Steuerungstheorie II*

Martin Brokate †

Inhaltsverzeichnis

*Vorlesungsskript, WS 1994/95

†Institut für Informatik und Praktische Mathematik, Universität Kiel, 24098 Kiel

1 Schur-Form und Singulärwertzerlegung

Die Matrizen eines Kontrollsystems (A, B, C) sind in der Regel unsymmetrisch. Für die numerische Behandlung spielen daher orthogonale Transformationen eine entscheidende Rolle (sie verändern die Kondition nicht).

Satz 1.1 (Schur-Form, komplexe Version)
Sei $A \in \mathbb{C}^{(n,n)}$. Dann gibt es $Q, R \in \mathbb{C}^{(n,n)}$ mit

$$A = QRQ^H\,, \tag{1.1}$$

wobei Q hermitesch und R obere Dreiecksmatrix ist.

Beweis: Lineare Algebra oder in den Übungen. □

Satz 1.2 (Schur-Form, reelle Version)
Sei $A \in \mathbb{R}^{(n,n)}$. Dann gibt es $Q, R \in \mathbb{R}^{(n,n)}$ mit

$$A = QRQ^T\,, \tag{1.2}$$

wobei Q orthogonal und R eine sogenannte quasi-obere Dreiecksmatrix ist, d.h. R hat die Form

$$R = \begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1m} \\ 0 & R_{22} & \cdots & R_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & R_{mm} \end{pmatrix}. \tag{1.3}$$

Für jedes R_{ii}, $1 \le i \le m$, gibt es zwei Möglichkeiten:

1. *$R_{ii} \in \mathbb{R}^{(1,1)} = \mathbb{R}$, oder*
2. *$R_{ii} \in \mathbb{R}^{(2,2)}$ mit konjugiert komplexen Eigenwerten.*

In jedem Fall gilt

$$\operatorname{spec}(A) = \bigcup_{1 \le i \le m} \operatorname{spec}(R_{ii})\,. \tag{1.4}$$

($\operatorname{spec}(A)$ bezeichnet das Spektrum von A.)

Beweis: Lineare Algebra oder in den Übungen. □

Definition 1.3 (Reelle Singulärwertzerlegung, SVD)
Sei $A \in \mathbb{R}^{(p,q)}$. Eine (reelle) Singulärwertzerlegung von A hat die Form

$$A = U\Sigma V^T = (U_1\ U_2) \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix}, \tag{1.5}$$

wobei $U \in \mathbb{R}^{(p,p)}$, $V \in \mathbb{R}^{(q,q)}$ orthogonale Matrizen sind (d.h. $U^{-1} = U^T$), und $S = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ Diagonalmatrix ist mit $0 \le r \le \min\{p, q\}$ und $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0$. Man setzt $\sigma_k = 0$ für $k > r$ und bezeichnet den kleinsten Singulärwert $\sigma_{\min\{p,q\}}$ auch mit $\sigma_{\min}(A)$.

Satz 1.4 (Existenz der SVD)
Jede Matrix $A \in \mathbb{R}^{(p,q)}$ hat eine Singulärwertzerlegung.

Beweis: Sei $A \neq 0$. Wir wählen $x \in \mathbb{R}^q$ mit $\|x\|_2 = 1$ und $\|Ax\|_2 = \|A\|_2$. Sei

$$y = \frac{Ax}{\|Ax\|_2}. \tag{1.6}$$

Falls $q > 1$, so ergänzen wir x zu einer orthogonalen Matrix $V = (x\ \tilde{V}) \in \mathbb{R}^{(q,q)}$; falls $p > 1$, ergänzen wir y zu einer orthogonalen Matrix $U = (y\ \tilde{U}) \in \mathbb{R}^{(p,p)}$. Für

$$\Sigma := U^T A V = \begin{pmatrix} \|A\|_2 & w^T \\ 0 & * \end{pmatrix}, \quad w = \frac{\tilde{V}^T A^T A x}{\|A\|_2}, \tag{1.7}$$

gilt dann

$$\Sigma \begin{pmatrix} \|A\|_2 \\ w \end{pmatrix} = \begin{pmatrix} \|A\|_2^2 + \|w\|_2^2 \\ * \end{pmatrix}, \tag{1.8}$$

also

$$\|A\|_2^2 + \|w\|_2^2 \leq \left\| \Sigma \begin{pmatrix} \|A\|_2 \\ w \end{pmatrix} \right\|_2 \leq \|\Sigma\|_2 \sqrt{\|A\|_2^2 + \|w\|_2^2}. \tag{1.9}$$

Aus $\|\Sigma\|_2 = \|A\|_2$ folgt nun $w = 0$. Die Behauptung folgt mit einem Induktionsschluß (über den rechten unteren Teil von Σ. Falls $p = 1$ oder $q = 1$, so ist man bereits fertig). □

Satz 1.5 (Elementare Eigenschaften der SVD)
Sei $A = U\Sigma V^T$ Singulärwertzerlegung wie in Definition 1.3, seien u_i, v_i die Spalten von U bzw. V. Dann gilt

(i) *Σ ist eindeutig bestimmt (i.a. aber nicht U und V).*

(ii) $\operatorname{rang}(A) = \operatorname{rang}(\Sigma) = r$.

(iii) *$Av_i = \sigma_i u_i$ und $A^T u_i = \sigma_i v_i$ für alle i.*

(iv) $\|A\|_2 = \sigma_1$, $\|A\|_F^2 = \sum_{i=1}^r \sigma_i^2$.

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

Beweis: Σ ist eindeutig bestimmt, da $\Sigma^T\Sigma$ die Diagonalisierung von A^TA darstellt. Die anderen Behauptungen folgen aus der Definition durch spaltenweises Betrachten der Gleichungen

$$AV = U\Sigma, \quad A^T U = V\Sigma, \tag{1.10}$$

sowie daraus, daß orthogonale Matrizen längentreu abbilden. □

Satz 1.6
Für $A, E \in \mathbb{R}^{(p,q)}$ und alle k gilt

$$|\sigma_k(A+E) - \sigma_k(A)| \leq \sigma_1(E) = \|E\|_2. \tag{1.11}$$

Beweis: Folgt aus der Störungstheorie für Eigenwerte. (Siehe Buch von Golub/van Loan bzw. Wilkinson.) □

Satz 1.6 bedeutet, daß die Berechnung der singulären Werte ein gut konditioniertes Problem ist.

Satz 1.7
Sei $A \in \mathbb{R}^{(p,q)}$, $k < r = \text{ rang}\,(A)$. *Dann gilt*

$$\sigma_{k+1} = \min_{\text{rang}\,(B)\le k} \|A - B\|_2 = \|A - A_k\|_2 \,, \tag{1.12}$$

wobei

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T \,, \qquad \text{rang}\,(A_k) = k \,. \tag{1.13}$$

Beweis: Es ist $\text{rang}\,(A_k) = k$, da $U^T A_k V = \text{ diag}\,(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$. Ferner ist

$$\|A - A_k\|_2 = \|U^T(A - A_k)V\|_2 = \|\text{ diag}\,(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_r, 0, \ldots, 0)\|_2 = \sigma_{k+1} \,. \tag{1.14}$$

Sei nun $B \in \mathbb{R}^{(p,q)}$ mit $\text{rang}\,(B) \le k$. Dann gilt

$$\dim(H) > 0 \,, \quad H := \ker(B) \cap \text{ span}\,(v_1, \ldots, v_{k+1})) \,. \tag{1.15}$$

Sei $z \in H$, $\|z\|_2 = 1$. Dann gilt

$$z = \sum_{i=1}^{k+1} (v_i^T z) v_i \,, \quad Az = \sum_{i=1}^{r} \sigma_i u_i v_i^T z = \sum_{i=1}^{k+1} \sigma_i u_i v_i^T z \,, \tag{1.16}$$

$$\|A - B\|_2^2 \ge \|(A - B)z\|_2^2 = \|Az\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 (v_i^T z)^2 \ge \sigma_{k+1}^2 \sum_{i=1}^{k+1} (v_i^T z)^2 = \sigma_{k+1}^2 \,. \tag{1.17}$$

□

Bemerkung 1.8 (Rangberechnung)
Satz 1.7 zeigt, daß die Singulärwertzerlegung den Abstand einer gegebenen Matrix A zur Menge der Matrizen vom Rang k liefert. Man kann also den numerischen Rang von A definieren als

$$\tilde{r} = \max\{k : \sigma_{k+1} \ge \delta\} + 1 \,. \tag{1.18}$$

Hierbei ist $\delta > 0$ eine vorgegebene (kleine) Größe, die in Abhängigkeit von Maschinengenauigkeit und Datenfehler in A gewählt werden muß. Probleme ergeben sich, falls einer oder mehrere singuläre Werte in der Größenordnung von δ liegen.

Bemerkung 1.9 (Zu A gehörende Unterräume)
Sei $A \in \mathbb{R}^{(p,q)}$ mit der Singulärwertzerlegung

$$A = U\Sigma V^T = (U_1\ U_2) \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix} \,, \tag{1.19}$$

Dann gilt

$$\text{im}\, A = \text{ im}\, U_1 \,, \quad (\text{ im}\, A)^\perp = (\text{ im}\, U_1)^\perp = \text{ im}\, U_2 \,, \tag{1.20}$$

$$\ker A = \ker V_1^{\perp} = (\ \mathrm{im}\, V_1)^{\perp} = \ \mathrm{im}\, V_2\,, \quad (\ker A)^{\perp} = (\ \mathrm{im}\, V_1)^{\perp\perp} = \ \mathrm{im}\, V_1\,, \tag{1.21}$$

und die zugehörigen Orthogonalprojektionen sind gegeben durch

$$U_1U_1^T\ ,\ U_2U_2^T\ ,\ V_2V_2^T\ ,\ V_1V_1^T\,. \tag{1.22}$$

Es ist z.B.

$$U_1U_1^T x = \sum_{i=1}^{r} u_i u_i^T x = \sum_{i=1}^{r} \langle u_i, x\rangle u_i\,. \tag{1.23}$$

2 Householdertransformation und Givensrotation

Bemerkung 2.1 (Householdertransformation)
Sei $w \in \mathbb{R}^n$, $w \neq 0$. Dann wird durch

$$Q = I - 2\frac{ww^T}{w^T w} \tag{2.1}$$

eine orthogonale Matrix $Q \in \mathbb{R}^{(n,n)}$ definiert. Q repräsentiert die Spiegelung an der auf w orthogonalen Hyperebene und heißt *elementare Householdertransformation*. Ist $x \in \mathbb{R}^n$ mit $x \neq 0$ gegeben, und setzt man

$$w = x \pm \|x\|_2 e_1 \,, \tag{2.2}$$

so ergibt sich

$$Qx = \mp\|x\|_2 e_1 \,. \tag{2.3}$$

Ist $A \in \mathbb{R}^{(n,m)}$ gegeben und wählt man für x die erste Spalte von A, so erhält man

$$QA = \begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \cdots & * \end{pmatrix} \tag{2.4}$$

mit einem Aufwand von $O(nm)$ Operationen. Ist $A \in \mathbb{R}^{(m,n)}$ gegeben und wählt man für x die erste Zeile von A, so erhält man

$$AQ^T = \begin{pmatrix} * & 0 & \cdots & 0 \\ * & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & * \end{pmatrix} \tag{2.5}$$

mit einem Aufwand von $O(nm)$ Operationen. Durch geeignete Einschränkung auf niederdimensionale Unterräume kann man auch Teile von Zeilen oder Spalten zu Null machen. Sei etwa $A \in \mathbb{R}^{(N,M)}$,

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad A_{11} \in \mathbb{R}^{(N-n,M-m)}, A_{22} \in \mathbb{R}^{(n,m)} \,. \tag{2.6}$$

Wählt man für x die erste Spalte von A_{22} und setzt

$$\tilde{Q} = \begin{pmatrix} I & 0 \\ 0 & Q \end{pmatrix}, \tag{2.7}$$

so ist

$$\tilde{Q}A = \begin{pmatrix} A_{11} & & A_{12} & & \\ & * & * & \cdots & * \\ & 0 & * & \cdots & * \\ QA_{21} & \vdots & \vdots & \ddots & \vdots \\ & 0 & * & \cdots & * \end{pmatrix} . \tag{2.8}$$

Bei der Transformation (2.8) bleiben alle Nullen von A_{11} und A_{12} sowie alle Nullspalten von A_{21} erhalten. Auf analoge Weise kann man erreichen, daß

$$A\tilde{Q}^T = \begin{pmatrix} A_{11} & A_{12}Q^T \\ & \begin{matrix} * & 0 & \cdots & 0 \\ * & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & * \end{matrix} \\ A_{21} & \end{pmatrix} . \tag{2.9}$$

Hierbei bleiben alle Nullen in A_{11} und A_{21} sowie Nullzeilen in A_{12} erhalten.

Bemerkung 2.2 (Givens-Rotation)
Die Matrix

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} , \quad c = \cos(\varphi) , \quad s = \sin(\varphi) , \tag{2.10}$$

repräsentiert im $\mathbb{R}^2$ die Drehung um den Winkel φ. Entsprechend repräsentiert die orthogonale Matrix $Q \in \mathbb{R}^{(n,n)}$,

$$\begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & c & & & s & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & -s & & & & c & \\ & & & & & & & & 1 \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{pmatrix} , \tag{2.11}$$

wobei die Elemente c und $\pm s$ in der i-ten und k-ten Spalte bzw. Zeile auftreten, die Drehung des zweidimensionalen Unterraums $\operatorname{span}\{e_i, e_k\}$ im $\mathbb{R}^n$ um den Winkel φ. Ist $x \in \mathbb{R}^n$ mit $x \neq 0$ gegeben und setzen wir

$$c = \frac{x_i}{\sqrt{x_i^2 + x_k^2}} , \quad s = \frac{x_k}{\sqrt{x_i^2 + x_k^2}} , \tag{2.12}$$

so wird

$$(Qx)_k = 0 , \quad (Qx)_j = x_j \quad \text{für } j \neq i, k . \tag{2.13}$$

Für eine Matrix $A \in \mathbb{R}^{(n,m)}$ bedeutet dies, daß wir einzelne Elemente zu Null machen können. Wählen wir für x die l-te Spalte von A und setzen wir

$$c = \frac{a_{il}}{\sqrt{a_{il}^2 + a_{kl}^2}} , \quad s = \frac{a_{kl}}{\sqrt{a_{il}^2 + a_{kl}^2}} , \tag{2.14}$$

so wird

$$(QA)_{kl} = 0 , \tag{2.15}$$

die Zeilen i und k ändern sich (sie werden Linearkombinationen der alten Zeilen i und k), alle anderen Zeilen bleiben unverändert. Es bleiben also alle Nullen außerhalb der Zeilen i und k und alle gemeinsamen (d.h. in derselben Spalte liegenden) Nullen der Zeilen i und k erhalten. Der Aufwand der Givens-Rotation ist $O(m)$ Operationen. Entsprechendes gilt für $A \in \mathbb{R}^{(m,n)}$ und AQ^T.

Algorithmus 2.3 (QR-Zerlegung)
Durch eine geeignete Folge von elementaren Householdertransformationen $Q_1, \ldots, Q_k$, $k = \min\{m, n\}$, erhält man aus $A \in \mathbb{R}^{(n,m)}$ die QR-Zerlegung

$$A = QR\,, \quad Q^T = Q_k Q_{k-1} \ldots Q_1\,, \tag{2.16}$$

wobei $R \in \mathbb{R}^{(n,m)}$ eine obere Dreiecksmatrix ist:

$$A = \begin{pmatrix} * & \cdots & \cdots & * \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ * & \cdots & \cdots & * \\ \vdots & & & \vdots \\ * & \cdots & \cdots & * \end{pmatrix} \mapsto Q_1 A = \begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & * & \cdots & * \end{pmatrix} \mapsto Q_2 Q_1 A = \begin{pmatrix} * & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ 0 & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & * & \cdots & * \end{pmatrix} \tag{2.17}$$

Der Aufwand beträgt $O(knm)$ Operationen. Die QR-Zerlegung ist bis auf Vorzeichenänderung der Zeilen von R eindeutig bestimmt. Durch Anwenden der elementare Householdertransformationen von rechts auf die Zeilen erhält man analog

$$A = LQ^T\,, \tag{2.18}$$

mit einer unteren Dreiecksmatrix L und einer orthogonalen Matrix Q.

Algorithmus 2.4 (Ähnlichkeitstransformation auf Hessenberg-Form)
Eine Matrix $A \in \mathbb{R}^{(n,n)}$ mit $a_{ij} = 0$ für $j + 1 < i$ heißt obere Hessenbergmatrix:

$$A = \begin{pmatrix} * & \cdots & \cdots & \cdots & * \\ * & \ddots & & & \vdots \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & * & * \end{pmatrix} \tag{2.19}$$

Jede Matrix $A \in \mathbb{R}^{(n,n)}$ läßt sich durch eine orthogonale Ähnlichkeitstransformation in eine obere Hessenberg-Matrix transformieren, d.h.

$$A = QHQ^T\,, \tag{2.20}$$

mit Q orthogonal, H obere Hessenberg-Matrix. Und zwar:

$$A = \left(\begin{array}{cccc} * & \cdots & \cdots & * \\ \hline * & \cdots & \cdots & * \\ \vdots & & & \vdots \\ * & \cdots & \cdots & * \end{array}\right) \quad \mapsto \quad Q_1 A = \left(\begin{array}{cccc} * & \cdots & \cdots & * \\ \hline * & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & * & \cdots & * \end{array}\right) \tag{2.21}$$

$$\mapsto \quad Q_1AQ_1^T = \begin{pmatrix} * & \cdots & \cdots & * \\ * & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & * & \cdots & * \end{pmatrix} \quad \mapsto \quad Q_2Q_1AQ_1^T = \left(\begin{array}{ccccc} * & \cdots & \cdots & \cdots & * \\ * & * & \cdots & \cdots & * \\ \hline 0 & * & * & \cdots & * \\ \vdots & 0 & * & \vdots & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & * & \cdots & * \end{array}\right) \tag{2.22}$$

und so weiter. Der Aufwand beträgt $\frac{5}{3}n^3$ Operationen, falls die Q_i einzeln gespeichert werden, bzw. $\frac{7}{3}n^3$ Operationen, falls deren Produkt Q berechnet wird. Zur numerischen Realisierung vgl. Golub/van Loan.

Algorithmus 2.5 (QR-Zerlegung einer Hessenberg-Matrix)
Ist $H \in \mathbb{R}^{(n,n)}$ eine obere Hessenberg-Matrix, so brauchen zur Berechnung der QR-Zerlegung von H keine Spiegelungen vorgenommen zu werden, sondern es genügen Givens-Rotationen:

$$H = \begin{pmatrix} * & * & \cdots & \cdots & * \\ * & * & & & \vdots \\ 0 & * & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & * & * \end{pmatrix} \quad \mapsto \quad Q_1H = \begin{pmatrix} * & * & \cdots & \cdots & * \\ 0 & * & * & & \vdots \\ 0 & * & * & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & * & * \end{pmatrix} \tag{2.23}$$

$$\mapsto \quad Q_2Q_1H = \begin{pmatrix} * & * & \cdots & \cdots & * \\ 0 & * & * & & \vdots \\ 0 & 0 & * & & \vdots \\ \vdots & \ddots & * & \ddots & \vdots \\ 0 & \cdots & 0 & * & * \end{pmatrix} \tag{2.24}$$

und so weiter. Der Aufwand ist $O(n^2)$ Operationen.

Algorithmus 2.6 (Berechnung der Übertragungsfunktion)
Sei $(A, B, C) \in S_{n,m,k}$ ein Kontrollsystem. Wir wollen die Übertragungsfunktion

$$G(s) = C(sI - A)^{-1}B\,, \quad s \in \mathbb{C}\,, \tag{2.25}$$

für viele verschiedene Werte von s berechnen (z.B. um gewisse Charakteristika der Übertragungsfunktion zu plotten). Wir machen uns dabei zunutze, daß sich die Übertragungsfunktion nicht ändert, wenn wir eine Ähnlichkeitstransformation des Kontrollsystems durchführen. Verfahren also:

1. Transformiere (A, B, C) orthogonal auf

$$(\tilde{A}, \tilde{B}, \tilde{C}) = (Q^TAQ, Q^TB, CQ)\,, \tag{2.26}$$

so daß $\tilde{A}$ eine obere Hessenberg-Matrix ist. Es ist dann

$$G(s) = \tilde{C}(sI - \tilde{A})^{-1}\tilde{B}\,. \tag{2.27}$$

2. Für gegebenes $s \in \mathbb{C}$ berechne die QR-Zerlegung

$$sI - \tilde{A} = Q(s)R(s)\,, \tag{2.28}$$

$Q \in \mathbb{C}^{(n,n)}$ hermitesch, $R \in \mathbb{C}^{(n,n)}$ obere Dreiecksmatrix. Berechne

$$X(s) = (sI - \tilde{A})^{-1}\tilde{B} \tag{2.29}$$

aus

$$R(s)X(s) = Q^H(s)\tilde{B} \tag{2.30}$$

spaltenweise durch Rückwärtseinsetzen, setze dann

$$G(s) = \tilde{C}X(s)\,. \tag{2.31}$$

Schritt 1 benötigt $O(n^3)$ Operationen, muß aber nur einmal durchgeführt werden. Schritt 2 muß für jeden Wert von s durchgeführt werden, benötigt aber nur $O(n^2m)$ Operationen.

3 Numerische Berechnung der SVD

Bemerkung 3.1 (SVD und Schur-Form)
Theoretisch kann man die Berechnung der SVD auf die Berechnung der Schur-Form zurückführen, da die Singulärwertzerlegung von A

$$A = U\Sigma V^T \tag{3.1}$$

sich aus der Schur-Form von A^TA bzw. AA^T,

$$A^TA = V(\Sigma^T\Sigma)V^T\,, \quad AA^T = U(\Sigma\Sigma^T)U^T\,, \tag{3.2}$$

ergibt. Aus der Sicht der Numerik ist dieses Vorgehen problematisch, da der Übergang von A zu A^TA die Kondition quadriert. Es ist besser, direkt auf A zu arbeiten. Das allgemein akzeptierte Verfahren zur Berechnung der SVD stammt von Golub/Kahan (1965) bzw. Golub/Reinsch (1970). Die zugrundeliegende Idee ist, anhand des Zusammenhangs von (3.1) und (3.2) das QR-Verfahren (welches die Schur-Form einer Matrix berechnet) geeignet zu modifizieren.

Bemerkung 3.2 (Allgemeine Struktur des Verfahrens)
Sei $A \in \mathbb{R}^{(n,n)}$. Das QR-Verfahren zur Berechnung der Schur-Form von A zerfällt in zwei Schritte:

1. Orthogonale Transformation auf obere Hessenbergform,

$$A = Q_0 H Q_0^T\,. \tag{3.3}$$

2. Berechnung der Schur-Form von H.

Ist A symmetrisch, so ist H ebenfalls symmetrisch, also Tridiagonalmatrix. Die Übertragung von Schritt 1 auf die SVD sieht folgendermaßen aus: Sei $A \in \mathbb{R}^{(m,n)}$ mit $m \geq n$. Wir bringen A durch elementare Householdertransformationen (von links und von rechts) auf obere Bidiagonalform

$$A = \begin{pmatrix} * & \cdots & \cdots & * \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ * & \cdots & \cdots & * \\ \vdots & & & \vdots \\ * & \cdots & \cdots & * \end{pmatrix} \mapsto \begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & * & \cdots & * \end{pmatrix} \mapsto \begin{pmatrix} * & * & 0 & \cdots & 0 \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & & & \vdots \\ 0 & * & \cdots & \cdots & * \\ \vdots & \vdots & & & \vdots \\ 0 & * & \cdots & \cdots & * \end{pmatrix} \tag{3.4}$$

$$\mapsto \begin{pmatrix} * & * & 0 & \cdots & 0 \\ 0 & * & * & \cdots & * \\ 0 & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & * & \cdots & * \end{pmatrix} \mapsto \quad \cdots \mapsto P_0^T A Q_0 = \begin{pmatrix} B \\ 0 \end{pmatrix} \tag{3.5}$$

mit

$$B = \begin{pmatrix} d_1 & f_2 & 0 & \cdots & 0 \\ 0 & d_2 & f_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & f_n \\ 0 & \cdots & \cdots & 0 & d_n \end{pmatrix} \in \mathbb{R}^{(n,n)} . \tag{3.6}$$

Es ist dann

$$A^T A = Q_0 B^T B Q_0^T , \tag{3.7}$$

d.h. B ist der Cholesky-Faktor der Tridiagonalmatrix

$$T = B^T B , \tag{3.8}$$

welche durch orthogonale Ähnlichkeitstransformation von $A^T A$ (statt A in (3.3)) entsteht.

