Numerische Methoden in der Steuerungstheorie

Vorlesung TU Chemnitz–Zwickau
SS 1995, Volker Mehrmann

Literatur:

(a) H. W. Knobloch, H. Kwakernaak
Lineare Kontrolltheorie, Springer 1980

(b) A. Linnemann
Numerische Methoden für lineare Regelungssysteme,
Bi–Wissenschaftsverlag 1993

(c) V. Mehrmann
The Autonomous Linear Quadratic Control Problem,
Springer Verlag 1991

(d) Golub/Van Loan
Matrix Computation,
Cambridge University Press 1989

Inhaltsverzeichnis

Kapitel 0

Einleitung

Wir wollen uns in dieser Veranstaltung mit Steuerungstheorie, oft auch Kontrolltheorie oder Regelungstheorie genannt, beschäftigen und zwar mit der Mathematik, die dazu gehört. Dies ist ein Gebiet auf der Grenze zwischen Mathematik/Ingenieurwissenschaft/Wirtschaftswissenschaft/ Informatik.

Unser Hauptaugenmerk liegt dabei auf der mathematischen Seite, ich werde aber versuchen, immer wieder Bezüge zur Praxis herzustellen, sowohl zu Anwendungsproblemen, als auch zur realen Implementierung der notwendigen Methoden.

Tabelle 0.1 *Einordnung des Themas*

<table>
<tr><th colspan="3">Allgemein:</th><th colspan="3">Beispiel:</th></tr>
<tr><td colspan="3">Praktisches Problem</td><td colspan="3">Steuerung einer Parabolantenne</td></tr>
<tr><td colspan="3">Entwicklung eines math. Modells</td><td colspan="3">lineares System $\dot{x} = Ax + Bu$
u Steuerungsfunktion</td></tr>
<tr><td>Analyse der math. Eigenschaften Lösbarkeit, Eindeutigkeit, Analyt. Methoden</td><td colspan="2">Entwicklung von numerischen Methoden</td><td>Existenz– und Eindeutigkeit, mögl. Steuer–funktionen, Transferfunktion,</td><td colspan="2">Num. Methoden für gew. Dgl., opt. Steuerungspr.</td></tr>
<tr><td></td><td>Analyse der Verfahren</td><td>Implement. auf Rechner</td><td></td><td>Fehler–analyse</td><td>Implem.</td></tr>
<tr><td>analyt. Lsg.</td><td colspan="2">num. Lsg.</td><td>analyt. Lsg.</td><td colspan="2">num. Lsg.</td></tr>
<tr><td colspan="3">Umsetzen der Lösung</td><td colspan="3">Umsetzen der Lösung</td></tr>
</table>

Beispiel 0.2 *Steuerung einer Parabolantenne, die immer auf einen Satelliten oder ein Raumfahrzeug gerichtet ist.*

Wir brauchen ein einfaches Modell für den Motor, siehe z.B.Kn./Kw. Dort wird ein Gleichstrommotor beschrieben, der über eine Eingangsspannung gesteuert wird

$$\text{Spannung } u \longrightarrow \boxed{\text{Verstärker}} \longrightarrow \boxed{\text{Gls.–Motor}} \longrightarrow \varphi \text{ Drehwinkel}$$

vereinfachte Bewegungsgleichungen: t Zeit

$$\left\{ \begin{array}{rcl} \frac{d\varphi}{dt} & = & \dot{\varphi}(t) = w(t) \\ j\dot{w}(t) & = & -rw(t) + ku(t) \end{array} \right. \quad \begin{array}{ll} w & \text{Winkelgeschwindigkeit} \\ j & \text{Trägheitsmoment des Motors} \\ r & \text{Reibungskoeffizient} \\ k & \text{Verstärkungsfaktor .} \end{array} \tag{0.3}$$

Wir wollen durch Anlegen der Eingangsspannung (Steuerungsfunktion $u(t)$) den Drehwinkel φ_1 zum Zeitpunkt t_1 erreichen, ausgehend davon, daß zum Zeitpunkt t_0 der Drehwinkel φ_0 ist.

(0.4)

Wir müssen also eine Funktion $u(t)$ finden, die man auch noch realisieren kann, so daß für die Lösung der Differentialgleichung (0.3) mit Anfangsbedingung

$$\varphi(t_0) = \varphi_0 \tag{0.5}$$

und diesem $u(t)$ gilt

$$\varphi(t_1) = \varphi_1. \tag{0.6}$$

Es wird im allgemeinen viele Lösungen geben, von denen wir eine aussuchen müssen, dazu brauchen wir Kriterien, typischerweise Minimalitätsbedingungen, z. B. Minimierung der Energie oder Minimierung von t_1, oder z. B.

$$f(\varphi, u) = Min! \tag{0.7}$$

Falls wir das Optimalitätskriterium angegeben haben, müssen wir dann das folgende Problem lösen, d.h.

(a) zeigen, daß es u gibt, so daß (0.5), (0.6), (0.7) gelten,

(b) zeigen, daß dies eindeutig ist,

(c) wenn es nicht eindeutig ist, (0.7) modifizieren,

(d) analytische Lösung/numerische Lösung, d.h. Bestimmung von $u(t)$, oft in **real time**,

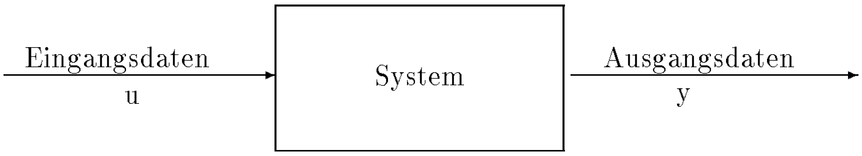
(e) Steuerung der Antenne.

Als Mathematiker machen wir das ganze natürlich so, daß es allgemein genug ist, um viele unterschiedliche Probleme gleichzeitig zu behandeln.
Es gibt noch viele andere Aufgaben in der Steuerungstheorie, aber dazu später.

Kapitel 1

Theoretische Grundlagen der Steuerungstheorie

Allgemein haben wir die folgende Beschreibung eines dynamischen Systems



Wir wissen typischerweise was wir in das System hereingeben und wir können typischerweise auch „beobachten", d.h. messen was herauskommt, aber in den meisten Fällen wissen wir nicht, was im System passiert.

Beispiel 1.1 *Die Bundesbank verändert den Leitzinssatz, das ist die Eingabe, eine Stunde später setzt an der Börse in Frankfurt ein hektisches Kaufen ein, dies ist ein Teil des Ausgangs.*
Was passiert innen drin? Niemand weiß es genau!

Um das System vernünftig beschreiben zu können, brauchen wir ein mathematisches Modell, das möglichst gut an die Realität angepaßt sein soll. Dies geht im allgemeinen nie exakt und hängt von ganz vielen Parametern ab.

Fragen: Was sind Zustandsgrößen?
Diskrete Zeit, kontinuierliche Zeit?
Welche Größen können vernachläßigt werden?
Wie ist der funktionale Zusammenhang?

Dies ist i.a. der schwierigste Teil, weil hier Praktiker und Mathematiker zusammen gefragt sind.

Wir machen hier nur den theoretischen Teil für lineare Systeme. Modellierung etc. können wir hier aus Zeitgründen nicht behandeln, obwohl das sehr interessant ist.

Ein lineares Steuerungsproblem wird dann wie folgt definiert.

Definition 1.2 *Ein lineares Steuerungsproblem hat die Form*

$$\dot{x} = \frac{dx}{dt} = A(t)x(t) + B(t)u(t), t \in [t_0, \infty) \qquad \textit{Zustandsgleichung} \tag{1.3}$$

$$x(t_0) = x^0 \qquad \textit{Anfangswert} \tag{1.4}$$

$$y(t) = C(t)x(t) + D(t)u(t), t \in [t_0, \infty) \qquad \textit{Ausgangsgleichung} \tag{1.5}$$

Dabei sind

- $x(t)$ *Zustand,* $x(t) \in \mathcal{X}$ *Zustandsraum*
- $y(t)$ *Ausgang,* $y(t) \in \mathcal{Y}$ *Ausgangsraum*
- $u(t)$ *Eingang,* $u(t) \in \mathcal{U}$ *Eingangsraum*

$\mathcal{X}, \mathcal{Y}, \mathcal{U}$ *sind Mengen von Funktionen, die auf* $[t_0, \infty)$ *definiert sind, mit den folgenden Dimensionen:*

$$\begin{array}{rcccll} x(t) & : & [t_0, \infty) & \longrightarrow & \mathbb{R}^n & (\mathbb{C}^n) \\ y(t) & : & [t_0, \infty) & \longrightarrow & \mathbb{R}^p & (\mathbb{C}^p) \\ u(t) & : & [t_0, \infty) & \longrightarrow & \mathbb{R}^m & (\mathbb{C}^m) \end{array}$$

und die Systemmatrizen $A(t), B(t), C(t), D(t)$ *sind matrixwertige Funktionen:*

$$\begin{array}{rcccll} A(t) & : & [t_0, \infty) & \longrightarrow & \mathbb{R}^{n,n} & (\mathbb{C}^{n,n}) \\ B(t) & : & [t_0, \infty) & \longrightarrow & \mathbb{R}^{n,m} & (\mathbb{R}^{n,m}) \\ C(t) & : & [t_0, \infty) & \longrightarrow & \mathbb{R}^{p,n} & (\mathbb{C}^{p,n}) \\ D(t) & : & [t_0, \infty) & \longrightarrow & \mathbb{R}^{p,m} & (\mathbb{C}^{p,m}). \end{array}$$

Bemerkung 1.6 *In vielen Anwendungen, z.B. Ökonomie, ist das Zeitinterval diskret, d.h. es gibt feste Zeitpunkte* t_i*, z.B. Monate, Tage, Jahre.*

Wir werden hier im wesentlichen nur kontinuierliche Systeme betrachten. Fast alles geht analog im diskreten Fall. Der Vollständigkeit halber hier die Gleichungen

$$\begin{array}{rcll} x_{k+1} & = & A_k x_k + B_k u_k, \quad k = 0, 1, 2, ... & \textit{Zustandsgleichung} \quad (1.7) \\ x_0 & = & x^0 & \textit{Anfangswert} \quad (1.8) \\ y_k & = & C_k x_k + D_k u_k, \quad k = 0, 1, 2, ... & \textit{Ausgangsgleichung.} \quad (1.9) \end{array}$$

Um Steuerungsprobleme mathematisch zu betrachten, werden Zustandsübergangsfunktionen bzw. Ausgangsfunktionen eingeführt.

Definition 1.10 *Gegeben* $x^0 \in \mathcal{X}$ *und* $u(\cdot) \in \mathcal{U}$ *so definiert man als* *<u>Zustandsübergangsfunktion</u>* *die Abbildung*

$$\left(t_0, t, x^0, u(t)\right) \longrightarrow x(t) \in \mathcal{X}$$

wobei $x(t)$ *die Lösung der Zustandsgleichung (1.3) mit Anfangswert (1.4) zum Zeitpunkt* t *ist, und als* *<u>Ausgangsfunktion</u>* *die Abbildung*

$$(t, x(t), u(t)) \longrightarrow y(t) \in \mathcal{Y}$$

wobei $y(t)$ *das Ergebnis von (1.5) ist.*

Die Zustandsübergangsfunktion heißt auch oft *<u>Transferfunktion</u>* *oder* *<u>Input–Output Relation.</u>*

Beispiel 1.11 *Elektrisches Netzwerk, das Netzwerk besteht aus einem Widerstand und einem Kondensator.*

Eingang ist die angelegte Spannung u; Ausgang ist die Stärke des fließenden Stroms.

Sei C die Kapazität des Kondensators, $q(t)$ die Ladung am Kondensator. Dann ist $\frac{q(t)}{C}$ die Kondensatorspannung. Es gelten die Kirchhoff'schen Gesetze

$$u(t) = R\dot{q}(t) + \frac{q(t)}{C}, \tag{1.12}$$

umgeformt:

$$\dot{q}(t) = -\frac{1}{RC}q(t) + \frac{1}{R}u(t) \qquad \textit{Ausgangsfunktion} \tag{1.13}$$

$$\Longrightarrow A = -\frac{1}{RC}, B = \frac{1}{R}, x = y = q, C = 1, D = 0.$$

Der Zustand $q(t)$ kann durch Lösung der Differentialgleichung (1.13) bestimmt werden.

$$q(t) = e^{-(t-t_0)/RC}q(t_0) + \frac{1}{R}\int_{t_0}^{t} e^{-(t-s)/RC}u(s)ds. \tag{1.14}$$

Für gegebene Anfangsladung $q^0 = q(t_0)$ ist also die Zustandsübergangsfunktion gegeben durch:

$$\Big(t_0, t, q^0, u(t)\Big) \longrightarrow e^{-(t-t_0)/RC}q^0 + \frac{1}{R}\int_{t_0}^{t} e^{-(t-s)/RC}u(s)ds. \tag{1.15}$$

Ausgangsfunktion ist gleich Zustandsfunktion.

Hier sieht man auch gut, daß wir bei der Modellbildung vieles vernachlässigt haben, denn im allgemeinen ist z. B. R nicht konstant.

Eine wichtige Spezialklasse der Systeme, die wir betrachten werden, sind die autonomen oder zeitinvarianten Systeme. Im linearen Fall bedeutet das, daß

$$A(t), B(t), C(t), D(t) \qquad \text{konstant in t sind .}$$

Satz 1.16 *Die Zustandsübergangsfunktion zum System (1.3) (1.4) ist gegeben durch*

$$x(t) = \Phi(t, t_0)x^0 + \int_{t_0}^{t} \Phi(t, s)B(s)u(s)ds \tag{1.17}$$

wobei die Übergangsmatrix $\Phi(t, s)$ die (Fundamental-)lösung der homogenen Matrixdifferentialgleichung

$$\frac{\partial \Phi}{\partial t}(t, s) = A(t)\Phi(t, s), \quad \Phi(s, s) = I \tag{1.18}$$

ist. Im zeitinvarianten Fall erhalten wir speziell $\Phi = e^{A(t-s)}$ also

$$x(t) = e^{A(t-t_0)}x^0 + \int_{t_0}^{t} e^{A(t-s)}Bu(s)ds. \tag{1.19}$$

Beweis: (1.17) (1.19) folgen sofort durch Ableiten. □

Zur Wiederholung: Für eine Matrix A ist $f(A)$ definiert als die Auswertung der Taylorreihe von $f(t)$ mit A für t eingesetzt.

$$e^A = I + \frac{1}{1!}A + \frac{1}{2!}A^2 + \cdots$$

Der nächste wichtige Begriff ist die Stabilität. Dies ist natürlich insbesondere wichtig vom Anwendungsstandpunkt aus gesehen, weil es oft wichtig ist, Systeme in stabile Gleichgewichtslagen zu steuern (Flugzeuge, politische Systeme, ...).

Definition 1.20 *Das System (1.3), (1.4) heißt stabil, falls eine (und damit jede) Lösung der homogenen Gleichung $\dot{x} = A(t)x$ auf einem (und damit auf jedem) Intervall (t_0, ∞) beschränkt ist. Das System heißt asymptotisch stabil, falls außerdem $\lim_{t\to\infty} ||x(t)|| = 0$ gilt, wobei $|| \bullet ||$ irgendeine Norm ist.*

Asymptotisch stabil heißt einfach, daß jede Lösung der homogenen Gleichung für $t \to \infty$ gegen 0 geht.

Die Charakterisierung von stabilen Systemen ist eine wichtige Aufgabe der Steuerungstheorie (Lyapunovtheorie)

Satz 1.21 *Betrachte das zeitinvariante System $\dot{x} = Ax + Bu$, $x(t_0) = x^0$, A, B konstant.*

i) Das System ist asymptotisch stabil genau dann, wenn alle Eigenwerte von A negativen Realteil besitzen.

ii) Das System ist stabil genau dann, wenn alle Eigenwerte von A einen Realteil ≤ 0 haben und alle Eigenwerte mit Realteil 0 gleiche algebraische wie geometrische Vielfachheit haben.

Beweis: Übung. □

Beispiel 1.22 *a) (Bsp 1.11)*

$$\dot{q}(t) = -\frac{1}{RC}q(t) + \frac{1}{R}u(t)$$

ist skalar, A ist konstant $-\frac{1}{RC} \leq 0 \Longrightarrow$ System ist asymptotisch stabil.

b) (Bsp 0.2)

$$\begin{aligned} \dot{\varphi}(t) &= \omega(t) \\ \dot{\omega}(t) &= -\frac{r}{j}\omega(t) + \frac{k}{j}u(t) \end{aligned}$$

Setze $x(t) = \begin{bmatrix} \varphi \\ \omega \end{bmatrix}$ *so ist*

$$\begin{aligned} y(t) &= Cx(t) + Du(t) := [1,0]\, x(t) + 0\, u(t) \\ \dot{x}(t) &= \begin{bmatrix} 0 & 1 \\ 0 & \frac{-r}{j} \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ \frac{k}{j} \end{bmatrix} u(t) =: Ax(t) + Bu(t) \end{aligned}$$

Eigenwerte von A *sind* $0, \frac{-r}{j}$, *also ist das System stabil, aber nicht asymptotisch stabil.*

Steuerungsprobleme im Frequenzraum

Wenn man die meisten Ingenieurbücher anschaut, sieht man die Steuerungstheorie nicht im Zustandsraum dargestellt, sondern im sogenannten Frequenzraum. Die Idee dabei ist im wesentlichen, daß man das System Laplace transformiert und dann alles in den transformierten Variablen weiterrechnet.

Dazu werden wir erst kurz etwas über die Laplacetransformation wiederholen.

Definition 1.23 *Sei* $f(t)$ *eine reellwertige Funktion definiert auf* $[0, \infty)$. *Falls*

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

existiert, so heißt $\mathcal{L}(f)$ *die* <u>*Laplacetransformation*</u> *von* f. *Es gelten folgende Rechenregeln*

$$\left\{ \begin{aligned} \mathcal{L}(af(t) + bg(t)) &= a\mathcal{L}(f(t)) + b\mathcal{L}(g(t)) \\ \mathcal{L}(f'(t)) &= s\mathcal{L}(f(t)) - f(0) \\ \mathcal{L}(\textstyle\int_0^t f(z)dz) &= \tfrac{1}{s}\mathcal{L}(f(t)) \end{aligned} \right. \tag{1.24}$$

Betrachte nun das zeitinvariante System

$$\left\{ \begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \quad x(0) = x^0 \\ y(t) &= Cx(t) + Du(t) \quad A, B, C, D \text{ konstant} \end{aligned} \right. \tag{1.25}$$

Laplacetransformation ergibt

$$\begin{aligned} s\hat{x}(s) &= A\hat{x}(s) + B\hat{u}(s) + x^0 \\ \hat{y}(s) &= C\hat{x}(s) + D\hat{u}(s) \end{aligned} \tag{1.26}$$

$$\begin{aligned} \Longrightarrow \hat{x}(s) &= (sI - A)^{-1}B\hat{u}(s) + (sI - A)^{-1}x^0 \\ \Longrightarrow \hat{y}(s) &= \left[C(sI - A)^{-1}B + D\right] \hat{u}(s) + C(sI - A)^{-1}x^0 \\ &=: H(s)\hat{u}(s) + C(sI - A)^{-1}x^0 \end{aligned}$$

Die Elemente von $H(s)$ sind echt gebrochen rationale Funktionen von s, (schreibe $(sI - A)^{-1}$ über die Adjungierte) deren Pole in den Eigenwerten von A liegen.

Definition 1.27 *$H(s)$ heißt Übertragungsmatrix, oder für $s = i\omega$ heißt $H(i\omega)$ Frequenzübergangsmatrix.*

Wir werden im wesentlichen nur im Zustandsraum arbeiten, weil das numerisch günstiger ist. Es gibt (fast) keine stabilen Algorithmen für die Behandlung von Steuerungsproblemen im Frequenzraum.

Kapitel 2

Steuerbarkeit, Zustandsrückführung

Eine der wichtigsten Fragestellungen der Steuerungstheorie ist die Frage, ob man ein System durch geeignete Wahl von $u(t)$ in jeden beliebigen Zustand steuern kann. Dabei ist es im allgemeinen wichtig, die Menge der zulässigen Steuerfunktionen zu kennen. Eine Einschränkung dieser Menge kann dazu führen, daß das System nicht mehr in jeden Zustand gesteuert werden kann.

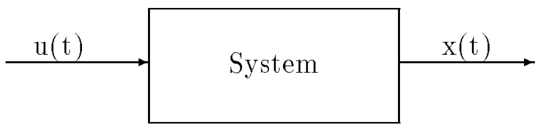
Wir betrachten zunächst das System

$$\dot{x} = A(t)x + B(t)u, \quad x(t_0) = x^0 \tag{2.1}$$

und eine Eingabemenge $\mathcal{U}$ von zulässigen Steuerfunktionen, dies sind typischerweise stetige oder stückweise stetige Funktionen.

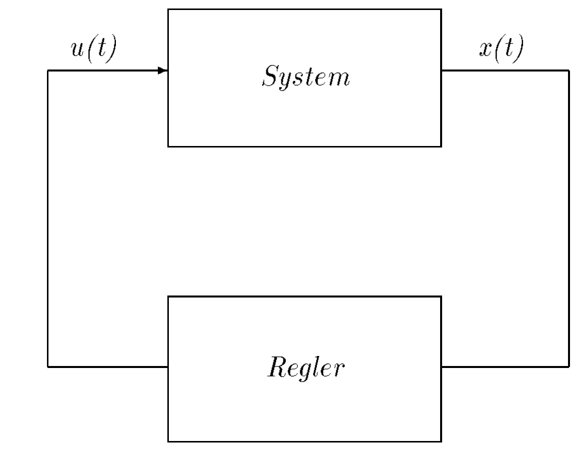
Definition 2.2 *Gegeben sei das System (2.1) und ein Endzustand x^1. Das Paar (t_0, x_0) heißt zur Zeit $t_1 > t_0$ nach x^1 steuerbar, falls es eine Steuerfunktion $u(t) \in \mathcal{U}$ gibt, so daß die Lösung von (2.1) mit dieser Steuerung $x(t_1) = x^1$ erfüllt. Das Paar (t_0, x_0) heißt nach x_1 steuerbar, wenn es zu irgendeiner Zeit $t_1, t_0 < t_1 < \infty$ nach x^1 steuerbar ist. Falls für jedes (t_0, x_0) und jedes x^1, das Paar (t_0, x_0) nach x^1 steuerbar ist, so heißt (2.1) vollständig steuerbar.*

Schematisch stellen wir System (2.1) dar als



Figur 2.3 *offener Kreis (open loop)*

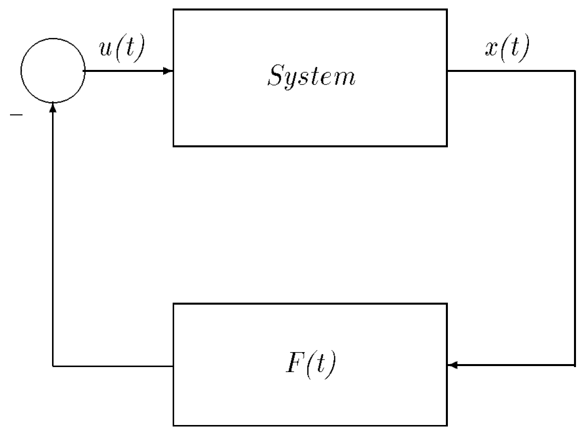
Dies ist ein sogenannter offener Kreis. Nun wählt man typischerweise $u(t)$ in Abhängigkeit vom Zustand, d.h. man macht eine Zustandsrückführung (feedback).



Figur 2.4 *Geschlossener Kreis (closed loop)*

Im linearen Fall macht man sogar gerne eine lineare Zustandrückführung $u(t) = -F(t)x(t)$ *und erhält dann für das geschlossene System die Differentialgleichung*

$$\dot{x} = (A(t) - B(t)F(t))\, x(t) \tag{2.5}$$



Figur 2.6 *Lineare Zustandsrückführung*

Wann ist nun ein System steuerbar?

Um dieses Problem zu lösen, betrachten wir das folgende analoge Problem:
Gegeben ein Zeitintervall $[t_0, t_1]$, bestimme alle x^0, so daß (t_0, x^0) nach $(t_1, 0)$ steuerbar ist. Sei $\mathcal{L}(t_0, t_1)$ die Menge dieser x^0, es muß also nach (1.17) gelten:

$$0 = x(t_1) = \Phi(t_1, t_0)x^0 + \int_{t_0}^{t_1} \Phi(t_1, s)B(s)u(s)ds \tag{2.7}$$

Ausklammern von $\Phi(t_1, t_0)$ von links ergibt

$$0 = x^0 + \int_{t_0}^{t_1} \Phi(t_0, s)B(s)u(s)\, ds, \tag{2.8}$$

denn wir können $\Phi(t_1, s)$ schreiben als

$$\Phi(t_1, s) = \Phi(t_1, t_0)\Phi(t_0, s) \tag{2.9}$$

denn $\Phi(t_1, s)$ ist der Übergang von $s \to t_1$ und $\Phi(t_0, s)$ ist der Übergang von $s \to t_0$ und da die Lösung von (1.18) eindeutig ist, so folgt (2.9).

Beispiel 2.10

$$\begin{aligned} \Phi(t_1, s) &= e^{A(t_1-s)} \\ \Phi(t_1, t_0)\Phi(t_0, s) &= e^{A(t_1-t_0)}e^{A(t_0-s)} \\ &= e^{A(t_1-s)} \end{aligned}$$

Dies heißt Halbgruppeneigenschaft der Lösung.
Also gilt $x^0 \in \mathcal{L}(t_0, t_1)$ genau dann, wenn (2.8) gilt.

Bevor wir zu einer Charakterisierung von $\mathcal{L}(t_0, t_1)$ kommen, brauchen wir das folgende wichtige Lemma.

Lemma 2.11 *Es sei $G(t)$ eine auf $(-\infty, +\infty)$ definierte matrixwertige stückweise stetige Funktion in $\mathbb{R}^{n,m}$ und sei $V = \int_{t_0}^{t_1} G(t)G(t)^T dt$ für $t_0 < t_1$. V ist eine symmetrische Matrix in $\mathbb{R}^{n,n}$. Es gilt: $x \in \mathbb{R}^n$ läßt sich genau dann als*

$$x = \int_{t_0}^{t_1} G(t)u(t)dt \tag{2.12}$$

mit einer stückweise stetigen Funktion $u(t) \in \mathbb{R}^m$ schreiben, wenn x im Bild von V liegt, d.h. es gibt ein $z \in \mathbb{R}^n$, so daß $x = Vz$.

Beweis: V ist positiv semidefinit, denn $G(t)G(t)^T$ ist positiv semidefinit und das Integral erhält diese Eigenschaft, denn es gilt für alle $x \in \mathbb{R}^n$

$$x^T V x = \int_{t_0}^{t_1} x^T G(t)G(t)^T x dt = \int_{t_0}^{t_1} ||G(t)^T x||_2^2 dt \geq 0. \tag{2.13}$$

Es gilt daher

$$Vx = 0 \Longleftrightarrow G(t)^T x = 0 \quad \forall t \in [t_0, t_1]. \tag{2.14}$$

Die Menge aller x, die sich mit geeignetem $u(t)$ in der Form (2.12) schreiben lassen, bilden wegen der Linearität des Integrals und der Multiplikation mit $G(t)$ einen linearen Raum $\mathcal{L}$. Der Bildraum von V ist in $\mathcal{L}$ enthalten. Bild $(V) \subset \mathcal{L}$. Dies folgt sofort aus der Definition von V mit der Wahl

$$u(t) = G(t)^T z, z = \text{ konstant }.$$

Was wir zeigen müssen, ist daß $\mathcal{L} =$ Bild (V), bzw. daß $\mathcal{L}\cap$ Kern $(V) = \{0\}$ ist. Es gilt für $x \in \mathcal{L}\cap$ Kern (V)

$$||x||^2 = x^T x = \int_{t_0}^{t_1} x^T (G(t)u(t))dt = \int_{t_0}^{t_1} \left(G(t)^T x\right)^T u(t)dt = 0 \tag{2.15}$$

$$\Longrightarrow x = 0.$$

□

Wir führen nun die folgende Matrix ein:

$$W(t_0,t_1) := \int_{t_0}^{t_1} \Phi(t_0,t)B(t)B(t)^T\Phi(t_0,t)^T dt \tag{2.16}$$

und erhalten den folgenden Satz:

Satz 2.17

(i) $\mathcal{L}(t_0,t_1)$ *ist der Bildraum der Matrix* $W(t_0,t_1)$ *d.h.*

$$\mathcal{L}(t_0,t_1) = \{x \mid \text{ es gibt ein } z \text{ mit } x = W(t_0,t_1)z\} \tag{2.18}$$

(ii) $W(t_0,t_1)x = 0 \iff x^T\Phi(t_0,t)B(t) \equiv 0\,\forall t \in [t_0,t_1]$.

Beweis: Lemma 2.11 (sowie (2.14)) mit $G(t) = \Phi(t_0,t)B(t)$. □

Mit diesem Satz können wir nun Steuerbarkeit charakterisieren. Dazu brauchen wir noch die zu $\dot{x} = Ax$ adjungierte Gleichung

$$\dot{z} = -A(t)^T z\ , \tag{2.19}$$

die die Übergangsmatrix $\Phi(\tau,t)^T$ hat.

Satz 2.20 *Das System (2.1) ist steuerbar genau dann, wenn folgende Aussage gilt: Ist* $z(t)$ *eine nichttriviale Lösung von (2.19), so verschwindet* $z(t)^TB(t)$ *auf keinem Intervall* $[t_0,\infty)$ *identisch.*

Beweis: Sei (2.1) steuerbar. Angenommen es gibt ein $z(t) \not\equiv 0$, Lösung von (2.19), so daß

$$z(t)^TB(t) = 0 \quad \forall t \geq t_0, \tag{2.21}$$

so gibt es Anfangswerte x^0, so daß (t_0,x_0) in keiner Zeit $t_1 > t_0$ steuerbar ist. Ist $x(t)$ eine Lösung von (2.1) für irgendein $u(t)$, so gilt

$$\frac{d}{dt}\Big(x(t)^Tz(t)\Big) = 0 \quad \forall t \geq t_0, \tag{2.22}$$

d.h. $x(t)^Tz(t) = x_0^Tz_0$ konstant, wobei $x_0 = x(t_0), z_0 = z(t_0)$. Man braucht nun x_0 nur so wählen, daß $x_0^Tz_0 \neq 0$, so ergibt sich ein Widerspruch.

Die Umkehrung ist komplizierter. Das geschieht in 3 Schritten.

<u>1. Schritt</u>
Wir beweisen zuerst: Zu jedem t_0 gibt es $t_1 > t_0$, so daß für jede nichttriviale Lösung von (2.19) $z(t)^TB(t) \not\equiv 0$ in $[t_0,t_1]$. Wäre dies falsch, so gäbe es eine Folge $t_v \to \infty$ und zugehörige Lösungen $z_v(t)$ von (2.19) mit

$$||z_v(t_0)|| = 1,\ z_v^T(t)B(t) = 0,\ \ \forall t \in [t_0,t_v] \tag{2.23}$$

Wir können o. B. d. A. annehmen, daß die Folge $z_v(t_0)$ konvergiert, ansonsten wählen wir eine konvergente Teilfolge nach Bolzano-Weierstraß aus.

Sei $z^*_\infty = \lim\limits_{v\to\infty} z_v(t_0)$. Dann ist $||z^*_\infty|| = 1$ und damit ist die Lösung von (2.19) mit $z_\infty(t_0) = z^*_\infty$ nicht trivial. Nach Voraussetzung gibt es nun $t_1^* > t_0$, so daß $z_\infty(t_1^*)^T B(t_1^*) \neq 0$. Wegen der stetigen Abhängigkeit der Lösung der Differentialgleichung von den Anfangswerten folgt, daß die Folge $z_v(t)$ gleichmäßig in t auf $[t_0, t_1^*)$ gegen $z_\infty(t)$ konvergiert. Daher gilt auch

$$z_v(t_1^*)^T B(t_1^*) \neq 0$$

für alle genügend großen v. Dies ist ein Widerspruch zu (2.23).

<u>2. Schritt</u>
Wir zeigen nun: Wenn für jede nichttriviale Lösung $z(t)$ von (2.19), $z(t)^T B(t)$ auf $[t_0, t_1]$ nicht identisch verschwindet, so ist $W(t_0, t_1)$ positiv definit. Da $W(t_0, t_1)$ positiv semidefinit ist, so reicht es zu zeigen, daß $W(t_0, t_1)$ nicht singulär ist. Sei x so, daß $W(t_0, t_1)x = 0 \Longrightarrow B(t)^T \Phi(t_0, t)^T x \equiv 0$ für alle $t \in [t_0, t_1]$. Setze $z(t) = \Phi(t_0, t)^T x$, so löst $z(t)$ (2.19), da Φ die Fundamentallösung ist, also muß $x = 0$ sein.

<u>3. Schritt</u>
Wenn $W(t_0, t_1)$ postiv definit ist, so ist jedes Paar (t_0, x^0) in jedem x^1 zur Zeit t_1 steuerbar und zwar mit

$$u(t) = B(t)^T \Phi(t_0, t)^T c \tag{2.24}$$

und c ist eindeutig bestimmt aus der Gleichung

$$\begin{aligned} x^0 &= \Phi(t_0, t_1)x^1 + \int_{t_1}^{t_0} \Phi(t_0, t)B(t)B(t)^T \Phi(t_0, t)^T c dt \\ &= \Phi(t_0, t_1)x^1 - W(t_0, t_1)c \end{aligned}$$

denn $W(t_0, t_1)$ ist nichtsingulär. $\square$

Korollar 2.25 *Das System 2.1 ist steuerbar genau dann, wenn es zu jedem t_0 ein t_1 gibt, so daß $W(t_0, t_1)$ postiv definit ist. Es ist dann in dieser Zeit t_1 jedes (t_0, x^0) in jedes x steuerbar.*

Nun noch zum zeitinvarianten Fall.

