Institut für Regelungstechnik, Technische Universität Graz

13. Steirisches Seminar über Regelungstechnik und Prozessautomatisierung

Tagungsband

Schloss Retzhof
Bildungshaus des Landes Steiermark

15. bis 18. September 2003

Herausgeber und Verleger:
Institut für Regelungstechnik
Technische Universität Graz
Inffeldgasse 16 c, A-8010 Graz
Tel.: +43 316 873 7021
Fax: +43 316 873 7028

ISBN 3-901439-05-6

Vorwort

Das Steirische Seminar über Regelungstechnik und Prozessautomatisierung findet alle zwei Jahre im Schloss Retzhof, dem Bildungshaus des Landes Steiermark in der Nähe der südsteirischen Stadt Leibnitz, statt. Bei dieser Veranstaltung, die vom Institut für Regelungstechnik der Technischen Universität Graz durchgeführt wird, werden grundlegende Probleme der Regelungs- bzw. Automatisierungstechnik in einem fruchtbringenden Dialog zwischen Angehörigen der Universitäten und der Industrie erörtert. Dabei besteht für die Vortragenden die Möglichkeit, ihre Beiträge im Rahmen eines Tagungsbandes zu veröffentlichen. Den Autoren sei an dieser Stelle für die Sorgfalt bei der Erstellung ihrer Beiträge gedankt.

Graz, im September 2003

Inhaltsverzeichnis

BEOBACHTER MINIMALER NORM

Alexander Weinmann, OVE, Senior Member IEEE
Technische Universität Wien,
Institut für Automatisierungs- und Regelungstechnik
Gusshausstrasse 27-29/376
A-1040 Wien / Österreich
Phone: +43 1 58801 37600, Fax: +43 1 58801 37699
email: weinmann@acin.tuwien.ac.at

April 29, 2003

Schlüsselwörter:
Kontrollbeobachter, Beobachter reduzierter Ordnung, minimale Matrix-Normen

Kurzfassung:

Um nummerischen Problemen bei Festkomma-Prozessoren vorzubeugen, sollten die Verstärkungsmatrizen im Normsinne möglichst klein sein. Dies kann erreicht werden, wenn die Frobenius-Norm aller Matrizen in die Realisierung im Zustandsraum mit einbezogen wird. Die Summe aller Matrixnormen wird minimiert. Die gewünschte Dynamik des Kontrollbeobachters wird durch Vorgabe einer Zustandsreglermatrix und einer Beobachtermatrix besorgt. In bestimmten Fällen ist höhere Dynamik erreichbar, obwohl die Norm der beteiligten Matrizen gesenkt wird.

Abstract:

To prevent numerical problems in fixed-point microcontrollers, the elements of controller and observer gain matrices should be as small as possible in norm sense. This is achieved by including the Frobenius norm of all the matrices when implementing the function observer in state space. This paper addresses the problem of obtaining the minimum sum of all the norms involved when implementing a closed-loop system of predetermined dynamical properties given by the matrices of a state controller and an observer. In specific cases better dynamics of the closed-loop control is achieved even by using smaller matrix norms.

1 Einleitung. Übersicht

Der vorliegende Aufsatz widmet sich der Realisierung von beobachtergestützten Reglern, bei denen die Koeffizientenmatrizen in Beobachter und Regler in ihrer Norm minimiert werden. Der Beobachter wird durch seine Koeffizientenmatrizen gegeben.

Das regelungstechnische Schrifttum kennt einige spezielle Arbeiten, die Berührungspunkte zu vorliegendem Aufsatz besitzen. Diese Arbeiten werden kurz beschrieben, wenn sie Normen der Verstärkungsmatrizen zu einer Limitierung heranziehen. Ohne Bezug auf Beobachter wird die Minimisierung der Norm der Zustandsreglermatrix bei gegebenen Eigenwerten des geschlossenen Regelkreises wird in [1] besorgt, unter zusätzlicher optimaler Konditionierung der Eigenvektormatrix des geschlossenen Regelkreises in [2]. Die iterative Zuweisung einiger Pole und Festhalten einer anderen Gruppe wird unter der Nebenbedingung der Minimierung des Regelungsaufwands durch Minimierung der Verstärkungsmatrix-Norm in [3] gezeigt; bei dyadischer Reglerstruktur in [4], für Zustands- und Ausgangsreglermatrix in [5]. In [6] wird diese Aufgabe LMI-basiert gelöst. Normausdrücke von Beobachterverstärkungsmatrizen werden auch in [7] eingesetzt, allerdings um einen Robustheitsrand zu begründen; Robustheit basierend auf einem quadratischen Konvergenzmaß.

Für zusätzlich geringe Empfindlichkeit gegenüber Parametervariationen und Unsicherheiten in den Streckenkoeffizienten schlägt [8] den Weg mit Mehrfachzielfunktionen und analytisch hergeleiteten Gradienten ein. Die Eigenwertvorgabe für Beobachter und die Abweichungsminimierung in LQ-Form wurde in [9] auch unter der Nebenbedingung der Normbeschränkung einer Hilfsmatrix gelöst, wobei nummerische Methoden zur Minimierung herangezogen werden, im Gegensatz zu vorliegendem Aufsatz, in dem die Lösung analytisch geschlossen erfolgt.

Seitens apparativer Anwendungen werden laut Schrifttum überwiegend Beobachter für Antriebsregelungen in Single-chip Festkomma-Mikrocontrollern und digitalen Signalprozessoren eingesetzt, insbesondere für Fluss-, Rotorstrom- und Drehzahl beobachtung [10-14]; auch für Lastmoment und Laststrom [15,16].

Die vorliegende Arbeit nimmt Bezug auf allgemeine Beobachterformulierungen [17-21]. Die Ergebnisse können aber auch auf Störungs- und Führungsmodellerweiterungen [22-29] ausgedehnt werden.

Der Standardfall des Beobachters stellt sich wie folgt dar. Die Regelstrecke ist durch

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t) \end{aligned}$$

beschrieben, wobei $\mathbf{x} \in \mathcal{R}^n$; $\mathbf{y} \in \mathcal{R}^r$; $\mathbf{u} \in \mathcal{R}^m$; $\mathbf{A} \in \mathcal{R}^{n\times n}$; $\mathbf{B} \in \mathcal{R}^{n\times m}$; $\mathbf{C} \in \mathcal{R}^{r\times n}$.

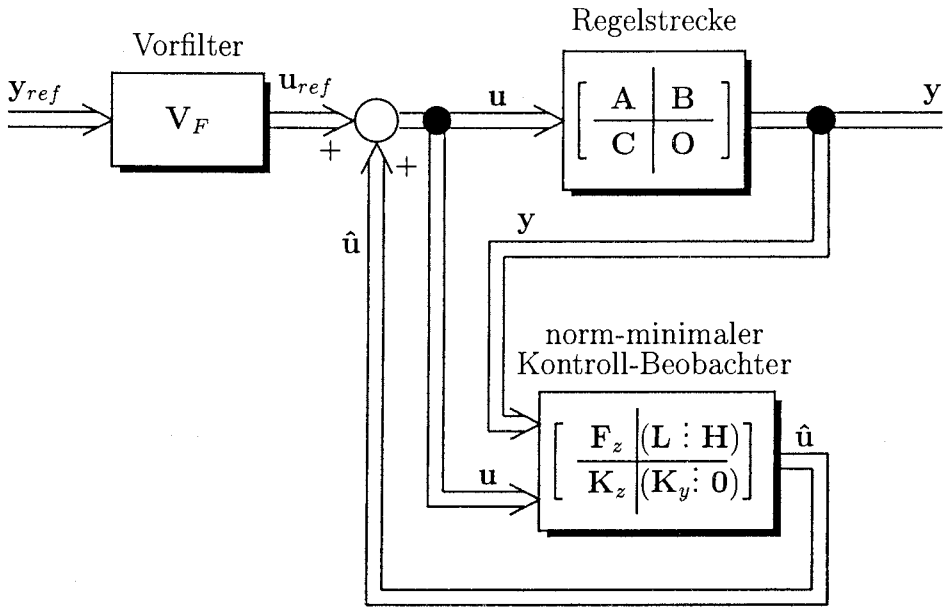


Figure 1: Regelung mit Beobachter unter Minimierung der Normsumme

Die Stellgröße $\mathbf{u}(t)$ wird aus $\mathbf{y}(t)$ über einen Ausgangsregler $\mathbf{K}_y$, aus einer geschätzten Beobachter-Zustandsgröße $\hat{\mathbf{z}}(t)$ mit dem Regler $\mathbf{K}_z$ und einem Referenzanteil $\mathbf{u}_{ref} = \mathbf{V}_F\mathbf{y}_{ref}$ bereitgestellt Mit Fig. 1 gilt

$$\begin{aligned} \mathbf{u}(t) &= \hat{\mathbf{u}}(t) + \mathbf{u}_{ref}(t) \\ \hat{\mathbf{u}}(t) &= \mathbf{K}_z\hat{\mathbf{z}}(t) + \mathbf{K}_y\mathbf{y}(t) \\ \mathbf{u}(t) &= \mathbf{K}_z\hat{\mathbf{z}}(t) + \mathbf{K}_y\mathbf{y}(t) + \mathbf{V}_F\mathbf{y}_{ref} \end{aligned}$$

unter $\mathbf{K}_z \in \mathcal{R}^{m\times(n-r)}$, $\mathbf{K}_y \in \mathcal{R}^{m\times r}$ und $\mathbf{V}_F \in \mathcal{R}^{m\times r}$. Die Zustandsgröße des Beobachters $\hat{\mathbf{z}}(t)$ und ihr Nominalwert $\mathbf{z}(t)$ befolgen

$$\dot{\hat{\mathbf{z}}}\,(t) = \mathbf{F}_z\hat{\mathbf{z}}(t) + \mathbf{L}\mathbf{y}(t) + \mathbf{H}\mathbf{u}(t) \tag{1}$$

$$\mathbf{z}(t) = \mathbf{T}\mathbf{x}(t) = \hat{\mathbf{z}}(t) + \tilde{\mathbf{z}}(t)$$

mit $\mathbf{z},\ \hat{\mathbf{z}}, \tilde{\mathbf{z}} \in \mathcal{R}^{n-r}$; $\mathbf{F}_z \in \mathcal{R}^{(n-r)\times(n-r)}, \mathbf{L} \in \mathcal{R}^{(n-r)\times r}, \mathbf{H} \in \mathcal{R}^{(n-r)\times m}$, $\mathbf{T} \in \mathcal{R}^{(n-r)\times n}$.

Die erforderlichen Beobachterbeziehungen lauten bekannterweise

$$\mathbf{F}_z\mathbf{T} - \mathbf{T}\mathbf{A} + \mathbf{L}\mathbf{C} = \mathbf{0}, \tag{2}$$

$$\mathbf{H} = \mathbf{T}\mathbf{B}, \tag{3}$$

$$(\mathbf{K}_z \vdots \mathbf{K}_y) = \mathbf{K}\begin{pmatrix} \mathbf{T} \\ \mathbf{C} \end{pmatrix}^{-1}, \qquad \text{rang}\begin{pmatrix} \mathbf{T} \\ \mathbf{C} \end{pmatrix} = n\ , \tag{4}$$

wobei $\mathbf{K} \in \mathcal{R}^{m \times n}$ gilt.

Für die Wahl eines Vorfilters $\mathbf{V}_F \in \mathcal{R}^{m \times r}$ zur Erzeugung von $\mathbf{u}_{ref} = \mathbf{V}_F \mathbf{y}_{ref}$ ist bei $m \geq r$ und stationär $\mathbf{y}_{ref} = \mathbf{y}_\infty$

$$-\mathbf{C}(\mathbf{A} + \mathbf{B}\mathbf{K}_z\mathbf{T} + \mathbf{B}\mathbf{K}_y\mathbf{C})^{-1}\mathbf{B}\mathbf{V}_F = \mathbf{I} \tag{5}$$

zu verlangen. Das Vorfilter $\mathbf{V}_F \in \mathcal{R}^{r \times m}$ ist von $\mathbf{L}$ nicht abhängig; denn in Gl.(5) steht nur der (in dieser Problemstellung unveränderliche) Ausdruck $\mathbf{K}_z\mathbf{T} + \mathbf{K}_y\mathbf{C} = \mathbf{K}$.

Die vorliegende Abhandlung ist der Realisierung eines Kontrollbeobachters gewidmet, mit dem das Regelungssystem eine Dynamik erhält, die durch die Zustandsreglermatrix $\mathbf{K}$ und Koeffizientenmatrix $\mathbf{F}_z$ des Beobachters vorgegeben wird. Die Zustandsreglermatrix $\mathbf{K}$ wird nicht realisiert, sie dient nur als Angabe. Bei dieser Realisierung soll eine minimale Normsumme aller Verstärkungsmatrizen $\mathbf{H}$, $\mathbf{K}_z$, $\mathbf{L}$ und $\mathbf{K}_y$ sicherstellt werden. Das Verfahren wird auch unter individueller Gewichtung der Matrixelemente besorgt.

Vor Vorteil sind derartige Entwürfe bei Festkomma-Mikrocontrollern zur Wahrung geeigneten Zahlenformats.

2 Problemstellung

Bei Low-Cost- oder Low-Power-Installationen und single-chip digitalen Signalprozessoren werden verschiedentlich Prozessoren mit Festkomma-Arithmetik eingesetzt. Bei ihr sind große Faktoren zu meiden. Erreicht wird dies durch Einbeziehung der Frobenius-Matrixnorm in die Realisierung.

Die zentrale Fragestellung lautet, welche Matrix $\mathbf{L}$ nach Gl.(2) eine Matrix $\mathbf{T}$ derart liefert, dass eine minimale Normsumme

$$\begin{aligned} f_N &\triangleq \|\mathbf{L}\|_F^2 + \|\mathbf{H}\|_F^2 + \|\mathbf{K}_z\|_F^2 + \|\mathbf{K}_y\|_F^2 && (6) \\ &= \|\mathbf{L}\|_F^2 + \|\mathbf{H}\|_F^2 + \|\mathbf{K}_z \vdots \mathbf{K}_y\|_F^2 && (7) \\ &\triangleq f_1 + f_2 + f_3 \rightarrow \min_{\mathbf{L}} . \end{aligned}$$

entsteht.

Die Aufgabe lässt sich nummerisch durch `fminsearch` oder `fmincon` in MATLAB realisieren, hier soll aber auf die analytische Lösung eingegangen werden.

2.1 Ableitung der Normsumme f_N nach colL

Aus einer Annahme von $\mathbf{L}$ folgt $\mathbf{T}$ aus Gl.(2) nach $\mathbf{T} = \texttt{lyap}(\mathbf{F}_z, -\mathbf{A}, \mathbf{LC})$. Damit lässt sich aus den Gln. (3) und (4) $\mathbf{H}$, $\mathbf{K}_z$ und $\mathbf{K}_y$ bestimmen. Alternativ kann auch aus

Gl.(2) unter Verwendung von

$$\mathbf{C}_I \triangleq \mathbf{C} \otimes \mathbf{I}_{n-r} = \mathcal{R}^{r(n-r)\times n(n-r)} \tag{8}$$

$$\mathbf{U} \triangleq (\mathbf{A}^T \otimes \mathbf{I}_{n-r} - \mathbf{I}_n \otimes \mathbf{F}_z)^{-1} \quad \in \mathcal{R}^{n(n-r)\times n(n-r)} \tag{9}$$

die Beziehung

$$\mathrm{col}\mathbf{T} = (\mathbf{A}^T \otimes \mathbf{I}_{n-r} - \mathbf{I}_n \otimes \mathbf{F}_z)^{-1}\mathrm{col}(\mathbf{LC}) = \mathbf{U}\mathbf{C}_I^T\mathrm{col}\mathbf{L}$$

entwickelt werden. Der Operator „col“ bedeutet die Vektorisierung einer Matrix, also die Aneinanderreihung der Spalten einer Matrix zu einem Vektor.

Um sowohl separieren als auch differenzieren zu können, wird folgender Weg beschritten. In der Umgebung der noch unbekannten Lösungsmatrix $\mathbf{L}$ wird ein Inkrement $\Delta\mathbf{L}$ angenommen. Die Funktion $f(\mathbf{L}+\Delta\mathbf{L})$ wird nach der Matrix-Abweichung $\Delta\mathbf{L}$ als $\partial f(\mathbf{L}+\Delta\mathbf{L})/\partial\Delta\mathbf{L}$ partiell differenziert und danach für den Optimalpunkt $\Delta\mathbf{L} = \mathbf{0}$ gesetzt. Auf solche Art gelingt es, auch komplizierte Abhängigkeiten $f(\mathbf{L})$ in einfache Beziehungen der niedrigen Elemente einer Taylor-Expansion zu entwickeln, solange $\Delta\mathbf{L}$ (in Normbetrachtung) klein bleibt. Demnach wird wie folgt entwickelt

$$f_1 \triangleq \|\mathbf{L}\|_F^2 \rightsquigarrow f_1 = \|\mathbf{L}+\Delta\mathbf{L}\|_F^2 \tag{10}$$

$$\frac{\partial f_1}{\partial \Delta\mathbf{L}}\Big|_{\Delta\mathbf{L}=\mathbf{0}} = 2(\mathbf{L}+\Delta\mathbf{L})\Big|_{\Delta\mathbf{L}=\mathbf{0}} \rightsquigarrow \frac{\partial f_1}{\partial \mathbf{L}} = 2\mathbf{L}$$

$$\frac{\partial f_1}{\partial \mathrm{col}\,\mathbf{L}} = 2\,\mathrm{col}\,\mathbf{L}\ .$$

Aus den Gln.(3) und (9) sowie

$$\mathbf{U}_1 \triangleq (\mathbf{B}^T \otimes \mathbf{I}_{n-r})\mathbf{U} \quad \in \mathcal{R}^{m(n-r)\times n(n-r)} \tag{11}$$

ergibt sich

$$\begin{aligned} f_2 \triangleq \|\mathbf{H}\|_F^2 &= \|\mathrm{col}\mathbf{H}\|_F^2 = \|(\mathbf{B}^T \otimes \mathbf{I}_{n-r})\mathrm{col}\ \mathbf{T}\|_F^2 \\ &= \|(\mathbf{B}^T \otimes \mathbf{I}_{n-r})\mathbf{U}\mathrm{col}(\mathbf{LC})\|_F^2 = \|\mathbf{U}_1\mathbf{C}_I^T\mathrm{col}\mathbf{L}\|_F^2\ . \end{aligned} \tag{12}$$

Daraus folgt der Differenzialquotient der Matrixnorm nach Gl.(74) zu

$$\frac{\partial f_2}{\partial \mathrm{col}\mathbf{L}} = 2(\mathbf{C} \otimes \mathbf{I}_{n-r})\mathbf{U}_1^T\mathbf{U}_1(\mathbf{C}^T \otimes \mathbf{I}_{n-r})\mathrm{col}\,\mathbf{L}\ .$$

Für

$$f_3 \triangleq \|\mathbf{K}\begin{pmatrix}\mathbf{T}\\ \mathbf{C}\end{pmatrix}^{-1}\|_F^2$$

gilt nach Gl.(75) mit der Hilfsmatrix $\mathbf{V}_L$

$$\mathbf{V}_L \triangleq \begin{pmatrix} \mathbf{T} \\ \mathbf{C} \end{pmatrix}^{-1} \quad \in \mathcal{R}^{n\times n} \tag{13}$$

$$\begin{pmatrix} \mathbf{T}+\Delta\mathbf{T} \\ \mathbf{C} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{T} \\ \mathbf{C} \end{pmatrix}^{-1} - \begin{pmatrix} \mathbf{T} \\ \mathbf{C} \end{pmatrix}^{-1} \begin{pmatrix} \Delta\mathbf{T} \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{T} \\ \mathbf{C} \end{pmatrix}^{-1} \tag{14}$$

$$= \mathbf{V}_L - \mathbf{V}_L \begin{pmatrix} \Delta\mathbf{T} \\ \mathbf{0} \end{pmatrix} \mathbf{V}_L \, . \tag{15}$$

Analog zu Gl.(10) wird der Ausdruck f_3 um $\Delta\mathbf{T}$ ergänzt

$$f_3 = \|\mathbf{K}\mathbf{V}_L - \mathbf{K}\mathbf{V}_L \begin{pmatrix} \Delta\mathbf{T} \\ \mathbf{0} \end{pmatrix} \mathbf{V}_L\|_F^2 \, .$$

Partitionieren von $\mathbf{K}\mathbf{V}_L \triangleq (\mathbf{U}_L \vdots \mathbf{U}_4)$ liefert

$$f_3 = \|\mathbf{K}\mathbf{V}_L - (\mathbf{U}_L \vdots \mathbf{U}_4) \begin{pmatrix} \Delta\mathbf{T} \\ \mathbf{0} \end{pmatrix} \mathbf{V}_L\|_F^2 = \|\mathbf{K}\mathbf{V}_L - \mathbf{U}_L\Delta\mathbf{T}\cdot\mathbf{V}_L\|_F^2 \tag{16}$$

$$= \|\mathrm{col}(\mathbf{K}\mathbf{V}_L) - (\mathbf{V}_L^T \otimes \mathbf{U}_L)\mathbf{U}(\mathbf{C}^T \otimes \mathbf{I}_{n-r})\mathrm{col}\Delta\mathbf{L}\|_F^2 \tag{17}$$

$$\frac{\partial f_3}{\partial \mathrm{col}\Delta\mathbf{L}} = -2(\mathbf{C}\otimes\mathbf{I}_{n-r})\mathbf{U}^T(\mathbf{V}_L\otimes\mathbf{U}_L^T)[\mathrm{col}(\mathbf{K}\mathbf{V}_L) \tag{18}$$

$$-(\mathbf{V}_L^T\otimes\mathbf{U}_L)\mathbf{U}(\mathbf{C}^T\otimes\mathbf{I}_{n-r})\mathrm{col}\Delta\mathbf{L}] \, . \tag{19}$$

Bei $\Delta\mathbf{L} = \mathbf{0}$ folgt

$$\frac{\partial f_3}{\partial \mathrm{col}\Delta\mathbf{L}} = -2(\mathbf{C}\otimes\mathbf{I}_{n-r})\mathbf{U}^T(\mathbf{V}_L\otimes\mathbf{U}_L^T)\mathrm{col}(\mathbf{K}\mathbf{V}_L) \, .$$

Die Addition der drei Terme $\frac{\partial f_i}{\partial \mathrm{col}\mathbf{L}}$ führt zu einer notwendigen Bedingung für das Minimum in Gl.(20), wobei $\mathbf{U}$ und $\mathbf{U}_L$ aus Gln.(9) und (22) resultieren.

Schließlich folgt das Ergebnis als Gleichung in $\mathbf{L}$ zu

$$\boxed{[\mathbf{I}_{(n-r)r} + \mathbf{C}_I\mathbf{U}_1^T\mathbf{U}_1\mathbf{C}_I^T]\,\mathrm{col}\,\mathbf{L} - \mathbf{C}_I\mathbf{U}_2^T[\mathbf{V}_L\otimes\mathbf{U}_L^T]\,\mathrm{col}[\mathbf{K}\mathbf{V}_L] = \mathbf{0}} \, , \tag{20}$$

wobei $\mathbf{V}_L$ und $\mathbf{U}_L$ von $\mathbf{L}$ abhängen

$$\mathbf{V}_L \triangleq \begin{bmatrix} \mathrm{loc}_{(n-r)\times n}(\mathbf{U}\ \mathrm{col}\,\mathbf{L}\mathbf{C}) \\ \mathbf{C} \end{bmatrix}^{-1} \tag{21}$$

$$\mathbf{U}_L^T \triangleq \left(\mathbf{I}_{n-r} \vdots \mathbf{0}_{n-r,r}\right) \mathbf{V}_L^T\,\mathbf{K}^T \, . \tag{22}$$

Die Matrizen $\mathbf{U}_1$ und $\mathbf{U}$ sind in Gl. 11 und 9 definiert, $\mathbf{C}_I$ in Gl.(8). Weiters ist „loc“ die Umkehrung der „col“-Operation, d.h. die Rückermittlung der dimensionsrichtigen Matrix aus der Spalte „col“ (siehe auch Anhang C).

Ohne Abkürzungen lautet die Bestimmungsgleichung in $\mathbf{L}$

$$[\mathbf{I}_{(n-r)r} + (\mathbf{C} \otimes \mathbf{I}_{n-r})\mathbf{U}^T\{(\mathbf{B}\mathbf{B}^T) \otimes \mathbf{I}_{n-r}\}\mathbf{U}](\mathbf{C}^T \otimes \mathbf{I}_{n-r})\mathrm{col}\mathbf{L} = \tag{23}$$

$$= (\mathbf{C} \otimes \mathbf{I}_{n-r})\mathbf{U}^T\Big\{ \begin{bmatrix} \mathrm{loc}_{(n-r)\times n}(\mathbf{U}\ \mathrm{col}\,\mathbf{L}\mathbf{C}) \\ \mathbf{C} \end{bmatrix}^{-1} \tag{24}$$

$$\otimes\{(\mathbf{I}_{n-r} \vdots \mathbf{0}_{n-r,r}) \begin{bmatrix} \mathrm{loc}_{(n-r)\times n}(\mathbf{U}\ \mathrm{col}\,\mathbf{L}\mathbf{C}) \\ \mathbf{C} \end{bmatrix}^{-T} \mathbf{K}^T\}\Big] \tag{25}$$

$$\times\mathrm{col}\Big[\mathbf{K} \begin{bmatrix} \mathrm{loc}_{(n-r)\times n}(\mathbf{U}\ \mathrm{col}\,\mathbf{L}\mathbf{C}) \\ \mathbf{C} \end{bmatrix}^{-1} \Big\}\,. \tag{26}$$

Die Dynamik des Gesamtsystems in $(\mathbf{x} \vdots \tilde{\mathbf{z}})^T$ resultiert unabhängig von der Wahl von $\mathbf{L}$ immer aus der Koeffizientenmatrix

$$\begin{pmatrix} \mathbf{A}+\mathbf{B}\mathbf{K} & -\mathbf{B}\mathbf{K}_z \\ \mathbf{0} & \mathbf{F}_z \end{pmatrix}\,. \tag{27}$$

Beispiel 1: Für eine Zweigrößenstrecke der Ordnung $n = 3$ mit $m = r = 2$ und

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -1 & -3 \end{pmatrix}\,, \quad \mathbf{B} = \begin{pmatrix} 0 & 0,5 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}\,, \tag{28}$$

$$\mathbf{C} = \begin{pmatrix} 1 & 2 & 1 \\ 1,5 & 1 & 2 \end{pmatrix}, \quad \mathbf{F}_z = -3,6595, \tag{29}$$

$$\mathbf{K} = \begin{pmatrix} -0,5675 & -0,8466 & -0,2115 \\ -0,4853 & -0,4952 & -0,1764 \end{pmatrix} \tag{30}$$

lautet das Ergebnis

$$\mathbf{L} = (-0,2016 \;\; 0,4060), \quad \mathbf{H} = (0,1914 \;\; 0,6043), \tag{31}$$

$$\mathbf{K}_z = \begin{pmatrix} 0,5512 \\ 0,5508 \end{pmatrix}, \quad \mathbf{K}_y = \begin{pmatrix} -0,4475 & -0,0571 \\ -0,2250 & -0,1507 \end{pmatrix}\,. \tag{32}$$

Zur Beurteilung, wie empfindlich das Ergebnis gegenüber einem veränderten Beobachterpol reagiert, ist das Ergebnis in verschiedene Beobachterpole eingebettet worden (Fig. 2). Die Matrix $\mathbf{F}_z$ wird zu $(\alpha/1,5)\cdot\mathbf{F}_z$ erweitert, wobei α als Parameter mit der Obergrenze 1,5 fungiert. Die minimale Normsumme $\min_{\mathbf{L}}\{f_N\}$ steigt nur bei größeren α-Werten mit α an, bei niedrigen bedeutet Steigerung der Dynamik sinkende Normsumme. Ein Minimum liegt bei α bei 0,585.

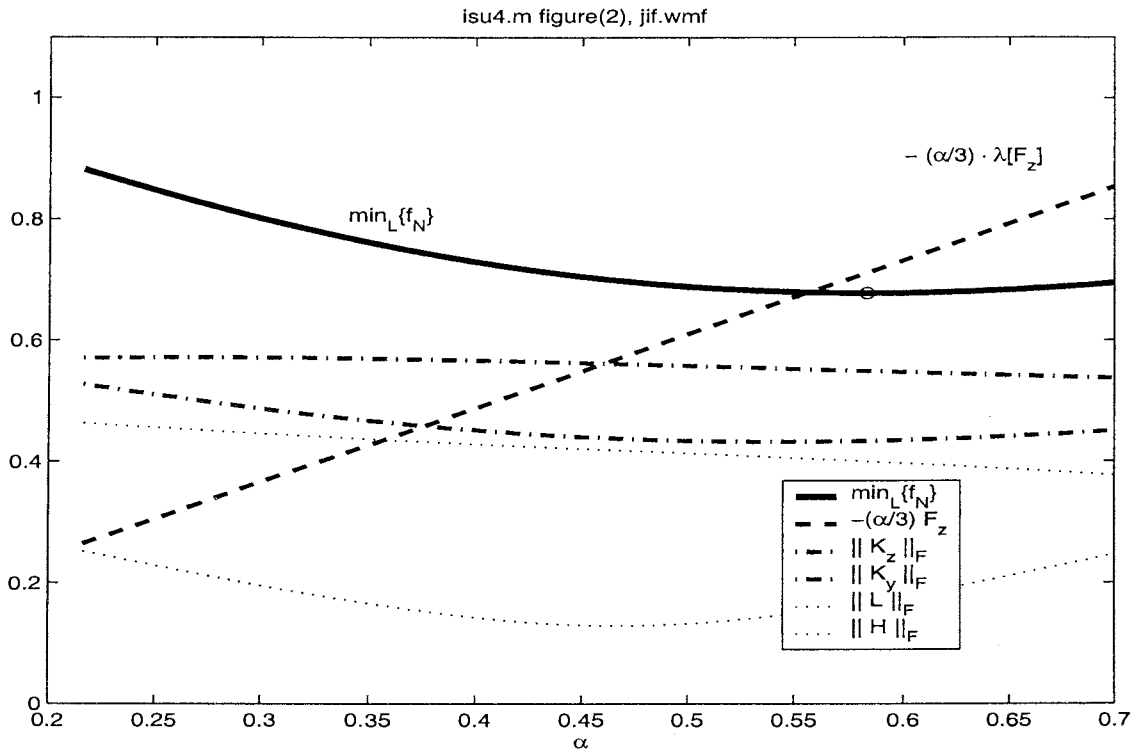


Figure 2: Verlauf von $\min_{\mathbf{L}} f_N$ sowie Normen der beteiligten Matrizen

3 Gewichtung der Matrix-Elemente bei der Norm-Minimierung

Für die Einführung von Gewichtsmatrizen bei der Normminimierung spricht: Treten im praktischen Betrieb Signale $\mathbf{y}$ und $\hat{\mathbf{z}}$ auf, die sich in der Größenordnung unterscheiden, dann erscheint es sinnvoll, Teile der Matrizen (z. B. Spalten von $\mathbf{K}_y$) mit unterschiedlichen Gewichten auszustatten, ohne dabei im Fixkomma-Rechner einen Overflow zu riskieren.

3.1 Ableitung der gewichteten Normsumme f_{Nw} nach colL

Soll aus einer beliebigen Reglermatrix $\mathbf{K}_x \in \mathcal{R}^{m\times n}$ nur das eine Element $K_x(i,j)$ mit w gewichtet werden, so ist $\mathbf{W}_v\mathbf{K}_x\mathbf{W}_n$ zu verwenden, und zwar mit Diagonalmatrizen $\mathbf{W}_v = \mathbf{0}_{m,m},\ W_v(i,i) = w$; $\mathbf{W}_n = \mathbf{0}_{n,n},\ W_n(j,j) = 1$. Für mehrere Elemente ist

$$\sum_k \mathbf{W}_{v,k}\mathbf{K}_x\mathbf{W}_{n,k} \tag{33}$$

zu bilden.

Für die drei gewichteten Komponenten f_{1w}, f_{2w}, f_{3w} wird auf Gl.(33) Bezug genommen.

In Summe bilden sie ein gewichtetes f_{Nw}. Man erhält für

$$f_{1w} \triangleq \|\mathrm{col}(\sum_k \mathbf{W}_{vLk}\mathbf{L}\mathbf{W}_{nLk})\|_F^2 = \|\sum_k \mathrm{col}(\mathbf{W}_{vLk}\mathbf{L}\mathbf{W}_{nLk})\|_F^2 \quad (34)$$

$$= \|(\sum_k \mathbf{W}_{nLk}^T \otimes \mathbf{W}_{vLk})\mathrm{col}\,\mathbf{L}\|_F^2 \quad (35)$$

mit Gl.(74) bei $\mathbf{N}_2 = \mathbf{I},\ \mathbf{N}_4 = \mathbf{0}$ und $\mathbf{N}_5 = \mathbf{0}$

$$\frac{\partial f_{1w}}{\partial \mathrm{col}\mathbf{L}} = 2\mathbf{W}_L \mathrm{col}\,\mathbf{L}\ ,$$

wobei $\mathbf{W}_L \in \mathcal{R}^{r(n-r)\times r(n-r)}$

$$\mathbf{W}_L \triangleq (\sum_k \mathbf{W}_{nLk} \otimes \mathbf{W}_{vLk}^T)(\sum_k \mathbf{W}_{nLk}^T \otimes \mathbf{W}_{vLk}).$$

Für f_{2w} resultiert nach Gln.(12) und (74)

$$f_{2w} \triangleq \|\sum_k \mathbf{W}_{vHk}\mathbf{H}\mathbf{W}_{nHk}\|_F^2 = \|(\sum_k \mathbf{W}_{nHk}^T \otimes \mathbf{W}_{vHk})\mathrm{col}\mathbf{H}\|_F^2 \quad (36)$$

$$= \|(\sum \mathbf{W}_{nHk}^T \otimes \mathbf{W}_{vHk})\mathbf{U}_1\mathbf{C}_I^T \mathrm{col}\ \mathbf{L}\|_F^2 \quad (37)$$

$$\frac{\partial f_{2w}}{\partial \mathrm{col}\,\mathbf{L}} = 2\mathbf{C}_I\mathbf{U}_1^T\mathbf{W}_H\mathbf{U}_1\mathbf{C}_I^T \mathrm{col}\,\mathbf{L} \quad (38)$$

unter $\mathbf{W}_H \in \mathcal{R}^{m(n-r)\times m(n-r)}$

$$\mathbf{W}_H \triangleq (\sum \mathbf{W}_{nHk} \otimes \mathbf{W}_{vHk}^T)(\sum \mathbf{W}_{nHk}^T \otimes \mathbf{W}_{vHk})\ .$$

Für f_{3w} schließlich gilt

$$f_{3w} \triangleq \|\sum_k \mathbf{W}_{vKk}\mathbf{K}\begin{pmatrix}\mathbf{T}\\ \mathbf{C}\end{pmatrix}^{-1}\mathbf{W}_{nKk}\|_F^2 \quad (39)$$

$$\rightsquigarrow \|\sum_k \mathbf{W}_{vKk}\mathbf{K}\mathbf{V}_L\mathbf{W}_{nKk} - \sum_k \mathbf{W}_{vKk}\mathbf{K}\mathbf{V}\begin{pmatrix}\Delta\mathbf{T}\\ \mathbf{0}\end{pmatrix}\mathbf{V}_L\mathbf{W}_{nKk}\|_F^2 \quad (40)$$

$$= \|\sum_k \mathbf{W}_{vKk}\mathbf{K}\mathbf{V}_L\mathbf{W}_{nKk} \quad (41)$$

$$-\sum \mathbf{W}_{vKk}\mathbf{K}\mathbf{V}_L\begin{pmatrix}\mathbf{I}_{n-r}\\ \mathbf{0}_{r\times(n-r)}\end{pmatrix} \times \Delta\mathbf{T}\cdot\mathbf{V}_L\mathbf{W}_{nKk}\|_F^2 \quad (42)$$

$$= \|(\sum_k \mathbf{W}_{nKk}^T \otimes \mathbf{W}_{vKk})\mathrm{col}(\mathbf{K}\mathbf{V}) \quad (43)$$

$$-(\sum \mathbf{W}_{nKk}^T \otimes \mathbf{W}_{vKk})[\mathbf{V}_L^T \otimes \mathbf{K}\mathbf{V}_L\begin{pmatrix}\mathbf{I}_{n-r}\\ \mathbf{0}_{r,n-r}\end{pmatrix}]\mathbf{U}\mathbf{C}_I^T\mathrm{col}\Delta\mathbf{L}\|_F^2 \quad (44)$$

$$\frac{\partial f_{3w}}{\partial \mathrm{col}\Delta\mathbf{L}} = -2\mathbf{C}_I\mathbf{U}^T[\mathbf{V}_L \otimes (\mathbf{I}_{n-r}\vdots\mathbf{0}_{n-r,r})\mathbf{V}_L^T\mathbf{K}^T]\mathbf{W}_K\mathrm{col}(\mathbf{K}\mathbf{V}_L) \quad (45)$$

unter $\mathbf{W}_K \in \mathcal{R}^{mn \times mn}$

$$\mathbf{W}_K \triangleq (\sum_k \mathbf{W}_{nKk} \otimes \mathbf{W}_{vKk}^T)(\sum_k \mathbf{W}_{nKk}^T \otimes \mathbf{W}_{vKk}) \ .$$

Aus der Summe $f_{Nw} = f_{1w} + f_{2w} + f_{3w}$ und deren Ableitungen resultiert

$$\boxed{(\mathbf{W}_L + \mathbf{C}_I \mathbf{U}_1^T \mathbf{W}_H \mathbf{U}_1 \mathbf{C}_I^T)\mathrm{col}\mathbf{L} - \mathbf{C}_I \mathbf{U}^T [\mathbf{V}_L \otimes \mathbf{U}_L^T] \mathbf{W}_K \mathrm{col}[\mathbf{K}\mathbf{V}_L] = \mathbf{0} \ ,} \qquad (46)$$

mit $\mathbf{U}_L$ und $\mathbf{V}_L$ als Funktionen von $\mathbf{L}$ laut Gln.(21) und (22).

Beispiel 2. Gewichtung: Verwendet werden die gleichen Angaben wie in Beispiel 1, ergänzt um $\mathbf{W}_{vL} = \mathbf{W}_{vH} = 1$,

$$\mathbf{W}_{nL} = \begin{pmatrix} 0,2 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{W}_{nH} = \begin{pmatrix} 1 & 0 \\ 0 & 0,2 \end{pmatrix},$$

$$\mathbf{W}_{vK} = \begin{pmatrix} 1 & 0 \\ 0 & 0,9 \end{pmatrix}, \qquad \mathbf{W}_{nK} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0,5 & 0 \\ 0 & 0 & 0,5 \end{pmatrix}.$$

Die zweite Komponente von $\mathbf{y}$ ist fünf mal so groß zu erwarten wie die erste, deshalb wird $\mathbf{L}$ (nach Gl.(1) der Vorfaktor von $\mathbf{y}$) mit Gewichtsfaktoren 0,2 und 1 ausgestattet. Die Matrizen $\mathbf{W}_L$, $\mathbf{W}_H$, $\mathbf{W}_K$ werden mit $k = 1$ ermittelt. Man erhält

$$\mathbf{L} = (-0,3648 \ \ 0,7250), \quad \mathbf{H} = (0,3386 \ \ 1,0760),$$

$$\mathbf{K}_z = \begin{pmatrix} 0,3092 \\ 0,3090 \end{pmatrix}, \ \mathbf{K}_y = \begin{pmatrix} -0,4470 & -0,0572 \\ -0,2245 & -0,1508 \end{pmatrix} .$$

Mit Einbettung in einen Parameter α findet man Fig. 3. Ausnützen wird man den Bereich $\alpha < 1$.

Beispiel 3. Partielle Gewichtung: Bei Beschränkung der Normen auf ein Matrix-Element von $\mathbf{L}$ und $\mathbf{H}$, und voller Freigabe von $\mathbf{K}_z$ und $\mathbf{K}_y$ erhält man ein Resultat nach Fig. 4.

4 Alternative Realisierung

Eine alternative Realisierung eines Kontrollbeobachters mit $\hat{\mathbf{u}}$ als Ausgang und $\mathbf{y}$ als einzigem Eingang lautet in der dafür maßgeblichen Systemmatrix

$$\left[\begin{array}{c|c} \mathbf{F}_z + \mathbf{H}\mathbf{K}_z & \mathbf{L} + \mathbf{H}\mathbf{K}_y \\ \hline \mathbf{K}_z & \mathbf{K}_y \end{array}\right] . \qquad (47)$$

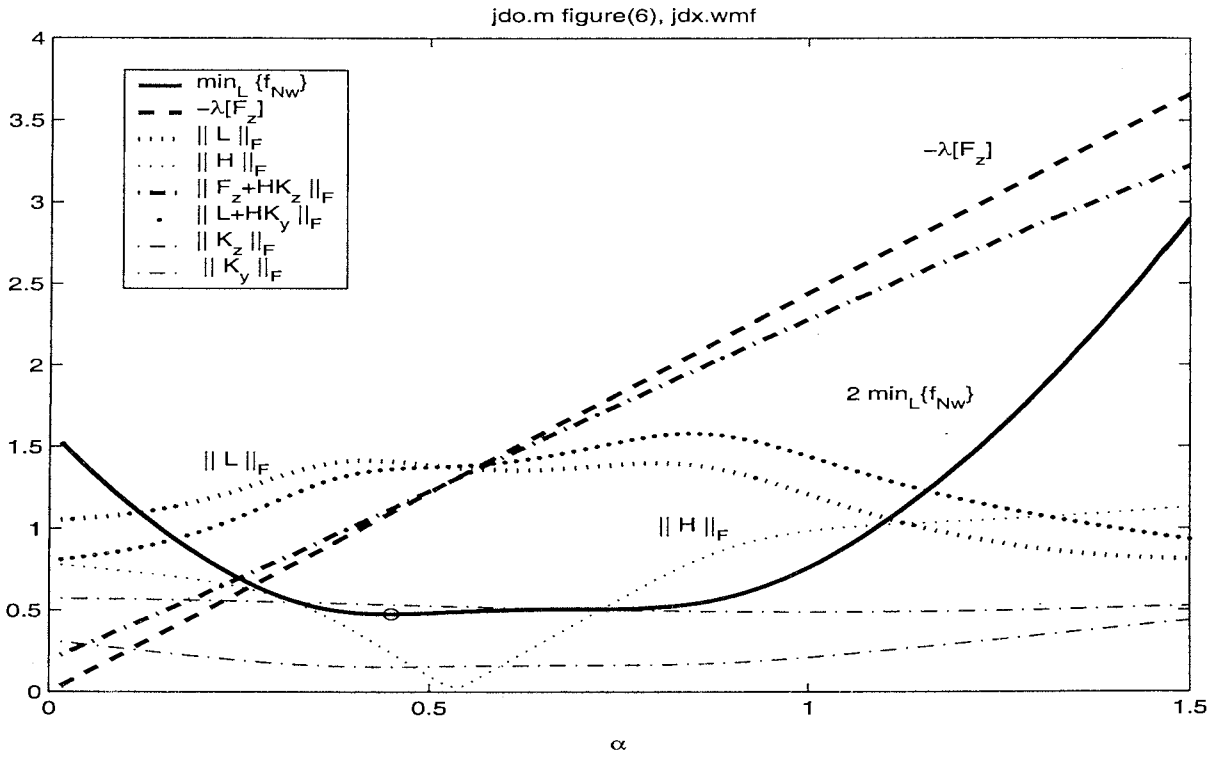


Figure 3: Normen bei Gewichtung der Verstärkungsmatrizen

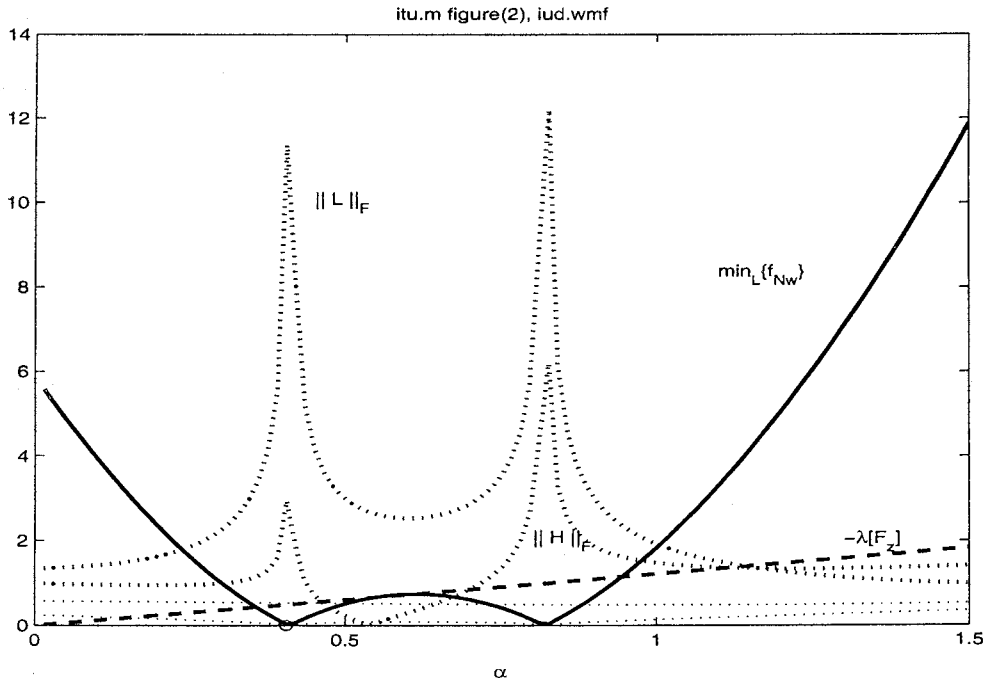


Figure 4: Partielle Normsummen-Gewichtung unter Freigabe von der Norm von $\mathbf{L}$ und $\mathbf{H}$ in Abhängigkeit von α

Zum Vergleich sind die Normen der Matrizen der ersten Zeile in Fig. 3 aufgenommen worden.

Wegen der allgemeinen Beziehung $\|\mathbf{MN}\|_F \leq \|\mathbf{M}\|_F \|\mathbf{N}\|_F$ ist sichergestellt, dass bei kleiner Norm von $\mathbf{H}$, $\mathbf{K}_z$, $\mathbf{L}$ und $\mathbf{K}_y$ auch alle Teilmatrizen in Gl.(47) beschränkt bleiben, was für Festkomma-Arithmetik zu begrüßen ist.

Ein gemeinsamer Vektor aus $\mathbf{x}(t)$ und $\hat{\mathbf{z}}(t)$ einerseits und $\mathbf{y}_{ref}(t)$ und $\mathbf{n}_r(t)$ andererseits erweist sich für eine Gesamtbetrachtung als zweckmäßig, wobei $\mathbf{n}_r(t)$ ein Rauschsignal ist, das dem Ausgang $\mathbf{y}(t)$ additiv zugesetzt wird. Dann lautet die Zustandsmatrix des geschlossenen Regelkreises

$$\begin{pmatrix} \mathbf{A}+\mathbf{B}\mathbf{K}_y\mathbf{C} & \mathbf{B}\mathbf{K}_z \\ \mathbf{L}\mathbf{C}+\mathbf{H}\mathbf{K}_y\mathbf{C} & \mathbf{F}_z+\mathbf{H}\mathbf{K}_z \end{pmatrix} . \tag{48}$$

Mit kleiner Matrixnorm von $\mathbf{K}_y$ gelingt es auch, der großen Verstärkung hochfrequenten Messrauschens $\mathbf{n}_r(t)$ in $\mathbf{y}(t)$ auf $\mathbf{u}(t)$ vorzubeugen.

5 Potenzen des Norm-Quadrats

Wird eine höhere Potenz der Frobenius-Norm als Basis der Minimierung gewählt,

$$f_{Np} = \|\mathbf{L}\|_F^{2p_L} + \|\mathbf{H}\|_F^{2p_H} + \|\mathbf{K}_z \vdots \mathbf{K}_y\|_F^{2p_K} ,$$

so erhält man ein zu Gl.(46) sehr ähnliches Ergebnis. Unter Verwendung von Gl.(76) gilt nämlich

$$\frac{\partial}{\partial \mathbf{K}} \|\mathbf{K}\|_F^{2p} = 2p\|\mathbf{K}\|_F^{2(p-1)}\mathbf{K} . \tag{49}$$

Gegenüber dem einfachen Normquadrat kommt also ein Skalar $p\|\mathbf{K}\|_F^{2(p-1)}$ als Faktor hinzu. Dies bedeutet, dass das Ergebnis Gl.(46) verwendet werden kann, wenn zusätzlich zu der Gewichtsmatrix $\mathbf{W}$ ein Faktor dieser Art eingebaut wird.

6 Nummerische Aspekte

Aufgrund der quadratischen Bewertung von $\mathbf{L}$, $\mathbf{H}$, $\mathbf{K}_z$ und $\mathbf{K}_y$ sind die Vorzeichen der Matrixelemente belanglos und es existiert insbesonders bei hoher Ordnung n eine Mehrzahl von Lösungen in lokalen Minima je nach Startwert, so wie auch überhaupt eine unendliche Zahl an Lösungen auch bei Eigenwertvorgabe bei MIMO Systemen existiert.

Die Anzahl der Unbekannten wächst in erster Linie mit der Ordnung n sehr stark an, auch die Anzahl der möglichen Lösungen. Die knapp im Gütekriterium nebeneinander

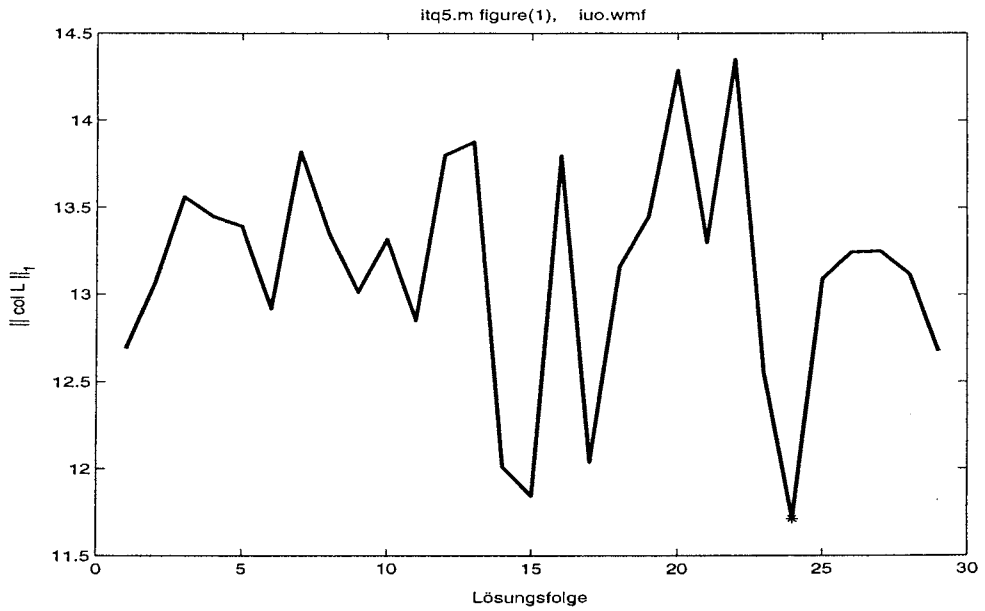


Figure 5: Lösungsfolge unter Beurteilung nach $\|\text{col}\mathbf{L}\|_1$ und Auswahl des Minimalpunkts ('$\star$')

liegenden Lösungen sind nicht auf geringfügige Unterschiede aus dem Nichterreichen des Minimums zurückzuführen.

Eine Nachjustierung ist auf verschiedene Art möglich.

Für ein- und dieselbe Lösungsfolge werden in den Fig. 5 und 6 zwei verschiedene zusätzliche Bewertungskriterien gewählt, nämlich nach den direkten Matrix-Normen $\|\text{col}\mathbf{L}\|_1$ (Summen-Norm) und $\|\text{col}\mathbf{K}_z\|_\infty$ (maximaler Wert der Beträge aller Elemente).

Beispiel 4. Nachjustierung: Für einen instabilen Prozess $n = 7,\ m = 3,\ r = 4$ mit den Polstellen bei $-1.34 \pm j0.95;\ -0.8;\ -0.06; +0.67;\ +1.04 \pm 1.21$ und den detaillierten Zahlenangaben

$$\mathbf{A} = \begin{pmatrix} -0.8379 & -0.8526 & 0.5398 & 0.7777 & 0.9991 & 0.3322 & -0.9653 \\ 0.7022 & -0.6003 & -0.2499 & -0.7968 & -0.5760 & -0.7381 & 0.6388 \\ 0.1241 & -0.9010 & 0.6468 & -0.8694 & -0.003180 & -0.8092 & 0.2423 \\ -0.3614 & 0.1334 & -0.9067 & -0.5314 & -0.4190 & -0.9703 & 0.1204 \\ -0.2502 & -0.7562 & 0.1958 & 0.8662 & 0.3455 & -0.4236 & -0.5119 \\ 0.7356 & 0.04422 & 0.8983 & -0.8737 & 0.9160 & 0.6335 & 0.6440 \\ -0.2556 & -0.7659 & -0.4224 & -0.4716 & 0.5331 & 0.9710 & -0.4736 \end{pmatrix} \tag{50}$$

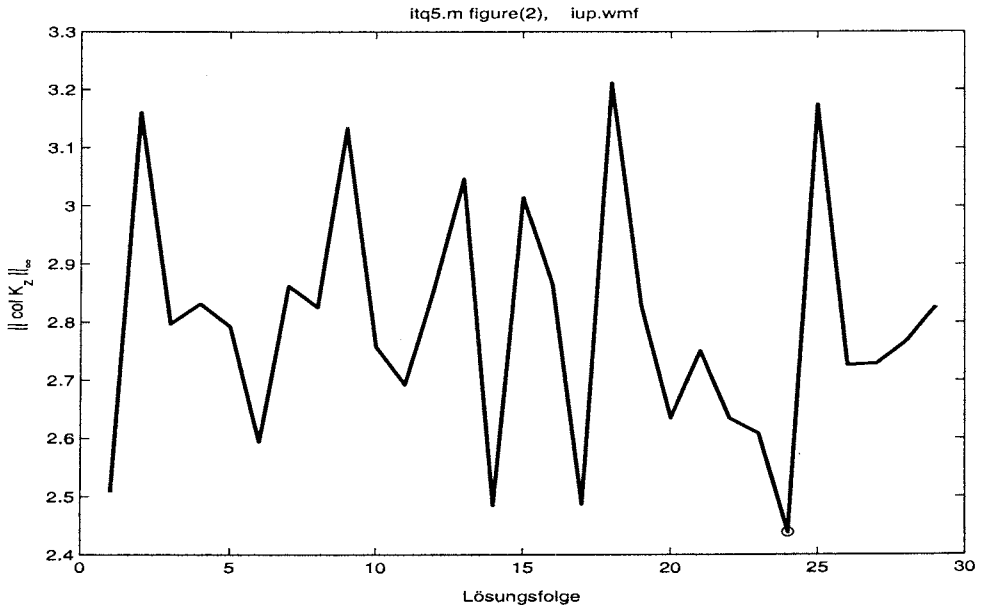


Figure 6: Lösungsfolge unter Beurteilung nach $\|\mathrm{col}\mathbf{K}_z\|_\infty$ und Auswahl des Minimalpunkts ('o')

$$\mathbf{B} = \begin{pmatrix} 0.5073 & 0.2731 & -0.9742 \\ 0.3193 & -0.6594 & -0.3792 \\ -0.5719 & 0.07920 & 0.5582 \\ 0.2042 & 0.2468 & -0.3854 \\ 0.2099 & 0.3718 & 0.8534 \\ 0.3190 & 0.3547 & 0.3574 \\ -0.6333 & 0.7537 & -0.8514 \end{pmatrix} \tag{51}$$

$$\mathbf{C} = \begin{pmatrix} -0.8587 & -0.08359 & -0.8514 & 0.5418 & 0.005752 & -0.7738 & -0.7558 \\ -0.9761 & 0.4064 & -0.6135 & -0.3721 & 0.8954 & 0.6243 & 0.5253 \\ -0.5457 & 0.1650 & -0.2408 & 0.2764 & 0.6561 & 0.8165 & 0.4436 \\ 0.03250 & 0.01841 & -0.4471 & 0.9731 & 0.8351 & -0.6872 & 0.3033 \end{pmatrix} \tag{52}$$

wird $\mathbf{K}$ mit $\mathbf{R} = 4\mathbf{I}_3$ und $\mathbf{Q} = \mathbf{I}_7$ als LQ-Regler vorgeschrieben

$$\mathbf{K} = \begin{pmatrix} 0.0395 & -0.5769 & 1.6205 & -0.7932 & -0.5127 & -0.3646 & -0.0118 \\ -0.6529 & 0.5475 & -0.9790 & 0.9250 & -1.4497 & -2.2597 & -0.6753 \\ -0.5864 & 1.1560 & -2.3248 & 1.0030 & -1.5420 & -1.4261 & -0.0083 \end{pmatrix}. \tag{53}$$

Die $n - r$ Eigenwerte von $\mathbf{F}_z$ werden um 50% weiter links als die kleinsten Eigenwerte

von $\mathbf{A}+\mathbf{BK}$ fest vorbestimmt, bezogen auf die Realteile

$$\mathbf{F}_z = \begin{pmatrix} -2.0475 & 0 & 0 \\ 0 & -2.0475 & 0 \\ 0 & 0 & -1.7588 \end{pmatrix} . \tag{54}$$

Man erhält

$$\mathbf{K}_z = \begin{pmatrix} -0.6960 & -2.4017 & -2.2775 \\ 1.4983 & -1.3734 & -1.9709 \\ 2.7266 & 0.7831 & -0.7115 \end{pmatrix} \tag{55}$$

$$\mathbf{K}_y = \begin{pmatrix} -0.7589 & -3.0606 & -0.8789 & 1.4590 \\ 2.2106 & -1.0397 & -0.6343 & 2.2956 \\ 2.8843 & 1.4621 & 1.1207 & 1.5284 \end{pmatrix} \tag{56}$$

$$\mathbf{L} = \begin{pmatrix} -1.8410 & -0.8159 & -0.3961 & -0.6678 \\ 1.5882 & -1.6755 & 0.1143 & -1.1069 \\ 1.1328 & -1.1188 & 0.6200 & 2.1672 \end{pmatrix} \tag{57}$$

$$\mathbf{H} = \begin{pmatrix} -0.5371 & -0.4294 & -0.3183 \\ -0.2120 & 0.1416 & 2.7186 \\ -0.0746 & -0.0806 & 0.1189 \end{pmatrix} \tag{58}$$

$$\mathbf{T} = \begin{pmatrix} 1.3446 & 0.5691 & 0.7523 & 0.0836 & -1.2541 & 0.3330 & 0.2987 \\ 0.7890 & -0.8549 & -0.0580 & -0.7930 & -1.0982 & -1.1665 & -0.4824 \\ 0.5690 & 0.4108 & 0.0344 & 1.9953 & 0.8431 & -0.0480 & -0.0084 \end{pmatrix} . \tag{59}$$

Mit den obigen Daten ist aus vielen Startwerten eine Folge von 29 verschiedene Lösungen gefunden worden, natürlich auch solche aus lokalen Minima. Auch bezüglich der Vorzeichen von $\mathbf{z}$ ergeben sich Alternativen, da das Vorzeichen der Verstärkungsmatrizenelemente nicht in die Norm eingeht. Die bestmögliche ist die mit Nummer 26 (Fig. 7).

In den Lösungen $\min_{\mathbf{L}}\{f_N\}$ gibt es drei Ergebnisniveaus, und zwar bei etwa 94,5; 107 und 112.

Für ein- und dieselbe Lösungsfolge werden in den Fig. 5 und 6 zwei verschiedene *zusätzliche* Bewertungskriterien gewählt, nämlich nach den direkten Matrix-Normen $\|\mathrm{col}\mathbf{L}\|_1$ (Summen-Norm) und $\|\mathrm{col}\mathbf{K}_z\|_\infty$ (nach dem maximalen Wert der Beträge aller Elemente).

Eine Minimumsuche über $\|\mathrm{col}\mathbf{L}\|_1$ zeigt die Lösung bei Lösungspunkt 24 in Fig. 5. Die Suche über $\|\cdot\|_\infty$ gibt ebenfalls Punkt 24 laut Fig. 6. In $\min_{\mathbf{L}}\{f_N\}$ ist zwar Punkt 26 der beste, aber zu Punkt 24 kaum ein Unterschied.

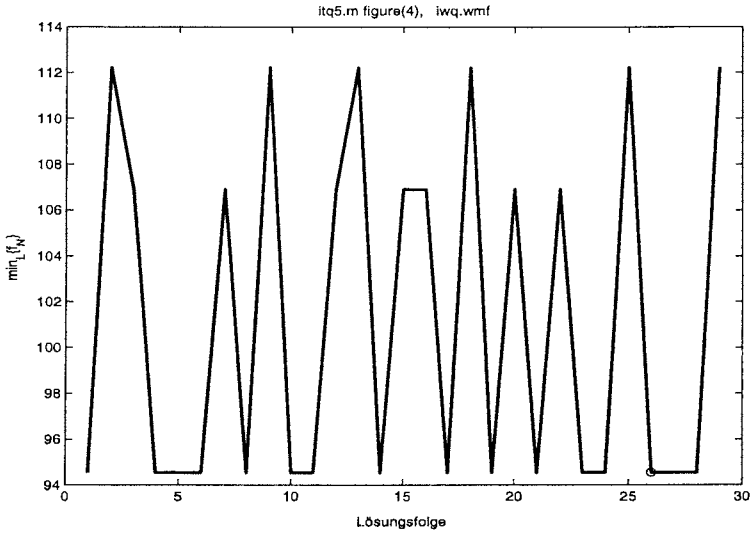


Figure 7: Optimalergebnisse von $\min_{\mathbf{L}}\{f_N\}$

Aufgrund der Mehrzahl der Lösungen, die oft in derselben Größenordnung liegen, ist die direkte Minimumsuche der analytischen Bestimmung der Minimalpunkte mit Aufsuchen der Nullstellen der ersten Ableitung unterlegen.

7 Zusammenfassung

Die Aufgabe wurde behandelt, die Normsumme aller Verstärkungs-Matrizen zu minimieren, die für die Beobachtung und Regelungsfunktion erforderlich ist. Dies ist vor allem bei Prozessoren mit Festkomma-Arithmetik empfehlenswert. Die Aufgabe lässt sich auf eine nichtlineare algebraische Gleichung in der Beobachter-Eingangsmatrix zurückführen. Die Lösung wurde auch für individuelle Auswahl oder Gewichtung der Matrixelemente besorgt. Bei praktischen Problemen kann man nicht die globale und elegante Lösungsmöglichkeit in den Vordergrund stellen. Hauptsache man erreicht überhaupt eine nützliche Lösung.

In bestimmten Fällen ist höhere Beobachterdynamik bei kleiner werdenden Beobachtermatrizen möglich.

Aufgrund einer Mehrzahl knapp nebeneinanderliegender Lösungen ist ein Nachjustieren empfehlenswert, weil es zusätzlich andere Kriterien zu optimieren ermöglicht, ohne das ursprünglich Ergebnis wesentlich zu beeinträchtigen.

Anhang

A Dimensionen und Abkürzungen

$$\mathbf{x} \in \mathcal{R}^n;\ \ \mathbf{z},\ \hat{\mathbf{z}} \in \mathcal{R}^{n-r};\ \ \mathbf{y} \in \mathcal{R}^r;\ \ \mathbf{u} \in \mathcal{R}^m\ ; \tag{60}$$

$$\mathbf{A} \in \mathcal{R}^{n\times n};\ \ \mathbf{B} \in \mathcal{R}^{n\times m};\ \ \mathbf{C} \in \mathcal{R}^{r\times n}\ ; \tag{61}$$

$$\mathbf{K} \in \mathcal{R}^{m\times n};\ \ \mathbf{K}_z \in \mathcal{R}^{m\times(n-r)};\ \ \mathbf{K}_y \in \mathcal{R}^{m\times r}\ ; \tag{62}$$

$$\mathbf{F}_z \in \mathcal{R}^{(n-r)\times(n-r)};\ \ \mathbf{T} \in \mathcal{R}^{(n-r)\times n}\ ; \mathbf{L} \in \mathcal{R}^{(n-r)\times r}\ ; \tag{63}$$

$$\mathbf{H} \in \mathcal{R}^{(n-r)\times m};\ \ \mathbf{C}_I \triangleq \mathbf{C} \otimes \mathbf{I}_{n-r} = \mathcal{R}^{r(n-r)\times n(n-r)}\ ; \tag{64}$$

$$\mathbf{U} \triangleq (\mathbf{A}^T \otimes \mathbf{I}_{n-r} - \mathbf{I}_n \otimes \mathbf{F}_z)^{-1} \ \in \mathcal{R}^{n(n-r)\times n(n-r)}\ ; \tag{65}$$

$$\mathbf{U}_1 \triangleq (\mathbf{B}^T \otimes \mathbf{I}_{n-r})\mathbf{U} \ \in \mathcal{R}^{m(n-r)\times n(n-r)}\ ; \tag{66}$$

$$\mathbf{U}_L^T \triangleq \left(\mathbf{I}_{n-r} \vdots 0_{n-r,r}\right) \mathbf{V}_L^T\, \mathbf{K}^T \in \mathcal{R}^{(n-r)\times m}\ . \tag{67}$$

$$\mathbf{V}_L \triangleq \begin{pmatrix} \mathbf{T} \\ \mathbf{C} \end{pmatrix}^{-1} \ \in \mathcal{R}^{n\times n}\ ; \tag{68}$$

$$\mathbf{W}_L \ \triangleq\ (\sum_k \mathbf{W}_{nLk} \otimes \mathbf{W}_{vLk}^T)(\sum_k \mathbf{W}_{nLk}^T \otimes \mathbf{W}_{vLk}) \ \in \mathcal{R}^{r(n-r)\times r(n-r)}\ ; \tag{69}$$

$$\mathbf{W}_H \triangleq (\sum \mathbf{W}_{nHk} \otimes \mathbf{W}_{vHk}^T)(\sum \mathbf{W}_{nHk}^T \otimes \mathbf{W}_{vHk}) \in \mathcal{R}^{m(n-r)\times m(n-r)}\ ; \tag{70}$$

$$\mathbf{W}_K \triangleq (\sum_k \mathbf{W}_{nKk} \otimes \mathbf{W}_{vKk}^T)(\sum_k \mathbf{W}_{nKk}^T \otimes \mathbf{W}_{vKk}) \ \in \mathcal{R}^{mn\times mn}\ . \tag{71}$$

B Benützte Rechenregeln

Formal werden folgende Rechenhilfsmittel herangezogen:

- Kroneckerprodukt $\otimes$ und Vektorisierung mittels „col“

$$\mathrm{col}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\mathrm{col}\ \mathbf{B}. \tag{72}$$

- Für die skalare Funtion $f(\mathbf{K})$ folgt aus einer geeigneten Separationsprozedur von $\mathbf{G}$ und für $\Delta\mathbf{K}$ klein im Normsinne,

$$f(\mathbf{K}+\Delta\mathbf{K}) - f(\mathbf{K}) \ \triangleq\ \mathrm{tr}[\mathbf{G}\Delta\mathbf{K}] \ \ \rightsquigarrow\ \ \frac{\partial f(\mathbf{K})}{\partial \mathbf{K}} = \mathbf{G}^T\ . \tag{73}$$

- Die Ableitung einer kombinierten Normgröße liefert bei $\mathbf{N}_i \in \mathcal{C},\ \mathbf{M} \in \mathcal{R}$ [30]

$$\begin{aligned} & \frac{\partial}{\partial \mathbf{M}} \|\mathbf{N}_1\mathbf{M}\mathbf{N}_2 + \mathbf{N}_3\mathbf{M}\mathbf{N}_4 + \mathbf{N}_5\|_F^2 = \\ = \ & 2\Re e\ \{\mathbf{N}_3^H(\mathbf{N}_1\mathbf{M}\mathbf{N}_2 + \mathbf{N}_3\mathbf{M}\mathbf{N}_4 + \mathbf{N}_5)\mathbf{N}_4^H \\ + \ & \mathbf{N}_1^H(\mathbf{N}_1\mathbf{M}\mathbf{N}_2 + \mathbf{N}_3\mathbf{M}\mathbf{N}_4 + \mathbf{N}_5)\mathbf{N}_2^H\} \ . \end{aligned} \tag{74}$$

- Frobenius-Norm im Quadrat

$$\|\mathbf{A}\|_F^2 = \sum_{i,j} |A_{ij}|^2 = \mathrm{tr}[\mathbf{A}^H\mathbf{A}] = (\mathrm{col}\mathbf{A})^H \mathrm{col}\mathbf{A},$$

  wobei ^{H} die konjugierte Transponierte darstellt.

- Weiters gilt

$$\begin{gathered} \|\mathbf{A} \vdots \mathbf{B}\|_F^2 = \|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 \ , \\ (\mathbf{A} \otimes \mathbf{B})^T = \mathbf{A}^T \otimes \mathbf{B}^T \ , \\ (\mathbf{R} + \Delta\mathbf{R})^{-1} \doteq \mathbf{R}^{-1} - \mathbf{R}^{-1}\Delta\mathbf{R} \cdot \mathbf{R}^{-1} \qquad (75) \\ \text{bei } \|\Delta\mathbf{R}\|_F \ll \|\mathbf{R}\|_F \ , \\ \frac{\partial \|\mathbf{A}\mathbf{M}\|_F^2}{\partial \mathbf{M}} = 2\mathbf{A}^T \ , \\ \frac{\partial a[b(\mathbf{M})]}{\partial \mathbf{M}} = \frac{\partial a}{\partial b}\frac{\partial b}{\partial \mathbf{M}} \ . \qquad (76) \end{gathered}$$

C Differenzierung einer Matrixnorm unter col-Darstellung, Kronecker-Rückzerlegung

Für die Angabe $\mathrm{col}\mathbf{T} = \mathbf{F}\mathrm{col}\mathbf{K}_x,\ \mathbf{K}_x \in \mathcal{R}^{m\times n},\ \mathrm{col}\mathbf{K}_x = \mathcal{R}^{mn}\ , \mathbf{T} \in \mathcal{R}^{k\times l}$ gilt

$$\frac{\partial \|\mathbf{T}\|_F^2}{\partial \mathbf{K}} = \sum_j^{nl} \mathbf{A}_j^T (\sum_i^{nl} \mathbf{A}_i\mathbf{K}_x\mathbf{D}_i)\mathbf{D}_j^T \ , \tag{77}$$

wenn $\mathbf{F} = \sum_i^{nl} \mathbf{D}_i^T \otimes \mathbf{A}_i$ bei $\mathbf{F} \in \mathcal{R}^{kl\times mn},\ \mathbf{A}_i \in \mathcal{R}^{k\times m},\ \mathbf{D}_i \in \mathcal{R}^{n\times l}$ vorzerlegt wurde. Es ist nämlich

$$\mathrm{col}\mathbf{T} = \sum_i^{nl} (\mathbf{D}_i^T \otimes \mathbf{A}_i)\mathrm{col}\mathbf{K}_x = \mathrm{col}[\sum_i^{nl} \mathbf{A}_i\mathbf{K}_x\mathbf{D}_i] \tag{78}$$

$$\mathbf{T} = \sum_i^{nl} \mathbf{A}_i\mathbf{K}_x\mathbf{D}_i \ . \tag{79}$$

Formel aus Gl.(74) liefert Gl.(77).

Werden die $\mathbf{D}_i$ derart gewählt, dass bis auf ein Eins-Element nur Null-Elemente auftreten, dann sind die $\mathbf{A}_i$ bloß entsprechend ausgewählte Submatrizen von $\mathbf{F}$.

Literatur

[1] KARBASSI, S.M., An algorithm for minimizing the norm of state feedback controllers in eigenvalue assignment. Computers & mathematics with applications **41** (2001), S. 1317-1326.

[2] KAUTSKY, J., NICHOLS, N.K., UND VAN DOOREN, P., Robust pole assignment in linear state feedback. Int. J. Control 41 (1985), S. 1129-1155.

[3] KOUVARITAKIS, B., AND CAMERON, R., Pole placement with minimized norm controllers. IEE Proc. Pt. D, **127** (1980), S. 32-36.

[4] CAMERON, R., AND KOUVARITAKIS, B., Minimizing the norm of output feedback controllers used in pole placement: a dyadic approach. Int. J. Control **32** (1980), S. 759-770.

[5] CAMERON, R.G., New pole assignment algorithm with reduced norm feedback matrix. IEE Proc. Pt. D, **135** (1988), S. 111-118.

[6] TAM, H.K., LAM, J. Robust deadbeat pole assignment with gain constraints: An LMI optimization approach. Optimal Control Applications and Methods **21** (2000), S. 243-256.

[7] BLACKWELL, C.C., YU, I.H.: The impact of the choice of gains on the robustness of control systems and observers. Proceedings of the American Control Conference (San Diego) 1990, S. 941-942.

[8] BURROWS, S.P., PATTON, R.J., Design of low-sensitivity modalized observers using left eigenstructure assignment. J. of Guidance, Control, and Dynamics **15** (1992), S. 779-782.

[9] RAMAR, K., GOURISHANKAR, V.: Optimal observers with specified eigenvalues. Int. J. Control **27** (1978), S. 239-244.

[10] GRIVA, G., EASTHAM, C., PROFUMO, J.F., VRANKA, P., High performance sensorless control of induction motor drives for industry applications. Proc. Power Conversion Conf., Part 2, S. 535-539.

[11] MARCHESONI, M., SEGARICH, P., SORESSI, E., Simple approach to flux and speed observation in induction motor drives. Proc. Industrial Electronics Conf. **1** (1994), S. 305-310.

[12] BEKKOUCHE, D., CHENAFA, M., MANSOURI, A., Nonlinear control of robot manipulators driven by induction motors. Proc. Int.Symp. Industrial Electronics, Part 2, (1998), S. 513-518.

[13] LOVATI, V., MARCHESONI, M., OBERTI, M., SEGARICH, P., Microcontroller-based sensorless stator flux-oriented asynchronous motor drive for traction application. IEEE Trans. Power Electronics **13** (1998), S. 777-784.

[14] MOYNIHAN, J.F., BOLOGNAM, S., KAVANAGH, R.C., EGAN, M.G., MURPHY, J.M.D., Single sensor current control of AC servodrives using digital signal processors. Proc. 5th European Conf. on Power Electronics and Applications 1993, S. 418-421.

[15] VOLCANJK, V., JEZERNIK, K., Induction motor control with PI-load estimator. Proc. 7th Mediterranean Electrotechnical Conf., Part 3 (1994), S. 1097-1100.

[16] HOFMANN, W., EL-BARBARI, S., Digital control of a four leg inverter for standalone photovoltaic system with unbalanced load. Proc. Industrial Electronic Conf., **1** (2000), S. 729-734.

[17] HIPPE, P., WURMTHALER, C.: Zustandsregelung. Springer Berlin. 1985.

[18] MAGNI, J.-F., Robust modal control with a toolbox for use with MATLAB, Kluwer Academic Publisher, New York, 2002.

[19] KORN, U., UND WILFERT, H.-H., Mehrgrößenregelungen. VEB Technik und Springer. 1982.

[20] O'REILLY, J., Observers for linear systems, Academic Press. London. 1983.

[21] OGATA, K., Modern Control Engineering, 4. Auflage. Prentice Hall, Upper Saddle River. 2002.

[22] DAVISON, E. J., The output control of linear time-invariant multivariable systems with unmeasurable arbitrary disturbances. IEEE-Transactions **AC-17** (1972), S. 621-630.

[23] DAVISON, E. J., A generalization of the output control of linear multivariable systems with unmeasurable arbitrary disturbances. IEEE-Transactions **AC-20** (1975), S. 621-630.

[24] HIPPE, P., WURMTHALER, C.: Ein Verfahren zum Entwurf von Zustandsreglern minimaler Ordnung mit Stör- und Führungsmodell. Regelungstechnik **29** (1981), S. 155-164.

[25] HIPPE, P., WURMTHALER, C.: Reduced-order observers for linear multivariable systems with disturbance and tracking signal accomodation. Int. J. Control **38** (1983), S. 831-842.

[26] MÜLLER, P.C., BREMER, H., UND BREINL,W., Tragregelsysteme mit Störgrößenkompensation für Magnetschwebefahrzeuge. Regelungstechnik **24** (1976), S. 257-265.

[27] MÜLLER, P.C., LÜCKEL, J.: Zur Theorie der Störgrößenaufschaltung in linearen Mehrgrößenregelsystemen. Regelungstechnik **25** (1977), S. 54-59.

[28] WEIHRICH, G.: Mehrgrößen-Zustandsregelung unter Einwirkung von Stör- und Führungssignalen. Regelungstechnik **25** (1977), S. 166-172 und S. 204-209.

[29] HIPPE, W., Ein einfaches Verfahren zum Entwurf von Zustandsreglern mit Störgrößenkompensation. Regelungstechnik 29 (1981), S. 91-96 und S.131-136

[30] WEINMANN, A., : Gradients of norms, traces and determinants for automatic control applications. Int. J. Automation Austria **9** (2001), S.36-50.

Ein Plädoyer für den Einsatz des Operatorenkalküls nach Mikusiński

Dietrich Dreyer
Einsteinufer 71B, D-10587 Berlin
Dietrich.Dreyer@alumni.TU-Berlin.DE

Zusammenfassung

Hier wird ein Plädoyer für den Einsatz des Operatorenkalküls nach Mikusiński gegeben, und zwar für einen Einsatz dieses Operatorenkalküls anstelle der Laplace-Transformation.

Hierzu wird zunächst eine kurze Bemerkung zur mathematischen Modellbildung gemacht. Es wird ausgeführt, daß analog dem ganz natürlichen Übergang von den ganzen zu den rationalen Zahlen bei Signalen eigentlich der Übergang von den reellen Funktionen zu den Faltungsquotienten oder Distributionen entsprechen müßte.

Dann wird an einem einfachen Beispiel die Historie des Operatorenkalküls dargestellt: Vom Operatorenkalkül nach Heaviside über die Laplace-Transformation zum Operatorenkalkül nach Mikusiński.

Danach erfolgt eine kurze Darstellung zu den mathematischen Grundlagen des Operatorenrechnung nach Mikusiński. Hierbei wird deutlich werden, daß diese Operatorenrechnung – bis auf die komplizierte Nomenklatur – nahezu trivial ist.

Danach erfolgt der Übergang zum Operatorenkalkül nach Mikusiński. Dieser Operatorenkalkül ist lediglich eine in der Schreibweise vereinfachte Operatorenrechnung, die aufgrund der verschiedenen erkennbaren möglichen Einbettungen der vorliegenden mathematischen Strukturen gerechtfertigt ist. Das Arbeiten mit diesem Operatorenkalkül wird hiernach an einem etwas ausführlicheren Beispiel demonstriert. Am sogenannten Differentiationssatz wird dann nochmals die verschiedene Interpretation der einzelnen Rechenschritte bei der Laplace-Transformation und dem Operatorenkalkül nach Mikusiński erläutert.

Abschließend wird die Diskussion einiger Gründe für (und gegen) den Einsatz des Operatorenkalküls nach Mikusiński gegeben. Hierbei sollte deutlich werden, daß (fast) alles für den Einsatz dieses Operatorenkalküls spricht.

1 Bemerkung zur mathematischen Modellierung

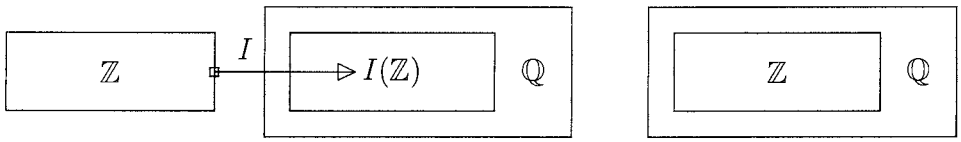
Es ist natürlich klar, daß sich jedes Meßergebnis mittels der ganzen Zahlen beschreiben läßt, so ist beispielsweise die an einem Widerstand gemessene Spannung jederzeit in ganzen Einheiten von Volt, Millivolt, Mikrovolt, oder sonst einer Einheit anzugeben. Wollte

man aber auch die Naturgesetze, beispielsweise für einen Widerstand R den Zusammenhang zwischen der Spannung U und dem Strom I mittels der ganzen Zahlen formulieren, so bekäme man sehr große Schwierigkeiten. Naturgesetze werden daher nicht mittels des Ringes der ganzen Zahlen $\mathbb{Z}$, sondern mittels des Körpers der rationale Zahlen $\mathbb{Q}$ (oder noch besser gleich mittels des Körpers $\mathbb{K}$, des Körpers $\mathbb{R}$ der reellen Zahlen $\mathbb{R}$ oder des Körpers $\mathbb{C}$ der komplexen Zahlen) formuliert, so daß zumindest die Quotienten zweier Größen stets gebildet werden können. Bei gegebenem Widerstand R und gemessener Spannung U kann somit der Strom I durch den Widerstand mittels des Gesetzes

$$I = \frac{U}{R} \tag{1}$$

innerhalb des gewählten Zahlenkörpers berechnet werden. Das Ergebnis der Rechnung, wird aber zumeist wieder – nach einer geeigneten Rundung – als ganze Zahl in einem Einheitensystem angegeben.

Dieses Vorgehen, die ganzzahligen Meßwerte in einen größeren Zahlenbereich „einzubetten“, in diesem größeren Zahlenbereich zu „rechnen“, und dann das Ergebnis unter Umständen wieder als ganzzahligen Wert zu interpretieren, ist so selbstverständlich, daß man es als ganz natürlich ansieht.



(a) Injektiver Homomorphismus I von $\mathbb{Z}$ in $\mathbb{Q}$ (b) $\mathbb{Z}$ in $\mathbb{Q}$ eingebettet

Abbildung 1: Einbettung des Ringes $\mathbb{Z}$ in den Körper $\mathbb{Q}$ (Kalkül des Zahlenrechnens)

Vom Standpunkt der Mathematik stellt man sich heute vor, daß diesem Vorgehen folgendes zugrunde liegt: Zunächst wurde zum Ring $\mathbb{Z}$ der ganzen Zahlen der sogenannte Quotientenkörper $\mathbb{Q}$ konstruiert, auf dem auch alle algebraischen Rechenoperationen erklärt werden konnten, und es wurde eine injektive Abbildung I vom Ring $\mathbb{Z}$ in den Körper $\mathbb{Q}$ angegeben, die mit allen algebraischen Rechenoperationen auf $\mathbb{Z}$ und $\mathbb{Q}$ verträglich war, ein sogenannter injektiver Homomorphismus. Vom algebraischen Standpunkt aus bestand dann kein Grund mehr dafür, den Ring $\mathbb{Z}$ und sein Bild $I(\mathbb{Z}) = \{I(z) \mid z \in \mathbb{Z}\}$, eine Teilmenge von $\mathbb{Q}$, auseinanderzuhalten, „es wurde $\mathbb{Z}$ mit seinem Bild $I(\mathbb{Z})$ in $\mathbb{Q}$ identifiziert“ oder „es wurde $\mathbb{Z}$ in $\mathbb{Q}$ eingebettet“, wie man zu sagen pflegt. Und demzufolge wird jetzt beispielsweise die ganze Zahl 3 und die rationale Zahl 3/1 oder $(3 \cdot 7)/7$ als „gleich“ angesehen. Die ganzzahligen Eingangsdaten werden also als im Körper $\mathbb{Q}$ liegend angesehen, in diesem Körper $\mathbb{Q}$ wird gerechnet, und die Rechenergebnisse können dann, unter Umständen nach einem Rundungsprozeß, wieder als ganzzahlige Ausgangsdaten vorliegen. Und für dieses Vorgehen ist meiner Meinung nach die Inklusion $\mathbb{Z} \subset \mathbb{Q}$ ganz entscheidend, denn bleibt man auf der Stufe (a) der Abbildung 1 stehen, so wird das ganze Arbeiten unerträglich kompliziert.

Und besteht das Meßergebnis eines Signal, beispielsweise für den Spannungsverlauf an einem Kondensator, aus einer Folge von Spannungswerten für die einzelnen Zeitpunkte der Messung, so könnte dieses Meßergebnis völlig ausreichend als Funktion in $\mathbb{Z} \times \mathbb{Z}$

beschrieben werden. In Anbetracht der oben angedeuteten Schwierigkeiten beim Arbeiten mit ganzen Zahlen werden die Meßergebnis eines Signals aber gleich als von einer Funktion $t \mapsto U(t)$ in $\mathbb{R} \times \mathbb{K}$ herrührend gedeutet. Beim Formulieren der Naturgesetze für Signale, beispielsweise für die Beziehung

$$I(t) = C \cdot \frac{dU(t)}{dt} \tag{2}$$

an einem Kondensator, gibt es aber trotzdem noch jede Menge Schwierigkeiten mit der Mathematik. Wie steht es beispielsweise um die Differenzierbarkeit des Signals $t \mapsto U(t)$.

Warum geht man auch hier nicht – wie bei den Zahlen – zu einem größeren Bereich über in dem alle mathematischen Operationen wie Differentiation usw. ohne Probleme stets ausführbar sind, und um dann das in diesem größeren Bereich errechnete Ergebnis unter Umständen wieder – nach einer geeigneten „Rundung" – als Funktion in $\mathbb{R} \times \mathbb{K}$ zu deuten?

Für diese Bereiche, in die die reellen bzw. die komplexwertigen Funktionen der Signale „eingebettet" werden können, und in denen alle benötigten mathematischen Operationen (wie differenzieren usw.) stets ausgeführt werden können, stehen heute im wesentlichen zwei Bereiche zur Verfügung:

- Zum einen der Bereich der Distributionen nach LAURENT SCHWARTZ. Für das Arbeiten in diesem Bereich sind gute Kenntnisse der Funktionalanalysis nötig.
- Zum anderen der Bereich der Faltungsquotienten nach JAN MIKUSIŃSKI. Für das Arbeiten in diesem Bereich sind demgegenüber recht bescheidene Kenntnisse aus der Algebra notwendig. Dies soll im folgenden verdeutlicht werden. (Wird allerdings über den Rahmen der folgenden Darstellung hinausgegangen, so sind im Zusammenhang mit Grenzwertfragen natürlich auch hier wieder einige Kenntnisse aus der Funktionalanalysis einzubringen.)

Es sei aber an dieser Stelle schon angemerkt, daß die aufgrund der Laplace-Transformation bestehende Abbildung von Funktionen aus $\mathbb{R} \times \mathbb{K}$ in die auf einer Halbebene von $\mathbb{C}$ holomorphen Funktionen nicht so einfach als eine Einbettung angesehen werden kann (bisher jedenfalls). Und so fehlt der Laplace-Transformation die Eigenschaft, die in der obigen Diskussion als für den Kalkül des Zahlenrechnens so entscheidend angesehene Eigenschaft $\mathbb{Z} \subset \mathbb{Q}$.

2 Historische Entwicklung anhand eines Beispiels

2.1 Das Beispiel

Betrachtet werde die Anfangswertaufgabe

$$\dot{x}(t) = a\,x(t) + b\,u(t) \quad \text{für} \quad t \geq 0 \quad \text{und} \quad x(0) = x_0, \tag{3}$$

wobei a, b, $t \mapsto u(t)$ auf $[0, \infty[$ und x_0 als gegeben und $t \mapsto x(t)$ auf $[0, \infty[$ als gesucht anzusehen sind.

Zunächst soll die einfachere Anfangswertaufgabe mit $a = 1$, $b = 1$, $u(t) = 1$ für alle $t \geq 0$ und $x_0 = 0$ betrachtet werden, also

$$\dot{x}(t) = x(t) + 1 \quad \text{für} \quad t \geq 0 \quad \text{und} \quad x(0) = 0. \tag{4}$$

2.2 Der Operatorenkalkül nach Heaviside

Nach HEAVISIDE schreibt man mit dem „Differentiationsoperator" p die Anfangswertaufgabe (4) zu

$$p\,x(t) = x(t) + 1, \quad x(0) = 0 \tag{5}$$

und

$$(p-1)x(t) = 1, \quad x(0) = 0 \qquad \text{bzw.} \qquad x(t) = \frac{1}{p-1}, \quad x(0) = 0. \tag{6}$$

Nun war für HEAVISIDE klar, daß

$$\frac{1}{p-1} = \frac{1}{p} \cdot \frac{1}{1-p^{-1}} = p^{-1}\left(1 + p^{-1} + p^{-2} + \cdots\right) = p^{-1} + p^{-2} + p^{-3} + \cdots \tag{7}$$

gilt, und wenn p der „Differentiationsoperator" ist, dann ist p^{-1} der „Integrationsoperator", so daß weiter gilt

$$p^{-1} = p^{-1}\,1 = \int 1\,dt = t = \frac{t}{1!}, \tag{8}$$

$$p^{-2} = p^{-1}\,p^{-1} = p^{-1}\,\frac{t}{1!} = \int \frac{t}{1!}\,dt = \frac{t^2}{1! \cdot 2} = \frac{t^2}{2!}, \tag{9}$$

$$p^{-3} = p^{-1}\,p^{-2} = p^{-1}\,\frac{t^2}{2!} = \int \frac{t^2}{2!}\,dt = \frac{t^3}{2! \cdot 3} = \frac{t^3}{3!}, \quad \ldots \tag{10}$$

Damit folgt für $t \geq 0$:

$$x(t) = \frac{1}{p-1} = \frac{t}{1!} + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots = \left(1 + \frac{t}{1!} + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots\right) - 1 = \mathrm{e}^t - 1. \tag{11}$$

Und es gilt tatsächlich $x(0) = \mathrm{e}^0 - 1 = 0$ sowie $\dot{x}(t) = \mathrm{e}^t = (\mathrm{e}^t - 1) + 1 = x(t) + 1$ für $t \geq 0$, also ist $t \mapsto x(t) = \mathrm{e}^t$-1 tatsächlich eine Lösung der Anfangswertaufgabe (4). Und für HEAVISIDE war diese Bestätigung als Rechtfertigung für seine „Operatorenrechnung" ausreichend. Nicht aber für die Mathematiker seiner Zeit, insbesondere die aus Cambridge, die diese Art der „Operatorenrechnung" zumeist einfach ignoriert haben (daneben jedoch dafür sorgten, daß HEAVISIDE nicht mehr so einfach Publizieren konnte), obwohl doch bereits LEIBNIZ, LAGRANGE und LAPLACE mit derartigen „Operatoren" gearbeitet haben. Und heute glaubt man auch zu wissen, daß der Autodidakt HEAVISIDE die betreffenden Arbeiten von LAGRANGE und LAPLACE gelesen haben muß und von dort auch die Namen „Operator" und „Operatorenrechnung" bzw. „Operatorenkalkül" übernommen hat.

2.3 Die Laplace-Transformation

In den zwanziger und dreißiger Jahren des letzten Jahrhunderts wurde dann die Laplace-Transformation zu einen schlagkräftigen Kalkül ausgebaut, so daß man glaubte, nichts anderes mehr zu benötigen. Denn fast alles was von HEAVISIDE mit seiner seltsamen „Operatorenrechnung" gelöst werden konnte, wurde jetzt mit einer relativ ordentlichen Mathematik auch erhalten (siehe z.B. bei DOETSCH in [2]).

Um die obige Anfangswertaufgabe (4)

$$\dot{x}(t) = x(t) + 1 \quad \text{für} \quad t \geq 0 \quad \text{und} \quad x(0) = 0. \tag{12}$$

zu lösen, mußte lediglich vorausgesetzt werden, daß die Lösung $t \mapsto x(t)$ existiert und nicht zu schnell wächst (was aus der Theorie der Differentialglreichungen bekannt ist), dann konnte die Differentialgleichung mit einem Faktor e^{-st} multipliziert und von 0 bis $+\infty$ integriert werden:

$$\int_0^\infty \mathrm{e}^{-st}\dot{x}(t)\,dt = \int_0^\infty \mathrm{e}^{-st}(x(t)+1)\,dt = \int_0^\infty \mathrm{e}^{-st}x(t)\,dt + \int_0^\infty \mathrm{e}^{-st}\,dt. \tag{13}$$

Hierbei ist s eine komplexe Veränderliche mit $\mathrm{Re}(s)$ hinreichend groß, so daß die hier auftretenden drei Integrale konvergieren. Speziell ist dann

$$\int_0^\infty \mathrm{e}^{-st}\,dt = -\frac{1}{s}\,\mathrm{e}^{-st}\Big|_{t=0}^{\infty} = \frac{1}{s}. \tag{14}$$

Und mit der Abkürzung $\mathcal{L}(f)(s) := F(s) := \int_0^\infty \mathrm{e}^{-st}f(t)\,dt$ (auch als $f(t) \circ\!-\!\bullet\, F(s)$ geschrieben) gilt somit $1 \circ\!-\!\bullet\, \mathcal{L}(\sigma)(s) = \frac{1}{s}$. Und für jede auf $[0,\infty[$ stetig differenzierbare Funktion g folgt mittels partieller Integration und der Voraussetzung $\lim_{t\to\infty} \mathrm{e}^{-st}g(t) = 0$ weiter

$$\begin{aligned}\mathcal{L}(\dot{g})(s) &= \int_0^\infty \mathrm{e}^{-st}\dot{g}(t)\,dt = \mathrm{e}^{-st}g(t)\Big|_{t=0}^{\infty} + s\int_0^\infty \mathrm{e}^{-st}g(t)\,dt \\ &= s\int_0^\infty \mathrm{e}^{-st}g(t)\,dt - g(0) = s\cdot\mathcal{L}(g)(s) - g(0).\end{aligned} \tag{15}$$

Also ist $\mathcal{L}(\dot{x})(s) = s\cdot\mathcal{L}(x)(s) - x(0) = s\cdot X(s) - x(0)$ und wegen $x(0) = 0$ und $\mathcal{L}(\sigma)(s) = \frac{1}{s}$ folgt daher

$$s\cdot X(s) = X(s) + \frac{1}{s} \qquad \text{sowie} \qquad (s-1)\cdot X(s) = \frac{1}{s}, \tag{16}$$

also mittels Partialbruchentwicklung:

$$X(s) = \frac{1}{(s-1)s} = \frac{1}{s-1} - \frac{1}{s}. \tag{17}$$

Von der Laplace-Transformation $f \mapsto \mathcal{L}(f) = F$ bzw. $f(t) \circ\!-\!\bullet\, F(s)$ ist jetzt bekannt, daß sie in gewisser Weise eineindeutig ist, d.h. genau nur dann wenn sich zwei „Originalfunktionen" f_1 und f_2 nur unwesentlichen unterscheiden, dann stimmen die beiden

„Bildfunktionen“ F_1 und F_2 überein. Nun wurde oben bereits berechnet: $1 \circ\!-\!\bullet\, \frac{1}{s}$. Und analog wird erhalten

$$\mathrm{e}^t \circ\!-\!\bullet \int_0^\infty \mathrm{e}^{-st}\mathrm{e}^t\,dt = \int_0^\infty \mathrm{e}^{-(s-1)t}\,dt = -\frac{1}{s-1}\mathrm{e}^{-(s-1)t}\Big|_{t=0}^{\infty} = \frac{1}{s-1}. \tag{18}$$

Also ist $\mathrm{e}^t - 1 \circ\!-\!\bullet \frac{1}{s-1} - \frac{1}{s}$, und damit muß im wesentlichen für alle $t \geq 0$ gelten:

$$x(t) = \mathrm{e}^t - 1. \tag{19}$$

Und da $t \mapsto x(t)$ die stetige Lösung einer Anfangswertaufgabe ist, ist diese Gleichung für alle $t \geq 0$ erfüllt, und nicht nur im wesentlichen.

Soweit so gut. In den dreißiger Jahren des letzten Jahrhunderts wurde dann das Arbeiten mit impulsförmigen Signalen zu einem festen Bestandteil der Anwendungen, vor allem im Bereich der Physik. Die Mathematik und damit auch die Laplace-Transformation konnten mit diesen impulsförmigen Signalen zunächst nichts anfangen, also wurden sie – ebenso wie früher die Operatorenrechnung nach HEAVISIDE – einfach ignoriert. Dann in der zweiten Hälfte der vierziger Jahre des letzten Jahrhunderts kam der Franzose LAURENT SCHWARTZ mit seinen Distributionen und machte die impulsförmigen Signalen zu einem festen Bestandteil der Mathematik. Jetzt konnte aber auch alles, wofür sich die Laplace-Transformation inzwischen für zuständig hielt, innerhalb der Theorie der Distributionen mit der guten alten Fourier-Transformation bearbeitet werden. Aber trotzdem versuchte man in den fünfziger Jahren des letzten Jahrhundert die Laplace-Transformation durch eine Art von „Klapperstorch-Theorie“ für das Arbeiten mit impulsförmigen Signalen zu erweitern. Und wenn man sich die seit dieser Zeit erschienene Literatur der Regelungstechnik ansieht, so stellt man fest, daß eigentlich durchweg mit dieser Art von „erweiterter“ Laplace-Transformation gearbeitet wird.

2.4 Der Operatorenkalkül nach Mikusiński

Dann in der zweiten Hälfte der vierziger und der ersten Hälfte der fünfziger Jahre des letzten Jahrhunderts kam der Pole JAN MIKUSIŃSKI mit seinen Faltungsquotienten, so daß das Arbeiten mit impulsförmigen Signalen nichts besonderes mehr war, und die Laplace-Transformation einfach überflüssig wurde. Um beispielsweise die obige Anfangswertaufgabe (4)

$$\dot{x}(t) = x(t) + 1 \quad \text{für} \quad t \geq 0 \quad \text{und} \quad x(0) = 0. \tag{20}$$

zu lösen, werden die Signale $x\colon t \mapsto x(t)$, $\dot{x}\colon t \mapsto \dot{x}(t)$, $\sigma\colon t \mapsto 1$ als Elemente einer nullteilerfreien Faltungsalgebra interpretiert, die dann in natürlicher Weise in die Divisionsalgebra ihrer Faltungsquotienten eingebettet werden kann. In dieser Divisionsalgebra (die gleichzeitig ein Körper ist), kann nun nämlich wie in einem Zahlenkörper gerechnet werden. So gibt es zum Beispiel in dieser Divisionsalgebra zu jedem von Null verschiedenen Element ein Inverses. Und betrachten wir beispielsweise die Sprungfunktion

$$\sigma\colon \mathbb{R} \to \mathbb{R}, \quad t \mapsto \sigma(t) = \begin{cases} 0 & \text{für} \quad t < 0 \\ 1 & \text{für} \quad t \geq 0, \end{cases} \tag{21}$$

so gilt für jede auf $[0,\infty[$ stetig differenzierbare Funktion g mit $g(t)=0$ für $t<0$ für die Faltung

$$(\sigma\,\dot{g})(t)=\int_0^t \sigma(t-\tau)\dot{g}(\tau)\,d\tau=\int_0^t \dot{g}(\tau)\,d\tau=g(t)-g(0)=g(t)-g(0)\sigma(t), \tag{22}$$

also ist $\sigma\,\dot{g}=g-g(0)\sigma$, und da σ nicht das Nullelement ist, existiert $\sigma^{-1}=:p$ mit $p\,\sigma=1$, und es ist

$$\dot{g}=p\,g-g(0). \tag{23}$$

Die obige Anfangswertaufgabe (4) wird damit in sogenannter „Operatorform“ in der Divisionsalgebra der Faltungsquotienten zu

$$p\,x-x(0)=\dot{x}=x+\sigma, \qquad \text{sowie wegen } x(0)=0 \text{ zu} \qquad (p-1)\,x=\sigma, \tag{24}$$

und damit schließlich zu

$$x=(p-1)^{-1}\sigma. \tag{25}$$

Wegen $\dot{g}=p\,g-g(0)$ für jede auf $[0,\infty[$ stetig differenzierbare Funktion g, gilt für die Funktion

$$\langle \mathrm{e}^t\rangle\colon \mathbb{R}\to\mathbb{R}, \quad t\mapsto \begin{cases} 0 & \text{für} \quad t<0 \\ \mathrm{e}^t & \text{für} \quad t\geq 0, \end{cases} \tag{26}$$

ganz einfach

$$\langle \mathrm{e}^t\rangle=p\,\langle \mathrm{e}^t\rangle-1 \tag{27}$$

und

$$(p-1)\,\langle \mathrm{e}^t\rangle=1 \qquad \text{sowie} \qquad \langle \mathrm{e}^t\rangle=(p-1)^{-1}. \tag{28}$$

Hiermit folgt sodann

$$x=(p-1)^{-1}\sigma=\langle \mathrm{e}^t\rangle\,\sigma=\sigma\,\langle \mathrm{e}^t\rangle. \tag{29}$$

Also gilt für alle $t\geq 0$ wieder

$$x(t)=(\sigma\,\langle \mathrm{e}^t\rangle)(t)=\int_0^t \sigma(t-\tau)\mathrm{e}^\tau\,d\tau=\int_0^t \mathrm{e}^\tau\,d\tau=\mathrm{e}^\tau\big|_{\tau=0}^t=\mathrm{e}^t-1. \tag{30}$$

Die Rechenschritte sind hier eigentlich die gleichen wie im Verfahren nach HEAVISIDE (oder bei der Laplace-Transformation), nur die Begründungen für die einzelnen Rechenschritte sind hier anders. Und wegen dieser neuen Begründungen nun ein kleiner Ausflug in die abstrakte Algebra.

3 Grundlagen der Operatorenrechnung

3.1 Einiges aus der abstrakten Algebra

Definition 1. Sei M eine nichtleere Menge eventuell versehen mit drei Rechenoperationen gemäß

$$+\colon M \times M \to M, \quad (a,b) \mapsto a+b, \quad (\textit{Addition}, \text{ auch als } \oplus \text{ geschrieben}) \tag{31}$$
$$*\colon M \times M \to M, \quad (a,b) \mapsto a*b, \quad (\textit{Multipl. (Faltung)}, \text{ auch als } \circledast \text{ geschr.}) \tag{32}$$
$$\cdot\colon \mathbb{K} \times M \to M, \quad (\alpha,a) \mapsto \alpha \cdot a, \quad (\textit{Vervielfachung}, \text{ auch als } \odot \text{ geschrieben}) \tag{33}$$

mit den folgenden Eigenschaften:

(a1) Für alle a, b, $c \in M$ ist $(a+b)+c = a+(b+c)$. (*Assoziativität* der Addition)

(a2) Für alle a, $b \in M$ ist $a+b = b+a$. (*Kommutativität* der Addition)

(a3) Es gibt ein $o \in M$, so daß für alle $a \in M$ stets $a+o = a$ ist. (Exist. eines *Nullele.*)

(a4) Zu jedem $a \in M$ gibt es ein $b \in M$ mit $a+b = o$. (Exist. eines *additiven Inversen*)

(r1) Für alle a, b, $c \in M$ ist $(a*b)*c = a*(b*c)$. (*Assoziativität* der Multiplikation)

(r2) Für alle a, $b \in M$ ist $a*b = b*c$. (*Kommutativität* der Multiplikation)

(r3) Für alle a, b, $c \in M$ ist $(a+b)*c = a*c+b*c$. (*Distributivität* der Add. und Mul.)

(r4) Es gibt ein $e \in M$, so daß für alle $a \in M$ stets $a*e = a$ ist. (Exist. eines *Einselem.*)

(r5) Zu jedem $a \in M$ mit $a \neq o$ gibt es ein $b \in M$ mit $a*b = e$. (Exist. eines *mult. Inv.*)

(s1) Für alle α, $\beta \in \mathbb{K}$ und alle $a \in M$ ist $\alpha \cdot (\beta \cdot a) = (\alpha\beta) \cdot a$.

(s2) Für alle $\alpha \in \mathbb{K}$ und alle a, $b \in M$ ist $\alpha \cdot (a+b) = \alpha \cdot a + \alpha \cdot b$.

(s3) Für alle $\alpha, \beta \in \mathbb{K}$ und alle $a \in M$ ist $(\alpha+\beta) \cdot a = \alpha \cdot a + \beta \cdot a$.