Bemerkung 3.3 (Schur-Form einer Tridiagonalmatrix: QR-Verfahren)
Gegeben ist die symmetrische Tridiagonalmatrix $T \in \mathbb{R}^{(n,n)}$,

$$T = \begin{pmatrix} a_1 & b_2 & 0 & \cdots & 0 \\ b_2 & a_2 & b_3 & \ddots & \vdots \\ 0 & b_3 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_n \\ 0 & \cdots & 0 & b_n & a_n \end{pmatrix} . \tag{3.9}$$

Gesucht ist die Schur-Form von T,

$$T = QDQ^T , \quad Q \text{ orthogonal } , D \text{ diagonal } . \tag{3.10}$$

Der grundlegende Iterationsschritt des QR-Verfahrens sieht folgendermaßen aus:

1. Wähle Shift-Parameter $\mu \in \mathbb{R}$.
2. Berechne eine QR-Zerlegung
$$T - \mu I = Q_+ R_+ . \tag{3.11}$$
3. Setze
$$T_+ = R_+ Q_+ + \mu I . \tag{3.12}$$

Die Matrix T_+ ist wieder eine symmetrische Tridiagonalmatrix (Übung). Der Übergang von T nach T_+ ist eine orthogonale Ähnlichkeitstransformation, da

$$Q_+ T_+ Q_+^T = Q_+ R_+ + \mu I = T , \quad T_+ = Q_+^T T Q_+ . \tag{3.13}$$

Der Shift-Parameter μ wird so gewählt, daß das Element b_{+n} von T_+ klein wird (Ziel: $b_{+n} = 0$, da dann a_{+n} ein Eigenwert von T_+ ist). Eine übliche Variante ist der Wilkinson-Shift: μ ist derjenige Eigenwert von

$$\begin{pmatrix} a_{n-1} & b_n \\ b_n & a_n \end{pmatrix} , \tag{3.14}$$

welcher dichter bei a_n liegt. Formel:

$$\mu = a_n + d - \operatorname{sign}(d)\sqrt{d^2 + b_n^2}\,, \quad d = \frac{a_{n-1} - a_n}{2}\,. \tag{3.15}$$

In der numerischen Realisierung wird der Shift implizit behandelt (und dabei der Iterationsschritt etwas modifiziert). Wir gewinnen $\tilde{Q}_+$ als Produkt

$$\tilde{Q}_+^T = Q_{n-1} \cdots Q_1 \tag{3.16}$$

von Givens-Rotationen Q_i. Q_1 ist die Givens-Rotation mit

$$Q_1 \begin{pmatrix} a_1 - \mu \\ b_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} * \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{3.17}$$

Ähnlichkeitstransformation mit Q_1 führt auf

$$T = \begin{pmatrix} a_1 & b_2 & 0 & \cdots & 0 \\ b_2 & a_2 & b_3 & \ddots & \vdots \\ 0 & b_3 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_n \\ 0 & \cdots & 0 & b_n & a_n \end{pmatrix} \mapsto Q_1 T = \begin{pmatrix} * & * & * & 0 & \cdots & 0 \\ * & * & * & 0 & & \vdots \\ 0 & * & * & * & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ & & & & & \end{pmatrix} \tag{3.18}$$

$$\mapsto Q_1 T Q_1^T = \begin{pmatrix} * & * & \boxed{*} & 0 & \cdots & 0 \\ * & * & * & 0 & & \vdots \\ \boxed{*} & * & * & * & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ & & & & & \end{pmatrix}. \tag{3.19}$$

Die beiden Elemente $\boxed{*}$ stören die Tridiagonalgestalt, also wird eine Givens-Rotation Q_2 auf die 2. und 3. Zeile angewendet, welche das linke untere der beiden Elemente zu Null macht. Bei anschließender Rechtsmultiplikation mit Q_2 verschwindet wegen der Symmetrie auch das andere störende Element, dafür entstehen zwei neue in der zweiten Zeile bzw. Spalte:

$$Q_2 Q_1 T Q_1^T = \begin{pmatrix} * & * & \boxed{*} & 0 & \cdots & 0 \\ * & * & * & \boxed{*} & 0 & \vdots \\ 0 & * & * & * & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ & & & & & \end{pmatrix} \mapsto Q_2 Q_1 T Q_1^T Q_2^T = \begin{pmatrix} * & * & 0 & 0 & \cdots & 0 \\ * & * & * & \boxed{*} & 0 & \vdots \\ 0 & * & * & * & \ddots & \vdots \\ 0 & \boxed{*} & * & \ddots & \ddots & \vdots \\ & & & & & \end{pmatrix} \tag{3.20}$$

Das Verfahren wird analog fortgesetzt, bis am Schluß sich eine Tridiagonalmatrix $\tilde{T}_+$ ergibt. Man kann nun beweisen: Ist T unreduziert, d.h. sind alle $b_i \neq 0$, so stimmt $\tilde{T}_+$ bis auf das Vorzeichen der Außendiagonalelemente b_{i+} mit der über die QR-Zerlegung definierten Matrix T_+ überein.

Algorithmus 3.4 (Iterationsschritt des QR-Verfahrens)
Gegeben ist die symmetrische Tridiagonalmatrix $T \in \mathbb{R}^{(n,n)}$,

$$T = \begin{pmatrix} a_1 & b_2 & 0 & \cdots & 0 \\ b_2 & a_2 & b_3 & \ddots & \vdots \\ 0 & b_3 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_n \\ 0 & \cdots & 0 & b_n & a_n \end{pmatrix} . \tag{3.21}$$

Sind anfangs bzw. während der Iteration gewisse Außendiagonalelemente Null, so zerfällt T in eine Blockdiagonalmatrix, deren Blöcke tridiagonal und unreduziert sind. Der eben beschriebene QR-Schritt wird dann auf einen einzelnen Block angewandt. Für die numerische Realisierung ist zu beachten, daß sehr kleine Außendiagonalelemente ebenfalls als Null zu behandeln sind. Es ergibt sich die folgende Formulierung für den Iterationsschritt: Sei $\varepsilon = c \text{ eps}$ mit $c > 1$, aber nicht sehr groß, gegeben. Definiere T' durch

$$b_i' = \begin{cases} 0\,, & \text{falls } |b_i| \le \varepsilon(|a_{i-1}| + |a_i|)\,, \\ b_i\,, & \text{andernfalls }, \end{cases} \tag{3.22}$$

$$a_i' = a_i\,. \tag{3.23}$$

Zerlege

$$T' = \begin{pmatrix} T_1' & 0 & 0 \\ 0 & T_2' & 0 \\ 0 & 0 & T_3' \end{pmatrix} , \tag{3.24}$$

so daß T_3' Diagonalmatrix und T_2' unreduzierte Tridiagonalmatrix maximaler Größe sind. Wende dann auf T_2' den in 3.3 beschriebenen Iterationsschritt an. Das Verfahren bricht ab, wenn alle Außendiagonalelemente Null geworden sind. Das Verfahren konvergiert immer und schnell. Zur Konvergenztheorie siehe Parlett und Golub/van Loan.

Bemerkung 3.5 (SVD-Verfahren für Bidiagonalmatrizen)
Ausgangspunkt ist die Bidiagonalmatrix $B \in \mathbb{R}^{(n,n)}$,

$$B = \begin{pmatrix} d_1 & f_2 & 0 & \cdots & 0 \\ 0 & d_2 & f_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & f_n \\ 0 & \cdots & \cdots & 0 & d_n \end{pmatrix} . \tag{3.25}$$

Die Idee des Verfahrens von Golub/Kahan ist, einen SVD-Schritt

$$B_+ = U_+^T B V_+\,, \quad U_+, V_+ \text{ orthogonal }, \tag{3.26}$$

so zu konstruieren, daß der Übergang von T nach T_+,

$$T_+ = B_+^T B_+ = V_+^T B^T B V_+\,, \quad T = B^T B\,, \tag{3.27}$$

ein QR-Schritt ist. Es ist

$$B^TB = \begin{pmatrix} d_1^2 & d_1f_2 & 0 & \cdots & 0 \\ d_1f_2 & d_2^2+f_2^2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & d_{n-1}f_n \\ 0 & \cdots & 0 & d_{n-1}f_n & d_n^2+f_n^2 \end{pmatrix} . \tag{3.28}$$

Als erstes wird der Shift μ festgelegt, etwa als Wilkinson-Shift für

$$\begin{pmatrix} d_{n-1}^2+f_{n-1}^2 & d_{n-1}f_n \\ d_{n-1}f_n & d_n^2+f_n^2 \end{pmatrix} . \tag{3.29}$$

Die erste Givens-Rotation Q_1 wird analog zu (3.17) festgelegt durch

$$Q_1 \begin{pmatrix} d_1^2-\mu \\ d_1f_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} * \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} . \tag{3.30}$$

Wie dort wird das unerwünschte Element entlang der Diagonalen gejagt:

$$B = \begin{pmatrix} * & * & 0 & 0 & \cdots & 0 \\ 0 & * & * & 0 & \ddots & \vdots \\ 0 & 0 & * & * & \ddots & \vdots \\ & & & & & \end{pmatrix} \mapsto BQ_1^T = \begin{pmatrix} * & * & 0 & 0 & \cdots & 0 \\ \boxed{*} & * & * & 0 & \ddots & \vdots \\ 0 & 0 & * & * & \ddots & \vdots \\ & & & & & \end{pmatrix} \tag{3.31}$$

$$\mapsto U_1BQ_1^T = \begin{pmatrix} * & * & \boxed{*} & 0 & \cdots & 0 \\ 0 & * & * & 0 & \ddots & \vdots \\ 0 & 0 & * & * & \ddots & \vdots \\ & & & & & \end{pmatrix} \mapsto U_1BQ_1^TV_2^T = \begin{pmatrix} * & * & 0 & 0 & \cdots & 0 \\ 0 & * & * & 0 & \ddots & \vdots \\ 0 & \boxed{*} & * & * & \ddots & \vdots \\ & & & & & \end{pmatrix} \mapsto \tag{3.32}$$

bis wir wieder die Bidiagonalform

$$\tilde{B}_+ = \tilde{U}^TB\tilde{V} = (U_{n-1}\cdots U_1)B(Q_1^TV_2^T\cdots V_{n-1}^T) \tag{3.33}$$

erreichen. Es ist dann

$$B = \tilde{U}\tilde{B}_+\tilde{V}^T , \quad \tilde{B}^T\tilde{B} = \tilde{V}^TB^TB\tilde{V} . \tag{3.34}$$

Wieder läßt sich zeigen, falls $T = B^TB$ unreduziert ist (d.h. falls alle $d_i, f_i \neq 0$), daß (modulo Vorzeichenstruktur) die Matrizen B_+ und $\tilde{B}_+$ übereinstimmen. Ist T nicht unreduziert, so muß es eine Null auf der Außendiagonale von T geben, d.h. $d_{i-1}f_i = 0$ für ein i. Ist $f_i = 0$, so zerfällt B in die Form

$$B = \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix} \tag{3.35}$$

mit Bidiagonalmatrizen B_1 und B_2. Ist $d_{i-1} = 0$, so läßt sich durch Givens-Rotationen eine Nullzeile herstellen:

$$\begin{pmatrix} * & * & & & & \\ & * & * & & & \\ & & 0 & * & 0 & 0 \\ & & & * & * & 0 \\ & & & 0 & * & * \\ & & & 0 & 0 & * \end{pmatrix} \mapsto \begin{pmatrix} * & * & & & & \\ & * & * & & & \\ & & 0 & 0 & * & 0 \\ & & & * & * & 0 \\ & & & 0 & * & * \\ & & & 0 & 0 & * \end{pmatrix} \mapsto \begin{pmatrix} * & * & & & & \\ & * & * & & & \\ & & 0 & 0 & 0 & * \\ & & & * & * & 0 \\ & & & 0 & * & * \\ & & & 0 & 0 & * \end{pmatrix} \tag{3.36}$$

$$\mapsto \begin{pmatrix} * & * & & & & \\ & * & * & & & \\ & & 0 & 0 & 0 & 0 \\ & & & * & * & 0 \\ & & & 0 & * & * \\ & & & 0 & 0 & * \end{pmatrix} \tag{3.37}$$

Hier werden Givens-Rotationen nacheinander auf die Zeilenpaare $(3,4)$, $(3,5)$, $(3,6)$ angewendet.

Algorithmus 3.6 (Iterationsschritt des SVD-Verfahrens)
Gegeben ist die Bidiagonalmatrix $B \in \mathbb{R}^{(n,n)}$,

$$B = \begin{pmatrix} d_1 & f_2 & 0 & \cdots & 0 \\ 0 & d_2 & f_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & f_n \\ 0 & \cdots & \cdots & 0 & d_n \end{pmatrix}, \tag{3.38}$$

sowie ein $\varepsilon = c \operatorname{eps}$, $c > 1$. Setze

$$f_i' = \begin{cases} 0, & \text{falls } |f_i| \le \varepsilon(|d_{i-1}| + |d_i|), \\ f_i, & \text{sonst .} \end{cases} \tag{3.39}$$

Zerlege die entstehende Bidiagonalmatrix B' in Blöcke

$$B' = \begin{pmatrix} B_1' & 0 & 0 \\ 0 & B_2' & 0 \\ 0 & 0 & B_3' \end{pmatrix}, \tag{3.40}$$

so daß B_3' Diagonalmatrix und B_2' Bidiagonalmatrix (mit Außendiagonalelementen ungleich Null) maximaler Größe sind. Bearbeite nun B_2' und setze

$$d_i'' = \begin{cases} 0, & \text{falls } |d_i'| \le \varepsilon(\|B_0\|), \\ d_i', & \text{sonst .} \end{cases} \tag{3.41}$$

$$f_i'' = f_i'. \tag{3.42}$$

Hierbei ist B_0 die Startmatrix aus dem ersten Iterationsschritt des SVD-Verfahrens. Sind alle Diagonalelemente der resultierenden Matrix B_2'' ungleich Null, so führe den in 3.5 beschriebenen SVD-Schritt durch. Andernfalls wähle eine Zeile i mit $d_i'' = 0$ und zerlege B_2'' in zwei Blöcke wie ebenfalls in 3.5 beschrieben.

Bemerkung 3.7 (Konvergenz des SVD-Verfahrens)
Die Konvergenz des SVD-Verfahrens wird auf die Konvergenz des QR-Verfahrens zurückgeführt. Solange keine weitere Zerlegung vorgenommen wird, iteriert das Verfahren auf einem festen Block B_2'. Da in $(B_2')^T B_2'$ mindestens ein Außendiagonalelement gegen Null konvergiert (wegen der Konvergenz des QR-Verfahrens), muß irgendwann ein f_i' oder ein d_i' die Abbruchschranke unterschreiten. Da das nur endlich oft passieren kann, ist irgendwann $B' = B_3'$ diagonal und damit die SVD-Zerlegung der ursprünglichen Bidiagonalmatrix hergestellt.

4 Polvorgabe

Wir betrachten das Kontrollsystem

$$\dot{x} = Ax + Bu\,, \quad A \in \mathbb{R}^{(n,n)}\,, \quad B \in \mathbb{R}^{(n,m)}\,. \tag{4.1}$$

Das Problem der Polvorgabe besteht darin, zu n gegebenen komplexen Zahlen $\lambda_1, \ldots, \lambda_n$ eine Matrix $F \in \mathbb{R}^{(m,n)}$ zu finden, so daß die Systemmatrix $A+BF$ des rückgekoppelten Systems diese Zahlen als Eigenwerte hat. Der Polverschiebungssatz besagt: Ist (A, B) steuerbar, so gibt es zu jedem Polynom

$$p(x) = x^n - \sum_{k=1}^{n} c_k x^{k-1}\,, \quad c_k \in \mathbb{R}\,, \tag{4.2}$$

eine Rückkopplungsmatrix $F \in \mathbb{R}^{(m,n)}$ mit

$$\chi_{A+BF} = p\,. \tag{4.3}$$

Wir behandeln zunächst skalare Kontrollen, d.h. $m = 1$.

Bemerkung 4.1 (Änderung eines einzelnen Eigenwerts)
Sei (A, b) steuerbar, sei $A \in \mathbb{R}^{(n,n)}$ in reeller Schur-Form mit einem Eigenwert $\alpha \in \mathbb{R}$ rechts unten, d.h.

$$A = \begin{pmatrix} \hat{A} & * \\ 0 & \alpha \end{pmatrix}\,, \quad b = \begin{pmatrix} * \\ \beta \end{pmatrix}\,, \quad \hat{A} \in \mathbb{R}^{(n-1,n-1)}\,, \quad \alpha, \beta \in \mathbb{R}\,. \tag{4.4}$$

Da (A, b) steuerbar ist, muß $\beta \neq 0$ sein. Wir setzen

$$f = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \eta \end{pmatrix}\,, \quad \eta = \frac{\lambda - \alpha}{\beta}\,. \tag{4.5}$$

Dann ist

$$A + bf^T = \begin{pmatrix} \hat{A} & * \\ 0 & \lambda \end{pmatrix}\,, \tag{4.6}$$

d.h. der Eigenwert α ist zu λ abgeändert, aber alle anderen Eigenwerte sind unverändert, und die Schur-Form bleibt erhalten.

Bemerkung 4.2 (Änderung eines Eigenwertpaares)
Sei (A, b) steuerbar, sei $A \in \mathbb{R}^{(n,n)}$ in reeller Schur-Form. Wir partitionieren

$$A = \begin{pmatrix} \hat{A} & & * \\ & \tilde{\alpha}_{11} & \tilde{\alpha}_{12} \\ 0 & \tilde{\alpha}_{21} & \tilde{\alpha}_{22} \end{pmatrix}\,, \quad b = \begin{pmatrix} * \\ \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix}\,, \quad \hat{A} \in \mathbb{R}^{(n-2,n-2)}\,. \tag{4.7}$$

Ziel ist, rechts unten die Eigenwerte $\lambda_{1,2} = r \pm is$ zu bekommen. Sei Q eine Givens-Rotation im $\mathbb{R}^2$ mit

$$Q \begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix} = \begin{pmatrix} \beta \\ 0 \end{pmatrix}\,, \tag{4.8}$$

sei

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} = Q \begin{pmatrix} \tilde{\alpha}_{11} & \tilde{\alpha}_{12} \\ \tilde{\alpha}_{21} & \tilde{\alpha}_{22} \end{pmatrix} Q^T . \tag{4.9}$$

Wir setzen

$$f = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ Q^T \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \end{pmatrix} , \tag{4.10}$$

wobei η_1 , η_2 so definiert werden sollen, daß die Matrix

$$M = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} + \begin{pmatrix} \beta \\ 0 \end{pmatrix} \begin{pmatrix} \eta_1 & \eta_2 \end{pmatrix} \tag{4.11}$$

die Eigenwerte $\lambda_{1,2} = r \pm is$ hat. Es ist dann nämlich

$$A + bf^T = \begin{pmatrix} \hat{A} & * \\ 0 & Q^T M Q \end{pmatrix} , \tag{4.12}$$

d.h. $A + bf^T$ ist wieder in reeller Schur-Form, und nur die zu dem rechten unteren 2×2-Block gehörenden Eigenwerte sind in der gewünschten Form abgeändert worden. Die Formeln für η_1 und η_2 sind

$$\eta_1 = \frac{2r - \alpha_{11} - \alpha_{22}}{\beta} , \quad \eta_2 = \frac{\alpha_{22}}{\alpha_{21}} \eta_1 + \frac{\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21} - r^2 - s^2}{\alpha_{21}\beta} . \tag{4.13}$$

Bemerkung 4.3 (Vertauschen von Eigenwerten entlang der Diagonalen)
(Siehe Stewart: Algorithm 506, ACM Trans. Math. Software 2 (1976), 275 – 280.)
Sei $A \in \mathbb{R}^{(2,2)}$,

$$A = \begin{pmatrix} \lambda_1 & a_{12} \\ 0 & \lambda_2 \end{pmatrix} , \quad \lambda_1 \neq \lambda_2 . \tag{4.14}$$

Der zu λ_2 gehörende Eigenvektor ist

$$x = \begin{pmatrix} a_{12} \\ \lambda_2 - \lambda_1 \end{pmatrix} . \tag{4.15}$$

Sei Q Givens-Rotation mit

$$Q^T x = \begin{pmatrix} * \\ 0 \end{pmatrix} . \tag{4.16}$$

Dann gilt

$$(Q^T A Q) e_1 = Q^T (A(Qe_1)) = Q^T \lambda_2 Q e_1 = \lambda_2 e_1 , \tag{4.17}$$

also

$$Q^T A Q = \begin{pmatrix} \lambda_2 & \tilde{a}_{12} \\ 0 & \lambda_1 \end{pmatrix} , \quad \tilde{a}_{12} = \pm a_{12} , \tag{4.18}$$

da $Q^T AQ$ dieselben Eigenwerte und dieselbe Frobenius-Norm wie A hat. Ist nun $A \in \mathbb{R}^{(n,n)}$,

$$A = \begin{pmatrix} * & * & * \\ 0 & \begin{matrix} a_{kk} \\ 0 \end{matrix} & \begin{matrix} a_{k,k+1} \\ a_{k+1,k+1} \end{matrix} & * \\ 0 & 0 & * \end{pmatrix}, \tag{4.19}$$

so wählen wir Q als Givens-Rotation auf $\operatorname{span}\{e_k, e_{k+1}\}$ und erhalten

$$Q^T AQ = \begin{pmatrix} * & * & * \\ 0 & \begin{matrix} a_{k+1,k+1} \\ 0 \end{matrix} & \begin{matrix} \pm a_{k,k+1} \\ a_{kk} \end{matrix} & * \\ 0 & 0 & * \end{pmatrix}. \tag{4.20}$$

Hier wird die Givens-Rotation also zum Vertauschen zweier Diagonalelemente benutzt (nicht zum Erzeugen einer 0).

Die Behandlung von 2×2-Blöcken stellen wir hier nicht dar.

Algorithmus 4.4 (Polverschiebung in der Schur-Form)
Dieses Verfahren stammt von Varga (IEEE Trans. Automatic Control 26 (1981), 517 - 519). Wir stellen hier nur den Fall $m = 1$ (skalare Kontrolle) dar. Sei ein steuerbares Kontrollsystem (A, b) gegeben.

Schritt 0: Wir transformieren A orthogonalähnlich auf obere reelle Schur-Form A_0 (z.B. mit dem QR-Verfahren),

$$A_0 = Q_0 A Q_0^T, \quad b_0 = Q_0 b, \tag{4.21}$$

und initialisieren

$$f_0 = 0, \quad P_0 = I. \tag{4.22}$$

Schritt k: Wir wählen eine orthogonale Transformation Q_k so, daß ein zu verändernder 1×1- oder 2×2-Block auf der Diagonale von A_{k-1} in die rechte untere Ecke wandert (mehrfaches Anwenden von Bemerkung 4.3) und setzen

$$\tilde{A}_k = Q_k A_{k-1} Q_k^T, \quad b_k = Q_k b_{k-1}, \quad P_k = Q_k P_{k-1}. \tag{4.23}$$

Dann wählen wir $\tilde{f}_k$ so, daß

$$A_k = \tilde{A}_k + b_k \tilde{f}_k^T \tag{4.24}$$

rechts unten den gewünschten geänderten Eigenwert bzw. 2×2-Block hat und die anderen Eigenwerte von $\tilde{A}_k$ nicht verändert werden (wie in den Bemerkungen 4.1 und 4.2 beschrieben). Wir akkumulieren die Rückkopplungsvektoren für A_0 mit

$$f_k = f_{k-1} + P_k^T \tilde{f}_k. \tag{4.25}$$

Wir hören auf, wenn alle Eigenwerte von A_k die gewünschte Größe haben. Da wir im folgenden Lemma zeigen, daß

$$A_0 + b_0 f_k^T = P_k^T A_k P_k \,, \tag{4.26}$$

haben wir erreicht, daß $A_0 + b_0 f_k^T$ die gewünschten Eigenwerte hat.

Lemma 4.5 *Für die in Algorithmus 4.4 definierten Größen gilt*

$$A_0 + b_0 f_k^T = P_k^T A_k P_k \,. \tag{4.27}$$

Beweis: Mit Induktion über k. Klar für $k = 0$. Wir führen den Induktionsschritt von $k-1$ nach k durch. Es gilt

$$\begin{aligned} P_k^T A_k P_k &= P_k^T (\tilde{A}_k + b_k \tilde{f}_k^T) P_k \\ &= P_k^T Q_k A_{k-1} Q_k^T P_k + P_k^T b_k \tilde{f}_k^T P_k \\ &= P_{k-1}^T A_{k-1} P_{k-1} + b_0 (\tilde{f}_k^T - \tilde{f}_{k-1}^T) \,, \qquad \text{da } b_k = P_k b_0 \,, \\ &= A_0 + b_0 \tilde{f}_{k-1}^T + b_0 \tilde{f}_k^T - b_0 \tilde{f}_{k-1}^T \,, \qquad \text{nach Induktionsvoraussetzung} \,, \\ &= A_0 + b_0 f_k \,. \end{aligned} \tag{4.28}$$

□

Wir betrachten nun den Fall einer vektorwertigen Kontrolle, d.h. $m > 1$. Die Rückkopplungsmatrix F hat nm Elemente. Schreibt man die Eigenwerte von $A + BF$ vor, so sind dies n Bedingungen. Die verbleibenden Freiheitsgrade in F kann man nutzen, um weitere wünschenswerte Eigenschaften des rückgekoppelten Systems herzustellen. Eine Möglichkeit ist es, die Sensitivität der Eigenwerte hinsichtlich Veränderungen der Systemmatrix möglichst klein zu machen ("robuste Polvorgabe").

Bemerkung 4.6 (Sensitivität von Eigenwerten, motivierende Rechnung)
Sei $A \in \mathbb{C}^{(n,n)}$, sei $\lambda \in \mathbb{C}$ ein einfacher Eigenwert von A, sei

$$Ax = \lambda x \,, \quad x \in \mathbb{C}^n \,,\ x \neq 0 \,. \tag{4.29}$$

Wir betrachten die Eigenwertgleichung für die gestörte Matrix $A + \varepsilon H$, wobei $H \in \mathbb{C}^{(n,n)}$ mit $\|H\|_2 = 1$ sein soll, also

$$(A + \varepsilon H) x(\varepsilon) = \lambda(\varepsilon) x(\varepsilon) \,, \quad x(0) = x \,,\ \lambda(0) = \lambda \,. \tag{4.30}$$

Man kann mit dem Satz über implizite Funktionen beweisen, daß die Funktion $\varepsilon \mapsto \lambda(\varepsilon)$ differenzierbar ist (ebenso $x(\varepsilon)$, falls man auf 1 normiert). Differenzieren ergibt

$$(A + \varepsilon H) x'(\varepsilon) + H x(\varepsilon) = \lambda'(\varepsilon) x(\varepsilon) + \lambda(\varepsilon) x'(\varepsilon) \,, \tag{4.31}$$

also

$$A x'(0) + H x = \lambda'(0) x + \lambda x'(0) \,. \tag{4.32}$$

Sei y^H Linkseigenvektor von A, also $y^H A = \lambda y^H$, dann gilt

$$y^H A x'(0) + y^H H x = y^H \lambda'(0) x + y^H \lambda x'(0) \,, \tag{4.33}$$

also

$$|\lambda'(0)| = \frac{|y^H H x|}{|y^H x|} \leq \frac{\|y^H\|_2 \|x\|_2}{|y^H x|} = \frac{1}{|\cos \angle(y^H, x)|} \,. \tag{4.34}$$

Definieren wir

$$s(\lambda) = \frac{\|y^H\|_2 \|x\|_2}{|y^H x|} \tag{4.35}$$

als die **Sensitivität** des Eigenwerts λ, so gilt also (für einfache Eigenwerte)

$$|\lambda(\varepsilon) - \lambda| \leq \varepsilon s(\lambda) + o(\varepsilon) \,. \tag{4.36}$$

Beispiel 4.7 (Sensitivität eines doppelten Eigenwerts)
Sei

$$A = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \,, \quad H = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \,. \tag{4.37}$$

Es ist

$$\chi_{A+\varepsilon H}(\lambda) = (1-\lambda)^2 - \varepsilon a \,, \quad \lambda_{1,2}(\varepsilon) = 1 \pm \sqrt{\varepsilon a} \,, \tag{4.38}$$

also

$$|\lambda(\varepsilon) - \lambda| = O(\sqrt{\varepsilon}) \,. \tag{4.39}$$

Der folgende Satz zeigt, daß die Situation (4.39) nur bei nichtdiagonalisierbaren Matrizen auftreten kann.

