bmatrix}\widetilde{a}\\ \widetilde{b}\\ \widetilde{\lambda}\end{bmatrix} - \begin{bmatrix}a\\ b\\ \lambda\end{bmatrix}\right\|^2\right),$$

kjer je

$$\begin{gathered} a=\operatorname{vec}(A), \quad b=\operatorname{vec}(B), \quad \lambda=[\lambda_1 \ \cdots \ \lambda_n]^T,\\ \widetilde{a}=\operatorname{vec}(\widetilde{A}), \quad \widetilde{b}=\operatorname{vec}(\widetilde{B}), \quad \widetilde{\lambda}=[\widetilde{\lambda}_1 \ \cdots \ \widetilde{\lambda}_n]^T,\\ \delta_k = \|\Phi(\widetilde{a}-a)+\Psi(\widetilde{b}-b)+Z(\widetilde{\lambda}-\lambda)\|,\\ \Delta_k = \|\Phi\|\cdot\|\widetilde{a}-a\| + \|\Psi\|\cdot\|\widetilde{b}-b\| + \|Z\|\cdot\|\widetilde{\lambda}-\lambda\|. \end{gathered}$$

Če označimo $\kappa_A(K)=\|\Phi\|$, $\kappa_B(K)=\|\Psi\|$, $\kappa_\lambda(K)=\|Z\|$, potem lahko definiramo absolutno pogojenostno število K kot

$$\kappa_K(K) = \left(\kappa_A(K)^2+\kappa_B(K)^2+\kappa_\lambda(K)^2\right)^{1/2}.$$

Veljata oceni

$$\begin{aligned} \kappa_A(K) &= \|\Phi\| \le \|Z\|_2\|X^{-1}\|_2,\\ \kappa_B(K) &= \|\Psi\| \le \max_j\|k_j\|_2\|Z\|_2\|X^{-1}\|_2. \end{aligned}$$

V Frobeniusovi normi dobimo relativno pogojenostno število

$$\kappa_K^{(r)}(K) = \left(\kappa_A^{(r)}(K)^2+\kappa_B^{(r)}(K)^2+\kappa_\lambda^{(r)}(K)^2\right)^{1/2},$$

kjer je

$$\kappa_A^{(r)}(K) = \kappa_A(K)\frac{\|A\|_F}{\|K\|_F}, \quad \kappa_B^{(r)}(K) = \kappa_B(K)\frac{\|B\|_F}{\|K\|_F}, \quad \kappa_\lambda^{(r)}(K) = \kappa_\lambda(K)\frac{\|\lambda\|_F}{\|K\|_F}.$$

Zgled 7.1 *Za podatke*

$$A = \begin{bmatrix} -4 & & & & \\ 0.001 & -3 & & & \\ & 0.001 & -3 & & \\ & & 0.001 & -1 & \\ & & & 0.001 & 0 \end{bmatrix}, \quad B = \begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix},$$

$$\Lambda = \{-2.992, -0.8808, -2, -1, 7.0032 \cdot 10^{-14}\}$$

dobimo

$$K = [\,3.12 \quad -1.67 \quad 7.45 \quad -2.8 \quad 0.37\,].$$

Če sedaj spremenimo prvo iskano lastno vrednosti iz -2.992 v -3 dobimo

$$\widehat{K} = [\,3.1199 \quad 0.0078 \quad 7.8345 \quad 0.0004 \quad 0.3701\,].$$

□

Zgled 7.2 *Za podatke*

$$A = \begin{bmatrix} 0.1 & & & \\ 0.01 & 0.1 & & \\ & 0.01 & 0.1 & \\ & & 1 & 0.1 \end{bmatrix}, \quad B = \begin{bmatrix} 100 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

$$\Lambda = \{-0.15, -0.20, -0.25, -0.30\}$$

dobimo

$$K = [\,0.013 \quad 0.6275 \quad 13.325 \quad 1.05\,].$$

Če sedaj spremenimo element a_{31} z 0 na -0.001, potem dobimo

$$\widehat{K} = [\,0.013 \quad 3.01 \quad 23.825 \quad 1.05\,].$$

Željeni poli so lepo separirani, tako da to ni razlog za velike motnje. □

Naslednji faktor, ki vpliva na odstopanje zaprtozančnih polov od predpisanih, je norma matrike K.

Kadar imajo elementi matrike velike absolutne vrednosti oziroma je norma $\|K\|$ velika, potem lahko za izračunano matriko velja, da je relativna napaka $\frac{\|\Delta K\|}{\|K\|}$ res majhna, ampak $\|\Delta K\|$ je lahko velika v primerjavi z $|\lambda|$. Potem imamo zaradi ocene iz Bauer-Fikeovega izreka lahko velika odstopanja polov $\widetilde{F}$ od ciljnih.

K ima lahko veliko normo kadar

- je par (A, B) blizu nevodljivemu paru,
- so predpisani poli daleč od polov odprtozančnega sistema.

Velja tudi naslednja ocena. Točne lastne vrednosti

$$\widehat{F} = A + \Delta A - (B + \Delta B)\widehat{K} = \widetilde{F} + \Delta A - \Delta B\widehat{K}$$

so želene lastne vrednosti Λ. Odstopanja lastnih vrednosti $\widetilde{F}$ od želenih je tako odvisno tudi od razmerja med $\|\Delta A - \Delta B\widehat{K}\|$ in $\|\widetilde{F}\|$.

Zgled 7.3 *Če vzamemo*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & a_{22} \end{bmatrix}, \qquad b = \begin{bmatrix} b_1 \\ \mu \end{bmatrix},$$

kjer je par (A_{11}, b_1) *vodljiv in je* μ *majhen, potem za* $k = [\, k_1 \quad k_2 \,]$ *dobimo*

$$A - bk = \begin{bmatrix} A_{11} - b_1 k_1 & A_{12} - b_1 k_2 \\ -\mu k_1 & a_{22} - \mu k_2 \end{bmatrix}.$$

Ko gre $\mu \to 0$*, se manjša oddaljenost* (A, b) *od nevodljivih sistemov, po drugi strani pa* $|k_2|$ *raste prek vseh meja, saj mora v primeru* $\lambda_2 \neq a_{22}$ *veljati* $k_2 \approx (\lambda - a_{22})/\mu$*.*

Ko se manjša oddaljenost od nevodljivega sistema, se veča norma matrike povratne zveze K*, torej, bolj ko je sistem nevodljiv, močnejši mora biti vhodni signal, da ga lahko kontroliramo.* □

Velika odstopanja lahko pričakujemo tudi v primerih, ko so lastne vrednosti $A - BK$ zelo občutljive.

Če je λ_i enostavna lastna vrednost $A - BK$, potem je njeno pogojenostno število enako

$$\frac{1}{y_i^* x_i},$$

kjer sta x_i in y_i normirana desni in levi lastni vektor za λ_i.

V primeru, ko so vse lastne vrednosti enostavne, je ocena za odstopanje tudi občutljivost matrike lastnih vektorjev X.

V primeru večkratnih lastnih vrednosti so motnje še večje, saj so niso več proporcionalne $\mathcal{O}(\epsilon)$, kjer je ϵ velikost motnje, temveč so lahko proporcionalne $\mathcal{O}(\epsilon^{1/k})$, kjer je k večkratnost lastne vrednosti. Tudi zaradi tega ponavadi ni najbolje, če želimo imeti v zaprtozančnem sistemu večkratne pole.

V primeru enovhodnega sistema je matrika K enolično določena z A, B in Λ. Pri fiksnem Λ tako ne moremo vplivati na občutljivost računanja K. Edino, na kar lahko vplivamo je, da poskušamo izbrati novo razporeditev polov tako, da bodo potem čimbolj neobčutljivi.

Drugače je pri večvhodnih sistemih, ko matrika K ni enolična. Tu imamo na voljo več svobode in lahko poiščemo tak K, da bo npr. matrika lastnih vektorjev zaprtozančnega sistema imela čim manjšo občutljivost. Druga možnost je, da npr. poskusimo poiskati K s čim manjšo normo, saj nam lahko tudi velika norma K povzroča težave.

Na koncu moramo upoštevati tudi to, da na začetku iskanja matrike povratne zveze K ponavadi matriki A in B transformiramo v enostavnejšo obliko. Če za to uporabimo neortogonalne transformacije, kot npr. pri Ackermanovi formuli, potem lahko pričakujemo velika odstopanja.

7.7 Robustno razporejanje polov

V primeru večvhodnega sistema matrika povratne zveze K ni več enolična. Zato jo poskusimo pri izbranem spektru Λ določiti tako, da bodo poli zaprtozančnega sistema čim manj občutljivi.

Naj se da $F = A - BK$ diagonalizirati, torej obstaja nesingularna matrika X, da je

$$X^{-1}(A - BK)X = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

Vemo, da je v primeru enostavne lastne vrednosti λ_j njeno pogojenostno število enako

$$c_j = \frac{1}{s_j} = \frac{\|x_j\|_2 \|y_j\|_2}{|y_j^* x_j|},$$

kjer sta x_j in y_j desni in levi lastni vektor za λ_j. Vemo tudi, da za vsak j velja

$$c_j \leq \|X\|_2 \|X^{-1}\|_2.$$

Tako dobimo nekaj variant za zmanjšanje občutljivosti:

1. $\nu_1 = \|c\|_\infty = |c_1| + \cdots + |c_n|$,
2. $\nu_2 = \|X\| \cdot \|X^{-1}\|$,
3. $\nu_3 = \|X^{-1}\|_F \sqrt{n} = \|C\|_2 \sqrt{n}$.