(s4) Für alle $a \in M$ und $1 \in \mathbb{K}$ ist $1 \cdot a = a$.

(s5) Für alle $\alpha \in \mathbb{K}$ und alle a, $b \in M$ ist $\alpha \cdot (a*b) = (\alpha \cdot a)*b = a*(\alpha \cdot b)$.

Nun wird genannt:

- *abelsche Gruppe* $(M,+,o)$, das ist M mit der Addition $+$ und dem Nullelement o mit den Eigenschaften (a1) bis (a4),
- *kommutativer Ring* $(M,+,o,*)$, das ist die abelsche Gruppe $(M,+,o)$ mit der Multiplikation $*$ mit den Eigenschaften (r1) bis (r3),
- *kommutativer Ring mit Einselement* $(M,+,o,*,e)$, das ist die abelsche Gruppe $(M,+,o)$ mit der Multiplikation $*$ und dem Einselement e mit den Eigenschaften (r1) bis (r4),

- *Körper* $(M, +, o, *, e)$, das ist die abelsche Gruppe $(M, +, o)$ mit der Multiplikation $*$ und dem Einselement e mit den Eigenschaften (r1) bis (r5),
- $\mathbb{K}$-*Vektorraum* $(M, +, o, \mathbb{K}, \cdot)$, das ist die abelsche Gruppe $(M, +, o)$ mit der Vervielfachung $\cdot$ mit Skalaren aus $\mathbb{K}$ mit den Eigenschaften (s1) bis (s4),
- *kommutative* $\mathbb{K}$-*Algebra* $(M, +, o, *, \mathbb{K}, \cdot)$, das ist der kommutative Ring $(M, +, o, *)$ mit der Vervielfachung $\cdot$ mit Skalaren aus $\mathbb{K}$ mit den Eigenschaften (s1) bis (s5),
- *kommutative* $\mathbb{K}$-*Algebra mit Einselement* $(M, +, o, *, e, \mathbb{K}, \cdot)$, das ist der kommutative Ring $(M, +, o, *, e)$ mit Einselement mit der Vervielfachung $\cdot$ mit Skalaren aus $\mathbb{K}$ mit den Eigenschaften (s1) bis (s5),
- $\mathbb{K}$-*Divisionsalgebra* $(M, +, o, *, e, \mathbb{K}, \cdot)$, das ist der Körper $(M, +, o, *, e)$ mit der Vervielfachung $\cdot$ mit Skalaren aus $\mathbb{K}$ mit den Eigenschaften (s1) bis (s5).

Bemerkung 2. Ist $(A, +, o, *, \mathbb{K}, \cdot)$ eine kommutative Algebra, so gilt: Das zu $a \in A$ existierende Element $b \in A$ mit $a + b = o$ ist eindeutig bestimmt und wird mit $-a$ bezeichnet, und es wird für alle $a, b \in A$ gesetzt: $a - b := a + (-b)$. Damit ist dann auch $-a = (-1) \cdot a$ für alle $a \in A$ und $a * (b - c) = a * b - a * c$ für alle $a, b, c \in A$.
Ist $(A, +, o, *, e, \mathbb{K}, \cdot)$ eine Divisionsalgebra, so gilt: Das zu $a \in A \backslash \{o\}$ existierende Element $b \in A$ mit $a * b = e$ ist eindeutig bestimmt und wird mit a^{-1} bezeichnet.

Definition 3. Seien $(A_1, +, o_1, *, \mathbb{K}, \cdot)$ und $(A_2, \oplus, o_2, \circledast, \mathbb{K}, \odot)$ zwei (kommutative) Algebren. Eine Abbildung

$$\Phi\colon A_1 \to A_2, \quad a \mapsto \Phi(a) \tag{34}$$

mit den Eigenschaften

(h1) für alle $a, b \in A_1$ ist $\Phi(a + b) = \Phi(a) \oplus \Phi(b)$, *(Additivität)*

(h2) für alle $a, b \in A_1$ ist $\Phi(a * b) = \Phi(a) \circledast \Phi(b)$, *(Multiplikativität)*

(h3) für alle $\alpha \in \mathbb{K}$ und alle $a \in A_1$ ist $\Phi(\alpha \cdot a) = \alpha \odot \Phi(a)$, *(Homogenität)*

wird ein *(Algebra-)Homomorphismus* genannt.
Ist der (Algebra-)Homomorphismus $\Phi\colon A_1 \to A_2$ bijektiv, so werden die beiden (kommutativen) Algebren $(A_1, +, o_1, *, \mathbb{K}, \cdot)$ und $(A_2, \oplus, o_2, \circledast, \mathbb{K}, \odot)$ *isomorph* genannt.
Ist der (Algebra-)Homomorphismus $\Phi\colon A_1 \to A_2$ injektiv, so wird die (kommutative) Algebra $(A_1, +, o_1, *, \mathbb{K}, \cdot)$ in die (kommutative) Algebra $(A_2, \oplus, o_2, \circledast, \mathbb{K}, \odot)$ *einbettbar* genannt. In diesem Fall sind die beiden (kommutativen) Algebren $(A_1, +, o_1, *, \mathbb{K}, \cdot)$ und $(\Phi(A_1), \oplus, o_2, \circledast, \mathbb{K}, \odot)$ isomorph.

Bemerkung 4. Sind $(A_1, +, o_1, *, \mathbb{K}, \cdot)$ und $(A_2, \oplus, o_2, \circledast, \mathbb{K}, \odot)$ zwei Algebren, und $\Phi\colon A_1 \to A_2$ ein Homomorphismus, so gilt für alle $a, b \in A_1$: $\Phi(a - b) = \Phi(a) \ominus \Phi(b)$.

Aussage 5 (Quotientenalgebra).
Ist eine kommutative Algebra $(A, +, o, *, \mathbb{K}, \cdot)$ nullteilerfrei, d.h. ist für alle a, $b \in A$ mit $a \neq o$ und $b \neq o$ auch stets $a * b \neq o$, und existiert ein Element $s \in A$ mit $s \neq o$, so gibt es eine Divisionsalgebra $(D, \oplus, 0_D, \circledast, 1_D, \mathbb{K}, \odot)$, in die die Algebra $(A, +, o, *, \mathbb{K}, \cdot)$ einbettbar ist.

Eine möglicher Divisionsalgebra dieser Art ist die Quotientenalgebra zur kommutativen Algebra $(A, +, o, *, \mathbb{K}, \cdot)$, die wie folgt erhalten werden kann: Für alle $a \in A$ und $b \in A \setminus \{o\}$ sei

$$a/b := \{(c, d) \in A \times A \setminus \{o\} \mid a * d = b * c\} \tag{35}$$

und damit dann

$$D := \{a/b \mid a \in A,\ b \in A \setminus \{o\}\} \tag{36}$$

sowie $0_D := o/s$, $1_D := s/s$ und

$$\oplus\colon D \times D \to D, \quad (a/b, c/d) \mapsto a/b \oplus c/d := (a * d + c * b)/(b * d), \tag{37}$$
$$\circledast\colon D \times D \to D, \quad (a/b, c/d) \mapsto a/b \circledast c/d := (a * c)/(b * d), \tag{38}$$
$$\odot\colon \mathbb{K} \times D \to D, \quad (\alpha, a/b) \mapsto \alpha \odot a/b := (\alpha \cdot a)/b. \tag{39}$$

Hierbei sind die beiden Abbildungen

$$\Phi\colon A \to D, \quad a \mapsto \Phi(a) := (a * s)/s \tag{40}$$
$$\Psi\colon \mathbb{K} \to D, \quad \alpha \mapsto \Psi(\alpha) := \alpha \odot 1_D \tag{41}$$

injektive (Algebra-)Homomorphismen.

Diese Aussagen werden in vielen Lehrbüchern der Algebra bewiesen.

Bemerkung 6. Ist $(D, \oplus, 0_D, \circledast, 1_D, \mathbb{K}, \odot)$ eine Divisionsalgebra, so gilt:
Das zu $a/b \in D$ existierende Element $c/d \in D$ mit $a/b \oplus c/d = 0_D$ ist eindeutig bestimmt und wird mit $\ominus a/b$ bezeichnet, und es wird für alle a/b, $c/d \in D$ gesetzt: $a/b \ominus c/d := a/b \oplus (\ominus c/d)$.
Das zu $a/b \in D \setminus \{0_D\}$ existierende Element $c/d \in D$ mit $a/b \circledast c/d = 1_D$ ist eindeutig bestimmt und wird mit $(a/b)^{-1}$ bezeichnet. Ansonsten gelten die üblichen Regeln für das Rechnen mit Potenzen.

3.2 Die Operatorenrechnung nach Mikusiński

Definition 7. Die Gesamtheit derjenigen meßbaren Funktionen $\mathbb{R} \to \mathbb{K}$, deren Träger in einem Intervall $[a, \infty[$ mit $a \in \mathbb{R}$ liegt und die auf $[a, \infty[$ lokal im wesentlichen beschränkt sind, werde mit $\mathscr{U}$ bezeichnet.

Definition 8. Die Gesamtheit derjenigen Funktionen aus $\mathscr{U}$, deren Träger im Intervall $[0, \infty[$ liegt und die auf $[0, \infty[$ stetig sind, werde mit $\mathscr{R}$ bezeichnet.

Bemerkung 9. Oft werden wir in die Situation kommen, daß eine Funktion aus $\mathscr{R}$ über eine Formel für die Funktionswerte gegeben ist. Ein wichtiges Beispiel wird für ein $s \in \mathbb{K}$ die Funktion

$$\mathbb{R} \to \mathbb{K}, \quad t \mapsto \begin{cases} 0 & \text{für} \quad t < 0, \\ \mathrm{e}^{st} & \text{für} \quad t \geq 0 \end{cases} \tag{42}$$

sein. Für diese Funktion aus $\mathscr{R}$ soll dann das Symbol $\langle \mathrm{e}^{st} \rangle$ verwendet werden (so wie oben bereits einmal geschehen), sofern aus dem Zusammenhang heraus klar ist, daß t und nicht s das Argument ist, und die Werte dieser Funktion für $t < 0$ gleich Null sind. Ein weiteres wichtiges Beispiel wird für ein $\alpha \in \mathbb{K}$ die „konstante Funktion"

$$\mathbb{R} \to \mathbb{K}, \quad t \mapsto \begin{cases} 0 & \text{für} \quad t < 0, \\ \alpha & \text{für} \quad t \geq 0 \end{cases} \tag{43}$$

sein, für die dann das Symbol $\langle \alpha \rangle$ verwendet werden soll. Also beispielsweise $\langle 0 \rangle$ für die Nullfunktion und $\langle 1 \rangle$ für die Sprungfunktion

$$\sigma \colon \mathbb{R} \to \mathbb{K}, \quad t \mapsto \sigma(t) = \begin{cases} 0 & \text{für} \quad t < 0, \\ 1 & \text{für} \quad t \geq 0. \end{cases} \tag{44}$$

Aussage 10 (Faltungsalgebra). *Mit den drei Operationen $+$, $*$ und $\cdot$ gemäß*

$$+ \colon \mathscr{R} \times \mathscr{R} \to \mathscr{R}, \ (f,g) \mapsto f+g, \ (f+g)(t) := f(t) + g(t), \qquad t \in \mathbb{R}, \tag{45}$$

$$* \colon \mathscr{R} \times \mathscr{R} \to \mathscr{R}, \ (f,g) \mapsto f*g, \ (f*g)(t) := \int_0^t f(t-\tau) g(\tau)\, d\tau, \ t \in \mathbb{R}, \tag{46}$$

$$\cdot \colon \mathbb{K} \times \mathscr{R} \to \mathscr{R}, \ (\alpha, f) \mapsto \alpha \cdot f, \ (\alpha \cdot f)(t) := \alpha \cdot f(t), \qquad t \in \mathbb{R}, \tag{47}$$

ist $(\mathscr{R}, +, \langle 0 \rangle, *, \mathbb{K}, \cdot)$ eine kommutative Algebra mit dem Nullelement $\langle 0 \rangle$ und ohne Einselement. Diese Algebra, die einfach die Faltungsalgebra *$\mathscr{R}$ genannt werden soll, ist nullteilerfrei, d.h. für alle f, $g \in \mathscr{R}$ mit $f \neq \langle 0 \rangle$ und $g \neq \langle 0 \rangle$ ist auch $f * g \neq \langle 0 \rangle$, und es ist $\sigma \in \mathscr{R}$ mit $\sigma \neq \langle 0 \rangle$.*

Bemerkung 11. Hier ist nur die Behauptung der Nullteilerfreiheit nicht trivial. Diese Nullteilerfreiheit der Faltungsalgebra $\mathscr{R}$ wurde als erstes in den zwanziger Jahren des letzten Jahrhunderts von TITCHMARSH angegeben und bewiesen. Einfache Versionen dieses Beweises können bei MIKUSIŃSKI in [3] und bei ERDÉLYI in [5] nachgelesen werden.

Die Nullteilerfreiheit der Faltungsalgebra $\mathscr{R}$ liefert nun den Zugang zum Operatorenkalkül nach Mikusiński. Denn wie oben bereits ausgeführt, kann eine nullteilerfreie kommutative Algebra in die Divisionsalgebra ihrer Quotienten eingebettet werden. In unserem Fall heißt dies, daß die Faltungsalgebra $\mathscr{R}$ in die Divisionsalgebra der sogenannten „Faltungsquotienten" eingebettet werden kann. Und diese sogenannten „Divisionsalgebra der Faltungsquotienten", ist dann gerade der Arbeitsbereich des Operatorenkalküls nach Mikusiński.

Aussage 12 (Divisionsalgebra der Faltungsquotienten).
Für $f \in \mathscr{R}$ und $g \in \mathscr{R} \setminus \{\langle 0\rangle\}$ sei zunächst

$$f/g := \{(h,k) \in \mathscr{R} \times \mathscr{R} \setminus \{\langle 0\rangle\} \mid f * k = h * g\} \tag{48}$$

sowie

$$\mathfrak{K} := \{f/g \mid f \in \mathscr{R},\ g \in \mathscr{R} \setminus \{\langle 0\rangle\}\} \tag{49}$$

gesetzt. Mit den drei Operationen $\oplus$, $\circledast$ und $\odot$ gemäß

$$\oplus\colon \mathfrak{K} \times \mathfrak{K} \to \mathfrak{K}, \quad (f/g, h/k) \mapsto f/g \oplus h/k := (f * k + h * g)/(g * k), \tag{50}$$
$$\circledast\colon \mathfrak{K} \times \mathfrak{K} \to \mathfrak{K}, \quad (f/g, h/k) \mapsto f/g \circledast h/h := (f * h)/(g * k), \tag{51}$$
$$\odot\colon \mathbb{K} \times \mathfrak{K} \to \mathfrak{K}, \quad (\alpha, f/g) \mapsto \alpha \odot f/g := (\alpha \cdot f)/g, \tag{52}$$

ist dann $(\mathfrak{K}, \oplus, 0_{\mathfrak{K}}, \circledast, 1_{\mathfrak{K}}, \mathbb{K}, \odot)$ eine Divisionsalgebra mit dem Nullelement $0_{\mathfrak{K}} := \langle 0\rangle/\sigma$ und dem Einselement $1_{\mathfrak{K}} := \sigma/\sigma$. Diese Divisionsalgebra soll einfach die Divisionsalgebra $\mathfrak{K}$ der Faltungsquotienten *genannt werden, und die Elemente aus $\mathfrak{K}$* Operatoren, Mikusiński-Operatoren *oder* Faltungsquotienten.
Die beiden Abbildungen

$$\Phi\colon \mathscr{R} \to \mathfrak{K}, \quad f \mapsto \Phi(f) := (f * \sigma)/\sigma \tag{53}$$
$$\Psi\colon \mathbb{K} \to \mathfrak{K}, \quad \alpha \mapsto \Psi(\alpha) := \alpha \cdot 1_{\mathfrak{K}} \tag{54}$$

sind injektive (Algebra-)Homomorphismen von der Faltungsalgebra $\mathscr{R}$ bzw. dem Zahlenkörper $\mathbb{K}$ in die Divisionsalgebra $\mathfrak{K}$ der Faltungsquotienten. Der Homomorphismus $\Phi\colon \mathscr{R} \to \mathfrak{K}$ werde der erste Homomorphismus der Operatorenrechnung *genannt und der Homomorphismus $\Psi\colon \mathbb{K} \to \mathfrak{K}$ der* zweite Homomorphismus der Operatorenrechnung.

Bei MIKUSIŃSKI in [3] ist die Faltungsalgebra $\mathscr{R}$ der „Faltungsring“ und die Divisionsalgebra $\mathfrak{K}$ der „Körper der Faltungsquotienten“, jeweils mit der zusätzlichen Eigenschaft einer Vervielfachung. Daher die Zeichen $\mathscr{R}$ für „Ring“ und $\mathfrak{K}$ für „Körper“.

Bemerkung 13. So wie oben für $\mathscr{R}$, d.h. für $(\mathscr{R}, +, \langle 0\rangle, *, \mathbb{K}, \cdot)$, so hätte auch für $\mathscr{U}$, d.h. für $(\mathscr{U}, +, \langle 0\rangle, *, \mathbb{K}, \cdot)$ gezeigt werden können, daß $(\mathscr{U}, +, \langle 0\rangle, *, \mathbb{K}, \cdot)$ mit den aus der Faltungsalgebra $\mathscr{R}$ übernommenen Rechenoperationen $+$ und $\cdot$ sowie mit

$$*\colon \mathscr{U} \times \mathscr{U} \to \mathscr{U}, \quad (f,g) \mapsto f * g, \quad (f * g)(t) := \int_{-\infty}^{\infty} f(t-\tau) g(\tau)\, d\tau,\ t \in \mathbb{R}, \tag{55}$$

ebenfalls eine Faltungsalgebra ist (die aber nur noch im wesentlichen nullteilerfrei ist). Und damit wäre dann die Abbildung

$$\widetilde{\Phi}\colon \mathscr{U} \to \mathfrak{K}, \quad f \mapsto \widetilde{\Phi}(f) := (f * \sigma)/\sigma \tag{56}$$

auch ein Algebra-Homomorphismus von der Faltungsalgebra $\mathscr{U}$ in die Divisionsalgebra $\mathfrak{K}$ der Faltungsquotienten, nun aber ein nicht injektiver Algebra-Homomorphismus, dessen Einschränkung $\widetilde{\Phi}|_{\mathscr{R}}$ auf die Faltungsalgebra $\mathscr{R}$ der Homomorphismus Φ der Operatorenrechnung ist. Damit die folgende Darstellung der Operatorenrechnung aber möglichst einfach bleibt, soll die Faltungsalgebra $\mathscr{U}$ und der Algebra-Homomorphismus $\widetilde{\Phi}\colon \mathscr{U} \to \mathfrak{K}$ jedoch nur am Rande betrachtet werden.

Bemerkung 14. Für jede Funktion $f \in \mathscr{R}$ werde $\Phi(f) \in \mathfrak{K}$ „die Darstellung der Funktion als *Funktionen-Operator*" genannt. Also ist $\Phi(\mathscr{R}) = \{\Phi(f) \mid f \in \mathscr{R}\}$, der Bildbereich des ersten Homomorphismus $\Phi\colon \mathscr{R} \to \mathfrak{K}$ der Operatorenrechnung, die Unteralgebra der Funktionen-Operatoren der Divisionsalgebra $\mathfrak{K}$ der Faltungsquotienten.

Aufgrund der Injektivität des ersten Homomorphismus Φ der Operatorenrechnung gibt es zu jedem Funktionen-Operator $q \in \Phi(\mathscr{R})$ genau eine Funktion $f \in \mathscr{R}$ mit $q = \Phi(f)$. Diese Funktion $f \in \mathscr{R}$ mit $\Phi(f) = q$ werde „die *Interpretation* des Funktionen-Operators als Funktion" genannt.

Bemerkung 15. Für jede Konstante $\alpha \in \mathbb{K}$ werde $\Psi(\alpha) \in \mathfrak{K}$ „die Darstellung der Konstanten als *Konstanten-Operator*" genannt. Also ist $\Psi(\mathbb{K}) = \{\Psi(\alpha) \mid \alpha \in \mathbb{K}\}$, der Bildbereich des zweiten Homomorphismus $\Psi\colon \mathbb{K} \to \mathfrak{K}$, die Unteralgebra der Konstanten-Operatoren der Divisionsalgebra $\mathfrak{K}$ der Faltungsquotienten.

Aufgrund der Injektivität des zweiten Homomorphismus Ψ der Operatorenrechnung gibt es zu jedem Konstanten-Operator $q \in \Psi(\mathbb{K})$ genau eine Konstante $\alpha \in \mathbb{K}$ mit $q = \Psi(\alpha)$. Diese Konstante $\alpha \in \mathbb{K}$ mit $\Psi(\alpha) = q$ werde „die *Interpretation* des Konstanten-Operators als Konstante" genannt.

Bemerkung 16. Der bis hierher aufgedeckte Zusammenhang zwischen den Algebren $\mathscr{R}$, $\Phi(\mathscr{R})$, $\mathscr{U}$, $\widetilde{\Phi}(\mathscr{U})$, $\mathbb{K}$, $\Psi(\mathbb{K})$ und $\mathfrak{K}$ und den Homomorphismen Φ, $\widetilde{\Phi}$ und Ψ ist in der folgenden Abbildung 2 dargestellt. In dieser Abbildung ist auch angedeutet, wo man sich die Laplace-Transformation $\mathfrak{L}$ vorstellen kann: Definiert auf einem Teil von $\mathscr{U}$ und mit Werten in den auf einer Halbebene holomorphen Funktionen. Aus dieser Abbildung aber bitte nicht zu viel herauslesen!

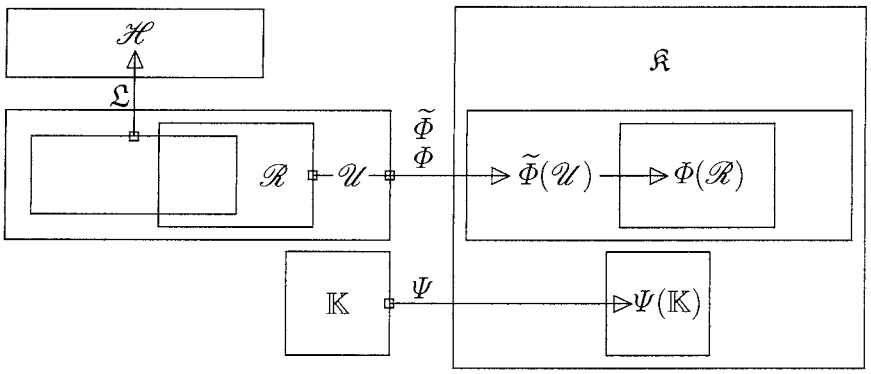


Abbildung 2: Die Verhältnisse bei der Operatorenrechnung nach Mikusiński

3.3 Einige Regeln der Operatorenrechnung nach Mikusiński

Definition 17. Mit Hilfe des ersten Homomorphismus $\Phi\colon \mathscr{R} \to \mathfrak{K}$ der Operatorenrechnung werde gesetzt:

$$\ell := \Phi(\sigma) = (\sigma * \sigma)/\sigma \qquad \text{und} \qquad p := \ell^{-1}. \tag{57}$$

Diese beiden Faltungsquotienten bzw. Operatoren aus $\mathfrak{K}$ haben besondere Namen, ℓ heißt *Integrationsoperator* und p heißt *Differentiationsoperator*.

Bemerkung 18. Der Name „Integrationsoperator" für den Faltungsquotienten ℓ ist dadurch begründet, daß für jede Funktion $f \in \mathscr{R}$ und alle $t \in \mathbb{R}$ natürlich

$$\begin{aligned}(\sigma * f)(t) &= \textstyle\int_0^t \sigma(t-\tau) f(\tau)\, d\tau && \text{Definition von } * \text{ in } \mathscr{R} \\ &= \textstyle\int_0^t f(\tau)\, d\tau && \text{wegen } \sigma(t-\tau) = 1 \text{ für } \tau \leq t \end{aligned} \tag{58}$$

gilt, also in $\mathscr{R}$ die Gleichung

$$\sigma * f = \big\langle \textstyle\int_0^t f(\tau)\, d\tau \big\rangle \tag{59}$$

erfüllt ist, und damit in $\mathfrak{K}$ schließlich gilt:

$$\ell \circledast \Phi(f) = \Phi(\sigma) \circledast \Phi(f) = \Phi(\sigma * f) = \Phi\big(\big\langle \textstyle\int_0^t f(\tau)\, d\tau \big\rangle\big). \tag{60}$$

Der Operator ℓ hat also in gewisser Weise etwas mit dem Integrieren zu tun. Daher der Name „Integrationsoperator". Und für die Operatoren ℓ und $p = \ell^{-1}$ gilt aufgrund ihrer Definition in $\mathfrak{K}$ natürlich

$$p \circledast \ell = \ell \circledast p = 1_{\mathfrak{K}}, \tag{61}$$

also ist in gewisser Weise auch der Name „Differentiationsoperator" für den Operator p, den „Inversen des Integrationsoperators", gerechtfertigt. Ein weiterer Grund für diese Namensgebung wird gleich deutlich werden.

Bemerkung 19. Für eine Funktion $g \in \mathscr{R}$ ist auch $\dot{g} \in \mathscr{R}$, wenn g auf dem offenen Intervall $]0, \infty[$ stetig differenzierbar ist, und wenn $\dot{g}(0)$ die rechtsseite Ableitung von g im Nullpunkt ist. Eine Funktion $g \in \mathscr{R}$ mit $\dot{g} \in \mathscr{R}$ wird auch „stetig differenzierbar" genannt.

Satz 20 (Differentiationssatz der Operatorenrechnung).
Für jede Funktionen $g \in \mathscr{R}$ mit $\dot{g} \in \mathscr{R}$ gilt in $\mathfrak{K}$ die Gleichung[1]

$$\Phi(\dot{g}) = p \circledast \Phi(g) \ominus g(0) \odot 1_{\mathfrak{K}}. \tag{62}$$

Beweis. Ist $g \in \mathscr{R}$ mit $\dot{g} \in \mathscr{R}$, dann sind auch $\sigma * \dot{g}$ und $g - g(0) \cdot \sigma \in \mathscr{R}$, also ist zunächst $(\sigma * \dot{g})(t) = (g - g(0) \cdot \sigma)(t)$ für alle $t < 0$. Für alle $t \geq 0$ gilt aber auch

$$\begin{aligned}&(\sigma * \dot{g})(t) = \textstyle\int_0^t \sigma(t-\tau)\dot{g}(\tau)\, d\tau \;\ldots\ldots \text{Definition von } * \text{ in } \mathscr{R} \\ &= \textstyle\int_0^t \dot{g}(\tau)\, d\tau \;\ldots\ldots\ldots\ldots\ldots\ldots \text{wegen } \sigma(t-\tau) = 1 \text{ für } \tau \leq t \\ &= g(\tau)\big|_{\tau=0}^{t} = g(t) - g(0) \;\ldots\ldots\ldots\ldots \text{HS der Differential- und Integralrechnung} \\ &= g(t) - g(0) \cdot \sigma(t) \;\ldots\ldots\ldots\ldots\ldots \text{wegen } \sigma(t) = 1 \text{ für } t \geq 0 \\ &= (g - g(0) \cdot \sigma)(t) \;\ldots\ldots\ldots\ldots\ldots\ldots \text{wegen Funktion und Funktionswert.} \end{aligned} \tag{63}$$

[1] Und später im Operatorenkalkül einfach $\dot{g} = p\, g - g(0)$.

Also in $\mathscr{R}$ ist $\sigma * \dot{g} = g - g(0) \cdot \sigma$ und in $\mathfrak{K}$:

$$\begin{aligned}
&\ell \circledast \Phi(\dot{g}) = \Phi(\sigma) \circledast \Phi(\dot{g}) \quad \ldots\ldots\ldots\ldots \text{Definition von } \ell \\
&= \Phi(\sigma * \dot{g}) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \Phi \text{ ist Homomorphismus} \\
&= \Phi(g - g(0) \cdot \sigma) \ldots\ldots\ldots\ldots\ldots\ldots \text{soeben } \sigma * f = g - g(0) \cdot \sigma \text{ gezeigt} \\
&= \Phi(g) \ominus g(0) \odot \Phi(\sigma) \ldots\ldots\ldots\ldots\ldots \Phi \text{ ist Homomorphismus und Bemerkung 4} \\
&= \Phi(g) \ominus g(0) \odot \ell \ldots\ldots\ldots\ldots\ldots\ldots \text{Definition von } \ell. \qquad (64)
\end{aligned}$$

Damit ergibt sich in $\mathfrak{K}$ weiter:

$$\begin{aligned}
&\Phi(\dot{g}) = 1_{\mathfrak{K}} \circledast \Phi(\dot{g}) = (p \circledast \ell) \circledast \Phi(\dot{g}) \quad \ldots \text{da } 1_{\mathfrak{K}} = p \circledast \ell \text{ Einselement in } \mathfrak{K} \\
&= p \circledast (\ell \circledast \Phi(\dot{g})) \ldots\ldots\ldots\ldots\ldots\ldots \text{Assoziativität von } \circledast \text{ in } \mathfrak{K} \\
&= p \circledast (\Phi(g) \ominus g(0) \odot \ell) \ldots\ldots\ldots\ldots \text{soeben } \ell \circledast \Phi(\dot{g}) = \Phi(g) \ominus g(0) \odot \ell \text{ gezeigt} \\
&= p \circledast \Phi(g) \ominus (p \circledast (g(0) \odot \ell)) \ldots\ldots\ldots \text{Distributivität in } \mathfrak{K} \text{ und Bemerkung 2} \\
&= p \circledast \Phi(g) \ominus (g(0) \odot (p \circledast \ell)) \ldots\ldots\ldots \text{Eigenschaft (s5) der Vervielfachung in } \mathfrak{K} \\
&= p \circledast \Phi(g) \ominus g(0) \odot 1_{\mathfrak{K}} \ldots\ldots\ldots\ldots\ldots \text{wegen } p \circledast \ell = 1_{\mathfrak{K}}. \qquad (65)
\end{aligned}$$

Damit ist der Satz bewiesen. □

Bemerkung 21. Für eine Funktion $g \in \mathscr{R}$ mit $\dot{g} \in \mathscr{R}$ und $g(0) = 0$ ist damit speziell: $\Phi(\dot{g}) = p \circledast \Phi(g)$.[2] Dies ist ein weiterer Grund dafür, warum der Operator p „Differentiationsoperator" genannt wird.

Satz 22. *Für jedes $\alpha \in \mathbb{K}$ gilt für die Funktion $\langle e^{\alpha t} \rangle \in \mathscr{R}$ in $\mathfrak{K}$ die Gleichung* [3]

$$\Phi\big(\langle e^{\alpha t} \rangle\big) = (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-1}. \qquad (66)$$

Beweis. Es ist $g := \langle e^{\alpha t} \rangle \in \mathscr{R}$ stetig differenzierbar mit $\dot{g}(t) = \alpha e^{\alpha t} = \alpha\, g(t)$ für alle $t \in \mathbb{R}$, also mit $\dot{g} = \alpha \cdot g$. Damit ist wegen $g(0) = 1$ aufgrund des Differentiationssatz

$$\Phi\big(\alpha \cdot g\big) = \Phi(\dot{g}) = p \circledast \Phi(g) \ominus 1 \odot 1_{\mathfrak{K}} = p \circledast \Phi(g) \ominus 1_{\mathfrak{K}}. \qquad (67)$$

Wegen $\Phi\big(\alpha \cdot g\big) = \alpha \odot \Phi(g) = 1_{\mathfrak{K}} \circledast (\alpha \odot \Phi(g)) = (\alpha \odot 1_{\mathfrak{K}}) \circledast \Phi(g)$ gilt damit

$$(\alpha \odot 1_{\mathfrak{K}}) \circledast \Phi(g) = p \circledast \Phi(g) \ominus 1_{\mathfrak{K}}. \qquad (68)$$

Diese Gleichung in $\mathfrak{K}$ nach $\Phi(g)$ auflösen:

$$(p \ominus \alpha \odot 1_{\mathfrak{K}}) \circledast \Phi(g) = 1_{\mathfrak{K}} \qquad \text{und} \qquad (69)$$

$$\Phi(g) = (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-1} \circledast 1_{\mathfrak{K}} = (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-1}. \qquad (70)$$

Also ist der Satz wegen $g = \langle e^{\alpha t} \rangle$ bewiesen. □

[2] Und später im Operatorenkalkül einfach $\dot{g} = p\, g$.

[3] Und später im Operatorenkalkül einfach $\langle e^{\alpha t} \rangle = (p - \alpha)^{-1}$.

Folgerung 23. *Es ist speziell*[4] $\Phi(\sigma) = p^{-1} = \ell$ (siehe auch Definition 17).

Beweis. Speziell mit $\alpha = 0$ ist $\langle \mathrm{e}^{\alpha t}\rangle = \langle \mathrm{e}^{0\cdot t}\rangle = \sigma$, und damit folgt aufgrund des letzten Satzes $\Phi(\sigma) = (p \ominus 0 \odot 1_{\mathfrak{K}})^{-1} = p^{-1}$, also ist wegen $\Phi(\sigma) = \ell$ alles klar. □

Satz 24. *Für jedes $\alpha \in \mathbb{K}$ und alle $n \in \mathbb{N}$ gilt für die Funktion $\langle t^{n-1}\mathrm{e}^{\alpha t}\rangle \in \mathscr{R}$ in $\mathfrak{K}$ die Gleichung*[5]

$$\Phi\big(\langle t^{n-1}\mathrm{e}^{\alpha t}\rangle\big) = (n-1)! \odot (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-n}. \tag{71}$$

Beweis. Sei $\alpha \in \mathbb{K}$ beliebig. Nun vollständige Induktion zu Hilfe nehmen.
Für $n = 1$ gilt mit $t^0 = 1$ und $(n-1)! = 0! = 1$ zufolge des vorhergehenden Satzes

$$\begin{aligned}\Phi\big(\langle t^{n-1}\mathrm{e}^{\alpha t}\rangle\big) &= \Phi\big(\langle t^0\mathrm{e}^{\alpha t}\rangle\big) = \Phi\big(\langle \mathrm{e}^{\alpha t}\rangle\big) = (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-1} \\ &= (n-1)! \odot (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-n}.\end{aligned} \tag{72}$$

Sei jetzt $n \in \mathbb{N}$ beliebig und angenommen, daß $\Phi\big(\langle t^{n-1}\mathrm{e}^{\alpha t}\rangle\big) = (n-1)! \odot (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-n}$ ist. Setze zur Abkürzung $g := \langle t^n\mathrm{e}^{\alpha t}\rangle$. Dann sind $g, \dot{g} \in \mathscr{R}$ mit

$$\dot{g}(t) = n \cdot t^{n-1} \cdot \mathrm{e}^{\alpha t} + \alpha \cdot t^n \cdot \mathrm{e}^{\alpha t} = n \cdot \big(t^{n-1}\mathrm{e}^{\alpha t}\big) + \alpha \cdot g(t) \tag{73}$$

für alle $t \in \mathbb{R}$ mit $t \geq 0$, also ist $\dot{g} = n \cdot \langle t^{n-1}\mathrm{e}^{\alpha t}\rangle + \alpha \cdot g \in \mathscr{R}$ mit $g(0) = 0$. Aufgrund des Differentiationssatz, d.h. mit $\Phi(\dot{g}) = p \circledast \Phi(g) \ominus g(0) \odot 1_{\mathfrak{K}}$, gilt damit

$$\begin{aligned}
&p \circledast \Phi(g) = \Phi(\dot{g}) && \text{Differentiationssatz mit } g(0) = 0 \\
&= \Phi\big(n \cdot \langle t^{n-1}\mathrm{e}^{\alpha t}\rangle + \alpha \cdot g\big) && \dot{g} = n \cdot \langle t^{n-1}\mathrm{e}^{\alpha t}\rangle + \alpha \cdot g \\
&= n \odot \Phi\big(\langle t^{n-1}\mathrm{e}^{\alpha t}\rangle\big) \oplus \alpha \odot \Phi(g) && \Phi \text{ ist Homomorphismus} \\
&= n \odot \big((n-1)! \odot (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-n}\big) \oplus \alpha \odot \Phi(g) && \text{zufolge der Induktionsannahme} \\
&= (n \cdot (n-1)!) \odot (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-n} \oplus \alpha \odot \Phi(g) && \text{Eigenschaft (s1) der Vervielfachung} \\
&= n! \odot (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-n} \oplus \alpha \odot \Phi(g) && n \cdot (n-1)! = n! \\
&= n! \odot (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-n} \oplus (\alpha \odot 1_{\mathfrak{K}}) \circledast \Phi(g) && \alpha \odot \Phi(g) = (\alpha \odot 1_{\mathfrak{K}}) \circledast \Phi(g).
\end{aligned} \tag{74}$$

Diese Gleichung in $\mathfrak{K}$ nach $\Phi(g)$ auflösen:

$$(p \ominus \alpha \odot 1_{\mathfrak{K}}) \circledast \Phi(g) = n! \odot (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-n} \qquad \text{und} \tag{75}$$

$$\Phi(g) = (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-1} \circledast \big(n! \odot (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-n}\big) = n! \odot (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-(n+1)}. \tag{76}$$

Also gilt wegen $g = \langle t^n\mathrm{e}^{\alpha t}\rangle$ auch $\Phi\big(\langle t^n\mathrm{e}^{\alpha t}\rangle\big) = n! \odot (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-(n+1)}$, und damit ist der Satz bewiesen. □

Folgerung 25. *Für alle $n \in \mathbb{N}$ ist speziell*[6] $\Phi(\langle t^{n-1}\rangle) = (n-1)! \odot p^{-n}$.

[4] Und später im Operatorenkalkül einfach $\sigma = p^{-1} = \ell$.
[5] Und später im Operatorenkalkül einfach $\langle t^{n-1}\mathrm{e}^{\alpha t}\rangle = (n-1)!\,(p-\alpha)^{-n}$.
[6] Und später im Operatorenkalkül einfach $\langle t^{n-1}\rangle = (n-1)!\,p^{-n}$.

Beweis. Sei $n \in \mathbb{N}$ beliebig. Speziell mit $\alpha = 0$ ist dann $\langle t^{n-1}\mathrm{e}^{\alpha t}\rangle = \langle t^{n-1}\mathrm{e}^{0\cdot t}\rangle = \langle t^{n-1}\rangle$, und damit folgt aufgrund des letzten Satzes $\Phi(\langle t^{n-1}\rangle) = (n-1)! \odot (p \ominus 0 \cdot 1_{\mathfrak{K}})^{-n} = (n-1)! \odot p^{-n}$. Damit ist die Folgerung bewiesen. □

Satz 26. *Für alle Konstanten $\alpha \in \mathbb{K}$ und alle Funktionen $f, g \in \mathscr{R}$ folgt aus $\dot{g} = \alpha \cdot g + f$ die Beziehung*

$$g = \langle \mathrm{e}^{\alpha t}\rangle \cdot g(0) + \langle \mathrm{e}^{\alpha t}\rangle * f. \tag{77}$$

Beweis. Zunächst eine Hilfsgleichung hergeleitet. In $\mathfrak{K}$ gilt:

$$\begin{aligned}
&(p \ominus \alpha \odot 1_{\mathfrak{K}})^{-1} \circledast (g(0) \odot 1_{\mathfrak{K}} \oplus \Phi(f)) \\
&= \Phi\big(\langle \mathrm{e}^{\alpha t}\rangle\big) \circledast (g(0) \odot 1_{\mathfrak{K}} \oplus \Phi(f)) \quad \ldots\ldots\ldots\ldots \text{Satz 22: } (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-1} = \Phi\big(\langle \mathrm{e}^{\alpha t}\rangle\big) \\
&= \Phi\big(\langle \mathrm{e}^{\alpha t}\rangle\big) \circledast (g(0) \odot 1_{\mathfrak{K}}) \oplus \Phi\big(\langle \mathrm{e}^{\alpha t}\rangle\big) \circledast \Phi(f) \ldots \text{Distributivität in } \mathfrak{K} \\
&= (g(0) \odot 1_{\mathfrak{K}}) \circledast \Phi\big(\langle \mathrm{e}^{\alpha t}\rangle\big) \oplus \Phi\big(\langle \mathrm{e}^{\alpha t}\rangle\big) \circledast \Phi(f) \ldots \text{Kommutativität von } \circledast \text{ in } \mathfrak{K} \\
&= g(0) \odot \Phi\big(\langle \mathrm{e}^{\alpha t}\rangle\big) \oplus \Phi\big(\langle \mathrm{e}^{\alpha t}\rangle\big) \circledast \Phi(f) \ldots\ldots\ldots 1_{\mathfrak{K}} \text{ ist Einselement in } \mathfrak{K} \\
&= \Phi\big(g(0) \cdot \langle \mathrm{e}^{\alpha t}\rangle\big) \oplus \Phi\big(\langle \mathrm{e}^{\alpha t}\rangle * f\big) \ldots\ldots\ldots\ldots \Phi \text{ ist Homomorphismus} \\
&= \Phi\big(g(0) \cdot \langle \mathrm{e}^{\alpha t}\rangle + \langle \mathrm{e}^{\alpha t}\rangle * f\big) \ldots\ldots\ldots\ldots\ldots \Phi \text{ ist Homomorphismus} \\
&= \Phi\big(\langle \mathrm{e}^{\alpha t}\rangle \cdot g(0) + \langle \mathrm{e}^{\alpha t}\rangle * f\big) \ldots\ldots\ldots\ldots\ldots g(0) \cdot \langle \mathrm{e}^{\alpha t}\rangle = \langle \mathrm{e}^{\alpha t}\rangle \cdot g(0).
\end{aligned} \tag{78}$$

Angenommen, es gelte $\dot{g} = \alpha \cdot g + f$. Dann ist wegen $f, g \in \mathscr{R}$ auch $\dot{g} \in \mathscr{R}$, und in $\mathfrak{K}$ gilt:

$$\begin{aligned}
&p \circledast \Phi(g) \ominus g(0) \odot 1_{\mathfrak{K}} = \Phi(\dot{g}) \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{Differentiationssatz} \\
&= \Phi(\alpha \cdot g + f) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{wegen } \dot{g} = \alpha \cdot g + f \\
&= \alpha \odot \Phi(g) \oplus \Phi(f) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \Phi \text{ ist Homomorphismus} \\
&= (\alpha \odot 1_{\mathfrak{K}}) \circledast \Phi(g) \oplus \Phi(f) \ldots\ldots\ldots\ldots\ldots\ldots 1_{\mathfrak{K}} \text{ ist Einselement in } \mathfrak{K}.
\end{aligned} \tag{79}$$

Diese Gleichung umordnen zu

$$(p \ominus \alpha \odot 1_{\mathfrak{K}}) \circledast \Phi(g) = g(0) \odot 1_{\mathfrak{K}} \oplus \Phi(f), \tag{80}$$

und nach $\Phi(g)$ auflösen. Unter Beachtung der obigen Hilfsgleichung folgt dann

$$\Phi(g) = (p \ominus \alpha \odot 1_{\mathfrak{K}})^{-1} \circledast (g(0) \odot 1_{\mathfrak{K}} \oplus \Phi(g)) = \Phi\big(\langle \mathrm{e}^{\alpha t}\rangle \cdot g(0) + \langle \mathrm{e}^{\alpha t}\rangle * f\big), \tag{81}$$

und aufgrund der Injektivität des Homomorphismus Φ ist somit $g = \langle \mathrm{e}^{\alpha t}\rangle \cdot g(0) + \langle \mathrm{e}^{\alpha t}\rangle * f$. Damit ist der Satz bewiesen. □

Satz 27 (Eine Anfangswertaufgabe). *Für die Konstanten a, b, $x_0 \in \mathbb{K}$, die Funktionen*

$$f_0 := \langle \mathrm{e}^{at}\rangle \cdot x_0, \qquad f := \langle \mathrm{e}^{at}\rangle \cdot b \in \mathscr{R} \tag{82}$$

und die Abbildungen

$$F_0 \colon \mathfrak{K} \setminus \{a \odot 1_{\mathfrak{K}}\} \to \mathfrak{K}, \quad q \mapsto F_0(q) := (q \ominus a \odot 1_{\mathfrak{K}})^{-1} \odot x_0, \tag{83}$$

$$F : \mathfrak{K} \setminus \{a \odot 1_{\mathfrak{K}}\} \to \mathfrak{K}, \quad q \mapsto F(q) := (q \ominus a \odot 1_{\mathfrak{K}})^{-1} \odot b, \tag{84}$$

gilt: $F_0(p) = \Phi(f_0), \quad F(p) = \Phi(f).$
Für jede Funktion $u \in \mathscr{R}$ *gilt für die Funktion*

$$x := f_0 + f * u \in \mathscr{R} \tag{85}$$

auch

$$\Phi(x) = F_0(p) \oplus F(p) \circledast \Phi(u) \tag{86}$$

sowie für alle $t \geq 0$

$$x(t) = \mathrm{e}^{at}x_0 + \int_0^t \mathrm{e}^{a(t-\tau)}b\,u(\tau)\,d\tau. \tag{87}$$

Die Funktion x *ist die eindeutig bestimmte Funktion aus* $\mathscr{R}$ *mit der Eigenschaft*

$$\dot{x}(t) = a\,x(t) + b\,u(t) \quad \textit{für} \quad t \geq 0 \qquad \textit{und} \qquad x(0) = x_0. \tag{88}$$

Beweis. Wegen $f_0 = \langle \mathrm{e}^{at} \rangle \cdot x_0$, $f = \langle \mathrm{e}^{at} \rangle \cdot b$ folgt mit dem Homomorphismus Φ aufgrund der „Korrespondenz“ $\Phi\big(\langle \mathrm{e}^{at} \rangle\big) = (p \ominus a \odot 1_{\mathfrak{K}})^{-1}$ aus Satz 22

$$\Phi(f_0) = \Phi\big(\langle \mathrm{e}^{at} \rangle \cdot x_0\big) = \Phi\big(\langle \mathrm{e}^{at} \rangle\big) \odot x_0 = (p \ominus a \odot 1_{\mathfrak{K}})^{-1} \odot x_0 = F_0(p), \tag{89}$$

$$\Phi(f) \;= \Phi\big(\langle \mathrm{e}^{at} \rangle \cdot b\big) \;= \Phi\big(\langle \mathrm{e}^{at} \rangle\big) \odot b \;= (p \ominus a \odot 1_{\mathfrak{K}})^{-1} \odot b \;= F(p). \tag{90}$$

Wegen $x = f_0 + f * u$ folgt mit dem Homomorphismus Φ und mit (89) und (90)

$$\Phi(x) = \Phi(f_0 + f * u) = \Phi(f_0) \oplus \Phi(f) \circledast \Phi(u) = F_0(p) \oplus F(p) \circledast \Phi(u). \tag{91}$$

Außerdem gilt wegen $x = f_0 + f * u$ für alle $t \in \mathbb{R}$ stets

$$x(t) = (f_0 + f * u)(t) = f_0(t) + (f * u)(t) = f_0(t) + \textstyle\int_0^t f(t-\tau)u(\tau)\,d\tau, \tag{92}$$

und damit wegen $f_0 = \langle \mathrm{e}^{at} \rangle \cdot x_0$, $f = \langle \mathrm{e}^{at} \rangle \cdot b$ für alle $t \geq 0$ stets

$$x(t) = f_0(t) + \textstyle\int_0^t f(t-\tau)u(\tau)\,d\tau = \mathrm{e}^{at}x_0 + \int_0^t \mathrm{e}^{a(t-\tau)}b\,u(\tau)\,d\tau, \tag{93}$$

also $x(0) = x_0$ sowie mittels Differentiation

$$\begin{aligned}\dot{x}(t) &= \frac{d}{dt}\big(\mathrm{e}^{at}x_0 + \textstyle\int_0^t \mathrm{e}^{a(t-\tau)}b\,u(\tau)\,d\tau\big) = a\,\mathrm{e}^{at}x_0 + \mathrm{e}^{a(t-t)}b\,u(t) + \int_0^t a\,\mathrm{e}^{a(t-\tau)}b\,u(\tau)\,d\tau \\ &= a\big(\mathrm{e}^{at}x_0 + \textstyle\int_0^t \mathrm{e}^{a(t-\tau)}b\,u(\tau)\,d\tau\big) + b\,u(t) = a\,x(t) + b\,u(t).\end{aligned} \tag{94}$$

Damit besitzt die Funktion x aus $\mathscr{R}$ die Eigenschaft (88).
Nun noch zu der Aussage, wonach diese Funktion die einzige Funktion aus $\mathscr{R}$ mit dieser Eigenschaften sein soll. Sei hierzu angenommen, daß die Funktion $\xi \in \mathscr{R}$ die Eigenschaft (88) besitzt, d.h. daß gilt:

$$\dot{\xi}(t) = a\,\xi(t) + b\,u(t) \quad \text{für} \quad t \geq 0 \qquad \text{und} \qquad \xi(0) = x_0. \tag{95}$$

Wegen $u \in \mathscr{R}$ gilt auch $\dot{\xi}(t) = a\,\xi(t) + b\,u(t)$ für $t < 0$, also ist $\dot{\xi} = a \cdot \xi + b \cdot u \in \mathscr{R}$, und zufolge Satz 26 ist damit

$$\xi = \langle \mathrm{e}^{at} \rangle \cdot \xi(0) + \langle \mathrm{e}^{at} \rangle * (b \cdot u) = \langle \mathrm{e}^{at} \rangle \cdot x_0 + \big(\langle \mathrm{e}^{at} \rangle \cdot b\big) * u = f_0 + f * u = x. \tag{96}$$

Also ist die Funktion x die einzige Funktion aus $\mathscr{R}$ mit der Eigenschaft (88). Damit ist der Beweis abgeschlossen. □

4 Der Operatorenkalkül nach Mikusiński

4.1 Von der Operatorenrechnung zum Operatorenkalkül

Der injektive Homomorphismus

$$\Phi\colon \mathscr{R} \to \mathfrak{K}, \quad f \mapsto \Phi(f) = (f * \sigma)/\sigma. \tag{97}$$

von der Faltungsalgebra $\mathscr{R}$ in die Divisionsalgebra $\mathfrak{K}$ der Faltungsquotienten liefert einen Isomorphismus zwischen der Faltungsalgebra $\mathscr{R}$ und dem Bildbereich $\Phi(\mathscr{R}) = \{\Phi(f) \mid f \in \mathscr{R}\}$ des Homomorphismus Φ, der Unteralgebra der Funktionen-Operatoren $\Phi(\mathscr{R})$ der Divisionsalgebra $\mathfrak{K}$. In der Mathematik wird – so wie oben bereits mehrfach erklärt – in einer derartigen Situation, wenn sich zwei Algebren, hier die beiden Algebren $\mathscr{R}$ und $\Phi(\mathscr{R})$, in allen algebraischen Eigenschaften gleich, also „isomorph" verhalten, und die Zugehörigkeit der Elemente zu einer der beiden Algebren nur noch durch die verschiedenen Namen der Elemente, hier f und $\Phi(f)$, deutlich wird, auch noch von den verschiedenen Namen abgesehen. Man sagt in diesem Fall, die beiden Algebren, hier $\mathscr{R}$ und $\Phi(\mathscr{R})$, „werden identifiziert", d.h. gleich gesetzt.

Wenn wir in unserem Fall also die beiden Algebren $\mathscr{R}$ und $\Phi(\mathscr{R})$ identifizieren, dann ist auszugehen von $\mathscr{R} = \Phi(\mathscr{R}) \subset \mathfrak{K}$ und

$$f = \Phi(f) = (f * \sigma)/\sigma \quad \text{für alle} \quad f \in \mathscr{R}. \tag{98}$$

(Und zumeist wird auch auf die für Funktionen $f, g \in \mathscr{U}$ eigentlich geltende Folgerung, daß aus der Gleichung $\widetilde{\Phi}(f) = \widetilde{\Phi}(g)$ lediglich „$f = g$ fast überall" folgt, kein großer Wert gelegt, und einfach davon gesprochen, daß f und g „fast gleich" sind.)

Analog kann aufgrund des injektiven Homomorphismus

$$\Psi\colon \mathbb{K} \to \mathfrak{K}, \quad \alpha \mapsto \Psi(\alpha) := \alpha \odot 1_{\mathfrak{K}} \tag{99}$$

der Körper $\mathbb{K}$ mit dem Bildbereich $\Psi(\mathbb{K}) = \{\Psi(\alpha) \mid \alpha \in \mathbb{K}\}$, der Unteralgebra der Konstanten-Operatoren der Divisionsalgebra $\mathfrak{K}$, identifiziert werden, d.h. wir können ausgehen von $\mathbb{K} = \Psi(\mathbb{K}) \subset \mathfrak{K}$ und

$$\alpha = \Psi(\alpha) = \alpha \odot 1_{\mathfrak{K}} \quad \text{für alle} \quad \alpha \in \mathbb{K}. \tag{100}$$

Durch alle diese Vereinbarungen wird der gesamte bisherige algebraische Formalismus ganz erheblich vereinfacht, wenn zusätzlich noch anstelle der Operationszeichen $\oplus$, $\ominus$, $\circledast$ und $\odot$ wieder die einfacheren Zeichen $+$, $-$, $*$ und $\cdot$ verwendet werden, und die Operationszeichen $*$ und $\cdot$ ganz unterdrückt werden, wenn dies zu keinen Mißverständnissen führt. So wird beispielsweise der komplizierten Ausdruck $p \circledast \Phi(g) \ominus g(0) \odot 1_{\mathfrak{K}}$ aus dem obigen Differentiationssatz einfach zu $p\,g - g(0)$, so daß der Differentiationssatz einfach lautet:

$$\text{Für alle } g \in \mathscr{R} \text{ mit } \dot{g} \in \mathscr{R} \text{ ist } \dot{g} = p\,g - g(0). \tag{101}$$

Hierbei ist jedoch zu beachten, daß die einzelnen Terme der rechten Seite der Gleichung $\dot{g} = p\,g - g(0)$ nur in der Divisionsalgebra $\mathfrak{K}$, nicht aber in der Faltungsalgebra $\mathscr{R}$, einen

Sinn besitzen, denn sowohl der Differentiationsoperator p, wie auch die Konstante $g(0)$, die hier ja eigentlich für den Konstanten-Operator $g(0) \odot 1_{\mathfrak{K}}$ steht, sind keine Funktionen-Operatoren. Und demzufolge ist auch $p\,g - g(0)$ nicht immer ein Funktionen-Operator, was ja eigentlich klar ist, da sonst jede Funktion g aus $\mathscr{R}$ differenzierbar mit einer Ableitung $\dot{g}$ aus $\mathscr{R}$ wäre.

Für das Rechnen in der Faltungsalgebra $\mathscr{R}$ und der Divisionsalgebra $\mathfrak{K}$ der Faltungsquotienten in dieser vereinfachten Schreibweise soll der Begriff *Operatorenkalkül nach Mikusiński* oder einfach *Operatorenkalkül* verwendet werden.

In diesem Operatorenkalkül werden die Faltungsalgebra $\mathscr{R}$ und der Konstantenkörper $\mathbb{K}$ also als Teilmengen der Divisionsalgebra $\mathfrak{K}$ angesehen. Die oben hergeleiteten „Korrespondenzen" nehmen in diesem Operatorenkalkül somit die folgenden Formen an:

$$\langle \mathrm{e}^{\alpha t} \rangle = (p-\alpha)^{-1} \qquad \text{für } \alpha \in \mathbb{K} \tag{102}$$

$$\sigma = p^{-1} = \ell \tag{103}$$

$$\langle t^{n-1} \mathrm{e}^{\alpha t} \rangle = (n-1)!\,(p-\alpha)^{-n} \qquad \text{für } \alpha \in \mathbb{K},\ n \in \mathbb{N} \tag{104}$$

$$\langle t^{n-1} \rangle = (n-1)!\,\ell^n \qquad \text{für } n \in \mathbb{N} \tag{105}$$

Im folgenden soll das Arbeiten mit diesem Operatorenkalkül an einen ausführlicheren Beispiel demonstriert werden.

4.2 Beispiel: Ein dynamisches System

(a) *Für die Konstanten a, b, c, $x_0 \in \mathbb{K}$, die Funktionen $f_0 = \langle \mathrm{e}^{at} \rangle x_0$, $f = \langle \mathrm{e}^{at} \rangle b$ gemäß (82), die Funktionen*

$$g_0 := c\langle \mathrm{e}^{at} \rangle x_0, \qquad g := c\langle \mathrm{e}^{at} \rangle b \in \mathscr{R}, \tag{106}$$

die Abbildungen F_0, F mit $F_0(q) = (q-a)^{-1}x_0$, $F(q) = (q-a)^{-1}b$ gemäß (83), (84) und die Abbildungen

$$G_0 \colon \mathfrak{K} \setminus \{a\} \to \mathfrak{K}, \quad q \mapsto G_0(q) := c(q-a)^{-1}x_0, \tag{107}$$

$$G \ : \mathfrak{K} \setminus \{a\} \to \mathfrak{K}, \quad q \mapsto G(q) \ := c(q-a)^{-1}b \tag{108}$$

gilt: $F_0(p) = f_0$, $F(p) = f$, $G_0(p) = g_0$, $G(p) = g$.

Denn wegen $(p-a)^{-1} = \langle \mathrm{e}^{at} \rangle$ ist dies klar. □

(b) *Für jede Funktion $u \in \mathscr{R}$ gilt für die Funktion $x = f_0 + f * u$ gemäß (85) und die Funktion*

$$y := g_0 + g * u \in \mathscr{R} \tag{109}$$

auch

$$x = F_0(p) + F(p)\,u, \qquad y = G_0(p) + G(p)\,u, \tag{110}$$

sowie für alle $t \geq 0$

$$x(t) = \mathrm{e}^{at}x_0 + \int_0^t \mathrm{e}^{a(t-\tau)}b\,u(\tau)\,d\tau, \qquad y(t) = c\,\mathrm{e}^{at}x_0 + \int_0^t c\,\mathrm{e}^{a(t-\tau)}b\,u(\tau)\,d\tau. \tag{111}$$

Die Funktionen x *und* y *sind die eindeutig bestimmten Funktionen aus* $\mathscr{R}$ *mit den Eigenschaften*

$$\dot{x}(t) = a\,x(t) + b\,u(t) \quad \textit{für} \quad t \geq 0 \quad \textit{und} \quad x(0) = x_0 \quad \textit{sowie} \tag{112}$$

$$y(t) = c\,x(t) \quad \textit{für} \quad t \geq 0. \tag{113}$$

Denn aufgrund von Satz 27 ist auch dies eigentlich alles klar. Der Nachweis der Eindeutigkeit der Lösung der Anfangswertaufgabe soll jetzt hier nochmals, nun aber mit dem Operatorenkalkül nach Mikusiński, gezeigt werden:
Die Funktion $x\colon t \mapsto x(t)$ sei auf $[0, \infty[$ eine Lösung der gegebenen Anfangswertaufgabe. Dann kann o.B.d.A. $x \in \mathscr{R}$, und wegen $u \in \mathscr{R}$ auch $\dot{x} \in \mathscr{R}$, angenommen werden, und es gilt in $\mathscr{R}$ und auch in $\mathfrak{K}$

$$\dot{x} = a\,x + b\,u, \tag{114}$$

Mit dem Differentiationssatz $\dot{x} = p\,x - x_0$ ist daher in $\mathfrak{K}$

$$p\,x - x_0 = a\,x + b\,u, \quad \text{also} \quad (p-a)x = x_0 + b\,u \quad \text{und}$$
$$x = (p-a)^{-1}x_0 + (p-a)^{-1}(b\,u), \tag{115}$$

und somit in $\mathscr{R}$ wegen $(p-a)^{-1} = \langle \mathrm{e}^{at} \rangle$ wieder:

$$x = \langle \mathrm{e}^{at} \rangle x_0 + \langle \mathrm{e}^{at} \rangle (b\,u). \tag{116}$$

Hiermit gilt für alle $t \geq 0$ in $\mathbb{C}$:

$$x(t) = \mathrm{e}^{at}x_0 + \textstyle\int_0^t \mathrm{e}^{a(t-\tau)}b\,u(\tau)\,d\tau. \tag{117}$$

Also ist jede Lösung der Anfangswertaufgabe, sofern sie existiert, von der angegebenen Form. Damit ist die Eindeutigkeit der Lösungsfunktion gezeigt. □

(c) *Für* $u = \langle \mathrm{e}^{st} \rangle \in \mathscr{R}$ *mit* $s \neq a$ *ist*

$$y = G_0(p) - G(p)(s-a)^{-1} + G(s)\,u = g_0 - g\,(s-a)^{-1} + G(s)\,u. \tag{118}$$

Mit $y_0 := g_0 - g\,(s-a)^{-1}$ *ist daher* $y = y_0 + G(s)\,u$ *und* $y(t) = y_0(t) + G(s)\,u(t)$ *für alle* $t \in \mathbb{R}$.

Denn zufolge Teil (b) ist $y = G_0(p) + G(p)\,u = g_0 + G(p)\,u$. Wegen $G(p) = c(p-a)^{-1}b$ und $u = \langle \mathrm{e}^{st} \rangle = (p-s)^{-1}$ ist dann mit den zunächst unbekannten Konstanten A, $B \in \mathbb{K}$

$$\begin{aligned} G(p)\,u &= c(p-a)^{-1}b(p-s)^{-1} = c(p-a)^{-1}bA + cBb(p-s)^{-1} \\ &= c(p-a)^{-1}bA(p-s)(p-s)^{-1} + c(p-a)^{-1}(p-a)Bb(p-s)^{-1} \\ &= c(p-a)^{-1}\big(bA(p-s) + (p-a)Bb\big)(p-s)^{-1}, \end{aligned} \tag{119}$$

und die Gleichung $b = bA(p-s) + (p-a)Bb$ ist für $A = -(s-a)^{-1}$ und $B = (s-a)^{-1}$ erfüllt, also folgt

$$\begin{aligned} G(p)\,u &= -c(p-a)^{-1}b(s-a)^{-1} + c(s-a)^{-1}b(p-s)^{-1} \\ &= -G_0(p)(s-a)^{-1} + G(s)u \end{aligned} \tag{120}$$

und damit

$$\begin{aligned} y &= G_0(p) - G(p)(s-a)^{-1} + G(s)u = g_0 - g(s-a)^{-1} + G(s)u \\ &= c\langle \mathrm{e}^{at}\rangle\big(x_0 - b(s-a)^{-1}\big) + G(s)u = y_0 + G(s)u. \end{aligned} \tag{121}$$

Also gilt für alle $t \in \mathbb{R}$:

$$y(t) = c\,\mathrm{e}^{at}\big(x_0 - b(s-a)^{-1}\big) + G(s)u(t) = y_0(t) + G(s)u(t). \tag{122}$$

Man beachte aber, daß hier in den Gleichungen für y und $y(t)$ der Faktor $G(s)$ eine Konstante aus $\mathbb{K}$ ist, die als Faktor vor u bzw. $u(t)$ steht. Die „Übertragungsfunktion" $G(s)$ tritt hier also lediglich als Faktor im „Zeitbereich" auf. □

(d) *Für* $u = \langle \mathrm{e}^{at}\rangle \in \mathscr{R}$ *ist*

$$y(t) = c\,\mathrm{e}^{at}(x_0 + b\,t) = g_0(t) + g(t)\,t. \tag{123}$$

Denn zufolge Teil (b) ist $y = G_0(p) + G(p)\,u = g_0 + G(p)\,u$. Wegen $G(p) = c(p-a)^{-1}b$ und $u = \langle \mathrm{e}^{at}\rangle = (p-a)^{-1}$ ist dann

$$y = g_0 + c(p-a)^{-1}b(p-a)^{-1} = g_0 + c(p-a)^{-2}b. \tag{124}$$

Mit der „Korrespondenz" $\langle t^{n-1}\mathrm{e}^{at}\rangle = (n-1)!(p-a)^{-n}$ ist für $n = 2$ einfach $\langle t\,\mathrm{e}^{at}\rangle = (p-a)^{-2}$ und

$$y = g_0 + c\langle t\,\mathrm{e}^{at}\rangle b = c\langle \mathrm{e}^{at}x_0\rangle + \langle c\,\mathrm{e}^{at}b\,t\rangle = c\langle \mathrm{e}^{at}(x_0 + b\,t)\rangle. \tag{125}$$

Also gilt für alle $t \geq 0$ stets $y(t) = c\,\mathrm{e}^{at}(x_0 + b\,t) = c\,\mathrm{e}^{at}x_0 + c\,\mathrm{e}^{at}b\,t = g_0(t) + g(t)t$. □

(e) *Speziell für* $a = b = c = x_0 = 1$ *und* $u = \langle (2t-1)\mathrm{e}^{t^2}\rangle$ *ist* $y = \langle \mathrm{e}^{t^2}\rangle$.

Denn für $v := \langle \mathrm{e}^{t^2}\rangle$ ist $v(0) = 1$ und $\dot{v}(t) = 2t\mathrm{e}^{t^2}$ für $t \geq 0$. Also gilt aufgrund des Differentiationssatzes zunächst

$$\langle 2t\mathrm{e}^{t^2}\rangle = \dot{v} = p\,v - v(0) = p\langle \mathrm{e}^{t^2}\rangle - 1 \tag{126}$$

und damit dann

$$u = \langle (2t-1)\mathrm{e}^{t^2}\rangle = \langle 2t\mathrm{e}^{t^2}\rangle - \langle \mathrm{e}^{t^2}\rangle = p\langle \mathrm{e}^{t^2}\rangle - 1 - \langle \mathrm{e}^{t^2}\rangle = (p-1)\langle \mathrm{e}^{t^2}\rangle - 1. \tag{127}$$

Weiter ist gemäß Teil (a)

$$G_0(p) = c(p-a)^{-1}x_0 = (p-1)^{-1} = \langle \mathrm{e}^{t}\rangle = c\langle \mathrm{e}^{at}\rangle x_0 = g_0 \tag{128}$$

$$G(p) \;= c(p-a)^{-1}b \;= (p-1)^{-1} = \langle \mathrm{e}^{t}\rangle = c\langle \mathrm{e}^{at}\rangle b \;= g \tag{129}$$

und damit gemäß Teil (b)

$$\begin{aligned} y &= g_0 + g\,u = G_0(p) + G(p)\,u = \langle \mathrm{e}^{t}\rangle + (p-1)^{-1}\big((p-1)\langle \mathrm{e}^{t^2}\rangle - 1\big) \\ &= \langle \mathrm{e}^{t}\rangle + \langle \mathrm{e}^{t^2}\rangle - (p-1)^{-1} = \langle \mathrm{e}^{t}\rangle + \langle \mathrm{e}^{t^2}\rangle - \langle \mathrm{e}^{t}\rangle = \langle \mathrm{e}^{t^2}\rangle. \end{aligned} \tag{130}$$

□

Bemerkung 28. Der Teil (e) demonstriert, daß die Operatorenrechnung nach Mikusiński einen größeren Anwendungsbereich als die Laplace-Transformation besitzt, denn die hier auftretende Funktion $u = \langle (2t-1)\mathrm{e}^{t^2} \rangle \in \mathscr{R}$ ist nicht Laplace-transformierbar, da das Integral

$$\int_0^\infty (2t-1)\mathrm{e}^{t^2} \cdot \mathrm{e}^{-st}\, dt = \lim_{T\to\infty} \int_0^T (2t-1)\mathrm{e}^{t^2} \cdot \mathrm{e}^{-st}\, dt \tag{131}$$

für kein $s \in \mathbb{C}$ existiert bzw. konvergiert. Derartige Konvergenzprobleme sind der Operatorenrechnung bzw. dem Operatorenkalkül nach Mikusiński jedoch völlig fremd.

4.3 Ein Vergleich von Laplace-Transf. und Operatorenkalkül

Nehmen wir einen kurzen Vergleich des Vorgehens beim Lösen der Anfangswertaufgabe anhand der beiden Formeln des Differentiationssatzes der Laplace-Transformation und des Operatorenkalküls nach Mikusiński vor.

Für eine auf $[0,\infty[$ definierte und stetig differenzierbare Funktion g gilt zum einen innerhalb der Laplace-Transformation

$$\mathcal{L}(\dot{g})(s) = s\,\mathcal{L}(g)(s) - g(0) \qquad \text{für} \quad s \in \mathbb{C},\ \mathrm{Re}(s) > \alpha \quad \text{mit einem} \quad \alpha \in \mathbb{R}, \tag{132}$$

und andererseits innerhalb des Operatorenkalküls nach Mikusiński

$$\dot{g} = p\,g - g(0). \tag{133}$$

Daß hier einmal der Buchstabe „s“ und einmal der Buchstabe „p“ auftritt ist unwesentlich. Aber ganz wesentlich ist:

- s ist hier eine komplexe Variable, und $s\,\mathcal{L}(g)(s)$ steht hier für das Produkt aus der komplexen Zahl s und dem Funktionswert $\mathcal{L}(g)(s)$ (der Laplace-Transformierten $\mathcal{L}(g)$ der Funktion g an der Stelle s) im Zahlenkörper $\mathbb{C}$.
- p ist hier der Differentiationsoperator des Operatorenkalküls nach Mikusiński, und $p\,g$ steht hier für das Produkt aus dem Differentiationsoperator p und dem Funktionen-Operator g (der Darstellung der Funktion g als Element in der Divisionsalgebra der Faltungsquotienten) in der Divisionsalgebra der Faltungsquotienten.

(In dem meisten Darstellungen des Operatorenkalküls nach Mikusiński wird übrigens für den Differentiationsoperator auch der Buchstabe „s“ verwendet, siehe bei MIKUSIŃSKI in [3] und bei ERDÉLYI in [5]). Ich habe hier lediglich das alternative Symbol „p“ benutzt, um beim Vergleich die Unterschiede von Laplace-Transformation und Operatorenkalkül nach Mikusiński klarer herausarbeiten zu können. Aber Vorsicht, der Buchstabe „p“ wird von einigen Autoren auch zur Kennzeichnung der komplexen Variablen der Laplace-Transformation verwendet.)

Wie aber kommen diese beiden Darstellungen des Differentiationssatzes zustande?

Bei der Laplace-Transformation ist es üblicherweise keine Folgerung aus dem Faltungssatz sondern das Ergebnis der partiellen Integration

$$\int_0^\infty \mathrm{e}^{-st}\dot{g}(t)\,dt = \mathrm{e}^{-st}g(t)\Big|_{t=0}^{\infty} + s\int_0^\infty \mathrm{e}^{-st}g(t)\,dt = s\int_0^\infty \mathrm{e}^{-st}g(t)\,dt - g(0) \tag{134}$$

unter der zusätzlichen Voraussetzung über die Grenzwerte

$$\mathrm{e}^{-st}g(t)\Big|_{t=0}^{\infty} = \lim_{t\to\infty} \mathrm{e}^{-st}g(t) - \lim_{t\to 0,t>0} \mathrm{e}^{-st}g(t) = 0 - g(0) = -g(0). \tag{135}$$

Hierbei ist $\lim_{t\to 0,t>0} \mathrm{e}^{-st}g(t) = g(0)$ aufgrund unserer Voraussetzung an g klar. Aber $\lim_{t\to\infty} \mathrm{e}^{-st}g(t) = 0$ ist eine weitere Forderung an die Funktion g, außer daß sie Laplace-transformierbar und auf $[0,\infty[$ stetig differenzierbar sein soll. Zumeist wird aber stattdessen gefordert, daß die Funktion g auf $[0,\infty[$ stetig differenzierbar und ihre Ableitung $\dot{g}$ Laplace-transformierbar sein soll, siehe beispielsweise bei DOETSCH in [2].

Beim Operatorenkalkül nach Mikusiński ist der Differentiationssatz dagegen eine Konsequenz des Hauptsatzes der Differential- und Integralrechnung

$$\int_0^t \dot{g}(\tau)\,d\tau = g(\tau)\Big|_{\tau=0}^{t} = g(t) - g(0) \qquad \text{für} \quad t \geq 0, \tag{136}$$

und der Schreibweise dieser Gleichung für die Funktionswerte als Gleichung für die Funktionen

$$\sigma\,\dot{g} = g - g(0)\,\sigma, \tag{137}$$

bzw. der Darstellung dieser Gleichung als Gleichung in der Divisionsalgebra der Faltungsquotienten, und der in dieser Divisionsalgebra erfolgten Multiplikation mit dem inversen Element p des Funktionen-Operators $\sigma \neq \langle 0\rangle$, also mit $p = \sigma^{-1}$ mit der Eigenschaft $p\sigma = 1$

$$\dot{g} = p\,g - g(0). \tag{138}$$

Hier gibt es also keine zusätzlichen Forderungen an die Funktion g, außer daß sie auf $[0,\infty[$ stetig differenzierbar sein soll:

Im Operatorenkalkül nach Mikusiński gibt es keine Probleme mit der Konvergrenz!

Und so geht der Bereich der Anwendbarkeit des Operatorenkalküls nach Mikusiński auch über den Bereich der Anwendbarkeit der Laplace-Transformation hinaus. Siehe hierzu im obigen Beispiel 4.2 den Teil (e).

5 Schlußfolgerungen

Ergebnisse (mit Bewertung):

(1) Die Laplace-Transformation ist völlig ausreichend. Es besteht kein Interesse an neuen Interpretationen alter Ergebnisse. Die ganze Welt arbeitet doch mit der Laplace-Transformation. $\ominus$

(2) Die hier ausgeführten Beispiele sollten gezeigt haben, daß zumindest alle Ergebnisse, die mittels der Laplace-Transformation erhalten werden können, auch mittels der

Operatorenrechnung nach Mikusiński zu erhalten sind. (Viel ausführlicher hat dies beispielsweise BERG in seinen beiden Lehrbüchern [9] und [10] demonstriert.) Und die aus der Laplace-Transformation bekannten Korrespondenzentafeln sind nach einer Umdeutung weiter verwendbar. Einige Autoren benutzen für den Differentiationsoperator das Symbol s und nicht p, so daß sie die Korrespondenzentafeln der Laplace-Transformation, beispielsweise die bei DOETSCH in [4] angegebenen, quasi direkt übernehmen können. ⊙

(3) Und natürlich kann mit Matrizen, deren Elemente der Faltungsalgebra $\mathscr{R}$ oder der Divisionsalgebra $\mathfrak{K}$ der Faltungsquotienten angehören, in ganz der üblichen Art und Weise gerechnet werden, so wie bei der Anwendung der Laplace-Transformation. ⊙

(4) Ergebnisse aus der Funktionentheorie sind, sofern sie überhaupt noch benötigt werden, auch leicht in die Operatorenrechnung nach Mikusiński zu übernehmen, da die komplexen Zahlen und die Konstanten-Operatoren ja isomorph sind, und die komplexen Zahlen daher als ein Teil der Divisionsalgebra $\mathfrak{K}$ der Faltungsquotienten angesehen werden können. Und natürlich ist $\mathbb{C}[p] \cong \mathbb{C}[X]$ sowie $\mathbb{C}(p) \cong \mathbb{C}(X)$, d.h. $\mathbb{C}[p]$, die Gesamtheit der Polynome in p mit Koeffizienten aus $\mathbb{C}$, ist isomorph zu $\mathbb{C}[X]$, der Gesamtheit der Polynome in der Unbestimmten X mit Koeffizienten aus $\mathbb{C}$, und $\mathbb{C}(p)$, die Gesamtheit der rationalen Ausdrücke in p mit Koeffizienten aus $\mathbb{C}$, ist isomorph zu $\mathbb{C}(X)$, der Gesamtheit der rationalen Ausdrücke in der Unbestimmten X mit Koeffizienten aus $\mathbb{C}$. Also ist alles was über das Rechnen in $\mathbb{C}(X)$ bekannt ist (z.B. die Teilbruchzerlegung usw.) auch in $\mathbb{C}(p)$ anwendbar.
Und sollte jemand mit der Fourier-Transformation arbeiten müssen, z.B. im Rahmen der Mustererkennung usw., so ist die Operatorenrechnung nach Mikusiński – einschließlich des später unter Punkt (8) angesprochenen Grenzwertbegriffes – ganz natürlich als ein Teil der Theorie der Distributionen und der Fourier-Transformation anzusehen. Siehe hierzu z.B. die Darstellung bei BERZ in [6]. ⊙

(5) Der Formalismus des Operatorenkalküls nach Mikusiński ist keineswegs komplizierter als der Formalismus der Laplace-Transformation. Für jemandem, der viele Jahre mit der Laplace-Transformation gearbeitet hat, unter Umständen nur etwas ungewohnt (weil so viel einfacher, denn die Korrespondenzen sind einfach Gleichungen). Für jemandem, der die Laplace-Transformation jedoch nicht kennt, so glaube ich fest, ist der Formalismus des Operatorenkalküls nach Mikusiński leichter zu erlernen als der Formalismus der Laplace-Transformation. ⊕

(6) Vom mathematischen Standpunkt aus ist die Operatorenrechnung nach Mikusiński gegenüber der Laplace-Transformation nahezu trivial. Denn bis auf die Aussage von TITCHMARSH zur Nullteilerfreiheit der Faltungsalgebra $\mathscr{R}$, sind die Beweise aller oben angegebenen Aussagen nach den ersten drei Wochen eines Algebra-Kurses ganz einfache Übungen. Und ich glaube nicht, daß heute noch jemand auf die Idee kommen würde, so etwas undurchsichtiges wie die Laplace-Transformation zu entwickeln. Denn diese besteht doch eigentlich nur aus einer Ansammlung unklaren Konvergenzen. ⊕

(7) Und wie oben bereits angedeutet wurde, ist auf die hier zumeist erfolgte Einschränkung auf das Arbeiten mit Funktionen aus $\mathscr{R}$ ohne Schwierigkeit zu verzichten. Denn ohne große Mühe ist es möglich, den Homomorphismus $\Phi\colon \mathscr{R} \to \mathfrak{K}$ auf die Faltungs-

algebra $\mathscr{U}$ aller meßbaren und lokal im wesentlichen beschränkten Funktionen $\mathbb{R} \to \mathbb{K}$ zu erweitern. $\oplus\oplus$

(8) Und wo ist die Delta-Funktion? Ja es gibt sie auch in der hier vorgestellten Operatorenrechnung nach Mikusiński nicht, sofern man erwartet, daß es einen Funktionen-Operator oder eine Funktion $\delta \in \mathscr{R}$ mit dem Funktionen-Operator $\Phi(\delta)$ und der Eigenschaft $f*\delta = f$ bzw. $\Phi(f)\circledast\Phi(\delta) = \Phi(f)$ für alle Funktionen $f \in \mathscr{R}$ gibt. Aber es gilt doch sehr wohl $\Phi(f) \circledast 1_{\mathfrak{K}} = \Phi(f)$ für alle $f \in \mathscr{R}$ mit dem Einselement $1_{\mathfrak{K}} = \sigma/\sigma$ der Divisionsalgebra $\mathfrak{K}$. Und da dieses Einselement eindeutig bestimmt und kein Funktionen-Operator ist, gibt es in dieser Operatorenrechnung also keine Delta-Funktion. Alles, was aber eigentlich von der Delta-Funktion gemacht werden sollte, wird in der Divisionsalgebra $\mathfrak{K}$ vom Einselement $1_{\mathfrak{K}}$ erreicht. Also gibt es in der Operatorenrechnung nach Mikusiński doch eine „Delta-Funktion", die jedoch keine Funktion bzw. kein Funktionen-Operator mehr, sondern ein Konstanten-Operator ist, nämlich der Faltungsquotient σ/σ. Einige Autoren bezeichnen diesen Faltungsquotienten daher auch mit dem Symbol δ und nicht mit dem Symbol $1_{\mathfrak{K}}$ (bzw. mit dem Symbol 1 im Operatorenkalkül).
Aber natürlich gibt es in den Faltungsalgebra $\mathscr{R}$ Funktionen, durch die in gewisser Weise die „Delta-Funktion" $1_{\mathfrak{K}}$ angenähert werden kann. Hierzu ist in der Divisionsalgebra $\mathfrak{K}$ jedoch ein Grenzwertbegriff notwendig, so daß z.B. für eine Folge $(q_n)_{n\in\mathbb{N}}$ aus $\mathfrak{K}$ gesagt werden kann, daß $\lim_{n\to\infty} q_n = 1_{\mathfrak{K}}$ ist. Und ein derartiger Grenzwertbegriff existiert natürlich, jedoch nicht in der obigen rein algebraischen Darstellung der Operatorenrechnung. Betrachtet man dann z.B.

$$\delta_n := \left\langle n\,\mathrm{e}^{-nt}\right\rangle \quad \text{mit} \quad q_n := \Phi(\delta_n) = n \odot \left(p \oplus n \odot 1_{\mathfrak{K}}\right)^{-1} = \left(1_{\mathfrak{K}} \oplus \tfrac{1}{n} \odot p\right)^{-1}, \tag{139}$$

so ist mit diesem Grenzwertbegriff in $\mathfrak{K}$

$$\lim_{n\to\infty} q_n = \lim_{n\to\infty} \Phi(\delta_n) = \lim_{n\to\infty} \left(1_{\mathfrak{K}} + \tfrac{1}{n} \odot p\right)^{-1} = 1_{\mathfrak{K}}, \tag{140}$$

und damit dann für jede Funktion f aus der Faltungsalgebra $\mathscr{R}$

$$\begin{aligned} \lim_{n\to\infty} \Phi(f * \delta_n) &= \lim_{n\to\infty} \Phi(f) \circledast \Phi(\delta_n) = \lim_{n\to\infty} \Phi(f) \circledast q_n = \Phi(f) \circledast \lim_{n\to\infty} q_n \\ &= \Phi(f) \circledast 1_{\mathfrak{K}} = \Phi(f). \end{aligned} \tag{141}$$

Also gilt in gewisser Weise $\lim_{n\to\infty} f * \delta_n = f$. Aber wohlgemerkt, für dieses Rechnen mit Grenzwerten ist über die oben rein algebraisch dargestellte Operatorenrechnung hinauszugehen. Im Rahmen der Laplace-Transformation kann etwas vergleichbares übriges nicht gemacht werden, da dort eine Funktion, die dem Einselement $1_{\mathfrak{K}}$ der Divisionsalgebra $\mathfrak{K}$ entspricht, nicht existiert. $\oplus\oplus$

Folgerungen:

Es spricht also vieles für den Einsatz des Operatorenkalküls anstelle der Laplace-Transformation! Wenn nur die menschliche Trägheit nicht wäre. Siehe hierzu auch die Bemerkungen bei FREUDENTHAL in [7] und bei Laugwitz in [8].

Man lasse sich auch nicht durch den Umfang der Lehrbücher zur Operatorenrechnung von ihrem Einsatz abschrecken. Denn in diesen Lehrbüchern geht es zum großen Teil (ca. 80 % und mehr) um die Verwendung der Operatorenrechnung bei der Untersuchung partieller Differentialgleichungen.

Und selbst wenn für eine gewisse Übergangszeit beide Begründungen für „das Arbeiten mit der Korrespondenzentafel" – der Operatorenkalkül nach Mikusiński und die Laplace-Transformation – parallel dargestellt werden müßten, so wäre dies meiner Meinung nach ein unschätzbarer Gewinn, da hierdurch deutlich werden würde, daß diese Rechenmethoden als Teile der mathematischen Modelle menschliche Erfindungen sind, die mehr oder weniger willkürlich ausgetauscht werden können, und keine kanonische Eigenschaften der physikalischen Realität.

Literatur

[1] HEAVISIDE, OLIVER, *Electromagnetic Theory,* London, 1893–1912.
Reprint: Chelsea Publishing Company, New York, 1971.

[2] DOETSCH, GUSTAV, *Handbuch der Laplace-Transformation, Band I: Theorie der Laplace-Transformation,* Birkhäuser Verlag, Basel und Stuttgart, 1950.

[3] MIKUSIŃSKI, JAN, *Operatorenrechnung,* VEB Deutscher Verlag der Wissenschaften, Berlin, 1957.

[4] DOETSCH, GUSTAV, *Anleitung zum praktischen Gebrauch der Laplace-Transformation,* Oldenbourg, München, 1961.

[5] ERDÉLYI, ARTHUR, *Operational Calculus and Generalized Functions,* Holt, Rinehart and Winston, New York, 1962.

[6] BERZ, EDGAR, *Verallgemeinerte Funktionen und Operatoren,* Bibliographisches Institut, Mannheim, 1967.

[7] FREUDENTHAL, HANS, *Operatorenrechnung – von Heaviside bis Mikusiński,* in Überblicke Mathematik, Band 2, herausgegeben von Detlef Laugwitz, Bibliographisches Institut, Mannheim und Zürich, 1969.

[8] LAUGWITZ, DETLEF, *Ein Jahrzehnt Operatorenrechnung,* in Überblicke Mathematik, Band 2, herausgegeben von Detlef Laugwitz, Bibliographisches Institut, Mannheim und Zürich, 1969.

[9] BERG, L., *Operatorenrechnung, I. Algebraische Methoden,* VEB Deutscher Verlag der Wissenschaften, Berlin, 1972.

[10] BERG, L., *Operatorenrechnung, II. Funktionentheoretische Methoden,* VEB Deutscher Verlag der Wissenschaften, Berlin, 1974.

„PosiCon Ball" – projektorientierter Unterricht in der Regelungstechnik anhand eines Beispiels

Wolfgang Werth
Fachhochschule Technikum Kärnten
Europastrasse 4, 9524 Villach
w.werth@cti.ac.at

Kurzfassung: *Wissen zu vermitteln und Studenten auf das Berufsleben vorzubereiten ist eine der Hauptaufgaben eines Lehrenden. Neben herkömmlichem Frontalunterricht und modernen e-Learning Methoden kann Lehrstoff auch anhand der Lösung von konkreten Aufgaben erworben werden. Dieser problemorientierte Lehransatz ist Gegenstand dieses Aufsatzes. Anhand eines Jahresprojektes, bei dem das Labormodell „PosiCon Ball" entwickelt und realisiert wurde, werden Erfahrungen, die mit dieser Lehrmethode gemacht wurden, dargestellt. Es sei erwähnt, dass das Hauptaugenmerk nicht auf die lückenlose Beschreibung des Modells und seine Regelung gelegt wurde. Es stehen vielmehr die Motivation, die Vorgehensweise, Schwierigkeiten und Erfolge im Mittelpunkt der Betrachtungen.*

1 Einleitung

Jemand, der heute im Bereich der Lehre, in welchem Fachbereich auch immer tätig ist, sieht sich ständig neuen Herausforderungen gestellt. Das reine „Vortragen" eines Lehrinhaltes aus einem Lehrbuch ist heute längst nicht mehr zeitgemäß. Verantwortlich für diesen Umstand ist zum einen die Tatsache, dass Studierende schon aus der Schule daran gewöhnt sind, dass das Erlernen eines Stoffgebietes leicht und unterhaltsam sein muss. Auf der anderen Seite bieten neue Medien, wie das Internet, jedem die Möglichkeit gerade im Selbststudium sich in unbekannte Wissensgebiete einzuarbeiten. Es ist daher einleuchtend, dass in Hinsicht einer veränderten Bildungslandschaft der Aufbereitung von Lehrinhalten große Bedeutung zukommt.

Selbstverständlich haben viele Lehrende auf diesen Umstand reagiert. Selbst in den Fällen, bei denen der Frontalvortrag nach wie vor als das geeignete Unterrichtsmittel zu sein scheint, sind Multimedia-Hilfsmittel intergaler Bestandteil der Lehreinheiten. Zusätzlich wird heute als ein Schritt in der Unterrichtsverbesserung oft viel Aufwand in ansprechende Unterrichtsunterlagen gesteckt.

Auf der anderen Seite ist in den letzten Jahren ein neuer Foschungsbereich entstanden, der der bestehenden Entwicklung Rechnung trägt, nämlich der Bereich „Einsatz neuer Medien im Unterricht" oder „e-Learning". Dieser Bereich versucht vor allem das Internet als Medium zur Verbreitung und Vermittlung von Wissen zu nutzen. E-Learing-Methoden zielen meist darauf ab, einer großen Anzahl von Studierenden Wissen eines bestimmten

Fachgebietes zu vermitteln. Lerneinheiten oder Übungen kann sich der Student dann ortsunabhängig, also auch zuhause, erarbeiten. Selbstverständlich ist dies speziell für berufstätige Studenten eine sinnvolle Methode. Dies erscheint heute vor allem deshalb möglich, da schnelle Internetanbindungen mittelweile auch im Privatbereich zu relativ günstigen Konditionen verfügbar sind. Auch im Bereich der Regelungstechnik gibt es zahlreiche Anwendungsmöglichkeiten, man denke nur an Vorbereitungsarbeiten zu anspruchsvollen Laborübungen.

Soll ein Unterrichtsgegenstand an eine realtiv kleine Gruppe von Studierenden – ca. 25 Personen – vermittelt werden, so bietet sich eine weitere sehr effektive Unterrichtsform an: das problemorientierte Lernen[1]. Anstelle der Konsumation einer großen Anzahl von herkömmlichen Lehrveranstaltungen, erarbeiten sich Studierende durch Lösen einer oder mehrerer konkreter Aufgaben die entsprechenden Kenntnisse und Fertigkeiten in einem bestimmten Fachgebiet. In unserem Fall ist die zu lösende Problemstellung durch eine Projektaufgabe gegeben. Es steht hierbei zur Bearbeitung einer solchen Aufgabe der Zeitraum eines Studienjahres zur Verfügung[2]. Aus diesem Grund sprechen wir vom so genannten „Projektjahr" im Rahmen des Studiums. Jede Gruppe von Studierenden des betreffenden Studienjahres verwendet dann etwa einen Tag pro Woche zur Abwicklung ihres Projektes.

1.1 Fahrplan

Um einen kleinen Einblick in die zeitliche Organisation unseres Projektjahres zu geben, sei ein grober Zeitplan erwähnt.

- Vorstellung der Projektthemen ca. 6 Monate von Projektstart
- Bewerbung und Zuteilung der Projekte an die Studierenden
- Start des Projektes im Oktober zu Beginn des Studienjahres
- Beabeitung der Konzeptionsphase des Projektes
- Zwischenpräsentation im Jänner
- Bearbeitung der Realisierungsphase
- Abschluss des Projektes durch eine Abschlusspräsentation

1.2 Allgemeine Ziele, was möchten wir erreichen?

Studierende, die mit dem Projektjahr beginnen, verfügen über Grundkenntnisse in den Bereichen Mathematik, Informatik, Systemtechnik und Elektrotechnik. Grundsätzliches Ziel des Unterrichtsfaches „Projektjahr" ist es daher, den Studierenden nach ihrer Grundlagenausbildung eine vertiefte Ausbildung in einem bestimmten Themengebiet zu geben,

[1] Problem based learning

[2] An der FH Kärnten im Studiengang Elektronik wird diese Unterrichtsmethode bereits zum dritten Mal (Jahr) erfolgreich angewendet.

beispielsweise in der Regelungstechnik. Gleichzeitig sollen die Studenten zum selbstständigen Arbeiten erzogen werden. Dies wird schon alleine dadurch erzwungen, indem zum Projektstart lediglich eine ungefähre Zielvorgabe (Spezifikation) gegeben ist. Die Bearbeitung eines Projektes im Vergleich zur „einfachen Lösung" einer Aufgabenstellung bietet auch die Möglichkeit weitere Lehrinhalte aufzunehmen. Neben den Gebieten Rhetorik und Präsentationstechniken – wie im Zeitplan erwähnt, ist die Präsentation der Ergebnisse ein wichtiger Bestandteil des Projektjahres – wird den Studierenden ein Einblick in die Thematik des Projektmanagements gegeben.

1.3 Abschließende Bemerkungen

Es hat sich gezeigt, dass vor allem Studierende eines technischen Fachbereiches zum Verstehen eines Lehrstoffes neben effektiver Betreuung vor allem eine praktische Umsetzung der theoretischen Erkenntnisse benötigen. Dies erklärt sich durch den relativ hohen Schwierigkeitsgrad des Lehrstoffes, dessen Aufnahme oftmals eine starke abstrake Denkweise erfordert. Es soll an dieser Stelle erwähnt werden, dass der projektorientierte Unterricht ohne Hinzunahme herkömmlicher Lehrveranstaltungen oder auch e-Learning-Methoden nicht realisierbar ist. Sowohl die begrenzte vorhandene Studierzeit der Studenten, als auch der endliche Betreuungsaufwand der Lehrkräfte sind als Ursachen dafür zu sehen. Die Auswahl der so genannten begleitenden Lehrveranstaltungen und vor allem die Abstimmung der zugehörigen Lehrinhalte ist allerdings sehr stark an das jeweilige Projekt gekoppelt.

Zusammenfassend kann gesagt werden, dass das sogenannte Projektjahr von der überwiegenden Anzahl der Studierenden als ein großer und wichtiger Ausbildungsschritt gesehen wird.

2 PosiCon Ball, das RT-Projekt

2.1 Die Idee und die Ziele

Überlegungen, die zu einem Projekt aus dem Bereich der Regelungstechnik geführt haben, lassen sich nun folgendermaßen zusammenfassen:

1. Einer Gruppe von Studenten soll regelungstechnisches Grundverständnis und natürlich einige Entwurfs- und Analyseverfahren der Regelungstechnik nähergebracht werden. Dies soll dadurch erreicht werden, dass ein „Gerät" von der Idee, über die Planung bis zur technischen Realisierung entwickelt wird. Selbstverständlich sind im Zuge der Realisierung des Projektes auch eine Reihe von unterschiedlichen Aufgaben zu lösen, die nicht unmittelbar der Regelungstechnik zuzuordnen sind.

2. Als Ergebnis soll ein regelungstechnischer Modellversuch zur Verfügung stehen, der sowohl zu Demonstrationszwecken, als auch als Laborversuch für nachfolgende Studenten verwendet werden kann.

Aus den obigen Überlegungen ergaben sich eine Reihe von Anforderungen an den zukünftigen Laborversuch:

- Prinzipielle Machbarkeit
- Sinnvolle Anwendbarkeit mehrerer Reglerentwurfsverfahren
- Gute Möglichkeit zur Erklärung regelungstechnischer Prinzipien
- Robustheit

Basierend auf obigen Überlegungen entstand das Projekt „PosiCon Ball"[3].

2.2 Spezifikation

Auf der Basis der erwähnten Grundanforderungen wurde eine entsprechende Aufgabe und daraus die erste technische Spezifikation formuliert.

Ein einseitig drehbar gelagerter Balken kann in seiner Neigung zur Horizontalen verändert werden. Dies soll über ein entsprechendes Getriebe mit Hilfe eines DC-Motors erfolgen. Auf dem Balken befindet sich eine Kugel, deren Position messtechnisch zugänglich sein muss. Über die Neigung des Balkens kann die Position der Kugel, die sich entlang des Balkens bewegen kann, verändert werden.

Ziel der Anordnung ist es, die Position der Kugel entlang des Balkens beliebig vorgeben zu können. Damit steht auch die zugehörige Regelungsaufgabe fest, nämlich eine geeignete digitale Positionsregelung der Kugel zu entwickeln. Der Sollwertverlauf soll dabei wahlweise durch den verwendeten Prozessrechner als auch durch eine externe Anordnung vorgegeben werden können. Zusätzlich sollen zwei unterschiedliche Konzepte der Reglerimplementierung realisiert werden. Zum einen wird eine PC-unabhängige Lösung mithilfe eines Mikrocontrollers gefordert. Andererseits soll das Programm LabVIEW zur Prozessregelung herangezogen werden. In der nachfolgenden Abbildung (Bild 1) ist nun die gewünschte prinzipielle Anordnung zu sehen.

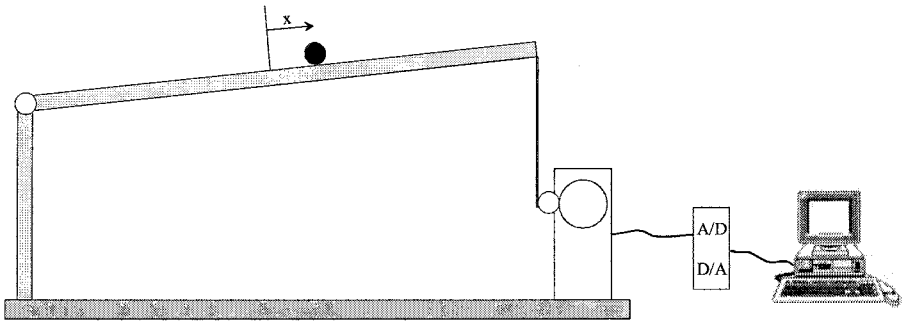


Bild 1: Prinzip der Anordnung (Vorgabe)

Aus dieser Aufgabenstellung ergeben sich unmittelbar die wesentlichen Teilaufgaben des Projektes:

- Entwicklung eines Realisierungskonzeptes mit Alternativen

[3]Der Projektname wurde als Abkürzung der Aufgabe „**Posi**tion **Con**trolled **Ball**" gewählt.

- Fertigung des mechanischen Aufbaus
- Auswahl einer geeigneten Antriebseinheit
- Entwicklung eines Leistungsteiles
- Auswahl der benötigten Sensorik
- Entwicklung einer Mikrocontrollereinheit
- Anbindung an die LabVIEW-Umgebung
- Entwicklung und Implementierung eines Reglerkonzeptes

Ausgehend von den oben formulierten Wünschen, hat eine Gruppe von 5 Studenten diese Aufgabe in Form eines Jahresprojektes gelöst. Wenn man die oben genannten wichtigsten Teilaufgaben betrachtet, ergeben sich ganz geradlinig jene konventionellen Lehrveranstaltungen, die das Projekt begleiten sollen: Sensorik, Leistungselektronik, vertiefende Mikrocontrollerprogrammierung. Es ist leicht einzusehen, dass dadurch eine vollständige Abdeckung aller benötigten Grundlagen nicht gewährleistet ist. Beispielsweise fehlt eine Ausbildung in Fächern der Mechanik, bedingt durch die Ausrichtung des Studienganges[4], nahezu vollständig. Damit ist die Notwendigkeit, Stoffgebiete direkt beim Lösen gewisser Teilaufgaben zu erlernen als eine Kernidee, nämlich „learning by doing", beim problemorientierten Unterricht, klar ersichtlich. Darin ist jedoch kein Nachteil der Methode, sondern ein arbeitsintensiver Aspekt zur realitätsnahen Vorbereitung auf das Berufsleben zu sehen.

3 Projektabwicklung

Entsprechend dem Vorgehen bei einer Projektabwicklung im Allgemeinen wurde der Arbeitszeitraum in bestimmte Phasen unterteilt. Ausgehend von der Aufgabenstellung wurde in der Definitionsphase eine technische Spezifikation des Gerätes entwickelt. Wesentlichen Teil des Projektes stellt dann die so genannte Alternativensuche dar. Dieser für die Studierenden besonders schwierige Teil soll die Entscheidungsgrundlage für die Auswahl der Komponenten der technischen Realisierung des Gerätes sein. Darunter fallen etwa Entscheidungen einer Detailaufgabe, wie die Auswahl eines geeigneten Sensors zur Bestimmung der Kugelposition, ebenso wie der prinzipielle Lösungsweg. Ausgehend von einem Konzept bildet die Realisierungsphase den dritten Teil des Projektes.

3.1 Von den ersten Schritten zum Konzept

Keines der Projektmitglieder hatte Erfahrungen in der Abwicklung eines Projektes dieser Größe. Im Rahmen von einzelnen Lehrveranstaltungen wurden lediglich kleinere Hausübungsprojekte durchgeführt. Den Umgang mit Messgeräten und Grundkenntnisse

[4]Elektronik

im elektronischen Gerätebau konnten die Studenten durch eine Reihe von Laborübungen erlernen. Mit welchen ersten Aufgaben und Schwierigkeiten waren daher die Studierenden zunächst konfrontiert?

- Die Suche nach Produkten, die es am Markt bereits gibt (Komplettlösungen, Sensoren, etc.)
- Beschaffung von Informationen (Wer hat schon Erfahrungen?)
- Erlernen von unbekannten Stoffgebieten (Mechanik, etc.)
- Aufteilen in Teilaufgaben und Festlegen der Verantwortungsbereiche
- Schaffung eines Kostenbewusstseins
- Erstellung einer Dokumentation

Nach einer intensiven Einarbeitungsphase entstand als Ergebnis diese Skizze (Bild 2) eines ersten Entwurfs des Labormodells. Erwähnenswert dabei ist, dass die Art des verwendeten Hubkonzepts, die Motor-Getriebe-Einheit, ebenso wie die geometrische Dimensionierung darin bereits weitgehend festgelegt wurden.

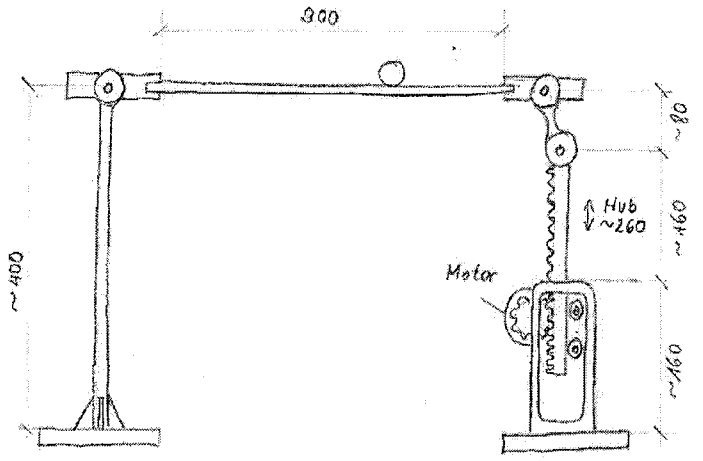


Bild 2: Skizze der ersten Idee

Neben Alternativen zum Antriebskonzept wurden vor allem verschiedene Möglichkeiten zur Messung der Kugelposition untersucht. Als robuste und vor allem kostengünstigste Lösung wurde eine Widerstandsfolie ausgewählt. Nach den zu lösenden Teilaufgaben wurden auch die einzelnen Arbeitsbereiche aufgeteilt: mechanische Konstruktion, Antriebs- und Leistungseinheit, Mikrocontrollerkonzept, LabVIEW-Konzept, Reglerentwurf. Jeder der Studenten hatte die Aufgabe, für seinen Bereich benötigtes „Know-how" zu erwerben und vor allem an die Gruppe weiterzugeben. Erwähnenswert ist hierbei, dass die Begeisterung der Studierenden erheblich größer und der Lernerfolg deutlich besser ist, als dies bei herkömmlichem Unterricht der Fall wäre. Es ist leicht verständlich, dass dabei der

Dokumentation der bewältigten Arbeitspakete sehr große Bedeutung zukommt. Dies ist für den Informationsaustausch in der Gruppe einerseits, für eine Nachvollziehbarkeit der einzelnen Schritte andererseits, unbedingt erforderlich. Als Ergebnis der ersten Monate konnte ein vollständiges Konzept erarbeitet werden. Als ein Teilergebnis gibt der zugehörige Strukturplan (Bild 3) Einblick in die Arbeitsbereiche der Projektmitglieder:

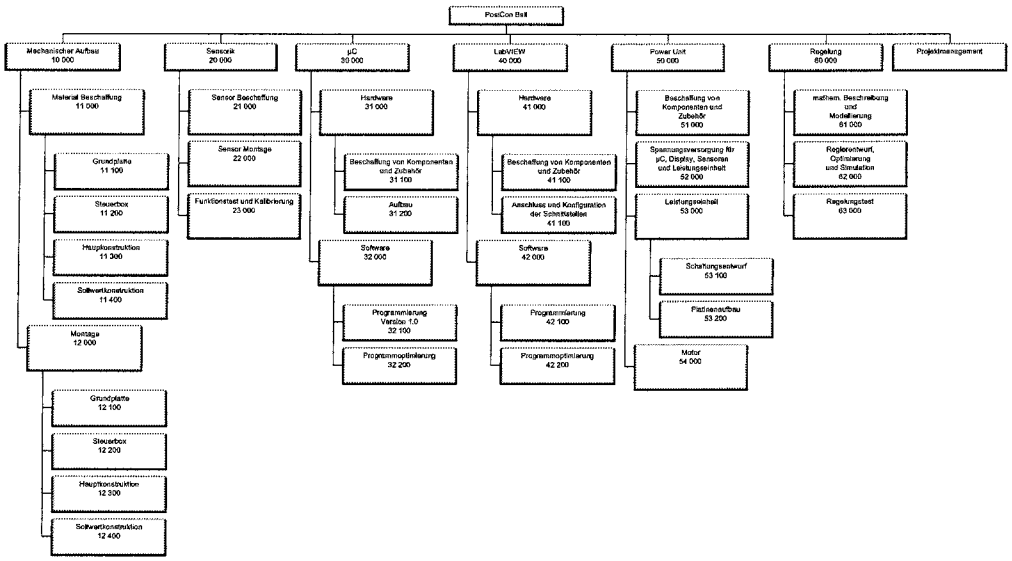


Bild 3: Projektstrukturplan

Einen wichtigen Teil der technischen Spezifikation stellen zwei verschiedene Varianten der Implementierung des Regelungskonzeptes dar. Zum einen war eine so genannte „Stand- alone"-Lösung gefordert. Die gesamte Regel- und Steuereinheit wird hierbei von einem Mikrocontroller, dem C167 der Firma Infineon, übernommen. Damit sollte eine schnelle und einfach bedienbare Demonstration des Laborversuchs möglich sein. Die Bedienung des Labormodells von einer graphischen Benutzeroberfläche aus bietet eine Implementierungslösung mit Hilfe von LabVIEW. Zum Ein- und Auslesen der benötigten Daten wird eine DAQ-Karte PCI-6024E der Firma National Instruments verwendet. Das umschaltbare Gesamtkonzept ist als Blockschaltbild in der Abbildung 4 zu sehen.

