Kontrolltheorie

Vorlesung TU Berlin
SS 2004, Volker Mehrmann

Literatur:

(a) G.E. Dullerud, F. Paganini
A Course in Robust Control Theory, Springer Verlag 1999

(b) G.H. Golub, C.F. Van Loan
Matrix Computation, 3. Aufl., Johns Hopkins University Press 1996

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

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

(e) A. Locatelli
Optimal Control, Birkhäuser Verlag 2001

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

(g) P.Hr. Petkov, N.D. Christov, M.M. Konstantinov
Computational Methods for Linear Control Systems, Prentice Hall 1991

(h) K. Zhou, J.C. Doyle, K. Glover
Robust and Optimal Control, Prentice Hall 1996

Inhaltsverzeichnis

Kapitel 0

Einleitung

Wir wollen uns in dieser Veranstaltung mit Kontrolltheorie (oder korrekter Steuerungstheorie oder Regelungstheorie), beschäftigen und zwar mit der mathematischen Theorie und entsprechenden numerischen Methoden dieses Gebietes. Dies Fachgebiet liegt auf der Grenze zwischen Mathematik, Ingenieurwissenschaften, Wirtschaftswissenschaften und Informatik.

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

0.1 Beispiele für Steuerungsprobleme

Wir wollen zuerst einige sehr einfache Beispiele betrachten.

Beispiel 0.1 Steuerung eines Zugs, modelliert als Massepunkt. Wir bezeichnen die Position zur Zeit t mit $s(t)$ und die entsprechende Geschwindigkeit mit $v(t)$ und den kombinierten Ort/Zeit-Vektor

$$\begin{bmatrix} s(t) \\ v(t) \end{bmatrix} = x(t)$$

als Zustand des Systems $x(t)$ zur Zeit t. Im allgemeinen haben wir einen gegebenen Anfangszeitpunkt t_0 und zugehörigen Anfangszustand

$$\begin{bmatrix} s(t_0) \\ v(t_0) \end{bmatrix} = x(t_0). \tag{1}$$

Als Steuerung haben wir eine (Beschleunigungs- bzw Brems-)Kraft

$$m\ddot{s}(t) = F(t) = u(t). \tag{2}$$

Die Gesamtkraft, die an den Zug angreift besteht aus mehreren Teilen;

- Ein Teil $u_f(s, v)$ hängt von Position und Geschwindigkeit des Zugs ab, die vom Zugführer nicht beeinflusst werden kann, wie z.B. Luftwiderstand oder Gravitationskräfte.
- Beschleunigungs- und Bremskräfte die durch den Zugführer beeinflusst werden, eingeschränkt durch die maximale Bremskraft u_{min} und die maximale Leistung P_{max} der Maschine.

Ein vereinfachtes Modell ergibt mit der Arbeit W erhalten wir die Leistung $P = \frac{\mathrm{d}W}{\mathrm{d}t}$ und mit $W = \int F \, \mathrm{d}s$ erhalten wir $P = Fv$. Mit konstanter Maximalleistung P_{max} der Maschine erhalten wir die maximale Beschleunigungskraft $u_{max}(v(t)) = P_{max}/v(t)$. Wir erhalten das Steuerungsproblem mit Nebenbedingungen

$$u = u_f(s, v) + u_v(t), \qquad u_v(t) \in [u_{min}, u_{max}(v(t))]. \tag{3}$$

Insgesamt haben wir das dynamische System

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{4}$$

oder in Matrixnotation

$$\begin{bmatrix} \dot{s}(t) \\ \dot{v}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} s(t) \\ v(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix} u(t), \tag{5}$$

mit Steuerbeschränkungen (3) und Zustandsbeschränkungen, wie

- einer Maximalgeschwindigkeit $v_{max}(s)$ in einem bestimmten Streckenbereich, so dass

$$v(s(t)) \in [0, v_{max}(s(t))] \quad \text{für } s \in [s^0, s^{end}], \tag{6}$$

- Fahrpläne, so dass der Zug einen bestimmten Bahnhof in einem Zeitintervall $t \in [t_a^i - \epsilon_a^i, t_a^i + \epsilon_a^i]$ so dass

$$s(t) = s^i \text{ and } u(t-0) < 0, \; v(t) = 0. \tag{7}$$

Um das Verhalten eines Zugs in einem Netzwerk zu simulieren brauchen wir Anfangsbedingungen für s, v und eine Steuerungsstrategie. Unglücklicherweise erhält man nur die Position an festen Stellen wo der Zug Sensoren passiert, aber man erhält im allgemeinen keine Informationen über die Geschwindigkeiten. D.h. wir erhalten keine Zustandsinformation sondern nur Ausgänge zu Zeiten $t = \bar{t}$.

$$y(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} x(t), \quad \text{für } t = \bar{t}. \tag{8}$$

Dabei sind diese Ausgänge oft nur ungenau ermittelbar.

Typische Kriterien solche Steuerungen zu entwickeln sind

- Zeit-optimale Steuerung
- Energie-optimale Steuerung

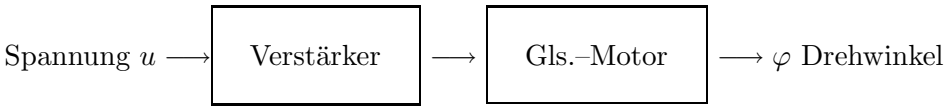
- Folgen eine Referenztrajektorie
- Minimale Abweichung vom Fahrplan
- Mixturen dieser Strategien.

□

Ein anderes Beispiel ist das folgende:

Beispiel 0.2 Steuerung einer Parabolantenne, die immer auf einen Satelliten oder ein Raumfahrzeug gerichtet ist. Wir brauchen ein einfaches Modell für den Motor, z.B. ein Gleichstrom-

motor, der über eine Eingangsspannung gesteuert wird.



Vereinfachte Bewegungsgleichungen: Sei t die Zeit.

$$\left\{\begin{array}{rcl} \frac{d\varphi}{dt} & = & \dot{\varphi}(t) = w(t) \\ j\dot{w}(t) & = & -rw(t) + ku(t) \end{array}\right. \qquad \begin{array}{ll} w & \text{Winkelgeschwindigkeit,} \\ j & \text{Trägheitsmoment des Motors,} \\ r & \text{Reibungskoeffizient,} \\ k & \text{Verstärkungsfaktor.} \end{array} \tag{3}$$

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. Wir müssen also eine Funktion $u(t)$ finden, so dass für die Lösung der Differentialgleichung (3) mit Anfangsbedingung

$$\varphi(t_0) = \varphi_0 \tag{4}$$

und diesem $u(t)$ gilt

$$\varphi(t_1) = \varphi_1. \tag{5}$$

Es wird im allgemeinen viele Lösungen geben, von denen wir eine aussuchen müssen, dazu brauchen wir wieder zusätzliche Kriterien, wie

- Zeit-optimale Steuerung,

- oder Energie-optimale Steuerung

so dass ingesamt

$$f(\varphi, u) = Min! \tag{6}$$

Falls wir das Optimalitätskriterium angegeben haben, müssen wir dann die folgenden Fragen behandeln.

- Zeige, dass es u gibt, so dass (4), (5), (6) gelten.
- Zeige, dass die Lösung für gegebenes u eindeutig ist.
- Wenn die Lösung nicht eindeutig ist, (6) modifizieren.
- Gibt es eine analytische Lösung/numerische Lösung, d.h. Bestimmung von $u(t)$, oft in Realzeit.
- Berechnung der Steuerung der Antenne.

□

Diese beiden Beispiele beschreiben zeit-kontinuierliche dynamische Systeme. In vielen Anwendungen, z.B. aus den Wirtschaftswissenschaften, sind die Systeme zeit-diskret, es gibt feste Zeitpunkte t_i, z.B. Monate, Tage, Jahre. Diskrete Systeme entstehen aber auch z.B. bei der Schrittweitensteuerung in der numerischen Lösung von Differentialgleichungen.

Beispiel 0.6 *Die Inflationsrate eines Landes wird in monatlichen Raten bestimmt. Um die Inflation niedrig zu halten kann die Zentralbank den Leitzins verändern. Sei also x_k die Inflationsrate im Monat k und sei u_k der Leitzins. Dann erhalten wir ein System*

$$x_{k+1} = f(x_k, u_k), \quad x_0 = x^0. \tag{7}$$

Wenn wir ein gutes mathematisches Modell für den Einfluss haben, d.h. die Funktion f, so können wir versuchen die Folge u_k so zu wählen, so dass die Inflationsrate unter einer politisch vorgegebenen Grenze bleibt.

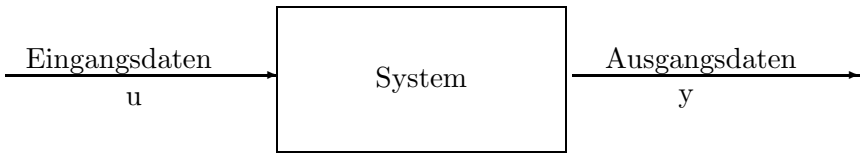
Es gibt noch viele andere Aufgaben in der Steuerungstheorie, aber dazu später.

Kapitel 1

Theoretische Grundlagen

Wir wollen nun die theoretischen Grundlagen betrachten.

Allgemein haben wir die folgende Beschreibung eines (zeit-kontinuierlichen) 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 genau im System passiert.

Beispiel 1.1 Beim Auto Fahren haben wir wenige Eingangsparameter, wie Gas geben, bremsen und lenken, und wir beobachten den Tacho, Drehzahlmesser, oder vielleicht noch die Temperaturanzeige, sowie die Umgebung des Autos, aber wir wissen im allgemeinen nicht was im Motor oder Getriebe passiert. □

Um ein System vernünftig beschreiben zu können, brauchen wir ein mathematisches Modell, z.B. eine Differentialgleichung (oder Differenzengleichung), das möglichst gut an die Realität angepasst sein soll, z.B.

$$\begin{aligned} \dot{x} &= f(x,u), \quad x(t_0) = x_0, \\ y &= g(x,u), \end{aligned} \tag{1.1}$$

wobei u eine Steuerfunktion ist, x den Systemzustand beschreibt und y die beobachtbaren oder messbaren Ausgangsgrößen. Solche Modelle sind im allgemeinen nie exakt und hängen

von ganz vielen (oft unbekannten) Parametern ab. Typische Fragen, die wir beantworten müssen sind:

Was sind Zustands-, Steuer- oder Ausgangsgrößen?

Diskrete Zeit, kontinuierliche Zeit?

Welche Größen können vernachlässigt werden?

Wie ist der funktionale Zusammenhang?

Dies ist i.a. der schwierigste Teil, weil hier die Modellierung eine interdisziplinäre Zusammenarbeit erfordert. Modellierung können wir hier aus Zeitgründen nicht behandeln. Wir betrachten in dieser Vorlesung im wesentlichen lineare zeit-kontinuierliche oder zeit-diskrete Steuerungsprobleme.

Definition 1.2 *Ein* (zeit-kontinuierliches) lineares Steuerungsproblem *hat die Form*

$$\begin{aligned} \dot{x} &= \frac{dx}{dt} = A(t)x(t) + B(t)u(t), t \in [t_0, \infty) \qquad \text{Zustandsgleichung ,} && (1.2)\\ x(t_0) &= x^0 \qquad\qquad \text{Anfangswert,} && (1.3)\\ y(t) &= C(t)x(t) + D(t)u(t), t \in [t_0, \infty) \qquad \text{Ausgangsgleichung} && (1.4) \end{aligned}$$

Dabei sind

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

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

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

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

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

In vielen Fällen ist $D = 0$, da sonst die Steuerung direkt den Ausgang beeinflusst. Wir werden in dieser Vorlesung immer $D = 0$ betrachten.

Im diskreten Fall haben wir:

Definition 1.3 *Ein* (diskretes) lineares Steuerungsproblem *hat die Form*

$$\begin{aligned} x_{k+1} &= A_k x_k + B_k u_k, \quad k = 0,1,2,... \quad \textit{Zustandsgleichung,} && (1.5)\\ x_0 &= x^0, \qquad\qquad\qquad\qquad \textit{Anfangswert,} && (1.6)\\ y_k &= C_k x_k + D_k u_k, \quad k = 0,1,2,... \quad \textit{Ausgangsgleichung.} && (1.7) \end{aligned}$$

Dabei sind

$x = (x_0, x_1, x_2, \ldots)$ *Zustand,* $x \in \mathcal{X}_d$ *Zustandsraum,*

$y = (y_0, y_1, y_2, \ldots)$ *Ausgang,* $y \in \mathcal{Y}_d$ *Ausgangsraum,*

$u = (u_0, u_1, u_2, \ldots)$ *Eingang,* $u \in \mathcal{U}_d$ *Eingangsraum.*

Die Räume $\mathcal{X}_d$, $\mathcal{Y}_d$ *und* $\mathcal{U}_d$ *sind Mengen von Folgen und die Systemmatrizen bilden Folgen* $(A_0, A_1, \ldots)$, $(B_0, B_1, \ldots)$, $(C_0, C_1, \ldots)$ *und* $(D_0, D_1, \ldots)$. *Die Dimension sind dabei wie im kontinuierlichen Fall.*

Wir werden auch im diskreten Fall in dieser Vorlesung immer $D_k = 0$, $k = 0, 1, 2, \ldots$ betrachten.

Eine wichtige Spezialklasse der Systeme, die wir intensiv betrachten werden, sind die *zeitinvarianten* Systeme. Im linearen kontinuierlichen Fall bedeutet dies, dass

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

Um Steuerungsprobleme mathematisch zu analysieren, werden *Zustandsübergangsfunktionen* bzw. *Ausgangsübergangsfunktionen* eingeführt. Wir betrachten zuerst wieder kontinuierliche Systeme.

Definition 1.4 *Gegeben* $t_0 \in \mathbb{R}$, $x^0 \in \mathbb{R}^n$ *und* $u \in \mathcal{U}_c$, *so definiert man als* Zustandsübergangsfunktion *die Abbildung*

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

wobei $x(t)$ *die Lösung der Zustandsgleichung (1.2) mit Anfangswert (1.3) zum Zeitpunkt* t *ist, und als* Ausgangsübergangsfunktion *die Abbildung*

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

wobei $y(t)$ *das Ergebnis von (1.4) ist.*

Die Übergangsfunktion heißt auch oft Transferfunktion *oder* Input–Output Relation.

Beispiel 1.5 Betrachte ein elektrisches Netzwerk, bestehend aus einem Widerstand und einem Kondensator.

Der Eingang ist die angelegte Spannung u; Ausgang und Zustand ist die Ladung $q(t)$ am Kondensator.

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

$$u(t) = R\dot{q}(t) + \frac{q(t)}{C}, \tag{1.8}$$

oder umgeformt:

$$\dot{q}(t) = -\frac{1}{RC}q(t) + \frac{1}{R}u(t) \tag{1.9}$$

und damit

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

Der Zustand $q(t)$ kann durch Lösung der Differentialgleichung (1.8) bestimmt werden und es gilt

$$q(t) = e^{-(t-t_0)/RC}q(t_0) + \frac{1}{R}\int_{t_0}^{t} e^{-(t-s)/RC}u(s)ds. \tag{1.10}$$

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

$$\big(t_0, t, q^0, u(t)\big) \longrightarrow e^{-(t-t_0)/RC}q^0 + \frac{1}{R}\int_{t_0}^{t} e^{-(t-s)/RC}u(s)ds. \tag{1.11}$$

Hier ist die Ausgangsfunktion gleich der Zustandsfunktion.

In diesem Beispiel sieht man auch gut, dass wir bei der Modellbildung vieles vernachlässigt haben, denn im allgemeinen ist z. B. R nicht konstant. □

Für allgemeine lineare Systeme erhalten wir:

Satz 1.6 *Die Zustandsübergangsfunktion zum System (1.2) (1.3) ist gegeben durch*

$$x(t) = \Phi(t, t_0)x^0 + \int_{t_0}^{t} \Phi(t, s)B(s)u(s)ds, \tag{1.12}$$

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

$$\frac{\partial \Phi}{\partial t}(t, s) = A(t)\Phi(t, s), \quad \Phi(s, s) = I \tag{1.13}$$

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

$$x(t) = e^{A(t-t_0)}x^0 + \int_{t_0}^{t} e^{A(t-s)}Bu(s)ds. \tag{1.14}$$

Beweis: Beide Formeln (1.12) und (1.14) folgen sofort durch Ableiten. □

Im Fall diskreter Systeme erhalten wir

Satz 1.7 *Die Zustandsübergangsfolge zum System (1.5), (1.6) ist gegeben durch*

$$x_k = (\prod_{l=0}^{k-1} A_l)x^0 + \sum_{j=0}^{k-1}(\prod_{l=j+1}^{k-1} A_l)B_j u_j \tag{1.15}$$

und die Ausgangsübergangsfolge zum System (1.5)–(1.7) ist gegeben durch

$$y_k = C_k\left((\prod_{l=0}^{k-1} A_l)x^0 + \sum_{j=0}^{k-1}(\prod_{l=j+1}^{k-1} A_l)B_j u_j\right) + D_k u_k. \tag{1.16}$$

Beweis: Formeln (1.15) und (1.16) folgen sofort durch Einsetzen. □

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

Definition 1.8

i) Das kontinuierliche System (1.2), (1.3) heißt c-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 c-stabil*, falls außerdem* $\lim_{t\to\infty} ||x(t)|| = 0$ *gilt, wobei* $||\ ||$ *irgendeine Norm ist.*

ii) Das diskrete System (1.5), (1.6) heißt d-stabil*, falls eine (und damit jede) Lösung der homogenen Gleichung* $x_{k+1} = A_k x_k$ *für* $k \to \infty$ *beschränkt ist. Das System heißt* asymptotisch d-stabil*, falls außerdem* $\lim_{k\to\infty} ||x_k|| = 0$ *gilt, wobei* $||\ ||$ *irgendeine Norm ist.*

Die Charakterisierung von stabilen Systemen ist eine wichtige Aufgabe der Steuerungstheorie und wir werden dieses Thema später genauer betrachten.

Satz 1.9 *Betrachte das zeitinvariante kontinuierliche System*

$$\dot{x} = Ax + Bu,\ x(t_0) = x^0, \tag{1.17}$$

mit A, B *konstant.*

i) Das System (1.17) ist asymptotisch c-stabil genau dann, wenn alle Eigenwerte von A *negativen Realteil besitzen.*

ii) Das System (1.17) ist c-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: Beweis mit Hilfe der Jordan'schen Normalform. □

Satz 1.10 *Betrachte das zeitinvariante System*

$$x_{k+1} = Ax_k + Bu_k,\ x_0 = x^0, \tag{1.18}$$

mit A, B *konstant.*

i) Das System (1.10)ist asymptotisch c-stabil genau dann, wenn alle Eigenwerte von A einen Betrag < 1 *besitzen.*

ii) Das System (1.10) ist c-stabil genau dann, wenn alle Eigenwerte von A einen Betrag ≤ 1 *haben und alle Eigenwerte mit Betrag* 1 *gleiche algebraische wie geometrische Vielfachheit haben.*

Beweis: Beweis mit Hilfe der Jordan'schen Normalform. □

Beispiel 1.11 *a) (Bsp 1.5)*

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

ist skalar, A ist konstant $-\frac{1}{RC} < 0$ *also ist das System asymptotisch c-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*

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

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

1.1 Steuerungsprobleme im Frequenzraum

Wenn man die meisten Bücher zur Regelungstechnik anschaut, sieht man die Steuerungstheorie nicht im Zustandsraum dargestellt, sondern im Frequenzraum.

Die Idee dabei ist im wesentlichen, dass man das System Laplace- oder Z-transformiert und dann alles in den transformierten Variablen weiterrechnet. Dies geht jedoch nur für zeitinvariante Systeme. Dazu werden wir erst kurz die Laplacetransformation wiederholen.

Definition 1.12 *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 Laplacetransformation *von f.*

Es gelten folgende Rechenregeln

$$\begin{cases} \mathcal{L}(af(t) + bg(t)) &= a\mathcal{L}(f(t)) + b\mathcal{L}(g(t)), \\ \mathcal{L}(f'(t)) &= s\mathcal{L}(f(t)) - f(0), \\ \mathcal{L}(\int_0^t f(z)dz) &= \frac{1}{s}\mathcal{L}(f(t)). \end{cases} \tag{1.19}$$

Betrachte nun das zeitinvariante System

$$\begin{cases} \dot{x}(t) &= Ax(t) + Bu(t), \quad x(0) = x^0, \\ y(t) &= Cx(t) + Du(t) \quad A, B, C, D \text{ konstant.} \end{cases} \tag{1.20}$$

Die Laplacetransformation ergibt

$$\begin{aligned} s\hat{x} &= A\hat{x} + B\hat{u} + x^0, \\ \hat{y} &= C\hat{x} + D\hat{u}, \end{aligned} \tag{1.21}$$

und damit folgt

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

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.

Im diskreten Fall verwenden wir für Folgen $x = (x_0, x_1, \ldots)$ den Shift-Operator $z(x_0, x_1, \ldots) = (x_1, x_2, \ldots)$ und erhalten für die Lösungsfolgen die Gleichungen

$$\begin{aligned} zx &= Ax + Bu, \\ y &= Cx + Du, \end{aligned}$$

und damit

$$\begin{aligned} x &= (zI - A)^{-1}Bu \\ y &= (C(zI - A)^{-1}B + D)u, \end{aligned}$$

Die Matrizen $H(s)$ bzw. $H(z)$ heißen *Übertragungsmatrizen.*

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

1.2 Steuerbarkeit

Eine der wichtigsten Fragestellungen der Steuerungstheorie ist die Frage, ob man ein System durch geeignete Wahl von $u(t)$ bzw $\{u_k\}$ von einem Zustand 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.

1.2.1 Steuerbarkeit für lineare zeit-kontinuierliche Systeme

Wir betrachten zunächst das zeit-kontinuierliche System

$$\dot{x} = A(t)x + B(t)u, \quad x(t_0) = x^0 \tag{1.22}$$

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

Definition 1.13 *Gegeben sei das System (1.22) 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 $u \in \mathcal{U}$ gibt, so dass die Lösung von (1.22) 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 (1.22) vollständig steuerbar*.*

Schematisch stellen wir System (1.22) dar wie in Abbildung 1.1.

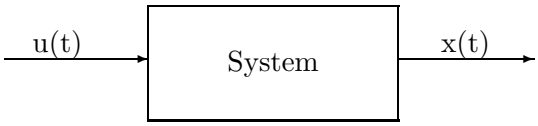


Abbildung 1.1: Offener Kreis (Steuerung)

Dies ist ein sogenannter *offener Kreis*. Nun wählt aber man typischerweise $u(t)$ in Abhängigkeit vom Zustand, d.h. man macht eine *Zustandsrückführung (feedback)* wie in Abbildung 1.2. Im linearen Fall macht man sogar gerne eine lineare Zustandrückführung $u(t) = -F(t)x(t)$

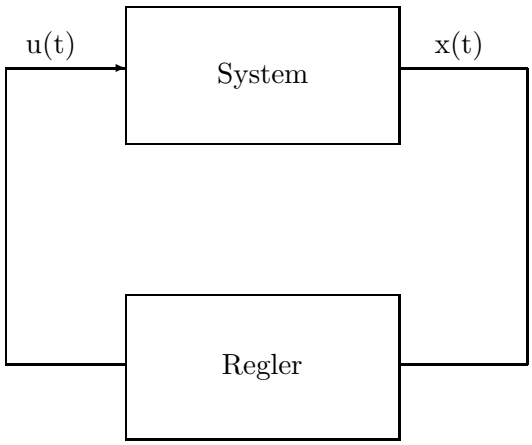


Abbildung 1.2: Geschlossener Kreis (Regelung)

bzw. $u_k = -F_k x_k$ und erhält dann für das geschlossene System die homogene Differentialgleichung

$$\dot{x} = (A(t) - B(t)F(t))\, x(t) \tag{1.23}$$

bzw. die homogene Differenzengleichung

$$x_{k+1} = (A_k - B_k F_k)\, x_k. \tag{1.24}$$

Um zu klären, wann ein System steuerbar ist, 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 (1.12) gelten:

$$0 = x(t_1) = \Phi(t_1, t_0)x^0 + \int_{t_0}^{t_1} \Phi(t_1, s)B(s)u(s)ds. \tag{1.25}$$

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

$$0 = x^0 + \int_{t_0}^{t_1} \Phi(t_0, s)B(s)u(s)\, ds, \tag{1.26}$$

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

$$\Phi(t_1, s) = \Phi(t_1, t_0)\Phi(t_0, s). \tag{1.27}$$

Dies heißt *Halbgruppeneigenschaft der Lösung.* Also gilt $x^0 \in \mathcal{L}(t_0, t_1)$ genau dann, wenn (1.26) gilt.

Die Funktion $\Phi(t_1, s)$ beschreibt den Übergang von $s \to t_1$ und $\Phi(t_0, s)$ den Übergang von $s \to t_0$ und da die Lösung von (1.13) eindeutig ist, so folgt (1.27).

Beispiel 1.14 Im zeitinvarianten Fall ergibt sich:

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

□

Bevor wir zu einer Charakterisierung von $\mathcal{L}(t_0, t_1)$ kommen, führen wir noch die *Gram'sche Matrix* eines Systems ein.

Es sei $G(t)$ eine auf $(-\infty, +\infty)$ definierte matrixwertige, stückweise stetige Funktion in $\mathbb{R}^{n,m}$. Dann heißt

$$V = \int_{t_0}^{t_1} G(t)G(t)^T dt \tag{1.28}$$

für $t_0 < t_1$ *Gram'sche Matrix.* Wir haben das folgende Lemma.

Lemma 1.15 *Ein $x \in \mathbb{R}^n$ lässt sich genau dann als*

$$x = \int_{t_0}^{t_1} G(t)u(t)dt \tag{1.29}$$

mit einer stückweise stetigen Funktion $u(t) \in \mathbb{R}^m$ schreiben, wenn x im Bild von V (wie in (1.28)) liegt, d.h. es gibt ein $z \in \mathbb{R}^n$, so dass $x = Vz$.

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

$$x^T V x = \int_{t_0}^{t_1} x^T G(t)G(t)^T x dt = \int_{t_0}^{t_1} ||G(t)^T x||_2^2 dt \geq 0. \tag{1.30}$$

Es gilt daher

$$Vx = 0 \Longleftrightarrow G(t)^T x = 0 \quad \forall t \in [t_0, t_1]. \tag{1.31}$$

Die Menge aller x, die sich mit geeignetem $u(t)$ in der Form (1.29) schreiben lassen, bilden wegen der Linearität des Integrals und der Multiplikation mit $G(t)$ einen linearen Raum L. Der Bildraum von V ist in L enthalten. $\text{Bild}(V) \subset L$. Dies folgt sofort aus der Definition von V mit der Wahl

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

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

$$||x||_2^2 = x^T x = \int_{t_0}^{t_1} x^T (G(t)u(t))dt = \int_{t_0}^{t_1} \left(G(t)^T x\right)^T u(t)dt = 0. \tag{1.32}$$

Also folgt $x = 0$. □

Wir führen nun die Gram'sche Matrix

$$W(t_0, t_1) := \int_{t_0}^{t_1} \Phi(t_0, t)B(t)B(t)^T \Phi(t_0, t)^T dt \tag{1.33}$$

ein und erhalten den folgenden Satz:

Satz 1.16 *Für die Matrix W aus (1.33) gilt:*

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

$$\mathcal{L}(t_0, t_1) = \{x \mid \text{ es gibt ein } z \text{ mit } x = W(t_0, t_1)z\}. \tag{1.34}$$

(ii) Es gilt $W(t_0, t_1)x = 0$ genau dann, wenn $x^T \Phi(t_0, t)B(t) \equiv 0$ für alle $t \in [t_0, t_1]$.