Satz 2.26 *Für ein zeitinvariantes System $\dot{x} = Ax + Bu$ ist $\mathcal{L}(t_0, t_1)$ von t_0, t_1 unabhängig und gleich dem von den Spalten von der <u>Steuerbarkeitsmatrix</u>*

$$K := (B, AB, A^2B, \cdots, A^{n-1}B)$$

aufgespannten Raum, d.h. $\mathcal{L}(t_0, t_1) = Bild(K)$.

Beweis: Aus Satz 2.17 folgt

$$\begin{aligned} \mathcal{L}(t_0, t_1) &= Bild(W(t_0, t_1)) \\ \Longleftrightarrow \mathcal{L}(t_0, t_1)^\perp &= Kern(W(t_0, t_1)) \end{aligned}$$

Wir müssen also zeigen

$$W(t_0, t_1)x = 0 \Longleftrightarrow x^T K = 0 \tag{2.27}$$

Satz 2.17 ii) besagt

$$W(t_0,t_1)x = 0 \Longleftrightarrow x^T e^{At} B = 0 \quad \forall t \in [t_0 - t_1, 0] \tag{2.28}$$
$$\Longleftrightarrow x^T A^v B = 0 \quad \forall v = 0,1,2,...$$

Der Satz von Cayley–Hamilton sagt, daß $\varphi(A) = 0$, (φ charakteristisches Polynom von A), daraus folgt

$$x^T A^v B = \sum_{j=0}^{n-1} \alpha_j^{(v)} x^T A^j B \quad \forall v \geq n$$

für geeignete $\alpha_1^{(v)}, \ldots, \alpha_{n-1}^{(v)} \in \mathbb{C}$. Also ist $x^T A^v B = 0 \quad \forall v = 0,1,2,3...$

$$\begin{aligned} &\Longleftrightarrow \quad x^T A^v B = 0 \quad \forall v = 0, \cdots, n-1 \\ &\Longleftrightarrow \quad x^T K = 0. \end{aligned}$$

Satz 2.29 *Die folgenden Aussagen sind äquivalent:*

i) Das zeitinvariante System $\dot{x} = Ax + Bu$ ist steuerbar;

ii) Rang $(K) = n$;

iii) ist $p \neq 0$ ein Eigenvektor zu A^T so gilt $p^T B \neq 0$;

iv) Rang $(\lambda I - A, B) = n \quad \forall \lambda \in \mathbb{C}$.

Beweis: Übungsaufgabe □

Beispiel 2.30 *Wir betrachten unsere Beispiele (0.2), (1.11).*
Beispiel 1.11. Da $A = \frac{-1}{RC}, B = \frac{1}{R}$ skalar $\neq 0$, so folgt Rang $\left(\frac{1}{R}\right) = 1 \Longrightarrow$ das System ist steuerbar.
Beispiel 0.2 $K = \begin{bmatrix} 0 & \frac{k}{j} \\ \frac{k}{j} & \frac{-rk}{j^2} \end{bmatrix}$ hat Rang 2, also ist das System steuerbar.

In vielen Fällen ist man nicht an der Steuerung in einen beliebigen Zustand interessiert, sondern an der Steuerung in eine Gleichgewichtslage. Das kann man sicher tun, wenn das System steuerbar ist, also kommt man mit weniger aus.

Definition 2.31 *Ein lineares System der Form (2.1) heißt <u>stabilisierbar</u>, wenn es zu jedem (t_0, x_0) eine für alle $t \geq t_0$ definierte, stückweise stetige Steuerfunktion $u(t)$ gibt, so daß die Lösung von (2.1) mit diesem $u(t)$ die Bedingung $\lim_{t\to\infty} x(t) = 0$ erfüllt.*

Ein notwendiges Kriterium für Stabilisierbarkeit ist der folgende

Satz 2.32 *Wenn das System (2.1) stabilisierbar ist, so gilt: Ist $z(t)$ eine Lösung der adjungierten Gleichung (2.19) die für $t \to \infty$ beschränkt ist, so kann nicht gelten $z(t)^T B(t) \equiv 0$ für alle $t \geq t_0$.*

Beweis: Angenommen (2.1) ist stabilisierbar, es gibt jedoch eine nichttriviale Lösung von (2.19) für die gilt:

$$z(t)^T B(t) \equiv 0 \quad \forall t \geq t_0 \quad \lim_{t\to\infty} ||z(t)|| \neq \infty \tag{2.33}$$

Wie beim Beweis von Satz (2.20) gilt dann für jede Wahl von $u(t)$ und die zugehörige Lösung von (2.1)

$$x(t)^T z(t) = x^{0^T} z(t_0) \quad \forall t \geq t_0$$

Aus $\lim_{t\to\infty} ||x(t)|| = 0$ und $\lim_{t\to\infty} ||z(t)|| \neq \infty$ folgt, daß es eine Folge $t_i \to \infty$ gibt, so daß $\lim_{i\to\infty} x(t_i)^T z(t_i) = 0$. Also muß $x^{0^T} z(t_0) = 0$ sein, d.h. nur solche Systeme können stabilisiert werden, für die $x^{0^T} z(t_0) = 0$ gilt. Dies steht aber in Widerspruch zur Stabilisierbarkeit. □

Allgemein ist eine hinreichende Bedingung für dieses Problem nicht bekannt. Im zeitinvarianten Fall erhalten wir

Satz 2.34 *Die folgenden Bedingungen sind äquivalent:*

i) Das System $\dot{x} = Ax + Bu$ mit A, B konstant, ist stabilisierbar;

ii) Ist λ mit $Re(\lambda) \geq 0$ Eigenwert von A mit Linkseigenvektor p, so gilt $p^T B \neq 0$;

iii) Rang $(\lambda I - A, B) = n \quad \forall \lambda$ mit $Re(\lambda) \geq 0$.

Beweis: i) $\longrightarrow$ ii)
Ist λ mit $Re(\lambda) \geq 0$ Eigenwert von A mit Linkseigenvektor p, so lösen $Re(e^{-\lambda t})p, Im(e^{-\lambda t})p$ (2.19) und sind beschränkt. Satz 2.32 impliziert $p^T B \neq 0$.
ii)$\Longrightarrow$ iii) Angenommen es gibt ein $\lambda \in \mathbb{C}, Re(\lambda) \geq 0$, so daß Rang $(\lambda I - A, B) < n \Longrightarrow$ es gibt ein $p \neq 0$, so daß $p^T(\lambda I - A, B) = 0 \Longrightarrow p$ ist Linkseigenvektor von A und $p^T B = 0$.
iii) $\Longrightarrow$ i) Rang $(\lambda I - A, B) = n \;\forall \lambda$ mit $Re(\lambda) \geq 0 \Longrightarrow$. Es gibt eine Matrix $F \in \mathbb{R}^{m,n}$, so daß alle Eigenwerte von $A - BF$ negativen Realteil haben.
Wäre das nicht so, so würde es ein λ mit $Re(\lambda) \geq 0$ geben und $p \neq 0$, so daß $p^T(A - BG - \lambda I) = 0 \quad \forall G \in \mathbb{R}^{m,n}$, das widerspricht iii). Also stabilisiert $u(t) = Fx(t)$ das System. □

Hier haben wir unser erstes numerisches Problem.

Aufgabe 2.35

i) Gegeben A, B finde $F \in \mathbb{R}^{m,n}$, so daß $Re(\lambda) < 0 \;\forall$ Eigenwerte von $A - BF$.

ii) Mache dies so, daß kleine Störungen von $A - BF$ die Eigenwerte nicht wieder in die rechte Halbebene bringen.

iii) Entscheide numerisch ob ein System steuerbar oder stabilisierbar ist.

Kapitel 3

Rekonstruierbarkeit, Ausgangsrückführung

Typischerweise können wir beim System

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \;\; y(t) = C(t)x(t), \;\; x(t_0) = x^0 \tag{3.1}$$

nicht den ganzen Zustandsraum sehen, d.h. er ist für Messungen nicht zugänglich, sondern nur Teile oder Funktionen des Zustands sind zugänglich. Daher ist eine der wichtigsten Fragestellungen, ob es möglich ist, so viel wie möglich an Informationen über den Zustand aus dem Ausgang herzuleiten bzw. mit dem Ausgang zu steuern. Wir werden zeigen, daß diese Konzepte dual zu denen im vorigen Kapitel sind.

Definition 3.2 *Das System (3.1) heißt <u>rekonstruierbar</u>, falls folgende Aussage für alle* t_0 *gilt:*
Sind $x, \tilde{x}$ *Lösungen von (3.1) mit der gleichen Steuerfunktion* $u(t)$ *und gilt*

$$C(t)x(t) = C(t)\tilde{x}(t) \quad \forall t \leq t_0, \tag{3.3}$$

so gilt

$$x(t) = \tilde{x}(t) \quad \forall t \leq t_0. \tag{3.4}$$

Man kann dies auch folgendermaßen ausdrücken:
Falls zwei Systeme die gleichen Eingänge und Ausgänge haben für $t \leq t_0$, so sind auch die Zustände gleich für $t \leq t_0$. Es ist klar, daß Rekonstruierbarkeit unabhängig vom Eingang $u(t)$ ist. (Bilde $x - \tilde{x}$.)
Man spricht in vielen Lehrbüchern auch oft von <u>Beobachtbarkeit</u> (Observability). Der Unterschied ist eigentlich nur bei zeitvariablen Sytemem relevant.

Definition 3.5 *Das System (3.1) heißt <u>beobachtbar</u>, falls folgende Aussage für alle* t_0 *gilt:*
Sind $x, \tilde{x}$ *Lösungen von (3.1) mit der gleichen Steuerfunktion* $u(t)$ *und gilt*

$$C(t)x(t) = C(t)\tilde{x}(t) \; \forall t \geq t_0 \tag{3.6}$$

so gilt

$$x(t) = \tilde{x}(t) \quad \forall t \geq t_0 \tag{3.7}$$

Rekonstruierbarkeit beschränkt sich auf die Vergangenheit, Beobachtbarkeit auf die Zukunft von t_0 aus gesehen.

Um zu sehen, daß die Konzepte von Rekonstruierbarkeit und Steuerbarkeit dual sind, haben wir den folgenden Satz.

Satz 3.8 *Das System (3.1) ist rekonstruierbar genau dann, wenn das System*

$$\dot{x} = A(-t)^T x + C(-t)^T u \tag{3.9}$$

steuerbar ist.
(Dualität $\Longleftrightarrow$ transponieren und Rollen von B und C vertauschen.)

Beweis: Bilde $\tilde{x} - x = z$. Es ist klar, daß (3.1) rekonstruierbar ist, genau dann wenn $C(t)z(t) = 0 \ \forall t \leq t_0$ impliziert, daß $z(t) \equiv 0 \ \forall t \leq t_0$, oder äquivalent: Falls $z(t) \not\equiv 0$ Lösung von $\dot{z} = A(t)z$ ist, so verschwindet $C(t)z(t)$ auf keinem Intervall $[-\infty, t_0]$ identisch. Dies ist aber äquivalent damit, daß (3.9) steuerbar ist, nach Satz 2.20. $\square$

Wir erhalten auch sofort Kriterien mit Hilfe der Übergangsmatrix:

Korollar 3.10 *System (3.1) ist rekonstruierbar genau dann, wenn es zu jedem t_1 ein $t_0 < t_1$ gibt, so daß*

$$\hat{W}(t_0, t_1) := \int_{t_0}^{t_1} \Phi(t, t_1)^T C(t)^T C(t) \Phi(t, t_1) dt \tag{3.11}$$

positiv definit ist. Dabei ist $\Phi(t, \tau)$ die Übergangsmatrix von $\dot{x} = A(t)x$. Es besteht die folgende Beziehung zwischen u, y für $t_0 \leq t \leq t_1$ und $x(t_1) = x^1$:

$$\int_{t_0}^{t_1} \Phi(t, t_1)^T C(t)^T y(t) dt = \hat{W}(t_0, t_1) x^1 + \int_{t_0}^{t_1} \left(\Phi(t, t_1)^T C(t)^T C(t) \int_{t_1}^{t} \Phi(t, s) B(s) u(s) ds \right) dt. \tag{3.12}$$

x^1 aus (3.12) ist für rekonstruierbare Systeme eindeutig bestimmt.

Beweis: Die erste Aussage ist klar nach Korollar 2.24 angewandt auf (3.9). (3.12) folgt aus der Formel für die Lösung von (3.1), mit „Anfangszustand" $x(t_1) = x^1$ bei t_1.

$$x(t) = \Phi(t, t_1) x(t_1) + \int_{t_1}^{t} \Phi(t, s) B(s) u(s) ds$$

durch Multiplikation mit $\Phi(t, t_1)^T C(t)^T C(t)$ und anschließender Integration von t_0 nach t_1. Die eindeutige Lösbarkeit folgt aus der Definitheit von $\hat{W}(t_0, t_1)$. $\square$

Wegen der Dualität folgt für zeitinvariante Systeme sofort das folgende Korollar.

Korollar 3.13 *Die folgenden Aussagen sind äquivalent:*

i) Das zeitinvariante System

$$\dot{x} = Ax + Bu, y = Cx \tag{3.14}$$

ist rekonstruierbar.

ii) Das zeitinvariante System (3.14) ist beobachtbar.

iii) *Rang* $(\tilde{K}) = n$, *wobei*
$\tilde{K} = \left[C^T, A^T C^T, \ldots (A^T)^{n-1} C^T\right]$.

iv) *Ist* $p \neq 0$ *ein Eigenvektor von* A, *so gilt* $C^T p \neq 0$.

v) *Rang* $\begin{bmatrix} \lambda I - A \\ C \end{bmatrix} = n \quad \forall \lambda \in \mathbb{C}$.

Beweis: Klar mit Dualität. □

Gibt es auch einen dualen Begriff zur Stabilisierbarkeit? Wir beschränken uns auf den zeitinvarianten Fall.

Definition 3.15 *Das zeitinvariante System*

$$\dot{x} = Ax + Bu, \quad y = Cx$$

heißt entdeckbar*, wenn die folgende Aussage gilt:*
Ist für $t_0 < 0$ z *im Kern von* $\hat{W}(t_0, 0)$ *so folgt* $\lim_{t \to \infty} \Phi(t,0)z = \lim_{t \to \infty} e^{At} z = 0$.

Es gilt sofort der folgende Satz:

Satz 3.16 *Betrachte das zeitinvariante System (3.14). Die folgenden Aussagen sind äquivalent.*

i) *Ist* $x(t)$ *Lösung von* $\dot{x} = Ax$ *und gilt* $Cx(t) \equiv 0$ *so folgt* $\lim_{t \to \infty} x(t) = 0$.

ii) *(3.14) ist entdeckbar.*

iii) *Das duale System*

$$\dot{x} = A^T x + C^T u \tag{3.17}$$

ist stabilisierbar.

iv) *Ist* λ *mit* $Re(\lambda) \geq 0$ *Eigenwert von* A *mit Eigenvektor* $p \neq 0$, *so gilt* $C^T p \neq 0$.

v) *Rang* $\left[\begin{smallmatrix} \lambda I - A \\ C \end{smallmatrix}\right] = n \ \forall \lambda$ *mit* $Re(\lambda) \geq 0$.

Ist λ *mit* $Re(\lambda) \geq 0$ *Eigenwert von* A *mit Eigenvektor* $p \neq 0$, *so gilt* $C^T p \neq 0$. *Rang* $\begin{bmatrix} \lambda I - A \\ C \end{bmatrix} = n \quad \forall \lambda$ *mit* $Re(\lambda) \geq 0$.

Beweis: i) $\Longleftrightarrow$ ii) ist Satz 2.33. ii) $\Longleftrightarrow$ iii) $\Longleftrightarrow$ iv) ist Satz 2.34. □

Hier kommt unser nächstes Problem.

Aufgabe 3.18

i) *Gegeben* A, C. *Finde* $F \in \mathbb{R}^{n,p}$, *so daß* $Re(\lambda) < 0$ $\forall$ *Eigenwerte von* $A - FC$.

ii) *Mache dies so, daß kleine Störungen von* $A - FC$ *die Eigenwerte nicht wieder in die rechte Halbebene bringen. „Robustheit"*

Eine noch komplexere Aufgabe, die wir später studieren werden, ist die Stabilisierung durch Ausgangsrückführung:

Aufgabe 3.19

i) *Gegeben* A, B, C. *Finde* $F \in \mathbb{R}^{m,p}$, *so daß* $Re(\lambda) < 0\ \forall$ *Eigenwerte von* $A - BFC$.

ii) *Mache dies, so daß kleine Störungen von* $A - BFC$ *diese Eigenschaft nicht wieder zerstören.*

iii) *Entscheide numerisch ob ein System rekonstruierbar oder entdeckbar ist.*

Kapitel 4

Äquivalenzumformungen

Wir betrachten nun zeitinvariante Systeme

$$\dot{x} = Ax + Bu, \quad y = Cx \tag{4.1}$$

und schauen uns an, welche linearen Transformationen wir mit dem System durchführen können, die die wichtigen Systemeigenschaften wie Steuerbarkeit und Stabilisierbarkeit erhalten.

Die wichtigsten dieser Transformationen sind die folgenden:

$$\begin{array}{llll}
x \longmapsto Px & P \in \mathbb{R}^{n,n} & \text{nichtsingulärer Basiswechsel} & (4.2)\\
u \longmapsto Qu & Q \in \mathbb{R}^{m,m} & \text{nichtsingulärer Basiswechsel} & (4.3)\\
y \longmapsto Ry & R \in \mathbb{R}^{p,p} & \text{nichtsingulärer Basiswechsel} & (4.4)\\
u \longrightarrow -Fx + u & F \in \mathbb{R}^{m,n} & \text{lineares Zustandsfeedback} & (4.5)\\
u \longrightarrow -Gy + u & G \in \mathbb{R}^{m,p} & \text{lineares Ausgangsfeedback} & (4.6)
\end{array}$$

Dies sind nicht alle linearen Transformationen die möglich sind, aber wir werden nur diese betrachten. Man kann z.B. auch

$$u \longrightarrow -F\dot{x} + u \quad F \in \mathbb{R}^{m,n} \text{ lineares Ableitungsfeedback} \tag{4.7}$$

oder

$$u \longrightarrow -G\dot{y} + u \quad G \in \mathbb{R}^{m,p} \text{ lineares Ableitungsausgangsfeedback} \tag{4.8}$$

machen. Eigentlich sind dieses alles lineare Transformationen, die auf der Blockmatrix

$$\begin{array}{cc}
\begin{array}{c} n \\ p \end{array} & \left[\begin{array}{cc} A & B \\ C & 0 \end{array}\right] \\
 & \begin{array}{cc} n & m \end{array}
\end{array} \tag{4.9}$$

operieren und zwar durch

$$\left[\begin{array}{cc} P^{-1} & 0 \\ 0 & R^{-1} \end{array}\right]\left[\begin{array}{cc} A & B \\ C & 0 \end{array}\right]\left[\begin{array}{cc} P & 0 \\ -F & Q \end{array}\right] =: \left[\begin{array}{cc} \tilde{A} & \tilde{B} \\ \tilde{C} & 0 \end{array}\right] \tag{4.10}$$

für (4.2), (4.3), (4.4) und (4.5) bzw.

$$\left[\begin{array}{cc} P^{-1} & 0 \\ 0 & R^{-1} \end{array}\right]\left[\begin{array}{cc} A & B \\ C & 0 \end{array}\right]\left[\begin{array}{cc} P & 0 \\ -GC & Q \end{array}\right] =: \left[\begin{array}{cc} \hat{A} & \hat{B} \\ \hat{C} & 0 \end{array}\right] \tag{4.11}$$

für (4.2), (4.3), (4.4) und (4.6).

Es ist klar, daß diese Transformationen das System äquivalent in ein anderes überführen, wichtig ist der folgende Satz:

Satz 4.12 *Falls das System (4.1)*

$$\left\{\begin{array}{l} \textit{steuerbar} \\ \textit{rekonstruierbar} \\ \textit{stabilisierbar} \\ \textit{entdeckbar} \end{array}\right\} \textit{ist, so sind auch}$$

die Systeme

$$\dot{\tilde{x}} = \tilde{A}\tilde{x} + \tilde{B}\tilde{u}, \quad \tilde{y} = \tilde{C}\tilde{x} \tag{4.13}$$

und

$$\dot{\hat{x}} = \hat{A}\hat{x} + \hat{B}\hat{u}, \quad \hat{y} = \hat{C}\hat{x} \tag{4.14}$$

mit $\tilde{A}, \tilde{B}, \tilde{C}$ *wie in (4.10) und* $\hat{A}, \hat{B}, \hat{C}$ *wie in (4.11)*

$$\left\{\begin{array}{l} \textit{steuerbar} \\ \textit{rekonstruierbar} \\ \textit{stabilisierbar} \\ \textit{entdeckbar} \end{array}\right\}$$

Beweis: Wir zeigen dies für steuerbar und (4.13), alle anderen Fälle gehen analog. Verwende Satz 2.29, iv).

$$\begin{aligned} \operatorname{Rang}[\lambda I - A, B] &= \operatorname{Rang} P^{-1}[\lambda I - A, B] \begin{bmatrix} P & 0 \\ -F & Q \end{bmatrix} \\ &= \operatorname{Rang}[\lambda I - \tilde{A}, \tilde{B}], \quad \text{da } P, Q \text{ nichtsingulär.} \end{aligned}$$

□

Wir können also unser System durch diese Transformationen in einfachere Formen transformieren, ohne die Eigenschaften zu ändern. (Beachte: die Variablen müssen natürlich auch transformiert werden.)

Nun betrachtet man für die theoretische Analyse Normalformen unter diesen Äquivalenztransformationen. Dies sind z.B. die Brunovsky Form in der A auf Jordanform und B auf reduzierte Form gebracht werden, siehe [Kn/Kw].

Diese Transformationen sind für die numerische Berechnung ungeeignet, da sie im allgemeinen nicht beschränkt sind. Transformation auf Jordanform ist numerisch nicht sinnvoll durchführbar, da das Problem schlecht konditioniert ist. Kleine Störungen können fundamentale Änderungen der Jordanblöcke bewirken. Das einzige was man typischerweise vernünftig machen kann, ist über die Größe von Jordan–Blöcken zu entscheiden. Um dies numerisch rückwärts stabil zu machen, brauchen wir orthogonale (bzw. im komplexen unitäre) Transformationen und wenden diese in der Form (4.2) bis (4.4) an, und wir versuchen auch normbeschränkte Feedbacks zu verwenden.

Um dies zu tun, verwenden wir die folgenden Zerlegungen:

Satz 4.15 *(Singulärwertzerlegung, SVD) Gegeben* $A \in \mathbb{R}^{n,m}$. *Dann gibt es orthogonale Matrizen* U, V *mit* $U \in \mathbb{R}^{n,n}, V \in \mathbb{R}^{m,m}$, *so daß* $U = [u_1, \cdots, u_n], V = [v_1, \cdots, v_m], A = U\Sigma V^T$ *und* $\Sigma \in \mathbb{R}^{n,m}$ *„diagonal" ist, d.h.*

$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \\ & 0 & \end{bmatrix} \text{ für } n > m, \text{ bzw. } \Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_n & \end{bmatrix}$$

falls $m \geq n$, *die* σ_i *sind geordnet als* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_s \geq 0,\ s = \min(m, n)$.

Beweis: Seien $x \in \mathbb{R}^m, y \in \mathbb{R}^n$, so daß $\|x\|_2 = \|y\|_2 = 1$ und $Ax = \sigma_1 y$, wobei $\sigma_1 = \|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$. Da dieses Supremum angenommen wird, so existiert x.
Ergänze x, y zu Orthonormalbasen von $\mathbb{R}^n, \mathbb{R}^m$.

$$V = [x, V_1] \in \mathbb{R}^{m,m}, \quad U = [y, U_1] \in \mathbb{R}^{n,n}.$$

Es folgt, daß

$$\hat{A} = U^T A V = \begin{array}{cc} \begin{bmatrix} \sigma_1 & w^T \\ 0 & A_1 \end{bmatrix} & \begin{array}{c} 1 \\ n-1 \end{array} \\ \begin{array}{cc} 1 & m-1 \end{array} & \end{array}$$

Da $\|\hat{A} \begin{bmatrix} \sigma_1 \\ w \end{bmatrix}\|_2^2 \geq (\sigma_1^2 + w^T w)^2$, so folgt

$$\|\hat{A}\|_2^2 \geq \sigma_1^2 + w^T w$$

Aber da $\sigma_1^2 = \|A\|_2^2 = \|\hat{A}\|_2^2$, so folgt $w = 0$. Per Induktion können wir nun A_1 behandeln. $\square$

Die

σ_i heißen Singulärwerte,
v_i heißen rechte Singulärvektoren,
u_i heißen linke Singulärvektoren.

Korollar 4.16 *Sei* $A = U\Sigma V^T$ *die SVD von A und gelte*
$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_s = 0$. *Dann gilt*

i) *Rang* $(A) = r$,

ii) *Kern* $(A) =$ *span* $\{v_{r+1}, \cdots, v_m\}$

iii) *Bild* $(A) =$ *span* $\{u_1, \cdots, u_r\}$

iv) $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$

v) $||A||_F^2 = \sum_{j=1}^{n} \sum_{i=1}^{m} | a_{ij} |^2 = \sigma_1^2 + \cdots + \sigma_s^2$

vi) $||A||_2 = \sigma_1$.

vii) Sei $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ *für* $k < r =$ *Rang* (A), *so gilt*

$$\min_{\substack{B \in \mathbb{R}^{n,m} \\ Rang\,(B)=k}} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1}.$$

Beweis: Übung □

Die Berechnung der SVD wird in den Ergänzungen gemacht. Die SVD ist die <u>beste</u> numerische Methode zur Bestimmung von Rängen.

Warum nimmt man nicht die QR–Zerlegung?

Beispiel 4.17 *Sei*

$$T_n(c) = \ diag\,(1, s, \cdots, s^{n-1}) \begin{bmatrix} 1 & -c & \cdots & -c \\ & \ddots & \ddots & \\ & & \ddots & -c \\ & & & 1 \end{bmatrix}$$

$$c^2 + s^2 = 1 \quad c, s > 0$$

für $T_{100}(.2)$ *ist* $s^{99} = 0.13$, *d.h. alle Diagonalelemente sind weit weg von* 0, *jedoch ist,* $\sigma_n(T_{100}(.2)) \approx 10^{-8}$, *d.h. z. B. in „single precision" ist diese Matrix singulär.*

Etwas besser als die QR Zerlegung, <u>insbesondere in der Praxis</u>, sind sogenannte QRP Zerlegungen der Form

$$A = QRP = \begin{array}{cc} \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix} & \begin{array}{c} r \\ n-r \end{array} \\ \begin{array}{cc} r & \quad m-r \end{array} & \end{array} \tag{4.18}$$

wobei $r =$ Rang (A), Q orthogonal, R_{11} obere $\triangle$–Matrix, P Permutationsmatrix.

P wird so gewählt, daß immer die Spalte mit der größten Norm zuerst eliminiert wird. Genaues in der Ergänzung. Es folgt auf jeden Fall, daß die Diagonalelemente von R_{11} in fallender Folge geordnet sind.

Beachte, daß Beispiel 4.17 auch die QRP Zerlegung betrügt.

Eine weitere wichtige Zerlegung ist die Schurform.

Satz 4.19 *(reelle Version) Sei* $A \in \mathbb{R}^{n,n}$ *so gibt es eine orthogonale Matrix* Q, *so daß*

$$Q^T A Q = R = \begin{bmatrix} R_{11} & \cdots & R_{1k} \\ & \ddots & \vdots \\ & & R_{kk} \end{bmatrix} \tag{4.20}$$

quasi obere $\triangle$*–Matrix ist und die Diagonalblöcke können in jeder beliebigen Reihenfolge auftreten.* R_{ii} *sind* 1×1 *oder* 2×2, *mit reellen oder komplexen Eigenwerten.*

Beweis: Wie SVD–Beweis per Induktion.
Sei R_{11} reeller Eigenwert von A mit Eigenvektor $x, ||x||_2 = 1$. Sei $Q = [x, Q_1]$ orthogonale Basis des $\mathbb{R}^n$, so gilt

$$Q^T A Q = \left[\begin{array}{c|c} R_{11} & w^T \\ \hline 0 & A_1 \end{array}\right]$$

Für ein Paar von Eigenwerten $a \pm ib$, $b \neq 0$ gibt es eine reelle 2×2 Matrix R_{11} und reelle Orthonormalbasis des Eigenraums zu den beiden Eigenwerten, so daß $A[x_1, x_2] = [x_1, x_2]R_{11}$. Sei dann $Q = [x_1, x_2, Q_1]$ orthogonale Basis des $\mathbb{R}^n$, so gilt ebenfalls

$$Q^T A Q = \left[\begin{array}{c|c} R_{11} & w^T \\ \hline 0 & A_1 \end{array}\right]$$

Rest per Induktion. $\square$

Wir brauchen noch eine weitere Zerlegung, die gleichzeitig 2 Matrizen zerlegt, dies ist die sogenannte verallgemeinerte Singulärwertzerlegung.

Satz 4.21 *Verallgemeinerte Singulärwertzerlegung (GSVD). Seien $B \in \mathbb{R}^{n,m}, C \in \mathbb{R}^{p,n}$, so existieren Orthogonale Matrizen $P \in \mathbb{R}^{p,p}, Q \in \mathbb{R}^{m,m}$ und eine nichtsinguläre Matrix $S \in \mathbb{R}^{n,n}$, so daß*

$$SBQ = \begin{bmatrix} \Sigma_B & 0 \\ 0 & 0 \end{bmatrix}, \quad PCS^{-1} = \begin{bmatrix} \Sigma_C & 0 \\ 0 & 0 \end{bmatrix},$$

mit Σ_B, Σ_C nichtsingulär diagonal.

Den Beweis machen wir hier nicht, sondern verweisen auf das Buch von Golub/Van Loan. Beachte: Da S nicht orthogonal ist, so ist im allgemeinen die Berechnung der GSVD nicht stabil. Man kann da Abhilfe schaffen, aber das sind technische Details.

Wir wollen nun diese Zerlegungen verwenden, um auf numerisch rückwärts stabile Weise zu entscheiden, ob ein zeitinvariantes System steuerbar, rekonstruierbar, stabilisierbar oder entdeckbar ist. Dazu brauchen wir das folgende Lemma bzw. den Algorithmus, der den Beweis gibt.

Lemma 4.22 *Gegeben Matrizen $A \in R^{n,n}, B \in \mathbb{R}^{n,m}$ so existieren orthogonale Matrizen P, Q mit $P \in \mathbb{R}^{n,n}, Q \in \mathbb{R}^{m,m}$, so daß*

$$PAP^T = \begin{array}{c} \left[\begin{array}{cccc|c} A_{11} & \cdots & \cdots & A_{1s-1} & A_{1,s} \\ A_{21} & \ddots & & \vdots & \vdots \\ & \ddots & \ddots & \vdots & \vdots \\ & & A_{s-1,s-2} & A_{s-1,s-1} & A_{s-1,s} \\ \hline 0 & \cdots & 0 & 0 & A_{ss} \end{array}\right] \begin{array}{c} n_1 \\ n_2 \\ \vdots \\ n_{s-1} \\ n_s \end{array} \\ \begin{array}{ccccc} n_1 & \ldots & n_{s-2} & n_{s-1} & n_s \end{array} \end{array}, \quad PBQ = \begin{array}{c} \begin{bmatrix} B_1 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} \begin{array}{c} n_1 \\ n_2 \\ \vdots \\ \vdots \\ n_s \end{array} \\ \begin{array}{cc} n_1 & m - n_1 \end{array} \end{array} \tag{4.23}$$

wobei $n_1 \geq n_2 \geq \cdots \geq n_{s-1} \geq n_s \geq 0, n_{s-1} > 0$

$$A_{i,i-1} = \begin{array}{c} \begin{bmatrix} \Sigma_{i,i-1} & 0 \end{bmatrix} n_i \\ \begin{array}{cc} n_i & n_{i-1} - n_i \end{array} \end{array} \quad i = 1, \cdots, s-1$$

$\Sigma_{i,i-1}$	*quadratisch, nicht singulär, $\Sigma_{s-1,s-2}$ diagonal*
B_1	*quadratisch nicht singulär.*

Beweis: Wir geben einen konstruktiven Beweis durch den folgenden Algorithmus:

Algorithmus 4.24 *„Staircase“ Algorithmus*

Input: $A \in \mathbb{R}^{n,n}, B \in \mathbb{R}^{n,m}$,
Output: PAP^T, PBQ in der Form (4.23), P, Q orthogonal.
Schritt 0: Führe eine SVD von B durch.

$$B = U_B \begin{bmatrix} \Sigma_B & 0 \\ 0 & 0 \end{bmatrix} V_B^T$$

mit Σ_B nichtsingulär, $n_1 \times n_1$.
Setze $P := U_B^T, Q := V_B$.

$$A := U_B^T A U_B = \begin{array}{cc} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} & n_1 \\ n_1 & \end{array} \qquad B := U_B^T B V_B = \begin{array}{cc} \begin{bmatrix} \Sigma_B & 0 \\ 0 & 0 \end{bmatrix} & n_1 \\ n_1 & \end{array}$$

Schritt 1
Führe SVD von A_{21} durch:

$$A_{21} = U_{21} \begin{bmatrix} \Sigma_{21} & 0 \\ 0 & 0 \end{bmatrix} V_{21}^T, \ \Sigma_{21} \ n_2 \times n_2 \text{ nichtsingulär, diagonal}$$

Setze

$$P_2 := \begin{bmatrix} V_{21}^T & 0 \\ 0 & U_{21}^T \end{bmatrix}, P := P_2 P$$

$$A := P_2 A P_2^T =: \begin{array}{cccc} \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ 0 & A_{32} & A_{33} \end{bmatrix} & \begin{matrix} n_1 \\ n_2 \\ \end{matrix}, \\ n_1 \quad n_2 & \end{array} \qquad \text{wobei } A_{21} = \begin{array}{c} [\Sigma_{21} \quad 0\ ] n_2 \\ n_2 \quad n_1 - n_2 \end{array},$$

$$B := P_2 B =: \begin{bmatrix} B_1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{matrix} n_1 \\ n_2 \\ \end{matrix},$$

wobei $B_1 := V_{21}^T \Sigma_B \quad n_1 \times n_1$ nichtsingulär.