3.2 Begleitung und Hilfe

Wenn man die umfangreichen Aufgaben des Projektes und den Ausbildungsstand der Studenten (5.Semester) betrachtet, ist es leicht einzusehen, dass ein erfolgreicher Projektabschluss die intensive Betreuung der Studentengruppe erforderte. Die Unterstützung läßt sich prinzipiell in 3 Gruppen unterteilen. Neben den begleitenden Lehrveranstaltungen – hier sind die Lehrenden natürlich direkt Ansprechpartner für Problemstellungen aus dem konkreten Fachgebiet – kommt den „Lehrveranstaltungen" Projektmanagement und Präsentationstechniken sehr große Bedeutung zu. Diese sind vollständig auf die zu bearbeitenden Projekte abgestimmt und begleiten die Studenten während des ganzen Jahres.

Die dritte unterstützende Gruppe bildet der sogenannte Projektsupervisor als übergeordneter Leiter der Projektgruppe und die Mitarbeiter des Technikums als Ansprechpartner für Probleme aus den jeweiligen Arbeitsgebieten. Prinzipiell wurde hierbei der Informationsaustausch und die Unterstützung durch wöchentliche Besprechungen und aber auch durch kurzfristige Fragestunden organisiert. Gelegentlich wurde auch Hilfe von „außen" herangezogen, etwa durch Fachleute entsprechender Partnerfirmen.

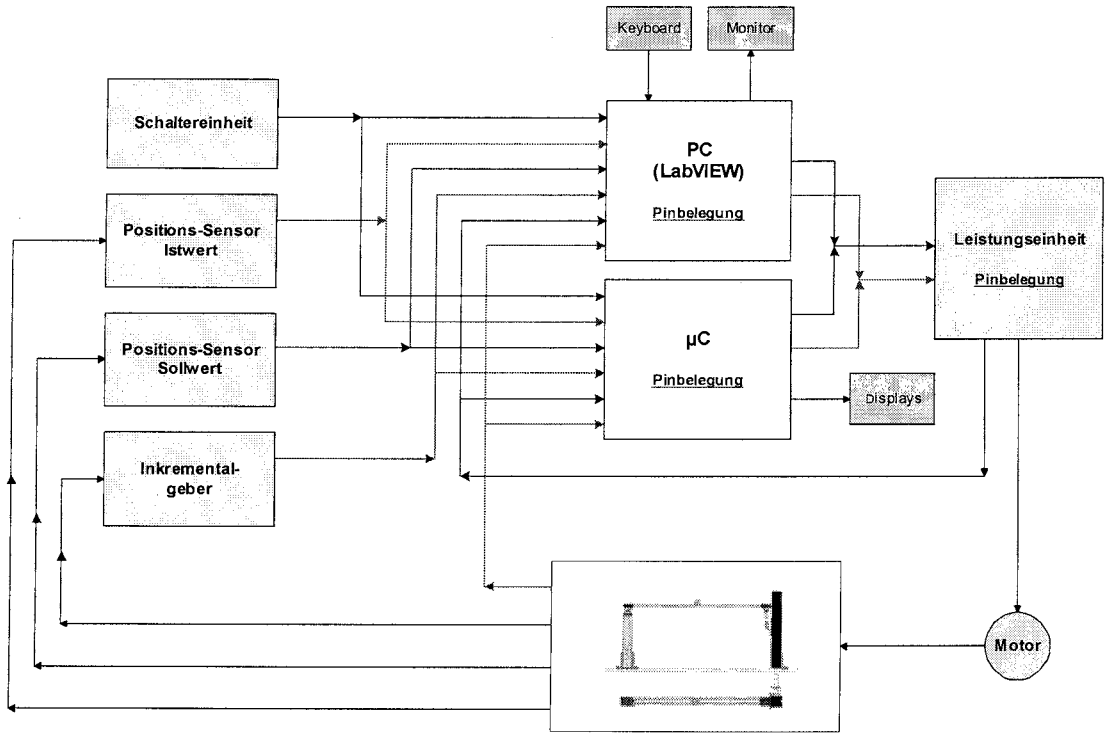


Bild 4: Blockschaltbild des Gesamtkonzeptes

3.3 Mathematische Modellierung und Reglerkonzepte

Neben der technischen Realisierung des Aufbaus, der in der folgenden Abbildung (Bild 5) als Zeichnung mit den wesentlichen Komponenten zu sehen ist, mußte zum Entwurf eines geeigneten Reglers eine mathematische Modellbildung durchgeführt werden.

Das Auffinden einer mathematischen Beschreibung des Labormodells läßt sich prinzipiell in zwei Abschnitte unterteilen. Zum einen waren physikalische Überlegungen zur Dimensionierung der elektrischen und mechanischen Komponenten erforderlich. Beispielhaft seien zwei Ergebnisse angeführt. Die maximal auftretende Beschleunigung der Kugel $a_{\max} = 1.2\ ms^{-2}$ bei einem Fahrweg von $s = 0.9\ m$ führte auf eine benötigte maximale Beschleunigung $a_{vert} = 5\ ms^{-2}$ der vertikalen Bewegung des Linearstellers. Die berechneten Werte wurden durch Versuchsreihen bestätigt. Hierbei ist zu sehen, wie beispielsweise mechanische Grundgesetze erlernt werden und unmittelbar am realen Modell Verwendung finden. Als Beispiel für Überlegungen aus dem elektrischen Bereich könnte der

erforderliche Haltestrom des Motors bei Mittelstellung des Balkens $i_{Halte} = 0.88\ A$ dienen. Die Größe i_{Halte} ist für die Auslegung des Leistungsteiles und die Auswahl des Motors ein wichtiger Parameter.

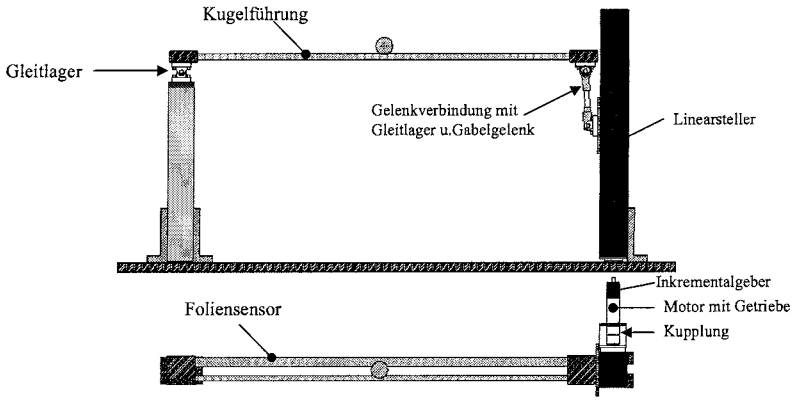


Bild 5: Zeichnung des realisierten Konzepts

Für einen Reglerentwurf ist jedoch eine Modellbildung durch Aufstellen entsprechender Differentialgleichungen unerlässlich. Bezeichnet man die Kugelposition mit x, so kann ein Zusammenhang in Form einer Differentialgleichung zwischen der Position und dem Winkel des Balkens zur Horizontalen α angegeben werden:

$$\ddot{x} = -\frac{5}{7} g \sin \alpha$$

Zwischen dem Hub des Linearstellers y, dem Winkel α und der tatsächlichen Balkenhöhe am rechten Ende h besteht folgender geometrischer Zusammenhang:

$$h = y + l_1 \cos\left[\arcsin\left(\frac{L - l\cos\alpha}{l_1}\right)\right] - l_1 \cos\left[\arcsin\left(\frac{L-l}{l_1}\right)\right]$$

Zusätzlich sind die aus der Standardliteratur [4] bekannten Differentialgleichungen für die Beschreibung der Motorbewegung – Zusammenhang zwischen Motorwinkel φ und Motorspannung u – verwendet worden. Dieses Differentialgleichungssystem stellt den Ausgangspunkt für die benützten Reglerentwurfskonzepte dar.

Wichtig aus der Sicht des Unterrichts ist folgender Punkt: Aus der Beschäftigung mit dem konkreten Modell und in weiterer Folge mit der Analyse der erhaltenen Gleichungen haben die Studierenden eine Reihe von Erkenntnissen aus dem Bereich der Regelungstechnik unmittelbar – anstelle oder als Ergänzung zu „Vorlesungserklärungen" – erworben:

- Erkennen des Nutzens in der Beschäftigung mit Differentialgleichungen (Mathematik).

- Auswirkung von Instabilitäten.
- Das Auftreten von Nichtlinearitäten.
- Sinnvolle oder notwendige Vernachlässigungen von manchen Größen (Idealisierungen).
- Das Zusammenschreiben der aufgestellten Gleichungen führt automatisch auf ein Zustandsraumkonzept.

Der Verständlichkeit halber seien folgende bereits vereinfachte Gleichungen dargestellt, die als mathematisches Modell für die weiteren Reglerentwürfe verwendet wurden:

$$\begin{aligned} \dot{x} &= v \\ \dot{v} &= -\frac{5}{7} g \frac{r}{L} \varphi \\ \dot{\varphi} &= \omega \\ \dot{\omega} &= -\frac{k_m^2}{R_m J} \omega + \frac{k_m}{k_g R_m J} u - \frac{1}{J} M_L \end{aligned}$$

Im Rahmen des Projektes wurden nur Reglerentwurfskonzepte für lineare Systeme verwendet.

Ziel des Regelungskonzeptes ist es, eine Führungsregelung für die Positionierung der Kugel zu realisieren. Im Besonderen hat sich das Konzept des Zustandsreglers als eine geeignete Lösung ergeben. Bekannterweise läßt sich ein solches Regelgesetz in folgender diskreter Form angeben:

$$u_k = \mathbf{k}^T \mathbf{x}_k + V r_k$$

Als Messgrößen stehen die Position der Kugel x und der Motorwinkel φ zur Verfügung, die benötitgen Ableitungen werden auf numerischem Wege ermittelt. Zur Überprüfung des Regelverhaltens wurde das Programm Simulink verwendet. Dies ist vor allem deswegen sinnvoll, da der Einfluß von Beschränkungen[5], die am realen Modell auftreten, jedoch im mathematischen Modell nicht berücksichtigt werden, in einfacher Weise untersucht werden kann.

Im Unterschied zu herkömmlichen Unterrichtsmethoden ist auch bei dieser Teilaufgabe zu erwähnen, dass Fragestellungen direkt aus der Beschäftigung mit dem Labormodell entstehen. Die Motivation der Studierenden theoretische Methoden zu erlernen, die zur Beantwortung der Fragen nötig sind, ist erheblich größer. Als Beispiel könnte man das Auftreten einer bleibenden Regelabweichung bei manchen Reglerkonzepten anführen. So zeigt sich das Verwenden eines Integrierers im offenen Regelkreis bei diesem Modell als ein lehrreicher Aspekt.

3.4 Die Implemetierung, μC und LabVIEW

In der folgenden Abbildung (Bild 6) ist das realisierte Labormodell zu sehen. Neben der Hauptkonstruktion (im Hintergrund) ist die Konstruktion zur Vorgabe der Sollposition und die Mikrocontrollereinheit zu erkennen.

[5] So ist beispielsweise der Fahrweg des Linearstellers auf $\pm 15cm$ beschränkt.

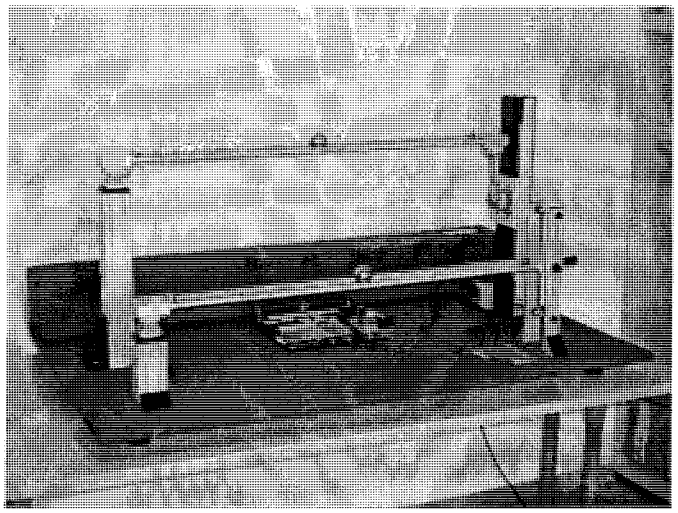


Bild 6: Labormodell „PosiCon Ball"

Entsprechend der Spezifikation wurde eine Implementierung mit Hilfe des Mikrocontrollers C167 realisiert. Die Studierenden hatten bereits Erfahrung im Umgang mit diesem Mikrocontroller aus den Grundvorlesungen. Dennoch zeigte sich, dass erst das Lösen einer konkreten Aufgabenstellung ohne Lehrbuchcharakter für das Beherrschen der Programmiertechnik unerlässlich ist. Mit diesem Controller ist zudem eine Ansteuerung des Motors über ein PWM-Signal einfach zu realisieren.

Als Alternative wurde eine Reglerimplementierung mit Hilfe der LabVIEW-Umgebung entwickelt. Die zugehörige Bedienoberfläche ist in der folgenden Abbildung (Bild 7) zu sehen.

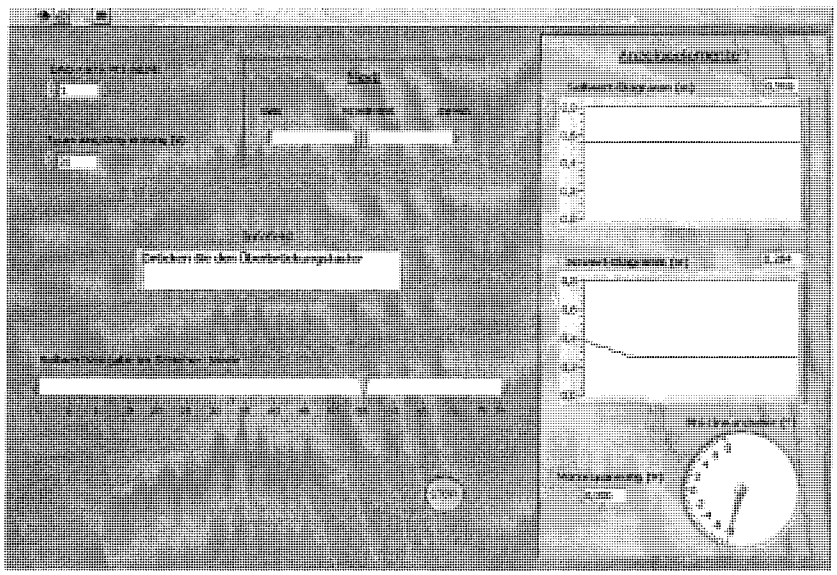


Bild 7: LabVIEW Bedienoberfläche

Erwähnenswert ist hierbei, dass die Entwicklung der LabVIEW-Regelung ohne zusätzliche Einschulungen von den Studenten durchgeführt wurde. In der folgenden Abbildung (Bild 8) sind die gemessenen Verläufe (Kugelposition x, Balkenwinkel α) eines typischen Versuchs dargestellt. Hierbei wurde der Sollwert sprungartig von 0 auf 78 cm geändert.

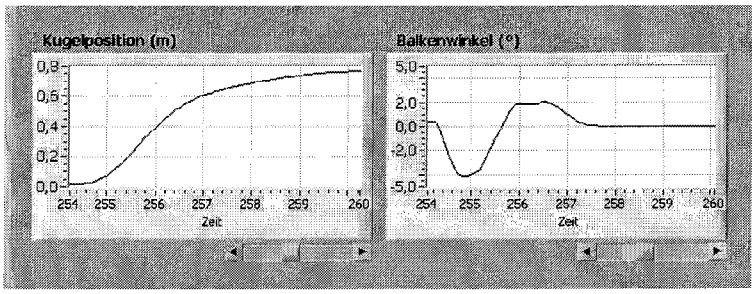


Bild 8: Messergebnisse

4 Diskussion der Methode, Ausblick

Wenn man versucht, Methoden des problemorientierten Unterrichts mit den herkömmlichen Lehrmethoden zu vergleichen, so kann man Folgendes zusammenfassen:

Aus der Sicht der Betreuer ist bei der beschriebenen Variante der Aufwand in der benötigten Unterstützung wesentlich höher anzusetzen. Dies betrifft sowohl den rein zeitlichen Aufwand als auch die Bereitschaft, sich in angrenzende Fachgebiete einzuarbeiten. Bei allen Unterrichtsprojekten dieser Art sind zwar Ziele definiert, aber Detaillösungen auch den Betreuern oftmals unbekannt.

Studenten bewerten diese Form des Unterrichts zum überwiegenden Teil sehr positiv. Selbstverständlich ist der benötigte Einsatz der Studierenden ebenfalls weit höher, als dies zum Erlernen von „Rezepten" für die Lösung von entsprechenden Problemen nötig wäre. Es sollte allerdings nicht unerwähnt bleiben, dass diese Form des Unterrichts eine begrenzte Anzahl von Studierenden voraussetzt.

Auch im nächsten Studienjahr werden eine Reihe von Studenten ihre Vorbereitung auf das Berufsleben in dieser Form des Unterrichts erhalten.

Literatur

[1] Grabner K., Groemansberger R., Gspandl P., Krainer D., Krainer F.: Projekthandbuch, „PosiCon Ball", Fachhochschule Technikum Kärnten, 2003

[2] Jamal R., Hagestedt A.: LabVIEW, das Grundlagenbuch, 3. Auflage, Addison-Wesely, 2001

[3] Infineon Technologies: C167CR Derivatives 16-Bit CMOS Microcontroller, User's Manual, V3.1, März 2000

[4] Chen C.T.: Control System Design, Stony Brook: Sauders College Publishing, 1993

Torque Control of an Induction Machine Based on Dynamic Equilibrium

Bojan Grčar, Peter Cafuta, Gorazd Štumberger
University of Maribor
Faculty of Electrical Engineering and Computer Science
Smetanova 17, 2000 Maribor
Slovenia
e-mail: bojan.grcar@uni-mb.si

Abstract: The paper proposes an IM torque control based on the stator current vector reference frame. The control structure is derived from the augmented dynamic equilibrium conditions assuring the existence and uniqueness of the inverse mapping between the required torque end control inputs. By manipulating simultaneously the magnitude and the rotation speed of the stator current vector in an appropriate way the rotor flux trajectories are forced to change only inside the stable sector under all operating conditions. Additional features as maximal torque per ampere steady state ratio, almost perfect command tracking at non-zero rotor flux and insensitivity to the bounded perturbation of the rotor time constant, characterize introduced control. Implementation of this control requires estimation of the machine torque and cascaded current controllers. Experimental results confirm the key expectations and show the potential and benefits of the proposed control.

Keywords: Induction machine, Non-linear transformations, Torque control, Dynamic equilibrium, Efficiency maximization

List of symbols

$i_{\alpha,\beta}, u_{\alpha,\beta}, \psi_{\alpha,\beta}$	components of stator current, stator voltage and rotor flux linkage vectors in stationary reference frame
$i_{d,q}, u_{d,q}, \psi_{d,q}$	components of stator current, stator voltage and rotor flux linkage vectors in synchronous reference frame
$\|i_s\|$	stator current vector magnitude
ω_m	rotor speed
ω_i	stator current vector rotation speed
$\omega_r = \omega_i - \omega_m$	relative rotation speed
θ_i	stator current vector angular position
M	mutual inductance
R_r, L_r	rotor resistance and self-inductance
$\tau_r = L_r/R_r$	rotor time constant
R_s, L_s	stator resistance and self-inductance

$L_\sigma = L_s - (M)^2/L_r$	
$R_\sigma = R_s + (M/L_r)^2 R_r$	
$\tau_\sigma = L_\sigma / R_\sigma$	
T_{el}, T_L	electrical torque, load torque
J	drive inertia
$\mathrm{J} = [0, -1; 1, 0]$	2×2 rotation matrix
k_p, k_i	torque controller parameters
γ	torque observer gain
v, v_1, v_2	outputs from integral control action
$k_{pi}, k_{ii}, k_{i\omega}$	current controllers parameters
$\star, \wedge$	reference values, estimated values
$\sim$	errors
η, ζ, ξ	compensating signals

1 Introduction

An induction machine is used as the "work horse" in the majority of industries demanding speed or torque controlled drives. Control design which satisfies an increasing number of different performance objectives is becoming a more and more important and challenging issue. Although the research community have proposed several more or less sophisticated control concepts, only two major schemes have, in fact, been accepted by the industry: the established field-oriented control (FOC) and the more recent direct torque control (DTC).

The popularity and wide applicability of FOC arises from its relatively simple, cascaded and decoupled (rotor flux linkage versus torque) structure [2]. Resulting from the powerful system theory (nonlinear transformations and feedback linearization [4],[11],[5], passivity [3], and flatness [9]) a deeper understanding of IM dynamics, formal stability proofs, along with the guidelines for controller parameter settings is now also available. The main characteristic of FOC, namely preserving the magnitude of the rotor flux linkage vector constant, enables a globally stable solution for the non-holonomic (double) integrator problem, describing in generally IM rotor dynamics [10]. As the mentioned concept requires implicit system inversion (mapping of the reference current vector $i_{d,q}$ into voltage vector $u_{d,q}$) the inner current loop is needed. An outer loop for speed and/or torque control is usually designed for the reduced model, assuming that the high gain current controllers attain perfect tracking of the current command. Sensitivity to the parameter variations is a unavoidable consequence as the flux linkage control is usually open-loop (IFOC). Introduction of rotor flux linkage observers (DFOC), however,

only transfers the problems from the control design into the observer design and additionally increases the complexity of the feedback system. No matter how the rotor flux angle estimate is obtained it must be known to perform the necessary coordinate transformations.

On other hand, DTC [6], [7], introduces a somewhat different control philosophy and is based on stator flux linkage. This concept is voltage-based and operates without current controllers [8], [12]. The authors claim that by introducing stator quantities into the torque control, substantial reduction in parameter sensitivity could be achieved. Control signal (stator voltage vector) is generated in accordance with the finite number of possible VSI states, resulting in a complex nonlinear hybrid feedback system. The switching logic can be expressed in the form of algebraic inequalities to enable stability analysis and determination of the feasible operation areas. Some additional performance criteria (minimal losses, best tracking,...) could be simultaneously satisfied by selecting different switching patterns.

In the paper we introduce a control concept which could not be placed in any of these categories, although a similar cascaded structure as in the FOC is adopted along with vector rotation concept resulting from DTC. The design is based on an IM model in a reference frame aligned with the stator current as presented in [13]. Such a choice is motivated by the following:

- as the stator current vector is measured corresponding coordinate transformation turns out to be a more reliable one and
- since rotation of the current vector is essential for torque generation, transient and steady state relations between the rotor field, current and the generated torque at different operating regimes are more obvious in the suggested rather than in other reference frames.

Further important characteristics of the proposed scheme are that no decoupling is gained and that rotor flux linkage magnitude and angle are simultaneously changed in accordance with the reference torque command. This strategy represents a crucial difference compared with FOC where the rotor flux magnitude only changes during the "field weakening mode". In our opinion forcing the machine to operate in such a "decoupled" mode, motivated only by the goal of simplifying the control problem, is somehow artificial and unnatural and must, therefore, be compensated with higher losses and energy consumption. In contrast, we propose that current vector magnitude and its relative rotation speed are manipulated in such a way that, for any feasible steady state condition, maximal torque per ampere ratio is achieved. This introduced concept results in a single input-two output system instead of the two input-two output system

common in other schemes. During transients the rotor flux is forced to change only inside the absolutely stable sector even in the case of the perturbed rotor time constant. This important property is achieved by introducing hard and variable bounds on the torque controller outputs. No singularities restrict the operation at zero torque, zero mechanical speed or even during active breaking. Assuming that the signal conditioning and estimation of the actual torque are solved adequately almost proportional torque responses could be obtained for the smooth or even step references. The problem of "field weakening" is solved simultaneously for all feasible steady state conditions since the machine operates with a minimal rotor flux magnitude needed to generate the required torque. The upper bounds of reachable torque and mechanical speed are given by the current and voltage restraints. Extension towards speed control could be achieved with an additional PI control loop. Proposed feedback structure is relatively simple, easy to implement on the standard hardware and is mostly based on physical rather than demanding control considerations.

The organization of the paper is as follows. First the IM model in fixed $\alpha - \beta$ reference frame is transformed into the reference frame aligned with the stator current vector by introducing a non-linear state transformation. A reduced model derived from the assumption of high gain current controllers is introduced next. A basic torque controller along with the key restrictions on states and control inputs, based on the resulting apparently simple second order non-liner system, is introduced next. Complex rotor flux linkage vector dynamics and, steady state operation points, as well as stability conditions, are explained for the nominal and perturbed rotor time constant. Next the inner current loop is presented where the mechanical speed is also considered as bounded time-variant disturbance. Essential modifications of the current controllers are introduced since the required mapping of the reference stator current vector into the machine voltage vector in polar coordinates is quite a nontrivial task. The final part of the paper includes different experimental results presenting the potential and performance characteristics of the proposed IM torque control.

2 IM model in stator current reference frame

Under standard modeling assumptions [1] for linear magnetics, and by choosing the stator current vector and the rotor flux linkage vector as the state variables, an $\alpha - \beta$ model in a fixed

reference frame for the two pole machine is obtained in the form of:

$$\begin{bmatrix} \dot{i}_{\alpha\beta} \\ \dot{\psi}_{\alpha\beta} \\ \dot{\omega}_m \end{bmatrix} = \begin{bmatrix} -\tau_\sigma^{-1} i_{\alpha\beta} + \frac{M}{L_\sigma L_r \tau_r}(\mathrm{J}^T \omega_m \tau_r + 1)\psi_{\alpha\beta} + \frac{1}{L_\sigma} u_{\alpha\beta} \\ -\tau_r^{-1}(\mathrm{J}^T \omega_m \tau_r + 1)\psi_{\alpha\beta} + \tau_r^{-1} M i_{\alpha\beta} \\ \frac{M}{JL_r} i_{\alpha\beta}^T \mathrm{J} \psi_{\alpha\beta} - T_L/J \end{bmatrix} \tag{1}$$

Introducing the new set of state variables $z = [\|i_s\|, \theta_i, \psi_d, \psi_q, \omega_m]^T$ along with the nonlinear state transformation:

$$z = T(x) = \begin{bmatrix} \sqrt{i_\alpha^2 + i_\beta^2} \\ \tan^{-1}(\frac{i_\beta}{i_\alpha}) \\ \cos(\theta_i)\psi_\alpha + \sin(\theta_i)\psi_\beta \\ -\sin(\theta_i)\psi_\alpha + \cos(\theta_i)\psi_\beta \\ \omega_m \end{bmatrix} \tag{2}$$

and with the new input vector:

$$\begin{bmatrix} u_d \\ u_q \end{bmatrix} = \begin{bmatrix} +\cos(\theta_i)u_\alpha + \sin(\theta_i)u_\beta \\ -\sin(\theta_i)u_\alpha + \cos(\theta_i)u_\beta \end{bmatrix} \tag{3}$$

transformed model in the stator current vector reference frame is obtained in the form of:

$$\begin{bmatrix} \|\dot{i}_s\| \\ \dot{\theta}_i \\ \dot{\psi}_d \\ \dot{\psi}_q \\ \dot{\omega}_m \end{bmatrix} = \begin{bmatrix} -\tau_\sigma^{-1}\|i_s\| + \frac{M}{L_\sigma L_r \tau_r}\psi_d + \omega_m \frac{M}{L_\sigma L_r}\psi_q + \frac{1}{L_\sigma} u_d \\ -\omega_m \frac{M}{L_\sigma L_r}\frac{\psi_d}{\|i_s\|} + \frac{M}{L_\sigma L_r \tau_r}\frac{\psi_q}{\|i_s\|} + \frac{1}{L_\sigma \|i_s\|} u_q \\ -\tau_r^{-1}\psi_d + (\omega_i - \omega_m)\psi_q + M\tau_r^{-1}\|i_s\| \\ -\tau_r^{-1}\psi_q - (\omega_i - \omega_m)\psi_d \\ -\frac{1}{J}(\frac{M}{L_r}\psi_q\|i_s\| - T_L) \end{bmatrix} \tag{4}$$

Assuming that, with some appropriate current controllers, the tracking problem of $\|i_s\|^\star, \omega_i^\star$ can be "perfectly" solved, the reduced second order IM model representing the rotor dynamics is derived as:

$$\begin{bmatrix} \dot{\psi}_d \\ \dot{\psi}_q \end{bmatrix} = \begin{bmatrix} -\tau_r^{-1}\psi_d + \omega_r \psi_q + M\tau_r^{-1}\|i_s\| \\ -\tau_r^{-1}\psi_q - \omega_r \psi_d \end{bmatrix} \tag{5}$$

along with the algebraic output equation for the generated electrical torque:

$$T_{el} = -\frac{M}{L_r}\psi_q\|i_s\| \tag{6}$$

In (5) we introduced relative speed as the difference between the rotation speed of the current vector and the rotor ($\omega_r = \omega_i - \omega_m$). Consequently, the rotor speed is simply interpreted as a timc varying parameter, the dynamics of the mechanical part could therefore be omitted in

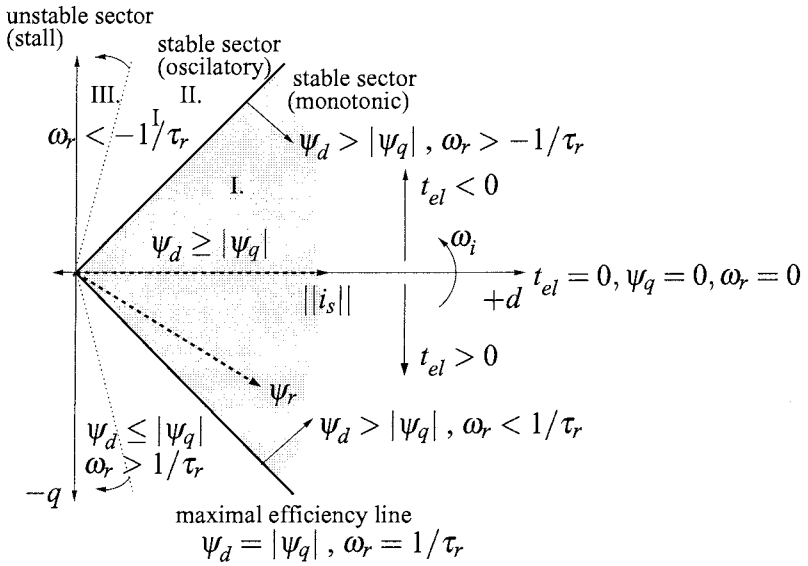


Figure 1: Admissible operating sectors

further analysis. For the simplicity reasons the assumption $\omega_m \geq 0$ will be considered through further analysis.

Considering the reduced model (5), along with the output equation (6) it can be seen that unique inverse mapping $T_{el} \rightarrow ||i_s||, \omega_i$ does not exists. Additional conditions must be therefore introduced to obtain required equilibrium. Manipulating only the first control input $||i_s||$, assuming that $\omega_r = 0$, only the rotor flux linkage component (ψ_d) will be changed. The second input ω_r must therefore also be changed in order to generate the required electrical torque. Since current vector rotation is required for all transient and steady-states (except at $T_{el} = 0$ and $\omega_m = 0$) corresponding dynamic equilibrium must be found. Different control strategies could, consequently, be employed in the torque generation. Keeping the current vector magnitude constant at some nominal value similar design to FOC could easily be introduced. On the other hand, if we keep the second input ω_r constant, only poor dynamics could be achieved. Yet the real control challenge lies in the question how to manipulate both control inputs simultaneously to achieve torque command tracking, global stability as well the maximal steady-state efficiency.

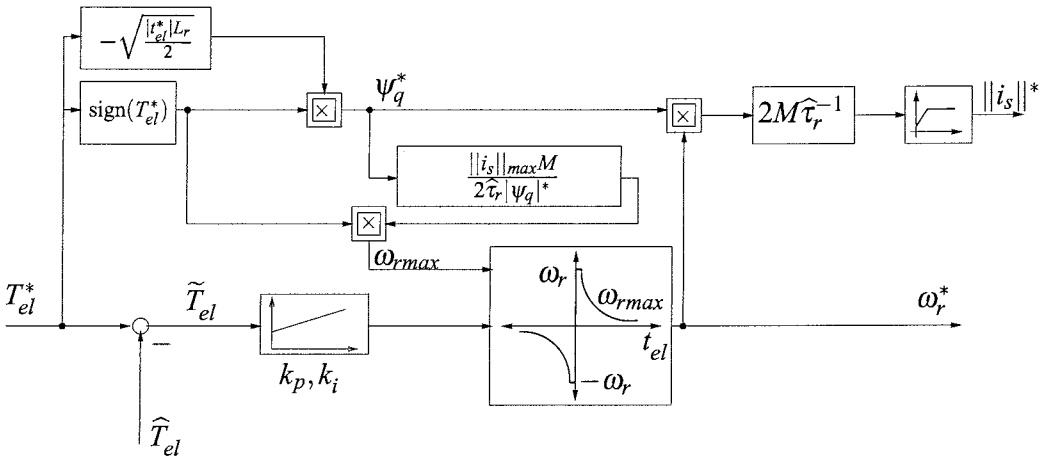


Figure 2: Torque controller structure

3 Dynamic equilibrium and torque control

Assuming that ω_r is a constant parameter then eq.(5) can be put in the linear state-space form:

$$\dot{\psi}_r = \begin{bmatrix} -\tau_r^{-1} & \omega_r \\ -\omega_r & -\tau_r^{-1} \end{bmatrix} \psi_r + \begin{bmatrix} -\tau_r^{-1}M \\ 0 \end{bmatrix} ||i_s|| \tag{7}$$

where $\psi_r = [\psi_d \psi_q]^T$. Calculating the corresponding transfer functions:

$$\begin{aligned} G_1(s) &= \frac{\psi_d}{||i_s||} = \frac{(s\tau_r+1)M}{s^2\tau_r^2+2s\tau_r+1+\omega_r^2\tau_r^2} \\ G_2(s) &= \frac{\psi_q}{||i_s||} = -\frac{\omega_r\tau_r M}{s^2\tau_r^2+2s\tau_r+1+\omega_r^2\tau_r^2} \end{aligned} \tag{8}$$

we can easily conclude that maximal steady-state gain between $||i_s||$ and ψ_q is obtained if dynamic equilibrium condition:

$$\omega_r = \tau_r^{-1} \tag{9}$$

is satisfied. Note that in the steady-state introduced ω_r equals the slip frequency. Since the quantities $\psi_q, ||i_s||$ are torque producing, this implies that maximal torque per ampere ratio is obtained at any steady-state resulting from equilibrium condition (9). It follows further from steady-state analysis that:

$$\begin{bmatrix} \psi_d \\ \psi_q \end{bmatrix} = \begin{bmatrix} \frac{M||i_s||}{1+\omega_r^2\tau_r^2} \\ -\frac{\omega_r\tau_r M||i_s||}{1+\omega_r^2\tau_r^2} \end{bmatrix} \tag{10}$$

from where important identity is obtained if (9) is considered:

$$\psi_d = |\psi_q| \tag{11}$$

where ψ_d is restricted to non-negative values. Furthermore, we can calculate the steady-state torque producing flux linkage vector component and the corresponding current vector magnitude for a given torque command $T_{el}^{\star}$:

$$\begin{aligned} \psi_q^{\star} &= -\text{sign}(T_{el}^{\star})\sqrt{\tfrac{|T_{el}^{\star}|L_r}{2}} \\ \|i_s\|^{\star} &= \tfrac{2}{M}|\psi_q^{\star}| \end{aligned} \tag{12}$$

When considering dynamic model (5) we can easily conclude that monotonic and stable torque responses will be obtained if we can assure that during transients the following general conditions are satisfied:

$$\begin{aligned} &\psi_d \geq |\psi_q| \quad ;\forall T_{el}^{\star} \\ &\dot{\psi}_q \leq 0 \quad ;T_{el}^{\star} > 0 \quad \text{or} \\ &\dot{\psi}_q \geq 0 \quad ;T_{el}^{\star} < 0 \quad \text{or} \\ &\psi_q = 0 \quad ;T_{el}^{\star} = 0 \end{aligned} \tag{13}$$

Fig. 1 gives a graphical interpretation of the different feasible operation sectors. The trajectories of the rotor flux vector inside the $\pi/2$ wide sector I, always ensure stable solutions with monotonic torque responese, yet only the steady-states on the boundaries additionally fulfill the maximal efficiency. In sector II we can employ accumulated energy in the field during transients, however, the torque responses become more and more oscillatory. The third sector must be avoided under any circumstances since the current limit prevents the generation of the required torque and, consequently, stall would occur.

The control design challenge lies in the question how the flux trajectories could be forced to change during transients only inside sector I and, partly in sector II, and that all steady-states will lie as close as possible to the maximal efficiency line. Considering the rotor time constant τ_r as the only uncertain parameter, and based on the presented statements, we therefore propose a torque controller in the form of:

$$\begin{aligned} &\omega_r^{\star} = k_p\widetilde{T}_{el} + v \quad ;|\omega_r^{\star}| \leq \omega_{r\max} \\ &\dot{v} = k_i\widetilde{t}_{el} \qquad ;v(0) = 0 \quad ;k_p, k_i > 0 \\ &\|i_s\|^{\star} = -2\tfrac{\widehat{\tau}_r}{M}\omega_r^{\star}\psi_q^{\star} \quad ; \quad 0 < \|i_s\|_{\min} \leq \|i_s\|^{\star} \leq \|i_s\|_{\max} \end{aligned} \tag{14}$$

where $\widetilde{T}_{el} = T_{el}^{\star} - \widehat{t}_{el}$ is the torque error between the reference and estimated torques, $\widehat{\tau}_r$ is the estimated rotor time constant and $\psi_q^{\star}$ is the required steady-state flux linkage vector component as given by (12). The structure of the proposed torque controller is given in Fig. 2. Stable and monotonic torque response is obtained in the sense that:

$$\lim_{t\to\infty}\widetilde{T}_{el} = 0 \quad ; \quad \lim_{t\to\infty}|\psi_q| \leq |\psi_q^{\star}| \tag{15}$$

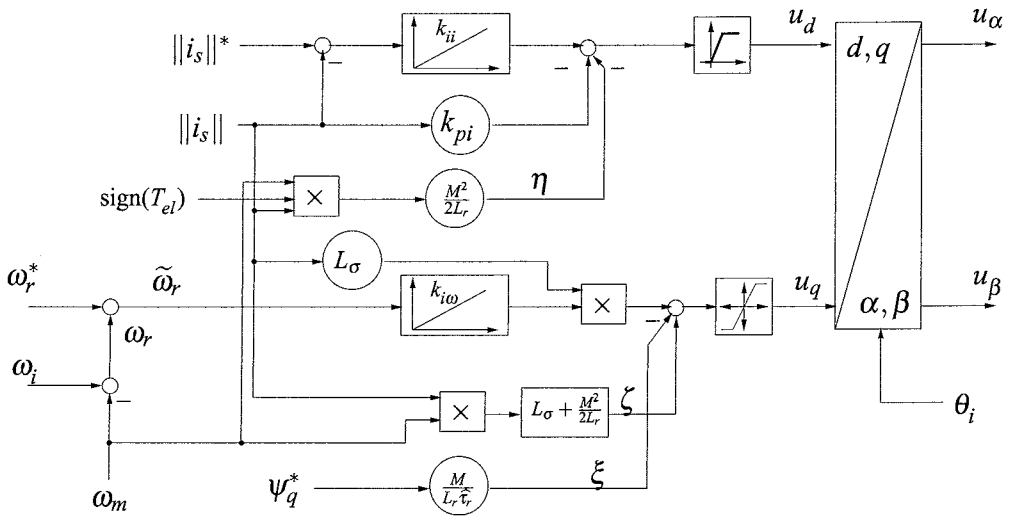


Figure 3: Current control realization scheme.

if following conditions are satisfied:

$$
\begin{aligned}
&\text{I. } \hat{\tau}_r \geq \tau_r \\
&\text{II. } \psi_{d\min} \geq 0 \\
&\text{III. } \psi_d \geq |\psi_q| \Leftrightarrow |\tfrac{d}{dt}(\psi_q|) \geq 0 \\
&\text{IV. } |\omega_r^\star| \leq \frac{||i_s||_{\max} M}{2\hat{\tau}_r |\psi_q^\star|} = \omega_{r\max} \quad ; |\psi_q^\star| \neq 0 \\
&\text{V. } \omega_r^\star = 0 \quad ; \psi_q^\star = 0
\end{aligned}
\tag{16}
$$

First condition is actually easy to satisfy since rotor resistance will generally increase during the operation. The second condition results from the lower current bound $||i_s||_{\min}$ introduced to avoid any numerical singularities. The third condition restricts the flux trajectories to the stable sector I, while condition four imposes variable torque-dependent bounds on $\omega_r^\star$ to admissible values in respect of the maximal reachable current magnitude. The final condition defines the operation at zero reference torque when the current vector rotates synchronous with rotor speed, and therefore influences just the magnetizing flux linkage component ψ_d. Assuming that machine torque is obtained with the first order estimator, the resulting fourth order feedback

system is obtained in the form:

$$\begin{aligned}
\dot{\psi}_d &= -\tau_r \psi_d + (k_p \tilde{T}_{el} + v)\psi_q + \\
&\quad + M\tau_r^{-1}(-M^{-1}2\hat{\tau}_r \psi_q^{\star}(k_p \tilde{T}_{el} + v)) \\
\dot{\psi}_q &= -\tau_r \psi_q - (k_p \tilde{T}_{el} + v)\psi_d \\
\dot{\hat{T}}_{el} &= -\gamma \hat{T}_{el} + \gamma \tfrac{M}{L_r} \psi_q (-M^{-1}2\hat{\tau}_r \psi_q^{\star}(k_p \tilde{T}_{el} + v)) \\
\dot{v} &= k_i \tilde{T}_{el}
\end{aligned} \tag{17}$$

with the initial conditions $[\psi_{d\min}, 0, 0, 0]^T$. As the torque error converge to zero ($\tilde{T}_{el} \to 0$), and assuming that $T_{el}^{\star} = \text{const!} > 0$, then it follows from the third equation of the feedback system (17) that:

$$\dot{\hat{T}}_{el} = 0 = T_{el}^{\star} - \underbrace{\frac{M}{L_r}\psi_q \frac{2\hat{\tau}_r}{M}\psi_q^{\star} v}_{=T_{el}^{\star}} \tag{18}$$

Considering that following is valid for the positive torque command:

$$|\psi_q| \le |\psi_q^{\star}| = -\sqrt{\frac{|T_{el}^{\star}|L_r}{2}} \tag{19}$$

then it follows that the integrator output will converge to the steady-state value of:

$$\lim_{t\to\infty} v \le \frac{1}{\tau_r} \tag{20}$$

Similar results are obtained for the negative reference torques except that: $\lim_{t\to\infty} v \ge -\frac{1}{\tau_r}$, is obtained. Note that in the nominal case ($\hat{\tau}_r \equiv \tau_r$) nominal steady-state integrator output $v = \omega_r = \tau_r^{-1}$ along with the maximal torque per ampere equilibrium $\psi_d = |\psi_q| = |\psi_q^{\star}|$ is obtained. The proportional gain k_p determine how fast the actual torque will converge to the reference value, meanwhile the integrator gain k_i determines how fast the efficiency line is approached. It should also be stressed that, in the case of perturbed rotor time constant, controller output gives its useful estimate in a steady-state which could easily be used in some adaptation procedure. In our design however, this property was not further employed. When analyzing the feedback system (17) further, we derive from the second equation that:

$$\lim_{t\to\infty} |\psi_q| = \underbrace{\frac{\tau_r}{\hat{\tau}_r}}_{\le 1} \psi_d \le |\psi_q^{\star}| \tag{21}$$

and from the first equation that:

$$\lim_{t\to\infty} \psi_d = \lim_{t\to\infty}(-\tau_r^{-1}\psi_d + \underbrace{(\psi_q - \frac{2\hat{\tau}_r}{\tau_r}\psi_q^{*})v}_{\ge 0}) \ge 0 \tag{22}$$

The proposed controller, therefore, satisfies the basic stability conditions (16). As regards the degree of efficiency, it is conditioned by the accuracy of rotor time constant estimate and becomes higher the closer we approach the nominal value τ_r. In the case of parameter mismatch the efficiency, therefore, obviously decreases yet the stable operation is preserved.

4 Current control

The task of the inner current controllers is quite demanding since reference tracking, disturbance suppression and voltage drop compensation of the VSI must be simultaneously achieved. Note that inverse mapping between the current vector magnitude $||i_s||$ and the voltage u_d is characterized by perturbed first-order dynamics and that inverse mapping between ω_r and voltage u_q is purely algebraic and nonlinear. Before introducing the current controllers some physical interpretation of the nominal mapping $||i_s||, \omega_r \rightarrow u_d, u_q$ should be given. If synchronous operation is assumed, as stator current vector and rotor flux linkage vector are aligned, characterized by the conditions: $\omega_r = 0 \rightarrow T_{el} = 0, \psi_q = 0$ and $\omega_m \neq 0, ||i_s|| \neq 0 \rightarrow \omega_i = \omega_m, \psi_d \neq 0$, required steady state voltage vector could be expressed as:

$$\begin{aligned} u_d &= (R_\sigma - \tfrac{M^2}{L_r \tau_r})||i_s|| \\ u_q &= (L_\sigma + \tfrac{M^2}{L_r})||i_s||\omega_m \end{aligned} \tag{23}$$

where the relation $\psi_d = M||i_s||$ valid at synchronous operation was considered. The control voltage u_d therefore influence only the flux linkage magnitude while u_q is needed to enable synchronous rotation of the current vector. On other hand, if the machine rotor is locked ($\omega_m = 0 \rightarrow \omega_i = \omega_r$) and if the maximal efficiency operation should be achieved ($\omega_r = \tau_r^{-1}$), the corresponding control voltage components are obtained in the form:

$$\begin{aligned} u_d &= (R_\sigma - \tfrac{M^2}{2L_r \tau_r})||i_s|| \\ u_q &= (L_\sigma + sign(T_{el})\tfrac{M^2}{2L_r})||i_s||\omega_r \end{aligned} \tag{24}$$

where the relations $\psi_d = ||i_s||M/2$ and $|\psi_q| = \psi_d$ were considered. In the general case when machine is rotated at some mechanical speed and an arbitrary electrical torque should be generated simultaneously satisfying maximal efficiency requirement the control voltage can not be simply expressed as a superposition of (23) and (24) since flux linkage vector changes its magnitude and relative position between any steady state operation point. Control voltage vector assuming

steady state operation at maximal efficiency $\omega_r = \tau_r^{-1}$ and considering that $\omega_i = \omega_m + \omega_r$ could be therefore calculated as:

$$\begin{aligned} u_d &= (R_\sigma + (sign(T_{el})\omega_m - \omega_r)\tfrac{M^2}{2L_r})\|i_s\| \\ u_q &= (L_\sigma + \tfrac{M^2}{2L_r})\|i_s\|\omega_m + (L_\sigma + sign(T_{el})\tfrac{M^2}{2L_r})\|i_s\|\omega_r \end{aligned} \tag{25}$$

Note that both voltage components in (25) ere expressed with known or measured quantities and that they also defines the steady state lower bound for u_d and upper bound for u_q if stable operation in sector I should be achieved. It is obvious that the control voltage component u_q is significant with respect to transient torque response, efficiency and stability. Poorly damped oscillatory torque responses are obtained if the control voltage u_q is too high. In an extreme case instability due to stall could even occur, as the angle between the stator current vector and the rotor flux linkage vector approaches $\pm\pi/2$. In the opposite case, as the rotation speed is too small, the angle between the stator current and the rotor flux linkage vectors is also small, machine is, therefore, forced to operate with an unnecessary large rotor flux. Poor efficiency is the obvious consequence although the required torque is generated.

By considering the first equation of model (4) along with (25) we therefore propose a PI state controller with feed-forward compensation in the form of:

$$\begin{aligned} u_d &= -k_{pi}\|i_s\| + v_1 + \eta \quad ; 0 \le u_d \le u_{d\max} \\ \dot{v}_1 &= k_{ii}(\|i_s\|^\star - \|i_s\|) \quad ; v_1 \ge 0 \\ \eta &= -sign(T_{el})\tfrac{M^2}{2L_r}\|i_s\|\omega_m \end{aligned} \tag{26}$$

where $\|i_s^\star\|$ is the reference current vector magnitude obtained from the torque controller (14), $k_{pi}, k_{ii} > 0$ are the state controller gains assuring the desired dynamics and η is the feed-forward term introduced to compensate the induced voltage due to rotor speed. By limiting controller output to only the positive values, the rotor field is built-up faster than it is destroyed. This useful property is specially important during the active breaking. The resulting feedback system is in the form of:

$$\begin{aligned} \dot{\|i_s\|} = \tfrac{1}{L_\sigma}(&-(R_\sigma + k_{pi})\|i_s\| + \tfrac{M}{L_r}\tau_r^{-1}\psi_d + \\ &+\omega_m\tfrac{M}{L_r}(\psi_q - \underbrace{sign(T_{el})\frac{M}{2}\|i_s\|}_{\to\psi_q^\star}) + v_1) \end{aligned} \tag{27}$$

In the steady-state the third summand vanishes as the maximal efficiency line is reached, resulting in a second order system with constant additive disturbance. As $\psi_d > M/2\|i_s\|$ controller

gain k_{pi} must be selected so that the condition:

$$(R_\sigma + k_{pi}) \geq \frac{M^2}{2L_r\widehat{\tau}_r} \tag{28}$$

is fulfilled. During the transients inside sector I compensation term is positively defined for positive torques and negatively defined for the negative torques or vice versa if ω_m change sign. Thus the current vector magnitude is adequately increased (or decreased) as required from the torque controller. In the case of perturbed rotor time constant, the current vector magnitude is slightly overcompensated which is corrected by integral control action. The feedback system, therefore, ensures that $\|i_s\| \to \|i_s\|^\star$ with the chosen dynamics, as long as the voltage restraints are not violated.

Looking at the second equation of (25) we conclude that the required voltage u_q could be represented as the sum of voltage components due to the rotor speed and due to required torque. For the nominal case as $\omega_r \to \text{sign}(T_{el}^\star)\tau_r^{-1}$, $\psi_q \to \psi_q^\star$ and $\psi_d \to |\psi_q^\star|$ the required steady-state voltage u_q can be expressed also as the function of ω_i in the form of:

$$\begin{aligned} u_q &= L_\sigma\|i_s\|\omega_i + \omega_m\tfrac{M}{L_r}|\psi_q^\star| + \omega_r\tfrac{M}{L_r}|\psi_q^\star| \\ &= (L_\sigma\|i_s\| + \tfrac{M}{L_r}|\psi_q^\star|)\omega_i \\ &= (\tfrac{2L_\sigma}{M} + \tfrac{M}{L_r})|\psi_q^\star|\omega_i = \psi_{sd}\omega_i \end{aligned} \tag{29}$$

Introducing the stator flux linkage vector component ψ_{sd} relation to the DTC could be established. In the nominal, as well as in the perturbed case, stable operation inside the sector I will be obtained if the selected controller will provide voltages that are lower or equal to those given with (29) or (25). To achieve the required steady-state conditions along, with the satisfactory transient operation, control design turns out to be quite demanding.

Due to algebraic mapping between $T_{el}^\star \to \omega_r^\star \to u_q$ we propose only integral control action to avoid algebraic loop along with additional compensation terms in the form of:

$$\begin{aligned} u_q &= L_\sigma\|i_s\|v_2 + \zeta + \xi \quad ; u_q \leq \sqrt{U^2 - u_{d\max}^2} \\ \dot{v}_2 &= k_{i\omega}(\omega_r^\star - \omega_r) \quad ; v_2(0) = 0; k_{i\omega} > 0 \\ \zeta &= (L_\sigma + \tfrac{M^2}{2L_r})\|i_s\|\omega_m \quad ; \xi = -\tfrac{M}{L_r\widehat{\tau}_r}\psi_q^\star \end{aligned} \tag{30}$$

where $\omega_r^\star$ is the speed reference obtained from the torque controller, $U = 0.86U_{DC}$ is the maximal voltage vector magnitude, $k_{i\omega}$ is the integrator gain. The resulting feedback is therefore obtained in the form of the output perturbed first order system:

$$\begin{aligned} \omega_r &= (\tfrac{M}{2}\|i_s\| - \psi_d)\tfrac{M}{L_\sigma L_r\|i_s\|}\omega_m + (\tfrac{\psi_q}{\tau_r} - \tfrac{\psi_q^\star}{\widehat{\tau}_r})\tfrac{M}{L_\sigma L_r\|i_s\|} + v_2 \\ &= \delta + v_2 \end{aligned} \tag{31}$$

where δ represents bounded output perturbation. Considering the nominal case it is obvious that, as $\psi_d/\|i_s\| \to M/2$ and $\psi_q \to \psi_q^\star$, the first and the second summand vanish and the integrator output v_2 converge to $\omega_r^\star$. The upper bound of the disturbance $\delta_{\max}$ at the perturbed case is given for $\psi_d = 0$, $\psi_q = 0$, $\omega_{m\max}$, $\|i_s\|_{\min}$ and must be compensated by the control action v_2. In order to achieve the required tracking as good as possible $k_{i\omega}$ must be a high gain to to dominate transient output perturbation at any operation condition. The structure of the current control is presented in Fig. 3.

5 Experimental results

The motor data, along with all controller parameters and current and voltage bounds, are given in Appendix I. The first few experiments were performed with the locked rotor ($\omega_m = 0 \to \omega_i = \omega_r$) to illustrate the proposed control performance more clearly. Although ω_i could be easily obtained directly by derivation of the current angle vector θ_i a second order resolver as proposed in [14] was used in our realization. At first the inner loop was tested for step and sinusoidal changes in the reference current $\|i_s\|^\star$ and relative speed $\omega_r^\star$. To obtain sufficiently accurate flux estimates we used torque sensor output T_{elm} to obtain $\widehat{\psi}_q$ first from where the $\widehat{\psi}_d$ was calculated as:

$$\begin{aligned} \widehat{\psi}_q &= -\tfrac{T_{elm}L_r}{M\|i_s\|} \\ \dot{\widehat{\psi}}_d &= -\widehat{\tau}_r^{-1}\widehat{\psi}_d - \tfrac{T_{elm}L_r}{M\|i_s\|}\omega_r + M\widehat{\tau}_r^{-1}\|i_s\| \end{aligned} \tag{32}$$

where $\widehat{\tau}_r$ can be obtained from the steady-state torque controller output v. It should be pointed out that these estimates are only used to evaluate the control performance and are not used as part of the feedback. We can see from the estimated rotor flux linkage trajectory that the optimal efficiency line is reached for each steady state as long as the condition $\omega_r \approx \tau_r^{-1}$ is satisfied. During transient the flux linkage trajectory changes only inside the stable sector I (Fig. 4).

In Fig. 5 an experiment is shown where current vector magnitude was kept constant meanwhile rotation speed ω_i was changed as a sin function with the magnitude between $\pm\tau_r^{-1}$. A similar operation could be observed as if FOC were employed. Since current magnitude is constant the rotor flux increases as ω_i decreases. The high stability margin and high reachable dynamics in torque generation are clearly advantages of this approach, yet efficiency is mainly lost.

The next two experiments show the responses of the torque controlled machine. Only positive torque reference was used during the first experiment in Fig. 6. The actual torque was

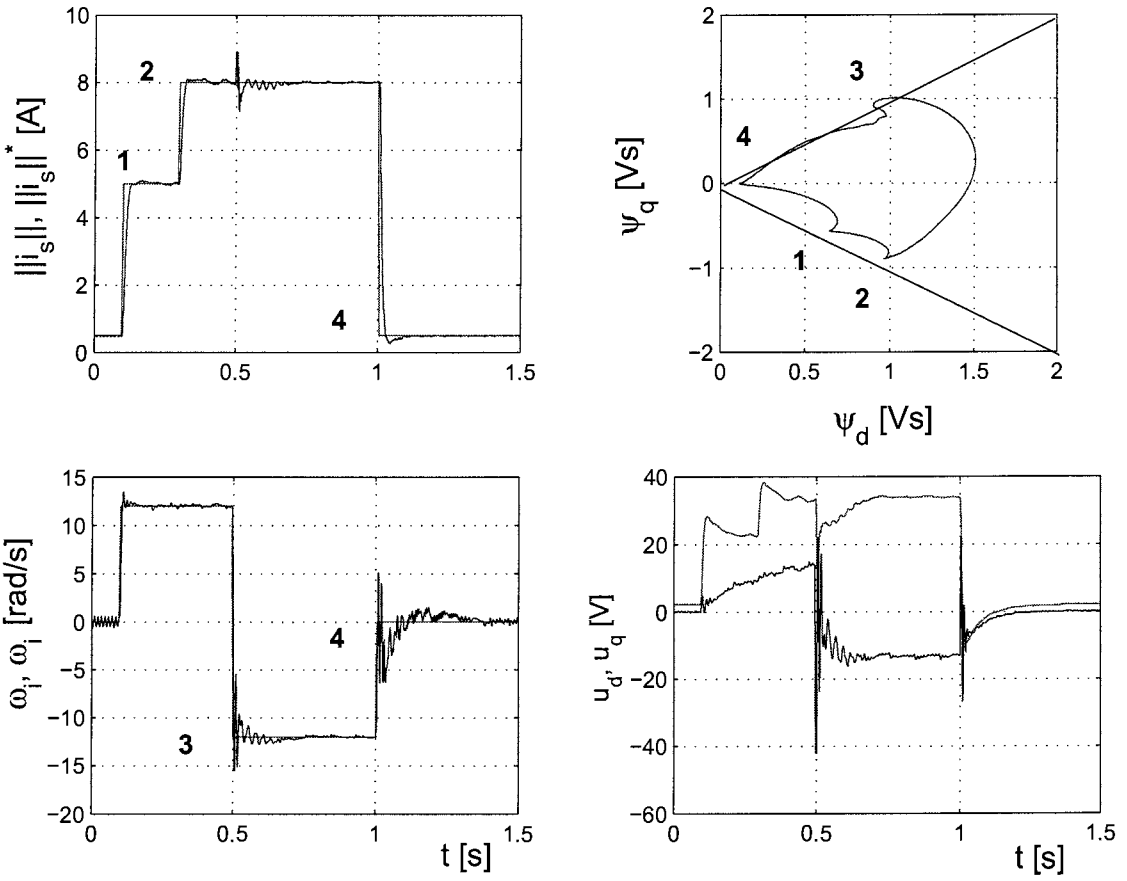


Figure 4: Current control step response.

built-up with $\approx$ 60ms delay, almost perfect tracking of the torque command can be observed after the rotor field is restored. The growth of ψ_d is slightly faster than growth ψ_q and the flux trajectory is, therefore, forced to change inside the desired sector I. As the magnitude of the flux linkage vector components equalize in the steady state the maximal efficiency line is also reached.

A more demanding tracking task is presented in the next experiment Fig. 7 where a comparison of the estimated and measured torque t_{elm} is also included. Since the measured and estimated electrical torque shows acceptable agreement also the estimated rotor flux using (32) is accurate enough to evaluate feedback performance. Stable operation inside the sector I, along with the required efficiency, is also preserved during this experiment.

Again the same delay as in the previous experiment could be observed whenever the flux linkage magnitude approaches zero. This delay could be theoretically reduced by increasing

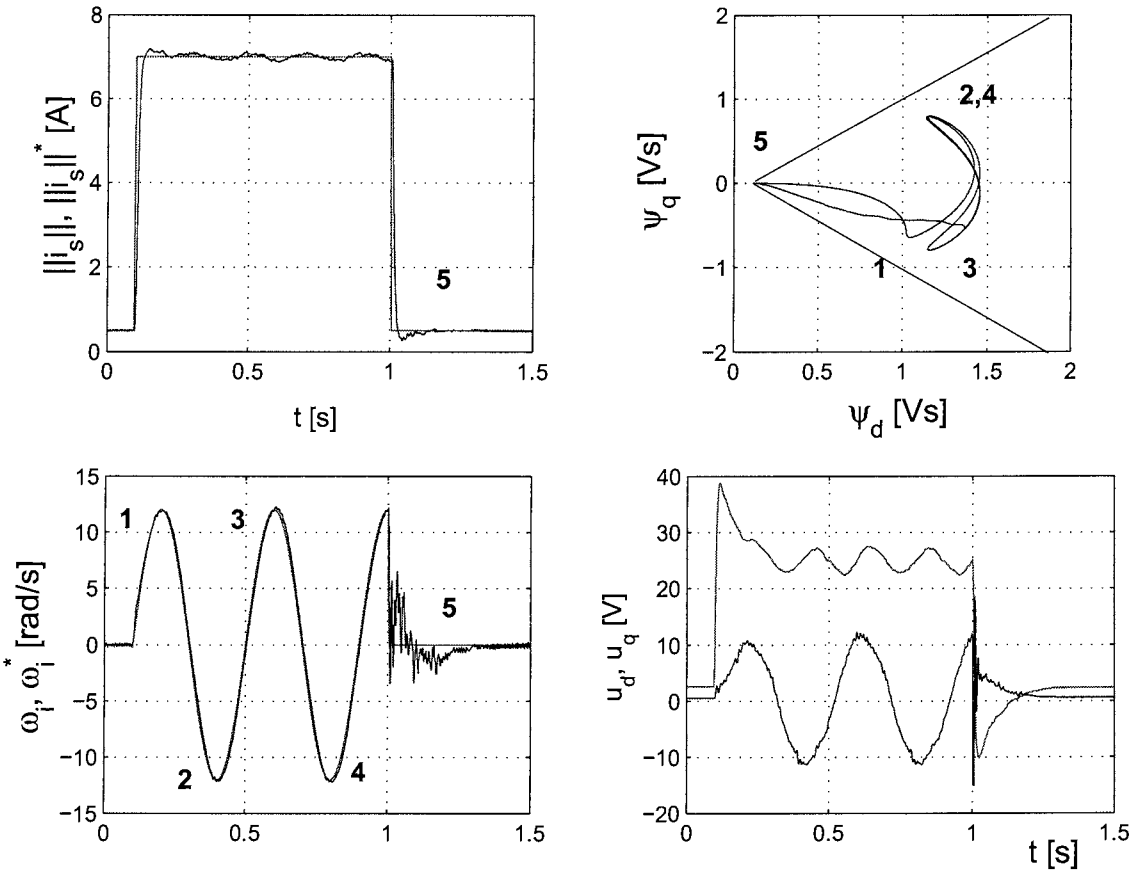


Figure 5: Current control response to sin changes of $\omega_i^\star$.

the torque controller gain k_p but the setting is limited by the dynamics of the ω_i inner loop. It was observed during the numerous experiments that the smoothness of the actual (measured) ω_i which is manly conditioned by the delay compensation of the VSI, actually restricts the reachable dynamics of the inner speed loop and, therefore, essentially influences the entire feedback performance. In spite of this, experiments with commercial available VSI, fulfill the main expected control objectives: torque tracking and maximal steady-state efficiency.

The last experiment in Fig. 8 shows the torque tracking for the unlocked rotor. To avoid the danger of escaping, the mechanical speed was bounded by the DC machine generating proportional breaking torque ($t_l = k\omega_m$) by actually simulating the effect of the increased linear friction. Insignificant differences were observed compared to experiments with locked rotor except that ω_i and u_q are additionally modulated with ω_m. To illustrate also the three phase

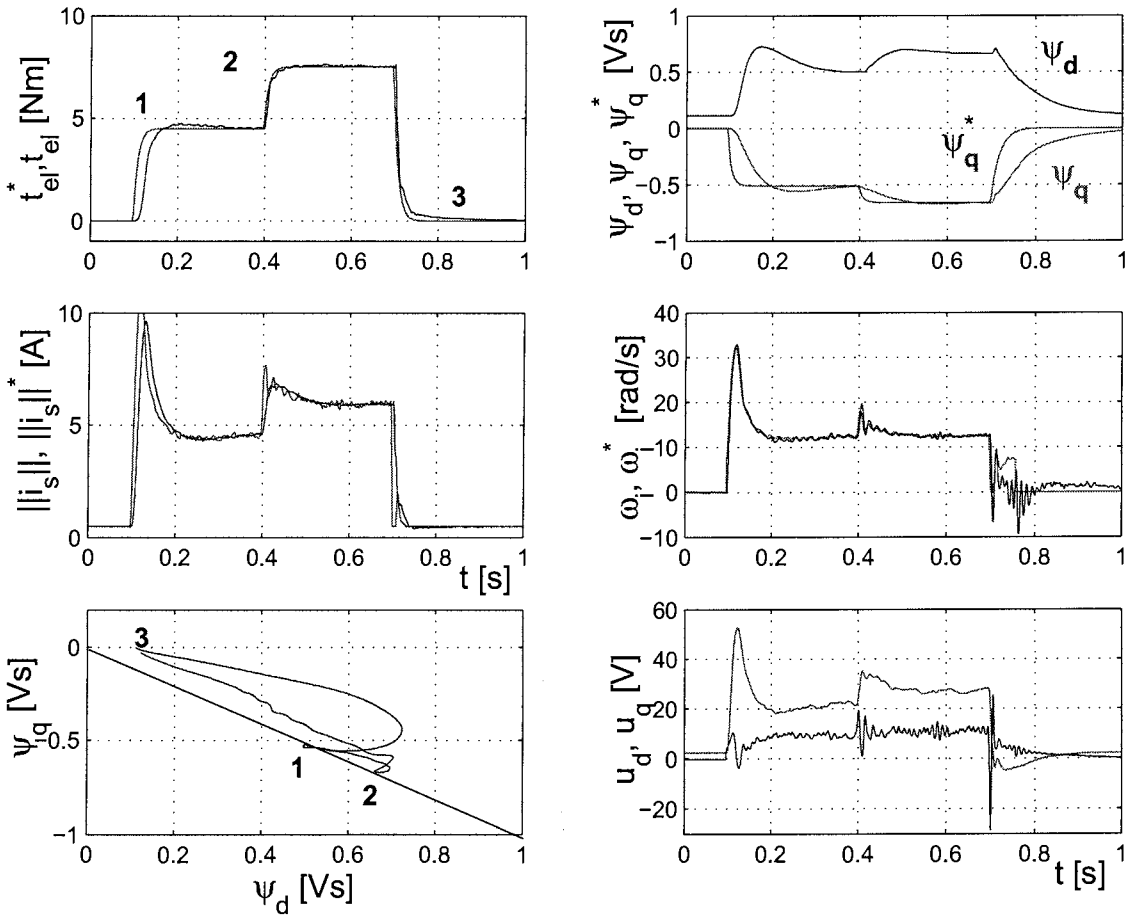


Figure 6: Torque tracking.

quantities measured phase currents are included in the figures.

6 Conclusion

In this paper the IM, torque control based on a stator current vector reference frame was presented. The overall design is motivated by physical interpretation of the transient and steady-state IM phenomena rather than by the complicated non-linear control theory. Nevertheless the derived IM control based on introduced dynamic equilibrium establishes almost perfect tracking, monotonous stability, sufficient robustness with respect to the rotor time constant perturbation and build in efficiency. These properties were achieved by rejecting the FOC paradigm of decoupled control with the constant rotor flux linkage vector magnitude. The key requirement that all rotor flux linkage vector trajectories should be restricted to stable sector I and that in a

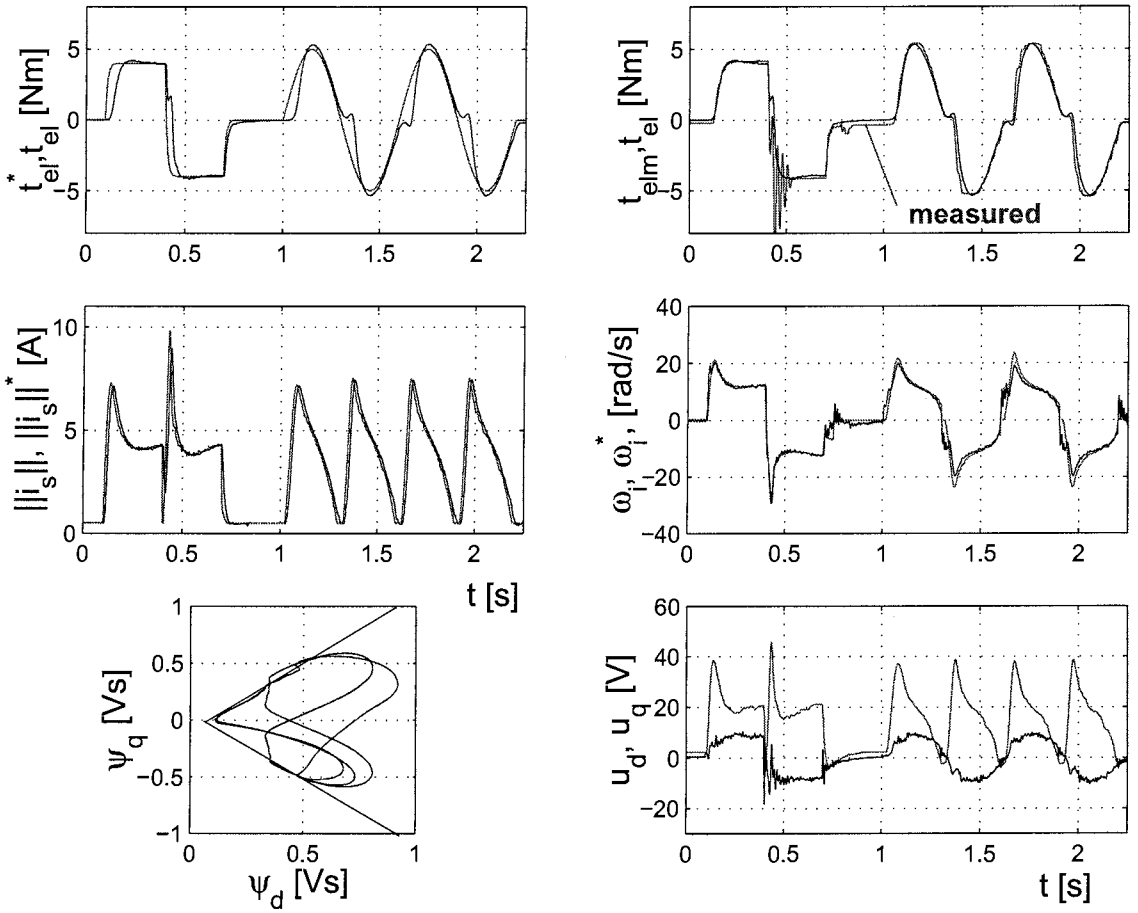


Figure 7: Tracking of the general torque reference.

steady-state the trajectories should end as close as possible to the maximal efficiency line, simultaneously fulfill the basic control objectives. High dynamic is preserved although the rotor flux linkage is forced to change according to the torque command. Proposed control inherently include "field weakening operation mode" without any additional control action. To implement the proposed control successfully, the machine torque must be estimated and the VSI must be delay compensated to obtain smooth current vector rotation.

Our further work will be oriented toward inclusion of the saturation effects, derivation of guidelines for the optimal controller settings and to the possibilities of adaptation mechanisms.

References

[1] P.C. Krause, "*Analysis of Electric Machinery*," McGraw-Hill,1986.

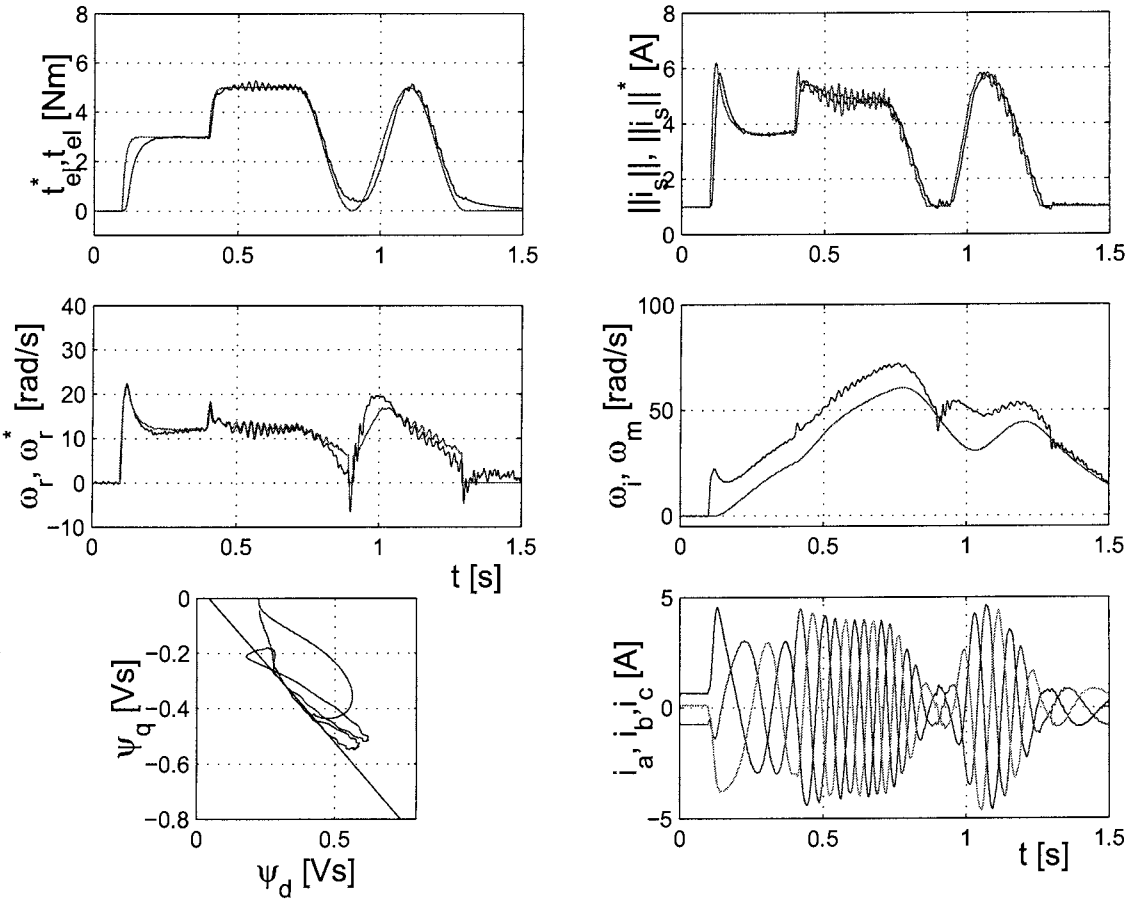


Figure 8: Torque control with the unlocked rotor.

[2] F. Blaschke, "Das Prinzip der Feldorientierung, die Grundlage für die Transvektor-Regelung von Drehfeldmaschinen," *Siemens Zeitschrift*, vol.10, no. 45, 1971.

[3] R. Ortega,A. Loria,P.J. Nicklasson and H. Siera-Ramirez "*Pasivity-based Control of Euler-Lagrange Systems*," Springer-Verlag, Berlin, 1998.

[4] R. Marino, S, Peresada and P. Tomei, "Global Adaptive Output Feedback Control of Induction Motors with Uncertain Rotor Resistance," *IEEE Trans. on Automatic Control*, vol.44, no. 5, pp.:967–980,1999.

[5] J. Chiasson, "A New Approach to Dynamic Feedbac Linearization Control of an Induction Motor," *IEEE Trans. on Automatic Control*, vol.43, no. 3, pp.:391–396,1998.

[6] I. Takagashi and T. Noguchi, "A New Quick-Response and High-Efficiency Control Strategy of an Induction Motor," *IEEE Trans. on Ind. Applications*, vol.IA-22, no. 5, pp.:820-827,1986.

[7] M. Depenbrock, "Direct Self-Control (DSC) of Inverter-Fed Induction Machine," *IEEE Trans. on Ind. Applications*, vol.IA-22, no. 5, pp.:820-827,1986.

[8] R. Ortega, N. Barabanov and G. Escobar, "Direct Torque Control of Induction Motors: Stability Analysis and Performance Improvement ," *IEEE Trans. on Automatic Control*, vol.46, no. 8, pp.:967–980,2000.

[9] P. Martin and P. Rouchon, "Flatness and Sampling Control of Induction Motor," *IFAC, 13th Triennial Worl Congres, San Francisco*, 2b 27 2, pp.:389–394,1996.

[10] R. Brockett, "Characteristic Phenomena and Model Problems in Nonlinear Control," *IFAC, 13th Triennial Worl Congres, San Francisco*, 2b II, pp.:257–262,1996.

[11] M. Bodson and J. Chiasson, "Differential-Geometric Methods for Control of Electric Motors ," *Int. Journal of Robust and Nonlinear Control*, no. 8, pp.:927–952,1998.

[12] C. Attaianese, V. Nardi, A. Perfetto and G. Tomasso, "Vectorial Torque Control: A Novel Approach to Torque and Flux Control of Induction Motor Drives ," *IEEE Trans. on Ind. Applications*, vol.35, no. 6, pp.:1399–1405,1999.

[13] Z. Beres and P. Vranka, "Sensorless IFOC of Induction Motor With Current Regulators in Current Reference Frame ," *IEEE Trans. on Ind. Applications*, vol.37, no. 4, pp.:1012–1018,2001.

[14] L. Harnefors and H.P. Nee, "A General Algorithm for Speed and Position Estimation of AC Motors," *IEEE Trans. on Ind. electronics*, vol.47, no. 1, pp.:77–83,2000.

Regelung von Petersen Spulen

Gernot Druml

Dohlenweg 3
D-90768 Fürth
g.druml@ieee.org

Kurzfassung

In diesem Beitrag werden zwei neue Verfahren zur Ermittlung der aktuellen Abstimmung eines gelöschten Netzes vorgestellt. Die Berechnung der Verstimmung erfolgt dabei ohne Verstellung der Petersen-Spule und ist auch für symmetrische Netze geeignet. Übliche Störeinflüsse auf die Verlagerungsspannung, verursacht durch Übersprechen des Mitsystems, werden bei diesen neuen Verfahren unterdrückt. Sowohl die Anzahl der erforderlichen Verstellungen der Petersen-Spule als auch die Zeit der Fehlabstimmungen werden durch die neuen Verfahren stark reduziert.
Im ersten Teil werden kurz die Schritte zur Herleitung der Ersatzschaltung des Nullsystems sowie die dabei durchgeführten Vernachlässigungen aufgezeigt. Im darauf folgenden Abschnitt werden die derzeit bekannten Berechnungs- und Regelverfahren sowie deren Vor- und Nachteile dargestellt. Im Anschluss werden die neuen Verfahren zur Berechnung der aktuellen Abstimmung vorgestellt, die als Grundlage für die eventuell erforderliche Nachregelung der Petersen-Spule verwendet werden können.

1 Einleitung

In vielen Ländern der EU ist die "Erdschlusslöschung", wie in Bild 1 dargestellt, eine der wichtigsten Optionen bei der Planung von Elektrizitätsverteilungs-Netzen, um eine optimale Qualität der Energieversorgung zu erreichen. Der Hauptvorteil dieser Art der Sternpunktsbehandlung ist, dass auch während des Erdschlusses die Energieversorgung aufrecht erhalten bleibt. Als Folge wird die Anzahl der Versorgungsunterbrechungen für den Kunden stark reduziert [1].

Weitere wesentliche Vorteile der Erdschlusslöschung sind:

- Reduzierung des Stromes über die Fehlerstelle
- Reduzierung der Zerstörung an der Fehlerstelle
- Reduzierung der Berühr- und Schrittspannung in der Umgebung der Fehlerstelle
- Langsamere Wiederkehr der Leiter-Erde Spannung nach dem Verlöschen des Lichtbogens
- Verlöschen des Lichtbogens in Freileitungsnetzen
- Verbesserung der Power Quality durch die Reduktion der Anzahl und Dauer

der Versorgungsunterbrechungen
- Reduktion der Gefahr von Kippschwingungen

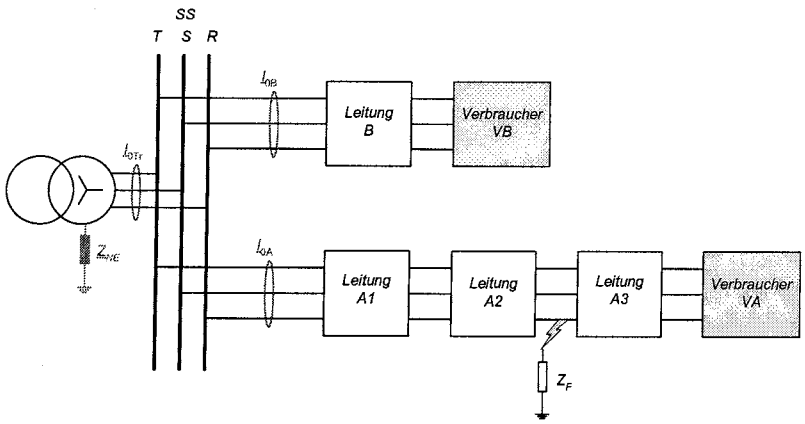


Bild 1: Entladung des fehlerhaften Leiters über die Erde

Damit das Löschen des Lichtbogens funktioniert, muss die Petersen-Spule unter Einhaltung von Randbedingungen [2], [3], [5], [10] gut abgestimmt sein.

Mit der Zunahme der Verkabelung in den Verteilnetzen werden einerseits die Maximalwerte der Resonanzkurven immer kleiner und anderseits wird die Kurvenform immer steiler. Die Ursachen für die kleiner werdenden Resonanzmaxima sind in dem wesentlich symmetrischeren Aufbau der Kabelnetze im Vergleich zu Freileitungen zu finden. Durch die Verkabelung ist die mögliche Wärmeabfuhr über das umgebende Erdreich reduziert, sodass die Kabelnetze so ausgelegt werden, dass sie wesentlich kleinere Verluste haben. Dies führt zu den steileren Resonanzkurven.

Ein erster Lösungsansatz wäre, die Empfindlichkeit der Messung der Verlagerungsspannung zu erhöhen. In [3] wurden die verschiedensten Störeinflüsse auf die Messung bei sehr symmetrischen Netzen untersucht und dargestellt.

Im vorliegenden Beitrag werden die bekannten Verfahren zur Abstimmung einer Petersen Spule gegenübergestellt und ein neues Verfahren aufgezeigt, dass auch für komplett symmetrische Netze geeignet ist. Zusätzlich ist mit dem neuen Verfahren eine Verstimmung der Petersenspule zur Berechnung der aktuellen Abstimmung nicht mehr erforderlich. Außerdem ist das neue Verfahren auch robust gegenüber den in [3] beschriebenen Störungen der Messung der Verlagerungsspannung.

Der Algorithmus wurde ausführlich an verschiedenen Modellen mit Hilfe von Matlab/Simulink/SimPowerSystems getestet, bevor er auf entsprechender Hardware implementiert wurde.

2 Grundlagen der Erdschlusslöschung

2.1 Prinzip der Erdschlusslöschung

Im gesunden Netz fließen bei einem symmetrischen Netzaufbau über die Leiter-Erdkapazitäten drei gleich große Ströme. Die Spannungsabfälle an den drei Leiter-Erde Kapazitäten sind somit auch gleich groß. Die sich dadurch einstellende Verlagerungsspannung $\underline{U}_{NE}$ ist Null bzw. sehr klein. Es fließt über die Petersenspule kein Strom bzw. nur ein sehr geringer Strom.

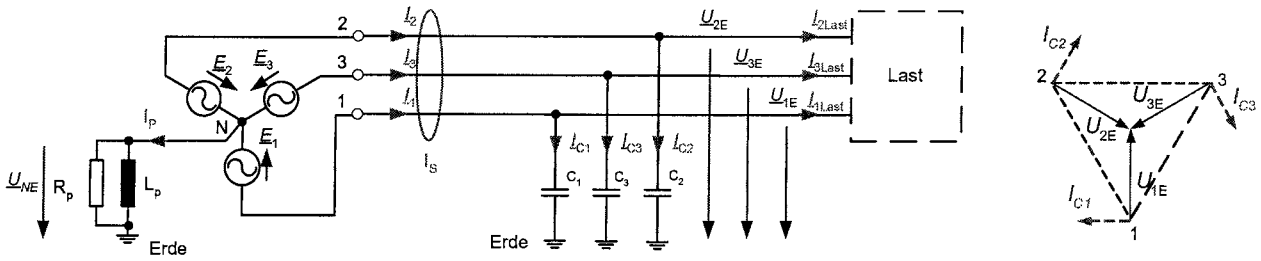


Bild 2: Gesundes Drehstromnetz mit Zeigerdiagramm

Durch den Erdschluss werden die beiden gesunden Leiter, wie in Bild 3 dargestellt, auf die verkettete Spannung angehoben.

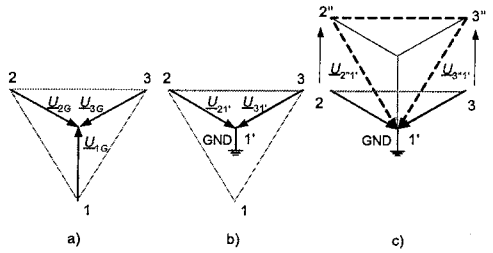


Bild 3: Aufladevorgang der beiden gesunden Leiter über Erde
a) gesundes Netz, b) Entladung des erdschlussbehafteten Leiters, c) Aufladung der beiden gesunden Leiter

Beim Erdschluss überlagern sich die drei folgenden Vorgänge, die gleichzeitig beginnen aber unterschiedlich lang andauern [4]:

- Entladung des fehlerhaften Leiters über die Erde
- Aufladung der beiden gesunden Leiter über die Erde
- Stationärer Zustand des Erdschlusses

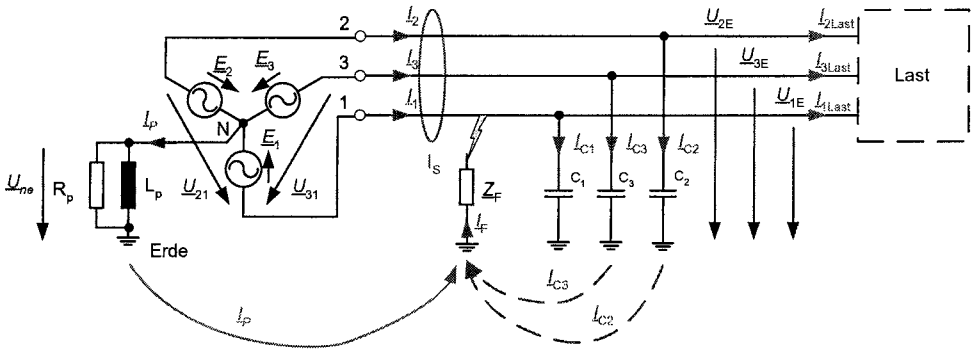


Bild 4: Erdschlussbehaftetes Netz

Im erdschlussbehafteten Netz fließen nun über die Leiter-Erde Kapazität große kapazitive Ladeströme, die über die Fehlerstelle zurück zum Transformator fließen. Abhängig von der Netzgröße können Ströme bis in den Bereich von 1000 A fließen. Dadurch wird die umgesetzte Energie an der Fehlerstelle groß und zulässige Grenzwerte für Berührspannung und Schrittspannung werden an der Fehlerstelle und in deren Umgebung meist stark überschritten. Eine Reduzierung des Stromes über die Fehlerstelle ist daher unbedingt erforderlich und kann durch Überlagerung eines Stromes mit umgekehrtem Vorzeichen Bild 5 erfolgen.

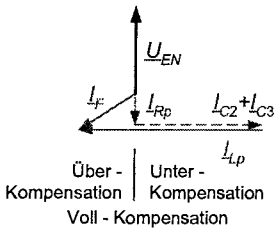


Bild 5: Zeigerdiagramm der Erdschlusskompensation

Von den unterschiedlichen Möglichkeiten der Anordnung der Kompensationsspulen hat sich als günstigste die von Waldemar Petersen vorgeschlagene Version erwiesen, die an den Sternpunkt des Trafos angeschlossen wird (Petersen-Spule). Bei idealer Kompensation des kapazitiven Stromes kann der Strom an der Fehlerstelle dadurch auf weniger als 3% reduziert werden. Es verbleibt im Prinzip nur der Wirkstrom des Netzes.

Die Beschreibung und Berechnung des Netzes kann mit Hilfe der symmetrischen Komponenten [6], [7], [8] erfolgen.

2.2 Symmetrische Komponenten

Im folgenden Bild 6 ist für einen Leitungsabschnitt aus Bild 1 das dreiphasige Ersatzschaltbild dargestellt.

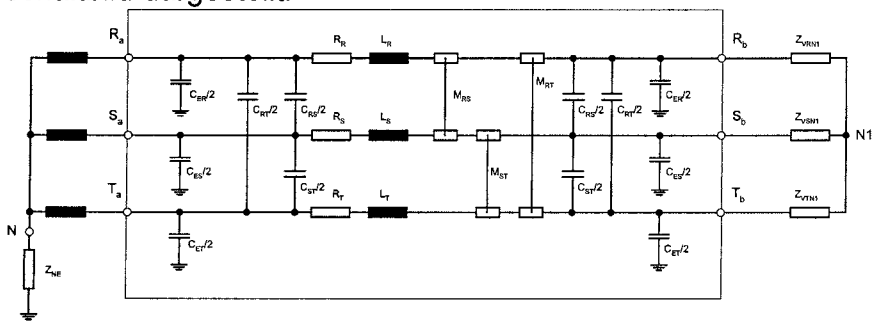


Bild 6: Ersatzschaltung eines Leitungselementes des Netzes im Bild 1

Die Anwendung der kompletten Ersatzschaltung der Leitung zur Beschreibung und Berechnung eines ausgedehnten Netzes wird schnell sehr komplex. Durch Transformation auf "Symmetrische Komponenten" erhält man für den gesunden Betrieb für den stationären Zustand das in Bild 7 dargestellte entkoppelte Netzwerk.

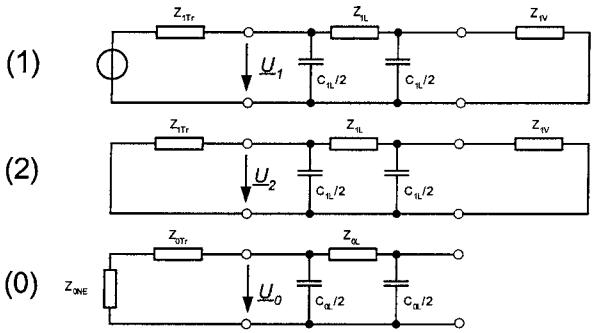


Bild 7: Darstellung des gesunden Netzes mit Symmetrische Komponenten

Im folgenden Bild ist das Netz von Bild 1 vollständig mit Hilfe der symmetrischen Komponenten beschrieben.

Beim einpoligen Erdschluss erfolgt, wie im Bild 8 dargestellt, eine Kopplung der drei Systeme an der Fehlerstelle.

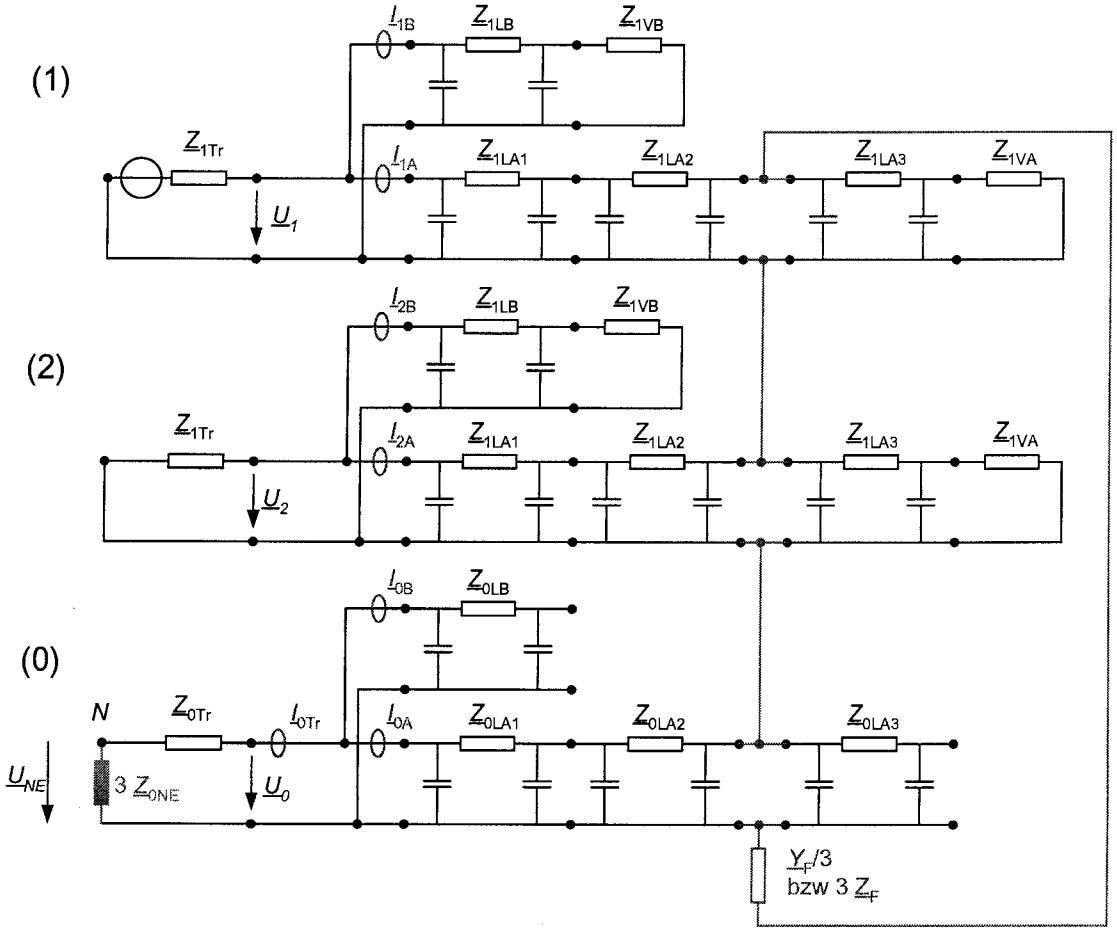


Bild 8: Darstellung des erdschlussbehafteten Netzes vom Bild 1 mit Symmetrischen Komponenten

Für die Betrachtung des Erdschlusses können weitere Vereinfachungen getroffen werden. Die Impedanz der Last ist wesentlich größer als die Summe der Mitimpedanz des Trafos und der Leitungsimpedanz vom Umspannwerk bis zur Fehlerstelle. Entlang der Leitung ist bei Nennbetrieb ein Spannungsabfall von 10% erlaubt. Dies bedeutet, dass die Impedanz der Last mindestens den zehnfachen Wert der Impedanz des Trafos und der Leitungslängsimpedanz hat. Die folgenden Impedanzen können daher zur Impedanz $\underline{Z}_i$ zusammengefasst werden:

- Mitimpedanz des Trafos
- Längsimpedanzen des Mitsystems der Leitung vom Trafo bis zur Fehlerstelle
- Gegenimpedanz des Trafos
- Längsimpedanz des Gegensystems der Leitung vom Trafo bis zur Fehlerstelle

Dies führt zu dem in Bild 9 dargestellten vereinfachten Ersatzschaltbild.

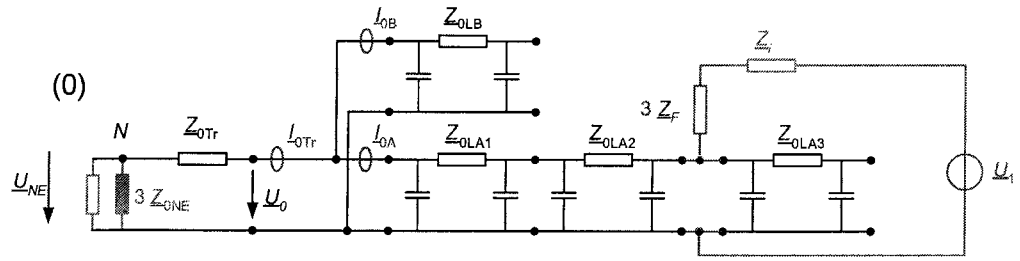


Bild 9: Reduktion des Mit- und Gegensystems

Im Bild 10 wurde zusätzlich die Leitungslängsimpedanz des Nullsystems der gesunden Abgänge vernachlässigt.

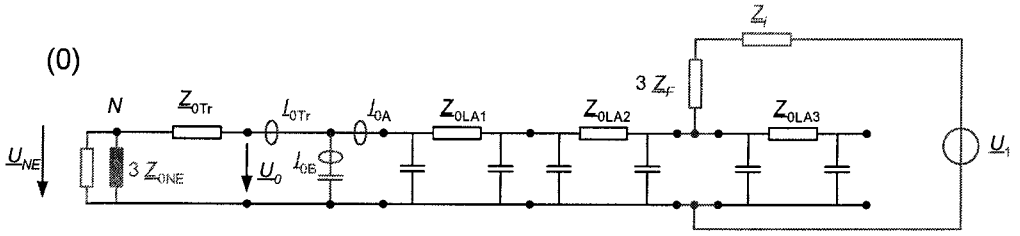


Bild 10: Reduktion der gesunden Abgänge

Für die weiteren Betrachtungen werden nun auch die Längs-Nullimpedanzen der Leitung im erdschlussbehafteten Abgang vernachlässigt. Die ohmschen Verluste in den Längsimpedanzen und die des Trafos werden über den Verlustwiderstand parallel zur Petersenspule berücksichtigt. Die Praxis zeigt, dass nur ca. 50 % der Verluste von der Petersenspule erzeugt werden. Der Rest wird durch die Verluste im Trafo und in den Leitungen verursacht. Dies führt zu der sehr einfachen Ersatzschaltung von Bild 11.

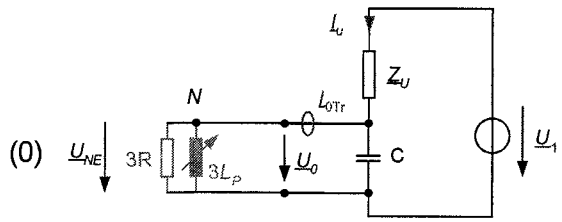


Bild 11: Resultierendes Ersatzschaltbild

Bei einem sehr niederohmigen Erdschluss ist die Spannung am Resonanzkreis mehr oder weniger eingeprägt. Der Strom $\underline{I}_u$ über die Fehlerstelle ist abhängig von der Art der Kompensation. Bei exakter Abstimmung v erbleibt nur der Wirkstrom.

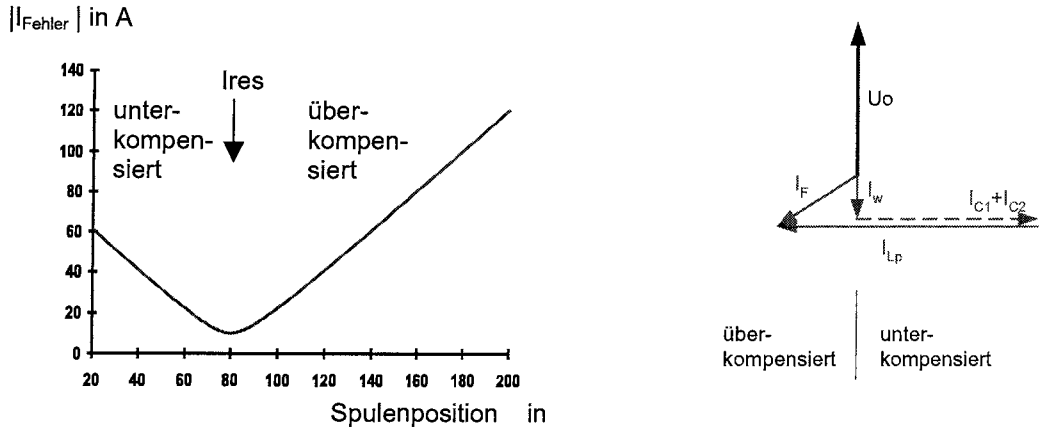


Bild 12: Stromverlauf über die Fehlerstelle bei niederohmigem Erdschluss

Der Vorteil dieser Ersatzschaltung ist, dass diese auch für den gesunden Netzbetrieb gültig ist. Im gesunden Betrieb ist $\underline{Y}_u$ durch die kapazitive Unsymmetrie des Netzes gegeben. Bei sehr symmetrischen Netzen wird $\underline{Y}_u$ sehr klein. Der Resonanzkreis wird vom Mitsystem entkoppelt. Eine Bestimmung der Resonanzkreises durch messen der Verlagerungsspannung ist nicht mehr möglich.

Für die weitere Berechnung wird das folgende Ersatzschaltbild von Bild 13 verwendet:

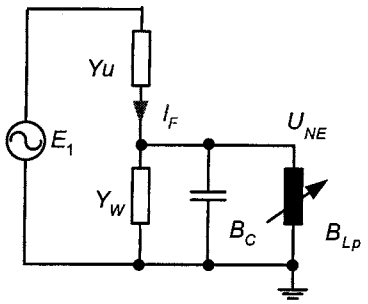


Bild 13: Vereinfachtes Ersatzschaltbild für die Berechnung

Der komplexe Verlauf der Verlagerungsspannung kann mit der folgenden Formel beschrieben werden.

$$\underline{U}_{res} = - \frac{\underline{Y}_U}{\underline{Y}_U + \underline{Y}_W} \underline{E}_1 \qquad (1)$$

bzw.:

$$\left|\frac{\underline{U}_{ne}}{\underline{U}_{res}}\right| = \frac{1}{\sqrt{2}} = \left|\frac{1}{1+\dfrac{j(B_C - B_{L,W})}{\underline{Y}_U + \underline{Y}_W}}\right| \tag{2}$$

$$\approx \left|\frac{1}{1+\dfrac{j(B_C - B_{L,W})}{\underline{Y}_W}}\right| \tag{3}$$

Die Resonanzkurve kann durch drei Punkte vollständig beschrieben werden:

- Ires: Spulenstellung mit dem Spannungsmaximum. Der kapazitive Stromanteil ist vollständig kompensiert.
- Ures: Maximum der Verlagerungsspannung im Resonanzmaximum
- Iw: Dämpfung der Resonanzkurve in A.

Der Wirkstrom kann durch den Punkt ermittelt werden, bei dem der Abfall der Verlagerungsspannung auf das $1/\sqrt{2}$ fache erfolgt. In diesem Punkt ist der Blindstrom und der Wirkstrom gleich groß. Die Phasenverschiebung zum Strom $\underline{I}_F$ durch die Unsymmetrie ist 45°.