Velja

$$1 \leq \nu_3 \leq \nu_1 \leq \nu_2 \leq n\nu_3,$$

tako da so vse mere med seboj ekvivalentne.

Problem *REVA* (Robust Eigenvalue Assignment) lahko podamo kot:

Za dani matriki $A \in \mathbb{R}^{n\times n}$ in $B \in \mathbb{R}^{n\times m}$, $m \leq n$, ki je polnega ranga, poišči $K \in \mathbb{R}^{m\times n}$ in nesingularno matriko X, da bo

$$(A - BK)X = X\Lambda$$

in bo ena izmed mer občutljivosti zaprtozančnih polov minimalna.

Izrek 7.4 *Če je X nesingularna, potem obstaja nesingularna matrika K, ki zadošča $(A-BK)X = X\Lambda$ natanko tedaj, ko je*

$$U_1^T(X\Lambda - AX) = 0,$$

kjer je

$$B = [\,U_0 \quad U_1\,] \begin{bmatrix} Z \\ 0 \end{bmatrix}$$

QR razcep matrike B, matrika Z je nesingularna, U_0 in U_1 pa imata ortonormirane stolpce. Potem je K podana eksplicitno z

$$K = Z^{-1}U_0^T(A - X\Lambda X^{-1}).$$

Sedaj lahko zapišemo algoritem za problem REVA. Vhodni podatki so $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $m \leq n$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, pri čemer je B polnega ranga, par (A, B) je vodljiv, Λ pa je zaprta za konjugiranje.

Izhodni podatki so matrika povratne zveze $K \in \mathbb{R}^{m\times n}$, določena tako, da ima $A - BK$ lastne vrednosti $\lambda_1, \ldots, \lambda_n$ in je občutljivost matrike lastnih vektorjev čim manjša.

1. QR razcep za B, da dobimo $B = [\,U_0 \quad U_1\,] \begin{bmatrix} Z \\ 0 \end{bmatrix}$ in od tod U_0, U_1 in Z.

   Za podprostor $\mathcal{S}_j = \ker(U_1^T(A - \lambda_j I))$ in njegov komplement $\widehat{\mathcal{S}}_j$ skontruiramo ortonormirano bazo, ki sestavlja stolpce S_j in $\widehat{S}_j$, za $j = 1, \ldots, n$.

2. Iz podprostorov $\mathcal{S}_j$ izberemo normirane vektorje $x_1, \ldots, x_n$, da je $X = [\,x_1 \quad \cdots \quad x_n\,]$ čim bolje pogojena.

3. Rešimo linearni sistem $FX = X\Lambda$, odtod dobimo F.

4. Izračunamo $K = Z^{-1}U_0^T(A - F)$.

Za dobljeni K velja

$$\|K\|_2 \leq (\|A\|_2 + \max_j |\lambda_j| \kappa_2(X)) \frac{1}{\sigma_{\min}(B)}.$$

V koraku 1. za računanje S_j in $\widehat{S}_j$ lahko uporabimo QR razcep ali pa singularni razcep.

a) Če uporabimo QR razcep, je $(U_1^T(A - \lambda_j I))^T = [\,\widehat{S}_j \quad S_j\,] \begin{bmatrix} R_j \\ 0 \end{bmatrix}$.

b) Če uporabimo singularni razcep, je $U_1^T(A - \lambda_j I) = T_j\,[\,\Gamma_j \quad 0\,]\,[\,\widehat{S}_j \quad S_j\,]^T$.

Opazimo lahko, da večkratnost predpisane lastne vrednosti λ_j ne more biti večja of m, saj je to maksimalno število linearno neodvisnih vektorjev, ki jih lahko izberemo za λ_j. Velja namreč

$$U^T\,[\,B \quad A - \lambda_j I\,] = \begin{bmatrix} Z & U_0^T(A - \lambda_j I) \\ 0 & U_1^T(A - \lambda_j I) \end{bmatrix}.$$

Ker je $\operatorname{rang}(Z) = m$ in $\operatorname{rang}([\,B \quad A - \lambda_j I\,]) = n - m$, potem je $\operatorname{rang}(U_1^T(A - \lambda_j I)) = n - m$ in enačba $U_1^T(A - \lambda_j I)v_j = 0$ ima lahko največ m linearno neodvisnih rešitev.

Glavni del algoritma je v koraku 2. Tukaj imamo različne pristope, ki so odvisni od količine, ki jo želimo zmanjšati. Najenostavnejši primer je, da želimo minimizirati $\kappa_2(X)$.

$\kappa_2(X)$ bo minimalna, če bodo koti med x_j in podprostori $\mathcal{T}_j = \operatorname{Lin}\{x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n\}$ čim večji.

Vektorje $x_1, \ldots, x_n$ izberemo naključno. V zanki gremo po $j = 1, \ldots, n$ in zamenjamo x_j z novim vektorjem iz $\mathcal{T}_j$, da bo kot med x_j in $\mathcal{T}_j$ maksimalen.

Izračunamo QR razcep $X_j = [\,Q_j \quad y_j\,] \begin{bmatrix} R_j \\ 0 \end{bmatrix}$, kjer je $X_j = [\,x_1 \quad \cdots \quad x_{j-1} \quad x_{j+1} \quad \cdots \quad x_n\,]$, in vzamemo

$$x_j = \frac{S_j S_j^T y_j}{\|S_j^T y_j\|_2}.$$

Ta izbira maksimizira $\frac{1}{|y_j^T x_j|}$. Tako gremo v ciklu skozi $j = 1, \ldots, n$. Če je na koncu cikla $\kappa_2(X)$ sprejemljiva, končamo, sicer pa nadaljujemo z novim ciklom. Na dolgi rok proces konvergira proti optimalni matriki K.

Zgled 7.4 *Za*

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, \quad B = \begin{bmatrix} 6 & 3 \\ 1 & 2 \\ 8 & 9 \end{bmatrix}, \qquad \Lambda = \{9, 5, 1\}$$

dobimo

$$K = \begin{bmatrix} -1.8988 & 0.0269 & 0.5365 \\ 2.4501 & 0.5866 & -0.1611 \end{bmatrix}$$

in $\kappa_2(K) = 6.32$.

□

7.8 Optimalno vodenje

Imamo sistem

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \qquad x(0) = x_0, \\ y(t) &= Cx(t). \end{aligned}$$

Če je sistem vodljiv, potem lahko določimo matriko povratne zveze K tako, da (do konjugiranja) poljubno razporedimo lastne vrednosti. Če vse pole prestavimo v levo polravnino $\mathbb{C}$, bo zaprtozančni sistem z matriko $A - BK$ stabilen.

Stabilnost pa je le ena izmed željenih lastnosti. Zanima nas, kako naj določimo določimo K, da bo imel zaprtozančni sistem še druge lastnosti. Dve možnosti sta:

- Že vnaprej poznamo območje, kjer naj ležijo poli zaprtozančnega sistema.
- Matriko K določimo tako, da poiščemo minimum izbranega optimizacijskega kriterija.

Denimo, da si želimo, da bi bile komponente vektorja stanja čim manjše po absolutni vrednosti. V tem primeru želimo minimizirati npr.

$$J_1(x) = \int_0^\infty x^T(t)x(t)dt,$$

obenem pa bi radi, da bo sistem tudi stabilen.

Če želimo, da bo izhod majhen, je naš cilj minimizirati

$$J_2(x) = \int_0^\infty y^T(t)y(t)dt = \int_0^\infty x^T(t)C^TCx(t)dt = \int_0^\infty x^T(t)Qx(t)dt,$$

kjer je $Q = C^TC$ simetrična pozitivno semidefinitna matrika.

Podobno lahko želimo, da vhod ne bo prevelik, zato bi radi minimizirali

$$J_3(x) = \int_0^\infty u^T(t)u(t)dt,$$

oziroma bolj splošno

$$J_4(x) = \int_0^\infty u^T(t)Ru(t)dt,$$

kjer je R simetrična pozitivno definitna matrika.

Izkaže se, da istočasno ne moremo zadostiti vsem kriterijem. Tako npr. ne moremo hkrati minimizirati J_1 in J_3, saj za minimizacijo J_1 potrebujemo močan, za minimizacijo J_3 pa čim šibkejši vhodni signal. Zato vzamemo neko konveksno kombinacijo J_1 in J_3 oziroma J_4. Tako pridemo do *problema linearne kvadratične optimizacije (LQR)*, kjer za dani matriki Q in R, kjer je Q simetrična pozitivno semidefinitna, R pa simetrična pozitivno definitna, iščemo vhod $u(t)$, ki minimizira

$$J_C(x) = \int_0^\infty \left(x^T(t)Qx(t) + u^T(t)Ru(t)\right) dt$$

pr pogoju $\dot{x}(t) = Ax(t) + Bu(t)$, $x(0) = x_0$.

Tega, kako izberemo Q in R, ne bomo obravnavali, saj je to odvisno od posameznih primerov. Matrika Q predstavlja utež za stanje, R pa za vhod. Če izberemo po normi velik R, bomo dobili majhen vhod. Z matriko Q lahko tudi izločimo nezaželjena stanja, če jih primerno utežimo.

Obstajajo še drugačne izbire kriterijske funkcije. Če npr. opazujemo le končno obdobje $[0, t_1]$, si lahko želimo, da je pri t_1 končno stanje čim bližje 0. V tem primeru vzamemo

$$J(x) = \frac{1}{2}x^T(t_1)^T Fx(t_1) + \frac{1}{2}\int_0^{t_1} \left(x^T(t)Qx(t) + u^T(t)Ru(t)\right) dt,$$

kjer je F simetrična pozitivno definitna matrika.