Satz 4.8 (Sensitivität der Eigenwerte diagonalisierbarer Matrizen)
Sei $A \in \mathbb{C}^{(n,n)}$ *diagonalisierbar,*

$$A = XDX^{-1} \,, \quad D \text{ diagonal} \,. \tag{4.40}$$

Dann gilt für jeden Eigenwert $\lambda(\varepsilon)$ *von* $A + \varepsilon H$, *wobei* $H \in \mathbb{C}^{(n,n)}$ *mit* $\|H\|_2 = 1$,

$$\min_{\mu \text{ EW von } A} |\lambda(\varepsilon) - \mu| \leq \varepsilon \, \mathrm{cond}_2(X) \,, \tag{4.41}$$

wobei

$$\mathrm{cond}_2(X) = \|X\|_2 \|X^{-1}\|_2 \,. \tag{4.42}$$

Beweis: Siehe etwa Golub/van Loan, S. 200. □

Bemerkung 4.9 (Konsequenzen für robuste Polvorgabe)
Wegen Satz 4.8 ist es offenbar sinnvoll, die Rückkopplungsmatrix F so zu wählen, daß $A+BF$ diagonalisierbar ist (das ist allerdings nicht immer möglich, siehe unten). Darüber hinaus ist es günstig, die Transformationsmatrix X mit

$$(A + BF)X = XD \,, \quad D = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n) \,, \tag{4.43}$$

so zu wählen, daß $\mathrm{cond}_2(X)$ möglichst klein wird (optimal wäre $\mathrm{cond}_2(X) = 1$, d.h. X orthogonal, das ist aber nur möglich, wenn $A + BF$ normal ist). Man beachte, daß in der i-ten Spalte von X der Eigenvektor von $A + BF$ zum Eigenwert λ_i steht.

Satz 4.10 *Sei* (A, B) *ein Kontrollsystem mit* $A \in \mathbb{R}^{(n,n)}$, $B \in \mathbb{R}^{(n,m)}$ *und* $m \leq n$, $\operatorname{rang}(B) = m$. *Sei*

$$B = Q \begin{pmatrix} R \\ 0 \end{pmatrix} = \begin{pmatrix} Q_0 & Q_1 \end{pmatrix} \begin{pmatrix} R \\ 0 \end{pmatrix}, \tag{4.44}$$

eine reelle QR-Zerlegung von B. *Seien* $X, D \in \mathbb{C}^{(n,n)}$, X *invertierbar. Dann gibt es eine Lösung* $F \in \mathbb{C}^{(m,n)}$ *von*

$$(A + BF)X = XD \tag{4.45}$$

genau dann, wenn

$$Q_1^T(AX - XD) = 0\,. \tag{4.46}$$

In diesem Fall ist F *gegeben durch*

$$F = R^{-1}Q_0^T(XDX^{-1} - A)\,. \tag{4.47}$$

Ist XDX^{-1} *reell, so ist auch* F *reell.*

Beweis: R^{-1} existiert, da $\operatorname{rang}(B) = m$. Es gilt

$$\begin{aligned}(A + BF)X = XD \;&\Leftrightarrow\; BF = XDX^{-1} - A \;\Leftrightarrow\; Q^TBF = Q^T(XDX^{-1} - A) \\ &\Leftrightarrow \begin{cases} RF = Q_0^T(XDX^{-1} - A) \\ 0 = Q_1^T(XDX^{-1} - A) \end{cases} \\ &\Leftrightarrow \begin{cases} F = R^{-1}Q_0^T(XDX^{-1} - A) \\ 0 = Q_1^T(XD - AX) \end{cases} \quad .\end{aligned} \tag{4.48}$$

□

Satz 4.11 *Sei* (A, B) *ein Kontrollsystem, welches die Voraussetzungen von 4.10 erfüllt, sei* $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, *sei*

$$S(\lambda_j) = \ker Q_1^T(A - \lambda_j I)\,. \tag{4.49}$$

Dann ist

$$\dim S(\lambda_j) = m + k_j\,, \quad k_j = \dim \ker \begin{pmatrix} B & A - \lambda_j I \end{pmatrix}^T, \tag{4.50}$$

Sei $\{x_1, \ldots, x_n\}$ *eine Basis von* $\mathbb{C}^n$, *sei* X *die Matrix mit den Spalten* x_j. *Dann existiert eine Lösung* $F \in \mathbb{C}^{(m,n)}$ *von (4.45) existiert genau dann, wenn* $x_j \in S(\lambda_j)$ *für alle* j.

Beweis: Wegen (4.46) ist nur (4.50) zu zeigen. Nach Definition von k_j gilt

$$\operatorname{rang} \begin{pmatrix} B & A - \lambda_j I \end{pmatrix} = n - k_j\,, \tag{4.51}$$

und es ist

$$Q^T \begin{pmatrix} B & A - \lambda_j I \end{pmatrix} = \begin{pmatrix} R & Q_0^T(A - \lambda_j I) \\ 0 & Q_1^T(A - \lambda_j I) \end{pmatrix}, \tag{4.52}$$

also

$$\operatorname{rang}(Q_1^T(A - \lambda_j I)) = n - k_j - m\,, \tag{4.53}$$

$$\dim S(\lambda_j) = n - \operatorname{rang}(Q_1^T(A - \lambda_j I)) = m + k_j\,. \tag{4.54}$$

□

Bemerkung 4.12
Durch Satz 4.11 wird die Berechnung von F auf die Berechnung einer Basis $\{x_1, \ldots, x_n\}$ mit $x_j \in S(\lambda_j)$ zurückgeführt, welche nun eine möglichst kleine Kondition haben soll. Wir untersuchen nicht im Detail, wann eine solche Basis existiert. Ist (A, B) steuerbar, so ist nach dem Hautus-Kriterium $k_j = 0$ für alle j, also $\dim S(\lambda_j) = m$. Jeder Wert λ_j darf also höchstens m mal in D auftreten.

Der folgende Satz liefert einen weiteren Grund, weshalb es günstig ist, die Kondition von X zu minimieren.

Satz 4.13 *Sei (A, B) ein Kontrollsystem, welches die Voraussetzungen in Satz 4.10 erfüllt, sei F eine Rückkopplungsmatrix, so daß $A + BF$ diagonalisierbar ist. Dann gilt*

$$\|F\|_2 \leq \frac{1}{\sigma_m(B)}(\|A\|_2 + \operatorname{cond}_2(X) \max_j\{|\lambda_j|\}), \tag{4.55}$$

und für jede Lösung des rückgekoppelten Systems gilt

$$\dot{x} = Ax + Bu, \quad u = Fx, \quad x(0) = x_0, \tag{4.56}$$

die Abschätzung

$$\|x(t)\|_2 \leq \operatorname{cond}_2(X) \cdot \max_j\{|e^{\lambda_j t}|\} \cdot \|x_0\|_2, \tag{4.57}$$

wobei $\lambda_1, \ldots, \lambda_n$ die Eigenwerte von $A + BF$ sind und X die Transformationsmatrix ist.

Beweis: Aus (4.47) folgt

$$\begin{aligned} \|F\|_2 &\leq \|R^{-1}\|_2(\|A\|_2 + \|X\|_2\|D\|_2\|X^{-1}\|_2) \\ &\leq \frac{1}{\sigma_m(B)}(\|A\|_2 + \operatorname{cond}_2(X) \max_j\{|\lambda_j|\}). \end{aligned} \tag{4.58}$$

Außerdem gilt

$$x(t) = e^{(A+BF)t}x_0 = Xe^{Dt}X^{-1}x_0. \tag{4.59}$$

□

Algorithmus 4.14 (Verfahren zur robusten Polvorgabe)
Dieses Verfahren stammt aus der Arbeit von Kautsky, Nichols, van Dooren (Int. J. Control 41 (1985), 1129 - 1155). Gegeben sind das Kontrollsystem (A, B) und gewünschte Eigenwerte $\lambda_1, \ldots, \lambda_n$ für das rückgekoppelte System.

1. Wir berechnen eine QR-Zerlegung

$$B = Q\begin{pmatrix} R \\ 0 \end{pmatrix} = \begin{pmatrix} Q_0 & Q_1 \end{pmatrix}\begin{pmatrix} R \\ 0 \end{pmatrix}, \tag{4.60}$$

   etwa mit dem Householder-Verfahren.

2. Wir berechnen eine ON-Basis für

$$S(\lambda_j) = \ker Q_1^T(A - \lambda_j I), \quad 1 \leq j \leq n, \tag{4.61}$$

   etwa durch Singulärwertzerlegung.

3. Wir berechnen $x_j \in S(\lambda_j)$ mit $\|x_j\|_2 = 1$, so daß

$$X = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix} \tag{4.62}$$

eine möglichst kleine Kondition hat. (Siehe unten.)

4. Wir berechnen $M = XDX^{-1}$, $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, durch Lösen der n linearen Gleichungssysteme

$$X^T M^T = DX^T . \tag{4.63}$$

(Kein Problem, wenn X gut konditioniert ist.)

5. Wir berechnen die Rückkopplungsmatrix F als Lösung der m linearen Gleichungssysteme

$$RF = Q_0^T (M - A) . \tag{4.64}$$

Sind die λ_j alle reell, so können wir vollständig im Reellen rechnen.

Algorithmus 4.15 (Realisierung von Schritt 3 in Algorithmus 4.14)
Wir beschreiben ein einfaches Verfahren, was allerdings nicht immer funktioniert. Wir nehmen an, daß es eine Basis $\{x_1^0, \ldots, x_n^0\}$ mit $x_j^0 \in S(\lambda_j)$ gibt. Dann gibt es auch eine solche, deren Elemente aus den in Schritt 2 berechneten ON-Basen stammen. Sei $X^{(0)}$ die daraus gebildete Matrix. Wir führen eine zyklische Iteration über die Spalten durch. Sei X mit den Spalten $x_1, \ldots, x_n$ die aktuelle Matrix und sei die Spalte j an der Reihe. Sei X_j die Matrix, die aus X durch Streichen der Spalte j entsteht. Wir bestimmen die neue Spalte $\tilde{x}_j$ so, daß $\tilde{x}_j \in S(\lambda_j)$ und der Winkel zwischen $\tilde{x}_j$ und dem Unterraum

$$U_j = \operatorname{im} X_j = \operatorname{span}\{x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n\} \tag{4.65}$$

maximiert wird. Hierzu bestimmen wir einen Vektor y mit $\|y\|_2 = 1$ und $y \perp U_j$. Wir erhalten y als letzte Spalte von $\tilde{Q}$ in der QR-Zerlegung von X_j,

$$X_j = \tilde{Q} \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix} . \tag{4.66}$$

Der gesuchte Vektor $\tilde{x}_j$ ist nun die Projektion von y auf $S(\lambda_j)$. (Die Projektion minimiert den Winkel zu y.) Sie wird berechnet aus der Formel

$$\tilde{x}_j = Z_j Z_j^T y , \tag{4.67}$$

wobei die Spalten von Z_j aus der ON-Basis von $S(\lambda_j)$ gebildet werden (Übung).

Zuverlässigere (aber kompliziertere) Verfahren finden sich in der oben zitierten Originalarbeit.

5 Sylvester-Gleichung, Ljapunov-Gleichung

Problem 5.1 (Sylvester-Gleichung und Ljapunov-Gleichung)
Die Sylvester-Gleichung hat die Form

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

wobei $A \in \mathbb{R}^{(n,n)}$, $B \in \mathbb{R}^{(m,m)}$, $C \in \mathbb{R}^{(n,m)}$ gegeben und $X \in \mathbb{R}^{(n,m)}$ gesucht sind. Im Spezialfall $B = A^T$, also $n = m$ und

$$AX + XA^T = C\,, \tag{5.2}$$

sprechen wir von der Ljapunov-Gleichung.

Satz 5.2 *Seien $A \in \mathbb{K}^{(n,n)}$, $B \in \mathbb{K}^{(m,m)}$ gegeben, $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$. Dann sind äquivalent:*

(i) *Die Sylvester-Gleichung*

$$AX - XB = C\,, \tag{5.3}$$

hat für jedes $C \in \mathbb{K}^{(n,m)}$ eine eindeutige Lösung $X \in \mathbb{K}^{(n,m)}$.

(ii) *A und B haben keinen gemeinsamen Eigenwert.*

Beweis: Die Aussage (i) gilt genau dann, wenn die homogene Matrixgleichung

$$AX - XB = 0\,, \tag{5.4}$$

nur die Lösung $X = 0$ hat. Sei zunächst $\lambda \in \mathbb{C}$ ein gemeinsamer Eigenwert von A und B. Dann gibt es $x \in \mathbb{C}^n$ und $y \in \mathbb{C}^m$ mit

$$Ax = \lambda x\,, \quad y^T B = \lambda y^T\,, \quad x \neq 0\,, \quad y \neq 0\,. \tag{5.5}$$

Es ist dann $X = xy^T \neq 0$ eine Lösung von (5.4). Falls $\mathbb{K} = \mathbb{R}$, so sind $\operatorname{Re} X$ und $\operatorname{Im} X$ Lösungen von (5.4), und mindestens eine der beiden Matrizen ist von Null verschieden. Damit ist "(i)⇒(ii)" bewiesen. Wir nehmen nun an, daß A und B keinen gemeinsamen Eigenwert haben. Die charakteristischen Polynome χ_A und χ_B sind dann teilerfremd. Aus einem Satz der Algebra folgt, daß es Polynome φ und ψ gibt mit

$$1 = \varphi\chi_A + \psi\chi_B\,. \tag{5.6}$$

Setzen wir als Argument die Matrix A ein, so ergibt sich

$$I = \varphi(A)\chi_A(A) + \psi(A)\chi_B(A) = \psi(A)\chi_B(A)\,, \tag{5.7}$$

da $\chi_A(A) = 0$ nach Hamilton-Cayley. Sei nun X eine Lösung von (5.4). Es gilt dann

$$A^k X = XB^k\,, \quad k \in \mathbb{N}\,, \tag{5.8}$$

was man per Induktion mit dem Induktionsschluß

$$A^k X = A(A^{k-1}X) = A(XB^{k-1}) = XB^k \tag{5.9}$$

beweist. Damit gilt

$$p(A)X = Xp(B) \tag{5.10}$$

für jedes Polynom p. Aus (5.7) folgt nun

$$X = \psi(A)\chi_B(A)X = \psi(A)X\chi_B(B) = 0\,, \tag{5.11}$$

wieder nach Hamilton-Cayley. □

Bemerkung 5.3 (Orthogonale Transformation der Sylvester-Gleichung)
Seien $A \in \mathbb{R}^{(n,n)}$, $B \in \mathbb{R}^{(m,m)}$, $C \in \mathbb{R}^{(n,m)}$ gegeben. Seien $U \in \mathbb{R}^{(n,n)}$, $V \in \mathbb{R}^{(m,m)}$ orthogonal. Wir setzen

$$\tilde{A} = U^T A U\,, \quad \tilde{B} = V^T B V\,, \quad \tilde{C} = U^T C V\,. \tag{5.12}$$

Die Sylvester-Gleichung

$$AX + XB = C \tag{5.13}$$

ist dann äquivalent zu

$$(U^T A U)(U^T X V) + (U^T X V)(V^T B V) = U^T C V\,, \tag{5.14}$$

also zu

$$\tilde{A}\tilde{X} + \tilde{X}\tilde{B} = \tilde{C}\,, \tag{5.15}$$

d.h. X ist Lösung von (5.13) genau dann, wenn

$$\tilde{X} = U^T X V \tag{5.16}$$

Lösung von (5.15) ist.

Algorithmus 5.4 (Hessenberg-Schur-Verfahren für die Sylvester-Gleichung)
Dieses Verfahren stammt von Golub, Nash, van Loan (IEEE Trans. Automatic Control 24 (1979), 909 - 913). Es stellt eine Verbesserung des Verfahrens von Bartels und Stewart (Comm. ACM 15 (1972), 820 - 826) dar. Die Idee besteht in beiden Fällen darin, durch orthogonale Ähnlichkeitstransformation die Matrizen A und B auf einfachere Gestalt zu bringen. Das Verfahren zur Lösung von

$$AX + XB = C\,, \quad A \in \mathbb{R}^{(n,n)}\,,\ B \in \mathbb{R}^{(m,m)}\,,\ C \in \mathbb{R}^{(n,m)}\,, \tag{5.17}$$

sieht folgendermaßen aus:

1. Bringe A auf obere Hessenberg-Form H,

   $$H = U^T A U\,, \quad U \text{ orthogonal}\ , \tag{5.18}$$

   etwa mit Householder-Transformationen.

2. Bringe B^T auf obere reelle Schur-Form S,

   $$S = V B^T V^T\,, \quad V \text{ orthogonal}\ , \tag{5.19}$$

   etwa mit dem QR-Verfahren.

3. Löse die Sylvester-Gleichung in Hessenberg-Schur-Form

   $$HY + Y S^T = F\,, \quad F = U^T C V\,, \tag{5.20}$$

4. Berechne

   $$X = U Y V^T\,. \tag{5.21}$$

Die Lösung $Y \in \mathbb{R}^{(n,m)}$ von Gleichung (5.20) wird spaltenweise von hinten nach vorn berechnet. Seien die Spalten $Y_m, \ldots, Y_{k+1}$ bereits berechnet. Man unterscheidet zwei Fälle:

1. Es ist $s_{k,k-1} = 0$. Die k-te Spalte von (5.20) hat die Form

$$HY_k + YS_k^T = F_k\,, \quad YS_k^T = \sum_{i=k}^{m} s_{ki} Y_i\,, \tag{5.22}$$

   also ist Y_k Lösung von

$$(H + s_{kk} I) Y_k = F_k - \sum_{i=k+1}^{m} s_{ki} Y_i\,. \tag{5.23}$$

2. Es ist $s_{k,k-1} \neq 0$. Dann ist $s_{k-1,k-2} = 0$, da S in reeller Schur-Form ist. Die beiden Spalten Y_k und Y_{k-1} werden gemeinsam berechnet aus

$$HY_k + YS_k^T = F_k\,, \quad HY_{k-1} + YS_{k-1}^T = F_{k-1}\,, \tag{5.24}$$

   also als Lösung des Gleichungssystems im $\mathbb{R}^{2n}$

$$(H + s_{kk} I) Y_k + s_{k,k-1} Y_{k-1} = F_k - \sum_{i=k+1}^{m} s_{ki} Y_i\,, \tag{5.25}$$

$$s_{k-1,k} Y_k + (H + s_{k-1,k-1} I) Y_{k-1} = F_{k-1} - \sum_{i=k+1}^{m} s_{k-1,i} Y_i\,. \tag{5.26}$$

   Numeriert man die Unbekannten in der Reihenfolge $y_{1k}, y_{1,k-1}, y_{2k}, y_{2,k-1}, \ldots$ und permutiert man die Gleichungen ebenso, so hat die Matrix des Gleichungssystems (5.25), (5.26) die Form

$$\begin{pmatrix} h_{11} + s_{kk} & s_{k,k-1} & * & \cdots \\ s_{k-1,k} & h_{11} + s_{k-1,k-1} & * & \cdots \\ h_{21} & 0 & h_{22} + s_{kk} & \cdots \\ 0 & h_{21} & s_{k-1,k} & \cdots \\ 0 & 0 & h_{32} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{5.27}$$

   d.h. unterhalb der Diagonale können nur die ersten beiden Nebendiagonalen von Null verschieden sein.

In beiden Fällen sind pro Spalte nur $O(n^2)$ Operationen erforderlich. Schätzt man den Aufwand des QR-Verfahrens zur Berechnung der Schur-Form S als $10m^3$, so beträgt der Gesamtaufwand (nur Terme der höchsten Ordnung)

$$\frac{5}{3} n^3 + 5n^2 m + \frac{5}{2} m^2 n + 10 m^3\,. \tag{5.28}$$

Falls $m > n$, ist es daher günstiger, die Sylvester-Gleichung in der transponierten Form

$$B^T X^T + X^T A^T = C^T \tag{5.29}$$

zu lösen, d.h. man bringt immer die kleinere der beiden Matrizen A und B auf Schur-Form und die größere auf Hessenberg-Form. Das ursprüngliche Verfahren von Bartels-Stewart transformiert sowohl A als auch B auf Schur-Form und ist insgesamt aufwendiger.

Bemerkung 5.5
Die einfache Anwendung eines Standardverfahrens zur Lösung linearer Gleichungssysteme auf

$$AX + XB = C \tag{5.30}$$

ist in der Regel aufwendiger als das Verfahren 5.4, da die Matrix des resultierenden Gleichungssystems $nm(n+m)$ zu berücksichtigende Elemente aufweist.

Satz 5.6 *Seien* $A, B \in \mathbb{R}^{(n,n)}$ *Stabilitätsmatrizen. Dann hat die Sylvester-Gleichung*

$$AX + XB = C \tag{5.31}$$

für jedes $C \in \mathbb{R}^{(n,n)}$ *die eindeutige Lösung*

$$X = -\int_0^\infty e^{tA} C e^{tB}\, dt\,. \tag{5.32}$$

Beweis: Die eindeutige Lösbarkeit folgt aus Satz 5.2, da die Eigenwerte von A negativen und die von $-B$ positiven Realteil haben. Es gilt

$$\frac{\mathrm{d}}{\mathrm{dt}}\left(e^{tA} C e^{tB}\right) = A e^{tA} C e^{tB} + e^{tA} C e^{tB} B\,, \tag{5.33}$$

$$\int_0^t A e^{sA} C e^{sB} + e^{sA} C e^{sB} B\, ds = e^{tA} C e^{tB} - C\,. \tag{5.34}$$

Wegen

$$\|e^{sA}\| \le e^{-\alpha s}\,, \quad \|e^{sB}\| \le e^{-\beta s}\,, \tag{5.35}$$

für geeignete $\alpha, \beta > 0$ ist der Grenzübergang $t \to \infty$ in (5.34) möglich, und

$$A \int_0^\infty e^{sA} C e^{sB}\, ds + \int_0^\infty e^{sA} C e^{sB}\, ds\, B = -C\,, \tag{5.36}$$

woraus die Behauptung folgt. □

Die Ljapunov-Gleichung hängt zusammen mit dem Begriff der Ljapunov-Funktion. Letztere geht auf Ljapunov zurück und erlaubt es, Stabilität und asymptotische Stabilität des nichtlinearen Systems

$$\dot{x} = f(x)\,, \quad f : \mathbb{R}^n \to \mathbb{R}^n \tag{5.37}$$

zu charakterisieren.

Definition 5.7 (Ljapunov-Funktion)
Sei $x_* \in \mathbb{R}^n$, U Umgebung von x_*, $f : U \to \mathbb{R}^n$ mit $f(x_*) = 0$. Eine Funktion $V : U \to \mathbb{R}$ heißt *Ljapunov-Funktion* für f, wenn gilt:

(i) V ist stetig in U und stetig differenzierbar in $U \setminus \{x_*\}$,

(ii) $V(x_*) = 0$, $V(x) > 0$ für alle $x \in U \setminus \{x_*\}$,

(iii) $(\operatorname{grad} V(x))^T f(x) \le 0$ für alle $x \in U \setminus \{x_*\}$.

V heißt *strikte Ljapunov-Funktion* für f wenn statt (iii) gilt

(iii') $(\operatorname{grad} V(x))^T f(x) < 0$ für alle $x \in U \setminus \{x_*\}$.

Bemerkung 5.8
Der Sinn der Definition von V in 5.7 besteht darin, daß entlang von Trajektorien von $\dot{x} = f(x)$ gilt

$$\frac{\mathrm{d}}{\mathrm{dt}} V(x(t)) = (\operatorname{grad} V(x(t)))^T f(x(t)) \leq (<)\, 0\,. \tag{5.38}$$

Satz 5.9 (Ljapunov-Funktionen und Stabilität)
Sei $x_* \in \mathbb{R}^n$, U Umgebung von x_*, $f : U \to \mathbb{R}^n$ lokal Lipschitz-stetig mit $f(x_*) = 0$. Sei V Ljapunov-Funktion für f. Dann ist x_* stabiler Gleichgewichtspunkt von

$$\dot{x} = f(x)\,. \tag{5.39}$$

Darüber hinaus ist x_* asymptotisch stabiler Gleichgewichtspunkt genau dann, wenn es eine Umgebung U_* von x_* gibt, so daß gilt: Die einzige Lösung von (5.39) mit Anfangswert in U_*, entlang derer V konstant ist, ist die Funktion $x(t) \equiv x_*$. Insbesondere ist x_* asymptotisch stabil, falls V strikte Ljapunov-Funktion ist.

Beweis: Sei $\varepsilon > 0$ beliebig, aber so klein, daß

$$B_\varepsilon = \{x : \|x - x_*\|_2 \leq \varepsilon\} \subset U\,, \tag{5.40}$$

sei

$$\alpha = \min\{V(x) : \|x - x_*\|_2 = \varepsilon\}\,, \qquad \text{also } \alpha > 0\,, \tag{5.41}$$

$$U_\alpha = \{x : V(x) < \alpha\} \cap \operatorname{int}(B_\varepsilon)\,. \tag{5.42}$$

Sei $x_0 \in U_\alpha$ beliebig, sei $x : I \to \mathbb{R}^n$ die Lösung von (5.39) zum Anfangswert $x(0) = x_0$ mit maximalem Definitionsintervall $I \subset \mathbb{R}_+$. Für $t \in I$ gilt

$$0 \leq V(x(t)) \leq V(x_0) < \alpha\,, \tag{5.43}$$

also ist $\|x(t) - x_*\|_2 = \varepsilon$ für kein $t \in I$, also gilt $\|x(t) - x_*\|_2 < \varepsilon$ für alle $t \in I$. Es folgen $I = \mathbb{R}_+$ und die Stabilität von x_*. Sei nun x_* asymptotisch stabil, sei U_* der Einzugsbereich von x_*. Für jede Lösung von (5.39) mit Anfangswert in U_* gilt

$$\lim_{t\to\infty} V(x(t)) = V(\lim_{t\to\infty} x(t)) = V(x_*) = 0\,. \tag{5.44}$$

Ist V konstant entlang x, so ist diese Konstante also 0 und damit $x(t) = x_*$ für alle t. Sei umgekehrt $x(t) = x_*$ die einzige Lösung von (5.39) mit Anfangswert in U_*, entlang derer V konstant ist. Durch Übergang von U zu $U \cap U_*$ können wir o.B.d.A. annehmen, daß $U_* = U$ gilt. Wir wollen zeigen, daß U_α zum Einzugsbereich von x_* gehört. Sei dazu $x_0 \in U_\alpha$ und $x : \mathbb{R}_+ \to \mathbb{R}^n$ die Lösung des zugehörigen Anfangswertproblems. Wir zeigen, daß die Annahme

$$\text{es gibt eine Folge } t_n \text{ mit } t_n \to \infty \text{ und } x(t_n) \not\to x_* \tag{5.45}$$

zum Widerspruch führt. Wegen der Kompaktheit von B_ε können wir weiter annehmen, daß $x(t_n) \to x_{**} \neq x_*$ gilt. Es ist dann $\lim V(x(t_n)) = V(x_{**})$. Sei $\tilde{x} : \mathbb{R}_+ \to \mathbb{R}^n$ die Lösung von (5.39) zum Anfangswert $\tilde{x}(0) = x_{**}$. Nach Voraussetzung gibt es ein $t > 0$

mit $V(\tilde{x}(t)) < V(x_{**})$. Da die Lösung von (5.39) stetig vom Anfangswert abhängt und (5.39) autonom ist, folgt aus $x(t_n) \to x_{**}$, daß

$$\lim_{t\to\infty} x(t_n + t) = \tilde{x}(t)\,, \tag{5.46}$$

also auch der Widerspruch

$$\lim_{t\to\infty} V(x(t_n + t)) = V(\tilde{x}(t)) < V(x_{**}) \leq V(x(\tau))\,, \qquad \text{für alle } \tau \geq 0\,. \tag{5.47}$$

□

Notation 5.10
Sei $A \in \mathbb{R}^{(n,n)}$. Wir schreiben $A > 0$ bzw. $A \geq 0$, falls A positiv definit bzw. positiv semidefinit ist, und $A < 0$ bzw. $A \leq 0$, falls $-A > 0$ bzw. $-A \geq 0$.

