

Institut für Regelungstechnik

RWTH Aachen University

Univ.-Prof. Dr.-Ing. Heike Vallery

PD Dr.-Ing. Lorenz Dörschel

Univ.-Prof. Dr.-Ing. Dirk Abel

Umdruck zur Vorlesung

Regelungstechnik

sowie Höhere Regelungstechnik
und weitere Ergänzungen

47. Auflage 2023

Nachdruck und Vervielfältigung nicht gestattet



Institut für
Regelungstechnik

RWTHAACHEN
UNIVERSITY

Institut für Regelungstechnik

Anschrift:

Institut für Regelungstechnik
RWTH Aachen University
Campus-Boulevard 30
52074 Aachen

Telefon: +49 241 80-27500

Telefax: +49 241 80-22296

Email: secretary@irt.rwth-aachen.de

Web: www.irt.rwth-aachen.de

Herausgeber:

© Aachener Forschungsgesellschaft Regelungstechnik e. V. (AFR)
Campus-Boulevard 30
52074 Aachen

Vorwort

Der vorliegende Umdruck enthält den Stoff der Vorlesungen „Regelungstechnik“ und „Höhere Regelungstechnik“ sowie Ergänzungen dazu. Auswahl und Darstellung des Stoffes orientieren sich soweit wie möglich am derzeitigen Lehrangebot speziell für den Bachelor of Science im Maschinenbau der RWTH Aachen, in dem „Regelungstechnik“ ein Pflichtfach ist. Der Umdruck soll als Arbeitsunterlage für das Studium dienen, wobei ein Kapitel näherungsweise dem Inhalt einer Vorlesungswoche entspricht. Er ersetzt nicht die aktive Teilnahme an Vorlesungen und Übungen, in denen der Stoff erläutert, vertieft und zur Lösung einschlägiger Aufgaben angewandt wird.

Die Ergänzungen sind für Leser gedacht, die sich in weiterführenden Lehrveranstaltungen, im Zusammenhang mit Bachelor- und Masterarbeiten oder während ihrer späteren Tätigkeit intensiver mit der Regelungstechnik befassen möchten. Die Inhalte der Vorlesung „Höhere Regelungstechnik“ (HRT) sind durch ein „H“ und der Ergänzungsstoff durch ein „E“ hinter der Seitenzahl gekennzeichnet.

Die Vorlesung „Regelungstechnik“ hat die gezielte Beeinflussung technischer Systeme mittels Rückkopplungen zum Inhalt. Für die Modelle, die das zu regelnde System beschreiben, gibt es dabei vor allem zwei unterschiedliche Ansätze. Die Beschreibung im sogenannten Bildbereich führt dabei oft auf graphische Darstellungen, denen sich ein vorwiegend zeichnerischer Reglerentwurf anschließt. Die Methoden im sogenannten Zustandsraum führen hingegen über die Matrizenrechnung auf algebraische Gleichungen. Beide Strömungen innerhalb der Regelungstechnik existieren seit Langem parallel. Bei der Gründung des Instituts für Regelungstechnik der RWTH Aachen und der Festlegung der Kerninhalte dieser Vorlesung standen dabei die graphischen Werkzeuge im Mittelpunkt, da sich diese in Zeiten ohne Rechnerunterstützung leichter umsetzen ließen. Heutzutage sind Computer aus dem Ingenieurberuf nicht mehr wegzudenken und speziell auf regelungstechnische Bedürfnisse zugeschnittene Software wie MATLAB gehören zum täglichen Handwerkzeugs des Regelungstechnikers. Da solche Software sowohl ein Vorgehen im Zustandsraum als auch den graphischen Entwurf unterstützt, handelt es sich bei beiden Ansätzen mittlerweile um zwei gleichberechtigte Säulen der Regelungstechnik.

Dieser Umdruck bis einschließlich Auflage 46 war in seiner Grundstruktur der Tradition der ersten Regelungstechnikvorlesungen verpflichtet, sodass von Anfang an die graphischen Methoden im Vordergrund standen und die Zustandsraummethoden in nachgelagerten Kapiteln behandelt wurden. Mit dieser 47. Auflage soll die Gleichberechtigung beider Ansätze in den Vordergrund gerückt werden, sodass die Zustandsraummethoden durchgängig in fast allen Kapiteln erscheinen und die Parallelen beider Ansätze aufgezeigt werden. Diese „Kernsanierung“ des Umdrucks sowie der zugehörigen Vorlesung und Übung wäre ohne die Unterstützung aller Mitarbeiterinnen und Mitarbeiter des Instituts für Regelungstechnik nicht möglich gewesen. Ihnen allen soll für ihre Mitwirkung an der Weiterentwicklung dieses Umdrucks und des entsprechenden Lehrangebots an dieser Stelle ausdrücklich gedankt werden.

Den Studierenden an der RWTH Aachen, deren Prüfungsordnung das Fach Regelungstechnik vorsieht, danken wir für das entgegengebrachte Interesse. Wir hoffen, dass sie neben der zu absolvierenden Studienleistung aus dem Fach auch neue Erkenntnisse und Betrachtungsweisen (z. B. zum Verständnis und zur Beherrschung komplexer, rückgekoppelter Systeme) gewinnen, die ihnen für die weiteren beruflichen Aufgaben von Nutzen sind.

Aachen, im August 2023

Heike Vallery

Lorenz Dörschel

Dirk Abel

Hinweise zur Notation

Innerhalb des Umdrucks werden besonders wichtige Sätze, Definitionen oder zentrale Sachverhalte optisch hervorgehoben und dadurch leichter auffindbar gemacht.

Optische Hervorhebung

Das ist ein Beispiel für eine optische Hervorhebung.

Diese Einrückungen sind als Teil des regulären Leseflusses zu verstehen.

Skalare Größen a werden durch kursive Schrift dargestellt, Vektoren \mathbf{a} werden fettgedruckt und Matrizen \mathbf{A} zusätzlich aufrecht gesetzt und nach Möglichkeit mit Großbuchstaben versehen. Vektoren werden grundsätzlich als Spaltenvektoren notiert und Zeilenvektoren über transponieren \mathbf{a}^T aus Spaltenvektoren gebildet. Die einzelnen Elemente von Matrizen oder Vektoren werden durch Indizierung des Symbols in Kursivschrift dargestellt. Somit bezeichnet a_i das i -te Element des Vektors \mathbf{a} . Davon zu unterscheiden sind steil gesetzte Indizes, die eine allgemeine Abkürzung darstellen. So ist y_m der m -te Systemausgang, während y_m die Messgröße bezeichnet.

Die Regelungstechnik beschreibt Prozesse über Signale, die typischerweise Kleinbuchstaben zur Bezeichnung erhalten. Diese Signale entsprechen realen physikalischen Größen, die in ihren jeweiligen Fachdisziplinen oft mit Großbuchstaben bezeichnet werden. In den Kapiteln 1.3 und 2, wo die physikalischen Größen im Vordergrund stehen, werden die Bezeichnungen der Fachdisziplinen verwendet. In den späteren Kapiteln werden nahezu ausschließlich Kleinbuchstaben verwendet und Großbuchstaben nur genutzt, um zwischen Absolut- und Abweichungsgrößen zu unterscheiden (siehe 3.3).

Inhaltsverzeichnis

1 Einführung	1
1.1 Steuerung und Regelung	1
1.2 Grundstruktur des Regelkreises	3
1.3 Beispiele technischer Regelungen	9
1.3.1 Tiefenregelung eines Unterwasserfahrzeuges	9
1.3.2 Regelung von Windkraftanlagen	12
1.3.3 Kraftregelung beim Fräsen	14
1.3.4 Regelung eines Bioreaktors	16
1.3.5 Regelung einer Dampfmaschine	18
2 Modellbildung	20
2.1 Allgemeines	20
2.2 Einführung in Differentialgleichungen	22
2.3 Darstellung von Differentialgleichungen im Zustandsraum .	28
2.4 Darstellung von Differentialgleichungen im Wirkungsplan .	33
2.5 Aufstellen von Differentialgleichungen	35
2.6 Beispiele für Modellbildung	37
2.6.1 Zerlegung in Teilsysteme	37
2.6.2 Rückwirkungen	41
2.6.3 Zusammenfassen von Teilsystemen im Wirkungsplan	44
2.6.4 Modellierung von Regelungen	45
2.7 Das Gesetz der Sparsamkeit	47
2.8 Einheiten	49
3 Autonome Systeme	52
3.1 Arbeitspunkte und Ruhelagen	52
3.2 Stabilität	56
3.3 Linearisierung	58
3.3.1 Linearisierung einer Funktion	59
3.3.2 Linearisierung einer Differentialgleichung	62
3.3.3 Linearisierung im Kennlinienfeld	64
3.4 Charakteristisches Polynom	66
3.5 Linearisierungstheorem	72
3.6 Analyse im Zustandsraum	74