Druga možnost je, da si želimo, da se stanje čim bolj prilega željenemu stanju $x_d(t)$. Sedaj iščemo minimum

$$J(x) = \frac{1}{2}(x(t_1) - x_d(t_1))^T F(x(t_1) - x_d(t_1))$$
$$+ \frac{1}{2}\int_0^{t_1} \left((x(t) - x_d(t))^T Q(x(t) - x_d(t)) + u^T(t)Ru(t)\right) dt.$$

Naslednji izrek prevede iskanje optimalnega vodenja na reševanje algebraične Riccatijeve enačbe.

Izrek 7.5 *Dan je sistem $\dot{x}(t) = Ax(t) + Bu(t)$, $x(0) = x_0$ in matriki Q, R, kjer je Q simetrična pozitivno semidefinitna, R pa simetrična pozitivno definitna. Naj bo par (A, B) stabilizabilen, par (A, Q) pa zaznaven. Potem obstaja enoličen vhod $\widetilde{u}(t)$, ki minimizira*

$$J_C(x) = \int_0^\infty \left(x^T(t)Qx(t) + u^T(t)Ru(t)\right) dt.$$

Podan je z $\widetilde{u}(t) = -Kx(t)$, kjer je $K = R^{-1}B^TP$ in je P enolična simetrična pozitivno semidefinitna rešitev zvezne algebraične Riccatijeve enačbe (CARE)

$$PA + A^TP + Q - PBR^{-1}B^TP = 0.$$

Še več, zaprtozančna matrika $A - BK$ je stabilna, minimalna vrednost $J_C(x)$ pa je enaka $x_0^TPx_0$.

To je glavni izrek tega razdelka. Dokazali ga bomo po kosih. Najprej bomo pokazali, da stabilizirajoča simetrična pozitivno semidefinitna rešitev CARE res minimizira J_C, obstoj in enoličnost take rešitve pa bomo dokazali kasneje.

Lema 7.6 *Pri predpostavkah izreka o zveznem LQR, naj bo P taka simetrična pozitivno semidefinitna rešitev CARE $PA + A^TP + Q - PBR^{-1}B^TP = 0$, da je zaprtozančni sistem $u(t) = -Kx(t)$, kjer je $K = R^{-1}B^TP$, stabilen. Potem $u(t)$ minimizira $J_C(x)$, katerega minimum je $x_0^TPx_0$.*

Dokaz.

$$\begin{aligned}\frac{d}{dt}(x^TPx) &= \dot{x}^TPx + x^TP\dot{x} = (Ax+Bu)^TPx + x^TP(Ax+Bu)\\ &= x^T(A^TP+PA)x + u^TB^TPx + x^TPBu\\ &= x^T(PBR^{-1}B^TP - Q)x + u^TB^TPx + x^TPBu\\ &= x^TPBR^{-1}B^TPx - x^TQx + u^TB^TPx + x^TPBu + u^TRu - u^TRu - x^TQx\\ &= (u^T + x^TPBR^{-1})R(u + R^{-1}B^TPx) - (x^TQx + u^TRu).\end{aligned}$$

Od tod sledi $x^TQx + u^TRu = -\frac{d}{dt}(x^TPx) + (u+R^{-1}B^TPx)^TR(u+R^{-1}B^TPx)$. To integriramo

$$\int_0^\tau (x^TQx + u^TRu)dt = x^T(\tau)Px(\tau) + x_0^TPx_0 + \int_0^\tau (u+R^{-1}B^TPx)^TR(u+R^{-1}B^TPx)dt.$$

Ko pošljemo $\tau \to \infty$ gre zaradi stabilnosti zaprtozančnega sistema $x(\tau) \to 0$ in ostane

$$\int_0^\infty (x^TQx + u^TRu)dt = x_0^TPx_0 + \int_0^\infty (u+R^{-1}B^TPx)^TR(u+R^{-1}B^TPx)dt.$$

Očitno je minimum enak $x_0^TPx_0$, dosežen pa je pri $u = -R^{-1}B^TPx$.

V nadaljevanju bomo pokazali, da je reševanje algebraičnih Riccatijevih enačb povezano s Hamiltonskimi matrikami. Za $2n \times 2n$ matriko

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}$$

pravimo, da je *Hamiltonska*, če velja $H_{11} = -H_{22}^T$, $H_{12} = H_{12}^T$ in $H_{21} = H_{21}^T$.

Ekvivalenten pogoj je $HJ + JH^T = 0$, kjer je $J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}$.

Iz $J^{-1}HJ = -H^T$ sledi, da sta matriki H in $-H^T$ podobni. Če je λ lastna vrednost H, je tudi lastna vrednost $-H^T$. Velja

Lema 7.7 *Če je H Hamiltonska matrika in je λ njena lastna vrednost, potem je $-\overline{\lambda}$ tudi lastna vrednost H z enako algebrajsko in geometrijsko večkratnostjo.*

Za spekter tako velja, da neničelne realne lastne vrednosti nastopajo v parih $\lambda, -\lambda$, čisto imaginarne lastne vrednosti nastopajo v parih $\lambda, \overline{\lambda}$, preostale pa v četvorkah $\lambda, -\lambda, \overline{\lambda}, -\overline{\lambda}$.

Naslednji izrek povezuje CARE in Hamiltonske matrike.

Izrek 7.8 *Dana je CARE*

$$PA + A^TP + Q - PSP = 0,$$

kjer je $S = BR^{-1}B^T$. *Matrika* P *je rešitev CARE natanko tedaj, ko stolpci* $\begin{bmatrix} I \\ P \end{bmatrix}$ *razpenjajo* n*-razsežni invariantni podprostor* $2n \times 2n$ *Hamiltonske matrike*

$$H = \begin{bmatrix} A & -S \\ -Q & -A^T \end{bmatrix}.$$

Dokaz. ($\Longrightarrow$) Denimo, da je P rešitev CARE. Potem je

$$H \begin{bmatrix} I \\ P \end{bmatrix} = \begin{bmatrix} A - SP \\ PA - PSP \end{bmatrix} = \begin{bmatrix} I \\ P \end{bmatrix} (A - SP),$$

kar pomeni, da stolpci $\begin{bmatrix} I \\ P \end{bmatrix}$ res razpenjajo invariantni podprostor H.

($\Longleftarrow$) Denimo, da obstaja taka $n \times n$ matrika L, da je

$$H \begin{bmatrix} I \\ P \end{bmatrix} = \begin{bmatrix} I \\ P \end{bmatrix} L \qquad \Rightarrow \qquad J^{-1} H \begin{bmatrix} I \\ P \end{bmatrix} = J^{-1} \begin{bmatrix} I \\ P \end{bmatrix} L,$$

od tod pa sledi

$$\begin{bmatrix} Q & A^T \\ A & -S \end{bmatrix} \begin{bmatrix} I \\ P \end{bmatrix} = \begin{bmatrix} -P \\ I \end{bmatrix} L$$

in

$$[\, I \quad P \,] \begin{bmatrix} Q & A^T \\ A & -S \end{bmatrix} \begin{bmatrix} I \\ P \end{bmatrix} = [\, I \quad P \,] \begin{bmatrix} -P \\ I \end{bmatrix} L = 0,$$

kar pomeni $Q + A^T P + PA - PSP = 0$.

Posledica 7.9 *Če stolpci* $\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$ *razpenjajo* n*-razsežni invariantni podprostor* $2n \times 2n$ *Hamiltonske matrike*

$$H = \begin{bmatrix} A & -S \\ -Q & -A^T \end{bmatrix},$$

ki ustreza CARE $PA + A^T P + Q - PSP = 0$ *in je* P_1 *obrnljiva, potem je* $P = P_2 P_1^{-1}$ *rešitev CARE.*

Dokaz.

$$\mathrm{Lin}\left(\begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \right) = \mathrm{Lin}\left(\begin{bmatrix} P_1 \\ P_2 \end{bmatrix} P_1^{-1} \right) = \mathrm{Lin}\left(\begin{bmatrix} I \\ P_2 P_1^{-1} \end{bmatrix} \right),$$

potem pa je po prejšnjem izreku $P_2 P_1^{-1}$ rešitev CARE.

Izrek 7.10 *Naj bo* P *simetrična rešitev CARE*

$$PA + A^T P + Q - PSP = 0,$$

kjer je $S = BR^{-1}B^T$. *Potem so lastne vrednosti pripadajoče Hamiltonske matrike* $H = \begin{bmatrix} A & -S \\ -Q & -A^T \end{bmatrix}$ *unija lastnih vrednosti matrik* $A - BK$ *in* $-(A - BK)^T$, *kjer je* $K = R^{-1}B^T P$.

Dokaz. Matrika $T = \begin{bmatrix} I & 0 \\ P & I \end{bmatrix}$ je nesingularna. Velja

$$\begin{aligned} T^{-1}HT &= \begin{bmatrix} I & 0 \\ -P & I \end{bmatrix} \begin{bmatrix} A & -S \\ -Q & -A^T \end{bmatrix} \begin{bmatrix} I & 0 \\ P & I \end{bmatrix} \\ &= \begin{bmatrix} A - SP & -S \\ -(A^TP + PA + Q - PSP) & -(A - SP)^T \end{bmatrix} \\ &= \begin{bmatrix} A - SP & -S \\ 0 & -(A - SP)^T \end{bmatrix}. \end{aligned}$$

Pravimo, da je P *stabilizirajoča rešitev* CARE, če je $A - BR^{-1}B^TP$ stabilna. Pokazali smo že, da če imamo stabilizirajočo rešitev, potem imamo optimalno rešitev LQR. Do polnega dokaza izreka LQR nam manjkata še obstoj in enoličnost stabilizirajoče rešitve.