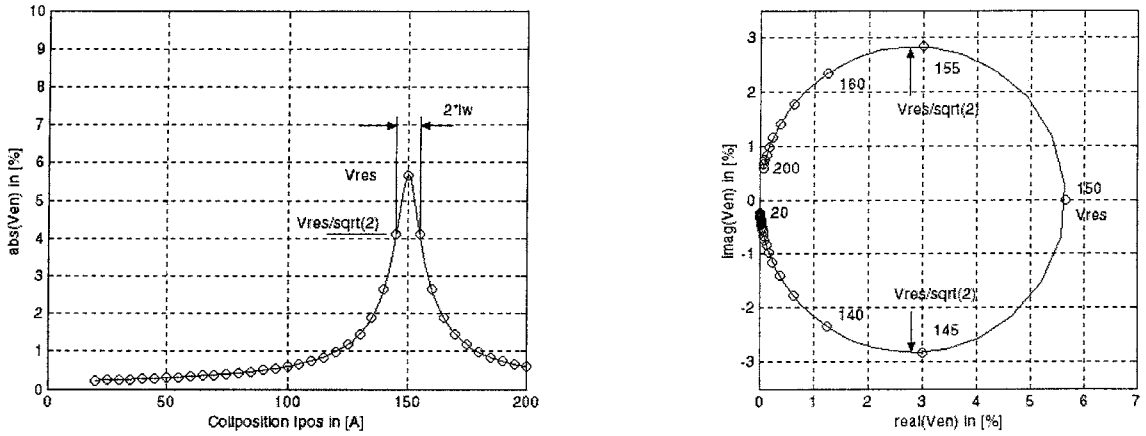


Bild 14: Resonanz- und Ortskurve des gesunden Netzes

Wird anstelle der Verlagerungsspannung die Inverse der Verlagerungsspannung aufgetragen, so ergibt sich die Formel (4) und die folgenden zugehörigen Bilder.

$$\frac{1}{\underline{U}_{ne}} = -\frac{\underline{Y}_U + \underline{Y}_W + j(B_C - B_L)}{\underline{Y}_U \underline{E}_1} \tag{4}$$

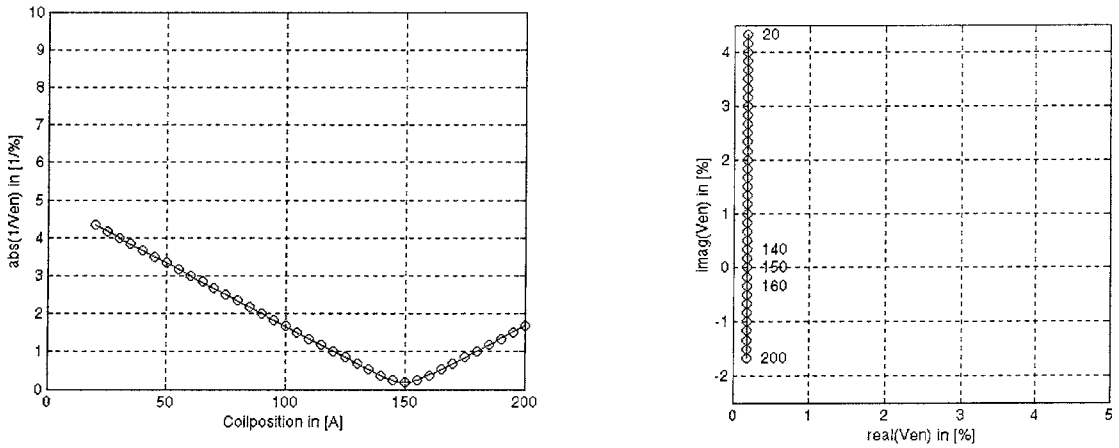


Bild 15: Inverse der Verlagerungsspannung als f(Ipos) und Ortskurve

Der Vorteil dieser Darstellung ist, dass wesentlich leichter und robuster eine Parameterschätzung für den Resonanzkreis [3], [9] durchgeführt werden kann.

Auf dieser Basis können die unterschiedlichen Regelungsverfahren verglichen werden. Außerdem wird ein neues Verfahren zur Berechnung der Parameter des Resonanzkreises vorgestellt, welches die Nachteile der bisher bekannten Verfahren umgeht.

3 Regelungsverfahren

Die Abstimmung der Petersen-Spule muss als Präventivmaßnahme bereits im gesunden Netz erfolgen. Eine Bestimmung der Parameter in einem Netz mit einem niederohmigen Erdschluss ist nicht mehr möglich. Es sind der Ort des Fehlers und der Übergangswiderstand an der Fehlerstelle unbekannt und einer Messung nicht zugänglich. Im Falle eines niederohmigen Fehlers ist die Verlagerungsspannung eingeprägt und die Messung des Nullstromes ist nur an der Sammelschiene möglich.

In der Vergangenheit haben sich die in diesem Abschnitt beschriebenen Regelungsverfahren entwickelt. Den meisten Verfahren ist gemeinsam, dass eine Verstellung der Petersen-Spule notwendig ist. Durch die immer besser werdende Symmetrie des Netzes werden die Nutzsignale immer kleiner. Anderseits wird das Übersprechen des Mitsystems auf das Nullsystem im gesunden Netz immer größer. Die Laststromänderungen, die auch ein Übersprechen bewirken, führen zum Auslösen eines Suchvorganges. Durch die Störungen werden einerseits die Bestimmung der Parameter erschwert und anderseits teilweise eine korrekte Abstimmung nicht mehr ermöglicht. Durch die Störungen werden immer häufigere und größere Verstellungen der Petersen-Spule notwendig. Die Auslegung der Motor-Antriebe der Petersen-Spulen ist aber nicht für derart häufige Regelvorgänge vorgesehen. Angestrebt werden maximal 10 Regelungen pro Tag.

Aus diesem Grund werden hier zwei Verfahren vorgestellt, mit denen eine Berechnung der Abstimmung einerseits **ohne Verstellung der Petersen-Spule** und anderseits **auch bei vollständig symmetrischen Netzen** erfolgen kann.

Die folgenden Verfahren zur Bestimmung der aktuellen Abstimmung werden bei den bisherigen Reglern zur Abstimmung der Petersenspule verwendet:

- Suche nach dem Maximum des Betrages von $\underline{U}_0$
- Parameterschätzung aus der Resonanzkurve
- Parameterschätzung aus der Inversen der Resonanzkurve
- Berechnung der Ortskurve ohne Stromeinspeisung
- Berechnung der Verstimmung mit 50 Hz Stromeinspeisung
- Auswertung des Ausgleichsvorganges nach einem Erdschluss

Die neuen Verfahren sind:

- Berechnung der Verstimmung mit Frequenzumrichter mit **veränderlichen** Frequenzanteilen
- Berechnung der Verstimmung mit Frequenzgenerator mit **fixen** Frequenzanteilen

Auf den folgenden Seiten erfolgt eine kurze Beschreibung der Verfahren, sowie eine Gegenüberstellung deren Vor- und Nachteile.

3.1 Bekannte Berechnungs- und Regelungsverfahren

Suche nach dem Maximum des Betrages von $\underline{U}_0$

Mit den ersten analogen Reglern wurde durch Verstellen der Petersen-Spule zuerst eine positive Steigung und danach das Maximum der Resonanzkurve gesucht. Im Resonanzmaximum wurde dann eine zeitabhängige Verstellung in Richtung der gewünschten Kompensation durchgeführt. (EA76, SRK)

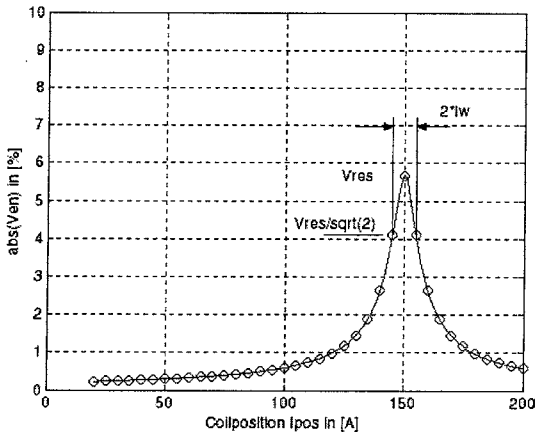


Bild 16: Resonanz- und Ortskurve des gesunden Netzes

Vorteile:

- Einfaches Verfahren
- Analogspeicher und einfache Zeitglieder sind für die Abstimmung ausreichend

Nachteile:

- Eine ausreichende Verlagerungsspannung ist notwendig
- Störanfällig bei kleinen Verlagerungsspannungen
- Probleme bei der Erkennung von Schalthandlungen während der Regelung => Mehrmaliger Suchvorgang ist erforderlich
- Erkennt nicht alle Schalthandlungen im abgestimmten Zustand

Parameterschätzung aus Resonanzkurve

Die Resonanzkurve ist prinzipiell durch drei Punkte festgelegt. Allerdings ist eine exakte Erfassung der Spulenstellung und der Verlagerungsspannung erforderlich. Bei diesem Verfahren wird angenommen, dass die Petersenspule ein konstantes Übersetzungsverhältnis über den gesamten Verstellbereich der Spule hat.

Vorteile:

- Schalthandlungen werden bereits während der Abstimmung erkannt.

Nachteile:

- Schwierigkeiten in der Umgebung des Wendepunkt der Resonanzkurve
- Empfindlich auf Störsignale
- Aufwendige Schätzalgorithmen

Parameterschätzung aus der Inversen der Resonanzkurve

Um einfachere und robustere Algorithmen zur Parameterschätzung verwenden zu können wird die Inverse der Verlagerungsspannung verwendet [3]. Dieses Verfahren ist wesentlich robuster als die Parameterschätzung mit der Resonanzkurve.

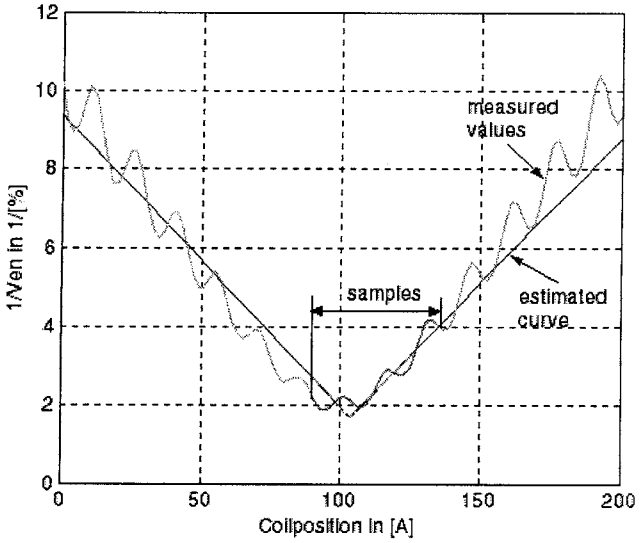


Bild 17: Inverse der Verlagerungsspannung mit Störungen

In den Bereichen weit entfernt vom Resonanzpunkt kann die Parameterschätzung mit einem Ausgleichspolynom 1. Ordnung (Gerade) und in der Nähe des Resonanzpunktes mit einem Ausgleichspolynom 2. Ordnung (Hyperbel) erfolgen.

Vorteil:

- Geringerer Rechenaufwand
- Robuster gegen Störsignale

Nachteil:

- Eine ausreichende Verlagerungsspannung ist notwendig

Berechnung der Ortskurve ohne Stromeinspeisung

Eine kurzzeitige Verstimmung des Resonanzkreises ist z.B. mit Hilfe eines Verstimmungskondensators möglich. Durch die definierte Verstimmung erhält man einen zweiten Punkt auf der Ortskurve. Der dritte Punkt der Ortskurve liegt im Koordinatenursprung.

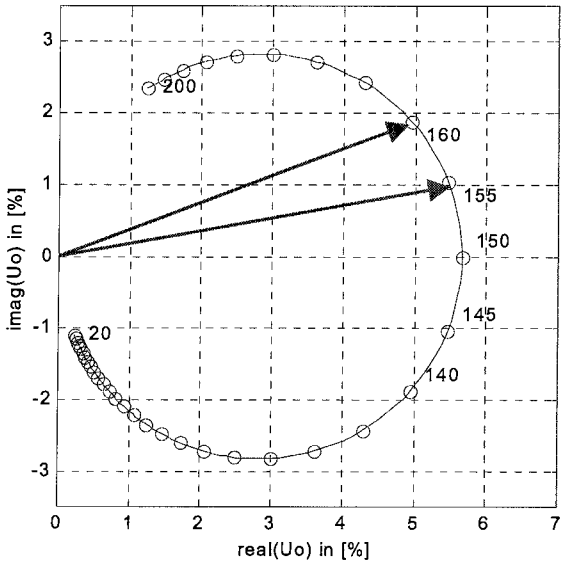


Bild 18: Bestimmung der Ortskurve mit Verstimmungskondensator

Vorteile:

- Schnelles Verfahren zur Berechnung

Nachteile:

- Eine ausreichende Verlagerungsspannung ist notwendig
- Exakte Betrags- und Winkelmessung bei kleinen Spannungen ist erforderlich
- Es erfolgt eine kurzzeitige Verstimmung durch Verstellen der P-Spule oder durch Zuschalten eines Kondensators
- Eine große Kapazität ist erforderlich
- Bei Verstimmung mit Hilfe einer Spule ist zu berücksichtigen, dass die ohmschen Anteile abhängig von der Spulenstellung sind
- Es werden konstante Verhältnisse für die Unsymmetrie, die Einstreuung und die Abstimmung während der Berechnung vorausgesetzt. Das Einsschwingen der Verlagerungsspannung muss abgewartet werden.

Berechnung der Verstimmung mit 50 Hz Stromeinspeisung

Bei sehr symmetrischen Netzen kann die notwendige Verlagerungsspannung mit Hilfe einer Stromeinspeisung direkt in den Sternpunkt des Netzes erzeugt werden. Zusätzlich kann durch zwei unterschiedliche Arbeitspunkte der Stromeinspeisung die aktuelle Verstimmung des Netzes ermittelt werden. Dazu ist eine komplexe Messung des Einspeisestromes vor und während einer Stromeinspeisung erforderlich.
Die Petersen-Spule kann dabei als Einphasentransformator verwendet werden.

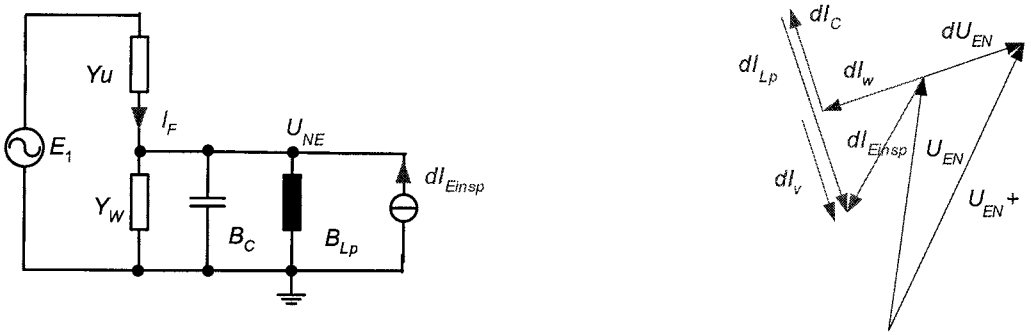


Bild 19: Berechnung der Abstimmung mit Hilfe der Stromeinspeisung

$$\underline{Yo} = \frac{d\underline{I}_{Einsp}}{d\underline{U}_{EN}} = Y_W + j(B_C - B_{Lp}) \tag{5}$$

Vorteile:

- Schnelles Verfahren zur Berechnung der Verstimmung
- Für sehr symmetrische Netze geeignet

Nachteile:

- Es werden konstante Verhältnisse für die Unsymmetrie, die Einstreuung und die Abstimmung während der Berechnung angenommen. Das Einsschwingen der Verlagerungsspannung muss abgewartet werden.
- Probleme bei großer Verstimmung; eine variable Stromeinspeisung ist erforderlich
- Große Probleme bei schwankendem $\underline{U}_{NE}$ verursacht durch Übersprechen aus dem Mitsystem

Auswertung des Ausgleichsvorganges nach einem Erdschluss

Wenn der Lichtbogen an der Fehlerstelle verlischt, schwingt der Resonanzkreis des Nullsystems mit seiner Eigenfrequenz aus. Aus dem Verlauf der Verlagerungsspannung ist eine Parameterschätzung möglich.

Der wesentliche Nachteil liegt darin, dass die Berechnung der Abstimmung erst nach dem Ende des Erdschlusses erfolgt. Die richtige Einstellung der Petersen-Spule ist jedoch eine Präventivmaßnahme für einen möglichen Erdschluss. Dieses Verfahren ist daher nur für eine nachträgliche Analyse des Abstimmungszustandes am Ende des Erdschlusses bzw. kurz vor Ende des Erdschlusses geeignet.

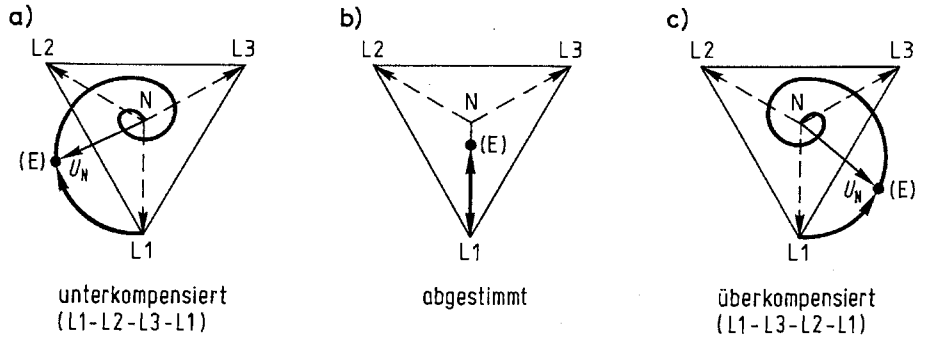


Bild 20: Ausgleichsvorgang am Ende des Erdschlusses

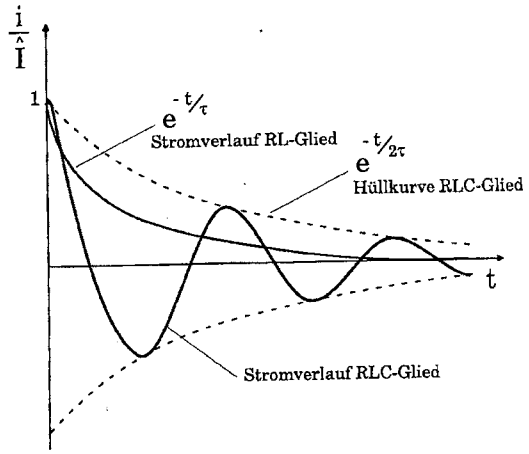


Bild 21: Ausgleichsvorgang der Verlagerungsspannung am Ende des Erdschlusses

Vorteile:

- Analyse des Schaltzustandes am Ende des Erdschlusses möglich. Auswirkungen von Schalthandlungen im Netz sind nachvollziehbar.

Nachteile:

- Erst am Ende des Erdschlusses durchführbar
- Transiente Aufzeichnung ist erforderlich
- Frequenz schwer bestimmbar
- Empfindlich auf Störsignale, bzw. Unsymmetrie des gesunden Netzes

3.2 Neue Berechnungsverfahren

Berechnung der Verstimmung mit Frequenzumrichter mit veränderlichen Frequenzanteilen

Aus den bisherigen Verfahren ist erkennbar, dass die Verlagerungsspannung durch die Unsymmetrie des Netzes hervorgerufen wird. Es wird bei den einzelnen Verfahren angenommen, dass das Übersprechen aus dem Mitsystem während der Messung konstant ist. Es gibt aber in der Realität einige Situationen wie z.B. im Einflußbereich von Schwerindustrie, wo einerseits sehr symmetrische Netze und anderseits starke Lastschwankungen vorliegen. In [3] wurde gezeigt, dass auch bei sehr symmetrischen kapazitiven Leitungen durch induktive Kopplungen ein Übersprechen des Laststromes erfolgt.

Diese Probleme können umgangen werden, wenn die Messung bei einer von 50 Hz abweichenden Frequenz erfolgt und die 50 Hz Komponenten unterdrückt werden.

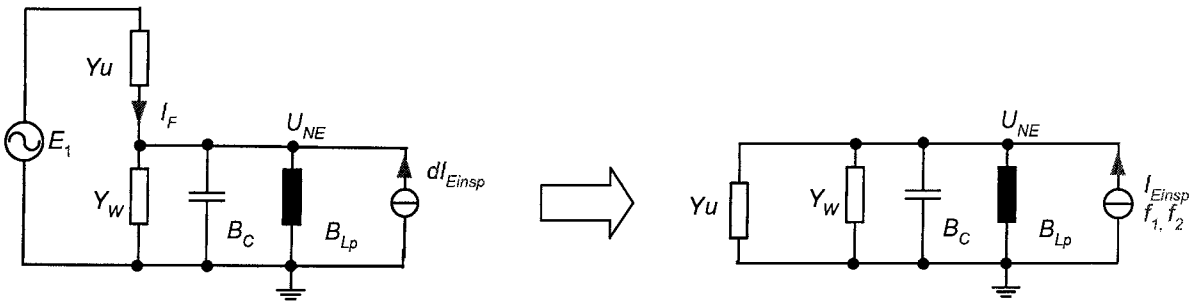


Bild 22: Ersatzschaltung für Frequenzen ≠ 50 Hz

Bei den neuen Verfahren erfolgt die Parameterschätzung des Resonanzkreises durch Bestimmung der Admittanzen bei zwei von 50 Hz unterschiedlichen Frequenzen. Für diese Frequenzen bildet das Mitsystem einen Kurzschluss, sodass die Unsymmetrieadmittanz zum Resonanzkreis parallel geschaltet wird. Der dadurch verursachte Fehler ist aber vernachlässigbar. Die Probleme sind, wie oben gezeigt wurde, im Bereich von sehr symmetrischen Netzen aufgetreten.

Durch das Ausweichen auf andere Frequenzen ist auch eine Unterdrückung der schwankenden Unsymmetrie bzw. des Übersprechen des Laststromes möglich.

Ein weiterer Vorteil liegt darin, dass mit einer einstellbaren Stromamplitude die notwendige bzw. maximale Änderung der Verlagerungsspannung, verursacht durch die Stromeinspeisung, überwacht und gesteuert werden kann. Gewünschte bzw. geforderte Randbedingungen können eingehalten werden. Eine Anpassung an unterschiedlichste Resonanzkurven ist leicht möglich.

Ein Problem bei der Berechnung der Abstimmung liegt auch darin, dass bei großen Verstimmungen nicht genügend Energie zur Verfügung steht, um die notwendige Änderung der Verlagerungsspannung zu erreichen. Mit dem neuen Verfahren besteht die Möglichkeit die Frequenz so zu ändern, dass die Einspeisung mit einer Frequenz in der Nähe des Resonanzpunktes erfolgt. Dadurch sind auch mit kleinen Strömen bereits genügend große Spannungsänderungen erzeugbar. Eine Verstellung der Petersen-Spule ist hierzu nicht erforderlich.

Durch die gleichzeitige Einspeisung und Messung von zwei Frequenzen erfolgt eine Berechnung zum gleichen Zeitpunkt. Ein stationärer Zustand wird nur über den Zeitrum von 0,2 ms gefordert, im Vergleich zu den 5 bis 20 Sekunden der bisher bekannten Verfahren.

Die Berechnung der Abstimmung erfolgt mit Hilfe der folgenden Gleichung.

$$\underline{Y}_{N\_f} = \frac{\underline{I}_f}{\underline{U}_{NE\_f}} = \frac{1}{R} + \frac{1}{j\omega_1 L_p} + j\omega_1 C_E \qquad (6)$$

Die Admittanzen werden bei zwei unterschiedlichen Frequenzen bestimmt. Durch Auswertung der Real- und Imaginärteile können die einzelnen Parameter des Modells bestimmt werden. Durch Beobachtung über einen längeren Zeitraum können mit Hilfe von Least Square Verfahren die zeitlichen Schwankungen minimiert werden.

Das folgende Bild zeigt eine Umsetzung des Algorithmus in Hardware

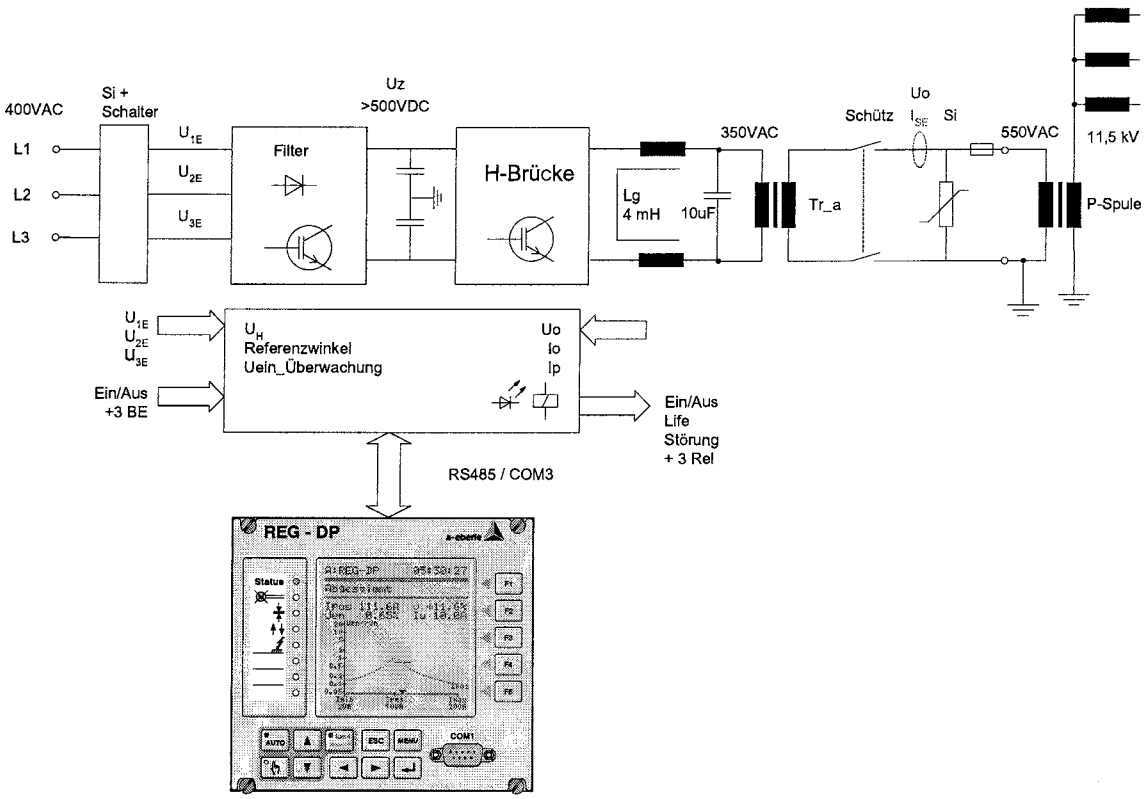


Bild 23: Realisierung der Stromeinspeisung mit Frequenzen ≠ 50 Hz

Vorteile:

- Keine Probleme mit der Berechnung auch bei großer Verstimmung; automatische Anpassung der Amplitude des Einspeisestromes ist möglich.
- Quasistationärer Zustand nur für 0,2 s gefordert
- Für sehr symmetrische Netze geeignet
- Automatische Anpassung des Einspeisestromes zur Einhaltung von Randbedingungen (Amplitude, Frequenz)
- Automatische Anpassung an unterschiedliche Resonanzkurven
- Robust gegen Übersprechen des Mitsystems (kapazitiv, induktiv)
- Power-Faktor Korrektur auf der Primärseite der Frequenzumrichters

Nachteile:

- Einphasen-Frequenzumrichter mit Vier-Quadranten Betrieb ist erforderlich

Berechnung der Verstimmung mit Frequenzgenerator mit fixen Frequenzanteilen

Wird auf eine variable Frequenz verzichtet, da keine Einstellung mit großer Verstimmung gefordert wird, so ist eine einfachere Schaltung ausreichend. Durch eine geeignete Ansteuerung der Thyristoren sind zwei zusätzliche Frequenzen ungleich 50 Hz erzeugbar.

Mit diesem Verfahren bleibt die Möglichkeit zur Einstellung der Stromamplitude erhalten.

Es gelten daher, bis auf die freie Frequenzwahl, die gleichen Vor- und Nachteile wie beim vorherigen Verfahren.

Im folgenden Bild ist eine Stromeinspeisung mit variabler Amplitude und zwei fixen Frequenzen dargestellt.

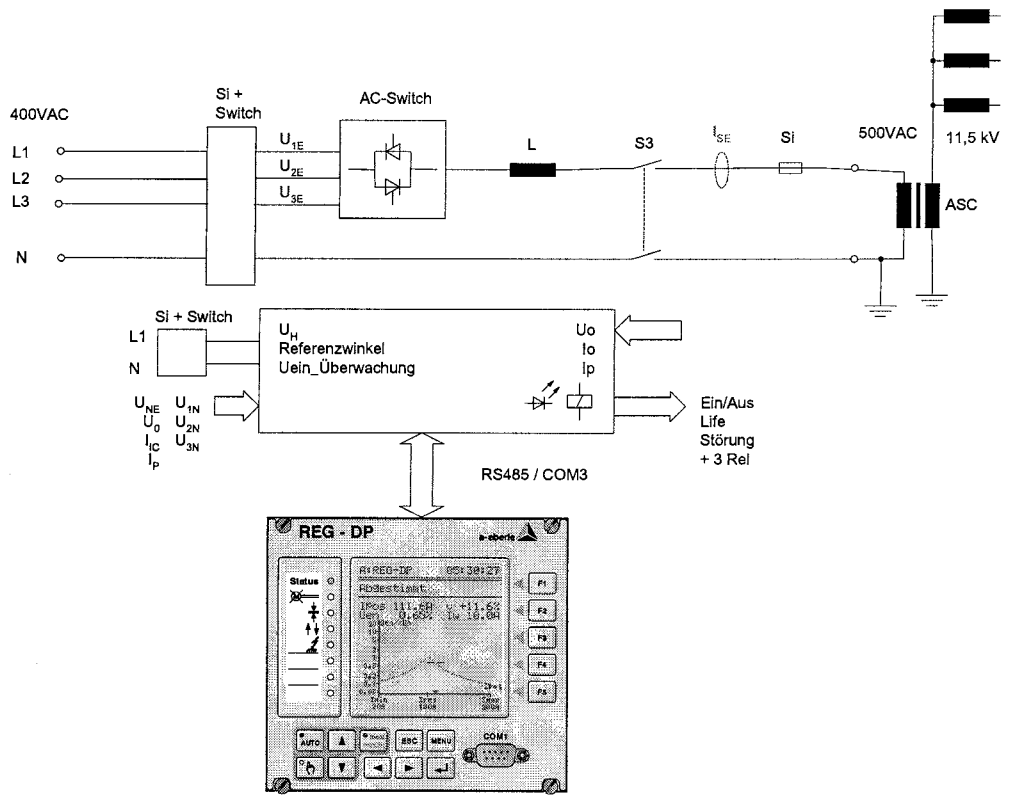


Bild 24: Realisierung der Stromeinspeisung mit fixen Frequenzen ≠ 50 Hz

4 Zusammenfassung

In diesem Beitrag wurden zwei neue Verfahren zur Ermittlung der aktuellen Abstimmung eines gelöschten Netzes vorgestellt. Die Berechnung der Verstimmung erfolgt dabei ohne Verstellung der Petersen-Spule und ist auch für symmetrische Netze geeignet. Durch die variable Stromeinspeisung erfolgt eine automatische Anpassung an unterschiedliche Netzformen. Eine Berechnung der aktuellen Verstimmung ist auch bei stark über- oder unterkompensierten Netzen möglich. Übliche Störeinflüsse auf die Verlagerungsspannung, verursacht durch Übersprechen des Mitsystems, werden bei diesen neuen Verfahren unterdrückt. Sowohl die Anzahl der erforderlichen Verstellungen der Petersen-Spule als auch die Zeit der Fehlabstimmungen werden durch die neuen Verfahren stark reduziert.

Die ersten Feldversuche verliefen erfolgreich und haben die Effizienz dieser neuen Verfahrens aufgezeigt.

5 Literaturverzeichnis

[1] Bergeal J., Berthet L., Grob O., Bertrand P., Lacroix B., *Single-phase faults on compensated neutral medium voltage networks*, CIRED93, vol.2, 2.9.1-2.9.5., 1993

[2] DIN VDE 0228, *Maßnahmen zur Beeinflussung von Fernmeldeanlagen durch Starkstromanlagen,* 1987.

[3] Druml G., Kugi A., Parr B., *Control of Petersen Coils*, XI. International Symposium on Theoretical Electrical Engineering, 2001, Linz

[4] Druml G., Kugi A., Seifert O., *A new directional transient relay for high ohmic earth faults*, CIRED 2003, vol 3, paper 3.50

[5] Druml G., *Resonanzregler REG-DP, Betriebsanleitung,* A.Eberle GmbH, Nürnberg, 2002.

[6] Grainger J., Stevenson W., Jr., *Power System Analysis*, McGraw-Hill, Singapore, 1994.

[7] Herold Gerhard, *Elektrische Energieversorgung III*, J.Schlembach Fachverlag, Weil der Stadt, 2002

[8] Koettnitz H., Pundt H., *Berechnung elektrischer Energieversorgungsnetze: Mathematische Grundlagen und Netzparameter,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig ,1973

[9] Ljung L., *System Identification: Theory for the User*, Prentice Hall, New Jersey, 1987

[10] Willheim R., Waters M., *Neutral Grounding in High-Voltage Transmission,* Elsevier Publishing Company, London, 1956.

DIE AUTOMATISCHE KUNSTUHR VON GAZA

Prof. Dr. Dimitri Kalligeropoulos, Dr. Soultana Vasileiadou
Fachhochschule Piraeus

In den Abhandlungen der Königlich Preussischen Akademie der Wissenschaften, Jahrgang 1917, steht unter dem Titel: "Über die von Prokop beschriebene Kunstuhr von Gaza", ein interessanter Artikel von Hermann Diels, dem bekannten deutschen Philologen, vor. Er beinhaltet einen seltenen griechischen Ausschnitt über eine komplizierte jedoch bemerkenswerte und monumentale automatische Uhr der Antike, die wir hier zu beschreiben und zu rekonstruieren versuchen werden.

Wir befinden uns am Anfang des 6ten n.Chr. Jahrhunderts. Es ist die Zeit des Kaisers Iustinians I (527 - 565 n. Chr.), der den Aufschwung großer technischer Werken im byzantinischen Reich beförderte: Es war die Konstruktion von Agia Sophia (532 - 537 n. Chr.) in Konstantinopel, die Konstruktion von großen Straßen, Mauern und Festungen im ganzen Imperium, die Konstruktion der ersten Seidenfabriken auf der langen Seidenstraße: Konstantinopel, Antiocheia, Tyros, Beirut.

Beschreibungen dieser großen technischen Werken schrieb auf griechisch ein Geschichtsschreiber,Jurist und Sekräter des byzantinischen Generals von Mesopotamien Belisars und spätterer Rat des Kaisers Iustinians, Prokop aus Kaesareia Palaestinas (490 - 562 n. Chr.).

In seinem Werk «Περί Κτισμάτων» (Aedificia, Über Bauten) beschreibt Prokop, sogar nach persönlichem Auftrage des Kaisers, die wichtigsten Baudenkmäler des byzantinischen Imperiums. In einem Ausschnitt aus diesem Werke ist uns die Beschreibung der "Kunstuhr von Gaza" unvollendet erhalten geblieben. Es handelt sich hier um ein monumentales Gebäude auf dem zentralen Platz der palaestinensischen Stadt, ein Gebäude jedoch das mit einem besonders komplizierten Uhrmechanismus ausgerüstet war.

Es ist schwer das Monument selbst zeitlich zu identifizieren. Es gehört wahrscheinlich zu dieser iustinianischer Periode des technologischen Aufschwungs im byzantinischen Reich (Anfang 6ten Jh. n.Chr.), aber es kann auch älter sein. Es hat jedenfalls seine Nachfolger und seine Vorfahren.

Mehr als vier Jahrhunderte später, während der Zeit der Kaiser Theophilos (829 - 842 n.Chr.) und Konstantinos Porphypogennetos (913 - 959 n.Chr.), finden wir Beschreibungen byzantinischer Uhrwerken, mit komplizierten automatischen Mechanismen ausgerüstet, wie auch jenen berühmten automatischen Solomonischen Thron von Magnaura.

Die Vorgeschichte aber der Konstruktion von Uhren, nämlich der durch Simulation von dynamischen Prozessen quantitativen Zeitmessung, geht ziemlich weit zurück. Sie weitet sich in der vorrigen hellenistischen Periode, wie auch in der älteren altgriechiscen Technik der Kostruktion von Uhrwerken.

Diese Technik kann man in drei Stadien aufteilen.

Das erste Stadium ist das der Sonnenuhren. Modell für die Zeitmessung ist der dynamische Bewegungsprozeß der Sonne. Instrument der "Gnomon", ein Stab auf einer Marmorplatte, die die Gravierungen der zwölf gleichlangen "Hores" (Stunden) zwischen Sonnenauf- und Sonnenuntergang enthält.

Das zweite Stadium ist das der hydraulischen Uhren. Modell hierfür ist der Prozeß des Wasserflußes. Erstes Instrument für die Messung von Zeitspannen die "Klepsydra". Gleich danach die großen klepsydraähnlichen Wasseruhren, die einen Schwimmer in einem Wasserspeicher langsam sinken und die Stunden auf einer Tafel anzeigen ließen.

Das dritte Stadium ist jedoch für uns von besonderer Bedeutung. Es ist das Stadium der Ausrüstung dieser Wasseruhren mit besonderen hydraulischen oder mechanischen Regelungsmechanismen für die Regelung des Wasserpegels und –flußes und die Sicherstellung einer gleichmässigen Bewegung der Uhrzeiger, wie auch der vielen komplizierten Beiwerken («Πάρεργα»), die die Uhren schmückten. Solche Regelmechanismen finden wir bei der Nachtuhr von Platon, die als Wecker diente und nach einer vorbestimmten Zeitspanne zu pfeifen anfing, bei der ersten Wasseruhr des hellenistischen Mechanikers Ktesibios (um 300 v. Chr.), die einen Behälter konststanten Flußes enthielt, bei dem komplizierten Uhrwerk von Archimedes (287 - 212 v. Chr.), wie auch bei den Wasserpegelregelungen von Philon aus Bysanz (um 250 v. Chr.) und von Heron aus Alexandria (um 100 v. Chr.).

Diese Mechanismen wurden, auch während der hellenistischen Zeiten, mit Zahnrädern und anderen mechanischen Transformationsmethoden ausgerüstet und führten zu mechanischen Modellen großer Komplexität, wie das Planetarium von Archimedes oder der gefundene Antikythera Mechanismus (um 80 v. Chr.). Sie

öffneten somit den Weg für die Konstruktion von vollmechanisierten Uhrwerken, wie die der islamischen Technik und des europäischen Mittelalters.

Die Beschreibung der Kunstuhr von Gaza, die uns hier vorliegt, beinhaltet keine Andeutung über ihre Mechanismen. Sie beschränkt sich nur auf ihren äußeren wundervollen Anblick und ihre verschiedenen Funktionen. Das erlaubt uns jedoch anzunehmen, daß wir hier mit einem Mechaniker zu tun haben, der die altgriechische und hellenistische Tradition der Automatentheater, wie auch die hydraulischen und mechanischen Regelmechanismen seiner Vorgänger vollausgenutzt hat.

Folgen wir nun die Worte dieser Beschreibung.

"Προκοπίου Σοφιστού Γάζης – Έκφρασις Ωρολογίου
Des Sophisten Prokopios aus Gaza – Beschreibung eines Uhrwerkes"

Die Beschreibung Prokops fängt mit einem Vorwort, das die Beziehung zwischen Worten und Werken betrifft, an. Und erzählt dann uralte technische Wunder, wie die ägyptischen Pyramiden, der babylonische Turm von Babel und die homerischen Automata von Hephaistos. Im Vergleich zu diesen technischen Wundern stellt Prokop folgendes für die vorliegende Uhr von Gaza fest:

"Dies galt mir nun freilich bis jetzt für eine Fabel und ein Märchen. Homer schien mir in der Kunst zu schwelgen, Dinge, die sich nie und nirgendwo begaben, straflos zu berichten. Wenn ich nun aber diese Kunstwerke unseres hier anwesenden Hephaistos betrachte..., muß ich zugeben, daß man auch jenen Wirklichkeit zugestehen dürfe."

Die eigentliche Beschreibung des Bauwerkes fängt Prokop mit einer Darstellung des umgebenden Raumes an.

"Da steht im Mittelpunkt der Stadt ein Bau mäßigen Umfangs. Gegenüber befindet sich die Königshalle, zur Linken ein freier Platz, im Sommer der Tummelplatz zahlloser Menschen.,,

Es handelt sich nun hier um eine große öffentliche Uhr, mitten im Zentralplatz der Stadt aufgebaut. Es besteht kaum einen Zweifel, daß sie eine hydraulische Uhr war, Nachfolger deren von Ktesibios und Archimedes. Die notwendige Infrastruktur mußte

nun eine Anlage für die Wasserförderung der Uhr aus einer überliegenden Zisterne, sowie eine für die Wasserabführung in einen nahefließenden Fluß enthalten.

Die Architektur des Bauwerkes wird folgenderweise beschrieben:

"Vor dem Gebäude steht ein doppeltes Säulenpaar. Die Säulen sind verschieden verteilt: die einen liegen nach Osten, die andern nach Westen. Sie bewirken, daß an das Gebäude in einem ziemlichen Abstand im Hintergrunde erblickt, so daß niemand eine Störung des Schauwerkes bewirken kann. Den Zwischenraum zwischen den Säulen füllen Marmorplatten aus. In diese sind eiserne Spitzen eingelassen, die solche, die etwa der Ehrgeiz treibt, frech die Schranken zu überklettern, zurückhalten."

Die Beschreibung der automatischen Einzelmechanismen fängt von oben an:

"Aber da ist auch eine Gorgo, die von oben herab mit fürchterlichem Blicke allen droht, die sich mit kühneren Absichten zu nahen wagen, indem sie zu allen Stunden des Tages die Augen verdreht..."

Diese Gorgo mit den beweglichen Augen errinnert uns an die antiken Schaustücken der stehenden Automaten, die Heron aus Alexandria in seine Automatentheater erwähnt:

"Die Aufführung, welcher sich die Alten bedient haben, ist ganz einfach. Wurde nämlich die Bühne geöffnet, so erschien darauf eine gemalte Maske. Diese bewegte die Augen, machte sie oft zu und wieder auf."

Es handelt sich nun um einen bekannten antiken Mechanismus von programmierten sich wiederholenden Bewegungen, die üblicherweise mit Schnurwicklungen bewirkt wurden.

Die Beschreibung fährt mit der Darstellung der Türen der Tagesstunden und der entsprechenden Türen des Nachtuhrwerkes fort.

"Da sind im Vordergrunde Türen angebracht. Ich glaube, die Türen des Tages verbergen etwas im Hintergrund. Die Türöffnungen der Nacht liegen zwar höher. Ich beschreibe sie aber noch nicht..."

Nach der Tradition der Sonnenuhren, waren die altgriechischen Tagesstunden (je 1/12 der Zeitspanne zwischen Sonnenauf- und Sonnenuntergang) von den Nachtstunden, wie auch die Sommer- von den Winterstunden, verschiedenlang. Die antiken Uhrwerke mußten sich also dieser Tradition anpassen.

Uhrwerke jedoch mit Türen als Anzeigen jeder Stunde haben, von technischer Ansicht her, ein doppeltes Interesse. Erstens, handelt es sich um eine digitale Darstellung der Zeit. Und zweitens, beinhalten sie zweifellos Mechanismen für die Wasserfluß- und somit die Geschwindigkeitsregelung der beweglichen Teile, Mechanismen die die lineare Funktion des Uhrwerkes gewährleisten.

Fahren wir nun mit der Beschreibung fort:

"Da stehen in einer Reihe eherne Adler, deren Zahl mit den darunter befindlichen Stunden übereinstimmt. Alle tragen Kränze. Sie schmücken nicht ihr eigenes Haupt, sie bezeichnen keinen eigenen Sieg. Vielmehr vereinigen sich die Spitzen ihrer Krallen und umklammern die Kränze. Jeder wartet gespannt auf den Augenblick, wenn der unter ihnen befindliche Herakles aus den ihn verdeckenden Türen heraustritt, sobald Helios, der davor steht, an ihnen vorbeigeht. Denn er ist es, der durch sein Vorbeischreiten die jedesmaligen Stunden abmißt."

Die gleichmäßige Bewegung des Sonnengottes Helios vor den diskreten Stundentüren simuliert hier die analoge, die kontinuierliche Darstellung der Zeit, sodaß wir sagen können, daß dieses Uhrwerk von Gaza gleichzeig analog und digital ist.

Der Mechanismus für die gleichmäßige Bewegung der Figur Athenas im letzten Akt des stehenden Automaten, das Heron in seine Automatentheater beschreibt, kann als Beispiel für die gleichmäßige geradlinige Bewegung des Helioswagens bei dieser Wasseruhr von Gaza gelten.

Die zwölf digitalen Tagesstunden dagegen werden durch zwölf Heraklesfiguren, den zwölf Herakleskämpfen ensprechend, dargestellt:

"Gespannt wartet also der Adler, bis der Zeussohn Herakles aus den ihn verdeckenden Türen heraustritt, um die Stunde anzusagen. Die erste Heraklesfigur kündet die erste, die übrigen der Reihe nach die übrigen Stunden an. Zwölf Stunden sind es und alle erscheinen in Herakles Gestalt...

...Das sind die Mühen und Arbeiten des Herakles. Darum die Stunden, die Kränze und das künstlich beflügelte Erz."

Jede stündliche Erscheinung einer solchen Heraklesfigur wird von einem besonders komplizierten automatischen Beiwerk begleitet:

"Wenn er nun die ehernen Türen aufstößt und mit der Kampfesbeute erscheint, folgt ihm von oben sofort ein Adler. Er entfallet seine Schwingen und legt mit beiden Fängen den passenden Kranz auf das passende Haupt. Dann verweilt er dort kurze Zeit, als ob er an dem Haupt des Heros sich laben wollte. Darauf überläßt er dem Herakles den Kranz, öffnet die beiden Fänge, schwingt sich in die Höhe und nimmt seinen alten Platz ein, legt seine Schwingen wieder an die Seiten an und zieht sich auf sich selbst zurück...

...Nunmehr verneigt sich Herakles vor den Kränzen, um sich allen zu zeigen wie mitten auf der Rennbahn. Dann geht er wieder auf seinen Platz zurück."

Hierzu eine entsprechende Wasseruhr in Fom einer Burg, mit Türen für die Tages- und die Nachtstunden, und verschiedene Beiwerke, die der Araber Mechaniker Al Jazzari im 13ten n. Chr. Jahrhundert in seinem Buch Kitab al Hiyal al Handasiyya beschreibt.

Die optische Darstellung der Stundenfolge, durch die verschiedenen Türen, die Heraklesfiguren, die Adler und den Helioswagen, werden in der Uhr von Gaza mit einer akustischen Ankundigung der digitalen Stunden durch Herakles' Trommelschlägen ergänzt.

"Auf einem andern, die Mitte einnehmenden Platze, der geräumig und dem übrigen vorgebaut ist und dadurch die Aufmerksamkeit auf sich zu ziehen weiß, erscheint der Held wiederum. Noch wächst kein Flaum ihm um die Wangen. Nackt steht er da. Nur an den Schultern ist die Löwenhaut befestigt, die ihm auf den Rücken fällt. Mit der Linken hält er ein Schallgefäß empor. Der Künstler nennt es seinen Löwen. Er hängt von der Mitte herab und schwankt hin und her. Herakles' Rechte ist mit der Keule bewehrt, die er jenem auch als Antwort auf sein Gebrüll verabfolgt. Er hebt sie in die Höhe und gibt dem Erze damit einen Schlag. Doch der "Löwe" schwebt in der Luft, brüllt laut auf, von dem mächtigen Schlage getroffen, und läßt den Schall lang nach dröhnen."

Hierzu der entsprechende automatische Mechanismus für den rythmischen Schlag der Hände der Danaer im stehenden Automaten Herons.

Die Beschreibung Prokops über die Funktion dieser digitalen akustischen Uhr fährt folgenderweise fort:

"Beim ersten Kampf gibt es nun einen Schlag, bei der zweiten Stunde einen Doppelschlag. So summiert er bei jedem Schlage die Zahl der vorangegangenen Kampfe, bis er in dieser Weise sechs vollendet hat, um nicht, wenn er noch mehr Stundenschlage gäbe, durch ihre Überzahl das Gehör abzustumpfen. So zählt er den siebenten wieder als den ersten, indem er nur einmal den Ton erschallen läßt, dann zweimal, und so fort, bis das Erz, sooft als es nötig ist, für die zweite Hexade seine Stimme erhebt und dadurch die übrigen Stunden kenntlich macht. Man kann aber die Schläge weiter hören als die Schweite beträgt. So ist es für alle leicht, auch für die, die in der Ferne ihren Standort haben, die Stundenzeit zu erfahren."

Entsprechende Mechanismen für die Erzeugung von Tönen finden wir bei Herons mobilen Automaten.

Die prokopische Beschreibung fährt mit der Darstellung von mehreren automatischen Beiwerken, die das eigentliche Uhrwerk begleiten, fort.

Erstes Beiwerk, die sich drehende Figur von Pan.

"Auch Pan hört den Ton. Man erkennt ihn an seinem Zottelbart und dem Doppelhorn auf der Stirn. Er war von Liebe ergriffen zur Echo, die ihn durch ihr Spiel dem Gelächter preisgibt. Er schmachtet nach seiner Echo. Da hört er den Ton des Erzes und aufgeregt erhebt er die Rechte und dreht sein Gesicht herum, ob er irgendwie das Mädchen entdecken könne, wobei er sich hin und her wendet, wie das nur eine unglückliche Liebe verursacht. Man könnte freilich auch sagen, er staune über Herakles, was das für ein Held sei."

Zweites Beiwerk, die Satyren als Pans Begleiter.

"Wenn Pan da ist, dürfen auch die Satyren nicht fehlen. Sie lachen ihn aus und umringen ihn zum Spott, da sie sein liebeerfülltes und wilderregtes Antlitz sehen und die Mischung von Sanftheit und Schroffheit in seiner Stimmung. Aber diese Gruppe

befindet sich über der Spitze des Tempels, in dem der Sohn der Alkmene nackt sichtbar wird."

Drittes Beiwerk, Diomedes mit der Trompete.

"Der Tydide aber, der seinen Standort zur Rechten gefunden hat, bleibt natürlich auch jetzt seiner Trompete treu. Denn er bläst sie zum Schluß zu Ehren des Herakles, wenn er zu seinem letzten Gang antritt, wie er das in Skyros tat, als er den Peliden entdeckt hatte. Denn auch damals blies er so oft, indem er die Zahl seiner Trompetenstöße gleich der der Tagesstunden machte."

Viertes Beiwerk, die Diomedes' Diener.

"Nun bringt ein Diener, der (dies Schlußzeichen) hört, seinem Herrn die Zurüstung zum Bad; selbstverständlich ist für die Speisen bereits gesorgt. Ein andrer bringt sie nämlich bei Anbruch des Tages in aller Eile vom Markte heim. Es scheint, die beiden haben bei einem strengen Herrn Dienst. Denn sonst liefen sie nicht in solchem Galopp."

Fünftes Beiwerk, der Hirt.

"Auf der entgegengesetzten Seite hat jener Hirte seinen Stand gefunden, der seinen Krummstab der Linken übergeben (um mich poetisch auszudrücken) und vergnügt und lächelnd die Rechte staunend erhebt.

Soweit nun also der Lärm, das Staunen und der Trompetenschall um die Herakleskämpfe!"

Sechstes Beiwerk, Herakles als Bogenschütze.

"Nun hat aber unser Meister hier den zwölften der oben beschriebenen Kämpfe sich noch einmal zum Vorwurf genommen und gibt ihn einein noch gewaltigeren Herakles.

Da steht er nun als Bogenschütze unter dem eben beschriebenen Diomedes. Der Kampf dreht sich auch hier wieder um die goldnen Äpfel. Da hat er nun gerade den Pfeil auf die Sehne gelegt. Mit der rechten Hand zieht, er sie zur Brust zurück. Die Linke stößt dagegen den Bogen in der Mitte zwischen den beiden Händen mit dem Pfeile so weit vorwärts, daß nur noch die vorragende Pfeilspitze nach außen

übersteht. Er zieht das Auge zu scharfem Zielen zusammen. Denn das Ziel, auf das er den Pfeil abschießen will, ist nur klein ...”

Hier bricht die erhaltene Beschreibung Prokops plötzlich ab.

Ein Vergleich mit dem Beispiel 41 der Pneumatik Herons aus Alexandria, in dem der Mechanismus für den automatischen Wurf eines Pfeiles vom Bogenschützen Herkules gegen eine Schlange dargestellt wird, erlaubt uns die Vermutung, daß auch dieser Bogenschütze Herkules automatisch funktionieren konnte.

Nun kann man sich eine globale Rekonstruktion dieser wunderbaren automatischenWasseruhr von Gaza zutrauen. Und gleichzeitig behaupten, daß sie die hellenistische mit der byzantinischen und der islamischen Uhrwerktechnologie, wie die der frühen euroräischen Renaissance, überbrückt.

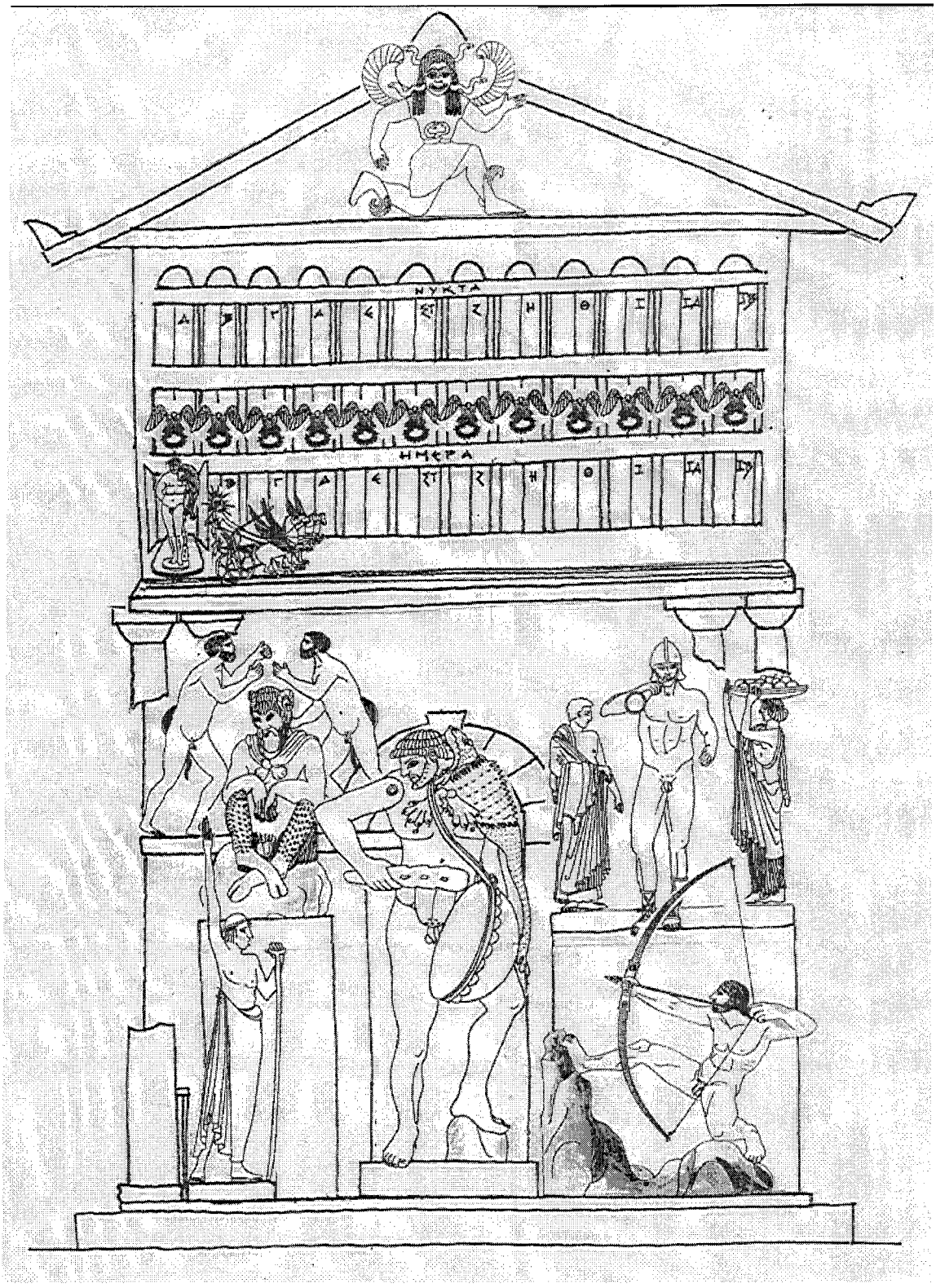


Bild 1: *Rekonstruktion des Uhrwerkes von Gaza*

Stabilität von Regelkreisen mit periodisch veränderlicher Totzeit

Rudolf Tracht, Marc Thorausch
Universität Duisburg-Essen

Regelkreise mit variabler Totzeit wurden in den letzten Jahren sehr eingehend hinsichtlich ihres Stabilitätsverhaltens untersucht. Meist wird dazu die Stabilitätsanalyse nach der direkten Methode nach Ljapunov durchgeführt. Um die Stabilitätsgrenzen in Abhängigkeit von Systemparametern zu erhalten, muss eine Matrixungleichung untersucht werden. Da die Ljapunov-Methode nur hinreichende Bedingungen für die Stabilität liefert, interessiert, wie weit die berechnete Stabilitätsgrenze von der tatsächlichen Stabilitätsgrenze entfernt ist. Im Folgenden wird für spezielle Totzeitverläufe eine Methode beschrieben, die es gestattet, die tatsächliche Stabilitätsgrenze und deren Abhängigkeit vom zeitlichen Verlauf der Totzeit zu bestimmen.

1. Einleitung

Wenn in Regelkreisen zwischen Regler und Regelstrecke eine größere Entfernung zu überbrücken ist, werden dazu häufig Kommunikationseinrichtungen verwendet, durch die eine zeitlich veränderliche Totzeit in den Regelkreis eingefügt wird [1]. Meist werden für die Regelung digitale Regler eingesetzt, so dass als mathematisches Modell ein Abtastsystem mit variabler Totzeit verwendet werden muss. Wenn vorausgesetzt wird, dass die Abtastzeit im Vergleich zu den Zeitkonstanten der Regelstrecke klein ist, kann von einem quasikontinuierlichen Regelungssystem ausgegangen werden. Wenn außerdem vorausgesetzt wird, dass die Regelstrecke durch ein lineares mathematisches Modell beschrieben werden kann, ist es in diesem Fall zulässig, das in Abb. 1 betrachtete Ersatzsystem für die Beurteilung der Stabilität zu untersuchen.

Zur Stabilitätsuntersuchung von kontinuierlichen Regelkreisen mit variabler Totzeit wurden in den letzten Jahren Methoden entwickelt, die auf der Stabilitätsanalyse nach Ljapunov basieren [2,3]. Durch Wahl einer geeigneten Ljapunov-Funktion kann eine lineare Matrixungleichung hergeleitet werden. Wenn diese Ungleichung nicht verletzt wird, ist der Regelkreis stabil. Als Parameter gehen die maximale Totzeit und die maximale Änderungsgeschwindigkeit der Totzeit ein. Mit Hilfe der Matlab-LMI-Toolbox können die Maximalwerte dieser Parameter numerisch bestimmt werden.

Es ist bekannt, dass Stabilitätsuntersuchungen nach Ljapunov manchmal zu sehr konservativen Ergebnissen führen. Es ist daher von Interesse, die tatsächliche

Stabilitätsgrenze für spezielle Totzeitverläufe zu ermitteln, um damit Aussagen über die praktische Nutzbarkeit der LMI-Methode gewinnen zu können.

Im Folgenden wird für eine spezielle Klasse von Zeitfunktionen eine Methode zur Bestimmung der Stabilitätsgrenze vorgestellt. Und zwar wird verlangt, dass die Totzeit sich periodisch verändert, wobei der zeitliche Verlauf einer Dreiecksfunktion entspricht. Im nächsten Abschnitt wird zunächst das Problem genauer formuliert. Es folgt dann im dritten Abschnitt eine Untersuchung des Regelkreises im Frequenzbereich. In Abschnitt 4 wird dann eine Bedingung für die Stabilitätsgrenze hergeleitet. Damit kann dann in Abschnitt 5 anhand von numerischen Beispielen gezeigt werden, wie die Stabilitätsgrenze von der Änderungsgeschwindigkeit der Totzeit und den Maximal-Minimalwerten der Totzeit abhängt. Es folgt ein Vorschlag zur Verallgemeinerung.

2. Aufgabenstellung

Für den in Abb. 1 dargestellten Regelkreis sollen folgende Bedingungen gelten:

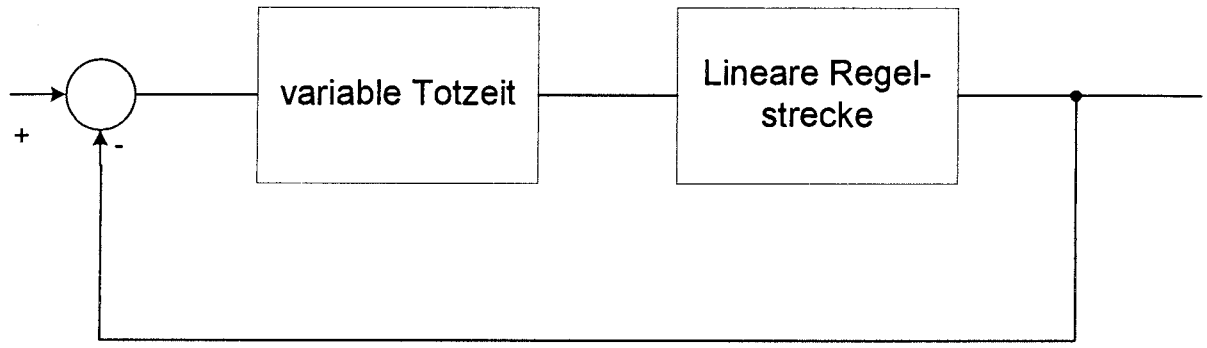


Abb. 1: Untersuchter Regelkreis

Die Totzeit $T_t(t)$ soll periodisch um einen Mittelwert T_{tm} schwanken.

$$T_t(t) = T_{tm} + \eta(t) \tag{1}$$

Dabei ist η(t) eine periodische Dreiecksfunktion:

$$\begin{aligned} \eta(t) &= p_T t \quad \textit{für} \quad 0 \le t \le \frac{1}{4} T_p \\ &= \Delta T_t - p_T t \quad \textit{für} \quad \frac{1}{4} T_p \le t \le \frac{3}{4} \; T_p \\ &= -\Delta T_t + p_T t \quad \textit{für} \quad \frac{3}{4} T_p \le t \le T_p \end{aligned} \tag{2}$$

Für die Änderungsgeschwindigkeit der Totzeit soll gelten

$$p_T = \frac{4\Delta T_t}{T_p} \tag{3}$$

Der zeitliche Verlauf der Totzeit ist in Abb. 2 dargestellt. Für den Spezialfall $\Delta T_t=1$ wird die Zeitfunktion $\eta(t)$ im Folgenden mit der Bezeichnung $trg(T_p)$ abgekürzt.

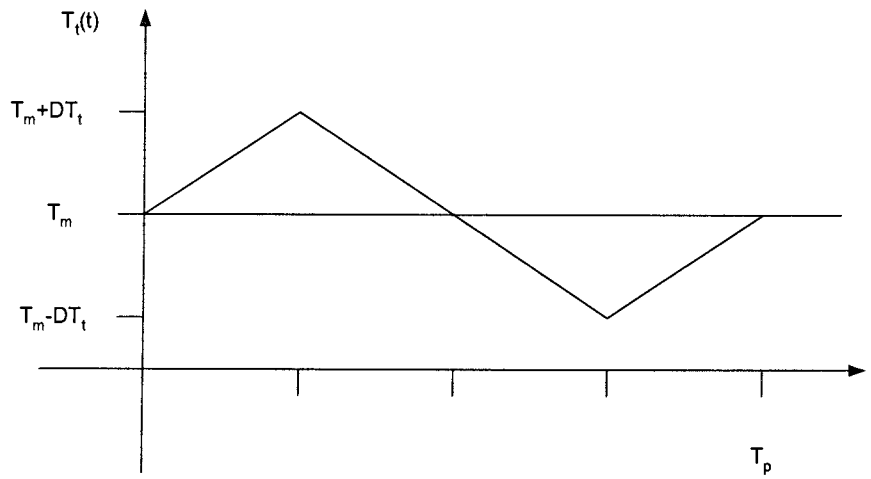


Abb.2: Zeitverlauf der variablen Totzeit

Die kontinuierliche Regelstrecke soll durch eine rationale Übertragungsfunktion G(s) beschrieben werden können. Gesucht wird die Stabilitätsgrenze, wenn der Verstärkungsfaktor der Übertragungsfunktion variiert wird.

3. Frequenzgang des offenen Kreises

Zunächst soll untersucht werden, wie der aufgeschnittene Regelkreis auf eine sinusförmig veränderliche Eingangsgröße reagiert. Die Kreisfrequenz des Eingangssignals sei ω_f. Durch das Totzeitglied wird die Phase des Eingangssignals verändert. Bei variierender Totzeit erhält man ein phasenmoduliertes Signal. Für die spezielle periodische Dreiecksschwingung ergibt sich ein Signal, bei dem zwischen zwei Frequenzen umgeschaltet wird, wie man leicht erkennt, wenn die Totzeit in die Sinusfunktion eingesetzt wird:

$$\begin{aligned} y(t) &= \sin(\omega(t-(T_{tm}+\eta(t)))) \\ &= \sin(\omega(1\mp p_T)t-\omega T_{tm}) \end{aligned} \qquad (4)$$

Je nachdem, ob die Steigung der Dreieckschwingung positiv oder negativ ist, muss im entsprechenden Zeitintervall das Vorzeichen von p_T positiv oder negativ gewählt werden.
Man erkennt, dass dieses modulierte Signal eine periodische Zeitfunktion mit der Periodendauer T_p ist. Diese periodische Funktion kann nun in eine Fourier-Reihe entwickelt werden. Wenn man die komplexe Darstellung der Fourier-Reihe wählt, erhält man:

$$y(t)=\sum_{k=-\infty}^{+\infty} c_k e^{jk\omega_p t} \qquad (5)$$

Es wird nun vorausgesetzt, dass die Kreisfrequenz ω_f der Eingangsgröße ein ganzzahliges Vielfaches der Kreisfrequenz ω_p ist, mit der sich die Totzeit ändert:

$$s(t) = \hat{s} \cdot e^{jl\omega_p t} \tag{6}$$

Die Fourier-Integrale liefern dann die in den folgenden Gleichungen dargestellten komplexen Größen.

Für $l(1-p_T) \neq k$ gilt:

$$c_{kl} = \frac{e^{-jl\omega_p T_t}}{2j(l(1-p_T)-k)\pi}(e^{j((1-p_T)l-k)\pi} - 1) \tag{7}$$

Für $l(1-p_T) = k$ gilt:

$$c_{kl} = \frac{1}{2}e^{-jl\omega_p T_t} \tag{8}$$

Für $l(1+p_T) \neq k$ gilt:

$$c_{k2} = \frac{e^{-jl\omega_p(T_t + p_T T_p)}}{2j(l(1+p_T)-k)\pi}(e^{j((1+p_T)l-k)2\pi} - e^{j((1+p_T)l-k)\pi}) \tag{9}$$

Für $l(1+p_T) = k$ gilt:

$$c_{k2} = \frac{1}{2}e^{-lj\omega_p(T_t - p_T T_p)} \tag{10}$$

Der komplexe Koeffizient der Fourier-Reihe ergibt sich aus den beiden Anteilen zu:

$$\begin{aligned} c_k &= c_{kl} + c_{k2} \\ &= c(k,l) \end{aligned} \tag{11}$$

Der Betrag dieser Koeffizienten entspricht dem Amplitudenspektrum des periodischen Ausgangssignals des Totzeitgliedes.

Wenn anstelle des sinusförmigen Eingangssignals eine beliebige periodische Funktion mit der Periodendauer T_p aufgeschaltet wird, dann kann diese Eingangsfunktion in eine Fourier-Reihe zerlegt werden. Jeder Summand wird durch das Totzeitglied mit der variablen Totzeit in eine periodische Funktion abgebildet, die durch eine Fourier-Reihe repräsentiert werden kann. Ein Vektor mit den Fourier-Koeffizienten des Eingangssignals wird daher in einen Vektor mit den Fourier-Koeffizienten des Ausgangssignals abgebildet. Dieser Vektor kann durch Multiplikation des Eingangsvektors mit einer Matrix, deren Elemente aus den Koeffizienten nach Gleichung (11) gebildet werden,

$$\mathbf{H} = [c(k,l)] \tag{12}$$

berechnet werden.

Wenn das Ausgangssignal des Totzeitgliedes auf den Eingang einer linearen Regelstrecke, die durch eine rationale Übertragungsfunktion G(s) beschrieben werden kann, geschaltet wird, dann ist das Ausgangssignal der Regelstrecke wieder durch eine Reihe aus Sinusfunktionen darstellbar. Die Koeffizienten dieser Reihe erhält man durch Multiplikation des Ausgangsvektors des Totzeitgliedes mit einer Diagonalmatrix der Form

$$\mathbf{G} = diag\left\{G(jk\frac{2\pi}{T_p})\right\} \tag{13}$$

Da technische Regelstrecken immer Tiefpassverhalten aufweisen, genügt es, eine vergleichsweise geringe Anzahl an Elementen der Fourier-Reihe des Ausgangssignals zu berücksichtigen. Die Dimension der Vektoren der Fourier-Koeffizienten kann also klein gehalten werden. Mit den oben eingeführten Bezeichnungen erhält man für den Vektor des Ausgangssignals der Regelstrecke in Abhängigkeit vom Vektor des Eingangssignals des offenen Kreises folgende Beziehung:

$$\mathbf{s}_a = \mathbf{G} \cdot \mathbf{H}_T \cdot \mathbf{s}_e \tag{14}$$

Diese Beziehung, die numerisch nur näherungsweise ausgewertet werden kann, wird nun dazu verwendet, um Aussagen über die Stabilitätsgrenze des geschlossenen Regelkreises zu gewinnen.

4. Stabilitätsgrenzen für Regelkreise mit periodisch variierenden Totzeiten

Ähnlich wie bei der Methode der harmonischen Balance wird das Totzeitglied näherungsweise durch den oben hergeleiteteten linearen Operator ersetzt. Es wird angenommen, dass der Regelkreis für kleine Streckenverstärkungen stabil ist. Wenn der Verstärkungsfaktor vergrößert wird, verringert sich der Abstand zur Stabilitätsgrenze. Im Grenzfall erhält man eine stationäre Schwingung, d.h., dass die Eingangsgröße bei Berücksichtigung der Vorzeichendrehung im Regelkreis gleich der negierten Ausgangsgröße ist. Folglich muss gelten:

$$\begin{aligned} \mathbf{s}_e &= -\mathbf{s}_a \\ &= -\mathbf{G} \cdot \mathbf{H}_T \cdot \mathbf{s}_e \end{aligned} \tag{15}$$

Der Vektor $\mathbf{s}_e$ ist also ein Eigenvektor der komplexen Matrix $\mathbf{GH}_T$ zum Eigenwert –1. Wenn eine periodisch veränderliche Totzeit existiert, so dass die resultierende Matrix $\mathbf{GH}_T$ einen Eigenwert besitzt, der zu einer Phasenverschiebung von 180° führt, dann kann durch

geeignete Wahl des Verstärkungsfaktors die Stabilitätsgrenze ermittelt werden. Die Vorgehensweise soll nun an einem numerischen Beispiel erläutert werden.

5. Numerisches Beispiel

Es wird nun ein Regelkreis nach Abb. 1 betrachtet. Die Periodendauer der Dreieckschwingung soll T_p=15 betragen. Die mittlere Totzeit sei T_{tm}=1. Für die Schwankungsbreite wird ΔT_t=0.5 gewählt. Dies führt auf eine Änderungsgeschwindigkeit p_T=0.133 der Totzeit. Die Regelstrecke soll durch eine Übertragungsfunktion zweiter Ordnung beschrieben werden können:

$$G(s) = \frac{k}{s^2 + 2\zeta\omega_0 s + \omega_0^2} \tag{16}$$

Für die Parameter der Regelstrecke soll gelten: k=1.5, ζ=0.5 und ω_0=1. Die komplexen Matrizen $\mathbf{H}_T$ und $\mathbf{G}$ können mit den Gleichungen (7)-(11) bzw. mit Gleichung (13) z.B. mit Matlab berechnet werden. Approximiert man die unendlich dimensionalen Vektoren der Fourier-Koeffizienten durch Vektoren der Dimension 8, dann können mit der Matlab-Funktion „eig" leicht die Eigenwerte und die Eigenvektoren des Matrixproduktes $\mathbf{G}\cdot\mathbf{H}_T$ bestimmt werden. Die 8 komplexen Eigenwerte können in der komplexen Ebene dargestellt werden (Abb. 3).

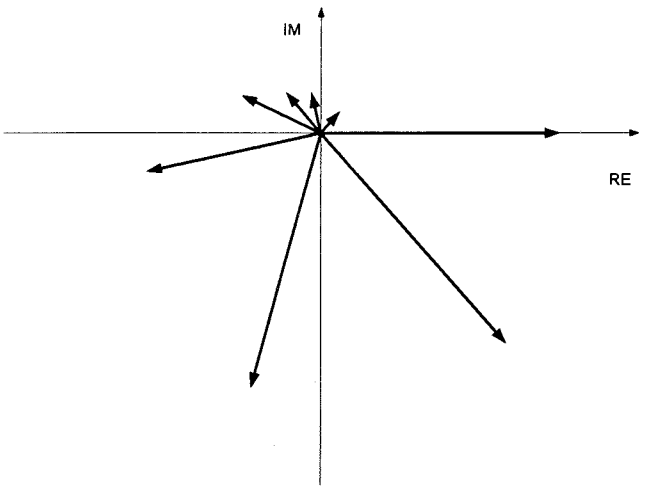


Abb. 3: Komplexe Eigenwerte der Matrix $\mathbf{G}\cdot\mathbf{H}_T$

Die Eigenwerte hängen von der Periodendauer T_p ab. Wählt man T_p=18.231, dann wandert der 5. Eigenwert in den Punkt

$$\lambda_5 = -0.856 - j0.004.$$

Durch eine Vergrößerung der Streckenverstärkung kann dieser Eigenwert mit k=1.753 in den kritischen Punkt −1 verschoben werden. Simulationsexperimente mit einem Simulink-Modell zeigen, dass man für diese Streckenparameter nach einer kurzen Einschwingphase eine stationäre Schwingung am Ausgang der Regelstrecke erhält. Der Zeitverlauf ist in Abb. 4 dargestellt.

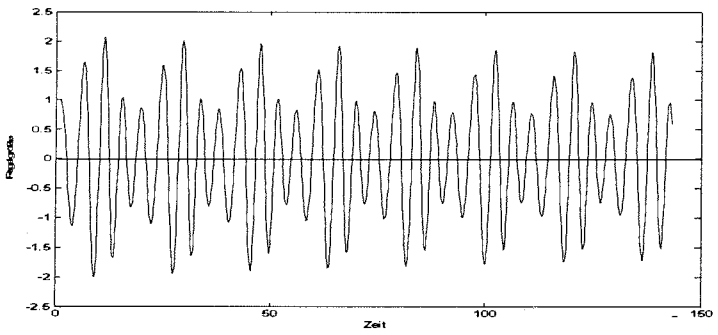


Abb. 4: Zeitverlauf der Regelgröße

Eine Spektralanalyse dieses Signals führt auf ein Spektrum, das mit den Komponenten des entsprechenden Eigenvektors übereinstimmt.

6. Parameterabhängigkeit der Stabilitätsgrenze

Mit der oben beschriebenen Methode kann nun überprüft werden, wie sich die Änderungsgeschwindigkeit und die Schwankungsbreite der Totzeit auf die Stabilitätsgrenze auswirken. Zunächst soll die Abhängigkeit von der Periodendauer überprüft werden. Wenn man die Periodendauer halbiert, die Änderungsgeschwindigkeit p_T also verdoppelt, dann wandert der dritte Eigenwert anstelle des fünften Eigenwertes in den Punkt −1. Der Zeitverlauf der dadurch entstehenden stationären Schwingung am Streckenausgang ändert sich zwar, die Stabilitätsgrenze wird jedoch nicht beeinflusst.

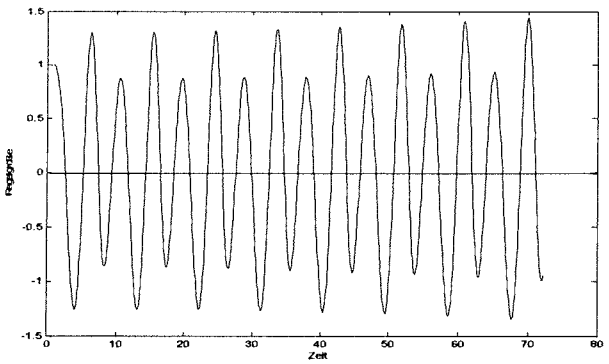


Abb. 5: Zeitverlauf der Regelgröße bei Verdopplung der Änderungsgeschwindigkeit p_T

Man kann daraus schließen, dass die Änderungsgeschwindigkeit für das Stabilitätsverhalten von untergeordneter Bedeutung ist. Ähnliches lässt sich von der Schwankungsbreite der Totzeit sagen. Wenn ΔT_t halbiert wird, ändert sich dadurch die Stabilitätsgrenze praktisch nicht. Auch eine Änderung der Form der Dreieckschwingung hat nur geringen Einfluss auf die Lage der Eigenwerte. Entscheidend ist der Mittelwert der Totzeit.

7. Sinusförmig veränderliche Totzeiten

Bisher wurden variable Totzeiten mit dreiecksförmiger Abhängigkeit betrachtet. Dies scheint auf den ersten Blick den Anwendungsbereich sehr stark einzuschränken. Wenn Untersuchungsmethoden für sinusförmig veränderliche Totzeiten verfügbar wären, könnte jeder periodische Totzeitverlauf in eine Fourier-Reihe zerlegt werden. Da eine Summe von Totzeiten durch Hintereinanderschalten einzelner Totzeitglieder nachgebildet werden kann, ist es möglich, für jeden Sinusterm eine Matrix $\mathbf{H}_T$ zu berechnen. Die entsprechende Matrix für das Gesamtsystem erhält man durch Multiplikation aus den Teilmatrizen.
Für die Fourier-Reihe einer Dreieckschwingung gilt:

$$trg(T_p) = \frac{4}{\pi}(\sin(\frac{2\pi}{T_p}) - \frac{1}{9}\sin(3\frac{2\pi}{T_p}) + \frac{1}{25}\sin(5\frac{2\pi}{T_p}) + \ldots \qquad (18)$$

Wenn diese Beziehung auch für höhere Harmonische dargestellt wird, dann erhält man ein lineares Gleichungssystem für die Sinusfunktionen, das aufgrund seiner Struktur leicht aufgelöst werden kann. Man erhält für die Grundschwingung

$$\sin\left(\frac{2\pi}{T_p}t\right) = \frac{\pi}{4}\left(trg(T_p) + \frac{1}{9}trg(\frac{1}{3}T_p) - \frac{1}{25}trg(\frac{1}{5}T_p) + \ldots\right) \qquad (19)$$

Man sieht, dass diese Reihe sehr schnell konvergiert. Es genügt daher, die ersten drei Terme zu berücksichtigen.
Simulationsexperimente zeigen, dass die Stabilitätsgrenzen bei sinusförmiger Variation der Totzeit nur geringfügig verschieden von den Stabilitätsgrenzen der Grunddreiecksfunktion sind. Da es sich bei der vorgeschlagenen Methode aufgrund der endlichen Dimension der Vektoren und Matrizen um ein Näherungsverfahren handelt, genügt es im allgemeinen, dreiecksförmige Totzeitvariationen zu untersuchen. Die tatsächliche Stabilitätsgrenze kann dann leicht durch Simulationsexperimente gefunden werden. Bei nichtperiodischen Totzeitverläufen kann wie üblich für ein endliches Beobachtungsintervall die Zeitfunktion periodisch fortgesetzt werden und dann mit der diskreten Fourier-Transformation untersucht werden.

8. Reglerentwurf

Die Untersuchung des totzeitbehafteten Regelkreises im Frequenzbereich erlaubt die Anwendung der klassischen Reglerentwurfsverfahren. Mit Lead-Lag-Gliedern kann der Abstand des Frequenzganges vom kritischen Punkt –1 durch Erzeugung einer Phasenreserve vergrößert werden und damit das Stabilitätsverhalten verbessert werden.

9. Zusammenfassung und Ausblick

Für spezielle zeitveränderliche Totzeiten wurde ein Verfahren zur Bestimmung der Stabilitätsgrenze vorgeschlagen. Mit diesem Verfahren konnte der Einfluss von Systemparametern auf die Stabilitätsgrenze untersucht werden. Dabei zeigte sich, dass entgegen den Erwartungen, die aus anderen Verfahren resultieren, die Änderungsgeschwindigkeit der Totzeit nur marginalen Einfluss auf die Stabilitätsgrenze besitzt. Entscheidend scheint vielmehr der zeitliche Mittelwert der Totzeit zu sein. Dabei muss selbstverständlich berücksichtigt werden, dass die Totzeit durch die Variation nicht in den negativen Bereich wandert und die Änderungsgeschwindigkeit kleiner 1 bleibt, da sonst die Vergangenheit des Zeitverlaufes der Totzeit vor Simulationsbeginn berücksichtigt werden müsste. Die theoretische Behandlung des Totzeitverlaufes beschränkt sich auf periodische Funktionen mit einem dreiecksförmigen Zeitverlauf. Durch Wahl der Periodendauer des Totzeitverlaufes wird erzwungen, dass die Frequenz der stationären Schwingung an der Stabilitätsgrenze ein ganzzahliges Vielfaches der Frequenz der periodischen Zeitfunktion des Totzeitsignals ist. Die Erweiterung für andere Frequenzen bedarf noch der genaueren Untersuchung.

Da im Prinzip ein verallgemeinerter Frequenzgang betrachtet wird, kann das Verfahren auch für den Reglerentwurf genutzt werden.

10. Literatur

1. Zhang W., Branicky M.S., Philips S.M.: Stability of Networked Control Systems. IEEE Control System Magazine February 2001

2. Jin Hoon Kin: Delay and ist Time-Derivative Dependent Robust Stability of Time-Delayed Linear Systems with Uncertainty. IEEE Transactions on Automatic Control, Vol. 46, No 5, May 2001

3. Eymann Th.: Reglerentwurf für Regelkreise in Feldbussystemen. VDI-Verlag, Düsseldorf 2001

PCH-basierte Regelung hydraulischer Antriebe mit einer Anwendung in Stahlwalzwerken

G. Grabmair* , K. Schlacher*,**

* Christian Doppler Laboratorium

Automatisierung mechatronischer Systeme in der Stahl Industrie

** Institut für Regelungstechnik und elektrische Antriebe

Johannes Kepler Universität, Linz, Altenbergerstr. 69, 4040, *Austria.*

Phone: ++43-732-2468-9730 Fax:++43-732-2468-9734

Email: gernot.grabmair@jku.at , kurt.schlacher@jku.at

Kurzfassung

In diesem Beitrag wird ein Energie basierter Reglerentwurf für hydraulische Aktuatoren beschrieben. Ausgehend von den Grundgesetzen der Thermodynamik werden die Modellgleichungen für einen doppelt wirkenden Hydraulikzylinder (DAP) hergeleitet. Es wird gezeigt, dass das mathematische Modell die Struktur eines PCHD-Systems aufweist. Aufbauend auf dieser speziellen Systemstruktur wird ein Regelgesetz so entworfen, dass der geschlossene Kreis passiv und im gesamten Arbeitsbereich asymptotisch stabil ist. Die so zu erreichende Güte des geregelten Systems wird abschließend anhand industrieller Messergebnisse demonstriert.

1 Einführung

Die besondere Bedeutung hydraulischer Systeme in der Stahlindustrie ist allgemein bekannt. Insbesondere in Großanlagen, wo erhebliche Massen mit leichten und kleinen Aktuatoren bewegt werden müssen, sind hydraulische Systeme vielfach im Einsatz. Als Beispiele seien hier nur die hydraulische Walzenanstellung für die Dickenregelung bei Flachprodukten (Bänder, Platten, etc.) oder der Verstellmechanismus der sogenannten Schlingenheber, sie dienen bei mehreren aufeinanderfolgenden Walzgerüsten als Bandspeicher und zur relativ freien Vorgabe des Bandzuges, angeführt. Eine Herausforderung stellt die Regelung dieser Systeme dar, wenn obige Aufgaben mit hoher Präzision und einer über den gesamten Arbeitsbereich – der insbesondere bei den Schlingenhebern einen Großteil des möglichen Verfahrweges der Kolben ausmachen kann – gleichbleibenden Dynamik und Regelgüte gelöst werden sollen. Es werden in diesem Beitrag passivitätsbasierte Regelungsverfahren angewandt, um Anforderungen für Strecken mit nichtlinearem Charakter gerecht zu werden. Zudem gewinnt man mit diesem Entwurfsverfahren eine gewisse Robustheit des Regelgesetzes. Für den hier vorgestellten Reglerentwurf werden spezielle

Systemeigenschaften sogenannter PCHD-Systeme (*Port Controlled Hamiltonian Systems with Dissipation*, z.B. [4] oder [8] und die Referenzen darin) ausgenützt. PCHD-Systeme sind eine Verallgemeinerung von in der Mechanik wohlbekannten Hamiltonsystemen, die es ermöglichen, Leistungsflüsse innerhalb des Systems und über dessen Grenzen besonders einfach zu beschreiben. Sie stellen zudem eine Mischung von Hamiltonschen Systemen und Gradientensystemen dar.

Beginnend mit einer kurzen Einführung in die mathematische Struktur von PCHD-Systemen wird das mathematische Modell von hydraulischen Aktuatoren hergeleitet. Für die Herleitung der energetischen Beziehungen wird hierfür auf die Gleichgewichtsthermodynamik zurückgegriffen. Sobald die Energiebilanz für stofflich offene Systeme geklärt ist, wird diese auf einen typischen hydraulischen Aktuator angewandt. In der Folge kann für diesen ein energiebasiertes Regelgesetz gefunden werden, welches ein asymptotisch stabiles Verhalten des Systems im gesamten Arbeitsbereich zur Folge hat.

2 PCHD-Systeme

Es seien mit $\partial_x f$ die Jakobimatrix von f bezüglich x und mit $\dot{x}$ die totale Ableitung von x nach der Zeit bezeichnet. Ein (verallgemeinertes) Hamiltonsystem kann in lokalen Koordinaten als

$$\dot{x} = J(x)\,\partial_x H(x)^T \tag{1}$$

mit $x \in \mathbb{R}^n$, der Hamiltonfunktion $H(x)$ und der schiefsymmetrischen *Strukturmatrix* $J(x)$ geschrieben werden. Falls die Jakobi-Identität erfüllt ist und für $\operatorname{rank}(J(x)) = 2m$ gilt, garantiert eine Folgerung aus dem Theorem von Darboux (siehe [6]) die Existenz *kanonischer Koordinaten* $x = (q, p, z)$, $q, p \in \mathbb{R}^m$, $z \in \mathbb{R}^{n-2m}$ so, dass

$$\begin{bmatrix} \dot{q} \\ \dot{p} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} 0 & I_m & 0 \\ -I_m & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \partial_x H(x)^T \tag{2}$$

mit der $m \times m$ Einheitsmatrix I_m zumindest lokal gilt. Die Größen $C(x) = z$ von (2) werden als *Casimirfunktionen* bezeichnet, und für sie gilt

$$\partial_x C(x)\,(J(x)) = 0\ .$$

Diese sind Invarianten (Erhaltungsgrößen) des Systems unabhängig von der speziellen Hamiltonfunktion H, sie bleiben konstant entlang der Lösungen von (2) und natürlich auch von (1). Den Casimirfunktionen kommt eine besondere Bedeutung bei der Analyse von (1) zu, da mit ihnen z.B. die Ordnung eines gewöhnlichen Differentialgleichungssystems reduziert werden kann. Zudem können sie dazu verwendet werden, die Hamiltonfunktion des Systems zu verändern, da die modifizierte Hamiltonfunktion

$$\tilde{H}(x) = H(x) + \tilde{f}(C_1, \ldots, C_{n-2m})$$

dieselben Differentialgleichungen erzeugt.

Um eine größere Klasse von physikalischen Systemen zu erfassen, kann der Begriff eines Hamiltonsystems erweitert werden. Dies führt zu sogenannten verallgemeinerten PCHD(Port Controlled Hamiltonian)-Systemen mit Dissipation (siehe [8]). Es seien mit $u \in \mathbb{R}^l$ und $y \in \left(\mathbb{R}^l\right)^*$ der l-dimensionale Eingang bzw. Ausgang bezeichnet. Dann kann ein PCHD-System als

$$\begin{aligned} \dot{x} &= \left(J\left(x\right) - R\left(x\right)\right) \partial_x H\left(x\right)^T + G\left(x\right) u \\ y^T &= \partial_x H\left(x\right) G\left(x\right) \end{aligned} \tag{3}$$

geschrieben werden. Die (verallgemeinerten) Casimirfunktionen werden jetzt als Lösungen von $\partial_x C\left(x\right)\left(J\left(x\right) - R\left(x\right)\right) = 0$ definiert. Zusätzlich gilt für die zeitliche Änderung der Hamiltonfunktion

$$\dot{H}\left(x\left(t\right)\right) = \langle y, u \rangle - \partial_x H\left(x\left(t\right)\right) R\left(x\left(t\right)\right) \partial_x H\left(x\right)^T$$

mit dem kanonischen Produkt $\langle y, u \rangle$, was die Passivität dieser Systemklasse sicherstellt. Falls die innere Energie eines Systems gleichzeitig die Hamiltonfunktion ist, steht J in Verbindung mit den internen Energieflüssen und die positiv semi-definite Matrix R mit der intern dissipierten Energie. Das Produkt $\langle y, u \rangle$ gibt den Leistungsfluss über die Systemgrenzen an. Dies ist auch der Grund, weshalb y als der zu u kollokierte Ausgang bezeichnet wird. Auf diese Systemstruktur sind spezielle Reglerentwurfsmethoden, wie Dämpfungsinjektion oder IDA-PBC (interconnection and damping assignement - passivity based control) zugeschnitten. Im Falle des IDA-PBC Entwurfs wird ein im Allgemeinen nichtlineares Zustandsregelgesetz so berechnet, dass das geregelte System wieder PCHD-Struktur mit vorgegebenen Größen R_d, J_d und H_d aufweist. Darüber hinaus können auch weiterführende Fragestellungen, wie obiges Zustandsregelgesetz $u = u\left(\hat{y}\right)$, $\hat{y} = \hat{y}\left(x\right)$ so zu entwerfen, dass es nur von messbaren Größen $\hat{y} \in \mathbb{R}^k$, $k < l$ abhängt. Dieses Problem ist insbesondere bei hydraulischen Systemen von Bedeutung (siehe [3]).

2.1 PCHD-Struktur hydraulischer Antriebe

Als praktisches Beispiel soll in der Folge ein doppelt wirkender hydraulischer Aktuator (DAP) dienen. Aus Gründen der Übersichtlichkeit ist dieser hier mit einem linearen Masse-Feder-Dämpfersystem nach Bild 1 verbunden ist. Zum Zwecke der Modellierung werden vorerst hydraulische Systeme mit Massenfluss über deren Systemgrenze aus energetischer Sicht behandelt werden.

2.1.1 Thermodynamische Ausführungen

Wie bereits erwähnt betonen PCHD-Systeme mit ihrer Beschreibung die internen und externen Leistungsflüsse. Bei hydraulischen Systemen sind die Leistungsflüsse teilweise an einen Massentransport gebunden. Hierbei soll die Massenzu- bzw. abflüsse so hinreichend langsam vonstatten gehen, dass deren kinetische Energie vernachlässigt werden kann, und dass die Systeme im Rahmen der Gleichgewichtsthermodynamik behandelt werden

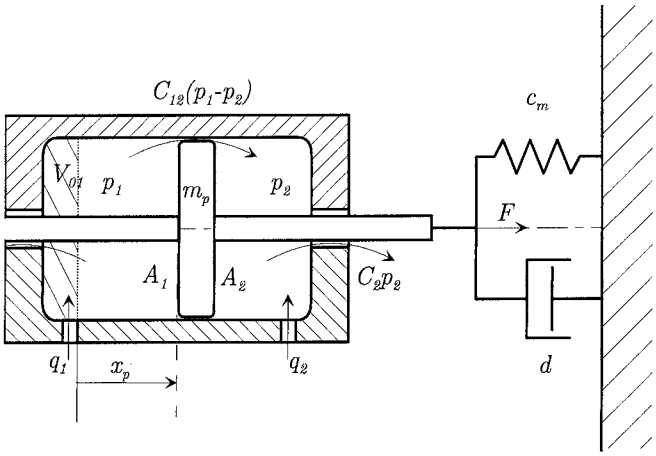


Bild 1: Der hydraulische Aktuator.

können. Diese basiert auf dem Prinzip der Energieerhaltung und ist deshalb bestens für die Herleitung einer energetisch basierten Systembeschreibung geeignet.

Die *Fundamentalgleichung*, welche die innere Energie U mit den anderen Zustandsgrößen – hier im thermodynamischen Sinne gemeint – verbindet, stellt eine mögliche Beschreibung eines thermodynamischen Systems dar. Oft, z.B. für ideale bzw. van der Waals Gase oder isentrope Fluide, wird von einem thermodynamischen System allerdings nur die *Zustandsgleichung* angegeben. Im Folgenden werden *offene thermodynamische Systeme* mit der Einschränkung auf *reversible Prozesse* betrachtet. Es seien (U, T, S, p, V, h, m), die interne Energie, Temperatur, Entropie, Druck, Volumen, die (freie) spezifische Enthalpie h und die Masse des Fluids, als thermodynamische Koordinaten gewählt. Mit den unabhängigen Variablen S, V und m lässt sich der erste Hauptsatz der Thermodynamik als

$$\omega = \mathrm{d}U - T\mathrm{d}S + p\mathrm{d}V - h\mathrm{d}m^1 \tag{4}$$

darstellen. Der Zusatzterm $h\mathrm{d}m$ berücksichtigt die Änderung der inneren Energie zufolge des zu- bzw. abfließenden Fluids. Dem Zufluss ist hierbei das positive Vorzeichen zugewiesen. Wir verfolgen hier den für Fluide üblichen Zugang, sich auf isentrope Prozesse ohne jeglichen Wärmezufluss $T\mathrm{d}S$ einzuschränken. Damit vereinfacht sich (4) zu

$$\omega = \mathrm{d}U + p\mathrm{d}V - h\mathrm{d}m\ . \tag{5}$$

[1]Indem man diese und die folgenden Formen gleich Null setzt, ergibt sich die in den Ingenieurwissenschaften verbreitete Notation. Die äußeren Formen generieren ein Kontaktideal auf einer siebendimensionalen Mannigfaltigkeit $\mathcal{E}$. Zusätzliche legt man mit der Wahl der unabhängigen und abhängigen Koordinaten das Bündel $(\mathcal{E}, \pi, \mathcal{B})$ mit der Projektion $\pi : \mathcal{E} \to \mathcal{B}$, $(S, V, m, U, T, p, h) \longmapsto (S, V, m)$, der Basismannigfaltigkeit $\mathcal{B}$ und der totalen Mannigfaltigkeit $\mathcal{E}$ (siehe [1]) fest. Ein gültiger Schnitt $\psi : (S, V, m) \longmapsto (S, V, m, U, T, p, h)$ muss nun $\psi^*\omega = 0$ erfüllen, um ein thermodynamisches System zu beschreiben.

Es werden nun als unabhängige Variablen die zwei Größen ρ (die Dichte des Fluids) und m gewählt und als abhängige Variablen die spezifischen innere Energie u_s und p. Die gewählten Größen erfüllen die Beziehungen

$$U = u_s m, \quad V = \rho^{-1} m \; . \tag{6}$$

Man erhält aus (5) die Form

$$\left(u_s + \frac{p}{\rho} - h \right) \mathrm{d}m + m \left(\frac{\partial u_s}{\partial \rho} - \frac{p}{\rho^2} \right) \mathrm{d}\rho \; .$$

Daraus ergeben sich weiter die Beziehungen (vergleiche, z.B. mit [2])

$$h = u_s + \frac{p}{\rho}, \qquad p = \rho^2 \frac{\partial u_s}{\partial \rho} \; . \tag{7}$$

Die erste Gleichung definiert die spezifische (freie) Enthalpie h des zufließenden Fluids. Dessen spezifische innere Energie ist deshalb gleich der inneren Energie des Fluids innerhalb der Systemgrenzen. Die zweite Gleichung stellt eine Bedingung für die Fundamentalgleichung des Materials dar. Wiederum für Fluide ist die Annahme gerechtfertigt, dass die – möglicherweise implizite – konstitutive Beziehung (die Zustandsgleichung) $f(p, \rho, T) = 0$ unabhängig von der Temperatur T ist. In der Hydraulik charakterisiert vielfach der isotherme Kompressionsmoduls E (siehe [5])

$$\frac{\partial p}{\partial \rho} = \frac{E}{\rho} \tag{8}$$

die Kompressibilität des Mediums. Weiters ist für hydraulische Fluide die Annahme eines konstanten Kompressionsmoduls E gerechtfertigt, solange die Druckänderungen moderat bleiben. Gleichung (8) kann nun integriert werden

$$p = \ln \left(\frac{\rho}{\rho_0} \right) E + p_0 \; , \tag{9}$$

wobei p_0 den Referenzdruck und ρ_0 die Dichte des Fluids unter Referenzdruck darstellen. Diese Integration erfolgt natürlich aufgrund obiger Annahmen entlang einer Isentropen. Da $\mathrm{d}U$ ein totales Differential darstellt, ist sein Integral wegunabhängig. Damit kann die Integration zuerst entlang $m = \text{const}$ und dann entlang $\rho = \text{const}$ durchgeführt werden. Als Ergebnis erhält man

$$\begin{aligned} \Delta U = U_1 - U_0 &= - \int_{\rho_0}^{\rho_1} p \mathrm{d} \left(\frac{m_0}{\rho} \right) + \int_{m_0}^{m_1} h(\rho_1) \, \mathrm{d}m \\ &= \frac{m_0}{\rho_1} \left(E \ln \frac{\rho_0}{\rho_1} + (E + p_0) \left(\frac{\rho_1}{\rho_0} - 1 \right) \right) + \int_{m_0}^{m_1} h \mathrm{d}m \; . \end{aligned} \tag{10}$$

Mit Δ soll ausgedrückt werden, dass nur Änderungen der inneren Energie von Belang sind. Für $m_1 = m_0$ erhält man unmittelbar die Fundamentalgleichung des abgeschlossenen thermodynamischen Systems. Offensichtlich führt obige Vorgangsweise auch zum Ziel, wenn E durch eine Funktion des Druckes p und damit der Dichte ρ ersetzt wird. Allerdings wird die Auswertung der Integrale im Allgemeinen aufwendiger.

In der technischen Hydraulik wird der Fluss des Fluids üblicherweise mit dem Volumenfluss q angegeben. Aufgrund der Massenerhaltung ergibt sich für die Gesamtmasse m des Fluids innerhalb der Kammer $\dot{m} = \rho q$ und man erhält für die zeitliche Änderung von m

$$\dot{m} = \rho q = \frac{m}{V} q \; . \tag{11}$$

Das Integral in (10) liefert mit $\mathrm{d}m = \dot{m}\mathrm{d}t$ die Beziehung

$$\int_{m_0}^{m_1} h \mathrm{d}m = \int_{t_0}^{t_1} h\rho q \mathrm{d}t = \int_{t_0}^{t_1} h\dot{m}\mathrm{d}t \; .$$

Dies stellt natürlich nichts anderes als das zeitliche Integral der Leistung des Flusses $P_h = \dot{m}h$ dar. Im Folgenden wird der Index 1 aus Gründen der Übersichtlichkeit unterdrückt.

2.1.2 PCHD-Struktur für den einfach und den doppelt wirkenden Aktuator

Der DAP nach Bild 1 stellt einen üblichen hydraulischen Aktuator dar. Im Sinne der Übersichtlichkeit werden die internen und externen Leckagen hier vernachlässigt ($C_1 = C_2 = C_{12} = 0$). Es sei aber angemerkt, dass sich diese ohne Weiteres in die Dissipationsmatrix einfügen lassen. Zudem seien die beiden Eingänge q_1 und q_2 über zwei getrennte Servoventile ansteuerbar, welche als bereits servokompensiert angenommen werden. Für nicht servokompensierte Ventile kann ihre statische Nichtlinearität

$$\begin{aligned} q &= K_d x_s \sqrt{p_s - p_1} \; , \quad x_s \geq 0 \\ q &= K_d x_s \sqrt{p_1 - p_t} \; , \quad x_s < 0 \end{aligned}$$

mit dem Ventilkoeffizienten K_d, dem Versorgungsdruck p_s und dem Tankdruck p_t ebenfalls berücksichtigt werden, wobei die Ventilauslenkung x_s den neuen Stelleingriff darstellt. Der in industriellen Anwendungen teilweise vorkommende Fall, dass ein DAP über nur ein 4-Wege-Servoventil angesteuert wird, erfordert eine genauere Analyse der sich einstellenden Ruhelagen. Er wird hier nicht behandelt.

Die Fluidmenge, die sich in der jeweiligen Kammer mit geometrischem Volumen $V_1 = V_{01} + A_1 x_p$ bzw. $V_2 = V_{02} - A_2 x_p$ (mit dem Offsetvolumen V_{0i} und dem effektiven Querschnitt A_i) befindet, wird durch ihre Masse m_i angegeben. Es wird zwischen zwei Koordinatensätzen, nämlich den kanonischen Koordinaten $x^T = [x_p, p_p, m_1, m_2]$ mit dem Impuls $p_p = m_p v_p$ und der Geschwindigkeit $v_p = \dot{x}_p$, und den Sensorkoordinaten $\breve{x}^T = [x_p, v_p, p_1, p_2]$ unterschieden werden. Dies sind Größen, die mit Hilfe herkömmlicher Sensorik (zum Großteil) direkt messbar sind.

Das dynamische Modell für die beiden hydraulischen Kammern ergibt sich jeweils bereits aus (11). Zusätzlich wird noch die Impulsbilanz

$$m_p \dot{v}_p = -F_c - F_d + A_1 p_1 - A_2 p_2 \tag{12}$$

für das mechanische Teilsystem mit der Masse des Kolbens und der Last m_p, der Feder- und der Dämpferkraft (F_c, F_d) und den Kammerdrücken benötigt. Aus den bereits erwähnten Gründen wurde die Masse des Fluids in (12) vernachlässigt. Schlussendlich seien $F_c = c\,(x_p - x_{p0})$ und $F_d = d v_p$ mit den Koeffizienten c, d. In den Koordinaten $x^T = [x_p, p_p, m_1, m_2]$ lässt sich die Energie des Gesamtsystems als

$$\begin{aligned} H &= \Delta H + \frac{p_p^2}{2m_p} + \frac{c\,(x_p - x_{p,0})^2}{2} \\ \Delta H &= V_1 E \ln \frac{V_1}{\rho_0^{-1} m_1} + (E + p_0)\left(\rho_0^{-1} m_1 - V_1\right) + \\ &\qquad V_2 E \ln \frac{V_2}{\rho_0^{-1} m_2} + (E + p_0)\left(\rho_0^{-1} m_2 - V_2\right) \end{aligned} \tag{13}$$

schreiben, welche gleichzeitig als Hamiltonfunktion dient. Mit (6, 7, 10, 11, 12, 13) ergeben sich die Struktur- und Dissipationsmatrizen zu

$$J = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad R = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{14}$$

und die Torbeziehungen für den Eingang $[q_1, q_2]^T$ zu

$$\begin{aligned} G^T &= \begin{bmatrix} 0 & 0 & m_1 V_1^{-1} & 0 \\ 0 & 0 & 0 & m_2 V_2^{-1} \end{bmatrix} \\ y^T &= \begin{bmatrix} \rho_1 h_1 & \rho_2 h_2 \end{bmatrix}. \end{aligned} \tag{15}$$

Offensichtlich liegt (14) bereits in der Form (2) vor und damit ist der Begriff kanonische Koordinaten gerechtfertigt. Weiters sind m_1 und m_2 Casimirfunktionen, welche konstant bleiben, solange die Zuflüsse q_1 und q_2 verschwinden. Hier zeigt sich, wie physikalische Überlegungen die Suche nach kanonischen Koordinaten erleichtern können. Davon abgesehen kann die Hamiltonfunktion als ein Kandidat für eine Lyapunov-Funktion zur Stabilitätsuntersuchung der ungeregelten Strecke verwendet werden.