4 Verhalten bei allgemeiner Anregung	85
4.1 Homogene und partikuläre Lösung	85
4.2 Übergangsfunktion	88
4.3 Faltung	91
4.4 Laplace-Transformation	97
4.4.1 Laplace-Transformation von Zeitfunktionen	97
4.4.2 Laplace-Transformation von Operationen	101
4.4.3 Bestimmung des Zeitverlaufes linearer Systeme	104
4.5 Übertragungsfunktion	107
4.6 Grenzwertsätze	112
5 Verhalten bei sinusförmiger Anregung	115
5.1 Frequenzgang	115
5.2 Ortskurve	123
5.3 Bode-Diagramm	126
5.4 Fourier-Transformation	130
5.5 Filter	133
6 Verschaltungen von Systemen	137
6.1 Zusammenfassen von Teilsystemen	137
6.2 Zerlegung in einfache Elemente	145
6.3 Zerlegung nicht-minimalphasiger Systeme	153
7 Typische Übertragungsglieder	157
7.1 Übersicht	157
7.2 Grundlegende Reglertypen	164
7.2.1 P-Element	164
7.2.2 I-Element	164
7.2.3 D-Element	168
7.2.4 PI, PD und PID	170
7.3 Verzögerungsglieder	175
7.3.1 PT_1	175
7.3.2 PT_2	177
7.3.3 PT_n	186
7.4 Kombinationen	186
7.4.1 IT_1	187
7.4.2 DT_1	189
7.4.3 PIT	191

7.4.4	PPT ₁ und PDT ₁	191
7.5	Nicht-minimalphasige Systeme	193
7.5.1	PA ₁	193
7.5.2	PT _t , PT ₁ T _t	196
8	Identifikation linearer Regelkreisglieder	200
8.1	Allgemeines	200
8.2	Nicht-parametrische Identifikation	203
8.3	Parametrische Identifikation	205
8.3.1	Überanpassung	205
8.3.2	Graphische Parameteridentifikation	206
8.3.3	Methode der kleinsten Fehlerquadrate	208
9	Stabilitätsprüfung	214
9.1	Problemstellung	214
9.2	Algebraische Stabilitätskriterien	215
9.2.1	Grundidee	215
9.2.2	Stabilitätskriterien nach Routh und Hurwitz	216
9.2.3	Beispiele	221
9.3	Nyquist-Kriterium	224
9.3.1	Vollständiges Nyquist-Kriterium	224
9.3.2	Beispiele	232
9.3.3	Anwendung bei Polen am Stabilitätsrand	236
9.3.4	Vereinfachtes Nyquist-Kriterium	238
9.3.5	Amplituden- und Phasenreserve	247
9.4	Sonderfälle	252
9.4.1	Pol-Nullstellen-Kürzungen	252
9.4.2	Unstetige Polstellen	255
10	Einführung in den Reglerentwurf	257
10.1	Ziele und Lösungsansätze	257
10.1.1	Motivation	257
10.1.2	Gütemaße und Kennwerte	260
10.1.3	Ansätze des Reglerentwurfs	262
10.2	Statischer Reglerentwurf	265
10.3	Abwägungen bei der Reglerverstärkung	268
10.3.1	Vorteile hoher Verstärkungen	268
10.3.2	Nachteile hoher Verstärkungen	269

10.4 Einstellregeln	275
10.4.1 Einstellung mittels T_u - T_g -Ersatzmodell	276
10.4.2 Einstellung mittels Schwingversuch	277
11 Grundlegende modellbasierte Reglerentwurfsverfahren	279
11.1 Frequenzkennlinienverfahren	279
11.1.1 Grundidee	279
11.1.2 Hohe Verstärkung bei niedrigen Frequenzen	281
11.1.3 Übergangsbereich	283
11.1.4 Niedrige Verstärkung bei hohen Frequenzen	288
11.2 Betragskriterium und Symmetrisches Kriterium	291 E
11.3 Polvorgabe	297
11.3.1 Polvorgabe für Ausgangsrückführungen	297
11.3.2 Polvorgabe für Zustandsrückführungen	300
11.3.3 Steuerbarkeit	302
11.4 Beobachterentwurf	306
11.4.1 Zustandsschätzung	306
11.4.2 Luenberger-Beobachter	310
11.4.3 Beobachtbarkeit und Dualität	312
11.4.4 Beispiel	318
11.5 Wurzelortskurven	321 H
11.5.1 Grundidee	321 H
11.5.2 Konstruktionsregeln	325 H
11.5.3 Beispiel	329 H
12 Vermischte Regelkreise	333
12.1 Erweiterung des Einfachregelkreises	333
12.2 Vorsteuerung	334
12.3 Führungsgrößenfilter	339
12.4 Störgrößenaufschaltung	342
12.5 Kaskadenregelung	344
12.6 Hilfsstellgröße	349
13 Mehrgrößenregelung	351 H
13.1 Zentrale vs. dezentrale Regelung	351 H
13.2 Eigenschaften von Mehrgrößensystemen	355 H
13.2.1 Verschaltungen von Mehrgrößensystemen	355 H
13.2.2 Querkopplungen	357 H

13.2.3	Polstellen von Mehrgrößensystemen	359 H
13.2.4	Richtungsabhängige Verstärkung	362 H
13.3	Verfahren der dezentralen Regelung	364 H
13.3.1	Relative Gain Array	364 H
13.3.2	MIMO-Nyquist und Diagonaldominanz	367 H
13.4	Verfahren der zentralen Regelung	371 H
13.4.1	Zentrale Regelung im Zustandsraum	371 H
13.4.2	Entkopplungsregler	372 H
14	Zeitdiskrete Systeme	375
14.1	Abtastregelungen	375
14.1.1	Definitionen	375
14.1.2	Abtaster und Halteglied	377
14.1.3	Aliasing	378
14.1.4	Verschaltung zu hybriden Systemen	382
14.2	Einführung in Differenzengleichungen	384
14.3	Autonome zeitdiskrete Systeme	386
14.4	Umrechnen von Differenzen- und Differentialgleichungen	389
14.4.1	Rückwärtsdifferenzen	389
14.4.2	Analytische Lösung	393
14.5	Quasikontinuierlicher Reglerentwurf	395
14.6	Zeitdiskreter Bildbereich	399 H
14.6.1	\mathcal{Z} -Transformation	399 H
14.6.2	Zeitdiskrete Übertragungsfunktion	405 H
14.6.3	Zeitdiskreter Frequenzgang	409 H
14.6.4	Zeitdiskrete Modelle zeitkontinuierlicher Systeme	410 H
14.7	Bilineare Transformation	412 H
14.8	Klassischer zeitdiskreter Reglerentwurf	414 H
14.9	Regler mit endlicher Einstellzeit	418 E
14.9.1	Entwurf	418 E
14.9.2	Stabilität	422 E
14.9.3	Beispiel	424 E
15	Kalmanfilter	426
15.1	Allgemeines	426
15.2	Herleitung	427
15.3	Auslegung und Beispiel	436
15.4	Limitierungen und Erweiterungen	441

16 Nichtlineare Systeme	447 H
16.1 Phasenportraits	447 H
16.2 Einzugsbereich	450 H
16.3 Lyapunov-Funktionen	453 H
16.3.1 Direkte Methode nach Lyapunov	453 H
16.3.2 Beispiel und Anwendungshinweise	454 H
16.4 Grenzzyklen	457 H
16.4.1 Definition	457 H
16.4.2 Grenzzyklen durch schaltende Komponenten	458 H
16.4.3 Grenzzyklen durch Integrator-Windup	461 H
16.5 Stabilitätsanalyse von Grenzzyklen	462 H
16.5.1 Beschreibungsfunktionen	462 H
16.5.2 Zwei-Ortskurven-Kriterium	469 H
16.5.3 Maßnahmen gegen Grenzzyklen	475 H
17 Nichtlineare Regelung	479 H
17.1 Exakte Linearisierung	479 H
17.2 Interne Dynamik	482 H
17.3 Flachheit	486 H
17.4 Normalformen	490 H
17.5 Flachheitsbasierte Steuerung und Regelung	492 H
17.5.1 Entwurf für flache Systeme	492 H
17.5.2 Beispiel	495 H
17.5.3 Umgang mit interner Dynamik	496 H
17.6 Integrator Backstepping	500 H
17.6.1 Reglerentwurf über Lyapunov	500 H
17.6.2 Schrittweises Erstellen des Stellgesetzes	502 H
17.7 Sliding Mode Regelung	505 H
17.7.1 Der eindimensionale Fall	505 H
17.7.2 Chatter	508 H
17.7.3 Systeme höherer Ordnung	510 H
18 Lineare Optimale Regelung	513 H
18.1 Allgemeines	513 H
18.2 Linear-quadratische Regler	515 H
18.2.1 Herleitung	515 H
18.2.2 Wahl der Gewichtungsmatrizen	519 H
18.2.3 Linear-quadratisch-Gaußsche Regler	520 H

18.3 \mathcal{H}_∞ -Regelung	522 H
18.3.1 Closed Loop Shaping	522 H
18.3.2 Generalized Plant	528 H
18.3.3 Mixed Sensitivity	532 H
19 Modellprädiktive Regelung (MPR)	534 H
19.1 Allgemeines	534 H
19.2 Quadratische Programme	541 H
19.3 Unbeschränkte lineare MPR	543 H
19.4 Beschränkte lineare MPR	545 H
19.5 Einstellparameter	548 H
19.6 Stationäre Genauigkeit	550 H
19.7 Stabilität	554 E
19.8 Nichtlineare MPR	555 E
20 Iterativ Lernende Regelung	561 H
20.1 Allgemeines	561 H
20.2 Systemtheoretische Betrachtung	562 H
20.3 Entwurf des Lernoperators	569 H
20.4 Normoptimale ILR	572 H
Literaturverzeichnis	573
Bezeichnungen	578
Index	580

1 Einführung

1.1 Steuerung und Regelung

Die Regelungstechnik beschäftigt sich mit der gezielten Beeinflussung von Prozessen, damit diese in einer gewünschten Art und Weise ablaufen. Zur Einführung in regelungstechnische Fragestellungen soll dabei ein Beispiel betrachtet werden: In Bild 1-1 besteht die Aufgabe darin, die Temperatur y eines Wohnraumes auf einen vorgegebenen Wert w zu bringen. Um das zu erreichen, muss die Stellung des Heizungsventils u passend gewählt werden. Allerdings wirken auch andere Einflüsse wie eine Änderung der Außentemperatur z_1 oder das Öffnen und Schließen von Fenstern und Türen z_2 auf die Temperatur y . Diese sollen die Temperatur nicht oder nicht wesentlich beeinflussen.