Schritt 2
i = 3
DO WHILE ($n_{i-1} > 0 \quad$ OR $\quad A_{i,i-1} \neq 0$).
Führe SVD von $A_{i,i-1}$ durch:
$A_{i,i-1} = U_{i,i-1} \begin{bmatrix} \Sigma_{i,i-1} & 0 \\ 0 & 0 \end{bmatrix} V_{i,i-1}^T$ mit
$\Sigma_{i,i-1} \ n_i \times n_i$ nichtsingulär, diagonal.
Setze

$$P_i := \begin{bmatrix} I_{n_1} & & & & \\ & \ddots & & & \\ & & I_{n_{i-2}} & & \\ & & & V_{i,i-1}^T & \\ & & & & U_{i,i-1}^T \end{bmatrix}, P := P_i P$$

$$A := P_i A P_i^T =: \begin{array}{cc} \begin{bmatrix} A_{11} & & \cdots & & A_{1,i+1} \\ A_{21} & \ddots & & & A_{2,i+1} \\ & \ddots & \ddots & & \vdots \\ & & A_{i,i-1} & A_{i,i} & \vdots \\ 0 & & & A_{i+1,i} & A_{i+1,i+1} \end{bmatrix} & \begin{matrix} n_1 \\ n_2 \\ \vdots \\ n_i \\ \end{matrix} \\ n_1 \qquad\quad n_{i-1} \quad n_i \qquad\qquad & \end{array}$$

wobei $A_{i,i-1} = [\Sigma_{i,i-1}, 0]$
$i := i + 1$
END
$s := i$
Es ist klar, daß der Algorithmus entweder mit $n_i = 0$ oder $A_{i,i-1} = 0$ abbricht, denn solange eine der beiden Bedingungen gilt, kann man den nächsten Schritt ausführen. Andererseits wird in jedem Schritt der Restblock mindestens um 1 kleiner, solange Rang$A_{i,i-1} > 1$ ist, so daß der Algorithmus nach maximal $n-1$ Schritten abbricht. □

Bemerkung 4.25 *Algorithmus (4.24) in etwas veränderter Form geht auf P. van Dooren zurück und heißt Staircase oder Treppen–Algorithmus. Es gibt viele Varianten dieses Algorithmus, z. B. mit QRP–Zerlegungen oder anderen Ordnungen. Er läßt sich numerisch stabil implementieren, kann jedoch $\mathcal{O}(n^4)$ flops brauchen. Kritischer Punkt ist natürlich in jedem Schritt die Rangentscheidung in $A_{i,i-1}$, bzw. B. Die Form (4.23) ist sehr eng verwandt mit der Kronecker Normalform für das Büschel (pencil) $\lambda[I\ \ 0] - [A\ \ B]$.*
Dies ist eine Verallgemeinerung der Jordan–Normalform für Matrizenpaare.

Als direkte Konsequenz aus Lemma 4.22 erhalten die folgende numerisch stabil nachprüfbaren Kriterien für Steuerbarkeit, Rekonstruierbarkeit, Stabilisierbarkeit und Entdeckbarkeit.

Satz 4.26 *Gegeben sei ein lineares zeitinvariantes System der Form*

$$\dot{x} = Ax + Bu. \quad y = Cx \tag{4.27}$$

i) Das System (4.27) ist steuerbar genau dann, wenn in der Treppenform von (A, B) $n_s = 0$ ist.

ii) Das System (4.27) ist rekonstruierbar genau dann, wenn in der Treppenform von (A^T, C^T) $n_s = 0$ ist.

iii) Das System (4.27) ist stabilisierbar genau dann, wenn in der Treppenform von (A, B) alle Eigenwerte von A_{ss} negativen Realteil haben.

iv) Das System (4.27) ist entdeckbar genau dann, wenn in der Treppenform von (A^T, C^T) alle Eigenwerte von A_{ss} negativen Realteil haben.

Beweis:

i) Wir können o.B.d.A. wegen Satz 4.12 und Lemma 4.22 annehmen, daß (A, B) in Treppenform vorliegt. Betrachte

$$[B, \lambda I - A] = \begin{array}{c} \left[\begin{array}{cc|ccccc} B_1 & 0 & \lambda I - A_{11} & -A_{12} & \cdots & \cdots & -A_{1,s} \\ 0 & 0 & -A_{21} & \lambda I - A_{22} & -A_{23} & & \vdots \\ & & & \ddots & \ddots & \ddots & -A_{s-2,s} \\ & & & & -A_{s-1,s-2} & \lambda I - A_{s-1,s-1} & -A_{s-1,s} \\ & & & & & 0 & \lambda I - A_{s,s} \end{array}\right] \end{array} \begin{array}{c} n_1 \\ n_2 \\ \vdots \\ \vdots \\ n_s \end{array}$$

Da alle Matrizen $B_1, A_{21}, \cdots, A_{s-1,s-2}$ vollen Rang haben, so gilt

$$\begin{aligned} \text{rang } [\lambda I - A, B] &= n \text{ für alle } \lambda \in \mathbb{C} \text{ genau dann wenn} \\ \text{rang } [\lambda I - A_{ss}] &= n \text{ für alle } \lambda \in \mathbb{C} \text{ genau dann wenn} \\ n_s &= 0. \end{aligned}$$

ii) Mit Dualität und i).

iii) Rang $[\lambda I - A, B] = n \quad \forall \lambda \in \mathbb{C}, Re(\lambda) \geq 0$ genau dann, wenn

$$\text{Rang } [\lambda I - A_{ss}] = n \quad \forall \lambda \in \mathbb{C}, Re(\lambda) \geq 0$$

genau dann, wenn alle Eigenwerte von A_{ss} negativen Realteil haben.

iv) mit Dualität und iii).

□

Bemerkung 4.28 *In vielen Lehrbüchern wird eine Aufspaltung des Raumes in steuerbare und nicht steuerbare Unterräume gemacht. Diese Aufspaltung erhält man natürlich sofort aus der Treppenform. Wir verwenden das hier nicht weiter, siehe [Knobloch/Kwakernaak].*

Die Treppenform von (A, B) bzw. (A^T, C^T) ist noch nicht ganz befriedigend, weil man nicht gleichzeitig Steuerbarkeit und Stabilisierbarkeit ablesen kann.

Man sollte hier noch kurz den Spezialfall erwähnen, daß $m = 1$ oder $p = 1$ ist. In diesem Fall erhalten wir sofort aus Lemma 4.22, daß $n_1 = n_2 = \cdots = n_{s-1} = 1$ und damit ist die Staircaseform einfach die sogenannte <u>System Hessenberg–Form</u>.

$$PAP^T = \begin{bmatrix} a_{11} & & \cdots & a_{1,n-1} & a_{1,n} \\ a_{21} & \ddots & & & \vdots \\ & \ddots & \ddots & & \vdots \\ & & a_{n-1,n-2} & a_{n-1,n-1} & a_{n-1,n} \\ 0 & & & a_{n,n-1} & a_{n,n} \end{bmatrix} \quad Pb = \begin{bmatrix} b_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4.29}$$

(Dabei ist noch A_{ss} auf Hessenbergform gebracht worden.) Wir erhalten als Korollar aus Satz 4.26, daß das System <u>steuerbar</u> ist genau dann, wenn die Hessenbergmatrix unreduziert ist $(a_{i,i-1} \neq 0 \quad \forall_i = 2, \ldots, n)$ und $b_1 \neq 0$.

Kapitel 5

Polvorgabe und QR–Algorithmus

Wir hatten schon im 3. Kapitel gesehen, daß es möglich ist, für gegebenes Paar $A, B, \quad (A, B)$ steuerbar, F so zu wählen, daß $A + BF$ beliebiges vorgegebenes Spektrum hat. Es gibt tausende von Algorithmen für dieses Problem, die sich im wesentlichen durch ihre Stabilitätseigenschaften unterscheiden. Wir werden hier nur zwei dieser Algorithmen betrachten, einen für den Fall $m = 1$. Falls $m > 1$ ist, so müssen wir zusätzliche Bedingungen stellen, damit das Problem eindeutig lösbar wird. Dies kann man auf viele verschiedene Arten tun, eine Möglichkeit ist, die Wahl von F so zu gestalten, daß die Eigenwerte robust gegen Störungen sind. Aber zuerst zum Fall $m = 1$. Um diesen Fall und auch den anderen studieren zu können, müssen wir uns erst einmal intensiv mit dem QR–Algorithmus zur Lösung des Eigenwertproblems beschäftigen. Dazu brauchen wir zuerst das folgende Lemma:

Lemma 5.1 *Sei $A \in \mathbb{R}^{n,n}$, dann gibt es eine orthogonale Matrix Q, so daß*

$$H = Q^T A Q = \begin{bmatrix} * & \cdots & \cdots & * \\ * & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & * & * \end{bmatrix}$$

obere Hessenbergmatrix ist.

Beweis: Siehe Golub/Van Loan □

Eine obere Hessenberg–Matrix, bei der alle Elemente auf der unteren Nebendiagonalen $\neq 0$ sind, heißt <u>unreduziert</u>. Die Transformation auf Hessenbergform ist nicht eindeutig. Wir können im Prinzip die erste Spalte der Transformationsmatrix direkt vorgeben:

Satz 5.2 *Seien $Q = [q_1, \ldots, q_n]$ und $V = [v_1, \ldots, v_n]$ orthogonale Matrizen, so daß*

$$Q^T A Q = H \text{ und } V^T A V = G$$

beide obere Hessenbergmatrizen in $\mathbb{R}^{n,n}$ sind. Sei k der kleinste Index für den $h_{k+1,k} = 0$ ist ($k = n$ falls H unreduziert ist). Falls $v_1 = q_1$, so gilt $v_i = \pm q_i$ und $|h_{i,i-1}| = |g_{i,i-1}|$ für $i = 2 : k$. Falls $k < n$, so ist $g_{k+1,k} = 0$.

Beweis: Sei $W = [w_1, \ldots, w_n] := V^T Q$, so gilt $GW = WH$, denn $VGV^T = QHQ^T$. Also gilt für $i = 2 : k$

$$h_{i,i-1} w_i = G w_{i-1} - \sum_{j=1}^{i-1} h_{j,i-1} w_j.$$

Es ist $v_1 = q_1$. Somit gilt $w_{11} = v_1^T v_1 = 1$. Da W aber wieder eine orthogonale Matrix ist, so muß bereits $w_1 = e_1$ gelten. Hieraus folgt (mit Induktion), daß $[w_1, \dots, w_k]$ obere $\triangle$–Matrix ist und da W orthogonal, so gilt $w_i = \pm e_i$ für $i = 2 : k$. Da $w_i = V^T q_i$ und $h_{i,i-1} = w_i^T G w_{i-1}$, so folgt $v_i = \pm q_i$ und $\mid h_{i,i-1} \mid = \mid g_{i,i-1} \mid$ für $i = 2 : k$. Falls $h_{k+1,k} = 0$, so folgt

$$\begin{aligned} |g_{k+1,k}| = \left| e_{k+1}^T G e_k \right| &= \left| e_{k+1}^T G W e_k \right| \\ = \left| (e_{k+1}^T W)(H e_k) \right| &= \left| e_{k+1}^T \sum_{i=1}^{k} h_{ik} W e_i \right| \\ = \left| \sum_{i=1}^{k} h_{ik} e_{k+1}^T e_i \right| &= 0 \end{aligned}$$

□

Also folgt für unreduzierte Hessenbergmatrizen, daß die Vorgabe der 1. Spalte die Hessenbergzerlegung im wesentlichen eindeutig macht. Diese Freiheit können wir nun nutzen, um ein Matrix–Vektor–Paar $[A, b]$ auf System–Hessenberg–Form zu bringen. Die Existenz dieser Form folgt aus Lemma 4.22. Wir geben nun einen Algorithmus an:
Als erstes geben wir eine Routine an, die für den Algorithmus nachher benötigt wird. Es handelt sich dabei um eine Spiegelung, mit welcher ein vorhandener Vektor auf ein Vielfaches des ersten Einheitsvektors abgebildet wird. Eine solche Transformationsmatrix heißt Householder–Matrix.

Algorithmus 5.3 *Berechnung einer Householder–Matrix P, so daß für gegebenes $x \in \mathbb{R}^n$*

$$Px = \begin{bmatrix} * \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \textit{wobei } P = I - 2vv^T / v^T v$$

$$\textit{und } v(1) = 1$$

Berechne v:
function $v = house(x)$
  $n = length\ (x)$; $\mu = \|x\|_2$; $v = x$;
  if $\mu \neq 0$
    $\beta = x(1) +\ sign\ (x(1)) * \mu$
    $v(2:n) = v(2:n)/\beta$.
  end
  $v(1) = 1$
end house.

<u>*Kosten:*</u> *$3n$ flops.*
<u>*Fehler:*</u> $\|\tilde{v} - v\|_2 = \mathcal{O}(eps)$

Algorithmus 5.4 *Multiplikation mit Householder–Matrix von links. Gegeben $n \times m$ Matrix A und n–Vektor $v \neq 0$ mit $v(1) = 1$, so überschreibt diese Methode A mit PA wobei $P = I - 2vv^T / v^T v$.*
function $A = row.house(A, v)$
  $\beta = -2/v^T * v$
  $w = \beta * A^T * v$

$A = A + v * w^T$
end row.house

Kosten: $4mn$ *flops.*
Fehler: $fl(\tilde{P}A) = P(A+E)$, $\|E\|_2 \le \mathcal{O}(eps\,\|A\|_2)$

Algorithmus 5.5 *Multiplikation mit Householder–Matrix von rechts. Gegeben* $n \times m$ *Matrix* A *und* m*–Vektor* v *mit* $v(1) = 1$*, so überschreibt dieser Algorithmus* A *mit* AP *wobei* $P = I - 2vv^T/v^Tv$.

function $A = col.house(A, v)$
$\beta = -2/v^T * v$
$w = \beta * A * v$
$A = A + w * v^T$
end col.house

Kosten: $4nm$ *flops.*
Fehler: $fl(A\tilde{P}) = (A+E)P$, $\|E\|_2 = \mathcal{O}(eps\|A\|_2)$.

Nun zum eigentlichen Algorithmus:

Algorithmus 5.6 *(System–Hessenberg–Reduktion)*

Input: $A \in \mathbb{R}^{n,n}, b \in \mathbb{R}^n$
Output: *Orthogonale Matrix* $Q \in \mathbb{R}^{n,n}$ *und System* $\tilde{A}, \tilde{b}$ *in System–Hessenberg–Form,* $Q^TAQ =: \tilde{A}$ *obere Hessenbergmatrix*

$$Q^Tb =: \tilde{b} = \begin{bmatrix} * \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Schritt 0 *Bestimme* Q_0 *orthogonal so, daß* $Q_0^Tb = \begin{bmatrix} * \\ 0 \\ \vdots \\ 0 \end{bmatrix}$.

$v = house\ (b)$.
$b = row.house\ (b, v)$
Anwendung auf A:
$A = row.house\ (A, v)$
$A = col.house\ (A, v)$

Schritt 1 *Transformiere A mit dem nachfolgenden Algorithmus 5.7 auf Hessenbergform, wobei die erste Spalte der Transformationsmatrix der erste Einheitsvektor* e_1 *ist.*

Algorithmus 5.7 *Householder–Hessenberg–Reduktion*

Input: $A \in \mathbb{R}^{n,n}$

Output: $Q \in \mathbb{R}^{n,n}$ *Produkt von Householdermatrizen* $P_1, \ldots, P_{n-1}$*, so daß* Q^TAQ *obere Hessenbergmatrix ist.*

$FOR \quad k = 1 : n-2$

$$
\begin{array}{lcl}
V(k+1:n) & = & house\ (A(k+1:n,k)) \\
A(k+1:n,k:n) & = & row.house\ (A(k+1:n,k:n), V(k+1:n)) \\
A(1:n,k+1:n) & = & col.house\ (A(1:n,k+1:n), V(k+1:n)) \\
A(k+2:n,k) & = & V(k+2,n)
\end{array}
$$

END

Dabei sind house, row.house, col.house die subroutines, die die Householdermatrix erzeugen und von links bzw. rechts anwenden (siehe oben).

Kosten: $\frac{10}{3}n^3$ *flops*

Fehler: *Berechnete* $\tilde{A}, \tilde{b}$ *erfüllen:*

$$\tilde{A} = Q^T(A+E)Q, \quad \tilde{b} = Q^T(b+e) \quad mit \quad Q^TQ = I$$

und

$$\|E\|_F \le cn^2\, eps\, \|A\|_F, \quad \|e\|_2 \approx eps\, \|b\|_2$$

Die Idee des Polvorgabe–Algorithmus von Miminis und Paige ist nun den QR–Algorithmus auf A, b in System–Hessenbergform anzuwenden, aber sozusagen rückwärts. Dazu müssen wir uns zuerst noch mal den QR–Algorithmus für eine unreduzierte Hessenberg–Matrix anschauen. Es gibt 2 wesentliche Varianten, man kann explizit und implizit arbeiten:

Algorithmus 5.8 *Impliziter Francis QR–Schritt*

Input: *Unreduzierte obere Hessenberg–Matrix* $A \in \mathbb{R}^{n,n}$, *deren untere* 2×2 *Ecke die Eigenwerte* a_1, a_2 *hat.*

Output: *Orthogonale Matrix* Z *(Produkt von Householder Transformationen* $P_1, \ldots, P_{n-2}$ *so, daß* $Z^T(A - a_1I)(A - a_2I)$ *obere* $\triangle$*–Matrix ist) und* A *wird überschrieben mit* Z^TAZ.
$l = n-1$

% { *Berechne 1. Spalte von* $(A - a_1I)(A - a_2I)$ }
$s = A(l,l) + A(n,n)$
$t = A(l,l)A(n,n) - A(l,n)A(n,l)$
$x = A(1,1) * A(1,1) + A(1,2) * A(2,1) - sA(1,1) + t$
$y = A(2,1) * (A(1,1) + A(2,2) - s)$
$z = A(2,1) * A(3,2)$
For $k = 0 : n-3$

% *Überschreibe* A *mit* $P_kAP_k^T$, *wobei* $P_k = \begin{bmatrix} I_k & & \\ & \tilde{P}_k & \\ & & I_{n-k-3} \end{bmatrix}$.

$v = house \left(\begin{bmatrix} x \\ y \\ z \end{bmatrix} \right)$

$A(k+1:k+3, k+1:n) = row.house\ (A(k+1:k+3, k+1:n), v)$

$r = \min\{k+4, n\}$

$A(1:r, k+1:k+3) = col.house\ (A(1:r, k+1:k+3), v)$

$x = A(k+2, k+1)$

$y = A(k+3, k+1)$
if $k < n-3$
$z = A(k+4, k+1)$
end
end
% $A = P_{n-2} A P_{n-2}^T$, *wobei* $P_{n-2} = \begin{pmatrix} I_{n-2} & \\ & \tilde{P}_{n-2} \end{pmatrix}$
$A(n-1:n, n-2:n) =$*row.house* $(A(n-1:n, n-2:n), v)$
$A(1:n, n-1:n) =$ *col.house* $(A(1:n, n-1:n), v)$

<u>Kosten:</u> $10n^2$ *flops ohne* Z *mit Akkumulation von* Z $10n^2$ *weitere flops.*

Es gibt auch eine explizite Variante, die allerdings sehr viel aufwendiger ist. Dort wird explizit eine QR–Zerlegung von $(A - a_1 I)(A - a_2 I) = QR$ durchgeführt und dann $Q^T AQ$ gebildet. Dazu muß natürlich $(A - a_1 I)(A - a_2 I)$ gebildet und dann QR zerlegt werden. Das gibt $O(n^3)$ flops. Also wäre der Algorithmus insgesamt $O(n^4)$ und das ist zu teuer.

Der gesamt QR–Algorithmus sieht dann wie folgt aus:

Algorithmus 5.9 <u>*QR–Algorithmus für Hessenbergmatrizen*</u>

<u>Input:</u> $A \in \mathbb{R}^{n,n}$ *in oberer Hessenbergform,* $Q_0 \in \mathbb{R}^{n,n}$ *orthogonal,* $tol > eps$.

<u>Qutput:</u> *Reelle Schurform von* $Q^T AQ = T$, *falls gewünscht auch* Q.

Wenn T *und* Q *gesucht sind, so wird* T *in* A *gespeichert. Wenn nur die Eigenwerte gesucht sind, so werden die entsprechenden Blöcke von* T *in* A *gespeichert.*

Do until $q = n$
Setze alle Subdiagonalelemente von A *welche*

$$|a_{i,i-1}| \leq \; tol \; (|a_{ii}| + |a_{i-1,i-1}|) \tag{5.10}$$

erfüllen zu 0.

% *Deflation*
Bestimme größtes $q > 0$ *und kleinstes* $p > 0$ *so, daß*

$$A = \begin{array}{cl} \left[\begin{array}{ccc} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{array} \right] & \begin{array}{l} p \\ n-p-q \\ q \end{array} \\ \begin{array}{ccc} p & n-p-q & q \end{array} & \end{array}$$

wobei A_{33} *quasi obere* $\triangle$*–Matrix,* A_{22} *unreduzierte Hessenbergmatrix ist.*
If $q < n$
Wende Algorithmus 5.8 auf A_{22} *an:*
$A_{22} := Z^T A_{22} Z$
$A_{12} := A_{12} Z$
$A_{23} := Z^T A_{23}$
Falls Q *berechnet werden soll, so setze*

$$Q \; := Q \cdot \begin{bmatrix} I_p & & \\ & Z & \\ & & I_q \end{bmatrix}$$

end
end
end

Bringe alle 2×2 Blöcke mit reellen Eigenwerten auf $\triangle$–form mit einer Jacobi–Rotation.

<u>Kosten:</u> Exklusive Hessenbergreduktion $\approx 22n^3$ falls Q und T berechnet werden. Falls nur Eigenwerte benötigt werden $\approx \frac{20}{3}n^3$.
$\approx$ 2.4 Iterationen pro Eigenwert, Erfahrung.

<u>Fehler:</u> Die berechnete Schurform ist die exakte Schurform von $A + E$

$$Q^T(A+E)Q = \tilde{T}$$

mit $Q^TQ = I$ und $\|E\|_2 \approx eps\ \|A\|_2$. Das berechnete $\tilde{Q}$ erfüllt $\tilde{Q}^T\tilde{Q} = I + F$ mit $\|F\|_2 \approx eps$.

Die Ordnung der Eigenwerte ist beliebig, wir werden sehen, daß wir die Eigenwerte umordnen können und werden dies auch brauchen.

Was ist nun die Idee beim Polvorgabe–Algorithmus?

- Gegeben Menge Ω von n Eigenwerten (reell abgeschlossen) d.h. mit λ ist auch $\bar{\lambda}$ drin.
- Bringe (A, b) auf (P^TAP, P^Tb) System Hessenbergform.
- Bestimme $\tilde{f} \in \mathbb{R}^n$ so daß $\sigma(P^TAP - P^Tb\tilde{f}^T) = \Omega$.
- Bestimme $f = P\tilde{f}$.

Das schwierigste ist die Wahl von $\tilde{f}$. Die Idee des Algorithmus ist nun die folgende.

Sei A unreduzierte obere Hessenbergmatrix und λ ein Eigenwert von A, so ist in der RQ Zerlegung von

$$(A - \lambda I)Q^T = R$$

die erste Spalte von R gleich 0.

$$RQ = \begin{bmatrix} x & \cdots & \cdots & x \\ x & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & x & x \end{bmatrix}$$

Man eleminiere die Nebendiagonale von unten nach oben. Da $Rang(A - \lambda I) \geq n - 1$ (wegen Unreduziertheit) so folgt, daß die erste Spalte 0 ist. Der Trick ist nun das f sukzessive so zu wählen, daß für vorgegebenes λ_1 in

$$[A - \lambda_1 I - \beta e_1 f^T]Q^T = R$$

das obere Eckelement in R verschwindet. Dann geht man zu einem kleineren Problem über.

Setze $Q = \begin{bmatrix} y^T \\ \tilde{Q} \end{bmatrix}$, so folgt

$$[A - \lambda_1 I - \beta e_1 f^T]y = 0$$

$$\Longrightarrow e_1^T(A-\lambda_1 I)y = \beta f^T y$$

Daraus können wir noch nicht das ganze f berechnen. Also führen wir den RQ Schritt zu Ende und setzen

$$Q(A-\beta e_1 f^T)Q^T = QR + \lambda_1 I = \left[\begin{array}{c|c} \lambda_1 & * \\ \hline 0 & \tilde{Q}(A-\beta e_1 f^T)\tilde{Q}^T \end{array}\right]$$

Setze nun $\tilde{A} = \tilde{Q}A\tilde{Q}^T, \tilde{\beta} = e_2^T \beta \tilde{Q} e_1 = q_{21}\beta$ und $\tilde{f} = \tilde{Q}f$, so sind alle Dimensionen um 1 kleiner geworden und weil $q_{21} \neq 0$, so folgt $\tilde{\beta} \neq 0$ und wir können mit dieser reduzierten Form weitermachen. Insgesamt erhalten wir den folgenden Algorithmus:

Algorithmus 5.11 *Polvorgabe nach Miminis/Paige für einen Input, explizite reelle Version.*

Input: $A \in \mathbb{R}^{n,n}, b = \beta e_1 \in \mathbb{R}^n, (A,b)$ *in System–Hessenbergform, A unreduziert.*

$$\Omega = \{\lambda_1, \cdots, \lambda_n\} \subseteq \mathbb{R} \qquad \textit{Eigenwertmenge}$$

Output: $f \in \mathbb{R}^n$ *so, daß* $\sigma(A - bf^T) = \Omega$.

Schritt 1
Setze $A_1 = A, \beta_1 = \beta$
FOR $i = 1 : n-1$
    Berechne RQ–Zerlegung

$$(A_i - \lambda_i I)Q_i^T = R_i = \left[r_{lj}^{(i)}\right], Q_i = \left[q_{lj}^{(i)}\right]$$

    Berechne $\tau_i = \frac{r_{11}^{(i)}}{\beta_i}, \beta_{i+1} = q_{21}^{(i)}\beta_i$.
    Berechne für $Q_i = \begin{bmatrix} y_i^T \\ \tilde{Q}_i \end{bmatrix}$
    $A_{i+1} = \tilde{Q}_i A_i \tilde{Q}_i^T$
END

Schritt 2
Berechne $k_n = \frac{(A_n - \lambda_n)}{\beta_n}$.
FOR $i = n-1 : -1 : 1$
    Berechne $k_i = Q_i^T \begin{bmatrix} \tau_i \\ k_{i+1} \end{bmatrix}$
END
$f = k_1$

Man kann auch eine Variante dieses Algorithmus konstruieren, die auf impliziten Doppelshiftschritten aufgebaut ist, kommt später.

Beispiel 5.12 *Betrachte das folgende Problem*

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

Schritt 1
$i = 1$:

$$A_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \beta_1 = 1$$

Berechne RQ–Zerlegung von $A_1 - \lambda_1 I$

$$(A_1 - \lambda_1 I) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad Q_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

$$R_1 = (A_1 - \lambda_1 I) Q_1^T = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & +1 \\ -1 & 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 2 \\ 0 & 2 \end{bmatrix}$$

Setze $\tau_1 = 0, \quad \beta_2 = \frac{1}{\sqrt{2}}$
Transformiere A_1

$$A_2 = \frac{1}{2}[1 \ \ 1] \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 1$$

Schritt 2

$$k_2 = \frac{(A_2 - \lambda_2)}{\beta_2} = 3\sqrt{2}$$

$$k_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 3\sqrt{2} \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix} = f$$

Ergebnis:

$$A - bf^T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix} [3 \ \ 3] = \begin{bmatrix} -3 & -2 \\ 1 & 0 \end{bmatrix}$$

hat als Eigenwerte die Nullstellen von

$$\lambda^2 + 3\lambda + 2 = (\lambda + 2)(\lambda + 1).$$

Es gibt eine allerdings noch etwas technischere Version dieses Algorithmus, welche auch komplexe Shifts in reeller Arithmetik verarbeitet. Das machen wir hier nicht.

Der nächste Punkte ist nun der Fall, das $m \geq 1$ ist, hier möchte man die Freiheiten, die man hat, ausnutzen, um weitere Eigenschaften zu erhalten.

Dazu muß man sich nun die Freiheiten einmal anschauen.

Satz 5.13 *Sei $A \in \mathbb{R}^{n,n}, B \in \mathbb{R}^{n,m}$ und $\Omega = \{\lambda_1, \cdots, \lambda_n\} \subset \mathbb{C}$ abgeschlossen unter Konjugation. Falls (A, B) steuerbar ist, so existiert $F \in \mathbb{R}^{m,n}$, so daß $\sigma(A - BF) = \Omega$. Unter diesen Voraussetzungen ist F eindeutig genau dann, wenn $m = 1$.*

Beweis: Der erste Teil folgt sofort aus Satz 2.29. Der zweite Teil folgt aus der Tatsache, daß jedes Matrix–Vektorpaar (A, b) in folgende Form gebracht werden kann (allerdings instabil). Dies geschieht durch sukzessive Elimination der Elemente oberhalb der unteren Nebendiagonalen in der System–Hessenbergform.

$$A = \begin{bmatrix} a_{11} & \cdots & \cdots & a_{1n} \\ a_{21} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & a_{n,n-1} & a_{nn} \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

(1) Sei $T_0 = \begin{bmatrix} b_1^{-1} & & & & \\ & (a_{21}b_1)^{-1} & & & \\ & & (a_{32}a_{21}b_1)^{-1} & & \\ & & & \ddots & \\ & & & & (a_{n,n-1}\cdots a_{21}b_1)^{-1} \end{bmatrix}$

so ist

$$\left[T_0 A T_0^{-1}, T_0 b\right] = [\begin{bmatrix} \tilde{a}_{11} & \cdots & \cdots & \tilde{a}_{1n} \\ 1 & \ddots & \ddots & \vdots \\ & \ddots & \ddots & \vdots \\ & & 1 & \tilde{a}_{nn} \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}]$$

Wir bezeichnen die transformierte Matrix jeweils wieder mit A, b.
FOR $i = n : -1 : 2$

$$T_i = \begin{array}{cc} \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & 1 & a_{i,i} & \cdots & a_{in} \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} & i \\ \qquad\qquad i-1 \quad i & \end{array}$$

$$A := [a_{ij}] = T_i A T_i^{-1}, b := T_i b = b$$

end
Dann ist (A, b) in der Form

$$\begin{bmatrix} a_1 & \cdots & \cdots & a_n \\ 1 & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ & & 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{5.14}$$

und es gilt

$$\begin{aligned} \det(A - \lambda I) &= (-1)^n(\lambda^n - a_1\lambda^{n-1} - \cdots - a_n) \\ \det(A - bf^T - \lambda I) &= (-1)^n(\lambda^n - (a_1 - f_1)\lambda^{n-1} - \cdots - (a_n - f_n)) \end{aligned}$$

Man kann sich also $\lambda_1, \ldots, \lambda_n$ vorgeben, und bildet

$$p(\lambda) = (\lambda - \lambda_1)\cdots(\lambda - \lambda_n) = \lambda^n + b_1\lambda^{n-1} + \cdots + b_n$$

und setzt $f_i = a_i + b_i \quad i = 1, \ldots, n$. Es ist klar, daß für $m > 1$ die Lösung nicht eindeutig sein kann. □

Für den Fall $m > 1$ müssen wir noch weitere Bedingungen an B stellen, um die Aufgabe eindeutig lösbar zu machen. Erste Bedingung

$$\text{Rang } (B) = m \tag{5.15}$$

Falls dies nicht gilt, so erhalten wir aus der Staircase–form B in der Form $\begin{bmatrix} B_1 & 0 \\ 0 & 0 \end{bmatrix}$, und können n entsprechend aufteilen $n = \begin{bmatrix} n_1 \\ n_2 \end{bmatrix}$, und im folgenden einfach n_2 und die letzte Blockspalte von B weglassen und voraussetzen, daß B vollen Rang hat.

Wir wandeln nur die Aufgabe ab:

Aufgabe 5.16 *Gegeben $A \in \mathbb{R}^{n,n}, B \in \mathbb{R}^{n,m}$ und Menge $\Omega = \{\lambda_1, \cdots, \lambda_n\} \subset \mathbb{C}$ abgeschlossen unter Konjugation, Rang $B = m$.*
Bestimme Matrix $F \in \mathbb{R}^{m,n}$ und nichtsinguläre Matrix X so, daß

$$(A - BF)X = X\Lambda \tag{5.17}$$

$$\textit{wobei } \Lambda = \textit{ diag} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$$

Dabei wähle X so, daß irgendein „Maß für Robustheit“optimiert wird. (Kautsky/Nichols/van Dooren 1985)
Das Maß, welches wir wählen, ist die Kondition des Eigenwertproblems mit $A - BF$.