Satz 5.11 (Ljapunov-Gleichung und Ljapunov-Stabilität)
Sei $A \in \mathbb{R}^{(n,n)}$. Dann sind äquivalent:

(i) *A ist Stabilitätsmatrix.*

(ii) *Für jedes $C \in \mathbb{R}^{(n,n)}$ hat die Ljapunov-Gleichung*

$$AX + XA^T = -C \tag{5.48}$$

eine eindeutige Lösung $X \in \mathbb{R}^{(n,n)}$. Ist $C > 0$ bzw. C symmetrisch, so ist auch $X > 0$ bzw. X symmetrisch.

(iii) *Es gibt ein $X \in \mathbb{R}^{(n,n)}$ mit $X > 0$ und*

$$AX + XA^T < 0\,. \tag{5.49}$$

(iv) *Es gibt ein $X \in \mathbb{R}^{(n,n)}$ mit $X > 0$, so daß*

$$V(x) = x^T X x \tag{5.50}$$

eine strikte Ljapunov-Funktion für $\dot{x} = A^T x$ ist.

Beweis:

"(i)⇒ (ii)": Folgt aus Satz 5.6.

"(ii)⇒ (iii)": Wähle $C = I$ in (ii).

"(iii)⇒ (iv)": Ist X gemäß (iii) gewählt, und ist $x : \mathbb{R}_+ \to \mathbb{R}^n$ Lösung von $\dot{x} = A^T x$, so gilt

$$\begin{aligned} (\text{grad}\,(V(x(t)))^T \dot{x}(t) &= \frac{\mathrm{d}}{\mathrm{dt}} V(x(t)) = \dot{x}(t)^T X x(t) + x(t)^T X \dot{x}(t) \\ &= x(t)^T A X x(t) + x(t)^T X A^T x(t) < 0\,, \end{aligned} \tag{5.51}$$

falls $x(t) \neq 0$. Da $\dot{x}(t) = A^T x(t)$ und da durch jedes $\xi \in \mathbb{R}^n$ eine Trajektorie des Systems (5.39) verläuft, folgt die Behauptung.

"(iv)⇒ (i)": Folgt aus Satz 5.9.

□

Bemerkung 5.12 (Numerische Lösung der Ljapunov-Gleichung)
Zur Lösung der Ljapunov-Gleichung

$$AX + XA^T = -C\,, \quad A, C \in \mathbb{R}^{(n,n)}\,, \tag{5.52}$$

kann man natürlich das Verfahren 5.4 (bzw. das ursprüngliche Verfahren von Bartels und Stewart) verwenden. Liegt C bereits als Quadrat vor, d.h.

$$C = BB^T\,, \quad B \in \mathbb{R}^{(n,m)}\,, \tag{5.53}$$

so kann man das Verfahren 5.4 so modifizieren, daß es direkt auf dem Faktor B arbeitet, siehe S.J. Hammarling, IMA J. Numer. Anal. 2 (1982), 303 - 323. Eine Alternative stellt das folgende Verfahren dar.

Algorithmus 5.13 (Lösung der Ljapunov-Gleichung: SVD-Schur-Verfahren)
Gesucht ist eine Lösung $X \in \mathbb{R}^{(n,n)}$ von

$$AX + XA^T = -BB^T\,, \quad A \in \mathbb{R}^{(n,n)}\,, B \in \mathbb{R}^{(n,m)}\,, \tag{5.54}$$

wobei A Stabilitätsmatrix ist. Wir wollen X berechnen, ohne BB^T auszurechnen. Sei

$$B = U\Sigma V^T \tag{5.55}$$

eine Singulärwertzerlegung von B, dann ist X Lösung von (5.55) genau dann, wenn $\tilde{X} = U^T X U$ Lösung ist von

$$\tilde{A}\tilde{X} + \tilde{X}\tilde{A}^T = -\Sigma\Sigma^T\,, \quad \tilde{A} = U^T A U\,. \tag{5.56}$$

Es ergibt sich das folgende Verfahren.

1. Berechne die Singulärwertzerlegung

   $$B = U\Sigma V^T\,, \tag{5.57}$$

   setze $\tilde{A} = U^T A U$.

2. Bringe $\tilde{A}$ auf obere reelle Schur-Form

   $$S = Q^T \tilde{A} Q\,, \quad Q \text{ orthogonal }. \tag{5.58}$$

3. Löse

   $$SY + YS^T = -Q^T(\Sigma\Sigma^T)Q \tag{5.59}$$

   mit Algorithmus 5.4 bzw. dem Verfahren von Bartels und Stewart. Hierfür wird in (5.59) die rechte Seite ausmultipliziert, und zwar die innere Klammer zuerst.

4. Berechne

   $$X = UQYQ^T U^T\,. \tag{5.60}$$

6 Balancierte Realisierungen

Das Kontrollsystem (A, B, C) mit $A \in \mathbb{R}^{(n,n)}$, $B \in \mathbb{R}^{(n,m)}$ und $C \in \mathbb{R}^{(k,n)}$,

$$\dot{x} = Ax + Bu\,, \quad y = Cx\,, \tag{6.1}$$

hat die Übertragungsfunktion

$$G(s) = C(sI - A)^{-1}B \tag{6.2}$$

und die Impulsantwortmatrix

$$K(t, s) = Ce^{(t-s)A}B\,. \tag{6.3}$$

Umgekehrt können wir G bzw. K auf verschiedene Weise durch ein Kontrollsystem (A, B, C) realisieren. Aus einem früheren Satz wissen wir, daß eine solche Realisierung minimale Zustandsraumdimension hat genau dann, wenn (A, B) steuerbar und (A, C) beobachtbar ist.

Definition 6.1 (Mc-Millan-Grad)
Sei G eine Übertragungsfunktion. Die Zustandsraumdimension n einer minimalen Realisierung von G heißt der Mc-Millan-Grad von G.

Die minimale Realisierung (A, B, C) einer Übertragungsfunktion G ist nicht eindeutig bestimmt, z.B. ist jedes zu (A, B, C) ähnliche Kontrollsystem ebenfalls eine minimale Realisierung von G. Nicht alle minimalen Realisierungen eignen sich für praktische Zwecke.

Beispiel 6.2
Wir betrachten die Übertragungsfunktion

$$G(s) = \frac{2s + 3}{(s + 1)(s + 2)}\,. \tag{6.4}$$

Sowohl

$$A = \begin{pmatrix} -1 & 0 \\ 0 & -2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 1 \end{pmatrix}, \tag{6.5}$$

als auch

$$\tilde{A} = A\,, \quad \tilde{B} = \begin{pmatrix} 10^6 \\ 10^{-6} \end{pmatrix}, \quad \tilde{C} = \begin{pmatrix} 10^{-6} & 10^6 \end{pmatrix}, \tag{6.6}$$

sind minimale Realisierungen von G. Während es nun gegen (A, B, C) zunächst nichts einzuwenden gibt, liegt $(\tilde{A}, \tilde{B})$ ziemlich nahe an einem nicht steuerbaren und $(\tilde{A}, \tilde{C})$ ziemlich nahe an einem nicht beobachtbaren System (man muß nur 10^{-6} durch 0 ersetzen).

Der Unterschied zwischen (A, B) und $(\tilde{A}, \tilde{B})$ in Beispiel 6.2 kommt zum Vorschein, wenn man die Erreichbarkeitsmenge von (A, B)

$$\mathcal{R}_1(t) = \{\varphi(t, 0, 0, u) : u \in L^2(0, t)\,, \,, \|u\|_2 \leq 1\}\,, \tag{6.7}$$

von (A, B) unter der Kontrolleinschränkung $\|u\|_2 \leq 1$ betrachtet. Für $(\tilde{A}, \tilde{B})$ ist $\mathcal{R}_1(t)$ eine sehr flache Ellipse. Zur Berechnung von $\mathcal{R}_1(t)$ kann man die Gramsche Matrix

$$W_t = \int_0^t e^{sA}BB^T e^{sA^T}\,ds = \int_0^t e^{(t-s)A}BB^T e^{(t-s)A^T}\,ds \tag{6.8}$$

heranziehen.

Satz 6.3 (Charakterisierung von $\mathcal{R}_1(t)$)

$$\mathcal{R}_1(t) = \{W_t z : z \in \mathbb{R}^n\,,\; z^T W_t z \le 1\}\,. \tag{6.9}$$

Beweis: Zunächst bemerken wir, daß für alle $z \in \mathbb{R}^n$ gilt: Definieren wir eine Steuerung $u : (0,t) \to \mathbb{R}^m$ durch

$$u(s) = B^T e^{(t-s)A^T} z\,, \tag{6.10}$$

so gilt

$$W_t z = \int_0^t e^{(t-s)A} B B^T e^{(t-s)A^T}\, ds\, z = \int_0^t e^{(t-s)A} B u(s)\, ds = \varphi(t,0,0,u)\,, \tag{6.11}$$

und

$$z^T W_t z = \int_0^t z^T e^{(t-s)A} B B^T e^{(t-s)A^T} z\, ds = \int_0^t u(s)^T u(s)\, ds = \|u\|_2^2\,. \tag{6.12}$$

Hieraus folgt bereits die Inklusion $''\supset''$. Sei umgekehrt $\tilde{x} \in \mathcal{R}_1(t)$, also

$$\tilde{x} = \varphi(t,0,0,\tilde{u})\,, \quad \|\tilde{u}\|_2 \le 1\,,\; \tilde{u} \in L^2(0,t)\,. \tag{6.13}$$

Da $\operatorname{im} W_t$ der erreichbare Unterraum von (A,B) (ohne Kontrolleinschränkung) ist, gibt es ein $z \in \mathbb{R}^n$ mit

$$\tilde{x} = W_t z\,. \tag{6.14}$$

Wir definieren u gemäß (6.10). Dann gilt

$$\|\tilde{u}\|_2^2 = \|u\|_2^2 + \|\tilde{u} - u\|_2^2 + 2\langle \tilde{u} - u, u\rangle\,, \tag{6.15}$$

und

$$\begin{aligned} \langle u, \tilde{u} - u\rangle &= \int_0^t u(s)^T(\tilde{u}(s) - u(s))\, ds = z^T \int_0^t e^{(t-s)A} B(\tilde{u}(s) - u(s))\, ds \\ &= z^T(\varphi(t,0,0,\tilde{u}) - \varphi(t,0,0,u)) = 0\,. \end{aligned} \tag{6.16}$$

Es folgt

$$z^T W_t z = \|u\|_2^2 \le \|\tilde{u}\|_2^2 \le 1\,, \quad \tilde{x} = W_t z\,. \tag{6.17}$$

Damit ist auch $''\subset''$ bewiesen. □

Bemerkung 6.4 (Minimum-Norm-Kontrolle)
Wie der Beweis von Satz 6.3 zeigt, hat eine gemäß (6.14) und (6.10) konstruierte Steuerung u minimale L^2-Norm im Vergleich zu allen anderen Steuerungen, die dasselbe $\tilde{x}$ zur Zeit t erreichen.

Folgerung 6.5 (Charakterisierung von $\mathcal{R}_1(t)$ über die Schur-Form von W_t)
Sei

$$W_t = V \Sigma^2 V^T\,, \tag{6.18}$$

eine Schur-Form von W_t, d.h. V ist orthogonal und $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$. (Wir nehmen o.B.d.A. $\sigma_i \ge 0$ für alle i an.) Dann gilt

$$\mathcal{R}_1(t) = V \Sigma(B_1)\,, \tag{6.19}$$

wobei B_1 die Einheitskugel im $\mathbb{R}^n$ bezeichnet.

Beweis: Aus Satz 6.3 folgt

$$\begin{aligned}\mathcal{R}_1(t) &= \{W_t z : z \in \mathbb{R}^n\,,\; z^T W_t z \le 1\,\} \\ &= \{V\Sigma^2 V^T z : z \in \mathbb{R}^n\,,\; z^T V\Sigma^2 V^T z \le 1\,\} \\ &= \{V\Sigma x : x \in \mathbb{R}^n\,,\; x^T x \le 1\,\}\,. \end{aligned} \tag{6.20}$$

□

Bemerkung 6.6 (Zusammenhang zur Kondition von W_t)
Die Flachheit des Ellipsoids $\mathcal{R}_1(t)$ wird also durch die Kondition von W_t bezüglich der Spektralnorm,

$$\mathrm{cond}\,(W_t) = \mathrm{cond}\,(\Sigma^2) = \frac{\sigma_1^2}{\sigma_n^2} \tag{6.21}$$

beschrieben. Analoge (bzw. duale) Überlegungen hinsichtlich Beobachtbarkeit lassen sich für die Matrix

$$Y_t = \int_0^t e^{sA^T} C^T C e^{sA}\, ds \tag{6.22}$$

anstellen. Die Idee von B. Moore (IEEE Trans. Automatic Control 26 (1981), 17-32) bestand nun darin, die Schur-Form der Gramschen Matrix heranzuziehen zur Konstruktion einer minimalen Realisierung, welche die Kondition von W_t und von Y_t gegeneinander ausbalanciert.

Lemma 6.7 *Seien die Kontrollsysteme (A, B, C) und $(\tilde{A}, \tilde{B}, \tilde{C})$ ähnlich, d.h. sei*

$$\tilde{A} = T^{-1}AT\,, \quad \tilde{B} = T^{-1}B\,, \quad \tilde{C} = CT\,, \tag{6.23}$$

für ein invertierbares $T \in \mathbb{R}^{(n,n)}$. Für die zugeordneten Gramschen Matrizen gilt dann

$$\tilde{W}_t = T^{-1} W_t T^{-T}\,, \quad \tilde{Y}_t = T^T Y_t T\,. \tag{6.24}$$

Beweis: Einsetzen von (6.23) in die Definitionen. □

Definition 6.8 (Balancierte Realisierung)
Eine minimale Realisierung $(A, B, C) \in \mathcal{S}_{n,m,k}$ einer gegebenen Übertragungsfunktion heißt *balanciert* (zum Zeitpunkt $t > 0$), wenn es eine Diagonalmatrix D gibt mit

$$W_t = Y_t = D\,, \tag{6.25}$$

wobei

$$W_t = \int_0^t e^{sA} B B^T e^{sA^T}\, ds\,, \quad Y_t = \int_0^t e^{sA^T} C^T C e^{sA}\, ds\,. \tag{6.26}$$

Satz 6.9 (Existenz balancierter Realisierungen)
Jede Übertragungsfunktion, welche eine Realisierung hat, hat auch eine (von $t > 0$ abhängige) balancierte Realisierung.

Beweis: Sei (A, B, C) eine minimale Realisierung der Übertragungsfunktion G. Wir betrachten die Schur-Form der Gramschen Matrix W_t,

$$W_t = V_w \Sigma_w^2 V_w^T\,. \tag{6.27}$$

Da die Realisierung minimal ist, sind W_t und Y_t invertierbar. Setzen wir $T = V_w \Sigma_w$, so gilt

$$\tilde{W}_t = T^{-1} W_t T^{-T} = I\,, \tag{6.28}$$

und

$$\tilde{Y}_t = T^T Y_t T\,. \tag{6.29}$$

Sei

$$\tilde{Y}_t = V_y \Sigma^4 V_y^T \tag{6.30}$$

die Schur-Form von $\tilde{Y}_t$. Da mit Y_t auch $\tilde{Y}_t$ invertierbar ist, ist Σ ebenfalls invertierbar. Wir wenden auf das mit T transformierte System eine weitere Ähnlichkeitstransformation an, und zwar

$$\tilde{T} = V_y \Sigma^{-1}\,. \tag{6.31}$$

Es ergibt sich

$$\tilde{\tilde{W}}_t = \tilde{T}^{-1} \tilde{W}_t \tilde{T}^{-T} = \Sigma V_y^T V_y^{TT} \Sigma = \Sigma^2\,, \tag{6.32}$$

$$\tilde{\tilde{Y}}_t = \tilde{T}^T \tilde{Y}_t \tilde{T} = \Sigma^{-1} V_y^T V_y \Sigma^4 V_y^T V_y \Sigma^{-1} = \Sigma^2\,. \tag{6.33}$$

□

Satz 6.10 (Ähnlichkeit minimaler Realisierungen)
Zwei minimale Realisierungen $(A, B, C), (\tilde{A}, \tilde{B}, \tilde{C}) \in \mathcal{S}_{n,m,k}$ derselben Übertragungsfunktion sind ähnlich.

Beweis: Sei

$$O = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix}, \quad \tilde{O} = \begin{pmatrix} \tilde{C} \\ \tilde{C}\tilde{A} \\ \vdots \\ \tilde{C}\tilde{A}^{n-1} \end{pmatrix}, \tag{6.34}$$

$$R = \begin{pmatrix} B & AB & \cdots & A^{n-1}B \end{pmatrix}, \quad \tilde{R} = \begin{pmatrix} \tilde{B} & \tilde{A}\tilde{B} & \cdots & \tilde{A}^{n-1}\tilde{B} \end{pmatrix}. \tag{6.35}$$

Aus der Gleichheit der Hankel-Matrizen beider Kontrollsysteme folgt

$$OR = H_n = \tilde{H}_n = \tilde{O}\tilde{R}\,. \tag{6.36}$$

Da OAR aus H_{n+1} durch Streichen der ersten Zeile und der ersten Spalte entsteht, gilt weiter

$$OAR = \tilde{O}\tilde{A}\tilde{R}\,. \tag{6.37}$$

Da beide Kontrollsysteme steuerbar und beobachtbar sind, folgt

$$\operatorname{rang}(O) = \operatorname{rang}(\tilde{O}) = \operatorname{rang}(R) = \operatorname{rang}(\tilde{R}) = n\,, \tag{6.38}$$

also gibt es eine Linksinverse $\tilde{O}^{\#}$ und eine Rechtsinverse $\tilde{R}^{\#}$ mit

$$\tilde{O}^{\#}\tilde{O} = I\,, \quad \tilde{R}\tilde{R}^{\#} = I\,, \quad (\tilde{O}^{\#}O)(R\tilde{R}^{\#}) = I\,. \tag{6.39}$$

Die gesuchte Ähnlichkeitstransformation ist

$$T = \tilde{R}\tilde{R}^{\#} = (\tilde{O}^{\#}\tilde{O})^{-1}\,, \tag{6.40}$$

da aus (6.37) und (6.36) folgt

$$\tilde{A} = T^{-1}AT\,, \quad T^{-1}R = \tilde{R}\,, \quad OT = \tilde{O}\,, \tag{6.41}$$

und damit auch $T^{-1}B = \tilde{B}$ sowie $CT = \tilde{C}$. □

Wir beschäftigen uns jetzt mit der Frage, in welchem Sinne eine balancierte Realisierung "minimale Kondition" hat.

Lemma 6.11 *Wir definieren*

$$\mu(A) = \text{ cond}\,(A^TA) = \|A^TA\|_2\|(A^TA)^{-1}\|_2\,, \quad A \in \mathbb{R}^{(n,n)}\,. \tag{6.42}$$

Dann gelten

$$\mu(AB) \le \mu(A)\mu(B)\,, \quad \mu(A) = \mu(A^T)\,, \tag{6.43}$$

für alle $A, B \in \mathbb{R}^{(n,n)}$.

Beweis: Ist $A = U\Sigma_A V^T$ die Singulärwertzerlegung von A, so gilt

$$A^TA = V\Sigma_A^2V^T\,, \quad AA^T = U\Sigma_A^2U^T\,, \tag{6.44}$$

also

$$\mu(A) = \text{ cond}\,(A^TA) = \text{ cond}\,(\Sigma_A^2) = \text{ cond}\,(AA^T) = \mu(A^T)\,. \tag{6.45}$$

Zieht man außerdem die Singulärwertzerlegung $B = P\Sigma_B Q^T$ von B heran, so folgt durch Einsetzen und Dreiecksungleichung

$$\begin{aligned} \|(AB)^T(AB)\|_2 &\le \|\Sigma_B\|_2\|\Sigma_A\|_2^2\|\Sigma_B\|_2 = \|\Sigma_A^2\|_2\|\Sigma_B^2\|_2 \\ &= \|A^TA\|_2\|B^TB\|_2 \end{aligned} \tag{6.46}$$

Kombiniert man (6.46) mit der analogen Formel für die Inverse $(AB)^{-1}$ anstelle von AB, so ergibt sich die Ungleichung in (6.43). □

Satz 6.12 (Minimaleigenschaft einer balancierten Realisierung)
Sei $(A, B, C) \in \mathcal{S}_{n,m,k}$ *eine balancierte Realisierung einer Übertragungsfunktion* G *mit*

$$W_t = Y_t = D\,, \quad D = \text{ diag}\,(d_1, \ldots, d_n)\,. \tag{6.47}$$

Dann gilt für jede andere minimale Realisierung $(\tilde{A}, \tilde{B}, \tilde{C})$ *von* G

$$\text{cond}\,(D)^2 \le \text{ cond}\,(\tilde{W}_t)\ \text{cond}\,(\tilde{Y}_t)\,. \tag{6.48}$$

Falls die Diagonalelemente von D *der Größe nach geordnet sind, ist*

$$\text{cond}\,(D)^2 = \frac{d_1^2}{d_n^2}\,. \tag{6.49}$$

Beweis: Da (A, B, C) und $(\tilde{A}, \tilde{B}, \tilde{C})$ ähnlich sind, gibt es eine Transformation T mit

$$\tilde{A} = T^{-1}AT\,, \quad T^{-1}B = \tilde{B}\,, \quad CT = \tilde{C}\,. \tag{6.50}$$

Sei Σ Diagonalmatrix mit $D = \Sigma^2$. Es gilt dann wegen Lemma 6.11

$$\begin{aligned} \text{cond}\,(D)^2 &= \text{cond}\,(D^2) = \mu(D) = \mu(\Sigma TT^{-1}\Sigma) \le \mu(\Sigma T)\mu(T^{-1}\Sigma) \\ &= \mu(\Sigma T)\mu(\Sigma T^{-T}) = \text{ cond}\,(T^T\Sigma^2T)\ \text{cond}\,(T^{-1}\Sigma^2T^{-T}) \\ &= \text{cond}\,(\tilde{Y}_t)\ \text{cond}\,(\tilde{W}_t)\,. \end{aligned} \tag{6.51}$$

□

Algorithmus 6.13 (Numerische Berechnung einer balancierten Realisierung) Sei $(A, B, C) \in \mathcal{S}_{n,m,k}$ ein steuerbares und beobachtbares Kontrollsystem, sei A Stabilitätsmatrix. Wir suchen eine balancierte Realisierung für $t = \infty$ derselben Übertragungsfunktion, d.h. eine Transformation T des Zustandsraums, so daß für

$$(\tilde{A}, \tilde{B}, \tilde{C}) = (T^{-1}AT, T^{-1}B, CT) \tag{6.52}$$

gilt

$$\tilde{W}_\infty = \tilde{Y}_\infty = D\,, \tag{6.53}$$

wobei D eine (ebenfalls zu bestimmende) Diagonalmatrix ist und

$$W_\infty = \int_0^\infty e^{sA}BB^T e^{sA^T}\,ds\,, \quad Y_\infty = \int_0^\infty e^{sA^T}C^T C e^{sA}\,ds\,. \tag{6.54}$$

Das Verfahren sieht folgendermaßen aus:

1. Berechne W_∞ und Y_∞ als Lösung der Ljapunov-Gleichungen

$$AW_\infty + W_\infty A^T = -BB^T\,, \tag{6.55}$$

$$A^T Y_\infty + Y_\infty A = -C^T C\,, \tag{6.56}$$

   etwa mit dem Verfahren von Bartels und Stewart.

2. Berechne eine obere Dreiecksmatrix R mit

$$W_\infty = R^T R\,, \tag{6.57}$$

   etwa mit dem Cholesky-Verfahren.

3. Berechne die Schur-Form

$$RY_\infty R^T = VD^2V^T\,, \tag{6.58}$$

   von $RY_\infty R^T$, etwa mit dem QR-Verfahren.

4. Setze

$$T = R^T V D^{-\frac{1}{2}}\,. \tag{6.59}$$

Dann sind T und D die gesuchten Matrizen, da

$$\tilde{Y}_\infty = T^T Y_\infty T = D^{-\frac{1}{2}}V^T R Y_\infty R^T V D^{-\frac{1}{2}} = D \tag{6.60}$$

$$TDT^T = R^T V D^{-\frac{1}{2}} D D^{-\frac{1}{2}} V^T R = R^T R = W_\infty\,, \tag{6.61}$$

also

$$\tilde{W}_\infty = T^{-1}W_\infty T^{-T} = D\,. \tag{6.62}$$

7 Die algebraische Matrix-Riccati-Gleichung

Problem 7.1
Wir suchen eine Lösung $X \in \mathbb{R}^{(n,n)}$ mit

$$A^T X + XA - XMX + R = 0\,, \tag{7.1}$$

wobei $A, M, R \in \mathbb{R}^{(n,n)}$ gegeben sind. Wir setzen in diesem Abschnitt voraus, daß M und R symmetrisch sind.

Bemerkung 7.2 (Zusammenhang zum LQ-Problem)
Wir betrachten das LQ-Problem: Minimiere

$$J_\infty(x, u; \xi) = \int_0^\infty x(t)^T R x(t) + u(t)^T u(t)\, dt\,, \tag{7.2}$$

wobei

$$\dot{x} = Ax + Bu\,, \quad x(0) = \xi\,, \tag{7.3}$$

und $A, R \in \mathbb{R}^{(n,n)}$, $B \in \mathbb{R}^{(n,m)}$. Wir setzen

$$M = BB^T\,. \tag{7.4}$$

Wir haben u.a. gezeigt: Ist (A, B) steuerbar und $R \geq 0$, so gibt es eine symmetrische Lösung $X \geq 0$ von (7.1), mit

$$X = \lim_{t\to\infty} P(t)\,, \tag{7.5}$$

wobei $P(t)$ die Lösung der Anfangswertaufgabe

$$\dot{X} = A^T X + XA - XMX + R\,, \quad X(0) = 0\,, \tag{7.6}$$

ist. Ist darüber hinaus $R > 0$, so ist auch $X > 0$ und $A - MX$ Stabilitätsmatrix.

Lemma 7.3 *Seien $X_1, X_2 \in \mathbb{R}^{(n,n)}$ Lösungen von*

$$\dot{X} = A^T X + XA - XMX + R\,. \tag{7.7}$$

Dann gilt: Ist $X_1(0) \geq X_2(0)$, so ist auch $X_1(t) \geq X_2(t)$ für alle $t \geq 0$.