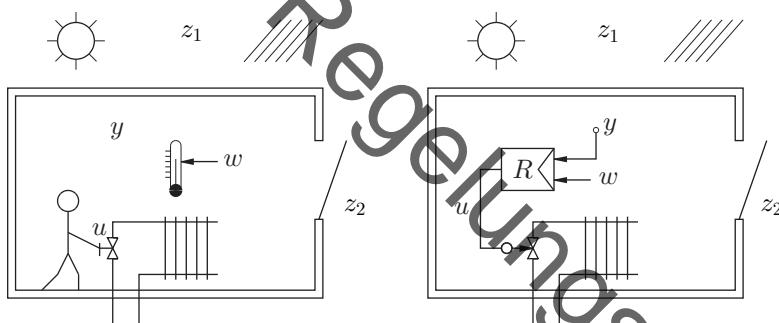


Bild 1-1: Raumtemperaturregelung von Hand und automatisch

Die intuitive Lösung besteht darin, die Temperatur y zu messen und das Heizungsventil entsprechend des gemessenen Wertes und des vorgegebenen Wertes w zu verstellen. Dabei wird man das Ventil öffnen, wenn die gemessene Temperatur y kleiner als die gewünschte Temperatur w ist. Zu große Messwerte der Temperatur y werden hingegen zu einem Schließen des Ventils führen. Dies kann von Hand geschehen oder durch ein zu entwerfendes Gerät, dass Regler genannt wird und in Bild 1-1 mit R bezeichnet ist.

Versucht man nun den in Bild 1-1 gegebenen Sachverhalt in seinen wirkungsmäßigen Zusammenhängen schematisch zu fassen, so erhält man eine

Darstellung wie in Bild 1-2. Aus dieser Darstellung ist zu erkennen, dass die Stellung des Heizungsventils u die Temperatur y beeinflusst, seinerseits aber auch auf Basis der Messung von y festgelegt wird. Folglich beeinflusst die Temperatur y über Regler und Heizungsventil sich selbst. Dieser in sich geschlossene Wirkungsablauf wird auch „geschlossener Regelkreis“ genannt und ist das zentrale Merkmal jeder Regelung.

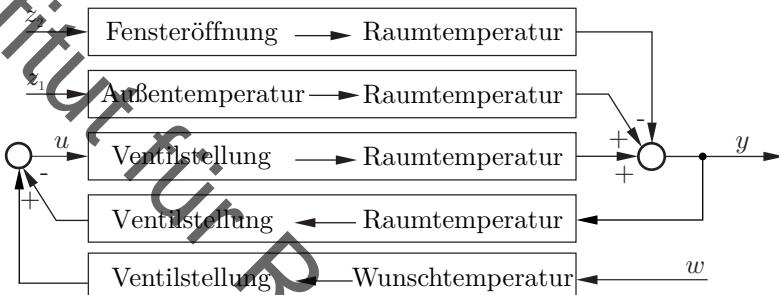


Bild 1-2: Schematische Darstellung der Raumtemperaturregelung

Vergleichend zum Konzept der Regelung soll ein anderer weit verbreiterter Lösungsansatz für das geschilderte Beispielproblem vorgestellt werden. So könnte man abweichend versuchen, das beschriebene Ziel durch z. B. das Messen der Außentemperatur z_1 und ein Verstellen des Heizungsventils entsprechend dieser Temperatur zu erreichen. Dieser Lösungsansatz und seine schematische Darstellung sind in Bild 1-3 und Bild 1-4 gezeigt. Offenbar liegt hier kein in sich geschlossener Wirkungsablauf vor. Zur Unterscheidung beider Ansätze spricht man in diesem Fall nicht von einer Regelung, sondern einer Steuerung, die entsprechend mit S bezeichnet ist. Aufgrund des fehlenden geschlossenen Wirkungsablaufes können größere Abweichungen der Temperatur y vom gewünschten Wert auftreten, wenn der nicht erfasste Einfluss z_2 (hier Öffnen von Fenstern usw.) entsprechende Werte annimmt oder wenn der im Steuergerät S enthaltene funktionale Zusammenhang zwischen Außentemperatur und Ventilstellung den Eigenschaften von Heizung und Gebäude nicht genau entspricht. Dies ist ein entscheidender Nachteil einer Steuerung gegenüber einer Regelung.

Neben dem Beispiel der Temperaturregelung gibt es zahllose weitere Anwendungsfälle, die auch nicht ausschließlich dem technischen Umfeld ent-

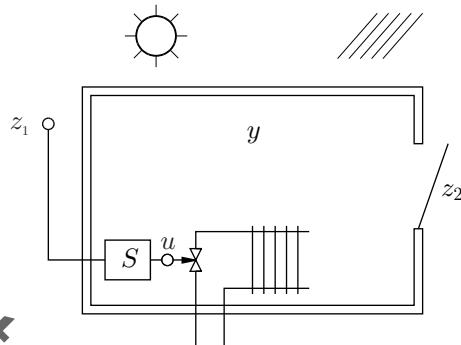


Bild 1-3: Raumtemperatursteuerung

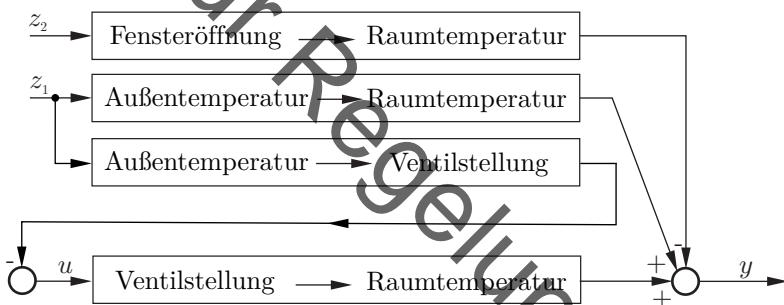


Bild 1-4: Schematische Darstellung der Raumtemperatursteuerung

stammen müssen. Ein Einblick in die Bandbreite möglicher Anwendungen wird dabei in Form von ausgewählten Beispielen aus dem Forschungsumfeld des Instituts für Regelungstechnik der RWTH Aachen in Abschnitt 1.3 gegeben. Trotz der sehr unterschiedlichen Aufgaben haben alle Regelungen eine ähnliche Struktur, die mit dem in sich geschlossenen Wirkungsablauf von Regelungen einhergeht.

1.2 Grundstruktur des Regelkreises

Der entscheidende Unterschied zwischen Regelung und Steuerung ist der geschlossene Wirkungsablauf im Falle einer Regelung, welcher besonders

leicht in der schematischen Darstellung der Wirkungszusammenhänge in den Bildern 1-2 und 1-4 identifiziert werden kann. Zur formalisierten Beschreibung einer Regelung bietet sich daher eine vergleichbare Darstellung, jedoch mit klaren Definitionen und festen Regeln und Bezeichnungen an. Eine solche Beschreibung ist der sogenannte Wirkungsplan, welcher in der DIN IEC 60050-351 Leittechnik [9] festgelegt ist.

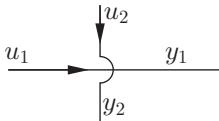
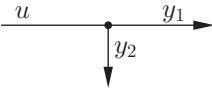
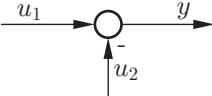
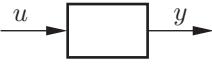
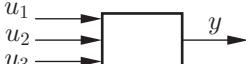
Bezeichnung	Symbol	Funktion
Wirkungslinien Signalübertragung		$y_1 = u_1$ $y_2 = u_2$
Verzweigungsstelle		$y_1 = u$ $y_2 = u$
Summenpunkt		$y = u_1 - u_2$
Übertragungsblock	 	$y = f(u)$ $y = f(u_1, u_2, u_3, \dots)$

Tabelle 1-1: Elemente des Wirkungsplans

Hierzu werden die jeweils interessierenden physikalischen Größen als dynamische Signale aufgefasst, die durch entsprechende dynamische Systeme verändert und miteinander verknüpft werden. Die Fachwörter „Signal“ und

„System“ werden je nach Fachdisziplin sehr unterschiedlich gebraucht. Die beiden Begriffe werden hier folgendermaßen verwendet:

Signale und Systeme

Ein Signal ist eine (physikalische) Größe, deren Wert einen Informationsgehalt besitzt. Ein Signal heißt dynamisch, wenn sich der Wert des Signals über die Zeit ändern kann.