Hier machen wir keine Forderungen mehr an die Steuerbarkeit von (A, B). Zwar gibt es dann Eigenwerte, die nicht gesetzt werden können, aber eventuell lassen sich trotzdem die Eigenvektoren X so wählen, daß die Kondition verbessert wird.

Satz 5.18 *Gegeben $A \in R^{n,n}, B \in \mathbb{R}^{n,m}$ Rang $(B) = m$ und $\Omega = \{\lambda_1, \cdots, \lambda_n\} \subset \mathbb{C}$ abgeschlossen unter Konjugation, sowie X nichtsingulär. Es existiert $F \in \mathbb{R}^{m,n}$ welches (5.17) löst genau dann, wenn*

$$U_1^T(AX - X\Lambda) = 0 \tag{5.19}$$

wobei $B = [U_0 \;\; U_1] \begin{bmatrix} Z \\ 0 \end{bmatrix}$ mit $[U_0 \;\; U_1]$ orthogonal und Z nichtsingulär. F ist dann gegeben durch

$$F = Z^{-1}U_0^T(X\Lambda X^{-1} - A). \tag{5.20}$$

Beweis: Da B vollen Rang hat, so existiert die Zerlegung von B (z. B. QR–Zerlegung). Aus (5.17) folgt

$$BF = A - X\Lambda X^{-1}. \tag{5.21}$$

Multiplikation mit $[U_0 \;\; U_1]^T$ von links gibt die beiden Gleichungen

$$\begin{aligned} ZF &= U_0^T(A - X\Lambda X^{-1}) \\ 0 &= U_1^T(A - X\Lambda X^{-1}), \end{aligned} \tag{5.22}$$

woraus die Behauptung folgt, da Z invertierbar ist. □

Es folgt, daß F existiert genau dann wenn

$$\text{Bild } (A - X\Lambda X^{-1}) \subset \text{ Bild } (B) = \text{ Bild } (U_0). \tag{5.23}$$

Korollar 5.24 *Der Eigenvektor $x_j \in \mathbb{C}^n$ von $A - BF$ zum vorgegebenen Eigenwert $\lambda_j \in \Omega$ muß in dem Raum*

$$\mathcal{S}_j = \textit{ Kern } \left(U_1^T(A - \lambda_j I)\right) \tag{5.25}$$

liegen. Die Dimension von $\mathcal{S}_j$ ist $m + k_j$ wobei

$$k_j = \textit{ dim } \left(\textit{ Kern } [B \mid A - \lambda_j I]^T \right) \tag{5.26}$$

Beweis: Wir haben aus (5.22) sofort

$$U_1^T(Ax_j - \lambda_j x_j) = 0 \quad \forall_j = 1, \cdots, n \tag{5.27}$$

also folgt die erste Behauptung. Weiter gilt

$$U^T [B \mid A - \lambda_j I] = \begin{bmatrix} Z & U_0^T(A - \lambda_j I) \\ 0 & U_1^T(A - \lambda_j I) \end{bmatrix}.$$

Aus (5.26) folgt

$$n - k_j = \text{ Rang } \left([B|A - \lambda_j I]^T\right)$$

und da Z invertierbar ist, folgt

$$n - m - k_j = \text{ Rang } \left[U_1^T(A - \lambda_j I)\right] \Longrightarrow \text{ Beh.}$$

□

Damit haben wir das Polvorgabe Problem verlagert auf die Wahl von Vektoren $x_j \in \mathcal{S}_j$ $\{j = 1, \ldots, n\}$, so daß das Eigenwertproblem (5.17) möglichst gut konditioniert ist.

Was für Bedingungen müssen nun notwendigerweise erfüllt sein? Aus Korollar 5.24 folgt:

Falls (A, B) steuerbar, so muß die Multiplizität von λ_j kleiner als m sein, denn die Maximalzahl unabhängiger Eigenvektoren von $A - BF$ zum Eigenwert λ_j ist $\dim(\mathcal{S}_j) = m$ da $k_j = 0$. Falls (A, B) nicht steuerbar, so können natürlich die Eigenwerte von A_{ss} in der Staircase form von (A, B) nicht verändert werden.

Sei $\Omega = \Omega_s \cup \Omega_u, \quad \mid \Omega_u \mid = n_s$ in (4.23) wobei Ω_u die nichtveränderbaren Eigenwerte angibt und Ω_s die steuerbaren.

Falls $\lambda_j \in \Omega_u$ so ist $k_j > 0$ und es gibt mindestens k_j linear unabhängige linke Eigenvektoren $y_l, l = 1, \ldots k_j$ für jede Wahl von F. Also müssen wir λ_j zumindest mit der Multiplizität k_j vorgeben um zu vermeiden, daß wir einen Jordan Block erhalten. (Dieser wäre schlecht konditioniert.) Diese Bedingungen reichen aber nicht aus.

Satz 5.28 *Eine notwendige Bedingung für die Existenz einer Lösung von Problem 5.16 bei der $A - BF$ keinen Jordan-Block hat ist, daß für jedes $\mu \in \mathbb{C}, s \in \mathbb{R}^n$ gilt:*

$$\left\{s^T B = 0, s^T(A - \mu I)B = 0, s^T(A - \mu I)^2 = 0\right\} \tag{5.29}$$

$$\Longrightarrow s^T(A - \mu I) = 0$$

Falls μ zur steuerbaren Menge gehört, so gilt dies immer.

Beweis: Falls $\mu \in \Omega_u$ mit zugehörigem $k = \dim \text{Kern } \left([B \mid A - \mu I]^T\right) > 0$, und es ein $s \neq 0$ gibt, so daß $s^T(A - \mu I)B = 0, s^T(A - \mu I)^2 = 0, s^T(A - \mu I) =: s_1 \neq 0$, so ist s_1^T ein Linkseigenvektor von $(A - BF)$ zum Eigenwert μ für alle F. Wenn auch noch gilt $s^T B = 0$ so gilt

$$s^T(A - BF - \mu I) = s^T(A - \mu I) \neq 0$$

aber

$$s^T(A - BF - \mu I)^2 = s^T\left[(A - \mu I)^2 - (A - \mu I)BF - BF(A - BF - \mu I)\right] = 0$$

Also ist s^T linker Hauptvektor von $A - BF$ und damit hat $A - BF$ einen Jordan Block. $\square$

Wir wollen aber auf jeden Fall X so wählen, daß es keine Jordan-Blöcke gibt, dies können wir erreichen, in dem wir die Konditonszahl von X minimieren.

$$cond_2(x) = \|X\|_2 \|X^{-1}\|_2. \tag{5.30}$$

Am besten wäre natürlich $cond_2(x) = 1$, d.h. X unitär, dann hätte man erreicht, daß $A - BF$ unitär diagonalisierbar wäre (normal).

Um zu betrachten, wie sich das System in Abhängigkeit von F bzw. X verhält, haben wir den folgenden Satz. Betrachte das „closed loop“ System

$$\dot{x}(t) = (A - BF)x(t), \quad x(0) = x^0 \tag{5.31}$$

Dann gilt der folgende Satz:

Satz 5.32 *Die Rückkopplungsmatrix F und die Lösung von (5.31) erfüllen*

$$\|F\|_2 \leq \left(\|A\|_2 + \max_j\{\mid \lambda_j \mid\} \cdot cond_2(X)\right) / \sigma_m(B) \tag{5.33}$$

wobei $\sigma_m(B)$ der kleinste Singulärwert von B ist.
Es gilt weiterhin

$$\|x(t)\|_2 \leq cond_2(X) \cdot \max_j\{\mid e^{\lambda_j t} \mid\} \cdot \|x^0\|_2 \tag{5.34}$$

Beweis: Aus (5.20) folgt

$$F = Z^{-1} U_0^T(-X\Lambda X^{-1} + A)$$

wobei $B = [U_0,\ U_1] \begin{bmatrix} \Sigma_1 V \\ 0 \end{bmatrix} = [U_0,\ U_1] \begin{bmatrix} Z \\ 0 \end{bmatrix}$ Singulärwertzerlegung von B ist.

$$\Longrightarrow \|F\|_2 \leq \|Z^{-1}\|_2 \|U_0^T\|_2 \left(\|X\|_2 \|X^{-1}\|_2 \|\Lambda\|_2 + \|A\|_2\right) \tag{5.35}$$

$$\|Z^{-1}\|_2 = \|V\|_2 \sigma_m^{-1}(B), \text{ aber } \|V\|_2 = 1 = \|U_0^T\|_2 \Longrightarrow (5.33).$$

Ungleichung (5.34) folgt aus

$$x(t) = e^{(A-BF)t} x^0 = X e^{\Lambda t} X^{-1} x^0. \tag{5.36}$$

$\square$

Aus diesem Satz folgt, daß eine Minimierung von $cond_2(X)$ auch obere Schranken für $\|F\|_2$ und das Wachstum von $\|x(t)\|_2$ minimiert. Wir schauen nun nach, wie sich $A - BF$ unter Störungen verhält.

Satz 5.37 *Angenommen $\Omega = \{\lambda_1, \ldots, \lambda_n\}$ ist so, daß $Re(\lambda_j) < 0 \ \ \forall i = 1, \ldots, n$ und F ist eine Matrix, die diese Eigenwerte in $A - BF$ zuweist. Dann ist $A - BF + \Delta$ stabil für alle Störungsmatrizen Δ für die gilt:*

$$\|\Delta\|_2 < \min_{s=i\omega} \sigma_n \left\{sI - (A - BF)\right\} =: \delta(F) \tag{5.38}$$

Eine untere Schranke für $\delta(F)$ ist gegeben durch

$$\delta(F) \geq \min_j\{Re(-\lambda_j)\} / cond_2(X) \tag{5.39}$$

Beweis: Für M nichtsingulär gilt:
$M+\Delta = M(I+M^{-1}\Delta)$ ist nichtsingulär, falls $\|M^{-1}\Delta\|_2 \leq \|M^{-1}\|_2\|\Delta\|_2 < 1$. Also folgt, daß $sI-(A-BF+\Delta)$ singulär auf der imaginären Achse ($s=i\omega$) ist nur dann, wenn $\|\Delta\|_2 \geq \delta(F)$. Wegen der Stetigkeit der Eigenwerte (in Abhängigkeit von den Koeffizienten) ist daher $A-BF+\Delta$ stabil, falls (5.38) gilt. Die andere Ungleichung folgt aus

$$\begin{aligned}\delta(F) &= \min_{s=i\omega}\left(\sigma_n(sI-X\Lambda X^{-1})\right) \\ &\geq \min_j Re(-\lambda_j)/\|X^{-1}\|_2\|X\|_2.\end{aligned} \tag{5.40}$$

□

Aus diesen Ergebnissen folgt, daß eine Minimierung von $cond_2(X)$ eine Maximierung der unteren Schranke für $\|\Delta\|_2$ ergibt, d.h. wir drücken damit $\|\Delta\|_2$ nach oben. Wir können aber $\mathrm{cond}_2(X)$ meistens nicht beliebig nahe an 1 bringen.

Wir wollten ja unsere Matrix X so wählen, daß die Spalten aus bestimmten Unterräumen sind und die Kondition minimal wird.

Sei $X=[X_1,\cdots,X_k]$, span $X_j \subseteq \mathcal{S}_j$, $\quad j=1,\ldots,k$ und $\mathcal{S}_1\dot{+}\mathcal{S}_2\dot{+}\cdots\dot{+}\mathcal{S}_k = \mathbb{R}^n$ oder $\mathbb{C}^n$. Sei $X_j \in \mathbb{R}^{n,r_j}, \dim\mathcal{S}_j = m_j$ und sei S_j eine Matrix mit orthonormalen Spalten, so daß $\mathrm{span}S_j = \mathcal{S}_j$. Dann gilt $X_j = S_jD_j$ mit $D_j \in \mathbb{R}^{m_j,r_j}$ und D_j hat vollen Spaltenrang, falls die Spalten von X_j linear unabhängig sind.

$$X = [S_1,\cdots,S_k]\begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_k\end{bmatrix} := SD \tag{5.41}$$

Wähle nun D so, daß ein Maß für die Kondition minimal wird. Wir könnten natürlich $\mathrm{cond}_2(X)$ nehmen, aber das Problem zu lösen ist relativ schwierig. Wir minimieren daher

$$v(D) = \frac{\|DX^{-1}\|_F}{\|D\|_F}. \tag{5.42}$$

Die Idee ist hier, die Matrix X additiv durch Rang 1 Modifikationen zu verändern. Wir lösen dabei nichtlineare kleinste Quadrate Probleme mit Nebenbedingungen.

Wir machen das hier nur für den Fall

$$r_j = 1 \quad j=1,\cdots,n$$

Sei X eine beliebige Startmatrix. Dann gehen wir suksessive durch die Matrix und bestimmen w_j, $\|w_j\|=1$ und minimieren $\|X^{-1}\|_F$ wobei $X=[x_1,\ldots,x_{j-1},\tilde{x}_j,x_{j+1},\ldots,x_n]$ und $\tilde{x}_j = S_jw_j$ und $\|w_j\|_2 = 1$.

Wir schreiben mit $X_j = [x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n]$

$$\|X^{-1}\|_F = \|[S_jw_j,\ X_j]^{-1}\|_F = \|[y_j,\ Y_j]^T\|_F = \|Y^T\|_F \tag{5.43}$$

Mittels einer QR–Zerlegung

$$X_j = [Q_j,\ q_j]\begin{bmatrix} R_j \\ 0\end{bmatrix} \tag{5.44}$$

und $Y^TX=I$ erhalten wir

$$\|Y^T\|_F = \left\|\begin{bmatrix} 0 & \rho_j \\ R_j^{-1} & -\rho_jR_j^{-1}Q_j^TS_jw_j\end{bmatrix}\right\|_F \tag{5.45}$$

wobei $\rho_j = \frac{1}{q_j^T S_j w_j}$.

Um also $\|Y^T\|_F$ zu minimieren, müssen wir

$$\rho_j^2 + \rho_j^2 \, \|R_j^{-1} Q_j^T S_j w_j\|_2^2 = \left\| \begin{bmatrix} R_j^{-1} Q_j^T S_j \\ I_m \end{bmatrix} \rho_j w_j \right\|_2^2$$

minimieren, wobei ρ_j ein Normalisierungsfaktor ist, den man noch rauswerfen kann, wie folgt: Bestimme unitäres $\tilde{P}_j$ so, daß

$$q_j^T S_j = \sigma_j e_m^T \tilde{P}_j^T \quad (Householdertransformation)$$

und sei p_j die m–te Spalte von $\tilde{P}_j = [P_j, \ p_j] \Longrightarrow \rho_j^{-1} = \sigma_j p_j^T w_j$. Setze $\tilde{w}_j = \rho_j P_j^T w_j$.

Dann folgt

$$\begin{aligned} \rho_j w_j &= \rho_j \tilde{P}_j \tilde{P}_j^T w_j = \tfrac{(P_j P_j^T w_j + p_j p_j^T w_j)}{(\sigma_j p_j^T w_j)} \\ &= \sigma_j^{-1} (P_j \tilde{w}_j + p_j). \end{aligned}$$

Also müssen wir das kleinste Quadrate Problem

$$\min \left\| \begin{bmatrix} R_j^{-1} Q_j^T S_j \\ I_m \end{bmatrix} (P_j \tilde{w}_j + p_j) \right\|_2 \quad \text{lösen, für } \tilde{w}_j \tag{5.46}$$

Dies geschieht mittels einer weiteren QR–Zerlegung.
Dann erhalten wir

$$\tilde{x}_j = S_j w_j = (\rho_j \sigma_j)^{-1} S_j (P_j \tilde{w}_j + p_j) \tag{5.47}$$

mit

$$\rho_j^2 = \rho_j^2 w_j^T w_j = \sigma_j^{-1} (\tilde{w}_j^T \tilde{w}_j + 1) \tag{5.48}$$

Dann nehmen wir das nächste x_{j+1}. Dies ist ein sehr teures Verfahren. Einmal durch X durch, kostet $\mathcal{O}(n^3 m) + \mathcal{O}(n^2 m^2)$ flops, aber man kann zeigen, daß es konvergiert.
Damit erhalten wir den folgenden allgemeinen Polvorgabe Algorithmus:

Algorithmus 5.49 *Polvorgabe nach Kautsky/Nichols/Van Dooren.*

Input: $A \in \mathbb{C}^{n,n}, B \in \mathbb{C}^{n,m}$, *Rang* $(B) = m$, $\Omega = \{\lambda_1, \cdots, \lambda_n\} \subset \mathbb{C}$.
Starteigenvektormatrix X (z.B. I)

Output: $F \in \mathbb{C}^{m,n}$, *so daß* $\sigma(A - BF) = \Omega$ *und* $X, X(A - BF)X^{-1}$ *diagonal,* $\|X^{-1}\|_F$ *optimiert.*

1. Schritt
Bestimme mit QR–Zerlegung oder SVD $\quad U = [U_0, \ U_1]$ *unitär, so daß* $B = [U_0, \ U_1] \begin{bmatrix} Z \\ 0 \end{bmatrix}$
sowie Orthonormalbasen für
$\mathcal{S}_j = Kern\left(U_1^T (A - \lambda_j I)\right)$ *und*
$\mathcal{S}_j^{\perp}, \ \ j = 1, \ldots, n \ \ \forall \lambda_j \in \mathbb{R}$.
z. B. QR–Zerlegung

$$\left[U_1^T (A - \lambda_j I)\right]^T = \left[\hat{S}_j, \ S_j\right] \begin{bmatrix} R_j \\ 0 \end{bmatrix}.$$

2. Schritt
Bestimme $X = [x_1, \ldots, x_n], \|x_j\| = 1, \quad x_j = S_j w_j \in \mathcal{S}_j$, *so daß X gut konditioniert (z. B. mit obiger Methode).*

3. Schritt
Bestimme $M = A + BF$ aus der Gleichung $MX = X\Lambda$ durch QR–Zerlegung oder LR–Zerlegung von X und berechne

$$F = Z^{-1}U_0^T(M - A)$$

Dies ist zur Zeit der beste Ansatz für dieses Problem, aber noch nicht befriedigend, da zu teuer.

Zusammenfassung:

Wir haben Algorithmen zur numerischen Bestimmung von $F \in \mathbb{R}^{m,n}$ betrachtet, so daß $A - BF$ vorgebene Pole hat.
Analog gehen wir für $A - FC$ vor. Für $A - BFC$ ist fast alles noch offen.
Man sollte hier beachten, daß Polvorgabe in den meißten Fällen nur verwendet wird, um das System zu stabilisieren. Aber dazu muß man Eigenwerte λ mit $Re(\lambda) < 0$ vorgeben.
Es ist nicht bekannt, wie man dieses am besten macht. Eine andere Möglichkeit zur Stabilisierung ist die Lösung eines Optimalsteuerungsprobems. Dies betrachten wir im nächsten Kapitel.

Kapitel 6

Optimale Steuerung

In diesem Kapitel beschäftigen wir uns nun mit einem der zentralen Themen der Steuerungstheorie, der Optimalsteuerung.

Wir haben wieder unsere Systemgleichung

$$\dot{x} = Ax + Bu \quad x(t_0) = x_0 \tag{6.1}$$
$$A \in \mathbb{R}^{n,n}, B \in \mathbb{R}^{n,m}$$

und wir wollen nun unter allen Steuerungen $u(t)$ diejenige heraussuchen, die ein Kostenfunktional minimiert und zwar

$$\begin{aligned} \mathcal{S}\left(x(t), u(t)\right) &= \frac{1}{2}\left(x(t_f)^T M x(t_f)\right) \\ &\quad + \int_{t_0}^{t_f} \{x(t)^T Q x(t) + u(t)^T R u(t) \\ &\quad + x(t)^T S u(t) + u(t)^T S^T x(t)\} dt. \\ t_0 < t_f \leq \infty, & \end{aligned} \tag{6.2}$$

wobei $M, Q \in \mathbb{R}^{n,n}, S \in \mathbb{R}^{n,m}, R \in \mathbb{R}^{m,m}$, M, Q, R symmetrisch.
Hier erlauben wir als Möglichkeit auch $t_f = \infty$. Dies bedeutet natürlich, daß wir ein asymptotisches Resultat wollen.

Beispiel 6.3 *Als Beispiel wird wieder der Gleichstrommotor betrachtet*

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 0 & b/J \end{bmatrix} x + \begin{bmatrix} 0 \\ k/J \end{bmatrix} u.$$

Als Kostenfunktional wählen wir

$$\mathcal{S}(x, u) := \frac{1}{2}\left\{ x(t_f)^T M x(t_f) + \int_{t_0}^{t_f} \left(x(t)^T \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} x(t) + u(t)^T \rho u(t) \right) dt \right\}$$

wobei M, ρ noch nicht spezifizierte, positiv definite Matrizen sind.
Es gibt also 3 Kriterien die „kosten“.

- *Die Abweichung des Endzustandes $x(t_f)$ von der Ruhelage.*
- *Das Einschwingverhalten des Zustands.*

– *Die gemittelte Amplitude der Eingangsspannung.*

Um dieses Problem zu lösen, verwenden wir das folgende Hamiltonische Prinzip: (Lagrange–Multiplikator–Methode).

Satz 6.4 *Betrachte das Optimalsteuerungsproblem (6.1), (6.2). Sei* $u_* \in U_m := \{u(t) \in \mathbb{R}^m, u(t)$ *stückweise stetig auf* $[t_0, t_f]\}$ *optimale Steuerung und sei* $x_*(t) \in \mathbb{R}^n$ *die zugehörige Lösung des geschlossenen Kreises, also die Lösung von*

$$\dot{x}(t) = Ax(t) + Bu_*(t) \quad x(t_0) = x^0 \tag{6.5}$$

Dann gibt es eine Kozustandsfunktion (Lagrange Multiplikator) $\mu(t) \in \mathbb{R}^n$*, so daß* $x_*(t), \mu(t), u_*(t)$ *das lineare Randwertproblem*

$$\begin{bmatrix} A & 0 & B \\ Q & A^T & S \\ S^T & B^T & R \end{bmatrix} \begin{bmatrix} x(t) \\ \mu(t) \\ u(t) \end{bmatrix} = \begin{bmatrix} I_n & 0 & 0 \\ 0 & -I_n & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ \dot{\mu}(t) \\ \dot{u}(t) \end{bmatrix} \tag{6.6}$$

$$x(t_0) = x^0, \ \mu(t_f) = Mx(t_f) \tag{6.7}$$

lösen.
(Beachte: $\dot{u}(t)$ *ist nur formal vorhanden.)*

Beweis: Dies ist eine Variante des Pontryagin'schen Maximumprinzips. Im Prinzip geht der allgemeine Beweis analog.

Sei u_* die Optimalsteuerung. Betrachte eine Störung erster Ordnung

$$u(t) = u_*(t) + \varepsilon v(t) \tag{6.8}$$

mit $u(t) \in U_m$. Dann erhalten wir aus (6.5)

$$\dot{x}(t) = Ax(t) + Bu_*(t) + \varepsilon Bv(t). \tag{6.9}$$

Dies hat die Lösung (siehe Kapitel 2)

$$\begin{aligned} x(t) &= e^{A(t-t_0)}x^0 + \int_{t_0}^{t} e^{A(t-s)}B\left(u_*(s) + \varepsilon v(s)\right) ds \\ &= x_*(t) + \varepsilon \int_{t_0}^{t} \left(e^{A(t-s)}Bv(s)\right) ds \\ &= x_*(t) + \varepsilon\varphi(t). \end{aligned} \tag{6.10}$$

Dabei erfüllt $\varphi(t)$ die Differentialgleichung

$$\dot{\varphi}(t) = A\varphi(t) + Bv(t), \varphi(t_0) = 0. \tag{6.11}$$

Wir führen nun den Vektor $\mu(t) \in \mathbb{R}^n$ und die Hamiltonfunktion $H(x, \mu, u)$ ein durch

$$\begin{aligned} H(x,\mu,u) &= x^T(t)Qx(t) + x(t)^T Su(t) + u(t)^T S^T x(t) \\ &\quad + u(t)^T Ru(t) + \mu(t)^T \left(Ax(t) + Bu(t)\right) \\ &\quad + \left(Ax(t) + Bu(t)\right)^T \mu(t) \end{aligned} \tag{6.12}$$

Dies ist der Integrand plus die rechte Seite der Differentialgleichung. Ähnlich wie Lagrange Multiplikator.

Ansatz: Wir können dann $\mathcal{S}(x,u)$ schreiben als

$$\mathcal{S}(x,u) = \frac{1}{2}\left\{x^T(t_f)Mx(t_f) + \int_{t_0}^{t_f}\left(H(x,\mu,u) - \mu^T\dot{x} - \dot{x}^T\mu\right)dt\right\} \tag{6.13}$$

und analog für u_*, x_*

$$\mathcal{S}(x_*,u_*) = \frac{1}{2}\left\{x_*^T(t_f)Mx_*(t_f) + \int_{t_0}^{t_f} H(x_*,\mu,u_*) - \mu^T\dot{x}_* - \dot{x}_*^T\mu)dt\right\} \tag{6.14}$$

und wir erhalten

$$\begin{aligned}\mathcal{S}(x,u) - \mathcal{S}(x_*,u_*) &= \frac{1}{2}\Big\{\left(x(t)^TMx(t) - x_*(t)^TMx_*(t)\right)|_{t=t_f} \\ &+ \int_{t_0}^{t_f}\left(H(x,\mu,u) - H(x_*,\mu,u_*)\right)dt \\ &+ \int_{t_0}^{t_f}\underbrace{(\mu^T\underbrace{(\dot{x}_*-\dot{x})}_{-\varepsilon\dot{\varphi}} + \underbrace{(\dot{x}_*-\dot{x})^T}_{-\varepsilon\dot{\varphi}^T}\mu)}_{=-2\varepsilon\mu^T\dot{\varphi}}dt\Big\}\end{aligned} \tag{6.15}$$

$$\begin{aligned}\frac{1}{2}(H(x,\mu,u) - H(x_*,\mu,u_*)) &= \frac{1}{2}\Big\{x^TQx + x^TSu + u^TS^Tx + u^TRu + \mu^T(Ax+Bu) \\ &+(Ax+Bu)^T\mu \\ &-x_*^TQx_* - x_*^TSu_* - u_*^TS^Tx_* - u_*^TRu_* - \mu^T(Ax_*+Bu_*) \\ &-(Ax_*+Bu_*)^T\mu\Big\}\end{aligned}$$

Wir verwenden jetzt die Tatsache, das

$$u = u_* + \varepsilon v, \quad x = x_* + \varepsilon\varphi$$

und erhalten

$$\begin{aligned}\frac{1}{2}(H(x,\mu,u) - H(x_*,\mu,u_*)) &= \frac{1}{2}\Big\{x_*^TQx_* + 2\varepsilon x_*^TQ\varphi + \varepsilon^2\varphi^TQ\varphi - x_*^TQx_* \\ &+\varepsilon^2\varphi^TSv + x_*^TSu_* + \varepsilon(x_*^TSv + \varphi^TSu_*) - x_*^TSu_* \\ &+\varepsilon^2v^TS^T\varphi + u_*^TS^Tx_* + \varepsilon(u_*^TS^T\varphi + v^TS^Tx_*) - u_*^TS^Tx_* \\ &+\varepsilon^2v^TRv + u_*^TRu_* + 2\varepsilon(u_*^TRv) - u_*^TRu_* \\ &+\varepsilon\mu^TA\varphi + \varepsilon\mu^TBv + \varepsilon\varphi^TA^T\mu + \varepsilon v^TB^T\mu\Big\} \\ &= \varepsilon(x_*^TQ\varphi + x_*^TSv + u_*^TS^T\varphi + u_*^TRv \\ &+\mu^TA\varphi + \mu^TBv) + \mathcal{O}(\varepsilon^2) \\ &= \varepsilon\Big\{[x_*^TQ + u_*^TS^T + \mu^TA]\varphi \\ &+[x_*^TS + u_*^TR + \mu^TB]v\Big\} + \mathcal{O}(\varepsilon^2)\end{aligned}$$

Weiter gilt mit partieller Integration

$$\begin{aligned}-\int_{t_0}^{t_f}\varepsilon\mu^T\dot{\varphi}dt &= -\varepsilon\mu^T\varphi\Big|_{t_0}^{t_f} + \varepsilon\int_{t_0}^{t_f}\dot{\mu}^T\varphi dt \\ &= -\varepsilon\mu^T(t_f)\varphi(t_f) + \varepsilon\int_{t_0}^{t_f}\dot{\mu}^T\varphi dt\end{aligned}$$

$$x^T M x - x_*^T M x_* = 2\varepsilon x_*^T M \varphi + \mathcal{O}(\varepsilon^2)$$

Zusammen erhalten wir also

$$\begin{aligned} \mathcal{S}(x,u) - \mathcal{S}(x_*,u_*) &= \mathcal{O}(\varepsilon^2) \\ &\quad + \varepsilon \left\{ \int_{t_0}^{t_f} \Big([x_*^T Q + u_*^T S^T + \mu^T A]\varphi + \dot{\mu}^T \varphi \right. \\ &\quad + [x_*^T S + u_*^T R + \mu^T B] v \Big)\, dt \\ &\quad \left. - \mu^T(t_f)\varphi(t_f) + x_*^T(t_f) M \varphi(t_f) \right\}. \end{aligned} \tag{6.16}$$

Da $\mathcal{S}(x,u) - \mathcal{S}(x_*,u_*) \geq 0 \quad \forall \varepsilon$ genügend klein (pos. oder negativ) so folgt, daß der Faktor von ε verschwinden muß für alle v und daraus resultierende φ. Wir wählen nun einfach μ als Lösung von

$$-\dot{\mu}(t) = A^T \mu(t) + Q x_* + S u_* \tag{6.17}$$

mit „Endbedingung“

$$\mu(t_f) = M x_*(t_f) \tag{6.18}$$

damit fällt bis auf den mittleren Term alles weg und wir erhalten

$$\int_{t_0}^{t_f} [x_*^T S + u_*^T R + \mu^T B] v dt = 0 \quad \forall v \in U_m \tag{6.19}$$

woraus sofort folgt, daß

$$x_*^T S + u_*^T R + \mu^T B \equiv 0 \quad \forall t \in [t_0, t_f] \tag{6.20}$$

Fassen wir nun die Gleichungen (6.20), (6.17), (6.18) und (6.1) zusammen, so erhalten wir das Zwei–Punkt–Randwertproblem (6.6),(6.7). □

Ein Punkt, der oft schwer nachzuvollziehen ist, ist die Wahl von μ in (6.17), (6.18). Diese kann man noch näher begründen durch geschickte Auswahl von v, φ. Aber das würde hier zu weit führen. Die Existenz der Lösung von (6.17), (6.18) reicht hier für diesen Satz.
Daß die Wahl vernünftig ist, zeigt uns auch der folgende Satz:

Satz 6.21 *Seien x_*, μ, u_* so gewählt, daß* $\begin{bmatrix} x_* \\ \mu \\ u_* \end{bmatrix}$ *Lösung des linearen Randwertproblems (6.6), (6.7) ist. Es gelte weiterhin, daß* $\mathcal{R} := \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix}$, *$M$ positiv semidefinit sind.*
Dann gilt

$$\mathcal{S}(x,u) \geq \mathcal{S}(x_*,u_*) \tag{6.22}$$

für alle x, u welche (6.1) erfüllen.