Aus dem Modell des doppelt wirkenden Aktuators lässt sich das Modell des einfach wirkenden Aktuators (SAP) auf folgende Weise gewinnen. Für die zweite Zylinderkammer wird $q_2 = 0$ und in der Hamiltonfunktion $m_2 = \rho_0 V_{2,0}$, $V_2 - V_{2,0} = \Delta V_2 = x_p A_2$ gesetzt. Mit dem Grenzübergang $\lim \frac{\Delta V_2}{V_2} \to 0$ ergibt sich für den Hydraulikanteil

$$\Delta H_{SAP} = V_1 E \ln \frac{V_1}{\rho_0^{-1} m_1} + (E + p_0)\left(\rho_0^{-1} m_1 - V_1\right) - x_p A_2 p_0 .$$

Die Differentialgleichung für die Koordinate m_2 ist in diesem Fall natürlich redundant.

3 Entwurf einer PCHD-basierten Regelung

Für das im vorangegangenen Abschnitt entwickelte PCHD-Modell des DAP soll nun ein Regelgesetz zur Positionierung der Last entwickelt werden. Es wird hierzu die physikalische Kopplungsstruktur des Modells ausgenützt. Aus (12) ist unmittelbar ersichtlich, dass nur die Differenz der mit den jeweiligen Querschnitten gewichteten Drücke – die resultierende Kraft –

$$F_h = A_1 p_1 - A_2 p_2$$

auf das mechanische Teilsystem wirkt. Um dieser Tatsache Rechnung zu tragen, werden neue kanonische Koordinaten $\xi = \left[\xi_1^T, \xi_2^T\right]^T$ gesucht, die sich durch eine Zustandstransformation $x = \phi(\xi)$ so ergeben, dass das System mit einem neuen Eingang $u = [u_1, u_2]^T$ und $q = \psi(u)$ in zwei entkoppelte Teilsysteme zerfällt. Allgemein muss für die Transformation ϕ zwischen den beiden Koordinatendarstellungen $\dot{\xi} = \bar{f}(\xi, u)$ und $\dot{x} = f(x, q)$

$$f \circ (\phi, \psi) = (\partial_\xi \phi)\, \bar{f}$$

gelten. Man wählt $\xi_1 = [x_p, p_p, z_1]^T$ und $\xi_2 = z_2$ und macht für die Transformation ϕ den Ansatz $\xi^T \mapsto [x_p, p_p, \phi_{m_1}(z_1, z_2), \phi_{m_2}(z_1, z_2)]^T$. Diese Wahl von ϕ lässt die Matrizen J und R aus (14) unverändert. Obige Bedingung ergibt für (13, 14, 15) die Beziehung

$$\left(\begin{bmatrix} f(x) \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ g_1(x)\, q_1 \\ g_2(x)\, q_2 \end{bmatrix}\right) \circ (\phi, \psi) = \begin{bmatrix} I_2 & 0 & 0 \\ 0 & \partial_{z_1}\phi_{m_1} & \partial_{z_2}\phi_{m_1} \\ 0 & \partial_{z_1}\phi_{m_2} & \partial_{z_2}\phi_{m_2} \end{bmatrix} \left(\begin{bmatrix} f(\xi_1) \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ u_1 \\ u_2 \end{bmatrix}\right)$$

mit $f(x) = \left[\partial_{p_p} H(x), -\partial_{x_p} H(x) - d\partial_{p_p} H(x)\right]^T$, $g_1(x) = m_1 V_1^{-1}$ und $g_2(x) = m_1 V_2^{-1}$, wobei die Struktur des Zielsystems entsprechend gewählt wurde. Weiters erhält man daraus die Bedingung

$$f(x) \circ \phi = f(\xi_1)$$

oder

$$\partial_{z_2}(f(x) \circ \phi) = 0\,. \tag{16}$$

Die Stellgrößentransformation ψ muss den Beziehungen

$$\begin{aligned} (g_1(x)\, q_1) \circ (\phi, \psi) &= \partial_{z_1}\phi_{m_1} u_1 + \partial_{z_2}\phi_{m_1} u_2 \\ (g_2(x)\, q_2) \circ (\phi, \psi) &= \partial_{z_1}\phi_{m_2} u_1 + \partial_{z_2}\phi_{m_2} u_2\,. \end{aligned}$$

genügen. Von (16) erhält man die Gleichung

$$\frac{EA_1}{\phi_{m_1}} \frac{\partial \phi_{m_1}}{\partial z_2} - \frac{EA_2}{\phi_{m_2}} \frac{\partial \phi_{m_2}}{\partial z_2} = 0\,, \tag{17}$$

und eine Lösung von (17) ist durch

$$\phi_{m_2} = F(z_1)\, (\phi_{m_1})^{\frac{A_1}{A_2}}$$

mit einer beliebigen Funktion $F(z_1)$ gegeben. Mit der speziellen Wahl der transformierten Casimirfunktionen

$$z_1 = EA_1 \ln\left(\frac{m_1}{\rho_0}\left(\frac{\rho_0}{m_2}\right)^{\frac{A_2}{A_1}}\right) + p_0(A_1 - A_2) \ , \qquad z_2 = EA_2 \ln\left(\frac{m_2}{V_{02}\rho_0}\right) + p_0 A2$$

und der transformierten Eingänge

$$u_1 = \frac{EA_1}{V_{01} + A_1 x_p} q_1 - \frac{EA_2}{V_{02} - A_2 x_p} q_2 \ , \qquad u_2 = \frac{EA_2}{V_{02} - A_2 x_p} q_2$$

erhält man ein PCHD-System, das aus zwei entkoppelten Teilsystemen besteht, mit der transformierten Hamiltonfunktion und Eingangsmatrix

$$H(\xi) = H(x) \circ \phi(\xi) \ , \qquad G^T = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \ , \tag{18}$$

dem Eingang $[u_1, u_2]^T$ und den ursprünglichen Matrizen (14). Als Regelgrößen werden die gewünschte Position $\mathring{x}_p$ und eine Sollgröße $\mathring{z}_2$ festgelegt. Damit ist aber die Ruhelage $\mathring{\xi} = [\mathring{x}_p, 0, \mathring{z}_1, \mathring{z}_2]^T$ des Systems (14, 18) eindeutig bestimmt. Das gesuchte Regelgesetz soll nun garantieren, dass die Abweichungen $\bar{\xi} = \xi - \mathring{\xi}$ von der Ruhelage gegen Null streben, und sich das geregelte System in den Koordinaten $\bar{\xi}$ aus zwei vollständig entkoppelten PCHD-Systemen zusammensetzt. Hierfür werden die gewünschte positiv definite Hamiltonfunktion des Zielsystems

$$\begin{aligned} H_d(\bar{\xi}) = H_{d1}(\bar{\xi}_1) + H_{d2}(\bar{\xi}_2) = c_m \frac{\bar{x}_p^2}{2} + \frac{p_p^2}{2m_p} + \Gamma_1 \frac{\bar{z}_1^2}{2} + \\ + E \int_0^{\bar{x}_p} \left(A_1 \ln\left(\frac{V_{01} + A_1(\mathring{x}_p + \tau)}{V_{01} + A_1 \mathring{x}_p}\right) + A_2 \ln\left(\frac{V_{02} - A_2 \mathring{x}_p}{V_{02} - A_2(\mathring{x}_p + \tau)}\right)\right) \mathrm{d}\tau + \Gamma_2 \frac{\bar{z}_2^2}{2} \end{aligned}$$

und neue Matrizen

$$J_d = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & \beta & 0 \\ 0 & -\beta & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \ R_d = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & d & -\beta & 0 \\ 0 & -\beta & \frac{\alpha_1}{\Gamma_1} & 0 \\ 0 & 0 & 0 & \frac{\alpha_2}{\Gamma_2} \end{bmatrix}$$

mit $\beta = (2\Gamma_1)^{-1}$ und $\Gamma_1 = \Gamma_2 = 1\mathrm{mN}^{-1}$ und den Reglerparametern α_1 und α_2 gewählt. Damit ist aber bereits das Regelgesetz im Sinne des IDA-PBC Entwurfs (siehe [7]) mit

$$q_1 = -\frac{\alpha_1 \bar{z}_1 + \alpha_2 \bar{z}_2}{EA_1}(V_{01} + A_1(\mathring{x}_p + \bar{x}_p)) \ , \qquad q_2 = -\frac{\alpha_2 \bar{z}_2}{EA_2}(V_{02} - A_2(\mathring{x}_p + \bar{x}_p))$$

bestimmt, welches den geschlossenen Kreis asymptotisch stabilisiert, soferne $d\alpha_1 - \frac{1}{4} > 0$ und $\alpha_2 > 0$ gilt.

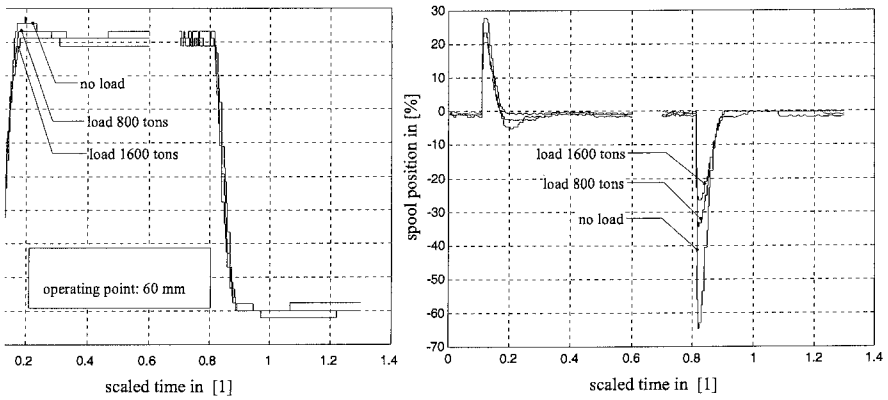


Bild 2: Sprungantworten mit dem nichtlinearen Regelgesetz unter verschiedenen Lasten.

Eine Variante dieses Regelgesetzes für den einfach wirkenden Aktuator ist bereits in einigen Stahlwalzwerken als der innerste Kreis einer Banddickenregelung im Einsatz. Durch den heuristischen Charakter des sogenannten Walzkraftmodells, welches die Beziehung zwischen der für den Walzprozess nötigen Kraft, der erzielten Banddicke und einigen anderen Einflussfaktoren zu erklären versucht, ist ein genaues Lastmodell für den Aktuator nicht in einfacher Weise zugänglich. Allerdings ist es wohlbekannt, dass sich das Walzgut in selbst für die Einhaltung kleinster Dickentoleranzen ausreichender Weise durch das hier verwendete Lastmodell beschreiben lässt. Für eine hohe Güte des Gesamtsystems ist aber das Verhalten des Aktuators im gesamten Arbeitbereich wesentlich. Hier sei auf den Unterschied zu einem um einen Arbeitspunkt entworfenen linearen Regler verwiesen. Die Hydraulik, also der Aktuator, wird voll nichtlinear behandelt, wodurch essentielle Steigerungen der Anstiegsgeschwindigkeit bei Positionssprüngen erzielt werden können.

Stellvertretend seien einige Messreihen aus einem Werk der Bethlehem Steel Corporation (Bethlehem, Pennsylvania) angeführt (Abb. 2), wobei die Amplitude des Quantisierungsrauschens bei etwa 1.1 μm liegt. Zu Demonstrationszwecken wurden hierfür die rotierenden (um Beschädigungen vorzubeugen) Arbeitswalzen in Kontakt gebracht und mit unterschiedlichen Belastungen Positionssprünge durchgeführt. Die Qualität der Regelung zeigt sich in der Lastinvarianz der Positionssignale. Die Ventilsignalverläufe zeigen das nichtlineare Verhalten der Regelung. Die relativ großen Variationen in den Kraftsignalen sind auf Störungen durch die leicht exzentrischen, rotierenden Stützwalzen zurückzuführen.

4 Zusammenfassung

In diesem Beitrag wird mit Hilfe der Gleichgewichtsthermodynamik gezeigt, wie sich Systeme mit einem massenstromgebundenen Leistungstransport über die Systemgrenzen in das Konzept der PCHD-System einfügen. Als spezielles Beispiel dient ein doppelt

wirkender hydraulischer Aktuator, welcher über zwei Servoventile angesteuert wird. Für dieses Mehrgrößensystem wir gezeigt, wie eine Koordinatentransformation gefunden werden kann, sodass es sich durch zwei getrennte PCHD-Systeme darstellen lässt. Basierend auf dieser entkoppelten Darstellung wird ein passivitätsbasiertes Positionsregelgesetz entwickelt, welches die Ruhelage des Aktuators asymptotisch stabilisiert. Abschließend werden noch Messresultate einer speziellen Variante des entwickelten Regelgesetzes aus der Stahlindustrie präsentiert. Dort ist dieser Regler bereits mehrfach im Einsatz.

5 Dank und Anerkennung

Dank gebührt unserem Partner VOEST ALPINE INDUSTRIEANLAGENBAU GMBH (VAI) für die Unterstützung im Rahmen des *Christian Doppler Laboratoriums (CDL) für Automatisierung mechatronischer Systeme der Stahlindustrie* sowie für die Bereitstellung von Daten und Messergebnissen. Ebenso bedanken sich die Autoren bei Herrn DI. Dr. Rainer Novak, der das hydraulische Regelgesetz das erste Mal in einer Anlage implementierte.

Literatur

[1] Burke, W. L. *Applied Differential Geometry.* Cambridge, 1997.

[2] Chorin, A. J. and Marsden, J. E. *A mathematical introduction to fluid mechanics.* Springer, 1990.

[3] Grabmair, G., Schlacher, K., and Kugi, A. Geometric Energy Based Analysis and Controller Design of Hydraulic Actuators Applied in Rolling Mills. In *ECC03 CD publication.* Cambridge, 2003.

[4] Kugi, A. and Schlacher, K. Dissipativitäts- und passivitätsbasierte Regelung nichtlinearer mechatronischer Systeme. *e&i*, 2001(1), pp. 40–48, 2001.

[5] Merritt, H. E. *Hydraulic control systems.* John Wiley & Sons, 1967.

[6] Olver, P. J. *Applications of Lie Groups to Differential Equations.* Springer, 2nd edn., 1993.

[7] Ortega, R., van der Schaft, A., et al. Interconnection and damping assignment passivity-based control of port-controlled Hamiltonian systems. *Automatica*, 38(4), 2002.

[8] van der Schaft, A. *L2-Gain and passivity techniques in nonlinear control.* Springer, 2000.

Wellendigitalsimulation mit passiven Zweischritt-Runge-Kutta-Verfahren

Dietrich Fränken
Fachgebiet Nachrichtentheorie
Universität Paderborn
fraenken@uni-paderborn.de

Karlheinz Ochs
Lehrstuhl für Nachrichtentechnik
Ruhr-Universität Bochum
ochs@nt.ruhr-uni-bochum.de

Kurzfassung: In diesem Beitrag wird der Einsatz von passiven Zweischritt-RUNGE-KUTTA-Verfahren im Rahmen des Wellendigitalkonzeptes diskutiert. Gemeinsamkeiten und Unterschiede zur Wellendigitalsimulation mit RUNGE-KUTTA-Verfahren werden aufgezeigt.

1 Einführung

Ursprünglich wurde das Wellendigitalkonzept eingeführt, um neuartige Digitalfilterstrukturen zu gewinnen, die sich durch auf einfache Weise zu gewährleistende Stabilitätseigenschaften auszeichnen [1, 2]. Später wurde erkannt, dass sich das Konzept auch zur digitalen Simulation physikalischer Systeme eignet [3]-[5], wobei die Passivität der eingesetzten Verfahren eine Vielzahl wünschenswerter numerischer Stabilitätseigenschaften sicherstellt [3, 6, 7]. Allerdings kamen im Rahmen dieses Konzeptes lange Zeit ausschließlich lineare Mehrschrittverfahren zum Einsatz, bei denen die Passivität gleichzeitig die erreichbare numerische Genauigkeit in bestimmten Anwendungsfällen stark beschränkt. Durch die Anwendung passiver RUNGE-KUTTA-Verfahren konnte diese Genauigkeitsgrenze, bei unverändert guten numerischen Stabilitätseigenschaften, deutlich erhöht werden [8]-[11].

Nach einer Einführung in die Prinzipien einer Wellendigitalsimulation mit linearen Mehrschrittverfahren und mit RUNGE-KUTTA-Verfahren wird in diesem Beitrag die Anwendung von passiven Zweischritt-RUNGE-KUTTA-Verfahren im Rahmen des Wellendigitalkonzeptes diskutiert. Insbesondere wird ein in [12] angegebenes Verfahren systematisch hergeleitet, welches allgemein eine höhere numerische Genauigkeit als ein vergleichbares RUNGE-KUTTA-Verfahren aufweist. Es wird aber auch bewiesen, dass durch den Einsatz der hier betrachteten Zweischritt-RUNGE-KUTTA-Verfahren die erreichbare Genauigkeit im Vergleich zu geeigneten passiven RUNGE-KUTTA-Verfahren insgesamt nicht erhöht werden kann.

2 Das Wellendigitalkonzept

Betrachtet wird das zeitkontinuierliche KIRCHHOFF-Netzwerk in Abbildung 1. Abbildung 2 verdeutlicht die übliche Vorgehensweise bei einer digitalen Simulation dieses Netzwerkes mit Hilfe des Wellendigitalkonzeptes. Hierzu wird eine Wellendigitalstruktur erzeugt, indem für ein gegebenes lineares Netzwerk mit (zeitlich) konstanten Elementen die folgenden Schritte durchgeführt werden:

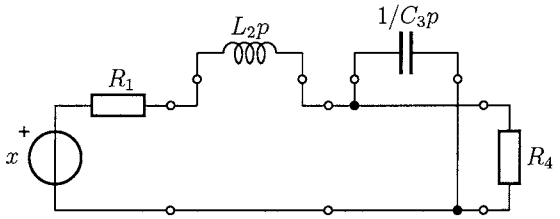


Abbildung 1: Zeitkontinuierliches KIRCHHOFF-Netzwerk.

Es wird ein *Referenznetzwerk* erzeugt, indem die gewöhnliche Frequenzvariable p durch $2\psi/T$ mit der *äquivalenten komplexen Frequenzvariablen*

$$\psi = \tanh(pT/2) = \frac{\mathrm{e}^{pT} - 1}{\mathrm{e}^{pT} + 1} \tag{1}$$

ersetzt wird, wobei T die zugehörige *Betriebsperiode* darstellt. Dabei bleiben sämtliche algebraischen Zusammenhänge, die dynamikfreie Elemente beschreiben, ebenso von diesem Übergang unbeeinflusst wie die KIRCHHOFF-Gleichungen. Lediglich die ursprünglich durch Differentialgleichungen beschriebenen reaktiven Elemente Kapazität und Induktivität werden jetzt durch Differenzengleichungen beschrieben. So erhält man bei einer Kapazität, die ursprünglich durch

$$\dot{\tilde{u}}(t) = \frac{1}{C}\,\tilde{i}(t) \quad \Longleftrightarrow \quad \tilde{u}(t) = \tilde{u}(t-T) + \frac{1}{C}\int_{t-T}^{t} \tilde{i}(\tau)\,\mathrm{d}\tau \tag{2}$$

und damit im Sonderfall einer festen Frequenz p mit $\tilde{i}(t) = \hat{i}\mathrm{e}^{pt}$ und $\tilde{u}(t) = \hat{u}\mathrm{e}^{pt}$ durch

$$\hat{u} = \frac{1}{pC}\hat{i} \tag{3}$$

beschrieben wird, nun für $i(t) = \hat{i}\mathrm{e}^{pt}$ und $u(t) = \hat{u}\mathrm{e}^{pt}$ die Beziehung

$$\hat{u} = \frac{\mathcal{R}}{\psi}\hat{i} \quad \text{mit} \quad \mathcal{R} := \frac{T}{2C} \tag{4}$$

bzw. allgemein

$$u(t) = u(t-T) + \mathcal{R}\left[i(t-T) + i(t)\right]\,. \tag{5}$$

Die Induktivität ist entsprechend zu behandeln. Sie wird aber im Weiteren nicht mehr immer gesondert diskutiert, da sie stets in Form eines mit einer Kapazität abgeschlossenen Gyrators realiert werden kann.

Im zweiten Schritt wird das auf diese Weise gewonnene Referenznetzwerk in *torweise verschaltete Komponenten* zerlegt, und so genannte *Wellengrößen* werden als Signalgrößen an den Toren eingeführt. Diese ergeben sich nach Festlegung eines geeigneten positiven *Torwiderstandes* R jeweils als Linearkombination von Torspannung u und Torstrom i in der Form

$$a = u + Ri \quad \text{und} \quad b = u - Ri\,. \tag{6}$$

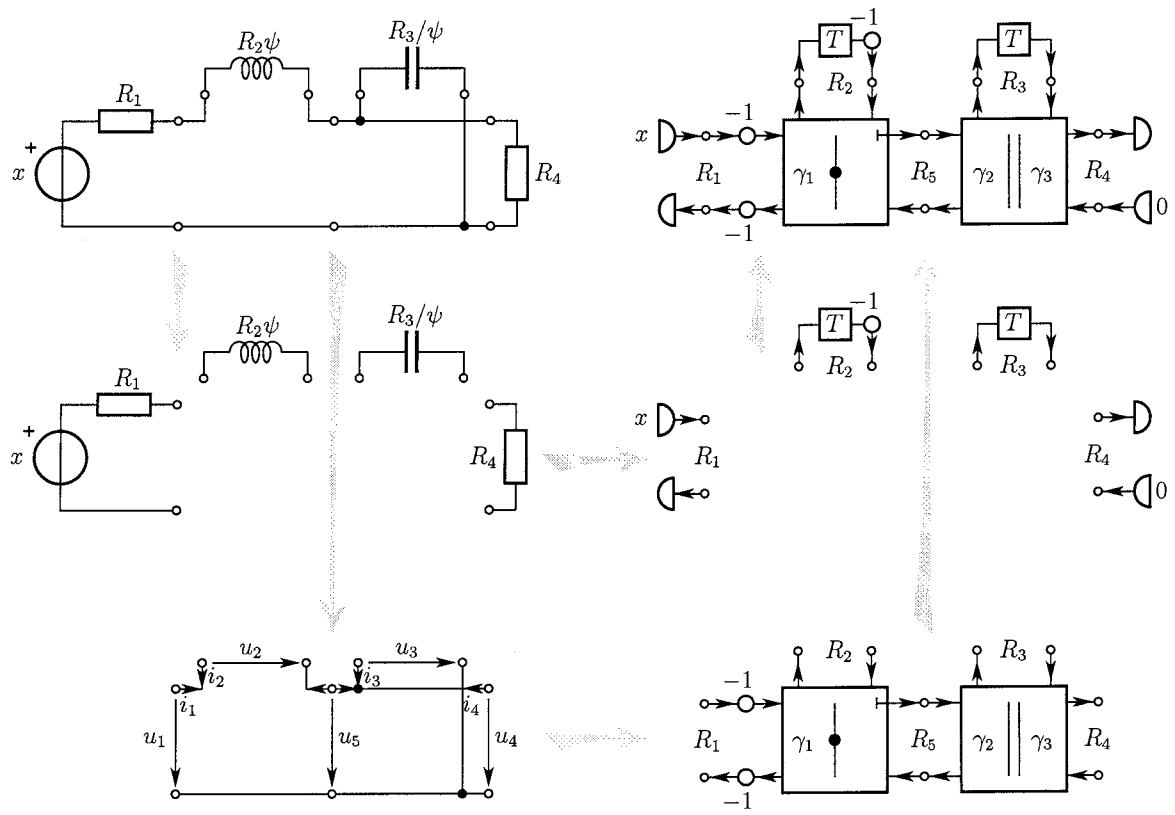


Abbildung 2: Referenznetzwerk zu Abbildung 1 und der Weg zur Wellendigitalstruktur: Torweise Zerlegung, Synthese von Quellen und Elementen, Nachbildung der Verbindungsstruktur mit Adaptoren. Zur Vermeidung verzögerungsfreier gerichteter Schleifen wurde der zunächst freie Torwiderstand R_5 zu $R_1 + R_2$ gewählt.

Die ursprünglich in Spannung und Strom gegebenen Elementebeziehungen werden nun in Abhängigkeit von den Wellengrößen formuliert. Es ergeben sich entsprechende *Wellendigitalbausteine.* Dabei bestimmt der Wunsch nach einer möglichst einfachen Realisierung die im vorigen Punkt angesprochene Wahl des Torwiderstandes. So erhält man insbesondere für eine Kapazität C aus (5) die Beziehung

$$b(t) = a(t - T) \tag{7}$$

und damit eine einfache Verzögerung als Wellendigitalrealisierung, wenn als Torwiderstand der Wert $\mathcal{R}$ gemäß (4) gewählt wird. Bei der Induktivität tritt, der Realisierung durch Gyrator und Kapazität entsprechend, ein zusätzlicher Vorzeicheninverter auf.

Zuletzt sind noch die KIRCHHOFF-Gleichungen in Abhängigkeit von den Wellengrößen zu formulieren. Die hieraus resultierenden Wellendigitalbausteine werden als *Adaptoren* bezeichnet. Dabei kann durch geeignete Wahl der Torwiderstände zwischen den Adaptoren das Auftreten verzögerungsfreier gerichteter Schleifen vermieden werden.

Die digitale Simulation besteht nun lediglich in einer Abarbeitung des entstandenen Signalflussgraphen. Aus den so zu äquidistanten diskreten Zeitpunkten bestimmten Größen

a und b können die eigentlich gesuchten Werte u und i als Näherungslösungen für die im ursprünglichen Netzwerk auftretenden Werte $\tilde{u}$ und $\tilde{i}$ aus

$$u = \frac{a+b}{2} \quad \text{sowie} \quad i = \frac{a-b}{2R} \tag{8}$$

ermittelt werden.

3 Passive LMS-Verfahren

Ein Vergleich der Beziehungen (2) und (5) zeigt, dass das im vorigen Abschnitt beschriebene Vorgehen einer Anwendung der Trapezregel auf das dem System zugrunde liegende differential-algebraische Gleichungssystem entspricht. Auch der Einsatz anderer linearer Mehrschrittverfahren (kurz *LMS-Verfahren*) wie etwa des impliziten EULER-Verfahrens (Einschritt-BDF) oder des GEAR-Verfahrens (Zweischritt-BDF) ist in diesem Kontext möglich. Anstelle von (5) sind dann die jeweiligen Differenzengleichungen

$$u(t) = u(t-T) + 2\mathcal{R}i(t) \quad \text{bzw.} \quad u(t) = \tfrac{4}{3}u(t-T) - \tfrac{1}{3}u(t-2T) + \tfrac{4}{3}\mathcal{R}i(t)\,. \tag{9}$$

zu verwenden, was bedeutet, dass in (4) statt $\mathcal{Z}(\psi) = \frac{1}{\psi}$ dann

$$\mathcal{Z}(\psi) = \frac{1}{\psi} + \psi \quad \text{bzw.} \quad \mathcal{Z}(\psi) = \frac{1}{\psi} + \frac{\psi}{1+2\psi} \tag{10}$$

als (normierte) *charakteristische Impedanz* des Verfahrens erscheint. Die drei hier genannten Verfahren zeichnen sich dadurch aus, dass die jeweilige charakteristische Impedanz eine *positive Funktion* ist (also der Implikation

$$\operatorname{Re}\psi \geq 0 \implies \operatorname{Re}\mathcal{Z}(\psi) \geq 0 \tag{11}$$

genügt) und somit ein passives (normiertes) *charakteristisches Eintor* repräsentiert. Dies stellt sicher, dass ein passives zeitkontinuierliches Netzwerk stets in einem ebenfalls passiven Referenznetzwerk resultiert, weshalb man in diesem Zusammenhang auch von *passiven LMS-Verfahren* spricht. Da durch die Passivität eines numerischen Integrationsverfahrens zahlreiche wünschenswerte numerische Stabilitätseigenschaften impliziert werden, werden bevorzugt solche Verfahren im Rahmen des Wellendigitalkonzeptes eingesetzt.

An dieser Stelle sei noch erwähnt, dass sich das Wellendigitalkonzept trotz seiner Einführung mittels einer Frequenzsubstitution und der Charakterisierung von Verfahren durch Impedanzen auch, wie das Beispiel Abschnitt 7 noch bestätigen wird, für die digitale Simulation nichtlinearer und/oder zeitvarianter Netzwerke einsetzen lässt. Es ist aber auch zu erwähnen, dass die Simulation mit dem eigentlich gewünschten LMS-Verfahren in der Regel erst dann erfolgen kann, nachdem mittels einer so genannten *Startphase* geeignete Anfangszustände für die Verzögerungen ermittelt worden sind, da die hierin gespeicherten Wellengrößen meist nicht mit der gesuchten Spannung übereinstimmen[1].

[1] Dies gilt sowohl für die Trapezregel als auch für das GEAR-Verfahren, nicht aber für das implizite EULER-Verfahren, das somit zur Durchführung der Startphase herangezogen werden kann.

4 Passive Runge-Kutta-Verfahren

Betrachtet werde nun zunächst ein einfaches elektrisches Netzwerkes, welches neben einer Kapazität C mit Spannung $\tilde{u}$ und Strom $\tilde{\imath}$ sowie einer Spannungsquelle mit Quellspannung x lediglich dynamik- und quellenfreie Elemente enthält und durch ein Anfangswertproblem der Form

$$\dot{\tilde{u}}(t) = \frac{1}{C}\tilde{\imath}(t) = f(\tilde{u}(t), x(t)), \quad \tilde{u}(t_0) = u_0 \tag{12}$$

beschrieben wird. Die Anwendung eines RUNGE-KUTTA-*Verfahrens* [13, 14] kann in diesem Fall in der vektoriellen Form

$$\boldsymbol{u}(t) = \boldsymbol{e}u(t-T) + \frac{T}{C}\boldsymbol{D}\boldsymbol{i}(t) \tag{13a}$$

$$u(t) = \quad u(t-T) + \frac{T}{C}\boldsymbol{g}^{\mathrm{T}}\boldsymbol{i}(t) \tag{13b}$$

$$\frac{1}{C}i_\sigma(t) = f(u_\sigma(t), x_\sigma(t)) \tag{14a}$$

und

$$x_\sigma(t) = x(t - k'_\sigma T) \quad \text{mit} \quad k'_\sigma = 1 - k_\sigma\,. \tag{14b}$$

ausgedrückt werden, wobei die reellen Größen $d_{\sigma\varrho}$, g_σ und k_σ mit $\sigma, \varrho = 1, \ldots, s$ geeignet gewählte Parameter des Verfahrens sind. Hierbei bezeichnet s die Zahl der *Stufen* des Verfahrens, die Vektoren $\boldsymbol{u}$ und $\boldsymbol{i}$ fassen jeweils die *Stufenspannungen* u_σ und *-ströme* i_σ zusammen, und $\boldsymbol{e} = [1, \ldots, 1]^{\mathrm{T}}$ ist ein Vektor passender Dimension.

Wie bei den LMS-Verfahren soll auch hier der Sonderfall einer festen Frequenz p betrachtet werden. Die vektorielle Differenzengleichung (13) führt dann mit Signalen der Form $\boldsymbol{u}(t) = \hat{\boldsymbol{u}}\mathrm{e}^{pt}$ und $\boldsymbol{i}(t) = \hat{\boldsymbol{\imath}}\mathrm{e}^{pt}$ bei Verwendung eines wie in (4) gewählten Bezugswiderstandes $\mathcal{R}$ und nach Übergang auf die komplexe äquivalente Frequenz ψ auf einen Zusammenhang der Form $\hat{\boldsymbol{u}} = \mathcal{R}\boldsymbol{\mathcal{Z}}(\psi)\hat{\boldsymbol{\imath}}$ mit der (normierten) *charakteristischen Impedanzmatrix*

$$\boldsymbol{\mathcal{Z}}(\psi) = \frac{1}{\psi}\boldsymbol{e}\boldsymbol{g}^{\mathrm{T}} + 2\boldsymbol{D} - \boldsymbol{e}\boldsymbol{g}^{\mathrm{T}}\,. \tag{15}$$

Umgekehrt sind, mit Ausnahme der zusätzlich anzugebenden *Knotenwerte* k_σ, durch $\boldsymbol{\mathcal{Z}}(\psi)$ die Parameter des Verfahrens eindeutig festgelegt, da $\boldsymbol{D}$ und $\boldsymbol{g}^{\mathrm{T}}$ aus

$$2\boldsymbol{D} = \boldsymbol{\mathcal{Z}}(1) \quad \text{bzw.} \quad s\boldsymbol{g}^{\mathrm{T}} = \boldsymbol{e}^{\mathrm{T}} \lim_{\psi\to 0} \psi\boldsymbol{\mathcal{Z}}(\psi) \tag{16}$$

unmittelbar zu ermitteln sind.

In Analogie zum Sprachgebrauch bei LMS-Verfahren soll ein RUNGE-KUTTA-Verfahren als *passiv* bezeichnet werden, wenn seine charakteristische Impedanzmatrix ein passives Mehrtor beschreibt. Dies ist dann der Fall, wenn die in Gleichung (15) definierte Impedanzmatrix $\boldsymbol{\mathcal{Z}}(\psi)$ eine *positive Matrix* ist, also die Beziehung

$$\mathrm{Re}\,\psi \geq 0 \quad \Longrightarrow \quad \boldsymbol{\mathcal{Z}}(\psi) + \boldsymbol{\mathcal{Z}}^*(\psi) \geq \boldsymbol{0} \tag{17}$$

erfüllt[2]. Diese Anforderung hat einerseits auf die frei wählbaren Parameter des Verfahrens und andererseits auf mögliche Realisierungen des *charakteristischen Mehrtores* Auswirkungen, die eng miteinander in Verbindung stehen.

Soll die Matrix $\boldsymbol{\mathcal{Z}}(\psi)$ nach (15) positiv sein, so müssen insbesondere die Grenzfälle

$$\lim_{\psi \to 0+} \psi \left[\boldsymbol{\mathcal{Z}}(\psi) + \boldsymbol{\mathcal{Z}}^*(\psi)\right] \quad \text{und} \quad \lim_{\psi \to \infty} \left[\boldsymbol{\mathcal{Z}}(\psi) + \boldsymbol{\mathcal{Z}}^*(\psi)\right]$$

positiv semidefinite Matrizen darstellen. Die Koeffizienten des Verfahrens müssen demnach den Ungleichungen

$$\boldsymbol{e}\boldsymbol{g}^{\mathrm{T}} + \boldsymbol{g}\boldsymbol{e}^{\mathrm{T}} \geq \boldsymbol{0} \quad \text{sowie} \quad 2\boldsymbol{D} + 2\boldsymbol{D}^{\mathrm{T}} - \boldsymbol{e}\boldsymbol{g}^{\mathrm{T}} - \boldsymbol{g}\boldsymbol{e}^{\mathrm{T}} \geq \boldsymbol{0} \tag{18}$$

genügen. Es lässt sich aber mittels der linken Ungleichung einfach nachweisen, dass der Vektor $\boldsymbol{g}^{\mathrm{T}}$ von der Form

$$\boldsymbol{g}^{\mathrm{T}} = r\boldsymbol{e}^{\mathrm{T}} \quad \text{mit} \quad r > 0 \tag{19}$$

sein muss [8]. Zerlegt man unter dieser Voraussetzung noch die Restmatrix

$$\boldsymbol{\mathcal{Q}} := 2\boldsymbol{D} - \boldsymbol{e}\boldsymbol{g}^{\mathrm{T}} = 2\boldsymbol{D} - r\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}} \tag{20}$$

in ihren symmetrischen Teil

$$\boldsymbol{\mathcal{R}} := \frac{1}{2}[\boldsymbol{\mathcal{Q}} + \boldsymbol{\mathcal{Q}}^{\mathrm{T}}] = \boldsymbol{D} + \boldsymbol{D}^{\mathrm{T}} - r\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}} \tag{21}$$

und ihren schiefsymmetrischen Teil

$$\boldsymbol{\mathcal{S}} := \frac{1}{2}[\boldsymbol{\mathcal{Q}} - \boldsymbol{\mathcal{Q}}^{\mathrm{T}}] = \boldsymbol{D} - \boldsymbol{D}^{\mathrm{T}}\,, \tag{22}$$

so wird aus der rechten Ungleichung in (18) die Forderung $2\boldsymbol{\mathcal{R}} \geq \boldsymbol{0}$. In Verbindung mit der Anforderung (19) bedeutet dies, dass sich die charakteristische Impedanzmatrix eines passiven RUNGE-KUTTA-Verfahrens in der Form

$$\boldsymbol{\mathcal{Z}}(\psi) = \frac{r}{\psi}\,\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}} + \boldsymbol{\mathcal{R}} + \boldsymbol{\mathcal{S}} \tag{23}$$

mit dem positiven Wert r schreiben lässt, wobei die symmetrische Matrix $\boldsymbol{\mathcal{R}}$ positiv semidefinit sein muss, während sich keine weiteren Restriktionen an die schiefsymmetrische Matrix $\boldsymbol{\mathcal{S}}$ ergeben. Dabei führt die Forderung nach Positivität der Matrix $\boldsymbol{\mathcal{R}}$ und somit nach stets nichtnegativen Hauptminoren dieser Matrix auf ein polynomiales Ungleichungssystem in den Einträgen $r_{\varrho\sigma}$.

Nun sind alle Parameter der schiefsymmetrischen Matrix $\boldsymbol{\mathcal{S}}$ bereits durch die Angabe einer strikt unteren Dreiecksmatrix $\boldsymbol{\mathcal{L}}$ eindeutig bestimmt, wenn man

$$\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{L}} - \boldsymbol{\mathcal{L}}^{\mathrm{T}} \quad \text{mit} \quad l_{\varrho\sigma} = 0 \quad \text{für} \quad \varrho \leq \sigma \tag{24}$$

[2]Für eine hermitesche Matrix $\boldsymbol{M}$ bedeutet die Notation $\boldsymbol{M} \geq \boldsymbol{0}$, dass $\boldsymbol{M}$ positiv semidefinit ist.

schreibt. Für die symmetrische Matrix $\boldsymbol{\mathcal{R}}$ wird zweckmäßigerweise die Parameterdarstellung

$$\boldsymbol{\mathcal{R}} = \sum_{\sigma=1}^{s} r_\sigma \boldsymbol{n}_\sigma \boldsymbol{n}_\sigma^{\mathrm{T}} \quad \text{mit} \quad n_{\varrho\sigma} = 0 \ \text{ für } \ \varrho < \sigma\,, \quad n_{\sigma\sigma} = 1 \tag{25}$$

gewählt, mit der die Forderung $\boldsymbol{\mathcal{R}} \geq \mathbf{0}$ in geeigneter Weise berücksichtigt werden kann. Es lässt sich nämlich zeigen, dass die Matrix $\boldsymbol{\mathcal{R}}$ genau dann positiv semidefinit ist, wenn sie sich in dieser Form schreiben lässt und dabei alle Werte r_σ nichtnegativ sind [15]. Mit den reellen Werten r_σ, $n_{\varrho\sigma}$ und $l_{\varrho\sigma}$ mit $\varrho > \sigma = 1, \ldots, s$ stehen insgesamt die benötigten s^2 freien reellen Parameter für eine Bestimmung der Matrix $\boldsymbol{\mathcal{Q}}$ zur Verfügung. Die Verwendung der Parameter r_σ und $n_{\varrho\sigma}$ zur Darstellung der Matrix $\boldsymbol{\mathcal{R}}$ anstelle der Einträge $r_{\sigma\varrho}$ bietet den Vorteil, auf das denkbar einfachste Kriterium zur Überprüfung der Passivität zu führen: Die Parameter r_σ müssen lediglich nichtnegativ sein.

Die gefundene Parameterdarstellung (23) korrespondiert unmittelbar mit einer Netzwerkrealisierung zur Impedanzmatrix $\boldsymbol{\mathcal{Z}}(\psi) = \boldsymbol{\mathcal{R}}\boldsymbol{\mathcal{Z}}(\psi)$. Dabei ist jeder der drei Summanden mit einem s-Tor zu identifizieren, und das vollständige Netzwerk resultiert dann entsprechend aus einer torweisen Serienverbindung dieser s-Tore. Hierbei lässt sich die Impedanzmatrix $[\mathcal{R}r/\psi]\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}$ mittels einer zu allen Toren parallel geschalteten Kapazität realisieren. Die Parameterdarstellung der symmetrischen Matrix $\boldsymbol{\mathcal{R}}$ aus (25) korrespondiert mit einem Netzwerk, das neben Widerständen $\mathcal{R}r_\sigma$ ideale Übertrager mit Übersetzungsverhältnissen $1 : \pm n_{\varrho\sigma}$ enthält. Die schiefsymmetrische Matrix $\boldsymbol{\mathcal{S}}$ aus (24) repräsentiert ein Netzwerk bestehend aus miteinander verschalteten Gyratoren mit Gyrationswiderständen $\pm\mathcal{R}l_{\varrho\sigma}$. Damit werden die genannten Einschränkungen, denen die Parameter eines passiven Verfahrens genügen müssen, unmittelbar physikalisch interpretierbar. Die Netzwerke für ein zweistufiges Verfahren mit

$$\boldsymbol{\mathcal{R}} = \begin{bmatrix} r_1 & n_{21}r_1 \\ n_{21}r_1 & n_{21}^2 r_1 + r_2 \end{bmatrix} \quad \text{und} \quad \boldsymbol{\mathcal{S}} = \begin{bmatrix} 0 & -l_{21} \\ l_{21} & 0 \end{bmatrix} \tag{26}$$

sind in Abbildung 3 dargestellt, wobei sämtliche Widerstandswerte auf $\mathcal{R}$ normiert wurden.

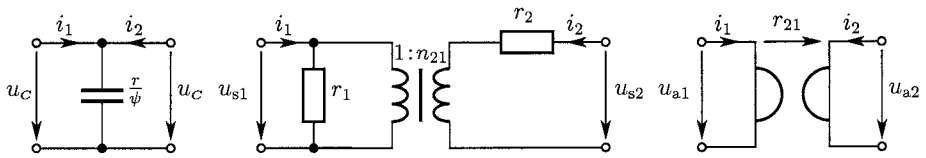


Abbildung 3: Netzwerke zur normierten charakteristischen Impedanzmatrix eines passiven 2-stufigen RUNGE-KUTTA-Verfahrens.

Zwar ist mit den bisherigen Ausführungen eine allgemeingültige Netzwerkrealisierung des charakteristischen Mehrtores für passive RUNGE-KUTTA-Verfahren gefunden, jedoch können die genannten Netzwerke nicht unmittelbar als Referenznetzwerke für eine Wellendigitalsimulation verwendet werden, da in der zugehörigen Wellendigitalstruktur verzögerungsfreie gerichtete Schleifen auftreten würden. Um dies zu verhindern, werden im Rahmen dieses Konzeptes nur so genannte *diagonal-implizite* RUNGE-KUTTA-*Verfahren*

betrachtet, die durch $d_{\sigma\varrho} = 0$ für alle $\varrho > \sigma$ gekennzeichnet sind und abkürzend mit DIRK-Verfahren (engl. *diagonally-implicit Runge-Kutta methods*) bezeichnet werden. Die Matrix $\boldsymbol{D}$ ist in diesem Fall also eine untere Dreiecksmatrix, und dasselbe gilt wegen (16) auch für $\boldsymbol{\mathcal{Z}}(1)$. Führt man nun noch eine Matrix $\boldsymbol{E}$ mit $e_{\sigma\varrho} = -1$ für alle $\varrho > \sigma$, $e_{\sigma\sigma} = 0$ sowie $e_{\sigma\varrho} = 1$ für alle $\varrho < \sigma$ ein, so kann (23) auch in der Form

$$\boldsymbol{\mathcal{Z}}(\psi) = \boldsymbol{\mathcal{Z}}_0(\psi) + \boldsymbol{\mathcal{Z}}_1 \quad \text{mit} \quad \boldsymbol{\mathcal{Z}}_0(\psi) = \frac{r}{\psi}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}} + r\boldsymbol{E} \quad \text{und} \quad \boldsymbol{\mathcal{Z}}_1 = \boldsymbol{\mathcal{R}} + \boldsymbol{\mathcal{S}} - r\boldsymbol{E} \tag{27}$$

geschrieben werden, wobei sowohl $\boldsymbol{\mathcal{Z}}_0(1)$ als auch $\boldsymbol{\mathcal{Z}}_1$ wiederum untere Dreiecksmatrizen sind. Das Mehrtor mit der Impedanz $\boldsymbol{\mathcal{Z}}(\psi)$ kann also durch torweise Serienverbindung zweier s-Tore mit den Impedanzmatrizen $\mathcal{R}\boldsymbol{\mathcal{Z}}_0(\psi)$ und $\mathcal{R}\boldsymbol{\mathcal{Z}}_1$ synthetisiert werden. Hierbei beschreibt $\mathcal{R}\boldsymbol{\mathcal{Z}}_0(\psi)$ einen $(s+1)$-Tor-Zirkulator mit Zirkulationswiderstand $\mathcal{R}r$, welcher an Tor $s+1$ mit einer Kapazität der Impedanz $\mathcal{R}r/\psi$ abgeschlossen ist. Eine ausführliche Beschreibung eines allgemein gültigen Syntheseverfahrens zur Matrix $\mathcal{R}\boldsymbol{\mathcal{Z}}_1$ ist in [16] angegeben, an dieser Stelle sei lediglich wieder ein zweistufiges Verfahren betrachtet, wobei zusätzlich $r_2 = 0$ gelte. Es ergeben sich die in Abbildung 4 gezeigten (wiederum auf $\mathcal{R}$ normierten) Netzwerke sowie die Wellendigitalstruktur in Abbildung 5.

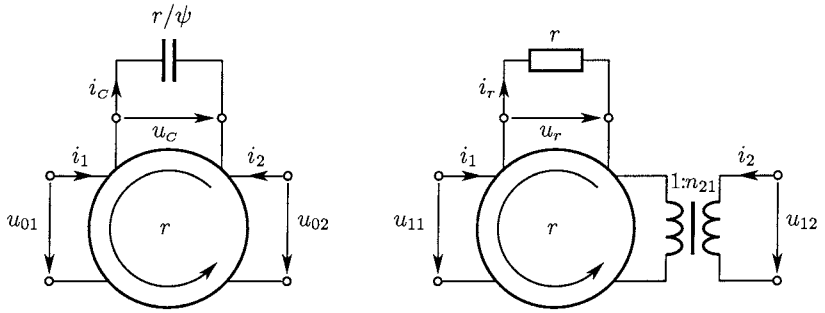


Abbildung 4: Netzwerke zur normierten charakteristischen Impedanzmatrix eines passiven DIRK-Verfahrens mit zwei Stufen.

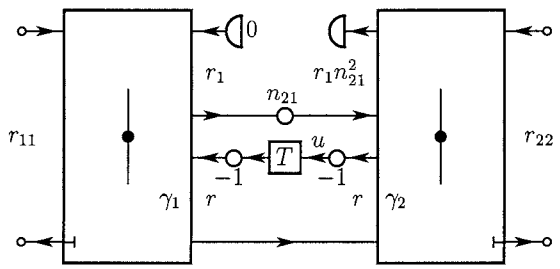


Abbildung 5: Wellendigitalstruktur zur normierten charakteristischen Impedanzmatrix eines passiven DIRK-Verfahrens mit zwei Stufen.

An dieser Stelle ist die Frage zu beantworten, wo in den Strukturen der Abbildungen 4 bzw. 5 die gesuchte Spannung $u(t)$ als Signalgröße auftritt. Durch Auswertung der

Zirkulatorbeziehungen $b_{02} = a_{01}$ sowie $b_{01} = u_C - \mathcal{R}ri_C$ und $u_C + \mathcal{R}ri_C = a_{02}$ findet man die Zusammenhänge

$$i_C = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{i}_0 \quad \text{und} \quad \boldsymbol{u}_0 = \boldsymbol{e}u_C + \mathcal{R}r\boldsymbol{E}\boldsymbol{i}_0\,. \tag{28}$$

Mit $\boldsymbol{i}_0 = \boldsymbol{i}_1 = \boldsymbol{i}$ und $\boldsymbol{u} = \boldsymbol{u}_0 + \boldsymbol{u}_1 = \boldsymbol{u}_0 + \mathcal{R}\boldsymbol{\mathcal{Z}}_1\boldsymbol{i}$ folgen weiter die Beziehungen

$$i_C = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{i} \quad \text{und} \quad \boldsymbol{u} = \boldsymbol{e}u_C + \mathcal{R}\boldsymbol{\mathcal{Q}}\boldsymbol{i}\,, \tag{29}$$

woraus schließlich mit der Differenzengleichung

$$b_C(t) = a_C(t-T) \quad \text{bzw.} \quad u_C(t) - \mathcal{R}ri_C(t) = u_C(t-T) + \mathcal{R}ri_C(t-T) \tag{30}$$

über $a_C(t-T) = a_C(t) - 2\mathcal{R}r\boldsymbol{e}^{\mathrm{T}}\boldsymbol{i}(t)$ bzw. $u_C(t) = a_C(t-T) + \mathcal{R}r\boldsymbol{e}^{\mathrm{T}}\boldsymbol{i}(t)$ die Gleichungen

$$\boldsymbol{u}(t) = \boldsymbol{e}a_C(t-T) + \mathcal{R}\left[r\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}} + \boldsymbol{\mathcal{Q}}\right]\boldsymbol{i}(t) \tag{31}$$

$$a_C(t) = \; a_C(t-T) + 2\mathcal{R}\,r\boldsymbol{e}^{\mathrm{T}}\boldsymbol{i}(t) \tag{32}$$

entstehen, die aber mit der aus (23) folgenden Parameterdarstellung

$$\boldsymbol{D} = \frac{\boldsymbol{\mathcal{Z}}(1)}{2} = \frac{1}{2}\left[r\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}} + \boldsymbol{\mathcal{Q}}\right] = \frac{1}{2}\left[r\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}} + \boldsymbol{\mathcal{R}} + \boldsymbol{\mathcal{S}}\right] \tag{33}$$

gerade den Beziehungen (13) entsprechen, wenn man die Zuweisung $a_C = u$ trifft. Die gesuchte Spannung stimmt also mit der einfallenden Welle an der Kapazität (im Referenznetzwerk) überein (und *nicht* etwa mit der Spannung über dieser Kapazität!). Passive DIRK-Verfahren sind in der hier realisierten Form also insbesondere *selbststartend in Wellengrößen*, der Speicherwert kann unmittelbar mit u_0 vorinitialisiert werden, und eine eigene Startphase ist nicht erforderlich.

Abbildung 6 zeigt das Vorgehen bei einer Wellendigitalsimulation mit den hier erörterten Prinzipien. Die Anzahl reaktiver Elemente im zu simulierenden Netzwerk bestimmt die Zahl der vorhandenen charakteristischen Mehrtore, deren Verknüpfung torweise über je eine statische Wellendigitalstruktur pro Stufe des Verfahrens erfolgt. Diese statischen Wellendigitalstrukturen leiten sich aus dem nichtreaktiven Teil des zu simulierenden Netzwerkes ab, weshalb die zugehörigen Signalflussgraphen von identischem Aufbau sind. Allerdings hängen darin enthaltene Werte wie etwa Adaptorkoeffizienten von den jeweiligen Torwiderständen an den Verbindungen zu den charakteristischen Mehrtoren ab, und die Werte einander zugeordneter Koeffizienten können sich, da an einem charakteristischen Mehrtor die äußeren Torwiderstände nicht zwangsweise gleich sind, zwischen den einzelnen Stufen durchaus voneinander unterscheiden. Ob dies tatsächlich der Fall ist, hängt einzig vom gewählten Verfahren ab, nicht aber von den Werten der ursprünglichen reaktiven Elemente.

Um verfahrensbedingte verzögerungsfreie gerichtete Schleifen zu vermeiden, ist dafür zu sorgen, dass die äußeren Tore der charakteristischen Mehrtore reflexionsfrei sind. Dazu ist der jeweilige normierte Torwiderstand so zu wählen, dass er mit dem entsprechenden Hauptdiagonalelement der Matrix $\boldsymbol{\mathcal{Z}}(1)$ übereinstimmt. In Abbildung 6 ist aber zu erkennen, dass diese Festlegung alleine das Auftreten verzögerungsfreier gerichteter Schleifen nicht vollständig zu unterbinden vermag. Derartige Schleifen entstehen nämlich

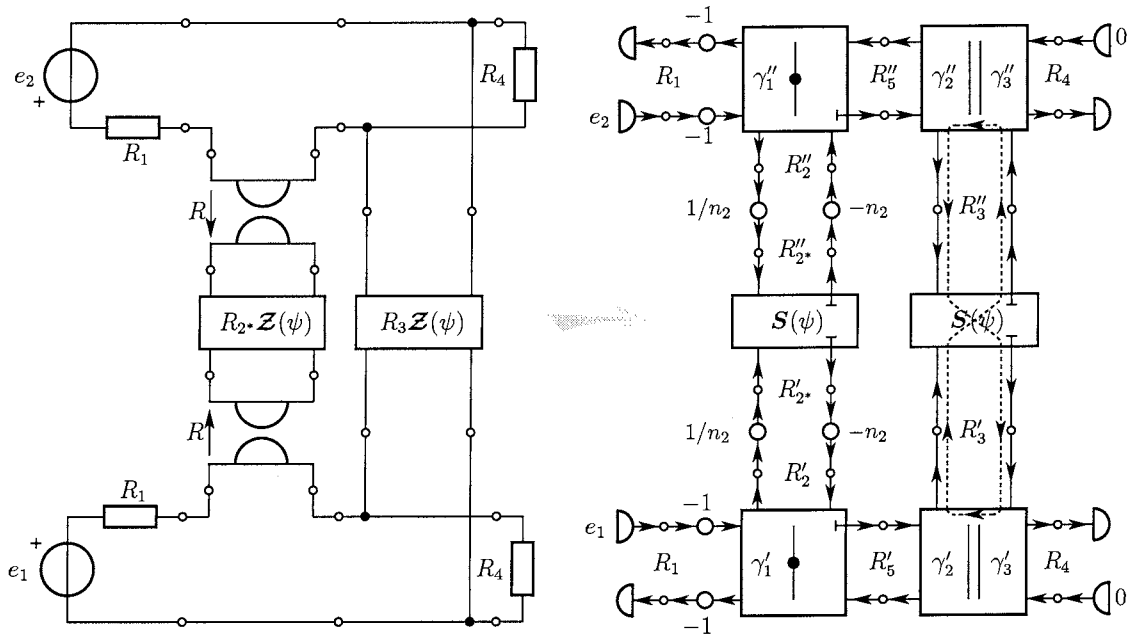


Abbildung 6: Referenznetzwerk zu Abbildung 1, basierend auf einem zweistufigen RUNGE-KUTTA-Verfahren mit der charakteristischen Impedanzmatrix $\boldsymbol{\mathcal{Z}}(\psi)$, und zugehörige Wellendigitalstruktur. Es gilt $e_1(t) = e(t - k_1'T)$ und $e_2(t) = e(t - k_2'T)$. Die gestrichelte Linie verdeutlicht das Entstehen von verzögerungsfreien gerichteten Schleifen bei Verwendung eines Verfahrens, welches nicht diagonal-implizit ist.

auch dann, wenn innerhalb der Wellendigitalrealisierung jedes charakteristischen Zweitors gleichzeitig verzögerungsfreie gerichtete Signalpfade von Tor 1 nach Tor 2 sowie von Tor 2 nach Tor 1 existieren. In der Abbildung sind hierfür maßgebliche Signalpfade durch gestrichelte Linien markiert. Besteht dagegen eine der beiden genannten Verbindungen zwischen den beiden Toren jeweils nicht, so entstehen verfahrensbedingt an keiner Stelle des Signalflussgraphen unerwünschte Schleifen. Dies führt dann auf die bereits oben erwähnte Betrachtung ausschließlich diagonal-impliziter Verfahren.

Bei einer programmtechnischen Realisierung des Wellendigitalkonzeptes mit RUNGE-KUTTA-Verfahren ist die Wellendigitalstruktur zum charakteristischen Mehrtor Bestandteil des Integrationsalgorithmus, während die Wellendigitalstrukturen zu den dynamikfreien Teilen in den einzelnen Stufen durch den Modellierer zur Verfügung zu stellen sind. Unter den getroffenen Voraussetzungen erlaubt die hier diskutierte Art der Umsetzung des charakteristischen Mehrtors ein sequentielles Abarbeiten dieser Teilstrukturen, die ja außerdem noch im Aufbau identisch sind. Damit ist seitens des Modellierers lediglich eine Routine zu erstellen, die die Umsetzung eines dieser Teilgraphen nach Vorgabe der verfahrens- und stufenabhängigen Parameter – dies sind speziell die (durch Elementwerte, normierte Torwiderstände und die Schrittweite bestimmten) Torwiderstände an den Verbindungstoren sowie die (durch den jeweiligen Knotenwert k_σ spezifizierten) Zeitpunkte,

zu denen die Quellsignale auszuwerten sind – vornimmt[3]. Genau diese Aufgabe ist aber auch bei einer Anwendung des auf der Trapezregel basierenden „klassischen" Wellendigitalkonzeptes zu bewältigen, sodass die Anwendung von RUNGE-KUTTA-Verfahren in dieser Hinsicht keine erhöhten Anforderungen an den Modellierer stellt.

Zuletzt sei hier noch angemerkt, dass wie schon bei den LMS-Verfahren auch bei den RUNGE-KUTTA-Verfahren die Passivität eine Vielzahl wünschenswerter numerischer Stabilitäteigenschaften impliziert. Insbesondere ist ein solches Verfahren stets algebraisch stabil [11]. Umgekehrt ist jedes algebraisch stabile Verfahren mit einem Vektor $\boldsymbol{g}^{\mathrm{T}}$ der Form (19) passiv [15].

5 Zweischritt-Runge-Kutta-Verfahren

In der Literatur zur numerischen Mathematik werden als Verallgemeinerung der RUNGE-KUTTA-Verfahren verschiedene Arten von Mehrschritt-RUNGE-KUTTA-Verfahren vorgeschlagen. In diesem Abschnitt werden Verfahren einer Form diskutiert, wie sie von JACKIEWICZ und TRACOGNA verwendet werden [17].

Der Einsatz dieser Verfahren im Rahmen des Wellendigitalkonzeptes erfolgt prinzipiell mit derselben Vorgehensweise wie bei den RUNGE-KUTTA-Verfahren, sobald auf geeignete Weise eine Wellendigitalstruktur zum charakteristischen Mehrtor gefunden ist. Gefordert wird wieder, dass die charakteristische Impedanzmatrix zu dem Verfahren ein passives Mehrtor beschreibt. Dieses Mehrtor ist nun aber im Gegensatz zum vorigen Kapitel ein System mit mehr als einer Zustandsgröße, sodass entsprechend angepasste Syntheseverfahren anzuwenden sind.

In Analogie zu den RUNGE-KUTTA-Verfahren werden auch bei den Zweischritt-RUNGE-KUTTA-Verfahren nach JACKIEWICZ und TRACOGNA Stufenspannungen und -ströme konstruiert, die nach wie vor über die algebraischen Gleichungen (14) miteinander verknüpft sind. Die vorgeschlagene Erweiterung besteht nun darin, in den Gleichungen (13) bei der Konstruktion der Stufenspannungen zusätzlich noch die Stufenströme eines weiteren (vergangenen) Zeitschrittes zu berücksichtigen[4]. Eine entsprechende Verfahrensvorschrift in vektorieller Form lautet

$$\boldsymbol{u}(t) = \boldsymbol{e}u(t-T) + \frac{T}{C}\left[\boldsymbol{D}\boldsymbol{i}(t) + \bar{\boldsymbol{D}}\boldsymbol{i}(t-T)\right] \tag{34a}$$

$$u(t) = \; u(t-T) + \frac{T}{C}\boldsymbol{g}^{\mathrm{T}}\boldsymbol{i}(t) \tag{34b}$$

gegeben. Wird in dieser Vorschrift die Matrix $\bar{\boldsymbol{D}}$ zu Null gesetzt, so erhält man offenbar die Vorschrift (13) eines gewöhnlichen RUNGE-KUTTA-Verfahrens, welches im weiteren Verlauf als *Basisverfahren* des Zweischritt-RUNGE-KUTTA-Verfahrens bezeichnet werden soll.

[3] Der im Falle einer Induktivität zu berücksichtigende Gyrator sollte zweckmäßigerweise als Teil des Integrationsalgorithmus realisiert werden.

[4] JACKIEWICZ und TRACOGNA schlagen zusätzlich auch ein entsprechendes Vorgehen bei der Konstruktion der gesuchten Spannung vor, wobei jeweils außerdem noch der vergangene Wert dieser Spannung mit berücksichtigt wird. Allerdings handelt es sich hierbei wohl um eine Überparametrisierung der Verfahren, mit der im Rahmen des Wellendigitalkonzeptes keine erhöhte Genauigkeit zu erzielen ist.

Auch zu Zweischritt-RUNGE-KUTTA-Verfahren kann wieder eine charakteristische Impedanzmatrix angegeben werden, die man in diesem Fall zu

$$\boldsymbol{\mathcal{Z}}(\psi) = \frac{1}{\psi}\boldsymbol{e}\boldsymbol{g}^{\mathrm{T}} + 2\boldsymbol{D} - \boldsymbol{e}\boldsymbol{g}^{\mathrm{T}} + \boldsymbol{\mathcal{Z}}_1(\psi) \quad \text{mit} \quad \boldsymbol{\mathcal{Z}}_1(\psi) = \frac{1-\psi}{1+\psi} 2\bar{\boldsymbol{D}}[\boldsymbol{1} - \boldsymbol{e}\boldsymbol{g}^{\mathrm{T}}] \tag{35}$$

berechnet. Die Impedanzmatrix $\boldsymbol{\mathcal{Z}}(\psi)$ eines im Rahmen des Wellendigitalkonzeptes anwendbaren Verfahrens muss wiederum für $\psi = 1$ eine untere Dreiecksform besitzen, sodass an dieser Stelle mit $\boldsymbol{\mathcal{Z}}(1) = 2\boldsymbol{D}$ ausschließlich diagonal-implizite Basisverfahren in Frage kommen.

Die Forderung nach einem passiven Zweischritt-RUNGE-KUTTA-Verfahren und damit nach einer positiven Impedanzmatrix (35) liefert zunächst wie bei den RUNGE-KUTTA-Verfahren die Forderung $\boldsymbol{g}^{\mathrm{T}} = r\boldsymbol{e}^{\mathrm{T}}$ mit $r > 0$. Außerdem muss in Erweiterung von (23) die Matrix

$$\boldsymbol{\mathcal{Z}}(\psi) - \frac{r}{\psi}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}} = 2\boldsymbol{D} - r\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}} + \frac{1-\psi}{1+\psi} 2\bar{\boldsymbol{D}}[\boldsymbol{1} - \boldsymbol{e}\boldsymbol{g}^{\mathrm{T}}] \tag{36}$$

ihrerseits positiv sein. Die Betrachtung des Sonderfalles $\psi = 1$ zeigt demnach, dass das Basisverfahren notwendigerweise ebenfalls passiv sein muss. Wie zuvor $\boldsymbol{\mathcal{Q}}$ in $\boldsymbol{\mathcal{R}}$ und $\boldsymbol{\mathcal{S}}$ zerlegt man nun noch die Matrix

$$\bar{\boldsymbol{\mathcal{Q}}} := 2\bar{\boldsymbol{D}}[\boldsymbol{1} - \boldsymbol{e}\boldsymbol{g}^{\mathrm{T}}] \tag{37}$$

in ihren symmetrischen Teil $\bar{\boldsymbol{\mathcal{R}}}$ und ihren schiefsymmetrischen Teil $\bar{\boldsymbol{\mathcal{S}}}$ gemäß

$$\bar{\boldsymbol{\mathcal{R}}} := \frac{1}{2}[\bar{\boldsymbol{\mathcal{Q}}} + \bar{\boldsymbol{\mathcal{Q}}}^{\mathrm{T}}] \quad \text{bzw.} \quad \bar{\boldsymbol{\mathcal{S}}} := \frac{1}{2}[\bar{\boldsymbol{\mathcal{Q}}} - \bar{\boldsymbol{\mathcal{Q}}}^{\mathrm{T}}] \tag{38}$$

und berücksichtigt den für reelle Frequenzen φ gültigen Zusammenhang

$$\left|\frac{1-\mathrm{j}\varphi}{1+\mathrm{j}\varphi}\right| = 1 \quad \text{bzw.} \quad \frac{1-\mathrm{j}\varphi}{1+\mathrm{j}\varphi} = \sigma_{\mathrm{r}} + \mathrm{j}\sigma_{\mathrm{j}} \quad \text{mit} \quad \sigma_{\mathrm{r}}^2 + \sigma_{\mathrm{j}}^2 = 1\,, \tag{39}$$

wobei σ_{r} und σ_{j} reele Größen sind. Damit liefert die Forderung nach Positivität von (36), ausgewertet für $\psi = \mathrm{j}\varphi$, die Bedingungen

$$\boldsymbol{\mathcal{R}} + \sigma_{\mathrm{r}}\bar{\boldsymbol{\mathcal{R}}} + \mathrm{j}\sigma_{\mathrm{j}}\bar{\boldsymbol{\mathcal{S}}} \geq \boldsymbol{0} \quad \text{für alle } \sigma_{\mathrm{r}}, \sigma_{\mathrm{j}} \in \mathbb{R} \text{ mit } \sigma_{\mathrm{r}}^2 + \sigma_{\mathrm{j}}^2 = 1\,, \tag{40}$$

die in Verbindung mit der speziellen Gestalt des Vektors $\boldsymbol{g}^{\mathrm{T}}$ für Passivität notwendig und hinreichend sind. Insbesondere enthalten sind die Bedingungen

$$\boldsymbol{\mathcal{R}} + \bar{\boldsymbol{\mathcal{R}}} \geq \boldsymbol{0} \quad \text{und} \quad \boldsymbol{\mathcal{R}} - \bar{\boldsymbol{\mathcal{R}}} \geq \boldsymbol{0}\,. \tag{41}$$

Im Gegensatz zur Matrix $\boldsymbol{\mathcal{S}}$ in der Parameterdarstellung des Basisverfahrens hat demnach die ebenfalls schiefsymmetrische Matrix $\bar{\boldsymbol{\mathcal{S}}}$ einen Einfluss auf die Passivität des Zweischritt-RUNGE-KUTTA-Verfahrens.

Anders als bei den RUNGE-KUTTA-Verfahren ist für Zweischritt-RUNGE-KUTTA-Verfahren ein allgemeiner Syntheseansatz zur Gewinnung eines Referenznetzwerkes und damit zur Parametrisierung von Verfahren, die der Randbedingung (40) genügen, nicht ohne

Weiteres erkennbar. Lediglich im Sonderfall $\bar{\boldsymbol{S}} = \boldsymbol{0}$ kann diese Aufgabe recht einfach gelöst werden. Die Impedanzmatrix kann dann in der Form

$$\boldsymbol{\mathcal{Z}}(\psi) = \frac{r}{\psi}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}} + \boldsymbol{\mathcal{S}} + \frac{1}{1+\psi}\left[\boldsymbol{\mathcal{R}} + \bar{\boldsymbol{\mathcal{R}}}\right] + \frac{\psi}{1+\psi}\left[\boldsymbol{\mathcal{R}} - \bar{\boldsymbol{\mathcal{R}}}\right] \tag{42}$$

geschrieben werden, wobei nach den obigen Betrachtungen die Matrizen $\boldsymbol{\mathcal{R}} + \bar{\boldsymbol{\mathcal{R}}}$ und $\boldsymbol{\mathcal{R}} - \bar{\boldsymbol{\mathcal{R}}}$ positiv semidefinit sind. Damit ist jeweils eine (25) entsprechende Zerlegung dieser Matrizen möglich. Anschließend sind noch die r_σ durch $r_\sigma/[1+\psi]$ bzw. durch $[r_\sigma\psi]/[1+\psi]$ zu ersetzen, was in den zugehörigen Netzwerken jeweils einem Austausch jedes Widerstandes gegen eine passende Parallelschaltung von Widerstand und Kapazität bzw. von Widerstand und Induktivität entspricht.

6 Entwurf passiver Verfahren

Um die numerische Genauigkeit, die beim Einsatz eines numerischen Integrationsverfahrens erzielt wird, abschätzen zu können, wird üblicherweise die so genannte *Konsistenzordnung* q eines Verfahrens betrachtet. Diese ist ein asymptotisches Maß für den resultierenden systematischen Diskretisierungsfehler, wobei eine hohe Konsistenzordnung einer hohen Approximationsgüte entspricht. Dies gilt allerdings nur dann, wenn das Verfahren gewisse numerische Stabilitätseigenschaften aufweist, die aber bei passiven Verfahren ausnahmslos gegeben sind.

Die Forderung nach Passivität beschränkt die mit den verschiedenen Klassen von numerischen Integrationsverfahren jeweils maximal erreichbare Konsistenzordnung, so beträgt diese etwa bei LMS-Verfahren Zwei[5]. Im weiteren Verlauf wird sich zeigen, dass dagegen, genau wie bei passiven diagonal-impliziten RUNGE-KUTTA-Verfahren, die maximal erreichbare Konsistenzordnung bei den hier betrachteten passiven Zweischritt-RUNGE-KUTTA-Verfahren (mit diagonal-impliziten Basisverfahren) Vier beträgt.

6.1 Konsistenzbedingungen

JACKIEWICZ und TRACOGNA haben zu ihren Zweischritt-RUNGE-KUTTA-Verfahren auch die entsprechenden Konsistenzbedingungen ermittelt. Unter der Voraussetzung, dass das Basisverfahren die *Knotenbedingung* nach BUTCHER und das Zweischritt-RUNGE-KUTTA-Verfahren außerdem die so genannte *Präkonsistenzbedingung* erfüllt, dass also

$$\boldsymbol{k} = \boldsymbol{D}\boldsymbol{e} \quad \text{und} \quad \boldsymbol{0} = \bar{\boldsymbol{D}}\boldsymbol{e} \overset{(37)}{\Longrightarrow} \bar{\boldsymbol{\mathcal{Q}}} = 2\bar{\boldsymbol{D}} \Longrightarrow \bar{\boldsymbol{\mathcal{Q}}}\boldsymbol{e} = \boldsymbol{0} \tag{43}$$

gilt, ergeben sich für Verfahren bis zur Ordnung Fünf die in Tabelle 1 angegebenen Konsistenzbedingungen. Hierbei wurden die Abkürzungen

$$\Gamma_\mu = \frac{1}{\mu!} - \frac{\boldsymbol{g}^{\mathrm{T}}\boldsymbol{k}^{\mu-1}}{[\mu-1]!} \quad \text{und} \quad \Delta\boldsymbol{\Gamma}_\mu = \boldsymbol{\Gamma}_\mu - \bar{\boldsymbol{\Gamma}}_\mu \tag{44a}$$

[5] Hierbei handelt es sich um die so genannte zweite DAHLQUIST-Schranke [18], denn bei LMS-Verfahren ist Passivität äquivalent zur A-Stabilität [3].

mit

$$\boldsymbol{\Gamma}_\mu = \frac{\boldsymbol{k}^\mu}{\mu!} - \frac{\boldsymbol{D}\boldsymbol{k}^{\mu-1}}{[\mu-1]!} \quad \text{und} \quad \bar{\boldsymbol{\Gamma}}_\mu = \frac{\bar{\boldsymbol{D}}[\boldsymbol{k}-\boldsymbol{e}]^{\mu-1}}{[\mu-1]!} \tag{44b}$$

verwendet. Hierbei ist das Produkt bzw. die Potenzierung von Spaltenvektoren elementweise zu verstehen, und die Bedingungen $\Gamma_\mu = 0$ lassen sich auch in der Form

$$\boldsymbol{g}^{\mathrm{T}}\boldsymbol{k}^{\mu-1} = \frac{1}{\mu} \tag{45}$$

schreiben. Die Bedingungen für ein entsprechendes RUNGE-KUTTA-Verfahren erhält man letztlich einfach dadurch, dass hierin durchweg $\bar{\boldsymbol{D}} = \boldsymbol{0}$ gesetzt wird[6].

	Allgemein	Vereinfacht
$q \geq 1:$	$\Gamma_1 = 0$	$\Gamma_1 = 0$
$q \geq 2:$	$\Gamma_2 = 0$	$\Gamma_2 = 0$
$q \geq 3:$	$\Gamma_3 = 0$	$\Gamma_3 = 0$
	$\boldsymbol{g}^{\mathrm{T}}\Delta\boldsymbol{\Gamma}_2 = 0$	$\boldsymbol{g}^{\mathrm{T}}\boldsymbol{\Gamma}_2 = 0$
$q \geq 4:$	$\Gamma_4 = 0$	$\Gamma_4 = 0$
	$\boldsymbol{g}^{\mathrm{T}}\Delta\boldsymbol{\Gamma}_3 = 0$	$\boldsymbol{g}^{\mathrm{T}}\boldsymbol{\Gamma}_3 = 0$
	$\boldsymbol{g}^{\mathrm{T}}[\boldsymbol{k}\Delta\boldsymbol{\Gamma}_2] = 0$	$\boldsymbol{g}^{\mathrm{T}}[\boldsymbol{k}\Delta\boldsymbol{\Gamma}_2] = 0$
	$\boldsymbol{g}^{\mathrm{T}}[\boldsymbol{D}+\bar{\boldsymbol{D}}]\Delta\boldsymbol{\Gamma}_2 = 0$	$\boldsymbol{g}^{\mathrm{T}}\boldsymbol{D}\Delta\boldsymbol{\Gamma}_2 = 0$
$q \geq 5:$	$\Gamma_5 = 0$	$\Gamma_5 = 0$
	$\boldsymbol{g}^{\mathrm{T}}\Delta\boldsymbol{\Gamma}_4 = 0$	$\boldsymbol{g}^{\mathrm{T}}\boldsymbol{\Gamma}_4 = 0$
	$\boldsymbol{g}^{\mathrm{T}}[\boldsymbol{k}\Delta\boldsymbol{\Gamma}_3] = 0$	$\boldsymbol{g}^{\mathrm{T}}[\boldsymbol{k}\Delta\boldsymbol{\Gamma}_3] = 0$
	$\boldsymbol{g}^{\mathrm{T}}[[\boldsymbol{D}+\bar{\boldsymbol{D}}]\Delta\boldsymbol{\Gamma}_3 - \bar{\boldsymbol{D}}\Delta\boldsymbol{\Gamma}_2] = 0$	$\boldsymbol{g}^{\mathrm{T}}\boldsymbol{D}\Delta\boldsymbol{\Gamma}_3 = 0$
	$\boldsymbol{g}^{\mathrm{T}}\Delta\boldsymbol{\Gamma}_2^2 = 0$	$\Delta\boldsymbol{\Gamma}_2 = \boldsymbol{0}$
	$\boldsymbol{g}^{\mathrm{T}}[[\boldsymbol{D}+\bar{\boldsymbol{D}}][\boldsymbol{k}\Delta\boldsymbol{\Gamma}_2] - \bar{\boldsymbol{D}}\Delta\boldsymbol{\Gamma}_2] = 0$	
	$\boldsymbol{g}^{\mathrm{T}}[\boldsymbol{D}+\bar{\boldsymbol{D}}]^2\Delta\boldsymbol{\Gamma}_2 = 0$	
	$\boldsymbol{g}^{\mathrm{T}}[\boldsymbol{k}[[\boldsymbol{D}+\bar{\boldsymbol{D}}]\Delta\boldsymbol{\Gamma}_2]] = 0$	
	$\boldsymbol{g}^{\mathrm{T}}[\boldsymbol{k}^2\Delta\boldsymbol{\Gamma}_2] = 0$	

Tabelle 1: Konsistenzbedingungen für Zweischritt-RUNGE-KUTTA-Verfahren bis zur Ordnung Fünf.

6.2 Verfahren bis zur Ordnung 3

Offenbar besitzen das Zweischritt-RUNGE-KUTTA-Verfahren und sein zugehöriges Basisverfahren bis zur Ordnung Zwei stets dieselbe Konsistenzordnung. Dabei gilt für konsis-

[6] Die so resultierenden Gleichungen stimmen nicht einzeln mit den von BUTCHER ermittelten Konsistenzbedingungen für ein RUNGE-KUTTA-Verfahren überein, sind aber in ihrer Gesamtheit hierzu äquivalent.

tente Verfahren ($q \geq 1$)

$$\boldsymbol{g}^{\mathrm{T}}\boldsymbol{e} = 1\,, \tag{46}$$

und für eine Konsistenzordnung $q \geq 2$ findet man mit der Symmetrie von $\boldsymbol{\mathcal{R}}$ und der Schiefsymmetrie von $\boldsymbol{\mathcal{S}}$ unter Verwendung der Parameterdarstellung (33) die Anforderung

$$r\boldsymbol{e}^{\mathrm{T}}\frac{1}{2}\left[r\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}} + \boldsymbol{\mathcal{R}} + \boldsymbol{\mathcal{S}}\right]\boldsymbol{e} = \frac{1}{2} \quad \Longrightarrow \quad \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\mathcal{R}}\boldsymbol{e} = 0 \quad \Longleftrightarrow \quad \boldsymbol{\mathcal{R}}\boldsymbol{e} = \boldsymbol{0}\,. \tag{47}$$

Die Äquivalenz der beiden letztgenannten Aussagen folgt dabei aus der Tatsache, dass die (symmetrische) Matrix $\boldsymbol{\mathcal{R}}$ infolge der Passivität ja positiv semidefinit sein muss. Damit kann aber nun aus der Darstellung

$$\bar{\boldsymbol{\mathcal{Q}}} = \bar{\boldsymbol{\mathcal{R}}} + \bar{\boldsymbol{\mathcal{S}}} \tag{48}$$

gefolgert werden, dass dann auch die Beziehungen

$$\bar{\boldsymbol{\mathcal{R}}}\boldsymbol{e} = \boldsymbol{0} \quad \Longrightarrow \quad \bar{\boldsymbol{\mathcal{S}}}\boldsymbol{e} = \boldsymbol{0} \tag{49a}$$

sowie schließlich

$$\boldsymbol{e}^{\mathrm{T}}\bar{\boldsymbol{\mathcal{Q}}} = \boldsymbol{0}^{\mathrm{T}} \quad \Longleftrightarrow \quad \boldsymbol{e}^{\mathrm{T}}\bar{\boldsymbol{D}} = \boldsymbol{0}^{\mathrm{T}} \tag{49b}$$

erfüllt sein müssen. Dabei ist zu beachten, dass aus (43) zwar wiederum $\boldsymbol{e}^{\mathrm{T}}\bar{\boldsymbol{\mathcal{R}}}\boldsymbol{e} = 0$ folgt, die Matrix $\bar{\boldsymbol{\mathcal{R}}}$ aber *nicht* notwendigerweise positiv semidefinit ist. Allerdings ist dies für die Matrix $\boldsymbol{\mathcal{R}} + \bar{\boldsymbol{\mathcal{R}}}$ der Fall, womit aus $\boldsymbol{e}^{\mathrm{T}}[\boldsymbol{\mathcal{R}} + \bar{\boldsymbol{\mathcal{R}}}]\boldsymbol{e} = 0$ auf $[\boldsymbol{\mathcal{R}} + \bar{\boldsymbol{\mathcal{R}}}]\boldsymbol{e} = \boldsymbol{0}$ und dann mit $\boldsymbol{\mathcal{R}}\boldsymbol{e} = \boldsymbol{0}$ doch auf (49) geschlossen werden kann. Weiterhin ist festzuhalten, dass mit (49) für ein passives Verfahren mit $q \geq 2$ stets

$$\boldsymbol{g}^{\mathrm{T}}\bar{\boldsymbol{\Gamma}}_{\mu} = \boldsymbol{0} \quad \text{bzw.} \quad \boldsymbol{g}^{\mathrm{T}}\Delta\boldsymbol{\Gamma}_{\mu} = \boldsymbol{g}^{\mathrm{T}}\boldsymbol{\Gamma}_{\mu} \tag{50}$$

und wegen (47)

$$\boldsymbol{k} = \boldsymbol{D}\boldsymbol{e} = \frac{1}{2}\left[\boldsymbol{e} + \boldsymbol{\mathcal{S}}\boldsymbol{e}\right] \quad \text{und} \quad \boldsymbol{e}^{\mathrm{T}}\boldsymbol{D} = \frac{1}{2}\left[\boldsymbol{e}^{\mathrm{T}} + \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\mathcal{S}}\right] \tag{51a}$$

und daher

$$\boldsymbol{e}^{\mathrm{T}}\boldsymbol{D} = \boldsymbol{e}^{\mathrm{T}} - \boldsymbol{k}^{\mathrm{T}} \tag{51b}$$

gilt, wodurch unmittelbar die Implikation

$$\Gamma_{\mu} = 0 \quad \wedge \quad \Gamma_{\mu-1} = 0 \quad \Longrightarrow \quad \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\Gamma}_{\mu-1} = 0 \tag{52}$$

zu zeigen ist.

Als Konsequenz der Beziehung (50) besitzt ein Zweischritt-RUNGE-KUTTA-Verfahren genau dann mindestens die Konsistenzordnung Drei, wenn dies auch für das Basisverfahren gilt. Die Konsistenzbedingung $\Gamma_3 = 0$ liefert dann unter Berücksichtigung der Parameterdarstellung (33) in Verbindung mit den bisher gefundenen Ergebnissen die Bestimmungsgleichung

$$\boldsymbol{g}^{\mathrm{T}}[\boldsymbol{\mathcal{S}}\boldsymbol{e}]^2 = \frac{1}{3}\,. \tag{53}$$

Werden die Parameter entsprechend gewählt, so ist damit in Verbindung mit $\Gamma_2 = 0$ wegen (52) automatisch auch die Bedingung $\boldsymbol{g}^{\mathrm{T}}\boldsymbol{\Gamma}_2 = 0$ erfüllt[7].

[7]Tatsächlich kann mit Hilfe der Parameterdarstellung (33) gezeigt werden, dass die allgemeinen Konsistenzbedingungen im Falle eines passiven RUNGE-KUTTA-Verfahren in hohem Maße redundant sind [15]. Diese Tatsache lässt sich zur Vereinfachung des Entwurfes nutzen.

6.3 Ein Zweischritt-Runge-Kutta-Verfahren der Ordnung 4

Sicher wird der Einsatz eines Zweischritt-RUNGE-KUTTA-Verfahrens anstelle eines RUNGE-KUTTA-Verfahrens im Rahmen des Wellendigitalkonzeptes nur dann sinnvoll sein, wenn bei ansonsten unveränderten Anforderungen eine höhere Genauigkeit erreicht werden kann. Nun kann aber gezeigt werden, dass sich zu einem Basisverfahren der Ordnung Drei, welches die zusätzliche Bedingung

$$\Gamma_4 = 0 \tag{54}$$

erfüllt, stets ein Zweischritt-RUNGE-KUTTA-Verfahren der Ordnung Vier konstruieren lässt. Dazu ist zunächst festzuhalten, dass in Verbindung mit $\Gamma_3 = 0$ wegen (52) dann auch die zweite Bedingung $\boldsymbol{g}^{\mathrm{T}}\boldsymbol{\Gamma}_3 = 0$ wieder automatisch erfüllt ist. Mit

$$\Delta\boldsymbol{\Gamma}_2 = \frac{\boldsymbol{k}^2}{2} - \boldsymbol{D}\boldsymbol{k} - \bar{\boldsymbol{D}}[\boldsymbol{k} - \boldsymbol{e}] \overset{(43)}{=} \frac{\boldsymbol{k}^2}{2} - [\boldsymbol{D} + \bar{\boldsymbol{D}}]\boldsymbol{k} \tag{55}$$

folgt aus der Anforderung $0 = \boldsymbol{g}^{\mathrm{T}}[\boldsymbol{k}\Delta\boldsymbol{\Gamma}_2] = r\boldsymbol{k}^{\mathrm{T}}\Delta\boldsymbol{\Gamma}_2$ unter Berücksichtigung der Parameterdarstellungen (33) und (48) in Verbindung mit $\bar{\boldsymbol{\mathcal{Q}}} = 2\bar{\boldsymbol{D}}$ (siehe (43))

$$0 = \frac{1}{2}\underbrace{\left[r\boldsymbol{e}^{\mathrm{T}}\boldsymbol{k}^3 - \left[r\boldsymbol{k}^{\mathrm{T}}\boldsymbol{e}\right]^2\right]}_{=0} - r\boldsymbol{k}^{\mathrm{T}}[\boldsymbol{\mathcal{R}} + \bar{\boldsymbol{\mathcal{R}}}]\boldsymbol{k} \quad \Longrightarrow \quad \boldsymbol{k}^{\mathrm{T}}[\boldsymbol{\mathcal{R}} + \bar{\boldsymbol{\mathcal{R}}}]\boldsymbol{k} = 0\,. \tag{56}$$

Mit einer Argumentation wie oben (die Matrix $\boldsymbol{\mathcal{R}} + \bar{\boldsymbol{\mathcal{R}}}$ ist positiv semidefinit) resultiert

$$[\boldsymbol{\mathcal{R}} + \bar{\boldsymbol{\mathcal{R}}}]\boldsymbol{k} = \boldsymbol{0}\,. \tag{57}$$

Werden die Parameter $\bar{\boldsymbol{\mathcal{R}}}$ entsprechend gewählt, so ist wegen (51b) dann auch

$$\boldsymbol{g}^{\mathrm{T}}\boldsymbol{D}\Delta\boldsymbol{\Gamma}_2 = \boldsymbol{g}^{\mathrm{T}}[\boldsymbol{k}\Delta\boldsymbol{\Gamma}_2] + \boldsymbol{g}^{\mathrm{T}}\boldsymbol{D}\Delta\boldsymbol{\Gamma}_2 = \boldsymbol{g}^{\mathrm{T}}\Delta\boldsymbol{\Gamma}_2 = \boldsymbol{g}^{\mathrm{T}}\boldsymbol{\Gamma}_2 = 0\,, \tag{58}$$

und das Zweischritt-RUNGE-KUTTA-Verfahren besitzt die gewünschte Konsistenzordnung Vier.

Mit dem gefundenen Ergebnis kann jetzt ein „sinvolles“ passives Zweischritt-RUNGE-KUTTA-Verfahren angegeben werden, dessen Konsistenzordnung die seines Basisverfahrens übertrifft. Als Basisverfahren wird das zweistufige passive DIRK-Verfahren mit der Konsistenzordnung Drei gewählt[8], welches die Parameter

$$\boldsymbol{D} = \begin{bmatrix} \frac{1}{2} + \frac{1}{2\sqrt{3}} & 0 \\ -\frac{1}{\sqrt{3}} & \frac{1}{2} + \frac{1}{2\sqrt{3}} \end{bmatrix}, \quad \boldsymbol{k} = \begin{bmatrix} \frac{1}{2} + \frac{1}{2\sqrt{3}} \\ \frac{1}{2} - \frac{1}{2\sqrt{3}} \end{bmatrix} \quad \text{und} \quad \boldsymbol{g}^{\mathrm{T}} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix} \tag{59a}$$

besitzt. Dieses Verfahren erfüllt die Anforderung (54). Mit der Wahl

$$\bar{\boldsymbol{\mathcal{Q}}} = \bar{\boldsymbol{\mathcal{R}}} = -\boldsymbol{\mathcal{R}} \quad \Longrightarrow \quad \bar{\boldsymbol{D}} = -\frac{\boldsymbol{\mathcal{R}}}{2} = \begin{bmatrix} -\frac{1}{4} - \frac{1}{2\sqrt{3}} & \frac{1}{4} + \frac{1}{2\sqrt{3}} \\ \frac{1}{4} + \frac{1}{2\sqrt{3}} & -\frac{1}{4} - \frac{1}{2\sqrt{3}} \end{bmatrix} \tag{59b}$$

[8] Die Forderung nach einem zweistufigen passiven DIRK-Verfahren mit der Konsistenzordnung Drei spezifiziert die Koeffizienten des Verfahrens eindeutig [8].

erhält man dann ein Zweischritt-RUNGE-KUTTA-Verfahren, das nicht nur (57) erfüllt und somit die Konsistenzordnung Vier besitzt, sondern auch wegen positiv semidefiniter Matrizen $\mathcal{R} - \bar{\mathcal{R}} = 2\mathcal{R}$ und $\mathcal{R} + \bar{\mathcal{R}} = \mathbf{0}$ sicher passiv ist.

Die Umsetzung des vorgeschlagenen Verfahrens in das Wellendigitalkonzept kann wegen $\bar{\mathcal{S}} = \mathbf{0}$ mit den in Abschnitt 5 angedeuteten Methoden erfolgen. Das zugehörige Netzwerk sowie die daraus resultierende Wellendigitalstruktur zur charakteristischen Impedanzmatrix sind in den Abbildungen 7 und 8 dargestellt.

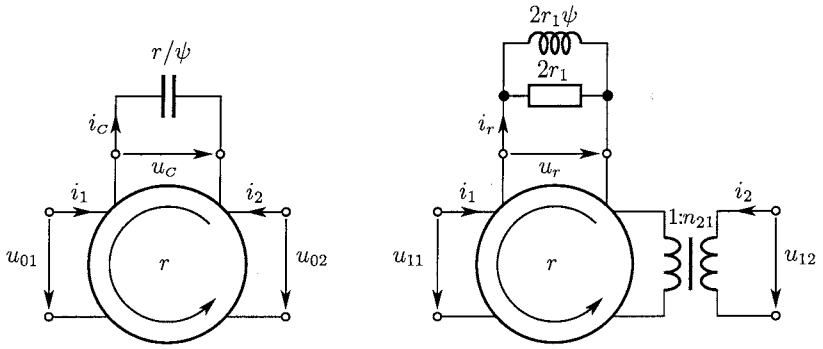


Abbildung 7: Netzwerk zur normierten charakteristischen Impedanzmatrix des passiven Zweischritt-RUNGE-KUTTA-Verfahrens mit den Parametern (59).

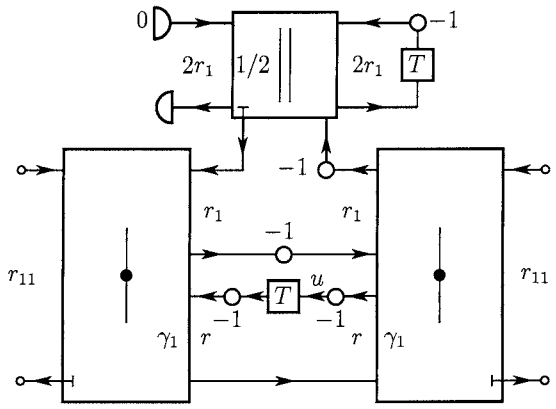


Abbildung 8: Wellendigitalstruktur zur normierten charakteristischen Impedanzmatrix des passiven Zweischritt-RUNGE-KUTTA-Verfahrens mit den Parametern (59).

Im Gegensatz zu den RUNGE-KUTTA-Verfahren, die ja selbststartend in Wellengrößen sind, wird es bei den Zweischritt-RUNGE-KUTTA-Verfahren aufgrund der erhöhten Zahl an Zustandsgrößen – wie schon bei den meisten LMS-Verfahren – erforderlich, die zugeordneten Verzögerungen in geeigneter Weise zu initialisieren. Mit einer ähnlichen Argumentation wie in Abschnitt 4 – dabei ist etwa der Zusammenhang $\hat{u}_{\mathrm{p}\sigma} = 2\mathcal{R}r_\sigma\psi/[1+\psi]\hat{i}_{\mathrm{p}\sigma}$, der eine hierbei auftretende Parallelschaltung von Induktivität und Widerstand beschreibt, in der Form $u_{\mathrm{p}\sigma}(t) = \mathcal{R}r_1\left[i_{\mathrm{p}\sigma}(t) - i_{\mathrm{p}\sigma}(t-T)\right]$ zu berücksichtigen – lässt sich zunächst zeigen, dass die gesuchte numerische Lösung $u(t)$ wiederum als Wellengröße an der Kapazität in

Abbildung 7 auftritt. Hiermit ist der Startwert der zugehörigen Verzögerung in Abbildung 8 naturgemäß zu u_0 zu wählen. Andererseits stimmt der Wert von $\boldsymbol{\mathcal{Z}}_1(1)$ in (35) mit der Matrix $\boldsymbol{\mathcal{Z}}_1$ aus (27) überein. Werden also zu Beginn der Simulation die übrigen Speicher zu Null initialisiert, wird der erste Simulationsschritt gerade mit dem (selbststartenden) Basisverfahren durchgeführt. Hierbei ist zu beachten, dass die Betrachtung bei $\psi = 1$ (dies entspricht dem Grenzfall $\mathrm{e}^{pT} \to \infty$) als Auftrennen an den Verzögerungen verstanden werden kann, welches in seiner Wirkung der hier vorgeschlagenen Initialisierung mit Null entspricht. Ein Vergleich der Abbildung 8 mit Abbildung 5 (hier gilt $n_{21} = -1$) verdeutlicht den Zusammenhang zwischen dem Zweischritt-RUNGE-KUTTA-Verfahren der Ordnung Vier sowie der „Startphase", die dann ohne die Notwendigkeit einer gesonderten Implementierung mit dem Basisverfahren der Ordnung Drei ausgeführt wird.

6.4 Maximal erreichbare Konsistenzordnung

Weder mit einem passiven DIRK-Verfahren noch mit einem passiven Zweischritt-RUNGE-KUTTA-Verfahren, dessen Basisverfahren diagonal-implizit ist, kann eine Konsistenzordung $q > 4$ erreicht werden. Verantwortlich hierfür ist die Tatsache, dass mit der Bedingung $\boldsymbol{g}^{\mathrm{T}} \Delta\boldsymbol{\Gamma}_2^2 = 0$ wegen durchweg positiver Einträge im Vektor $\boldsymbol{g}^{\mathrm{T}}$

$$\mathbf{0} = \Delta\boldsymbol{\Gamma}_2 \overset{(55)}{=} \frac{\boldsymbol{k}^2}{2} - [\boldsymbol{D} + \bar{\boldsymbol{D}}]\boldsymbol{k} \tag{60}$$

gelten muss. Diese Bedingung kommt der Forderung nach einer *Stufenkonsistenzordnung* von mindestens Zwei gleich und führt zusammen mit den übrigen bisherigen Ergebnissen zu den vereinfachten Konsistenzbedingungen in der rechten Spalte von Tabelle 1. Im Fall eines diagonal-impliziten Verfahrens entnimmt man der Bedingung, die dann

$$\boldsymbol{k}^2 - 2\boldsymbol{D}\boldsymbol{k} = \mathbf{0} \tag{61}$$

lautet, unmittelbar $d_{11}^2 = d_{11}^2/2$ und damit $d_{11} = 0$, was wegen $d_{11} = [r + r_1]/2$ unmittelbar auf einen Widerspruch zur geforderten Passivität führt[9].

Nimmt man für ein Zweischritt-RUNGE-KUTTA-Verfahren eine Konsistenzordnung $q \geq 5$ an, so erhält man aus den Parameterdarstellungen (33) und (48) in Verbindung mit (57) mit einer Vorgehensweise wie bei der Herleitung von (51)

$$\boldsymbol{k}^{\mathrm{T}}[\boldsymbol{D} + \bar{\boldsymbol{D}}] = [\boldsymbol{k}^{\mathrm{T}}\boldsymbol{e}]r\boldsymbol{e}^{\mathrm{T}} - \boldsymbol{k}^{\mathrm{T}}[\boldsymbol{D} + \bar{\boldsymbol{D}}]^{\mathrm{T}} \overset{(60)}{=} [\boldsymbol{k}^{\mathrm{T}}\boldsymbol{e}]r\boldsymbol{e}^{\mathrm{T}} - \frac{\boldsymbol{k}^{2\mathrm{T}}}{2}\,. \tag{62}$$

Mit

$$\Delta\boldsymbol{\Gamma}_3 = \frac{\boldsymbol{k}^3}{6} - \frac{\boldsymbol{D}\boldsymbol{k}^2}{2} - \frac{\bar{\boldsymbol{D}}[\boldsymbol{k} - \boldsymbol{e}]^2}{2} = \frac{\boldsymbol{k}^3}{6} - \frac{[\boldsymbol{D} + \bar{\boldsymbol{D}}]\boldsymbol{k}^2}{2} + \bar{\boldsymbol{D}}\boldsymbol{k} \tag{63}$$

erhält man also aus der Forderung $\boldsymbol{g}^{\mathrm{T}}[\boldsymbol{k}\Delta\boldsymbol{\Gamma}_3] = 0 \iff \boldsymbol{k}^{\mathrm{T}}\Delta\boldsymbol{\Gamma}_3 = 0$ in Verbindung mit $\Gamma_5 = 0 \iff r\boldsymbol{e}^{\mathrm{T}}\boldsymbol{k}^4 = 1/5$ die Bedingung

$$0 = \underbrace{\frac{\boldsymbol{k}^{\mathrm{T}}\boldsymbol{k}^3}{6} - \frac{[\boldsymbol{k}^{\mathrm{T}}\boldsymbol{e}][r\boldsymbol{e}^{\mathrm{T}}\boldsymbol{k}^2] - \dfrac{\boldsymbol{k}^{2\mathrm{T}}\boldsymbol{k}^2}{2}}{2}}_{=0} + \boldsymbol{k}^{\mathrm{T}}\bar{\boldsymbol{D}}\boldsymbol{k} = \frac{1}{2}\boldsymbol{k}^{\mathrm{T}}\bar{\boldsymbol{\mathcal{R}}}\boldsymbol{k} \tag{64}$$

[9] Dies ist der Sonderfall eines bekannten Ergebnisses, wonach die Konsistenzordnung eines algebraisch stabilen diagonal-impliziten Verfahren Vier nicht überschreiten kann [19].

Mit der bereits bekannten Argumentationskette (über $\boldsymbol{\mathcal{R}} + \bar{\boldsymbol{\mathcal{R}}}$ in (57)) ergibt sich daher die notwendige Bedingung

$$\bar{\boldsymbol{\mathcal{R}}}\boldsymbol{k} = \boldsymbol{0}\,. \tag{65}$$

Damit wird aber aus (60) die Bedingung

$$\bar{\boldsymbol{\mathcal{S}}}\boldsymbol{k} = \boldsymbol{k}^2 - 2\boldsymbol{D}\boldsymbol{k}\,, \tag{66}$$

wobei der Ausdruck $\bar{\boldsymbol{\mathcal{S}}}\boldsymbol{k}$ nicht Null sein kann, da anderenfalls das diagonal-implizite Basisverfahren ja der Beziehung (61) genügen müsste. Schreibt man nun andererseits $\boldsymbol{x}^*[\boldsymbol{\mathcal{R}} + \mathrm{j}\bar{\boldsymbol{\mathcal{S}}}]\boldsymbol{x}$ mit $\boldsymbol{x} = \boldsymbol{a} + \mathrm{j}\boldsymbol{b}$, so erhält man nach Ausmultiplizieren unter Berücksichtigung der Symmetrie von $\boldsymbol{\mathcal{R}}$ und der Schiefsymmetrie von $\bar{\boldsymbol{\mathcal{S}}}$ die Beziehung

$$\boldsymbol{x}^*\big[\boldsymbol{\mathcal{R}} + \mathrm{j}\bar{\boldsymbol{\mathcal{S}}}\big]\boldsymbol{x} = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{\mathcal{R}}\boldsymbol{a} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\mathcal{R}}\boldsymbol{b} + 2\boldsymbol{b}^{\mathrm{T}}\bar{\boldsymbol{\mathcal{S}}}\boldsymbol{a}\,. \tag{67}$$

Der Ansatz $\boldsymbol{a} = \alpha\boldsymbol{k}$ liefert damit den Ausdruck

$$\boldsymbol{x}^*\big[\boldsymbol{\mathcal{R}} + \mathrm{j}\bar{\boldsymbol{\mathcal{S}}}\big]\boldsymbol{x} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\mathcal{R}}\boldsymbol{b} + 2\alpha\boldsymbol{b}^{\mathrm{T}}\bar{\boldsymbol{\mathcal{S}}}\boldsymbol{k}\,, \tag{68}$$

der mit $\bar{\boldsymbol{\mathcal{S}}}\boldsymbol{k} \neq \boldsymbol{0}$ durch geeignete Wahl von $\boldsymbol{b}$ und anschließend α mit Sicherheit negativ gemacht werden kann.

Damit ist der Beweis geführt, dass, trotz der für ein zweistufiges Verfahren erzielten Erhöhung der Konsistenzordnung von Drei auf Vier, die Ordnung Vier durch den Einsatz der in diesem Beitrag behandelten passiven Zweischritt-RUNGE-KUTTA-Verfahren mit diagonal-impliziten Basisverfahren nicht übertroffen werden kann.

7 Simulationsbeispiel

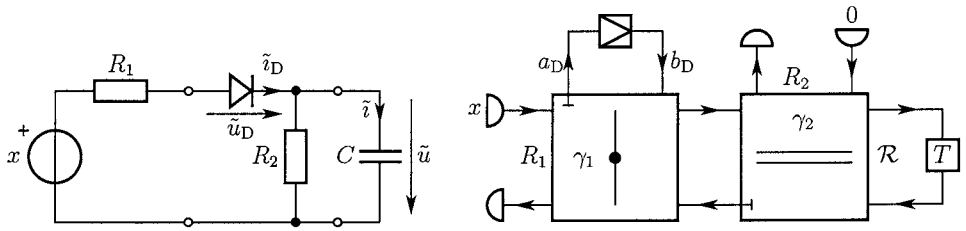


Abbildung 9: Beispiel für ein nichtlineares Netzwerk und zugehörige Wellendigitalstruktur bei Anwendung der Trapezregel.

Als Simulationsbeispiel wird die in Abbildung 9 dargestellte Anordnung betrachtet, die insbesondere eine Diode enthält. Zu der Simulation des Netzwerkes kann die (für den Fall einer Anwendung der Trapezregel) dargestellte Wellendigitalstruktur verwendet werden. Aus der Kennlinie

$$i_\mathrm{D} = i_\mathrm{D0}\left[\mathrm{e}^{\frac{u_\mathrm{D}}{u_\mathrm{D0}}} - 1\right]\,, \tag{69}$$

ist dabei eine (implizit gegebene) Abhängigkeit $b_\mathrm{D} = b_\mathrm{D}(a_\mathrm{D})$ zu ermitteln.

Um die Genauigkeit der verwendeten numerischen Integrationsverfahren bewerten zu können, wurde nun angenommen, dass die Spannung $\tilde{u}_\mathrm{D}$ über der Diode einen sinusförmigen Zeitverkauf besitzt, womit die FOURIER-Koeffizienten der exakten Lösung für die Spannung über der Kapazität in geschlossener Form angegeben werden können. Das Eingangssignal für die Schaltung wurde dann entsprechend gewählt. Netzwerk-Parameter waren $R_1 = 10\,\Omega$, $R_2 = 1\,\mathrm{k}\Omega$, $C = 1\,\mu\mathrm{F}$, $u_{\mathrm{D}0} = 30\,\mathrm{mV}$ und $i_{\mathrm{D}0} = 10\,\mathrm{pA}$. Amplitude und Frequenz der Spannung $\tilde{u}_\mathrm{D}$ wurden zu 650 mV bzw. 440 Hz gewählt.