Izrek 7.11 *Dan je sistem $\dot{x}(t) = Ax(t) + Bu(t)$, $x(0) = x_0$ in matriki Q, R, kjer je $Q \neq 0$ simetrična pozitivno semidefinitna, R pa simetrična pozitivno definitna. Ekvivalentno je*

1. *CARE*
$$PA + A^TP + Q - PSP = 0$$
in $S = BR^{-1}B^T$, ima enolično simetrično pozitivno semidefinitno stabilizirajočo rešitev P,

2. *Par (A, B) je stabilizabilen in pridružena Hamiltonska matrika*
$$H = \begin{bmatrix} A & -S \\ -Q & -A^T \end{bmatrix}$$
nima nobene čisto imaginarne lastne vrednosti.

Dokaz. (1. $\Longrightarrow$ 2.) Naj bo P stabilizirajoča rešitev CARE. Potem je $A - BK$, kjer je $K = R^{-1}B^TP$ stabilna, torej je (A, B) stabilizabilen par. Zaradi tega imajo vse lastne vrednosti $A - BK$ negativni realni del. Ker so po izreku 7.10 lastne vrednosti H enake lastnim vrednostim $A - BK$ in $-(A - BK)^T$, potem H ne more imeti nobene čisto imaginarne lastne vrednosti.

(2. $\Longrightarrow$ 1.) Ker H nima nobenih čisto imaginarnih lastnih vrednosti, ima natanko n stabilnih lastnih vrednosti, ki imajo realni del strogo negativen. Potem obstajajo stabilna matrika E in matriki P_1, P_2, da je

$$H \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} E$$

in stolpci $\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$ razpenjajo lastni podprostor H, ki pripada stabilnim lastnim vrednostim. Pokazali bomo, da je $P = P_2P_1^{-1}$ iskana enolična stabilizirajoča s.p.sd. rešitev.

Nadaljevanje dokaza bo sestavljeno iz naslednjih točk:

a) $P_2^TP_1$ je simetrična,

b) P_1 je obrnljiva,

c) $P = P_2P_1^{-1}$ je simetrična,

č) P je stabilizirajoča rešitev,

d) P je enolična,

e) P je s.p.sd.

a) Velja

$$\begin{aligned} AP_1 - SP_2 &= P_1E, \quad (7.5)\\ -QP_1 - A^TP_2 &= P_2E. \quad (7.6)\end{aligned}$$

Enačbi pomnožimo z P_2^T oz. P_1^T in dobimo (drugo potem še transponiramo)

$$\begin{aligned} P_2^TAP_1 - P_2^TSP_2 &= P_2^TP_1E,\\ -P_1^TQP_1 - P_2^TAP_1 &= E^TP_2^TP_1.\end{aligned}$$

Enačbi seštejemo v

$$-P_2^TSP_2 - P_1^TQP_1 = P_2^TP_1E + E^TP_2^TP_1.$$

Ker sta S in Q simetrični, je leva stran simetrična, torej mora biti tudi desna in velja

$$P_2^TP_1E + E^TP_2^TP_1 = E^TP_1^TP_2 + P_1^TP_2E$$

oziroma

$$E^T(P_2^TP_1 - P_1^TP_2) + (P_2^TP_1 - P_1^TP_2)E = 0.$$

Ker je E stabilna, ima zgornja enačba Ljapunova enolično rešitev, iz katere sledi $P_2^TP_1 - P_1^TP_2 = 0$ oziroma $(P_2^TP_1)^T = P_2^TP_1$.

b) Denimo, da P_1 ni obrnljiva. Potem obstaja vektor $d \neq 0$, da je $P_1d = 0$.
Iz (7.5) dobimo (pomnožimo s P_2d in d)

$$\begin{aligned} d^TP_2^TSP_2d &= -d^TEP_1^TP_2d + d^TP_1^TA^TP_2d\\ &= -d^TE^TP_2^TP_1d + d^TP_1^TA^TP_2d = 0.\end{aligned}$$

Ker je $S = BR^{-1}B^T$ in je R s.p.d., sledi $B^TP_2d = 0$, iz (7.5) pa dobimo $P_1Ed = 0$. To velja za vsak $d \in \ker(P_1)$, zato je $\ker(P_1)$ invarianten za E. Torej obstaja lastni par $(\mu, \widetilde{d})$, $\widetilde{d} \neq 0$, za E, da je

$$E\widetilde{d} = \mu\widetilde{d}, \quad P_1\widetilde{d} = 0.$$

Če pomnožimo (7.6) z $\widetilde{d}$, dobimo

$$(\mu I + A^T)P_2\widetilde{d} = 0.$$

Skupaj z $B^TP_2\widetilde{d} = 0$ dobimo

$$\begin{bmatrix} \mu I + A^T \\ B^T \end{bmatrix} P_2\widetilde{d} = 0.$$

Ker je par (A, B) stabilizabilen, mora biti $P_2\widetilde{d} = 0$. To pa pomeni, da $\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$ ni polnega ranga, torej mora biti P_1 obrnljiva.

c) Ker je P_1 nesingularna, je po posledici 7.9 $P = P_2P_1^{-1}$ rešitev CARE. P je tudi simetrična, saj velja

$$\begin{aligned} P^T - P &= P_1^{-T}P_2^T - P_2P_1^{-1} \\ &= P_1^{-T}(P_2^TP_1)P_1^{-1} - P_1^{-T}(P_1^TP_2)P_1^{-1} \\ &= P_1^{-T}(P_2^TP_1 - P_1^TP_2)P_1^{-1} = 0. \end{aligned}$$

č) Če pomnožimo (7.5) s P_1^{-1}, dobimo

$$A - SP_2P_1^{-1} = P_1EP_1^{-1}.$$

Ker je E stabilna, mora biti tudi $A - SP_2P_1^{-1} = A - SP$ in P je stabilizirajoča rešitev.

d) Naj bosta P_a in P_b dve stabilizirajoči rešitvi. Potem je

$$\begin{aligned} A^TP_a + P_aA - P_aSP_a + Q &= 0 \\ A^TP_b + P_bA - P_bSP_b + Q &= 0. \end{aligned}$$

Z odštevanjem dobimo

$$A^T(P_a - P_b) + (P_a - P_b)A + P_2SP_2 - P_1QP_1 = 0$$

oziroma

$$(A - SP_a)^T(P_a - P_b) + (P_a - P_b)(A - SP_b) = 0.$$

Dobili smo homogeno Sylvestrovo enačbo. Ker sta matriki $A - SP_a$ in $A - SP_b$ stabilni, mora biti $P_a = P_b$.

e) Riccatijevo enačbo

$$PA + A^TP + Q - PSP = 0$$

lahko zapišemo v obliki enačbe Ljapunova

$$(A - BK)^TP + P(A - BK) = -Q - PSP,$$

kjer je $K = R^{-1}B^TX$.

Matrika $A - BK = A - BR^{-1}B^TX = A - SP$ je stabilna, zato lahko P zapišemo v obliki

$$P = \int_0^\infty e^{(A-BK)^Tt}(Q + PSP)e^{(A-BK)t}.$$

Ker sta Q in S pozitivno definitni, je potem P pozitivno semidefinitna.

Zadnji del dokaza izreka LQR je, da pokažemo, da pri pogojih izreka LQR res obstaja rešitev Riccatijeve enačbe oziroma, da pridružena Hamiltonska matrika nima čisto imaginarnih lastnih vrednosti.

Izrek 7.12 *Dan je sistem $\dot{x}(t) = Ax(t) + Bu(t)$, $x(0) = x_0$, s.p.d. matrika R in s.p.sd. matrika Q, pri čemer je par (A, B) je stabilizabilen, (A, Q) pa zaznaven. Potem pridružena Hamiltonska matrika*

$$H = \begin{bmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{bmatrix}$$

nima nobene čisto imaginarne lastne vrednosti.

Dokaz. Denimo, da imamo lastno vrednost oblike $i\alpha$, pripadajoči lastni vektor pa je $\begin{bmatrix} r \\ s \end{bmatrix}$. Iz

$$\begin{bmatrix} s^* & r^* \end{bmatrix} H \begin{bmatrix} r \\ s \end{bmatrix} = i\alpha \begin{bmatrix} s^* & r^* \end{bmatrix} \begin{bmatrix} r \\ s \end{bmatrix}$$

dobimo

$$(s^*Ar - r^*A^Ts) - r^*Qr - s^*Ss = i\alpha(s^*r + r^*s),$$

kjer je $S = BR^{-1}B^T$. Realni del je $-r^*Qr - s^*Ss = 0$ in iz definitnosti R in Q sledi

$$B^Ts = Qr = 0.$$

Iz $H \begin{bmatrix} r \\ s \end{bmatrix} = i\alpha \begin{bmatrix} r \\ s \end{bmatrix}$ ostane še $Ar = i\alpha r$ in $-A^Ts = i\alpha s$. Dobimo

$$\begin{bmatrix} A - i\alpha I \\ Q \end{bmatrix} r = 0, \qquad \begin{bmatrix} A^T + i\alpha I \\ B^T \end{bmatrix} s = 0.$$

Zaradi zaznavnosti (A, Q) in stabilizabilnosti (A, B) dobimo protislovje $r = s = 0$.

Še enkrat zapišimo celoten izrek.