Beweis: : Übung. □

Lemma 7.4 *Seien $X, \tilde{X} \in \mathbb{R}^{(n,n)}$ symmetrisch, sei X Lösung von*

$$A^T X + XA - XMX + R = 0\,. \tag{7.8}$$

Dann ist

$$Y = X - \tilde{X} \tag{7.9}$$

eine Lösung von

$$\tilde{A}^T Y + Y\tilde{A} - YMY + \tilde{R} = 0\,, \tag{7.10}$$

wobei

$$\tilde{A} = A - M\tilde{X}\,, \tag{7.11}$$

$$\tilde{R} = A^T\tilde{X} + \tilde{X}A - \tilde{X}M\tilde{X} + R\,, \tag{7.12}$$

d.h. $\tilde{R}$ ist der Defekt von $\tilde{X}$, eingesetzt in (7.8). Insbesondere gilt

$$\tilde{A}^T Y + Y\tilde{A} - YMY = 0\,, \tag{7.13}$$

falls $\tilde{X}$ ebenfalls Lösung von (7.8) ist.

Beweis: Einsetzen ergibt

$$\begin{aligned}\tilde{A}^T Y + Y\tilde{A} - YMY + R &= -A^T\tilde{X} - \tilde{X}A - \tilde{X}M\tilde{X} + XM\tilde{X} + \tilde{X}MX \\ &= -(A^T\tilde{X} + \tilde{X}A - \tilde{X}M\tilde{X}) + YM\tilde{X} + \tilde{X}MY\,, \end{aligned} \tag{7.14}$$

woraus die Behauptung folgt. □

Satz 7.5 (Eindeutigkeit)
Sei $R > 0$ und $M > 0$. Dann hat (7.1) eine Lösung $X > 0$, welche eindeutig ist in der Menge aller symmetrischer positiv semidefiniter Matrizen, und $A - MX$ ist Stabilitätsmatrix.

Beweis: Wir zerlegen $M = BB^T$ mit $B \in \mathbb{R}^{(n,n)}$. B ist nichtsingulär, also ist (A, B) steuerbar. Wegen Bemerkung 7.2 bleibt nur noch die Eindeutigkeit zu zeigen. Sei $\tilde{X} \geq 0$ eine symmetrische Lösung von (7.1). Aus Lemma 7.3, angewendet auf $X_1(0) = \tilde{X}$ und $X_2(0) = 0$, folgt $X \leq \tilde{X}$. Die Differenz $Y = \tilde{X} - X$ ist nach Lemma 7.4 eine Lösung der Ljapunov-Gleichung

$$(A - MX)^T Y + Y(A - MX) = -C\,, \quad C = -YMY \leq 0\,. \tag{7.15}$$

Da $A - MX$ Stabilitätsmatrix ist, folgt $Y \leq 0$, also auch $Y = 0$. □

Bemerkung 7.6
Die Voraussetzungen an R und M lassen sich abschwächen. Sind

$$M = BB^T\,, \quad R = C^T C\,, \tag{7.16}$$

und sind (A, B) stabilisierbar und (A, C) entdeckbar, so gibt es eine symmetrische Lösung $X \geq 0$ von (7.1), welche eindeutig ist in der Menge aller symmetrischer positiv semidefiniter Matrizen. (Siehe etwa Knobloch/Kwakernaak.)

Algorithmus 7.7 (Newton-Verfahren)
Sei $\tilde{X}$ eine Näherungslösung der algebraischen Matrix-Riccati-Gleichung (7.1). Die Newtonkorrektur H in $\tilde{X}$ ergibt sich durch Linearisieren von

$$A^T(\tilde{X} + H) + (\tilde{X} + H)A - (\tilde{X} + H)M(\tilde{X} + H) + R = 0\,, \tag{7.17}$$

also als Lösung von

$$\tilde{A}^T H + H\tilde{A} + \tilde{R} = 0\,, \tag{7.18}$$

wobei

$$\tilde{A} = A - M\tilde{X}\,, \quad \tilde{R} = A^T\tilde{X} + \tilde{X}A - \tilde{X}M\tilde{X} + R\,. \tag{7.19}$$

Die neue Näherung X mit $X - \tilde{X} = H$ löst dann

$$\tilde{A}^T X + X\tilde{A} + R + \tilde{X}M\tilde{X} = 0\,. \tag{7.20}$$

Das Newton-Verfahren hat also folgende Form:

1. Wähle eine symmetrische Startnäherung $X_0 \geq 0$, so daß $A - MX_0$ Stabilitätsmatrix ist.

2. Berechne

$$A_j = A - MX_{j-1}\,, \quad R_j = R + X_{j-1}MX_{j-1}\,, \tag{7.21}$$

und X_j als Lösung der Ljapunov-Gleichung

$$A_j^T X_j + X_j A_j + R_j = 0\,, \tag{7.22}$$

etwa mit dem Verfahren von Bartels und Stewart.

Satz 7.8 (Konvergenz des Newton-Verfahrens)
Seien A, M, R wie in Problem 7.1 gegeben, sei zusätzlich $R > 0$ und $M \geq 0$, sei $X_0 \geq 0$ symmetrisch und $A - MX_0$ Stabilitätsmatrix. Dann ist das Newtonverfahren 7.7 wohldefiniert und gegen eine Lösung X von (7.1) konvergent. Die Konvergenz ist (vom zweiten Iterationsschritt an) monoton, d.h.

$$0 \leq X \leq \ldots \leq X_j \leq \ldots \leq X_1\,. \tag{7.23}$$

Es gilt $X > 0$ genau dann, wenn $A - MX$ Stabilitätsmatrix ist; in diesem Fall ist die Konvergenz quadratisch, d.h. es gibt ein $c > 0$ mit

$$\|X - X_j\|_2 \leq c\|X - X_{j-1}\|_2^2\,, \quad j \geq 1\,. \tag{7.24}$$

Beweis: Wohldefiniertheit: Sei $X_{j-1} \geq 0$ symmetrisch und $A - MX_{j-1}$ Stabilitätsmatrix. Dann hat die Ljapunov-Gleichung

$$(A - MX_{j-1})^T X_j + X_j(A - MX_{j-1}) = -R - X_{j-1}MX_{j-1} \tag{7.25}$$

eine eindeutig bestimmte symmetrische Lösung $X_j > 0$. Es gilt weiter

$$(A - MX_j)^T X_j + X_j(A - MX_j) = -R - X_j MX_j - (X_j - X_{j-1})M(X_j - X_{j-1}) < 0\,, \tag{7.26}$$

also ist $A - MX_j$ ebenfalls Stabilitätsmatrix. Zur Monotonie: Subtraktion von (7.26) (für $j-1$ statt j) und (7.25) ergibt, falls $j \geq 2$,

$$\begin{aligned}(A - MX_{j-1})^T(X_{j-1} - X_j) + (X_{j-1} - X_j)(A - MX_{j-1})\\ = -(X_{j-1} - X_{j-2})M(X_{j-1} - X_{j-2}) \leq 0\,,\end{aligned} \tag{7.27}$$

also ist $X_{j-1} - X_j \geq 0$. Zur Konvergenz: Für jedes $y \in \mathbb{R}^n$ ist die Folge

$$z_j = y^T X_j y\,, \quad j \geq 1\,, \tag{7.28}$$

monoton fallend (wegen $X_j - X_{j-1} \leq 0$) und nach unten beschränkt (nämlich durch 0), also konvergent. Wegen

$$X_{j,ik} = e_i^T X_j e_k = \frac{1}{2}\Big((e_i + e_k)^T X_j(e_i + e_k) - e_i^T X_j e_i - e_k^T X_j e_k\Big) \tag{7.29}$$

existiert also

$$X = \lim_{j\to\infty} X_j \tag{7.30}$$

und ist symmetrisch und positiv semidefinit. Grenzübergang in (7.25) ergibt

$$(A - MX)^T X + X(A - MX) = -R - XMX < 0\,, \tag{7.31}$$

also ist X Lösung von (7.1) und $X > 0$ genau dann, wenn $A - MX$ Stabilitätsmatrix ist. Es bleibt für diesen Fall die quadratische Konvergenz zu zeigen. Aus (7.31) folgt zunächst

$$(A - MX_j)^T X + X(A - MX_j) = -R - XMX + (X - X_j)MX + XM(X - X_j)\,, \quad (7.32)$$

und durch Subtraktion von (7.26)

$$\begin{aligned}(A - MX_j)^T(X - X_j) + (X - X_j)(A - MX_j) = \\ = (X - X_j)M(X - X_j) + (X_j - X_{j-1})M(X_j - X_{j-1})\,,\end{aligned} \quad (7.33)$$

und schließlich

$$\begin{aligned}(A - MX)^T(X - X_j) + (X - X_j)(A - MX) = \\ = -(X - X_j)M(X - X_j) + (X_j - X_{j-1})M(X_j - X_{j-1})\,.\end{aligned} \quad (7.34)$$

Die Lösungsdarstellung der Ljapunov-Gleichung liefert

$$X_j - X = \int_0^\infty e^{t(A-MX)^T}\Big[-(X-X_j)M(X-X_j)+(X_j-X_{j-1})M(X_j-X_{j-1})\Big]e^{t(A-MX)}\,dt\,, \quad (7.35)$$

also

$$0 \le X_j - X \le \int_0^\infty e^{t(A-MX)^T}(X_j - X_{j-1})M(X_j - X_{j-1})e^{t(A-MX)}\,dt\,, \quad (7.36)$$

und weiter

$$\begin{aligned}\|X_j - X\|_2 &\le \left\|\int_0^\infty e^{t(A-MX)^T}(X_j - X_{j-1})M(X_j - X_{j-1})e^{t(A-MX)}\,dt\right\|_2 \\ &\le \|X_j - X_{j-1}\|_2^2 \underbrace{\|M\|_2 \int_0^\infty \left\|e^{t(A-MX)^T}\right\|_2 \left\|e^{t(A-MX)}\right\|_2 dt}_{=:c} \\ &\le c\|X - X_{j-1}\|_2^2\,,\end{aligned} \quad (7.37)$$

da

$$0 \le X_{j-1} - X_j \le X_{j-1} - X\,. \quad (7.38)$$

□

Bemerkung 7.9
Die Hauptschwierigkeit liegt (wie meistens beim Newton-Verfahren) in der Beschaffung der Startnäherung X_0. Sogar wenn man ein X_0 gefunden hat, welches die Voraussetzungen des Konvergenzsatzes 7.8 erfüllt, kann die nächste Näherung X_1 sehr schlecht sein. Dieses Verhalten tritt schon im Skalaren ($n = 1$) auf, etwa bei der Lösung von

$$-x^2 + 1 = 0 \quad (7.39)$$

mit der Startnäherung $x_0 = \varepsilon$. Das Newton-Verfahren wird daher meistens in Kombination mit einem anderen Verfahren eingesetzt.

Lemma 7.10 *Sei* $J \in \mathbb{R}^{(2n,2n)}$ *definiert durch*

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}\,, \quad (7.40)$$

wobei $I = I_n$ *die Einheitsmatrix im* $\mathbb{R}^n$ *ist. Dann gilt*

$$J^T = J^{-1} = -J\,, \quad J^2 = -I_{2n}\,. \quad (7.41)$$

Beweis: Klar. □

Definition 7.11 (Hamiltonsche Matrix)
Eine Matrix $H \in \mathbb{R}^{(2n,2n)}$ heißt hamiltonsch, wenn

$$(JH)^T = JH\,. \tag{7.42}$$

Lemma 7.12 *Summen, skalare Vielfache und Inverse von hamiltonschen Matrizen sind wieder hamiltonsch. Ist H hamiltonsch und λ ein Eigenwert von H, so ist auch $-\lambda$ ein Eigenwert von H.*

Beweis: Für die Inverse: Sei H hamiltonsch und invertierbar. Dann gilt

$$(JH^{-1})^T = H^{-T}J^T = -H^{-T}J = -H^{-T}JHH^{-1} = -H^{-T}H^TJ^TH^{-1} = JH^{-1}\,. \tag{7.43}$$

Sei $\lambda \in \operatorname{spec}(H)$, x ein zugehöriger Eigenvektor. Dann gilt

$$\lambda Jx = JHx = (JH)^Tx = H^TJ^Tx = -H^TJx\,, \tag{7.44}$$

also ist $\lambda \in \operatorname{spec}(-H^T) = \operatorname{spec}(-H)$. □

Lemma 7.13 *Seien $A, M, R \in \mathbb{R}^{(n,n)}$ und M, R symmetrisch. Ein symmetrisches $X \in \mathbb{R}^{(n,n)}$ ist Lösung von*

$$A^TX + XA - XMX + R = 0 \tag{7.45}$$

genau dann, wenn

$$\begin{pmatrix} A & M \\ R & -A^T \end{pmatrix}\begin{pmatrix} I & 0 \\ -X & I \end{pmatrix} = \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix}\begin{pmatrix} A-MX & M \\ 0 & -(A-MX)^T \end{pmatrix}. \tag{7.46}$$

Die Matrix H,

$$H = \begin{pmatrix} A & M \\ R & -A^T \end{pmatrix}, \tag{7.47}$$

ist hamiltonsch. Ist $A - MX$ Stabilitätsmatrix, so hat H keine rein imaginären Eigenwerte.

Beweis: Man multipliziert (7.47) aus und betrachtet die entstehenden 4 Matrixgleichungen einzeln. Links unten steht (7.46). Die anderen drei sind identisch erfüllt. Die anderen Behauptungen sind klar. □

Bemerkung 7.14
Das im folgenden dargestellte Verfahren von Byers (1987) zur Lösung der algebraischen Riccati-Gleichung (7.45) beruht auf der Verwendung der Matrixsignumfunktion, und zwar wird zunächst $\operatorname{sign}(H)$ (für H aus (7.47)) und daraus X berechnet. Es stellt sich die Frage, wie man aus einer Funktion $f : \mathbb{R} \to \mathbb{R}$ bzw. $f : \mathbb{C} \to \mathbb{C}$ eine Funktion $f : \mathbb{R}^{(n,n)} \to \mathbb{R}^{(n,n)}$ bzw. $f : \mathbb{C}^{(n,n)} \to \mathbb{C}^{(n,n)}$ erhält. Hat f eine Potenzreihenentwicklung

$$f(z) = \sum_{k=0}^{\infty} c_k z^k\,, \tag{7.48}$$

so können wir im Innern des Konvergenzkreises f durch

$$f(A) = \sum_{k=0}^{\infty} c_k A^k \,, \tag{7.49}$$

definieren (Beispiel: Matrixexponentialfunktion). Eine allgemeinere Methode beruht auf der Integralformel von Cauchy

$$f(z) = \frac{1}{2\pi i} \oint_\Gamma \frac{f(\zeta)}{\zeta - z}\, d\zeta \,. \tag{7.50}$$

Hier ist Γ eine geschlossene Kurve, welche den Punkt z einmal im mathematisch positiven Sinn umläuft, und f analytisch in dem von Γ umschlossenen Gebiet.

Definition 7.15 (Matrixfunktionen)
Sei $G \subset \mathbb{C}$ offen und einfach zusammenhängend, $f : G \to \mathbb{C}$ analytisch. Wir definieren $f : \mathbb{C}^{(n,n)} \to \mathbb{C}^{(n,n)}$ durch

$$f(H) = \frac{1}{2\pi i} \oint_\Gamma f(\zeta)(\zeta I - H)^{-1}\, d\zeta \,, \tag{7.51}$$

wobei $\Gamma \subset G$ eine geschlossene Kurve ist, welche die Menge

$$\operatorname{spec}(H) \cap G \tag{7.52}$$

einmal im mathematisch positiven Sinn umläuft.

Lemma 7.16 *Die Definition von f in (7.51) hängt nicht von der Wahl von Γ ab (solange die in Definition 7.15 formulierten Bedingungen erfüllt sind). Für jede invertierbare Matrix $T \in \mathbb{C}^{(n,n)}$ gilt*

$$f(T^{-1}HT) = T^{-1} f(H) T \,, \tag{7.53}$$

und

$$HT = TH \quad \Rightarrow \quad f(H)T = Tf(H) \,. \tag{7.54}$$

Weiter gilt

$$H = \operatorname{diag}(H_1, \ldots, H_k) \quad \Rightarrow \quad f(H) = \operatorname{diag}(f(H_1), \ldots, f(H_k)) \,. \tag{7.55}$$

Beweis: Sei Γ fest gewählt. Die Gleichungen (7.53) und (7.55) folgen aus der Definition, (7.54) aus (7.53) mit der Rechnung

$$f(H)T = TT^{-1} f(H) T = T f(T^{-1}HT) = T f(T^{-1}TH) = T f(H) \,. \tag{7.56}$$

Da wir jede Matrix auf Jordansche Normalform transformieren können, ohne daß sich das Spektrum ändert, können wir annehmen, daß H in Jordanscher Normalform vorliegt, d.h. $H = \operatorname{diag}(H_1, \ldots, H_k)$, wobei jeder einzelne Block H_i die Form hat

$$\tilde{H} = \lambda I + E \,, \quad E = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix} \in \mathbb{C}^{(p,p)} \,, \quad \lambda \in \operatorname{spec}(H) \,. \tag{7.57}$$

Wir berechnen $f(\tilde{H})$. Für $\zeta \neq \lambda$ gilt, da $E^p = 0$,

$$\begin{aligned}(\zeta I - \tilde{H})^{-1} &= ((\zeta-\lambda)I - E)^{-1} = \frac{1}{\zeta-\lambda}\left(I - \frac{E}{\zeta-\lambda}\right)^{-1} \\ &= \sum_{k=0}^{p-1} \frac{E^k}{(\zeta-\lambda)^{k+1}}\,, \end{aligned} \tag{7.58}$$

also

$$f(\tilde{H}) = \frac{1}{2\pi i}\oint_\Gamma f(\zeta)(\zeta I - \tilde{H})^{-1}\,d\zeta = \sum_{k=0}^{p-1}\frac{1}{2\pi i}\oint_\Gamma \frac{f(\zeta)}{(\zeta-\lambda)^{k+1}}\,d\zeta\,E^k\,. \tag{7.59}$$

Falls $\lambda \notin G$, so sind alle Kurvenintegrale in (7.59) Null, also

$$f(\tilde{H}) = 0\,. \tag{7.60}$$

Falls $\lambda \in G$, so liegt λ im Innern von Γ, und

$$\frac{f^{(k)}(\lambda)}{k!} = \frac{1}{2\pi i}\oint_\Gamma \frac{f(\zeta)}{(\zeta-\lambda)^{k+1}}\,d\zeta\,, \tag{7.61}$$

also

$$f(\tilde{H}) = \begin{pmatrix} f(\lambda) & f'(\lambda) & \cdots & \cdots & \frac{f^{(p-1)}(\lambda)}{(p-1)!} \\ 0 & f(\lambda) & f'(\lambda) & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & f'(\lambda) \\ 0 & \cdots & \cdots & \cdots & f(\lambda) \end{pmatrix}. \tag{7.62}$$

In beiden Fällen ist also $f(\tilde{H})$ unabhängig von der Wahl von Γ. □

Bemerkung 7.17 (Signumfunktion auf $\mathbb{C}$)
Wir setzen $\mathbb{C} = \mathbb{C}_+ \cup \mathbb{C}_0 \cup \mathbb{C}_-$, wobei

$$\mathbb{C}_+ = \{z : \mathrm{Re}\,(z) > 0\}\,, \quad \mathbb{C}_- = \{z : \mathrm{Re}\,(z) < 0\}\,, \quad \mathbb{C}_0 = \{z : \mathrm{Re}\,(z) = 0\}\,, \tag{7.63}$$

und definieren sign^+, $\mathrm{sign} : \mathbb{C}_+ \cup \mathbb{C}_- \to \mathbb{R}$ durch

$$\mathrm{sign}^+(z) = \begin{cases} 1\,, & z \in \mathbb{C}_+\,, \\ 0\,, & z \in \mathbb{C}_- \end{cases}\,, \tag{7.64}$$

$$\mathrm{sign}\,(z) = 2\,\mathrm{sign}^+(z) - 1 = \begin{cases} 1\,, & z \in \mathbb{C}_+\,, \\ -1\,, & z \in \mathbb{C}_- \end{cases}\,. \tag{7.65}$$

Definition 7.18 (Matrixsignumfunktion)
Sei $H \in \mathbb{C}^{(n,n)}$ mit $\mathrm{spec}\,(H) \subset \mathbb{C}_+ \cup \mathbb{C}_-$, d.h. H hat keinen rein imaginären Eigenwert. Wir definieren $\mathrm{sign}^+(H)$ gemäß 7.15 mit $G = \mathbb{C}_+$ und $f = 1$, also

$$\mathrm{sign}^+(H) = \frac{1}{2\pi i}\oint_\Gamma (\zeta I - H)^{-1}\,d\zeta\,, \tag{7.66}$$

wobei Γ ein Weg in $\mathbb{C}_+$ ist, welcher alle Eigenwerte von H mit positivem Realteil einmal im mathematisch positiven Sinne umläuft, und setzen

$$\mathrm{sign}\,(H) = 2\,\mathrm{sign}^+(H) - I\,. \tag{7.67}$$

Lemma 7.19 (Elementare Eigenschaften der Matrixsignumfunktion) *Sei* $H \in \mathbb{C}^{(n,n)}$, *H habe keinen rein imaginären Eigenwert.*

(i) *Hat H die Form eines Jordan-Kästchens $H = \lambda I + E$ wie in (7.57), so gilt*

$$\operatorname{sign}(H) = \begin{cases} I\,, & \operatorname{Re}(\lambda) > 0\,, \\ -I\,, & \operatorname{Re}(\lambda) < 0\,. \end{cases} \tag{7.68}$$

(ii) *Ist* $H = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, *so gilt*

$$\operatorname{sign}(H) = \operatorname{diag}(\operatorname{sign}(\lambda_1), \ldots, \operatorname{sign}(\lambda_n))\,. \tag{7.69}$$

(iii)

$$\operatorname{sign}(H) = I\,, \qquad \text{falls } \operatorname{spec}(H) \subset \mathbb{C}_+\,, \tag{7.70}$$

$$\operatorname{sign}(H) = -I\,, \qquad \text{falls } \operatorname{spec}(H) \subset \mathbb{C}_-\,, \tag{7.71}$$

$$\operatorname{sign}(H)^2 = I\,. \tag{7.72}$$

Beweis: Folgt aus Betrachtung der Jordanschen Normalform und Formel (7.62), da $f = 1$ in $\mathbb{C}_+$. □

Bemerkung 7.20
Hat H keinen rein imaginären Eigenwert, so ist sign analytisch in einer Umgebung von H. Hat H mindestens einen rein imaginären Eigenwert, so ist sign i.a. unstetig in H (egal wie man $\operatorname{sign}(H)$ definiert).

Satz 7.21 *Seien $A, M, R \in \mathbb{R}^{(n,n)}$ und M, R symmetrisch, sei $X \in \mathbb{R}^{(n,n)}$ symmetrische Lösung von*

$$A^T X + XA - XMX + R = 0\,, \tag{7.73}$$

sei $A - MX$ Stabilitätsmatrix. Dann ist X eine Lösung des (überbestimmten) Gleichungssystems

$$\begin{pmatrix} W_{12} \\ W_{22} + I \end{pmatrix} X = \begin{pmatrix} I + W_{11} \\ W_{21} \end{pmatrix}, \tag{7.74}$$

wobei

$$\begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} = \operatorname{sign}(H)\,, \quad H = \begin{pmatrix} A & M \\ R & -A^T \end{pmatrix}. \tag{7.75}$$

Beweis: Aus Lemma 7.13 wissen wir, daß

$$H \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix} = \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix} \begin{pmatrix} A - MX & M \\ 0 & -(A - MX)^T \end{pmatrix}, \tag{7.76}$$

also gilt auch

$$\operatorname{sign}(H) \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix} = \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix} \operatorname{sign} \begin{pmatrix} A - MX & M \\ 0 & -(A - MX)^T \end{pmatrix}. \tag{7.77}$$

Sei Z die eindeutig bestimmte symmetrische Lösung von

$$(A - MX)Z + Z(A - MX)^T = -M\,. \tag{7.78}$$

Dann gilt (Ausmultiplizieren, rechts oben steht (7.78), sonst eine Identität)

$$\begin{pmatrix} I & -Z \\ 0 & I \end{pmatrix} \begin{pmatrix} A - MX & M \\ 0 & -(A - MX)^T \end{pmatrix} = \begin{pmatrix} A - MX & 0 \\ 0 & -(A - MX)^T \end{pmatrix} \begin{pmatrix} I & -Z \\ 0 & I \end{pmatrix}, \tag{7.79}$$

also auch

$$\begin{aligned} &\operatorname{sign} \begin{pmatrix} A - MX & M \\ 0 & -(A - MX)^T \end{pmatrix} = \\ &= \begin{pmatrix} I & -Z \\ 0 & I \end{pmatrix}^{-1} \operatorname{sign} \begin{pmatrix} A - MX & 0 \\ 0 & -(A - MX)^T \end{pmatrix} \begin{pmatrix} I & -Z \\ 0 & I \end{pmatrix} \\ &= \begin{pmatrix} I & Z \\ 0 & I \end{pmatrix} \begin{pmatrix} -I & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} I & -Z \\ 0 & I \end{pmatrix} \\ &= \begin{pmatrix} -I & 2Z \\ 0 & I \end{pmatrix}. \end{aligned} \tag{7.80}$$

Aus (7.77) und (7.80) folgt

$$\begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix} = \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix} \begin{pmatrix} -I & 2Z \\ 0 & I \end{pmatrix}. \tag{7.81}$$

Indem wir (7.81) ausmultiplizieren, erhalten wir (7.74) als die erste Spalte des Produkts. □

Algorithmus 7.22 (Lösung der Riccati-Gleichung, Matrixsignumfunktion)
Aus Satz 7.21 ergibt sich folgender Algorithmus zur Lösung der Matrix-Riccati-Gleichung (7.73):

1. Berechne
$$\begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} = \operatorname{sign}(H), \quad H = \begin{pmatrix} A & M \\ R & -A^T \end{pmatrix}. \tag{7.82}$$
2. Berechne X als Lösung des linearen Ausgleichsproblems
$$\min \left\| \begin{pmatrix} W_{12} \\ W_{22} + I \end{pmatrix} X - \begin{pmatrix} I + W_{11} \\ W_{21} \end{pmatrix} \right\|. \tag{7.83}$$

Bemerkung 7.23 (Berechnung der Matrixsignumfunktion)
In $\mathbb{C}$ gilt: Lösen wir die Gleichung

$$z^2 - 1 = 0 \tag{7.84}$$

mit dem Newton-Verfahren

$$z_{n+1} = g(z_n), \quad g(z) = \frac{1}{2}\left(z + \frac{1}{z}\right), \tag{7.85}$$

so gilt für $z_0 \in \mathbb{C}_+ \cup \mathbb{C}_-$

$$\lim_{n \to \infty} z_n = \operatorname{sign}(z_0). \tag{7.86}$$

Die Konvergenz ist zwar quadratisch, aber global langsam, da für große $|z_n|$ gilt, daß

$$|z_{n+1}| \approx \frac{1}{2}|z_n| \,. \tag{7.87}$$

Durch reelle Skalierung läßt sich das Konvergenzverhalten verbessern, da

$$\text{sign}\,(\alpha z_0) = \text{sign}\,(z_0)\,, \qquad \text{falls } \alpha > 0\,. \tag{7.88}$$

Beide Überlegungen lassen sich auf Matrizen übertragen.