Beweis: Verwende Lemma 1.15 (sowie (1.31)) mit $G(t) = \Phi(t_0, t)B(t)$. □

Mit diesem Satz können wir nun die Steuerbarkeit charakterisieren. Dazu brauchen wir noch die zu $\dot{x} = A(t)x$ *adjungierte Gleichung*

$$\dot{z} = -A(t)^T z. \tag{1.35}$$

Satz 1.17 *Das System (1.22) ist vollständig steuerbar genau dann, wenn folgende Aussage gilt:*
Ist $z(t)$ eine nichttriviale Lösung von (1.35), so verschwindet $z(t)^T B(t)$ auf keinem Intervall $[t_0, \infty)$ identisch.

Beweis: Sei (1.22) vollständig steuerbar. Angenommen, es gibt eine Lösung $z(t) \not\equiv 0$ von (1.35), so dass

$$z(t)^T B(t) = 0 \quad \forall t \geq t_0. \tag{1.36}$$

Dann gibt es Anfgangswerte x^0, so dass (t_0, x_0) in keiner Zeit $t_1 > t_0$ steuerbar ist. Dies folgt sofort, denn, ist $x(t)$ eine Lösung von (1.22) für ein $u(t)$, so gilt

$$\frac{d}{dt}\left(x(t)^T z(t)\right) = 0 \quad \forall t \geq t_0, \tag{1.37}$$

d.h. $x(t)^T z(t) = {x^0}^T z^0$ konstant, wobei $x^0 = x(t_0), z^0 = z(t_0)$. Man braucht nun x^0 nur so zu wählen, dass ${x^0}^T z^0 \neq 0$, so ergibt sich ein Widerspruch.

Die Umkehrung ist komplizierter. Das geschieht in mehreren Schritten.

Wir beweisen zuerst: Zu jedem t_0 gibt es $t_1 > t_0$, so dass für jede nichttriviale Lösung von (1.35) gilt, dass $z(t)^T B(t) \not\equiv 0$ in $[t_0, t_1]$. Wäre dies falsch, so gäbe es eine Folge $t_v \to \infty$ und zugehörige Lösungen $z_v(t)$ von (1.35) mit

$$||z_v(t_0)|| = 1,\ z_v^T(t)B(t) = 0,\ \ \forall t \in [t_0, t_v] \tag{1.38}$$

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 (1.35) mit $z_\infty(t_0) = z_\infty^*$ nichttrivial. 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 (1.38).

Als zweiten Schritt zeigen wir nun: Wenn für jede nichttriviale Lösung $z(t)$ von (1.35), $z(t)^T B(t)$ 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. Angenommen, es gibt $x \neq 0$, so dass $W(t_0, t_1)x = 0$. Dann folgt $B(t)^T \Phi(t_0, t)^T x \equiv 0$ für alle $t \in [t_0, t_1]$. Setze $z(t) = \Phi(t_0, t)^T x$, so löst $z(t)$ die adjungierte Gleichung (1.35), da Φ die Fundamentallösung ist, und damit folgt als Widerspruch $x = 0$.

Als letzten Schritt ergibt sich, dass wenn $W(t_0, t_1)$ positiv definit ist, so ist jedes Paar (t_0, x^0) in jedem x^1 zur Zeit t_1 steuerbar und zwar mit

$$u(t) = B(t)^T \Phi(t_0, t)^T c. \tag{1.39}$$

Dabei ist c 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. □

Korollar 1.18 *Das System (1.22) ist vollständig 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.*

Im zeitinvarianten Fall gibt es rein algebraische Charakterisierungen der Steuerbarkeit.

Satz 1.19 *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* Steuerbarkeitsmatrix

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

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

Beweis: Aus Satz 1.16 folgt

$$\mathcal{L}(t_0, t_1) = \text{Bild}\left(W(t_0, t_1)\right)$$

genau dann wenn

$$\mathcal{L}(t_0, t_1)^{\perp} = \text{Kern}\left(W(t_0, t_1)\right).$$

Wir müssen also zeigen

$$W(t_0, t_1)x = 0 \Longleftrightarrow x^T K = 0. \tag{1.40}$$

Satz 1.16 ii) besagt, dass $W(t_0, t_1)x = 0$ genau dann, wenn

$$x^T e^{At} B = 0 \quad \forall t \in [t_0 - t_1, 0] \tag{1.41}$$

und dies gilt genau dann, wenn

$$x^T A^v B = 0 \quad \forall v = 0, 1, 2, ...$$

Der Satz von Cayley–Hamilton sagt, dass $\varphi(A) = 0$, wobei φ das charakteristisches Polynom von A ist. 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 $x^T A^v B = 0 \quad \forall v = 0, 1, 2, 3...$ oder äquivalent

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

□

Wir fassen nun den zeitinvarianten Fall zusammen:

Satz 1.20 *Die folgenden Aussagen sind äquivalent:*

i) Das zeitinvariante System $\dot{x} = Ax + Bu$ ist vollständig 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: Die Äquivalenz von i) und ii) haben wir schon gezeigt. Die Äquivalenz von iii) und iv) ist trivial. Zeige noch, dass ii) äquivalent zu iii). Falls $p \neq 0$ ein Eigenvektor zu A^T ist und $p^T B = 0$, so folgt natürlich $p^T K = 0$ und damit ist der Rang von K nicht voll. Umgekehrt folgt, dass falls es $p \neq 0$ gibt mit $p^T K = 0$, so folgt $p^T A^j B = 0$ für alle $j = 0, 1, 2, \ldots$ und damit ist p Linkseigenvektor von A, ein Widerspruch zu iii). □

Beispiel 1.21 Wir betrachten wieder unsere Beispiele.

Beispiel 1.5: Da $A = \frac{-1}{RC}, B = \frac{1}{R}$ skalar $\neq 0$, so folgt Rang $\left(\frac{1}{R}\right) = 1$ also ist das System 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. □

1.2.2 Steuerbarkeit für diskrete Systeme

Für diskrete Systeme gelten ähnliche Ergebnisse. Führe in (1.15) die Abkürzung $\Phi_d(k, j) = \prod_{l=j}^{k-1} A_l$ ein und setze voraus, dass alle Systemmatrizen A_j invertierbar sind.

Wir betrachten wieder erst eine etwas andere Aufgabe.

Gegeben $k \in \mathbb{N}$, bestimme alle x^0, so dass $(0, x^0)$ nach $(k, 0)$ steuerbar ist. Sei $\mathcal{L}_d(0, k)$ die Menge dieser gesuchten x^0. Es muss also nach (1.15) gelten:

$$0 = x_k = \Phi_d(k, 0)x_0 + \sum_{j=0}^{k-1} \Phi_d(k-1, j+1)B_j u_j. \tag{1.42}$$

Ausklammern von $\Phi_d(k, 0)$ nach links ergibt

$$0 = x^0 + \sum_{j=0}^{k-1} \Phi_d(j, 0)^{-1} B_j u_j. \tag{1.43}$$

Auch bei diskreten Systemen haben wir eine analoge *Gram'sche Matrix*.

Es sei $\{G_j\}$ eine Matrixfolge in $\mathbb{R}^{n,m}$. Dann heißt

$$V = \sum_{j=0}^{k-1} G_j G_j^T \tag{1.44}$$

für $k > 0$ *Gram'sche Matrix*. Wir haben das folgende Lemma.

Lemma 1.22 *Ein $x \in \mathbb{R}^n$ lässt sich genau dann als*

$$x = \sum_{j=0}^{k-1} G_j u_j \tag{1.45}$$

mit einer Folge $\{u_j\}, u_j \in \mathbb{R}^m$ schreiben, wenn x im Bild von V liegt.

Beweis: Übungsaufgabe. □

Wir führen nun wieder die Gram'sche Matrix

$$W_d(0,k) := \sum_{j=0}^{k-1} \Phi_d(j,0) B_j B_j^T \Phi_d(j,0)^T \tag{1.46}$$

ein und erhalten den folgenden Satz:

Satz 1.23 *Für die Matrix W_d aus (1.46) gilt:*

(i) Die Menge $\mathcal{L}_d(0,k)$ ist der Bildraum der Matrix $W_d(0,k)$..

(ii) Es gilt $W_d(0,k)x = 0$ genau dann, wenn $x^T\Phi_d(0,j)B_j$ für alle $j = 0,\dots,k-1$.

Beweis: Übungsaufgabe. □

Mit diesem Satz können wir nun die Steuerbarkeit charakterisieren. Dazu brauchen wir wieder die zu $x_{k+1} = A_k x_k$ *adjungierte Gleichung*

$$z_{k+1} = -A_k^T z_k. \tag{1.47}$$

Satz 1.24 *Das System (1.5) ist vollständig steuerbar genau dann, wenn folgende Aussage gilt:*
Ist $\{z_j\}$ eine nichttriviale Lösung von (1.47), so verschwindet $z_j^T B_j$ nicht für alle $j = 0,\dots,k$.

Beweis: Übungsaufgabe. □

Korollar 1.25 *Das System (1.5) ist vollständig steuerbar genau dann, wenn es ein $k \in \mathbb{N}$ gibt, so dass $W_d(0,k)$ postiv definit ist. Es ist dann in dieser Zeit k in jedes x steuerbar.*

Im zeitinvarianten Fall gibt es auch wieder rein algebraische Charakterisierungen der Steuerbarkeit.

Satz 1.26 *Für ein zeitinvariantes System $x_{k+1} = Ax_k + Bu_k$ ist $\mathcal{L}_d(0,k)$ von k unabhängig und gleich dem von den Spalten der* Steuerbarkeitsmatrix

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

aufgespannten Raum, d.h. $\mathcal{L}_d(0,k) = \mathrm{Bild}(K)$.

Beweis: Übungsaufgabe. □

Wir fassen den zeitinvarianten Fall zusammen:

Satz 1.27 *Die folgenden Aussagen sind äquivalent:*

i) Das zeitinvariante System $x_{k+1} = Ax_k + Bu_k$ 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. □

1.3 Stabilisierbarkeit

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, aber kommt man mit weniger aus?

1.3.1 Stabilisierbarkeit bei kontinuierlichen Systemen

Definition 1.28 *Ein lineares System der Form (1.22) heißt* c-stabilisierbar*, 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 (1.22) mit diesem* $u(t)$ *die Bedingung* $\lim\limits_{t\to\infty} x(t) = 0$ *erfüllt. (Besser wäre hier der Begriff asymptotisch c-stabilisierbar.)*

Wir geben zuerst ein notwendiges Kriterium für Stabilisierbarkeit.

Satz 1.29 *Wenn das System (1.22) c-stabilisierbar ist, so gilt: Ist* $z(t)$ *eine Lösung der adjungierten Gleichung (1.35) 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 (1.22) ist c-stabilisierbar, es gibt jedoch eine nichttriviale Lösung von (1.35) für die gilt:

$$z(t)^T B(t) \equiv 0 \quad \forall t \geq t_0 \quad \lim_{t\to\infty} ||z(t)|| \neq \infty \tag{1.48}$$

Wie beim Beweis von Satz 1.17 gilt dann für jede Wahl von $u(t)$ und die zugehörige Lösung von (1.22), dass

$$x(t)^T z(t) = x_0^T z(t_0) \quad \forall t \geq t_0.$$

Aus $\lim\limits_{t\to\infty} ||x(t)|| = 0$ und $\lim\limits_{t\to\infty} ||z(t)|| \neq \infty$ folgt, dass es eine Folge $t_i \to \infty$ gibt, so dass $\lim\limits_{i\to\infty} x(t_i)^T z(t_i) = 0$. 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 c-Stabilisierbarkeit. □

Allgemein ist eine hinreichende Bedingung für dieses Problem nicht bekannt.

Im zeitinvarianten Fall erhalten wir:

Satz 1.30 *Die folgenden Bedingungen sind äquivalent:*

i) Das System $\dot{x} = Ax + Bu$ mit A, B konstant, ist c-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 $Re(e^{-\lambda t})p, Im(e^{-\lambda t})p$ (1.35) und sind beschränkt. Satz 1.29 impliziert $p^T B \neq 0$.
ii)$\Longrightarrow$ iii) Angenommen es gibt ein $\lambda \in \mathbb{C}, Re(\lambda) \geq 0$, so dass Rang $(\lambda I - A, B) < n$. Dann gibt es ein $p \neq 0$, so dass $p^T(\lambda I - A, B) = 0$, also ist p Linkseigenvektor von A und $p^T B = 0$.
iii) $\Longrightarrow$ i) Rang $(\lambda I - A, B) = n \; \forall \lambda$ mit $Re(\lambda) \geq 0$. Dann gibt es 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 $p^T(A - BG) = 0 \quad \forall G \in \mathbb{R}^{m,n}$, das widerspricht iii). Also stabilisiert $u(t) = Fx(t)$ das System. □

1.3.2 Stabilisierbarkeit bei diskreten Systemen

Definition 1.31 *Ein lineares System der Form (1.5) heißt* d-stabilisierbar*, wenn es zu jedem x_0 eine Folge $\{u_k\}$ gibt, so dass die Lösung von (1.5) mit dieser Eingangsfolge die Bedingung $\lim_{k\to\infty} x_k = 0$ erfüllt.*

Wir erhalten wieder ein notwendiges Kriterium für d-Stabilisierbarkeit.

Satz 1.32 *Wenn das System (1.5) d-stabilisierbar ist, so gilt: Ist $\{z_j\}$ eine Lösung der adjungierten Gleichung (1.47) die für $k \to \infty$ beschränkt ist, so kann nicht gelten $z_k^T B_k \equiv 0$ für alle $k \geq 0$.*

Beweis: Übungsaufgabe. □

Allgemein ist auch hier eine hinreichende Bedingung für dieses Problem nicht bekannt.

Im zeitinvarianten Fall erhalten wir analog:

Satz 1.33 *Die folgenden Bedingungen sind äquivalent:*

i) Das System $x_{k+1} = Ax_k + Bu_k$ mit A, B konstant, ist d-stabilisierbar;

ii) Ist λ mit $|\lambda| \geq 1$ Eigenwert von A mit Linkseigenvektor p, so gilt $p^T B \neq 0$;

iii) Rang $(\lambda I - A, B) = n \quad \forall \lambda$ mit $|\lambda| \geq 1$.

Beweis: Übungsaufgabe. □

Hier stellen sich gleich einige wichtige Fragen. Wie überprüfe ich die Steuerbarkeits- und Stabilisierbarkeitsbedingungen praktisch, z.B. mit einem numerischen Verfahren. Wie berechne ich entsprechende Steuerungen, die das System in einen gewünschten Zustand bringen oder das System stabilisieren. Und da wir ja davon ausgehen müssen, dass unsere Modelle mit Fehlern behaftet sind, wie sichere ich, dass meine berechnete Steuerung robust gegenüber kleinen Störungen ist. Diese Fragen werden wir später im Detail behandeln.

1.4 Rekonstruierbarkeit und Entdeckbarkeit

Wie wir schon bemerkt haben, können wir im allgemeinen nicht den ganzen Zustand eines System zur Steuerung verwenden, sondern haben nur Ausgangsgrößen. Können wir das System nun auf der Basis dieser partiellen Information immer noch in einen beliebigen Zustand oder in ein Gleichgewicht steuern?

1.4.1 Rekonstruierbarkeit und Entdeckbarkeit für kontinuierliche Systeme

Wir betrachten zeit-kontinuierliche Systeme mit Ausgang.

$$\dot{x}(t) = A(t)x(t) + B(t)u(t),\ y(t) = C(t)x(t),\ x(t_0) = x^0. \tag{1.49}$$

Definition 1.50 *Das System (1.49) heißt* rekonstruierbar*, falls folgende Aussage für alle t_0 gilt. Sind $x, \tilde{x}$ Lösungen von (1.49) mit der gleichen Steuerfunktion $u(t)$ und gilt*

$$C(t)x(t) = C(t)\tilde{x}(t) \quad \forall t \leq t_0, \tag{1.50}$$

so gilt

$$x(t) = \tilde{x}(t) \quad \forall t \leq t_0. \tag{1.51}$$

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, dazu bilde $x - \tilde{x}$.
Man spricht in vielen Lehrbüchern von *Beobachtbarkeit (Observability)* anstatt Rekonstruierbarkeit. Es gibt jedoch für zeitvariable Syteme eine Untersschied.

Definition 1.51 *Das System (1.49) heißt* beobachtbar*, falls folgende Aussage für alle t_0 gilt. Sind $x, \tilde{x}$ Lösungen von (1.49) mit der gleichen Steuerfunktion $u(t)$ und gilt*

$$C(t)x(t) = C(t)\tilde{x}(t) \quad \forall t \geq t_0 \tag{1.52}$$

so gilt

$$x(t) = \tilde{x}(t) \quad \forall t \geq t_0 \tag{1.53}$$

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

Die Konzepte der Rekonstruierbarkeit und Steuerbarkeit sind in gewissem Sinne dual, denn haben wir den folgenden Satz.

Satz 1.52 *Das System (1.49) ist rekonstruierbar genau dann, wenn das* duale System

$$\dot{x} = A(-t)^T x + C(-t)^T u \tag{1.54}$$

vollständig steuerbar ist.

Beweis: Bilde $z = \tilde{x} - x$. Es ist klar, dass (1.49) rekonstruierbar ist, genau dann wenn $C(t)z(t) = 0$ für alle $t \leq t_0$ impliziert, dass $z(t) = 0$ für alle $t \leq t_0$, 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 nach Satz 1.17 äquivalent damit, dass (1.54) vollständig steuerbar ist. □

Wir erhalten daher auch sofort Kriterien mit Hilfe der Übergangsmatrix:

Korollar 1.53 *System (1.49) ist rekonstruierbar genau dann, wenn es zu jedem t_1 ein $t_0 < t_1$ gibt, so dass*

$$\hat{W}(t_0, t_1) := \int_{t_0}^{t_1} \Phi(t, t_1)^T C(t)^T C(t) \Phi(t, t_1) dt \tag{1.55}$$

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 $x(t_1) = x^1$:

$$\int_{t_0}^{t_1} \Phi(t, t_1)^T C(t)^T y(t) dt = \hat{W}(t_0, t_1) x^1 + \int_{t_0}^{t_1} \left(\Phi(t, t_1)^T C(t)^T C(t) \int_{t_1}^{t} \Phi(t, s) B(s) u(s) ds \right) dt. \tag{1.56}$$

x^1 aus (1.56) ist für rekonstruierbare Systeme eindeutig bestimmt.

Beweis: Die erste Aussage ist klar nach Korollar 1.39 angewandt auf (1.54). (1.56) folgt aus der Formel für die Lösung von (1.49), 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 $\hat{W}(t_0, t_1)$. □

Für zeitinvariante Systeme haben wir sofort das folgende Korollar.

Korollar 1.54 *Die folgenden Aussagen sind äquivalent:*

i) Das zeitinvariante System

$$\dot{x} = Ax + Bu, \; y = Cx \tag{1.57}$$

ist rekonstruierbar.

ii) Das zeitinvariante System (1.57) 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 dazu auf den zeitinvarianten Fall.

Definition 1.55 *Das zeitinvariante System*

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

heißt c-entdeckbar*, wenn die folgende Aussage gilt: Ist für* $t_0 < 0$*,* $z \in \mathrm{Kern}\hat{W}(t_0, 0)$ *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 1.56 *Betrachte das zeitinvariante System (1.57). 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) (1.57) ist c-entdeckbar.

iii) Das duale System

$$\dot{x} = A^T x + C^T u \tag{1.58}$$

ist c-stabilisierbar.

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

v) $Rang\begin{bmatrix} \lambda I - A \\ C \end{bmatrix} = n$ *für alle* λ *mit* $Re(\lambda) \geq 0$.

Beweis: Klar mit Dualität. □

1.4.2 Rekonstruierbarkeit und Entdeckbarkeit für diskrete Systeme

Wir erhalten auch wieder analoge Ergebnisse für diskrete Systeme mit Ausgang.

$$x_{k+1} = A_k x_k + B_k u_k, \; y_k = C_k x_k, \; x_0 = x^0. \tag{1.59}$$

Definition 1.57 *Das System (1.59) heißt* rekonstruierbar*, falls folgende Aussage gilt: Sind* $\{x_j\}, \{\tilde{x}_j\}$ *Lösungsfolgen von (1.59) mit der gleichen Steuerfolge* $\{u_j\}$ *und gilt*

$$C_k x_k = C_k \tilde{x}_k \quad \forall k \leq 0 \tag{1.60}$$

so gilt

$$x_k = \tilde{x}_k \quad \forall k \geq 0. \tag{1.61}$$

Es ist wieder klar, daß Rekonstruierbarkeit unabhängig von der Eingangsfolge $\{u_k\}$ ist.

Definition 1.58 *Das System (1.59) heißt* beobachtbar*, falls folgende Aussage gilt: Sind $\{x_j\}, \{\tilde{x}_j\}$ Lösungsfolgen von (1.59) mit der gleichen Steuerfolge $\{u_j\}$ und gilt*

$$C_k x_k = C_k \tilde{x}_k \quad \forall k \geq 0 \tag{1.62}$$

so gilt

$$x_k = \tilde{x}_k \quad \forall k \geq 0. \tag{1.63}$$

Es gilt die analoge Dualität und im zeitinvarianten Fall erhalten wir:

Korollar 1.59 *Die folgenden Aussagen sind äquivalent:*

i) Das zeitinvariante System

$$x_{k+1} = Ax_k + Bu_k, \quad y = Cx_k \tag{1.64}$$

ist rekonstruierbar.

ii) Das zeitinvariante System (1.64) 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 $Cp \neq 0$.

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

Beweis: Übungsaufgabe. □

Auch der Begriff der Entdeckbarkeit folgt analog.

Satz 1.60 *Betrachte das zeitinvariante System (1.64). Die folgenden Aussagen sind äquivalent.*

i) Ist $\{x_k\}$ Lösungsfolge von $x_{k+1} = Ax_k$ und gilt $Cx_k = 0$ für alle k, so folgt $\lim_{k\to\infty} x_k = 0$.

ii) (1.64) ist d-entdeckbar.

iii) Das duale System

$$x_{k+1} = A^T x_k + C^T u_k \tag{1.65}$$

ist d-stabilisierbar.

iv) Ist λ mit $|\lambda| \geq 1$ Eigenwert von A mit Eigenvektor $p \neq 0$, so gilt $C^T p \neq 0$.

v) Rang$\begin{bmatrix} \lambda I - A \\ C \end{bmatrix} = n$ für alle λ mit $|\lambda| \geq 1$.

Beweis: Übungsaufgabe □

Es ergeben sich wieder analoge Fragestellungen wie z.B. die (robuste) Stabilisierung durch Ausgangsrückführung oder die numerische Entscheidung ob ein System rekonstruierbar oder beobachtbar ist.

Kapitel 2

Algebraische und geometrische Theorie

In diesem Abschnitt betrachten wir zeitinvariante lineare Systeme

$$\dot{x} = Ax + Bu, \quad y = Cx \tag{2.1}$$

$$x_{k+1} = Ax_k + Bu_k, \quad y_k = Cx_k. \tag{2.2}$$

Um solche Systeme zu analysieren, und deren Systemeigenschaften zu bestimmen, wie Steuerbarkeit, Rekonstruierbarkeit, Beobachtbarkeit, Stabilisierbarkeit und Entdeckbarkeit schauen wir uns an, welche linearen Transformationen wir mit dem System durchführen können.

Diese Transformationen sind die folgenden:

$$\begin{array}{lllll}
x & \longmapsto & Px, & P \in \mathbb{R}^{n,n} & \text{nichtsingulärer Basiswechsel} \\
u & \longmapsto & Qu, & Q \in \mathbb{R}^{m,m} & \text{nichtsingulärer Basiswechsel} \\
y & \longmapsto & Ry, & R \in \mathbb{R}^{p,p} & \text{nichtsingulärer Basiswechsel} \\
u & \longrightarrow & -Fx + v, & F \in \mathbb{R}^{m,n} & \text{lineares Zustandsfeedback} \\
u & \longrightarrow & -Gy + v, & G \in \mathbb{R}^{m,p} & \text{lineares Ausgangsfeedback}
\end{array}$$

Für diskrete Systeme sind die Transformationen analog.

Dies sind nicht alle linearen Transformationen die möglich sind. Man kann z.B. auch

$$u \longrightarrow -F\dot{x} + v, \quad F \in \mathbb{R}^{m,n} \text{ lineares Ableitungsfeedback} \tag{2.3}$$

oder

$$u \longrightarrow -G\dot{y} + v, \quad G \in \mathbb{R}^{m,p} \text{ lineares Ableitungsausgangsfeedback} \tag{2.4}$$

machen aber wir werden sehen, dass diese beiden Transformationen die Systemeigenschaften zerstören und damit mit Vorsicht zu genießen sind.

Dieses sind alles lineare Transformationen, die auf der Blockmatrix

$$\begin{array}{c} n \\ p \end{array} \begin{bmatrix} A & B \\ C & 0 \end{bmatrix} \tag{2.5}$$

$$\begin{matrix} n & m \end{matrix}$$

operieren und zwar durch

$$\begin{bmatrix} P^{-1} & 0 \\ 0 & R^{-1} \end{bmatrix} \begin{bmatrix} A & B \\ C & 0 \end{bmatrix} \begin{bmatrix} P & 0 \\ -F & Q \end{bmatrix} =: \begin{bmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & 0 \end{bmatrix} \tag{2.6}$$

für unsere Äquivalenztransformationen bzw.

$$\begin{bmatrix} P^{-1} & 0 \\ 0 & R^{-1} \end{bmatrix} \begin{bmatrix} A & B \\ C & 0 \end{bmatrix} \begin{bmatrix} P & 0 \\ -GC & Q \end{bmatrix} =: \begin{bmatrix} \hat{A} & \hat{B} \\ \hat{C} & 0 \end{bmatrix}. \tag{2.7}$$

Diese Transformationen erhalten die Systemeigenschaften, es gilt der folgende Satz:

Satz 2.8 *Falls das System (2.1)*

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

die Systeme

$$\dot{\tilde{x}} = \tilde{A}\tilde{x} + \tilde{B}\tilde{u}, \quad \tilde{y} = \tilde{C}\tilde{x} \tag{2.9}$$

und

$$\dot{\hat{x}} = \hat{A}\hat{x} + \hat{B}\hat{u}, \quad \hat{y} = \hat{C}\hat{x} \tag{2.10}$$

mit $\tilde{A}, \tilde{B}, \tilde{C}$ wie in (2.6) und $\hat{A}, \hat{B}, \hat{C}$ wie in (2.7)

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

Beweis: Wir zeigen dies für die Steuerbarkeit und (2.9), alle anderen Fälle gehen analog. Verwende Satz 1.20, iv).