Beweis: Definiere

$$\Phi(s) = \mathcal{S}\left(s x_*(t) + (1-s)x(t), s u_*(t) + (1-s)u(t)\right). \tag{6.23}$$

Die Behauptung des Satzes ist äquivalent zu der Aussage, daß $\Phi(s)$ sein Minimum bei $s = 1$ hat für alle $x(t), u(x)$ welche (6.1) erfüllen.
$\Phi(s)$ ist quadratisch in s, also hat $\Phi(s)$ ein Minimum für $s = 1$ genau dann, wenn

$$\left.\frac{d\Phi}{ds}\right|_{s=1} = 0, \quad \left.\frac{d^2\Phi}{ds^2}\right|_{s=1} \geq 0 \tag{6.24}$$

(Normalerweise $\left.\frac{d^2\Phi}{ds^2}\right|_{s=1} > 0$, aber da das Funktional quadratisch ist, reicht ≥ 0)

$$\begin{aligned}
\left.\frac{d\Phi}{ds}\right|_{s=1} &= \Big[\left.(x_* - x)^T M x_*\right|_{t=t_f} \\
&+\frac{1}{2}\int_{t_0}^{t_f} \Big\{(x_* - x)^T Q x_* + x_*^T Q(x_* - x) \\
&+u_*^T S^T(x_* - x) + (u_* - u)^T S^T x_* \\
&+(u_* - u)^T R u_* + u_*^T R(u_* - u) \\
&+ (x_* - x)^T S u_* + x_*^T S(u_* - u)\Big\}\, dt\Big]
\end{aligned} \tag{6.25}$$

Wenn wir die 2. Gleichung von (6.6) von links mit x_*^T multiplizieren und die anderen Gleichungen einsetzen, erhalten wir

$$\begin{aligned}
x_*^T Q x_* &= -x_*^T A^T \mu - x_*^T S u_* - x_*^T \dot{\mu} \\
\text{(1. Gl. (6.6))} \quad &= u_*^T B^T \mu - \dot{x}_*^T \mu - x_*^T S u_* - x_*^T \dot{\mu} \\
\text{(3. Gl. (6.6))} \quad &= -u_*^T S^T x_* - u_*^T R u_* - \dot{x}_*^T \mu - x_*^T S u_* - x_*^T \dot{\mu}
\end{aligned} \tag{6.26}$$

Analog nach Multiplikation mit x^T

$$\begin{aligned}
x^T Q x_* &= -x^T A^T \mu - x^T S u_* - x^T \dot{\mu} \\
\text{(Ausgangsgl. für } x) \quad &= u^T B^T \mu - \dot{x}^T \mu - x^T S u_* - x^T \dot{\mu} \\
\text{(2. Gl. für } B^T\mu) \quad &= -u^T S^T x_* - u^T R u_* - \dot{x}^T \mu - x^T S u_* - x^T \dot{\mu}
\end{aligned} \tag{6.27}$$

Einsetzen von (6.26), (6.27) in (6.25) ergibt

$$\begin{aligned}
\left.\frac{d\Phi}{ds}\right|_{s=1} &= \left.(x_* - x)^T M x_*\right|_{t=t_f} \\
&+\frac{1}{2}\int_{t_0}^{t_f} (x^T \dot{\mu} + \dot{x}^T \mu - x_*^T \dot{\mu} - \dot{x}_*^T \mu) dt \\
&= \left.(x_* - x)^T M x_*\right|_{t=t_f} + \left. x^T \mu\right|_{t=t_0}^{t=f_f} - \left. x_*^T \mu\right|_{t=t_0}^{t=t_f}
\end{aligned}$$

Nun gilt aber, daß für $t = t_0 \quad x(t_0) = x_*(t_0)$ und für $t = t_f \quad \mu(t_f) = M x_*(t_f) \Longrightarrow \left.\frac{d\Phi}{ds}\right|_{s=1} = 0$.

$$\begin{aligned}
\left.\frac{d^2\Phi}{ds^2}\right|_{s=1} &= \left.(x_* - x)^T M (x_* - x)\right|_{t=t_f} \\
&+ \int_{t_0}^{t_f} \left[(x_* - x)^T, (u_* - u)^T\right] \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} x_* - x \\ u_* - u \end{bmatrix} dt \\
&\geq 0, \text{ da } M, \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \text{ positiv semidefnit.}
\end{aligned}$$

□

Wir haben also den Zusammenhang hergestellt zwischen der Lösung des optimalen Steuerungsproblems und der Lösung eines linearen 2 Punkt Randwertproblems. Theoretisch könnten wir nun aufhören, wenn es nicht noch einen weiteren Trick gäbe, der die Lösung dieses Randwertproblems vereinfacht.

Zuerst einmal vereinfachen wir noch das Problem (6.6), (6.7) indem wir ausnutzen, daß R positiv definit ist. Also folgt aus

$$\begin{bmatrix} A & 0 & B \\ Q & A^T & S \\ S^T & B^T & R \end{bmatrix} \begin{bmatrix} x \\ \mu \\ u \end{bmatrix} = \begin{bmatrix} I_n & 0 & 0 \\ 0 & -I_n & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x} \\ \dot{\mu} \\ \dot{u} \end{bmatrix},$$

daß gilt

$$u(t) = -R^{-1}\left(S^T x(t) + B^T \mu(t)\right). \tag{6.28}$$

Einsetzen in die anderen Gleichungen ergibt das System

$$\begin{bmatrix} A - BR^{-1}S^T & -BR^{-1}B^T \\ -(Q - SR^{-1}S^T) & -A^T + SR^{-1}B^T \end{bmatrix} \begin{bmatrix} x \\ \mu \end{bmatrix} = \begin{bmatrix} \dot{x} \\ \dot{\mu} \end{bmatrix} \tag{6.29}$$

mit Randbedingungen

$$x(t_0) = x^0, \quad \mu(t_f) = Mx(t_f) \tag{6.30}$$

Wir betrachten die Matrix in (6.29) einmal näher. Da wir vorausgesetzt haben, daß

$$\mathcal{R} = \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \text{ semidefinit, } R \text{ positiv definit },\tag{6.31}$$

so folgt $\underline{Q - SR^{-1}S^T \text{ symmetrisch positiv semidefinit}}$ (Schur–Komplement). Weiter ist $BR^{-1}B^T$ symmetrisch positiv semidefinit mit Rang m, also hat (6.29) die Form

$$\begin{bmatrix} \dot{x} \\ \dot{\mu} \end{bmatrix} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \begin{bmatrix} x \\ \mu \end{bmatrix} = \mathcal{H} \begin{bmatrix} x \\ \mu \end{bmatrix} \tag{6.32}$$

mit H symmetrisch negativ semidefinit, G symmetrisch negativ semidefinit. Eine Matrix der From $\mathcal{H} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix}$ mit G, H symmetrisch heißt <u>Hamiltonische Matrix</u>. Hamiltonische Matrizen sind schiefsymmetrisch bezüglich eines indefiniten Skalarproduktes.

$$< x, y >_J = x^T J y \text{ wobei } J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}. \tag{6.33}$$

Beachte, daß gilt:

$$\begin{aligned} < x, \mathcal{H}y >_J &= x^T J \mathcal{H} y \\ &= x^T \begin{bmatrix} H & -F^T \\ -F & -G \end{bmatrix} y \\ &= x^T \begin{bmatrix} H^T & -F^T \\ -F & -G^T \end{bmatrix} y \\ &= x^T (J\mathcal{H})^T y \\ &= x^T \mathcal{H}^T J^T y \\ &= -x^T \mathcal{H}^T J y \\ &= - < x\mathcal{H}^T, y >_J \end{aligned} \tag{6.34}$$

Die Menge der Hamiltonischen Matrizen in $\mathbb{R}^{2n,2n}(\mathbb{C}^{2n,2n})$

$$\begin{aligned} \mathcal{H}_{2n}(\mathbb{R}) &= \left\{ \mathcal{H} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} | G = G^T, H = H^T, F \in \mathbb{R}^{n,n} \right\} \\ \mathcal{H}_{2n}(\mathbb{C}) &= \left\{ \mathcal{H} = \begin{bmatrix} F & G \\ H & -F^* \end{bmatrix} | G = G^*, H = H^*, F \in \mathbb{C}^{n,n} \right\} \end{aligned} \tag{6.35}$$

ist eine Lie–Algebra mit der normalen Matrixaddition und der Lie–Multiplikation:

$$[\mathcal{H}_1, \mathcal{H}_2] := \mathcal{H}_1\mathcal{H}_2 - \mathcal{H}_2\mathcal{H}_1 \tag{6.36}$$

denn

$$\begin{aligned}
&\begin{bmatrix} F_1 & G_1 \\ H_1 & -F_1^T \end{bmatrix}\begin{bmatrix} F_2 & G_2 \\ H_2 & -F_2^T \end{bmatrix} - \begin{bmatrix} F_2 & G_2 \\ H_2 & -F_2^T \end{bmatrix}\begin{bmatrix} F_1 & G_1 \\ H_1 & -F_1^T \end{bmatrix} \\
= &\begin{bmatrix} F_1F_2 + G_1H_2 & F_1G_2 - G_1F_2^T \\ H_1F_2 - F_1^TH_2 & H_1G_2 + F_1^TF_2^T \end{bmatrix} - \begin{bmatrix} F_2F_1 + G_2H_1 & F_2G_1 - G_2F_1^T \\ H_2F_1 - F_2^TH_1 & H_2G_1 + F_2^TF_1^T \end{bmatrix} \\
= &\begin{bmatrix} F_1F_2 - F_2F_1 + G_1H_2 - G_2H_1 & F_1G_2 - G_1F_2^T + G_2F_1^T - F_2G_1 \\ H_1F_2 - F_1^TH_2 - H_2F_1 + F_2^TH_1 & H_1G_2 + F_1^TF_2^T - H_2G_1 - F_2^TF_1^T \end{bmatrix} \in \mathcal{H}_{2n}
\end{aligned}$$

Die auf dieser Lie–Algebra operierende Lie–Gruppe ist die Menge der symplektischen Matrizen

$$\begin{aligned} \mathcal{S}_{2n}(\mathbb{R}) &:= \left\{S \in \mathbb{R}^{2n,2n} | S^TJS = J\right\} \\ \mathcal{S}_{2n}(\mathbb{C}) &:= \{S \in \mathbb{C}^{2n,2n} | S^*JS = J\} \end{aligned} \tag{6.37}$$

Dies sind die bzgl. $<,>_J$ orthogonalen (unitären) Matrizen, denn

$$< x, Sy >_J = x^TJSy = x^TS^{-T}Jy = < S^{-1}x, y >_J . \tag{6.38}$$

Daß dies eine Gruppe ist bzgl. der normalen Matrix–Multiplikation, folgt sofort aus der Definition (denn $S_1^TS_2^TJS_2S_1 = S_1^TJS_1 = J$).

Man beachte, daß gilt

$$|\det S| = 1, \tag{6.39}$$

denn $\det S^TJS = (\det S)^2 \det J = (\det S)^2 = 1$,
aber symplektische Matrizen sind im allgemeinen nicht normbeschränkt, denn es gilt

$$\begin{bmatrix} 1 & v \\ 0 & 1 \end{bmatrix} \in \mathcal{S}_2(\mathbb{R}) \quad \forall v \in \mathbb{R}. \tag{6.40}$$

$$\begin{bmatrix} 1 & 0 \\ v & 1 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} 1 & v \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & v \end{bmatrix}\begin{bmatrix} 1 & v \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Über die Eigenstruktur von Hamiltonischen Matrizen haben wir nun die folgende Aussage:

Satz 6.41 *Seien $z^{(0)}, z^{(1)}, \ldots, z^{(k)} \in \mathbb{C}^{2n}$ und $\lambda \in \mathbb{C}$ so daß für $\mathcal{H} \in \mathcal{H}_{2n}(\mathbb{R})$ (oder $\mathcal{H} \in \mathcal{H}_{2n}(\mathbb{C})$)*

(i) $(\mathcal{H} - \lambda I)z^{(0)} = 0$

(ii) $(\mathcal{H} - \lambda I)z^{(j)} = z^{(j-1)}, \ j = 1, \cdots, k$

so gilt für $w^{(j)} = Jz^{(j)}, \ j = 0, \cdots, k$

(iii) $w^{(0)^*}(\mathcal{H} + \bar{\lambda}I) = 0$

(iv) $w^{(j)^*}(\mathcal{H} + \bar{\lambda}I) = -w^{(j-1)^*}, \ j = 1, \cdots, k$

Beweis:

$$\begin{aligned}
(\mathcal{H}-\lambda I)z^{(0)}=0 &\iff (J\mathcal{H}-\lambda J)z^{(0)}=0\\
&\iff (-\mathcal{H}^*J-\lambda J)z^{(0)}=0\\
&\iff (\mathcal{H}^*+\lambda I)w^{(0)}=0\\
&\iff w^{(0)^*}(\mathcal{H}+\bar{\lambda}I)=0
\end{aligned}$$

und analog

$$\begin{aligned}
(\mathcal{H}-\lambda I)z^{(j)}=z^{(j-1)} &\iff (J\mathcal{H}-\lambda J)z^{(j)}=Jz^{(j-1)}\\
&\iff (-\mathcal{H}^*J-\lambda J)z^{(j)}=w^{(j-1)}\\
&\iff w^{(j)*}(\mathcal{H}+\bar{\lambda}I)=-w^{(j-1)*}
\end{aligned}$$

□

Beachte, daß im reellen Fall dann $\lambda, \bar{\lambda}, -\lambda, -\bar{\lambda}$ Eigenwerte sind. Aus diesem Satz folgt sofort, daß zu jedem Jordanblock von H zum Eigenwert λ ein gleichgroßer Jordanblock von H zum Eigenwert $-\bar{\lambda}$ gehört. Diese sind auf jeden Fall verschieden, falls $\lambda \neq -\bar{\lambda}$, d.h. $Im(\lambda) \neq 0$.

Lemma 6.42 *Sei* $\mathcal{H}=\begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \in \mathcal{H}_{2n}, \lambda \in \mathbb{C}, z=\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$, *so daß* $\mathcal{H}z=\lambda z$, *so gilt*

$$z_2^*Gz_2+z_1^*Hz_1=(\lambda+\bar{\lambda})z_1^*z_2=2Re(\lambda)z_1^*z_2 \tag{6.43}$$

Beweis: Aus $\mathcal{H}z=\lambda z$ folgt

$$\begin{array}{rl} z_2^*| & Fz_1+Gz_2 = \lambda z_1\\ z_1^*| & Hz_1-F^Tz_2 = \lambda z_2 \end{array}$$

$$\Longrightarrow \left\{ \begin{array}{rcl} z_2^*Fz_1+z_2^*Gz_2 &=& \lambda z_2^*z_1\\ z_1^*Hz_1-z_1^*F^Tz_2 &=& \lambda z_1^*z_2 \end{array}\right.$$

Konjugation der 2. Gleichung liefert

$$z_1^*Hz_1-z_2^*Fz_1=\bar{\lambda}z_2^*z_1$$

Addition der Gleichungen liefert

$$z_2^*Gz_2+z_1^*Hz_1=(\lambda+\bar{\lambda})z_1^*z_2$$

□

Satz 6.44 *Sei* $\mathcal{H}=\begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \in \mathcal{H}_{2n}, \mathcal{H}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}=\lambda\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$.

i) Falls $Re(\lambda)=0$ *so folgt, daß*

$$z_2^*Gz_2+z_1^*Hz_1=0$$

ii) Falls G, H *negativ semidefinit und* $Re(\lambda)=0$, *so folgt*

$$Gz_2=0,\ Hz_1=0$$

iii) Sei $\mathcal{H}$ wie in (6.29), $\mathcal{R} = \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix}$ positiv definit und (A, B) stabilisierbar. So hat $\mathcal{H}$ keine Eigenwerte mit Realteil 0.

Beweis:

i) klar aus Lemma 6.42.

ii) folgt aus i)

iii) Angenommen es gibt λ mit $Re(\lambda) = 0$ und $\mathcal{H}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \lambda \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$

Aus (6.43) folgt $z_2^* G z_2 + z_1^* H z_1 = 0$. Es gilt $G = -BR^{-1}B^T$ negativ semidefinit und $H = -(Q - SR^{-1}S^T)$ negativ semidefinit $\Longrightarrow Gz_2 = 0, Hz_1 = 0$, d.h.: $(Q - SR^{-1}S^T)z_1 = 0, BR^{-1}B^T z_2 = 0$ und da R positiv definit folgt sofort

$$B^T z_2 = 0 \Longrightarrow z_2^* B = 0.$$

Aus Satz 6.41 folgt, daß

$$[z_2^*, \; -z_1^*] \begin{bmatrix} A - BR^{-1}S^T & -BR^{-1}B^T \\ -(Q - SR^{-1}S^T) & -(A - BR^{-1}S^T)^T \end{bmatrix} = -\bar{\lambda}[z_2^*, \; -z_1^*],$$

mit $z_1^*(Q - SR^{-1}S^T) = 0, z_2^* B = 0$ folgt

$$z_2^*(A - BR^{-1}S^T) = z_2^* A = -\bar{\lambda} z_2^*$$

Nun hatten wir vorausgesetzt, daß $\begin{bmatrix} Q & S \\ S^T & R \end{bmatrix}$ positiv definit ist, dann ist auch

$$\begin{bmatrix} I & -SR^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} I & 0 \\ -R^{-1}S^T & 0 \end{bmatrix} = \begin{bmatrix} Q - SR^{-1}S^T & 0 \\ 0 & R \end{bmatrix}$$

positiv definit, also folgt $z_1 = 0 \Longrightarrow z_2 \neq 0$
$\Longrightarrow$ Widerspruch zu (A, B) stabilisierbar, denn wir haben $z_2 \neq 0, z_2^* A = -\bar{\lambda} z_2^* \quad (Re(\bar{\lambda}) = 0)$ und $z_2^* B = 0$. $\square$

Wir erhalten also, daß unter den sinnvollen Annahmen, das $\mathcal{R}$ positiv definit und (A, B) stabilisierbar ist, gilt daß $\mathcal{H} = \begin{bmatrix} A - BR^{-1}S^T & -BR^{-1}B^T \\ -(Q - SR^{-1}S^T) & -(A - BR^{-1}S^T)^T \end{bmatrix}$ keine rein imaginären Eigenwerte hat.
Falls $S = 0$ ist, so kann man "$\mathcal{R}$ positiv definit" abschwächen durch "R positiv definit und (A, Q) entdeckbar".

Satz 6.45 *Sei $\mathcal{H} \in \mathcal{H}_{2n}$ und $\mathcal{H}$ habe keine rein imaginären Eigenwerte. Dann gibt es eine symplektische Matrix $S \in \mathcal{S}_{2n}(\mathbb{C}) = \{S \in \mathbb{C}^{2n,2n} | S^* J S = J\}$.*

$$S^{-1}\mathcal{H}S = \begin{bmatrix} J_1 & 0 \\ 0 & -J_1^* \end{bmatrix} \tag{6.46}$$

und J_1 ist in Jordan'scher Normalform und hat nur Eigenwerte mit negativem Realteil.

Beweis:

Aus Satz 6.41 folgt, daß für alle Eigen– und Hauptvektoren gilt, daß mit

z (rechter) Eigen– oder Hauptvektor zum Eigenwert λ,
Jz (linker) Eigen– oder Hauptvektor zu $-\bar{\lambda}$ ist.

Es gibt deshalb ein V, so daß

$$V^{-1}\mathcal{H}V = \begin{pmatrix} J_1 & 0 \\ 0 & J_2 \end{pmatrix}$$

in Jordanscher Normalform ist, J_1 nur Eigenwerte in der linken und J_2 nur Eigenwerte in der rechten komplexen Halbebene hat. Wir können aufgrund von Satz 6.41 o.B.d.A. annehmen, daß $J_2 = -J_1^*$ ist. Wir müssen noch zeigen, daß diese Hamiltonische Form der Jordanschen Normalform auch mit einer symplektischen Matrix S zu erreichen ist. Sei

$$V = (V_1, \quad V_2)$$

Wir setzen

$$W = \begin{pmatrix} -(JV_2)^* \\ (JV_1)^* \end{pmatrix} = J^*V^*J$$

Dann gilt nach Satz 6.41:

$$W\mathcal{H} = \begin{pmatrix} J_1 & 0 \\ 0 & -J_1^* \end{pmatrix} W$$

Da die Hauptraumzerlegung eindeutig ist und J_1 und J_1^* keine gemeinsamen Eigenwerte haben, existiert eine nicht singuläre Block Diagonalmatrix

$$D = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix},$$

so daß

$$W = J^*V^*J = DV^{-1} \Longleftrightarrow V^*JV = JD = \begin{pmatrix} 0 & D_2 \\ -D_1 & 0 \end{pmatrix},$$

Weil V^*JV schief–Hermitesch ist, ist bereits $D_2 = D_1^*$. Setze

$$S := V \begin{pmatrix} D_1^{-1} & 0 \\ 0 & I \end{pmatrix},$$

$$\Longrightarrow S^*JS = \begin{pmatrix} D_1^{-*} & 0 \\ 0 & I \end{pmatrix} V^*JV \begin{pmatrix} D_1^{-1} & 0 \\ 0 & I \end{pmatrix} = \begin{pmatrix} D_1^{-*} & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} 0 & D_1^* \\ -D_1 & 0 \end{pmatrix} \begin{pmatrix} D_1^{-1} & 0 \\ 0 & I \end{pmatrix} = J$$

Dieses S ist symplektisch nach Konstruktion und es ist

$$S^{-1}\mathcal{H}S = \begin{pmatrix} J_1 & 0 \\ 0 & -J_1^* \end{pmatrix}$$

□

Wir haben also eine strukturerhaltende Jordan–Form. Aber diese ist nicht numerisch stabil berechenbar, denn die symplektischen Matrizen können beliebig große Norm haben. Also brauchen wir orthogonal symplektische Transformationen für die numerische Stabilität. Die Menge der orthogonal unitär symplektischen Matrizen bezeichnen wir mit

$$\begin{aligned} \mathcal{US}_{2n}(\mathbb{C}) &= \{Q \in \mathcal{S}_{2n}(\mathbb{C}) | Q^*Q = I\} \\ \mathcal{US}_{2n}(\mathbb{R}) &= \{Q \in \mathcal{S}_{2n}(\mathbb{R}) | Q^TQ = I\} \end{aligned} \qquad (6.47)$$

Es gilt das folgende Lemma:

Lemma 6.48 *Sei* $Q \in \mathcal{US}_{2n}(\mathbb{C})\,(\mathcal{US}_{2n}(\mathbb{R}))Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$ *mit* $Q_{ij} \in \mathbb{C}^{n,n}\,(\mathbb{R}^{n,n})$ *so folgt* $Q_{12} = -Q_{21}, Q_{22} = Q_{11}$.

Beweis:

$$
\begin{aligned}
Q \in \mathcal{S}_{2n}(\mathbb{C}) &\iff Q^*JQ = J, Q^*Q = I \\
&\implies JQ = Q^{-*}J = QJ \\
&\implies \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} = \begin{bmatrix} Q_{21} & Q_{22} \\ -Q_{11} & -Q_{12} \end{bmatrix} \\
&\quad = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}\begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} = \begin{bmatrix} -Q_{12} & Q_{11} \\ -Q_{22} & Q_{21} \end{bmatrix} \\
&\implies \text{Beh.}
\end{aligned}
$$

□

Was können wir nun mit unitär symplektischen Transformationsmatrizen erreichen?

Satz 6.49 *(Hamiltonische Schurform)*

(i) Sei $\mathcal{H} \in \mathcal{H}_{2n}(\mathbb{R})(\mathcal{H}_{2n}(\mathbb{C}))$. $\mathcal{H}$ *habe keine Eigenwerte mit Realteil* 0. *Dann gibt es*

$$Q \in \mathcal{US}_{2n}(\mathbb{C}), \ \textit{so daß}$$

$$Q^*\mathcal{H}Q = \begin{bmatrix} T & N \\ 0 & -T^* \end{bmatrix}, T, N \in \mathbb{C}^{n,n} \tag{6.50}$$

mit T *obere Dreiecksmatrix,* $N = N^*$ *und* T *kann so gewählt werden, daß alle Eigenwerte von* T *negativen Realteil haben.*

(ii) Sei $\mathcal{H} \in \mathcal{H}_{2n}(\mathbb{R})$. $\mathcal{H}$ *habe keine Eigenwerte mit Realteil* 0. *Dann gibt es*

$$Q \in \mathcal{US}_{2n}(\mathbb{R}), \textit{so daß}$$

$$Q^T\mathcal{H}Q = \begin{bmatrix} T & N \\ 0 & -T^T \end{bmatrix} \quad T, N \in \mathbb{R}^{n,n}, \tag{6.51}$$

T *quasiobere Dreiecksmatrix,* $N = N^T$ *und* T *kann so gewählt werden, daß alle Eigenwerte von* T *negativen Realteil haben.*

Beweis:

i) Sei $\lambda_1 \in \sigma(\mathcal{H})$ Re $(\lambda_1) < 0$, und $\mathcal{H}x_1 = \lambda_1 x_1, \|x_1\|_2 = 1$.
Sei $Q_1 \in \mathcal{US}_{2n}(\mathbb{C})$ so daß $Q_1^* x_1 = e_1$. Wir werden später noch Algorithmen entwickeln, die dies machen. Dann gilt

$$Q_1^*\mathcal{H}Q_1 = \left[\begin{array}{cc|cc} \lambda_1 & w_1^T & \rho_1 & w_3 \\ 0 & A_1 & w_3^* & G_1 \\ \hline 0 & 0 & -\lambda_1 & 0 \\ 0 & H_1 & -\bar{w}_1 & -A_1^* \end{array}\right]$$

wie man sofort nachrechnet und

$$\begin{bmatrix} A_1 & G_1 \\ H_1 & -A_1^* \end{bmatrix}$$

ist wieder Hamiltonisch. Per Induktion folgt dann die Behauptung.

ii) Sei $\lambda \in \sigma(\mathcal{H})$ Re $(\lambda) < 0, \lambda \in \mathbb{C} \setminus \mathbb{R}$. Für $\lambda \in \mathbb{R}$ verwende das Argument aus i). Seien $x_1, x_2 \in \mathbb{R}^{2n}$, so daß

$$\mathcal{H}[x_1,\ x_2] = [x_1,\ x_2]Z$$

mit $Z \in \mathbb{R}^{2,2}, \sigma(Z) = \lambda, \bar{\lambda}$.

Und sei $Q_1 \in \mathcal{US}_{2n}(\mathbb{R})$ so daß

$$Q_1^T\,[x_1,\ x_2] = \begin{bmatrix} \tilde{t}_{11} & \tilde{t}_{12} \\ \tilde{t}_{21} & \tilde{t}_{22} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}$$

Dann gilt $Q_1^T\mathcal{H}Q_1 = \left[\begin{array}{c|c|c|c} \begin{matrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{matrix} & w_1 & W_2 & w_3 \\ \hline 0 & A_1 & w_3^T & G_3 \\ \hline 0 & 0 & \begin{matrix} -t_{11} & -t_{21} \\ -t_{12} & -t_{22} \end{matrix} & 0 \\ \hline 0 & H_1 & w_1^T & -A_1^T \end{array}\right]$

und $\sigma\left(\begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix}\right) = \lambda, \bar{\lambda}$.
Der Induktion folgt der Rest. □

Wozu haben wir alle diese Betrachtungen über Hamiltonische und symplektische Matrizen gemacht? Wir werden diese Eigenschaften nun auf unser Zweipunkt–Randwertproblem (6.29), (6.30) anwenden. Der Trick ist der Folgende. Wir machen einen Ansatz: $\mu(t) = X(t)x(t)$, (6.30) impliziert dann $\mu(t_f) = X(t_f)x(t_f) = Mx(t_f)$, also z. B. $X(t_f) = M$. Es ergibt sich

$$\begin{bmatrix} \dot{x} \\ \dot{\mu} \end{bmatrix} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \begin{bmatrix} x \\ \mu \end{bmatrix}$$

$$\begin{bmatrix} \dot{x} \\ X(t)\dot{x} + \dot{X}(t)x \end{bmatrix} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \begin{bmatrix} x \\ X(t)x \end{bmatrix}$$

$$\dot{x} = Fx + GX(t)x$$

$$X(t)\dot{x} + \dot{X}(t)x = Hx - F^T X(t)x$$

$$\dot{X}(t)x - Hx + F^T X(t)x + X(t)Fx + X(t)GX(t)x = 0$$

Wenn $X(t)$ also die gewöhnliche Differentialgleichung

$$\dot{X}(t) = H - F^T X(t) - X(t)F - X(t)GX(t) \qquad (6.52)$$

$$X(t_f) = M \qquad (6.53)$$

erfüllt, so ist x die Lösung von

$$\dot{x} = [F + GX(t)]\,x \text{ und } \mu = X(t)x \qquad (6.54)$$

Die Gleichung (6.52), (6.53) ist eine Anfangswertaufgabe für eine Matrix–Riccati Differentialgleichung. Diese hat mit der Theorie gewöhnlicher Differentialgleichungen eine eindeutige Lösung. Damit haben wir dann auch die Lösung der Randwertaufgabe, denn da es mit diesem Ansatz eine eindeutige Lösung gibt, die die Randbedigungen erfüllt, so ist auch die Lösung der Randwertaufgabe eindeutig bestimmt.
Der andere Fall, den wir betrachten ist der Fall $t_f = \infty, M = 0$. Ansatz ist dann $\mu(t) = Xx(t)$ mit X konstant. Es ergibt sich dann analog

$$H - F^T X - XF - XGX = 0 \tag{6.55}$$

Dies ist eine algebraische Riccatigleichung und da $\lim\limits_{t\to\infty} x(t) = 0$ sein soll, so muß $x(t)$ asymptotisch stabil sein, d. h. $x(t)$ muß eine Linearkombination der Elemente des invarianten Unterraums von $\mathcal{H}$ zu den Eigenwerten mit negativem Realteil sein. Und hier kommt dann unsere Theorie der Hamiltonischen und symplektischen Matrizen ins Spiel.

In beiden Fällen erhalten wir jedenfalls aus (6.28) die lineare Rückkopplung

$$\begin{aligned} u(t) &= -R^{-1}\left(S^T x(t) + B^T \mu(t)\right) \\ &= -R^{-1}\left(S^T + B^T X\right) x(t) \end{aligned} \tag{6.56}$$

und X löst entweder die algebraische Riccatigleichung (6.54) oder die Riccati–Differentialgleichung (6.52), (6.53).

Nun ist im Gegensatz zu (6.52), (6.53) die algebraische Riccatigleichung (6.54) nicht eindeutig lösbar, aber da wir wollen, daß die Lösung $x(t)$ asymptotisch stabil ist, so muß gelten, daß die Lösung des geschlossenen Kreises

$$\dot{x} = Ax + Bu = \left[A - BR^{-1}(S^T + B^T X)\right] x \tag{6.57}$$

asymptotisch stabil ist und damit muß

$$(A - BR^{-1}S^T) - BR^{-1}B^T X \tag{6.58}$$

Eigenwerte mit negativem Realteil haben.
Nun gilt aber

$$\begin{aligned} \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \begin{bmatrix} I \\ X \end{bmatrix} &= \begin{bmatrix} I \\ X \end{bmatrix} (F + GX) \\ &= \begin{bmatrix} I \\ X \end{bmatrix} (A - BR^{-1}S^T - BR^{-1}B^T X) \end{aligned} \tag{6.59}$$

also müssen die Spalten von $\begin{bmatrix} I \\ X \end{bmatrix}$ den stabilen invarianten Unterraum aufspannen. Wie können wir diesen erhalten? Nun ganz einfach mit Hilfe der Hamiltonischen Jordan- oder Schurform, denn falls $S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \in \mathcal{S}_{2n}(\mathbb{R}), \quad S^{-1}\mathcal{H}S = \begin{bmatrix} T_{11} & T_{12} \\ 0 & -T_{11}^T \end{bmatrix}$ mit $\sigma(T_{11})$ in der linken komplexen Halbebene, so spannen die Spalten von $\begin{bmatrix} S_{11} \\ S_{21} \end{bmatrix}$ diesen Unterraum auf und wir werden zeigen, daß gilt: S_{11} ist invertierbar, $X = S_{21}S_{11}^{-1}$ ist symmetrisch, positiv semidefinite Lösung von (6.54).

Dazu zeigen wir zuerst, daß jede Lösung der algebraischen Riccatigleichung zu einem invarianten Unterraum von $\mathcal{H}$ gehört.

Satz 6.60

i) Sei X eine symmetrische Lösung der algebraischen Riccatigleichung

$$0 = H - F^T X - XF - XGX. \tag{6.61}$$

Dann spannen die Spalten von $\begin{bmatrix} I_n \\ X \end{bmatrix}$ einen n–dimensionalen invarianten Unterraum von

$$\mathcal{H} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \tag{6.62}$$

auf.

ii) Sei X symmetrisch, so daß die Spalten von $\begin{bmatrix} I_n \\ X \end{bmatrix}$ einen invarianten Unterraum von $\mathcal{H}$ wie in (6.62) aufspannen. Dann ist X eine Lösung von (6.61).

Beweis: $\begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \begin{bmatrix} I_n \\ X \end{bmatrix} = \begin{bmatrix} F + GX \\ H - F^T X \end{bmatrix} = \begin{bmatrix} I_n \\ X \end{bmatrix} Z$
$\iff H - F^T X = XZ = X(F + GX) \iff X$ erfüllt (6.61). □

Die Frage die vor allem bleibt ist, ob es so einen invarianten Unterraum der Form $\begin{bmatrix} I_n \\ X \end{bmatrix}$ mit X symmetrisch gibt. Dann erhalten wir mit Satz 6.60 eine Beziehung zwischen den Lösungen der algebraischen Riccatigleichung (6.61) und den invarianten Unterräumen von $\mathcal{H}$.

Satz 6.63 *Sei $\mathcal{H} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \in \mathcal{H}_{2n}(\mathbb{R})$ wie in (6.29).*
$F = A - BR^{-1}S^T, G = -BR^{-1}B^T, H = -(Q - SR^{-1}S^T)$, (A, B) stabilisierbar, $\mathcal{R}$ positiv definit. Sei $S \in \mathcal{S}_{2n}(\mathbb{R}), S^{-1}\mathcal{H}S = \begin{bmatrix} T_{11} & T_{12} \\ 0 & -T_{11}^T \end{bmatrix}, T_{11}$ stabil. Dann ist S_{11} invertierbar, $S_{21}S_{11}^{-1}$ symmetrisch, $X = S_{21}S_{11}^{-1}$ positiv semidefinite Lösung von
$0 = H - XF - F^T X - XGX$.

Beweis:

$$\begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} \\ 0 & -T_{11}^T \end{bmatrix}$$

$$\begin{aligned}
&\Longrightarrow \left\{ \begin{array}{rcl} FS_{11} + GS_{21} &=& S_{11}T_{11} \\ HS_{11} - F^T S_{21} &=& S_{21}T_{11} \Longrightarrow S_{11}^T H - S_{21}^T F = T_{11}^T S_{21}^T \end{array} \right. \\
&\Longrightarrow \left\{ \begin{array}{rcl} S_{21}^T F S_{11} + S_{21}^T G S_{21} &=& S_{21}^T S_{11} T_{11} \\ S_{11}^T H S_{11} - S_{21}^T F S_{11} &=& T_{11}^T S_{21}^T S_{11} \end{array} \right. \\
&\Longrightarrow S_{21}^T G S_{21} + S_{11}^T H S_{11} = S_{21}^T S_{11} T_{11} + T_{11}^T S_{21}^T S_{11} \quad \text{Lyapunov–Gleichung}
\end{aligned}$$

$$\begin{aligned}
S \in \mathcal{S}_{2n}(\mathbb{R}) \quad &\Longrightarrow \quad [I_n,\ 0]\, S^T J S \begin{bmatrix} I_n \\ 0 \end{bmatrix} = 0 \\
&\iff \quad S_{21}^T S_{11} = S_{11}^T S_{21}
\end{aligned}$$

T_{11} stabil, $S_{21}^T G S_{21} + S_{11}^T H S_{11}$ negativ semidefinit $\Longrightarrow S_{21}^T S_{11}$ positiv semidefinit.(Satz von Lyapunov, siehe Kapitel 7. Wir werden dieses Resultat noch beweisen im Zusammenhang mit dem Newtonverfahren für die algebraische Riccatigleichung)
Nun müssen wir noch zeigen, daß S_{11} invertierbar ist.