Izrek 7.13 *Dan je sistem $\dot{x}(t) = Ax(t) + Bu(t)$, $x(0) = x_0$ in matriki Q, R, kjer je Q simetrična pozitivno semidefinitna, R pa simetrična pozitivno definitna. Naj bo par (A, B) stabilizabilen, par (A, Q) pa zaznaven. Potem obstaja enoličen vhod $\widetilde{u}(t)$, ki minimizira*

$$J_C(x) = \int_0^\infty \left(x^T(t)Qx(t) + u^T(t)Ru(t)\right) dt.$$

Podan je z $\widetilde{u}(t) = -Kx(t)$, kjer je $K = R^{-1}B^TP$ in je P enolična simetrična pozitivno semidefinitna rešitev CARE

$$PA + A^TP + Q - PBR^{-1}B^TP = 0.$$

Rešitev P je enaka $P = P_2P_1^{-1}$, kjer stolpci matrike $\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$ razpenjajo invariantni podprostor stabilnih lastnih vrednosti pridružene Hamiltonske matrike $H = \begin{bmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{bmatrix}$. Zaprtozančna matrika $A - BK$ je stabilna, minimalna vrednost $J_C(x)$ pa je enaka $x_0^TPx_0$.

7.8.1 Diskretni sistemi

Imamo diskretni sistem

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k, \\ y_k &= Cx_k. \end{aligned}$$

Sedaj želimo minimizirati

$$J_D(x) = \sum_{k=0}^{\infty} (x_k^TQx_k + u_k^TRu_k).$$

Rešitev je povezana z *diskretno algebraično Riccatijevo enačbo (DARE)*

$$A^TPA - P + Q - A^TPB(R + B^TPB)^{-1}B^TPA = 0.$$

Izrek 7.14 *Dan je sistem $x_{k+1} = Ax_k + Bu_k$, $y_k = Cx_k$ in matriki Q, R, kjer je Q simetrična pozitivno semidefinitna, R pa simetrična pozitivno definitna. Naj bo par (A, B) stabilizabilen, par (A, Q) pa zaznaven. Potem obstaja enoličen vhod $\widetilde{u}$, ki minimizira*

$$J_D(x) = \sum_{k=0}^{\infty} (x_k^T Q x_k + u_k^T R u_k).$$

Podan je z $\widetilde{u}_k = -Kx_k$, kjer je $K = (R + B^T P B)^{-1} B^T P A$ in je P enolična simetrična pozitivno semidefinitna rešitev DARE

$$A^T P A - P + Q - A^T P B (R + B^T P B)^{-1} B^T P A = 0.$$

Zaprtozančna matrika $A - BK$ je konvergentna, minimalna vrednost $J_D(x)$ pa je enaka $x_0^T P x_0$.

Vlogo Hamiltonskih matrik imajo sedaj simplektične matrike. Za $2n \times 2n$ matriko

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

pravimo, da je *simplektična*, če velja

$$J^{-1} M^T J = J^T M^T J = M^{-1}.$$

Iz $J^{-1} M J = -M^T$ sledi, da sta matriki M in $-M^T$ podobni. Če je λ lastna vrednost M, je tudi lastna vrednost $-H^T$. Velja

Lema 7.15 *Če je M simplektična matrika in je λ njena lastna vrednost, potem je $1/\lambda$ tudi lastna vrednost M z enako algebrajsko in geometrijsko večkratnostjo.*

Izrek 7.16 *Dana je DARE*

$$A^T P A - P + Q - A^T P B (R + B^T P B)^{-1} B^T P A = 0.$$

Matrika P je rešitev DARE natanko tedaj, ko stolpci $\begin{bmatrix} I \\ P \end{bmatrix}$ razpenjajo n-razsežni invariantni podprostor $2n \times 2n$ simplektične matrike

$$M = \begin{bmatrix} A + SA^{-T}Q & -SA^{-T} \\ -A^{-T}Q & A^{-T} \end{bmatrix}.$$

Izrek 7.17 *Naj bo par (A, B) stabilizabilen, par (A, Q) pa zaznaven, Q je simetrično pozitivno semidefiniten, R pa simetrično pozitivno definitna. Potem simplektična matrika*

$$M = \begin{bmatrix} A + SA^{-T}Q & -SA^{-T} \\ -A^{-T}Q & A^{-T} \end{bmatrix}$$

nima nobene lastne vrednoti na enotski krožnici.

Izrek 7.18 *Naj bo par (A, B) stabilizabilen, par (A, Q) pa zaznaven, Q je simetrično pozitivno semidefiniten, R pa simetrično pozitivno definitna. Potem ima DARE*

$$A^T PA - P + Q - A^T PB(R + B^T PB)^{-1} B^T PA = 0$$

enolično simetrično pozitivno semidefinitno stabilizirajočo rešitev P, ki je podana z $P = P_2 P_1^{-1}$, kjer stolpci matrike $\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$ razpenjajo n-razsežen invariantni podprostor pridružene simplektične matrike

$$M = \begin{bmatrix} A + SA^{-T}Q & -SA^{-T} \\ -A^{-T}Q & A^{-T} \end{bmatrix},$$

ki pripada lastnim vrednostim z absolutnimi vrednostmi pod 1.

Poglavje 8

Numerično reševanje Riccatijeve enačbe

8.1 Uvod

Obravnavali bomo numerične metode za zvezno algebraično Riccatijevo enačbo (CARE)

$$XA + A^T X + Q - XSX = 0,$$

kjer je $S = BR^{-1}B^T$.

Prav tako se bomo ukvarjali z diskretno algebraično Riccatijevo enačbo (DARE)

$$Q - X + A^T X(I + SX)^{-1}A = 0.$$

Riccatijeva enačba je nelinearna in ima lahko mnogo rešitev.

Če Riccatijeva enačba izvira iz LQR problema, potem je Q simetrična pozitivno semidefinitna, R pa simetrična pozitivno definitna. Če je par (A, B) stabilizabilen, par (A, Q) pa zaznaven, potem obstaja enolična simetrična pozitivno semidefinitna rešitev X za CARE in zaprtozančna matrika $A - BR^{-1}B^T X$ je stabilna.

8.2 Občutljivost Riccatijeve enačbe

Iščemo stabilizirajočo s.p.sp. rešitev CARE

$$XA + A^T X + Q - XSX = 0.$$

Zanima nas, kako občutljiv je problem oziroma kako močno lahko motnje matrik A, Q in S vplivajo na spremembo X. Pri reševanju z obratno stabilnim algoritmom lahko pričakujemo, da za izračunani $\widetilde{X}$ velja

$$(A + \Delta A)^T \widetilde{X} + \widetilde{X}(A + \Delta A) + Q + \Delta Q - \widetilde{X}(S + \Delta S)\widetilde{X} = 0, \tag{8.1}$$

kjer je $\|\Delta A\| \leq \epsilon\|A\|$, $\|\Delta Q\| \leq \epsilon\|Q\|$ in $\|\Delta S\| \leq \epsilon\|S\|$.

Razlika $\|\widetilde{X} - X\|/\|X\|$ je lahko zelo velika, če je sistem občutljiv.

Naj bo $\widetilde{X} = X + \Delta X$. Če zanemarimo vse kvadratne Δ člene v (8.1), dobimo

$$(A - SX)^T \Delta X + \Delta X (A - SX) + \Delta A^T X + X \Delta A + \Delta Q - X \Delta S X = 0. \tag{8.2}$$

To je enačba Ljapunova za ΔX. Matrika $A - SX = A - BR^{-1}B^T X$ je stabilna, zato je enačba rešljiva. Če za enačbo Ljapunova uporabimo ocene iz poglavja 4, dobimo:

$$\frac{\|\Delta X\|_F}{\|X\|_F} \leq \frac{2\|\Delta A\|_F \ + \frac{\|\Delta Q\|_F}{\|X\|_F} + \|\Delta S\| \|X\|_F}{\mathrm{sep}\Big((A - SX)^T, -(A - SX)\Big)},$$

pri čemer je $\mathrm{sep}((A-SX)^T, -(A-SX))$ minimalna singularna vrednost matrike $I \otimes (A-SX)^T + (A - SX)^T \otimes I$. Če velja $\|\Delta A\| \leq \epsilon \|A\|$, $\|\Delta Q\| \leq \epsilon \|Q\|$ in $\|\Delta S\| \leq \epsilon \|S\|$, potem dobimo

$$\frac{\|\Delta X\|_F}{\|X\|_F} \leq \epsilon \frac{2\|A\|_F \ + \frac{\|Q\|_F}{\|X\|_F} + \|S\| \|X\|_F}{\mathrm{sep}\Big((A - SX)^T, -(A - SX)\Big)}.$$

Iz ocene sledi, da je *pogojenostno število CARE* enako

$$c_R = \frac{2\|A\|_F \ + \frac{\|Q\|_F}{\|X\|_F} + \|S\| \|X\|_F}{\mathrm{sep}\Big((A - SX)^T, -(A - SX)\Big)}.$$

Riccatijeva enačba je občutljiva, kadar imamo matrike velikih norm (velja za vsako izmed matrik A, Q, X, S) in pa kadar je $\mathrm{sep}((A - SX)^T, -(A - SX))$ blizu 0.

Podobno lahko za diskretno Riccatijevo enačbo $Q - X + A^T X (I + SX)^{-1} A = 0$ izpeljemo oceno

$$\frac{\|\Delta X\|_F}{\|X\|_F} \leq \frac{2\|A\|_F \|\Delta A\|_F \ + \frac{\|\Delta Q\|_F}{\|X\|_F} + \|A\|_F^2 \|\Delta S\|_F \|X\|_F}{\mathrm{sep}_d(A_c^T, A_c)},$$

kjer je

$$A_c = (I + SX)^{-1} A = A - B(R + B^T X B)^{-1} B^T X A$$

in

$$\mathrm{sep}_d(A_c^T, A_c) = \min_{X \neq 0} \frac{\|A_C^T X A_C - X\|_F}{\|X\|_F}.$$

Pri predpostavkah $\|\Delta A\| \leq \epsilon \|A\|$, $\|\Delta Q\| \leq \epsilon \|Q\|$ in $\|\Delta S\| \leq \epsilon \|S\|$ je

$$c_D = \frac{2\|A\|_F^2 + \frac{\|Q\|_F}{\|X\|_F} + \|A\|_F^2 \|S\|_F \|X\|_F}{\mathrm{sep}_d(A_c^T, A_c)},$$

pogojenostno število DARE in velja $\|\Delta X\|_F / \|X\|_F \leq c_D \epsilon$.