Ein System Σ ist eine durch eine Systemgrenze von der Umgebung abgegrenzte Einheit, die über Signale mit der Umgebung Informationen austauschen kann. Ein dynamisches System nutzt dabei dynamische Signale und man unterscheidet aufgrund des Ursache-Wirkungs-Prinzip zwischen Eingangssignalen $u(t)$, die auf das System einwirken, und Ausgangssignalen $y(t)$, die die Reaktion des Systems auf die Eingangssignale darstellen.

Zur Beschreibung des Systems nutzt man auch die Schreibweise $u(t) \mapsto y(t)$. Diese kennzeichnet, dass das beschriebene System dem Eingangssignal $u(t)$ das Ausgangssignal $y(t)$ zuordnet.

Die Elemente des Wirkungsplans stellen also gerichtete Operationen dar, wobei die an den Signalen angegebene Pfeilrichtung die Ursache-Wirkungs-Richtung angibt. Somit sind auf Systeme einwirkenden Eingangsgrößen als Ursachen aufzufassen, während die hervorgerufenen Ausgangsgrößen als Wirkungen zu verstehen sind. In diesem Sinne wird jedes einzelne Teilsystem für sich genommen als rückwirkungsfrei angesehen, d. h. dass Änderungen der Ausgangsgröße eines Elementes keinen Einfluss auf die zugehörige Eingangsgröße haben, sofern die Rückwirkung nicht durch andere Elemente des Wirkungsplans dargestellt wird. Weitere Elemente des Wirkungsplans sind die Verzweigung von Signalen an Verzweigungsstellen und die Addition von Signalen unter Beachtung von Vorzeichen an Summenpunkten. Zur besseren Unterscheidung von Verzweigungsstellen und sich kreuzenden Wirkungslinien wird empfohlen, letztgenannte mit einem Halbkreis zu kennzeichnen. Das positive Vorzeichen an Summenpunkten darf i. Allg. fortgelassen werden. Zur Vermeidung von Doppeldeutigkeiten sind negative Vorzeichen stets am Summenpunkt in *Pfeilrichtung rechts* vom Pfeil anzutragen (siehe Tab. 1-1). Durch zusätzliche Angaben in oder an den Blöcken sowie an den Signalen kann das System oder das Signal näher bezeichnet werden. Eine Zusammenstellung dieser Elemente des Wirkungs-

plans gibt Tab. 1-1.

Durch die Abstraktion physikalischer Größen und technischer Prozesse als Signale und Systeme wird die Regelungstechnik unabhängig von speziellen Eigenschaften des jeweiligen technischen Problems. Dadurch wird es möglich, die Regelungsaufgabe von einem konkreten Beispiel zu abstrahieren und die Grundstruktur eines Regelkreises allgemeingültig zu beschreiben. Je nach Verwendungszweck der im Regelkreis auftretenden Systeme haben sich dabei weitestgehend synonyme Bezeichnungen für „System“ etabliert.

Bezeichnungen für Systeme

Die folgenden Bezeichnungen werden in der Regelungstechnik weitestgehend synonym verwendet: System, Glied, Übertragungssystem, dynamisches System, Übertragungsblock, Regelkreisglied, Übertragungsglied, Regelkreiselement, Übertragungselement

Im Allgemeinen ist es das Ziel einer Regelung, bestimmte Ausgangsgrößen – die sogenannten *Regelgrößen* – eines (technischen) Prozesses – der sogenannten *Regelstrecke* – an vorgegebene *Führungsgrößen* anzugeleichen. Die Regelgrößen sollen sowohl Änderungen der Führungsgrößen möglichst gut folgen als auch von Störungen – auch *Störgrößen* genannt –, die auf die Regelstrecke einwirken, möglichst wenig beeinflusst werden. Die genannten Ziele werden dadurch angestrebt, dass die Regelgrößen gemessen und die Messergebnisse zusammen mit den Führungsgrößen einer zu entwerfenden Einheit, die *Regler* genannt wird, zur Verfügung gestellt werden. Der Regler leitet hieraus Eingriffe in den Prozess – die sogenannten *Stellgrößen* – ab, die geeignet sind, diese Ziele zu erreichen, d. h. die Differenz zwischen Regelgröße und Führungsgröße zu vermindern. Deswegen findet innerhalb des Reglers ein Abgleich zwischen aktuell gemessener Regelgröße (Istwert) und aktueller Führungsgröße (Sollwert) statt. Durch diese *Rückführung* (oder auch *Rückkopplung* genannt) der Regelgrößen auf die Stellgrößen, die wiederum die Regelgrößen beeinflussen, entsteht ein geschlossener Wirkungsablauf, der als *Regelkreise* bezeichnet wird. Bild 1-5 zeigt anhand eines Wirkungsplans diese einfachste Grundstruktur des Regelkreises und führt wichtige Begriffe und Bezeichnungen im Zusammenhang mit Regelungen ein, wie sie in DIN IEC 60050-351 Leittechnik [9] genormt sind.

Typische Benennungen der Signale sind y für die Regelgröße, z für die Störgröße, u für die Stellgröße und w für die Führungsgröße. Die Differenz zwi-

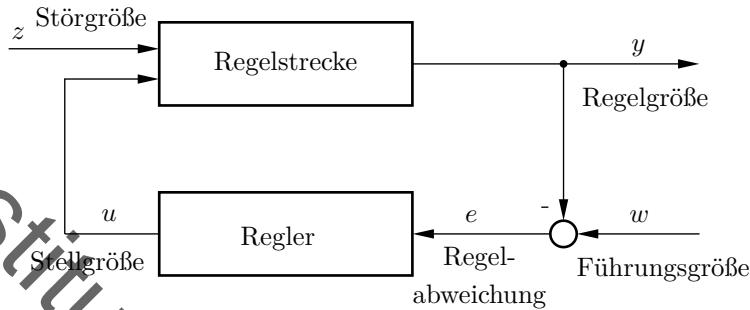


Bild 1-5: Einfachste Grundstruktur des Regelkreises

schen Regel- und Führungsgröße wird auch als Regelabweichung e bezeichnet und bildet im Standardfall die Eingangsgröße des Reglers. Komplexere Regelungen können hiervon abweichende Eingangsgrößen besitzen. So ist es möglich, die Messgröße und die Führungsgröße innerhalb der Regelung getrennt voneinander zu verarbeiten. Das wird insbesondere dann notwendig, wenn die Messgrößen y_m aufgrund von bspw. einer durch Messrauschen n verfälschten Messung nicht den eigentlichen Regelgrößen y entsprechen und eine zusätzliche Filterung der Messwerte notwendig wird. Diese allgemeine Grundstruktur ist in Bild 1-6 gezeigt.

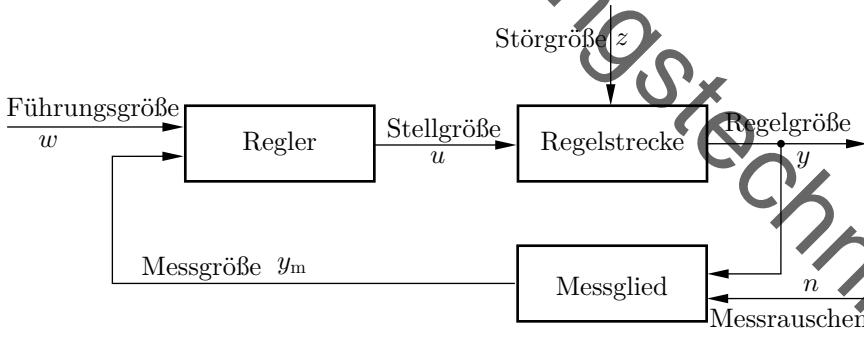


Bild 1-6: Allgemeine Grundstruktur des Regelkreises

Hieraus ergeben sich abschließend die folgenden Definitionen:

Definition der Signale und Systeme im Regelkreis

Die *Regelgröße* y ist die Ausgangsgröße der Regelstrecke, die auf einem vorgegebenen konstanten oder veränderlichen Wert gehalten werden soll.

Die *Führungsgröße* w (auch Sollwert genannt) ist eine der Regelung von außen zugeführte Größe, der die Regelgröße folgen soll.

Die *Stellgröße* u ist die Ausgangsgröße des Reglers und beeinflusst die Regelgröße über die Regelstrecke, z. B. um sie der Führungsgröße anzugelichen.

Störgröße z ist jede nicht beeinflussbare Größe, die von außen auf die Regelstrecke und damit auf die Regelgröße wirkt.

Die *Messgröße* y_m ist die Eingangsgröße der Regelung, die vom Messglied bereitgestellt wird und näherungsweise der Regelgröße entspricht.

Messrauschen n ist jede nicht beeinflussbare Größe, die von außen auf die Messgröße wirkt.

Als *Regelstrecke* wird ein System oder Prozess bezeichnet, dessen Ausgangsgrößen geregelt werden, indem eine oder mehrere Eingangsgrößen verändert werden.

Der *Regler* ist ein System oder Gerät, das Messgröße und Führungsgröße miteinander vergleicht und hieraus die Stellgröße bildet.

Das *Messglied* ist ein System oder Gerät, das unter Messrauschen eine Messung der Regelgröße bereitstellt.