Sei $w \neq 0$, so daß $S_{11}w = 0$

$$\begin{aligned}
&\Longrightarrow \quad w^T S_{21}^T G S_{21} w + w^T S_{11}^T H S_{11} w = 0 \\
&\Longrightarrow \quad G S_{21} w = 0 \text{ aus der Lyapunovgleichung} \\
&\Longrightarrow \quad B^T S_{21} w = 0, \text{ weil } R \text{ positiv definit ist.}
\end{aligned}$$

Aber wegen $FS_{11} + GS_{21} = S_{11}T_{11}$ gilt dann

$$S_{11}(T_{11}w) = 0 \Longrightarrow T_{11}w \in Ker(S_{11})$$

D.h. $Ker(S_{11})$ ist ein invarianter Unterraum von T_{11}. Da in jedem invarianten Unterraum einer Matrix mindestens ein Eigenwert λ und zugehöriger Eigenvektor z liegt,

$$\Longrightarrow \exists z \in Ker(S_{11}) z \neq 0, \text{ so daß } T_{11}z = \lambda z$$

für ein $\lambda \in \sigma(T_{11})$

$$\begin{aligned}
\Longrightarrow H S_{11} z - F^T S_{21} z &= S_{21} T_{11} z \\
0 - F^T S_{21} z &= \lambda S_{21} z \\
\Longrightarrow (F^T + \lambda I) S_{21} z &= 0
\end{aligned}$$

Es ist $S_{21}z \neq 0$, weil $\binom{S_{11}}{S_{21}}$ vollen Rang hat. Also ist $S_{21}z$ Eigenvektor von F^T zum Eigenwert $-\lambda$, $Re(-\lambda) > 0$.
Aber da $z \in$ Ker (S_{11}) ist, folgt

$$G S_{21} z = 0 \Longrightarrow B^T S_{21} z = 0$$

$$\Longrightarrow \text{Rang} \begin{bmatrix} F^T - (-\lambda) I \\ B^T \end{bmatrix} < n \quad \text{für } \lambda \text{ mit } \mathrm{Re}(-\lambda) > 0$$

$\Longrightarrow (F, B) = (A - BR^{-1}S^T, B)$ nicht stabilisierbar
$\Longrightarrow (A, B)$ nicht stabilisierbar. Widerspruch! □

Wir fassen nun alle Ergebnisse dieses Abschnitts zusammen.

Satz 6.64 *Betrachte das optimale Steuerungsproblem (6.1), (6.2).*

i) Falls $t_f < \infty$, so existiert die Optimalsteuerung und sie ist eine lineare Zustandsrückkopplung der Form:

$$u(t) = -R^{-1}\left(S^T + B^T X(t)\right) x(t) \tag{6.65}$$

wobei $X(t)$ die eindeutige Lösung der Matrix Riccati–Differentialgleichung

$$\begin{aligned}
\dot{X}(t) &= -(Q - SR^{-1}S^T) - (A - BR^{-1}S^T)^T X(t) \\
&\quad -X(t)(A - BR^{-1}S^T) + X(t)BR^{-1}B^T X(t) \\
X(t_f) &= M
\end{aligned} \tag{6.66}$$

ist.

ii) Sei (A, B) stabilisierbar, $t_f = \infty, M = 0$, $\mathcal{R}$ positiv definit.
Dann ist die eindeutige Lösung von (6.1) gegeben durch die Rückkopplungssteuerung

$$u(t) = -R^{-1}(S^T + B^T X)x(t) \tag{6.67}$$

wobei X die eindeutige positiv semidefinite Lösung der algebraischen Riccati–Gleichung

$$0 = (Q - SR^{-1}S^T) + (A - BR^{-1}S^T)^T X + X(A - BR^{-1}S^T) - XBR^{-1}B^T X \tag{6.68}$$

ist. In diesem Fall gilt für den geschlossenen Kreis

$$\lim_{t\to\infty} x(t) = 0. \tag{6.69}$$

Beweis: Den Großteil des Beweises haben wir bereits vorweggenommen.

i) Aus der Theorie der gewöhnlichen Differentialgleichungen folgt, daß (6.66) eine eindeutige Lösung hat. Also erhalten wir eine Lösung der Form (6.65) und aus Satz 6.21 folgt, daß wir eine Minimallösung haben. Es bleibt noch die Frage der Eindeutigkeit der Lösung von (6.29),(6.30). Die Lösung der 2 Punkt Randwertaufgabe ist

$$\begin{bmatrix} x(t) \\ \mu(t) \end{bmatrix} = e^{\mathcal{H}(t-t_0)} \begin{bmatrix} q_0 \\ v_0 \end{bmatrix} \tag{6.70}$$

Da $Z(t) := e^{\mathcal{H}(t-t_0)}$ für alle t nichtsingulär ist folgt, daß bei einer Aufteilung

$$Z(t) = [Z_1(t),\ Z_2(t)] \ \text{ mit } Z_i(t) \in \mathbb{C}^{2n,n} \quad i = 1, 2 \tag{6.71}$$

$Z_i(t)$ vollen Rang hat, also haben

$$x(t_0) = Z_1(t_0)q_0 = x^0 \tag{6.72}$$

und

$$\mu(t_f) = Mx(t_f) = Z_2(t_f)v_0 = MZ_1(t_f)q_0 \tag{6.73}$$

falls Lösungen existieren (und das wissen wir schon), eindeutige Lösungen. Damit ist i) bewiesen.

ii) Aus der Stabilisierbarkeit von (A, B) folgt mit $\mathcal{R}$ positiv definit nach Satz 6.44, daß

$$\mathcal{H} = \begin{bmatrix} A - BR^{-1}S^T & -BR^{-1}B^T \\ -(Q - SR^{-1}S^T) & -(A - BR^{-1}S^T)^T \end{bmatrix}$$

keine Eigenwerte mit Realteil 0 hat.

Wir haben bereits gezeigt, daß die Rückkopplungssteuerung (6.67) mit X Lösung von (6.68), zu einer Lösung von (6.29), (6.30) für $t_f = \infty, M = 0$ führt, falls $\lim_{t\to\infty} \mu(t) = 0$ ist. Es bleibt die Frage der Eindeutigkeit.

Da $\mu(t) = Xx(t)$ für X konstant so muß auch $\lim_{t\to\infty} x(t) = 0$ gelten. Also muß

$$A - BR^{-1}(S^T + B^T X) = F + GX$$

stabil sein. Wenn $\begin{bmatrix} I \\ X \end{bmatrix}$ den stabilen invarianten Unterraum aufspannt (und dieser ist eindeutig) so folgt, daß

$$F + GX = T_{11} \text{ stabil ist.}$$

$\square$

Bemerkung 6.74 *Aus den vorigen Überlegungen folgt sofort, daß wir falls nur (6.1) gegeben ist und wir das System stabilisieren wollen, wir einfach eine Riccatigleichung vorgeben können z. B.* $S = 0, Q = I, R = I$, *und diese lösen, dann ist*

$$\begin{aligned} u(t) &= -R^{-1}(S^T + B^T X)x(t) \\ &= -(B^T X)x(t) \end{aligned} \tag{6.75}$$

stabilisierendes feedback. X *löst die Riccatigleichung*

$$0 = I + A^T X + XA - XBB^T X. \tag{6.76}$$

Dies liefert also eine andere Methode, um das System zu stabilisieren.

Kapitel 7

Numerische Lösung von Riccatigleichungen

In diesem Kapitel betrachten wir numerische Methoden für algebraische und differentielle Riccatigleichungen.

Hier gibt es nun verschiedene Ansätze. Wir haben gesehen (Satz 6.63), daß wir die Lösung der algebraischen Riccatigleichung

$$0 = H + F^T X + XF - XGX \tag{7.1}$$

über die Berechnung des stabilen invarianten Unterraums der Hamiltonischen Matrix

$$\mathcal{H} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \tag{7.2}$$

bestimmen können. Wir werden aber zuerst den direkten Zugang über (7.1) betrachten. Dies ist eine quadratische Matrixgleichung und man kann sie mit dem Newton–Verfahren lösen. Das Problem ist, die richtige Lösung zu erwischen.

7.1 Das Newton Verfahren

Die Idee ist nun ganz einfach: Angenommen X_0 ist eine Startnäherung, dann schauen wir uns an, welche Gleichung für die Differenz $P = X - X_0$ gilt:

Satz 7.3 *Sei X eine symmetrische Lösung von (7.1) und X_0 eine symmetrische Näherungslösung. Sei $P = X - X_0$,*
$\tilde{F} = F - GX_0$,
$\tilde{H} = H + F^T X_0 + X_0 F - X_0 G X_0$.
Dann erfüllt P die algebraische Riccatigleichung

$$0 = \tilde{H} + P\tilde{F} + \tilde{F}^T P - PGP. \tag{7.4}$$

Beweis:

$$\begin{aligned} 0 &= F^T(P + X_0) + (P + X_0)F + H - (P + X_0)G(P + X_0) \\ &= (F^T - X_0 G)P + P(F - GX_0) - PGP \\ &\quad + F^T X_0 + X_0 F + H - X_0 G X_0 \\ &= \tilde{F}^T P + P\tilde{F} - PGP + \tilde{H}. \end{aligned}$$

□

Das heißt, daß der Defekt der Lösung eine Gleichung des gleichen Typs erfüllt. Dies kann sehr gut zur Nachiteration verwendet werden, falls man eine ungenaue Lösung berechnet hat.

Für das Newton Verfahren nimmt man an, daß X_0 gute Näherung und damit $P = X - X_0$ klein ist. Dann können wir quadratische Terme in P vernachlässigen und erhalten aus (7.4)

$$0 \approx \tilde{H} + P\tilde{F} + \tilde{F}^T P \tag{7.5}$$

Dies ist eine Lyapunovgleichung, die wir mit dem Bartels–Stewart Algorithmus (Siehe Golub/Van Loan) lösen können.

Wenn wir P aus (7.5) ausgerechnet haben, setzen wir

$$X_1 := X_0 + P \tag{7.6}$$

$\Longrightarrow X_1$ löst die Gleichung

$$\begin{aligned}
0 &= \tilde{H} + (X_1 - X_0)\tilde{F} + \tilde{F}^T(X_1 - X_0) \\
&= H + X_0F + F^TX_0 - X_0GX_0 \\
&\quad X_1(F - GX_0) + (F - GX_0)^TX_1 \\
&\quad -X_0F + X_0GX_0 - F^TX_0 + X_0GX_0 \\
&= H + X_1(F - GX_0) + (F - GX_0)^TX_1 + X_0GX_0 \\
&= H + X_0GX_0 + X_1\tilde{F} + \tilde{F}^TX_1
\end{aligned}$$

Algorithmus 7.7 *Newton–Verfahren für (7.1)*
Input: $F, G, H \in \mathbb{R}^{n,n}, G = G^T, H = H^T$
Startnäherung $X_0 = X_0^T \in \mathbb{R}^{n,n}$.

Output: Lösung X *von (7.1)*
Setze $H_0 = H, F_0 = F$
FOR $i = 1, 2, \ldots$ *UNTIL SATISFIED*

Setze $F_i := F - GX_{i-1}$
$H_i := H + X_{i-1}GX_{i-1}$
Löse

$$0 = H_i + X_iF_i + F_i^TX_i \tag{7.8}$$

(mit Bartels/Stewart)
END FOR
END.

Wir wollen Konvergenz gegen eine positiv semidefinite Lösung zeigen. Bevor wir den Konvergenzsatz betrachten, machen wir wie versprochen den Satz von Lyapunov.

Satz 7.9

a) Die eindeutige Lösung von

$$A^TX + XA = C \tag{7.10}$$

für A stabil, ist

$$X = -\int_0^\infty e^{A^Tt}Ce^{At}\,dt \tag{7.11}$$

b) Falls A stabil und C positiv (semi-) definit ist, so hat (7.10) eine eindeutige negativ (semi-) definite Lösung X.

c) Falls X Lösung von (7.10) und C positiv definit ist, so ist A stabil.

Beweis:

a) Betrachte Dgl.

$$\dot{Z} = A^T Z + ZA, \quad Z(0) = C$$

Lösung: $Z(t) = e^{A^T t} C e^{At}$ durch nachrechnen.
Da A stabil $\Longrightarrow \lim_{t\to\infty} e^{A^T t} C e^{At} = 0$, $\int_0^\infty \|e^{A^T t} C e^{At}\| \, dt$ beschränkt.

$$\begin{aligned}
\Longrightarrow Z(\infty) - Z(0) &= \int_0^\infty \dot{Z}(t)\, dt \\
&= A^T \left(\int_0^\infty Z(t)\, dt \right) + \left(\int_0^\infty Z(t)\, dt \right) A
\end{aligned}$$

$$\begin{aligned}
\Longrightarrow C &= \left(- \int_0^\infty Z(t)\, dt \right) A + A^T \left(- \int_0^\infty Z(t)\, dt \right) \\
\Longrightarrow X &= - \int_0^\infty e^{A^T t} C e^{At}\, dt
\end{aligned}$$

Die Eindeutigkeit folgt aus der Tatsache, das

$$A^T X + XA = C \Longleftrightarrow \left[(I \otimes A^T) + (A^T \otimes I) \right] \mathrm{vec}(X) = \mathrm{vec}(C) \tag{7.12}$$

wobei

$$(B \otimes A) \;=\; \begin{bmatrix} b_{11}A & \cdots & b_{1n}A \\ \vdots & & \vdots \\ b_{n1}A & \cdots & b_{nn}A \end{bmatrix} \quad \text{und} \quad \left[(I \otimes A^T) + (A^T \otimes I) \right]$$

hat Eigenwerte $\lambda_j + \lambda_l$, wobei $\lambda_j, \lambda_l \in \sigma(A)$
$\Longrightarrow \left[(I \otimes A^T) + (A^T \otimes I) \right]$ nicht singulär, weil A stabil, d.h. $\lambda_j + \lambda_l \neq 0$.

b) Aus a) folgt

$$X = - \int_0^\infty e^{A^T t} C e^{At}\, dt$$

$$y \neq 0 \Longrightarrow y^* X y = - \int_0^\infty y^* e^{A^T t} C e^{At} y \, dt \begin{cases} < 0, & \text{falls } C > 0 \\ \geq 0, & \text{falls } C \leq 0 \end{cases}$$

Dabei haben wir im Falle der Definitheit von C ausgenutzt, daß $e^{At} y \neq 0$, weil e^{At} nicht singulär ist.

c) Sei $Ax = \lambda x \Longrightarrow x^* A^* = \bar{\lambda} x$
(7.10) $\Longrightarrow 0 < x^* C x = x^*(A^T X + XA)x = x^* X x(\bar{\lambda} + \lambda)$
$X < 0 \Longrightarrow \bar{\lambda} + \lambda < 0. \Longrightarrow$ Beh.

$\square$

Damit können wir nun die Konvergenz des Newtonverfahrens beweisen.

Satz 7.13 *Seien G, H positiv semidefinit und X_0 so gewählt, daß $F - GX_0$ stabil ist. Dann gilt für die Folge der X_i, die mit Algorithmus 7.7 berechnet wurden:*

a) $0 \le X \le X_{j+1} \le X_j \le \cdots \le X_0$ *wobei X die positiv semidefinite Lösung von (7.1) ist. (Hier: $A \ge B \Longleftrightarrow A - B$ positiv semidefinit)*

b) $A - GX_j$ *ist stabil für alle j.*

c) Es gibt eine Konstante γ, so daß

$$\|X_j - X\|_2 \le \gamma \|X_{j-1} - X\|_2^2 \tag{7.14}$$

d.h. wir haben quadratische Konvergenz.

Beweis: (7.8) ist äquivalent zu

$$X_j(F - GX_{j-1}) + (F - GX_{j-1})^T X_j = -H - X_{j-1}GX_{j-1} \le 0 \tag{7.15}$$

Außerdem gilt

$$X_j(F - GX_j) + (F - GX_j)^T X_j = -H - X_jGX_j - (X_j - X_{j-1})G(x_j - X_{j-1}) \tag{7.16}$$

Also zusammen

$$\begin{aligned} &(F - GX_j)^T(X_j - X_{j+1}) + (X_j - X_{j+1})(F - GX_{j-1}) \\ &\quad = -(X_{j-1} - X_j)G(X_{j-1} - X_j) \le 0 \quad j = 1, 2, \ldots \end{aligned} \tag{7.17}$$

Mit Satz 7.9 folgt aus (7.15): $X_j \ge 0$.
$\Longrightarrow F - GX_j$ stabil in (7.16). $\Longrightarrow X_j \ge X_{j+1}$ aus (7.17).
Per Induktion folgt die Behauptung.

Aus der Tatsache, daß die Folge der X_j monoton fallend und nach unten beschränkt ist, folgt mit Bolzano/Weierstraß, daß es einen Grenzwert $X \ge 0$ gibt.

Aus (7.16) folgt im Limes

$$(F - GX)^T X + X(F - GX) = -H - XGX \tag{7.18}$$

$\Longrightarrow X$ löst (7.1) , $X \ge 0$.
Aus (7.18) folgt

$$(F - GX_j)^T X + X(F - GX_j) = -H - XGX + (X - X_j)GX + XG(X - X_j) \tag{7.19}$$

Differenz (7.19) und (7.16) ergibt

$$\begin{aligned} & (F - GX_j)^T(X - X_j) + (X - X_j)(F - GX_j) \\ = \; & -XGX + (X - X_j)GX + XG(X - X_j) + X_jGX_j \\ & +(X_j - X_{j-1})G(X_j - X_{j-1}) \\ = \; & (X - X_j)G(X - X_j) + (X_j - X_{j-1})G(X_j - X_{j-1}) \end{aligned}$$

$$\begin{aligned} \Longrightarrow \quad & (F - GX)^T(X - X_j) + (X - X_j)(F - GX) \\ = \; & (F - GX_j)^T(X - X_j) + (X - X_j)(F - GX_j) \\ & -(G(X - X_j))^T(X - X_j) - (X - X_j)G(X - X_j) \\ = \; & -(X - X_j)G(X - X_j) + (X_j - X_{j-1})G(X_j - X_{j-1}) \end{aligned} \tag{7.20}$$

Somit läßt sich $X - X_j$ schreiben als

$$
\begin{aligned}
X - X_j &= \int_0^\infty e^{t(F-GX)^T} \left(\underbrace{-(X - X_j)G(X - X_j)}_{\leq 0} + (X_j - X_{j-1})G(X_j - X_{j-1}) \right) e^{t(F-GX)}\, dt \\
&\leq \int_0^\infty e^{t(F-GX)^T} (X_j - X_{j-1})G(X_j - X_{j-1}) e^{t(F-GX)}\, dt
\end{aligned}
$$

$$
\begin{aligned}
\Longrightarrow \|X_j - X\|_2 &\leq \int_0^\infty \|e^{t(F-GX)^T}(X_j - X_{j-1})G(X_j - X_{j-1})e^{t(F-GX)}\|_2\, dt \\
&\leq \|X_j - X_{j-1}\|_2^2 \int_0^\infty \|e^{t(F-GX)^T}\|_2 \|G\|_2 \|e^{t(F-GX)}\|_2 dt \\
&=: \|X_j - X_{j-1}\|_2^2 \gamma \\
&\leq \gamma \|X - X_{j-1}\|_2^2
\end{aligned}
$$

da $X_{j-1} - X \geq X_{j-1} - X_j \geq 0$. □

Um die Lyapunov–Gleichung, die in jedem Schritt vorkommt, zu lösen , bringen wir in jedem Schritt mittels des QR–Algorithmus

$$Q_i^T F_i Q_i = T_i \tag{7.21}$$

auf Schurform (reelle Schurform), bilden

$$Q_i^T H_i Q_i = \underbrace{Q_i^T X_i Q_i}_{Y} \underbrace{Q_i^T F_i Q_i}_{T_i} + Q_i^T F_i^T Q_i Y \tag{7.22}$$

und lösen mit Bartels/Stewart.

Das große Problem beim Newtonverfahren ist die stabile Startnäherung, denn wir müssen ein X_0 finden, so daß $F - GX_0$ stabil ist. Dies ist im Prinzip nur durch Lösen einer Riccati–Gleichung möglich, daher wird das Newtonverfahren bevorzugt als Nachiterationsverfahren verwendet. Man kann aber für die Nachiteration auch direkt Satz 7.3 verwenden, indem man den Löser für die Riccatigleichung immer wieder auf die Defektgleichung (7.4) anwendet und jeweils den stabilen invarianten Unterraum berechnet. Wir haben nämlich die folgenden Sätze:

Lemma 7.23 *Betrachte die Riccatigleichung (7.1) mit* $F = A - BR^{-1}B^T, G = -BR^{-1}B^T, H = -(Q - SR^{-1}S^T)$.
Sei (A, B) *stabilisierbar und* X_0 *eine symmetrische Näherungslösung von (7.1). Dann ist* $(\tilde{A}, B)$ *stabilisierbar wobei*

$$\tilde{A} = A - BR^{-1}B^T X_0 - BR^{-1}S^T$$

Beweis: Dies ist eine Rückkopplung. Also folgt die Behauptung mit der Invarianz von Stabilisierbarkeit unter Feedback. □

Satz 7.24 *Wenn die Spalten von* $\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \in \mathbb{C}^{2n,n}$ *den invarianten Unterraum zu den stabilen Eigenwerten von*

$$\begin{bmatrix} \tilde{F} & G \\ \tilde{H} & -\tilde{F}^T \end{bmatrix} = \begin{bmatrix} A - BR^{-1}B^T X_0 - BR^{-1}S^T & G \\ \tilde{H} & -\tilde{F}^T \end{bmatrix}$$

$$\begin{aligned} \textit{mit } \tilde{H} &= H + F^T X_0 + X_0 F - X_0 G X_0 \\ &= Q - SR^{-1}S^T + A^T X_0 - S^T R^{-1} B^T X_0 \\ &\quad + X_0 A - X_0 B R^{-1} S^T - X_0 G X_0, \end{aligned}$$

aufspannen und $P = Z_2 Z_1^{-1}$, *so ist* $X = X_0 + P$ *die eindeutige positiv semidefinite Lösung von (7.1).*

Beweis:

$$P = Z_2 Z_1^{-1}$$

$$\begin{bmatrix} \tilde{F} & G \\ \tilde{H} & -\tilde{F}^T \end{bmatrix} \begin{bmatrix} I \\ P \end{bmatrix} = \begin{bmatrix} I \\ P \end{bmatrix} Z_1 Z Z_1^{-1}$$

$Z_1 Z Z_1^{-1}$ ist stabil wenn Z stabil ist.
Nun ist aber

$$\tilde{F} + GP = Z_1 Z Z_1^{-1}$$

$$\begin{aligned} \Longrightarrow \tilde{F} + GP &= A - BR^{-1}B^T X_0 - BR^{-1}S^T - BR^{-1}B^T P \\ &= A - BR^{-1}B^T(X_0 + P) - BR^{-1}S^T. \end{aligned}$$

stabil. Die Eindeutigkeit folgt damit aus Satz über die Riccatigleichung. □

Wir können damit den folgenden Defekt–Korrekturalgorithmus durchführen.

Algorithmus 7.25 *Defektkorrektur für die algebraische Riccatigleichung*

Input: A, B, Q, R, S, *Toleranz tol für Residuum*

Output: *Lösung der Riccati-Gleichung (7.1) mit* $F = A - BR^{-1}S^T, H = -(Q - SR^{-1}S^T)$, $G = -BR^{-1}B^T$ *sowie Fehlerabschätzung für die Güte der Lösung.*

(1) Verwende irgendeine Lösungsmethode, um die stabilisierende Lösung $\tilde{X}$ *von (7.1) zu berechnen.*

(2) Setze $P = \tilde{X}, \tilde{X} = 0$

(3) WHILE $\|P\| > Tol$

Setze $\tilde{X} := \tilde{X} + P$,

$H := H + F^T\tilde{X} + \tilde{X}F - \tilde{X}G\tilde{X}$,

$F := A - BR^{-1}(B^T\tilde{X} + S^T)$,

verwende irgendeine Lösungsmethode, um die stabilisierende Lösung von

$$H + F^T P + PF - PGP = 0 \tag{7.26}$$

zu berechnen.

END WHILE

END.

In manchen Fällen ist es notwendig, das neue Residuum H mit erhöhter Genauigkeit zu berechnen (wie bei Gauss Nachiteration).

Wenn die in (1) verwendete Methode die ersten Stellen korrekt berechnet, so ist nach 1–2 Schritten die Lösung korrekt bis zur Maschinengenauigkeit.

Jede Methode zur Lösung von algebraischen Riccatigleichungen sollte einen Schritt Defekt–Korrektur nach sich ziehen.

Zur Lösung von (7.26) wird bevorzugt das Newton–Verfahren verwendet.

7.2 Die Signum–Funktions–Methode

Sei $A \in \mathbb{R}^{n,n}(\mathbb{C}^{n,n})$ $\quad S = P^{-1}AP$ die Jordan-Form von $A, S = D+N, D = \text{diag}\,(d_1, \cdots, d_n), N$ nilpotent. Angenommen, A habe keine Eigenwerte mit Realteil 0, dann definiere

$$\text{sign}\ (A) = P\Sigma P^{-1} \tag{7.27}$$

wobei

$$\Sigma = \ \text{diag}\ (\sigma_1, \cdots, \sigma_n), \sigma_i = \text{sign (Realteil } d_i) \tag{7.28}$$

Für Matrizen mit rein imaginären Eigenwerten ist $\text{sign}(A)$ nicht definiert.

Es gilt

$$\text{sign}\ (A) = \lim_{k\to\infty} Z_k, \tag{7.29}$$

wobei

$$Z_0 = A, \quad Z_{k+1} = \frac{Z_k + Z_k^{-1}}{2} \tag{7.30}$$

Dies ist ein Newtonverfahren für $Z^2 - I = 0$. Sei nun X die stabilisierende Lösung der Riccatigleichung (7.1). Dann gilt

$$\mathcal{H} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} = \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} \begin{bmatrix} F + GX & G \\ 0 & -(F+GX)^T \end{bmatrix} \begin{bmatrix} I & 0 \\ X & I \end{bmatrix}^{-1} \tag{7.31}$$

Da X die stabilisierende Lösung ist, gilt $F - GX$ stabil

$$\Longrightarrow \ \text{sign}\ (\mathcal{H}) =: \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} \begin{bmatrix} -I & Z \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ -X & I \end{bmatrix} \tag{7.32}$$

$\Longrightarrow$

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} -I \\ -X \end{bmatrix} \tag{7.33}$$

$\Longleftrightarrow$

$$\begin{bmatrix} W_{12} \\ W_{22} \end{bmatrix} X + \begin{bmatrix} W_{11} \\ W_{21} \end{bmatrix} + \begin{bmatrix} I \\ X \end{bmatrix} = 0 \tag{7.34}$$

$\Longleftrightarrow$

$$\begin{bmatrix} W_{12} \\ W_{22} + I \end{bmatrix} X = - \begin{bmatrix} W_{11} + I \\ W_{21} \end{bmatrix} \tag{7.35}$$

Damit löst X dieses überbestimmte System von Gleichungen und dies kann mit Hilfe der QR–Zerlegung gelöst werden.
Wir erhalten dann den folgenden Algorithmus

Algorithmus 7.36 *Sign–Funktions–Methode*

Input: $F, G, H \in \mathbb{R}^{n,n}$ *so daß*
$\begin{bmatrix} F & G \\ H & -F^T \end{bmatrix}$ *keine rein imaginären Eigenwerte hat.*

Output: *Lösung der Riccatigleichung 7.1 und Fehlerabschätzung.*

(1) $Z_0 = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix}$

(2) FOR $k = 0, 1, 2, \ldots$ UNTIL Z_k *konvergent*

$$Z_k := |\det Z_k|^{-\frac{1}{2n}} Z_k$$
$$Z_{k+1} := Z_k - \tfrac{1}{2}\left(Z_k - (JZ_k)^{-1}J\right)$$

END FOR

(3) Mit $Z_\infty = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$, $W_{ij} \in \mathbb{R}^{n,n}$ *löse*

$$\begin{bmatrix} W_{12} \\ W_{22} + I \end{bmatrix} X = - \begin{bmatrix} W_{11} + I \\ W_{21} \end{bmatrix}$$

(4) Verwende Defekt–Korrektur Algorithmus 7.25 mit diesem Startwert X zur Berechnung von genauerem X und Fehlerschätzer P.

END.

Kosten: $\mathcal{O}(n^3 k_{\max})+$ *Defektkorrektur,* $k_{\max} = \sharp$ *Iterationen.*

Die Determinante erhält man als Nebenprodukt bei der Berechnung der Inversen von JZ_k bzw. der Lösung des Gleichungssystems mit JZ_k über die QR oder LR Zerlegung. Beachte, daß

$$JZ_k \text{ symmetrisch indefinit ist.}$$

Die Sign–Funktionsmethode als solche ist NICHT numerisch stabil, es gibt große Probleme, wenn Eigenwerte nahe an der imaginären Achse liegen, und natürlich auch wenn Z_k fast singulär (d.h. reelle Eigenwerte nahe an der imaginären Achse).

Ohne Defektkorrektur nicht praktikabel.
Wird sehr viel verwendet.

Newtonverfahren und Sign–Funktionsmethode sind anders als die anderen Methoden nicht direkt über die Berechnung des invarianten Unterraums der Hamiltonischen Matrix erklärt.

7.3 Laubs Schur–Methode

Dies ist der am häufigsten verwendete Ansatz:
In der letzten Übung haben wir gesehen, daß wenn wir die reelle Schurform der Hamiltonischen Matrix $\mathcal{H} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix}$ berechnen, d.h.

$$\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}^T \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \tag{7.37}$$

und die Eigenwerte von T_{11} haben alle negativen Realteil, so ist

$$X = Q_{21} Q_{11}^{-1} \tag{7.38}$$

die gesuchte Lösung der algebraischen Riccatigleichung (7.1).
Wir können darauf also den folgenden Algorithmus aufbauen.

Algorithmus 7.39 *Laub's Schurmethode*

Input: *$\mathcal{H}$ wie in (7.2), $F, G, H \in \mathbb{R}^{n,n}, G = G^T, H = H^T, \mathcal{H}$ habe keine Eigenwerte mit Realteil* 0.

Output: *Stabilisierende positiv semidefinite Lösung X der algebraischen Riccatigleichung 7.1.*

1. Schritt: *Bestimme $Q \in \mathbb{R}^{2n,2n}$ orthogonal, so daß*

$$Q^T \mathcal{H} Q = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}, \quad Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$$

$T_{11}, T_{22} \in \mathbb{R}^{n,n}$ quasi obere $\triangle$–Matrix, $Re(\lambda) < 0$ für alle $\lambda \in \sigma(T_{11}), Q_{ij} \in \mathbb{R}^{n,n}$. Dies kann mit Hilfe des QR–Algorithmus gemacht werden, gefolgt von anschliessender Umordnung der Eigenwerte mit Hilfe des Bartels/Stewart Algorithmus.

2. Schritt: *Löse $Q_{11} X = Q_{21}$ mit Hilfe der QR– (oder LR–) Zerlegung (mit Pivotisierung).*

3. Schritt: *Verwende Defekt–Korrektur Algorithmus 7.25 mit diesem Startwert X zur Berechnung von Fehlerschranke und eventuell verbessertem X.*

Kosten: *$\mathcal{O}(n^3)$ plus Defekt Korrektur.*

Numerisch stabil aber nicht strukturerhaltend, denn $Q^T H Q$ ist nicht Hamiltonisch und die Lösung ist nur theoretisch symmetrisch. Ohne ein strukturangepaßtes Defektkorrekturverfahren und viele kleine Details ist das Verfahren nicht so gut. Dies ist das am meisten verwendete Verfahren, implementiert in Paketen wie NAG, MATLAB–Control toolbox.

7.4 Der Multishift–Algorithmus(OSMARE)

Es gibt noch einen weiteren Ansatz. Dieser beruht auf der Tatsache, daß wir ja eigentlich nur den invarianten Unterraum berechnen wollen. Auf dem letzten Übungsblatt haben wir gesehen, daß für $\mathcal{H} \in \mathcal{H}_{2n}(\mathbb{R}), \sigma(\mathcal{H}) = \{\lambda_1, \ldots, \lambda_{2n}\}$ und $\lambda_1, \ldots, \lambda_n$ die instabilen Eigenwerte

$$x = (\mathcal{H} - \lambda_1 I)(\mathcal{H} - \lambda_2 I) \cdots (\mathcal{H} - \lambda_n I) e_1 \tag{7.40}$$

in dem n–dimensionalen Lagrange–Unterraum zu den stabilen Eigenwerten $\lambda_{n+1}, \ldots, \lambda_{2n}$ liegt. Wir können also einen QR–Algorithmus mit n gleichzeitigen Shifts wie folgt konstruieren.