8.3 Newtonova metoda

Kot prvo bomo za reševanje CARE

$$XA + A^T X + Q - XSX = 0$$

obravnavali Newtonovo metodo. Naj bo X_0 približek za rešitev CARE. Če označimo $X = X_0 + \Delta X_0$ in to vstavimo v CARE, dobimo

$$(A - SX_0)^T X + X(A - SX_0) + Q + X_0 S X_0 - \Delta X_0 S \Delta X_0 = 0.$$

Ob predpostavki, da imamo dober začetni približek, zanemarimo člen $\Delta X_0 S \Delta X_0$. Tako dobimo enačbo Ljapunova za naslednji približek X_1.

$$(A - SX_0)^T X_1 + X_1(A - SX_0) = -Q - X_0 S X_0.$$

Newtonova metoda za CARE

izberi X_0, da je $A - SX_0$ stabilna matrika
za $k = 0, 1, \ldots$
    reši $(A - SX_k)^T X_{k+1} + X_{k+1}(A - SX_k) = -Q - X_k S X_k$.

Alternativno lahko Newtonovo metodo izpeljemo tako, da definiramo funkcijo ostanka

$$R(X) = XA + A^T X + Q - XSX.$$

Fréchetov odvod $R(X)$ je

$$R'_X(Z) = (A - SX)^T Z + Z(A - SX),$$

od tod pa sledi en korak Newtonove metode za enačbo $R(X) = 0$:

$$R'_{X_k}(\Delta_k) = -R(X_k), \qquad X_{k+1} = X_k + \Delta_k.$$

Izboljšana različica Newtonove metode je

izberi $X_0 = X_0^T$, da je $A - SX_0$ stabilna matrika
za $k = 0, 1, \ldots$
    $A_k = A - SX_k$
    $R(X_k) = X_k A + A^T X_k + Q - X_k S X_k$
    reši $A_k^T \Delta_k + \Delta_k A_k + R(X_k) = 0$
    $X_{k+1} = X_k + \Delta_k$

Ker računamo s popravki, je metoda natančnejša.

Izrek 8.1 *Naj bo X_0 stabilizirajoči približek, X pa enolična s.p.sd. rešitev za CARE. Za matrike A_k in X_k, $k = 0, 1, \ldots$, ki jih dobimo pri Newtonovi metodi, velja*

a) vse matrike A_k so stabilne,

b) $X \le \cdots \le X_{k+1} \le X_k \le \cdots \le X_1$,

c) zaporedje matrik $X_0, X_1, \ldots$ konvergira proti enolični s.p.sd. rešitvi CARE X,

d) obstaja taka konstanta $c > 0$, da je $\|X_{k+1} - X_k\| \le c\|X_k - X\|^2$ za $k \ge 1$, konvergenca je torej kvadratična.

Ustavitveni kriterij:

- končamo, ko je $\|X_{k+1} - X_k\|_F \le \|X_k\|_F \, \epsilon$ za dovolj majhen ϵ,
- končamo, ko prekoračimo maksimalno število korakov,
- če imamo na voljo oceno za pogojenostno število CARE, lahko nehamo takrat, ko je $\|X_{k+1} - X_k\|_F \le \|X_k\|_F \, \kappa_{\mathrm{CARE}} \, u$, kjer je u osnovna zaokrožitvena napaka.

Metoda je numerično stabilna in v primeru konvergence vrne zelo natančen rezultat.

Za konvergenco je na začetku potrebno imeti potrebno imeti dober začetni približek X_0, drugače lahko metoda sploh ne konvergira.

Izrek 8.2 *Naj bo $R = I$. Naj bo X_0 tak začetni približek, da je $A - BB^T X_0$ stabilna in*

$$\|X - X_0\| \le 1/(3\|B\|^2 \|\Omega^{-1}\|),$$

kjer je

$$\Omega(Z) = (A - BB^T X)^T Z + Z(A - BB^T X).$$

Potem velja

$$\|X_1 - X_0\| \le \|X - X_0\|,$$

enakost pa je dosežena natanko pri $X_0 = X$.

Ponavadi Newtonovo metodo uporabljamo za iterativno izboljšanje rezultata, ki smo ga dobili s kakšno drugo metodo.

Z relaksacijo Newtonove metode lahko pospešimo konvergenco. Namesto, da v vsaakem koraku vzamemo $X_{k+1} = X_k + \Delta_k$, vzamemo $X_{k+1} = X_k + \lambda_k \Delta_k$, kjer je parameter $\lambda_k \in [0, 2]$ izbran tako, da minimizira $\|R(X_k + \lambda_k \Delta_k)\|_F$. Velja

$$\begin{aligned} \|R(X_k + \lambda_k \Delta_k)\|_F^2 &= \mathrm{sled}(R(X_k + \lambda_k \Delta_k)^2) \\ &= a_k^2 (1 - \lambda_k)^2 - 2 b_k (1 - \lambda_k) \lambda_k^2 + c_k \lambda_k^4, \end{aligned}$$

kjer je $a_k = \mathrm{sled}(R(X_k)^2)$, $b_k = \mathrm{sled}(R(X_k) V_k)$, $c_k = \mathrm{sled}(V_k^2)$ in $V_k = \Delta_k S \Delta_k$.

Relaksirana različica Newtonove metode je

izberi $X_0 = X_0^T$, da je $A - SX_0$ stabilna matrika
za $k = 0, 1, \ldots$
$A_k = A - SX_k$
$R(X_k) = X_k A + A^T X_k + Q - X_k S X_k$
reši $A_k^T \Delta_k + \Delta_k A_k + R(X_k) = 0$

določi $\lambda_k \in [0,2]$, ki minimizira $\|R(X_k + \lambda_k \Delta_k)\|_F$
$X_{k+1} = X_k + \lambda_k \Delta_k$

Podobno lahko izpeljemo tudi Newtonovo metodo za DARE.

Sedaj definiramo

$$R_D(X) = A^T X A - X + Q + A^T X B (R + B^T X B)^{-1} B^T X A.$$

Newtonova metoda za DARE:

izberi $X_0 = X_0^T$, da je $A - B(R + B^T X_0 B)^{-1} B^T X_0 A$ stabilna matrika
za $k = 0, 1, \ldots$
$K_k = (R + B^T X_k B)^{-1} B^T X_k A$
$A_k = A - B K_k$
$R_D(X_k) = A^T X_k A - X_k + Q + A^T X_k B (R + B^T X_k B)^{-1} B^T X_k A$
reši $A_k^T \Delta_k A_k - \Delta_k + R_D(X_k) = 0$
$X_{k+1} = X_k + \Delta_k$

V vsakem koraku metode moramo rešiti diskretno enačbo Ljapunova.

8.4 Uporaba matričnega predznaka

Naj ima matrika A Jordanovo formo $X^{-1} A X = D + N$, kjer je X nesingularna matrika, $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ je diagonalna matrika lastnih vrednosti, N pa je nilpotentna matrika.

Matrični predznak matrike A je matrika

$$\mathrm{sign}(A) = X \mathrm{diag}(\mathrm{sign}(\lambda_1), \ldots, \mathrm{sign}(\lambda_n)) X^{-1},$$

kjer je

$$\mathrm{sign}(\lambda_k) = \begin{cases} 1, & \mathrm{Re}(\lambda_k) > 0, \\ -1, & \mathrm{Re}(\lambda_k) < 0. \end{cases}$$

Matrični predznak ni definiran, če ima matrika lastno vrednost 0 ali kakšno strogo imaginarno lastno vrednost.

Matrični predznak ima naslednje lastnosti:

1. $\mathrm{sign}(A)^2 = I$,
2. $\mathrm{sign}(\alpha A) = \mathrm{sign}(\alpha)\,\mathrm{sign}(A)$ za $\alpha \in \mathbb{C}$,
3. $\mathrm{sign}(T A T^{-1}) = T\,\mathrm{sign}(A) T^{-1}$,
4. $A\,\mathrm{sign}(A) = \mathrm{sign}(A) A$,
5. lastne vrednosti $\mathrm{sign}(A)$ so ± 1,
6. $\mathrm{sign}(A)$ ima isti stabilni invariantni podprostor kot A,
7. $\mathrm{Lin}(\mathrm{sign}(A) - I)$ je stabilni invariantni podprostor A,

8. lastni vektorji $\operatorname{sign}(A)$ so lastni vektorji in korenski vektorji A.