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

□

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.)

Bemerkung 2.11 Für Ableitungsfeedback wie in (2.3), gelten diese Aussagen nicht, denn erstens ist das Ergebnis kein Standardsystem mehr sondern ein *Deskriptorsystem* der Form

$$E\dot{x} = Ax + Bv, \tag{2.11}$$

mit $E = I + BF$ und es kann dabei sein, dass E singulär ist.

Als Beispiel betrachten wir das System $\dot{x} = u$ und das Ableitungsfeedback $u = \dot{x} + v$ und erhalten das rückgekoppelte System $0 = v$, welches offensichtlich eine vollkommen andere Struktur hat und für das die Sätze über die Systemeigenschaften so nicht gelten. □

Für die theoretische Analyse kann man Normalformen unter diesen Äquivalenztransformationen betrachten und damit die genauen Systemeigenschaften charakterisieren. Dies sind z.B. die Brunovsky Form in der A auf Jordanform und B auf reduzierte Form gebracht wird, siehe z.B. das Buch von Knobloch/Kwakernaak. Diese Transformationen wie auch noch einige andere System-Normalformen sind für die praktische Verwendung ungeeignet, da sie im allgemeinen nicht numerisch ausgerechnet werden können und auch nicht robust unter kleinen Störungen sind. Schon die Transformation auf Jordanform ist numerisch nicht sinnvoll durchführbar. Kleine Störungen können fundamentale Änderungen der Jordanblöcke bewirken.

Daher verwendet man in der Praxis heute Transformationen die numerisch rückwärts stabil umsetzbar sind. Dazu brauchen wir orthogonale (bzw. im komplexen unitäre) Transformationen und wir versuchen auch normbeschränkte Feedbacks zu verwenden.

Um dies zu tun, verwenden wir die folgenden bekannten Zerlegungen. Wir betrachten hier nur die reelen Versionen, es gibt jeweils analoge komplexe Versionen.

Satz 2.12 *(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 $\mathbb{R}^n$ bzw. $\mathbb{R}^m$.

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

Es folgt, daß

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

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. □

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

Korollar 2.13 *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: Siehe Vorlesung PMII oder Buch von Golub/Van Loan. □

Die SVD ist die *beste* numerische Methode zur Bestimmung von Rängen.

Warum nimmt man nicht die QR–Zerlegung?

Beispiel 2.14 *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, *insbesondere in der Praxis*, sind sogenannte QRP Zerlegungen der Form

$$A = QRP = \begin{array}{c} \\ \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix} \\ r \quad m-r \end{array} \begin{array}{c} r \\ n-r \end{array} \tag{2.15}$$

wobei $r = \text{Rang } (A)$, Q orthogonal, R_{11} obere Dreiecksmatrix und P Permutationsmatrix. Dabei wird P so gewählt, dass immer die Spalte mit der größten Norm zuerst eliminiert wird. Es folgt dann, dass die Diagonalelemente von R_{11} in fallender Folge geordnet sind. Beachte jedoch, dass Beispiel 2.14 auch die QRP Zerlegung betrügt.

Eine weitere wichtige Zerlegung ist die (reelle) Schur-Form.

Satz 2.16 *(Reelle Version) Sei $A \in \mathbb{R}^{n,n}$ so gibt es eine orthogonale Matrix Q, so dass*

$$Q^T A Q = R = \begin{bmatrix} R_{11} & \cdots & R_{1k} \\ & \ddots & \vdots \\ & & R_{kk} \end{bmatrix} \tag{2.17}$$

quasi obere Dreiecksmatrix ist. Die Diagonalblöcke R_{ii} sind 1×1 oder 2×2, mit reellen oder komplexen Eigenwerten und können in jeder beliebigen Reihenfolge auftreten.

Beweis: 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

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

Der Beweis folgt dann per Induktion. □

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

Satz 2.18 *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.

Beweis: Übungsaufgabe. □

Da S nicht orthogonal ist, so ist im allgemeinen die Berechnung der GSVD nicht numerisch stabil. Man kann da jedoch durch einige technische Tricks trotzdem einen brauchbaren Algorithmus erzeugen, siehe LAPACK, SLICOT.

Wir wollen nun diese Zerlegungen verwenden, um 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 2.19 *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*

$$PAP^T = \left[\begin{array}{cccc|c} A_{11} & \cdots & \cdots & A_{1s-1} & A_{1,s} \\ A_{21} & \ddots & & \vdots & \vdots \\ & \ddots & \ddots & \vdots & \vdots \\ & & A_{s-1,s-2} & A_{s-1,s-1} & A_{s-1,s} \\ \hline 0 & \cdots & 0 & 0 & A_{ss} \end{array}\right] \begin{array}{c} n_1 \\ n_2 \\ \vdots \\ n_{s-1} \\ n_s \end{array}, \quad PBQ = \left[\begin{array}{cc} B_1 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ \vdots & \vdots \\ 0 & 0 \end{array}\right] \begin{array}{c} n_1 \\ n_2 \\ \vdots \\ \vdots \\ n_s \end{array}$$
$$\begin{array}{cccccc} n_1 & \ldots & n_{s-2} & n_{s-1} & n_s \end{array} \qquad \begin{array}{cc} n_1 & m-n_1 \end{array} \tag{2.20}$$

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

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

$\Sigma_{i,i-1}$ *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 2.21 „Staircase“ Algorithmus

Input: $A \in \mathbb{R}^{n,n}, B \in \mathbb{R}^{n,m}$,
Output: PAP^T, PBQ in der Form (2.20), P, Q orthogonal.
Schritt 0: Führe eine SVD von B durch.

$$B = U_B \left[\begin{array}{cc} \Sigma_B & 0 \\ 0 & 0 \end{array}\right] V_B^T$$

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

$$A := U_B^T A U_B = \left[\begin{array}{cc} A_{11} & A_{12} \\ A_{21} & A_{22} \end{array}\right], \; B := U_B^T B V_B = \left[\begin{array}{cc} \Sigma_B & 0 \\ 0 & 0 \end{array}\right]$$

mit A_{11} der Größe $n_1 \times n_1$.

Schritt 1: Führe ein SVD von A_{21} durch:

$$A_{21} = U_{21} \left[\begin{array}{cc} \Sigma_{21} & 0 \\ 0 & 0 \end{array}\right] V_{21}^T, \quad \text{mit } \Sigma_{21} \; n_2 \times n_2 \text{ nichtsingulär und diagonal.}$$

Setze

$$P_2 := \left[\begin{array}{cc} V_{21}^T & 0 \\ 0 & U_{21}^T \end{array}\right], \; P := P_2 P$$

sowie

$$A := P_2 A P_2^T =: \left[\begin{array}{ccc} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ 0 & A_{32} & A_{33} \end{array}\right], \; B := P_2 B =: \left[\begin{array}{cc} B_1 & 0 \\ 0 & 0 \\ 0 & 0 \end{array}\right],$$

wobei $A_{21} = [\Sigma_{21} \quad 0]$ mit Σ_{21} nichtsingulär, diagonal von der Größe $n_2 \times n_2$ und $B_1 := V_{21}^T \Sigma_B$ nichtsingulär.

Schritt 2:
i = 3
DO WHILE ($n_{i-1} > 0$ OR $A_{i,i-1} \neq 0$).
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 und diagonal.
Setze

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

sowie

$$A := P_i A P_i^T =: \begin{bmatrix} A_{11} & & \cdots & & A_{1,i+1} \\ A_{21} & \ddots & & & A_{2,i+1} \\ & \ddots & \ddots & & \vdots \\ & & A_{i,i-1} & A_{i,i} & \vdots \\ 0 & & & A_{i+1,i} & A_{i+1,i+1} \end{bmatrix}$$

wobei $A_{i,i-1} = [\Sigma_{i,i-1} \quad 0]$.
$i := i + 1$
END
$s := i$

Es ist klar, dass dieser Algorithmus entweder mit $n_i = 0$ oder $A_{i,i-1} = 0$ abbricht, denn solange eine der beiden Bedingungen nicht gilt, kann man den nächsten Schritt ausführen. Andererseits wird in jedem Schritt der Restblock mindestens um 1 kleiner, solange $\text{Rang} A_{i,i-1} > 1$ ist, so dass der Algorithmus nach maximal $n - 1$ Schritten abbricht. □

Bemerkung 2.22 Algorithmus (2.21) in etwas veränderter Form geht auf 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 im schlimmsten Fall $\mathcal{O}(n^4)$ flops brauchen. Kritischer Punkt ist natürlich in jedem Schritt die Rangentscheidung in $A_{i,i-1}$, bzw. B. Die Form (2.20) ist sehr eng verwandt mit der Kronecker Normalform für das Büschel $\lambda[I \;\; 0] - [A \;\; B]$ und liefert die Information über die Größen der entsprechenden Blöcke in der Normalform.

Als direkte Konsequenz aus Lemma 2.19 erhalten die folgenden (numerisch stabil nachprüfbaren) Kriterien für Steuerbarkeit, Rekonstruierbarkeit, Stabilisierbarkeit und Entdeckbarkeit.

Satz 2.23 *Gegeben sei ein lineares zeitinvariantes System der Form (2.1) oder (2.2).*

i) Das System (2.1) oder (2.2) ist (vollständig) steuerbar genau dann, wenn in der Treppenform von (A, B) gilt, dass $n_s = 0$.

ii) Das System (2.1) oder (2.2) ist rekonstruierbar genau dann, wenn in der Treppenform von (A^T, C^T) gilt, dass $n_s = 0$.

iii) Das System (2.1) oder (2.2) ist stabilisierbar genau dann, wenn in der Treppenform von (A, B) alle Eigenwerte von A_{ss} negativen Realteil haben.

iv) Das System (2.1) oder (2.2) 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 2.8 und Lemma 2.19 annehmen, dass (A, B) in Treppenform vorliegt. Betrachte die Matrix

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

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

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

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 2.24 *In vielen Lehrbüchern wird eine Aufspaltung des Raumes in steuerbare und nicht steuerbare Unterräume gemacht. Diese Aufspaltung erhält man natürlich sofort aus der Treppenform, siehe z.B. 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 2.19, daß $n_1 = n_2 = \cdots = n_{s-1} = 1$ und damit ist die Staircaseform einfach die sogenannte *System Hessenberg–Form*.

$$PAP^T = \begin{bmatrix} a_{11} & & \cdots & a_{1,n-1} & a_{1,n} \\ a_{21} & \ddots & & & \vdots \\ & \ddots & \ddots & & \vdots \\ & & a_{n-1,n-2} & a_{n-1,n-1} & a_{n-1,n} \\ 0 & & & a_{n,n-1} & a_{n,n} \end{bmatrix} \quad PB = \begin{bmatrix} b_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{2.25}$$

(Dabei ist noch A_{ss} auf Hessenbergform gebracht worden.) Wir erhalten als Korollar aus Satz 2.23, daß das System *steuerbar* 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 3

Polvorgabe

Eine der beliebtesten Techniken zur Veränderung von Systemeigenschaften ist die Polvorgabe, d.h. die Veränderung der Eigenwerte der Systemmatrix A bzw. der Pole der Transferfunktion $C(sI - A)^{-1}B + D$ durch Rückkopplung.

Wir betrachten zuerst die folgende Fragestellung. Seien Matrizen $A \in \mathbb{R}^{n,n}$ und $B \in \mathbb{R}^{n,m}$ gegeben. Für eine vorgegebene Menge komplexer Zahlen

$$\mathcal{P} = \{\lambda_1, \ldots, \lambda_n\} \subset \mathbb{C}, \tag{3.1}$$

wobei wir annehmen, dass die Menge abgeschlossen unter komplexer Konjugation ist, bestimme eine Matrix $F \in \mathbb{R}^{m,n}$, so dass das Spektrum von $\lambda(A - BF)$ gleich $\mathcal{P}$ ist.

Damit kann man natürlich das Verhalten des Systems beliebig beeinflussen und auch z.B. das System asymptotisch stabil machen.

Wir wollen untersuchen, unter welchen Bedingungen dies möglich ist und wie die Lösung dieser Aufgabe aussieht.

Satz 3.2 *Seien Matrizen $A \in \mathbb{R}^{n,n}$ und $B \in \mathbb{R}^{n,m}$ gegeben. Es gibt für jede Menge $\mathcal{P}$ wie in (3.1) eine Matrix $F \in \mathbb{R}^{m,n}$, so dass das Spektrum von $\lambda(A - BF)$ gleich $\mathcal{P}$ ist, genau dann wenn das zugehörige lineare System (kontinuierlich oder diskret) vollständig steuerbar ist.*

Beweis: Zeige zuerst, dass die vollständige Steuerbarkeit notwendig für die beliebige Polvorgabe ist. Angenommen, dass System ist nicht steuerbar aber jede Polmenge $\mathcal{P}$ kann zugewiesen werden. Dann gibt es $\hat{\lambda} \in \mathbb{C}$ und $p \neq 0$, so dass $p^T B = 0$ und $p^T(\hat{\lambda}I - A) = 0$. Sei nun $\mathcal{P}$, so dass $\hat{\lambda} \notin \mathcal{P}$, so ergibt sich sofort ein Widerspruch, denn $p^T(\hat{\lambda}I - A + BF) = 0$.

Die Umkehrung beweisen wir im folgenden konstruktiv durch Angabe der Rückkopplungsmatrix. □

Dazu betrachten wir zuerst die Staircase form (2.20) und reduzieren diese noch weiter.

Lemma 3.3 *Seien $A \in R^{n,n}$, $B \in R^{n,m}$, (A, B) steuerbar und* $\mathrm{Rang}(B) = m$. *Dann gibt es nicht-singuläre Matrizen $S \in R^{n,n} T \in R^{m,m}$, so dass*

$$\hat{A} \;:=\; S^{-1}AS$$

$$
\begin{array}{rcl}
 & & \begin{array}{cccccccc} d_1 & n_2 & d_2 & n_3 & \dots & d_{s-1} & n_s & d_s \end{array} \\
 & \begin{array}{c} n_1 \\ n_2 \\ n_3 \\ \vdots \\ n_{s-1} \\ n_s \end{array} & \left[\begin{array}{cccccccc}
\hat{A}_{1,1} & 0 & \hat{A}_{1,2} & 0 & \dots & \hat{A}_{1,s-1} & 0 & \hat{A}_{1,s} \\
0 & I_{n_2} & \hat{A}_{2,2} & 0 & \dots & \hat{A}_{2,s-1} & 0 & \hat{A}_{2,s} \\
 & & 0 & I_{n_3} & \dots & \hat{A}_{3,s-1} & 0 & \hat{A}_{3,s} \\
 & & & & \ddots & \vdots & \vdots & \vdots \\
 & & & & & \hat{A}_{s-1,s-1} & 0 & \hat{A}_{s-1,s} \\
 & & & & & 0 & I_{n_s} & \hat{A}_{s,s}
\end{array}\right], \\
\hat{B} & := & S^{-1}BT = \left[\begin{array}{c} I_{n_1} \\ 0 \end{array}\right],
\end{array} \tag{3.2}
$$

wobei für die Indizes n_i und d_i gilt, dass

$$d_i := n_i - n_{i+1},\ i = 1, \dots, s-1,\ d_s := n_s, \tag{3.4}$$

Beweis: Übung. □

Setze noch

$$\pi_i := d_1 + \ldots + d_i = m - n_{i+1},\ i = 1, \dots, s-1, \quad \pi_s = m. \tag{3.5}$$

Damit führen wir noch Krylov-Matrizen der Form

$$K_k := [B, AB, \dots, A^{k-1}B],\ \hat{K}_k := [\hat{B}, \hat{A}\hat{B}, \dots, \hat{A}^{k-1}\hat{B}], \tag{3.6}$$

und auch noch Block-Matrizen

$$\hat{X}_k := \left[\begin{array}{ccc} \hat{X}_{1,1} & \dots & \hat{X}_{1,k} \\ & \ddots & \vdots \\ & & \hat{X}_{k,k} \end{array}\right] \in \mathbb{R}^{km,\pi_k}, \tag{3.7}$$

$$X_k := \left[\begin{array}{ccc} X_{1,1} & \dots & X_{1,k} \\ & \ddots & \vdots \\ & & X_{k,k} \end{array}\right] := \mathrm{Diag}(T, \dots, T)\hat{X}_k \in \mathbb{R}^{km,\pi_k}, \tag{3.8}$$

$$\hat{R}_k := [\hat{A}_{1,1}, \hat{A}_{1,2}, \dots, \hat{A}_{1,k}], \quad R_k := T\hat{R}_k \in \mathbb{R}^{m,\pi_k} \tag{3.9}$$

ein, wobei

$$
\begin{array}{rcl}
 & & \begin{array}{c} d_i \end{array} \\
\hat{X}_{i,i} & := & \begin{array}{c} \pi_{i-1} \\ d_i \\ n_{i+1} \end{array} \left[\begin{array}{c} 0 \\ I_{d_i} \\ 0 \end{array}\right],\ i = 1, \dots, k, \\
 & & \begin{array}{c} d_j \end{array} \\
\hat{X}_{i,j} & := & \begin{array}{c} \pi_i \\ n_{i+1} \end{array} \left[\begin{array}{c} 0 \\ -\hat{A}_{i+1,j} \end{array}\right],\ i = 1, \dots, k-1,\ j = i+1, \dots, k, \\
X_{i,j} & := & T\hat{X}_{i,j},\ i = 1, \dots, k,\ j = i, \dots, k.
\end{array}
$$

Mit den Abkürzungen $X := X_s$, $R := R_s$, $K := K_s$ können wie dann den Nullraum von $[A, B]$ charakterisieren.

Lemma 3.7 *Seien X_k, $\hat{X}_k$, R_k, $\hat{R}_k$, K_k und $\hat{K}_k$ wie in (3.6)–(3.9). Dann gilt*

$$AK_kX_k = BR_k,\ \hat{A}\hat{K}_k\hat{X}_k = \hat{B}\hat{R}_k,\ k = 1, \ldots, s \tag{3.8}$$

und die Spalten von

$$\begin{bmatrix} U_0 \\ -V_0 \end{bmatrix} = \begin{bmatrix} KX \\ -R \end{bmatrix}$$

spannen den Kern von $[A, B]$ auf.

Beweis: Der Beweis ergibt sich sofort aus den Formeln $AK_kX_k = S(\hat{A}\hat{K}_k\hat{X}_k)$ und $BR_k = S(\hat{B}\hat{R}_k)$ sowie der speziellen Struktur der Blöcke in $\hat{K}_k$, denn für $1 \le l \le s$ gilt

$$\hat{A}^{l-1}\hat{B} = \begin{array}{c} \\ n_1 \\ \vdots \\ n_{l-1} \\ n_l \\ n_{l+1} \\ \vdots \\ n_s \end{array} \begin{array}{c} \begin{array}{ccccc} d_1 & \ldots & d_{l-2} & d_{l-1} & n_l \end{array} \\ \begin{bmatrix} * & \ldots & * & \hat{A}_{1,l-1} & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ * & \ldots & * & \hat{A}_{l-1,l-1} & 0 \\ 0 & \ldots & 0 & 0 & I_{n_l} \\ 0 & \ldots & 0 & 0 & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & 0 & 0 \end{bmatrix} \end{array}, \tag{3.9}$$

durch Ausmultiplikation beider Seiten der Gleichung in (3.8). Beachte, dass die vollständige Steuerbarkeit und die Staircase-Form implizieren, dass der volle Kern für $k = s$ erreicht wird, da dann die Dimension des Raumes der durch Spalten von

$$\begin{bmatrix} KX \\ -R \end{bmatrix}$$

aufgespannt wird, $m = n_1$ ist, welches die Dimension des Kerns von $[A, B]$ ist. □

Wir brauchen noch weitere Notation. Sei

$$\Theta_{i,j} := \sum_{l=i}^{j} A^{l-i}BX_{l,j},\ \hat{\Theta}_{i,j} := \sum_{l=i}^{j} \hat{A}^{l-i}\hat{B}\hat{X}_{l,j},\ i = 1, \ldots, s,\ j = i, \ldots, s, \tag{3.10}$$

und sei

$$\begin{aligned} W_i &:= [\Theta_{i,i}, \ldots, \Theta_{i,s}] \in \mathbb{C}^{n,n_i},\ i = 1, \ldots, s, \\ W &:= [W_1, W_2, \ldots, W_s] \in \mathbb{C}^{n,n}, \\ Y_i &:= [X_{i,i}, \ldots, X_{i,s}] \in \mathbb{C}^{m,n_i},\ i = 1, \ldots, s, \\ Y &:= [Y_1, Y_2, \ldots, Y_s] \in \mathbb{C}^{m,n}. \end{aligned} \tag{3.11}$$

Definiere weiterhin

$$\mathcal{I}_{i,j} := \begin{array}{c} \\ n_i \end{array} \begin{array}{c} \begin{array}{cc} n_j - n_i & n_i \end{array} \\ \begin{bmatrix} 0 & I_{n_i} \end{bmatrix} \end{array},\ i \ge j, \tag{3.11}$$

und

$$N := \begin{bmatrix} 0 & & & & \\ \mathcal{I}_{2,1} & 0 & & & \\ 0 & \mathcal{I}_{3,2} & \ddots & & \\ \vdots & & \ddots & 0 & \\ 0 & \dots & 0 & \mathcal{I}_{s,s-1} & 0 \end{bmatrix}, \; \tilde{N} = \begin{bmatrix} 0 & I_m & & & \\ & \ddots & \ddots & & \\ & & \ddots & I_m & \\ & & & 0 & \end{bmatrix}.$$

Dann gilt das folgende Lemma.

Lemma 3.12 *Die Matrizen W, W_1 aus (3.11) haben die folgenden Eigenschaften.*

i)

$$W_1 = KX, \quad W = K\tilde{X}, \quad \tilde{X} = [X, \tilde{N}X, \dots, \tilde{N}^{s-1}X]. \tag{3.13}$$

ii)

$$W = AWN + BY. \tag{3.14}$$

iii) W ist nicht-singulär.

Beweis:

i) folgt direkt aus der Definition von W_1 und W.

ii) Aus der Frm von W und N folgt, dass

$$\begin{aligned} AWN &= A[W_1, W_2, \dots, W_s]N \\ &= A[0, W_2; \dots; 0, W_{s-1}; 0] \\ &= [0, A\Theta_{2,2}, \dots, A\Theta_{2,s}; \dots; 0, A\Theta_{s,s}; 0] \\ &= [0, \Theta_{1,2}, \dots, \Theta_{1,s}; \dots; 0, \Theta_{s-1,s}; 0] \\ &\quad -B[0, X_{1,2}, \dots, X_{1,s}; \dots; 0, X_{s-1,s}; 0] \\ &= W - BY. \end{aligned}$$

iii) Wir haben

$$\begin{aligned} W &= S(S^{-1}W) \\ &= S[\hat{\Theta}_{1,1}, \dots, \hat{\Theta}_{1,s}; \dots; \hat{\Theta}_{s-1,s-1}, \hat{\Theta}_{s-1,s}, \hat{\Theta}_{s,s}] \\ &= S \begin{bmatrix} I_{n_1} & * & \dots & * \\ & I_{n_2} & \dots & \vdots \\ & & \ddots & * \\ & & & I_{n_s} \end{bmatrix} P \end{aligned}$$

für eine passende Permutationsmatrix P. Dies folgt direkt aus der Definition von $\hat{\Theta}_{i,j}$ in (3.10). Also ist W nicht-singulär. □

Bemerkung 3.15 Falls $m = 1$ so ist $X = [a_1, \ldots, a_{n-1}, 1]^T$, $R = -a_0$. Wenn $K = [B, \ldots, A^{n-1}B]$ nicht-singulär ist so folgt aus $AKX = BR$, dass $a_0, \ldots, a_{n-1}$ die Koeffizienten des charakteristischen Polynomials von A, d.h.,

$$\xi(\lambda) := \lambda^n + \sum_{k=0}^{n-1} a_k \lambda^k = \det(\lambda I_n - A)$$

sind. Mit $\mathrm{adj}(\lambda I_n - A) := \sum_{k=0}^{n-1} A_k \lambda^k$, folgt $W_1 = A_0 B$ and $W = [A_0 B, \ldots A_{n-1} B]$.

Nun können wir den Nullraum von $[A - \lambda I, B]$ bestimmen für beliebiges λ.

Satz 3.16 *Sei* $E_{\lambda,k} := (I - \lambda N)^{-1} \begin{bmatrix} I_{\pi_k} \\ 0 \end{bmatrix}$. *Dann spannen die Spalten von*

$$\begin{bmatrix} U_{\lambda,k} \\ -V_{\lambda,k} \end{bmatrix} := \begin{bmatrix} WE_{\lambda,k} \\ -(R_k - \lambda Y E_{\lambda,k}) \end{bmatrix}, \ k = 1, 2, \ldots, s, \tag{3.17}$$

die Unterräume $\mathbb{N}_{\lambda,k}$ *der Dimension* π_k *von* $\mathrm{Kern}[A - \lambda I, B]$ *auf. Insbesondere erhalten wir für* $k = s$ *den gesamten Kern* $\mathbb{N}_\lambda$, *aufgespannt von den Spalten von*

$$\begin{bmatrix} U_\lambda \\ -V_\lambda \end{bmatrix} := \begin{bmatrix} WE_{\lambda,s} \\ -(R - \lambda Y E_{\lambda,s}) \end{bmatrix}. \tag{3.18}$$

Dieser Raum hat Dimension $\pi_s = m$ *und es gilt* $(A - \lambda I)U_\lambda = BV_\lambda$.