Berechne die Eigenwerte von $\mathcal{H}$ zum Beispiel mit QR, oder der $SQRED$–Methode aus den Ergänzungen, bilde x wie in (7.40) und berechne dann $Q \in \mathcal{US}_{2n}(\mathbb{R})$ so daß $Q^T x = \alpha e_1$. Dann bilde $Q_1^T \mathcal{H} Q_1$ und bringe dies in die Paige/Van Loan–Form mit Algorithmus 7.46 oder Algorithmus 4.7 aus den Ergänzungen, d.h. berechne $Q_2 \in \mathcal{US}_{2n}(\mathbb{R})$, so daß

$$Q_2^T Q_1^T \mathcal{H} Q_1 Q_2 = \begin{bmatrix} T_{11} & T_{12} \\ 0 & -T_{11}^* \end{bmatrix},$$

und T_{11} hat nur Eigenwerte mit negativem Realteil. Damit haben wir folgenden Algorithmus Ammar/Benner/Mehrmann.

Algorithmus 7.41 *Multishift Algorithmus*

Input: *$\mathcal{H} \in \mathcal{H}_{2n}(\mathbb{R}), \mathcal{H}$ habe nur Eigenwerte mit Realteil $\neq 0$. Toleranz tol*

Output: $Q \in \mathcal{US}_{2n}(\mathbb{R})$, *so daß*

$$Q^T \mathcal{H} Q = \begin{bmatrix} F_{11} & G_{11} \\ 0 & -F_{11}^T \end{bmatrix}$$

und F_{11} *stabil ist.*

Setze $Q = I_{2n}$

(1) Berechne die Eigenwerte von $\mathcal{H}$ *(mit QR oder SQRED) und wähle die instabilen aus als*

$$\lambda_1, \ldots, \lambda_n.$$

(2) Bilde $x = \alpha(\mathcal{H} - \lambda_1 I) \cdots (\mathcal{H} - \lambda_n I) e_1$ *(nach jeder Multiplikation skalieren, reell rechnen).*

(3) Berechne $Q_1 \in \mathcal{US}_{2n}(\mathbb{R})$, *so daß*

$$Q_1^T x = \pm \|x\|_2 e_1.$$

(4) Transformiere $Q_1^T \mathcal{H} Q_1$ *auf die Form*

$$\begin{bmatrix} * & \cdots & \cdots & * & * & \cdots & \cdots & * \\ * & \ddots & & \vdots & \vdots & & & \vdots \\ & \ddots & \ddots & \vdots & \vdots & & & \vdots \\ & & * & * & * & \cdots & \cdots & * \\ * & & & & * & * & & \\ & \ddots & & & \vdots & \ddots & \ddots & \\ & & \ddots & & \vdots & & \ddots & * \\ & & & * & * & \cdots & \cdots & * \end{bmatrix}$$

(5) $p = 0$

WHILE $p \neq n$

FOR $i = 1 : n$

IF $|h_{n+i,i}| <$ *tol, setze* $h_{n+i,i}$ *zu* 0.

END FOR

FOR $i = 2 : n$

IF $|h_{i,i-1}| <$ *tol, setze* $h_{i,i-1}$ *zu* 0.

END FOR

Sei $p \geq 0$ *die größte ganze Zahl in* $\{1, \ldots, n\}$, *so daß* $h_{p+1,p} = 0$ *und* $h_{n+k,k} = 0$ *für* $k = 1, \ldots, p$.

Partitioniere $\mathcal{H}$ *also*

$$\begin{array}{cc} \begin{bmatrix} F_{11} & F_{12} & G_{11} & G_{12} \\ 0 & F_{22} & G_{21} & G_{22} \\ 0 & 0 & -F_{21}^T & 0 \\ 0 & H_{22} & -F_{12}^T & -F_{22}^T \end{bmatrix} & \begin{array}{c} p \\ n-p \\ p \\ n-p \end{array} \end{array}$$

IF $p < n$ *THEN*

Setze $H_{22} = \begin{bmatrix} F_{22} & G_{22} \\ H_{22} & -F_{22}^T \end{bmatrix}$,

$x_2 = \alpha_2(H_{22} - \lambda_1 I)\cdots(H_{22} - \lambda_n I)e_1$
und wende Schritt (3), (4) auf H_{22}, x_2 an, d.h. bilde $Q_2 \in \mathcal{US}_{2(n-p)}(\mathbb{R})$, so daß $Q_2^T x_2 = \pm\|x_2\|_2 e_1$ und dann berechne $P_2 \in \mathcal{US}_{2(n-p)}(\mathbb{R})$, so daß

$$P_2^T Q_2^T \mathcal{H} Q_2 P_2 \hat{=} \left[\begin{array}{cccc|cccc} * & \cdots & \cdots & * & * & \cdots & \cdots & * \\ * & \ddots & & \vdots & \vdots & & & \vdots \\ & \ddots & \ddots & \vdots & \vdots & & & \vdots \\ & & * & * & * & \cdots & \cdots & * \\ * & & & & * & * & & \\ & \ddots & & & \vdots & \ddots & \ddots & \\ & & \ddots & & \vdots & & \ddots & * \\ & & & * & * & \cdots & \cdots & * \end{array}\right]$$

Setze $Q_2 P_2 = \begin{bmatrix} U_2 & V_2 \\ -V_2 & U_2 \end{bmatrix}$

$$Q_p = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & U_2 & 0 & V_2 \\ 0 & 0 & I & 0 \\ 0 & -V_2 & 0 & U_2 \end{bmatrix}$$

$$\mathcal{H} := Q_p^T \mathcal{H} Q_p,\ Q := QQ_p$$

END IF

END WHILE

END.

Kritische Stellen sind die Berchnung der Eigenwerte und die Deflation. Man muß auch diesen Algorithmus mit Defektkorrektur verbinden und dann noch bei der Deflation aufpassen. Dieser Algorithmus liefert für schlecht konditionierte Probleme die besten Ergebnisse.

7.5 Byers Hamiltonischer QR–Algorithmus

Damit bleibt die Frage ob es möglich ist, einen strukturerhaltenden Algorithmus anzugeben, der die Hamiltonische Schur–form erzeugt. Solch ein Algorithmus wurde von Byers für den Fall entwickelt, daß die Hamiltonische Matrix in Hamiltonischer Hessenbergform ist, d.h.

$$\mathcal{H} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \hat{=} \left[\begin{array}{cccc|cccc} * & \cdots & \cdots & * & * & \cdots & \cdots & * \\ * & \ddots & & \vdots & \vdots & & & \vdots \\ & \ddots & \ddots & \vdots & \vdots & & & \vdots \\ & & * & * & * & \cdots & \cdots & * \\ & & & & * & * & & \\ & & & & \vdots & \ddots & \ddots & \\ & & & & \vdots & & \ddots & * \\ & & & * & * & \cdots & \cdots & * \end{array}\right] \tag{7.42}$$

mit F obere Hessenbergmatrix und $H = h_{nn} e_n e_n^T$. Diese Form entspricht der Hessenbergform im Standard–Fall. Es ist sogar eine Permutation der normalen Hessenbergform. Mit

$$P = \left[\begin{array}{c|cccc} I & & & 0 & \\ \hline & & & & 1 \\ & & & \cdot & \\ 0 & & \cdot & & \\ & \cdot & & & \\ & 1 & & & \end{array}\right] \tag{7.43}$$

ist $P^T \mathcal{H} P$ obere Hessenbergmatrix. $\mathcal{H}$ in Form 7.42 heißt <u>unreduziert</u> falls $h_{n,n} \neq 0$ und $f_{i+1,i} \neq 0, \ i = 1, \ldots, n-1$.

Das große Problem ist einen $\mathcal{O}(n^3)$ Algorithmus zu finden (möglichst endlich), der die Form (7.42) erzeugt durch Ähnlichkeitstransformation mit orthogonal symplektischen Matrizen. Dies ist aber im allgemeinen sehr schwer, denn es gilt der folgende entmutigende Satz:

Satz 7.44 *Sei $\mathcal{H} \in \mathcal{H}_{2n}(\mathbb{R})$. Dann gibt es eine Matrix $Q \in \mathcal{US}_{2n}(\mathbb{R})$, so daß $Q^T \mathcal{H} Q$ unreduzierte Hamiltonische Hessenbergform hat genau dann, wenn das nichtlineare System*

$$x^T J \mathcal{H}^{2i-1} x = 0 \quad i = i, \ldots, n-1, \quad x^T x = 1 \tag{7.45}$$

eine Lösung hat, die nicht in einem invarianten Unterraum von $\mathcal{H}$ der Dimension $\leq n$ liegt.

Beweis: Sei $Q = [q_1, \ldots, q_{2n}] \in \mathcal{US}_{2n}(\mathbb{R})$, so daß $\tilde{\mathcal{H}} = Q^T \mathcal{H} Q$ unreduzierte obere Hessenbergform hat. Da $\mathcal{H}$ unreduziert ist folgt, daß q_1 nicht in einem invarianten Unterraum der Dimension $\leq n$ liegt. Es gilt $q_1^T q_1 = 1$ und

$$\begin{array}{rcl} q_1^T J \mathcal{H}^{2i-1} q_1 & = & q_1^T J Q (Q^T \mathcal{H} Q)^{2i-1} Q^T q_1 \\ (Q \in \mathcal{US}_{2n}) & = & q_1^T Q J (Q^T \mathcal{H} Q)^{2i-1} e_1 \\ & = & e_{n+1}^T (Q^T \mathcal{H} Q)^{2i-1} e_1 \\ P \text{ wie in } (7.43) & = & e_{2n}^T (P^T Q^T \mathcal{H} Q P)^{2i-1} e_1. \end{array}$$

$P^T Q^T \mathcal{H} Q P$ ist obere Hessenbergmatrix und erst die $(2n-1)$. Potenz hat ein Element in der Position $(2n, 1) \Longrightarrow$ (7.45). Für die Rückrichtung sei $q_1 \in \mathbb{R}^{2n}$, so daß für q_1 (7.45) gilt, und q_1 nicht in einem invarianten Unterraum der Dimension $\leq n$ von $\mathcal{H}$. Der folgende orthogonal symplektische Arnoldi Algorithmus bestimmt dann Q. Die Idee ist dabei auszunutzen, daß $Q \in \mathcal{US}_{2n}(\mathbb{R})$ und $\mathcal{H} Q = Q \tilde{\mathcal{H}}$ sein soll.

Algorithmus 7.46
<u>Input:</u> *$\mathcal{H} \in \mathcal{H}_{2n}(\mathbb{R}), q_1 \in \mathbb{R}^{2n}$, welches (7.45) erfüllt:*

<u>Output:</u> *$Q \in \mathcal{US}_{2n}(\mathbb{R}), \tilde{\mathcal{H}} \in \mathcal{H}_{2n}(\mathbb{R})$, so daß $Q^T \mathcal{H} Q = \tilde{\mathcal{H}}$ in unreduzierter Hamiltonischer Hessenbergform.*
Setze $\tilde{\mathcal{H}} = 0$
  $q_{n+1} = J q_1$
  FOR $i = 1 : n-1$
    FOR $j = 1 : i$
      $\tilde{h}_{j,i} = q_j^T \mathcal{H} q_i$
    END FOR
    $\tilde{h}_{i+1,i} = \|(\mathcal{H} - \tilde{h}_{i,i} I) q_i - \sum_{j=1}^{i-1} \tilde{h}_{j,i} q_j \|_2$

$$q_{i+1} = \frac{1}{\tilde{h}_{i+1,i}} \left((\mathcal{H} - \tilde{h}_{i,i} I) q_i - \sum_{j=1}^{i-1} \tilde{h}_{j,i} q_j \right)$$
$q_{n+i+1} = J q_{i+1}$

END FOR

FOR $j = 1 : n$

$\tilde{h}_{j,n} = q_j^T \mathcal{H} q_n$

END FOR

$\tilde{h}_{2n,n} = q_n^T J^T \mathcal{H} q_n$

FOR $i = 1 : n$

FOR $j = 1 : n$

$\tilde{h}_{j,n+i} = q_j^T \mathcal{H} q_{n+i}$

END FOR

END FOR

$\tilde{\mathcal{H}}(n+1 : 2n, n+1 : 2n) = -\tilde{\mathcal{H}}(1 : n, 1 : n)^T$

END.

□

Dieser Algorithmus als solcher ist nicht stabil wegen der Division durch $\tilde{h}_{i+1,i}$ und die erzwungene Orthogonalität geht verloren während der Prozedur.

Wie kommt man auf diesen Algorithmus?

$$\mathcal{H}Q = Q\tilde{\mathcal{H}} = Q \left[\begin{array}{cccc|cccc} \tilde{h}_{11} & \cdots & \cdots & \tilde{h}_{1n} & * & \cdots & \cdots & * \\ \tilde{h}_{21} & \ddots & & \vdots & \vdots & & & \vdots \\ & \ddots & \ddots & \vdots & \vdots & & & \vdots \\ 0 & & \tilde{h}_{n,n-1} & \tilde{h}_{n,n} & * & \cdots & \cdots & * \\ \hline 0 & \cdots & \cdots & 0 & * & \cdots & \cdots & * \\ \vdots & & & \vdots & \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 & \vdots & & & \vdots \\ 0 & \cdots & 0 & \tilde{h}_{2n,n} & * & \cdots & \cdots & * \end{array} \right]$$

$$\begin{array}{rcl} \Longrightarrow \mathcal{H}q_1 & = & \tilde{h}_{11} q_1 + \tilde{h}_{21} q_2 \\ \Longrightarrow q_1^T \mathcal{H} q_1 & = & \tilde{h}_{11} \\ \tilde{h}_{21} q_2 & = & \mathcal{H} q_1 - \tilde{h}_{11} q_1 \\ \Longrightarrow |\tilde{h}_{21}| & = & \|\mathcal{H} q_1 - \tilde{h}_{11} q_1\|_2 \neq 0 \\ & & \text{ da } q_1 \text{ nicht in invariantem Unterraum} \\ 0 & = & q_1^T J \mathcal{H} q_1 = q_{n+1}^T \mathcal{H} q_1 \quad \text{ aus (7.45)} \\ \mathcal{H} q_2 & = & \tilde{h}_{12} q_1 + \tilde{h}_{22} q_2 + \tilde{h}_{32} q_3 \\ \Longrightarrow \tilde{h}_{12} & = & q_1^T \mathcal{H} q_2 \\ \Longrightarrow \tilde{h}_{22} & = & q_2^T \mathcal{H} q_2 \\ \Longrightarrow \tilde{h}_{32} q_3 & = & \mathcal{H} q_2 - \tilde{h}_{12} q_1 - \tilde{h}_{22} q_2 \\ \Longrightarrow |\tilde{h}_{32}| & = & \|\mathcal{H} q_2 - \tilde{h}_{12} q_1 - \tilde{h}_{22} q_2\|_2 \neq 0 \quad \text{ usw.} \end{array}$$

Daß die Einträge $\tilde{h}_{n+i,i} = 0$ sind, folgt aus (7.45). Wegen Satz 6.47 folgen die anderen Vektoren aus der Symplektizität.

Das Problem ist also so ein x zu finden, welches (7.45) erfüllt. In Spezialfällen ist dies einfach (siehe auch Übung). Ist in unserem Steuerungsproblem $m = 1$, d. h. $B = b \in \mathbb{R}^n$, so folgt für

$$\mathcal{H} = \begin{bmatrix} A - BR^{-1}S^T & -BR^{-1}B^T \\ -(Q - SR^{-1}S^T) & -(A - BR^{-1}S^T)^T \end{bmatrix} =: \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix},$$

daß G Rang 1 hat. Sei $Q_1 \in \mathcal{US}_{2n}(\mathbb{R})$, $Q_1 = \begin{bmatrix} 0 & Q_{12} \\ -Q_{12} & 0 \end{bmatrix}$, wobei

$$Q_{12}^T B = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \alpha \end{bmatrix}, \quad Q_{12}^T F^T Q_{12} \hat{=} \begin{bmatrix} * & \cdots & \cdots & * \\ * & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & * & * \end{bmatrix} \quad \text{(System Hessenbergform transponiert).}$$

Dann gilt für die erste Spalte q_1 von Q_1, $q_1 = Q_1 e_1$:

$$\begin{aligned} q_1^T J \mathcal{H}^{2i-1} q_1 &= e_1^T Q_1^T J \mathcal{H}^{2i-1} Q_1 e_1 \\ &= e_1^T J (Q_1^T \mathcal{H} Q_1)^{2i-1} e_1 \\ &= e_{n+1}^T (Q_1^T \mathcal{H} Q_1)^{2i-1} e_1 \\ &= e_{n+1}^T \begin{bmatrix} -Q_{12}^T F^T Q_{12} & Q_{12}^T H Q_{12} \\ \tilde{\alpha} e_n e_n^T & Q_{12}^T F Q_{12} \end{bmatrix}^{2i-1} e_1 \\ &\hat{=} e_{n+1}^T \begin{bmatrix} * & \cdots & \cdots & * & * & \cdots & \cdots & * \\ * & \ddots & & \vdots & \vdots & & & \vdots \\ & \ddots & \ddots & \vdots & \vdots & & & \vdots \\ & & * & * & * & \cdots & \cdots & * \\ & & & & * & * & & \\ & & & & \vdots & \ddots & \ddots & \\ & & & & \vdots & & \ddots & * \\ & & & * & * & \cdots & \cdots & * \end{bmatrix}^{2i-1} e_1 \\ &= e_{2n}^T \begin{bmatrix} * & \cdots & \cdots & * \\ * & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & * & * \end{bmatrix}^{2i-1} e_1 = 0 \end{aligned}$$

Hier ist im Zwischenschritt über die System–Hessenbergform von (A, B) schon die richtige Form erzielt worden.

Dies ist ein wichtiger Spezialfall für den wir also die Reduktion haben.

Wenn wir also im Fall $m = 1$ vorab (A, B) auf System Hessenbergform bringen, ist die Hamiltonische Hessenbergform schon erzeugt.

Bevor wir nun den Hamiltonischen QR–Algorithmus machen können, müssen wir noch die Reduktion mit Matrizen aus $\mathcal{US}_{2n}(\mathbb{R})$ besprechen.

Wir haben in den Ergänzungen schon

<u>Householder symplektische Matrizen</u>

$\begin{bmatrix} H & 0 \\ 0 & H \end{bmatrix}$ mit H Householder Matrix in $\mathbb{R}^{n,n}$
und Givens symplektische Matrizen

$$\begin{bmatrix} C & S \\ -S & C \end{bmatrix} \quad C = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & c & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} \leftarrow j \quad S = \begin{bmatrix} 0 & & & & & & \\ & \ddots & & & & & \\ & & 0 & & & & \\ & & & s & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix} \leftarrow j$$

betrachtet.

Mit Hilfe von orthogonal symplektischen Transformationen mit solchen Matrizen wollen wir nun einen orthogonal symplektischen QR–Algorithmus für Hamiltonische Matrizen konstruieren. Dieser stammt von R. Byers aus dem Jahre 1983.

Die Grundlage dieser Methoden ist in dem folgenden schematischen Algorithmus dargestellt:

Algorithmus 7.47 *Basis des Hamiltonischen QR–Algorithmus*

Input: Hamiltonische Matrix $\mathcal{H} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \in \mathcal{H}_{2n}(\mathbb{R})\,(\mathbb{C})$.

Output: $S \in \mathcal{US}_{2n}(\mathbb{R})\,(\mathbb{C})$, *so daß* $S^*\mathcal{H}S = T$ *in reeller Hamiltonischer Schurform (Hamiltonischer Schurform).*

$Q := I_{2n}, \mathcal{H}_0 = \mathcal{H}$
FOR $k = 1, 2, 3, \ldots$ *UNTIL* T *ungefähr in reeller Hamiltonischer Schurform.*

a) Wähle meromorphe Funktion $f_k(z)$ *und bilde die Zerlegung*

$$f_k(\mathcal{H}_{k-1}) = Q_k R_k, \tag{7.48}$$

wobei $Q_k \in \mathcal{US}_{2n}(\mathbb{R}), R_k$ *„Dreiecksähnlich“*

b) Setze $T = \mathcal{H}_k = Q_k^T \mathcal{H}_{k-1} Q_k, Q := QQ_k$

END FOR
END.

Es ist klar, daß

$$Q^*\mathcal{H}Q = \mathcal{H}_k \tag{7.49}$$

in jedem Schritt, wir erhalten also eine Folge von Ähnlichkeitstransformationen. Wahl von f: Im normalen QR-Algorithmus wählt man $f = z - \lambda$ oder $f = (z - \lambda_1)(z - \lambda_2)$ wobei λ_1, λ_2 Näherung. Hier wählt man den sogenannten Cayley-Shift

$$f_k(z) = \begin{cases} (z - \lambda_k)(z + \bar{\lambda}_k)^{-1} & \text{Einzelshift} \\ (z - \lambda_k)(z - \bar{\lambda}_k)/((z + \bar{\lambda}_k)(z + \lambda_k)) & \text{Doppelshift} \end{cases} \tag{7.50}$$

wobei λ_k eine Eigenwertnäherung ist.

Aus der Paarungseigenschaft für die Eigenwerte von Hamiltonischen Matrizen folgt, daß mit λ_k auch $-\bar{\lambda}_k$ Eigenwert ist. Für Hamiltonische Matrizen $\mathcal{H}$ folgt nun mit diesem f_k

$$f_k(\mathcal{H}) \in \mathcal{S}_{2n}(\mathbb{R}), \text{ bzw. } \mathcal{S}_{2n}(\mathbb{C}). \tag{7.51}$$

denn

$$\begin{array}{rrcl} & f_k(\mathcal{H})^* J f_k(\mathcal{H}) & = & J \\ \Longleftrightarrow & (\mathcal{H} - \lambda_k I)^* J (\mathcal{H} - \lambda_k I) & = & (\mathcal{H} + \bar{\lambda}_k I)^* J (\mathcal{H} + \bar{\lambda}_k I) \\ \Longleftrightarrow & -\bar{\lambda}_k J\mathcal{H} - \lambda_k \mathcal{H}^* J & = & \lambda_k J\mathcal{H} + \bar{\lambda}_k \mathcal{H}^* J \\ \overset{J\mathcal{H} = -\mathcal{H}^* J}{\Longleftrightarrow} & 0 & = & 0 \end{array}$$

analog für den Doppelshift.
Aus (7.51) folgt, daß auch R_k symplektisch ist, da $\mathcal{US}_{2n}$ Untergruppe von $\mathcal{S}_{2n}$. Wie sieht nun eine Dreiecksähnlichkeitsform für symplektische Matrizen aus.

$$S = \begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{11}^{-*} \end{bmatrix} \in \mathcal{S}_{2n}(\mathbb{R})\,(\mathcal{S}_{2n}(\mathbb{C})) \tag{7.52}$$

heißt symplektische Dreiecksmatrix. Wir sehen sofort, daß wir nur die ersten n Spalten auf Dreiecksform bringen müssen, die restlichen Spalten folgen dann aus der Tatsache, daß

$$R_k \in \mathcal{S}_{2n}(\mathbb{R}).$$

In der in 7.47 gegebenen Form ist der Algorithmus zwar numerisch stabil (stark stabil) und Struktur erhaltend, aber da in jedem Schritt $\mathcal{O}(n^3)$ flops ausfallen, wird er insgesamt $\mathcal{O}(n^4)$ flops brauchen. Wir müssen daher annehmen, daß $\mathcal{H}$ bereits in Hamiltonischer Hessenbergform ist, dann zeigen, daß diese Form invariant ist unter der Iteration in Algorithmus 7.47.

Um das letztere zu erreichen, verwenden wir wieder implizite Iteration, d.h. wir berechnen nur die erste Spalte von

$$f_k(\mathcal{H}_{k-1}) \text{ und daß } Q_k, \text{ welches } \tilde{Q}_k^* \mathcal{H}_{k-1} \tilde{Q}_k$$

wieder auf Hamiltonische Hessenbergform bringt, wobei $\tilde{Q}_k$ die Matrix ist, die in der ersten Spalte von $f_k(\mathcal{H}_k)$ alles bis auf das $(1,1)$ Element eleminiert. Aus Satz 7.44 folgt im wesentlichen die Eindeutigkeit der Transformationsmatrix. Das einzige Problem was bleibt (Übung), ist nun die Transformation auf Hamiltonische Hessenbergform in jedem Iterationsschritt, da dies ja, wie gesehen, Probleme macht. Sei also

$$\mathcal{H} = \begin{bmatrix} F & G \\ H & -F^* \end{bmatrix} \hat{=} \begin{bmatrix} * & \cdots & \cdots & * & * & \cdots & \cdots & * \\ * & \ddots & & \vdots & \vdots & & & \vdots \\ & \ddots & \ddots & \vdots & \vdots & & & \vdots \\ & & * & * & * & \cdots & \cdots & * \\ & & & & * & * & & \\ & & & & \vdots & \ddots & \ddots & \\ & & & & \vdots & & \ddots & * \\ & & & * & * & \cdots & \cdots & * \end{bmatrix}$$

und sei

$$f(\mathcal{H}) = (\mathcal{H} - \mu I)(\mathcal{H} + \bar{\mu})^{-1}$$

Der Trick ist nun der folgende. Würden wir $f(\mathcal{H})$ explizit bilden, so wäre die Matrix voll, aber wir können zwischen die beiden Terme immer noch etwas zwischen schieben, und machen dies so, daß wir Z berechnen (möglicherweise nicht orthogonal), so daß für

$$Q^T f(\mathcal{H}) = Q^T(\mathcal{H}-\mu I)Z\left((\mathcal{H}+\bar{\mu}I)Z\right)^{-1} \tag{7.53}$$

gilt, daß sowohl $Q^T(\mathcal{H}-\mu I)Z$ als auch $(\mathcal{H}+\bar{\mu}I)Z$ die Form $\begin{bmatrix} * & * \\ 0 & * \end{bmatrix}$ haben, wobei $Q \in \mathcal{US}_{2n}$ aber Z nur invertierbar ist.

$$\left.\begin{aligned} (\mathcal{H}-\mu I) &= \begin{bmatrix} F-\mu I & G \\ H & -F^T-\mu I \end{bmatrix} \\ (\mathcal{H}+\bar{\mu} I) &= \begin{bmatrix} F+\bar{\mu} I & G \\ H & -F^T+\bar{\mu} I \end{bmatrix} \end{aligned}\right\} \hat{=} \begin{bmatrix} * & \cdots & \cdots & * & * & \cdots & \cdots & * \\ * & \ddots & & \vdots & \vdots & & & \vdots \\ & \ddots & \ddots & \vdots & \vdots & & & \vdots \\ & & * & * & * & \cdots & \cdots & * \\ & & & & * & * & & \\ & & & & \vdots & \ddots & \ddots & \\ & & & & \vdots & & \ddots & * \\ & & & * & * & \cdots & \cdots & * \end{bmatrix}$$

Sei $\tilde{Z}_1$ orthogonal, so daß $(-F^T+\bar{\mu}I)\tilde{Z}_1^* = R_1$ untere $\triangle$–Matrix, $\tilde{Z}_1$ ist dann untere Hessenberg–Matrix, denn $-F^T+\bar{\mu}I \hat{=} \begin{bmatrix} * & * & & \\ \vdots & \ddots & \ddots & \\ \vdots & & \ddots & * \\ * & \cdots & \cdots & * \end{bmatrix} \hat{=} R_1\tilde{Z}_1 \hat{=} \begin{bmatrix} * & & \\ \vdots & \ddots & \\ * & \cdots & * \end{bmatrix} \tilde{Z}_1$.

Bilde $Z_1 = \begin{bmatrix} I & 0 \\ 0 & \tilde{Z}_1^* \end{bmatrix}$, so folgt daß

$$\begin{aligned} (\mathcal{H}+\bar{\mu}I)Z_1 \;&\hat{=}\; \begin{bmatrix} * & \cdots & \cdots & * & * & \cdots & \cdots & * \\ * & \ddots & & \vdots & \vdots & & & \vdots \\ & \ddots & \ddots & \vdots & \vdots & & & \vdots \\ & & * & * & * & \cdots & \cdots & * \\ & & & & * & * & & \\ & & & & \vdots & \ddots & \ddots & \\ & & & & \vdots & & \ddots & * \\ & & & * & * & \cdots & \cdots & * \end{bmatrix} \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & 0 & & \\ & & 1 & & & & \\ & & & * & * & & \\ & & & \vdots & \ddots & \ddots & \\ & 0 & & \vdots & & \ddots & * \\ & & & * & \cdots & \cdots & * \end{bmatrix} \\ &= \begin{bmatrix} * & \cdots & \cdots & * & * & \cdots & \cdots & * \\ * & \ddots & & \vdots & \vdots & & & \vdots \\ & \ddots & \ddots & \vdots & \vdots & & & \vdots \\ & & * & * & * & \cdots & \cdots & * \\ & & & & * & & & \\ & & & & \vdots & \ddots & & \\ & & & & \vdots & & \ddots & \\ & & & h & * & \cdots & \cdots & t \end{bmatrix} \end{aligned} \tag{7.54}$$

Sei

$$Z_2 = \begin{bmatrix} C_1 & S_1 \\ -S_1 & C_1 \end{bmatrix}, \quad C_1 = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & c_1 \end{bmatrix}, \quad S_1 = \begin{bmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & s_1 \end{bmatrix}$$

eine orthogonale Matrix, die bei der Bildung von $(\mathcal{H} + \bar{\mu} I)Z_1 Z_2$ den (2,1) Block eleminiert. Dann gilt

$$(\mathcal{H} + \bar{\mu} I)Z_1 Z_2 \hat{=} \begin{bmatrix} * & \cdots & \cdots & * & * & \cdots & \cdots & * \\ * & \ddots & & \vdots & \vdots & & & \vdots \\ & \ddots & \ddots & \vdots & \vdots & & & \vdots \\ & & * & * & * & \cdots & \cdots & * \\ & & & & * & & & \\ & & & & \vdots & \ddots & & \\ & 0 & & & \vdots & & \ddots & \\ & & & & * & \cdots & \cdots & * \end{bmatrix} \tag{7.55}$$

Nun gilt aber

$$\begin{aligned} \begin{bmatrix} \tilde{Z}_1 & 0 \\ 0 & \tilde{Z}_1 \end{bmatrix} (\mathcal{H} - \mu I) \begin{bmatrix} I & 0 \\ 0 & \tilde{Z}_1^* \end{bmatrix} &= \begin{bmatrix} \tilde{Z}_1(F - \mu I) & \tilde{Z}_1 G \tilde{Z}_1^* \\ \tilde{Z}_1 H & -\tilde{Z}_1(F^T + \mu I)\tilde{Z}_1^* \end{bmatrix} \\ &= \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \end{aligned} \tag{7.56}$$

$$\tilde{Z}_1(F - \mu I) = -\left[(-F^T + \bar{\mu} I)\tilde{Z}_1^*\right]^* = -R_1^* \hat{=} \begin{bmatrix} * & \cdots & * \\ & \ddots & \vdots \\ & & * \end{bmatrix}$$

$$\tilde{Z}_1 H = \left[\tilde{Z}_{ij}\right]_{i,j} \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ \vdots & & & \vdots \\ \vdots & & & 0 \\ 0 & \cdots & 0 & h \end{bmatrix} = \begin{bmatrix} 0 & & \cdots & & 0 \\ & & & & \vdots \\ \vdots & & & & 0 \\ & & & & \tilde{z}_{n-1,n}h \\ 0 & \cdots & \cdots & 0 & \tilde{z}_{n,n}h \end{bmatrix}$$

$$\begin{aligned} \tilde{Z}_1(F^T + \mu I)\tilde{Z}_1^* &= \tilde{Z}_1(F^T - \bar{\mu} I)\tilde{Z}_1^* + 2Re(\mu) I \\ &\hat{=} \begin{bmatrix} * & * & & \\ \vdots & \ddots & \ddots & \\ \vdots & & \ddots & * \\ * & \cdots & \cdots & * \end{bmatrix} + \begin{bmatrix} * & & & \\ & \ddots & & \\ & & \ddots & \\ & & & * \end{bmatrix} \\ &= \begin{bmatrix} * & * & & \\ \vdots & \ddots & \ddots & \\ \vdots & & \ddots & * \\ * & \cdots & \cdots & * \end{bmatrix}. \end{aligned}$$

Die beiden Elemente in der letzten Spalte sind $\begin{bmatrix} \tilde{z}_{n-1,n}t \\ \tilde{z}_{n,n}t + 2Re(\mu) \end{bmatrix}$.
Dann folgt, daß

$$Y \;:=\; \begin{bmatrix} \tilde{Z}_1 & 0 \\ 0 & \tilde{Z}_1 \end{bmatrix} (\mathcal{H} - \mu I) \begin{bmatrix} I & 0 \\ 0 & \tilde{Z}_1^* \end{bmatrix} \begin{bmatrix} C_1 & S_1 \\ -S_1 & C_1 \end{bmatrix} \tag{7.57}$$

$$\hat{=} \begin{bmatrix} * & \cdots & \cdots & * & * & \cdots & \cdots & * \\ & \ddots & & \vdots & \vdots & & & \vdots \\ & & \ddots & \vdots & \vdots & & & \vdots \\ & & & * & * & \cdots & \cdots & * \\ & & & & * & * & & \\ & & & & \vdots & \ddots & \ddots & \\ & & & & \vdots & & \ddots & * \\ & & & * & * & \cdots & \cdots & * \end{bmatrix} \tag{7.58}$$

Wähle nun

$$Z_3 = \begin{bmatrix} C_2 & S_2 \\ -S_2 & C_2 \end{bmatrix} \quad C_2 = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & c_2 \end{bmatrix} \quad S_2 = \begin{bmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & s_2 \end{bmatrix}$$

so daß

$$Z_3 Y \hat{=} \begin{bmatrix} * & \cdots & \cdots & * & * & \cdots & \cdots & * \\ & \ddots & & \vdots & \vdots & & & \vdots \\ & & \ddots & \vdots & \vdots & & & \vdots \\ & & & * & * & \cdots & \cdots & * \\ & & & & * & * & & \\ & 0 & & & \vdots & \ddots & \ddots & \\ & & & & \vdots & & \ddots & * \\ & & & & * & \cdots & \cdots & * \end{bmatrix} \tag{7.59}$$

Dann folgt

$$\begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} \;:=\; Z_3 \begin{bmatrix} \tilde{Z}_1 & 0 \\ 0 & \tilde{Z}_1 \end{bmatrix} (\mathcal{H} - \bar{\mu} I) Z_1 Z_2 \left[(\mathcal{H} + \bar{\mu} I) Z_1 Z_2\right]^{-1} \tag{7.60}$$

$$\hat{=} \begin{bmatrix} * & \cdots & \cdots & * & * & \cdots & \cdots & * \\ & \ddots & & \vdots & \vdots & & & \vdots \\ & & \ddots & \vdots & \vdots & & & \vdots \\ & & & * & * & \cdots & \cdots & * \\ & & & & * & * & & \\ & 0 & & & \vdots & \ddots & \ddots & \\ & & & & \vdots & & \ddots & * \\ & & & & * & \cdots & \cdots & * \end{bmatrix} \begin{bmatrix} * & \cdots & \cdots & * & * & \cdots & \cdots & * \\ * & \ddots & & \vdots & \vdots & & & \vdots \\ & \ddots & \ddots & \vdots & \vdots & & & \vdots \\ & & * & * & * & \cdots & \cdots & * \\ & & & & * & & & \\ & 0 & & & \vdots & \ddots & & \\ & & & & \vdots & & \ddots & \\ & & & & * & \cdots & \cdots & * \end{bmatrix}^{-1}$$

$$
= \left[\begin{array}{cccc|cccc}
* & \cdots & \cdots & * & * & \cdots & \cdots & * \\
\vdots & & & \vdots & \vdots & & & \vdots \\
\vdots & & & \vdots & \vdots & & & \vdots \\
* & \cdots & \cdots & * & * & \cdots & \cdots & * \\
 & & & & * & * & & \\
 & & 0 & & \vdots & \ddots & \ddots & \\
 & & & & \vdots & & \ddots & * \\
 & & & & * & \cdots & \cdots & *
\end{array}\right]
$$

Sei nun $\tilde{Z}_4$ so, daß $Y_{22}\tilde{Z}_4^*$ untere Dreiecksmatrix und bilde

$$
\begin{aligned}
\left[\begin{array}{cc} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{array}\right]\left[\begin{array}{cc} \tilde{Z}_4^* & 0 \\ 0 & \tilde{Z}_4^* \end{array}\right] &= Z_3 \left[\begin{array}{cc} \tilde{Z}_1 & 0 \\ 0 & \tilde{Z}_1 \end{array}\right](\mathcal{H}-\mu I)Z_1Z_2\left[(\mathcal{H}+\bar{\mu}I)Z_1Z_2\right]^{-1}\left[\begin{array}{cc} \tilde{Z}_4^* & 0 \\ 0 & \tilde{Z}_4^* \end{array}\right] \\
&\hat{=} \left[\begin{array}{cccc|cccc}
* & \cdots & \cdots & * & * & \cdots & \cdots & * \\
\vdots & & & \vdots & \vdots & & & \vdots \\
\vdots & & & \vdots & \vdots & & & \vdots \\
* & \cdots & \cdots & * & * & \cdots & \cdots & * \\
 & & & & * & & & \\
 & & 0 & & \vdots & \ddots & & \\
 & & & & \vdots & & \ddots & \\
 & & & & * & \cdots & \cdots & *
\end{array}\right]
\end{aligned} \tag{7.61}
$$

aber diese Matrix ist symplektisch und daher folgt auch $Y_{11}\tilde{Z}_4 \hat{=} \left[\begin{array}{ccc} * & \cdots & * \\ & \ddots & \vdots \\ & & * \end{array}\right]$.