Der Unterschied zwischen Steuerung und Regelung wird bei der Betrachtung des Wirkungsplan der Steuerung in Bild 1-7 besonders deutlich. Hierbei ist es möglich, dass die Steuerung wie in Bild 1-3 die Stellgröße auf Basis einer (gemessenen) Störgröße bildet, aber auch, dass die Berechnung der Stellgröße auf Basis der Führungsgröße erfolgt. Zentraler Unterschied ist, dass bei der Steuerung die Ausgangsgröße y nicht in die Berechnung der Stellgröße miteinfließt und somit keine Rückführung und kein geschlossener Wirkungsablauf vorliegt. Daher wird im Fall der Steuerung auch kein Messglied zur Erfassung der Regelgröße benötigt. Davon ausgehend haben sich die folgenden Definitionen eingebürgert.

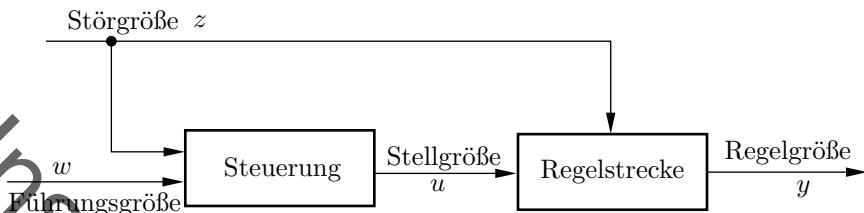


Bild 1-7: Allgemeine Grundstruktur einer Steuerung

Offener und Geschlossener Regelkreis

Das Übertragungsverhalten in Bild 1-6 mit Rückführung wird als *geschlossener Regelkreis* bezeichnet. Das Übertragungsverhalten in Bild 1-7 ohne Rückführung wird als *offener Regelkreis* bezeichnet.

Im Englischen wird bei einer Steuerung von *open-loop control* gesprochen, während eine Regelung als *closed-loop control* oder *feedback control* bezeichnet wird. Dies macht direkt deutlich, dass eine Regelung im Gegensatz zur Steuerung einen Regelkreis durch Rückführung der Ausgangsgrößen schließt. Der Begriff *control* ohne Zusatz verweist üblicherweise auf eine Regelung, ist an sich aber uneindeutig.

1.3 Beispiele technischer Regelungen

Die abstrakte Beschreibung technischer Prozesse als signalverarbeitende Systeme ermöglicht den universellen Einsatz der regelungstechnischen Methodik in zahllosen Anwendungsfeldern. Die folgenden Beispiele sollen einen unvollständigen Einblick in mögliche Anwendungen geben und die Anschauungskraft der Signalnamen stärken. Dabei sind die Anwendungen dem Forschungsumfeld des Instituts für Regelungstechnik der RWTH Aachen entnommen.

1.3.1 Tiefenregelung eines Unterwasserfahrzeugs

Die Existenz von extraterrestrischem Leben ist eine der großen ungeklärten Fragen der Menschheit. Vielversprechende Orte zum Auffinden von extraterrestrischem Leben in unserem Sonnensystem sind dabei die Eismonde

Europa und Enceladus, die Jupiter bzw. Saturn umrunden. Dort sind die zwei wichtigsten Voraussetzungen für Leben gegeben: Unter einer Wassereisdecke befinden sich große Ozeane aus Wasser und mit vulkanischer Aktivität ist eine Energiequelle vorhanden.

Im Forschungsprojekt TRIPLE [62] wird ein Erkundungssystem gemäß der Prinzipskizze in Bild 1-8 erforscht. Ausgehend von einer Oberflächenstation wird mit einem Schmelzroboter ein Kanal durch das Eis geschmolzen. Sobald das Wasserreservoir erreicht ist, wird ein Miniaturunterwasserfahrzeug (nanoAUV) ausgesetzt, welches die Umgebung erkundet, Proben sammelt und diese an den Schmelzroboter übermittelt.

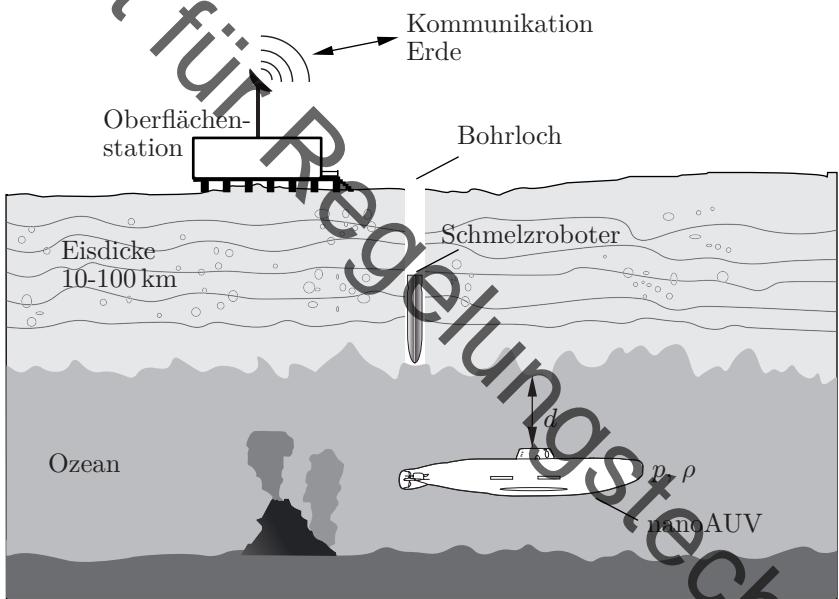


Bild 1-8: Übersicht über das TRIPLE Projekt.

Eine zentrale Herausforderung dieser Raumfahrtmission ist, dass alle Komponenten ihre Aufgaben selbstständig ohne menschlichen Eingriff erledigen müssen. Das liegt daran, dass keine Menschen auf die Mission gesandt werden können und aufgrund des großen Abstandes der Monde zur Erde die Zeitverzögerungen zu groß für eine terrestrische Fernsteuerung sind: Trifft

das nanoAUV beispielsweise auf unerwartete vertikale Strömungen, so muss es sein Tauchverhalten schnell anpassen, damit es nicht mit der Eisdecke oder dem Grund kollidiert. Die Eingabe eines menschlichen Bedieners auf der Erde käme hier zu spät. Stattdessen muss eine Tiefenregelung selbstständig derartige Störungen ausgleichen und eine von der Erde kommandierte Tauchtiefe des Fahrzeuges einstellen.

Bei der Umsetzung einer solchen Regelung ist zu beachten, dass die Tauchtiefe sich nicht direkt messen lässt, sondern aus dem hydrostatischen Druck p berechnet werden muss. Um die Tauchtiefe zu beeinflussen, kann das Unterwasserfahrzeug unter anderem über Tauchzellen seine Auftriebskraft F_A ändern, da diese das Volumen des vom Unterwasserfahrzeug verdrängten Wassers verändern. Hierzu wird durch eine Pumpe mit Spannung U der Druck in den Tauchzellen verändert. Der Auftrieb hängt aber nicht nur von den einstellbaren Tauchzellen ab, sondern wird auch durch nicht einstellbare Effekte beeinflusst. So variiert mit dem Salzgehalt des Wassers auch dessen Dichte ρ , was ebenfalls den Auftrieb verändert. Das Regelziel und die Führungsgröße w , die Regelgröße y , die Stellgröße u und die Störgrößen z lassen sich also tabellarisch wie folgt zusammenstellen.

Ziel: Sorge dafür, dass das nanoAUV eine vorgegebene Tauchtiefe einhält.

$w(t)$: Von der Erde aus vorgegebene Solltiefe

$y(t)$: Aus dem hydrostatischen Druck p geschätzte Tauchtiefe d

$u(t)$: Spannung U der Pumpe in den Tauchzellen

$z(t)$: Vertikale Strömungen; Wasserdichte ρ

Diese Tiefenregelung muss dabei in ein Regelungskonzept mit vielen Komponenten eingebettet werden. So muss das nanoAUV nicht nur seine Tauchtiefe regeln, sondern sich auch zielgerichtet vorwärts bewegen und lenken. Die Regelung der Vorwärtsbewegung geschieht über Propeller, die Lenkung durch eine bewegliche Masse zur Schwerpunktsänderung. Da auch die Vorgabe der Sollwerte für diese Regelkreise von der Erde aus zu lange benötigen würde, werden diese Regelungen in einer Gesamtlösung kombiniert, welche eine abzufahrende Trajektorie in Form von gewünschten Sollgrößen für die einzelnen Regelgrößen bestimmt. Somit kann mit steigendem Grad der Automatisierung der notwendige Eingriff des Menschen immer weiter zurückgefahren werden, sodass man am Ende ein selbstständig agierendes Explorationssystem erhält. Insofern kommt der Regelungstechnik

eine Schlüsselrolle in der Raumfahrt zu.

1.3.2 Regelung von Windkraftanlagen

Eine der ältesten Methoden, die Energie der Natur dem Menschen nutzbar zu machen, ist die Nutzung von Windenergie. Historisch trieb der Wind dabei zunächst mechanische Windmühlen an. Seit Mitte des 20. Jahrhunderts erzeugen Windkraftanlagen im großen Stil und werden stetig weiterentwickelt, sodass sie heute eine zentrale Rolle bei der Gewinnung regenerativer Energien einnehmen[4].