Beweis: Aus (3.14) folgt

$$(A - \lambda I)W = AW - \lambda W = AW - \lambda AWN - \lambda BY = AW(I - \lambda N) - \lambda BY.$$

Da $I - \lambda N$ nicht-singulär ist, erhalten wir

$$(A - \lambda I)W(I - \lambda N)^{-1} = AW - \lambda BY(I - \lambda N)^{-1}$$

und nach Multiplikation mit $\begin{bmatrix} I_{\pi_k} \\ 0 \end{bmatrix}$ von rechts erhalten wir

$$(A - \lambda I)WE_{\lambda,k} = AW \begin{bmatrix} I_{\pi_k} \\ 0 \end{bmatrix} - \lambda BY E_{\lambda,k}.$$

Mit Lemma 3.7 und (3.13) haben wir das $AW \begin{bmatrix} I_{\pi_k} \\ 0 \end{bmatrix} = BR_k$ und damit folgt die erste Behauptung. Die Dimension von $\mathbb{N}_{\lambda,k}$ erhalten wir aus

$$\mathrm{Rang} U_{\lambda,k} = \mathrm{Rang} WE_{\lambda,k} = \mathrm{Rang} E_{\lambda,k} = \pi_k.$$

□

Mittels der expliziten Darstellungen für den Kern von $[A - \lambda I, B]$ erhalten wir nun explizite Lösungsformeln für die Matrix F, welche das Polvorgabeproblem löst. Setze

$$\mathcal{U}_{\lambda,k} := \mathrm{Bild} U_{\lambda,k}, \quad \mathcal{V}_{\lambda,k} := \mathrm{Bild} V_{\lambda,k}, \quad k = 1, \ldots, s, \tag{3.19}$$

wobei $U_{\lambda,k}, V_{\lambda,k}$ wie in (3.17) definiert sind. Setze speziell $\mathcal{U}_\lambda := \text{Bild}U_\lambda$ ($U_\lambda = U_{\lambda,s}$), $\mathcal{V}_\lambda := \text{Bild}V_\lambda$ ($V_\lambda = V_{\lambda,s}$).

Sei (λ, g) ein Eigenpaar von $A - BF$, d.h.,

$$(A - BF)g = \lambda g \text{ oder } (A - \lambda I)g = BFg =: Bz.$$

Mittels der Darstellung des Kerns von $[A - \lambda I, B]$ in (3.18) gibt es einen Vektor $\phi \in \mathbb{C}^m$, so dass $g = U_\lambda \phi$, $z = V_\lambda \phi$. Und damit ist $\mathcal{U}_\lambda$ der Raum der möglichen Eigenvektoren von $A - BF$ zum Eigenwert λ.

Betrachte zuerst einen einzelnen Jordanblock $J_p = \lambda I + N_p$, wobei

$$N_p := \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ & 0 & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & 0 \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix}_{p,p}.$$

Lemma 3.20 *Angenommen $A - BF$ hat einen Jordanblock der Größe $p \times p$ zum Eigenwert λ und eine Hauptvektorkette $g_1, \dots, g_p$, d.h.,*

$$(A - BF)[g_1, \dots, g_p] = [g_1, \dots, g_p]J_p. \tag{3.21}$$

Seien $G_p := [g_1, \dots, g_p]$ und $Z_p =: FG_p =: [z_1, \dots, z_p]$, dann existieren Matrizen $\Phi_p = [\phi_1, \dots, \phi_p] \in \mathbb{C}^{m,p}$ und $\Gamma_p \in \mathbb{C}^{n,p}$, so dass

$$G_p = W\Gamma_p, \; Z_p = R\Phi_p - Y\Gamma_p J_p, \tag{3.22}$$

wobei für

$$\Gamma_p = \begin{bmatrix} \Phi_p \\ \mathcal{I}_{2,1}\Phi_p J_p \\ \vdots \\ \mathcal{I}_{s,1}\Phi_s J_p^{s-1} \end{bmatrix} \tag{3.23}$$

gilt, dass $\text{Rang}\Gamma_p = p$. *(Die Matrizen $\mathcal{I}_{i,1}$ sind dabei wie in (3.11).)*

Beweis: Addition von $-\lambda WN$ auf beiden Seiten von (3.14) ergibt

$$W(I - \lambda N) = (A - \lambda I)WN + BY.$$

Damit folgt

$$W = (A - \lambda I)WN(I - \lambda N)^{-1} + BY(I - \lambda N)^{-1}. \tag{3.24}$$

Sei $E = \begin{bmatrix} I_{n_1} \\ 0 \end{bmatrix}$ dann zeigen wir per Induktion, dass es Vektoren $\phi_j \in \mathbb{C}^m$ gibt, so dass für g_k, z_k gilt, dass

$$g_k = W \sum_{j=1}^{k} N^{j-1}(I - \lambda N)^{-j} E\phi_{k+1-j}, \tag{3.25}$$

$$z_k = V_\lambda \phi_k - Y \sum_{j=2}^{k} N^{j-2}(I - \lambda N)^{-j} E\phi_{k+1-j}, \tag{3.26}$$

für $k = 1, 2, \ldots, p$.

Für $k = 1$ folgt aus (3.21), dass g_1 ein Eigenvektor von $A - BF$ ist. Also gibt es $\phi_1 \in \mathbb{C}^m$, so dass

$$g_1 = WE_{\lambda,s}\phi_1 = W(I - \lambda N)^{-1}E\phi_1,\ z_1 = V_\lambda \phi_1. \tag{3.25}$$

Angenommen, dass (3.25) und (3.26) für k gelten, dann zeigen wir, dass dies auch für $k+1$ gilt. Aus (3.21) folgt, dass $(A - \lambda I)g_{k+1} = Bz_{k+1} + g_k$ und mit (3.25), (3.24), dass

$$\begin{aligned} g_k &= (A - \lambda I)W \sum_{j=1}^{k} N^j (I - \lambda N)^{-(j+1)} E\phi_{k+1-j} \\ &+ BY \sum_{j=1}^{k} N^{j-1}(I - \lambda N)^{-(j+1)} E\phi_{k+1-j}. \end{aligned}$$

Dann existiert $\phi_{k+1} \in \mathbb{C}^m$, (beachte, dass $N^k = 0$ für $k \geq s$,) so dass

$$\begin{aligned} g_{k+1} &= W\{(I - \lambda N)^{-1}E\phi_{k+1} + \sum_{j=1}^{k} N^j (I - \lambda N)^{-(j+1)} E\phi_{k+1-j}\} \\ &= W \sum_{j=1}^{k+1} N^{j-1}(I - \lambda N)^{-j} E\phi_{k+2-j} \end{aligned}$$

und

$$\begin{aligned} z_{k+1} &= V_\lambda \phi_{k+1} - Y \sum_{j=1}^{k} N^{j-1}(I - \lambda N)^{-(j+1)} E\phi_{k+1-j} \\ &= V_\lambda \phi_{k+1} - Y \sum_{j=2}^{k+1} N^{j-2}(I - \lambda N)^{-j} E\phi_{k+2-j}. \end{aligned}$$

Mit (3.25) und (3.26) erhalten wir

$$\begin{aligned} G_p &= W \sum_{j=1}^{p} N^{j-1}(I - \lambda N)^{-j} E\Phi_p N_p^{j-1} =: W\Gamma_p, \\ Z_p &= V_\lambda \Phi_p - Y \sum_{j=2}^{p} N^{j-2}(I - \lambda N)^{-j} E\Phi_p N_p^{j-1}. \end{aligned}$$

Aus

$$N^{j-1}(I - \lambda N)^{-j} = \sum_{k=j}^{s} \binom{k-1}{j-1} \lambda^{k-j} N^{k-1},$$

ergibt sich

$$\Gamma_p = \sum_{j=1}^{s} (\sum_{k=j}^{s} \binom{k-1}{j-1} \lambda^{k-j} N^{k-1}) E\Phi_p N_p^{j-1}$$

$$
\begin{aligned}
&= \sum_{j=1}^{s}(\sum_{k=j}^{s}\binom{k-1}{j-1}\lambda^{k-j}\begin{bmatrix}0\\ \mathcal{I}_{k,1}\Phi_p\\ 0\end{bmatrix})N_p^{j-1}\\
&= \sum_{k=1}^{s}\begin{bmatrix}0\\ \mathcal{I}_{k,1}\Phi_p\\ 0\end{bmatrix}(\sum_{j=1}^{k}\binom{k-1}{j-1}\lambda^{k-j}N_p^{j-1})\\
&= \sum_{k=1}^{s}\begin{bmatrix}0\\ \mathcal{I}_{k,1}\Phi_p(\lambda I_p+N_p)^{k-1}\\ 0\end{bmatrix} = \begin{bmatrix}\Phi_p\\ \mathcal{I}_{2,1}\Phi_p J_p\\ \vdots\\ \mathcal{I}_{s,1}\Phi_p J_p^{s-1}\end{bmatrix}.
\end{aligned}
$$

Da

$$\sum_{j=2}^{p} N^{j-2}(I-\lambda N)^{-j}E\Phi_p N_p^{j-1} = (I-\lambda N)^{-1}\Gamma_p N_p,$$

erhalten wir $Z_p = V_\lambda\Phi_p - Y(I-\lambda N)^{-1}\Gamma_p N_p$, und dann ergibt sich mit $V_\lambda = R - \lambda Y(I-\lambda N)^{-1}E$, dass

$$Z_p = R\Phi_p - Y(I-\lambda N)^{-1}\begin{bmatrix}\Phi_p J_p\\ \mathcal{I}_{2,1}\Phi_p J_p N_p\\ \vdots\\ \mathcal{I}_{s,1}\Phi_p J_p^{s-1}N_p\end{bmatrix}.$$

Es ist einfach nachzurechen, dass $Z_p = R\Phi_p - Y\Gamma_p J_p$ indem man die explizite Inverse von $(I-\lambda N)^{-1}$ verwendet und die Blöcke von oben nach unten berechnet. Das Rang$\Gamma_p = p$ folgt aus Rang$W = n$ und Rang$G_p = p$. □

Kombination der Ergebnisse für einzelne Jordanblöcke ergibt den Satz.

Satz 3.26 *Sei*

$$J = \mathrm{Diag}(J_{1,1},\ldots,J_{1,r_1},\ldots,J_{q,1},\ldots,J_{q,r_q}), \tag{3.27}$$

mit $J_{ij} = \lambda_i I_{p_{ij}} + N_{p_{ij}}$. Dann gibt es ein F so dass J die Jordanform von $A-BF$ ist, genau dann wenn es eine Matrix $\Phi \in \mathbb{C}^{m,n}$ gibt, so dass

$$\Gamma := \begin{bmatrix}\Phi\\ \mathcal{I}_{2,1}\Phi J\\ \vdots\\ \mathcal{I}_{s,1}\Phi J^{s-1}\end{bmatrix} \tag{3.28}$$

nicht-singulär ist. Falls so ein Γ existiert, so ist mit $G := W\Gamma$ und $Z := R\Phi - Y\Gamma J$, die Matrix $F = ZG^{-1}$ eine Rückkopplungsmatrix die die gewünschte Eigenstruktur zuweist und es gilt $A-BF = GJG^{-1}$.

Beweis: Die Notwendigkeit folgt direkt aus Lemma 3.20. Für die Rückrichtung folgt mit (3.13), (3.8) und (3.14), dass

$$AW\Gamma \;=\; AW_1\Phi + A[W_2,\ldots,W_s]\begin{bmatrix}\mathcal{I}_{2,1}\Phi J\\ \vdots\\ \mathcal{I}_{s,1}\Phi J^{s-1}\end{bmatrix}$$

$$\begin{aligned} &= AW_1\Phi + A[0, W_2; \ldots; 0, W_s; 0]\Gamma J \\ &= BR\Phi + AWN\Gamma J \\ &= BR\Phi + W\Gamma J - BY\Gamma J = BZ + W\Gamma J. \end{aligned}$$

Da Γ und W invertierbar sind, folgt

$$A - BZ(W\Gamma)^{-1} = W\Gamma J(W\Gamma)^{-1}$$

und daher ist $F = Z(W\Gamma)^{-1}$ die gewünschte Rückkopplungsmatrix. □

Nun können wir endlich die Rückrichtung in Satz 3.26 beweisen.

Korollar 3.29 *Betrachte das Polvorgabeproblem aus Satz 3.26, mit J wie in (3.27). Eine notwendige Bedingung für die Existenz einer Matrix F, so dass J die Jordanform von $A-BF$ ist, ist dass Φ die Eigenschaft hat, dass das Paar (J^H, Φ^H) steuerbar ist. Eine hinreichende Bedingung ist die Existenz eines $\Psi \in \mathbb{C}^{m,n}$, so dass (J^H, Ψ^H) steuerbar ist, und die gleichen Indizes n_k wie (A, B) hat.*

Beweis: Die Notwendigkeit ist klar. Für die hinreichende Bedingung verwenden wir einfach die Form (3.2). □

Aus Satz 3.26 erhalten wir alle Rückkopplungen, die eine gewünschte Jordan Form erzeugen.

Korollar 3.30 *Die Menge aller Rückkopplungen F die eine gegebene Jordanform (3.27) erzeugen ist gegeben durch*

$$\{F = ZG^{-1} = (R\Phi - Y\Gamma J)(W\Gamma)^{-1} | \det\Gamma \neq 0,\ \Gamma \text{ wie in} (3.28)\}. \tag{3.31}$$

Wir müssen dabei in Satz 3.26 J nicht in Jordan-Form wählen, dies geht für jede beliebige Matrix J. Denn mit beliebigem nicht-singulären Q, gilt

$$\Gamma Q = \begin{bmatrix} \Phi Q \\ \mathcal{I}_{2,1}\Phi Q(Q^{-1}JQ) \\ \vdots \\ \mathcal{I}_{s,1}\Phi Q(Q^{-1}JQ)^{s-1} \end{bmatrix} = \begin{bmatrix} \hat{\Phi} \\ \mathcal{I}_{2,1}\hat{\Phi}\hat{J} \\ \vdots \\ \mathcal{I}_{s,1}\hat{\Phi}\hat{J}^{s-1} \end{bmatrix},$$

wobei $\hat{\Phi} = \Phi Q$, $\hat{J} = Q^{-1}JQ$. Inbesondere können wir damit natürlich auch J in reeller Jordan-Form wählen und ein reeles Φ bekommen.

Bemerkung 3.32 Im Fall $m = 1$, darf die Jordan-Form nicht degeneriert sein, d.h. wir brauchen $r_1 = \ldots = r_q = 1$. Let $\Phi = [\phi_1, \ldots, \phi_q]$ and $\phi_k = [\phi_{k,1}, \ldots, \phi_{k,p_k}] \in \mathbb{C}^{1,p_k}$, let $\xi(\lambda) = \det(\lambda I_n - A)$, $\Xi(\lambda) = \mathrm{adj}(\lambda I_n - A)$, wie in Bemerkung 3.15. Dann folgt sofort

$$G = W\Gamma = [G_1, \ldots, G_q]\mathrm{Diag}(\hat{\Phi}_1, \ldots, \hat{\Phi}_q),$$

$$Z = -[Z_1, \ldots, Z_q]\mathrm{Diag}(\hat{\Phi}_1, \ldots, \hat{\Phi}_q),$$

where

$$\begin{aligned} G_k &= [\Xi(\lambda_k)B, \Xi^{(1)}(\lambda_k)B, \ldots, \Xi^{(p_k-1)}(\lambda_k)B], && (3.32)\\ Z_k &= [\xi(\lambda_k), \xi^{(1)}(\lambda_k), \ldots, \xi^{(p_k-1)}(\lambda_k)], && (3.33)\\ \hat{\Phi}_k &= \sum_{j=0}^{p_k-1} \phi_{k,j+1} N_{p_k}^j. \end{aligned}$$

Hier sind $\xi^{(k)}$ und $\Xi^{(k)}$ die k-ten Ableitungen bzgl. λ. Wir brauchen natürlich $\hat{\Phi}_k$ nicht singulär für $1 \leq k \leq q$, und damit ergibt sich

$$G := [G_1, \ldots, G_q], \quad , F = -[Z_1, \ldots, Z_q]G^{-1},$$

mit G_k, Z_k wie in (3.32) und (3.33).

Obwohl wir eine beliebige Polmenge zuweisen können, so gilt dies nicht für die Jordanstruktur. Dazu brauchen wir ein invertierbares Γ wie in (3.28).

Können wir vielleicht immer ein diagonalisierbares $A - BF$ bekommen? Es ist klar, dass das System nur dann numerisch robust wird, wenn dies gilt, und keine mehrfachen Eigenwerte auftauchen.

Lemma 3.33 *Sei (A, B) vollständig steuerbar. Gegeben eine Menge $\lambda_1, \ldots, \lambda_k$, und eine natürliche Zahl l mit $1 \leq l \leq s$. Für jedes λ_i wähle $g_i \in \mathcal{U}_{\lambda_i,l}$, wobei $\mathcal{U}_{\lambda_i,l}$ wie in (3.19) ein Unterraum des Kerns von $[A-\lambda_i I, B]$ ist. Falls $k > \sum_{i=1}^{l} d_i i$, so sind die $g_1, \ldots, g_k$ linear unabhängig.*

Beweis: Da $g_i \in \mathcal{U}_{\lambda_i,l}$, so existiert $\phi_i = \begin{bmatrix} \hat{\phi}_i \\ 0 \end{bmatrix}$, mit $\hat{\phi}_i \in \mathbb{C}^{\pi_l}$, so dass $g_i = U_{\lambda_i,s}\phi_i$. Sei $\Phi_k := [\phi_1, \ldots, \phi_k]$, $\Lambda_k := \mathrm{Diag}(\lambda_1, \ldots, \lambda_k)$ und $\Gamma_k = \begin{bmatrix} \Phi_k \\ \mathcal{I}_{2,1}\Phi_k\Lambda_k \\ \vdots \\ \mathcal{I}_{s,1}\Phi_k\Lambda_k^{s-1} \end{bmatrix}$. Nach Lemma 3.20, ist$G_k = [g_1, \ldots, g_k] = W\Gamma_k$ und, da W invertierbar, so gilt $\mathrm{Rang}G_k = \mathrm{Rang}\Gamma_k$. Durch eine Zeilenvertauschung erhalten wir, dass Γ_k zu $\begin{bmatrix} \hat{\Gamma}_k \\ 0 \end{bmatrix}$ transformiert werden kann mit $\hat{\Gamma}_k = \begin{bmatrix} \hat{\Phi}_{k,1} \\ \hat{\Phi}_{k,2}\Lambda_k \\ \vdots \\ \hat{\Phi}_{k,l}\Lambda_k^{l-1} \end{bmatrix}$ wobei $\hat{\Phi}_{k,1} = [\hat{\phi}_1, \ldots, \hat{\phi}_k]$, $\hat{\Phi}_{k,i}$ die untere $(\pi_l - \pi_{i-1}) \times k$ Untermatrix von $\hat{\Phi}_{k,1}$ ist. Weil die Anzahl der Zeilen von $\hat{\Gamma}_k$ gerade $\sum_{i=1}^{l}(\pi_l - \pi_{i-1}) = \sum_{i=1}^{l} d_i i$ ist, so folgt

$$\mathrm{Rang}G_k = \mathrm{Rang}\Gamma_k = \mathrm{Rang}\hat{\Gamma}_k \leq \sum_{i=1}^{l} d_i i.$$

Und damit impliziert $k > \sum_{i=1}^{l} d_i i$, dass $g_1, \ldots, g_k$ linear unabhängig sind. ∎

Satz 3.34 *Sei (A,B) vollständig steuerbar. Gegeben Pole $\lambda_1,\ldots,\lambda_q$ mit Multiplizitäten $r_1,\ldots,r_q$, so dass $r_1 \geq r_2 \geq \cdots \geq r_q$. Dann gibt es F so dass das Spektrum von $A-BF$ gegeben ist durch $\{\lambda_1,\ldots,\lambda_q\}$. Weierhin ist $A-BF$ diagonalisierbar genau dann wenn*

$$\sum_{i=1}^{k} r_i \leq \sum_{i=1}^{k} n_i,\ k=1,\ldots,q. \tag{3.35}$$

Beweis: Angenommen so ein F und eine invertierbares G existieren, so dass (3.43) gilt. Partitioniere $G := [G_1,\ldots,G_q]$, so dass $G_i \in \mathbb{C}^{n,r_i}$ mit Bild$G_i \subseteq \mathcal{U}_{\lambda_i}$. Wir beweisen (3.35) per Induction.

Falls $k=1$, so folgt aus Theorem 3.16, dass $\dim\mathcal{U}_{\lambda_1} = m = n_1$. Da Bild$G_1 \subseteq \mathcal{U}_{\lambda_1}$, so gilt Rang$G_1 \leq n_1$, aber da G invertierbar ist so folgt Rang$G_1 = r_1$ und damit $r_1 \leq n_1$.

Nun gelte (3.35) für k. Falls (3.35) nicht für $k+1$ gilt so folgt aus der Induktionsvoraussetzung, dass $r_1 \geq \ldots \geq r_{k+1} > n_{k+1}$. Da G_i vollen Spaltenrang hat und nach Satz 3.16, $n_{k+1} = m - \pi_k = \dim\mathcal{U}_{\lambda_i} - \dim\mathcal{U}_{\lambda_i,k}$, so folgt $l_i := \dim(\text{Bild}G_i \cap \mathcal{U}_{\lambda_i,k}) \geq r_i - n_{k+1}$, $i=1,\ldots,k+1$. Sei $g_{i,1},\ldots,g_{i,l_i}$ eine Basis von Bild$G_i \cap \mathcal{U}_{\lambda_i,k}$. Da

$$\sum_{i=1}^{k+1} l_i \geq \sum_{i=1}^{k+1}(r_i - n_{k+1}) > \sum_{i=1}^{k}(n_i - n_{k+1}) = \sum_{i=1}^{k} d_i i,$$

so folgt mit Lemma 3.33, dass $g_{1,1},\ldots,g_{1,l_1},\ldots,g_{k+1,1},\ldots,g_{k+1,l_{k+1}}$ linear abhängig sind, also gibt es eine Vektor $\nu \neq 0$ so dass $[G_1,\ldots,G_{k+1}]\nu = 0$ und dies ein Widerpsruch.

Für die Umkehrung verwende Satz 3.26, und konstruiere $\Phi \in \mathbb{C}^{m,n}$, so dass

$$\Gamma = \begin{bmatrix} \Phi \\ \mathcal{I}_{2,1}\Phi\Psi \\ \vdots \\ \mathcal{I}_{s,1}\Phi\Psi^{s-1} \end{bmatrix}$$

invertierbar ist und Ψ diagonal mit den Polen auf der Diagonale, d.h. $P\Lambda P^T$ mit Λ wie in (3.43) und P Permutationsmatrix. Sei

$$\Phi := \begin{array}{c} \\ d_1 \\ d_2 \\ \vdots \\ d_s \end{array} \begin{array}{c} \begin{array}{cccc} d_1 & 2d_2 & \ldots & sd_s \end{array} \\ \begin{bmatrix} \Phi_{1,1} & \Phi_{1,2} & \ldots & \Phi_{1,s} \\ & \Phi_{2,2} & \ldots & \Phi_{2,s} \\ & & \ddots & \vdots \\ & & & \Phi_{s,s} \end{bmatrix} \end{array}, \text{ mit } \Phi_{i,i} = \begin{bmatrix} \phi_{1,1}^{(i)} & \ldots & \phi_{1,d_i}^{(i)} \\ & \ddots & \vdots \\ & & \phi_{d_i,d_i}^{(i)} \end{bmatrix}$$

und $\phi_{j,j}^{(i)} = \begin{bmatrix} \omega_1^{(i,j)},\ldots,\omega_i^{(i,j)} \end{bmatrix} \in \mathbb{C}^{1,i}$ with $\omega_l^{(i,j)} \neq 0$ for all $i=1,\ldots,s$, $j=1,\ldots,d_i$, $l=1,\ldots,i$. Partitioniere Ψ als

$$\begin{array}{c} \\ d_1 \\ 2d_2 \\ \vdots \\ sd_s \end{array} \begin{array}{c} \begin{array}{cccc} d_1 & 2d_2 & \ldots & sd_s \end{array} \\ \begin{bmatrix} \Psi_1 & & & \\ & \Psi_2 & & \\ & & \ddots & \\ & & & \Psi_s \end{bmatrix} \end{array}, \text{ mit } \Psi_i = \begin{bmatrix} \psi_{i,1} & & \\ & \ddots & \\ & & \psi_{i,d_i} \end{bmatrix}$$

und $\psi_{i,j} = \mathrm{Diag}(\nu_1^{(i,j)}, \ldots, \nu_i^{(i,j)})$.

Dann erhalten wir

$$\Gamma = \begin{bmatrix} \Phi_{1,1} & \Phi_{1,2} & \ldots & \ldots & \Phi_{1,s} \\ & \Phi_{2,2} & & & \Phi_{2,s} \\ & & \ddots & & \vdots \\ & & & \ddots & \Phi_{s,s} \\ & \Phi_{2,2}\Psi_2 & \ldots & \ldots & \Phi_{2,s}\Psi_s \\ & & \ddots & & \vdots \\ & & & & \Phi_{s,s}\Psi_s \\ & & & & \vdots \\ & & & \Phi_{s-1,s-1}\Psi_{s-1}^{s-2} & \Phi_{s-1,s}\Psi_s^{s-2} \\ & & & 0 & \Phi_{s,s}\Psi_s^{s-2} \\ & & & & \Phi_{s,s}\Psi_s^{s-1} \end{bmatrix}.$$

Aus der Form von $\Phi_{i,i}$ folgt, dass durch Anwendung einer Zeilenpermutation Γ transformiert werden kann als

$$\hat{\Gamma} = \begin{bmatrix} \hat{\Gamma}_1 & * & \ldots & * \\ & \hat{\Gamma}_2 & & \vdots \\ & & \ddots & \vdots \\ & & & \hat{\Gamma}_s \end{bmatrix}, \quad \text{mit } \hat{\Gamma}_i = \begin{bmatrix} \hat{\Gamma}_{1,1}^{(i)} & * & \ldots & * \\ & \hat{\Gamma}_{2,2}^{(i)} & & \vdots \\ & & \ddots & \vdots \\ & & & \hat{\Gamma}_{d_i,d_i}^{(i)} \end{bmatrix}$$

und

$$\hat{\Gamma}_{j,j}^{(i)} = \begin{bmatrix} 1 & \ldots & 1 \\ \nu_1^{(i,j)} & \ldots & \nu_i^{(i,j)} \\ \vdots & & \vdots \\ (\nu_1^{(i,j)})^{i-1} & \ldots & (\nu_i^{(i,j)})^{i-1} \end{bmatrix} \mathrm{Diag}(\omega_1^{(i,j)}, \ldots, \omega_i^{(i,j)}).$$

Da $\hat{\Gamma}$ Block-obere Dreieecksmatrix ist und da jedes $\hat{\Gamma}_{j,j}^{(i)}$ Produkt einer nichtsingulären Diagonalmatrix und einer Vandermonde Matrix ist, welche invertierbar ist wenn $\nu_1^{(i,j)}, \ldots, \nu_i^{(i,j)}$ paarweise verschieden sind, so ist $\hat{\Gamma}$ und damit Γ nicht singulär. Es bleibt zu zeigen, dass $\nu_j^{(i,j)}$ aus den Eigenwerten so gewählt werden kann, dass alle auftretenden Vandermonde Matrizen invertierbar sind. Dies ist durch (3.50) garantiert. □

Wie gut sind nun diese Ergebnisse bei kleinen Störungen durch Rundungs-, Linearisierungs- oder Modellfehler.