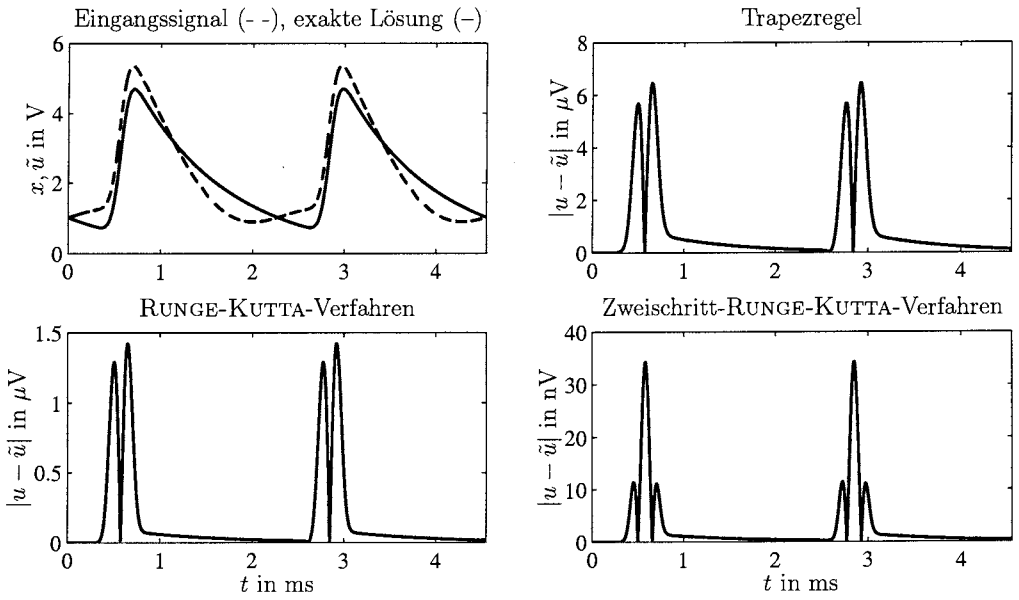


Abbildung 10: Eingangssignal $x(t)$, exakte Lösung $\tilde{u}(t)$, sowie der erzielte Simulationsfehler mit der Trapezregel, dem RUNGE-KUTTA-Verfahren mit den Parametern (59a) bzw. dem Zweischritt-RUNGE-KUTTA-Verfahren mit den Parametern (59).

Abbildung 10 zeigt die Simulationsergebnisse, die sich bei Anwendung der Trapezregel (Konsistenzordnung 2), des zweistufigen RUNGE-KUTTA-Verfahren mit den Parametern (59a) (Konsistenzordnung 3) sowie des zweistufigen Zweischritt-RUNGE-KUTTA-Verfahrens mit den Parametern (59) (Konsistenzordnung 4) ergaben. Der Zugewinn an numerischer Genauigkeit ist jeweils gut zu erkennen. Dabei betrug die Schrittweite für das RUNGE-KUTTA- und das Zweischritt-RUNGE-KUTTA-Verfahren jeweils $T = 1\,\mu\mathrm{s}$, während bei der Trapezregel $T = 500\,\mathrm{ns}$ gewählt wurde, um einen etwa vergleichbaren Rechenaufwand zu erhalten (hier ist in jedem Zeitschritt nur eine „Stufe“ statt zweier durchzurechnen).

8 Fazit

Mit den passiven linearen Mehrschrittverfahren, RUNGE-KUTTA-Verfahren und Zweischritt-RUNGE-KUTTA-Verfahren steht eine große Zahl verschiedener Integrationsverfahren zur Verfügung, die zur numerischen Lösung der Gleichungen eines KIRCHHOFF-Netz-

werkes im Rahmen des Wellendigitalkonzeptes angewendet werden können. Dabei kann durch Anwendung der RUNGE-KUTTA-Verfahren und Zweischritt-RUNGE-KUTTA-Verfahren die erreichbare Genauigkeit im Vergleich zu den LMS-Verfahren deutlich erhöht werden. Allerdings ist bekannt, dass die numerische Genauigkeit bei sehr steifen Problemen maßgeblich von der Stufenkonsistenzordnung des Verfahrens beeinflusst wird [20]. Diese ist aber bei den hier betrachteten passiven diagonal-impliziten Verfahren ja nie größer als Eins, sodass die Suche nach geeigneten Verfahren sicher noch nicht abgeschlossen ist.

Die Auswahl eines geeigneten Verfahrens hängt von verschiedenen Faktoren ab, und kann im Einzelfall erhebliche Auswirkungen auf die Qualität des numerisch ermittelten Ergebnisses aufweisen. Hier ist, trotz aller wünschenswerter numerischer Stabilitätseigenschaften, die passive Integrationsverfahren auszeichnen, in letzter Instanz die Erfahrung des Modellierers gefordert, um die Verlässlichkeit der gefundenen Lösungen bewerten und im Zweifelsfall durch Anwendung eines geeigneteren Verfahrens genauere Aussagen treffen zu können.

Literatur

[1] A. Fettweis: „Digital filter structures related to classical filter networks", *AEÜ Intern. J. Electronics and Communications*, vol. 25, no. 2, pp. 79-89, 1971.

[2] A. Fettweis: „Wave digital filters: Theory and practice", *Proceedings of the IEEE*, vol. 74, no. 2, pp. 270-327, 1986.

[3] H. D. Fischer: „Wave digital filters for numerical integration", *NTZ Archiv*, vol. 6, no. 2, pp. 37-40, 1984.

[4] K. Meerkötter; R. Scholz: „Digital simulation of nonlinear circuits by wave digital filter principles", *Proc. IEEE Intern. Symp. Circuits and Systems (ISCAS)*, pp. 720-723, 1989.

[5] T. Felderhoff: „Simulation of nonlinear circuits with periodic doubling and chaotic behavior by wave digital filter principles", *IEEE Trans. Circuits and Systems*, vol. 41, no. 1, pp. 485-491, 1994.

[6] G. Nitsche: *Numerische Lösung partieller Differentialgleichungen mit Hilfe von Wellendigitalfiltern*, Dissertation, Fakultät für Elektrotechnik, Ruhr-Universität Bochum 1994.

[7] T. Felderhoff: „A new approach to design A-stable linear multistep integration algorithms", *Proc. IEEE Intern. Symp. Circuits and Systems (ISCAS)*, vol. 5, pp. 133-136, 1994.

[8] K. Ochs: „Passive integration methods: Fundamental theory", *AEÜ Intern. J. Electronics and Communications*, vol. 55, no. 3, pp. 153-163, 2001.

[9] K. Ochs: *Passive Integrationsmethoden*, Dissertation, Fachbereich Elektrotechnik & Informationstechnik, Universität Paderborn, 2001.

[10] D. Fränken; K. Ochs: „Synthesis and design of passive Runge-Kutta methods", *AEÜ Intern. J. Electronics and Communications*, vol. 55, no. 6, pp. 417-425, 2001.

[11] D. Fränken; K. Ochs: „Numerical stability properties of passive Runge-Kutta methods", *Proc. IEEE Intern. Symp. Circuits and Systems (ISCAS)*, vol. 3, pp. 473-476, 2001.

[12] D. Fränken; K. Ochs; M. Schmidt: „A passive two-step Runge-Kutta method for the simulation of nonlinear electrical networks", *Proc. Intern. Symp. Nonlinear Theory and its Applications (NOLTA)*, pp. 347-350, 2002.

[13] J. C. Butcher: *The numerical analysis of ordinary differential equations*, John Wiley & Sons, New York, 1987.

[14] E. Hairer; S. P. Nørsett; G. Wanner: *Solving Ordinary Differential Equations I*, Springer, New York, 1993.

[15] D. Fränken; K. Ochs: „Passive Runge-Kutta methods – Properties, parametric representation, and order conditions", zur Veröffentlichung angenommen in *BIT Numerical Mathematics*.

[16] D. Fränken: *Digitale Simulation physikalischer Systeme mit Methoden der Netzwerktheorie*, als Habilitationsschrift eingereicht in der Fakultät für Elektrotechnik, Informatik & Mathematik der Universität Paderborn.

[17] Z. Jackiewicz; S. Tracogna: „A general class of two-step Runge-Kutta methods for ordinary differential equations", *SIAM J. Num. Anal.*, vol. 32, pp. 1390-1427, 1995.

[18] G. Dahlquist: „A special stability problem for linear multistep methods", *BIT Numerical Mathematics*, vol. 3, pp. 27-43, 1963.

[19] E. Hairer: „Highest possible order of algebraically stable diagonally implicit Runge–Kutta methods", *BIT Numerical Mathematics*, vol. 20, pp. 254-256, 1980.

[20] E. Hairer; G. Wanner: *Solving Ordinary Differential Equations II*, Springer, New York, 1996.

SLIDING MODE MOTION CONTROL WITH FUZZY DISTURBANCE ESTIMATION

Andreja Rojko, Karel Jezernik
University of Maribor FERI, Smetanova 17, 2000 Maribor, SLOVENIA
E-Mail: andreja.rojko@uni-mb.si, karel.jezernik@uni-mb.si

This paper describes a motion control scheme that was successfully applied to control of a direct drive (DD) robot. The scheme belongs to the class of a sliding mode control with disturbance estimation. Adaptive fuzzy disturbance estimator works as an identifier of unknown and variable part of a system dynamics. Adaptation algorithm is derived by using the Lyapunov stability theory and provides global asymptotic stability of the state errors. For the implementation on the robot, physically based fuzzy logic subsystems are introduced. With this estimator's transparency is enhanced and complexity reduced. If linguistic knowledge is available, subsystems enable its systematic inclusion.

1 Introduction

The aim of this work was to develop and implement a motion controller suitable for DD robots. Specific mechanical construction of those robots means that the dynamic model is nonlinear and very complex. So for the control a model-free control is preferred. Sliding mode control, fuzzy control, neural networks and their combinations are between possible choices. Especially when combining sliding mode and fuzzy methods some very good motion controllers were developed (survey in [3]). For example, an adaptive fuzzy logic controller was used for approximation of the desired sliding mode control law with a boundary layer [1]. Controller based on fuzzy logic with an addition of discontinuous term of sliding mode control was described in [9]. Fuzzy logic was also used for tuning of the parameters of sliding mode controller [10]. However high complexity of most of those algorithms limits their applicability in practice.

In this paper proposed control method includes fuzzy logic for an estimation of almost whole system dynamics in the sliding mode control scheme. With use of the physically based subsystems it was possible to reduce rule base to only few rules, without influencing the estimation accuracy or robustness.

Paper is organized as follows. In the Section 2. a sliding mode control for DD robot is developed. An adaptive fuzzy logic system (FLS) for disturbance estimation is proposed in the section 3. In the Section 4. the validity of the proposed control scheme is verified by the experiments on the DD robot. The tracking accuracy for an average movement is shown. Summary of the work is given in the Section 5.

2 Sliding mode control design

Dynamics of k-th robot axis for a direct drive robot with m-degrees of freedom is:

$$\tau_k = J_{kk}(\boldsymbol{q})\ddot{q}_k + \sum_{j=1, j\neq k}^{m} J_{kj}(\boldsymbol{q})\ddot{q}_j + \sum_{j=1}^{m}\sum_{l=1}^{m} C_{jl,k}(\boldsymbol{q})\dot{q}_j\dot{q}_l + G_k(\boldsymbol{q}) + \tau_{k,t}(\boldsymbol{q},\dot{\boldsymbol{q}}) + \tau_{k,d} \tag{1}$$

where τ_k is robot axis's motor torque, $J(\boldsymbol{q})$ is inertia, $C_k(\boldsymbol{q}), G_k(\boldsymbol{q})$ are Coriollis and gravitation torques, $\tau_{k,t}(\boldsymbol{q},\dot{\boldsymbol{q}})$ and $\tau_{k,d}$ are friction and external disturbances torques. $q_k, \dot{q}_k, \ddot{q}_k$ are axis position, velocity and acceleration. As disturbance let us define the whole dynamic effects (2), with exception of constant part of inertia.

$$w_k = \Delta J_{kk}(\boldsymbol{q})\ddot{q}_k + \sum_{j=1, j\neq k}^{m} J_{kj}(\boldsymbol{q})\ddot{q}_j + \sum_{j=1}^{m}\sum_{l=1}^{m} C_{jl,k}(\boldsymbol{q})\dot{q}_j\dot{q}_l + G_k(\boldsymbol{q}) + \tau_{t,k}(\boldsymbol{q},\dot{\boldsymbol{q}}) + \tau_{b,k} + \tau_{d,k} \tag{2}$$

Now the dynamics (1) can be rewritten as:

$$\tau_k = \bar{J}_{kk}\ddot{q}_k + w_k(\boldsymbol{q},\dot{\boldsymbol{q}},\ddot{\boldsymbol{q}}), \tag{3}$$

where $\bar{J}_{kk}$ denotes the constant part of inertia.

For the control design we employ a decentralized switching scheme. To include tracking requirements, all m sliding surfaces have to be chosen as a function of acceleration, velocity and position errors:

$$\sigma_k = \left(\ddot{q}_k^d - \ddot{q}_k\right) + g_{2,k}\left(\dot{q}_k^d - \dot{q}_k\right) + g_{1,k}\left(q_k^d - q_k\right), \tag{4}$$

where $g_{1,k}$ and $g_{2,k}$ are positive constants and superscript d denotes reference values. Because this switching function is a function of unknown actual acceleration, it cannot be calculated. To solve this problem the method from [2] can be used. This method proposes a transformation of the switching functions, so that no actual acceleration is needed. Transformed switching functions still define the same sliding manifolds as original switching functions. Let the transformed switching functions be:

$$\sigma_{transf,k} = i_{a,k}^c - i_{a,k}, \tag{5}$$

with reference current $i_{a,k}$ and actual current are defined as $i_{a,k}^c$

$$i_{a,k} = \frac{\bar{J}_{kk}}{K_{M,k}}\cdot\ddot{q}_k + \frac{1}{K_{M,k}}\cdot w_k, \quad i_{a,k}^c = \frac{\bar{J}_{kk}}{K_{M,k}}\cdot\ddot{q}_k^c + \frac{1}{K_{M,k}}\cdot\hat{w}_k \tag{6}$$

In actual current is actual acceleration replaced by calculated one $\ddot{q}_k^c$ (7), derived from the condition of sliding regimee $\sigma_k = 0$. Also the disturbance is in actual current replaced by estimated one $\hat{w}$.

$$\sigma_k = 0 \;\rightarrow\; \ddot{q}_k^c = \ddot{q}_k^d + g_{2,k}\left(\dot{q}_k^d - \dot{q}_k\right) + g_{1,k}\left(q_k^d - q_k\right) \tag{7}$$

If we now rewrite (5) with using (6) we get :

$$\sigma_{transf,k} = i_{a,k}^c - i_{a,k} = \frac{\bar{J}_{kk}}{K_{M,k}} \cdot \sigma_k + \frac{1}{K_{M,k}} \cdot \left(\hat{w}_k - w_k\right). \tag{8}$$

From (8) can be seen that when $\lim_{t\to\infty}\left(\hat{w}_k - w_k\right) = 0$, *k=1..m*, the transformed switching functions define the same sliding manifolds as original switching functions. The importance of good disturbance estimation can be also shown by error equation on $\sigma_{transf,k} = 0$, (9).

$$\left(\ddot{q}_k^d - \ddot{q}_k\right) + g_{2,k}\left(\dot{q}_k^d - \dot{q}_k\right) + g_{1,k}\left(q_k^d - q_k\right) = \bar{J}_{kk}^{-1}\left[\hat{w}_k - w_k\right] \tag{9}$$

For disturbance estimation a lot of different schemes can be used. Some of them [2] were tested on our three degree of freedom direct drive robot, but failed to provide a good tracking accuracy specially for high speed movements [5]. So in the next section of the paper an emphasis is put on a design of a good disturbance estimator.

3 Fuzzy disturbance estimation

One multiple input - single output FLS for a disturbance estimation (under disturbance we consider all dynamic effects with exception of the constant part of inertia) was designed for each robot axis. A vector of FLS input variables is defined as $\boldsymbol{x_k} = \left[q_k, \dot{q}_k, \ddot{q}_k^d\right]$, where $\ddot{q}_k^d$ is desired acceleration used instead of an unknown actual acceleration. Fuzzy rule base of FLS on *k*-th robot axis consist of IF-THEN rules R_k^l in the following form:

$$R_k^l: \text{ if } q_k = X_k^{q,l} \text{ and } \dot{q}_k = X_k^{\dot{q},l} \text{ and } \ddot{q}_k^d = X_k^{\ddot{q}^d,l} \text{ then } \hat{w}_k = \tau_k^l \tag{10}$$

Superscript *l* refers to the *l*-th rule *l=1..M*. $X_k^{q,l}$, $X_k^{\dot{q},l}$, $X_k^{\ddot{q}^d,l}$ are fuzzy sets in an input universe of discourse, $\hat{w}_k$ are output linguistic variables and τ_k^l are singleton fuzzy sets in the output universe of discourse.

The structure of FLS was chosen as: singleton output membership functions, singleton fuzzifier, product-operation rule of fuzzy implication, sum-operation for fuzzy aggregation

and center of average deffuzifier. Bell shaped function form was chosen for input membership functions. The output of the resulting FLS can be calculated as:

$$\hat{w}_k = \frac{\sum_{l=1}^{M} \bar{y}_k^l \prod_{i=1}^{n} \mu_{R_k^{i,l}}(x_k^i)}{\sum_{l=1}^{M} \prod_{i=1}^{n} \mu_{R_k^{i,l}}(x_k^i)}, \quad \mu_{R_k^{i,l}}(x_k^i) = \left(1 \Big/ \left(1 + \left| \frac{x_k^i - \bar{x}_k^{i,l}}{\sigma_k^{i,l}} \right|^{2 \cdot b_k^{i,l}} \right) \right), \tag{11}$$

where $\bar{x}_k^{i,l}$ are the centers of the input membership functions, $\bar{y}_k^l$ are positions of the output membership functions, $\sigma_k^{i,l}$ determine the width of the bell function and $b_k^{i,l}$ its slope. x_k^i refers to the i-th element of the vector of input variables. FLS in the form of (11) are universal approximators, capable to approximate any nonlinear continuous function on a compact set to arbitrary accuracy, [8].

After choosing the structure of FLS its parameters have to be determined. To find suitable parameters of input membership functions is an easy task. First we set their number and then determine their positions, widths and slopes, so that they cover all possible values of the input variables. These values are physically limited by the positions of the limit switchers and by the maximal motor torques. More pretentious is to set the positions of the output membership functions. First, there is no sufficient linguistic knowledge for this. Second, by fixing them, the tracking accuracy would decay rapidly in the presence of unpredictable dynamic changes such as varying payload. So the positions of the singleton output membership functions $\bar{y}_k^l$s have to be adaptive.

Let us put all adjustable parameters into parameter vector $\hat{\boldsymbol{\theta}}_k = [\bar{y}_k^1, .., \bar{y}_k^M]^T$ and the rest of the expression (11) into a vector of known nonlinear functions $\xi_k(\boldsymbol{x}_k) = [\xi_k^1(\boldsymbol{x}_k), .., \xi_k^l(\boldsymbol{x}_k), .., \xi_k^M(\boldsymbol{x}_k)]^T$, where $\xi_k^l(\boldsymbol{x}_k)$ are defined with

$$\xi_k^l = \left(\prod_{i=1}^{n} \mu_{R_k^{i,l}}(x_k^i) \right) \Big/ \left(\sum_{l=1}^{M} \prod_{i=1}^{n} \mu_{R_k^{i,l}}(x_k^i) \right). \tag{12}$$

With using these two vectors, the FLS (11) can be written in the parameter vector – regressor form (13), known from the classical theory of system identification [4].

$$\hat{w}_k = \hat{\boldsymbol{\theta}}_k^T \cdot \xi_k(\boldsymbol{x}_k) \tag{13}$$

Next the adaptation law has to be determined, so that the difference between the optimal parameter vector $\boldsymbol{\theta}_k^T$ (which guarantee perfect perturbance estimation) and used estimated parameter vector $\hat{\boldsymbol{\theta}}_k^T$, which is denoted by $\tilde{\boldsymbol{\theta}}_k^T$, is minimized:

$$\Delta w_k = \hat{w}_k - w_k = \left(\hat{\boldsymbol{\theta}}_k^T - \boldsymbol{\theta}_k^T \right) \cdot \xi_k(\boldsymbol{x}_k) = \tilde{\boldsymbol{\theta}}_k^T \cdot \xi_k(\boldsymbol{x}_k). \tag{14}$$

The adaptation law must assure a global asymptotic stability of equilibrium point $\boldsymbol{e}_k = [e_k, \dot{e}_k]^T = \boldsymbol{0}$ and can be derived by using Lyapunov stability theory. Let us chose following Lyapunov function:

$$V = \frac{1}{2}\left(\boldsymbol{e}_k^T \cdot \boldsymbol{A}_k \cdot \boldsymbol{e}_k + \frac{\bar{J}_{kk}}{\alpha_k} \cdot \tilde{\boldsymbol{\theta}}_k^T \cdot \tilde{\boldsymbol{\theta}}_k \right), \quad \boldsymbol{A}_k = \begin{bmatrix} a_{2,k} & a_{1,k} \\ a_{1,k} & a_{2,k} \end{bmatrix}. \tag{15}$$

where $\boldsymbol{A_k}$ is a symmetric, positive definite. It can be proven, that the derivative of this Lyapunov function is negative semi-definite and therefore equilibrium point stable, if we use adaptation law (16). Further it can be shown by using the Barbalat theorem, that the equilibrium point is asymptotically stable (details in [5]).

$$\bar{y}_k^l(n+1) = \bar{y}_k^l(n) - \boldsymbol{a}_k^T \boldsymbol{e}_k \xi_k^l(\boldsymbol{x}_k), \quad \boldsymbol{a}_k = [\alpha_{e,k}, \alpha_{\dot{e},k}]^T \tag{16}$$

In the derivation process it was assumed that $\dot{\tilde{\boldsymbol{\theta}}} = \dot{\hat{\boldsymbol{\theta}}}$. This is a usual in the design of adaptive systems and it means that the optimal parameter vector is constant. However the derived adaptation law can be also used on the systems with time varying parameters, if the adaptation is much faster than parameter changes.

Next we try to reduce the complexity of FLS. FLS on each robot axis was divided to three FLSB, each for estimation of one part of dynamic influencing this robot axis. First FLSB is designed to estimate the difference between the average inertia, which is already used in the control scheme, and the actual inertia. Its inputs are position and desired acceleration, $\boldsymbol{x}_k = [q_k, \ddot{q}_k^d]$.

$$\hat{w}_{1.FLS,k} = \left(J_{kk}(q_k) - \bar{J}_{kk} \right) \cdot \ddot{q}_k^d. \tag{17}$$

Second FLSB inputs are position and velocity, $\boldsymbol{x}_k = [q_k, \dot{q}_k]$, and it is designed for an estimation of the gravitation, Coriollis, centrifugal forces and velocity and position dependent friction effects:

$$\hat{w}_{2.FLS,k} = C_k(q_k, \dot{q}_k,) + G_k(q_k) + \tau_{k,t}(q_k, \dot{q}_k). \tag{18}$$

Third FLSB is designed for the estimation of other effect and eventual residue of dynamics that should be estimated by the first two FLSB. Its inputs are actual position, velocity and desired acceleration, $\boldsymbol{x}_k = [q_k, \dot{q}_k, \ddot{q}_k^d]$.

Estimator's transparency is by using FLSB preserved, regardless of the complexity of its task (estimating the dynamic of a direct drive robot). Also the number of inputs into single rule is reduced. Additionally the decision which rule of all possible rules should be included in the rule base is much easier. This all together reduces the complexity of FLS without negatively influencing its performance and enables a real-time implementation of the controller.

4 Application of the motion control algorithm

Our control plant is a three-degree of freedom Puma like configuration direct drive robot, Figure 1. Controller scheme is shown on Figure 2.

Three membership functions have been used for each of the inputs of FLS, same for all three axes, Figure 3. All used rules are shown in Table 1. Sliding manifold parameters were chosen as $g_{1,1}$=1000, $g_{1,2}$=2400, $g_{1,3}$=1200, $g_{2,1}$=63.25, $g_{2,2}$=97.97, $g_{2,3}$=69.28, parameters of diagonal inertia matrix as $\bar{J}_{11}$=3.5, $\bar{J}_{22}$=2.5, $\bar{J}_{33}$=0.13 kgm^2. Learning parameters were $a_{1,k=1}$=220, $a_{2,k=1}$=1.5, $a_{1,k=2}$=250, $a_{2,k=2}$=5, $a_{1,k=3}$=35, $a_{2,k=3}$=1 and the learning rates as $\alpha_{e,k=1,2,3} = \alpha_{\dot{e},k=1,2,3} = 0.2$.

Experimental results are shown for an average movement. A reference trajectory was a point-to-point movement; same for all three axes. Reference position and velocity of robot tip is shown in Figure 4. The resulting robot tip's position error with a peak error of 2.6 mm and zero positioning error is shown in Figure 5. Figure 6. shows nominal torques and torques form FLS for all robots' axes for this movement. On second and third axis most of the torque origins from FLS, which confirms the quality of disturbance estimation.

5 Conclusion

In this paper the development and implementation of a motion controller for a direct drive robot has been presented. Sliding mode control with fuzzy disturbance estimation has been used. To cope with highly non-linear dynamics of robot manipulator and incomplete linguistic knowledge, an on-line adjustment of some FLS's parameters has been used. The control scheme is completely decentralized and besides some basic linguistic knowledge requires only known average inertia matrix. Its simple structure and small number of linguistic rules enable easy implementation of proposed control on almost every controller hardware.

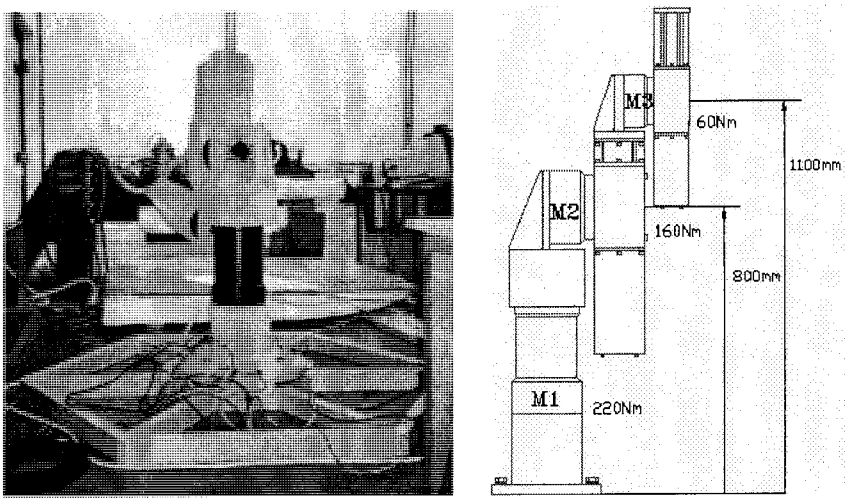


Figure 1. Control object, direct drive robot

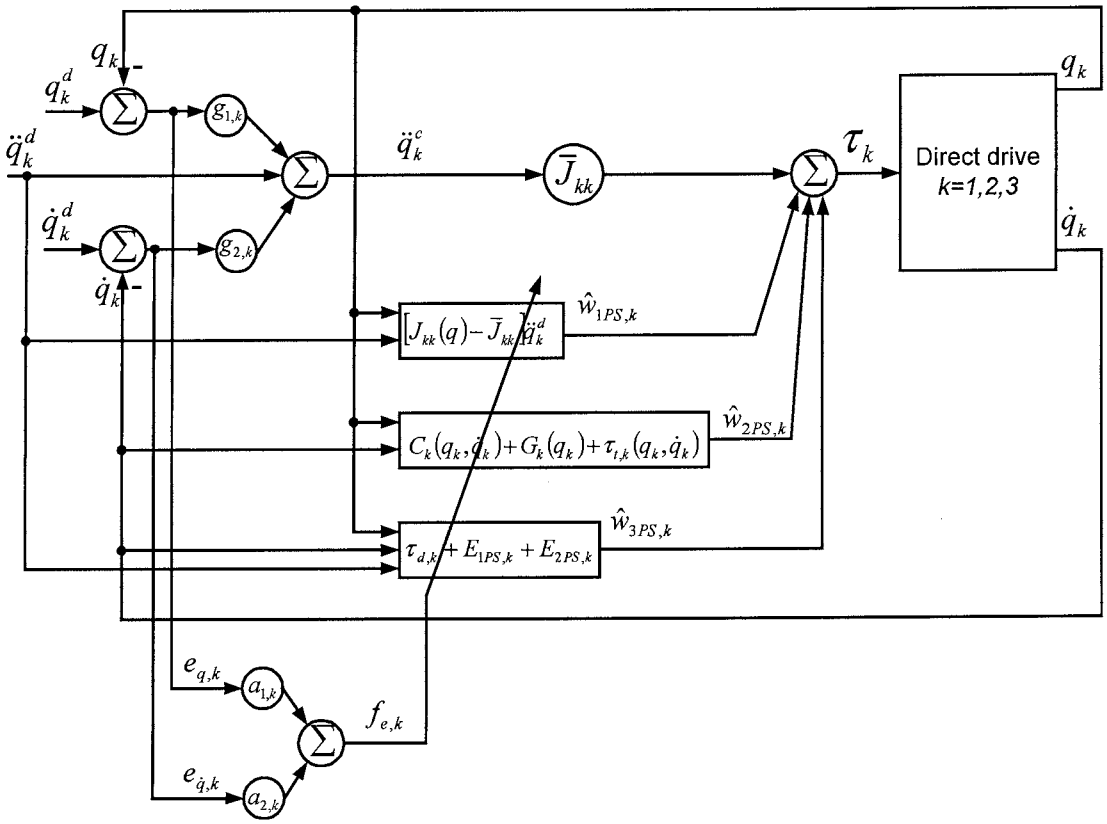


Figure 2. Control scheme

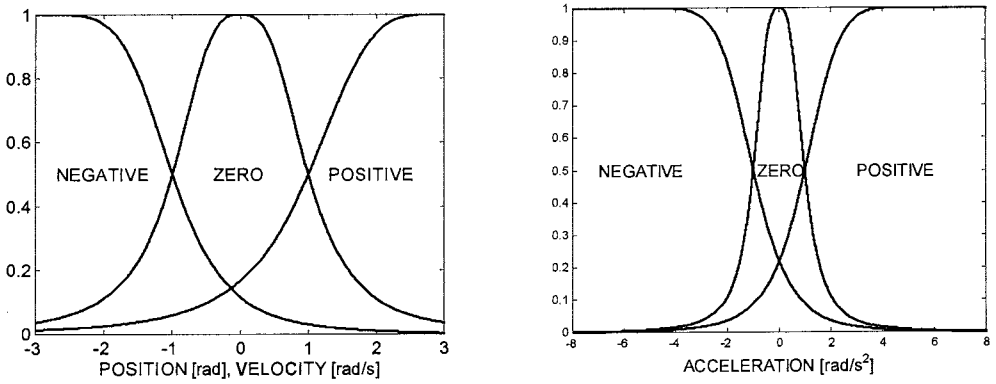


Figure 3. Input membership functions

	RULE	POSITION	VELOCITY	ACCELERATION
1. subsystem	R^1	negative	-	negative
	R^2	zero	-	zero
	R^3	positive	-	positive
2. subsystem	R^4	negative	negative	-
	R^5	zero	zero	-
	R^6	positive	positive	-
3. subsystem	R^7	positive	zero	zero
	R^8	negative	zero	zero
	R^9	zero	negative	positive
	R^{10}	zero	zero	zero
	R^{11}	zero	zero	positive
	R^{12}	zero	negative	negative
Not implemented in 1. axis	R^{13}	positive	zero	positive
	R^{14}	negative	negative	negative
	R^{15}	positive	positive	positive

Table 1: Rules in FLS

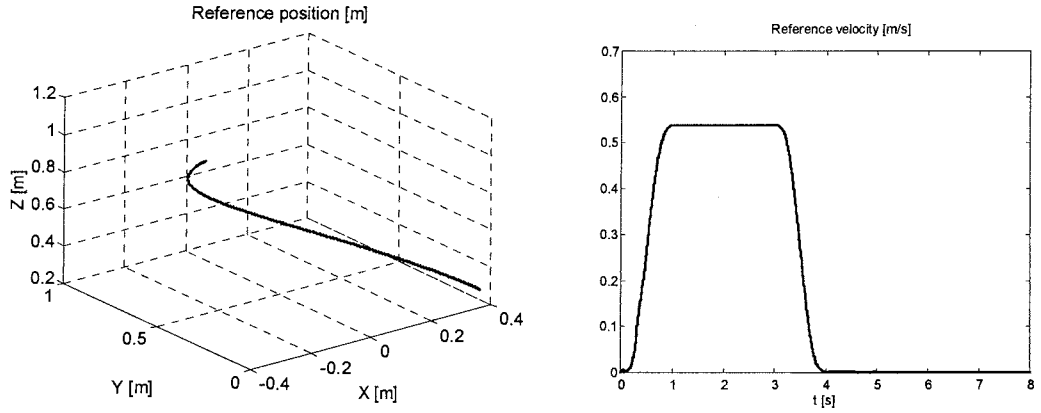


Figure 4. Reference trajectory

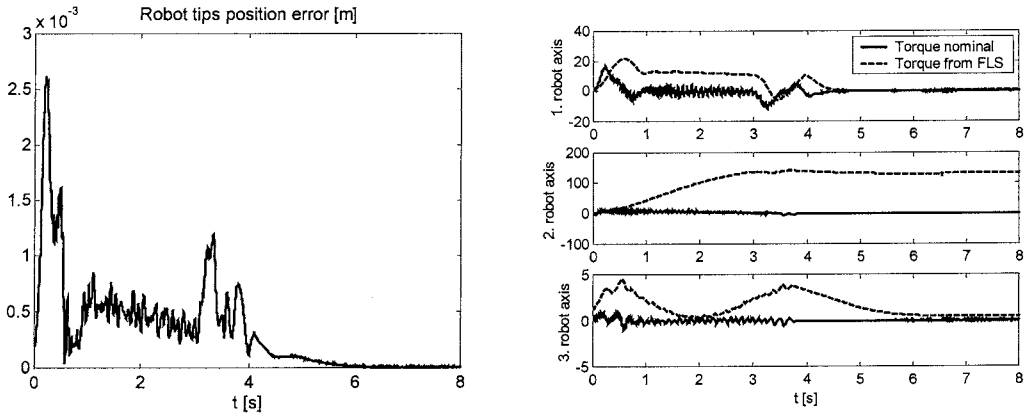


Figure 5. Robot tip's position error and applied torques

References

1. R. Berstecher, R. Palm, H. Unbehauen, "An Adaptive Fuzzy Sliding-Mode Controller", *IEEE Trans. Industrial Electronics,* Vol.48, No.1, pp.18-31, (2001).
2. K. Jezernik, B. Curk, J. Harnik, "Observer Based Sliding Mode Control of Robotic Manipulator", *Robotica*, Vol.12, pp. 443-448, (1994).
3. O. Kaynak, K. Erbatur, M. Ertugrul, "The Fusion of Computationally Intelligent Methodologies and Sliding-Mode Control - A Survey", *IEEE Trans. Industrial Electronics,* Vol. 48, No.1, pp. 4-17, (2001).
4. L. Ljung, *System identification*, Prentice- Hall, 1987.
5. A. Rojko, K. Jezernik "Sliding mode control of direct drive robot using fuzzy disturbance estimation". *Proc. of IECON '01*, Denver, pp. 335-340, (2001).
6. J. Slotine, W. Li, *Applied Nonlinear Control*, Prentice-Hall, Englewood Cliffs, (1991).
7. V. I. Utkin, *Sliding Modes in Control Optimization*, Springer Verlag, (1981).
8. L. X. Wang, *Adaptive Fuzzy Systems and Control*, Prentice Hall, Englewood Cliffs, (1994).
9. B. K. Yoo and W. C. Ham, "Adaptive Control of Robot Manipulator Using Fuzzy Compensator", *IEEE Trans. Fuzzy Syst.*, vol. 8, pp. 186-199, (2000)
10. A. Šabanovič A., K. Jezernik, K. Erbatur and O. Kaynak, "Soft Computing techniques in Discrete Time Sliding Mode Control Systems", *Automatika* 38, pp.7-14, (1997).

Optimierter Betrieb einer Energiequelle mit Zwischenspeicher

G. Steinmaurer
Institut für Regelungstechnik und elektrische Antriebe
Universität Linz
Altenbergerstraße 69, 4040 Linz
gerald.steinmaurer@jku.at

Kurzfassung

In vielen praktischen Anwendungsbeispielen tauchen Probleme auf, bei denen 2 Leistungsquellen eine vorab bekannte Leistung liefern müssen. Eine Quelle dient meist der Abdeckung der Hauptlast, während eine zweite Quelle zur Energiezwischenspeicherung verwendet werden kann. Dadurch wird der Einsatz eines Energie-Management Systems nötig, der die Leistungsaufteilung der beiden Quellen steuert. In diesem Aufsatz wird eine Methode zum Entwurf einer solchen Steuerung gezeigt.

1 Einleitung

Leistungsquellen mit Zwischenspeicher sind Systeme, bei denen die zu liefernde Energie aus zwei unterschiedlichen Quellen, der eigentlichen Energiequelle und einem Speicher kommen kann, wobei auch die Möglickeit besteht, den Speicher sowohl als Leistungsquelle als auch als Leistungssenke zu betreiben. Ein typischer Vertreter eines solchen Systems stellt das Hybridfahrzeug dar, welches seine für den Antrieb benötigte Energie sowohl aus einer Verbrennungskraftmaschine (oder ähnlichem) und einer Batterie in Kombination mit einem Elektromotor bezieht. Eine Schlüsselrolle bei speicherfähigen Energiequellen nimmt das Energie-Mangagement ein, das die Aufgabe der Steuerung der beiden Quellen übernimmt.

Das Standardziel dieses Energie-Management ist die Minimierung einer Kostenfunktion (meist der Gesamtverbrauch der Hauptquelle) unter Berücksichtigung von Beschränkungen und Nebenbedingungen, wie z.B. Ober- und Untergrenzen der Speicherkapazität, Leistungsgrenzen etc., die meist physikalisch motiviert sind oder auch vom Gesetzgeber vorgegeben werden können (z.b. Obergenze der im Betrieb entstehenden Gesamtemissionen).

In den meisten Fällen von speicherfähigen Energiequellen ist eine Quelle für die Hauptenergieversorgung zuständig, deren Speicherfähigkeit ist, verglichen mit der Kapazität der zweiten Quelle, gering. Beide Quellen und die Last liefern bzw. beziehen ihre Leistung von einem gemeinsamen Leistungsbus (siehe Abb. 1). Die gemeinsame Größe am Bus kann abhängig von der Art der Energiequelle eine elektrische Spannung, ein elektrischer Strom, eine Temperatur oder ein Druck sein.

Es wird davon ausgegangen, dass der zeitliche Belastungsverlauf $P_{load}(t)$ bereits bekannt ist. Dies erscheint auf den ersten Blick eine starke Restriktion zu sein, bei weiterer Betrachtung stellt sich diese Annahme als durchaus sinnvoll heraus.

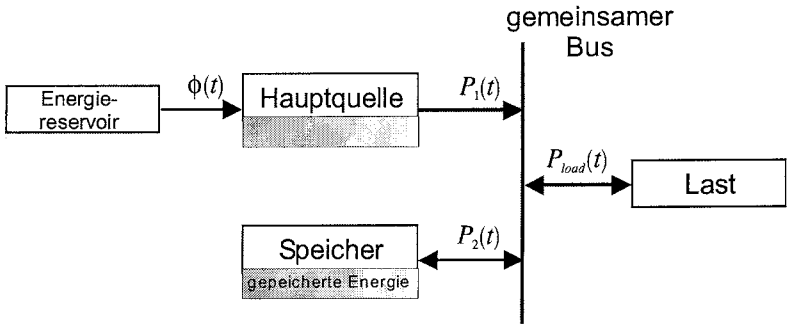


Abb. 1. Typische Grundstruktur einer Leistungsquelle mit Speicher

Im Fall von Hybridfahrzeugen werden Emissions- und Verbrauchstest anhand eines gesetzlich festgelegten Normzyklus durchgeführt, bei dem der Geschwindigkeitsverlauf vorgegeben ist. Bei solch einem Fahrzyklus kann aus dem Geschwindigkeitsprofil und den Fahrzeugdaten die zur Absolvierung des Zyklus notwendige Antriebsleistung errechnet werden.

Auch bei elektrischen Energieversorgungs-Systemen kann von einem Sollprofil der zu liefernden Leistung ausgegangen werden. Diese Profile können aus historischen Daten gewonnen werden und verändern sich nur in sehr geringem Maße.

$$P_{load}(t) = \text{bekannt}$$

Elektrischer Bus

Ein Anwendungsbeispiel eines elektrischen Bussystems stellt ein hydraulisches (Speicher-)Kraftwerk dar. Hier wird die überschüssige Energie dazu verwendet, Wasser über eine Pumpe auf ein höhergelegenes Reservoir zu pumpen. Die dabei verwendete Pumpe kann meist auch wieder in der Gegenrichtung als Turbine verwendet werden. Ein weiterer typischer Anwendungsfall eines elektrischen Busses ist das Serienhybridfahrzeug, wo meist die elektrische Spannung die gemeinsame Größe am Bus ist.

Mechanischer Bus

Eine typische Anwendung einer Energiequelle mit Zwischenspeicher an einem mechanischen Bussystem ist das Parallelhybridfahrzeug, bei dem die Verbrennungskraftmaschine und eine elektrische Maschine an einer Welle (=Bus) unterschiedliche Drehmomente anbringen (Abbildung 2). Die gemeinsame Größe an der Welle ist die Drehzahl, die elektrische Maschine kann sowohl als Motor als auch als Generator betrieben werden.

2 Problemformulierung

Zur Systembeschreibung wird auf die in der Literatur [1] übliche Verwendung von Bond-Graphen zurückgegriffen, um eine möglichst allgemeine Formulierung zu finden. Betrachtet man bei einem mathematischen Modell eines physikalischen Systems die Energie als

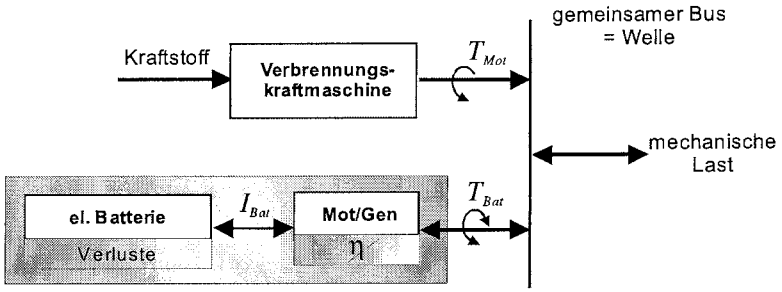


Abb. 2. Quellen mit Energiespeicher, mechanischer Bus

Austauschgröße, so führt das zur Verwendung von sogenannten Co-Variablen *effort (e)* und *flow(f)*, wobei die Energie sich durch

$$E(t) = \int_0^t e(\tau) \cdot f(\tau)\, d\tau$$

ergibt. Dieses *effort-flow Paar* ist auch unter dem Begriff *Leistungs-konjugierte Variablen* oder *Co-Variablen* bekannt. Diese effort-flow Klassifizierung erlaubt eine Systembeschreibung, die weitgehend von der physikalischen Natur des Systems unabhängig bleibt. Das Produkt dieser beiden Variablen ist gleich der Leistung des Systems.

$$P_i(t) = e_i(t) \cdot f_i(t)$$

2.1 Speicher

Behandelt wird ein ideales Speichersystem, das durch eine lineare Zustandbeschreibung

$$\begin{aligned} \dot{x} &= \frac{1}{C} f_1 \\ x(0) &= x_0, \end{aligned}$$

darstellbar ist, wobei f_1 den *flow* des Speichers repräsentiert. Die Leistung des Speichers ist vom Zustand unabhängig und ergibt sich aus $P_1 = e_1(t) \cdot f_1(t)$. Die Leistung des Speichers kann zeitlichen Beschränkungen unterworfen sein.

$$P_{1,lb}(t) \leq P_1(t) \leq P_{1,ub}(t)$$

2.2 Hauptquelle

Die Hauptenergiequelle wird als statisches System betrachtet, deren Ausgangsleistung sich aus

$$P_2(t) = e_2(t) \cdot f_2(t)$$

als Produkt einer *effort*- und einer *flow*-Variable ergibt, wobei diese Leistung beschränkt ist

$$P_{2,lb}(t) \leq P_2(t) \leq P_{2,ub}(t)$$

Es wird angenommen, dass der für die Bereitstellung der Ausgangsleistung P_2 benötigte Energieverbrauch der Hauptquelle ϕ nur vom Arbeitspunkt (e_2, f_2) der Hauptquelle abhängt

$$\phi(t) = \phi(e_2, f_2).$$

2.3 Optimierungsproblem

Aufgrund der Busstruktur müssen entweder *effort* oder *flow* der beiden Leistungsquellen gleich sein. Im weiteren Verlauf wird

$$e_1(t) = e_2(t) = e = const.$$

angenommen. Beide Quellen müssen nun so betrieben werden, dass die gesamte Ausgangsleistung einer vorab bekannten Last $P_{load}(t)$ entspricht

$$P_{load}(t) \equiv P_1(t) + P_2(t) = e \cdot (f_1(t) + f_2(t)) . \tag{1}$$

Dadurch ist man in der Lage, den Verbrauch der Hauptquelle durch $f_1(t)$ auszudrücken

$$\phi(e(t), f_2(t)) = \phi\left(e, \frac{P_{load}(t)}{e} - f_1(t)\right).$$

Es soll nun der Energieverbrauch der Hauptquelle minimiert werden

$$\min_{f_1(t)} \int_0^T \phi(e, P_{load}(t), f_1(t))dt \tag{2}$$

unter den Randbedingungen

$$\begin{aligned} \dot{x} &= \frac{1}{C} f_1 \\ x(T) - x(0) &= \Delta \\ \mathbf{h}(\mathrm{P_{load}(t)}, \mathrm{f_1(t)}) &\leq \mathbf{0} \end{aligned} \tag{3}$$

wobei $x \in R$ den Speicherzustand des Zwischenspeichers, Δ die gewünschte Änderung des Speicherzustandes und $\mathbf{h} \in \mathrm{R}^m$ einen Satz von Systembeschränkungen beschreiben.

3 Lösungsansatz

Das vorliegende Optimierungsproblem ist auf analytischem Weg in allgemeiner Form kaum lösbar, zumal es bei praktischen Problem oftmals vorkommt, dass der Momentanverbrauch $\phi(e_2, f_2)$ aber auch die Last $P_{load}(t)$ nur in Form von Messdaten und nicht durch analytische Funktionen gegeben sind. Ein üblicher Lösungsweg besteht daher in der Diskretisierung der Aufgabenstellung, also dem Suchen nach einer numerischen Lösung für

$$\min_{f_{1,k}} \sum_{k=1}^{N} T_k \cdot \phi(P_{load,k}, f_{1,k}) \tag{4}$$

mit den Randbedingungen

$$\begin{aligned} x_{k+1} &= x_k + T_k \frac{1}{C} f_{1,k} \\ x_N - x_0 &= \Delta \\ \mathbf{h}\,(\mathbf{f_{1,k}}) &\leq \mathbf{0} \end{aligned}$$

und mit T_k als den zeitlichen, nicht notwendigerweise äquidistanten, Diskretisierungintervallen. Dieses Optimierungsproblem stellt sich bei realen Anwendungen meist nichtkonvex und von großer Ordnung dar, wodurch das Finden einer globalen Lösung natürlich deutlich erschwert wird.

3.1 Bekannte Lösungsansätze

Das Problem der Erzeugung von Leistungen mit mehreren Energiequellen mit ähnlichen Aufgabenstellung wurde in der Literatur bereits ausführlich behandelt und ist unter den Begriffen *economic dispatch problem* und *optimal power flow* [2] bekannt. Hierbei wird dabei ausgegangen, dass eine endliche Anzahl von Energiequellen zu jedem Zeitpunkt eine geforderte Gesamtleistung (auch unter Berücksichtigung von Leitungsverlusten) produzieren muss. Die daraus resultierende Optimierungsaufgabe nimmt die gleiche Gestalt wie Problem (4) an, oftmals auch mit einer Abhängigkeit der Kostenfunktion und der Beschränkungen vom Zustand x.

Das klassische *optimal power flow problem* stellt eine nichtlineare Optimierungsaufgabe mit einer nichtlinearen Kostenfunktion und einem Satz von nichtlinearen Gleichungs- und Ungleichungsnebenbedingungen dar. Das Ziel ist die Minimierung einer oder mehrerer Kostenfunktionen unter Einhaltung der meist physikalisch oder sicherheitstechnisch begründeten Beschränkungen.

Bei diesem Problem geht man davon aus, dass ein verlustbehaftetes Netzwerk mit mehreren Energiequellen und Verbrauchern existiert und durch die Lösung der Leistungsanteil jeder Quelle bestimmt wird, um die Kostenfunktion zu minimieren. Die Aufgabenstellung sieht folgendermaßen aus [3]

$$\begin{aligned} & & \min \quad & f(\mathbf{u}, \mathbf{x}) \\ \text{so dass.} \quad & \mathbf{h}(\mathbf{u}, \mathbf{x}) &= \; & \mathbf{0} \\ & \mathbf{g}(\mathbf{u}, \mathbf{x}) &\leq \; & 0, \end{aligned}$$

wobei $\mathbf{u}$ den Vektor der Steuergrößen, $\mathbf{x}$ den Vektor der abhängigen Variablen , $f(\mathbf{u},\mathbf{x})$ die skalare Kostenfuntion, $\mathbf{h}(\mathbf{u},\mathbf{x})$ die Beschränkungen des Leistungsflusses und $\mathbf{g}(\mathbf{u},\mathbf{x})$ die Grenzen der Steuergrößen und Betriebsbereiche beschreiben.

Das traditionelle *economic dispatch problem* nimmt an, dass die gesamte, von einer gegebenen Menge von Quellen abgegebene Leistung in einem betrachteten Zeitinverall konstant bleibt und minimiert dementsprechend die Gesamtkosten der Energieerzeugung bei Betrachtung des statischen Verhaltens und unter Berücksichtigung der Leistungsgrenzen der Quellen. Dies bedeutet, die Dimension des Optimierungsproblems ist gleich der Anzahl Leistungsquellen. Die einfache Problemstellung sieht dann folgendermaßen aus [4]

$$\begin{aligned} & \sum_{i=1}^{N} F_i(P_i), \\ \sum_{i=1}^{N} P_i &= P_D \\ P_{i,lb} &\leq P_i \leq P_{i,lb} \end{aligned}$$

wobei i den Index der Leistungsquellen, P_i die Leistung einer Quelle, F_i die anteiligen Kostenfunktion der i-ten Quelle und P_D den gesamten Leistungsbedarf bezeichnen.

Dynamic economic dispatch [5] stellt eine Erweiterung zur konventionellen Aufgabenstellung dar, indem auch zeitliche Beschränkungen im Leistungsbedarf und in der Leistungserzeugung berücksichtigt werden. Der Hintergrund der Beschränkung im Leistungsanstieg der Quellen resultiert aus Lebensdauerberechnungen, wobei versucht wird, thermische Gradienten innerhalb des Systems (z.B. Turbinen) in gewissen Grenzen zu halten.

Es existieren eine große Anzahl von Methoden der Optimierung, die auch für *economic dispatch* und *optimal power flow* Aufgaben verwendet werden. Diese Methoden können grob in folgende Klassen eingeteilt werden [6], [7]

- *Nichtlineare Programmierung* (NLP) behandelt Problemstellungen mit nichtlinearen Kostenfunktionen und Beschränkungsfunktionen. Die Beschränkungen können sowohl aus Gleichungen als auch aus Ungleichungen bestehen. Solch eine Methode kann z.B. *Sequential Unconstrained Minimization Technique* (SUMT) sein.
- Ein *Quadratisches Programm* stellt eine Sonderform des Nichtlinearen Programmes dar, wobei die Kostenfunktion eine quadratische Form besitzt.
- Bei *Newton-basierenden Lösungen* werden die notwendigen Bedingungen üblicherweise als Kuhn-Tucker Bedingungen formuliert. Im allgemeinen benötigt dieser Ansatz iterative Methoden zur Lösungsfindung.
- Bei der *Linearen Programmierung* (LP) werden Probleme behandelt, in denen sowohl die Kostenfunktion als auch die Beschränkungen in linearer Form auftreten. Zur Lösung wird dabei sehr oft die Simplex-Methode aufgrund ihrer Effizienz eingesetzt.
- *Mixed-Integer Programme* (MIP) stellen eine Sonderform des Linearen Programmes dar, bei der Beschränkungen und/oder die Zielvariable zum Teil nur Integer-Werte annehmen darf. Die Zahl der auftretenden diskreten Werte ist dabei ein wichtiger Indikator, wie schwierig es sein wird, das MIP zu lösen.

• *Innere-Punkt Methoden* lösen LP noch schneller als der Simplex-Algorithmus und wurden auch auf Nichtlineare und Qudratische Programme mit großem Erfolg und sehr guter Qualität angewandt.
• *Heuristische Methoden* sind Verfahren, deren Wirkungsweise typischen Vorgängen in Physik und Biologie nachempfunden sind. Vertreter dieser Verfahren sind *Simulated Annealing*, *Tabu search* oder *Genetische Algorithmen.*

4 Alternative Lösung

Vorgestellt wird hier eine Methode, bei der das Problem nur bezüglich der Lastanforderung $P_{load}(t)$ diskretisiert wird und für jede Last ein optimaler Verlauf von $f_1(t)$ gesucht wird,also

$$\min_{f_1(t)} \sum_{i=1}^{N} \left(\int_{t_{i-1}}^{t_i} \phi(P_{load,i}, f_1(t))\, dt \right)$$

mit den Randbedingungen

$$\begin{aligned} \dot{x} &= \frac{1}{C} f_1 \\ x(T) - x(0) &= \Delta \\ \mathbf{h}(\mathbf{f_1}(\mathbf{t})) &\leq \mathbf{0}. \end{aligned}$$

4.1 Optimierung eines Einzelschrittes

Im Sinne des Bellman'schen Optimalitätsprinzips, wonach beim Durchfahren der optimalen Trajektorie auch Teilstücke davon in optimaler Weise absolviert werden, ist nun ein optimaler Verlauf innerhalb des $i-ten$ Schrittes, wobei die Gleichungsnebenbedingung

$$x_i - x_{i-1} = \delta_i \tag{5}$$

unter Einbeziehung von Leistungsgrenzen

$$\mathbf{h}(\mathbf{f_1}(\mathbf{t})) \leq \mathbf{0} \tag{6}$$

die Zielfunktion

$$\min_{f_1(t)} \int_{t_{i-1}}^{t_i} \phi(P_{load,i}, f_1(t))\, dt \tag{7}$$

gesucht.

4.1.1 Analytische Funktionen

Es lässt sich mit einer Reihe von notwendigen Bedingungen (Euler-Lagrange Bedingung, Legendre Bedingung, Weierstrass-Erdmann-Eck Bedingungen zeigen [8], dass das Problem (7) unter der Nebenbedingung (5) und unter der Ungleichungsnebenbedingung (6)

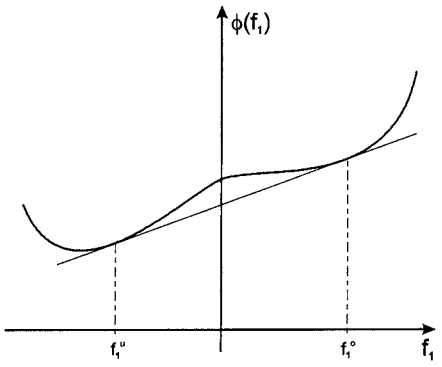


Abb. 3. Typischer Verlauf von $\phi_i(f_1(t))$ bei konstanter Last P_{load}

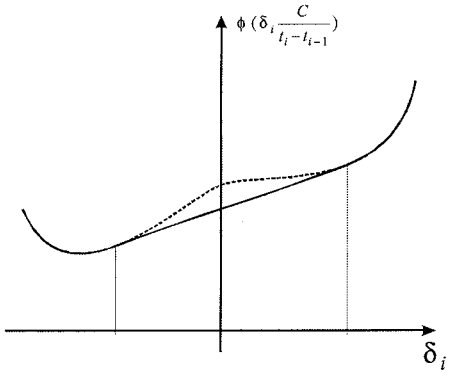


Abb. 4. Kosten in Abhängigkeit von δ_i bei konstanter Last

zu Kosten im $i-ten$ Abschnitt führen, die konvex bezüglich der gewünschten Zustandsänderung δ_i sind. Diese Bedingungen endet auch in einer sehr anschaulichen geometrischen Interpretation aufgrund der Charakteristik von $\phi_i(\delta_i) = \phi_i(\frac{t_i - t_{i-1}}{C} f_1)$ (Abbildungen 3 und 4), also

$$\phi^*(P_{load,i}, \delta_i) = \phi_i^* = konvex$$

Um ein konvexes Optimierungsproblem zu erhalten, muss neben einer konvexen Kostenfunktion auch durch eventuelle Gleichungs- und Ungleichungsnebenbedingungen eine konvexe Menge festgelegt werden. Dies ist bei diesem Problem der Fall, da durch Leistungsbeschränkungen bei konstantem *effort* e direkt auf Ober- und Untergrenzen des *flow* f_1 geschlossen werden kann [1].

4.1.2 Nicht-analytische Kostenfunktion

Es wurde bisher von einer Kostenfunktion ausgegangen, die in analytischer Form vorliegt. Bei der Behandlung von technischen Aufgabenstellungen, wie z.B. der Minimierung des

[1]Es kann auch gezeigt werden, dass bei einem nicht idealen Speicher (z.B. elektrische Spannungsquelle e_0 mit Innenwiderstand R mit $P_1 = (e_0 + R \cdot f_1) \cdot f_1$) die Leistungsbeschränkung zu einem konvexen zulässigen Bereich $f_1 \in X$ führt.

Kraftstoffverbrauchs eines Verbrennungsmotors ist es somit nötig, die Kostenfunktion durch eine analytische Funktion zu approximieren. Oftmals ist es jedoch so, dass die Kostenfunktion nur als Messgröße zu einer endlichen Anzahl von Betriebspunkten vorliegt und die Kosten zwischen diesen Messpunkten (linear) interpoliert werden. Dies führt aber zu einer Kostenfunktion, die im Allgemeinen nicht mehr glatt ist, d.h. es existieren Punkte der Kostenfunktion, deren Ableitung nicht mehr existiert. Die notwendigen Bedingungen für ein Optimum werden aber durch die Ableitungen der Kostenfunktion gebildet.

Da aber die optimale Lösung im analytischen Fall auf der konvexen Hülle der Kostenfunktion (skaliert mit der Zeit des jeweiligen Diskretisierungsabschnittes und der Speicherkapazität) zu liegen kommt, liegt die Vermutung nahe, dass dies bei nichtanalytischen Kostenfunktionen ebenfalls der Fall ist.

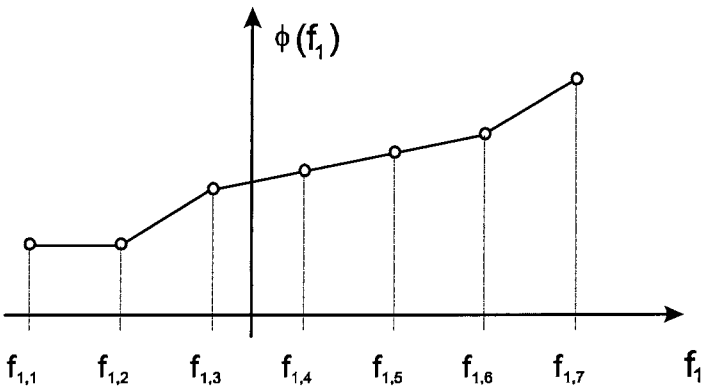


Abb. 5. Aus Messwerten gebildete Kostenfunktion

Lösung des Optimierungsproblems

Im $i-ten$ Intervall $[t_{i-1}, t_i]$ soll die Kostenfunktion (bestehend aus M diskreten Messwerte-Paaren $(f_{1,j}, \phi_{i,j} = \phi_i(f_{1,j}))$ unter der Nebenbedingung

$$\frac{1}{C} \int_{t_{i-1}}^{t_i} f_1(t)dt = \delta_i \tag{8}$$

minimiert werden. Wird jetzt angenommen, dass nur diese diskreten Messpunkte in der Suche nach einem optimalen Verlauf berücksichtigt werden, so kann jedem Messpunkt eine auf die gesamte Intervalllänge normierte Zeitdauer $t_{i,j} \in [0, 1]$ zugeordnet werden, wobei die Gesamtzeit aller Verweilzeiten in einem Messpunkt der Gleichung

$$\sum_{j=1}^{M} t_{i,j} = 1$$

genügen muss. Um nun die Nebenbedingung der Optimierung (8) zu erfüllen, muss weiters

$$\frac{\Delta T_i}{C} \sum_{j=1}^{M} t_{i,j} \cdot f_{1,j} = \delta_i$$

mit $\Delta T_i = (t_i - t_{i-1})$ gelten. Die Kostenfunktion der Optimierung kann durch

$$J_i = \sum_{j=1}^{M} \phi_{i,j} \cdot t_{i,j}$$

gefunden werden. Durch das Einführen der Vektoren

$$\begin{aligned} \mathbf{t_i} &= [t_{i,1}, t_{i,2}, ...t_{i,M}]^T \\ \boldsymbol{\phi_i} &= [\phi_i(f_{1,1}), \phi(f_{1,2}), ..., \phi(f_{1,M})]^T \end{aligned}$$

kann das Problem in vektorieller Schreibweise dargestellt werden

$$\min_{\mathbf{t_i}} \boldsymbol{\phi}_\mathbf{i}^\mathbf{T} \cdot \mathbf{t_i}$$

so, dass

$$\begin{aligned} \underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ f_{1,1} & f_{1,2} & \cdots & f_{1,M} \end{bmatrix}}_{\mathbf{A}} \mathbf{t_i} &= \underbrace{\begin{bmatrix} 1 \\ \frac{\delta_i \cdot C}{\Delta T_i} \end{bmatrix}}_{\mathbf{b}}. \\ \mathbf{t_i} &\geq \mathbf{0} \\ \mathbf{t_i} &\leq \mathbf{1} \end{aligned} \tag{9}$$

Diese Darstellung eines vektoriellen Optimierungsproblems entspricht aber genau dem eines linearen Programms in Standardform (ausgenommen die Bedingung $\mathbf{t_i} \leq \mathbf{1}$). Ein lineares Programm besitzt folgende Struktur

$$\begin{aligned} & \min_{\mathbf{t_i}} \boldsymbol{\phi}_\mathbf{i}^\mathbf{T} \mathbf{t_i} \\ \mathbf{A} \cdot \mathbf{t_i} &= \mathbf{b} \\ \mathbf{t_i} &\geq \mathbf{0}, \end{aligned}$$

wobei die Matrix $\mathbf{A}$ die Dimension $(m \cdot n)$ besitzt. Durch Bildung der Matrix $\mathbf{B}$ aus m unabhängigen Spaltenvektoren von $\mathbf{A}$ und

$$\mathbf{B} \cdot \mathbf{t_{i,B}} = \mathbf{b}$$

kann die sogenannte Basislösung des Optimierungsproblems gebildet werden

$$\mathbf{t_i} = \begin{bmatrix} \mathbf{t_{i,B}} \\ \mathbf{0} \end{bmatrix} \text{... Basislösung von } \mathbf{A} \cdot \mathbf{t_i} = \mathbf{b}. \tag{10}$$

Aufgrund des Fundamentalsatzes, dass das Problem eine optimale zulässige Basislösung besitzt, wenn eine optimale zulässige Lösung existiert, müssen nur Basislösungen des linearen Programmes untersucht werden.

Nimmt man nun an, dass bei der Kostenfunktion zumindest 2 Werte von $f_{1,j}$, (mit $j = 1..M$) unterschiedlich sind (was natürlich der Fall ist, wenn die Kostenfunktion zumindest aus 2 Punkten besteht), dann folgt

$$Rang(\mathbf{B}) = \mathbf{2}.$$

Somit ist

$$\dim(\mathbf{t_{i,B}}) = \mathbf{2}$$

und daher besteht aufgrund von (10) die Basislösung $\mathbf{t_{i,B}}$ nur mehr aus (höchstens) 2 Einträgen, die ungleich 0 sind. Somit ergibt sich für jede Basislösung die Kostenfunktion als Linearkombination von 2 Messpunkten k, l der Kostenfunktion

$$J_i = t_{i,k} \cdot \phi_i(f_{1,k}) + t_{i,l} \cdot \phi_i(f_{1,l}), \tag{11}$$

mit

$$t_{i,l} = 1 - t_{i,k}$$

Wird jetzt noch die Beschränkung (9) in die Optimierung mit aufgenommen, so folgt $t_k, t_l \in [0, 1]$. Aus Gleichung (11) ist dann leicht zu ersehen, dass die optimale Basislösung

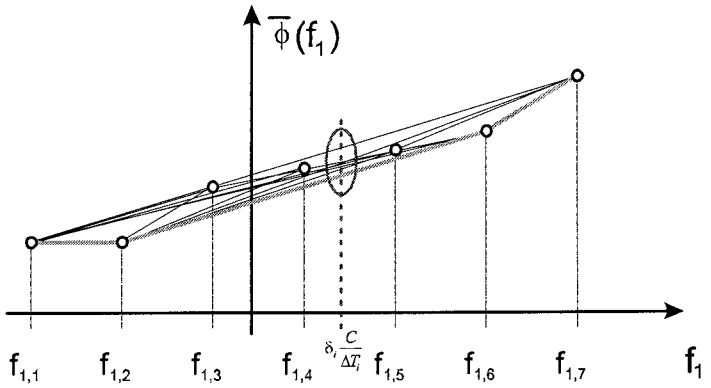


Abb. 6. Basislösungen des Optimierungsproblems

des Optimierungsproblemes auf der konvexen Hülle der Verbindungspunkte der Messwerte zu liegen kommt. Abbildung (6) zeigt in einem Beispiel, dass die optimale Lösung zum Erreichen der Nebenbedingung $\frac{\Delta T_i}{C} \sum_{j=1}^{M} t_{i,j} \cdot f_{1,j} = \delta_i$ sich als Linearkombination von $f_{1,2}$ und $f_{1,6}$ ergibt.

4.1.3 Ordnungsreduktion

Der vorangegangene Abschnitt zeigte, dass die optimale Lösung in einem Abschnitt konstanter Leistung entweder aus einem konstanten Wert für $f_1(t)$ oder aus einem Schalten zwischen 2 Werten besteht. Durch die leistungsmäßige Diskretisierung können auch gleiche Leistungen P_{load} zu unterschiedlichen Zeitpunkten des Optimierungszeitraums auftreten.

Angenommen, $\boldsymbol{\delta}^* = [\overline{\boldsymbol{\delta}}_1^*, \overline{\boldsymbol{\delta}}_2^*, ..., \overline{\boldsymbol{\delta}}_N^*]$, mit $\overline{\delta}_i = \frac{\delta_i}{\Delta T_i}$ als die auf das i-te Zeitintervall bezogene Zustandsänderung, sei die Lösung der Optimierungsaufgabe

$$\min_{\delta} \sum_{i=1}^{N} \Delta T_i \cdot \phi_i(P_{load,i}, \overline{\delta}_i) \tag{12}$$

so dass

$$\sum_{i=1}^{N} \Delta T_i \cdot \overline{\delta}_i = \Delta.$$

Man kann nun zeigen, dass, falls die Kostenfunktion die konvexe Eigenschaft besitzt, bei einem Auftreten gleicher Belastungen (und somit gleicher Kostenfunktionen $\phi_k(P_{load,k}, \overline{\delta}_i) = \phi_l(P_{load,l}, \overline{\delta}_i)$ zu verschiedenen Zeitpunkten k und l

$$P_{load,k} = P_{load,l}, \ k \neq l$$

für die optimale Lösung

$$\overline{\delta}_k^* = \overline{\delta}_l^*$$

gilt.

Beweis. Die *notwendige Bedingung 1.Ordnung* der optimalen Lösung zu (12) ist, dass die partielle Ableitung der Lagrange-Fuktion L

$$L = \sum_{i=1}^{N} \Delta T_i \cdot \phi_i(P_{load,i}, \overline{\delta}_i) + \lambda \left(\sum_{i=1}^{N} \Delta T_i \overline{\delta}_i - \Delta \right)$$

bezüglich der gesuchten optimalen Lösung $\boldsymbol{\delta}^*$ und bei einem optimalen skalaren Wert λ^* für ein reguläres lokales Minimum verschwindet

$$\frac{\partial L}{\partial \overline{\delta}_i^*} = \Delta T_i \frac{\partial \phi_i}{\partial \overline{\delta}_i^*} + \lambda^* \Delta T_i = 0, \qquad i = 1..N.$$

Somit folgt λ^*

$$\lambda^* = -\frac{\partial \phi_i}{\partial \overline{\delta}_i^*}, \tag{13}$$

das der negativen Steigung der Kostenfunktionen an der Stelle der optimalen Lösung $\overline{\delta}_i^*$ entspricht. Treten nun zu den Zeitpunkten k und l gleiche Belastungen $P_{load,k} = P_{load,l}$ auf, dann sind auch die Kostenfunktionen $\phi_k = \phi_l$ gleich. Für die optimale Lösung muss (13) erfüllt sein, das bedeutet aber, dass die Ableitungen der Kostenfunktionen an der Stelle der optimale Lösung gleich sein muss

$$\frac{\partial \phi_k}{\partial \overline{\delta}_k^*} = \frac{\partial \phi_l}{\partial \overline{\delta}_l^*}$$

- Fall 1: Ist nun die Kostenfunktion **strikt konvex**, so kann diese Forderung aber nur im Punkt $\overline{\delta}_k^* = \overline{\delta}_l^*$ erfüllt sein, da die Steigung $-\lambda^*$ nur in einem Punkt auftritt (Abbildung 7).
- Fall 2: Bei einer **konvexen, aber nicht strikt konvexen Kostenfunktion** (Kostenfunktion mit stückweisen linearen Abschnitten) existieren entlang eines linearen Abschnitts der Kostenfunktion unendlich viele Stellen mit gleicher Steigung $-\lambda^*$.
- Fall 3: Hingegen können bei einer **nicht konvexen Kostenfunktion** mehrere Stellen der Kostenfunktion exisitieren, die die optimale Steigung $-\lambda^*$ besitzen (Abbildung 8),

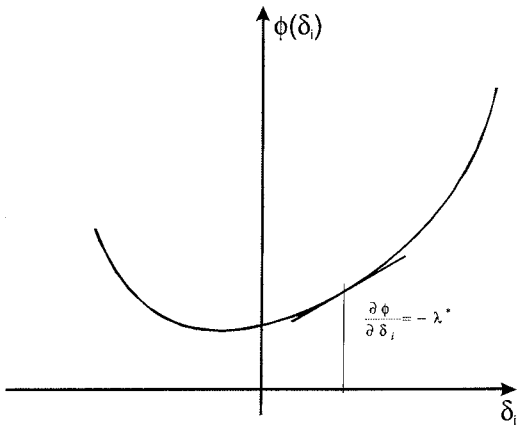


Abb. 7.

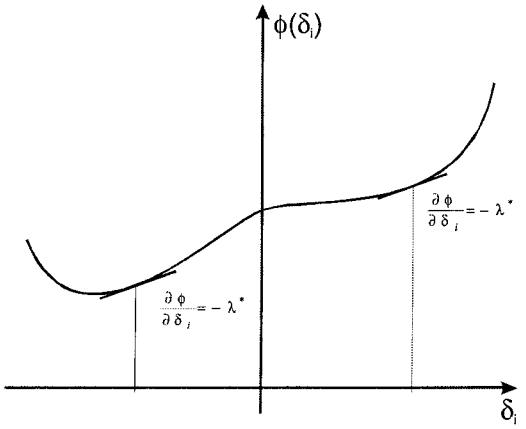


Abb. 8.

d.h. bei der optimalen Lösung kann der Fall $\overline{\delta}_k^* \neq \overline{\delta}_l^*$ auftreten und Kostenfunktion der Abschnitte k und l, ϕ_k und ϕ_l, dürfen somit nicht mehr zusammengefasst werden.

■

Da nun das optimale Schaltverhalten nur von der Last, nicht aber von der Zeit abhängt, ist es nicht nötig, den genauen Zeitverlauf der Last zu kennen. Aufgrund der konvexen Struktur des Optimierungsproblems ergibt sich für zwei gleiche Leistungen, die aber zu unterschiedlichen Zeiten auftretenden

$$P_{load,k} = P_{load,l}$$

die gleiche Lösung

$$\overline{\delta}_k^* = \overline{\delta}_l^*$$

Daher ist es nicht nötig, die (üblicherweise hohe) Dimension N des Optimierungsproblems zu betrachten, sondern gleiche Leistungen zusammenzufassen und somit nur mehr statistische Leistungsverteilung zur Optimierung zu verwenden.

4.2 Gesamtlösung

Um eine Lösung des Gesamtproblems zu erhalten, gilt es

$$\min_{\delta} \sum_{i=1}^{N} \Delta T_i \cdot \phi_i^*$$

unter der Nebenbedingung

$$\sum_{i=1}^{N} \Delta T_i \cdot \overline{\delta}_i = \Delta$$

zu erzielen. Da die Zielfunktion aus Summe von konvexen Funktionen wieder eine konvexe Funktion darstellt und eine lineare Nebenbedingung vorliegt, so ist das Gesamtproblem ebenfalls konvex.

5 Anwendungsbeispiel

Für die Energieversorgung existieren ein kalorisches Kraftwerk und als Speicherquelle ein Wasserkraftwerk (Beispiel aus [2]). Der Stundenverbrauch $\phi(P_k)$ des kalorischen Kraftwerks kann in Abhängigkeit seiner erzeugten Energie P_k in analytischer Form durch

$$\phi(P_k) = 500 + 10 \cdot P_k - 0.01 \cdot P_k^2 + 10^{-8} P_k^4$$

angegeben werden. Die Leistungsbereich des kalorischen Kraftwerks reicht von

$$100MW \leq P_k \leq 1000MW.$$

Der Leistungsbeitrag des Speicherkraftwerkes P_H ergibt sich (bei Vernachlässigung des Einflusses der Wasserhöhe x) in Abhängigkeit des Wasserflusses

$$q = \dot{x}$$

durch

$$P_H = \frac{1}{25} q - 10,$$

wobei sowohl die maximale als auch die minimale Wasserentnahmemenge mit

$$q \in [250, 22750] \frac{acre - ft}{h}$$

begrenzt ist[2]. Der Leistungsbereich des Wasserkraftwerkes ist durch

$$0\,MW \leq P_H \leq 900\,MW$$

[2] $1\ acre - ft$ ist eine im amerikanischen Raum gebräuchliche Einheit des Wasservolumens und ist direkt aus dem angegebenen Beispiel übernommen worden. $1\ acre - ft \triangleq 1233{,}5 m^3$

Abschnitt i	Zeitintervall	Last P_{load}
-	h	MW
1	$0-4$	800
2	$4-8$	1000
3	$8-12$	500

Tabelle 1
ZEITLICHER VERLAUF DER BELASTUNG

limitert. Die beiden Kraftwerke müsssen additiv ein vorgegebenes Lastprofil (Tabelle 1) liefern, das Wasserreservoir soll dabei beginnend mit einer Wassermenge von

$$x_0 = x(t=0) = 75 \cdot 10^3 acre - ft$$

am Ende vollständig entleert sein. Die beiden Leistungen sind jeweils in MW einzusetzten. Das geforderte Lastprofil ist vorab bekannt und besteht aus 3 Abschnitten mit einer Dauer von je $4h$. Unter diesen Rahmenbedingungen soll nun jener zeitlicher Verlauf der beiden Energiequellen gefunden werden, bei dem die Kosten

$$\int_{t=0}^{12h} \phi(t)dt$$

minimiert werden und dabei der Wasserspeicher vollständig entleert werden, d.h.

$$x_T = x(t=12h) = 0$$

5.1 Aufstellen der Gleichungen

Zu jedem Zeitpunkt muss die Leistung beider Energiequellen der geforderten Last entsprechen

$$P_{load}(t) = P_k(t) + P_H(t). \tag{14}$$

Diskretisiert man nun die Last in 3 Abschnitte (gekennzeichnet durch Index i) konstanten Leistungsbedarfes, so gilt weiter

$$P_{load,i} = P_{k,i} + P_{H,i}$$

und somit kann der Leistungsbedarf des kalorischen Kraftwerkes in Abhängigkeit des Wasserflusses beschrieben werden

$$\begin{aligned} P_{k,i} &= P_{load,i} - P_{H,i} \\ &= P_{load,i} - \frac{1}{25}q_i + 10. \end{aligned}$$

Dadurch folgt auch der Energieverbrauch $\phi_i(P_{load,i}, q)$ in Abhängigkeit des Wasserflussses und der Last

$$\phi_i = 600 + 10 \cdot P_{load,i} - 0.4 \cdot q - 0.01 \cdot (P_{load,i} - \frac{1}{25}q + 10)^2 + 10^{-8} \cdot (P_{load,i} - \frac{1}{25}q + 10)^4$$

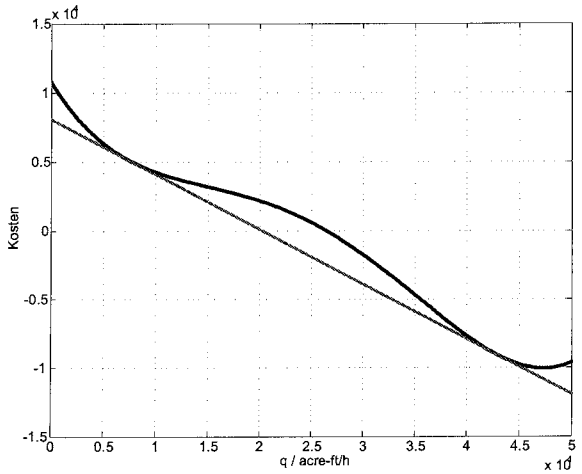


Abb. 9. Kostenfunktion im Abschnitt 2 in Abhängigkeit des Wasserflusses q

Die folgende Prozedur muss für jeden Abschnitt $i, i = 1..3$ durchgeführt werden. Als Beispiel wird hier der Abschnitt $i = 2$ mit $P_{load}(t) = 1000MW$, $t \in (4h, 8h)$ angeführt.

5.2 Unbeschränkte Formulierung

Der Verlauf des zeitlichen Energieverbrauches in Abhängigkeit des Wasserflusses ist in Abbildung 9 zu sehen.Unter der Annahme eines konstanten Lastbedarfs $P_{load,i}$ im Intervall $i = 2$ ergibt sich eine Kostenfunktion

$$\phi(x, \dot{x}, t) = 10600 - 0.4 \cdot \dot{x} - 0.01 \cdot (1010 - 0.04 \cdot \dot{x})^2 + 10^{-8} \cdot (1010 - 0.04 \cdot \dot{x})^4.$$

Der Anfangszustand der Wasserhöhe zum Zeitpunkt $t = 4h$ wird mit $x(4) = x_0$ festgelegt. Gesucht ist nun der optimale Verlauf von $x(t)$, bzw $\dot{x}(t) = q(t)$ im Intervall $i = 2$, so dass gilt

$$\begin{aligned} & \min_{x(t)} \int_{4h}^{8h} \phi(\dot{x}) dt \\ x(4h) \;=\; & x_0, \qquad x(8h) = x_0 + \delta_2, \end{aligned}$$

wobei δ_2 der Änderung der Wasserhöhe im zweiten Abschnitt zwischen $t = 4h$ und $t = 8h$ entspricht. Zum Finden des optimalen Verlaufs für ein gegebenes δ_2 muss vorerst die notwendige Euler-Lagrange Bedingung

$$\phi^*_{i,x} - \frac{d}{dt}\phi^*_{i,\dot{x}} = 0 \qquad \vee t \in [t_{i-1}, t_i], \tag{15}$$

erfüllt sein. Es gilt

$$\phi_{2,x} = 0$$

und somit

$$\frac{d}{dt}\phi_{2,\dot{x}} = 0, \tag{16}$$

das

$$\phi_{2,\dot{x}} \equiv c_1 \equiv konst. \qquad \forall t \in [4h, 8h]$$

impliziert. Die Weierstrass-Erdman Eckbedingungen

$$\phi^*_{i,\dot{x}}\big|_{t=\tau^-} = \phi^*_{i,\dot{x}}\big|_{t=\tau^+} \tag{17}$$

$$\left[\phi^*_i - \dot{x}^{*T}\phi^*_{i,\dot{x}}\right]_{t=\tau^-} = \left[\phi^*_i - \dot{x}^{*T}\phi^*_{i,\dot{x}}\right]_{t=\tau^+}, \tag{18}$$

garantieren, dass selbst bei einem Knick im Verlauf $x(t)$ eben diese Gleichungen einen kontinuierlichen Verlauf besitzen. Zur Bestimmung es optimalen Verlaufs werden nun 3 Fälle unterschieden. Unter Anwendung der Legendre-Bedingung

$$\phi_{i,\dot{x}\dot{x}}(x^*, \dot{x}^*, t) \geq 0 \qquad \vee\, t \in [t_{i-1}, t_i]. \tag{19}$$

ergibt sich ein erlaubter Bereich für $\dot{x}(t)$ mit

$$\dot{x}^2(t) - 50500 \cdot \dot{x}(t) + 5.334 \cdot 10^8 \geq 0 \tag{20}$$

und somit

$$\begin{aligned} \dot{x}(t) &\geq 35456.2 \text{ oder} \\ \dot{x}(t) &\leq 15043.8 \end{aligned}$$

1. Im ersten Fall wird willkürlich angenommen, dass $c_1 = -0.4$ ist und $x(t)$ keinen Knick im Verlauf besitzt. Es gilt

$$\phi_{2,\dot{x}} = c_1 = -0.4$$

$$0.408 - 3.2 \cdot 10^{-5}\dot{x}(t) - 1.6 \cdot 10^{-9} \cdot (1010 - 0.04 \cdot \dot{x}(t))^3 = -0.4$$

Dies ist nur bei $\dot{x}(t) = 7572.3$, $\dot{x}(t) = 25250$ oder $\dot{x}(t) = 42927.7$ erfüllt. Mit der notwendigen Bedingung (20) fällt auch die zweite Lösung weg. Daraus folgt ein optimaler Verlauf $x(t) = x_0 + \frac{\delta_2}{4h}t$, wobei $\delta_2 = 7572.3 \cdot 4h$ oder $\delta_2 = 42927.7 \cdot 4h$ sein kann.
2. Im zweiten Fall wird neben $c_1 = -0.4$ noch angenommen,dass $x(t)$ mindestens einen Knick besitzt. Durch Anwendung der notwendigen Bedingungen (17) und (18) sieht man, dass die optimale Lösung nur zwischen $\dot{x}(t) = 7572.3$ und $\dot{x}(t) = 42927.7$ springen kann, die Lösung $\dot{x}(t) = 25250$ ist wieder aufgrund der Legendre-Bedingung nicht optimal. Aufgrund der Sprünge kann jeder Wert δ_2 zwischen

$$7572.3 \cdot 4h < \delta_2 < 42927.7 \cdot 4h$$

erreicht werden, wobei für jedes δ_2 durch eine unendliche Anzahl von verschiedenen Lösungen realisiert werden kann.
3. Im dritten Fall wird $c_1 \neq -0.4$ angenommen, und man kann zeigen, dass die optimale Lösung keine Sprünge enthält: Im optimalen Fall ist ein kontinuierlicher Verlauf von Gleichung (17) und (18) gefordert. Somit folgt auch, dass

$$\phi_{2,\dot{x}}(t, x, \dot{x}(t+0)) = \phi_{2,\dot{x}}(t, x, \dot{x}(t-0)) = c_1 \neq -0.4$$

und unter Verwendung der Abkürzungen

$$\begin{aligned} \dot{x}(t+0) &= \dot{x}_+ \\ \dot{x}(t-0) &= \dot{x}_- \end{aligned}$$

ergibt sich

$$\begin{aligned} (\dot{x}(t+0) - \dot{x}(t-0))\phi_{2,\dot{x}} &= \phi_2(t,x,\dot{x}(t+0)) - \phi_2(t,x,\dot{x}(t-0)) \\ (\dot{x}_+ - \dot{x}_-)c_1 &= \phi_2(t,x,\dot{x}_+) - \phi_2(t,x,\dot{x}_-) \end{aligned}$$

Beweis. Nimmt man nun an, dass die optimale Lösung doch einen Sprung von $\dot{x}(t)$ bei $c_1 \neq -0.4$ besitzt, somit

$$\dot{x}_+ \neq \dot{x}_- \tag{21}$$

ist, ergibt sich

$$c_1 = \frac{\phi_2(t,x,\dot{x}_+) - \phi_2(t,x,\dot{x}_-)}{(\dot{x}_+ - \dot{x}_-)} \tag{22}$$

Der Wert c_1 entspricht der Steigung der Kostenfunktion an der Sprungstelle $\phi_{2,\dot{x}}$ und darf nach Glg. (16) zeitlich nicht veränderlich sein

$$c_1 = \phi_{2,\dot{x}}(t,x,\dot{x}_+) = \phi_{2,\dot{x}}(t,x,\dot{x}_-).$$

Setzt man nun zur Berechnung von c_1 in die Ableitung der analytischen Kostenfunktion ein

$$\begin{aligned} c_1 &= 0.408 - 3.2 \cdot 10^{-5}\dot{x}_+ - 1.6 \cdot 10^{-9} \cdot (1010 - 0.04 \cdot \dot{x}_+)^3 = \\ &= 0.408 - 3.2 \cdot 10^{-5}\dot{x}_- - 1.6 \cdot 10^{-9} \cdot (1010 - 0.04 \cdot \dot{x}_-)^3. \end{aligned}$$

Diese kubische Gleichung ergibt 3 Lösungen für $\dot{x}_+$

$$\begin{aligned} \dot{x}_{+,1} &= \dot{x}_- \\ \dot{x}_{+,2/3} &= 37875 - \frac{\dot{x}_-}{2} \pm 0.5 \cdot \sqrt{-3 \cdot \dot{x}_-^2 + 151500 \cdot \dot{x}_- - 662687500} \end{aligned} \tag{23}$$

Die 1. Lösung wurde bereits durch Gleichung (21) ausgeschlossen, die beiden restlichen Lösungen liefern nur im Bereich

$$4837.59 \leq \dot{x}_- \leq 45662.41$$

reelle Lösungswerte. Durch Einsetzen der Lösung von (23) mit dem größeren Maximalwert ist Gleichung (22) in den Punkten $\dot{x}_+ = 7572.33$ und $\dot{x}_+ = 35456.21$ erfüllt. Im erstgenannten Punkt ist jedoch die Steigung $c_1 = -0.4$, welche in dieser Fallunterscheidung ausgenommen wurde, im letztgenannten Punkt ergibt sich $\dot{x}_- = \dot{x}_+ = 35456.21$ und somit ist eine der beide Annahmen ($c_1 \neq -0.4$ bzw. $\dot{x}_- \neq \dot{x}_+$) in dieser Fallunterscheidung verletzt. Gleiches gilt für die Lösung mit dem kleineren Maximalwert aus 23. ■

5.3 Berücksichtigung von Beschränkungen

Die Lösung des vorherigen Abschnitts berücksichtigt nur unbeschränkte Optimierungsprobleme. Bei diesem Beispiel treten aber aufgrund des physikalischen Hintergrunds sowohl Limitierungen des kalorischen als auch des Wasserkrafwerkes auf. Diese Beschränkungen führen dann zu den Ungleichungsnebenbedingungen

$$\begin{aligned} h_1 &= 250 - q \leq 0 \\ h_2 &= q - 22750 \leq 0, \end{aligned}$$

die in der erweiterten Hamilton-Funktion, die die UNB einschließt,

$$\tilde{H}(x(t), u(t), \lambda(t), t) = \phi_i(x(t), u(t), t) + \lambda(t) f(x(t), u(t), t) + \boldsymbol{\mu}(\mathrm{t})\mathbf{h}(\mathrm{x,u,t}). \tag{24}$$

berücksichtigt werden können

$$\tilde{H} = \phi(q) + \lambda q + \mu_1 h_1 + \mu_2 h_2$$

mit

$$\dot{x} = q.$$

Aus der notwendigen Bedingung $\tilde{H}_x = -\dot{\lambda}$ folgt

$$\lambda(t) = const. \qquad \forall t \in [4h, 8h]$$

Fallunterscheidung

1. Unter der Annahme, dass der unterer Randwert $q = 250$ ein optimaler Wert ist, kann folgendes Gleichungssystem aufgestellt werden

$$\begin{aligned} H(q &= 250) = \phi(250) + \lambda \cdot 250 = H(q = q_1) = \phi(q_1) + \lambda \cdot q_1 \\ H_q(q &= 250) = 0 = \phi_q(250) + \lambda - \mu_1 \\ H_q(q &= q_1) = 0 = \phi_q(q_1) + \lambda, \end{aligned}$$

wobei q_1 ein weiterer optimaler Steuerwert ist. Es zeigt sich, dass dieses Gleichungssystem unter der Annahme $\mu_1 > 0$ nur bei $q_1 = 250$ erfüllt ist und somit dieser Randwert nicht Teil einer Lösung ist, bei der zwischen 2 Werten gesprungen wird.
2. Unter der Annahme, dass der obere Randwert $q = 22750$ ein optimaler Wert ist, kann folgendes Gleichungssystem aufgestellt werden

$$\begin{aligned} H(q &= 22750) = \phi(22750) + \lambda \cdot 22750 = H(q = q_1) = \phi(q_1) + \lambda \cdot q_1 \\ H_q(q &= 22750) = 0 = \phi_q(22750) + \lambda + \mu_2 \\ H_q(q &= q_1) = 0 = \phi_q(q_1) + \lambda. \end{aligned}$$

Dieses Gleichungssystem liefert mit $\mu_2 > 0$ eine Lösung für $q_1 = 11698$. Somit besteht die optimale Lösung für q im Bereich

$$11698 < q < 22750 \tag{25}$$

im Springen zwischen dieses Werten.
3. Unter der Annahme, dass kein Randwert beim optimalen Verlauf mitberücksichtigt ist ($\mu_1 = \mu_2 = 0$) und unter der Annahme, dass zwischen 2 Werten q_1 und q_2 gesprungen wird und bei Ausnutzung der Eigenschaft der (erweiterten) Hamilton-Funktion ist, dass für die optimale Lösung

$$\tilde{H}(x(t), u(t), \lambda(t), \boldsymbol{\mu}(\mathrm{t}), \mathrm{t}) = \text{konst.} \quad \forall \mathbf{t} \in [\mathbf{t_i}, \mathbf{t_{i+1}}] \tag{26}$$

gilt, wenn in (24) die Funktionen ϕ_i, f und $\mathbf{h}$ nicht explizit zeitabhängig sind, kann ein Gleichungssystem aufgestellt werden

$$\begin{aligned} H(q &= q_1) = \phi(q_1) + \lambda \cdot q_1 = H(q = q_2) = \phi(q_2) + \lambda \cdot q_2 \\ H_q(q &= q_1) = 0 = \phi_q(q_1) + \lambda \\ H_q(q &= q_2) = 0 = \phi_q(q_2) + \lambda. \end{aligned}$$

Dieses Gleichungssystem liefert wieder nur Lösungen $q_1 = q_2$ und somit besteht der optimale Verlauf von q ausserhalb des Bereiches von Gleichung (25) nur in einem konstanten Wert

$$q(t) = const. \qquad \forall t \in [4h, 8h]$$

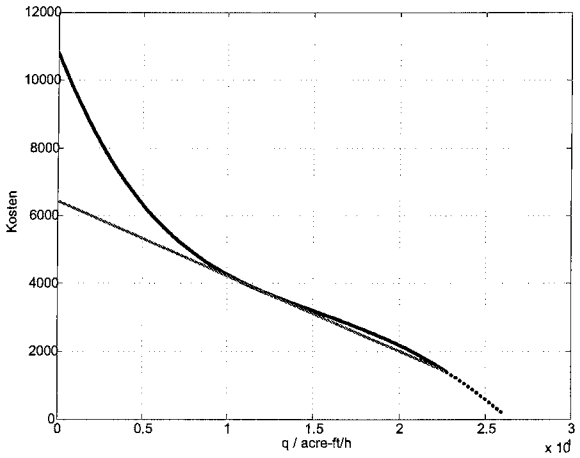


Abb. 10. Kostenfunktion im Abschnitt 2 in Abhängigkeit des Wasserflusses q unter Berücksichtigung von Beschränkungen

5.4 Zusammenfassen aller 3 Abschnitte

Die im vorherigen Kapitel durchgeführte Prozedur entspricht einer Konvexifizierung der Kostenfunktionen in den einzelnen Abschnitten und erleichtert die Lösungsfindung der Optimierungsaufgabe. Die Optimierungsaufgabe gerät dadurch zu einer Suche von

$$\min \sum_{i=1}^{3} \phi_i(\delta_i),$$

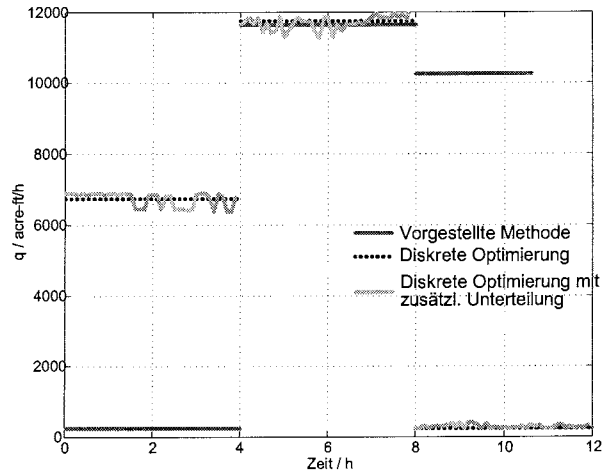


Abb. 11. Zeitliche Wasserentnahme während des Optimierungszeitraums

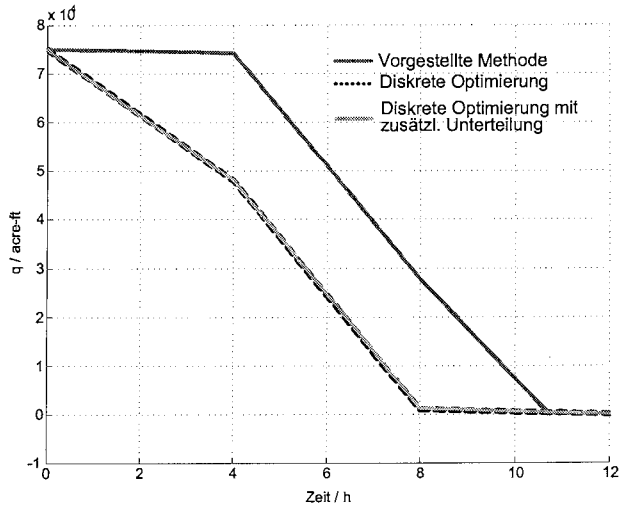


Abb. 12. Verlauf des Wasserpegels innerhalb des Optimierungszeitraums

so dass

$$\sum_{i=1}^{3} \delta_i \cdot 4h = 75 \cdot 10^3 acre - ft$$

und unter den in der Einführung beschriebenen Beschränkungen. Die optimale Lösung der konvexifizierten Aufgabe ergibt sich (durch Lösung mit MATLAB - fmincon- Funktion). Verglichen wird dabei eine Lösung mit einer feineren Diskretisierung (40 Intervalle je konstanter Last = 6 min Subintervalldauer). Tabelle 2 zeigt die Ergebnisse von verschiedenen Optimierungsvarianten dieses Beispiels.

	Optimierte Kosten	Rechenzeit	Optimierungs-parameter
Vorgestellte Methode	45143	$128ms$	3
Diskrete Optimierung	45174	$78ms$	3
Diskrete Optimierung mit zusätzlicher Unterteilung	45175	$45s$	120
ohne Optimierung konstanter Wasserfluss	48309	–	–

Tabelle 2
ERGEBNIS DER OPTIMIERUNG

6 ZUSAMMENFASSUNG UND AUSBLICK

In dieser Arbeit wurde eine Methode vorgestellt, die es erlaubt, ein nichtkonvexes Optimierungsproblem, das eventuell in nicht-analytischer Form vorliegt, durch entsprechende Diskretisierung auf ein konvexes Problem überzuführen. Diese Vorgehensweise wurde sowohl bei der Verbrauchsminimierung von Serienybridfahrzeugen [9] als auch bei Fahrzeugen mit einem Kurbelwellen-Startergenerator [10] bereits eingesetzt, wodurch sich ein Verbrauchsreduktionspotential bei einem Serienmotor eines BMW 320d von bis zu 5% bei einem Normfahrzyklus ergibt.

Die weitere Arbeit betreffen Erweiterungen in der Optimierung, bei denen auch der Zustand des Zwischenspeichers in die Kostenfunktion eingehen kann sowie Berücksichtigungen von nicht idealen Speichermodellen.

LITERATURVERZEICHNIS

[1] Smith-L. Gawthrop-G., "Metamodelling: Bond graphs and dynamic systems", 1996.

[2] Wollenberg-B.F. Wood-A.J., *Power Generation, Operation and Control, Second Edition*, John Wiley and Sons, Inc., 1996.

[3] Shaoyun-G; Chung-TS, "Optimal active power flow incorporating facts devices with power flow control constraints", *International-Journal-of-Electrical-Power-and-Energy-Systems. vol.20, no.5; June 1998; p.321-6*, 1998.

[4] Chun-Lung-Chen; Nanming-Chen, "Direct search method for solving economic dispatch problem considering transmission capacity constraints", *IEEE-Transactions-on-Power-Systems. vol.16, no.4; Nov. 2001; p.764-9*, 2001.

[5] Han-XS; Gooi-HB; Kirschen-DS, "Dynamic economic dispatch: feasible and optimal solutions", *IEEE-Transactions-on-Power-Systems. vol.16, no.1; Feb. 2001; p.22-8*, 2001.

[6] Momoh-J.-A.; El-Hawary-M.-E.; Adapa-R., "A review of selected optimal power flow literature to 1993. i. nonlinear and quadratic programming approaches", *IEEE-Transactions-on-Power-Systems. Feb. 1999; 14(1): 96-104*, 1999.

[7] Momoh-J-A; El-Hawary-M-E; Adapa-R., "A review of selected optimal power flow literature to 1993. ii. newton, linear programming and interior point methods", *IEEE-Transactions-on-Power-Systems. Feb. 1999; 14(1): 105-11*, 1999.

[8] Steinmaurer-G; Del-Re-L, "Optimal control of dual power sources", *Proceedings of the 2001 IEEE International Conference on Control Applications, Mexico City*, pp. 422–7, 2001.

[9] Steinmaurer-G; Del-Re-L, "Development and use of a series hybrid vehicle control strategy for the ADVISOR simulation tool", *Proceedings of the Advisor User Conference*, 2001.

[10] Steinmaurer-G; Del-Re-L, *Downsizing Opportunities for Drivelines with 42V Starter/Alternator, Tagungsband Haus der Technik, '42V PowerNet: Stepping Into Mass Production*, chapter 16, pp. 150–167, expert-Verlag, 2002.

Robustes Backstepping für den Nichtlinearen Entwurf

Alexander Viehweider
Institut für Automatisierungs- und Regelungstechnik
Technische Universität Wien
Gusshausstrasse 27-29/E376
1040 Wien
viehweider@acin.tuwien.ac.at

Kurzfassung: *Der vorliegende Beitrag zeigt, wie anhand der Backstepping-Methodik durch geeignete Erweiterung robuste Verfahren zur Stabilisierung und Folgeregelung nichtlinearer Systeme entworfen werden können. Insbesondere wird auf die Methodik der Nichtlinearen Dämpfung, der robusten Erweiterung durch weich schaltende Anteile und der Einbindung von Integralanteilen in den Backstepping-Regler zur Erhöhung der Robustheit eingegangen. Durch Kombination der Verfahren und gegebenenfalls durch geeignete Erweiterungen lassen sich Verbesserungen erzielen. Als Anwendungsbeispiel dient eine magnetische Schweberegelung. Einige Aspekte wie etwa die Empfindlichkeit der Nichtlinearen Dämpfung gegenüber Messrauschen und Methoden zur Verbesserung der Stelldynamik werden näher untersucht.*

1 Einleitung

Die Regelung nichtlinearer Systeme stellt für den Entwurfsingenieur eine große Herausforderung dar. Zum einen können nichtlineare Systeme nicht allgemein einer einheitlichen, geschlossenen Systembetrachtung zugeführt werden (Die Betrachtung erfolgt vielmehr systemklassenbezogen. Diese Systemklassen definieren sich z.B. durch Topologie des Systems oder Eigenschaften der im System auftretenden Nichtlinearitäten.), zum anderen können Phänomene auftreten, die von der linearen Systemtheorie her unbekannt sind.
In Abhängigkeit von der Systemklasse haben sich daher unterschiedliche Verfahren herausgebildet, so unter anderem das Backstepping-Verfahren zur Stabilisierung und Folgeregelung von nichtlinearen Systemen in sog. unterer Dreiecksform.
Die große Bandbreite des durch diese Methodik abgedeckten Spektrums von realen Anwendungen (Regelung von Schiffen, Flugkörpern, Weltraumsonden, Gleichstrom- und Asynchronmotoren, biochemischer Prozesse, ...) ergibt sich durch die Tatsache, dass durch Abwandlungen, Ergänzungen und durch ein dem Backstepping dualen Verfahren („Forwarding", [SeJa97]) die Systemklasse bedeutend erweitert werden kann, dass viele elektromechanische Systeme untere Dreiecksform oder annähernd untere Dreiecksform aufweisen und nicht zuletzt durch die dem Ansatz inherente Designflexibilität, was heißen soll, das durch den Ansatz in seiner allgemeinsten Form kein Regelgesetz, sondern eine Klasse von Regelgesetzen geschaffen wird, die sich in unterschiedlichsten Eigenschaften wie z.B. Robustheit und Stellgrößenbedarf [ZhKa00] unterscheiden können.

Inhalt dieses Beitrages ist die Robustheit des sich ergebenden Regelgesetzes und geeignete Abwandlungen des Regelgesetzes zur Erhöhung der Robustheit. Die Robustheit wird gegenüber an unterschiedlicher Stelle des Systems eingespeisten Störsignalen verstanden. Modellunsicherheiten können nach geeigneter Interpretation mitbetrachtet werden.

Abschnitt 2 stellt das Backstepping-Verfahren vor, wobei die Darstellung auf die in diesem Beitrag benötigten notwendigsten Grundlagen beschränkt ist. Eine ausführlichere Einführung kann in [SeJa97], [KrKa95] oder [Vieh03a] gefunden werden. In Abschnitt 3 werden die verschiedenen Möglichkeiten zur robusten Erweiterung des Verfahrens unter einheitlichen Gesichtspunkten vorgestellt. Abschnitt 4 erweitert die Betrachtung aus Abschnitt 3 hinsichtlich der Aspekte Empfindlichkeit gegenüber Meßrauschen, Kombination der Verfahren und Verbesserung der Stelldynamik. Die Freiheitsgrade des Entwurfs werden durch ein abschließendes Anwendungsbeispiel einer Magnetschweberegelung in Abschnitt 5 aufgezeigt.

2 Backstepping als Entwurfsmethodik

Backstepping ist ein rekursives Lyapunov-basiertes Entwurfsverfahren. Ausgangspunkt sind *„strict feedback"*- Systeme in unterer Dreiecksform (Siehe Anhang 1).