Satz 7.24 (Newton-Verfahren zur Berechnung von sign(H))
Sei $H \in \mathbb{C}^{(n,n)}$, H habe keinen rein imaginären Eigenwert. Dann ist die Newton-Iteration

$$Z_{k+1} = \frac{1}{2}(Z_k + Z_k^{-1})\,, \quad Z_0 = H\,, \tag{7.89}$$

wohldefiniert, und es gilt

$$\lim_{k\to\infty} Z_k = \text{sign}\,(H)\,. \tag{7.90}$$

Beweis: Zusammen mit Z_k betrachten wir

$$F_k = (Z_k - \text{sign}\,(H))(Z_k + \text{sign}\,(H))^{-1}\,. \tag{7.91}$$

Wir zeigen mit vollständiger Induktion (7.92)-(7.98),

$$Z_k H = H Z_k\,, \tag{7.92}$$

$$Z_k + \text{sign}\,(H) \text{ ist invertierbar}, \tag{7.93}$$

$$F_k = F_{k-1}^2\,, \tag{7.94}$$

$$\rho(F_k) < 1\,, \qquad (\,\rho = \text{ Spektralradius})\quad, \tag{7.95}$$

$$Z_k = (I - F_k)^{-1}(I + F_k)\text{sign}\,(H)\,, \tag{7.96}$$

$$Z_k \text{ ist invertierbar}\quad, \tag{7.97}$$

$$Z_k^{-1} H = H Z_k^{-1}\,. \tag{7.98}$$

Wir verwenden dabei die Rechenregeln aus 7.16 und 7.19.
Induktionsanfang $k = 0$: (7.92), (7.97) und (7.98) sind trivial, (7.94) entfällt. Es gilt (Beweis: Übung)

$$\text{spec}\,(H \pm \text{sign}\,(H)) = \{\lambda \pm \text{sign}\,(\lambda) : \lambda \in \text{spec}\,(H)\}\,, \tag{7.99}$$

$$\text{spec}\,\left((H - \text{sign}\,(H))(H + \text{sign}\,(H))^{-1}\right) = \left\{\frac{\lambda - \text{sign}\,(\lambda)}{\lambda - \text{sign}\,(\lambda)} : \lambda \in \text{spec}\,(H)\right\}\,. \tag{7.100}$$

Hieraus folgen (7.93) und (7.95). Zum Beweis von (7.96): Aus (7.91) für $k = 0$ folgt

$$F_0(Z_0 + \text{sign}\,(H)) = Z_0 - \text{sign}\,(H)\,, \tag{7.101}$$

also

$$(I + F_0)\text{sign}\,(H) = (I - F_0)\text{sign}\,(H)\,, \tag{7.102}$$

also, da $\rho(F_0) < 1$,

$$Z_0 = (I - F_0)^{-1}(I + F_0)\text{sign}\,(H)\,. \tag{7.103}$$

Induktionsschritt $k \to k+1$: (7.92) folgt aus

$$Z_{k+1}H = \frac{1}{2}(Z_k + Z_k^{-1})H = \frac{1}{2}H(Z_k + Z_k^{-1}) = HZ_{k+1}\,. \tag{7.104}$$

(7.93) folgt aus

$$\begin{aligned} Z_{k+1} + \text{sign}\,(H) &= \frac{1}{2}(Z_k + Z_k^{-1}) + \text{sign}\,(H) = \frac{1}{2}Z_k^{-1}\Big[Z_k^2 + I + 2Z_k\text{sign}\,(H)\Big] \\ &= \frac{1}{2}Z_k^{-1}\Big[Z_k + \text{sign}\,(H)\Big]^2\,. \end{aligned} \tag{7.105}$$

Eine analoge Rechnung liefert

$$Z_{k+1} - \text{sign}\,(H) = \frac{1}{2}Z_k^{-1}\Big[Z_k - \text{sign}\,(H)\Big]^2\,. \tag{7.106}$$

Hieruas ergeben sich (7.94) und (7.95) wegen

$$\begin{aligned} F_{k+1} &= (Z_{k+1} - \text{sign}\,(H))(Z_{k+1} + \text{sign}\,(H))^{-1} \\ &= (Z_{k+1} - \text{sign}\,(H))^2(Z_{k+1} + \text{sign}\,(H))^{-2} = F_k^2\,. \end{aligned} \tag{7.107}$$

(7.96) folgt genauso wie für $k = 0$ (Index 0 durch $k+1$ ersetzen). (7.97) folgt aus (7.96), (7.98) aus (7.104). Damit ist die Induktion vollständig.
Da alle Z_k invertierbar sind, ist die Newton-Iteration wohldefiniert. Aus (7.94) und (7.95) folgt

$$\lim_{k\to\infty} F_k = 0\,, \tag{7.108}$$

und hieraus folgt wegen (7.96) die Behauptung (7.90). □

Bemerkung 7.25 (Skalierung in der Newtoniteration)
Anstatt

$$Z_{k+1} = \frac{1}{2}(Z_k + Z_k^{-1})\,, \quad Z_0 = H \tag{7.109}$$

setzen wir

$$Z_k = \alpha_k W_k\,, \quad \alpha_k > 0 \text{ geeignet }\,, \tag{7.110}$$

$$W_{k+1} = \frac{1}{2}(Z_k + Z_k^{-1})\,, \quad W_0 = H \in \mathbb{R}^{(2n,2n)}\,. \tag{7.111}$$

Der Skalierungsparameter α_k soll erreichen, daß die Eigenwerte von Z_k betragsmäßig möglichst nahe bei 1 sind. Eine Möglichkeit ist es, α_k als Lösung des Minimierungsproblems

$$\min_{\alpha>0} J(\alpha) = \sum_{i=1}^{2n}(\ln|\alpha\lambda_i|)^2\,, \tag{7.112}$$

zu wählen, wobei λ_i die Eigenwerte von W_k bezeichnet. Differenzieren und Nullsetzen ergibt

$$0 = J'(\alpha) = \sum_{i=1}^{2n} 2\frac{\ln|\alpha\lambda_i|}{\alpha} = \frac{2}{\alpha}\ln\left(\prod_{i=1}^{2n}\alpha|\lambda_i|\right)\,, \tag{7.113}$$

also

$$\alpha = \left(\prod_{i=1}^{2n}|\lambda_i|\right)^{-\frac{1}{2n}} = \Big|\det(W_k)\Big|^{-\frac{1}{2n}}\,. \tag{7.114}$$

Bemerkung 7.26 (Ausnutzen der Symmetrie)
Wenn wir die Riccati-Gleichung lösen wollen, ist die Startmatrix $Z_0 = H$ hamiltonsch. Wegen Lemma 7.12 sind dann alle Iterierten W_k, Z_k hamiltonsch, und es gilt mit J aus (7.40)

$$W_k^{-1} = (JW_k)^{-1} J\,, \tag{7.115}$$

wobei JW_k symmetrisch ist, sowie

$$|\det(JW_k)| = |\det J||\det W_k| = 1\,. \tag{7.116}$$

Algorithmus 7.27 (Berechnung der Matrixsignumfunktion)
Sei $H \in \mathbb{R}^{(2n,2n)}$ hamiltonsch, H habe keine rein imaginären Eigenwerte.

1. Setze $W_0 = H$.

2. Iteriere über k: Setze

$$\beta_k = \left|\det(JW_k)\right|^{\frac{1}{2n}}, \quad Z_k = \frac{1}{\beta_k} W_k\,. \tag{7.117}$$

   Berechne $Y_k = (JW_k)^{-1} J$ als Lösung von

$$(JW_k) Y_k = J\,, \tag{7.118}$$

   etwa mit dem Cholesky-Verfahren. Setze

$$W_{k+1} = \frac{1}{2}(Z_k + \beta_k Y_k)\,. \tag{7.119}$$

   Breche die Iteration ab, falls $\|W_{k+1} - Z_k\|$ klein ist.

Bemerkung 7.28
Das vorgestellte Verfahren (Lösung der Matrix-Riccati-Gleichung über die Matrixsignumfunktion von H) kann schlecht konditioniert sein, wenn H Eigenwerte nahe der imaginären Achse hat.

8 Identifizierung: Adaptive dynamische Beobachter

Problem 8.1 (Identifizierungsproblem)
Wir betrachten das Kontrollsystem

$$\dot{x} = Ax + Bu\,, \quad y = Cx\,, \quad A \in \mathbb{R}^{(n,n)}\,, B \in \mathbb{R}^{(n,m)}\,, C \in \mathbb{R}^{(k,n)}\,. \tag{8.1}$$

Gesucht sind die Übertragungsfunktion G bzw. die Impulsantwortmatrix K. Nicht bekannt sind die Matrizen A, B, C, wir nehmen aber an, daß n, m, k bekannt sind. Zur Verfügung stehen gewisse Informationen über das Input-Output-Verhalten (etwa ein zusammengehöriges Paar $(u, y) : \mathbb{R}_+ \to \mathbb{R}^m \times \mathbb{R}^k$). Wir werden uns beschränken auf single-input-single-output-(SISO-)Systeme, d.h. $m = k = 1$.

Bemerkung 8.2 (Approximationsproblem im Zeitbereich)
Wir betrachten das Beispiel

$$\dot{x} = -ax + bu\,, \quad y = x\,, \quad a > 0\,, b \in \mathbb{R}\,. \tag{8.2}$$

Es ist

$$G(s) = \frac{b}{s+a}\,, \quad K(t) = be^{-at}\,. \tag{8.3}$$

Ist etwa $x(0) = 0$ und $u = 1$, so hat der zugehörige Output (die sogenannte Sprungantwort) die Form

$$y(t) = \frac{b}{a}(1 - e^{-at})\,. \tag{8.4}$$

Messung von y zu N verschiedenen (sinnvollerweise $N > 2$) Zeitpunkten ergibt ein überbestimmtes Gleichungssystem, aus dem die Parameter a und b durch Lösung eines Ausgleichsproblems (z.B. least squares) bestimmt werden. Für $n > 1$ ergibt sich das Problem, aus gegebenen Funktionswerten $y_i = y(t_i)$, $1 \le i \le N$, die Parameter a_i, b_i im Ansatz

$$y(t) = \sum_{i=1}^{n} b_i e^{-a_i t} \tag{8.5}$$

zu bestimmen. Die Bestimmung der b_i ist unproblematisch, falls die a_i (d.h. die Eigenwerte des Systems) bekannt sind; die Bestimmung der a_i führt allerdings auf schlecht konditionierte nichtlineare Approximationsprobleme.

Bemerkung 8.3 (Approximationsproblem im Frequenzbereich)
Man kann auch ausnutzen, daß Inputs $u(t) = \sin \omega t$ vermittels der Beziehung

$$\lim_{t\to\infty}(y(t) - r\sin(\omega t + \varphi)) = 0\,, \quad G(i\omega) = re^{i\varphi}\,, \tag{8.6}$$

die Bestimmung einzelner Funktionswerte von G ermöglicht. (Pro Funktionswert ist ein Input-Output-Paar erforderlich.) Es ergibt sich das Problem, die Koeffizienten einer rationalen Funktion zu bestimmen.

Den beiden in 8.2 und 8.3 beschriebenen Ansätzen ist gemeinsam, daß die gesuchten Systemparameter mit einem numerischen Verfahren separat bestimmt werden. Im Gegensatz dazu ist es das Ziel des im folgenden dargestellten Ansatzes, das dynamische System so zu erweitern, daß gewisse hinzugefügte Zustandskomponenten gegen die Werte der gesuchten Parameter konvergieren.

Beispiel 8.4 (Konstruktion eines adaptiven Beobachters, $n = 1$)
Wir betrachten wieder

$$\dot{x} = -ax + bu\,, \quad y = x\,, \quad a, b \in \mathbb{R}\,, \tag{8.7}$$

mit der Übertragungsfunktion

$$G(s) = \frac{b}{s+a}\,, \tag{8.8}$$

wobei a und b unbekannt sind. Wir nehmen an, daß zum Zeitpunkt t die Werte $u(t)$ und $y(t)$ zur Verfügung stehen. Wir wählen ein $\lambda > 0$ und definieren einen Beobachter mit dem Zustand $w = (w_1, w_2)$ und dem Output y_I durch

$$\dot{w}_1 = -\lambda w_1 + u\,, \tag{8.9}$$

$$\dot{w}_2 = -\lambda w_2 + y\,, \tag{8.10}$$

$$y_I = \theta_1 w_1 + \theta_2 w_2\,, \quad \theta_1\,,\, \theta_2 \in \mathbb{R}\,. \tag{8.11}$$

Es gilt

$$\begin{aligned} \hat{y}_I(s) &= \theta_1 \hat{w}_1(s) + \theta_2 \hat{w}_2(s) = \frac{\theta_1}{s+\lambda}\hat{u}(s) + \frac{\theta_2}{s+\lambda}\hat{y}(s) \\ &= \left(\frac{\theta_1}{s+\lambda} + \frac{\theta_2}{s+\lambda}\cdot\frac{b}{s+a}\right)\hat{u}(s)\,. \end{aligned} \tag{8.12}$$

Falls

$$\theta_1 = b\,, \quad \theta_2 = \lambda - a\,, \tag{8.13}$$

so gilt für die Übertragungsfunktion G_I des Beobachters

$$G_I(s) = \frac{b}{s+a} = G(s)\,. \tag{8.14}$$

Da die richtigen Parameter aus (8.13) unbekannt sind, setzen wir θ_1, θ_2 als zeitabhängige Funktionen an und suchen nach einer Adaptionsregel

$$\dot{\theta}_i = g_i(t, u, w, y, y_I)\,, \tag{8.15}$$

so daß

$$\lim_{t\to\infty} \theta_1(t) = b\,, \quad \lim_{t\to\infty} \theta_2(t) = \lambda - a\,. \tag{8.16}$$

Wir betrachten $e : \mathbb{R}^2 \times \mathbb{R}_+ \to \mathbb{R}$,

$$e(\theta, t) = \theta_1 w_1(t) + \theta_2 w_2(t) - y(t)\,. \tag{8.17}$$

Es ist

$$e(\theta(t), t) = y_I(t) - y(t)\,. \tag{8.18}$$

Eine Idee ist nun, die Adaptionsregel zu definieren in Analogie zum Gradientenverfahren, angewendet auf $J : \mathbb{R}^2 \to \mathbb{R}$,

$$J(\theta) = \frac{1}{2}e(\theta, t)^2\,. \tag{8.19}$$

Es ergibt sich der Ansatz

$$\dot{\theta} = -\gamma \nabla J(\theta) = -\gamma e(\theta, t) \frac{\partial e}{\partial \theta}(\theta, t)\,, \quad \gamma > 0\,, \tag{8.20}$$

also

$$\begin{pmatrix} \dot{\theta}_1(t) \\ \dot{\theta}_2(t) \end{pmatrix} = -\gamma (y_I(t) - y(t)) \begin{pmatrix} w_1(t) \\ w_2(t) \end{pmatrix}. \tag{8.21}$$

Bemerkung 8.5
Die Konvergenz des Verfahrens ist nicht offensichtlich und auch nicht immer zu erwarten, z.B. nicht für $u = 0$. Die Gleichungen (8.9) - (8.11) stellen bei Parameterwahl gemäß (8.13) eine Realisierung der Übertragungsfunktion dar, bei der die zu bestimmenden Parameter in den Vektor c verlagert sind. Der konstruierte Beobachter ist stabil.

Durch die zweite Gleichung $y = x$ in (8.7) ist die minimale Realisierung von (8.8) bereits eindeutig festgelegt. Ist der Mc-Millan-Grad n der Übertragungsfunktion größer als 1, so gibt es eine große Vielfalt minimaler Realisierungen (die durch Ähnlichkeitstransformationen auseinander gewonnen werden). Ein adaptiver Beobachter kann natürlich zwischen verschiedenen minimalen Realisierungen nicht unterscheiden. Er identifiziert die Parameter einer Normalform, die wir uns entweder als Koeffizienten des Zähler- und des Nennerpolynoms der Übertragungsfunktion oder als Parameter der (hier: skalaren) Regelungsnormalform vorstellen können. Wir betrachten ein SISO-System

$$y(t) = \int_0^t K(t-\tau) u(\tau)\, d\tau\,, \quad \hat{y}(s) = G(s)\hat{u}(s)\,, \quad G = \hat{K}\,. \tag{8.22}$$

G ist unbekannt, aber zum Zeitpunkt t stehen die Werte $u(t)$ und $y(t)$ zur Verfügung.

Voraussetzung 8.6
Das zu identifizierende System sei ein SISO-Kontrollsystem vom Mc-Millan-Grad n mit der Übertragungsfunktion

$$G(s) = \frac{p(s)}{q(s)}\,, \tag{8.23}$$

wobei p und q teilerfremde Polynome der Form

$$p(s) = \sum_{k=0}^{n-1} \beta_{k+1} s^k\,, \quad q(s) = s^n + \sum_{k=0}^{n-1} \alpha_{k+1} s^k\,, \tag{8.24}$$

sind.

Bemerkung 8.7 (Konstruktion eines adaptiven Beobachters)
Wir wählen ein Stabilitätspolynom (d.h. alle Nullstellen haben negativen Realteil)

$$q_I(s) = s^n + \sum_{k=0}^{n-1} \alpha_{k+1}^I s^k\,, \tag{8.25}$$

und definieren einen Beobachter mit dem Zustand $(w_1, w_2) \in \mathbb{R}^n \times \mathbb{R}^n$ und dem skalaren Output y_I durch

$$\dot{w}_1 = A_I\, w_1 + b_I\, u\,, \tag{8.26}$$

$$\dot{w}_2 = A_I\, w_2 + b_I\, y\,, \tag{8.27}$$

$$y_I = \theta^T w = \theta_1^T w_1 + \theta_2^T w_2 \,, \tag{8.28}$$

wobei

$$A_I = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ -\alpha_1^I & -\alpha_2^I & \cdots & \cdots & -\alpha_n^I \end{pmatrix}, \quad b_I = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \quad \theta_i = \begin{pmatrix} \theta_{i1} \\ \theta_{i2} \\ \vdots \\ \vdots \\ \theta_{in} \end{pmatrix}, \; i = 1, 2 \,. \tag{8.29}$$

Ist θ konstant, so wird das Input-Output-Verhalten der beiden Teilsysteme beschrieben durch (siehe den Satz über die Realisierung skalarer Übertragungsfunktionen)

$$\widehat{\theta_1 w_1}(s) = G_1(s)\hat{u}(s) \,, \quad G_1 = \frac{p_1}{q_I} \,, \tag{8.30}$$

$$\widehat{\theta_2 w_2}(s) = G_2(s)\hat{y}(s) \,, \quad G_2 = \frac{p_2}{q_I} \,, \tag{8.31}$$

wobei

$$p_i(s) = \sum_{k=0}^{n-1} \theta_{i,k+1} s^k \,, \quad i = 1, 2 \,. \tag{8.32}$$

Sei G_I die Übertragungsfunktion des Beobachters, d.h.

$$\hat{y}_I(s) = G_I(s)\hat{u}(s) \,. \tag{8.33}$$

Es gilt dann (θ ist immer noch konstant)

$$G_I = G_1 + G_2 \cdot G = \frac{p_1}{q_I} + \frac{p_2}{q_I} \cdot \frac{p}{q} = \frac{p_1 q + p p_2}{q_I \, q} \,, \tag{8.34}$$

also ist

$$G_I = G \,, \tag{8.35}$$

falls

$$p_1 = p \,, \quad p_2 = q_I - q \,. \tag{8.36}$$

Um die Identifizierung durchzuführen, setzen wir wieder θ_1, θ_2 als zeitabhängige Funktionen an und suchen nach einer Adaptionsregel

$$\dot{\theta}_i = g_i(t, u, w, y, y_I) \,, \tag{8.37}$$

so daß die nun mit t parametrisierten Polynome p_1 und p_2 im Limes $t \to \infty$ die Gleichung (8.36) erfüllen, also

$$\lim_{t\to\infty} \theta_1(t) = \theta_1^* = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}, \quad \lim_{t\to\infty} \theta_2(t) = \theta_2^* = \begin{pmatrix} \alpha_1^I - \alpha_1 \\ \vdots \\ \alpha_n^I - \alpha_n \end{pmatrix}. \tag{8.38}$$

Da die α_k^I aus der Konstruktion des Beobachters bekannt sind, liefert der Limesvektor $\theta^* = (\theta_1^*, \theta_2^*)$ alle Parameter der gesuchten Übertragungsfunktion G, falls (8.38) gilt.

Lemma 8.8 *Sei* $u \in L^\infty_{loc}(\mathbb{R}_+;\mathbb{R}^m)$, *sei* $y:\mathbb{R}_+\to\mathbb{R}$ *der zugehörige Output des Kontrollsystems*

$$\dot{x} = Ax + bu\,, \quad x(0) = 0\,, \quad y = c^T x\,, \tag{8.39}$$

sei der Beobachter (8.26) - (8.28) gemäß 8.6 und 8.7 konstruiert mit $\theta = \theta^*$ *aus (8.38). Dann gilt für den Output* y_I^* *des Beobachters zum Anfangswert* $w(0) = 0$

$$y_I^*(t) = \theta^{*T} w(t) = y(t)\,. \tag{8.40}$$

Beweis: Sei K die Impulsantwort von (8.39), sei K_I^* die Impulsantwort des Gesamtsystems (8.39), (8.26) - (8.28) für $\theta = \theta^*$, also

$$y(t) = \int_0^t K(t-\tau)u(\tau)\,d\tau\,, \quad y_I^*(t) = \int_0^t K_I^*(t-\tau)u(\tau)\,d\tau\,. \tag{8.41}$$

Nach Konstruktion gilt

$$\hat{K}_I^* = G_I^* = G = \hat{K}\,, \tag{8.42}$$

also auch $K_I^* = K$. □

Definition 8.9
Seien $y, z : \mathbb{R}_+ \to \mathbb{R}$. Wir sagen, daß z exponentiell gegen y konvergiert, wenn es Konstante $c_1, c_2 > 0$ gibt mit

$$|z(t) - y(t)| \le c_1 e^{-c_2 t}\,, \qquad \text{für alle } t > 0\,. \tag{8.43}$$

Bemerkung 8.10
Falls in Lemma 8.8 die Anfangswerte nicht Null sind, so gilt statt (8.40), daß y_I^* exponentiell gegen y konvergiert. (Für $w(0) \neq 0$ folgt dies aus der Stabilität des Beobachters, für $x(0) \neq 0$ aus ...).

Definition 8.11
Sei (u, y) ein Input-Output-Paar, sei $\theta : \mathbb{R}_+ \to \mathbb{R}^{2n}$. Wir definieren den Parameterfehler $\varphi : \mathbb{R}_+ \to \mathbb{R}^{2n}$ und den Beobachterfehler $e : \mathbb{R}_+ \to \mathbb{R}$ als

$$\varphi(t) = \theta(t) - \theta^*\,, \quad e(t) = y_I(t) - y(t)\,. \tag{8.44}$$

Die Adaptionsregel

$$\dot{\theta} = -\gamma \frac{ew}{1 + \kappa w^T w}\,, \tag{8.45}$$

mit den Konstanten $\gamma > 0$ und $\kappa \ge 0$ heißt Gradientenregel für $\kappa = 0$ und normalisierte Gradientenregel für $\kappa > 0$.

Lemma 8.12 *Sei* w *Lösung von (8.26) - (8.27), sei* θ *Lösung von (8.45). Der Beobachterfehler hat die Form*

$$e(t) = w(t)^T \varphi(t) + \varepsilon(t)\,, \quad t \ge 0\,, \tag{8.46}$$

wobei ε *exponentiell gegen 0 konvergiert. Der Parameterfehler ist eine Lösung von*

$$\dot{\varphi} = -\gamma \frac{ww^T\varphi}{1 + \kappa w^T w} - \gamma\varepsilon \frac{w}{1 + \kappa w^T w}\,. \tag{8.47}$$

Beweis: Folgt aus den Definitionen, Lemma 8.8 und Bemerkung 8.10. □

Bemerkung 8.13
Für die normalisierte Gradientenregel ist die rechte Seite von (8.47) global Lipschitz-stetig in φ.