Satz 3.36 *(Mehrmann/Xu 1998) Gegeben ein vollständig steuerbares Paar (A, B), und eine Polmenge $\mathcal{P} = \{\lambda_1, \ldots, \lambda_n\}$. Betrachte ein gestörtes System $(\hat{A}, \hat{B})$ welches immer noch vollständig steuerbar ist, und eine gestörte Polmenge $\hat{\mathcal{P}} = \{\hat{\lambda}_1, \ldots, \hat{\lambda}_n\}$. Setze $\hat{A} - A =: \delta A$, $\hat{B} - B =: \delta B$ and $\hat{\lambda}_k - \lambda_k =: \delta\lambda_k$, $k = 1, \ldots, n$. Angenommen beide Polvorgabeprobleme haben Lösungen mit diagonalisierbarer Matrix des geschlossenen Kreises.*

Mit

$$\epsilon := \|[\delta A, \delta B]\|. \tag{3.37}$$

und unter der Voraussetzung, dass

$$\max_i \frac{\epsilon + |\delta\lambda_i|}{\sigma_n([A-\lambda_i I, B])} < \frac{3}{4}, \tag{3.38}$$

so gibt es Rückkopplungsmatrix $\hat{F} := F + \delta F$ *für* $(\hat{A}, \hat{B})$ *so dass*

$$\|\delta F\| < \frac{5\sqrt{n}}{4}\kappa\sqrt{1+\|\hat{F}\|^2}\max_i\{\frac{\sqrt{1+(\|B^\dagger(A-\lambda_i I)\|)^2}\,(\epsilon+|\delta\lambda_i|)}{\sigma_n([A-\lambda_i I, B])}\}, \tag{3.39}$$

$\lambda(\hat{A}-\hat{B}\hat{F}) = \hat{\mathcal{P}}$ *und* $\hat{A}-\hat{B}\hat{F}$ *ist diagonalisierbar.*

Weierthin gilt für jeden Eigenwert μ_i *von* $A - B\hat{F}$*, (d.h., die gestörte Rückkopplung wird für das ungestörte Systeme verwendet) ein* $\lambda_i \in \mathcal{P}$ *so dass*

$$|\mu_i - \lambda_i| < |\delta\lambda_i| + \epsilon\hat{\kappa}\sqrt{1+\|\hat{F}\|^2}. \tag{3.40}$$

Hier sind κ *und* $\hat{\kappa}$ *die skalierten spektralen Konditionszahlen von* $A-BF$ *und* $\hat{A}-\hat{B}\hat{F}$*.* $\sigma_n(A)$ *ist der kleinste Singulärwert von* A*, und* $B^\dagger$ *ist die Moore-Penrose Pseudoinverse von* B*.*

Beweis: Siehe Orginalarbeit. □

Der Hauptfaktor in der Störungstheorie ist $\mathcal{S} := \kappa\sqrt{1+\|F\|^2}$. Dazu gibt es in der Schranke für F noch den Faktor $d := 1/\min_i \sigma_n[A-\lambda_i I, B]$ welcher eng mit dem *Abstand des Systems zur Nichsteuerbarkeit*

$$d_u(A,B) = \min_{\lambda\in\mathbb{C}} \sigma_n[A-\lambda I, B], \tag{3.41}$$

zusammenhängt. Falls $d_u(A,B)$ klein ist, so kann d sehr groß sein und damit ist die Berechnung von F schlecht konditioniert. Falls $d_u(A,B)$ groß ist, so ist d klein und damit ist die Schranke auch klein in d.

Die Analyse von $\mathcal{S}$ ist sehr viel schwieriger. Falls $m = 1$, so ist z.B. $\mathcal{S}$ die Konditionszahl der Cauchy Matrix $C = [\frac{1}{\nu_i - \lambda_j}]$, die mit n sehr schnell wächst. Aus (3.39) folgt, dass $\mathcal{S}$ sich verhält, wie κ^2 und auch dies kann sehr schnell wachsen. Heuristisch gesprochen folgt mit $\frac{n}{m} \le s \le n-m+1$ aus der Staircase Form, dass für n groß auch groß sein muss, damit das Problem der Polvorgabe gut konditionert ist.

Beispiel 3.42 Sei $A = \mathrm{Diag}(1,\ldots,20)$, $\mathcal{P} = \{-1,\ldots,-20\}$ und sei B eine Matrix die aus den ersten M Spalten einer zufälligen 20×20 Orthogonalmatrix gebildet wird.

Die folgenden Resulte wurden auf einem Pentium-s PC mit $eps = 2.22\times 10^{-16}$, unter Matlab Version 4.2 erzielt mit Hilfe des Polvorgabe-Algithmus von Miminis/Paige, s.u.. Für $m = 1,\ldots,20$ und jeweils 20 Tests ergaben sich die folgenden geometrischen Mittel von $\hat{\kappa}$, $\hat{F}$, bound, err, wobei bound=$eps\|[A,B]\|\hat{\kappa}\sqrt{1+\|\hat{F}\|^2}$, und err=$\max_{1\le i\le 20}|\mu_i-\lambda_i|$, mit λ_i und den realteilen von μ_i in aufsteigender Folge. In der zweiten Spalte steht das mittel von s über 20 Tests für jedes m. Für alle 400 Tests variierte $\min_i \sigma_n([A-\lambda_i I, B])$ von 2.0 bis 2.24.

m	s	$\hat{\kappa}$	$\hat{F}$	Bound	Err
1	20	3.5×10^{9}	1.1×10^{14}	1.7×10^{9}	7.3×10^{4}
2	10	1.8×10^{11}	5.0×10^{9}	3.9×10^{6}	2.7×10^{2}
3	7	2.1×10^{10}	2.4×10^{10}	2.2×10^{6}	1.4×10^{2}
4	5	7.4×10^{11}	5.8×10^{7}	1.9×10^{5}	2.4×10^{1}
5	4	1.2×10^{14}	1.3×10^{5}	7.3×10^{4}	1.0×10^{1}
6	4	2.1×10^{14}	2.6×10^{4}	2.5×10^{4}	5.8
7	3	1.7×10^{14}	4.2×10^{4}	3.1×10^{4}	2.0
8	3	1.7×10^{14}	1.1×10^{4}	8.6×10^{3}	7.8×10^{-1}
9	3	2.4×10^{14}	9.0×10^{3}	9.8×10^{3}	6.6×10^{-1}
10	2	2.1×10^{14}	2.6×10^{3}	2.9×10^{3}	3.8×10^{-1}
11	2	1.8×10^{13}	7.9×10^{2}	6.5×10^{1}	1.0×10^{-4}
12	2	9.2×10^{12}	5.0×10^{2}	2.0×10^{1}	3.6×10^{-3}
13	2	5.7×10^{11}	4.5×10^{2}	1.1	1.5×10^{-4}
14	2	2.1×10^{11}	3.2×10^{2}	3.0×10^{-1}	6.7×10^{-5}
15	2	3.4×10^{10}	2.8×10^{2}	4.2×10^{-2}	1.3×10^{-5}
16	2	5.9×10^{8}	2.6×10^{2}	6.7×10^{-4}	3.0×10^{-7}
17	2	3.1×10^{7}	2.2×10^{2}	3.0×10^{-5}	1.6×10^{-8}
18	2	1.6×10^{5}	2.0×10^{2}	1.4×10^{-7}	1.0×10^{-10}
19	2	7.0×10^{2}	1.9×10^{2}	5.9×10^{-10}	9.9×10^{-13}
20	1	1.0	3.5×10^{1}	1.5×10^{-13}	2.6×10^{-14}

Table 1

Nun zu Polvorgabe-Algorithmen, soweit das nach den obigen Betrachtungen noch Sinn macht. Es gibt viele 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 3.43 *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 *unreduziert*. Die Transformation auf Hessenbergform ist nicht eindeutig. Wir können im Prinzip die erste Spalte der Transformationsmatrix direkt vorgeben:

Satz 3.44 *Seien $Q = [q_1, \ldots, q_n]$ und $V = [v_1, \ldots, v_n]$ orthogonale Matrizen, so daß*

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

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 $i = 2 : k$

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

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 $[w_1, \ldots, w_k]$ 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

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

□

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

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

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

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.

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

Algorithmus 3.46 *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* $P = I - 2vv^T/v^Tv$.
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: $fl(\tilde{P}A) = P(A + E),\ \|E\|_2 \leq \mathcal{O}(eps\,\|A\|_2)$

Algorithmus 3.47 *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 3.48 *(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 $Q_0^T b = \begin{bmatrix} * \\ 0 \\ \vdots \\ 0 \end{bmatrix}$.

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

Schritt 1 Transformiere A mit dem nachfolgenden Algorithmus 3.49 auf Hessenbergform, wobei die erste Spalte der Transformationsmatrix der erste Einheitsvektor e_1 ist.

Algorithmus 3.49 *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* $Q^T AQ$ *obere Hessenbergmatrix ist.*

FOR $k = 1 : n-2$

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

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

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

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 3.50 *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 $Z^T(A - a_1 I)(A - a_2 I)$ obere $\triangle$–Matrix ist) und A wird überschrieben mit $Z^T AZ$.
$l = n - 1$

% { Berechne 1. Spalte von $(A - a_1 I)(A - a_2 I)$ }
$s = A(l, l) + A(n, n)$
$t = A(l, l)A(n, n) - A(l, n)A(n, l)$
$x = A(1, 1) * A(1, 1) + A(1, 2) * A(2, 1) - sA(1, 1) + t$
$y = A(2, 1) * (A(1, 1) + A(2, 2) - s)$
$z = A(2, 1) * A(3, 2)$
For $k = 0 : n - 3$

% Überschreibe A mit $P_k A P_k^T$, 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 $P_{n-2} = \begin{pmatrix} I_{n-2} & \\ & \tilde{P}_{n-2} \end{pmatrix}$
$A(n - 1 : n, n - 2 : n) = row.house\ (A(n - 1 : n, n - 2 : n), v)$
$A(1 : n, n - 1 : n) = col.house\ (A(1 : n, n - 1 : n), v)$

Kosten: $10n^2$ 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 3.51 *QR–Algorithmus für Hessenbergmatrizen*

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

Qutput: 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

$$|a_{i,i-1}| \leq \ tol \ (|a_{ii}| + |a_{i-1,i-1}|) \tag{3.52}$$

erfüllen zu 0.

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

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix} \begin{matrix} p \\ n-p-q \\ q \end{matrix}$$
$$\begin{matrix} p & n-p-q & q \end{matrix}$$

wobei A_{33} *quasi obere* $\triangle$*–Matrix,* A_{22} *unreduzierte Hessenbergmatrix ist.*
If $q < n$
Wende Algorithmus 3.50 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.*

Kosten: 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.*

Fehler: Die berechnete Schurform ist die exakte Schurform von $A + E$

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

mit $Q^T Q = I$ *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.

3.0.3 Polvorgabe Algorithmen

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 $(P^T AP, P^T b)$ System Hessenbergform.
- Bestimme $\tilde{f} \in \mathbb{R}^n$ so daß $\sigma(P^T AP - P^T b\tilde{f}^T) = \Omega$.
- 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

$$\begin{aligned} &[A - \lambda_1 I - \beta e_1 f^T]y = 0 \\ \Longrightarrow\ &e_1^T(A - \lambda_1 I)y = \beta f^T y \end{aligned}$$

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 3.53 *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.*

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

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.

Beispiel 3.54 *Betrachte das folgende Problem*

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

Schritt 1
$i = 1$:

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

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

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

$$\text{Rang } (B) = m \tag{3.55}$$

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 der Polvorgabe ab:

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

$$(A - BF)X = X\Lambda \tag{3.56}$$

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

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 3.57 *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 (3.56) löst genau dann, wenn*

$$U_1^T(AX - X\Lambda) = 0 \tag{3.58}$$

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*

$$F = Z^{-1} U_0^T (X \Lambda X^{-1} - A). \tag{3.59}$$

Beweis: Da B vollen Rang hat, so existiert die Zerlegung von B (z. B. QR–Zerlegung). Aus (3.56) folgt

$$BF = A - X \Lambda X^{-1}. \tag{3.60}$$

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

$$\begin{aligned} ZF &= U_0^T (A - X \Lambda X^{-1}) \\ 0 &= U_1^T (A - X \Lambda X^{-1}), \end{aligned} \tag{3.61}$$

woraus die Behauptung folgt, da Z invertierbar ist. □

Es folgt, daß F existiert genau dann wenn

$$\text{Bild } (A - X \Lambda X^{-1}) \subset \text{ Bild } (B) = \text{ Bild } (U_0). \tag{3.62}$$

Korollar 3.63 *Der Eigenvektor* $x_j \in \mathbb{C}^n$ *von* $A - BF$ *zum vorgegebenen Eigenwert* $\lambda_j \in \Omega$ *muß in dem Raum*

$$\mathcal{S}_j = \; Kern \; \left(U_1^T (A - \lambda_j I) \right) \tag{3.64}$$

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

$$k_j = \; dim \; \left(\; Kern \; [B \mid A - \lambda_j I]^T \right) \tag{3.65}$$

Beweis: Wir haben aus (3.61) sofort

$$U_1^T (A x_j - \lambda_j x_j) = 0 \quad \forall_j = 1, \cdots, n \tag{3.66}$$

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 (3.65) 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 (3.56) möglichst gut konditioniert ist.

Was für Bedingungen müssen nun notwendigerweise erfüllt sein? Aus Korollar 3.63 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 $\Omega = \Omega_s \cup \Omega_u, \quad \mid \Omega_u \mid = n_s$, 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 $y_l, l = 1, \dots k_j$ 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 3.67 *Eine notwendige Bedingung für die Existenz einer Lösung von Problem 3.56 bei der $A - BF$ keinen Jordan–Block hat ist, dass für jedes $\mu \in \mathbb{C}, s \in \mathbb{R}^n$ gilt:*

$$\left\{ s^T B = 0, s^T (A - \mu I) B = 0, s^T (A - \mu I)^2 = 0 \right\} \tag{3.68}$$
$$\Longrightarrow s^T (A - \mu I) = 0$$

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. □

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.

$$cond_2(x) = \|X\|_2 \|X^{-1}\|_2. \tag{3.69}$$

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

$$\dot{x}(t) = (A - BF) x(t), \quad x(0) = x^0 \tag{3.70}$$

Dann gilt der folgende Satz:

Satz 3.71 *Die Rückkopplungsmatrix F und die Lösung von (3.70) erfüllen*

$$\|F\|_2 \leq \left(\|A\|_2 + \max_j \{ \mid \lambda_j \mid \} \cdot cond_2(X) \right) / \sigma_m(B) \tag{3.72}$$

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

$$\|x(t)\|_2 \leq cond_2(X) \cdot \max_j \{ \mid e^{\lambda_j t} \mid \} \cdot \|x^0\|_2 \tag{3.73}$$

Beweis: Aus (3.59) 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.

$$\Longrightarrow \|F\|_2 \leq \|Z^{-1}\|_2\|U_0^T\|_2\left(\|X\|_2\|X^{-1}\|_2\|\Lambda\|_2 + \|A\|_2\right) \tag{3.74}$$

$$\|Z^{-1}\|_2 = \|V\|_2\sigma_m^{-1}(B),\ \text{aber}\ \|V\|_2 = 1 = \|U_0^T\|_2 \Longrightarrow (3.72).$$

Ungleichung (3.73) folgt aus

$$x(t) = e^{(A-BF)t}x^0 = Xe^{\Lambda t}X^{-1}x^0. \tag{3.75}$$

□

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 3.76 *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:*

$$\|\Delta\|_2 < \min_{s=i\omega} \sigma_n \{sI - (A - BF)\} =: \delta(F) \tag{3.77}$$

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

$$\delta(F) \geq \min_j\{Re(-\lambda_j)\}/cond_2(X) \tag{3.78}$$

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 (3.77) gilt. Die andere Ungleichung folgt aus

$$\begin{aligned} \delta(F) &= \min_{s=i\omega}\left(\sigma_n(sI - X\Lambda X^{-1})\right) \\ &\geq \min_j Re(-\lambda_j)/\|X^{-1}\|_2\|X\|_2. \end{aligned} \tag{3.79}$$

□

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 $\text{cond}_2(X)$ 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

span$S_j = \mathcal{S}_j$. Dann gilt $X_j = S_j D_j$ 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.

$$X = [S_1, \cdots, S_k] \begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_k \end{bmatrix} := SD \tag{3.80}$$

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

$$v(D) = \frac{\|DX^{-1}\|_F}{\|D\|_F}. \tag{3.81}$$

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 $\tilde{x}_j = S_j w_j$ und $\|w_j\|_2 = 1$.

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

$$\|X^{-1}\|_F = \|[S_j w_j,\ X_j]^{-1}\|_F = \|[y_j,\ Y_j]^T\|_F = \|Y^T\|_F \tag{3.82}$$

Mittels einer QR–Zerlegung

$$X_j = [Q_j,\ q_j] \begin{bmatrix} R_j \\ 0 \end{bmatrix} \tag{3.83}$$

und $Y^T X = I$ erhalten wir

$$\|Y^T\|_F = \left\| \begin{bmatrix} 0 & \rho_j \\ R_j^{-1} & -\rho_j R_j^{-1} Q_j^T S_j w_j \end{bmatrix} \right\|_F \tag{3.84}$$

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

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

Also müssen wir das kleinste Quadrate Problem

$$\min \left\| \begin{bmatrix} R_j^{-1} Q_j^T S_j \\ I_m \end{bmatrix} (P_j \tilde{w}_j + p_j) \right\|_2 \quad \text{lösen, für } \tilde{w}_j \tag{3.85}$$

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

$$\tilde{x}_j = S_j w_j = (\rho_j \sigma_j)^{-1} S_j (P_j \tilde{w}_j + p_j) \tag{3.86}$$

mit

$$\rho_j^2 = \rho_j^2 w_j^T w_j = \sigma_j^{-1}(\tilde{w}_j^T \tilde{w}_j + 1) \tag{3.87}$$

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 3.88 *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 $U = [U_0,\ U_1]$ *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.

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 vieles noch

offen. Man sollte hier beachten, daß Polvorgabe in den meisten Fällen nur verwendet wird, um das System zu stabilisieren. Eine andere Möglichkeit zur Stabilisierung ist die Lösung eines Optimalsteuerungsprobems. Dies betrachten wir im nächsten Kapitel.

Kapitel 4

Optimale Steuerung

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

Wir haben wieder unsere Systemgleichung

$$\dot{x} = Ax + Bu \quad x(t_0) = x_0 \tag{4.1}$$

$$A \in \mathbb{R}^{n,n}, B \in \mathbb{R}^{n,m}$$

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

$$\begin{aligned} \mathcal{S}\left(x(t), u(t)\right. &= \frac{1}{2}\left(x(t_f)^T M x(t_f)\right) \\ &\quad + \int_{t_0}^{t_f} \{x(t)^T Q x(t) + u(t)^T R u(t) \\ &\quad + x(t)^T S u(t) + u(t)^T S^T x(t)\} dt. \end{aligned} \tag{4.2}$$

$$t_0 < t_f \leq \infty,$$

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 4.2 *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 4.3 *Betrachte das Optimalsteuerungsproblem (4.1), (4.2). 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*

$$\dot{x}(t) = Ax(t) + Bu_*(t) \quad x(t_0) = x^0 \tag{4.4}$$

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

$$\begin{bmatrix} A & 0 & B \\ Q & A^T & S \\ S^T & B^T & R \end{bmatrix} \begin{bmatrix} x(t) \\ \mu(t) \\ u(t) \end{bmatrix} = \begin{bmatrix} I_n & 0 & 0 \\ 0 & -I_n & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ \dot{\mu}(t) \\ \dot{u}(t) \end{bmatrix} \tag{4.5}$$

$$x(t_0) = x^0, \ \mu(t_f) = Mx(t_f) \tag{4.6}$$

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

$$u(t) = u_*(t) + \varepsilon v(t) \tag{4.7}$$

mit $u(t) \in U_m$. Dann erhalten wir aus (4.4)

$$\dot{x}(t) = Ax(t) + Bu_*(t) + \varepsilon Bv(t). \tag{4.8}$$

Dies hat die Lösung (siehe Kapitel 2)

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

Dabei erfüllt $\varphi(t)$ die Differentialgleichung

$$\dot{\varphi}(t) = A\varphi(t) + Bv(t), \varphi(t_0) = 0. \tag{4.9}$$

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

$$\begin{aligned} H(x,\mu,u) &= x^T(t)Qx(t) + x(t)^T Su(t) + u(t)^T S^T x(t) \\ &\quad + u(t)^T Ru(t) + \mu(t)^T \left(Ax(t) + Bu(t)\right) \\ &\quad + \left(Ax(t) + Bu(t)\right)^T \mu(t) \end{aligned} \tag{4.10}$$

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

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

$$\mathcal{S}(x,u) = \frac{1}{2}\left\{ x^T(t_f)Mx(t_f) + \int_{t_0}^{t_f} \left(H(x,\mu,u) - \mu^T\dot{x} - \dot{x}^T\mu\right) dt \right\} \tag{4.10}$$

und analog für u_*, x_*

$$\mathcal{S}(x_*,u_*) = \frac{1}{2}\left\{ x_*^T(t_f)Mx_*(t_f) + \int_{t_0}^{t_f} H(x_*,\mu,u_*) - \mu^T\dot{x}_* - \dot{x}_*^T\mu)dt \right\} \tag{4.11}$$

und wir erhalten

$$\begin{aligned} \mathcal{S}(x,u) - \mathcal{S}(x_*,u_*) &= \frac{1}{2}\left\{ \left(x(t)^T Mx(t) - x_*(t)^T Mx_*(t)\right) |_{t=t_f} \right. \\ &\quad + \int_{t_0}^{t_f} \left(H(x,\mu,u) - H(x_*,\mu,u_*)\right) dt \\ &\quad \left. + \int_{t_0}^{t_f} \underbrace{\left(\mu^T \underbrace{(\dot{x}_* - \dot{x})}_{-\varepsilon\dot{\varphi}} + \underbrace{(\dot{x}_* - \dot{x})^T}_{-\varepsilon\dot{\varphi}^T}\mu\right)}_{=-2\varepsilon\mu^T\dot{\varphi}} dt \right\} \end{aligned} \tag{4.12}$$

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

Wir verwenden jetzt die Tatsache, das

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

und erhalten

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

Weiter gilt mit partieller Integration

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

$$x^TMx - x_*^TMx_* = 2\varepsilon x_*^TM\varphi + \mathcal{O}(\varepsilon^2)$$

Zusammen erhalten wir also

$$\begin{aligned} \mathcal{S}(x,u) - \mathcal{S}(x_*,u_*) &= \mathcal{O}(\varepsilon^2) \\ &\quad +\varepsilon\Big\{\int_{t_0}^{t_f} \big([x_*^TQ + u_*^TS^T + \mu^TA]\varphi + \dot{\mu}^T\varphi \\ &\quad +[x_*^TS + u_*^TR + \mu^TB]v\big)\,dt \\ &\quad -\mu^T(t_f)\varphi(t_f) + x_*^T(t_f)M\varphi(t_f)\Big\}. \end{aligned} \tag{4.13}$$

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

$$-\dot{\mu}(t) = A^T\mu(t) + Qx_* + Su_* \tag{4.12}$$

mit „Endbedingung“

$$\mu(t_f) = Mx_*(t_f) \tag{4.13}$$

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

$$\int_{t_0}^{t_f} [x_*^TS + u_*^TR + \mu^TB]vdt = 0 \quad \forall v \in U_m \tag{4.14}$$

woraus sofort folgt, dass

$$x_*^TS + u_*^TR + \mu^TB \equiv 0 \quad \forall t \in [t_0, t_f] \tag{4.15}$$

Fassen wir nun die Gleichungen (4.15), (4.12), (4.13) und (4.1) zusammen, so erhalten wir das Zwei–Punkt–Randwertproblem (4.5),(4.6). □

Ein Punkt, der oft schwer nachzuvollziehen ist, ist die Wahl von μ in (4.12), (4.13). Diese kann man noch näher begründen durch geschickte Auswahl von v, φ. Aber das würde hier zu weit führen. Die Existenz der Lösung von (4.12), (4.13) reicht hier für diesen Satz.
Dass die Wahl vernünftig ist, zeigt uns auch der folgende Satz:

Satz 4.16 *Seien x_*, μ, u_* so gewählt, dass $\begin{bmatrix} x_* \\ \mu \\ u_* \end{bmatrix}$ Lösung des linearen Randwertproblems (4.5), (4.6) ist. Es gelte weiterhin, dass $\mathcal{R} := \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix}$, M positiv semidefinit sind. Dann gilt*

$$\mathcal{S}(x,u) \geq \mathcal{S}(x_*,u_*) \tag{4.17}$$

für alle x, u welche (4.1) erfüllen.