Bilde nun als neue Iterierte

$$
\begin{aligned}
\mathcal{H} &=: \left[\begin{array}{cc} \tilde{Z}_4 & 0 \\ 0 & \tilde{Z}_4 \end{array}\right] Z_3 \left[\begin{array}{cc} \tilde{Z}_1 & 0 \\ 0 & \tilde{Z}_1 \end{array}\right] \mathcal{H} \left[\begin{array}{cc} \tilde{Z}_1^* & 0 \\ 0 & \tilde{Z}_1^* \end{array}\right] Z_3^* \left[\begin{array}{cc} \tilde{Z}_4^* & 0 \\ 0 & \tilde{Z}_4^* \end{array}\right] \\
&\hat{=} \left[\begin{array}{cccc|cccc}
* & \cdots & \cdots & * & * & \cdots & \cdots & * \\
* & \ddots & & \vdots & \vdots & & & \vdots \\
 & \ddots & \ddots & \vdots & \vdots & & & \vdots \\
 & & * & * & * & \cdots & \cdots & * \\
 & & & & * & * & & \\
 & & & & \vdots & \ddots & \ddots & \\
 & & & & \vdots & & \ddots & * \\
 & & & * & * & \cdots & \cdots & *
\end{array}\right]
\end{aligned} \tag{7.62}
$$

Man kann dies alles natürlich auch für den Doppelshift machen, dann wird alles nur noch komplizierter, aber die Grundidee ist dieselbe.

Damit erhalben wir den folgenden impliziten Hamiltonischen QR–Schritt. Implizite Version des obigen Algorithmus:

Algorithmus 7.63 *Byers impliziter Hamiltonischer QR–Schritt*

Input: $\mathcal{H} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \in \mathcal{H}_{2n}(\mathbb{C})(\mathcal{H}_{2n}(\mathbb{R}))$ *in Hamiltonischer Hessenbergform*

$$\mathcal{H} \hat{=} \begin{bmatrix} * & \cdots & \cdots & * & * & \cdots & \cdots & * \\ * & \ddots & & \vdots & \vdots & & & \vdots \\ & \ddots & \ddots & \vdots & \vdots & & & \vdots \\ & & * & * & * & \cdots & \cdots & * \\ & & & & * & * & & \\ & & & & \vdots & \ddots & \ddots & \\ & & & & \vdots & & \ddots & * \\ & & & * & * & \cdots & \cdots & * \end{bmatrix},$$

Shift μ.

Output: $Z \in \mathcal{US}_{2n}(\mathbb{C})$, *so daß* $Z^*\mathcal{H}Z \in \mathcal{H}_{2n}(\mathbb{C})$ *in Hamiltonsicher Hessenbergform und*

$$Z^*(\mathcal{H} - \mu I)(\mathcal{H} + \bar{\mu} I)^{-1} Z \hat{=} \begin{bmatrix} * & \cdots & * & * & \cdots & * \\ & \ddots & \vdots & \vdots & & \vdots \\ & & * & * & \cdots & * \\ & & & * & & \\ & 0 & & \vdots & \ddots & \\ & & & * & \cdots & * \end{bmatrix}$$

$$\begin{aligned} x &= (F - \mu I)e_1 \\ v &= \text{house}(x(1:2)) \end{aligned}$$

$$\begin{aligned}
F(1:2,:) &= \text{row.house}(F(1:2,:), v) \\
G(1:2,:) &= \text{row.house}(G(1:2,:), v) \\
H(1:2,n) &= \text{row.house}(H(1:2,n), v) \\
F(:,1:2) &= \text{col.house}(F(:,1:2), v) \\
G(:,1:2) &= \text{col.house}(G(:,1:2), v) \\
H(n-1:n,1:2) &= \text{col.house}(H(n-1:n,1:2), v) \\
Z(:,1:2) &= \text{col.house}(Z(:,1:2), v) \\
Z(:,n+1:n+2) &= \text{col.house}(Z(:,n+1:n+2), v)
\end{aligned}$$

FOR $k = 1 : n-2$

$$\begin{aligned}
v &= \text{house}(F(k+1:k+2,k)) \\
F(k+1:k+2,k:n) &= \text{row.house}(F(k+1:k+2,k:n), v) \\
G(k+1:k+2,:) &= \text{row.house}(G(k+1:k+2,:), v) \\
H(k+1:k+2,n) &= \text{row.house}(H(k+1:k+2,n), v) \\
F(:,k+1:k+2) &= \text{col.house}(F(:,k+1:k+2), v) \\
G(:,k+1:k+2) &= \text{col.house}(G(:,k+1:k+2), v) \\
H(n-1:n,k+1:k+2) &= \text{col.house}(H(n-1:n,k+1:k+2), v) \\
Z(:,k+1:k+2) &= \text{col.house}(Z(:,k+1:k+2), v) \\
Z(:,n+k+1:n+k+2) &= \text{col.house}(Z(:,n+k+1:n+k+2), v)
\end{aligned}$$

END FOR

$$\% \qquad F \hat{=} \begin{bmatrix} * & \cdots & \cdots & * \\ * & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & * & * \end{bmatrix}, \quad G \hat{=} \begin{bmatrix} * & \cdots & * \\ \vdots & & \vdots \\ * & \cdots & * \end{bmatrix}, \quad H \hat{=} \begin{bmatrix} 0 & \cdots & \cdots & 0 & 0 \\ & & & \vdots & \vdots \\ \vdots & & & 0 & 0 \\ & & & * & * \\ 0 & \cdots & 0 & * & * \end{bmatrix}$$

$$Sei \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2(Re(F(n,n)H(n-1,n-1)) + Re(H(n,n-1)\bar{F}(n-1,n)) \\ H(n-1,n-1)G(n,n) + |F(n-1,n)|^2 \end{bmatrix}$$

$$\begin{aligned}
[c,s] &= \text{givens}(x_1, x_2) \\
A &= \begin{bmatrix} F(n,:) \\ H(n,:) \end{bmatrix} \\
B &= \begin{bmatrix} G(n,:) \\ -F(:,n)^* \end{bmatrix} \\
A &= \text{row.rot}(A,c,s) \\
F(n,:) &= A(1,:) \\
H(n,:) &= A(2,:) \\
B &= \text{row.rot}(B,c,s) \\
G(n,:) &= B(1,:) \\
A &= [F(:,n),\ G(:,n)] \\
B &= [H(:,n),\ -F(n,:)^*] \\
A &= \text{col.rot}(A,c,s) \\
F(:,n) &= A(:,1) \\
G(:,n) &= A(:,2) \\
B &= \text{col.rot}(B,c,s) \\
H(:,n) &= B(:,1) \\
B &= [Z(:,n),\ Z(:,2n)] \\
B &= \text{col.rot}(B,c,s) \\
Z(:,n) &= B(:,1) \\
Z(:,2n) &= B(:,2).
\end{aligned}$$

$$\% \qquad F \hat{=} \begin{bmatrix} * & \cdots & \cdots & * \\ * & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & * & * \end{bmatrix}, \quad G \hat{=} \begin{bmatrix} * & \cdots & * \\ \vdots & & \vdots \\ * & \cdots & * \end{bmatrix}, \quad H \hat{=} \begin{bmatrix} 0 & \cdots & \cdots & 0 & 0 \\ & & & & \vdots \\ \vdots & & & \vdots & 0 \\ & & & & * \\ 0 & \cdots & \cdots & 0 & * \end{bmatrix}$$

Setze $x = F^* e_n$

$v = \text{house}(x(n-1:n))$

$$\begin{aligned}
F(n-1:n, n-1:n) &= \text{row.house}(F(n-1:n,n-1:n), v) \\
G(n-1:n,:) &= \text{row.house}(G(n-1:n,:), v) \\
H(n-1:n,n-1:n) &= \text{row.house}(H(n-1:n,n-1:n), v) \\
F(:,n-1:n) &= \text{col.house}(F(:,n-1:n), v) \\
G(:,n-1:n) &= \text{col.house}(G(:,n-1:n), v) \\
H(n-1:n,n-1:n) &= \text{col.house}(H(n-1:n,n-1:n), v) \\
Z(:,n-1:n) &= \text{col.house}(Z(:,n-1:n), v) \\
Z(:,2n-1:2n) &= \text{col.house}(Z(:,2n-1:2n), v)
\end{aligned}$$

$$\% \qquad F \hat{=} \begin{bmatrix} * & \cdots & \cdots & * \\ * & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & * & * & * \end{bmatrix}, \quad G \hat{=} \begin{bmatrix} * & \cdots & * \\ \vdots & & \vdots \\ * & \cdots & * \end{bmatrix}, \quad H \hat{=} \begin{bmatrix} 0 & \cdots & \cdots & 0 & 0 \\ & & & & \vdots \\ \vdots & & & \vdots & \vdots \\ & & & & 0 \\ 0 & \cdots & \cdots & 0 & * \end{bmatrix}$$

FOR $k = n : -1 : 3$

$$\textit{Setze } x = \begin{bmatrix} & & & 1 \\ & & \cdot & \\ & \cdot & & \\ \cdot & & & \\ 1 & & & \end{bmatrix} F^* e_k$$

$$v = \text{house } (x(n-k+2 : n-k+3))$$

$$v = \begin{bmatrix} & & & 1 \\ & & \cdot & \\ & \cdot & & \\ \cdot & & & \\ 1 & & & \end{bmatrix} v$$

$$\begin{aligned}
F(k-2:k-1,:) &= \text{row.house}(F(k-2:k-1,:),v) \\
G(k-2:k-1,:) &= \text{row.house}(G(k-2:k-1,:),v) \\
F(1:k,k-2:k-1) &= \text{col.house}(F(1:k,k-2:k-1),v) \\
G(:,k-2:k-1) &= \text{col.house}(G(:,k-2:k-1),v) \\
Z(:,k-2:k-1) &= \text{col.house}(Z(:,k-2:k-1),v) \\
Z(:,n+k-2:n+k-1) &= \text{col.house}(Z(:,n+k-2:n+k-1),v)
\end{aligned}$$

END FOR

Ich verzichte hier auf die Doppelshiftvariante, weil es noch viel technischer wird. Insgesamt erhalten wir dann den folgenden Hamiltonischen QR–Algorithmus.

Algorithmus 7.64 *Byers Hamiltonischer QR–Algorithmus*

Input: $\mathcal{H} \in \mathcal{H}_{2n}(\mathbb{C}), \mathcal{H} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix}$,

$\mathcal{H}$ habe keine Eigenwerte mit Realteil 0 und sei in Hamiltonischer Hessenbergform

$$\mathcal{H} \hat{=} \begin{bmatrix} * & \cdots & \cdots & * & * & \cdots & \cdots & * \\ * & \ddots & & \vdots & \vdots & & & \vdots \\ & \ddots & \ddots & \vdots & \vdots & & & \vdots \\ & & * & * & * & \cdots & \cdots & * \\ & & & & * & * & & \\ & & & & \vdots & \ddots & \ddots & \\ & & & & \vdots & & \ddots & * \\ & & & * & * & \cdots & \cdots & * \end{bmatrix}, \quad Q \in \mathcal{US}_{2n}(\mathbb{C})$$

Output: $Q \in \mathcal{US}_{2n}(\mathbb{C})$ *so daß* $Q^*\mathcal{H}Q$ *in Hamiltonischer Schurform mit den stabilen Eigenwerten im* $(1,1)$ *Block,* $q = 0$

(1) FOR $j = 1, 2, \ldots$ UNTIL $q = n-1$

Setze alle Subdiagonal Elemente von F, die

$$|f(i,i-1)| \leq \; tol\,(|f(i,i)| + |f(i-1,i-1)|) \tag{7.65}$$

erfüllen, zu 0. Bestimme größtes $q \geq 0$, so daß

$$\mathcal{H} = \begin{bmatrix} F_{11} & F_{12} & G_{11} & G_{12} \\ 0 & F_{22} & G_{21} & G_{22} \\ 0 & 0 & -F_{11}^* & 0 \\ 0 & H_{22} & -F_{12}^* & -F_{22}^* \end{bmatrix} \begin{matrix} \}q \\ \}n-q \\ \}q \\ \}n-q \end{matrix}$$

wobei F_{22} unreduzierte obere Hessenbergmatrix.

IF $q = n-1$ THEN

$x = F_{22}(n,n) - \bar{F}_{22}(n,n)$

$y = -|F_{22}(n,n)|^2 - G_{22}(n,n)H_{22}(n,n)$

$\mu_1 = -\frac{x}{2} + \sqrt{(\frac{x}{2})^2 - y}$

$\mu_2 = -\frac{x}{2} - \sqrt{(\frac{x}{2})^2 - y}$

IF $Re(\mu_1) < 0$ THEN

$\mu = \mu_2$

ELSE

$\mu = \mu_1$

END IF

$$\begin{array}{lcl} [c,s] & = & \text{givens } [F_{22}(n,n) - \mu, H_{22}(n,n)] \\ A & = & \begin{bmatrix} F(n,1:n) \\ H(n,1:n) \end{bmatrix} \\ B & = & \begin{bmatrix} G(n,1:n) \\ -F(1:n,n)^* \end{bmatrix} \\ A & = & \text{row.rot } (A,c,s) \\ B & = & \text{row.rot } (B,c,s) \\ F(n,1:n) & = & A(1,:) \\ H(n,1:n) & = & A(2,:) \\ G(n,1:n) & = & B(1,:) \end{array}$$

Wende den QR–Algorithmus auf F_{11} an und berechne $\tilde{Q}_1 \in \mathbb{C}^{n-1,n-1}$ orthogonal, so daß $\tilde{Q}_1^* F_{11} \tilde{Q}_1 \hat{=} \begin{bmatrix} * & \cdots & * \\ & \ddots & \vdots \\ & & * \end{bmatrix}$.

Setze $Q_1 = \begin{bmatrix} \tilde{Q}_1 & \\ & 1 \end{bmatrix}$ *und bilde*

$F = Q_1^* F Q_1, \quad G = Q_1^* G Q_1.$

ELSE

Setze $\mathcal{H} = \begin{bmatrix} F_{22} & G_{22} \\ H_{22} & -F_{22}^* \end{bmatrix}$ *und $\tilde{Z} = I \in \mathbb{C}^{n-q,n-q}$ und wende Algorithmus 7.63 auf $\tilde{\mathcal{H}}, \tilde{Z}$ an.*

Setze $Z = \begin{bmatrix} I_q & & & \\ & \tilde{Z}_{11} & 0 & \tilde{Z}_{12} \\ & & I_q & \\ & \tilde{Z}_{21} & 0 & \tilde{Z}_{22} \end{bmatrix}$ *wobei $\tilde{Z}_{ij}$ die Blöcke von $\tilde{Z}$ sind.*

Setze $\mathcal{H} = Z^*\mathcal{H}Z, Q = QZ$.
END IF
END FOR.

(2) Vertausche die Eigenwerte sukzessive mit symplektischen Givensrotationen, bis der $(1,1)$ *Block lauter Eigenwerte vom Realteil* < 0 *hat.*

END.

Kosten: $40n^2$ *pro Iterationsschritt*

Fehler: *Die berechnete Hamiltonische Schurform ist die exakte Hamiltonische Schurform von* $\mathcal{H} + \mathcal{E}$, *wobei* $\mathcal{E} \in \mathcal{H}_{2n}(\mathbb{C})$ *und* $\|\mathcal{E}\| \le c\gamma(n)\|\mathcal{H}\|$. *(numerisch stark stabil)*

Problem: *Reduktion auf Hamiltonische Hessenbergform*

Es gibt bisher im wesentlichen folgende Methoden

	Kosten	Vorteile	Nachteile
Newton	$\mathcal{O}(n^3)$ pro Schritt	monotone quadr. Konvergenz	Berechnung des Startwerts
Sign–Funktion	$\mathcal{O}(n^3)$ pro Schritt	einfach , parallelisierbar,	instabil, insbes. bei Eigenwerte nahe der Imaginären Achse
Schur–Vektor–Methode	$\mathcal{O}(n^3)$	numerisch stabil	keine Strukturerhaltung
Hamiltonischer QR (Byers)	$\mathcal{O}(n^3)$	numerisch stabil, strukturerhaltend	vorab Reduktion auf Hamiltonsiche Hessenbergform i.a. nicht gegeben.
SR Bunse–Gerstner/Mehrmann	$\mathcal{O}(n^3)$	strukturerhaltend	instabil.
OSMARE Multi–Shift Amar/Benner/Mehrmann	$\mathcal{O}(n^3)$	strukturerhaltend numerisch stabil	wesentlich langsamer als QR

Kapitel 8

Zeitdiskrete Systeme

Wir haben bisher im wesentlichen zeitkontinuierliche Systeme betrachtet. Fast alle Ergebnisse aus dem bisherigen Teil der Vorlesung lassen sich direkt auf zeitdiskrete Systeme übertragen. Dabei können wir fast immer mit Hilfe der Cayley Transformation hin- und herspringen, um die Ergebnisse zu erhalten. Das allgemeine zeitdiskrete System hat die Form

$$x_{k+1} = A_k x_k + B_k u_k,\ x_0 = x^0,\quad k = 1, 2, 3, \ldots \tag{8.1}$$

mit Ausgang

$$y_k = C_k x_k \quad k = 1, 2, 3, \ldots \tag{8.2}$$

eventuell mit $A_k \equiv A, B_k \equiv B, C_k \equiv C \quad \forall k = 1, 2, 3, \ldots$. Wenn wir Optimalsteuerung betrachten wollen, so ist das entsprechende Kostenfunktional

$$\mathcal{S}(\{x_k\}, \{u_k\}) = \frac{1}{2}\left(x_{k_f}^T M x_{k_f} + \sum_{k=k_0}^{k_f} [x_k^T,\ u_k^T] \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix} \right) \tag{8.3}$$

Diskrete Modelle spielen besonders in der Ökonomie eine wichtige Rolle „Controlling", Betriebsorganisation,....

Definition 8.4 *Gegeben sei das System (8.1) und ein Endzustand x^1. Das Paar $(0, x^0)$ heißt* <u>*zum Zeitpunkt $k_1 > 0$ nach x^1 steuerbar*</u>*, falls es eine Steuerfolge $\{u_k\}$ gibt, so daß die Lösung* <u>*von (8.1) mit dieser Steuerfolge*</u> *$x_{k_1} = x^1$ erfüllt.*

Das Paar $(0, x^0)$ heißt <u>*nach x^1 steuerbar*</u>*, falls es zu irgendeinem Zeitpunkt $0 < k_1 < \infty$ nach x^1 steuerbar ist. Falls für jedes $(0, x^0)$ und jedes x^1 das Paar $(0, x_0)$ nach x^1 steuerbar ist, so heißt (8.1)* <u>*vollständig*</u> *steuerbar.*

Satz 8.5 *Das System (8.1) ist vollständig steuerbar genau dann wenn folgende Aussage gilt: Ist $(z_0, z_1, z_2, \ldots)$ eine nichttriviale Lösungsfolge des dualen Systems*

$$x_{k+1} = A_k^T x_k \tag{8.6}$$

so ist die Folge $z_k^T B_k,\ k = 1, 2, \ldots$ nicht die Folge $(0, 0, 0, \ldots)$.

Beweis: Analog zum Beweis von Satz 2.20. □

Im zeitinvarianten Fall erhalten wir die analoge Charakterisierung über Ränge von Matrizen.

Satz 8.7 *Für das zeitinvariante System*

$$x_{k+1} = Ax_k + Bu_k, \quad x_0 = x^0 \tag{8.8}$$

sind die folgenden Aussagen äquivalent:

i) *(8.8) ist steuerbar;*

ii) *Rang* $(K) =$ *Rang* $([B, AB, A^2B, \ldots, A^{n-1}B]) = n$;

iii) *Ist* $p \neq 0$ *ein Eigenvektor zu* A^T *so gilt* $p^T B \neq 0$;

iv) *Rang* $(\lambda I - A, B) = n, \forall \lambda \in \mathbb{C}$

Beweis: Vollkommen analog zu Satz 2.26,2.29. □

Definition 8.9 *Ein System der Form (8.1) heißt stabilisierbar, wenn es zu jedem* $(0, x^0)$ *eine für* $k \geq 0$ *definierte Steuerfolge* $\{u_k\}$ *gibt, so daß die Lösungsfolge von (8.1) mit dieser Steuerung gegen* 0 *konvergiert für* $k \to \infty$.

Satz 8.10 *Wenn (8.1) stabilisierbar ist so gilt: Ist* $z_k; k = 0, 1, 2, \ldots$ *eine Lösungsfolge von (8.6) die für* $k \to \infty$ *beschränkt ist, so kann nicht gelten*

$$z_k^T B_k = 0 \quad \forall k = 0, 1, 2, \ldots$$

Beweis: Analog Satz 2.32 □

Satz 8.11 *Für das zeitinvariante System (8.8) sind folgende Aussagen äquivalent:*

i) *(8.8) ist stabilisierbar;*

ii) *Ist* λ *mit* $|\lambda| \geq 1$ *Eigenwert von* A *mit Linkseigenvektor* p, *so gilt* $p^T B \neq 0$;

iii) *Rang* $(\lambda I - A, B) = n \quad \forall \lambda \in \mathbb{C}$ *mit* $|\lambda| \geq 1$.

Beweis: Analog zu Satz 2.34. □

Man sieht, daß alle diese Ergebnisse fast gleich sind bis auf die Tatsache, daß sich die Stabilität auf den Einheitskreis bezieht und nicht auf die imaginäre Achse. Aber dies ist klar, denn was ist die Lösung von (8.1):

$$\begin{aligned} x_{k+1} &= A_k x_k + B_k u_k \\ &= A_k (A_{k-1} x_{k-1} + B_{k-1} u_{k-1}) + B_k u_k \quad \text{usw.} \end{aligned}$$

also

$$x_{k+1} = \left(\prod_{j=0}^{k} A_j \right) x^0 + \left(+ \sum_{j=0}^{k} (\prod_{i=j+1}^{k} A_i) B_j u_j \right) \tag{8.12}$$

(leeres Produkt ist 1) $\qquad k = 0, 1, 2, \ldots$

bzw. für zeitinvariante Systeme

$$x_{k+1} = A^{k+1} x^0 + \sum_{j=0}^{k} (A^{k-j} B) u_j. \tag{8.13}$$

Es ist klar, daß für $k \to \infty \quad A^k \to \infty$, falls Eigenwerte von A einen Betrag > 1 haben, und daß asymptotische Stabilität ($x_k \to 0$) nur gelten kann wenn $|\lambda| < 1$ für alle Eigenwerte von A. Für Systeme mit Ausgang haben wir analog

Definition 8.14 *Das System (8.1), (8.2) heißt* <u>*rekonstruierbar*</u>*, falls gilt:*

Sind $\{x_k\}, \{\tilde{x}_k\}$ *Lösungsfolgen von (8.1) für die gleiche Steuerfolge* $\{u_k\}$ *und gilt*

$$C_k x_k = C_k \tilde{x}_k \quad \forall k = 0, 1, 2, \ldots, k_0 \tag{8.15}$$

so gilt $\quad x_k = \tilde{x}_k \quad \forall k = k_0, k_{0+1}, \ldots$.

Das System heißt <u>*beobachtbar*</u>*, falls aus*

$$C_k x_k = C_k \tilde{x}_k \quad \forall k = 0, 1, 2, \ldots. \tag{8.16}$$

folgt, daß

$$x_k = \tilde{x}_k \quad \forall k = 0, 1, 2, \ldots. \tag{8.17}$$

Satz 8.18 *Das System (8.1), (8.2) ist rekonstruierbar, falls das duale System*

$$x_{k+1} = A_k^T x_k + C_k^T u_k \quad x_0 = x^0 \tag{8.19}$$

steuerbar ist.

Beweis: Analog zu Satz 3.8. □

Satz 8.20 *Für das zeitinvariante System (8.8) mit Ausgang*

$$y_k = C x_k \tag{8.21}$$

sind die folgenden Aussagen äquivalent:

(i) Das System (8.8), (8.21) ist rekonstruierbar;

(ii) Das System (8.8), (8.21) ist beobachtbar;

(iii) Rang $(K) =$ *Rang* $\left([C^T, A^T, C^T, \ldots, (A^T)^{n-1} C^T]\right) = n$*;*

(iv) Ist $p \neq 0$ *Eigenvektor von* A *so gilt* $C^T p \neq 0$*;*

(v) Rang $\begin{bmatrix} \lambda I - A \\ C \end{bmatrix} = n, \ \forall \lambda \in \mathbb{C}$.

Beweis: Klar mit Dualität. □

Definition 8.22 *Das zeitinvariante System (8.8), (8.21) heißt* <u>*entdeckbar*</u>*, wenn das duale System (8.19) stabilisierbar ist.*

Satz 8.23 *Die folgenden Aussagen sind äquivalent:*

(i) Das System (8.8), (8.21) ist entdeckbar;

(ii) Ist λ *mit* $|\lambda| \geq 1$ *Eigenwert von* A *mit Eigenvektor* $p \neq 0$*, so gilt* $C^T p \neq 0$*;*

(iii) Rang $\begin{pmatrix} \lambda I - A \\ C \end{pmatrix} = n$ *für alle* λ *mit* $|\lambda| > 1$.

Beweis: Mit Dualität. □

Ebenso kann man natürlich optimale Steuerung machen und Polvorgabe. Die Polvorgabe geht natürlich genau so wie vorher. Bei der optimalen Steuerung erhalten wir die folgenden Sätze.

Satz 8.24 *Betrachte das Optimalsteuerungsproblem der Minimierung von (8.3) mit Nebenbedingung (8.8). (also konstante Koeffizienten)*
Sei $\{u_k^*\}_{k=0}^{k_f}$ *die Optimalsteuerung und* $\{x_k^*\}_{k=0}^{k_f}$ *die zugehörige Lösung des geschlossenen Kreises, also die Lösung von*

$$x_{k+1} = Ax_k + Bu_k^*, \quad k = 0, 1, 2, \dots \quad x_0 = x^0 \tag{8.25}$$

Dann gibt es eine Kozustandsfolge $\{\mu_k\}_{k=0}^{k_f}$ *so daß* $\{x_k^*\}, \{\mu_k\}, \{u_k^*\}$ *das lineare Randwertproblem*

$$\begin{bmatrix} A & 0 & B \\ Q & -I & S \\ S^T & 0 & R \end{bmatrix} \begin{bmatrix} x_k \\ \mu_k \\ u_k \end{bmatrix} = \begin{bmatrix} I & 0 & 0 \\ 0 & -A^T & 0 \\ 0 & -B^T & 0 \end{bmatrix} \begin{bmatrix} x_{k+1} \\ \mu_{k+1} \\ u_{k+1} \end{bmatrix} \tag{8.26}$$

$$x_0 = x^0, \quad \mu_{k_f} = Mx_{k_f} \tag{8.27}$$

lösen.

Beweis: Vollkommen analog zum Beweis von Satz 6.4. □

Satz 8.28 *Seien* $\{x_k^*\}, \{\mu_k\}, \{u_k^*\}$ *die Lösungsfolgen für das lineare Randwertproblem (8.26), (8.27). Seien* $\mathcal{R} = \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix}$, *M positiv semidefinit. Dann gilt*

$$\mathcal{S}(\{x_k\}, \{u_k\}) \geq \mathcal{S}(\{x_k^*\}, \{u_k^*\}) \tag{8.29}$$

für alle $\{x_k\}, \{u_k\}$ *welche (8.8) lösen.*

Beweis: Vollkommen analog zum Beweis von Satz 6.21. □

In (8.26) kann man wieder nach $\{u_k\}$ auflösen und erhält:

$$\begin{bmatrix} A - BR^{-1}S^T & 0 \\ -(Q - SR^{-1}S^T) & I \end{bmatrix} \begin{bmatrix} x_k \\ \mu_k \end{bmatrix} = \begin{bmatrix} I & BR^{-1}B^T \\ 0 & (A - BR^{-1}S^T)^T \end{bmatrix} \begin{bmatrix} x_{k+1} \\ \mu_{k+1} \end{bmatrix} \tag{8.30}$$

oder äquivalent:

$$\begin{bmatrix} F & 0 \\ H & I \end{bmatrix} \begin{bmatrix} x_k \\ \mu_k \end{bmatrix} = \begin{bmatrix} I & -G \\ 0 & F^T \end{bmatrix} \begin{bmatrix} x_{k+1} \\ \mu_{k+1} \end{bmatrix} \tag{8.31}$$

Beweis: Vollkommen analog zum Beweis von Satz 6.20. □

Es gilt nun, daß

$$\begin{bmatrix} I & -G \\ 0 & F^T \end{bmatrix}^{-1} \begin{bmatrix} F & 0 \\ H & I \end{bmatrix} = \begin{bmatrix} F + GF^{-T}H & GF^{-T} \\ F^{-T}H & F^{-T} \end{bmatrix} \in \mathcal{S}_{2n}(\mathbb{R}) \tag{8.32}$$

und das Büschel in (8.31)

$$\alpha\mathcal{F} - \beta\mathcal{G} \equiv \alpha \begin{bmatrix} F & 0 \\ H & I \end{bmatrix} - \beta \begin{bmatrix} I & -G \\ 0 & F^T \end{bmatrix} \tag{8.33}$$

erfüllt

$$\mathcal{G}J\mathcal{G}^T = \mathcal{F}J\mathcal{F}^T \tag{8.34}$$

Büschel, die diese Gleichung erfüllen, heißen symplektische Büschel.

Man kann alle die Algorithmen und Ergebnisse für Hamiltonische Matrizen auf symplektische Matrizen oder Büschel übertragen. Allerdings ist die symplektische Struktur implizit gegeben und damit nicht so einfach zu überprüfen wie die Hamiltonische Struktur, die durch explizite Symmetrie gegeben ist.