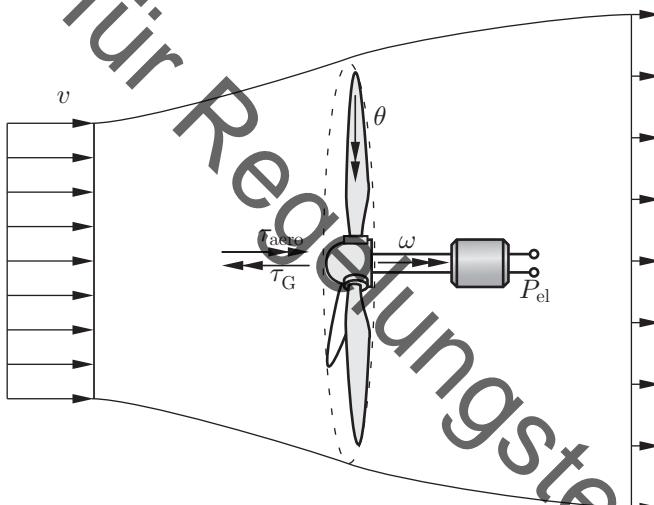


Bild 1-9: Vereinfachte Windkraftanlage bestehend aus den Komponenten Rotor, Antriebsstrang und Generator

Moderne Windkraftanlagen wandeln einen Teil der kinetischen Energie des Windes erst in kinetische Energie eines Rotors und anschließend in elektrische Energie um. Der Rotor mit seiner Massenträgheit J dreht sich mit der Winkelgeschwindigkeit ω . Er wird dabei durch ein aerodynamisches Moment τ_{aero} beschleunigt und durch ein aufgeprägtes Generatormoment τ_G gebremst.

Während das Generatormoment τ_G in bestimmten Betriebsgrenzen frei eingestellt werden kann, hängt das aerodynamische Moment τ_{aero} von drei anderen Größen ab. Die Windgeschwindigkeit v erhöht das Moment τ_{aero} . Hingegen wird eine nicht optimale Rotordrehzahl ω oder ein ungünstiger Anstellwinkel θ der Rotorblätter τ_{aero} reduzieren. Die optimale Drehzahl ω_{opt} , bei der die meiste Energie umgewandelt werden kann, hängt dabei direkt von der Windgeschwindigkeit v ab.

Ziel: Wandle innerhalb der Betriebsgrenzen möglichst viel kinetische Energie des Windes in elektrische Energie um.

$w(t)$: Nenndrehzahl / Optimale Drehzahl ω_{opt}

$y(t)$: Drehzahl ω des Rotors

$u(t)$: Drehmoment τ_G des Generators; Anstellwinkel θ der Rotorblätter

$z(t)$: Windgeschwindigkeit v

Der Windgeschwindigkeit kommt hier also eine doppelte Rolle zu: Zum einen bestimmt sie die optimale Drehzahl und hat somit einen direkten Einfluss auf die Führungsgröße. Zum anderen ist sie naturgemäß Schwankungen unterworfen und damit eine Störgröße.

Aus dem Ziel, die maximale Energie innerhalb der Betriebsgrenzen zu entnehmen, ergibt sich eine zweiteilige Aufgabe für eine Regelung: Bei kleinen Windgeschwindigkeiten, wenn weniger Energie bereitsteht als der Generator umwandeln darf, muss die Drehzahl der Windgeschwindigkeit nachgeführt werden, um stets die maximale Energie zu entnehmen. Bei höheren Windgeschwindigkeiten übersteigt das verfügbare aerodynamische Moment das maximal mögliche abtriebende Generatormoment. Dieser Einfluss der Windgeschwindigkeit auf das aerodynamische Moment muss dann durch den Anstellwinkel der Rotorblätter kompensiert werden.

Über das klassische Regelziel der Leistungsmaximierung können mit modernen Regelstrategien weitere untergeordnete und zum Teil gegenläufige Ziele verfolgt werden. Diese können beispielsweise netzstützende Dienste sein, bei der die Rotordrehzahl kurzzeitig bewusst verändert wird, um mehr oder weniger Leistung ins Stromnetz zu speisen. Zusätzlich zur Leistung lassen sich auch mechanische Lasten aktiv regeln, um die Lebensdauer von Windkraftanlagen zu verlängern. [4]

1.3.3 Kraftregelung beim Fräsen

Das Fräsen ist ein weit verbreitetes und flexibles Verfahren zur Herstellung komplexer Geometrien mit hoher Genauigkeit. Während des FräSENS wird ein rotierendes Werkzeug gegen ein Bauteil verfahren, sodass gezielt Material abgenommen wird, bis das Bauteil eine gewünschte Geometrie erhält. Insbesondere beim Schrubbfräsen soll dabei in möglichst kurzer Zeit so viel Material wie möglich abgetragen werden, damit die Endgeometrie schnellstmöglich hergestellt werden kann. Dieses Ziel kann durch eine möglichst hohe Vorschubgeschwindigkeit des Fräserwerkzeugs erreicht werden. Mit der Vorschubgeschwindigkeit wächst aber auch die sogenannte Aktivkraft F auf das Werkzeug, die beschreibt, mit welcher Kraft in der Arbeitsebene Material abgetragen wird. Eine zu große Aktivkraft führt dabei zu einer unerwünschten Verformung oder gar zum Bruch des Werkzeugs.

Fräsmaschinen existieren in verschiedenen Bauweisen, die sich maßgeblich in der Anzahl und der Anordnung ihrer Bewegungssachsen unterscheiden. Bild 1-10 zeigt das Prinzip schematisch für eine einzelne linearen Bewegungssachse gezeigt: Die Motorspannung U regelt die Vorschubgeschwindigkeit v des Fräswerkzeugs, wobei auch Störeinflüsse wie Schlupf auf der Linearachse die Vorschubgeschwindigkeit beeinflussen.

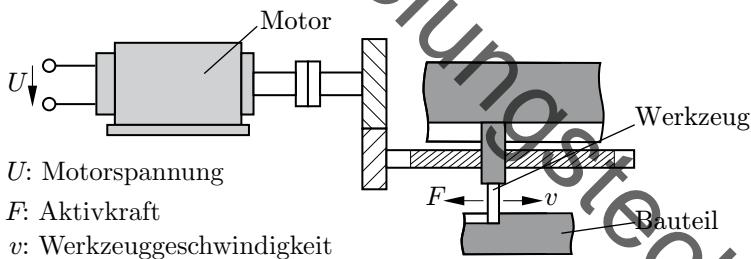


Bild 1-10: Schematische Darstellung der Werkzeugpositionierung beim Fräsen

Nimmt man in einem ersten Schritt an, dass die Aktivkraft nur von der Vorschubgeschwindigkeit abhängt, so kann man aus der zulässigen Maximalkraft eine zulässige Maximalgeschwindigkeit ableiten. Diese kann dann von einer Geschwindigkeitsregelung eingestellt werden. So minimiert sich die Prozesszeit bei gleichzeitig Schutz des Werkzeugs.

Ziel: Minimiere die Prozesszeit, ohne die Maximalkraft zu überschreiten.

$w(t)$: Sollgeschwindigkeit

$y(t)$: Werkzeugvorschubgeschwindigkeit v

$U(t)$: Motorspannung U

$z(t)$: Schlupf

Mit dieser Strukturierung der Regelungsaufgabe ist es in der Praxis allerdings nicht getan, da die Aktivkraft nicht nur von der Werkzeugvorschubgeschwindigkeit abhängt. So wirken auch weitere Faktoren wie beispielsweise die Fliehkraft auf das Werkzeug. Somit kann die Werkzeugvorschubgeschwindigkeit nicht gut geführt werden, ohne ebenfalls die Prozesskraft zu regeln und umgekehrt [55]. Daher bietet sich eine kombinierte Kraft- und Geschwindigkeitsregelung an, wie sie in Bild 1-11 vereinfacht dargestellt ist. Hierbei werden zwei verschachtelte Regler genutzt, wobei die Stellgröße des Kraftregelkreises der Führungsgröße des Geschwindigkeitsregelkreises entspricht. Diese Struktur tritt bei komplexeren Regelungsaufgaben häufig auf und wird in Kapitel 12 intensiver beleuchtet.

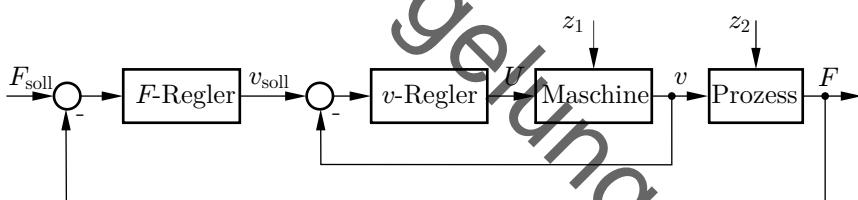


Bild 1-11: Aufbau des Regelungssystems aus Kraft- und Geschwindigkeitsregelung

Ziel: Minimiere die Prozesszeit ohne die Maximalkraft zu überschreiten.