Beweis: Definiere

$$\Phi(s) = \mathcal{S}\left(sx_*(t) + (1-s)x(t), su_*(t) + (1-s)u(t)\right). \tag{4.18}$$

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 (4.1) erfüllen.
$\Phi(s)$ ist quadratisch in s, also hat $\Phi(s)$ ein Minimum für $s = 1$ genau dann, wenn

$$\left.\frac{d\Phi}{ds}\right|_{s=1} = 0, \quad \left.\frac{d^2\Phi}{ds^2}\right|_{s=1} \geq 0 \tag{4.19}$$

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

$$\begin{aligned}
\left.\frac{d\Phi}{ds}\right|_{s=1} &= \Big[(x_* - x)^T M x_*\big|_{t=t_f} \\
&\quad + \frac{1}{2}\int_{t_0}^{t_f} \big\{ (x_* - x)^T Q x_* + x_*^T Q (x_* - x) \\
&\quad + u_*^T S^T (x_* - x) + (u_* - u)^T S^T x_* \\
&\quad + (u_* - u)^T R u_* + u_*^T R (u_* - u) \\
&\quad + (x_* - x)^T S u_* + x_*^T S (u_* - u) \big\}\, dt \Big]
\end{aligned} \tag{4.20}$$

Wenn wir die 2. Gleichung von (4.5) von links mit x_*^T multiplizieren und die anderen Gleichungen einsetzen, erhalten wir

$$\begin{aligned}
x_*^T Q x_* &= -x_*^T A^T \mu - x_*^T S u_* - x_*^T \dot{\mu} \\
(\text{ 1. Gl. (4.5) }) &= u_*^T B^T \mu - \dot{x}_*^T \mu - x_*^T S u_* - x_*^T \dot{\mu} \\
(\text{ 3. Gl. (4.5) }) &= -u_*^T S^T x_* - u_*^T R u_* - \dot{x}_*^T \mu - x_*^T S u_* - x_*^T \dot{\mu}
\end{aligned} \tag{4.21}$$

Analog nach Multiplikation mit x^T

$$\begin{aligned}
x^T Q x_* &= -x^T A^T \mu - x^T S u_* - x^T \dot{\mu} \\
(\text{ Ausgangsgl. für } x) &= u^T B^T \mu - \dot{x}^T \mu - x^T S u_* - x^T \dot{\mu} \\
(\text{ 2. Gl. für } B^T\mu) &= -u^T S^T x_* - u^T R u_* - \dot{x}^T \mu - x^T S u_* - x^T \dot{\mu}
\end{aligned} \tag{4.22}$$

Einsetzen von (4.21), (4.22) in (4.20) ergibt

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

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

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

□

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 (4.5), (4.6) 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

$$u(t) = -R^{-1}\left(S^T x(t) + B^T \mu(t)\right). \tag{4.20}$$

Einsetzen in die anderen Gleichungen ergibt das System

$$\begin{bmatrix} A - BR^{-1}S^T & -BR^{-1}B^T \\ -(Q - SR^{-1}S^T) & -A^T + SR^{-1}B^T \end{bmatrix} \begin{bmatrix} x \\ \mu \end{bmatrix} = \begin{bmatrix} \dot{x} \\ \dot{\mu} \end{bmatrix} \tag{4.21}$$

mit Randbedingungen

$$x(t_0) = x^0, \quad \mu(t_f) = Mx(t_f) \tag{4.22}$$

Wir betrachten die Matrix in (4.21) einmal näher. Da wir vorausgesetzt haben, dass

$$\mathcal{R} = \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \text{ semidefinit, } R \text{ positiv definit }, \tag{4.23}$$

so folgt $Q - SR^{-1}S^T$ *symmetrisch positiv semidefinit* (Schur–Komplement). Weiter ist $BR^{-1}B^T$ symmetrisch positiv semidefinit mit Rang m, also hat (4.21) die Form

$$\begin{bmatrix} \dot{x} \\ \dot{\mu} \end{bmatrix} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \begin{bmatrix} x \\ \mu \end{bmatrix} = \mathcal{H} \begin{bmatrix} x \\ \mu \end{bmatrix} \tag{4.24}$$

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 *Hamiltonische Matrix*. Hamiltonische Matrizen sind schiefsymmetrisch bezüglich eines indefiniten Skalarproduktes.

$$< x, y >_J = x^T J y \text{ wobei } J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}. \tag{4.25}$$

Beachte, dass gilt:

$$\begin{aligned} < x, \mathcal{H}y >_J &= x^T J \mathcal{H} y \\ &= x^T \begin{bmatrix} H & -F^T \\ -F & -G \end{bmatrix} y \\ &= x^T \begin{bmatrix} H^T & -F^T \\ -F & -G^T \end{bmatrix} y \\ &= x^T (J\mathcal{H})^T y \\ &= x^T \mathcal{H}^T J^T y \\ &= -x^T \mathcal{H}^T J y \\ &= - < x\mathcal{H}^T, y >_J \end{aligned} \tag{4.26}$$

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

$$\begin{aligned}\mathcal{H}_{2n}(\mathbb{R}) &= \left\{\mathcal{H} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} | G = G^T, H = H^T, F \in \mathbb{R}^{n,n}\right\} \\ \mathcal{H}_{2n}(\mathbb{C}) &= \left\{\mathcal{H} = \begin{bmatrix} F & G \\ H & -F^* \end{bmatrix} | G = G^*, H = H^*, F \in \mathbb{C}^{n,n}\right\}\end{aligned} \tag{4.26}$$

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

$$[\mathcal{H}_1, \mathcal{H}_2] := \mathcal{H}_1\mathcal{H}_2 - \mathcal{H}_2\mathcal{H}_1 \tag{4.27}$$

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

$$\begin{aligned}\mathcal{S}_{2n}(\mathbb{R}) &:= \left\{S \in \mathbb{R}^{2n,2n} | S^TJS = J\right\} \\ \mathcal{S}_{2n}(\mathbb{C}) &:= \{S \in \mathbb{C}^{2n,2n} | S^*JS = J\}\end{aligned} \tag{4.28}$$

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

$$< x, Sy >_J = x^TJSy = x^TS^{-T}Jy = < S^{-1}x, y >_J . \tag{4.29}$$

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

$$|\det S| = 1, \tag{4.30}$$

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

$$\begin{bmatrix} 1 & v \\ 0 & 1 \end{bmatrix} \in \mathcal{S}_2(\mathbb{R}) \quad \forall v \in \mathbb{R}. \tag{4.31}$$

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

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

Satz 4.32 *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 4.33 *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*

$$z_2^* G z_2 + z_1^* H z_1 = (\lambda + \bar{\lambda}) z_1^* z_2 = 2Re(\lambda) z_1^* z_2 \tag{4.34}$$

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

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

$$\Longrightarrow \left\{ \begin{array}{rcl} z_2^* F z_1 + z_2^* G z_2 & = & \lambda z_2^* z_1 \\ z_1^* H z_1 - z_1^* F^T z_2 & = & \lambda z_1^* z_2 \end{array} \right.$$

Konjugation der 2. Gleichung liefert

$$z_1^* H z_1 - z_2^* F z_1 = \bar{\lambda} z_2^* z_1$$

Addition der Gleichungen liefert

$$z_2^* G z_2 + z_1^* H z_1 = (\lambda + \bar{\lambda}) z_1^* z_2$$

□

Satz 4.35 *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*

$$z_2^* G z_2 + z_1^* H z_1 = 0$$

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

$$G z_2 = 0,\ H z_1 = 0$$

iii) Sei $\mathcal{H}$ *wie in (4.21),* $\mathcal{R} = \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix}$ *positiv definit und* (A, B) *stabilisierbar. So hat* $\mathcal{H}$ *keine Eigenwerte mit Realteil* 0*.*

Beweis:

i) klar aus Lemma 4.33.

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 (4.34) 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 $\Longrightarrow G z_2 = 0, H z_1 = 0$, 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 4.32 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 $z_2 \neq 0, z_2^* A = -\bar{\lambda} z_2^* \quad \big(Re(\bar{\lambda}) = 0\big)$ und $z_2^* B = 0$. ∎

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

$$S^{-1}\mathcal{H}S = \begin{bmatrix} J_1 & 0 \\ 0 & -J_1^* \end{bmatrix} \tag{4.37}$$

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

Beweis:

Aus Satz 4.32 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 4.32 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 4.32:

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

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

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

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

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

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

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

$$\begin{aligned} \mathcal{US}_{2n}(\mathbb{C}) &= \{Q \in \mathcal{S}_{2n}(\mathbb{C}) | Q^*Q = I\} \\ \mathcal{US}_{2n}(\mathbb{R}) &= \{Q \in \mathcal{S}_{2n}(\mathbb{R}) | Q^TQ = I\} \end{aligned} \tag{4.38}$$

Es gilt das folgende Lemma:

Lemma 4.39 *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:

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

□

Was können wir nun mit unitär symplektischen Transformationsmatrizen erreichen?

Satz 4.40 *(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*

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

$$Q^*\mathcal{H}Q = \begin{bmatrix} T & N \\ 0 & -T^* \end{bmatrix}, T, N \in \mathbb{C}^{n,n} \tag{4.41}$$

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

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

$$Q^T \mathcal{H} Q = \begin{bmatrix} T & N \\ 0 & -T^T \end{bmatrix} \quad T, N \in \mathbb{R}^{n,n}, \tag{4.42}$$

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

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

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 $Q_1 \in \mathcal{U}S_{2n}(\mathbb{R})$ 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 $Q_1^T \mathcal{H} Q_1 = \left[\begin{array}{cc|c|cc|c} t_{11} & t_{12} & w_1 & \multicolumn{2}{c|}{W_2} & w_3 \\ t_{21} & t_{22} & & & & \\ \hline \multicolumn{2}{c|}{0} & A_1 & \multicolumn{2}{c|}{w_3^T} & G_3 \\ \hline \multicolumn{2}{c|}{0} & 0 & -t_{11} & -t_{21} & 0 \\ & & & -t_{12} & -t_{22} & \\ \hline \multicolumn{2}{c|}{0} & H_1 & \multicolumn{2}{c|}{w_1^T} & -A_1^T \end{array}\right]$

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 (4.21), (4.22) anwenden. Der Trick ist der Folgende. Wir machen einen Ansatz: $\mu(t) = X(t)x(t)$, (4.22) 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

$$\dot{X}(t) = H - F^T X(t) - X(t)F - X(t)GX(t) \tag{4.43}$$

$$X(t_f) = M \tag{4.44}$$

erfüllt, so ist x die Lösung von

$$\dot{x} = [F + GX(t)]\, x \text{ und } \mu = X(t)x \tag{4.43}$$

Die Gleichung (4.43), (4.44) 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

$$H - F^T X - XF - XGX = 0 \tag{4.44}$$

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 (4.20) die lineare Rückkopplung

$$\begin{aligned} u(t) &= -R^{-1}\left(S^T x(t) + B^T \mu(t)\right) \\ &= -R^{-1}\left(S^T + B^T X\right) x(t) \end{aligned} \tag{4.45}$$

und X löst entweder die algebraische Riccatigleichung (4.43) oder die Riccati–Differentialgleichung (4.43), (4.44).

Nun ist im Gegensatz zu (4.43), (4.44) die algebraische Riccatigleichung (4.43) nicht eindeutig lösbar, aber da wir wollen, dass die Lösung $x(t)$ asymptotisch stabil ist, so muß gelten, dass die Lösung des geschlossenen Kreises

$$\dot{x} = Ax + Bu = \left[A - BR^{-1}(S^T + B^T X)\right] x \tag{4.45}$$

asymptotisch stabil ist und damit muß

$$(A - BR^{-1}S^T) - BR^{-1}B^T X \tag{4.46}$$

Eigenwerte mit negativem Realteil haben.
Nun gilt aber

$$\begin{aligned} \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \begin{bmatrix} I \\ X \end{bmatrix} &= \begin{bmatrix} I \\ X \end{bmatrix} (F + GX) \\ &= \begin{bmatrix} I \\ X \end{bmatrix} (A - BR^{-1}S^T - BR^{-1}B^T X) \end{aligned} \tag{4.47}$$

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 $S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \in \mathcal{S}_{2n}(\mathbb{R})$, $S^{-1}\mathcal{H}S = \begin{bmatrix} T_{11} & T_{12} \\ 0 & -T_{11}^T \end{bmatrix}$ 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 (4.43).

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

Satz 4.47

i) Sei X eine symmetrische Lösung der algebraischen Riccatigleichung

$$0 = H - F^T X - XF - XGX. \tag{4.48}$$

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

$$\mathcal{H} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \tag{4.49}$$

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 (4.49) aufspannen. Dann ist X eine Lösung von (4.48).

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 (4.48). □

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 4.47 eine Beziehung zwischen den Lösungen der algebraischen Riccatigleichung (4.48) und den invarianten Unterräumen von $\mathcal{H}$.

Satz 4.50 *Sei* $\mathcal{H} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \in \mathcal{H}_{2n}(\mathbb{R})$ *wie in (4.21).*
$F = A - BR^{-1}S^T, G = -BR^{-1}B^T, H = -(Q - SR^{-1}S^T)$, (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*
$0 = H - XF - F^TX - XGX$.

Beweis:

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

$$\begin{aligned}
\Longrightarrow & \begin{cases} FS_{11} + GS_{21} &= S_{11}T_{11} \\ HS_{11} - F^TS_{21} &= S_{21}T_{11} \end{cases} \Longrightarrow S_{11}^TH - S_{21}^TF = T_{11}^TS_{21}^T \\
\Longrightarrow & \begin{cases} S_{21}^TFS_{11} + S_{21}^TGS_{21} &= S_{21}^TS_{11}T_{11} \\ S_{11}^THS_{11} - S_{21}^TFS_{11} &= T_{11}^TS_{21}^TS_{11} \end{cases} \\
\Longrightarrow & S_{21}^TGS_{21} + S_{11}^THS_{11} = S_{21}^TS_{11}T_{11} + T_{11}^TS_{21}^TS_{11} \quad \text{Lyapunov–Gleichung}
\end{aligned}$$

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

T_{11} stabil, $S_{21}^TGS_{21} + S_{11}^THS_{11}$ negativ semidefinit $\Longrightarrow S_{21}^TS_{11}$ 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 $S_{11}w = 0$

$$\begin{aligned}
&\Longrightarrow \quad w^TS_{21}^TGS_{21}w + w^TS_{11}^THS_{11}w = 0 \\
&\Longrightarrow \quad GS_{21}w = 0 \text{ aus der Lyapunovgleichung} \\
&\Longrightarrow \quad B^TS_{21}w = 0, \text{ weil } R \text{ positiv definit ist.}
\end{aligned}$$

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 $\lambda \in \sigma(T_{11})$

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

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

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

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

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

Wir fassen nun alle Ergebnisse dieses Abschnitts zusammen.

Satz 4.51 *Betrachte das optimale Steuerungsproblem (4.1), (4.2).*

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

$$u(t) = -R^{-1}\left(S^T + B^T X(t)\right) x(t) \tag{4.52}$$

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

$$\begin{aligned} \dot{X}(t) &= -(Q - SR^{-1}S^T) - (A - BR^{-1}S^T)^T X(t) \\ &\quad -X(t)(A - BR^{-1}S^T) + X(t)BR^{-1}B^T X(t) \\ X(t_f) &= M \end{aligned} \tag{4.53}$$

ist.

ii) Sei (A, B) stabilisierbar, $t_f = \infty, M = 0$, $\mathcal{R}$ positiv definit.
Dann ist die eindeutige Lösung von (4.1) gegeben durch die Rückkopplungssteuerung

$$u(t) = -R^{-1}(S^T + B^T X)x(t) \tag{4.53}$$

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

$$0 = (Q - SR^{-1}S^T) + (A - BR^{-1}S^T)^T X + X(A - BR^{-1}S^T) - XBR^{-1}B^T X \tag{4.54}$$

ist. In diesem Fall gilt für den geschlossenen Kreis

$$\lim_{t\to\infty} x(t) = 0. \tag{4.55}$$

Beweis: Den Großteil des Beweises haben wir bereits vorweggenommen.

i) Aus der Theorie der gewöhnlichen Differentialgleichungen folgt, dass (4.53) eine eindeutige Lösung hat. Also erhalten wir eine Lösung der Form (4.52) und aus Satz 4.16 folgt, dass wir eine Minimallösung haben. Es bleibt noch die Frage der Eindeutigkeit der Lösung von (4.21),(4.22). Die Lösung der 2 Punkt Randwertaufgabe ist

$$\begin{bmatrix} x(t) \\ \mu(t) \end{bmatrix} = e^{\mathcal{H}(t-t_0)} \begin{bmatrix} q_0 \\ v_0 \end{bmatrix} \tag{4.56}$$

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

$$Z(t) = [Z_1(t),\ Z_2(t)] \text{ mit } Z_i(t) \in \mathbb{C}^{2n,n} \quad i = 1, 2 \tag{4.57}$$

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

$$x(t_0) = Z_1(t_0)q_0 = x^0 \tag{4.58}$$

und

$$\mu(t_f) = Mx(t_f) = Z_2(t_f)v_0 = MZ_1(t_f)q_0 \tag{4.59}$$

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 4.35, 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 (4.53) mit X Lösung von (4.54), zu einer Lösung von (4.21), (4.22) für $t_f = \infty, M = 0$ führt, falls $\lim\limits_{t\to\infty} \mu(t) = 0$ ist. Es bleibt die Frage der Eindeutigkeit.

Da $\mu(t) = Xx(t)$ für X konstant so muß auch $\lim\limits_{t\to\infty} x(t) = 0$ 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

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

□

Bemerkung 4.60 *Aus den vorigen Überlegungen folgt sofort, dass wir falls nur (4.1) 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*

$$\begin{aligned} u(t) &= -R^{-1}(S^T + B^T X)x(t) \\ &= -(B^T X)x(t) \end{aligned} \tag{4.60}$$

stabilisierendes feedback. X löst die Riccatigleichung

$$0 = I + A^T X + XA - XBB^T X. \tag{4.61}$$

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

Wir haben bisher in diesem kapitel nur zeitkontinuierliche Systeme betrachtet. Fast alle Ergebnisse lassen sich wieder direkt auf zeitdiskrete Systeme übertragen. Dabei können wir fast immer mit Hilfe der Cayley Transformation hin- und herspringen, um die Ergebnisse zu erhalten. Das zeitdiskrete System hat die Form

$$x_{k+1} = Ax_k + Bu_k,\ x_0 = x^0,\ \ k = 1, 2, 3, \ldots \tag{4.62}$$

mit Ausgang

$$y_k = Cx_k \quad k = 1, 2, 3, \ldots \tag{4.63}$$

Wenn wir Optimalsteuerung betrachten wollen, so ist das entsprechende Kostenfunktional

$$\mathcal{S}\left(\{x_k\}, \{u_k\}\right) = \frac{1}{2}\left(x_{k_f}^T M x_{k_f} + \sum_{k=k_0}^{k_f} [x_k^T,\ u_k^T] \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix}\right) \tag{4.64}$$

Bei der optimalen Steuerung erhalten wir die folgenden Sätze.

Satz 4.65 *Betrachte das Optimalsteuerungsproblem der Minimierung von (4.64) mit Nebenbedingung (4.62). 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*

$$x_{k+1} = Ax_k + Bu_k^*, \quad k = 0, 1, 2, \ldots \quad x_0 = x^0 \tag{4.66}$$

Dann gibt es eine Kozustandsfolge $\{\mu_k\}_{k=0}^{k_f}$ so daß $\{x_k^*\}, \{\mu_k\}, \{u_k^*\}$ das lineare Randwertproblem

$$\begin{bmatrix} A & 0 & B \\ Q & -I & S \\ S^T & 0 & R \end{bmatrix} \begin{bmatrix} x_k \\ \mu_k \\ u_k \end{bmatrix} = \begin{bmatrix} I & 0 & 0 \\ 0 & -A^T & 0 \\ 0 & -B^T & 0 \end{bmatrix} \begin{bmatrix} x_{k+1} \\ \mu_{k+1} \\ u_{k+1} \end{bmatrix} \tag{4.67}$$

$$x_0 = x^0, \quad \mu_{k_f} = M x_{k_f} \tag{4.68}$$

lösen.

Beweis: Vollkommen analog zum Beweis von Satz 3.40. □

Satz 4.69 *Seien $\{x_k^*\}, \{\mu_k\}, \{u_k^*\}$ die Lösungsfolgen für das lineare Randwertproblem (4.67), (4.68). Seien $\mathcal{R} = \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix}, M$ positiv semidefinit. Dann gilt*

$$\mathcal{S}\left(\{x_k\}, \{u_k\}\right) \geq \mathcal{S}\left(\{x_k^*\}, \{u_k^*\}\right) \tag{4.70}$$

für alle $\{x_k\}, \{u_k\}$ welche (4.62) lösen.

Beweis: Vollkommen analog zum Beweis für kontinuierliche Systeme. □

In (4.67) kann man wieder nach $\{u_k\}$ auflösen und erhält:

$$\begin{bmatrix} A - BR^{-1}S^T & 0 \\ -(Q - SR^{-1}S^T) & I \end{bmatrix} \begin{bmatrix} x_k \\ \mu_k \end{bmatrix} = \begin{bmatrix} I & BR^{-1}B^T \\ 0 & (A - BR^{-1}S^T)^T \end{bmatrix} \begin{bmatrix} x_{k+1} \\ \mu_{k+1} \end{bmatrix} \tag{4.71}$$

oder äquivalent:

$$\begin{bmatrix} F & 0 \\ H & I \end{bmatrix} \begin{bmatrix} x_k \\ \mu_k \end{bmatrix} = \begin{bmatrix} I & -G \\ 0 & F^T \end{bmatrix} \begin{bmatrix} x_{k+1} \\ \mu_{k+1} \end{bmatrix} \tag{4.72}$$

Beweis: Vollkommen analog zum Beweis vom analogen Satz für kontinuierliche Systeme. □

Es gilt nun, daß

$$\begin{bmatrix} I & -G \\ 0 & F^T \end{bmatrix}^{-1} \begin{bmatrix} F & 0 \\ H & I \end{bmatrix} = \begin{bmatrix} F + GF^{-T}H & GF^{-T} \\ F^{-T}H & F^{-T} \end{bmatrix} \in \mathcal{S}_{2n}(\mathbb{R}) \tag{4.73}$$

und das Büschel in (4.72)

$$\alpha\mathcal{F} - \beta\mathcal{G} \equiv \alpha \begin{bmatrix} F & 0 \\ H & I \end{bmatrix} - \beta \begin{bmatrix} I & -G \\ 0 & F^T \end{bmatrix} \tag{4.74}$$

erfüllt

$$\mathcal{G}J\mathcal{G}^T = \mathcal{F}J\mathcal{F}^T \tag{4.75}$$

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.

Kapitel 5

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 4.50), dass wir die Lösung der algebraischen Riccatigleichung

$$0 = H + F^T X + XF - XGX \tag{5.1}$$

über die Berechnung des stabilen invarianten Unterraums der Hamiltonischen Matrix

$$\mathcal{H} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \tag{5.2}$$

bestimmen können. Wir werden aber zuerst den direkten Zugang über (5.1) betrachten. Dies ist eine quadratische Matrixgleichung und man kann sie mit dem Newton–Verfahren lösen. Das Problem ist, die richtige Lösung zu erwischen.

5.1 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 5.3 *Sei X eine symmetrische Lösung von (5.1) 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

$$0 = \tilde{H} + P\tilde{F} + \tilde{F}^T P - PGP. \tag{5.4}$$

Beweis:

$$\begin{aligned} 0 &= F^T(P+X_0) + (P+X_0)F + H - (P+X_0)G(P+X_0) \\ &= (F^T - X_0G)P + P(F - GX_0) - PGP \\ &\quad + F^TX_0 + X_0F + H - X_0GX_0 \\ &= \tilde{F}^TP + P\tilde{F} - PGP + \tilde{H}. \end{aligned}$$

□

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 (5.4)

$$0 \approx \tilde{H} + P\tilde{F} + \tilde{F}^TP \tag{5.5}$$

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

Wenn wir P aus (5.5) ausgerechnet haben, setzen wir

$$X_1 := X_0 + P \tag{5.6}$$

$\Longrightarrow X_1$ 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 5.7 *Newton–Verfahren für (5.1)*
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 (5.1)
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

$$0 = H_i + X_iF_i + F_i^TX_i \tag{5.8}$$

(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 5.9

a) Die eindeutige Lösung von

$$A^T X + XA = C \tag{5.10}$$

für A stabil, ist

$$X = -\int_0^\infty e^{A^T t} C e^{At}\, dt \tag{5.11}$$

b) Falls A stabil und C positiv (semi-) definit ist, so hat (5.10) eine eindeutige negativ (semi-) definite Lösung X.

c) Falls X Lösung von (5.10) 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

$$A^T X + XA = C \Longleftrightarrow \left[(I \otimes A^T) + (A^T \otimes I) \right] \mathrm{vec}(X) = \mathrm{vec}(C) \tag{5.12}$$

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

□

Damit können wir nun die Konvergenz des Newtonverfahrens beweisen.

Satz 5.13 *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 5.7 berechnet wurden:*

a) $0 \leq X \leq X_{j+1} \leq X_j \leq \cdots \leq X_0$ wobei X die positiv semidefinite Lösung von (5.1) ist. (Hier: $A \geq B \Longleftrightarrow A - B$ positiv semidefinit)

b) $A - GX_j$ ist stabil für alle j.

c) Es gibt eine Konstante γ, so dass

$$\|X_j - X\|_2 \leq \gamma \|X_{j-1} - X\|_2^2 \tag{5.14}$$

d.h. wir haben quadratische Konvergenz.

Beweis: (5.8) ist äquivalent zu

$$X_j(F - GX_{j-1}) + (F - GX_{j-1})^T X_j = -H - X_{j-1} G X_{j-1} \leq 0 \tag{5.15}$$

Außerdem gilt

$$X_j(F - GX_j) + (F - GX_j)^T X_j = -H - X_j G X_j - (X_j - X_{j-1}) G (x_j - X_{j-1}) \tag{5.16}$$

Also zusammen

$$\begin{aligned} &(F - GX_j)^T (X_j - X_{j+1}) + (X_j - X_{j+1})(F - GX_{j-1}) \\ &\quad = -(X_{j-1} - X_j) G (X_{j-1} - X_j) \leq 0 \quad j = 1, 2, \ldots \end{aligned} \tag{5.17}$$

Mit Satz 5.9 folgt aus (5.15): $X_j \geq 0$.
$\Longrightarrow F - GX_j$ stabil in (5.16). $\Longrightarrow X_j \geq X_{j+1}$ aus (5.17).
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 $X \geq 0$ gibt.

Aus (5.16) folgt im Limes

$$(F-GX)^TX+X(F-GX)=-H-XGX \tag{5.18}$$

$\Longrightarrow X$ löst (5.1) , $X\geq 0$.
Aus (5.18) folgt

$$(F-GX_j)^TX+X(F-GX_j)=-H-XGX+(X-X_j)GX+XG(X-X_j) \tag{5.19}$$

Differenz (5.19) und (5.16) ergibt

$$\begin{aligned}
& (F-GX_j)^T(X-X_j)+(X-X_j)(F-GX_j)\\
= \;& -XGX+(X-X_j)GX+XG(X-X_j)+X_jGX_j\\
& +(X_j-X_{j-1})G(X_j-X_{j-1})\\
= \;& (X-X_j)G(X-X_j)+(X_j-X_{j-1})G(X_j-X_{j-1})
\end{aligned}$$

$$\begin{aligned}
\Longrightarrow \quad & (F-GX)^T(X-X_j)+(X-X_j)(F-GX)\\
= \;& (F-GX_j)^T(X-X_j)+(X-X_j)(F-GX_j)\\
& -(G(X-X_j))^T(X-X_j)-(X-X_j)G(X-X_j)\\
= \;& -(X-X_j)G(X-X_j)+(X_j-X_{j-1})G(X_j-X_{j-1})
\end{aligned} \tag{5.20}$$

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

$$Q_i^TF_iQ_i=T_i \tag{5.21}$$

auf Schurform (reelle Schurform), bilden

$$Q_i^TH_iQ_i=\underbrace{Q_i^TX_iQ_i}_{Y}\underbrace{Q_i^TF_iQ_i}_{T_i}+Q_i^TF_i^TQ_iY \tag{5.22}$$

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 5.3 verwenden, indem man den Löser für die Riccatigleichung immer wieder auf die Defektgleichung (5.4) anwendet und jeweils den stabilen invarianten Unterraum berechnet. Wir haben nämlich die folgenden Sätze:

Lemma 5.23 *Betrachte die Riccatigleichung (5.1) 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 (5.1). 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 5.24 *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*

$$\begin{bmatrix} \tilde{F} & G \\ \tilde{H} & -\tilde{F}^T \end{bmatrix} = \begin{bmatrix} A - BR^{-1}B^T X_0 - BR^{-1}S^T & G \\ \tilde{H} & -\tilde{F}^T \end{bmatrix}$$

$$\begin{aligned} \text{mit } \tilde{H} &= H + F^T X_0 + X_0 F - X_0 G X_0 \\ &= Q - SR^{-1}S^T + A^T X_0 - S^T R^{-1} B^T X_0 \\ &\quad + X_0 A - X_0 B R^{-1} S^T - X_0 G X_0, \end{aligned}$$

aufspannen und $P = Z_2 Z_1^{-1}$, *so ist* $X = X_0 + P$ *die eindeutige positiv semidefinite Lösung von (5.1).*

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 5.25 *Defektkorrektur für die algebraische Riccatigleichung*

Input: A, B, Q, R, S, *Toleranz tol für Residuum*

Output: Lösung der Riccati–Gleichung (5.1) 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 (5.1) 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

$$H + F^T P + PF - PGP = 0 \tag{5.26}$$

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 (5.26) wird bevorzugt das Newton–Verfahren verwendet.