Satz 8.14 *Sei $w : \mathbb{R}_+ \to \mathbb{R}^N$ meßbar und beschränkt auf kompakten Intervallen, sei $\varepsilon : \mathbb{R}_+ \to \mathbb{R}$ exponentiell gegen 0 konvergent. Dann hat (8.47) zu jedem Anfangswert $\varphi(0)$ eine eindeutige Lösung $\varphi \in L^\infty(\mathbb{R}_+; \mathbb{R}^N)$. Für das Gradientenverfahren gilt darüber hinaus*

$$e \in L^2(\mathbb{R}_+)\,, \tag{8.48}$$

und für das normalisierte Gradientenverfahren

$$\dot\varphi\,,\ \frac{e}{\sqrt{1+\kappa w^T w}}\,,\ \frac{\varphi^T w}{1+\|w\|_{t,\infty}}\ \in L^2(\mathbb{R}_+)\cap L^\infty(\mathbb{R}_+)\,, \tag{8.49}$$

wobei

$$\|w\|_{t,\infty} = \sup_{0\le\tau\le t}\|w(\tau)\|\,. \tag{8.50}$$

Beweis: Wir definieren $V : \mathbb{R}_+ \times \mathbb{R}^N$ durch

$$V(t,\varphi) = \frac{1}{2}\varphi^T\varphi + \frac{\gamma}{4}\int_t^\infty \frac{\varepsilon(\tau)^2}{1+\kappa w(\tau)^T w(\tau)}\,d\tau\,. \tag{8.51}$$

Aus den Voraussetzungen an w und ε folgt, daß jede Lösung φ von (8.47) sich auf $\mathbb{R}_+$ fortsetzen läßt. Für

$$v(t) = V(t,\varphi(t))\,,\quad t\in\mathbb{R}_+\,, \tag{8.52}$$

gilt dann

$$\begin{aligned}
\dot v(t) &= \dot\varphi(t)^T\varphi(t) - \frac{\gamma}{4}\frac{\varepsilon(t)^2}{1+\kappa w(\tau)^T w(\tau)}\\
&= -\gamma\left(\frac{(w^T\varphi)^2}{1+\kappa w^T w} + \frac{w^T\varphi}{1+\kappa w^T w}\varepsilon + \frac{\varepsilon(t)^2}{4(1+\kappa w^T w)}\right)\\
&= -\gamma\left(\frac{w^T\varphi}{\sqrt{1+\kappa w^T w}} + \frac{\varepsilon}{2\sqrt{1+\kappa w^T w}}\right)^2\,,
\end{aligned} \tag{8.53}$$

also

$$v(t)\ge 0\,,\quad \dot v(t)\le 0\,, \tag{8.54}$$

also ist v und damit auch φ beschränkt. Da v monoton fällt, existiert

$$v(\infty) = \lim_{t\to\infty} v(t)\,, \tag{8.55}$$

und es folgt weiter

$$0\le\int_0^\infty\left(\frac{w^T(t)\varphi(t)}{\sqrt{1+\kappa w(t)^T w(t)}} + \frac{\varepsilon(t)}{2\sqrt{1+\kappa w(t)^T w(t)}}\right)^2 dt = -\frac{1}{\gamma}(v(\infty)-v(0)) \le \frac{1}{\gamma}v(0)\,. \tag{8.56}$$

Hieraus folgen alle Aussagen, da $\varepsilon\in L^2\cap L^\infty$. □

Lemma 8.15 *Sei* $z \in L^p(\mathbb{R}_+)$ *für ein* $p \in [1, \infty)$, *sei* $\dot{z} \in L^\infty(\mathbb{R}_+)$. *Dann gilt*

$$\lim_{t\to\infty} z(t) = 0\,. \tag{8.57}$$

Beweis: Sei $p = 1$. Für jedes $t \geq 0$ gilt

$$\int_t^\infty |z(\tau)|\, d\tau \geq \frac{|z(t)|^2}{2\,\|\dot{z}\|_\infty}\,. \tag{8.58}$$

Da $z \in L^1(\mathbb{R}_+)$, konvergiert die linke Seite gegen 0 für $t \to \infty$. Für $p > 1$ genügt es, das eben Bewiesene auf die Funktion $t \mapsto |z(t)|^p$ anzuwenden. □

Satz 8.16 (Stabilität des Identifizierers)
Sei das Input-Output-Paar (u, y) *beschränkt auf* $\mathbb{R}_+$, *sei als Adaptionsregel die Gradientenregel oder die normalisierte Gradientenregel gewählt. Dann gilt für den Beobachterfehler*

$$e \in L^2(\mathbb{R}_+) \cap L^\infty(\mathbb{R}_+)\,, \quad \lim_{t\to\infty} e(t) = 0\,, \tag{8.59}$$

und für den Parameterfehler

$$\varphi \in L^\infty(\mathbb{R}_+; \mathbb{R}^{2n})\,,\ \dot{\varphi} \in L^2(\mathbb{R}_+; \mathbb{R}^{2n}) \cap L^\infty(\mathbb{R}_+; \mathbb{R}^{2n})\,, \quad \lim_{t\to\infty} \dot{\varphi}(t) = 0\,. \tag{8.60}$$

Beweis: Da der Beobachter stabil ist, folgt aus der Beschränktheit von u und y auch die Beschränktheit von w und $\dot{w}$. Aus Satz 8.14 folgt, daß φ und $\dot{\varphi}$ beschränkt sind, also auch e und $\dot{e}$. Wegen Lemma 8.15 gilt für $t \to \infty$ auch $e(t) \to 0$, also auch $w(t)^T\varphi(t) \to 0$ und damit auch $\dot{\varphi}(t) \to 0$. □

Bemerkung 8.17
Ist das zu identifizierende System stabil, so genügt es, die Beschränktheit von u anzunehmen, da dann y ebenfalls beschränkt ist.

Definition 8.18
Eine Funktion $w \in L^\infty_{loc}(\mathbb{R}_+; \mathbb{R}^N)$ heißt "persistently exciting", falls es $c_1, c_2, \delta > 0$ gibt, so daß

$$c_1 I \leq \int_t^{t+\delta} w(\tau)w(\tau)^T\, d\tau \leq c_2 I \tag{8.61}$$

für alle $t \geq 0$.

Lemma 8.19 *Sei* t *fest. Dann gilt (8.61) genau dann, wenn*

$$c_1 \leq \int_t^{t+\delta} (x^T w(\tau))^2\, d\tau \leq c_2 \tag{8.62}$$

gilt für alle $x \in \mathbb{R}^N$ *mit* $\|x\|_2 = 1$.

Beweis: Multipliziere (8.61) von rechts und links mit x. □

Lemma 8.20
Seien $c_1, c_2, \gamma, \delta > 0$, $t \geq 0$, *seien* $w, \varphi : [t, t+\delta] \to \mathbb{R}^N$ *mit*

$$\dot{\varphi} = -\gamma w w^T \varphi \tag{8.63}$$

auf $[t, t+\delta]$, *und es gelte*

$$c_1 I \leq \int_t^{t+\delta} w(\tau) w(\tau)^T \, d\tau \leq c_2 I \,. \tag{8.64}$$

Dann gilt

$$\int_t^{t+\delta} (w(\tau)^T \varphi(\tau))^2 \, d\tau \geq \frac{c_1}{(1+\gamma c_2 \sqrt{N})^2} \|\varphi(t)\|^2 \,. \tag{8.65}$$

Beweis: Wir definieren

$$z(\tau) = -\gamma w(\tau) w(\tau)^T \varphi(\tau) \,, \quad \tau \in [t, t+\delta] \,. \tag{8.66}$$

Es gilt dann

$$\begin{aligned}
\Big(w(\tau)^T(\varphi(\tau) - \varphi(t))\Big)^2 &= \left(\int_t^\tau w(\tau)^T z(\sigma) \, d\sigma\right)^2 \\
&\leq \left(\int_t^\tau w(\tau)^T \frac{z(\sigma)}{\|z(\sigma)\|} \gamma \|w(\sigma)\| |w(\sigma)^T \varphi(\sigma)| \, d\sigma\right)^2 \\
&\leq \gamma^2 \int_t^\tau (w(\sigma)^T \varphi(\sigma))^2 \, d\sigma \cdot \int_t^\tau \left(w(\tau)^T \frac{z(\sigma)}{\|z(\sigma)\|}\right)^2 \|w(\sigma)\|^2 \, d\sigma \,.
\end{aligned} \tag{8.67}$$

Mit Lemma 8.19 folgt weiter

$$\begin{aligned}
&\int_t^{t+\delta} \Big(w(\tau)^T(\varphi(\tau) - \varphi(t))\Big)^2 \, d\tau = \\
&\qquad \leq \gamma^2 \int_t^{t+\delta} (w(\sigma)^T \varphi(\sigma))^2 \, d\sigma \cdot \int_t^{t+\delta} \int_t^\tau \left(w(\tau)^T \frac{z(\sigma)}{\|z(\sigma)\|}\right)^2 \|w(\sigma)\|^2 \, d\sigma \, d\tau \\
&\qquad \leq \gamma^2 \int_t^{t+\delta} (w(\sigma)^T \varphi(\sigma))^2 \, d\sigma \cdot \int_t^{t+\delta} \|w(\sigma)\|^2 \int_\sigma^{t+\delta} \left(w(\tau)^T \frac{z(\sigma)}{\|z(\sigma)\|}\right)^2 d\tau \, d\sigma \\
&\qquad \leq \gamma^2 \int_t^{t+\delta} (w(\sigma)^T \varphi(\sigma))^2 \, d\sigma \cdot \int_t^{t+\delta} \|w(\sigma)\|^2 \, d\sigma \cdot c_2 \,.
\end{aligned} \tag{8.68}$$

Aus (8.61) folgt

$$c_1 \varphi(t)^T \varphi(t) \leq \int_t^{t+\delta} \Big(w(\tau)^T \varphi(t)\Big)^2 \, d\tau \,, \tag{8.69}$$

also

$$\begin{aligned}
\sqrt{c_1} \|\varphi(t)\| &\leq \left(\int_t^{t+\delta} \Big(w(\tau)^T \varphi(t)\Big)^2 \, d\tau\right)^{\frac{1}{2}} \\
&\leq \left(\int_t^{t+\delta} \Big(w(\tau)^T (\varphi(t) - \varphi(\tau))\Big)^2 \, d\tau\right)^{\frac{1}{2}} + \left(\int_t^{t+\delta} (w(\sigma)^T \varphi(\sigma))^2 \, d\sigma\right)^{\frac{1}{2}} \\
&\leq \left(\int_t^{t+\delta} (w(\sigma)^T \varphi(\sigma))^2 \, d\sigma\right)^{\frac{1}{2}} \left[1 + \gamma \left(c_2 \int_t^{t+\delta} \|w(\sigma)\|^2 \, d\sigma\right)^{\frac{1}{2}}\right] \,,
\end{aligned} \tag{8.70}$$

also

$$\int_t^{t+\delta} (w(\sigma)^T \varphi(\sigma))^2 \, d\sigma \geq \frac{c_1 \|\varphi(t)\|^2}{\left(1 + \gamma\sqrt{c_2}\left(\int_t^{t+\delta} \|w(\sigma)\|^2 \, d\sigma\right)^{\frac{1}{2}}\right)^2} \,. \tag{8.71}$$

Es gilt außerdem (e_i = i-ter Einheitsvektor)

$$\int_t^{t+\delta} \|w(\sigma)\|^2 \, d\sigma = \int_t^{t+\delta} \sum_{i=1}^{N} (w(\sigma)^T e_i)^2 \, d\sigma = \sum_{i=1}^{N} \int_t^{t+\delta} (w(\sigma)^T e_i)^2 \, d\sigma \leq N c_2 \,. \tag{8.72}$$

Setzt man die beiden letzten Ungleichungen zusammen, so folgt die Behauptung. □

Satz 8.21 *Wir betrachten den in 8.7 konstruierten Beobachter mit der Adaptionsregel (8.45), d.h. mit der Gradientenregel oder der normalisierten Gradientenregel. Sei* (u, y) *ein auf* $\mathbb{R}_+$ *beschränktes Input-Output-Paar des zu identifizierenden Systems und* $w : \mathbb{R}_+ \to \mathbb{R}^{2n}$ *der zugehörige Zustand des Beobachters. Alle Anfangswerte seien Null. Sei* w *persistently exciting. Dann konvergiert der Parameterfehler* φ *exponentiell gegen 0.*

Beweis: Wir setzen

$$\tilde{w}(t) = \frac{w(t)}{\sqrt{1 + \kappa w(t)^T w(t)}} \,. \tag{8.73}$$

Dann ist

$$\dot{\varphi}(t) = -\gamma \tilde{w}(t) \tilde{w}(t)^T \,. \tag{8.74}$$

Da (u, y) beschränkt ist, ist wegen der Stabilität von A_I auch w auf $\mathbb{R}_+$ beschränkt, und es gibt $\delta, \tilde{c}_1, \tilde{c}_2 > 0$ mit

$$\begin{aligned} \tilde{c}_1 &\leq \frac{1}{1 + \kappa \|w\|_\infty} \int_t^{t+\delta} (x^T w(\tau))^2 \, d\tau \leq \int_t^{t+\delta} (x^T \tilde{w}(\tau))^2 \, d\tau \\ &\leq \int_t^{t+\delta} (x^T w(\tau))^2 \, d\tau \leq \tilde{c}_2 \,, \end{aligned} \tag{8.75}$$

für alle $x \in \mathbb{R}^{2n}$ mit $\|x\|_2 = 1$. Also ist auch $\tilde{w}$ persistently exciting, und aus Satz 8.20 folgt für alle $t \geq 0$

$$\int_t^{t+\delta} (\tilde{w}(\tau)^T \varphi(\tau))^2 \, d\tau \geq \frac{\tilde{c}_1}{(1 + \gamma \tilde{c}_2 \sqrt{2n})^2} \|\varphi(t)\|_2^2 \,. \tag{8.76}$$

Wir setzen

$$v(t) = \frac{1}{2} \varphi(t)^T \varphi(t) \,. \tag{8.77}$$

Es ist dann

$$\dot{v}(t) = \dot{\varphi}(t)^T \varphi(t) = -\gamma (\tilde{w}(t)^T \varphi(t))^2 \leq 0 \,, \tag{8.78}$$

also

$$v(t + \delta) - v(t) \leq -\frac{2\tilde{c}_1}{\gamma(1 + \gamma \tilde{c}_2 \sqrt{2n})^2} v(t) \,, \tag{8.79}$$

also

$$v(t + \delta) \leq (1 - c) v(t) \leq e^{-c} v(t) \,, \quad c = \frac{2\tilde{c}_1}{\gamma(1 + \gamma \tilde{c}_2 \sqrt{2n})^2} \,, \tag{8.80}$$

falls $c < 1$ (andernfalls ist $v(t+\delta) = 0$, da $v \geq 0$). Es folgt also für $t = n\delta + \tau$, $n \in \mathbb{N}$, $0 \leq \tau < t$,

$$v(t) \leq e^{-nc} v(\tau) = e^{-\frac{c}{\delta}t} e^{\frac{c\tau}{\delta}} \sup_{0 \leq s \leq \delta} |v(s)| . \tag{8.81}$$

Also konvergiert v und damit auch φ exponentiell gegen 0. □

Es erhebt sich die Frage, welche Inputs u zu einem Beobachterzustand w mit der gewünschten Eigenschaft "persistently exciting" führen.

Lemma 8.22 (Stationäres Signal) *Sei $z \in L^\infty(\mathbb{R}_+; \mathbb{R}^N))$. Existiert*

$$\lim_{T \to \infty} \frac{1}{T} \int_{t_0}^{t_0+T} z(\tau) z(t+\tau)^T \, d\tau \tag{8.82}$$

für ein $t_0 \in \mathbb{R}$, so existiert er für alle $t_0 \in \mathbb{R}$ und ist unabhängig von t_0. (Wir setzen $z = 0$ auf $\mathbb{R}_-$.) Ist dieser Limes darüber hinaus gleichmäßig in t_0, so heißt z stationär, und die Funktion $R_z : \mathbb{R} \to \mathbb{R}^{N,N}$,

$$R_z(t) = \lim_{T \to \infty} \frac{1}{T} \int_{t_0}^{t_0+T} z(\tau) z(t+\tau)^T \, d\tau , \tag{8.83}$$

heißt Autokovarianz von z.

Beweis: Für $t, T, t_0, t_1 \in \mathbb{R}$ gilt

$$\begin{aligned} & \frac{1}{T} \int_{t_1}^{t_1+T} z(\tau) z(t+\tau)^T \, d\tau - \frac{1}{T} \int_{t_0}^{t_0+T} z(\tau) z(t+\tau)^T \, d\tau = \\ = \; & \frac{1}{T} \left(\int_{t_1}^{t_0} + \int_{t_0+T}^{t_1+T} \right) z(\tau) z(t+\tau)^T \, d\tau , \end{aligned} \tag{8.84}$$

und die rechte Seite ist beschränkt durch

$$\frac{1}{T} 2 |t_1 - t_0| \|z\|_\infty^2 . \tag{8.85}$$

□

Definition 8.23 *Sei $z \in L^\infty(\mathbb{R}_+; \mathbb{R}^N)$ stationär, sei $\mu = (\mu_{jk})_{1 \leq j,k \leq N}$ eine Matrix beschränkter Maße auf $\mathbb{R}$. Falls*

$$R_z(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega t} \, d\mu(\omega) , \tag{8.86}$$

gilt, heißt μ Spektralmaß von z.

Bemerkung 8.24
Formel (8.86) besagt, daß das Spektralmaß μ die inverse Fouriertransformierte der Autokovarianz R_z und damit R_z die Fouriertransformierte von μ ist. Die Autokovarianz R_z ist stetig, und es gilt

$$R_z(-t) = R_z(t)^T . \tag{8.87}$$

Ein auf Bochner zurückgehender Satz besagt, daß jede stationäre Funktion ein Spektralmaß besitzt.

Beispiel 8.25
Sei $N = 1$, $z(t) = \sin \omega t$. Dann gilt

$$\begin{aligned} \frac{1}{T}\int_{t_0}^{t_0+T} z(\tau)z(t+\tau)\,d\tau &= \frac{1}{T}\int_{t_0}^{t_0+T} \sin(\tau)\sin(t+\tau)\,d\tau \\ &= \frac{1}{T}\int_{t_0}^{t_0+T} \frac{1}{2}\cos(\omega t) - \frac{1}{2}\cos(\omega(t+2\tau))\,d\tau \\ &= \frac{1}{2}\cos(\omega t) - \frac{1}{4\omega T}\sin(\omega(t+2\tau))\Big|_{t_0}^{t_0+T}, \end{aligned} \tag{8.88}$$

also ist z stationär und hat die Autokovarianz

$$R_z(t) = \frac{1}{2}\cos(\omega t)\,, \tag{8.89}$$

und das Spektralmaß

$$\mu = \frac{\pi}{2}(\delta_\omega + \delta_{-\omega})\,, \tag{8.90}$$

wobei δ_ω das Dirac-Maß im Punkt ω bezeichnet. Allgemeiner gilt: Sind ω_j, $1 \le j \le J$, paarweise verschiedene nichtnegative Zahlen, so ist die Funktion

$$z(t) = \sum_{j=1}^{J} \alpha_j \sin(\omega_j t)\,, \quad \alpha_j \in \mathbb{R}\,, \tag{8.91}$$

stationär mit der Autokovarianz

$$R_z(t) = \sum_{j=1}^{J} \frac{1}{2}\alpha_j^2 \cos(\omega_j t)\,, \tag{8.92}$$

und dem Spektralmaß

$$\mu = \sum_{j=1}^{J} \frac{\pi}{2}\alpha_j^2(\delta_{\omega_j} + \delta_{-\omega_j})\,. \tag{8.93}$$

(Beweis: Übung.)

Satz 8.26
Sei (u, y) ein Input-Output-Paar zur Impulsantwortmatrix $K : \mathbb{R}_+ \to \mathbb{R}^{(k,m)}$, d.h.

$$y(t) = \int_0^t K(\tau)u(t-\tau)\,d\tau\,, \quad t \ge 0\,. \tag{8.94}$$

Ist $u \in L^\infty(\mathbb{R}_+; \mathbb{R}^m)$ stationär und K integrierbar, so ist auch $y \in L^\infty(\mathbb{R}_+; \mathbb{R}^k)$ stationär und hat die Autokovarianz

$$R_y(t) = \int_0^\infty \int_0^\infty K(\tau_1)R_u(t+\tau_1-\tau_2)K(\tau_2)^T\,d\tau_2\,d\tau_1\,. \tag{8.95}$$

Beweis: Wir vereinbaren $K(t) = u(t) = y(t) = 0$ für $t < 0$. Dann gilt (8.94) für alle $t \in \mathbb{R}$, und weiter für alle t_0, T

$$\frac{1}{T}\int_{t_0}^{t_0+T} y(\tau)y(t+\tau)^T\,d\tau =$$

$$= \frac{1}{T}\int_{t_0}^{t_0+T}\int_0^\infty K(\tau_1)u(\tau-\tau_1)\,d\tau_1\cdot\int_0^\infty u(t+\tau-\tau_2)^T K(\tau_2)^T\,d\tau_2\,d\tau$$
$$= \int_0^\infty\int_0^\infty K(\tau_1)\left[\frac{1}{T}\int_{t_0}^{t_0+T} u(\tau-\tau_1)u(t+\tau-\tau_2)^T\,d\tau\right]K(\tau_2)^T\,d\tau_2\,d\tau_1\,.$$
$$= \int_0^\infty\int_0^\infty K(\tau_1)\underbrace{\left[\frac{1}{T}\int_{t_0-\tau_1}^{t_0-\tau_1+T} u(\sigma)u(t+\sigma+\tau_1-\tau_2)^T\,d\sigma\right]}_{=:f(T,\tau_1,\tau_2)}K(\tau_2)^T\,d\tau_2\,d\tau_1\,. \quad (8.96)$$

Da der Ausdruck in eckigen Klammern wegen

$$\|f(T,\tau_1,\tau_2)\| \le \|u\|_\infty^2 \quad (8.97)$$

gleichmäßig beschränkt ist, können wir nach dem Satz von Lebesgue den Grenzübergang $T\to\infty$ mit den Integralen vertauschen und erhalten die Behauptung. □

Bemerkung 8.27
Ist K die Impulsantwortmatrix eines Kontrollsystems (A,B,C) und ist A Stabilitätsmatrix, so ist K integrierbar, d.h. Satz 8.26 läßt sich auf stabile Kontrollsysteme anwenden.

Satz 8.28 *Es liege die Situation von Satz 8.26 vor mit $m=1$ (skalare Kontrolle), sei μ Spektralmaß von u. Dann gilt*

$$R_y(t) = \frac{1}{2\pi}\int_{-\infty}^\infty e^{i\omega t}\overline{G(i\omega)}G(i\omega)^T\,d\mu(\omega)\,,\quad G=\hat{K}\,. \quad (8.98)$$

Beweis: Einsetzen von

$$R_u(t) = \frac{1}{2\pi}\int_{-\infty}^\infty e^{i\omega t}\,d\mu(\omega)\,, \quad (8.99)$$

in (8.95) ergibt

$$\begin{aligned} R_y(t) &= \frac{1}{2\pi}\int_0^\infty\int_0^\infty K(\tau_1)\int_{-\infty}^\infty e^{i\omega(t+\tau_1-\tau_2)}\,d\mu(\omega)K(\tau_2)^T\,d\tau_2\,d\tau_1 \\ &= \frac{1}{2\pi}\int_{-\infty}^\infty e^{i\omega t}\left(\int_0^\infty e^{i\omega\tau_1}K(\tau_1)\,d\tau_1\right)\left(\int_0^\infty e^{-i\omega\tau_2}K(\tau_2)^T\,d\tau_2\right)\,d\mu(\omega) \\ &= \frac{1}{2\pi}\int_{-\infty}^\infty e^{i\omega t}\hat{K}(-i\omega)\hat{K}(i\omega)^T\,d\mu(\omega)\,. \end{aligned} \quad (8.100)$$

Da K reell ist, ist

$$G(-i\omega) = \int_0^\infty e^{i\omega t}K(t)\,dt = \overline{G(i\omega)}\,. \quad (8.101)$$

□

Definition 8.29 *Sei $u\in L^\infty(\mathbb{R}_+)$ stationär, u habe das Spektralmaß μ. Wir sagen, daß u den Spektralwert $\omega\in\mathbb{R}$ enthält, falls es ein $\lambda>0$ gibt mit*

$$\mu \ge \lambda\delta_\omega\,. \quad (8.102)$$

Bemerkung 8.30
Mit (8.102) ist gemeint, daß $\mu-\lambda\delta_\omega$ ein (nichtnegatives) Maß definiert. Äquivalent dazu ist, daß

$$\int_{-\infty}^\infty f(x)\,d\mu(x) \ge \lambda f(\omega) \quad (8.103)$$

für jede stetige beschränkte Funktion $f\ge 0$ gilt.

Lemma 8.31 *Wir betrachten die Übertragungsfunktion G_w des Beobachterzustandes,*

$$G_w(s) = \begin{pmatrix} (sI - A_I)^{-1} b_I \\ (sI - A_I)^{-1} b_I G(s) \end{pmatrix} \in \mathbb{C}^{2n} . \tag{8.104}$$

Es gilt: Sind $\omega_1, \ldots, \omega_{2n} \in \mathbb{R}$ paarweise verschieden, so sind die Vektoren $G_w(i\omega_j)$, $1 \le j \le 2n$, linear unabhängig im $\mathbb{C}^{2n}$.

Beweis: Andernfalls gibt es ein $\theta \in \mathbb{C}^{2n}$ mit $\theta \neq 0$ und

$$\theta^T G_w(i\omega_j) = 0 , \quad 1 \le j \le 2n . \tag{8.105}$$

Wir betrachten den zu diesem θ gehörenden Beobachteroutput

$$y_I = \theta^T w . \tag{8.106}$$

Nach Konstruktion des Beobachters gilt (siehe (8.34)

$$G_I = \theta^T G_w = G_1 + G_2 \cdot G = \frac{p_1}{q_I} + \frac{p_2}{q_I} \cdot \frac{p}{q} = \frac{p_1 q + p p_2}{q_I \, q} . \tag{8.107}$$

Der Zähler hat nach (8.105) $2n$ Nullstellen, aber höchstens den Grad $2n - 1$. Es folgt

$$p_1 q + p p_2 = 0 . \tag{8.108}$$

Nach Voraussetzung sind p und q teilerfremd. Es gilt also: q teilt p_2. Dies aber ist nicht möglich, da q den Grad n und p_2 einen Grad kleiner als n hat. □

Satz 8.32 *Sei $u \in L^\infty(\mathbb{R}_+)$ ein stationärer Input mit Spektralmaß μ, welcher mindestens $2n$ verschiedene Spektralwerte enthält, sei $w : \mathbb{R}_+ \to \mathbb{R}^{2n}$ der zugehörige Zustand des Beobachters. Dann ist die Matrix $R_w(0)$ positiv definit.*

Beweis: Sei $\theta \in \mathbb{R}^{2n}$, $\theta \neq 0$. Aus Satz 8.28 folgt

$$\theta^T R_w(0) \theta = \frac{1}{2\pi} \theta^T \int_{-\infty}^{\infty} \overline{G_w(i\omega)} G_w(i\omega)^T \, d\mu(\omega) \, \theta^T = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\theta^T G_w(i\omega)|^2 \, d\mu(\omega) . \tag{8.109}$$

Nach Voraussetzung gibt es Zahlen $\omega_j \in \mathbb{R}$, $\lambda_j > 0$, $1 \le j \le 2n$, mit $\mu \ge \lambda_j \delta_{\omega_j}$, also

$$\mu \ge \frac{1}{2n} \sum_{j=1}^{2n} \lambda_j \delta_{\omega_j} , \tag{8.110}$$

wobei die ω_j paarweise verschieden sind. Aus (8.109) folgt dann

$$\theta^T R_w(0) \theta \ge \frac{1}{2\pi} \frac{1}{2n} \sum_{j=1}^{2n} \lambda_j |\theta^T G_w(i\omega_j)|^2 \, d\mu(\omega) > 0 , \tag{8.111}$$

da nach Lemma 8.31 die Vektoren $G_w(i\omega_j)$ linear unabhängig sind. □

Satz 8.33 *Sei $u \in L^\infty(\mathbb{R}_+)$ ein stationärer Input mit Spektralmaß μ, welcher mindestens $2n$ verschiedene Spektralwerte enthält. Dann ist der zugehörige Zustand $w : \mathbb{R}_+ \to \mathbb{R}^{2n}$ des Beobachters "persistently exciting".*

Beweis: Wir wählen ε mit $0 < \varepsilon < \lambda_{\min}$, wobei $\lambda_{\min}$ der kleinste Eigenwert von $R_w(0)$ ist (möglich wegen Lemma 8.32). Wir wählen $\delta > 0$ so, daß

$$\Big\| \frac{1}{\delta} \int_t^{t+\delta} w(\tau) w(\tau)^T \, d\tau - R_w(0) \Big\| < \varepsilon \,, \qquad \text{für alle } t \geq 0 \,. \tag{8.112}$$

Es gilt dann für alle $x \in \mathbb{R}^{2n}$ mit $\|x\|_2 = 1$ und alle $t \geq 0$

$$\frac{1}{\delta} \int_t^{t+\delta} (x^T w(\tau))^2 \, d\tau \geq x^T R_w(0) x - \varepsilon \leq \lambda_{\min} - \varepsilon > 0 \,. \tag{8.113}$$

Andererseits gilt

$$\int_t^{t+\delta} (x^T w(\tau))^2 \, d\tau \leq 2n\delta \|w\|_\infty^2 \,. \tag{8.114}$$

□

Zusammenfassend gilt also

Satz 8.34 *Für den in 8.7 konstruierten Beobachter mit der (normalisierten) Gradientenregel als Adaptionsregel und Anfangswerten 0 gilt: Ist (u, y) ein auf $\mathbb{R}_+$ beschränktes Input-Output-Paar des zu identifizierenden Systems, und enthält u mindestens $2n$ Spektralwerte, so konvergiert der Parameterfehler exponentiell gegen 0.*

Beweis: Folgt aus Satz 8.21 und Satz 8.33. □

Bemerkung 8.35
In Beispiel 8.25 haben wir gesehen, daß

$$u(t) = \sum_{j=1}^{n} \alpha_j \sin(\omega_j t) \tag{8.115}$$

das Spektralmaß

$$\mu = \sum_{j=1}^{n} \frac{\pi}{2} \alpha_j^2 (\delta_{\omega_j} + \delta_{-\omega_j}) \tag{8.116}$$

hat, falls $\omega_1, \ldots, \omega_n > 0$ paarweise verschieden sind. Es genügen also n verschiedene Frequenzen im Input, um $2n$ verschiedene Spektralwerte zu erhalten.