Naj bo CARE $XA+A^TX+Q-XSX=0$ pridružena Hamiltonska matrika $H=\begin{bmatrix} A & -S \\ -Q & -A^T \end{bmatrix}$. Vemo

$$H=\begin{bmatrix} I & 0 \\ X & I \end{bmatrix}\begin{bmatrix} A-SX & -S \\ 0 & -(A-SX)^T \end{bmatrix}\begin{bmatrix} I & 0 \\ -X & I \end{bmatrix}.$$

Ker je matrika $A-SX$ stabilna, je matrični predznak matrike $\operatorname{sign}(H)$ dobro definiran in velja

$$\operatorname{sign}(H)=\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}=\begin{bmatrix} I & 0 \\ X & I \end{bmatrix}\begin{bmatrix} -I & Z \\ 0 & I \end{bmatrix}\begin{bmatrix} I & 0 \\ -X & I \end{bmatrix}. \tag{8.3}$$

Iz $H\operatorname{sign}(H)=\operatorname{sign}(H)H$ dobimo za Z enačbo Ljapunova $(A-SX)Z+Z(A-SX)^T=2S$. Dobimo tudi

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}\begin{bmatrix} I \\ X \end{bmatrix}=-\begin{bmatrix} I \\ X \end{bmatrix} \tag{8.4}$$

oziroma

$$MX=-N,$$

kjer sta $M=\begin{bmatrix} W_{12} \\ W_{22}+I \end{bmatrix}$ in $N=\begin{bmatrix} W_{11}+I \\ W_{21} \end{bmatrix}$.

Dobili smo konsistentni sistem $2n^2$ enačb za n^2 elementov matrike X. Iz (8.3) sledi

$$\begin{bmatrix} W_{12} \\ W_{22} \end{bmatrix}=\begin{bmatrix} Z \\ XZ+I \end{bmatrix} \quad\Longrightarrow\quad M=\begin{bmatrix} Z \\ XZ+2I \end{bmatrix} \quad\Longrightarrow\quad [\,-X \quad I\,]\,M=2I.$$

M je torej polnega ranga in X lahko lahko stabilno izračunamo iz $MX=-N$, npr. s QR razcepom.

Dokazali smo naslednji izrek.

Izrek 8.3 *Naj bo par (A,B) stabilizabilen, (A,Q) pa zaznaven. Naj bo*

$$H=\begin{bmatrix} A & -S \\ -Q & -A^T \end{bmatrix}$$

Hamiltonska matrika, pridružena CARE

$$XA+A^TX+Q-XSX=0.$$

Če je

$$\operatorname{sign}(H)=\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix},$$

potem je stabilizirajoča rešitev CARE rešitev konsistentnega predoločenega sistema

$$\begin{bmatrix} W_{12} \\ W_{22}+I \end{bmatrix}X=-\begin{bmatrix} W_{11}+I \\ W_{21} \end{bmatrix}.$$

Sedaj moramo samo še pogledati, kako lahko ekonomično izračunamo matrični predznak matrike H. Za računanje matričnega predznaka obstaja več numeričnih metod.

Osnovni algoritem je iteracija

$$W_{k+1} = \frac{1}{2}(W_k + W_k^{-1})$$

za $k = 0, 1, \ldots$, ki se začne z $W_0 = H$. Ta iteracija temelji na dejstvu, da je $\operatorname{sign}(H)$ kvadratni koren identitete.

Če je H Hamiltonska matrika, sta to tudi H^{-1} in $\frac{1}{2}(H + H^{-1})$, kar pomeni, da se v zgornjem algoritmu ta lastnost matrike ohranja.

Izkaže se, da uporaba matričnega predznaka ni numerično stabilna metoda za reševanje Riccatijeve enačbe. Če pa metodo kombiniramo z iterativnimi izboljšavami po Newtonovi metodi, potem v praksi dobimo učinkovito in stabilno metodo.

Osnovni algoritem za računanje $\operatorname{sign}(H)$ je

$W_0 = H$
$k = 0, 1, \ldots$
  $W_{k+1} = \frac{1}{2}(W_k + W_k^{-1})$

Konvergenca osnovnega algoritma je v bližini rešitve kvadratična, na začetku pa je lahko dokaj počasna. Pospešimo jo lahko tako, da v vsakem koraku pomnožimo W_k s skalarjem, ki pošlje lastne vrednosti W_k čim bližje ± 1. Idealni koeficient je enak obratni vrednosti geometrijskega povprečja lastnih vrednosti W_k, ki je $|\det(W_k)|^{1/(2n)}$.

Novi algoritem je

$W_0 = H$
$k = 0, 1, \ldots$
  $c_k = |\det(W_k)|^{1/(2n)}$
  $W_{k+1} = \frac{1}{2c_k}(W_k + c_k^2 W_k^{-1})$

V algoritmu moramo izračunati inverz matrike W_k. Ker je JW_k simetrična matrika, to izračunamo kot $W_k^{-1} = (JW_k)^{-1}J$, da lahko izkoristimo simetrijo in je algoritem še učinkovitejši.

Če na začetku vzamemo $W_0 = JH$ in računamo $W_{k+1} = \frac{1}{2c_k}(W_k + c_k^2 JW_k^{-1}J)$ za $k = 0, 1, \ldots$ potem se simetrija ohranja, matrike W_k pa skonvergirajo proti $J\operatorname{sign}(H)$.

Novi algoritem, ki ohranja simetrijo, je

$Y_0 = JH$
$k = 0, 1, \ldots$
  $c_k = |\det(Y_k)|^{1/(2n)}$
  $Y_{k+1} = \frac{1}{2c_k}(Y_k + c_k^2 JY_k^{-1}J)$

Če je $Y = \lim_{k\to\infty} Y_k$, potem na koncu velja

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} = \operatorname{sign}(H) = J^T Y = \begin{bmatrix} -Y_{21} & -Y_{22} \\ Y_{11} & Y_{12} \end{bmatrix}$$

in za X rešimo predoločeni sistem

$$\begin{bmatrix} Y_{22} \\ Y_{12}+I \end{bmatrix} X = \begin{bmatrix} I-Y_{21} \\ -Y_{11} \end{bmatrix}.$$

DARE

$$A^TXA - X + Q - A^TXB(R+B^TXB)^{-1}B^TXA = 0$$

je povezana s simplektično matriko

$$M = \begin{bmatrix} A+SA^{-T}Q & -SA^{-T} \\ -A^{-T}Q & A^{-T} \end{bmatrix},$$

ki ima lastne vrednosti $\lambda_1, \ldots, \lambda_n$ in $\lambda_1^{-1}, \ldots, \lambda_n^{-1}$.

Če uporabimo transformacijo $H = (M+I)^{-1}(M-I)$, dobimo Hamiltonsko matriko, ki ima lastne vrednosti

$$\pm\frac{\lambda_i - 1}{\lambda_i + 1}$$

za $i = 1, \ldots, n$.

H je povezana z zvezno Riccatijevo enačbo, ki ima isto rešitev kot DARE. Če uporabimo zvezne algoritme na H, dobimo potem tudi rešitev za DARE. Težava pri DARE je, da imamo v M inverz matrike A, torej težave, če je A singularna ali blizu singularni matriki.

Matriko

$$M = \begin{bmatrix} A+SA^{-T}Q & -SA^{-T} \\ -A^{-T}Q & A^{-T} \end{bmatrix}$$

lahko zapišemo kot

$$M = N^{-1}P,$$

kjer sta $N = \begin{bmatrix} I & S \\ 0 & A^T \end{bmatrix}$ in $P = \begin{bmatrix} A & 0 \\ -Q & I \end{bmatrix}$.

Matrika $N+P$ je obrnljiva in H lahko potem zapišemo kot

$$H = (P+N)^{-1}(P-N).$$

Na ta način se izognemo računanu A^{-1}.

8.5 Metoda lastnih vektorjev

Pri metodi lastnih vektorjev upoštevamo dejstvo, da je stabilizirajoča rešitev CARE povezana s stabilnim podprostorom pripadajoče Hamiltonske matrike H.

Naj se da H diagonalizirati in naj bo

$$H = V \begin{bmatrix} -\overline{\Lambda} & 0 \\ 0 & \Lambda \end{bmatrix} V^{-1},$$

kje je $V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$ matrika lastnih vektorjev in $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, kjer je $\mathrm{Re}(\lambda_k) > 0$ za $k = 1, \ldots, n$.

Potem je $X = V_{21} V_{11}^{-1}$ stabilizirajoča rešitev CARE.

Ta metoda ni numerično stabilna, saj imamo lahko težave s slabo pogojeno matriko V_{11}, prav tako ni nujno, da se da matriko H diagonalizirati. Sicer bi teoretično lahko uporabili tudi korenske vektorje, toda numerično to ni najbolj stabilno.

Podobno metodo bi lahko razvili za DARE, tu pa imamo težave že pri formiranju simplektične matrike M, če je A singularna ali slabo pogojena.

8.6 Uporaba Schurove forme

Če namesto lastnega razcepa uporabimo Schurovo formo, se znebimo težav, ki se pojavijo zaradi zelo občutljive matrike lastnih vektorjev oziroma dejstva, da se matrike ne da diagonalizirati.

Schurova forma mora biti urejena tako, da imamo v prvem bloku vse stabilne lastne vrednosti, torej

$$T = U^T H U = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix},$$

kjer so vse lastne vrednosti z negativnim realnim delom v T_{11}, tiste s pozitivnim realnim delom pa v T_{22}.

Če je $U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}$, potem stolpci $\begin{bmatrix} U_{11} \\ U_{21} \end{bmatrix}$ razpenjajo stabilni invariantni podprostor H in $X = U_{21} U_{11}^{-1}$ je stabilizirajoča s.p.sp. rešitev CARE.

Lastne vrednosti v Schurovi formo preuredimo tako, kot smo to že omenili pri Schurovi metodi za razporejanje polov, torej z zamenjavami sosednjih 1×1 ali 2×2 diagonalnih blokov.

Schurov algoritem je sestavljen iz naslednjih korakov:

1. zgeneriraj Hamiltonsko matriko $H = \begin{bmatrix} A & -S \\ -Q & -A^T \end{bmatrix}$,
2. H spremeni v urejeno Schurovo formo $U^T H U = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}$, kjer so lastne vrednosti z negativnim realnim delom vse v T_{11},
3. bločno zapiši $U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}$,
4. rešitev je $X = U_{21} U_{11}^{-1}$.