5.2 Die Signum–Funktions–Methode

Sei $A \in \mathbb{R}^{n,n}(\mathbb{C}^{n,n})$ $S = P^{-1}AP$ 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

$$\text{sign}\,(A) = P\Sigma P^{-1} \tag{5.27}$$

wobei

$$\Sigma = \text{ diag}\,(\sigma_1, \cdots, \sigma_n), \sigma_i = \text{sign}\,(\text{Realteil}\, d_i) \tag{5.28}$$

Für Matrizen mit rein imaginären Eigenwerten ist sign(A) nicht definiert.

Es gilt

$$\text{sign}\,(A) = \lim_{k\to\infty} Z_k, \tag{5.29}$$

wobei

$$Z_0 = A, \quad Z_{k+1} = \frac{Z_k + Z_k^{-1}}{2} \tag{5.30}$$

Dies ist ein Newtonverfahren für $Z^2 - I = 0$. Sei nun X die stabilisierende Lösung der Riccatigleichung (5.1). Dann gilt

$$\mathcal{H} = \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} = \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} \begin{bmatrix} F+GX & G \\ 0 & -(F+GX)^T \end{bmatrix} \begin{bmatrix} I & 0 \\ X & I \end{bmatrix}^{-1} \tag{5.31}$$

Da X die stabilisierende Lösung ist, gilt $F - GX$ stabil

$$\implies \text{sign}\,(\mathcal{H}) =: \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} \begin{bmatrix} -I & Z \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ -X & I \end{bmatrix} \tag{5.32}$$

$\implies$

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} -I \\ -X \end{bmatrix} \tag{5.33}$$

$\iff$

$$\begin{bmatrix} W_{12} \\ W_{22} \end{bmatrix} X + \begin{bmatrix} W_{11} \\ W_{21} \end{bmatrix} + \begin{bmatrix} I \\ X \end{bmatrix} = 0 \tag{5.34}$$

$\iff$

$$\begin{bmatrix} W_{12} \\ W_{22} + I \end{bmatrix} X = - \begin{bmatrix} W_{11} + I \\ W_{21} \end{bmatrix} \tag{5.35}$$

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 5.36 *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 5.1 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*
$Z_k := |\det Z_k|^{-\frac{1}{2n}} Z_k$
$Z_{k+1} := Z_k - \frac{1}{2}\left(Z_k - (JZ_k)^{-1}J\right)$
END FOR

(3) Mit $Z_\infty = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}, \quad W_{ij} \in \mathbb{R}^{n,n}$ *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 5.25 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.

5.3 Laub's 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.

$$\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}^T \begin{bmatrix} F & G \\ H & -F^T \end{bmatrix} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \tag{5.37}$$

und die Eigenwerte von T_{11} haben alle negativen Realteil, so ist

$$X = Q_{21} Q_{11}^{-1} \tag{5.38}$$

die gesuchte Lösung der algebraischen Riccatigleichung (5.1).
Wir können darauf also den folgenden Algorithmus aufbauen.

Algorithmus 5.39 *Laub's Schurmethode*

Input: $\mathcal{H}$ *wie in (5.2),* $F, G, H \in \mathbb{R}^{n,n}, G = G^T, H = H^T, \mathcal{H}$ *habe keine Eigenwerte mit Realteil* 0.

Output: Stabilisierende positiv semidefinite Lösung X *der algebraischen Riccatigleichung 5.1.*

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 5.25 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 $Q^T HQ$ 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.

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 6

Singuläre Steuerungsprobleme

Für singuläre Steuerungsprobleme

$$E\dot{x} = Ax + Bu, y = Cx \quad x(t_0) = x^0. \tag{6.1}$$

brauchen wir eine Abschwächung der Begriffe vollständig steuerbar, vollständig beobachtbar.

Definition 6.2

i) Das System (6.1) heisst vollständig steuerbar, *falls*

$$Rang\ [\alpha E - \beta A, B] = n \quad \forall(\alpha, \beta) \in \mathbb{C}^2\{(0,0)\} \tag{6.3}$$

ii) Das System (6.1) heisst vollständig rekonstruierbar (beobachtbar) *falls*

$$Rang \begin{bmatrix} \alpha E - \beta A \\ C \end{bmatrix} = n \quad \forall(\alpha, \beta) \in \mathbb{C}\{(0,0)\} \tag{6.4}$$

iii) Das System (6.1) heisst stark steuerbar*, falls*

$$Rang\ [\lambda E - A, B] = n \quad \forall\lambda \in \mathbb{C} \tag{6.5}$$

und

$$Rang\ [E, AS_\infty, B] = n \tag{6.6}$$

wobei S_∞ eine Matrix ist, deren Spalten Kern (S) aufspannen. Bedingung (6.6) heisst Steuerbarkeit bei ∞ *oder* Impulssteuerbarkeit.

iv) Das System (6.1) heisst stark rekonstruierbar *(beobachtbar) falls*

$$Rang \begin{bmatrix} \lambda E - A \\ C \end{bmatrix} = n \quad \forall\lambda \in \mathbb{C} \tag{6.7}$$

und

$$Rang \begin{bmatrix} E \\ T_\infty^T A \\ C \end{bmatrix} = n, \tag{6.8}$$

wobei T_∞ eine Matrix ist, deren Spalten Kern (E^*) aufspannen. (6.8) heisst Rekonstruierbarkeit (Beobachtbarkeit) bei ∞ *oder Impuls–Beobachtbarkeit.*

v) Das System (6.1) heisst Zustands–regulariserbar*, falls es eine lineare Zustands–Rückkopplungssteuerung* $u(t) = Fx(t)$ *gibt, so dass* $\alpha E - \beta(A + BF)$ regulär *ist.*

vi) Das System (6.1) heisst Ausgangs–regulisierbar*, falls es eine lineare Ausgangsrückkopplungssteuerung* $u(t) = Fy(t)$ *gibt, so dass* $\alpha E - \beta(A + BFC)$ regulär *ist*

vii) Das System (6.1) heisst erreichbar *falls Rang* $(E, B) = n$.

Zuerst einige Beispiele:

Beispiel 6.9 *(a)* $E = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

$Rang\ [\alpha E - \beta A, B] = Rang \begin{bmatrix} 0 & \alpha & \beta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = 2$ *für* $(\alpha, \beta) \neq (0, 0)$

$\Longrightarrow$ *System ist* vollständig steuerbar*, obwohl* $\alpha E - \beta A$ *nicht regulär.*

$S_\infty = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Longrightarrow Rang\ [E, AS_\infty, B] = Rang \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = 2.$

$\Longrightarrow$ *System ist* steuerbar bei ∞

$\Longrightarrow$ *System ist* stark steuerbar

$Rang\ [E, B] = Rang \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 2$

$\Longrightarrow$ ist erreichbar.

b) $E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

$Rang\ [\alpha E - \beta A, B] = Rang \begin{bmatrix} \alpha & -\beta & 0 & 0 \\ 0 & 0 & -\beta & 0 \\ 0 & 0 & \alpha & 1 \end{bmatrix} < 3.$

für $\beta = 0$.

$\Longrightarrow$ *nicht* vollständig steuerbar.

$Rang\ [\lambda E - A, B] = Rang \begin{bmatrix} \lambda & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & \lambda & 1 \end{bmatrix} = 3$

$S_\infty = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \Longrightarrow Rang\ [E, AS_\infty, B] = Rang \left[\begin{array}{cc|c|c} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{array}\right] < 3$

$\Longrightarrow$ nicht steuerbar bei ∞

$\Longrightarrow$ nicht stark steuerbar.

$Rang\ [E, B] = Rang \left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{array}\right] = 2$

$\Longrightarrow$ nicht erreichbar

$\alpha E - \beta A$ singulär.

c) $E = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

$Rang\ [\alpha E - \beta A, B] = Rang \begin{bmatrix} 0 & -\beta & 0 \\ \alpha - \beta & 0 & 1 \end{bmatrix} = 1$ *für* $\beta = 0$

$\Longrightarrow$ *nicht* vollständig steuerbar

$$Rang\ [\lambda E - A, B] = Rang \begin{bmatrix} 0 & 1 & 0 \\ \lambda - 1 & 0 & 1 \end{bmatrix} = 2\ \forall \lambda$$

$$Rang\ [E, AS_\infty, B] = Rang \left[\begin{array}{c|c|c} 0 & 1 & 0 \\ 1 & 0 & 1 \end{array}\right] = 2$$

$\Longrightarrow$ stark steuerbar.

$Rang\ [E, B] = 1$

$\Longrightarrow$ nicht erreichbar.

$\alpha E - \beta A$ regulär.

Wir wollen nun auch wieder numerisch berechenbare kondensierte Formen entwickeln, aus denen die Systemeigenschaften abgelesen werden können.

Dazu müssen wir anschauen, welche Transformationen wir an dem System durchführen können ohne die Lösungseigenschaften zu verändern.

Es gibt zwei Sorten von Transformationen für

$$E\dot{x} = Ax + Bu, y = Cx, E, A \in \mathbb{C}^{n,l}, B \in \mathbb{C}^{n,m}, C \in \mathbb{C}^{p,l} \tag{6.10}$$

a) Zustandsraumtransformationen

$$\begin{cases} \tilde{E} = PEQ, & \tilde{A} = PAQ, & \tilde{B} = PBR \\ & \tilde{C} = SCQ \end{cases} \tag{6.11}$$

wobei $P \in \mathbb{C}^{n,n}, Q \in \mathbb{C}^{l,l}, R \in \mathbb{C}^{m,m}, S \in \mathbb{C}^{p,p}$ nichtsingulär sind.
Die Quadrupel $(E, A, B, C), (\tilde{E}, \tilde{A}, \tilde{B}, \tilde{C})$ sind äquivalent.

b) Rückkopplungen

$$\begin{cases} i) & \tilde{E} = E + BF, \tilde{A} = A, \tilde{B} = b, \tilde{C} = C & \text{Zustandsableitungs–feedback} \\ ii) & \tilde{E} = E, \tilde{A} = A + BF, \tilde{B} = B, \tilde{C} = C & \text{Zustandsfeedback} \\ iii) & \tilde{E} = E + BFC, \tilde{A} = A, \tilde{B} = B, \tilde{C} = C & \text{Ausgangsableitungs–feedback} \\ iv) & \tilde{E} = E, \tilde{A} = A + BFC, \tilde{B} = B, \tilde{C} = C & \text{Ausgangsfeedback} \end{cases} \tag{6.12}$$

Man kann sich nun anschauen, was die Normalformen unter (6.10), (6.11) oder beiden zusammen sind und damit sämtliche Systemeigenschaften charakterisieren. Wir fordern statt dessen, dass P, Q, R, S unitär sind, so dass es numerisch stabile Verfahren gibt diese zu berechnen. Wir haben das folgende Lemma

Lemma 6.13 *Es gibt eine Zustandsraumtransformation der Form (6.11), so dass*

$$PEQ = \begin{bmatrix} E_{11} & 0 & E_{13} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{matrix} r \\ s \\ n-r-s=q \end{matrix}, PAQ = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix}, PB = \begin{bmatrix} B_1 \\ B_2 \\ 0 \end{bmatrix}$$

$$CQ = [C_1\ C_2\ C_3]$$

mit $r = Rang\ (E)$, $s = Rang\ (B_2)$.

Beweis: Per Konstruktion. Mache eine Zeilenkompression

$$P[E,B,A] = \begin{bmatrix} \tilde{E}_{11} & \tilde{E}_{12} & \tilde{E}_{13} & \tilde{B}_1 & \tilde{A}_{11} & \tilde{A}_{12} & \tilde{A}_{13} \\ 0 & 0 & 0 & \tilde{B}_2 & \tilde{A}_{21} & \tilde{A}_{22} & \tilde{A}_{23} \\ 0 & 0 & 0 & 0 & \tilde{A}_{31} & \tilde{A}_{32} & \tilde{A}_{33} \end{bmatrix} \begin{matrix} r \\ s \\ q = n-r-s \end{matrix}$$

mit $r = \text{Rang } (E)$ $s = \text{Rang } (B_2)$. Falls $\text{span}(B) \subset \text{span}(E)$ so ist $B = 0$. Die Zeilenkompression kann mit QR Zerlegung oder Singulärwertzerlegung gemacht werden. Mache dann eine Spaltenkompression, angewendet auf die Untermatrix $\begin{bmatrix} \tilde{A}_{33} & \tilde{A}_{31} & \tilde{A}_{32} \\ \tilde{E}_{13} & \tilde{E}_{11} & \tilde{E}_{12} \end{bmatrix}$ so dass

$$\begin{bmatrix} \tilde{A}_{33} & \tilde{A}_{31} & \tilde{A}_{32} \\ \tilde{E}_{13} & \tilde{E}_{11} & \tilde{E}_{12} \end{bmatrix} \tilde{Q} = \begin{bmatrix} A_{33} & 0 & 0 \\ E_{13} & E_{11} & 0 \end{bmatrix} \begin{matrix} q \\ r \end{matrix}$$

und setze $Q = \begin{bmatrix} 0 & 0 & I \\ I & 0 & 0 \\ 0 & I & 0 \end{bmatrix} \tilde{Q} \begin{bmatrix} 0 & I & 0 \\ 0 & 0 & I \\ I & 0 & 0 \end{bmatrix}$ so folgt mit diesem P, Q die Behauptung. □

In Lemma 6.13 ist nichts über den Rang von E_{11}, A_{33} ausgesagt. Falls Rang (E_{11}) nicht voll ist , so wiederholen wir die Konstruktion aus dem Beweis von Lemma 6.13 so lange, bis wir E_{11} nichtsingulär erhalten.

Satz 6.14 *Betrachte ein quadratisches Problem der Form (6.10) mit $l = n$. Es gibt eine Zustandsraumtransformation der Form (6.11), so dass*

$$\left\{ \begin{array}{l} PEQ = \begin{bmatrix} E_{11} & 0 & E_{13} \\ 0 & 0 & E_{23} \\ 0 & 0 & E_{33} \end{bmatrix} \begin{matrix} t_1 \\ n-t_1-t_3 = t_2 \\ t_3 \end{matrix}, PAQ = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix} \begin{matrix} t_1 \\ t_2 \\ t_3 \end{matrix} \\ PB = \begin{bmatrix} B_1 \\ B_2 \\ 0 \end{bmatrix} \begin{matrix} t_1 \\ t_2 \\ t_3 \end{matrix}, \quad CQ = \begin{bmatrix} C_1 & C_2 & C_3 \end{bmatrix} \end{array} \right. \tag{6.15}$$

und

i) Rang $(E_{11}) = t_1$

ii) Rang $(B_2) = t_2$

iii) A_{33} ist Block–obere–Dreiecksmatrix mit quadratischen Diagonalblöcken

iv) E_{33} ist Block–obere–Dreiecksmatrix mit 0 Diagonalblöcken und der gleichen Einteilung wie A_{33}.

Beweis: Induktive Anwendung von Lemma 6.13.
Für den Anfangsschritt wende Lemma 6.13 und erhalte

$$P^{(1)}EQ^{(1)} = \begin{bmatrix} E_{11}^{(1)} & 0 & E_{13}^{(1)} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{matrix} r \\ s \\ q \end{matrix}, \quad P^{(3)}B = \begin{bmatrix} B_1^{(1)} \\ B_2^{(1)} \\ 0 \end{bmatrix}$$

$$P^{(1)}AQ^{(1)} = \begin{bmatrix} A_{11}^{(1)} & A_{12}^{(1)} & A_{13}^{(1)} \\ A_{21}^{(1)} & A_{22}^{(1)} & A_{23}^{(1)} \\ 0 & 0 & A_{23}^{(1)} \end{bmatrix} \quad CQ^{(1)} = \begin{bmatrix} C_1^{(1)} & C_2^{(1)} & C_3^{(1)} \end{bmatrix}$$

Für den Induktionsschritt sei angenommen, dass wir $P^{(k)}, Q^{(k)}$ gefunden haben, so dass das System die Form (6.15) hat mit der Ausnahme, dass der Rang von (E_{11}) nicht voll ist.

Wende Lemma 6.13 auf

$$\tilde{E} = \begin{bmatrix} E_{11}^{(1)} & 0 \\ 0 & 0 \end{bmatrix}, \quad \tilde{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \tilde{B} = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$$
$$\tilde{C} = [C_1 \; C_2] \quad \text{an}$$

und erhalte $\tilde{P}, \tilde{Q}$. Setze

$$P^{(k+1)} = \begin{bmatrix} \tilde{P} & 0 \\ 0 & I \end{bmatrix}^{(k)}, Q^{(k+1)} = Q^{(k)} \begin{bmatrix} \tilde{Q} & 0 \\ 0 & I \end{bmatrix}.$$

Entweder ist der (1.1) Block von $P^{(k+1)}EQ^{(k+1)}$ nicht singulär, so haben wir die gewünschte Form oder wir können den Prozess wiederholen.

In jedem Schritt wird die Dimension des (1.1) Blocks um mindestens 1 kleiner. Also ist man nach maximal n Schritten fertig. □

Die Hauptkonsequenz aus Satz 6.14 ist, dass durch diese Transformation die Teile, die nicht steuerbar bei ∞ sind von den anderen getrennt werden.

Korollar 6.16 *Sei das System (6.10) transformiert auf die Form (6.15), so ist das Subsystem*

$$\begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} n \tag{6.17}$$

steuerbar bei ∞.

Beweis: Da $S_\infty = \begin{bmatrix} 0 \\ I \end{bmatrix}$, so ist Rang $[E, AS_\infty, B]$ = Rang $\begin{bmatrix} E_{11} & A_{12} & B_1 \\ 0 & A_{22} & B_2 \end{bmatrix}$ voll, da Rang (E_{11}) und Rang (B_2) voll. □

In vielen Fällen müssen wir die Ausgangsvariablen zur Steuerung verwenden, also kommt C auch noch ins Spiel:

Satz 6.18 *Betrachte ein quadratisches Problem (6.10) ($l = n$). Es gibt eine Zustandsraumstransformation mit unitären Matrizen $\tilde{P}, \tilde{Q}$ welche die Systemmatrizen wie folgt transformieren:*

$$\begin{cases} \tilde{P}E\tilde{Q} = \begin{bmatrix} \tilde{E}_{11} & 0 & 0 & \tilde{E}_{14} \\ 0 & 0 & 0 & \tilde{E}_{24} \\ \tilde{E}_{31} & \tilde{E}_{32} & \tilde{E}_{33} & \tilde{E}_{34} \\ 0 & 0 & 0 & \tilde{E}_{44} \end{bmatrix} \begin{matrix} \tilde{t}_1 \\ \tilde{t}_2 \\ \tilde{t}_3 \\ \tilde{t}_4 \end{matrix} \\ \tilde{P}A\tilde{Q} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} & 0 & \tilde{A}_{14} \\ \tilde{A}_{21} & \tilde{A}_{22} & 0 & \tilde{A}_{24} \\ \tilde{A}_{31} & \tilde{A}_{32} & \tilde{A}_{33} & \tilde{A}_{34} \\ 0 & 0 & 0 & \tilde{A}_{44} \end{bmatrix} \begin{matrix} \tilde{t}_1 \\ \tilde{t}_2 \\ \tilde{t}_3 \\ \tilde{t}_4 \end{matrix} \\ \tilde{P}B = \begin{bmatrix} \tilde{B}_1 \\ \tilde{B}_2 \\ \tilde{B}_3 \\ 0 \end{bmatrix} \begin{matrix} \tilde{t}_1 \\ \tilde{t}_2 \\ \tilde{t}_3 \\ \tilde{t}_4 \end{matrix}, C\tilde{Q} = \begin{bmatrix} \tilde{C}_1 & \tilde{C}_2 & 0 & \tilde{C}_4 \end{bmatrix} \end{cases} \tag{6.19}$$

Dabei gilt

(1) Rang $(\tilde{E}_{11}) = \tilde{t}_1$

(2) Rang $(\tilde{C}_2) = \tilde{T}_2$

(3) $\tilde{A}_{33}, \tilde{A}_{44}^\top$ ist Block untere $\triangle$–Matrix

(4) $\tilde{E}_{33}, \tilde{E}_{44}^\top$ ist Block untere $\triangle$–Matrix mit 0 Diagonalblöcken der gleichen Dimension wie der Diagonalblöcke $\tilde{A}_{33}(\tilde{A}_{44}^\top)$.

(5) das Untersystem bestehend aus den ersten beiden Blockzeilen und Spalten ist steuerbar und beobachtbar bei ∞.

(6) Das Untersystem bestehend aus den ersten 3 Blockzeilen und Spalten ist steuerbar bei ∞.

Beweis: Bestimme zuerst P_1, Q_1, so dass die Form (6.15) entsteht und wende dann den Satz (6.14) auf

$$\hat{E} = \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix}, \hat{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^H, \hat{B} = \begin{bmatrix} C_1 & C_2 \end{bmatrix}^H$$

an, d.h. bestimme $\hat{P}, \hat{Q}$ welche das Subsystem transformieren. Setze $P_2 = \begin{bmatrix} \hat{Q} & 0 \\ 0 & I_{t_3} \end{bmatrix}, Q_2 = \begin{bmatrix} \hat{P}^H & 0 \\ 0 & I_{t_3} \end{bmatrix}$
$\tilde{P} = P_2 P_1, \tilde{Q} = Q_1 Q_2$. Dann erhalten wir die Form (6.19). Eigenschaften 1.–4. folgen sofort aus Satz 6.14 Teile 5), 6) folgen durch sukzessives Anwenden auf am Satz 6.14. □

Diese beiden Sätze sagen einem sofort ob es überhaupt möglich ist durch Rückkopplung das Büschel zu regularisieren.

Satz 6.20 *Sei System (6.10) in der Form von Satz 6.15 gegeben, so existiert ein $F \in \mathbb{C}^{m,n}$, so dass $\alpha E - \beta(A + BF)$ regulär ist genau dann wenn A_{33} nichtsingulär ist.*

Beweis: Sei $F \in \mathbb{C}^{m,n}$ partitioniert als $F = \begin{bmatrix} F_1 & F_2 & F_3 \end{bmatrix}$, dann ist

$$\alpha E - \beta(A + BF) = \tag{6.21}$$
$$\alpha \begin{bmatrix} E_{11} & 0 & E_{13} \\ 0 & 0 & E_{23} \\ 0 & 0 & E_{33} \end{bmatrix} - \beta \left(\begin{bmatrix} A_{21} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \\ 0 \end{bmatrix} [F_1 F_2 F_3] \right)$$

$\Longrightarrow \det[\alpha E - \beta(A + BF)] =$

$$\underbrace{\det\left[\alpha E_{33} - \beta A_{33}\right]}_{=-\beta \det A_{33}} \det \alpha \begin{bmatrix} E_{21} & 0 \\ 0 & 0 \end{bmatrix} - \beta \left(\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} [F_1 F_2] \right).$$

Falls A_{33} singulär, so ist wegen der speziellen Form
$\left(E_{33} = \begin{bmatrix} 0 & * & & x \\ & \ddots & \ddots & \\ & & \ddots & x \\ & & & 0 \end{bmatrix} A_{33} = \begin{bmatrix} * & x & \cdots & x \\ & \ddots & & x \\ & & \ddots & \\ & & & x \end{bmatrix} \right)$ and $\det(\alpha E_{33} - \beta A_{33}) = 0$ unabhängig von α, β also ist das Büschel singulär, unabhängig von F. Angenommen A_{33} ist nichtsingulär. Da B_2 vollen Rang hat so gibt es F_2 so dass $A_{22} + B_2F_2$ nichtsingulär. Sei $F \begin{bmatrix} 0 & F_2 & 0 \end{bmatrix}$ so folgt $\alpha E - \beta(A + BF) = \alpha \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix} - \beta \begin{bmatrix} A_{11} & A_{12} + B_1F_2 \\ A_{21} & A_{22} + B_2F_2 \end{bmatrix}$. Die Regularität hierfür ist äquivalent dazu dass $\alpha \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix} - \beta \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & A_{22} + B_2F_2 \end{bmatrix}$ regulär ist denn die Determinante ist $\det(\alpha E_{11} - \beta \tilde{A}_{11}) \cdot \det(-\beta(A_{22} + B_2F_2)) \neq 0$ für $\beta \neq 0, (\alpha, \beta) \notin \sigma(E_{11}\tilde{A}_{11})$, da $E_{11}, A_{22} + B_2F_2$ nichtsingulär. □

Wir haben schon gesehen dass $\mathrm{ind}(E, A) \leq 1$ sein sollte, weil wir sonst Ableitungen von $u(t)$ in der Lösung erhalten. Was ist nun der minimale Index den wir erreichen mit feedback können?

Satz 6.22 *Sei System (6.10) in der Form von Satz 6.15 gegeben und A_{33} nichtsingulär. Dann existiert $F \in \mathbb{C}^{m,n}$, so dass $\alpha E - \beta(A + BF)$ regulär und*

$$ind_\infty(E, A + BF) = ind_\infty\left(\begin{bmatrix} 0 & E_{23} \\ 0 & E_{33} \end{bmatrix} \right) \tag{6.23}$$

Beweis: Wähle $F_1 \in \mathbb{C}^{m,t_1}$, so dass $A_{21} + B_2F_1 = 0$ und wähle $F_2 \in \mathbb{C}^{m,t_2}$ so dass $A_{22} + B_2F_2$ nichtsingulär. Beides ist möglich da B_2 vollen Spaltenrang hat. Setze

$F = \begin{bmatrix} F_1 & F_2 & 0 \end{bmatrix}$. Dann ist $\alpha E - \beta(A + BF)$

block oberes $\triangle$–Büschel mit Diagonalblöcken

$$\begin{cases} \alpha E_{11} - \beta(A_{11} + B_1F_1), \\ \alpha \begin{bmatrix} 0 & E_{23} \\ 0 & E_{33} \end{bmatrix} - \beta \begin{bmatrix} A_{22} + B_2F_2 & A_{33} \\ 0 & A_{33} \end{bmatrix} \end{cases} \tag{6.24}$$

Der erste Block ist regulär und hat Index 0, da E_{11} nicht singulär ist. Beim zweiten Block ist die rechte Seite invertierbar, also ist der Index der von $\begin{bmatrix} 0 & E_{23} \\ 0 & E_{33} \end{bmatrix}$. □

Da wir aber im allgemeinen wollen, dass das System index 1 ist, so muss gelten $\begin{bmatrix} E_{23} \\ E_{33} \end{bmatrix} = 0$ und A_{33} invertierbar. Was passiert aber wenn dies nicht gilt, d.h. $\begin{bmatrix} E_{23} \\ E_{33} \end{bmatrix} \neq 0$ oder A_{33} nicht invertierbar? Falls A_{33} nicht invertierbar so ist das System nicht regularisierbar, also auf jeden Fall schlecht. (Wahrscheinlich ein Modellfehler.)