$w(t)$: Sollgeschwindigkeit (siehe $u(t)$); Sollkraft

$y(t)$: Werkzeugvorschubgeschwindigkeit v ; Aktivkraft F

$u(t)$: Motorspannung U ; Sollgeschwindigkeit (siehe $w(t)$)

$z(t)$: Schlupf (z_1); Fliehkraft (z_2)

Die Minimierung der Prozesszeiten wird dabei dadurch erreicht, dass im Prozess diejenige Kraft eingestellt wird, die das Werkzeug maximal aus-

hält ohne sich zu verformen. Zu diesem Zweck muss im Prozess eine entsprechende maximale Geschwindigkeit eingestellt werden. So trägt sich die Regelungsaufgabe in der gegebenen verschachtelten Darstellung von außen nach innen.

1.3.4 Regelung eines Bioreaktors

Das Herz ist ein stark beanspruchtes Organ, das ständigen Druckwechseln ausgesetzt ist. Obgleich das menschliche Herz für diese Aufgabe in seiner Konstruktion mit sehr stabilen aber auch flexiblen Herzklappen bestens vorbereitet ist, kommt es immer wieder zu Ausfällen, die eine Herzoperation und ein Ersetzen der Herzklappen notwendig machen. Als Ersatz werden derzeit biologische Transplantate vom Mensch oder Tier und Implantate aus Metall oder Keramik genutzt. Sowohl Transplantate als auch künstliche Herzklappen haben gewichtige Nachteile wie Abstoßungsreaktionen, Verwachsungen, Bruch und viele mehr. Ein moderner Versuch, die Vorteile beider Ansätze zu vereinen, ist das Tissue-Engineering. Dabei lässt man körpereigene Zellen auf strukturgebendem Textilgewebe zu Herzklappen reifen. Die Zellen reifen in einem Bioreaktor wie in Bild 1-12, der möglichst identische Bedingungen wie im Herzen bieten soll. Neben den Eigenschaften des Nährmediums (Temperatur, pH-Wert, O₂ bzw. CO₂-Sättigung) beeinflussen insbesondere auch biomechanische Stimuli wie Druckgradienten und Scherkräfte das Wachstum der Zellen.

Im Folgenden wird nur das mechanische System näher betrachtet. Ein Herzzyklus wird im Wesentlichen durch die Verläufe der Drücke p_v vor und p_h hinter der Herzklappe charakterisiert, woraus sich der Druckunterschied $\Delta p = p_v - p_h$ ergibt. Ist Δp positiv, öffnet sich die Herzklappe, sodass das Nährmedium als Volumenstrom Q durch die geöffnete Herzklappe fließt und Scherkräfte auf diese ausübt. Ist Δp negativ, schließt sich die Herzklappe und der Fluss des Nährmediums wird unterbrochen. Folglich gilt $Q = 0$ und es entsteht ein Druckgradient entlang der Zellen der Herzklappe. Das Medium fließt durch das Quetschventil in die Kammer vor der Herzklappe zurück.

Die Funktionsweise des Bioreaktors in seinem mechanischen Aufbau entspricht näherungsweise dem des menschlichen Herzens. Für eine gute Nachbildung des Herzens müssen aber auch die Verläufe der Drücke und des Vo-

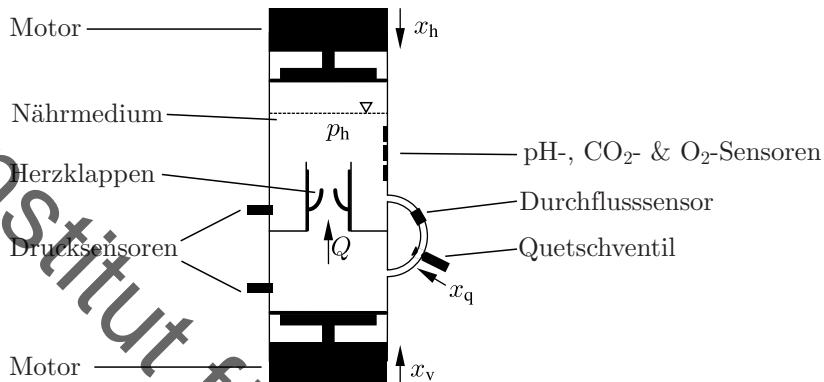


Bild 1-12: Regelbarer Bioreaktor in Anlehnung an Voß et al. [61].

lumenstroms im Bioreaktor den Verläufen eines echten Herzzyklus gleichen. Um p_v , p_h und Q einstellen zu können, stehen dabei die beiden Motoren mit Position x_v und x_h sowie die Position des Quetschventils x_q zur Verfügung. Um diese Größen reproduzierbar einstellen zu können, reicht dabei eine reine Steuerung nicht aus. Durch das Wachstum der Herzklappe wird nämlich der Zusammenhang zwischen Δp und Q nicht fest gegeben sein, sondern sich mit der Zeit verändern. Hiermit ändern sich auch die mechanischen Eigenschaften. Diese Modellunsicherheiten lassen sich als Störungen interpretieren, die zum Ausgleich eine Regelung erfordern.

Ziel: Sorge dafür, dass die Herzklappe biomechanischen Bedingungen wie im echten Herzen ausgesetzt ist.

w(t): Verläufe für p_v , p_h und Q wie im menschlichen Herzen

y(t): Drücke p_v und p_h und Durchfluss v

u(t): Positionierung der Motoren x_v und x_h ; Öffnung x_q des Quetschventils

z(t): Unbekanntes, zeitveränderliches Verhalten der Herzklappe

Aus dieser strukturellen Auflistung lassen sich bereits einige praktische Anforderungen an die Ausstattung des zu regelnden Systems ableiten. So wird es nicht möglich sein, drei Regelgrößen p_v , p_h und Q mit nur zwei Stellgrößen unabhängig voneinander einzustellen. Daher wird neben dem Motor vor der Herzklappe und dem Quetschventil auch ein weiterer Motor hinter der

Herzklappe benötigt. Außerdem ist die Regelgröße Q leider nicht messbar, da kein Sensor direkt an der Herzklappe angebracht werden kann. Dies erschwert die Regelungsaufgabe: Anstelle eines Durchflusssensors am Herzen muss ein Durchflusssensor am Quetschventil angebracht werden. Aus dem Durchfluss am Quetschventil und den gemessenen Drücken lässt sich dann der Durchfluss Q schätzen.

1.3.5 Regelung einer Dampfmaschine

Von besonderem historischen Interesse ist das letzte Beispiel für technische Regelungen, das abweichend nicht dem Umfeld der RWTH entstammt, sondern seine Wurzeln in Großbritannien besitzt. Der sogenannte Fliehkraftregler wurde James Watt¹ 1788 für Dampfmaschinen entwickelt. Dieser gilt – aufbauend auf älteren Reglern für Windmühlen – als der erste industriell genutzte Regler und als Geburtsstunde der modernen Regelungstechnik.

Das Wirkprinzip einer Dampfmaschine besteht darin, dass über den Aufbau von Druck eine Antriebswelle in Drehung versetzt wird. Dies wird in Bild 1-13 visualisiert. Die Dampfmaschinen wurden im 18. Jahrhundert als universelle Kraftquelle für zahlreiche andere Maschinen verwendet, indem man diese nach Bedarf mit einem Riemen mit der Antriebswelle verband. Der Anschluss dieser Maschinen an die Antriebswelle sorgt für ein Lastmoment τ , welches ohne weiteren Eingriff zu einem unvorteilhaften Abfall der Drehzahl bzw. Winkelgeschwindigkeit ω führt. Den gleichen Effekt hat eine Veränderung des Frischdampfdrucks p im Zulauf der Dampfmaschine. Um die Drehzahl der Dampfmaschine konstant zu halten, musste daher der Maschinenbediener das Dampfeinlassventil entsprechend öffnen oder schließen.

Ziel: Halte die Drehzahl der Dampfmaschine auf einem konstanten Wert

$w(t)$: Solldrehzahl der Dampfmaschine (üblicherweise konstant)

$y(t)$: Winkelgeschwindigkeit ω der Dampfmaschine

$u(t)$: Verstellwinkel des Dampfeinlassventsils

$z(t)$: Frischdampfdruck p ; auf Dampfmaschine wirkendes Lastmoment τ

Der von Watt entwickelte Fliehkraftregler in Bild 1-13 erledigt durch einen geschickten mechanischen Aufbau die beschriebene Aufgabe automatisch.