9 Adaptive Regelung: Tracking Problem

(Dieser Abschnitt ist unvollständig.)

Problem 9.1 (Tracking Problem)
Gegeben sind ein unbekanntes System $\mathcal{P}$, etwa ein SISO-Kontrollsystem mit Übertragungsfunktion G, und ein Referenzsystem $\mathcal{M}$, beschrieben durch eine bekannte Übertragungsfunktion G_M. Gesucht ist ein Regler, der aus einem Referenzsignal $r : \mathbb{R}_+ \to \mathbb{R}$ einen Input $u : \mathbb{R}_+ \to \mathbb{R}$ erzeugt, so daß der Output y von $\mathcal{P}$ für $t \to \infty$ gegen den Output y_M von $\mathcal{M}$ konvergiert. Der Regler darf außer r auch y und y_M als Eingangsgrößen verwenden.

Das Referenzsystem $\mathcal{M}$ beschreibt das gewünschte Verhalten. $\mathcal{P}$ kann etwa das Ergebnis einer Implementation von $\mathcal{M}$ sein.

Beispiel 9.2 (Skalarer Zustand)
Sei $\mathcal{M}$ beschrieben durch

$$\dot{y}_M = -a_M y_M + b_M r\,, \quad a_M > 0\,,\ b_M \in \mathbb{R}\,, \tag{9.1}$$

und $\mathcal{P}$ beschrieben durch

$$\dot{y} = -ay + bu\,. \tag{9.2}$$

Die Koeffizienten $a, b \in \mathbb{R}$ sind unbekannt, aber wir nehmen an, daß wir wissen, daß $b > 0$ ist. Wir setzen eine adaptive Regelung an mit

$$u(t) = c_0(t)r(t) + d_0(t)y(t)\,, \tag{9.3}$$

wobei die Funktionen c_0 und d_0 durch eine Adaptionsregel definiert werden sollen. Wären a, b bekannt, so könnten wir setzen

$$c_0(t) = c_0^* := \frac{b_M}{b}\,, \quad d_0(t) = d_0^* := \frac{a - a_M}{b}\,. \tag{9.4}$$

Mit (9.4) ergibt sich nämlich

$$\begin{aligned} \dot{y} &= -ay + b(c_0^* r + d_0^* y) = -(a - bd_0^*)y + bc_0^* r \\ &= -a_M y + b_M r\,. \end{aligned} \tag{9.5}$$

Es ist also $y(t) = y_M(t)$ für alle t, falls die Anfangsbedingungen übereinstimmen, bzw. y konvergiert exponentiell gegen y_M andernfalls. Sind a und b unbekannt, so betrachten wir den Outputfehler

$$e_y(t) = y(t) - y_M(t) \tag{9.6}$$

und die Adaptionsregel

$$\dot{c}_0 = -\gamma e_y(t) r(t)\,, \tag{9.7}$$

$$\dot{d}_0 = -\gamma e_y(t) y(t)\,, \tag{9.8}$$

wobei $\gamma > 0$ ein Parameter ist. Das Gesamtsystem wird also beschrieben durch (9.1) - (9.3) und (9.6) - (9.8). Um seine Stabilität und Konvergenz zu untersuchen, betrachten wir die Differentialgleichungen für den Outputfehler e_y und den Parameterfehler

$$\varphi(t) = \begin{pmatrix} \varphi_c(t) \\ \varphi_d(t) \end{pmatrix} = \begin{pmatrix} c_0(t) - c_0^* \\ d_0(t) - d_0^* \end{pmatrix}. \tag{9.9}$$

Sie lauten

$$\dot{e}_y = -a_M e_y + b(\varphi_c r + \varphi_d e_y + \varphi_d y_M)\,, \tag{9.10}$$

$$\dot{\varphi}_c = -\gamma e_y r\,, \tag{9.11}$$

$$\dot{\varphi}_d = -\gamma e_y^2 - \gamma e_y y_M\,. \tag{9.12}$$

Wir definieren die Ljapunow-Funktion

$$V(e_y, \varphi_c, \varphi_d) = \frac{1}{2}e_y^2 + \frac{b}{2\gamma}(\varphi_c^2 + \varphi_d^2)\,. \tag{9.13}$$

Für

$$v(t) = V(e_y(t), \varphi_c(t), \varphi_d(t)) \tag{9.14}$$

gilt dann

$$\dot{v} = e_y\dot{e}_y + \frac{b}{\gamma}(\varphi_c\dot{c}_0 + \varphi_d\dot{d}_0) = -a_M e_y^2 \leq 0\,. \tag{9.15}$$

Da der Grenzwert $v(\infty) = \lim_{t\to\infty} v(t)$ existiert, folgt wieder $e_y \in L^2(\mathbb{R}_+) \cap L^\infty(\mathbb{R}_+)$ und aus (9.10) auch $\dot{e}_y \in L^\infty(\mathbb{R}_+)$. Aus Lemma ... folgt

$$\lim_{t\to\infty} e_y(t) = 0\,, \tag{9.16}$$

daß heißt, der Ansatz (9.3) in Verbindung mit der Adaptionsregel (9.7) - (9.8) löst das Tracking Problem.

Voraussetzung 9.3
Sowohl das Referenzsystem $\mathcal{M}$ als auch das unbekannte System $\mathcal{P}$ seien SISO-Kontrollsysteme vom Mc-Millan-Grad n mit den Übertragungsfunktionen

$$G(s) = \frac{p(s)}{q(s)}\,, \quad G_M(s) = \frac{p_M(s)}{q_M(s)}\,, \tag{9.17}$$

wobei $p \in P_m$ und $q \in P_n$ teilerfremde Polynome der Form

$$p(s) = \sum_{k=0}^{m} \beta_{k+1} s^k\,, \quad q(s) = s^n + \sum_{k=0}^{n-1} \alpha_{k+1} s^k\,, \tag{9.18}$$

sind. Es gelte $m < n$, $\beta_{m+1} > 0$, und p sei Stabilitätspolynom. Für $\mathcal{M}$ gelte analog, daß $p_M \in P_m$ und $q_M \in P_n$ teilerfremde Polynome der Form

$$p_M(s) = \sum_{k=0}^{m} \beta_{k+1}^M s^k\,, \quad q_M(s) = s^n + \sum_{k=0}^{n-1} \alpha_{k+1}^M s^k\,, \tag{9.19}$$

sind, daß $\beta_{m+1}^M > 0$ und p_M, q_M Stabilitätspolynome sind.

Bemerkung 9.4 (Struktur des adaptiven Reglers)
Sei $q_B \in P_{n-1}$ ein Stabilitätspolynom der Form

$$q_B(s) = s^{n-1} + \sum_{k=0}^{n-2} \alpha_{k+1}^B s^k\,. \tag{9.20}$$

Wir definieren einen Beobachter mit dem Zustand $(w_1, w_2) \in \mathbb{R}^{n-1} \times \mathbb{R}^{n-1}$ durch

$$\dot{w}_1 = A_B\, w_1 + b_B\, u\,, \tag{9.21}$$
$$\dot{w}_2 = A_B\, w_2 + b_B\, y\,, \tag{9.22}$$

wobei

$$A_B = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ -\alpha_1^B & -\alpha_2^B & \cdots & \cdots & -\alpha_{n-1}^B \end{pmatrix}, \quad b_B = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{pmatrix}. \tag{9.23}$$

Wir definieren den Regler durch

$$u = c_0 r + c^T w_1 + d_0 y + d^T w_2\,, \tag{9.24}$$

wobei $c_0, d_0 \in \mathbb{R}$ und $c, d \in \mathbb{R}^{n-1}$ die Parameter des Reglers darstellen. Durch (9.21) - (9.24) zusammen mit

$$\hat{y}(s) = G(s)\hat{u}(s) \tag{9.25}$$

wird die Übertragungsfunktion G_R des Gesamtsystems mit Input r und Output y,

$$\hat{y}(s) = G_R(s)\hat{r}(s)\,, \tag{9.26}$$

festgelegt (sie wird im nächsten Lemma berechnet). Ziel ist es, durch eine geeignete Adaptionsregel die Parameter (c_0, c, d_0, d) so einzustellen, daß

$$G_R = G_M \tag{9.27}$$

gilt.

Lemma 9.5 (Übertragungsfunktion des Reglers)
Wir definieren $p_c, p_d \in P_{n-2}$ durch

$$p_c(s) = \sum_{k=0}^{n-2} c_{k+1} s^k\,, \quad p_d(s) = \sum_{k=0}^{n-2} d_{k+1} s^k\,. \tag{9.28}$$

Dann gilt

$$G_R = \frac{c_0\, G}{1 - G_c - G_d G} = \frac{c_0\, p\, q_B}{q(q_B - p_c) - (d_0\, q_B + p_d)p}\,. \tag{9.29}$$

Beweis: Das Teilsystem $u \to c^T w_1$ hat die Übertragungsfunktion

$$G_c(s) = \frac{p_c(s)}{q_B(s)}\,, \tag{9.30}$$

das Teilsystem $y \to d_0 y + d^T w_2$ hat die Übertragungsfunktion

$$G_d(s) = d_0 + \frac{p_d(s)}{q_B(s)}\,. \tag{9.31}$$

Aus (9.24) folgt

$$\hat{u}(s) = c_0\hat{r}(s) + G_c(s)\hat{u}(s) + G_d(s)\hat{y}(s)\,, \quad \hat{y}(s) = G(s)\hat{u}(s)\,, \tag{9.32}$$

also

$$\hat{u} = \frac{c_0}{1 - G_c - G_dG}\,\hat{r}\,, \quad \hat{y} = \frac{c_0\,G}{1 - G_c - G_dG}\,\hat{r}\,. \tag{9.33}$$

Einsetzen von (9.30), (9.31) und (9.17) liefert die Behauptung. □

Lemma 9.6
Seien p, q, p_M, q_M, q_B in der Form (9.18), (9.19) und (9.20) gegeben, es gelte $p_M \,|\, q_B$ und $\beta_{m+1}, \beta^M_{m+1} > 0$. Dann gibt es eindeutig bestimmte $c_0^*, d_0^* \in \mathbb{R}$ und $p_c^*, p_d^* \in P_{n-2}$, so daß

$$G_R = G_M \tag{9.34}$$

gilt.

Beweis: Sei

$$q_0 = \frac{q_B}{p_M}\,, \quad q_B = q_0\,p_M\,. \tag{9.35}$$

Wegen (9.29) gilt $G_R = G_M$ genau dann, wenn

$$c_0\,p\,q_0\,q_M = q(q_B - p_c) - (d_0\,q_B + p_d)p \tag{9.36}$$

gilt. Zur Existenz: Das Polynom q_0 hat den Grad $n - m - 1$. Wir zerlegen

$$q_0\,q_M = g\,q + r\,, \quad g \in P_{n-m-1}\,,\, r \in P_{n-1}\,. \tag{9.37}$$

Aus (9.36) wird dann

$$c_0\,p\,g\,q + c_0\,p\,r = q(q_B - p_c) - (d_0\,q_B + p_d)p\,. \tag{9.38}$$

Wir betrachten die beiden Gleichungen

$$p_c = q_B - c_0\,p\,g\,, \quad c_0\,r = -d_0\,q_B - p_d\,. \tag{9.39}$$

Da g den Grad $n - m - 1$ hat, können wir c_0^* so wählen, daß $p_c \in P_{n-2}$ gilt. Dieses p_c wählen wir als p_c^*. Wir wählen weiter d_0^* so, daß $-d_0^*$ gleich dem höchsten Koeffizienten von $c_0^*\,r$ ist, und definieren p_d^* durch die zweite Gleichung in (9.39). Es sind dann beide Gleichungen in (9.39) und damit auch (9.36) erfüllt.
Zur Eindeutigkeit: Sei $(\tilde{c}_0, \tilde{p}_c, \tilde{d}_0, \tilde{p}_d)$ die Differenz zweier Lösungen von (9.36). Dann gilt

$$\tilde{c}_0\,p\,q_0\,q_M = -q\,\tilde{p}_c - (\tilde{d}_0\,q_B + \tilde{p}_d)p \tag{9.40}$$

Die rechte Seite von (9.41) liegt in P_{2n-2}, während $p\,q_0\,q_M$ den Grad $n - m - 1$ hat. Also ist $\tilde{c}_0 = 0$. Es folgt

$$q\tilde{p}_c = -(\tilde{d}_0\,q_B + \tilde{p}_d)p\,. \tag{9.41}$$

Da p und q teilerfremd sind, muß $\tilde{p}_c = 0$ sein (andernfalls würde $q \in P_n$ das Polynom $\tilde{d}_0 q_B + \tilde{p}_d \in P_{n-1}$ teilen). Aus (9.41) folgt weiter, daß auch $\tilde{d}_0 = \tilde{p}_d = 0$. □

Bemerkung 9.7
Lemma 9.6 zeigt, daß die Konstruktion aus Bemerkung 9.4 sinnvoll sein kann. Ob sie es tatsächlich ist, stellt sich heraus, wenn man eine Adaptionsregel für die Parameter angeben kann, so daß das gestellte Ziel ($y(t) - y_M(t) \to 0$) erreicht wird. Als Adaptionsregel kann man wieder die Gradientenregel (oder eine Variante) nehmen. Es stellt sich allerdings heraus, daß es günstiger ist, nicht am Output-Fehler $e_y = y - y_M$ anzusetzen (siehe Sastry/Bodson für eine Diskussion), sondern am Input-Fehler

$$e_r(t) = r_P(t) - r(t)\,, \tag{9.42}$$

wobei r_P diejenige Funktion ist, die als Input für $\mathcal{M}$ verwendet den Output y erzeugt, also

$$\hat{y}(s) = G_M(s)\hat{r}_P(s)\,, \quad \hat{r}_P(s) = \frac{1}{G_M(s)}\,\hat{y}(s)\,. \tag{9.43}$$

Die Funktion r_P kann man aber nicht stabil aus y gewinnen. Ein solches Vorgehen entspräche $n-m$ Differentiationen. Man geht daher zu einer Funktion über, die einem $n-m$-fachen Integrieren des Input-Fehlers (in geeigneter Kombination mit dem Beobachter-Output) entspricht.

Bemerkung 9.8 (Modifizierter Inputfehler, Konstruktion des Identifizierers)
Sei q_E Stabilitätspolynom vom Grad $n - m$. Der Identifizierer für den modifizierten Inputfehler wird konstruiert als Kontrollsystem mit Input (y, w_1, w_2, u), Zustand $\nu, \nu_u \in \mathbb{R}^{2n} \times \mathbb{R}$ und Output $e_\nu \in \mathbb{R}$,

$$e_\nu(t) = \theta^T \nu(t) - \nu_u(t)\,, \tag{9.44}$$

wobei (ν, ν_u) komponentenweise gemäß den Übertragungsfunktionen

$$\hat{\nu} = \left(\frac{1}{q_E\, G_M}\,\hat{y}\,,\ \frac{1}{q_E}\,\hat{w}_1\,,\ \frac{1}{q_E}\,\hat{y}\,,\ \frac{1}{q_E}\,\hat{w}_2 \right)\,, \quad \hat{\nu_u} = \frac{1}{q_E}\,\hat{u}\,, \tag{9.45}$$

aus (y, w_1, w_2, u) gewonnen werden. Der erste Term in (9.45) entspricht dem aufintegrierten Input r_P. Die zugehörige Übertragungsfunktion

$$\frac{1}{q_E\, G_M} = \frac{q_M}{q_E\, p_M} = a + \frac{\tilde{q}_M}{q_E\, p_M}\,, \quad a \in \mathbb{R}\,,\ \tilde{q}_M \in P_{n-1}\,, \tag{9.46}$$

läßt sich durch ein stabiles Kontrollsystem realisieren (p_M ist nach Voraussetzung 9.3 Stabilitätspolynom).

Bemerkung 9.9 (Adaptionsregel)
Standardbeispiel für eine Adaptionsregel ist wieder die normalisierte Gradientenregel, gewonnen aus dem Gradienten von

$$J(\theta) = \frac{1}{2} e_\nu(t;\theta)^2\,, \tag{9.47}$$

also

$$\dot{\theta} = -\gamma \frac{e_\nu\, \theta}{1 + \kappa \nu^T \nu}\,, \quad \gamma\,,\ \kappa > 0\,. \tag{9.48}$$

Um die Konvergenz des adaptiven Reglers zu garantieren, muß man die erste Kompontente c_0 von θ von 0 weg beschränken, d.h.

$$c_0(t) \geq c_{\min} > 0 \tag{9.49}$$

garantieren. Dies erreicht man, indem man die erste Komponente in (9.48) ersetzt durch

$$\dot{c}_0(t) = 0\,, \tag{9.50}$$

falls $c_0(t) = c_{\min}$ und $e_\nu(t) > 0$ (diese Variante heißt projizierte Gradientenregel).

Lemma 9.10
Wir betrachten das SISO-System $u \to e_\nu$, bestehend aus dem System $\mathcal{P}$, den Beobachtergleichungen

$$\dot{w}_1 = A_B\, w_1 + b_B\, u\,, \tag{9.51}$$
$$\dot{w}_2 = A_B\, w_2 + b_B\, y\,, \tag{9.52}$$

und dem Identifizierer (9.44), (9.45). Wir bezeichnen die zugehörige Übertragungsfunktion mit G_E, also

$$\hat{e}_\nu(s) = G_E(s)\hat{u}(s)\,. \tag{9.53}$$

Werden die Parameter $\theta \in \mathbb{R}^{2n}$ gemäß Lemma 9.6 so gewählt, daß $G_R = G_M$ gilt, so ist

$$G_E = 0\,. \tag{9.54}$$

Beweis: Für beliebige Reglerparameter θ gilt

$$\begin{aligned}\hat{e}_\nu &= \theta^T\hat{\nu} - \hat{\nu}_u = \frac{1}{q_E}\left(\frac{c_0}{G_M}\hat{y} + c^T\hat{w}_1 + d_0\hat{y} + d^T\hat{w}_2 - \hat{u}\right)\\ &= \frac{1}{q_E}\left(\frac{c_0\,G}{G_M} + G_c + G_d G - 1\right)\hat{u} = \frac{1}{q_E}c_0 G\left(\frac{1}{G_M} - \frac{1}{G_R}\right)\hat{u}\,,\end{aligned} \tag{9.55}$$

also $G_E = 0$ falls $G_R = G_M$. □

Lemma 9.11 (Darstellung des modifizierten Input-Fehlers) *Sei $\theta : \mathbb{R}_+ \to \mathbb{R}^{2n}$ beliebig, sei $u \in L^\infty(\mathbb{R}_+)$ ein beliebiger Input des SISO-Systems aus Lemma 9.10, seien alle Anfangswerte 0. Für*

$$e_\nu(t) = \theta(t)^T\nu(t) - \nu_u(t) \tag{9.56}$$

gilt dann

$$e_\nu(t) = (\theta(t) - \theta^*)^T\nu(t) = \varphi(t)^T\nu(t)\,, \tag{9.57}$$

wobei $\varphi(t) = \theta(t) - \theta^*$ wieder den Parameterfehler bezeichnet.

Beweis: Aus Lemma 9.10 folgt, daß

$$\theta^{*T}\nu(t) - \nu_u(t) = 0\,, \quad t \geq 0\,. \tag{9.58}$$

Bemerkung 9.12
Man beachte, daß die Darstellung (9.57) weder von der Adaptionsregel (θ ist eine beliebige Funktion von t) noch von der gewählten Reglergleichung (u ist ebenfalls eine beliebige Funktion von t) abhängt.

Sind die Anfangswerte nicht 0, so gilt statt (9.57)

$$e_\nu(t) = \varphi(t)^T \nu(t) + \varepsilon(t)\,, \tag{9.59}$$

wobei ε exponentiell gegen 0 konvergiert (siehe Sastry/Bodson).

Bemerkung 9.13 (Gesamtsystem)
Um die Stabilität und Konvergenz des adaptiven Reglers zu untersuchen, muß man das Gesamtsystem betrachten. Sei (A_p, b_p, c_p^T) eine minimale Realisierung von G, d.h. des unbekannten Systems $\mathcal{P}$. Das Gesamtsystem hat die Form

$$\dot{x} = A_p x + b_p u\,, \quad y = c_p^T x\,, \tag{9.60}$$

$$\dot{w}_1 = A_B w_1 + b_B u\,, \tag{9.61}$$

$$\dot{w}_2 = A_B w_2 + b_B y\,, \tag{9.62}$$

$$u = \theta^T \begin{pmatrix} r \\ w_1 \\ y \\ w_2 \end{pmatrix} = c_0 r + c^T w_1 + d_0 y + d^T w_2\,, \tag{9.63}$$

$$e_\nu = \theta^T \nu - \nu_u = \varphi^T \nu\,, \quad \varphi = \theta - \theta^*\,, \tag{9.64}$$

$$\dot{\theta} = \dot{\varphi} = -\gamma \frac{e_\nu \nu}{1 + \kappa \nu^T \nu} = -\gamma \frac{\nu \nu^T \varphi}{1 + \kappa \nu^T \nu}\,, \tag{9.65}$$

und ν sowie ν_u ergeben sich gemäß (9.45) als Output eines stabilen Kontrollsystems mit Input (w_1, y, w_2) bzw. u. Wir setzen hier und im folgenden voraus, daß der Referenzinput r global beschränkt ist,

$$r \in L^\infty(\mathbb{R}_+)\,, \tag{9.66}$$

und daß alle Anfangswerte Null sind.

Lemma 9.14 *Das Gesamtsystem hat eine auf ganz $\mathbb{R}_+$ definierte eindeutige Lösung, und es gilt*

$$\varphi, \theta \in L^\infty(\mathbb{R}_+)\,, \quad \dot{\varphi}, \beta \in L^2(\mathbb{R}_+) \cap L^\infty(\mathbb{R}_+)\,, \tag{9.67}$$

wobei

$$\beta = \frac{\varphi^T \nu}{1 + \|\nu\|_{t,\infty}}\,. \tag{9.68}$$

Beweis: Wir betrachten eine Lösung auf einem beschränkten Intervall $I = (0, T)$. Ersetzen wir in Lemma ... und Satz ... ,die Funktion w durch ν, so schließen wir, daß (9.67) auf I gilt. Damit ist die Lösung sowie deren Ableitung auf $(0, T)$ beschränkt und läßt sich über T hinaus fortsetzen. Das maximale Existenzintervall ist daher $\mathbb{R}_+$, und (9.67) gilt auf $\mathbb{R}_+$ nach dem zitierten Satz. □

Notation 9.15 *Wir schreiben*

$$w = (r, \bar{w}) = (r, w_1, y, w_2)\,, \quad \theta = (c_0, \bar{\theta})\,, \tag{9.69}$$

und bezeichnen mit G_c^*, G_d^*, G_R^* *die zu* θ^* *gehörenden Übertragungsfunktionen.*

Lemma 9.16 *Es gelten*

$$u = \theta^T w = \varphi^T w + \theta^{*T} w = \varphi^T w + c_0^* r + \bar{\theta}^{*T} \bar{w}\,, \tag{9.70}$$

$$\hat{e}_y = \hat{y} - \hat{y}_M = \frac{1}{c_0^*} G_M \widehat{\varphi^T w}\,, \tag{9.71}$$

d.h. verwenden wir

$$\frac{1}{c_0^*}\varphi^T w \tag{9.72}$$

als Input von $\mathcal{M}$*, so erhalten wir* $e_y = y - y_M$ *als Output.*

Beweis: Gleichung (9.70) folgt unmittelbar aus den Definitionen. Es gilt weiter

$$\begin{aligned} \widehat{\varphi^T w} &= \hat{u} - c_0^* \hat{r} - \bar{\theta}^{*T} \hat{\bar{w}} = \hat{u} - c_0^* \hat{r} - (G_c^* + G_d^* G)\hat{u} \\ &= (1 - G_c^* - G_d^* G)\hat{u} - c_0^* \hat{r} = \frac{1}{G_R^*} c_0^* G \hat{u} - c_0^* \hat{r} \\ &= c_0^* \left(\frac{G}{G_M} \hat{u} - \hat{r} \right)\,, \end{aligned} \tag{9.73}$$

woraus die Behauptung folgt . □

Lemma 9.17 *Es gibt eine Konstante* $k > 0$ *mit*

$$|y(t)| \le k \|\varphi^T w\|_{t,\infty} + k\,, \quad t \ge 0\,, \tag{9.74}$$

$$|\dot{y}(t)| \le k \|\varphi^T w\|_{t,\infty} + k\,, \quad t \ge 0\,. \tag{9.75}$$

Beweis: Folgt aus Lemma 9.16, da $\mathcal{M}$ stabil ist und r, also auch y_M und $\dot{y}_M$, beschränkt sind. □

Lemma 9.18 *Es gibt eine Konstante* k*, so daß gilt*

$$\|f(t)\| \le k \|\bar{\varphi}^T \bar{w}\|_{t,\infty} + k\,, \quad t \ge 0\,, \tag{9.76}$$

für jede der Funktionen $w_1, w_2, y, u, \dot{w}_1, \dot{w}_2, \dot{y}$ *an der Stelle von* f*.*

Beweis: Wir schreiben die Differentialgleichung für w_1 um zu

$$\dot{w}_1 = A_B w_1 + b_B(\theta^{*T} w + \varphi^T w)\,. \tag{9.77}$$

Mit

$$\dot{w}_2 = A_B w_2 + b_B y\,, \tag{9.78}$$

ergibt sich aus der Schranke für y in (9.74) und der Stabilität von A_B dieselbe Schranke (mit einer anderen Konstanten) für w_1, w_2 und deren Ableitungen. Da die Differenz

$$\varphi^T w - \bar{\varphi}^T \bar{w} = (c_0 - c_0^*) r \tag{9.79}$$

beschränkt ist, folgt die Behauptung. □