In jedem Schritt des Verfahrens wird für ein i-tes Teilsystem ein sogenanntes virtuelles Regelgesetz α_i entworfen. Würde der virtuelle Eingang des Teilsystems diesem Regelgesetz entsprechen, so wäre die Stabilisierung des Teilsystems garantiert. An folgendem einfachen SISO-System in unterer Dreiecksform ohne Unsicherheiten soll die Vorgehensweise demonstriert werden:

$$\begin{aligned} \dot{x}_1 &= f_1(x_1) + x_2 \\ \dot{x}_2 &= f_2(x_1.x_2) + x_3 \\ &\vdots \\ \dot{x}_n &= f_n(x_1, \cdots, x_n) + u \\ y &= x_1 \quad . \end{aligned} \tag{1}$$

Die erste Koordinatentransformation beträgt im Falle einer Stabilisierung $z_1 = x_1$ bzw. im Falle einer Folgeregelung $z_1 = x_1 - x_S$, wobei x_S den Sollwert darstellt. Die Koordinaten des Backstepping-Reglers werden in diesem Beitrag mit z_i bezeichnet wie in der Literatur häufig der Fall. Das im ersten Schritt zu betrachtetende Teilsystem ergibt sich aus der ersten Zeile von Gl.(1). x_2 wird als virtueller Eingang des Teilsystems bezeichnet. Virtuell, da er nicht unmittelbar durch den Systemeingang u festgelegt werden kann. α_1 bezeichnet ein virtuelles Regelgesetz, welches z.B. in seiner einfachsten Form folgendermaßen lauten kann:

$$\alpha_1 = \dot{x}_S - f_1(x_1) - c_1 z_1, \quad c_1 > 0 \quad . \tag{2}$$

Die z_2-Koordinate bezeichnet die Abweichung des virtuellen Eingangs vom virtuellen Regelgesetz, also $z_2 = x_2 - \alpha_1$. Damit ergibt sich

$$\dot{z}_1 = -c_1 z_1 + z_2 \ . \tag{3}$$

Nun wird das Teilsystem betrachtet, dass sich aus den beiden ersten Zeilen von Gl.(1) ergibt. x_3 stellt nun den virtuellen Eingang dar. Wird die z_2-Koordinate nach der Zeit abgeleitet, so ergibt sich

$$\begin{aligned} \dot{z}_2 &= \dot{x}_2 - \dot{\alpha}_1 = f_2(x_1, x_2) + x_3 - \dot{\alpha}_1 = f_2(x_1, x_2) + x_3 - (\ddot{x}_S - \frac{\partial f_1(x_1)}{\partial x_1}\dot{x}_1 - c_1 \dot{z}_1) = \\ &= f_2(x_1, x_2) + x_3 - (\ddot{x}_S - \frac{\partial f_1(x_1)}{\partial x_1}(f_1(x_1) + x_2) - c_1(-c_1 z_1 + z_2)) \ . \end{aligned} \tag{4}$$

Das virtuelle Regelgesetz α_2 kann folgendermaßen festgelegt werden

$$\alpha_2 = -c_2 z_2 - z_1 + \dot{\alpha}_1 = -c_2 z_2 - z_1 - f_2(x_1, x_2) + (\ddot{x}_S - \frac{\partial f_1(x_1)}{\partial x_1}(f_1(x_1) + x_2) - c_1(-c_1 z_1 + z_2)), \ c_2 > 0 \ . \tag{5}$$

Zur Berechnung des virtuellen Regelgesetzes α_2 wird die Ableitung des vorhergehenden Regelgesetzes $\dot{\alpha}_1$ benötigt. Dies ist für den Backstepping-Entwurf charakteristisch. Dies führt zu einer hohen Anzahl von Termen, insbesondere bei Strecken höherer Ordnung. Die aufgezeigte Vorgehensweise wird für ein System n-ter Ordnung $n-1$ mal angewandt. Im letzten Entwurfschritt ergibt sich das eigentliche Regelgesetz. Die $\mathbf{z}$-Koordinaten des Backsteppingreglers bezeichnen jeweils die Abweichung der virtuellen Eingänge von den virtuellen Regelgesetzen. Wird nach folgender Methodik fortgefahren, so ergibt sich für das Gesamtsystem Backstepping-Regler und Strecke ohne Unsicherheiten folgendes lineares Zustandsgleichungssystem

$$\dot{\mathbf{z}} = \begin{bmatrix} -c_1 & 1 & 0 & \cdots & 0 \\ -1 & -c_2 & \ddots & \ddots & \vdots \\ 0 & \ddots & -c_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & -1 & -c_n \end{bmatrix} \mathbf{z} \ . \tag{6}$$

Die Stabilität dieses System kann über die Lyapunovfunktion $V = \frac{1}{2}\mathbf{z}^T\mathbf{z}$ nachgewiesen werden.

Es muß hierbei erwähnt werden, dass die Wahl der virtuellen Regelgesetze einen Freiheitsgrad des Verfahrens darstellen. In dieser Herleitung wurde das „Exakte" Backstepping verwendet. Hierbei können unter Umständen „nützliche" Nichtlinearitäten des Systems exakt kompensiert werden. Auch die Lyapunovfunktion für das Gesamtsystem ergibt sich rekursiv durch die Erweiterung der Lyapunovfunktionen für die Teilsysteme. Unterschiedliche Gestaltungen derselben können zu unterschiedlichen Eigenschaften des Regelgesetzes führen (Stellgrößenaufwand, Robustheit, ...) (Siehe z.B. [ZhKa98]).

Die Parameter c_i stellen die Backstepping-Reglerparameter dar. Sie beeinflußen Stellgrößenverlauf, Robustheit und evtl. Störgrößenunterdrückung. Deren optimale Festlegung ist nicht Gegenstand dieses Beitrages. Untersuchungen und Verfahren hierzu können z.B. in [Vieh03c] nachgelesen werden.

3 Das Robuste Backstepping-Design

Die Berücksichtigung von Unsicherheiten im Entwurf erfordert in der Regel besondere Parameterwahl [ZhKa98] oder kann durch Abwandlung und geeignete Ergänzung des Regelgesetzes erfolgen. Es sei folgendes einfaches System 1.Ordnung betrachtet

$$\begin{aligned} \dot{x} &= u + \tau(x)\Delta(t) \\ y &= x \ , \end{aligned} \tag{7}$$

wobei x den Zustand und u den Systemeingang darstellen. $\tau(x)$ ist eine bekannte Funktion und $\Delta(t)$ ein unbekanntes Störsignal. Die Unsicherheit ist vom Typ „matched uncertainty", was heißen soll, dass die Unsicherheit und die Stellgröße „gleichen Eingriff"[1] in das System haben.
Da Backstepping ein Regelgesetz zur Stabilisierung eines Teilsystems bereits voraussetzt, soll das Beispielsystem aus Gl.(7) nur als Ausgangspunkt zur Verdeutlichung der robusten Erweiterung des Regelgesetzes dienen. Das System nach Gl.(7) wird in einem weiteren Schritt um einen Integrator ergänzt und auf das Gesamtsystem dann das Integrator-Backstepping Verfahren angewandt.
Zur Stabilisierung des Systems nach Gl.(7) wird die Lyapunovfunktion $V = \frac{1}{2}x^2$ verwandt. Für $V > 0$ mit $\dot{V} < 0$ für $x \neq 0$ist global asymptotische Stabilität des Systems gegeben. Ist eine obere Grenze des Ausdrucks $\tau(x)\Delta(t)$ bekannt, also $|\tau(x)\Delta(t)| \leq \chi(x)$, so ergibt sich durch ein Regelgesetz der Form (Robuste Erweiterung durch weichschaltende Anteile - REWA)

$$\text{REWA: } u = -cx - sign(x)\chi(x), \tag{8}$$

die Ableitung der Lyapunovfunktion zu

$$\dot{V} = -cx^2 - xsign(x)\chi(x) + x\tau(x)\Delta(t)). \tag{9}$$

Da der zweite Term $xsign(x)\chi(x)$ den dritten mit Unsicherheit behafteten Term $x\tau(x)\Delta(t)$ dominiert, ist die Funktion $\dot{V}$ negativ definit.
Es sei hier hervorgegriffen, dass bedingt durch die Methodik des Backsteppings keine in den Zustandsgrößen unstetigen Funktionen in den einzelnen virtuellen Regelgesetzen auftreten dürfen, da in jedem Rekursionsschritt die zeitiche Ableitung des Regelgesetzes des vorhergehenden Schrittes benötigt wird. Es wird hierzu die „harte" signum-Funktion durch

[1] Zur allgemeinen koordinatenunabhängigen Charakterisierung von „matched uncertainties" sei auf [MaTo96] S.96 verwiesen.

die „weiche" Sättigungsfunktionen $sign_n(x,\varpi)$ ersetzt, welche folgende Eigenschaften aufweisen: $sign_n(x,\varpi)$ ist eine monoton steigende Funktion mit $sign_n(0,\varpi)=0$, $\lim_{x\to+/-\infty} sign_n(x,\varpi)=+/-1$ und $sign_n(x,\varpi)$ ist n–fach ableitbar. ϖ ist ein Formparameter, der den Übergang von der negativen in die positive Sättigung beschreibt.
Für gewisse Werte des Formparameters ϖ kann die sich dadurch ergebende Lyapunovfunktion $V=\frac{1}{2}x^2$ Gebiete in x – in Abhängigkeit von der Unsicherheit – aufweisen, in welchen die Ableitung der Lyapunovfunktion $\dot{V}$ positiv wird. In Abb.1 wird hierzu die Unsicherheit mit $\tau(x)=x^8$ und $|\Delta(t)|\le 1$ angegeben. Damit gilt $\chi(x)=x^8$. Betrachtet wird die Ableitung der Lyapunovfunktion bei unterschiedlichen „Schaltfunktionen", zum einen bei der hartschaltenden Funktion $sign(x)$ und zum anderen bei der weichschaltenden Funktion $\tanh(1.4x)$.
Bei der weichschaltenden Funktion ergeben sich in Abhängigkeit vom Formparameter Gebiete, in welchen die Ableitung der Lyapunovfunktion positiv wird. Da außerhalb dieser Gebiete die negative Definitheit der Ableitung der Lyapunovfunktion jedoch erfüllt ist, findet eine Konvergenz gegen die – im allgemeinen Fall zeitvarianten – Gebietsgrenzen statt.

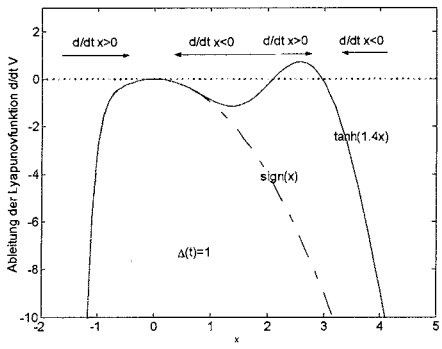


Abbildung 1: Ableitung der Lyapunovfunktion In Abhängigkeit vom Formparameter

Diese Form der Stabilität, bei dem sich der Zustandsvektor nach einer gewissen Systemzeit innerhalb eines Gebietes bewegt, wird als *Ultimate Bounded Stability* bezeichnet. Der Formparameter beeinflußt die Stelldynamik wesentlich, hohe Werte deselben führen zwar zu guten Konvergenzeigenschaften, jedoch unnötig hoher Stelldynamik. Es muß daher der für die jeweilige Anwendung geeignete Formparameter gefunden werden. Über den Einfluß der Form der Lyapunovfunktion auf die Stelldynamik wird noch in Abschnitt 4 gesprochen.

Eine andere Möglichkeit zur Beherrschung der Unsicherheit bietet die Methodik der Nichtlinearen Dämpfung: Die Nichtlineare Dämpfung geht von der Beschränkung der Unsicherheit $\Delta(t)$ in Gl.(7) aus. Häufig kann eine Beschränkung der Form $\|\Delta(t)\|_\infty \le \Delta_{max}$ für die Unsicherheit aus Gl.(7) angegeben werden, weiters kann $\tau(x)$ als bekannt vorausgesetzt werden.
Wird ein Regelgesetz der Form (Nichtlineare Dämpfung - NLD)

$$\text{NLD: } u=-cx-d(x)x=-cx-\kappa\tau^2(x)x\ ,\ \kappa\in\Re^+ \tag{10}$$

verwandt, wobei κ den Verstärkungsfaktor der Nichtlinearen Dämpfung bezeichnet, so ergibt sich die Ableitung der Lyapunovfunktion $V=\frac{1}{2}x^2$ zu

$$\dot{V}=-cx^2-\kappa x^2\tau(x)^2+x\tau(x)\Delta(t)=-cx^2-\kappa(x\tau(x)-\frac{\Delta(t)}{2\kappa})^2+\frac{\Delta^2(t)}{4\kappa}. \tag{11}$$

Aus der quadratischen Ergänzung in Gl.(11) oder durch Anwendung der Young`schen Ungleichung (Siehe Anhang A2) ergibt sich die Ungleichung

$$\dot{V} \leq -cx^2 + \frac{\Delta^2(t)}{4\kappa}. \tag{12}$$

Aus Gl.(12) ist ersichtlich, dass im Gebiet $|x| \geq \frac{\|\Delta\|_\infty}{2\sqrt{c\kappa}}$ die Ableitung der Lyapunovfunktion sicher negativ ist. Der Verstärkungsfaktor κ der nichtlinearen Dämpfung schränkt also das Gebiet, in welchem der Zustandsvektor nach einer gewissen Zeit zu liegen kommt, ein. Große Werte führen zu kleinerer Regelabweichung, können sich aber z.B. ungünstig auf das Rauschverhalten des Reglers auswirken. Es wurde bis jetzt nur eine Fixpunktregelung mit dem Fixpunkt $x = 0$ betrachet, die Erweiterung auf eine Folgeregelung kann sehr einfach durch die Fehlervariable $z_1 = x - x_S$ erfolgen, hierbei ergibt sich als Regelgesetz für die Nichtlineare Dämpfung

$$u = -cz_1 - z_1 \kappa \tau^2(x) = -cz_1 - z_1 d(x) \tag{13}$$

mit zugehöriger Lyapunovfunktion $\frac{1}{2} z_1^2$.

Die beiden Regelgesetze (Robuste Erweiterung mit weichschaltenden Anteilen und die Nichtlineare Dämpfung) sollen als Ausgangspunkt für einen Backstepping-Schritt dienen. Hierzu wird das System nach Gl.(7) um einen Integrator erweitert:

$$\begin{aligned} \dot{x}_1 &= x_2 + \tau(x_1)\Delta(t) \\ \dot{x}_2 &= u \\ y &= x_1 \quad . \end{aligned} \tag{14}$$

Mit der Definition $z_1 = x_1 - x_S$ wird im ersten Schritt des Verfahrens ein geeignetes virtuelles Regelgesetz für das Teilsystem entworfen, das sich aus der ersten Zeile von Gl.(14) ergibt. Die zeitliche Ableitung von z_1 führt zu

$$\dot{z}_1 = \dot{x}_1 - \dot{x}_S = x_2 + \tau(x_1)\Delta(t) - \dot{x}_S . \tag{15}$$

Es wird nun die Koordinate $z_2 = x_2 - \alpha_1$ eingeführt, die die Abweichung des virtuellen Eingangs x_2 vom virtuellen Regelgesetz α_1 darstellt. α_1 wird in Anlehnung an Gl.(8) für den REWA-Fall folgendermaßen gewählt

$$\alpha_1 = \dot{x}_S - c_1 z_1 - sign_1(z_1, \varpi_1)\chi(x_1) . \tag{16}$$

Der zweite Schritt erfordert die „Rückführung des Regelgesetzes hinter den Integrator". Hierzu werden die Ableitungen der entsprechenden Terme benötigt. Formal geschieht dies durch Ableiten von z_2:

$$\dot{z}_2 = \dot{x}_2 - \dot{\alpha}_1 = u - (\ddot{x}_S - c_1\dot{z}_1 - \frac{d}{dt} sign_1(z_1, \varpi_1)\chi(x_1)) = \tag{17}$$

$$= u - (\ddot{x}_S - c_1(-c_1 z_1 - sign(z_1, \varpi_1)\chi(x_1) + \tau(x_1)\Delta(t)) - \frac{d}{dt}(sign_1(z_1, \varpi_1)\chi(x_1))).$$

Der Ausdruck $\frac{d}{dt}(sign_1(z_1, \varpi_1)\chi(x))$ kann analytisch zwar bestimmt werden, aber es taucht wiederum die Störgröße $\Delta(t)$ auf. Es wird hier angenommen, dass die Möglichkeit besteht, die Ableitung direkt (durch numerische Differentiation) zu berechnen. Dies kann bei sehr starkem Rauschem im System nicht zulässig sein. Man wird hierbei versuchen, im Regelgesetz den Anteil, der sich analytisch exakt angeben läßt, zu kompensieren und den restlichen Anteil geeignet zu dominieren. Aus Gl.(17) wird klar, dass die Unsicherheit, die durch die Gestaltung des Regelgesetzes u im zweiten Entwurfsschritt beherrscht werden muß, durch folgenden Ausdruck beschrieben wird:

$$c_1\tau(x_1)\Delta(t)) \ . \tag{18}$$

Es fällt auf, dass die Unsicherheit im Gegensatz zum ersten Entwurfsschritt um den Ausdruck c_1 verstärkt auftritt. c_1 ist ein Entwurfsparameter des Backstepping-Reglers. Hohe Werte dieses Parameters führen zu einer schnellen erwünschten Konvergenz des Folgefehlers gegen Null. Da mit der Erhöhung von c_1 auch die effektive „sichtbare" Unsicherheit im zweiten Entwurfsschritt anwächst, sind der Erhöhung von c_1 Grenzen gesetzt, da bei hohen Werten die notwendige Stelldynamik dadurch drastisch zunimmt.

Das Stellgesetz für das System nach Gl.(14) ergibt sich nun zu

$$u = \ddot{x}_S - c_2 z_2 - z_1 - c_1(-c_1 z_1 + z_2 + sign_1(z_1, \varpi_1)\chi(x_1)) + \frac{d}{dt}\left(\left(sign_1(z_1, \varpi_1)\chi(x_1)\right)\right) + sign_1(z_2, \varpi_2)c_1\chi(x_1) \ . \tag{19}$$

Das Backstepping-Verfahren, das sich dieser Form der Unterdrückung von Unsicherheiten bedient (siehe Gl.(19)) wird im folgenden als Backstepping mit robuster Erweiterung durch weichschaltende Anteile bezeichnet (REWA-BS). Die Einbeziehung der Unsicherheit kann dabei auf zweierlei Arten geschehen. Entweder werden jene Funktionen, die die Unsicherheit dominieren, in der hier gezeigten Abhängigkeit von den Systemzuständen gewählt, oder es wird x_1 durch $z_1 + x_S(t)$ ersetzt und die dominierenden Funktionen in Abhängigkeit von den $\mathbf{z}$-Koordinaten entsprechend angepaßt. Sie werden dabei in Abhängigkeit vom Referenzsignal gewählt und müssen an das jeweilige Referenzsignal angepasst werden. Dies bietet für gewisse Varianten des Verfahrens rechentechnische Vorteile. Ohne weitere Abwandlungen jeodch ist von der Stelldynamik her erstere Vorgehensweise vorteilhafter, da das Einbinden des Referenzsignals meist zu sehr konservativen Abschätzungen führt. Der Unsicherheitsterm wird dadurch zu stark dominiert und „schaltende" Bewegungen der Regelgröße können die Folge sein.

Die automatische Anwendung dieses Verfahrens mit jeweiligem Backstepping-Schritt und Zusammenfassen der sich neu ergebenden Unsicherheit zu einer „effektiven" Unsicherheit führt zu Regelgesetzen, die sich durch immer „härter" werdende Regelgesetze auszeichnen. Dies kann bei Systemen höherer Ordnung dazu führen, dass sich das Regelgesetz aufgrund seiner hohen Dynamik für den zur Verfügung stehenden Aktuator nicht eignet. Die Verringerung des Formparamters der weich schaltenden Anteile kann zum Verlust des guten Folgeverhaltens führen und ist nur bedingt eine Lösung. In Abb.2 sind die Stellgrößen und das Ausgangssignal eines Regelkreises unter Verwendung von REWA-BS dargestellt. Die Formparameter ϖ_1, ϖ_2 wurden gleich groß gewählt, mit $\varpi_1 = \varpi_2 = 5$ und $\varpi_1 = \varpi_2 = 50$. Die Strecke wird beschrieben durch Gl.(14) mit $\tau(x_1) = x_1^2$ und Störsignal $\Delta(t) = 2\sin(3t + 0.5)$.

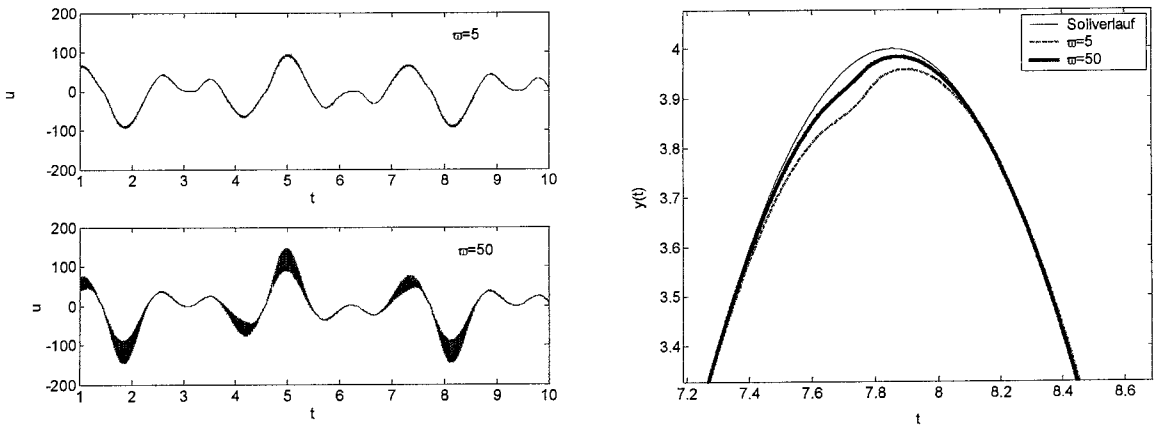


Abbildung 2: REWA-BS mit unterschiedlichen Formparametern $\varpi_1 = \varpi_2$

Erkennbar ist, dass bei höheren Formparametern ϖ_1, ϖ_2 das Folgeverhalten etwas besser ist. Die erforderliche Stelldynamik ist jedoch bedeutend höher. Unter Umständen kann dieses Regelgesetz nicht anwendbar sein. Abhilfe hierzu ist die Kombination dieses Reglerentwurfs mit der dynamischen Erweiterung des Backstepping-Reglers (Abschnitt 4.2) und/oder einer verbesserten Gestaltung der Lyapunovfunktion (Abschnitt 4.3).

Anstelle robust weich schaltender Anteile kann für das System nach Gl.(14) auch die Nichtlineare Dämpfung herangezogen werden. Hierbei ergibt sich anstelle von Gl.(16) folgendes virtuelles Regelgesetz

$$\alpha_1 = \dot{x}_S - c_1 z_1 - \kappa_1 z_1 \tau^2(x_1) . \tag{20}$$

Analog wie beim REWA-BS wird auch hier im zweiten Entwurfsschritt die Ableitung des virtuellen Regelgesetzes benötigt

$$\begin{aligned} \dot{z}_2 &= \dot{x}_2 - \dot{\alpha}_1 = u - \dot{\alpha}_1 = u - (\ddot{x}_S - c_1 \dot{z}_1 - \kappa_1 \dot{z}_1 \tau^2(x_1) - \kappa_1 z_1 2\tau(x_1)\dot{x}_1) = \\ &= u - (\ddot{x}_S + (-c_1 - \kappa_1 \tau^2(x_1))(-c_1 z_1 + z_2 - \kappa_1 z_1 \tau^2(x_1)) - 2\kappa_1 z_1 x_2 \tau(x_1) + . \\ &\quad + \Delta(t)(-c_1 \tau(x_1) - 2\kappa_1 z_1 \tau^2(x_1) - \kappa_1 \tau^3(x_1))) . \end{aligned} \tag{21}$$

Damit ergibt sich folgendes Regelgesetz

$$u = -c_2 z_2 - z_1 + \ddot{x}_S + (-c_1 - \kappa_1 \tau^2(x_1))(-c_1 z_1 + z_2 - \kappa_1 z_1 \tau^2(x_1)) - 2\kappa_1 z_1 x_2 \tau(x_1) + \kappa_2 z_2 (-c_1 \tau(x_1) - 2\kappa_1 z_1 \tau(x_1)^2 - \kappa_1 \tau^3(x_1))^2 \quad . \tag{22}$$

Diese Art der Begegnung der Unsicherheiten wird im folgenden kurz als NLD-RBS (Nichtlineare Dämpfung Robust Backstepping) bezeichnet.

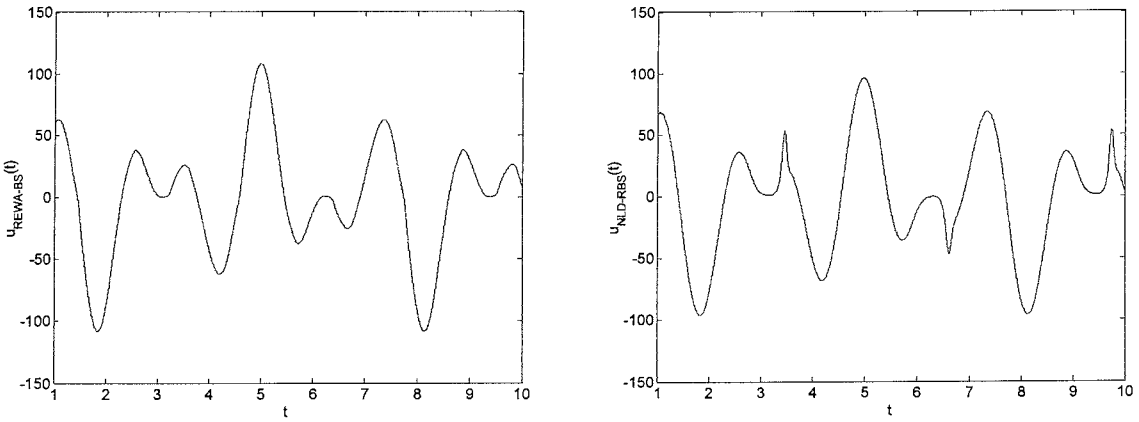


Abbildung 3: Vergleich der Stellgrößen bei REWA-BS und NLD-RBS

In Abb.(3) ist für das System nach Gl.(14) mit $\tau(x_1) = x_1^2$ und Störsignal $\Delta(t) = 2\sin(3t + 0.5)$ und mit den Reglerparameter $c_1 = 10$, $c_2 = 10$ der Stellgrößenverlauf für das REWA-BS Verfahren (links) und für das NLD-RBS Verfahren (rechts) für das Tracking eines Sollsignals $\sin(t)$ aufgetragen. Der Stellgrößenverlauf der beiden Verfahren unterscheidet sich deutlich. Bei Erhöhung der Parameter (ϖ_1, ϖ_2 für REWA-BS und κ_1, κ_2 für NLD-RBS) für bessere Folgegenauigkeit können bei beiden Verfahren in gewissen Bereichen hohe Stelldynamik bzw. sehr unruhige Bewegungen auftreten. In Abschnitt 4.2 und 4.3 werden Methoden zur Verbesserung der Stelldynamik und des Stellbereiches aufgezeigt.

Eine weitere Möglichkeit zur Erhöhung der Robustheit ist die Einbindung eines dynamischen Anteiles in den Backsteppingregler. Die einzelnen Reglerkoordinaten $z_{i=1..n-1}$ werden dabei aufintegriert $\gamma_i = \int_0^t z_i(\tau) d\tau$ und diese Integralanteile in den Backstepping-Regler aufgenommen. Es ist dies für langsam sich ändernde Störgrößen oder für sprungartig sich ändernde, aber in einzelnen Zeitabschnitten konstante Parameter der Strecke durchaus probates Mittel, das auch in Kombination mit den bereits erwähnten Verfahren (REWA-BS oder NLD-RBS) Vorteile bringen kann. Auch die Folgegüte kann hierdurch verbessert werden. Zur Berücksichtigung des Integralanteils wird folgende Lyapunovfunktion, welche die neuen Zustandsgrößen des Reglers mitberücksichtigt, vorgeschlagen:

$$V = \frac{1}{2} z_1^2 + \ldots + \frac{1}{2} z_n^2 + \frac{d_1}{2} \gamma_1^2 + \ldots + \frac{d_{n-1}}{2} \gamma_{n-1}^2 . \tag{23}$$

Der Integralanteil wird in das Regelgesetz eingebunden, so wird für den Nominalfall Gl.(2) ersetzt durch

$$\alpha_1 = \dot{x}_S - f(x_1) - c_1 z_1 - d_1 \gamma_1 . \tag{24}$$

Die Berücksichtigung dieses Integralanteiles in den weiteren Schritten des Reglerentwurfs geschieht durch die Ableitung von α_1 . Im Gegengesatz zu einem klassischen aus dem linearen Regelungsentwurf bekannten PI-Regler ergeben sich durch das Backstepping-Verfahren zusätzliche Anteile. In [Vieh03b] sind die sich ergebenden Anteile – es sind im wesentlichen zusätzliche Vorwärtszweige – im Blockschaltbild dargestellt.

Die Anwendung des Backstepping –Verfahrens mit dynamischen Anteilen auf das Problem nach Gl.(14) mit $\tau(x_1) = x_1^2$ und Störsignal $\Delta(t) = 2\sin(3t + 0.5)$ führt zu ruhigeren Stellgrößenverlaufen als jene in Abb.3 von REWA-BS und NLD-RBS . Es muß jedoch berücksichtigt werden, dass, um eine zu diesen Verfahren vergleichbare Folgegüte zu erhalten, der Verstärkungsfaktor d_i des Integralanteils sehr hoch gewählt werden muß. Dies führt zu äußerst schlechter Messrauschunterdrückung bezüglich Stell- und Regelgröße. Der Einschluss eines geringen Messrauschens vermindert die Regelgüte beträchtlich und kann auch zu Instabilität führen.

Bereits in [PaDa96] wurde gezeigt, dass der Einschluss eines zusätzlichen Integralanteils in ein lineares Regelgesetz für eine nichtlineare Strecke den Stabilitätsbereich im Zustandsraum bezüglich im Entwurf nicht beachteter Nichtlinearitäten bedeutend erweitern kann. In Simulationen (z.B. [Vieh03b]) hat sich ergeben, dass insbesondere bei Variationen der Strecke bzw. der Streckenparameter sich Verbesserungen durch Einbinden eines Integralanteils in Verbindung mit REWA-BS und NLD-RBS ergeben.

4 Erweiterungen, Kombination der Verfahren und Verbesserung der Stelldynamik

4.1 Verbesserung der Stellgrößendynamik bei Nichtlinearer Dämpfung

Die theoretische Herleitung der Nichtlinearen Dämpfung berücksichtigt keine zusätzlichen Rauschanteile der Zustandsgrößen. Diese können zum einen durch das Rauschen des Meßsensors bedingt sein oder aber sich aus der zumeist zeitdiskreten Implementierung des Reglers und damit durch die Abtastung der Meßsignale und dem damit verbundenen Abtastfehler ergeben.
Bei Anwendung der nichtlinearen Dämpfung mit verrauschten Meßssignalen kann simulativ gezeigt werden, daß die Stellgröße hochfrequente Anteile aufweist und damit stark verrauscht ist. Dies tritt auch bei niederen Verstärkungsfaktoren der Nichtlinearen Dämpfung auf.

Ein Mittel zur Verringerung der hochfrequenten Anteile der notwendigen Stelldynamik ist die Einführung einer Totzone in den ersten Faktor des Dämpfungsterms in Gl.(13) der

Nichtlinearen Dämpfung.. Der Ausdruck $-z_i d_i(\mathbf{x}_{[1..i]}, \mathbf{z}_{[1..i]})$ wird ersetzt durch $-\sigma(z_i, l_i) d_i(\mathbf{x}_{[1..i]}, \mathbf{z}_{[1..i]})$, wobei die Funktion σ definiert ist als

$$\sigma(x,l) = \begin{cases} x+l & x \le -l \\ 0 & -l < x < +l \\ x-l & x \ge +l. \end{cases} \tag{25}$$

Das Totzonenglied aus Gl.(25) sorgt dafür, dass bei kleinen Werten der **z**-Koordinaten des Backstepping-Reglers die Anteile der Nichtlinearen Dämpfung keinen Beitrag zur Stellgröße u liefern. Es wird also bei kleinen Fehlerwerten (Man beachte, z_1 ist als Folgefehler definiert) vermieden, dass die Nichtlineare Dämpfung hochfrequente Anteile zur Stellgröße beitragen kann, die Stellgröße ist hierbei nur durch die „linearen Anteile" des Backsteppingsgesetzes gegeben. Die Totzonen l_i sollen ausreichend groß sein, um ausreichende Unterdrückung der hochfrequenten Anteile der Stellgröße zu bewirken, müssen jedoch klein genug gehalten werden, um weiterhin gutes Folgeverhalten der Ausgangsgröße zu garantieren. Insofern kann durch diese Methodik nur eine Unterdrückung der Auswirkung des Meßrauschens auf die Stellgröße in einem gewissen Rahmen bewirkt werden.

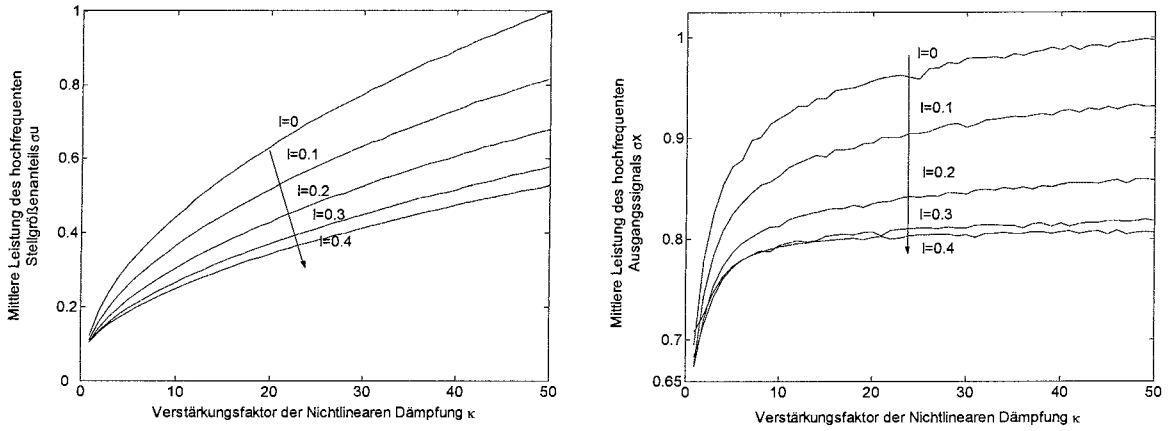


Abbildung 4: Mittlere Leistung der hochfrequenten Anteile der Stellgröße und der Ausgangsgröße in Abhängigkeit von der Größe l der Totzone

Die ideale Größe der Totzonen und der Verstärkungsfaktoren der Nichtlinearen Dämpfung ist abhängig von Art und Leistung des Rauschsignals $n(t)$ und von den Eigenschaften des Störsignals $\Delta(t)$. Für die Strecke aus Gl.(7) mit $\tau(x) = x^2$ und $\Delta(t) = 2\sin(3t + 0.5)$ wurde der Einfluss des Meßrauschens $n(t)$ mit $\sigma_n = 0.1$ auf die Stellgröße und die Ausgangsgröße in Abhängigkeit von dem Verstärkungsfaktor der nichtlinearen Dämpfung und der Größe der Totzone untersucht. Wiederum sollte die Strecke einem sinusförmigen Sollwertsignal folgen. In Abb.(4) sind die mittleren Leistungen der hochfrequenten Anteile von Stellgröße und Ausgangsgröße dargestellt. Diese sollen möglichst gering sein, was durch eine Vergrößerung der Totzone erreicht werden kann. Dem sind natürlich Grenzen gesetzt, da die Folgegüte beibehalten werden soll. Hierzu ist in Abb.(5) der mittlere quadratische Folgefehler e_m aufgetragen, der mit größer werdender Totzone auch ansteigt. Ein vernünftiger Kompromiß

kann z.B. bei $l = 0.2$ und $\kappa = 5$ erreicht werden. Hierbei ist die Auswirkung des Meßrauschens auf Stellgröße und Ausgangsgröße bedeutend verringert und dennoch gutes Folgeverhalten gegeben.

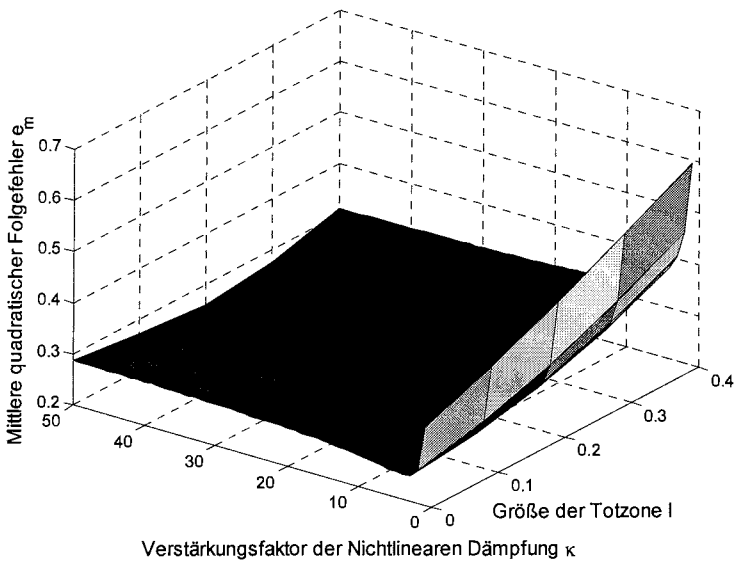


Abbildung 5:Mittler quadratischer Folgefehler in Abhängigkeit von der Größe l derTotzone und dem Verstärkungsfaktor der nichtlinearen Dämpfung

Ähnliche Ergebnisse können auch für Systeme höherer Ordnung mit NLD-RBS Design gewonnen werden. Die Größe der Totzonen muß simulativ ermittelt werden. In Abhängigkeit von den beschreibenden, stochastischen Parametern des Rauschsignals lassen sich vernünftige Startwerte bestimmen. Die Vorgehensweise läßt sich auf REWA-BS übertragen.

4.2 Kombination der Verfahren

Die Kombination von Verfahren kann zu Verbesserungen führen, so kann die dynamische Erweiterung des Backstepping-Reglers um einen Integralanteil mit REWA-BS oder NLD-RBS bei geeigneter Parameterwahl zu Verbesserungen führen.

In [Vieh03b] ist hierzu ein hybrides Verfahren zur Regelung eines einfachen Manipulatormodells mit sich sprungartig ändernder Last vorgestellt. Den einzelnen Zuständen der Strecke wird dabei noch Rauschen hinzugefügt. Es kann dabei gezeigt werden, dass durch die alleinige Hinzufügung eines dynamischen Anteiles in den Backstepping-Regler bei hohem Verstärkungsfaktor desselben die Unsicherheit durch die sich sprungartig verändernde Last beherrscht werden kann. Der Stellgrößenaufwand ist hierbei jedoch sehr hoch und die Rauschunterdrückung des Reglers schlecht. Aus diesem Grunde wird zusätzlich REWA-BS verwendet und eine niedrige Verstärkung des dynamischen Anteiles gewählt. Hierbei wird gutes Folgeverhalten bei vernünftiger Stelldynamik und Stellbereich, sowie gute Rauschunterdrückung erzielt (Siehe Anhang A4).

4.3 Verbesserung der Stelldynamik und Reduzierung der Verstärkung der Unsicherheit

Der Problematik der durch die rekursive Art des Backsteppingsdesigns mit höherer Ordnung zunehmenden Stelldynamik und der Verstärkung der „effektiven" Unsicherheit kann durch besondere Gestaltung der Lyapunovfunktion begegnet werden. Bereits in [ZhKa98] wurden durch unterschiedliche Gestaltung derselben unterschiedliche Ergebnisse und Verbesserungen in Hinblick auf die Stelldynamik erzielt. Die Idee einer ständigen Erweiterung einer Lyapunovfunktion für ein Teilsystem für ein jeweils größeres Teilsystem soll beibehalten werden, die einzelnen Anteile jedoch in Hinblick auf ein günstigeres Regelgesetz abgeändert werden. Hierzu wird ein in [FrKo93] ansatzweise vorgestelltes Verfahren erweitert und für Folgeregelungen adaptiert.

Die Form der Lyapunovfunktion beeinflußt das sich ergebende Regelgesetz und damit die Stelldynamik. Die Bauart der in Abschnitt 1 vorgestellten Lyapunovfunktion für das Backstepping-Design mit $V = \frac{1}{2}(z_1^2 + z_2^2 + ... + z_n^2)$ stellt eine Aufsummierung der quadratischen Abweichung der Systemzustände $\mathbf{x}$ von den virtuellen Regelgesetzen $\boldsymbol{\alpha}$ dar. Das Minimum ist genau dann erreicht, wenn die virtuellen Eingänge der einzelnen Teilsysteme (im besonderen Fall von System Gl.(1) die Systemzustände) und die virtuellen Regelgesetze übereinstimmen. Diese Forderung wird nun aufgeweicht. Es gibt Gebiete im $\mathbf{z}$-Zustandsraum, in welchen die Ableitung der Lyapunovfunktion des ersten Teilsystems negativ ist, aber die Zustände $z_2, z_3, ..., z_n$ nicht verschwinden Es soll das Minimum der Lyapunovfunktion immer dann erreicht sein, wenn die Ausgangsgröße mit dem Sollverlauf übereinstimmt und die restlichen Koordinaten innerhalb dieser Gebiete zu liegen kommen.

Hierzu wird folgende Lyapunovfunktion betrachtet

$$
\begin{aligned}
V &= \frac{1}{2}(z_1^2 + g_1(z_1, z_2) + ... + g_{n-1}(z_1, ..., z_n)) \\
g_1 &= \begin{cases} \frac{1}{2}(z_2 - r_{1o}(z_1))^2 & z_2 > r_{1o}(z_1) \\ 0 & r_{1u}(z_1) \le z_2 \le r_{1o}(z_1) \\ \frac{1}{2}(z_2 - r_{1u}(z_1))^2 & z_2 \le r_{1u}(z_1) \end{cases} \\
&\vdots \\
g_{n-1} &= \begin{cases} \frac{1}{2}(z_n - r_{n-1,o}(z_1, ..., z_{n-1}))^2 & z_2 > r_{n-1,o}(z_1, ..., z_{n-1}) \\ 0 & r_{n-1.u}(z_1, ..., z_{n-1}) \le z_n \le r_{n-1,o}(z_1, ..., z_{n-1}) \\ \frac{1}{2}(z_n - r_{n-1,u}(z_1, ..., z_{n-1}))^2 & z_n \le r_{n-1,u}(z_1, ..., z_{n-1}), \end{cases}
\end{aligned}
\tag{26}
$$

wobei $r_{i,o}$ bzw. $r_{i,u}$ obere, bzw. untere Gebietsgrenzen darstellen.

Es wird das System nach Gl.(14) mit $\tau(x_1) = x_1^2$ und $\Delta(t) = 2\sin(3t + 0.5)$ betrachtet . Es wird ein Regler mit der REWA-BS Methodik mit zusätzlicher Einbindung von dynamischen Anteilen entworfen. Die die Unsicherheit dominierenden Terme wurden in Abhängigkeit von z_1 und

nicht von x_1 entworfen. Sie sind dadurch auch abhängig von den Eigenschaften des Referenzsignals $x_S(t)$.

Für das System zweiter Ordnung müssen im Zustandsraum nach Gl.(26) geeignete Gebietsgrenzen angegeben werden. Die Festlegung der Gebietsgrenzen ist allgemein bis auf gewisse Fundamentaleigenschaften noch nicht bestimmt und Gegenstand weiterer Untersuchung. In Abbildung 6 ist eine mögliche Aufteilung des Zustandsraumes angegeben.

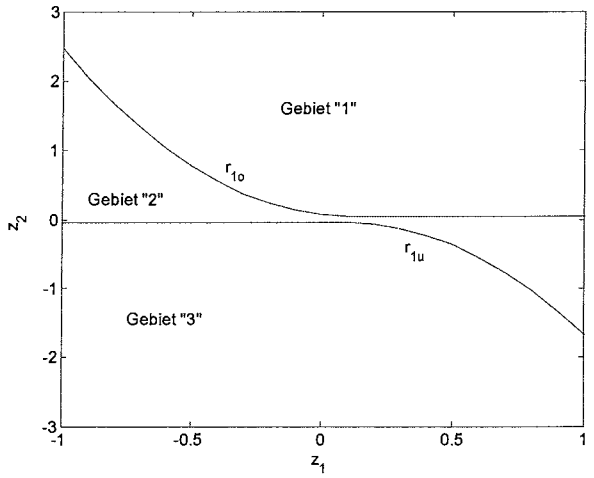


Abbildung 6: Darstellung der Gebiete für die Gestaltung der Lyapunovfunktion im z-Zustandsraum

Aus der Lyapunovfunktion ergeben sich die jeweiligen Regelgesetze, die abhängig sind von der Lage im $\mathbf{z}$-Zustandsraum. So ergibt sich für das Stellsignal bei bereits bekanntem virtuellen Regelgesetz für das erste Teilsystem im Gebiet 1

$$u_{R1} = -z_1 - c_2 z_2 + \ddot{x}_S + \left(\frac{\partial \alpha_1(z_1)}{\partial z_1} + \frac{\partial r_{1,O}(z_1)}{\partial z_1} \right) \{\dot{z}_1\}_{certain} - s_{2,R1}(z_1, z_2) \quad , \tag{27}$$

im Gebiet 2

$$u_{R2} = -z_1 - c_2 z_2 + \ddot{x}_S + \left(\frac{\partial \alpha_1(z_1)}{\partial z_1} \right) \{\dot{z}_1\}_{certain} - s_{2,R2}(z_1, z_2) \tag{28}$$

und im Gebiet 3

$$u_{R3} = -z_1 - c_2 z_2 + \ddot{x}_S + \left(\frac{\partial \alpha_1(z_1)}{\partial z_1} + \frac{\partial r_{1,U}(z_1)}{\partial z_1} \right) \{\dot{z}_1\}_{certain} - s_{2,R3}(z_1, z_2) \, , \tag{29}$$

wobei $\{\dot{z}_1\}_{certain}$ jenen Teil der Ableitung von z_1, in welchem die Unsicherheit $\Delta(t)$ nicht auftritt und $s_{2,R1}(z_1, z_2)$, $s_{2,R2}(z_1, z_2)$ $s_{2,R3}(z_1, z_2)$ geeignete Anteile zur Beherrschung der auftretenden „effektiven" Unsicherheiten darstellen (entweder mit NLD oder REWA). In der Simulation hat sich gezeigt, dass die Anwendung dieser besonderen Gestaltung der Lyapunovfunktion für REWA-BS ohne dynamischen Anteil eine Reduzierung der Stellleistung um den Faktor 3.6

ergeben kann. Besonders gute Ergebnisse konnten in Verbindung mit einem dynamischen Anteil erzielt werden. Hierzu ist in Abbildung 7 der Stellgrößenverlauf bei Verwendung einer Lyapunovfunktion der Form $V = \frac{1}{2}\mathbf{z}^T\mathbf{z}$ und zusätzlich der Vergleich mit der modifizierten Ausgestaltung derselben angegeben. Im Bereich zwischen 3 und 4 sec und 9 und 10 sec fällt eine wesentliche Verbesserung der notwendigen Stelldynamik ins Auge.

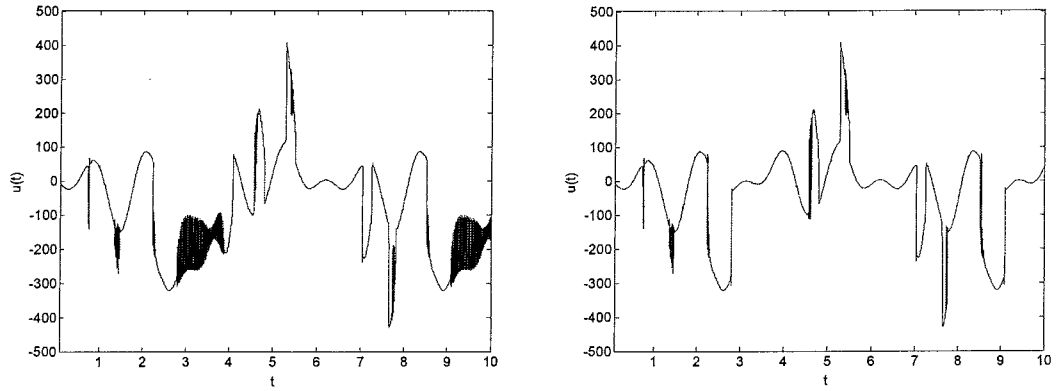
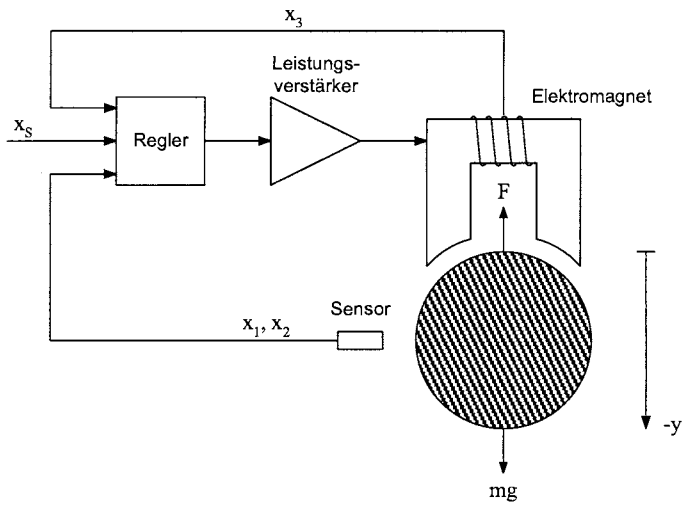


Abbildung 7: Stellgrößenverlauf bei herkömmlicher Gestaltung der Lyapunovfunktion und modifizierter Ausgestaltung der Lyapunovfunktion

5 Anwendungsbeispiel

Als Anwendungsbeispiel zur robusten Folgeregelung mit dem Backstepping-Verfahren soll eine magnetische Schweberegelung herangezogen werden. Sie kann beschrieben werden - samt elektrischen Eigenschaften der Anordnung – mit



$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -g + \xi(x_1, \theta_1, \theta_2)x_3 \\ \dot{x}_3 &= -\theta_3 x_3 + u \\ y &= x_1 \quad , \end{aligned} \tag{30}$$

Abbildung 8: Darstellung der Magnetschweberegelung und des beschreibenden Gleichungssystems

wobei x_1, x_2 die Lage und die Geschwindigkeit der schwebenden Masse bezeichnen und x_3 den Strom in der Spule.

Die Funktion, welche die Abhängigkeit der magnetischen Kraft vom Abstand zum Magneten beschreibt, sei mit $\xi(x_1) = \frac{\theta_1}{1+\theta_2 x_1^2}$ angenommen. Abb. 9 links zeigt unterschiedliche Ausprägungen der Unsicherheit $\xi(x_1)$ bei unterschiedlichen Parameterwerten θ_1, θ_2. Für die Parameter $\theta_1, \theta_2, \theta_3$ seien jeweils obere und untere Grenzen bekannt (Abb. 9 rechts). Man beachte, dass in diesem Fall die Unsicherheit aus der Einkopplung des virtuellen Eingangs x_3 und des unsicheren Rückkoppelparameters θ_3 für x_3 besteht. Nichtsdestoweniger kann diese Form der Unsicherheit in das in Abschnitt 3 eingeführte Konzept eingebracht werden.

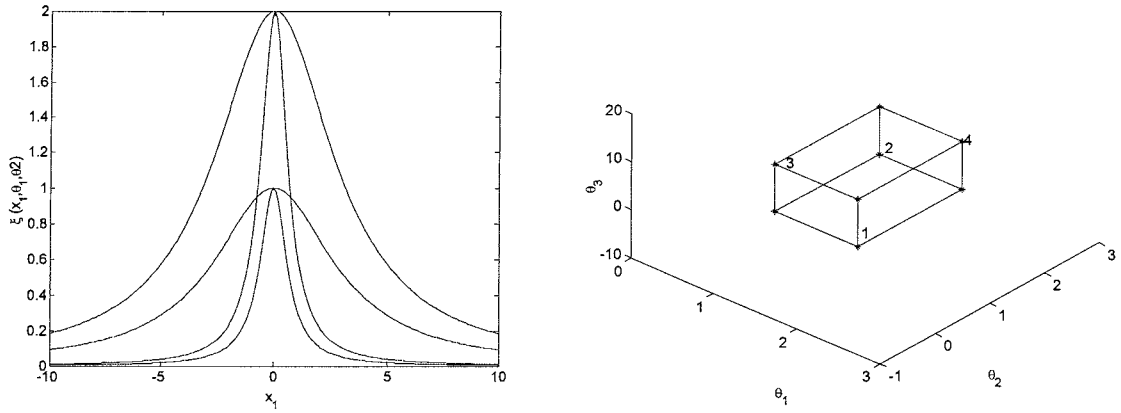


Abbildung 9: Darstellung der Unsicherheit und der Unsicherheitsintervalle im Parameterraum

Der Entwurf des Backstepping-Reglers für den Nominalfall bereitet keine größeren Schwierigkeiten. Aus diesem Grund werden hierbei nur die Ergebnisse des Entwurfs vorgestellt und auf Abschnitt 2 verwiesen. Ergebnis sind folgende virtuelle Regelgesetze für den Nominalfall:

$$\begin{aligned} \alpha_1 &= -c_1 z_1 + \dot{x}_S \\ \alpha_2 &= \xi^{-1}(x_1,\theta_1,\theta_2)(g - (c_1+c_2)z_2 + (c_1^2 - 1)z_1 + \ddot{x}_S) \\ \alpha_3 &= u = \theta_3 x_3 - c_3 z_3 - \xi(x_1,\theta_1,\theta_2) z_2 - \dot{\alpha}_2 \quad . \end{aligned} \tag{31}$$

Diese führen zu folgendem Ersatzsystem für den Nominalfall in z-Koordinaten

$$\dot{\mathbf{z}} = \begin{bmatrix} -c_1 & -1 & 0 \\ 1 & -c_2 & \xi(x_1,\theta_1,\theta_2) \\ 0 & -\xi(x_1,\theta_1,\theta_2) & -c_3 \end{bmatrix} \mathbf{z} \quad . \tag{32}$$

Der robuste Entwurf erfordert die Ergänzung des Regelgesetzes, hierbei wird im zweiten und dritten Entwurfsschritt das virtuelle Regelgesetz jeweils um einen robusten Anteil dermaßen erweitert, dass entweder die Unsicherheit dominiert wird oder aber die sich ergebende Ableitung der Lyapunovfunktion in einem möglichst kleinen Gebiet im $\mathbf{z}$-Zustandsraum indefinit ist.

Wird versucht die Unsicherheit zu dominieren (REWA-BS), so ergibt sich für die robuste Erweiterung des virtuellen Regelgesetzes im zweiten Entwurfsschritt

$$\alpha_{2,ROB,REWA-BS} = -sign_1(z_2,\varpi_1)\left|\frac{\theta_{1,\max}}{1+\theta_{2,\min}x_1^2}\frac{1+\theta_{2,nom}x_1^2}{\theta_{1,nom}}-1\right|\left(\frac{1+\theta_{2,\max}x_1^2}{\theta_{1,\min}}\right)\left|x_S+\left(c_1^2-1\right)z_1-\left(c_1+c_2\right)z_2+g\right|. \quad (33)$$

Das erweiterte Regelgesetz gewährleistet, dass bei beliebiger Lage der Parameter θ_1,θ_2 im erlaubten Bereich des Parameterraums die Konvergenz der Zustandsgrößen in z-Koordinaten gewährleistet ist. Wird die Methodik der Nichtlinearen Dämpfung angewandt, so ergibt sich für den zweiten Entwurfsschritt folgendes Regelgesetz

$$\alpha_{2,ROB-NLD} = -\kappa_1 z_2\left(x_S+\left(c_1^2-1\right)z_1-\left(c_1+c_2\right)z_2+g\right)^2, \quad (34)$$

wobei κ_1 den Parameter der nichtlinearen Dämpfung darstellt.
Man beachte: In Gl.(33) und Gl.(34) sind nur die *zusätzlichen* Anteile zur Unterdrückung der Unsicherheiten angegeben. Das gesamte virtuelle Regelgesetz ergibt sich nach Aufaddierung zu den Nominalanteilen in Gl.(31).

Da das System nach Gl.(30) dritter Ordnung ist, ist noch ein weiterer Rekursionsschritt vonnöten. Die sich in diesem Entwurfsschritt ergebende „effektive" Unsicherheit besteht dabei aus der durch den unsicheren Parameter θ_3 induzierten Unsicherheit und die durch die Ableitung des Regelgesetzes α_2 hervorgehenden Unsicherheiten (Siehe hierzu auch Anhang A3).
Auf die explizite Darstellung des dritten Entwurfsschrittes wird an dieser Stelle verzichtet.

In Abb.10 sind die Stellgrößenverläufe jeweils für REWA-BS und NLD-RBS fur unterschiedliche Konfigurationen im Parameterraum angegeben. Die Bezeichnung der Parameterkonfiguration („1","2",...) stimmt mit jener in Abb.9 rechts überein. Abb.11 zeigt die dazugehörigen Verläufe der Regelgröße. Die Verläufe der Regelgröße $y(t)$ unterscheiden sich nur wenig (Abb.11), die dazugehörigen Stellgrößenverläufe in Abhängigkeit von der Parameterkonfiguration doch deutlich (Abb.10).

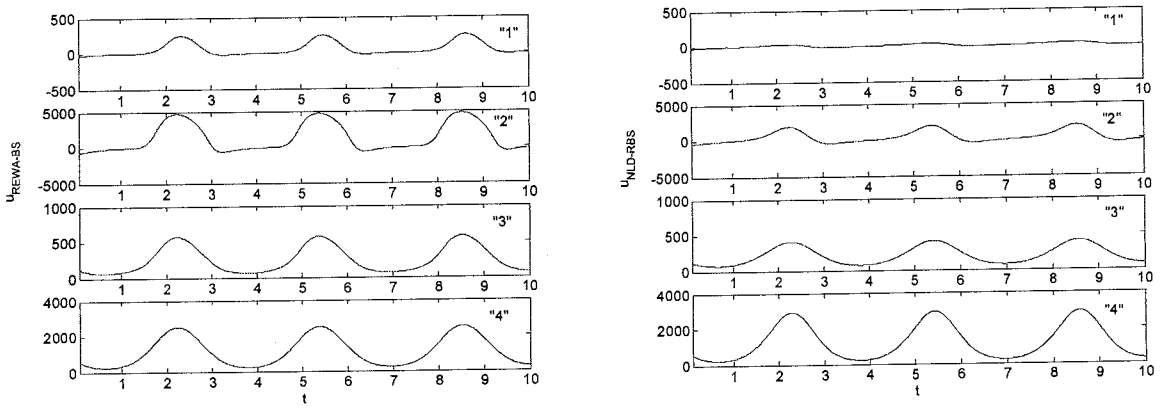


Abbildung 10: Stellgrößenverläufe in Abhängigkeit von der Parameterkonfiguration

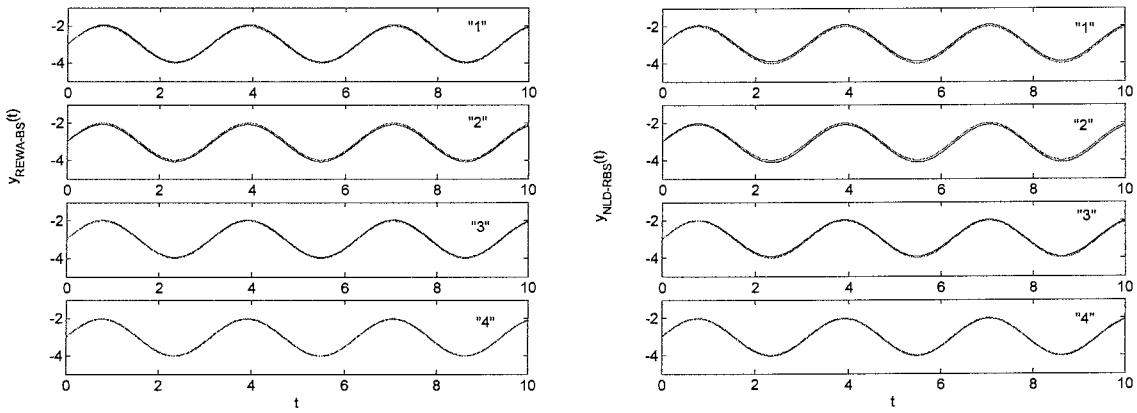


Abbildung 11: Ausgangsgröße der Schweberegelung in Abhängigkeit von der Parameterkonfiguration und dem verwendeten Verfahren (REWA-BS oder NLD-RBS)

In Abb.12 ist die Ausgangsgröße des Regelkreises bei Verwendung von REWA-BS mit einem zusätzlichen Integralanteil dargestellt. Es ergibt sich hierbei zu Beginn ein zusätzlicher Einschwingvorgang. Nach diesem Vorgang wird eine höhere Folgegüte bei allen Parameterkonfigurationen erzielt.

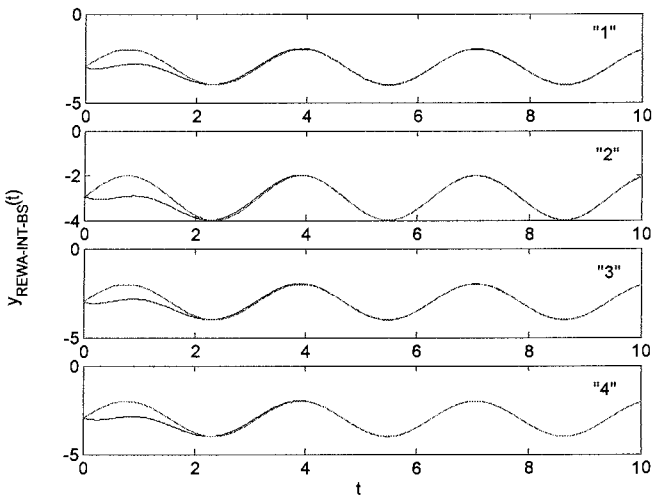


Abbildung 12: Ausgangsgröße in Abhängigkeit der Parameterkonfiguration bei REWA-BS mit integralem Anteil

6 Zusammenfassung und Ausblick

Der Robuste Backstepping-Entwurf setzt den Umgang mit vielen Freiheitsgraden voraus. Hierzu muß der Umgang mit grundsätzlichen Entwurfsmöglichkeiten bekannt sein. Gewisse Eigenschaften des entworfenen Systems können analytisch bestimmt und festgelegt werden, manche wie z.B. Rauschverhalten, ... sind jedoch bedingt durch die auftretenden Nichtlinearitäten nur simulativ zu ermitteln. Die Kombination von Verfahren zur Erhöhung der Robustheit kann zu Verbesserungen (Z.B. niedriger Stellgrößenbedarf bei verbesserter

Folgegüte) führen. Eine Systematik zur optimalen oder suboptimalen Festlegung der Reglerparameter und der Anteile des Regelgesetzes zur Unterdrückung der Unsicherheiten mit Berücksichtigung des Stellgrößenaufwandes, der Rauschunterdrückung und des Folgefehlers ist noch weitgehend unbekannt und wird Gegenstand zukünftiger Forschung sein.

Literaturverzeichnis

[FrKo93] R. A. Freeman, P. V. Kokotovic, *Design of 'Softer' Robust Nonlinear Control Laws*, Automatica, Vol. 29, pp. 1425-1437, 1993.

[JoDa03] Johan Dahlgren, *Robust Nonlinear Control Design for a Missile using Backstepping*, Thesis, University of Technology Linköping, January 2003.

[KrKa95] M. Kristic, I. Kanellakopoulos, P. V. Kokotovic, *Nonlinear and Adaptive Control* Design, Wiley Interscience, 1995.

[MaTo96] R. Marino, P. Tomei, *Nonlinear Control Design, Geometric, Adaptive and Robust*, Prentice Hall, 1996.

[PaDa96] M. Pachter, J. J. D`Azzo, M. Veth, *Proportional and integral control of nonlinear systems,* Int. J. Control, Vol. 64, pp. 679-692, 1996.

[SeJa97] R. Sepulchre, M. Jankovic, P.V. Kokotovic, *Constructive Nonlinear Control*, Springer, 1997.

[Vieh03a] A. Viehweider, *Backstepping – Ein Rekursives Entwurfsverfahren für den Nichtlinearen Regelungsentwurf, Teil 1: Grundlagen und Einführung*, International Journal of Automation Austria, Heft 1, Jahrgang 11, S.21-35, 2003.

[Vieh03b] A. Viehweider, *Backstepping – Ein Rekursives Entwurfsverfahren für den Nichtlinearen Regelungsentwurf, Teil 2: Ein Hybrider Robuster Ansatz zur Regelung von Manipulatoren* , International Journal of Automation Austria, Heft 2, Jahrgang 11, 2003.

[Vieh03c] A. Viehweider, *A Design and Optimization Method for Robust Backstepping Control with Integral Action, in Vorbereitung*, 2003.

[YaMi01] Zi-Jiang Yang, Masayuiki Minashima, *Robust Nonlinear Control of Feedback Linearizable Voltage-Controlled Magnetic Levitation System,* T.IEE Japan, Vol. 121-C, No. 7, 2001.

[ZhKa98] J. Zhao, I. Kanellakopoulos, *Flexible Backstepping Design For Tracking And Disturbance Attenuation*, Int. J. Robust Nonlinear Control 8, 1998.

[ZhQu93] Zhihua Qu, *Robust Control of Nonlinear Uncertain Systems Under Generalized Matching Conditions*, Automatica, Vol. 29, pp.985-998, 1993.

Anhang

A1
Systeme in unterer Dreiecksform („*strict feedback systems*“)

Ein System hat untere Dreiecksform („*strict-feedback*“-Form) , wenn es sich folgendermaßen darstellen läßt:

$$\begin{aligned}
\dot{x}_1 &= a_1(x_1) + b(x_1)x_2 \\
\dot{x}_2 &= a_2(x_1, x_2) + b_2(x_1, x_2)x_3 \\
&\vdots \\
\dot{x}_{n-1} &= a_{n-1(}(x_1, \ldots, x_{n-1}) + b_{n-1}(x_1, \ldots, x_{n-1})x_n \\
\dot{x}_n &= a_n(x_1, \ldots, x_n) + b_n(x_1, \ldots, x_n)u \\
y &= x_1 .
\end{aligned} \tag{35}$$

A2
Young´sche Ungleichung

Es gilt allgemein

$$xy \le \frac{\alpha^p}{p}|x|^p + \frac{1}{q\alpha^q}|y|^q, \quad \alpha > 0, \tag{36}$$

wobei $(x, y) \in \Re^2$ und $p > 1,\ q > 1$ mit $(p-1)(q-1) = 1$ sind.

Im besonderen gilt

$$xy \le \beta x^2 + \frac{1}{4\beta} y^2, \quad \beta > 0. \tag{37}$$

A3
Die „Effektive“ Unsicherheit

Gegeben sei ein System mit Unsicherheiten der Form

$$\begin{aligned}
\dot{\mathbf{x}} &= \mathbf{f}(\mathbf{x}) + \Delta\mathbf{f}(\mathbf{x}) + (\mathbf{g}(\mathbf{x}) + \Delta\mathbf{g}(\mathbf{x}))x_{n+1} \\
\dot{x}_{n+1} &= a_1(\mathbf{x}, x_{n+1}) + \Delta a_1(\mathbf{x}, x_{n+1}) + b_1(\mathbf{x}, x_{n+1})x_{n+2} \\
&\vdots \qquad\qquad .
\end{aligned} \tag{38}$$

Das erste Teilsystem, das sich aus der ersten Zeile von Gl.(38) ergibt, sei durch das virtuelle Regelgesetz $\alpha_1 = \alpha_{1,NOM} + \alpha_{1,ROB}$ global stabilisierbar, wobei sich dieses aus einem Anteil für den Nominalfall und einem robusten Anteil zusammensetzt. Die für die weitere Herleitung

des Backstepping-Verfahrens benötigte zeitliche Ableitung setzt sich nun aus zwei Anteilen zusammen:

$$\begin{aligned}
\frac{d\alpha_1(t)}{dt} &= \frac{d\alpha_{1,NOM}}{dt} + \frac{d\alpha_{1,ROB}}{dt} = \\
&= \frac{\partial \alpha_{1,NOM}(\mathbf{x})}{\partial \mathbf{x}} \left(\mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})(z + \alpha_1(\mathbf{x}))\right) + \frac{\partial \alpha_{1,NOM}(\mathbf{x})}{\partial x} \left(\Delta\mathbf{f}(\mathbf{x}) + \Delta\mathbf{g}(\mathbf{x})(z_1 + \alpha_1(\mathbf{x}))\right) + \\
&\quad \frac{\partial \alpha_{1,ROB}(\mathbf{x})}{\partial \mathbf{x}} \left(\mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})(z + \alpha_1(\mathbf{x}))\right) + \frac{\partial \alpha_{1,ROB}(\mathbf{x})}{\partial \mathbf{x}} \left(\Delta\mathbf{f}(\mathbf{x}) + \Delta\mathbf{g}(\mathbf{x})(z_1 + \alpha_1(\mathbf{x}))\right) = \\
&= \left(\frac{\partial \alpha_{1,NOM}(\mathbf{x})}{\partial \mathbf{x}} + \frac{\partial \alpha_{1,ROB}(\mathbf{x})}{\partial \mathbf{x}}\right) \left(\mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})(z_1 + \alpha_1(\mathbf{x}))\right) + \\
&\quad \left(\frac{\partial \alpha_{1,NOM}(\mathbf{x})}{\partial \mathbf{x}} + \frac{\partial \alpha_{1,ROB}(\mathbf{x})}{\partial \mathbf{x}}\right) \left(\Delta\mathbf{f}(\mathbf{x}) + \Delta\mathbf{g}(\mathbf{x})(z_1 + \alpha_1(\mathbf{x}))\right) = \\
&= \left(\frac{d\alpha(t)}{dt}\right)_{certain} + \left(\frac{d\alpha(t)}{dt}\right)_{uncertain},
\end{aligned} \tag{39}$$

wobei $z_1 = x_{n+1} - \alpha_1$ ist. Im nächsten Entwurfsschritt wird $z_2 = x_{n+2} - \alpha_2$ eingeführt und z_1 abgeleitet:

$$\dot{z}_1 = \dot{x}_{n+1} - \dot{\alpha}_1 = a_1(\mathbf{x}, x_{n+1}) + \Delta a_1(\mathbf{x}, x_{n+1}) + b_1(\mathbf{x}, x_{n+1})(z_2 + \alpha_2) - \left(\frac{d\alpha_1(t)}{dt}\right)_{certain} - \left(\frac{d\alpha_1(t)}{dt}\right)_{uncertain}, \tag{40}$$

Die *„Effektive" Unsicherheit*, die durch die Wahl von α_2 im zweiten Entwurfsschritt des Backstepping-Verfahrens beherrscht werden muß, ergibt sich zu

$$\Delta a_1(\mathbf{x}, x_{n+1}) - \left(\frac{d\alpha_1(t)}{dt}\right)_{uncertain} = \Delta a_1(\mathbf{x}, x_{n+1}) - \left(\frac{\partial \alpha_{1,NOM}(\mathbf{x})}{\partial \mathbf{x}} + \frac{\partial \alpha_{1,ROB}(\mathbf{x})}{\partial \mathbf{x}}\right) \left(\Delta\mathbf{f}(\mathbf{x}) + \Delta\mathbf{g}(\mathbf{x})(z_1 + \alpha_1(\mathbf{x}))\right). \tag{41}$$

Aus Gl.(41) ist ersichtlich, dass sich die Unsicherheit im ersten Teilsystem für die Gestaltung des Regelgesetzes des zweiten Teilsystems α_2 (ergibt sich aus den ersten beiden Zeilen von Gl.(38)) auswirkt. Die Form der Auswirkung wird durch das Regelgesetz für das erste Teilsystem α_1 wesentlich beeinflußt.

A4
Vergleich Backstepping mit dynamischem Anteil, REWA-BS und Kombination der beiden Verfahren.

Die folgenden Ergebnisse sind aus [Vieh03b] übernommen. Der Gelenkwinkel einer einfachen Manipulatoranordnung sollte einem sinusförmigen Verlauf folgen. Während dieser Regelaufgabe treten Lastsprünge auf. Es wird der Backstepping-Regler mit dynamischen Anteil, mit einem REWA-Backstepping-Regler, sowie einer Kombination der beiden Verfahren verglichen. Es sind in den Abbildung 13-15 jeweils die Lastsprünge, das Ausgangssignal und die Stellgröße aufgetragen.

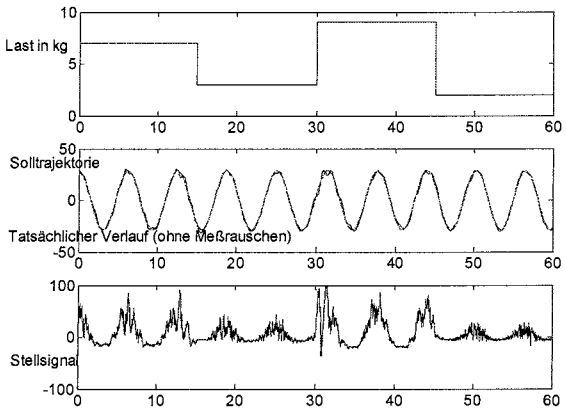


Abbildung 13: Backstepping-Regler mit dynamischem Anteil

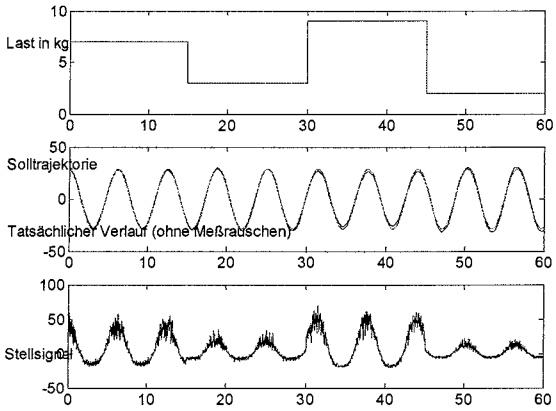


Abbildung 14:REWA-BS

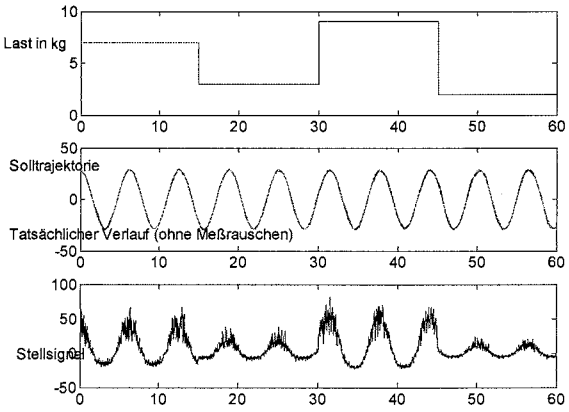


Abbildung 15: REWA-BS mit dynamischem Anteil

Über die Reziprozität bilinearer Systeme und ihrer quadratischen Erweiterungen

R. Starkl
Institut für Regelungstechnik und elektrische Antriebe
Universität Linz
Altenbergerstraße 69, 4040 Linz
reinhard.starkl@jku.at

Kurzfassung

Die vorliegende Arbeit entwickelt ein weitgehend analytisch orientiertes Lösungsverfahren für QLS-Zustandsgleichungen, wobei die QLS eine Verallgemeinerung der bilinearen Systeme ("BLS") bezeichnen. Der Gültigkeitsrahmen dieser Theorie ist ziemlich weitgespannt und kann durch spezielle Algebrensymmetrien beschrieben werden. Es wird die praktische Bedeutung des Lösungsverfahrens im Zusammenhang mit der Realisierungstheorie diskutiert.

I. Einleitung

Ein wichtiges Problem der Systemtheorie ist die Realisierung bzw. die Approximation analytischer, linear steuerbarer Systeme ("ALS") durch einfacher strukturierte nichtlineare Systeme. Bekanntlich existiert ein derartiger Realisierungssatz unter Zuhilfenahme bilinearer Systeme ("BLS") (Carleman-Linearisierung bzw. L_2-Linearisierung; sh. dazu auch [3] bzw. [1]). Nachteilig zeigt sich bei dieser BLS-Realisierung die Tatsache, daß die Dimension des erhaltenen BLS in der Regel wesentlich höher ist als jene des ALS:

$$\dim \text{ BLS} >> \dim \text{ ALS}.$$

Diese für die Praxis ungünstigen Verhältnisse legitimieren die Frage nach der Existenz eines erweiterten Realisierungssatzes, der anstelle von BLS eine komplexer strukturierte Systemklasse verwendet. Wie in [4] und [5] gezeigt, kann ein derartiger Realisierungssatz tatsächlich angegeben werden, wobei als approximierende Systemklasse die QLS $\supseteq$ BLS (das sind im Zustandsvektor quadratische Systeme mit linearer Steuerung) fungiert. Dabei ist die Dimension der QLS-Darstellungen maximal halb so groß wie jene der korrespondierenden BLS-Darstellung:

$$\dim \text{ QLS } \leq \frac{1}{2} \dim \text{ BLS}.$$

Dieser praktische Vorteil der QLS-Realisierung legitimiert den Wunsch nach einem genaueren Studium dieser Systemklasse. Nun sind die BLS in der Literatur sehr ausführlich, die QLS hingegen vergleichsweise dürftig untersucht. Dies betrifft vor allem die Lösungsverhältnisse der i.a. recht komplizierten QLS-Zustandsgleichungen. Aus diesem Grund stellt die vorliegende Arbeit die

Frage I.1:

Lässt sich eine QLS-Lösungstheorie entwickeln?

Wie wir sehen werden, kann für eine sehr umfangreiche QLS-Teilmenge - die wir hier mit $\sum_Q^{(*)}$ bezeichnen wollen - eine einfach strukturierte Lösungstheorie aufgebaut werden. Konkret ergeben sich folgende Aussagen:

1. Die Teilmenge $\sum_Q^{(*)}$ verhält sich "reziprok" zu den BLS in dem Sinne, daß jeder QLS-Zustandsvektor durch eine (nichtlineare) stationäre Transformation T umkehrbar eindeutig auf einen BLS-Zustandsvektor abgebildet werden kann.
2. Diese Transformation T (im folgenden als "Reziprozitätstransformation" bezeichnet) kann mathematisch "griffig" charakterisiert werden. Im einfachsten Fall zeigen sich die korrespondierenden BLS/QLS-Zustandsvektoren als zueinander inverse Elemente einer abelschen Gruppe, d.h. es gilt $x_{QLS} = (x_{BLS})^{-1}$. Darüberhinaus existieren auch Lösungen in allgemeineren Algebren.

Um die praktische Bedeutung dieser Ergebnisse einzusehen, betrachten wir die Fig. 1

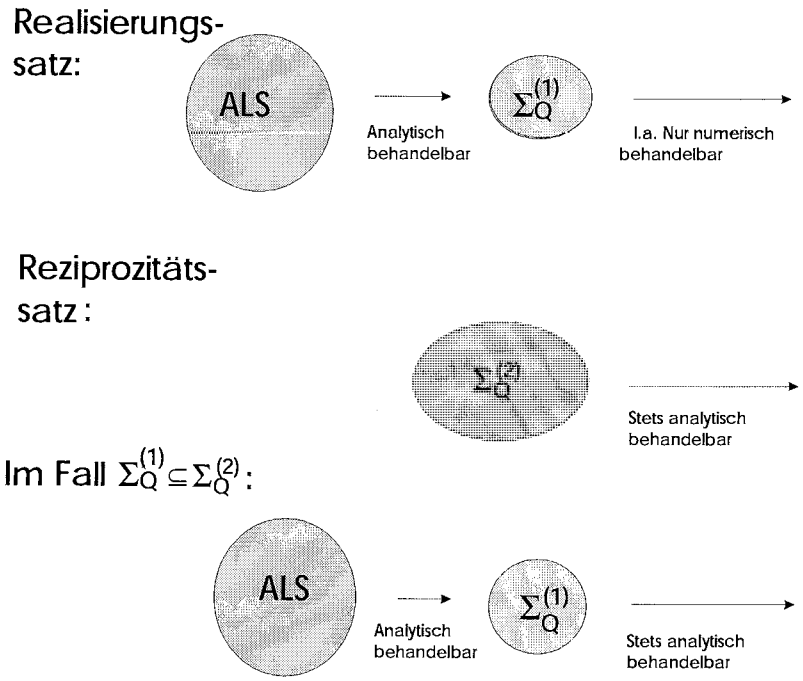


Fig. 1.

Dabei bezeichnet $\sum_Q^{(2)} \subseteq \sum_Q^{(*)}$ jene QLS-Teilmenge, die nicht bloß reziprok zu den BLS, sondern verstärkt reziprok zu den LS (=lineare Systeme mit konstanten Koeffizienten) ist. Da die Systemklasse LS analytisch behandelbar sind, ist es - bei Beherrschung der Reziprozitätstransformation - auch die Systemklasse $\sum_Q^{(2)}$. Wie in ([4]) gezeigt, führt nun die Realisierung von ALS stets auf eine QLS-Teilmenge

$$\sum_Q^{(1)}, \text{ die große formale Ähnlichkeit mit } \sum_Q^{(2)} \text{ hat.}$$

Für die Lösung von AI-Zustandsgleichungen stellt sich somit die

Frage I.2: Gilt

$$\sum_Q^{(1)} \subseteq \sum_Q^{(2)}, \quad \text{(a)} \quad \text{bzw.} \quad \sum_Q^{(1)} \cap \sum_Q^{(2)} \neq \{0\} \text{ (b)} \quad ?$$

Bei positiver Antwort kann der Realisierungssatz mit dem Reziprozitätssatz zu einem weitgehend analytisch orientierten Lösungsverfahren für allgemeine ALS (Fall a) bzw. für eine ALS-Teilmenge (Fall b) kombiniert werden: Das konkrete AI wird realisiert durch ein Element von $\sum_Q^{(1)}$, das im analytisch lösbaren QLS-Bereich liegt. Es gilt also

$$\text{Realisierungssatz} + \text{Reziprozitätssatz} \implies \text{analytisches Lösungsverfahren für AI.}$$

Diese Überlegungen sollen die Sinnhaftigkeit der Fragestellung 1.1. unterstreichen, deren Beantwortung das Ziel der vorliegenden Arbeit ist.[1] Die benötigten algebrentheoretischen Grundlagen sind im Anhang kurz zusammengefaßt. Die in der Einleitung aus aus Gründen der Einfachheit verwendete Bezeichnung der QLS-Teilmengen wird modifiziert.

II. Die Systemklasse $S_{QLS}(V)$ und ihre Spezialisierungen

Der Rahmen unserer Untersuchungen wird durch die Systemklasse S_{QLS} gegeben, die durch das quadratische Auftreten von Zustandsvektoren charakterisiert ist.[2] Zu ihrer exakten Definition erinnern wir an die Auffassung eines Tensors als Element des Vektorraumes V_l^k, der durch wiederholte Tensorproduktbildung der Vektorräume V (k mal) und V^* (l mal) entsteht (Tensor vom Typ (k,l)). Konkret: Bezeichne $\{g_i\}$eine Basis in V, $\{g^j\}$ eine Basis im Dualraum V^*, so gilt für $M \in V_l^k$:

$$M = M^{i_1 i_2 \ldots i_k}{}_{j_1 j_2 \ldots j_l} g_{i_1} \otimes g_{i_2} \ldots\ldots g_{i_k}\ g^{j_1} \otimes g^{j_2} \ldots\ldots\ldots \otimes g^{j_l}.$$

Definition II.1: Sei V ein Vektorraum. Dann definieren wir die Systemklasse $S_{QLS}(V)$ durch die Zustandsgleichung

$$\dot{x} = r + Ax + P(x \otimes x) + u(t)\{s + Nx + Q(x \otimes x)\}, \text{ (a)} \tag{1}$$

$$y = h(x), \text{ (b)}$$

mit $r, s, x \in V$, $A, N \in V_1^1$, $P, Q \in V_2^1$. Für die durch $P = Q = 0$ hervorgehende Systemklasse schreiben wir $S_{BLS}(V)$.

Die Elemente von $S_{QLS}(V)$ bezeichnen wir als QLS (zustandsquadratische, linear steuerbare Systeme), die Elemente der Klasse $S_{BLS}(V)$ als BLS (bilineare Systeme). Den Spezielfall $r = s = 0$ notieren wir als $S_{QLSh}(V)$ bzw. $S_{BLSh}(V)$ ("homogene" Systeme). Eine für die späteren Untersuchungen wichtige Teilmenge von $S_{QLS}(V)$ verdeutlicht die

Definition 1: Sei $X(V, *)$ eine beliebige Algebra. Dann definieren wir die Teilmenge $S_{QLS}(X, *) \subset S_{QLS}(V)$ durch Zustandsgleichungen der Gestalt

$$\dot{x} = r + a * x + p * x * x + \ u(t)(s + n * x + q * x * x), \quad r, a, p, s, n, q \in X(V, *). \tag{2}$$

[1] Es sei bemerkt, daß die Fragestellung 1.2 auf sehr fundamentale differentialgeometrische Probleme führt, worüber zu einem späteren Zeitpunkt berichtet wird.

[2] In der Einleitung mit QLS bezeichnet.

Die Zustandsgleichung läßt sich also ausschließlich aus Vektorprodukten von Vektoren aufbauen. Die Frage nach der Gestalt der Systemmatrizen klärt

Satz II.1: Ein QLS ist genau dann Element von $S_{QLS}(X,*)$, wenn seine Systemmatrizen den folgenden Relationen genügen:

$$A^{\alpha}_{\ \beta} = a^{\mu}C_{\mu\beta}^{\ \ \alpha},\ N^{\alpha}_{\ \beta} = n^{\mu}C_{\mu\beta}^{\ \ \alpha},\ P^{\alpha}_{\ \beta\gamma} = p^{\mu}C_{\mu\gamma}^{\ \ \sigma}C_{\sigma\beta}^{\ \ \alpha},\ Q^{\alpha}_{\ \beta\gamma} = q^{\mu}C_{\mu\gamma}^{\ \ \sigma}C_{\sigma\beta}^{\ \ \alpha}. \quad (3)$$

Beweis. Ein Vergleich mit 2 mit 1 zeigt die Bedingungen $a*x = A.x$, $n*x = Nx$, $p*x*x = P(x\otimes x)$, $q*x*x = Q(x\otimes x)$ als notwendig und hinreichend für die Übereinstimmung beider Zustandsgleichungen. Nun gelten für die Vektoren Ax und Nx die Darstellungen

$$Ax = A^{\alpha}_{\ \beta}x^{\beta}\,g_{\alpha},\quad Nx = n^{\alpha}_{\ \beta}x^{\beta}\,g_{\alpha},$$

während eine entsprechende Entwicklung von $a*x$ und $n*x$ auf

$$a*x = (a^{\mu}g_{\mu})*(x^{\beta}g_{\beta}) = a^{\mu}x^{\beta}C_{\mu\beta}^{\ \ \alpha}\,g_{\alpha},\quad n*x = (n^{\mu}g_{\mu})*(x^{\beta}g_{\beta}) = n^{\mu}x^{\beta}C_{\mu\beta}^{\ \ \alpha}\,g_{\alpha}$$

führt. Der Vergleich liefert dann wegen der linearen Unabhängigkeit der Vektorbasis die ersten beiden Relationen von 2. Zum Nachweis der letzten Beziehung beachten wir die Gültigkeit der Darstellungen

$$P(x\otimes x) = (P^{\alpha}_{\beta\gamma}g^{\beta}\otimes g^{\gamma}\otimes g_{\alpha})(x^{\mu}g_{\mu}\otimes x^{\nu}g_{\nu}) =$$

$$P^{\alpha}_{\beta\gamma}x^{\mu}x^{\nu}(g^{\beta}.g_{\mu})(g^{\gamma}.g_{\nu})g_{\alpha} = P^{\alpha}_{\beta\gamma}x^{\mu}x^{\nu}\delta^{\beta}_{\mu}\,\delta^{\gamma}_{\nu}g_{\alpha} = P^{\alpha}_{\beta\gamma}x^{\beta}x^{\gamma}g_{\alpha},$$

$$p*x*x = ((p^{\mu}g_{\mu})*(x^{\alpha}g_{\alpha}))*(x^{\beta}g_{\beta}) = p^{\mu}x^{\alpha}x^{\beta}((g_{\mu}*g_{\alpha})*g_{\beta}) =$$

$$p^{\mu}x^{\alpha}x^{\beta}C_{\mu\alpha}^{\ \ \gamma}(g_{\gamma}*g_{\beta}) = p^{\mu}x^{\alpha}x^{\beta}C_{\mu\alpha}^{\ \ \gamma}C_{\gamma\beta}^{\ \ \sigma}g_{\sigma} = p^{\mu}x^{\gamma}x^{\beta}C_{\mu\gamma}^{\ \ \alpha}C_{\alpha\beta}^{\ \ \sigma}g_{\sigma},$$

wobei im letzten Schritt eine Indexumbenennung durchgeführt wurde. Ein Vergleich beider Darstellungen liefert dann die vorletzte Aussage von 3, und Gleiches gilt für die letzte Aussage. ■

Bemerkung II.1: Die Systemmatrizen der Elemente von $S_{QLS}(X,*)$ sind durch die Bedingungen 3 eingeschränkt, wobei ihre konkrete Gestalt durch jene der gewählten Algebra $X(V,*)$ mitbestimmt wird. Je mehr Symmetrie von der Algebra verlangt wird, desto eingeschränkter ist die Formgebung der Matrizen.

Eine fundamentale Aussage über die Systemmatrizen der Klasse $S_{QLS}(X,*)$ für assoziative Algebren liefert das

Lemma 1: Die Systemmatrizen jedes Elementes von $S_{QLS}(A_{sign(m)},*)$ genügen den Relationen

$$[A,C^{(3)\gamma}]_{-} = [N,C^{(3)\gamma}]_{-} = 0,$$

$$[A,N]_{-} = ma^{\alpha}n^{\beta}([C^{(2)}_{\alpha},C^{(2)}_{\beta}]_{-})^{T}, \quad (4)$$

Beweis. Die Definition des Operators $[A,B]_{-sign(m)}$ lautet in Komponentenschreibweise

$$([A,B]_{-sign(m)})_{\nu}^{\ \sigma} = A^{\mu}_{\ \nu}B^{\sigma}_{\ \mu} - mB^{\mu}_{\ \nu}A^{\sigma}_{\ \mu} \quad (5)$$

(Tausch der Buchstaben A und B bei festgehaltenem Indexmuster). Nach Senken des Index σ erhält man $([A,B]_{-sign(m)})_{\nu\sigma} = A^{\mu}{}_{\nu}B_{\sigma\mu} - mB^{\mu}{}_{\nu}A_{\sigma\mu}$. Setzt man für $B_{\sigma\mu}$ die quadratische Matrix $C^{(3)\gamma}$bei festgehalten gedachtem Index γ ein, so gilt

$$([A,C^{(3)\gamma}]_{-sign(m)})_{\nu\sigma} = A^{\mu}{}_{\nu}C_{\mu\sigma}{}^{\gamma} - mC_{\mu\nu}{}^{\gamma}A^{\mu}{}_{\sigma},$$

bzw. eine gleichlautende Beziehung für die Matrix N, und es ist zu zeigen, daß die rechten Seiten der Gleichungen identisch verschwinden. Einsetzen der Darstellung aus 3 für A liefert

$$([A,C^{(3)\gamma}]_{-sign(m)})_{\nu\sigma} = a^{\alpha}(C_{\alpha\nu}{}^{\mu}C_{\mu\sigma}{}^{\gamma} - mC_{\mu\nu}{}^{\gamma}C_{\alpha\sigma}{}^{\mu}).$$

Im Fall $m = 1$ verschwindet gemäß 23a,b der geklammerte Term, d.h. A kommutiert tatsächlich mit allen Matrizen $C^{(3)\gamma}$. Im Fall $m = -1$ erhält man für den Antikommutator einen i.a. nichtverschwindenden Term. Der Nachweis für die Matrix N verläuft identisch. Zum Nachweis von 4 bilden wir den Kommutator von A und N gemäß

$$([A,N]_{-sign(m)})_{\nu}{}^{\sigma} = A^{\mu}{}_{\nu}N^{\sigma}{}_{\mu} - mN^{\mu}{}_{\nu}A^{\sigma}{}_{\mu}.$$

Einsetzen der Darstellungen $A^{\mu}{}_{\nu} = a^{\alpha}C_{\alpha\nu}{}^{\mu}$, $N^{\sigma}{}_{\mu} = n^{\beta}C_{\mu\beta}{}^{\sigma}$ führt auf

$$([A,N]_{-sign(m)})_{\nu}{}^{\sigma} = a^{\alpha}n^{\beta}(C_{\alpha\nu}{}^{\mu}C_{\mu\beta}{}^{\sigma} - mC_{\nu\beta}{}^{\mu}C_{\mu\alpha}{}^{\sigma}),$$

und die Berücksichtigung von $C_{\alpha\nu}{}^{\mu} = mC_{\nu\alpha}{}^{\mu}$ liefert

$$([A,N]_{-sign(m)})_{\nu}{}^{\sigma} = ma^{\alpha}n^{\beta}(C_{\alpha\nu}{}^{\mu}C_{\mu\beta}{}^{\sigma} - C_{\mu\alpha}{}^{\sigma}C_{\nu\beta}{}^{\mu}).$$

Der geklammerte Term ist unabhängig von m und ein Vergleich mit der Kommutationsaussage 5 zeigt die Gültigkeit der Satzaussage. ■

III. Die Reziprozitätssätze

Nach den einleitenden Bemerkungen stellt sich nun die Frage, ob bzw. unter welchen Voraussetzungen eine bijektive *stationäre* Abbildung $T : S_{QLS}(V) \leftrightarrow S_{BLS}(V)$ existiert. Es ist einsichtig, daß die Beantwortung dieser Frage empfindlich von der Gestalt des zugrundeliegenden Vektorraumes abhängen muß.

A. Der Reziprozitätssatz für Gruppenstrukturen

Die einfachste Beantwortung der obigen Frage ergibt sich im Zusammenhang mit Gruppenstrukturen, d.h. für die Teilmengen $S_{QLSh}(G,*)$, $S_{BLS}(G,*)$. Bezeichne x_{QLS} den Zustandsvektor eines QLS, x_{BLS} den Zustandsvektor eines BLS, $I_{(*)}$ das Einselement einer Algebra vom Typ G. Dann gilt der

Satz III.1: Die Transformation

$$x_{QLS} * x_{BLS} = I_{(*)} \tag{6}$$

definiert eine bijektive, stationäre Abbildung $T : S_{QLSh}(G,*) \leftrightarrow S_{BLS}(G,*)$.

Beweis. Die Zeitunabhängigkeit der Transformation ist evident, da $I_{(*)}$ stationär ist. Zu zeigen bleibt also, daß der 6 definierte Vektor x_{QLS} eine QLSh-Zustandsgleichung erfüllt, wenn der Vektor x_{BLS} einer BLS-Zustandsgleichung genügt und umgekehrt. Differenziert man 6 nach der Zeit, so folgt

$$\dot{x}_{QLS} * x_{BLS} = -x_{QLS} * \dot{x}_{BLS}. \tag{7}$$

Vektormultiplikation der linken Seite mit mit x_{QLS} :

$$(\dot{x}_{QLS} * x_{BLS}) * x_{QLS} = -(x_{QLS} * \dot{x}_{BLS}) * x_{QLS}.$$

1. Wir zeigen, daß x_{QLS} einer QLSh-Zustandsgleichung genügt, wenn x_{BLS} einer BLS-Zustandsgleichung genügt.

Die linke Seite der obigen Beziehung vereinfacht sich wegen der Assoziativität und Kommutativität gemäß

$$(\dot{x}_{QLS} * x_{BLS}) * x_{QLS} = \dot{x}_{QLS} * (x_{BLS} * x_{QLS}) = \dot{x}_{QLS} * I_{(*)} = \dot{x}_{QLS},$$

wobei wir im vorletzten Schritt die Transformation 6 benutzt haben. Für die rechte Seite ergibt sich wegen Assoziativität und Kommutativität

$$-(x_{QLS} * \dot{x}_{BLS}) * x_{QLS} = -\dot{x}_{BLS} * (x_{QLS} * x_{QLS}),$$

und nach Einsetzen der BLS-Zustandsgleichung 2 ($p = q = 0$):

$$= -\{r + a * x_{BLS} + (u\ t)(s + n * x_{BLS})\} * (x_{QLS} * x_{QLS}) =$$

$$= -\{r * x_{QLS} + a * (x_{BLS} * x_{QLS}) + u(t)(s * x_{QLS} + n * (x_{BLS} * x_{QLS})\} * x_{QLS}.$$

Berücksichtigung der Transformation 6:

$$= -\{r * x_{QLS} * x_{QLS} + a * x_{QLS} + (u\ t)(s * x_{QLS} * x_{QLS} + n * x_{QLS})\}.$$

Wir erhalten somit aus 7 die Zustandsgleichung

$$\dot{x}_{QLS} = -\{r * x_{QLS} * x_{QLS} + a * x_{QLS} + u(t)(s * x_{QLS} * x_{QLS} + n * x_{QLS})\}, \tag{8}$$

also tatsächlich ein QLSh, im Einklang mit der Satzaussage.

2. Den umgekehrten Fall, daß x_{BLS} einer BLS-Zustandsgleichung genügt, wenn x_{QLS} einer QLSh-Zustandsgleichung genügt, zeigt man völlig analog, worauf verzichtet werden kann. ■

Die beiden Zustandsvektoren x_{QLS} und x_{BLS} sind zueinander *invers* (*reziprok*), weshalb wir die obige Transformation 6 auch als *Reziprozitätstransformation* und den Satz als *speziellen Reziprozitätssatz* bezeichnen. "Speziell" deshalb, weil er sich auf Gruppenstrukturen bezieht. Bezeichnet man die QLS-Systemmatrizen mit dem Index "QLS", die BLS-Systemmatrizen mit dem Index "BLS", so erkennt man durch einen Vergleich von 8 und 2 die Gültigkeit von

$$a_{QLS} = -a_{BLS},\ n_{QLS} = -n_{BLS},\quad p_{QLS} = -r_{BLS},\ q_{QLS} = -s_{BLS}. \tag{9}$$

Die linearen Anteile beider Systeme sind also bis auf das geänderte Vorzeichen identisch.

Bemerkung III.1: Die Singularität der Reziprozitätstransformation im Nullpunkt ist bedeutungslos, da durch eine geeignete Koordinatentransformation des Zustandsraumes immer reguläre Verhältnisse erzwungen werden können.