Satz 6.25 *Sei das System (6.10) in der Form von Satz 6.11 und regularisierbar, d. h.*

$$\begin{bmatrix} E_{11} & 0 & E_{13} \\ 0 & 0 & E_{23} \\ 0 & 0 & E_{33} \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \\ 0 \end{bmatrix} n \tag{6.26}$$

und A_{33} nichtsingulär, so folgt $x_3 \equiv 0$.

Beweis: Die dritte Gleichung lautet

$$E_{33}\dot{x} = A_{33}x_3$$

und damit folgt

$$A_{33}^{-1}E_{33}\dot{x}_3 = x_3.$$

Aber $A_{33}^{-1}E_{33}$ ist block–obere–$\triangle$ Matrix mit 0 Blöcken auf der Diagonale also folgt $\begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix}$

$\begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} \Longrightarrow x_3 \equiv 0.$ □

Es folgt, dass der Teil des Systems der nicht bei ∞ steuerbar ist 0 ist, also unproblematisch denn er ist stabil. Nach Satz 6.14 ist der Rest des Systems steuerbar bei ∞. Man kann also diesen Teil des Systems einfach weglassen.

Beispiel 6.27 *Betrachte ein klassisches Rollringgetriebe.*

$$\begin{bmatrix} I_3 & 0 & 0 \\ 0 & M & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & I_3 & 0 \\ Q & -P & G^\top \\ H & G & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ S \\ 0 \end{bmatrix} u \tag{6.28}$$

wobei $M = \begin{bmatrix} I_R & & \\ & m_g & \\ & & m_z \end{bmatrix}, P = \begin{bmatrix} 0 & 0 & 0 \\ 0 & +d_1 & -d_1 \\ 0 & -d_1 & d_1 \end{bmatrix}, Q = \begin{bmatrix} 0 & 0 & 0 \\ 0 & c_1 & -c_1 \\ 0 & -c_1 & c_1 \end{bmatrix}$

$$S = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \begin{matrix} I_R & - & \textit{Trägheitsmomente} \\ m_g & - & \textit{Getriebemasse} \end{matrix}$$

$$G = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}, H = \begin{bmatrix} -v_n & 0 & 0 \end{bmatrix} Z = \begin{bmatrix} \varphi, & z_G, & z_t^\top \end{bmatrix}$$

$$x = \begin{bmatrix} z_0 \\ z \\ \mu \end{bmatrix} \quad \begin{array}{lll} v_n & - & \textit{Wellenumfangsgeschwindigkeit} \\ \varphi & - & \textit{Schwenkwinkel} \end{array}$$

m_t Zusatzmasse, z_z Auslenkung der Zusatzmasse, z_G Auslenkung des Getriebes. M_3 ist invertierbar.

Das System ist nicht steuerbar bei ∞, da $S_\infty = e_7$ und $\begin{bmatrix} E, & AS_\infty, & B \end{bmatrix} = \begin{bmatrix} I_3 & 0 & 0 \\ M & G^\top & S \\ 0 & 0 & 0 \end{bmatrix}$.

Wir kommen zur Ausgangs–Rückkopplung.

Satz 6.29 *Sei System (6.10) in der Form (6.15) gegeben, A_{33} sei nichtsingulär und* $\begin{bmatrix} A_{22} \\ C_2 \end{bmatrix}$ *habe vollen Spaltenrang. Dann ist das System mittels Ausgangs–Rückkopplung regularisierbar, d. h. F so dass $\alpha E - (A + BFC)$ regulär und*

$$ind\ (\alpha E - (A + BFC)) = \ ind \begin{bmatrix} 0 & E_{23} \\ 0 & E_{33} \end{bmatrix} \tag{6.30}$$

Beweis: Da B_2 vollen Zeilenrang und $\begin{bmatrix} A_{22} \\ C_2 \end{bmatrix}$ vollen Spaltenrang haben, so gibt es F, so dass $A_{22} + B_2FC_2$ nichtsingulär. Dann können wir die gleichen Argumente wie im Beweis von Satz 6.20 und Satz 6.22 verwenden. □

Wir haben auch ein Analogon zu Satz 6.25 für den Ausgangsfall.

Satz 6.31 *Sei (6.10) in der Form (6.19) aus Satz 6.18. Falls das System regularisierbar ist so sind A_{33} und A_{44} nichtsingulär und $x_4 \equiv 0$.*

Beweis:

$$\begin{aligned} \det\left[\alpha E - \beta(A + B) \neq C)\right] \ &= \ \det\left\{\alpha \begin{bmatrix} \tilde{E}_{21} & 0 \\ 0 & 0 \end{bmatrix} - \beta\left(\begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix}\right.\right. \\ &\left.\left. + \begin{bmatrix} \tilde{B}_1 \\ \tilde{B}_2 \end{bmatrix} F \begin{bmatrix} \tilde{C}_1 & \tilde{C}_2 \end{bmatrix}\right)\right\} \det(\alpha E_{33} - \beta A_{33}) \det(\alpha E_{44} - \beta A_{44}) \end{aligned}$$

Wegen der Struktur von $\alpha E - \beta A_{33}, \alpha E_{44} - \beta A_{44}$ ist die Determinante identisch 0 falls A_{33} oder A_{44} singulär sind . Der Rest folgt genau wie vorher. Man erhält

$$E_{44}\dot{x}_4 = A_{44}x_4 \tag{6.32}$$

und aus der Struktur folgt $x_4 \equiv 0$. □

Die Komponenten die zum 3. Block gehören sind gefährlich, da es falls der Index nicht 0 ist d. h. $\tilde{t}_3 > 0$ und $\tilde{E}_{33} \neq 0$, dann Ableitungen von u in dieser Komponente gibt. Diese werden aber nicht beobachtet.

Wir nehmen daher in Zukunft an, dass das System steuerbar und beobachtbar bei ∞ ist, d. h. die 0 Komponenten sind schon eliminiert. (Falls das nicht gilt, so muss man diese Komponenten mittels der Reduktion auf die Form (6.15) oder (6.19) entfernen. Dies ist aber sehr gefährlich unter dem Einfluss von Rundungsfehlern (Störungen wie Rauschen) können diese Komponenten trotzdem noch Ärger bereiten.) Das dürfte aber ein Modellierungsfehler sein, daher nehmen wir an, dass das System steuerbar und beobachtbar bei ∞ ist.

Wir betrachten nun weitere kondensierte Formen.

Satz 6.33 *Seien $E, A \in \mathbb{C}^{n,n}, B \in \mathbb{C}^{m,n}, C \in \mathbb{C}^{p,n}$ so gibt es unitäre Matrizen $U, V \in \mathbb{C}^{n,n}, w \in \mathbb{C}^{m,m}, Y \in \mathbb{C}^{p,p}$ so dass*

$$\left\{\begin{array}{lcl} U^*EV &=& \begin{bmatrix} \sum_E & 0 \\ 0 & 0 \end{bmatrix} \begin{matrix} t_1 \\ t-t_1 \end{matrix}, \quad U^*BW = \begin{bmatrix} B_{11} & B_{12} & 0 \\ \hat{B}_{21} & 0 & 0 \end{bmatrix} \begin{matrix} t_1 \\ n-t_1 \end{matrix} \\ Y^*CV &=& \begin{bmatrix} C_{11} & \hat{C}_{12} \\ C_{21} & 0 \\ 0 & 0 \end{bmatrix} \begin{matrix} l_1 \\ l_2 \\ l_3 \end{matrix}, \quad U^*AV = \begin{bmatrix} A_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{bmatrix} \begin{matrix} t_1 \\ n-t_1 \end{matrix} \end{array}\right. \tag{6.34}$$

wobei

$$\left\{\begin{array}{lcl} \hat{A}_{22} &=& \begin{bmatrix} A_{22} & A_{23} & A_{24} & 0 & 0 \\ A_{32} & A_{33} & A_{34} & \sum_{35} & 0 \\ A_{42} & A_{43} & \sum_{44} & 0 & 0 \\ 0 & \sum_{53} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{matrix} t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \end{matrix}, \hat{B}_{21} = \begin{bmatrix} B_{21} \\ B_{31} \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{matrix} t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \end{matrix} \\ \hat{A}_{21} &=& \begin{bmatrix} A_{21} \\ A_{31} \\ A_{41} \\ A_{51} \\ A_{61} \end{bmatrix} \begin{matrix} t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \end{matrix}, \hat{A}_{12} = \begin{bmatrix} A_{12} & A_{13} & A_{14} & A_{15} & A_{16} \end{bmatrix} t_1 \\ \hat{C}_{12} &=& \begin{bmatrix} C_{12} & C_{13} & 0 & 0 & 0 \end{bmatrix} l_1, l_1 = s_2 t t_s \end{array}\right. \tag{6.35}$$

$\sum_E, \sum_{35}, \sum_{44}, \sum_{53}$ sind nichtsinguläre Diagonalmatrizen, B_{12} hat vollen Spaltenrang, C_{21} hat vollen Zeilenrang und $\begin{bmatrix} B_{21} \\ B_{31} \end{bmatrix}, [C_{12}\ C_{13}]$ sind nichtsingulär.

Vor dem Beweis einige Folgerungen:

Korollar 6.36 *Seien E, A in der Form (6.34) (6.35). Dann sind die folgenden Aussagen äquivalent.*

(i) $\alpha E - \beta A$ *regulär und* $ind(E, A) = 1$

(ii) $\hat{A}_{22}$ *nichtsingulär*

(iii) $s_6 = t_6 = 0$ *und* A_{22} *ist nichtsingulär*

(iv) *Rang* $[E, AS_\infty] = n$

(v) *Rang* $\begin{bmatrix} E \\ T_\infty^* A \end{bmatrix} = n.$

Beweis: Die Äquivalenz von ii), iii), iv) und v) folgt aus der Form von $\hat{A}_{22}$ und aus

$$[E, AS_\infty] = \begin{bmatrix} \sum_E & 0 & \hat{A}_{12} \\ 0 & 0 & \hat{A}_{22} \end{bmatrix}, \begin{bmatrix} E \\ T_\infty^* A \end{bmatrix} = \begin{bmatrix} \sum_E & 0 \\ 0 & 0 \\ \hat{A}_{21} & \hat{A}_{22} \end{bmatrix}. \tag{6.37}$$

ii) $\Longrightarrow$ i) Aus ii) folgt das $\alpha E - \beta A$ äquivalent zu
$\alpha \begin{bmatrix} \sum_E & 0 \\ 0 & 0 \end{bmatrix} - \beta \begin{bmatrix} A_{11} - \hat{A}_{12}\hat{A}_{22}^{-1}\hat{A}_{21} & 0 \\ 0 & I \end{bmatrix}$ ist, welches natürlich index 1 hat. i) $\Longrightarrow$ iv) Aus i) folgt, dass $\alpha E - \beta A$ äquivalent zu $\alpha \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} - \beta \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$ ist. $\Longrightarrow$ Rang $[E, AS_\infty] = n$. □

Korollar 6.38 *Seien* (E, A, B, C) *in der Form (6.34). (6.35)*

(i) *Das System ist steuerbar bei* ∞ *genau dann wenn* $t_6 = 0$.

(ii) *Das System ist beobachtbar bei* ∞ *ganu dann wenn* $s_6 = 0$.

Beweis: Folgt sofort aus der kondensierten Form. □

Korollar 6.39 *Seien* (E, A, B, C) *in der Form (6.34), (6.35).*

i) *Rang* $[E, B] = n$ *genau dann wenn* $t_4 = t_5 = t_2 = 0$

ii) *Rang* $\begin{bmatrix} E \\ C \end{bmatrix} = n$ *genau dann wenn* $t_4 = t_3 = s_6 = 0$.

Beweis: Folgt sofort aus kondensierter Form. □

Beachte dass Rang $[\alpha E - \beta A, B] = n \ \forall (\alpha, \beta) \neq (0, 0)$

$\Longrightarrow$ Rang $[E, B] = n$ und
Rang $\begin{bmatrix} \alpha E - \beta A \\ C \end{bmatrix} = n \ \forall (\alpha, \beta) \neq (0, 0)$
$\Longrightarrow$ Rang $\begin{bmatrix} E \\ C \end{bmatrix} = n.$

Korollar 6.39 liefert also notwendige Bedingungen für vollständige Steuerbarkeit und Erreichbarkeit. Wir können im folgenden annehmen, dass in der kondensierten Form $k_3 = l_3 = 0$, denn diese Teile können wir einfach weglassen durch Definition eines neuen u oder y.

Nun zum Beweis der kondensierten Form. Wir machen das konstruktiv mit dem folgenden Algorithmus.

Algorithmus 6.40
Input: *Matrizen* $E, A \in \mathbb{C}^{n,n}, B \in \mathbb{C}^{n,m}, C \in \mathbb{C}^{p,n}$

Output: *unitäre Matrizen* $u, v \in \mathbb{C}^{n,n}, w \in \mathbb{C}^{m,m}$

> *$y \in C^{p,p}$, so dass $U^*EV, U^*AV, U^*BW, y^*CV$ in kondensierter Form (6.34) (6.35).*

Setze $U := I_n, V := I_n, W := I_n, Y = I_p$.

Schritt 1: *Berechne SVD* $E = U_E \begin{bmatrix} \sum_E & 0 \\ 0 & 0 \end{bmatrix} V_E^H$ *mit* $\sum_E$ *der Grösse* $t_1 \times t_1$ *und nichtsingulär. Setze*

$$\begin{cases} E & := & U_E^* E V_E = \begin{bmatrix} \sum_E & 0 \\ 0 & 0 \end{bmatrix} \\ A & := & U_E^* A V_E := \begin{bmatrix} A_{11}^{(1)} & A_{12}^{(1)} \\ A_{21}^{(1)} & A_{22}^{(1)} \end{bmatrix} \\ B & := & U_E^* B = \begin{bmatrix} B_1^{(1)} \\ B_2^{(1)} \end{bmatrix} C := C V_E = \begin{bmatrix} C_1^{(1)} & C_2^{(1)} \end{bmatrix} \\ U & := & U U_E, V := V V_E \end{cases}$$

Schritt 2: *Berechne SVD's*

$$B_2^{(1)} = U_B \begin{bmatrix} \sum_B & 0 \\ 0 & 0 \end{bmatrix} V_B^* \quad C_2^{(1)} = U_C \begin{bmatrix} \sum_C & 0 \\ 0 & 0 \end{bmatrix} V_C^*$$

mit $\sum_B$ *der Grössek*$_1 \times k_1$ *und* $\sum_C$ *der Grösse* $l_1 \times l_1$ *und nicht singulär. Setze*

$$E := \begin{bmatrix} I_{t_1} & 0 \\ 0 & U_B^* \end{bmatrix} E \begin{bmatrix} I_{t_1} & 0 \\ 0 & V_C \end{bmatrix} = \begin{bmatrix} \sum_E & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$A := \begin{bmatrix} I_{t_1} & 0 \\ 0 & U_B^* \end{bmatrix} A \begin{bmatrix} I_{t_1} & 0 \\ 0 & V_C \end{bmatrix} = \begin{bmatrix} A_{11}^{(2)} & A_{12}^{(2)} & A_{13}^{(2)} \\ A_{21}^{(2)} & A_{22}^{(2)} & A_{23}^{(2)} \\ A_{31}^{(2)} & A_{32}^{(2)} & A_{33}^{(2)} \end{bmatrix}$$

$$B := \begin{bmatrix} I_{t_k} & 0 \\ 0 & U_B^* \end{bmatrix} B V_B = \begin{bmatrix} B_{11}^{(2)} & B_{12}^{(2)} \\ \sum_B & 0 \\ 0 & 0 \end{bmatrix}$$

$$C := U_C^* C \begin{bmatrix} I_{t_1} & 0 \\ 0 & V_C \end{bmatrix} = \begin{bmatrix} C_{11}^{(2)} & \sum_C & 0 \\ C_{21}^{(2)} & 0 & 0 \end{bmatrix},$$

$$U := U \begin{bmatrix} I_{t_1} & 0 \\ 0 & U_B \end{bmatrix}, V := V \begin{bmatrix} I_{t_1} & \\ & V_C \end{bmatrix}, Y = Y U_C, W = W V_B$$

Schritt 3: *Berechne SVDs*

$$B_{12}^{(2)} = U_{12} \begin{bmatrix} \sum_{12} & 0 \\ 0 & 0 \end{bmatrix} V_{12}^*, C_{21}^{(2)} = U_{21} \begin{bmatrix} \sum_{21} & 0 \\ 0 & 0 \end{bmatrix} V_{21}^*$$

mit $\sum_{21}$ *der Grösse* $k_2 \times k_2$, *und* $\sum_{12}$ *der Grösse* $l_2 \times l_2$ *nicht singulär und setze*

$$B := B \begin{bmatrix} I_{k_1} & 0 \\ 0 & V_{12} \end{bmatrix} = \begin{bmatrix} B_{11}^{(3)} & B_{12}^{(3)} & 0 \\ \sum_B & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, W := W \begin{bmatrix} I_{k_1} & 0 \\ 0 & V_{12} \end{bmatrix}$$

$$C := \begin{bmatrix} I_{l_1} & 0 \\ 0 & U_{21}^* \end{bmatrix} C = \begin{bmatrix} C_{11}^{(3)} & \sum_C & 0 \\ C_{21}^{(3)} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, Y := Y \begin{bmatrix} I_{l_1} & 0 \\ 0 & U_{21} \end{bmatrix}.$$

Schritt 4: *Berechne SVD*

$$A_{33}^{(2)} = U_A \begin{bmatrix} \sum_{44} & 0 \\ 0 & 0 \end{bmatrix} V_A^*$$

mit $\sum_{44}$ *der Grösse* $t_4 \times t_4$ *nichtsingulär und setze*

$$E := \begin{bmatrix} I_{t_1} & 0 & 0 \\ 0 & I_{k_1} & 0 \\ 0 & 0 & U_A^* \end{bmatrix} E \begin{bmatrix} I_{t_1} & 0 & 0 \\ 0 & I_{l_1} & 0 \\ 0 & 0 & V_A \end{bmatrix} =: \begin{bmatrix} \sum_E & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$A := \begin{bmatrix} I_{t_1} & 0 & 0 \\ 0 & I_{k_1} & 0 \\ 0 & 0 & U_A^* \end{bmatrix} A = \begin{bmatrix} I_{t_1} & 0 & 0 \\ 0 & I_{l_1} & 0 \\ 0 & 0 & V_A \end{bmatrix} =: \begin{bmatrix} A_{21}^{(3)} & A_{12}^{(3)} & A_{13}^{(3)} & A_{14}^{(3)} \\ A_{21}^{(3)} & A_{22}^{(3)} & A_{23}^{(3)} & A_{24}^{(3)} \\ A_{31}^{(3)} & A_{32}^{(3)} & \sum_{44} & 0 \\ A_{41}^{(3)} & A_{42}^{(3)} 0 & 0 & \end{bmatrix}$$

$$B := \begin{bmatrix} I_{t_1} & 0 & 0 \\ 0 & I_{k_1} & 0 \\ 0 & 0 & U_A^* \end{bmatrix} B := \begin{bmatrix} B_{11}^{(3)} & B_{12}^{(3)} & 0 \\ \sum_B 0 & 0 & \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, U := U \begin{bmatrix} I_{t_1} & & \\ & I_{k_1} & \\ & & U_A \end{bmatrix}$$

$$C := C \begin{bmatrix} I_{t_1} & 0 & 0 \\ 0 & I_{l_1} & 0 \\ 0 & 0 & V_A \end{bmatrix} = \begin{bmatrix} C_{11}^{(3)} & \sum_C & 0 & 0 \\ C_{21}^{(3)} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$V := V \begin{bmatrix} I_{t_1} & 0 & 0 \\ 0 & I_{l_1} & 0 \\ 0 & 0 & V_A \end{bmatrix}$$

Schritt 5: *Berechne permutierte SVDs*

$$A_{42}^{(3)} = U_{42} \begin{bmatrix} 0 & \sum_{53} \\ 0 & 0 \end{bmatrix} V_{42}^*, \; A_{24}^{(3)} = U_{24} = \begin{bmatrix} 0 & 0 \\ \sum_{35} & 0 \end{bmatrix} V_{24}^*$$

mit $\sum_{53}$ *der Grösse* $t_5 \times t_5$ *und* $\sum_{35}$ *der Grösse* $t_3 \times t_3$ *nichtsingulär und setze*

$$E := \begin{bmatrix} I_{t_1} & & & \\ & U_{24}^* & & \\ & & I_{t_4} & \\ & & & U_{42}^* \end{bmatrix} E \begin{bmatrix} I_{t_1} & & & \\ & V_{42} & & \\ & & I_{t_4} & \\ & & & V_{24} \end{bmatrix} = \begin{bmatrix} \sum_E & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{matrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \end{matrix}$$

$$A := \begin{bmatrix} I_{t_1} & & & \\ & U_{24}^* & & \\ & & I_{t_4} & \\ & & & U_{42}^* \end{bmatrix} E \begin{bmatrix} I_{t_1} & & & \\ & V_{42} & & \\ & & I_{t_4} & \\ & & & V_{24} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} & A_{15} & A_{16} \\ A_{21} & A_{22} & A_{23} & A_{24} & 0 & 0 \\ A_{31} & A_{32} & A_{33} & A_{34} & \sum_{35} & 0 \\ A_{41} & A_{42} & A_{43} & \sum_{44} & 0 & 0 \\ A_{51} & 0 & \sum_{53} & 0 & 0 & 0 \\ A_{61} & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$B := \begin{bmatrix} I_{t_1} & & & \\ & U_{24}^* & & \\ & & I_{t_4} & \\ & & & U_{42}^* \end{bmatrix} B = \begin{bmatrix} B_{11} & B_{12} & 0 \\ B_{21} & 0 & 0 \\ B_{31} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} .U := U \begin{bmatrix} I_{t_1} & & & \\ & U_{24} & & \\ & & I_{t_4} & \\ & & & U_{42} \end{bmatrix}$$

$$C := C \begin{bmatrix} I_{t_1} & & & \\ & V_{42} & & \\ & & I_{t_4} & \\ & & & V_{24} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{13}0 & 0 & 0 & \\ C_{21} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{matrix} l_1 \\ l_2 \\ l_3 \end{matrix}$$

$$V := V \begin{bmatrix} I_{t_1} & 0 & 0 & 0 \\ 0 & V_{42}0 & 0 & \\ 0 & 0 & I_{t_4} & 0 \\ 0 & 0 & 0 & V_{24} \end{bmatrix}.$$

Wir können also die kondensierte Form durch 8 SVDs berechnen. Falls wir nur Zustandsrückkopplungen betrachten wollen und keine Ausgangsrückkopplungen, so können wir $C = I$ setzen und die Transformationen, die auf C operieren einfach weglassen.

Mit Hilfe von Satz 6.33 können wir nun Rückkopplungssteuerungen konstruieren:

Satz 6.41 *Seien* E, A, B, C *in der Form (6.34), (6.35) gegeben. Falls das System steuerbar und beobachtbar bei* ∞ *ist, d. h.*

$$Rang\ [E, AS_\infty, B] = Rang \begin{bmatrix} E \\ T_\infty^* A \\ C \end{bmatrix} = n \tag{6.42}$$

so gibt es für alle $s \in N \quad 0 \le s \le t_2 = s_2$ *Matrizen* $F, G \in \mathbb{C}^{m,p}$*, so dass*

$$\alpha(E + BGC) - \beta(A + BFC) \text{ regulär, index } 1 \text{ und } Rang\ (E + BGC) = t_1 + s. \tag{6.43}$$

i) Falls $s = t_2$, dann brauchen wir dazu nur Ausgangsableitungsrückkopplung ($F = 0$).

ii) Falls $s = 0$, dann erreichen wir das mit nur Ausgangsrückkopplung ($G = 0$). In diesem Fall gilt auch die Umkehrung, d. h.: Falls es F gibt, so dass (6.43) gilt mit $G = 0$ so ist das System steuerbar und beobachtbar bei ∞.

Korollar 6.44 *Seien E, A, B, C wie in Satz 6.41. Falls Rang $(\lambda E - A, B) = n \forall \lambda \in \mathbb{C}$, und Rang $[E, AS_\infty, B] =$ Rang $\begin{bmatrix} E \\ T_\infty^* A \\ C \end{bmatrix} = n$, so gibt es $F, G \in \mathbb{C}^{m,p}$ und eine Steuerung*

$$u = Fy - G\dot{y} + v, \tag{6.45}$$

so dass das neue „closed loop" System, welches durch $(E + BGC, A + BFC, B, C)$ gegeben ist, stark steuerbar und beobachtbar ist, mit index ≤ 1 und Rang $(E + BGC) = t_1 + s$.

Beweis: Aus der Steuerbarkeit und Beobachtbarkeit bei ∞ folgt die Existenz von B, C so dass $\alpha(E + BGC) - \beta(A + BFC)$ regulär index 1, und immer noch steuerbar und beobachtbar bei ∞. Die anderen beiden Bedingungen sind invariant unter Rückkopplung $\Longrightarrow$ Beh. □

Korollar 6.46 *Seien E, A, B, C wie in Satz 6.41.*

i) Es gibt G, wso dass $E + BGC$ nichtsingulär genau dann wenn

$$\text{Rang } [E, B] = \text{ Rang } \begin{bmatrix} E \\ C \end{bmatrix} = n.$$

ii) Es gibt $G \in \mathbb{C}^{p,m}$ und Steuerung $u = -G\dot{y} + v$ so dass der geschlossene Kreis mit $(E + BGC, A, BC)$ vollständig steuerbar und beobachtbar ist mit Rang $(E + BGC) = n$ genau dann wenn

$$\text{Rang } [\alpha E - \beta A, B] = \text{ Rang } \begin{bmatrix} \alpha E - \beta B \\ C \end{bmatrix} = n \text{ für alle } (\alpha, \beta) \neq (0, 0). \tag{6.47}$$

Beweis: Wähle $s = t_2$ in Satz 6.41. Für die Umkehrung folgt Rang $\begin{bmatrix} E \\ C \end{bmatrix}$ = Rang $[EB] = n$ aus $t_3 = t_4 = t_5 = t_6 = s_6 = 0$. (6.47) ist invariant unter Rückkopplung $\Longrightarrow$ Beh. □

Man kann die vorhandene Freiheit in der Wahl von F, G nutzen um die Kondition des Problems zu verbessern, d. h., falls das Problem in der Form

$$\begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A12 \\ A_{21} & A_{22} \end{bmatrix}, \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \begin{bmatrix} C_1 & C_2 \end{bmatrix} \tag{6.48}$$

ist, so soll E_{11}, A_{22} beide gut konditioniert für Inversion sind.