Poleg tega, da je Schurov algoritem stabilnejši od uporabe lastnih vektorjev, je tudi cenejši, saj je Schurova forma vmesni korak pri izračunu lastnih vektorjev.

Časovna zahtevnost Schurovega algoritem je primerljiva s kombinacijo matričnega predznaka in iterativnega izboljšanja z Newtonovo metodo.

8.6.1 Analiza zaokrožitvenih napak

Za izračunano $\widetilde{T}$ in ortogonalno matriko $\widetilde{U}$ velja, da je to točna Schurova forma za $H+E$, kjer je $\|E\| \le c_1 u\|H\|$ in je c_1 je reda n^k za majhen k.

Za $\widetilde{U}_{11}$ velja $\widetilde{U}_{11} = (U_{11}+U_{12}Z)(I+Z^TZ)^{-1/2}$, kjer je $\|Z\| \le \frac{2c_1u}{\delta}\|H\|$ in

$$\delta = \mathrm{sep}(T_{11}, T_{22}) - c_1\|H\|u.$$

$\widetilde{X}$ dobimo iz $X = U_{21}U_{11}^{-1}$ in lahko ocenimo

$$\|\widetilde{X} - X\| \le c_2\kappa(\widetilde{U}_{11})\|\widetilde{X}\|u + (1+\|X\|)\|Z\|\|\widetilde{U}_{11}^{-1}\|,$$

kjer je c_2 konstanta povezana s pivotno rastjo pri računanju $\widetilde{X}$.

Če velja $\delta > 0$ in

$$c_1u\|H\|^2(1+c_1u) \le \frac{1}{4}\delta^2,$$

potem za izračunano rešitev $\widetilde{X}$ po Schurovi metodi velja

$$\frac{\|\widetilde{X}-X\|}{\|X\|} \le 2\frac{c_1u}{\delta}\left(1+\frac{1}{\|X\|}\right)(2\|A\|+\|Q\|+\|S\|)\|U_{11}^{-1}\| + c_2\kappa_2(U_{11})u\frac{\|\widetilde{X}\|}{\|X\|}.$$

Analiza pravi, da je napaka odvisna od $\mathrm{sep}(T_{11}, T_{22})$. Za primerjavo, pogojenostno število CARE je

$$c_R = \frac{2\|A\|_F \ + \frac{\|Q\|_F}{\|X\|_F} + \|S\|\|X\|_F}{\mathrm{sep}((A-SX)^T, -(A-SX))}.$$

Kljub temu, da imata matriki T_{11} in $(A-SX)$ oziroma T_{22} in $-(A-SX)$ enake lastne vrednosti, je lahko $\mathrm{sep}(T_{11}, T_{22})$ dosti manjša kot $\mathrm{sep}((A-SX)^T, -(A-SX))$ in v teh primerih je Schurova metoda nestabilna.

To se lahko zgodi v primeru, ko v Riccatijevi enačbi nastopajo matrike z normami zelo različnih velikosti. Izračunana rešitev $\widetilde{X}$ je točna rešitev za malo zmoteno Hamiltonovo matriko $H+E$. Sicer vemo, da je razmerje $\|E\|/\|H\|$ majhno, a v primeru, ko so norme matrik A, Q in S zelo različne, je katera izmed relativnih motenj $\|\Delta A\|/\|A\|$, $\|\Delta Q\|/\|Q\|$ ali $\|\Delta S\|/\|S\|$ lahko velika.

Tu je problem tudi ta, da uporabljamo algoritme, ki ne upoštevajo strukture, zato ne moremo pri obratni stabilnosti opazovati motenj posameznih blokov matrike E, temveč le matriko v celoti.

Videli smo, da imamo lahko težave, če so norme matrik A, Q in S preveč neizenačene.

S *skaliranjem* lahko poskrbimo, da bodo norme bolj izenačene in napake manjše. Namesto CARE $XA + A^TX + Q - XSX = 0$ pišemo

$$X_\rho A_\rho + A_\rho^T X_\rho + Q - X_\rho S_\rho X_\rho = 0,$$

kjer je ρ pozitivna konstanta in

$$A_\rho = \rho A, \quad A_\rho^T = (\rho A)^T, \quad X_\rho = X_\rho/\rho, \quad S_\rho = \rho^2 S,$$

kjer ρ izberemo kot

$$\rho = \begin{cases} \frac{\|S\|}{\|Q\|}, & \|Q\| > \|S\|, \\ \frac{\|A\|}{\|S\|}, & \|Q\| \le \|S\| \text{ in } \|Q\|\|S\| \le \|A\|^2, \\ 1 & \text{sicer.} \end{cases}$$

Pogojenost CARE se s skaliranjem ne spremeni.

8.6.2 Schurova metoda za DARE

Na podoben način lahko pri DARE

$$A^T X A - X + Q - A^T X B (R + B^T X B)^{-1} B^T X A = 0$$

uporabimo Schurovo metodo na simplektični matriki

$$M = \begin{bmatrix} A + SA^{-T}Q & -SA^{-T} \\ -A^{-T}Q & A^{-T} \end{bmatrix}.$$

Sedaj jo moramo preurediti tako, da bodo v prvem bloku vse lastne vrednosti, ki so po absolutni vrednosti pod 1, ostale pa v drugem bloku.

Če je $U^T M U = \begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{22} \end{bmatrix}$ in $U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}$, kjer so v bloku S_{11} vse lastne vrednosti z absolutnimi vrednostmi pod 1, je rešitev

$$X = U_{21} U_{11}^{-1}.$$

Tudi tu lahko pričakujemo velike napake, če je A singularna oz. slabo pogojena. Temu se lahko izognemo, če problem prevedemo na posplošeni problem lastnih vrednosti.

8.7 Posplošeni problem lastnih vrednosti in DARE

DARE $A^T X A - X + Q - A^T X B (R + B^T X B)^{-1} B^T X A = 0$ lahko prevedemo na posplošeni problem lastnih vrednosti

$$Px = \lambda N x,$$

kjer sta $N = \begin{bmatrix} I & S \\ 0 & A^T \end{bmatrix}$ in $P = \begin{bmatrix} A & 0 \\ -Q & I \end{bmatrix}$.

Posplošeni problem lastnih vrednosti ima tudi simplektične lastnosti v smislu, da je $\lambda \neq 0$ lastna vrednost natanko tedaj, ko je lastna vrednost tudi λ^{-1}. Ko je A singularna, je vsaj ena lastna vrednost enaka 0. Če je večkratnost 0 kot lastne vrednosti r, potem imamo tudi lastno vrednost ∞ z večkratnostjo r, torej je $2n - r$ lastnih vrednosti končnih.

Lastne vrednosti so

$$0, \ldots, 0, \lambda_{r+1}, \ldots, \lambda_n, \lambda_n^{-1}, \ldots, \lambda_1^{-1}, \infty, \ldots, \infty,$$

kjer je $0 < |\lambda_i| < 1$ za $i = r+1, \ldots, n$. Za rešitev DARE potrebujemo bazo za stabilni podprostor.

Posplošeni problem lastnih vrednosti $Px = \lambda Nx$ rešimo s pomočjo QZ algoritma. Poiščemo ortogonalni matriki U in V, da je

$$\widetilde{N} = UNV = \begin{bmatrix} N_{11} & N_{12} \\ 0 & N_{22} \end{bmatrix}$$

in

$$\widetilde{P} = UPV = \begin{bmatrix} P_{11} & P_{12} \\ 0 & P_{22} \end{bmatrix}.$$

Pri tem mora biti Schurova forma preurejena tako, da so lastne vrednosti para (P_{11}, N_{11}) znotraj enotskega kroga.

Če je $V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$, potem je $X = V_{21}V_{11}^{-1}$.

Literatura

[1] S. Barnett, R. G. Cameron: *Introduction to Mathematical Control Theory*, Second Edition, Clarendon Press, Oxford, 1985.

[2] B. N. Datta: *Numerical Methods for Linear Control Systems*, Elsevier Academic Press, San Diego, 2004.

[3] P. M. Van Dooren: *Graduate Course on Numerical Linear Algebra for Signals Systems and Control*, Draft notes prepared for the Graduate School in Systems and Control, University of Louvain, 2003.
dosegljivo na: `http://www.inma.ucl.ac.be/~vdooren/PVDnotes.pdf`

[4] F. W. Fairman: *Linear Control Theory: The State Space Approach*, John Wiley & Sons, New York, 1998.

[5] G. H. Golub, C. F. Van Loan: *Matrix Computations*, Third editiion, John Hopkins, Baltimore, 1996.

[6] N. J. Higham: *The Scaling and Squaring Method for the Matrix Exponential Revisited*, SIAM J. Matrix Anal. Appl. **26**, 2005, str. 1179–1193.

[7] B. Kisačanin, G. C. Agarwal: *Linear Control Systems: with solved problems and MATLAB examples*, Kluwer Academic, New York, 2001.

[8] C. Moler, C. Van Loan: *Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later*, SIAM Review **45**, 2003, str. 3–49.

[9] P. Hr. Petkov, N. D. Christov, M. M. Konstantinov: *Computational Methods for Linear Control Systems*, Prentice Hall, 1991.

[10] C. F. Van Loan: *The sensitivity if the matrix exponential*, SIAM J. Numer. Anal. **14**, 1971, str. 971–981.

[11] S. H. Żak: *Systems and Control*, Oxford University Press, Oxford, 2003.