B. Der allgemeine Reziprozitätssatz

Für die praktische Anwendung sind möglichst allgemeine Systemmatrizen wünschenswert. Es stellt sich daher die Frage, ob man anstelle der Gruppenstruktur weniger einschneidende Forderungen voraussetzen kann. Dies ist tatsächlich möglich. Die einfachste Verallgemeinerung besteht in der Zulassung von Algebren A_-, in der nun allerdings kein Einselement, sondern bestenfalls ein "Links- oder rechtsneutrales Element" existiert. Man zeigt leicht, daß in diesem Fall eine zur obigen analoge Vorgangsweise möglich ist. Wesentlich interessanter noch ist die Frage, ob sich die Assoziativitätsbedingung umgehen läßt, da diese für den Strukturkonstantentensor viel einschränkender ist, als die (Anti)Kommutationsforderung. In diesem Zusammenhang beachten wir den

Satz III.2: Sei $X(V,*)$ eine beliebige Algebra, $\lambda \in V$, $F \in V_2^1$ beliebige konstante Tensoren. Dann definiert die Transformation

$$x_{QLS} * x_{BLS} = \lambda \tag{10}$$

genau dann eine bijektive Abbildung $T : S_{QLSh}(X,*) \leftrightarrow S_{BLS}(X,*)$, falls für die Systemmatrizen die folgenden Relationen gelten:

$$(A^{\nu}{}_{\sigma})_{BLS} C_{\mu\nu}{}^{\beta} + (A^{\alpha}{}_{\mu})_{QLS} C_{\alpha\sigma}{}^{\beta} = 0, \text{ (a)} \quad (N^{\nu}{}_{\sigma})_{BLS} C_{\mu\nu}{}^{\beta} + (N^{\alpha}{}_{\mu})_{QLS} C_{\alpha\sigma}{}^{\beta} = 0, \text{ (b)} \tag{11}$$

$$P^{\alpha}_{\mu\upsilon} C_{\alpha\sigma}{}^{\beta} = F^{\beta}_{\mu\alpha} C_{\nu\sigma}{}^{\alpha}, \text{ (a)} \qquad Q^{\alpha}_{\mu\upsilon} C_{\alpha\sigma}{}^{\beta} = F^{\beta}_{\mu\alpha} C_{\nu\sigma}{}^{\alpha}, \text{ (b)} \tag{12}$$

$$r^{\nu} C_{\mu\nu}{}^{\beta} + F^{\beta}_{\mu\nu}\, \lambda^{\nu} = 0, \text{ (a)} \qquad s^{\nu} C_{\mu\nu}{}^{\beta} + F^{\beta}_{\mu\nu}\, \lambda^{\nu} = 0. \text{ (b)} \tag{13}$$

Beweis. Da wir die früheren Symmetrien (Kommutativität, Assoziativität, Einselement) nicht zur Verfügung haben, bringt die symbolische Vektorschreibweise keine Vorteile, weshalb wir die allgemeine Indexschreibweise verwenden. Wir gehen also von der Komponentenversion von 10:

$$x^{\mu}_{QLS} C_{\mu\nu}{}^{\gamma} x^{\nu}_{BLS} = \lambda^{\gamma} \tag{14}$$

aus. Mit Hilfe der Abkürzung

$$D_{\mu}{}^{\gamma}(x_{BLS}) := C_{\mu\nu}{}^{\gamma} x^{\nu}_{BLS} \tag{15}$$

(Achtung auf die Stellung des Summationsindex) schreibt sich die Transformationsgleichung als

$$x^{\mu} D_{\mu}{}^{\gamma}(x_{BLS}) = \lambda^{\gamma}.$$

Überschiebung mit der durch $D_{\mu}{}^{\gamma}(x_{QLS}) D_{*\gamma}{}^{\alpha}(x_{QLS}) = \delta^{\alpha}_{\mu}$ definierten inversen Matrix $D_{*\gamma}{}^{\alpha}(x_{BLS})$ ergibt die explizite Darstellung

$$x^{\alpha}_{QLS} = \lambda^{\gamma} D_{*\gamma}{}^{\alpha}(x_{BLS}). \tag{16}$$

Zur Berechnung der zeitlichen Ableitung des Zustandsvektors x_{QLS} differenzieren wir 14 nach der Zeit und erhalten

$$\dot{x}^{\mu}_{QLS} x^{\nu}_{BLS} C_{\mu\nu}{}^{\gamma} + x^{\mu}_{QLS} \dot{x}^{\nu}_{BLS} C_{\mu\nu}{}^{\gamma} = 0,$$

bzw.

$$\dot{x}^{\mu}_{QLS} D_{\mu}{}^{\gamma}(x_{BLS}) = -x^{\mu}_{QLS} \dot{x}^{\nu}_{BLS} C_{\mu\nu}{}^{\gamma}.$$

Überschiebung mit $D_{*\gamma}{}^{\alpha}(x_{BLS})$ liefert dann

$$\dot{x}^{\alpha}_{QLS} = -x^{\mu}_{QLS} \dot{x}^{\nu}_{BLS} C_{\mu\nu}{}^{\gamma} D_{*\gamma}{}^{\alpha}(x_{BLS}). \tag{17}$$

Voraussetzungsgemäß ist x_{BLS} Zustandsvektor eines BLS:

$$\dot{x}^{\nu}_{BLS} = G^{\nu}{}_{\beta}(t)\ x^{\beta}_{BLS} + w^{\nu}(t),$$

mit

$$(G^{\nu}{}_{\beta})_{BLS}(t) := (A^{\nu}{}_{\beta})_{BLS} + u(t)(N^{\nu}{}_{\beta})_{BLS}, \quad w^{\nu}(t) := r^{\nu} + (u\ t)s^{\nu}.$$

Daraus folgt

$$\dot{x}^{\alpha}_{QLS} = -x^{\mu}_{QLS}(G^{\nu}{}_{\beta})_{BLS}(t) x^{\beta}_{BLS} C_{\mu\nu}{}^{\gamma} D_{*\gamma}{}^{\alpha}(x_{BLS}) - x^{\mu}_{QLS} w^{\nu}(t) C_{\mu\nu}{}^{\gamma} D_{*\gamma}{}^{\alpha}(x_{BLS}). \tag{18}$$

Damit x_{QLS} ein QLS-Zustandsvektor ist, muß

$$\dot{x}^{\alpha}_{QLS} = (G^{\alpha}{}_{\mu})_{QLS}(t) x^{\mu}_{QLS} + (P^{\alpha}{}_{\mu v} + (u\ t) Q^{\alpha}{}_{\mu v}) x^{\mu}_{QLS} x^{\nu}_{QLS}$$

gelten, mit

$$(G^{\nu}{}_{\beta})_{QLS}(t) := (A^{\nu}{}_{\beta})_{QLS} + (u\ t)(N^{\nu}{}_{\beta})_{QLS}.$$

Ein Vergleich der beiden Zustandsgleichungen liefert für den ersten Term auf der rechten Seite von 18:

$$-x^{\mu}_{QLS}(G^{\nu}{}_{\beta})_{BLS}(t) x^{\beta}_{BLS} C_{\mu\nu}{}^{\gamma} D_{*\gamma}{}^{\alpha}(x_{BLS}) = \hat{G}^{\alpha}{}_{\mu}(t) x^{\mu}_{QLS}.$$

Nach Überschiebung mit $D_{\alpha}{}^{\beta}(x_{BLS})$ erhält man wegen der linearen Unabhängigkeit der Zustandsvektoren

$$\begin{aligned} -(G^{\nu}{}_{\sigma})_{BLS}(t) x^{\sigma}_{BLS} C_{\mu\nu}{}^{\gamma} D_{*\gamma}{}^{\alpha}(x_{BLS}) D_{\alpha}{}^{\beta}(x_{BLS}) &= -(G^{\nu}{}_{\sigma})_{BLS}(t) x^{\sigma}_{BLS} C_{\mu\nu}{}^{\gamma} \delta^{\beta}_{\gamma} = \\ &= (G^{\alpha}{}_{\mu})_{QLS}(t) D_{\alpha}{}^{\beta}(x_{BLS}), \end{aligned}$$

und somit

$$-(G^{\nu}{}_{\sigma})_{BLS}(t) x^{\sigma}_{BLS} C_{\mu\nu}{}^{\beta} = (G^{\alpha}{}_{\mu})_{QLS}(t) D_{\alpha}{}^{\beta}(x_{BLS}) = (G^{\alpha}{}_{\mu})_{QLS}(t) C_{\alpha\nu}{}^{\beta} x^{\nu}_{BLS}. \tag{19}$$

Wegen der linearen Unabhängigkeit der x_{BLS}-Komponenten folgt

$$(G^{\nu}{}_{\sigma})_{BLS} C_{\mu\nu}{}^{\beta} + (G^{\alpha}{}_{\mu})_{QLS} C_{\alpha\sigma}{}^{\beta} = 0,$$

was für die zeitunabhängigen Systemmatrizen

$$(A^{\nu}{}_{\sigma})_{BLS} C_{\mu\nu}{}^{\beta} + (A^{\alpha}{}_{\mu})_{QLS} C_{\alpha\sigma}{}^{\beta} = 0, \qquad (N^{\nu}{}_{\sigma})_{BLS} C_{\mu\nu}{}^{\beta} + (N^{\alpha}{}_{\mu})_{QLS} C_{\alpha\sigma}{}^{\beta} = 0 \tag{20}$$

impliziert. Betrachten wir nun den zweiten Term in 18. Für eine Übereinstimmung mit dem EBLS-Term muß

$$-x^{\mu}_{QLS} w^{\nu}(t) C_{\mu\nu}{}^{\gamma} D_{*\gamma}{}^{\alpha}(x_{BLS}) = +(P^{\alpha}{}_{\mu v} + u(t) Q^{\alpha}{}_{\mu v}) x^{\mu}_{QLS} x^{\nu}_{QLS}$$

gelten. Überschiebung mit $D_{\alpha}^{\beta}(x_{BLS})$ liefert dann wegen der linearen Unabhängigkeit der x_{QLS}-Komponenten für den von $u(t)$ unabhängigen Teil:

$$-r^{\nu}C_{\mu\nu}{}^{\gamma}D_{*\gamma}{}^{\alpha}(x_{BLS})D_{\alpha}^{\beta}(x_{BLS}) = -r^{\nu}C_{\mu\nu}{}^{\gamma}\delta_{\gamma}^{\beta} = -r^{\nu}C_{\mu\nu}{}^{\beta} = P_{\mu\nu}^{\alpha}x^{\nu}D_{\alpha}^{\beta}(x_{BLS}),$$

und weiter

$$-r^{\nu}C_{\mu\nu}{}^{\beta} = P_{\mu\nu}^{\alpha}x_{QLS}^{\nu}C_{\alpha\sigma}{}^{\beta}x_{BLS}^{\sigma}.$$

Für den von u(t) abhängigen Teil erhält man analog

$$-s^{\nu}C_{\mu\nu}{}^{\beta} = Q_{\mu\nu}^{\alpha}x_{QLS}^{\nu}C_{\alpha\sigma}{}^{\beta}x_{BLS}^{\sigma}.$$

Da die linken Seiten dieser Gleichungen von den Zustandsvektoren unabhängig sind, muß dies auch für die rechten Seiten gelten. Man erkennt die rechten Seiten der obigen Beziehungen als konstant, falls Relationen der Gestalt

$$P_{\mu\nu}^{\alpha}C_{\alpha\sigma}{}^{\beta} = F_{\mu\rho}^{\beta}C_{\nu\sigma}{}^{\rho}, \quad Q_{\mu\nu}^{\alpha}C_{\alpha\sigma}{}^{\beta} = F_{\mu\rho}^{\beta}C_{\nu\sigma}{}^{\rho}, \tag{21}$$

mit einer beliebigen Matrix $F_{\mu\rho}^{\beta}$ gilt. Für die erste Gleichung folgt damit unter Berücksichtigung von 14

$$P_{\mu\nu}^{\alpha}C_{\alpha\sigma}{}^{\beta}x_{QLS}^{\nu}x_{BLS}^{\sigma} = F_{\mu\rho}^{\beta}C_{\nu\sigma}{}^{\rho}x_{QLS}^{\nu}x_{BLS}^{\sigma} = F_{\mu\rho}^{\beta}v^{\rho},$$

und somit

$$r^{\nu}C_{\mu\nu}{}^{\beta} + F_{\mu\nu}^{\beta}\lambda^{\nu} = 0.$$

Analog dazu erhält man für die zweite Gleichung in 21:

$$s^{\nu}C_{\mu\nu}{}^{\beta} + F_{\mu\nu}^{\beta}\lambda^{\nu} = 0,$$

womit alles bewiesen ist. ■

Bemerkung III.2: Man beachte, daß sich die Gruppenstruktur zusammen mit 9 als einfachste Lösung der Kommutationsbeziehungen 11-13 ergibt.

IV. Zusammenfassung und Ausblick

Die vorliegende Arbeit entwickelt eine Theorie zur Lösung von QLS-Zustandsgleichungen im Rahmen des "Reziprozitätsproblems". Die wesentlichen Ergebnisse lassen sich zusammenfassen wie folgt:

1. Eine umfangreiche Teilmenge der QLS verhält sich "reziprok" zu den BLS, d.h. jeder QLS-Zustandsvektor kann durch eine (nichtlineare) stationäre Transformation (="Reziprozitätstransformation") umkehrbar eindeutig auf einen BLS-Zustandsvektor abgebildet werden. Damit folgt aus der rechentechnischen Beherrschung des BLS jene des dazu reziproken QLS.
2. Diese Reziprozitätstransformation ist nicht eindeutig festgelegt; Alle Lösungen lassen sich als Multiplikation in einer geeigneten algebraischen Struktur beschreiben. Im Spezialfall einer abelschen Gruppe zeigen die Zustandsvektoren von BLS und QLS eine

fundamentale Symmetrie: sie sind interpretierbar als zueinander inverse Elemente der Gruppe.

Auf der Basis der hier vorgenommenen gruppentheoretischen Charakterisierung stellen sich folgende grundlegende Fragen, die in einer späteren Arbeit behandelt werden sollen:

- Die Frage nach der expliziten Konstruktion von Reziprozitätstransformationen, d.h. nach der expliziten Konstruktionsmöglichkeit von Strukturkonstanten von G bzw. allgemeinerer algebraischer Strukturen.
- Die Frage nach der praktischen Relevanz der hier präsentierten QLS-Teilmenge (sh. dazu die Überlegungen in der Einleitung). Ihre Beantwortung setzt die explizite Konstruktion möglichst allgemeiner algebraischer Strukturen (sh. den obigen Punkt) voraus.

V. Anhang: Elemente der Algebrentheorie

Der vorliegende Abschnitt erinnert an die Grundlagen der Algebrentheorie und entwickelt die in der Arbeit benötigten Begriffsbildungen. Sei $X(V,*)$ eine durch die Produktoperation $*$ über dem Vektoraum V definierte Algebra. Im Falle der Gültigkeit des Assoziativitätsgesetzes schreiben wir dafür $A(V,*)$. Falls $A(V,*)$ zusätzlich kommutativ ist, verdeutlichen wir dies durch $A_+(V,*)$, bzw. im Falle der zusätzlichen Antikommutativität durch $A_-(V,*)$. Falls für $A_+(V,*)$ zusätzlich ein Einselement $I_{(*)}$ existiert, so erhält $A_+(V,*)$ eine abelsche Gruppenstruktur, wofür wir $G(V,*)$ schreiben. Zusammenfassend gilt also mit $m = \pm 1$:

$$A_{sign(m)}(V;*): \ (x*y)*z = x*(y*z), \ \text{(a)} \qquad x*y = my*x, \ \text{(b)} \tag{22}$$

$$G(V,*): \ m = +1, \quad \exists I_{(*)} : x*I_{(*)} = I_{(*)}*x = x, \quad \forall x \in V. \ \text{(c)}$$

Falls Mißverständnisse ausgeschlossen sind verwenden wir die Kurzformen $X, A, A_{sign(m)}, G$, ohne Angabe von $V, *$.

Lemma 2: Die Strukturkonstanten der Algebren $A, A_{sign(m)}, G$ genügen den Relationen

$$A: C_{ij}{}^{m} C_{mk}{}^{l} - C_{jk}{}^{m} C_{im}{}^{l} = 0, \ \text{(a)} \quad A_{sign(m)}: \text{zusätzlich } C_{ij}{}^{k} = mC_{ji}{}^{k}, \ \text{(b)} \tag{23}$$

$$G: \text{zusätzlich} \quad I^{m}_{(*)} C_{mh}{}^{k} = \delta^{k}_{h}. \ \text{(c)}$$

Beweis. Für Algebren A schreiben wir das Assoziativgesetz in der Gestalt

$$(g_i * g_j) * g_k = g_i * (g_j * g_k),$$

wobei $\{g_i\}$ eine Vektorbasis von V bezeichnet. Der Strukturkonstantentensor ist per definitionem gegeben durch $g_i * g_j = C_{ij}{}^{m} g_m$. Eine Entwicklung der linken Seite der obigen Gleichung liefert dann

$$(C_{ij}{}^{m} g_m) * g_k = C_{ij}{}^{m}(g_m * g_k) = C_{ij}{}^{m} C_{mk}{}^{l} g_l,$$

und jene der rechten Seite

$$g_i * (g_j * g_k) = g_i * (C_{jk}{}^{m} g_m) = C_{jk}{}^{m}(g_i * g_m) = C_{jk}{}^{m} C_{im}{}^{l} g_l,$$

woraus wegen der linearen Unabhängigkeit der Basisvektoren

$$C_{ij}{}^{m}C_{mk}{}^{l} = C_{jk}{}^{m}C_{im}{}^{l}$$

folgt, also 23a. Der Nachweis von 23b ist trivial. Zum Nachweis von 23c gehen wir von der Komponentenschreibweise der Bedingung $x * I_{(*)} = I_{(*)} * x = x$ aus. Mit der Komponentendarstellung von $I_{(*)}$ gemäß $I_{(*)} = I^{m}_{(*)}g_m$ lautet die Definitionsgleichung

$$(x^h g_h) * (I^m_{(*)}g_m) = x^h I^m_{(*)}C_{hm}{}^{k}g_k = x^k g_k,$$

und nach Koeffizientenvergleich 23c. ■

Nun ist die Artikulation von Algebrensymmetrien mit Hilfe des Strukturkonstantentensors in der obigen Komponentenschreibweise nachteilig. Die Verhältnisse werden transparenter, wenn man folgende Matrixoperatoren verwendet:

Definition 2: Seien A und B beliebige quadratische Matrizen. Dann definieren wir die Symmetrieoperationen

$$[A,B]_{-sign(m)} := AB - mBA, \quad \text{(a)} \qquad \{A,B\}_{-sign(m)} := AB - m(AB)^T. \quad \text{(b)} \qquad (24)$$

Dabei bezeichnet der hochgestellte Index T die Transpositionsoperation. Im Fall $m = 1$ repräsentieren $[A,B]_-$ den *Kommutator* der Matrizen A, B und $\{A,B\}_-$ die *Symmetrisierung* des Produktes AB. Im Fall $m = -1$ liefert $[A,B]_+$ den *Antikommutator* der Matrizen A, B und $\{A,B\}_+$ die *Antisymmetrisierung* des Produktes AB. Man erkennt die Übereinstimmung von $[,]_{-sign(m)}$ mit $\{,\}_{-sign(m)}$ für symmetrische Operanden $A = A^T, B = B^T$, denn es gilt in diesem Fall
$\{A,B\}_{-sign(m)} = AB - mB^TA^T = AB - mBA = [A,B]_{-sign(m)}$. Mit Hilfe dieser Operationen läßt sich die Struktur des in $A_{sign(m)}$ gültigen Assoziativgesetzes formal transparenter machen. Bezeichnen $C^{(j)}_m$ bzw. $C^{(j)m}$, $j = 1,2,3$ die durch Festhalten des j-ten Index m aus dem Strukturkonstantentensor hervorgehenden quadratischen Matrizen.[3] Dann gilt das

Lemma 3: In den Algebren $A, A_{sign(m)}$ genügt der Strukturkonstantentensor stets den Relationen

$$A : [C^{(1)}_i, C^{(2)}_k]_- = 0, \quad \text{(a)} \qquad (25)$$

$$A_{sign(m)} : C^{(1)}_i = mC^{(2)}_i, \quad \text{(b)} \qquad [C^{(1)}_i, C^{(1)}_k]_- = [C^{(2)}_i, C^{(2)}_k]_- = 0. \quad \text{(c)}$$

$$A_+ : \{C^{(2)}_i, C^{(3)}_k\}_- = 0. \quad \text{(d)}$$

Beweis. Wir denken uns in der Assoziativitätsbedingung 23 die Indizes i und k als festgehalten, und setzen $A_j{}^m = C_{ij}{}^m = (C^{(1)}_i)_j{}^m, B_m{}^l = C_{mk}{}^l = (C^{(2)}_k)_m{}^l$ und erhalten so $A_j{}^m B_m{}^l - B_j{}^m A_m{}^l = 0$. Dies bedeutet (Vertauschung der Buchstaben A und B bei gleichbleibender Indexstellung): $AB = BA$, bzw. $[A,B]_- = 0$. Rücktransformation auf die Strukturkonstanten liefert dann die Aussage 25a. Zum Nachweis von 25b beachte man, daß für (anti)kommutative Algebren gemäß 23b stets $C^{(3)l} = m(C^{(3)l})^T$ gilt, und als direkte Folge

$$C^{(1)}_i = mC^{(2)}_i.$$

[3]Wenn Mißverständnisse ausgeschlossen sind, schreiben wir kürzer $C^{(j)}$ anstelle von $C^{(j)m}$▷

Einsetzen in 25a liefert dann 25c. Für den Nachweis von 25d schreiben wir die allgemeine Assoziativbedingung 23a in der Form

$$C_{ij}{}^{m}\, C_{mk}{}^{l} - C_{jk}{}^{m}\, C_{mi}{}^{l} = 0,$$

(Tausch der Indizes m,i imzweiten Summanden), was in A_+ gestattet ist. Wir denken uns nun die Indizes j, l festgehalten und setzen $A_i{}^{m} = C_{ij}{}^{m} = (C_j^{(2)})_i{}^{m}, B_{mk} = C_{mk}{}^{l} = (C^{(3)l})_{mk}^{l}$. Dann lautet die Bedingung $A_i{}^{m}\, B_{mk} - A_k{}^{m}\, B_{mi} = 0$, bzw. nach Heben und Senken der entsprechenden Indizes $A_i{}^{m}\, B_m^{k} - B_i^{m} A^{k}{}_{m} = 0$. Im Unterschied zu oben erscheinen hier nicht bloß die Buchstaben A,B sondern auch die Indexstellungen vertauscht, was den Übergang zur transponierten Matrix (Tausch von Zeilen und Spalten) bedeutet. In symbolischer Schreibweise gilt somit $AB - B^T A^T = 0$, bzw.

$$\{A, B\}_- = 0.$$

Einsetzen der Strukturkonstanten liefert dann 25d. ■

Literaturverzeichnis

[1] Henneberger, Klaus: "Regelung nichtlinearer Systeme auf der Basis bilinearer Approximationsmodelle", Fortschrittberichte VDI, Reihe 8, Nr. 595, Erlangen; 1996, p.78-87

[2] H. J. Sussmann, "Semigroup representations, bilinear approximation of input-output maps, and generalized inputs", in: Proceedings of the International Symposium on Mathematical Systems Theory. Springer-Verlag, Berlin, West Germany; 1976; p.172-91

[3] H. Schwarz, "Nichtlineare Regelungssysteme", Oldenbourg, 1991

[4] R. Starkl, "Finite order representations of infinite dimensional bilinear models of nonlinear systems", Proceedings of the ACC, Denver, June 4-6, 2003, PaperNr. 0-7803-7896-2/03, p.131-36

[5] R. Starkl, "Low order equivalent forms to finite dimensional bilinear representations of nonlinear systems", to appear on 42^{nd} CDC, Hawaii, De. 9.12.2003

Symbolische Kalman Zerlegung

W. Kleissl, M. Hofbaur
TU-Graz, Institut für Regelungstechnik, 8010 Graz, Inffeldgasse 16c
kleissl@irt.tu-graz.ac.at*, hofbaur@irt.tu-graz.ac.at

Zusammenfassung

Wir präsentieren einen Algorithmus, der es durch effiziente Ausnutzung der Struktur des Problems erlaubt, symbolisch die kanonische Aufteilung eines linearen, zeitinvarianten mathematischen Zustandsraummodells bezüglich Steuerbarkeit und Beobachtbarkeit (*Kalman Zerlegung*) zu ermitteln. Das gegebene Modell darf hierzu auch symbolische Parameter enthalten. Im Zuge unserer Berechnungen werden wir uns der Berechnung sogenannter *Gröbner Basen* als Werkzeug zur effizienten und systematischen Lösung der auftretenden Gleichungssysteme bedienen.

1 Einleitung

Die kanonische Zerlegung eines mathematischen Modells bezüglich seiner steuerbaren und beobachtbaren Anteile (*Kalman Zerlegung*) wurde von Kalman in den 1960'ern eingeführt [1] und erfreut sich heute unter Regelungstechnikern praktisch uneingeschränkter Bekanntheit. Auch viele Lehrbücher ([2][3][4][5], um nur einige zu nennen) greifen das Thema "*Kalman Zerlegung*" auf. Jedoch beschränken sie sich häufig darauf, zu erwähnen, dass eine derartige Zerlegung für jedes beliebige mathematische Modell gefunden werden kann und dass man das Modell auf die gewünschte Struktur bringen kann, indem man eine konstante, reguläre Zustandstransformation darauf anwendet. Sie bleiben jedoch häufig schuldig, eine algorithmische Vorschrift anzugeben, wie die dazu nötige Transformationsmatrix $\mathbf{T}$ ermittelt werden kann. Der Leser muss sich damit zufrieden geben, zu erfahren, dass "eine derartige Transformationsmatrix im Allgemeinen nicht leicht gefunden werden kann". Trotzdem stellt auch eine Reihe von Autoren, beginnend mit Kalman selbst [1], fest, dass diese Matrix $\mathbf{T}$ einige interessante Eigenschaften aufweist: Ihre Spalten bilden (blockweise) eine Basis des steuerbaren und beobachtbaren, des steuerbaren aber nicht beobachtbaren, des nicht steuerbaren aber beobachtbaren und des weder steuerbaren noch beobachtbaren Unterraumes des Zustandsraumes des zu zerlegenden mathematischen Modells. Aus diesem Wissen heraus wird zum Beispiel in [6]

*Korrespondenz bitte an diese Adresse

ein detaillierter Algorithmus zur numerischen Ermittlung einer solchen Matrix **T** angegeben. Dieser Algorithmus verwendet Singulärwertzerlegung (SVD) als Werkzeug, um die 4 erwähnten Unterräume aus dem Zustandsraummodell zu konstruieren. Ebenfalls bedient sich [7] der Singulärwertzerlegung, um eine Transformationsmatrix zu bestimmen, die bezüglich einer bestimmten Konditionszahl optimal ist.

Wir hingegen wollen hier einen grundsätzlich anderen Ansatz wählen, um das Problem der *Kalman Zerlegung* zu lösen. Wir werden nicht Wissen über die Struktur der *Transformationsmatrix* nutzen, um einen Algorithmus zu ihrer Ableitung zu entwickeln, sondern wir werden danach trachten, Wissen über die *Lösung*, also über die *Struktur des zerlegten Modells* zu unserem Vorteil zu nutzen. Das wird es uns gestatten, einen Algorithmus zu entwickeln, der auf rein symbolischen Berechnungen aufbaut und der es deshalb ermöglicht, auch Modelle zu zerlegen, die symbolische Parameter enthalten (parametrierte Modelle). Wir werden also die Möglichkeit schaffen, den Einfluss verschiedener Modellparameter anhand der *kalmanzerlegten* Struktur des Modells zu untersuchen, was in vielen Fällen zu größerer Einsicht führen kann. Zusätzlich wird unser Algorithmus aufgrund seiner symbolischen Natur auch mit numerisch beliebig schlecht konditionierten Modellen perfekt umgehen können und deshalb für solche Modelle bekannten numerischen Methoden auch ohne das Auftreten symbolischer Parameter überlegen sein. Der Algorithmus wird darauf aufbauen, einen Satz nichtlinearer, aber -salopp gesprochen- "angenehm" strukturierter Gleichungen aufzustellen, die dann durch die Berechnung sogenannter *Gröbner Basen* ([8] [9][10][11]) einfach gelöst werden können.

Diese Arbeit gliedert sich wie folgt: Abschnitt 2 wird in kurzen Worten die prinzipielle Problemstellung und unseren "lösungsbasierten" Ansatz vorstellen, bevor in Abschnitt 3 ein Beispiel veranschaulichen soll, wie das Problem algorithmisch durch die Berechnung von *Gröbner Basen* gelöst werden kann. Diese Abschnitte sollen hauptsächlich dazu dienen, eine Motivation dafür zu liefern, eine symbolische *Kalman Zerlegung* unter Zuhilfenahme der Berechnung von *Gröbner Basen* durchzuführen. In Abschnitt 4 werden wir die gesammelten Ideen sodann benutzen, um Schritt für Schritt von Anfang an einen Algorithmus zu entwickeln und im Detail zu beschreiben, der es uns gestatten wird, symbolisch eine *Kalman Zerlegung* für beliebige (eventuell parametrierte) mathematische Zustandsraummodelle zu ermitteln. In Abschnitt 5 werden wir schließlich ein Beispiel dazu präsentieren, bevor die Arbeit in Abschnitt 6 mit einer Zusammenfassung abgeschlossen wird. Obwohl eine Erweiterung auf Modelle mit mehreren Ein- und Ausgängen prinzipiell möglich wäre, werden wir uns im Rahmen dieser Arbeit auf Modelle mit nur jeweils einem Ein- und Ausgang beschränken, um die vorteilhafte Ausnutzung der Struktur des zerlegten Modells besser zeigen zu können.

Da die Arbeit mit *Gröbner Basen* unserer Meinung nach noch nicht zu den Standardwerkzeugen eines typischen Regelungstechnikers gehört, werden wir zusätzlich in Appendix A eine kurze Einführung in die Rechnung mit *Gröbner Basen* geben, soweit dies für das genaue Verständnis unseres Algorithmus notwendig ist.

2 Problemformulierug

Wir gehen von einem linearen, zeitinvarianten (LZI) mathematischen Zustandsraummodell mit einem Eingang u und einem Ausgang y

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A} \cdot \mathbf{x} + \mathbf{b} \cdot u \\ y &= \mathbf{c}^T \cdot \mathbf{x} \end{aligned} \tag{1}$$

aus. Dieses wollen wir in seine bezüglich ihrer Steuerbarkeit und Beobachtbarkeit unterschiedlichen Teilmodelle aufspalten (*Kalman Zerlegung*) und so ein Modell

$$\begin{aligned} \dot{\mathbf{z}} &= \tilde{\mathbf{A}} \cdot \mathbf{z} + \tilde{\mathbf{b}} \cdot u \\ y &= \tilde{\mathbf{c}}^T \cdot \mathbf{z} \end{aligned} \tag{2}$$

erhalten, wobei dementsprechend gelten muss

$$\begin{aligned} \mathbf{z} &= \begin{bmatrix} \mathbf{z}_{sb} & \mathbf{z}_{s-} & \mathbf{z}_{-b} & \mathbf{z}_{--} \end{bmatrix}^T \\ \tilde{\mathbf{A}} &= \begin{bmatrix} \tilde{\mathbf{A}}_{sb} & \mathbf{0} & \tilde{\mathbf{A}}_{13} & \mathbf{0} \\ \tilde{\mathbf{A}}_{21} & \tilde{\mathbf{A}}_{s-} & \tilde{\mathbf{A}}_{23} & \tilde{\mathbf{A}}_{24} \\ \mathbf{0} & \mathbf{0} & \tilde{\mathbf{A}}_{-b} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \tilde{\mathbf{A}}_{43} & \tilde{\mathbf{A}}_{--} \end{bmatrix} \\ \tilde{\mathbf{b}} &= \begin{bmatrix} \tilde{\mathbf{b}}_{sb}^T & \tilde{\mathbf{b}}_{s-}^T & 0^T & 0^T \end{bmatrix}^T \\ \tilde{\mathbf{c}}^T &= \begin{bmatrix} \tilde{\mathbf{c}}_{sb}^T & \mathbf{0} & \tilde{\mathbf{c}}_{-b}^T & \mathbf{0} \end{bmatrix}. \end{aligned} \tag{3}$$

Die hier auftretenden Null-Blöcke wollen wir im folgenden mit "*strukturelle Nullen*" bezeichnen. Dieses *kalmanzerlegte* Modell zeichnet sich dadurch aus, dass die bezüglich ihrer Steuerbarkeit/Beobachtbarkeit unterschiedlichen Teilsysteme auf den ersten Blick erkannt werden können, da diese Eigenschaften schon in der Modellstruktur widergespiegelt werden. Diese Struktur ist in Abbildung 1 dargestellt, wobei die dicken Linien die

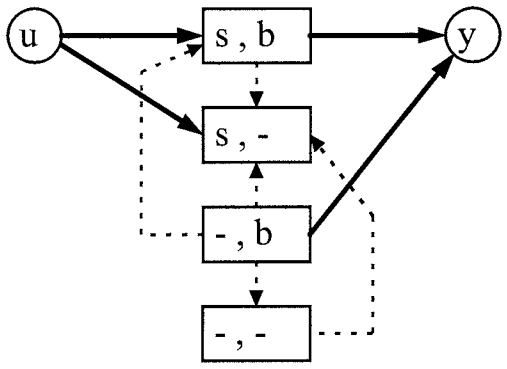


Abbildung 1: Struktur des *kalmanzerlegten* Modells

Vektoren $\tilde{\mathbf{b}}$ und $\tilde{\mathbf{c}}$ im Modell repräsentieren und die strichlierten Linien die Einflüsse der

verschiedenen Teilsysteme untereinander (Ausserdiagonalelemente von $\tilde{\mathbf{A}}$) kennzeichnen. Unser Problem besteht nun darin, eine konstante, reguläre Matrix $\mathbf{T}$ zu finden, sodass die Zustandstransformation

$$\begin{aligned} \mathbf{x} &= \mathbf{T} \cdot \mathbf{z} \\ \tilde{\mathbf{A}} &= \mathbf{T}^{-1} \cdot \mathbf{A} \cdot \mathbf{T} \\ \tilde{\mathbf{b}} &= \mathbf{T}^{-1} \cdot \mathbf{b} \\ \tilde{\mathbf{c}}^T &= \mathbf{c}^T \cdot \mathbf{T} \end{aligned} \tag{4}$$

zum gewünschten Ergebnis führt. Dazu wollen wir die Information ausnutzen, die wir bereits über die gewünschte Lösung unseres Problems haben - die Struktur des zerlegten Modells. Indem wir das tun, erhalten wir eine Problembeschreibung, die intern ein großes Maß an Struktur zeigt, was uns helfen wird, den Berechnungsaufwand bei der Ermittlung von $\mathbf{T}$ drastisch zu reduzieren.

Ohne uns vorerst noch mit strukturell bedingten Vereinfachungen zu beschäftigen, könnte ein intuitiver Ansatz zur symbolischen Berechnung einer geeigneten Transformationsmatrix $\mathbf{T}$ der sein, ein Gleichungssystem wie (4) mit einem $\mathbf{T}$, das nur symbolische Werte enthält

$$\mathbf{T} = \begin{bmatrix} t_{11} & t_{12} & \cdots \\ t_{21} & t_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{5}$$

aufzustellen und mit einem geeigneten Lösungsverfahren die Elemente t_{ij} der Transformationsmatrix zu bestimmen. Hier lässt sich auch sofort eine Möglichkeit erkennen, wie die Lösungsstruktur sinnvoll zur Vereinfachung der Berechnungen eingesetzt werden kann: Es müssen nur die Gleichungen, die den *strukturellen Nullen* des transformierten Modells entsprechen, ausgewertet werden, da das Vorhandensein dieser Nullen schon hinreichend für das Vorliegen eines *kalmanzerlegten* Modells ist.

Es zeigt sich jedoch, dass die hier auftretenden Gleichungen selbst bei relativ kleinen Problemen (Modelle mit vier oder fünf Zustandsvariablen) schon extrem lang werden, da die nötige symbolische Inversion der Matrix $\mathbf{T}$ zu riesigen Ausdrücken führt. Wir werden das Problem deshalb etwas umformulieren

$$\begin{aligned} \mathbf{0} &= \mathbf{A} \cdot \mathbf{T} - \mathbf{T} \cdot \tilde{\mathbf{A}} \\ \mathbf{0} &= \mathbf{b} - \mathbf{T} \cdot \tilde{\mathbf{b}} \\ \mathbf{0} &= \mathbf{c}^T \cdot \mathbf{T} - \tilde{\mathbf{c}}^{\mathbf{T}}, \end{aligned} \tag{6}$$

sodass diese Inversion umgangen wird und die einzelnen Gleichungen extrem verkürzt werden. Das handelt uns leider den Nachteil ein, dass wir *alle* Gleichungen von (6) auswerten müssen, da die Zuordnung einzelner Gleichungen zu *strukturellen Nullen* verloren geht. Jedoch ist auch dieser Nachteil nicht so groß, wie es auf den ersten Blick scheinen mag! Die *strukturellen Nullen* vereinfachen die Gleichungen zusätzlich noch, indem sie einige Parameter t_{ij} mit null multiplizieren und so ihren Einfluss auf die Gleichung eliminieren. Zudem erfolgt diese Eliminierung noch in einer strukturierten Art und Weise,

sodass es uns möglich wird, das gesamte Gleichungssystem in Teile aufzuspalten, die wir nacheinander auswerten können. Mehr dazu in Abschnitt 4.

Prinzipiell haben wir also schon eine Möglichkeit gefunden, wie wir die *Kalman Zerlegung* für ein gegebenes Modell symbolisch berechnen können. Was uns allerdings noch fehlt ist eine algorithmische Methode zur Ermittlung der Lage der *strukturellen Nullen* im zerlegten Modell. Und schließlich müssen wir uns auch noch mit der systematischen Lösung des Gleichungssystems auseinandersetzen. Zwei nichttriviale Aufgaben, die wir aber im Folgenden durch die Verwendung eines Werkzeugs zur Berechnung von *Gröbner Basen* bewältigen werden.

3 Einführendes Beispiel

3.1 Einfaches Beispiel

In diesem Abschnitt wollen wir anhand eines akademischen Beispiels zeigen, welcher Art die Gleichungen (6) sind und wie wir diese systematisch durch die Berechnung von *Gröbner Basen* lösen können. Wir stellen uns also die Aufgabe, eine *Kalman Zerlegung* zum Modell

$$\begin{aligned} \dot{\mathbf{x}} &= \begin{bmatrix} 1 & -1 \\ -2 & 0 \end{bmatrix} \cdot \mathbf{x} + \begin{bmatrix} 1 \\ 2 \end{bmatrix} \cdot u \\ y &= \begin{bmatrix} 2 & 1 \end{bmatrix} \cdot \mathbf{x} \end{aligned} \tag{7}$$

zu bestimmen. Dazu stellen wir zuerst einmal fest (zum Beispiel durch Berechnung der Steuerbarkeits- und Beobachtbarkeitsmatrix), dass das Modell beobachtbar ist (Beobachtbarkeitsmatrix hat den Rang 2), dass Steuerbarkeit allerdings nicht gegeben ist (Steuerbarkeitsmatrix hat nur den Rang 1). Wir schließen daraus, dass das *kalmanzerlegte* Modell ein steuerbares und beobachtbares Teilmodell und ein beobachtbares aber nicht steuerbares Teilmodell aufweisen wird. Die Struktur des zerlegten Modells wird dementsprechend wie folgt aussehen

$$\begin{aligned} \begin{bmatrix} z_{sb} \\ z_{-b} \end{bmatrix} &= \begin{bmatrix} a_{sb} & a_{12} \\ 0 & a_{-b} \end{bmatrix} \cdot \begin{bmatrix} z_{sb} \\ z_{-b} \end{bmatrix} + \begin{bmatrix} b_{sb} \\ 0 \end{bmatrix} \cdot u \\ y &= \begin{bmatrix} c_{sb} & c_{-b} \end{bmatrix} \cdot \begin{bmatrix} z_{sb} \\ z_{-b} \end{bmatrix}. \end{aligned} \tag{8}$$

Wir setzen nun mit symbolischen Parametern eine Transformationsmatrix

$$\mathbf{T} = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} \tag{9}$$

an, mit der wir das Gleichungssystem (6) aufstellen

$$\begin{aligned}
\mathbf{0} &= \begin{bmatrix} 1 & -1 \\ -2 & 0 \end{bmatrix} \cdot \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} - \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} \cdot \begin{bmatrix} a_{sb} & a_{12} \\ 0 & a_{-b} \end{bmatrix} \\
\mathbf{0} &= \begin{bmatrix} 1 \\ 2 \end{bmatrix} - \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} \cdot \begin{bmatrix} b_{sb} \\ 0 \end{bmatrix} \\
\mathbf{0} &= \begin{bmatrix} 2 & 1 \end{bmatrix} \cdot \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} - \begin{bmatrix} c_{sb} & c_{-b} \end{bmatrix}.
\end{aligned} \tag{10}$$

Sehen wir uns die hier auftretenden Gleichungen genauer an, stellen wir fest, dass auf der rechten Seite des Gleichheitszeichens ausschließlich Polynome stehen. Wir müssen uns also ausschließlich mit Gleichungen der Form

$$0 = p(\chi_1, ..., \chi_j) \tag{11}$$

befassen, wobei p ein Polynom kennzeichnet und χ_1 bis χ_j für die symbolischen Parameter in (10) stehen. Das ist eine äußerst zufriedenstellende Erkenntnis, da die *Gröbner Basen* Theorie die algorithmische Lösung eines solchen Gleichungssystems ermöglicht, sofern diese Lösung existiert. Ganz am Ziel sind wir allerdings immer noch nicht, da wir durch den Ansatz der Transformationsmatrix $\mathbf{T}$ nach (9) zwar deren Konstanz festgelegt haben, allerdings noch keinerlei Garantie für deren Regularität erhalten haben. Um dies zu erreichen, fordern wir zusätzlich, dass

$$\det(\mathbf{T}) \neq 0. \tag{12}$$

Hierbei handelt es sich leider um eine Ungleichung, die nicht in das Schema (11) passt. Wir machen uns nun aber den Umstand zunutze, dass, wie schon von Kalman [1] gezeigt, die Transformation eines gegebenen Modells in seine *kalmanzerlegte* Form nicht eindeutig ist, sondern lediglich die Struktur des zerlegten Modells eindeutig bestimmt ist. Hat man z.B. eine Transformationsmatrix gefunden, die auf die gewünschte Zerlegung führt, so führen auch Transformationen mit Vielfachen der Transformationsmatrix auf dieselbe Struktur. (Multipliziere Gleichungen (10) mit einer Konstanten ungleich 0!) Diese Erkenntnis ermöglicht uns -ohne Einschränkung(!)- die Determinante der Transformationsmatrix auf einen beliebigen Wert (z.B. 1) festzulegen. Damit können wir aus der Ungleichung (12) eine Gleichung machen

$$0 = t_{11} \cdot t_{22} - t_{12} \cdot t_{21} - 1, \tag{13}$$

die sich in das Schema (11) einordnen lässt.

3.2 Lösung durch Berechnung von Gröbner Basen

Eine kurze Einführung in die Berechnung von *Gröbner Basen* findet sich in Appendix A. Umfassendere Informationen liefern [8], [9], [11], und eine Reihe von Anwendungsmöglichkeiten von *Gröbner Basen* im Bereich der Regelungstechnik werden in [10] vorgestellt.

Für unsere Zwecke ist es ausreichend, die Berechnung von *Gröbner Basen* als das polynomische Äquivalent zur bekannten Gauß-Elimination für lineare Gleichungssysteme zu betrachten. Im Vorigen haben wir gezeigt, dass wir zur Bestimmung der Transformationsmatrix ein Gleichungssystem der Form

$$\begin{aligned} 0 &= p_1(\chi_1, ..., \chi_j) \\ 0 &= p_2(\chi_1, ..., \chi_j) \\ &\vdots \\ 0 &= p_k(\chi_1, ..., \chi_j) \end{aligned} \tag{14}$$

lösen müssen. Dieses wollen wir in ein anderes Gleichungssystem

$$\begin{aligned} 0 &= g_1(\chi_1) \\ 0 &= g_2(\chi_1, \chi_2) \\ &\vdots \\ 0 &= g_l(\chi_1, ..., \chi_j) \end{aligned} \tag{15}$$

überführen, das die selben Lösung wie das ursprüngliche besitzen soll. g_1 bis g_l werden dabei als die *Gröbner Basen* der Polynome p_1 bis p_k bezeichnet und können wie in Appendix A beschrieben berechnet werden. Die Gleichungen (15) können wir nun eine nach der anderen auswerten und müssen sie jeweils nur nach einer Variablen lösen. Wir wollen an dieser Stelle nicht näher darauf eingehen, wie solche *Gröbner Basen* berechnet werden. Wir wollen uns lediglich damit auseinandersetzen, was für uns als Benutzer eines Werkzeugs zur Berechnung von *Gröbner Basen* (z.B. Maple) wichtig ist. Hier ist zunächst zu erwähnen, dass wir innerhalb der Variablen χ_1 bis χ_j eine "Rangordnung" festlegen müssen. Diese ist wichtig, da sie bestimmt, wie die Variablen nacheinander in den *Gröbner Basen* g_1 bis g_l auftreten und dadurch deren Aussehen beeinflussen. Eine naheliegende Möglichkeit ist die Ordnung

$$\chi_1 \prec \chi_2 \prec \cdots \prec \chi_j. \tag{16}$$

Diese wird in der Literatur mit *rein lexicographische Ordnung* bezeichnet. Damit legen wir fest, dass χ_1 einen niedrigeren Rang hat als χ_2 usw. Diese Definition der Rangordnung impliziert weiters, dass

$$\chi_1 \prec \chi_1^2 \prec \cdots \prec \chi_1^n \prec \chi_2 \prec \chi_1\chi_2 \prec \cdots, \tag{17}$$

also dass jeder Term, der in irgendeiner Form die Variable χ_2 enthält einen höheren Rang hat (erst in einer späteren *Gröbner Basis* g_i auftreten wird), als jeder Term, der nur die Variable χ_1 enthält. Damit erreichen wir genau *Gröbner Basen*, wie sie in (15) auftreten.

Wichtig ist auch noch, festzuhalten, dass der von Buchberger präsentierte (in Appendix A kurz zusammengefasste) Algorithmus - sofern (14) lösbar ist - immer *Gröbner Basen* g_1 bis g_l wie in (15) liefern wird, sodass (14) und (15) dieselben Lösungen haben. Die Existenz einer Lösung zu (14) können wir nach Kalman [1] voraussetzen, was uns garantiert, auch entsprechende *Gröbner Basen* zu erhalten.

3.3 Lösung des Beispiels

Bezüglich des oben eingeführten Beispiels erfolgt die Lösung [Gleichungen (10) und (13)] wie folgt: Wir definieren innerhalb der Variablen eine Ordnung, welche die Variablen in **T** rangniedriger einordnet, als jene in **A**, **b** und **c**. Diese werden intern jeweils nach Zeilennummer und Spaltennummer geordnet

$$t_{11} \prec t_{12} \prec t_{21} \prec t_{22} \prec a_{sb} \prec a_{12} \prec a_{-b} \prec b_{sb} \prec c_{sb} \prec c_{-b}.$$

Mit dieser Rangordnung unter Zuhilfenahme des "grobner" Pakets in Maple V berechnen wir *Gröbner Basen* zu

$$\begin{aligned} 0 &= t_{21} - 2 \cdot t_{11} \\ 0 &= -1 - 2 \cdot t_{12} \cdot t_{11} + t_{11} \cdot t_{22} \\ 0 &= 1 + a_{sb} \\ 0 &= a_{12} + t_{22}^2 - t_{22} \cdot t_{12} - 2 \cdot t_{12}^2 \\ 0 &= a_{-b} - 2 \\ 0 &= 2 \cdot t_{12} - t_{22} + b_{sb} \\ 0 &= c_{sb} - 4 \cdot t_{11} \\ 0 &= -2 \cdot t_{12} - t_{22} + c_{-b}. \end{aligned} \tag{18}$$

Wir stellen fest, dass die erste Gleichung - auf den ersten Blick anders als erwartet - mehr als eine Variable enthält. Müssen wir doch Gleichungen nach mehr als einer Variablen lösen? Die Antwort ist nein! Die Sache ist lediglich die, dass das ursprüngliche Gleichungssystem nicht wohlbestimmt ist und mehrere Lösungen besitzt. Genauso besitzt auch das "neue" Gleichungssystem mehrere Lösungen. Sind wir - wie in unserem Fall - lediglich an irgendeiner Lösung interessiert, ist es uns gestattet, verschiedene Parameter nach eigenem Gutdünken zu wählen. Beginnen wir mit der "obersten" Gleichung, wählen alle bis auf einen Parameter ungleich null und lösen nach dem verbleibenden, bevor wir zur nächsten Gleichung weitergehen, so können wir nach der Theorie der *Gröbner Basen* sicher sein, dass alle weiteren Gleichungen ebenfalls lösbar sein werden und die schlussendlich gefundene Lösung auch das ursprüngliche Gleichungssystem lösen wird. Wie wir die Parameter wählen, ist uns grundsätzlich freigestellt. Sie algorithmisch als 1 zu wählen hat in den bisher von uns durchgeführten Zerlegungen gut funktioniert.

Wir wählen in der ersten Gleichung $t_{11} = 1$ und berechnen anschließend $t_{21} = 2$. Die zweite Gleichung vereinfacht sich damit zu $0 = -1 - 2 \cdot t_{12} + t_{22}$, wobei wir wieder einen Parameter wählen ($t_{12} = 1$) und den verbleibenden berechnen ($t_{22} = 3$). Auf diese Weise setzen wir fort, bis wir als endgültiges Ergebnis für das *kalmanzerlegte* Modell

$$\begin{aligned} \begin{bmatrix} z_{sb} \\ z_{-b} \end{bmatrix} &= \begin{bmatrix} -1 & -4 \\ 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} z_{sb} \\ z_{-b} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \cdot u \\ y &= \begin{bmatrix} 4 & 5 \end{bmatrix} \cdot \begin{bmatrix} z_{sb} \\ z_{-b} \end{bmatrix} \end{aligned} \tag{19}$$

mit der entsprechenden Transformationsmatrix

$$\mathbf{T} = \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix} \tag{20}$$

vorliegen haben. Wir haben unser Beispiel damit gelöst. Es ist allerdings noch anzumerken, dass der Berechnungsaufwand zur Ermittlung von *Gröbner Basen* (unter "worst case" Bedingungen) exponentiell mit der Problemgröße ansteigt, weshalb es für umfangreichere Problemstellungen nötig sein wird, weitere Maßnahmen zu ergreifen, um das Gesamtproblem in mehrere (kleinere) Teilprobleme zu unterteilen.

4 Algorithmus

In Abschnitt 4 werden wir die bisher gesammelten Ideen benutzen, um Schritt für Schritt von Anfang an einen Algorithmus zu entwickeln und im Detail zu beschreiben, der es uns gestatten wird, symbolisch eine *Kalman Zerlegung*

$$\begin{aligned} \dot{\mathbf{z}} &= \tilde{\mathbf{A}} \cdot \mathbf{z} + \tilde{\mathbf{b}} \cdot u \\ y &= \tilde{\mathbf{c}}^T \cdot \mathbf{z} \end{aligned} \tag{21}$$

mit

$$\begin{aligned} \mathbf{z} &= \begin{bmatrix} \mathbf{z}_{sb} & \mathbf{z}_{s-} & \mathbf{z}_{-b} & \mathbf{z}_{--} \end{bmatrix}^T \\ \tilde{\mathbf{A}} &= \begin{bmatrix} \tilde{\mathbf{A}}_{sb} & \mathbf{0} & \tilde{\mathbf{A}}_{13} & \mathbf{0} \\ \tilde{\mathbf{A}}_{21} & \tilde{\mathbf{A}}_{s-} & \tilde{\mathbf{A}}_{23} & \tilde{\mathbf{A}}_{24} \\ \mathbf{0} & \mathbf{0} & \tilde{\mathbf{A}}_{-b} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \tilde{\mathbf{A}}_{43} & \tilde{\mathbf{A}}_{--} \end{bmatrix} \\ \tilde{\mathbf{b}} &= \begin{bmatrix} \tilde{\mathbf{b}}_{sb}^T & \tilde{\mathbf{b}}_{s-}^T & 0^T & 0^T \end{bmatrix}^T \\ \tilde{\mathbf{c}}^T &= \begin{bmatrix} \tilde{\mathbf{c}}_{sb}^T & \mathbf{0} & \tilde{\mathbf{c}}_{-b}^T & \mathbf{0} \end{bmatrix}. \end{aligned} \tag{22}$$

für beliebige (eventuell parametrierte) mathematische Zustandsraummodelle

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A} \cdot \mathbf{x} + \mathbf{b} \cdot u \\ y &= \mathbf{c}^T \cdot \mathbf{x} \end{aligned} \tag{23}$$

und die zugehörige Transformation $\mathbf{x} = \mathbf{T} \cdot \mathbf{z}$ zu ermitteln. Dazu werden wir uns zuerst in Abschnitt 4.1 damit beschäftigen, die Lage der *strukturellen Nullen* (Nullen in 22) zu ermitteln. Wir werden dazu nicht, wie intuitiv vielleicht erwartet, ausschließlich mit Steuerbarkeits- und Beobachtbarkeitsmatrizen arbeiten, sondern auch mit charakteristischen Polynomen. Dies wird es uns ermöglichen, einige Parameter im Gleichungssystem sofort festzulegen und darüber hinaus in Abschnitt 4.2 eine Problembeschreibung zu erarbeiten, die eine Unterteilung in mehrere Teilprobleme zulässt, bevor wir das gestellte Problem schließlich in Abschnitt 4.3 algorithmisch lösen.

4.1 Ermittlung der verschiedenen Modellteile

4.1.1 Ermittlung der Dimensionen der Modellteile

Die Dimension des steuerbaren Teilmodells ($\dim\left(\left[\begin{array}{cc}\mathbf{z}_{sb} & \mathbf{z}_{s-}\end{array}\right]^T\right)$ in (22)) können wir durch die Bestimmung des Ranges der Steuerbarkeitsmatrix

$$\mathbf{S}_u = \left[\begin{array}{cccc}\mathbf{b} & \mathbf{A}\cdot\mathbf{b} & \cdots & \mathbf{A}^{n-1}\cdot\mathbf{b}\end{array}\right]$$

des gegebenen Modells (23) ermitteln. Analog dazu erhalten wir die Dimension des beobachtbaren Teilmodells ($\dim\left(\left[\begin{array}{cc}\mathbf{z}_{sb} & \mathbf{z}_{-b}\end{array}\right]^T\right)$) durch die Bestimmung des Ranges der Beobachtbarkeitsmatrix

$$\mathbf{B}_y = \left[\begin{array}{c}\mathbf{c}^T \\ \mathbf{c}^T\cdot\mathbf{A} \\ \vdots \\ \mathbf{c}^T\cdot\mathbf{A}^{n-1}\end{array}\right].$$

Unbekannt ist uns aber immer noch die Dimension des steuerbaren und beobachtbaren Teilmodells ($\dim\left([\mathbf{z}_{sb}]^T\right)$). Diese könnten wir durch die Bestimmung des Ranges von $\mathbf{B}_y\cdot\mathbf{S}_u$ ermitteln und so die Dimensionen der 4 gesuchten Teilmodelle

$$\begin{aligned} n_{sb} &: = \dim\left([\mathbf{z}_{sb}]^T\right) \\ n_{s-} &: = \dim\left([\mathbf{z}_{s-}]^T\right) \\ n_{-b} &: = \dim\left([\mathbf{z}_{-b}]^T\right) \\ n_{--} &: = \dim\left([\mathbf{z}_{--}]^T\right) \end{aligned}$$

bestimmen.

Alternativ zur Rangbestimmung von $\mathbf{B}_y\cdot\mathbf{S}_u$ wollen wir uns aber einer anderen Methode bedienen, die es uns in weiterer Folge gestatten wird, einige Parameter im Gleichungssystem sofort festzulegen.

4.1.2 Steuerbares und beobachtbares Teilmodell

Es ist eine wohlbekannte Tatsache, dass durch eine (konstante) Rückkopplung

$$u = k\cdot y \tag{24}$$

nur steuerbare und beobachtbare Eigenwerte des Modells (23) beeinflusst werden können. In einem ersten Schritt versuchen wir also, die Eigenwerte der Dynamikmatrix $\mathbf{A}$ des ursprünglichen Modells (23) und der Dynamikmatrix $(\mathbf{A}+k\cdot\mathbf{b}\cdot\mathbf{c}^\mathbf{T})$ des mit (24) rückgekoppelten Modells zu vergleichen. Jene Eigenwerte von $\mathbf{A}$, die in $(\mathbf{A}+k\cdot\mathbf{b}\cdot\mathbf{c}^\mathbf{T})$ nicht mehr angetroffen werden, tragen folglich zum steuerbaren und beobachtbaren Teilmodell

bei. Selbstverständlich ändert eine konstante, reguläre Zustandstransformation keine Eigenwerte, weshalb diese zum steuerbaren und beobachtbaren Teilmodell beitragenden Eigenwerte auch Eigenwerte von $\tilde{\mathbf{A}}_{sb}$ in (22) sein müssen.

Symbolische Berechnung von Eigenwerten stellt jedoch für Modelle höherer Ordnung oft ein Problem dar. Dieses Problem können wir aber leicht umgehen, da wir ja im Grunde weniger an den Eigenwerten s_i selbst, als vielmehr an deren Anzahl n_{sb} (und als Vorarbeit für Abschnitt 4.2 am zugehörigen charakteristischen Polynom $\prod_{i=1}^{n_{sb}}(s - s_i)$) interessiert sind (Wir werden in Zukunft das charakteristische Polynom einer Matrix $\mathbf{M}$ mit

$$\Delta(\mathbf{M}) := \det(s\mathbf{E} - \mathbf{M}) \tag{25}$$

abkürzen). Wir vergleichen also nicht die Eigenwerte selbst, sondern die charakteristischen Polynome $\Delta(\mathbf{A})$ und $\Delta(\mathbf{A} + k \cdot \mathbf{b} \cdot \mathbf{c}^{\mathrm{T}})$, indem wir deren größten gemeinsamen Teiler (ggT) suchen. Algorithmen hierzu sind in der Literatur wohlbekannt und in der Praxis erprobt ([12],[13]). Einer der ältesten und einfachsten Algorithmen ist der sogenannte Euklidische Algorithmus, der für unsere Zwecke allerdings ausreichend ist und auch eine gute Einsicht in den Einfluss verschiedener symbolischer Modellparameter auf Steuerbarkeit und Beobachtbarkeit bietet.

```
Euklidischer Algorithmus:
definiere Polynome p1, p2
r2 ← p1, r3 ← p2
solange (r3 nicht null)
    r1 ← r2, r2 ← r3
    euklidische Division:  r1 = q · r2 + r3
ende solange
r2 ist der ggT von p1, p2
STOP
```

Der dadurch ermittelte ggT entspricht dem Anteil von $\Delta(\mathbf{A})$, der in $\Delta(\mathbf{A} + k \cdot \mathbf{b} \cdot \mathbf{c}^{\mathrm{T}})$ wiedergefunden wurde, also dem Produkt $\prod_i(s - s_i)$ über all jene Eigenwerte, die durch die Rückkopplung *nicht* geändert werden konnten.

Um nun das zu $\tilde{\mathbf{A}}_{sb}$ in (22) gehörige charakteristische Polynom $\Delta(\tilde{\mathbf{A}}_{sb})$ zu erhalten, müssen wir $\Delta(\mathbf{A})$ noch durch den gefundenen ggT dividieren. Während der Algorithmus fortschreitet und die euklidischen Divisionen $r_1 = q \cdot r_2 + r_3$ durchgeführt werden, erhält man durch Auswertung der Bedingung $r_3 = 0$ Informationen darüber, welche Bedingungen die Modellparameter nicht erfüllen dürfen, damit das Modell nicht an Steuerbarkeit oder Beobachtbarkeit verliert. Der Rest r_3 wird im Allgemeinen ein Polynom in s sein, dessen Koeffizienten gebrochen rationale Funktionen in den Modellparametern sind. Um die Bedingung $r_3 = 0$ zu erfüllen, müssen die Zählerpolynome(!) aller Koeffizienten dieses Restpolynoms gleich 0 sein. Das entsprechende Gleichungssystem kann deshalb durch Berechnung von *Gröbner Basen* vereinfacht und ausgewertet werden.

4.1.3 Restliche Teilmodelle

In ganz ähnlicher Weise können wir eventuell zusätzlich noch die zu den weiteren Teilmodellen gehörigen charakteristischen Polynome $\Delta(\tilde{\mathbf{A}}_{s-})$, $\Delta(\tilde{\mathbf{A}}_{-b})$ und $\Delta(\tilde{\mathbf{A}}_{--})$ bestim-

men. Sind wir z.B. an $\Delta(\tilde{\mathbf{A}}_{s-})$ interessiert, so verwenden wir ebenfalls eine konstante Rückkopplung der Form

$$u = k \cdot y,$$

allerdings arbeiten wir nun mit einem modifizierten Ausgangsvektor des Modells (23), der rein aus symbolischen Werten besteht

$$\mathbf{c}^T_{steuerbar} = \left[\begin{array}{ccc} c_1 & \cdots & c_n \end{array} \right].$$

Durch diese Modifikation versuchen wir, das Modell künstlich beobachtbar zu machen. Ob dies erfolgreich war, werden wir an späterer Stelle einfach überprüfen. Setzen wir nun voraus, dass es uns gelungen ist, das gesamte Modell beobachtbar zu machen, so werden wir in $\Delta(\mathbf{A}+\mathbf{b} \cdot \mathbf{c}^{\mathbf{T}}_{\mathbf{steuerbar}})$ nur noch den Anteil von $\Delta(\mathbf{A})$ wiederfinden, der den nicht steuerbaren Eigenwerten entspricht. Der Quotient von $\Delta(\mathbf{A})$ und $ggT(\Delta(\mathbf{A}+\mathbf{b} \cdot \mathbf{c}^{\mathbf{T}}_{\mathbf{steuerbar}}),\Delta(\mathbf{A}))$ entspricht also dem Produkt $\prod_i (s - s_i)$ über alle *steuerbaren* Eigenwerte von $\mathbf{A}$. Bevor wir diese Information allerdings verwenden dürfen, ist es nötig zu überprüfen, ob unsere Annahme der Beobachtbarkeit des mit $\mathbf{c}^T_{steuerbar}$ modifizierten Modells auch zutreffend ist. Das ist genau dann der Fall, wenn der Rang dieses Polynoms dem Rang von $\mathbf{B}_y$ des ursprünglichen Modells entspricht. Ist diese Bedingung nicht erfüllt, so müssen wir das gefundene Polynom verwerfen. Ist sie allerdings erfüllt, so müssen wir das Polynom nur noch durch $\Delta(\tilde{\mathbf{A}}_{sb})$ dividieren (auch das zu den steuerbaren und beobachtbaren Eigenwerten gehörige charakteristische Polynom ist selbstverständlich enthalten), um $\Delta(\tilde{\mathbf{A}}_{s-})$ zu erhalten.

$\Delta(\tilde{\mathbf{A}}_{-b})$ ermitteln wir genauso, nur dass wir diesmal anstelle von $\mathbf{c}^T_{steuerbar}$ einen symbolischen Eingangsvektor $\mathbf{b}_{beobachtbar}$ verwenden. $\Delta(\tilde{\mathbf{A}}_{--})$ erhalten wir anschließend, indem wir $\Delta(\mathbf{A})$ nacheinander durch $\Delta(\tilde{\mathbf{A}}_{sb})$, $\Delta(\tilde{\mathbf{A}}_{s-})$ und $\Delta(\tilde{\mathbf{A}}_{-b})$ dividieren.

Auf diese Art und Weise haben wir sowohl die Größen der Teilmodelle festgelegt (und damit die Lage der *strukturellen Nullen* bestimmt), als auch die zu den Teilmodellen gehörigen charakteristischen Polynome erhalten. Als zusätzliche Information haben wir auch noch Bedingungen erhalten, welche die symbolischen Modellparameter erfüllen müssen, damit der steuerbare bzw. beobachtbare Modellteil möglichst groß ist.

4.1.4 Beispiel

Rekapitulieren wir noch einmal das Beispiel aus Abschnitt 3, nur soll der Ausgangsvektor diesmal einen symbolischen Parameter c enthalten.

$$\begin{aligned} \dot{\mathbf{x}} &= \left[\begin{array}{cc} 1 & -1 \\ -2 & 0 \end{array} \right] \cdot \mathbf{x} + \left[\begin{array}{c} 1 \\ 2 \end{array} \right] \cdot u \\ y &= \left[\begin{array}{cc} c & 1 \end{array} \right] \cdot \mathbf{x} \end{aligned}$$

Wir wollen beobachten, wie unser Algorithmus die Größe des steuerbaren und beobachtbaren Teilmodells ermittelt. Zuerst einmal bestimmen wir dazu das charakteristische Polynom

$$\Delta\left(\left[\begin{array}{cc} 1 & -1 \\ -2 & 0 \end{array} \right] \right) = s^2 - s - 2 \tag{26}$$

der Dynamikmatrix. Als nächstes ermitteln wir die Dynamikmatrix des rückgekoppelten Modells

$$\begin{bmatrix} 1 & -1 \\ -2 & 0 \end{bmatrix} + k \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} c & 1 \end{bmatrix} = \begin{bmatrix} 1+k\cdot c & -1+k \\ -2+2k\cdot c & 2k \end{bmatrix}$$

und das zugehörige charakteristische Polynom

$$\Delta\left(\begin{bmatrix} 1+k\cdot c & -1+k \\ -2+2\cdot k\cdot c & 2\cdot k \end{bmatrix}\right) = s^2 + (-k\cdot c - 2k - 1)\cdot s - 2 + 2k\cdot c + 4k$$

Nun führen wir die euklidische Division durch und erhalten

$$s^2 + (-k\cdot c - 2k - 1)\cdot s - 2 + 2k\cdot c + 4k =$$

$$= 1\cdot\left(s^2 - s - 2\right) \quad + \quad (-k\cdot c - 2k)\cdot s + 2k\cdot c + 4k.$$

Der Rest dieser Division beträgt $(-k\cdot c - 2k)\cdot s + 2k\cdot c + 4k$. Damit dierser unabhängig von k gleich null wird, muss c die Bedingungen

$$\begin{aligned} -k\cdot c - 2k &= 0 \\ 2k\cdot c + 4k &= 0, \end{aligned}$$

vereinfacht

$$c = -2$$

erfüllen. Gilt also $c = -2$, so ist der Rest der euklidischen Division null und der ggT der beiden charakteristischen Polynome ist gleich $s^2 - s - 2$. Dieses entspricht genau (26), ihr Quotient ist 1 und damit enthält unter der Annahme $c = -2$ das Modell kein steuerbares und beobachtbares Teilmodell.

Nehmen wir aber an, dass $c \neq -2$, so ist der Rest der euklidischen Division ungleich null und der Algorithmus setzt mit der nächsten euklidischen Division

$$\left(s^2 - s - 2\right) = \left(-\frac{1}{k(c+2)}\cdot s - \frac{1}{k(c+2)}\right)\cdot\left((-k\cdot c - 2k)\cdot s + 2k\cdot c + 4k\right) \quad + \quad 0$$

fort. Diesmal ist der Rest unabhängig von c gleich null und wir erhalten

$$\left(-\frac{1}{k(c+2)}\cdot s - \frac{1}{k(c+2)}\right)$$

als größten gemeinsamen Teiler der beiden charakteristischen Polynome. Dividieren wir (26) durch diesen und normieren das resultierende Polynom auf einen Koeffizienten der höchsten Potenz von s von 1, so erhalten wir als charakteristisches Polynom des steuerbaren und beobachtbaren Teilmodells

$$s + 1,$$

und stellen damit fest, dass das gegebene Modell für $c \neq -2$ ein steuerbares und beobachtbares Teilmodell erster Ordnung enthält (wie wir ja für den Spezialfall $c = 1$ in Abschnitt 3 bereits festgestellt haben).

4.2 Aufstellen des Gleichungssystems

Im vorigen Abschnitt haben wir uns damit beschäftigt, wie wir algorithmisch die Struktur, sprich die Lage der *strukturellen Nullen*, des zerlegten Modells ermitteln können. Zusätzlich haben wir auch die charakteristischen Polynome der Teilmodelle des *kalmanzerlegten* Modells bestimmt. Bevor wir damit fortfahren, das Gleichungssystem zur Bestimmung der Transformationsmatrix aufzustellen, wollen wir annehmen, wir würden die Lösung - also das *kalmanzerlegte* Modell (21) und (22) - schon kennen. Wir stellen fest, dass wir in diesem Fall durch eine weitere Transformation mit einer Matrix der Struktur

$$\mathbf{T}_2 = \begin{bmatrix} \mathbf{T}_{sb} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_{s-} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{T}_{-b} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{T}_{--} \end{bmatrix} \tag{27}$$

die Teilmodelle $(\tilde{\mathbf{A}}_{sb},\ \tilde{\mathbf{b}}_{sb},\ \tilde{\mathbf{c}}_{sb}^T)$, $(\tilde{\mathbf{A}}_{s-},\ \tilde{\mathbf{b}}_{s-},\ \mathbf{0}^T)$, $(\tilde{\mathbf{A}}_{-b},\ \mathbf{0},\ \tilde{\mathbf{c}}_{-b}^T)$ und $(\tilde{\mathbf{A}}_{--},\ \mathbf{0},\ \mathbf{0})$ getrennt voneinander beeinflussen können, ohne die Struktur des Gesamtmodells (die *strukturellen Nullen*) zu beeinflussen. Da es nach Kalman [1] immer möglich ist, ein beliebiges Modell durch eine reguläre, konstante Zustandstransformation auf die gewünschte zerlegte Struktur zu bringen, heißt das für uns mit der Erkenntnis der gezielten Beeinflussbarkeit der Teilmodelle, dass wir uns für diese auch *intern* vorteilhafte Strukturen wünschen können. So ist es zum Beispiel immer möglich, die beiden steuerbaren Teilmodelle $(\tilde{\mathbf{A}}_{sb},\ \tilde{\mathbf{b}}_{sb},\ \tilde{\mathbf{c}}_{sb}^T)$ und $(\tilde{\mathbf{A}}_{s-},\ \tilde{\mathbf{b}}_{s-},\ \mathbf{0}^T)$ in erster Normalform

$$\begin{aligned} \dot{\mathbf{x}} &= \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_{n1} & -a_{n2} & -a_{n3} & \cdots & -a_{nn} \end{bmatrix} \cdot \mathbf{x} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \cdot u \\ y &= \begin{bmatrix} c_1 & c_2 & c_3 & \cdots & c_n \end{bmatrix} \cdot \mathbf{x} \end{aligned} \tag{28}$$

und das beobachtbare aber nicht steuerbare Teilmodell in zweiter Normalform

$$\begin{aligned} \dot{\mathbf{x}} &= \begin{bmatrix} 0 & 0 & 0 & \cdots & -a_{1n} \\ 1 & 0 & 0 & \cdots & -a_{2n} \\ 0 & 1 & 0 & \cdots & -a_{3n} \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ 0 & 0 & 0 & 1 & -a_{nn} \end{bmatrix} \cdot \mathbf{x} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix} \cdot u \\ y &= \begin{bmatrix} 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \cdot \mathbf{x} \end{aligned} \tag{29}$$

zu erhalten. Diese Normalformen bieten dreierlei Vorteile: Erstens beinhalten sie nur eine minimale Anzahl an Parametern, was die Anzahl der Variablen im noch aufzustellenden Gleichungssystem reduziert, zweitens enthalten sie eine große Anzahl von Nullen, die später die Berechnung der *Gröbner Basen* vereinfachen werden und drittens entsprechen die Parameter a_{ni} bzw. a_{in} in ihren Dynamikmatrizen genau den Koeffizienten der bereits

bekannten charakteristischen Polynome. Es ist uns also durch Wahl dieser Normalformen gelungen, die Diagonalblöcke von $\tilde{\mathbf{A}}$ in (22) bis auf $\tilde{\mathbf{A}}_{--}$ vollständig zu bestimmen. Darauf, auch für $\tilde{\mathbf{A}}_{--}$ eine derartige Normalform zu wählen (z.B. Block-Jordanform, was allerdings explizite Eigenwertberechnung verlangen würde) wollen wir verzichten, da im Allgemeinen damit zu rechnen ist, dass der weder steuerbare noch beobachtbare Anteil eines Modells klein ist und $\tilde{\mathbf{A}}_{--}$ deshalb ohnehin nur wenige Variablen enthält.

Auch die Eingangs- und Ausgangsvektoren des transformierten Modells (21) $\tilde{\mathbf{b}}$ und $\tilde{\mathbf{c}}$ sind durch die Wahl dieser Normalformen bereits fast vollständig definiert. Das einzige, was noch fehlt, ist der Anteil $\tilde{\mathbf{c}}_{sb}$ des Ausgangsvektors. Um diesen zu ermitteln, nutzen wir die wohlbekannte Tatsache, dass das Produkt aus Beobachtbarkeits- und Steuerbarkeitsmatrix eines Modells invariant gegenüber regulären Zustandstransformationen ist. Weiters hängt - wie sich durch berechnen dieses Produkts mit dem zerlegten Modell (22) leicht überprüfen lässt - dieses Produkt nur vom steuerbaren und beobachten Anteil des Modells ab. Das heißt, es gilt

$$\mathbf{c}^T \cdot \mathbf{A}^i \cdot \mathbf{b} = \tilde{\mathbf{c}}_{sb}^T \cdot \tilde{\mathbf{A}}_{sb}^i \cdot \tilde{\mathbf{b}}_{sb}, \tag{30}$$

wobei uns $\tilde{\mathbf{A}}_{sb}^i$ und $\tilde{\mathbf{b}}_{sb}$ schon bekannt sind. Aufgrund der speziell gewählten Struktur von $\tilde{\mathbf{A}}_{sb}$ und $\tilde{\mathbf{b}}_{sb}$ ergeben sich damit für die einzelnen Elemente von

$$\tilde{\mathbf{c}}_{sb}^T := \begin{bmatrix} c_1 & c_2 & \cdots & c_n \end{bmatrix} \tag{31}$$

eine Reihe von linearen Gleichungen

$$\begin{aligned} \mathbf{c}^T\mathbf{b} &= c_n \\ \mathbf{c}^T\mathbf{A}\mathbf{b} &= c_{n-1} - c_n a_{nn} \\ \mathbf{c}^T\mathbf{A}^2\mathbf{b} &= c_{n-2} - c_{n-1}a_{nn-1} + c_n \cdot (-a_{nn-1} + a_{nn}^2) \\ &\vdots, \end{aligned} \tag{32}$$

die nacheinander leicht nach jeweils einer Variablen gelöst werden können.

Somit ist es uns gelungen, durch den (immer erfüllbaren!) Wunsch, das zerlegte Modell (21) in einer speziellen Form vorliegen zu haben, bereits einen guten Teil der Dynamikmatrix $\tilde{\mathbf{A}}$ sowie die vollständigen Ein- und Ausgangsvektoren $\tilde{\mathbf{b}}$ und $\tilde{\mathbf{c}}$ des zerlegten Modells durch bekannte (numerische bzw. symbolische) Werte festlegen zu können.

4.3 Ermittlung der Lösung

Wir haben damit das Gleichungssystem zur Bestimmung der Transformationsmatrix $\mathbf{T}$ vollständig aufgestellt und dabei Freiheitsgrade, die uns bei der Ermittlung einer solchen Transformationsmatrix zur Verfügung stehen, sinnvoll eingesetzt, um das jetzt zu lösende Gleichungssystem möglichst klein und "angenehm strukturiert" darstellen zu können. Es scheint jetzt an der Zeit, etwas genauer zu definieren, was wir mit "angenehm strukturiert" denn meinen.

4.3.1 Ausnutzung der Struktur des zerlegten Modells

Stellen wir die Transformationsmatrix in Block-Spaltenform

$$\mathbf{T} =: \left[\begin{array}{cccc} \mathbf{T}_{sb} & \mathbf{T}_{s-} & \mathbf{T}_{-b} & \mathbf{T}_{--} \end{array} \right] \tag{33}$$

dar, wobei die einzelnen Spaltenblöcke in ihrer Breite jeweils der Ordnung des korrespondierenden Teilmodells entsprechen. Das Gleichungssystem $\mathbf{0} = \mathbf{A} \cdot \mathbf{T} - \mathbf{T} \cdot \tilde{\mathbf{A}}$ aus (6) können wir damit wie folgt anschreiben, wobei wir der Einfachheit halber lediglich zwischen bekannten numerischen bzw. symbolischen Werten "$*$", unbekannten numerischen bzw. symbolischen Werten "?", Parametern der Transformationsmatrix $\mathbf{T}$ und *strukturellen Nullen* "$\mathbf{0}$" unterscheiden wollen:

$$\mathbf{0} = \left[\begin{array}{c} * \end{array} \right] \left[\begin{array}{cccc} \mathbf{T}_{sb} & \mathbf{T}_{s-} & \mathbf{T}_{-b} & \mathbf{T}_{--} \end{array} \right] - \left[\begin{array}{cccc} \mathbf{T}_{sb} & \mathbf{T}_{s-} & \mathbf{T}_{-b} & \mathbf{T}_{--} \end{array} \right] \left[\begin{array}{cccc} * & \mathbf{0} & ? & \mathbf{0} \\ ? & * & ? & ? \\ \mathbf{0} & \mathbf{0} & * & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & ? & ? \end{array} \right] \tag{34}$$

Hiervon wollen wir die zweite Blockspalte genauer betrachten

$$\mathbf{0} = \left[\begin{array}{c} * \end{array} \right] \cdot \left[\begin{array}{c} \mathbf{T}_{s-} \end{array} \right] - \left[\begin{array}{cccc} \mathbf{T}_{sb} & \mathbf{T}_{s-} & \mathbf{T}_{-b} & \mathbf{T}_{--} \end{array} \right] \left[\begin{array}{c} \mathbf{0} \\ * \\ \mathbf{0} \\ \mathbf{0} \end{array} \right]. \tag{35}$$

Wir stellen fest, dass in diesem Teil des Gleichungssystems nur Parameter der zweiten Blockspalte von $\mathbf{T}$ ($\mathbf{T}_{s-}$) vorkommen. Diese angenehme Tatsache ist eine Eigenschaft der Struktur des zerlegten Modells. Wir können also in einem ersten Schritt ein Teilgleichungssystem aufstellen, das nur einen kleinen Teil der Variablen des Gesamtgleichungssystems enthält, was uns bei der Berechnung von *Gröbner Basen* Vorteile bringen wird.

4.3.2 Ausnutzung der festgelgten Normalformen der Teilmodelle

Damit aber noch nicht genug. Bisher haben wir lediglich die Struktur der Lösung ausgenutzt, die interne Struktur des steuerbaren aber nicht beobachtbaren Teilmodells in 1. Normalform hingegen noch nicht berücksichtigt. Werfen wir also noch einmal einen genaueren Blick auf das Teilgleichungssystem (35), wobei wir die Matrix

$$\mathbf{T}_{s-} =: \left[\begin{array}{cccc} \mathbf{t}_1 & \cdots & \mathbf{t}_{j-1} & \mathbf{t}_j \end{array} \right] \tag{36}$$

in Spaltenform darstellen:

$$\mathbf{0} = \left[\begin{array}{c} * \end{array} \right] \cdot \left[\begin{array}{cccc} \mathbf{t}_1 & \cdots & \mathbf{t}_{j-1} & \mathbf{t}_j \end{array} \right] - \left[\begin{array}{cccc} \mathbf{t}_1 & \cdots & \mathbf{t}_{j-1} & \mathbf{t}_j \end{array} \right] \cdot \left[\begin{array}{cccc} \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{1} \\ * & \cdots & * & * \end{array} \right]. \tag{37}$$

Von diesem Gleichungssystem betrachten wir nun zunächst nur die letzte Spalte

$$\mathbf{0} = \begin{bmatrix} * \end{bmatrix} \cdot [\mathbf{t}_j] - \begin{bmatrix} \mathbf{t}_{j-1} & \mathbf{t}_j \end{bmatrix} \cdot \begin{bmatrix} \vdots \\ \mathbf{0} \\ \mathbf{1} \\ * \end{bmatrix} \tag{38}$$

und stellen fest, dass wir damit die Anzahl der Variablen weiter auf die Parameter der letzten beiden Spalten von $\mathbf{T_{s-}}$ einschränken können. Zudem stellen wir fest, dass dieses Teilgleichungssystem ausschließlich lineare Gleichungen beinhaltet, weshalb die Berechnung von *Gröbner Basen* gänzlich entfallen kann und wir auf Standardlösungsverfahren für lineare Gleichungssysteme - wie Gauß-Elimination - zurückgreifen können.

Das ausschließliche Auftreten linearer Gleichungen ist allerdings nur für die Gleichungen der zweiten Blockspalte von $\mathbf{0} = \mathbf{A} \cdot \mathbf{T} - \mathbf{T} \cdot \tilde{\mathbf{A}}$, sowie der Gleichungssysteme $\mathbf{0} = \mathbf{b} - \mathbf{T} \cdot \tilde{\mathbf{b}}$ und $\mathbf{0} = \mathbf{c}^\mathbf{T} \cdot \mathbf{T} - \tilde{\mathbf{c}}^\mathbf{T}$ charakteristisch. Die übrigen Gleichungen enthalten auch Produkte von Variablen und sind daher über die Berechnungen von *Gröbner Basen* auszuwerten.

4.3.3 Lösung des Gleichungssystems

Wie das Lösungsverfahren auch immer sein mag, Tatsache bleibt, dass es uns möglich ist, jeweils nur einen Teil der gesamten Gleichungen (6), der nur eine Untermenge der Unbekannten dieses Gleichungssystems enthält, herauszugreifen und auszuwerten. Wir stellen also in einem ersten Schritt nur die Gleichungen für die letzte Spalte der zweiten Blockspalte der Gleichungen $\mathbf{0} = \mathbf{A} \cdot \mathbf{T} - \mathbf{T} \cdot \tilde{\mathbf{A}}$ auf und werten diese aus (z.B. durch Gauß-Elimination oder auch die Berechnung von *Gröbner Basen*, was selbstverständlich auch für lineare Gleichungen funktioniert). Typischerweise werden nun einige Variablen eindeutig bestimmt sein (z.B. $t_{ij} = 0$) oder in eindeutiger Beziehung zu anderen Variablen stehen (z.B. $t_{ij} = t_{kl} + 1$). Die derart bestimmten Variablen können wir mit diesen Ergebnissen in allen weiteren Gleichungen des Gesamtgleichungssystems (6) ersetzen, was die noch folgenden Berechnungen weiter vereinfachen wird. Sollte das "umgeformte" Gleichungssystem (Ergebnisgleichungssystem der Berechnungen von *Gröbner Basen*) auch Gleichungen enthalten, die noch nicht dazu geeignet sind, eine Variable eindeutig festzulegen (z.B. $0 = t_{ij}^2 - 4$), so lassen wir sie für den Moment unberücksichtigt, merken sie uns aber, um sie im weiteren Verlauf an geeigneter Stelle erneut zu behandeln.

Im nächsten Schritt setzen wir z.B. die Gleichungen der zweitletzten Spalte der zweiten Blockspalte von $\mathbf{0} = \mathbf{A} \cdot \mathbf{T} - \mathbf{T} \cdot \tilde{\mathbf{A}}$ an. In diesen Gleichungen sind nun einige der Variablen schon bekannt, was unsere Rechnung vereinfachen wird. Wieder werten wir die Gleichungen mit einem geeigneten Lösungsverfahren (Gauß-Elimination oder Berechnung von *Gröbner Basen*) aus, lösen einige Gleichungen direkt nach einer Variablen und merken uns diejenigen, bei denen das nicht möglich ist. So arbeiten wir sukzessive das gesamte Gleichungssystem (6) ab, wobei wir uns jeweils an folgendes Schema halten:

- `Ermittlung eines Teils des Gleichungssystems (6), das nur eine Untermenge aller vorkommenden Variablen enthält (Eine System zur Ermittlung dieser Teilgleichungssysteme werden wir noch angeben)`

- Auswertung des Teilgleichungssystems durch Gauß-Elimination oder Berechnung von Gröbner Basen, wobei
  - Resultierende Gleichungen (zu den Gröbner Basen korrespondierende Gleichungen) nach Möglichkeit dazu benutzt werden, eine Variable eindeutig festzulegen oder
  - Sofern das nicht möglich ist, die Gleichung vorerst unberücksichtigt bleibt, wir sie uns aber für später merken

Zur Ermittlung der erwähnten Teilgleichungssysteme von

$$\begin{aligned} \mathbf{0} &= \mathbf{A}\cdot\mathbf{T}-\mathbf{T}\cdot\tilde{\mathbf{A}} && (39)\\ \mathbf{0} &= \mathbf{b}-\mathbf{T}\cdot\tilde{\mathbf{b}} && (40)\\ \mathbf{0} &= \mathbf{c}^T\cdot\mathbf{T}-\tilde{\mathbf{c}}^{\mathbf{T}}, && (41) \end{aligned}$$

hat sich in den von uns durchgeführten Beispielen das folgende Schema in Bezug auf die nötige Rechenzeit als vorteilhaft erwiesen. Wir führen zur einfacheren Beschreibung die Werte r_{sb}, r_{s-}, r_{-b} und r_{--} für die (Spalten- oder Zeilen-) Dimension von $\tilde{\mathbf{A}}_{sb}$, $\tilde{\mathbf{A}}_{s-}$, $\tilde{\mathbf{A}}_{-b}$ und $\tilde{\mathbf{A}}_{--}$ ein.

- Die $r_{sb}+r_{s-}-te$ Spalte von (39) und diejenigen Spalten von (41), die nur dieselben Variablen enthalten
- Jeweils die nächstlinke Spalte von (39) und diejenigen Spalten von (41), die nur dieselben ''neuen'' Variablen enthalten, bis die $r_{sb}+1-te$ Spalte erreicht ist
- Gleichungen (40)
- Die $r_{sb}-te$ Spalte von (39) und diejenigen Spalten von (41), die nur dieselben ''neuen'' Variablen enthalten
- Jeweils die nächstlinke Spalte von (39) und diejenigen Spalten von (41), die nur dieselben ''neuen'' Variablen enthalten, bis die erste Spalte erreicht ist
- Alle Gleichungen von der $r_{sb}+r_{s-}+r_{b-}+1-ten$ bis zur letzten Spalte von (39) und (41)
- Die $r_{sb}+r_{s-}+1-te$ Spalte von (39) und diejenigen Spalten von (41), die nur dieselben ''neuen'' Variablen enthalten
- Jeweils die nächstrechte Spalte von (39) und diejenigen Spalten von (41), die nur dieselben ''neuen'' Variablen enthalten, bis die $r_{sb}+r_{s-}+r_{b-}-te$ Spalte erreicht ist
- Die Gleichung für die Determinante von T gemeinsam mit allen Gleichungen, die wir uns bisher merken mussten

Nachdem wir das letzte Gleichungssystem aufgestellt haben, berechnen wir ein letztes Mal die zugehörigen *Gröbner Basen.* Jetzt haben wir alle Gleichungen berücksichtigt. Wir können uns daher sicher sein, dass alle Freiheitsgrade, die uns das erhaltene Ergebnis offen lässt, vollständig nach unseren Wünschen genutzt werden können. Wir wollen an dieser Stelle wieder auf die aus Abschnitt 3 bekannte Methode zur Lösung solch eines unterbestimmten Gleichungssystems zurückgreifen: Wir beginnen bei der niederwertigsten Gleichung und wählen alle Variablen bis auf eine zu 1. Damit lösen wir dann die Gleichung nach der verbleibenden Variablen.

Bei der Ermittlung der *Gröbner Basen* hat es sich als vorteilhaft herausgestellt, die Rangordnung zwischen den Variablen derart zu wählen, dass Parameter von $\tilde{\mathbf{A}}$ ranghöher sind, als Parameter von $\mathbf{T}$, wobei innerhalb dieser Ordnung jeweils "neu hinzugekommene" Variable ranghöher sein sollen, als bereits verwendete.

Wir haben damit unser Problem vollständig gelöst und sowohl das zerlegte Modell (21) als auch die zugehörige Transformationsmatrix $\mathbf{T}$ ermittelt.

5 Beispiel

Wir wollen hier zur Veranschaulichung ein einfaches akademisches Beispiel mit einem symbolischen Parameter präsentieren. Wir werden dabei zuerst in Abhängigkeit dieses Parameters p diejenigen Lösungen ermitteln, für die das Modell einen möglichst großen steuerbaren und beobachtbaren Teil enthält, wobei der Algorithmus Bedingungen liefern wird, denen p hierzu genügen muss. Anschließend werden wir p so wählen, dass diese Bedingungen verletzt werden und das jetzt in rein numerischer Form vorliegende Modell erneut zerlegen.

Gegeben:

$$\begin{aligned} \dot{\mathbf{x}} &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & p \end{bmatrix} \cdot \mathbf{x} + \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} \cdot u \\ y &= \begin{bmatrix} 1 & 1 & 0 & 1 \end{bmatrix} \cdot \mathbf{x} \end{aligned}$$

Gesucht: *Kalman Zerlegung* (21) und Zustandstransformation $\mathbf{x} = \mathbf{T} \cdot \mathbf{z}$, die dazu führt.

5.1 Intuitive Lösung

Wir haben das Beispiel derart gewählt, dass wir die gestellte Aufgabe auch intuitiv schnell lösen können und so die algorithmisch erhaltenen Ergebnisse gut überprüfen können. Für das vorliegende Beispiel genügt es nämlich, die zweite und vierte Zustandsvariable zu vertauschen, um für "günstiges p" sofort das Modell in *kalmanzerlegter* Form vorliegen zu haben. Setzen wir für $p = 1$ ein, so stellen wir fest, dass im zerlegten Modell eine Zustandsvariable vom steuerbaren und beobachtbaren Teilmodell (bisher Ordnung 2) in das (für "günstiges p" nicht vorhandene) weder steuerbare noch beobachtbare Teilmodell

wandern muss. Weniger auffällig ist, dass auch der Fall $p = 2$ gesondert behandelt werden muss. Aber das wird uns die Algorithmische Berechnung noch zeigen...

5.2 Lösung für "günstiges p"

Wir lassen den im vorigen Abschnitt beschriebenen Algorithmus annehmen, p würde allen Bedingungen genügen, sodass sowohl das steuerbare als auch das beobachtbare Teilmodell möglichst groß wird. Wir erhalten damit das Ergebnis

$$\begin{aligned}\dot{\mathbf{z}}_p &= \begin{bmatrix} 0 & 1 & 0 & \frac{p^2-3p+3}{(p-1)^2} \\ 1-p & 1+p & 0 & p-1 \\ 2-p & 1 & 1 & \frac{p-2}{(p-1)^2} \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \mathbf{z}_p + \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \cdot u \\ y &= \begin{bmatrix} p+1 & 2 & 0 & 1 \end{bmatrix} \cdot \mathbf{z}_p \end{aligned}$$

mit einer zugrundeliegenden Zustandstransformation

$$\mathbf{x} = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{1-p} \\ 1-p & 1 & 0 & \frac{1}{p-1} \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \cdot \mathbf{z}_p.$$

Zusätzlich erhalten wir die Bedingungen

$$p \neq 1 \quad \textit{und} \quad p \neq 2$$

die erfüllt werden müssen, damit obige Zerlegung gilt.

Wir stellen fest, dass sich das algorithmisch ermittelte Ergebnis vom intuitiv gefundenen unterscheidet, da wir ja z.B. gefordert haben, dass das steuerbare und beobachtbare Teilmodell in 1. Normalform vorliegen soll. Nichtsdestotrotz sind selbstverständlich beide Lösungen korrekt und führen auf dieselbe *kalmanzerlegte* Struktur.

Die Berechnungen zur algorithmischen Ermittlung des zerlegten Modells benötigten auf einem PII-Rechner mit 400Mhz und einer MapleV Implementierung des beschriebenen Verfahrens inklusive aller Bildschirmausgaben etwa 1 Sekunde.

5.3 Lösungen für p=1 und p=2

Wir erhalten für $p = 1$ das Ergebnis

$$\begin{aligned}\dot{\mathbf{z}}_{p=1} &= \begin{bmatrix} 2 & 0 & \frac{3}{4} & 0 \\ 2 & 1 & \frac{7}{4} & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -\frac{3}{4} & 0 \end{bmatrix} \cdot \mathbf{z}_{p=1} + \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \cdot u \\ y &= \begin{bmatrix} 2 & 0 & 1 & 0 \end{bmatrix} \cdot \mathbf{z}_{p=1} \\ \mathbf{x} &= \begin{bmatrix} 0 & 0 & \frac{1}{2} & 0 \\ 1 & 0 & -\frac{1}{2} & -1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix} \cdot \mathbf{z}_{p=1}. \end{aligned}$$

Wie intuitiv erwartet, verliert eine der ursprünglich steuerbaren und beobachtbaren Zustandsvariablen sowohl ihre Steuerbarkeit als auch ihre Beobachtbarkeit.

Für $p = 2$ erhalten wir das folgende Ergebnis

$$\begin{aligned}
\dot{\mathbf{z}}_{p=2} &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \mathbf{z}_{p=2} + \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \cdot u \\
y &= \begin{bmatrix} 1 & 1 & 0 & 1 \end{bmatrix} \cdot \mathbf{z}_{p=2} \\
\mathbf{x} &= \begin{bmatrix} 0 & 0 & -1 & -1 \\ -1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \cdot \mathbf{z}_{p=2}
\end{aligned}$$

und stellen fest, dass zwar die Anzahl der steuerbaren und beobachtbaren Zustandsvariablen erhalten bleibt, alle übrigen Zustandsvariablen jedoch nunmehr weder steuerbar noch beobachtbar sind, was intuitiv nicht so leicht zu erkennen war.

Es gestaltet sich durchaus interessant, den Rechengang für die verschiedenen Lösungen dieses Beispiels einmal händisch unter Verwendung eines Werkzeugs zur Berechnung von *Gröbner Basen* durchzugehen. Auf eine explizite Ausarbeitung dieses Rechenganges wird hier jedoch aus Gründen der Übersichtlichkeit verzichtet.

Die Rechenzeit unterscheidet sich für die jetzt rein numerisch vorliegenden Modelle nur unwesentlich von der, die für die Zerlegung des Modells mit symbolischem Parameter p nötig war.

6 Ergebnis

Wir haben im Detail einen Algorithmus vorgestellt, der es uns ermöglicht, symbolisch die *Kalman Zerlegung* [1] für beliebige lineare, zeitinvariante Zustandsraummodelle, die auch symbolische Parameter enthalten können, zu ermitteln. Um noch einmal den Vorteil des hier vorgestellten Verfahrens herauszustreichen, blicken wir auf das Beispiel in Abschnitt 5 zurück. Die Anwendung eines numerischen Algorithmus wie in [6] würde für dieses Beispiel sicher Probleme machen, wenn wir den Parameter p sehr nahe an 1 oder 2 wählen. Unser Algorithmus löst diese Aufgabe nicht nur ohne Probleme für beliebige Werte von p, nein er gibt zusätzlich sogar noch Bedingungen an, die p erfüllen muss, damit das Modell nicht an Steuerbarkeit/Beobachtbarkeit einbüßt. Darüber hinaus ermöglicht er uns auch, den Einfluss dieses Parameters anhand der zerlegten Modellstruktur zu analysieren.

Wir haben unsere Methode hier nur für relativ kleine Modelle mit einem Eingang und einem Ausgang vorgeführt. Es sei aber nochmals darauf hingewiesen, dass prinzipiell in gleicher Weise auch größere Modelle mit mehreren Ein- und Ausgängen zerlegt werden können. Allerdings können wir im Fall solcher Mehrgrößenmodelle für die Diagonalelemente der Dynamikmatrix des zerlegten Modells nicht mehr derart einfache Normalformen, wie hier verwendet, ansetzen. Ähnliche, wenn auch ein wenig unvorteilhaftere, Formen sind aber auch für Mehrgrößenmodelle bekannt [4].

Ein vielleicht wichtigerer Nachteil der hier vorgestellten Methode darf aber auch nicht unerwähnt bleiben. Wir werden im Allgemeinen nur brauchbare Ergebnisse erhalten, wenn unser gegebenes Modell aus physikalischen Beziehungen abgeleitet und nicht numerisch durch irgendeine Form der Systemidentifikation ermittelt wurde. Da es uns nicht möglich ist, Toleranzen anzugeben und da ein numerisch ermitteltes Modell immer numerische Ungenauigkeiten aufweisen wird, wird unser Algorithmus im Allgemeinen ein solches Modell als vollständig steuerbar und beobachtbar (die Zerlegung nach [1] enthält keine Teilmodelle, die nicht sowohl steuerbar als auch beobachtbar sind) erkennen. Und das ohne Rücksicht darauf, wir "*gut*" oder "*schlecht*" steuerbar oder beobachtbar es ist. (Wir wollen diese Ausdrücke hier nicht näher definieren. Verschiedene Ideen dazu finden sich in [14], [15], [16].) Das wird üblicherweise nicht die Antwort sein, die wir uns erhoffen - obwohl sie zugegebenermaßen richtig ist.

Wir werden unsere Methode daher vorzugsweise für die Zerlegung von Modellen einsetzen, die entweder symbolische Parameter enthalten oder numerisch schlecht konditioniert sind, wobei wir alle steuerbaren und beobachtbaren Modellteile erkennen wollen, egal wie "*schlecht*" steuerbar/beobachtbar diese sind.

In weiterer Folge könnte daran gedacht werden, die Methode derart zu erweitern, dass wir Toleranzen vorgeben können, innerhalb derer ein Modellteil als nicht steuerbar/beobachtbar eingeordnet werden kann, obwohl er das streng betrachtet eigentlich wäre. Ein erster Ansatzpunkt dazu könnten Ideen aus [17] sein. Auch eine Adaption des Algorithmus zur Berechnung von *Gröbner Basen* in Hinblick auf die hier auftretenden Gleichungen könnte überlegt werden.

Literatur

[1] R. E. Kalman, "Mathematical description of linear systems," *SIAM Journal of Control*, vol. 1, no. 2, pp. 152–192, 1963.

[2] T. Kailath, *Linear Systems.* Englewood Cliffs, NJ: Prentice Hall, 1980.

[3] W. L. Brogan, *Modern Control Theory.* Englewood Cliffs, NJ: Prentice Hall, 1991.

[4] C. T. Chen, *Introduction to Linear System Theory.* NY: Holt, Rinehart and Winston Inc., 1970.

[5] J. Lunze, *Regelungstechnik 2.* Berlin Heidelberg: Springer-Verlag, 1997.

[6] S. Bingulac and H. F. VanLandingham, *Algorithms for Computer-Aided Design of Multivariable Control Systems.* NY: Marcel Dekker, 1993.

[7] D. Boley, "Computing the kalman decomposition: An optimal method," *IEEE Transactions on Automatic Control*, vol. 29, no. 1, pp. 51–53, 1984.

[8] B. Buchberger, "An algorithm for finding the bases elements of the residue class ring modulo a zero dimensional polynomial ideal," *PhD thesis, Univ. of Innsbruck, Austria*, 1965.

[9] B. Buchberger, "An algorithmical criterion for the solvability of algebraic systems of equations," *Aequationes Mathematicae*, vol. 4, no. 3, pp. 347–383, 1970.

[10] B. Buchberger and F. Winkler, "Groebner bases and applications," *Proceedings of the International Conference "33 Years of Groebner Bases" vol. 251 of London Mathematical Society Series. Cambridge University Press*, 1998.

[11] B. Buchberger, "Groebner bases and system theory," *Special Issue on Applications of Groebner Bases in Multidimensional Systems and Signal Processing*, vol. 12, no. 3, pp. 223–235, 2001.

[12] H. C. Liao and R. Fateman, "Evaluation of the heuristic polynomial gcd," *Proceedings of the International Symposium on Symbolic and Algebraic Computation*, 1995.

[13] E. Kaltofen and M. Monagan, "On the genericity of the modular polynomial gcd algorithm," *Proceedings of the International Symposium on Symbolic and Algebraic Computation*, 1999.

[14] R. E. Kalman, Y. C. Ho, and K. S. Narendra, "Controllability of linear dynamical systems," *Contributions to Differential Equations*, vol. 1, pp. 189–213, 1963.

[15] J. Lückel and P. C. Müller, "Analyse von steuerbarkeits-, beobachtbarkeits- und störbarkeitsstrukturen linearer, zeitinvarianter systeme," *Regelungstechnik*, vol. 5, pp. 163–171, 1975.

[16] R. Eising, "The distance between a system and the set of uncontrollable systems," *Proceedings of the MTNS, Beer-Sheva, Israel*, pp. 303–314, 1983.

[17] R. Eising, "Between controllable and uncontrollable," *Systems and Control Letters*, vol. 4, pp. 263–264, 1894.

A Berechnung von Gröbner Basen

Wir betrachten Polynomgleichungen wie z.B.

$$\begin{aligned} 0 &= -1 + 3xy^3 + 2x^2y \\ 0 &= xy^2 + x^2. \end{aligned} \tag{42}$$

Dieses Gleichungssystem ist nicht leicht zu lösen. Die Idee ist nun die, das Gleichungssystem so umzuformen, dass man eine Gleichung erhält, die nur x enthält und eine, die sowohl x als auch y enthält. (vgl. Gauß-Elimination für lineare Gleichungssysteme). Allgemeiner gesprochen wollen wir ein Gleichungssystem

$$\begin{aligned} 0 &= p_1(x_1, x_2, \ldots, x_j) \\ 0 &= p_2(x_1, x_2, \ldots, x_j) \\ &\vdots \\ 0 &= p_k(x_1, x_2, \ldots, x_j) \end{aligned}$$

in ein anderes

$$\begin{aligned} 0 &= g_1(x_1) \\ 0 &= g_2(x_1, x_2) \\ &\vdots \\ 0 &= g_l(x_1, x_2, \ldots, x_j) \end{aligned}$$

umformen, das die selben Lösungen besitzt, aber leichter zu lösen ist. Genau das werden wir durch die Berechnung sogenannter *Gröbner Basen* g_1 bis g_l zu den Polynomen p_1 bis p_k erreichen.

In diesem Appendix werden wir ein "Rezept" dafür angeben, wie solche *Gröbner Basen* berechnet werden können. In unserem Beispiel betrachten wir die Polynome

$$\begin{aligned} p_1 &: -1 + 3xy^3 + 2x^2y \\ p_2 &: xy^2 + x^2. \end{aligned}$$

Wichtig ist nun, die darin vorkommenden Terme 1, xy^3, x^2y, xy^2 und x^2 einer Rangordnung zu unterwerfen. Wir definieren dazu

$$x \prec y.$$

Das bedeutet, x wird als von kleinerem Rang angesehen als y. In der Literatur bezeichnet man so eine Rangordnung als *lexicographisch.* Diese Ordnung impliziert weiters, dass jeder Term, der in irgendeiner Weise y enthält von höherem Rang ist, als jeder Term, der nur x, egal in welcher Form, enthält, dass also

$$x \prec x^2 \prec \ldots \prec x^n \prec y \prec xy \prec \ldots \prec x^n y \prec y^2 \prec xy^2 \prec \ldots \prec x^n y^2 \prec \ldots \prec x^n y^m$$

gilt. Mit der zusätzlichen Vereinbarung, dass Konstante immer von niedrigstem Rang sind, erhalten wir unsere Polynome in geordneter Form

$$\begin{aligned} p_1 &: 3xy^3 + 2x^2y - 1 \\ p_2 &: xy^2 + x^2. \end{aligned}$$

Der *Führungsterm* (der Term mit dem höchsten Rang) eines jeden Polynoms bestimmt nun das weitere Vorgehen. Falls möglich - das heißt, falls der *Führungsterm* eines Polynoms ein Vielfaches des *Führungsterms* eines anderen Polynoms ist - werden diese beiden Polynome gegeneinander *reduziert*, so dass sich diese Terme aufheben. Wir werden das an unserem Beispiel demonstrieren: Der *Führungsterm* von p_1 ($3xy^3$) ist das $3y$-fache dessen von $p_2(xy^2)$. Wir reduzieren daher p_1 mit Hilfe von p_2 zu einem neuen Polynom - nennen wir es p_{11} - indem wir

$$p_{11} \leftarrow p_1 - 3y \cdot p_2 = -x^2y - 1$$

bilden. Nach der von uns definierten Rangordnung ist der neue *Führungsterm* (x^2y) von niedrigerem Rang, als der alte (xy^3). Daher der Ausdruck *reduzieren.* Der neue Satz von Polynomen lautet

$$\begin{aligned} p_{11} &: -x^2y - 1 \\ p_2 &: xy^2 + x^2. \end{aligned}$$

Nun ist keiner der *Führungsterme* mehr ein Vielfaches des anderen und keine weiteren *Reduktionen* sind möglich. Wir müssen deshalb mit einer anderen Prozedur fortsetzen.

Die Idee ist nun die, aus dem Satz von Polynomen zwei beliebige auszuwählen und Vielfache davon so zu addieren, dass das neu entstehende Polynom (*S-Polynom* - Summenpolynom) durch beide alten reduziert werden kann. Für unser Beispiel müssen wir dieses neue Polynom selbstverständlich aus p_{11} und p_2 bilden. Das kleinste gemeinsame Vielfache ihrer *Führungsterme* ist x^2y^2. Dieses wollen wir als *Führungsterm* für das neue Polynom erhalten und bilden daher

$$p_3 \leftarrow -y \cdot p_{11} + x \cdot p_2 = 2x^2y^2 + y + x^3.$$

Dieses neue Polynom *reduzieren* wir nun durch die beiden alten, wobei die Reihenfolge freigestellt ist. Wir beginnen mit p_{11}.

$$p_{31} \leftarrow p_3 - (-2y) \cdot p_{11} = -y + x^3.$$

Eine weitere *Reduktion* durch p_2 ist nicht mehr möglich. (Anmerkung: Dies steht *nicht* im Gegensatz zur Forderung, dass das *S-Polynom* durch beide alten Polynome *reduzierbar* sein muss!) Nach Buchbergers Algorithmus fügen wir dieses *reduzierte S-Polynom* -sofern es wie hier nicht zu null *reduziert* werden kann- dem bisherigen Satz von Polynomen bei. Wäre es möglich gewesen, das *S-Polynom* zu null zu *reduzieren*, so hätten wir es verwerfen und ein anderes Paar von "alten" Polynomen wählen müssen. Gelingt es, für jedes beliebige Paar das entsprechende *S-Polynom* zu null zu *reduzieren*, so sind wir am Ende und der momentane Satz von Polynomen entspricht den *Gröbner Basen* des ursprünglichen Satzes von Polynomen.

In unserem Beispiel lautet der verbleibende Rechengang daher wie folgt:

$$\begin{aligned} p_{11} &: -x^2y - 1 \\ p_2 &: xy^2 + x^2 \\ p_{31} &: -y + x^3, \end{aligned}$$

Diese Polynome *reduzieren* wir gegeneinander

$$p_{12} \leftarrow p_{11} - x^2 \cdot p_{31} = -x^5 - 1$$

und

$$p_{21} \leftarrow p_2 - (-xy) \cdot p_{31} = x^4y + x^2$$

sowie weiters

$$p_{22} \leftarrow p_{21} - (-x^4) \cdot p_{31} = x^2 + x^7$$

und schließlich

$$p_{23} \leftarrow p_{22} - (-x^2) \cdot p_{12} = 0.$$

Damit verschwindet p_{23} und es verbleiben.

$$\begin{aligned} p_{12} &: -x^5 - 1 \\ p_{31} &: -y + x^3. \end{aligned} \tag{43}$$

Eine weitere *Reduktion* ist nicht möglich und das einzige mögliche *S-Polynom* lautet

$$p_4 \leftarrow +(-y \cdot p_{12}) + (-x^5 \cdot p_{31}) = 2yx^5 + y - x^8.$$

Dieses *reduzieren* wir zu

$$\begin{aligned} p_{41} &\leftarrow p_4 - (-x^5 \cdot p_{31}) = y + x^8 \\ p_{42} &\leftarrow p_{41} - (-p_{31}) = x^8 + x^3 \\ p_{43} &\leftarrow p_{42} - (-x^3 \cdot p_{12}) = 0 \end{aligned}$$

Damit sind wir mit unseren Berechnungen am Ende und die Polynome (43) sind die gesuchten *Gröbner Basen*. Das entsprechende Gleichungssystem

$$\begin{aligned} 0 &= -x^5 - 1 \\ 0 &= -y + x^3 \end{aligned} \tag{44}$$

ist leicht zu lösen und es ist weiters leicht überprüfbar, dass die gefundenen Lösungen auch die ursprünglichen Gleichungen (42) lösen.