¹James Watt (1736 - 1819), schottischer Erfinder [36]

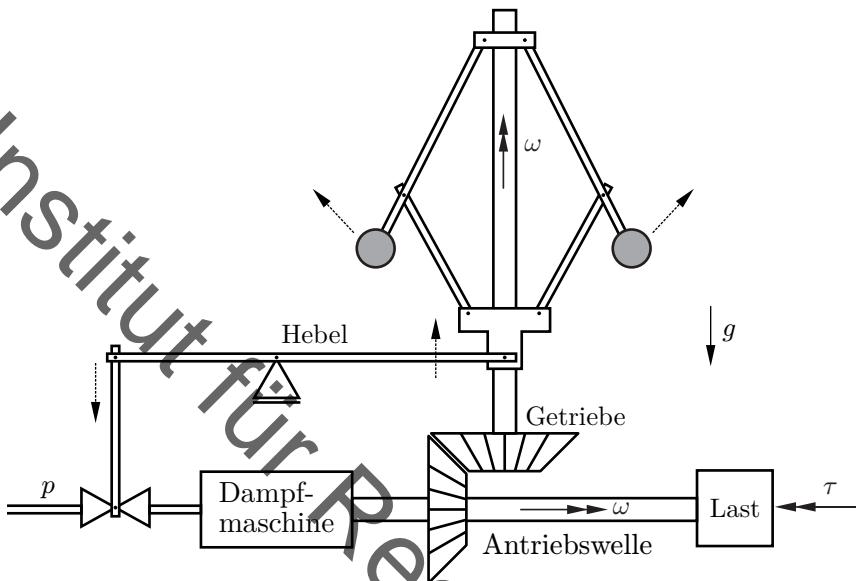


Bild 1-13: Dampfmaschine mit Fliehkraftregler

Hierzu wird über ein Getriebe die Drehzahl der Welle auf die Drehzahl eines Fliehkraftpendels übersetzt. Steigt die Drehzahl der Dampfmaschine, so wird auch die Fliehkraft größer und die beiden befestigten Massen werden nach außen gehoben (gestrichelte Pfeile in Bild 1-13). Hierdurch wird über eine mechanische Verbindung das Dampfeinlassventil geschlossen, wodurch die zugeführte Dampfmenge und damit die Drehzahl verringert wird. Läuft die Dampfmaschine hingegen zu langsam, fallen die Kugeln und das Ventil öffnet sich. Die Konstruktion des Fliehkraftreglers ist folglich die mechanische Realisierung eines Reglers.

Die Wirkweise des Reglers lässt sich durch eine Veränderung des Auflagepunktes des Hebels verändern. Heutzutage werden Regler in den allermeisten Fällen digital auf einem Mikrocontroller umgesetzt, während andere Umsetzungen nur in wenigen Anwendungen wie beispielsweise sicherheitskritischen Bauteilen der Raumfahrt verwendet werden.

2 Modellbildung

2.1 Allgemeines

Aus dem in sich geschlossenen Wirkungsablauf von Regelungen entsteht eine Reihe von Herausforderungen, die allen Anwendungsfällen gemeinsam sind. Zu Gunsten einer allgemein, d. h. in allen Fachdisziplinen anwendbaren Methodik setzen die entsprechenden Analyse- und Entwurfsverfahren der Regelungstechnik auf einer mathematischen Beschreibung der betrachteten realen Prozesse auf, die von deren spezieller technischer Ausprägung abstrahiert und eine Modellbildung voraussetzt.

Modell

Ein Modell ist eine Beschreibung, die nur einen Teil der Eigenschaften des Originals wiedergibt. Ein richtig gewähltes Modell zeichnet sich dadurch aus, dass es alle wichtigen Eigenschaften des Originals widerspiegelt und gleichzeitig auf überflüssige Eigenschaften verzichtet.

Die Modellbildung realer Prozesse verfolgt somit eigentlich zwei unterscheidbare Ziele: Das ist zum einen das Schaffen einer für die Analyse- und Entwurfsverfahren handhabbaren mathematischen Beschreibung. Es umfasst zum anderen aber auch das Herausarbeiten der für eine Regelung relevanten Eigenschaften der Prozesse. Einerseits können bei dieser Abstraktion gegenüber den in den jeweiligen Fachdisziplinen gebräuchlichen Prozessbeschreibungen oft zulässige Vereinfachungen vorgenommen werden, da diese für Regelungstechnische Zwecke nur eine untergeordnete Rolle spielen. Andererseits kann es sein, dass für den Reglerentwurf wichtige Informationen – insbesondere zu den dynamischen Eigenschaften – in gängigen Modellen fehlen. Das Erlernen und die Anwendung der Regelungstechnik erfordert daher neben mathematischem Rüstzeug und Regelungstechnischem Verständnis auch ein Grundwissen in den klassischen Ingenieurdisziplinen Maschinenbau, Verfahrenstechnik und Elektrotechnik.

Bild 2-1 zeigt schematisch, wie Regelungstechnische Probleme mit Hilfe von Modellen zu lösen sind. Es wird ein Modell des Prozesses gebildet und ein Modell des Reglers anhand des Prozessmodells entworfen. Der reale Regler – in früheren Zeiten wie beim Beispiel des Fliehkraftreglers in 1.3 in analoger Ausführung, heutzutage meist als Algorithmus auf einem Mikrocontroller

- ergibt sich dann aus der Rückübertragung des Modells in die Realität.

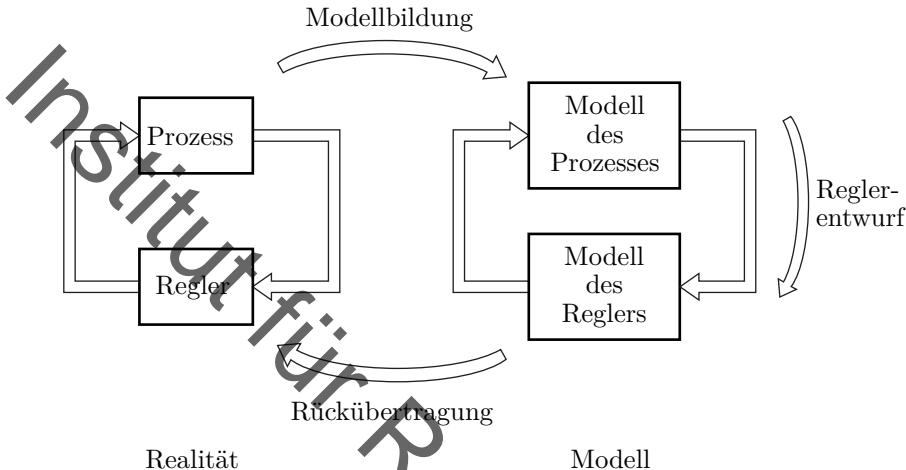


Bild 2-1: Lösung technischer Probleme mit Modellen

Grundauffassung der regelungstechnischen Modellierung ist die Vorstellung von Signalen und Systemen, die auch dem Wirkungsplan zugrunde liegt. Jedes Teilsystem hat eine oder mehrere Eingangsgrößen und eine oder mehrere Ausgangsgrößen. Das Modell beschreibt dann den Zusammenhang der Eingangs- und Ausgangsgrößen für das jeweilige Teilsystem. Quantitative Aussagen über Regelungen und Steuerungen lassen sich dabei nur mit quantitativen mathematischen Modellen der dabei zusammenwirkenden Systeme gewinnen. Daher werden nur solche Modelle im Folgenden behandelt.

Quantitative Modelle für Systeme können im Prinzip auf zwei gänzlich verschiedenen Wegen gewonnen werden. Eine Möglichkeit ist die Messung und das Experiment am realen System, wobei aber nicht alle Systeme für Experimente zur Verfügung stehen. Das Verfahren wird als experimentelle Modellbildung oder Identifikation (siehe auch Kapitel 8) und das Ergebnis gelegentlich als „black-box-Modell“ bezeichnet. Die andere Möglichkeit ist die Nutzung von Einsicht in die Wirkungsweise und Ansetzen und Verknüpfen der entsprechenden physikalischen oder auch chemischen Grundgleichungen. Dieses Verfahren wird als theoretische Modellbildung und das Ergebnis manchmal als „white-box-Modell“ bezeichnet. Im Gegensatz zur

experimentellen Modellbildung wird ein reales System nicht benötigt. Oft wird eine Kombination beider Verfahren benutzt und nur in den Bereichen gemessen und experimentiert, für die auf anderen Wegen keine Modelle zu gewinnen sind („grey-box-Modell“).

2.2 Einführung in Differentialgleichungen

Da die Ausgangsgrößen sich auch bei unverändertem Eingangssignal mit der Zeit verändern können, muss eine Modellform gewählt werden, die die Zeitaabhängigkeit des Systems berücksichtigt. Die grundlegende Modellform hierfür ist die Differentialgleichung. Aus Gründen der Übersichtlichkeit werden dabei zunächst nur Systeme mit einer Eingangsgröße $u(t)$ und einer Ausgangsgröße $y(t)$ betrachtet.

Eine gewöhnliche Differentialgleichung ist eine mathematische Gleichung für eine gesuchte Funktion in der Zeit, in der auch die Zeitableitungen dieser Funktion vorkommen. In expliziter Darstellung lässt sich eine gewöhnliche Differentialgleichung mit Eingangsgröße $u(t)$ und Ausgangsgröße $y(t)$ wie folgt schreiben:

$$y^{(n)}(t) = f \left(y^{(n-1)}(t), \dots, \ddot{y}(t), \dot{y}(t), y(t), u^{(m)}(t), \dots, u(t), t \right) . \quad (2.1)$$

Hier bezeichnet $y^{(k)}(t)$ die k -te Ableitung von $y(t)$ nach der Zeit. Ableitungen niedriger Ordnung werden durch entsprechende Punkte gekennzeichnet: $\dot{y}(t) = \frac{dy}{dt}$.

Systemordnung

Die *Ordnung des Systems* entspricht n und somit der höchsten Ableitung der Ausgangsgröße.

Als Beispiel zur Erläuterung der mathematischen Definitionen diene der in Bild 2-2 skizzierte Einmassenschwinger mit Feder-Dämpferauflage, dessen Auslenkung y um seine Ruhelage (= Ausgangsgröße) in Folge einer einwirkenden Kraft f (= Eingangsgröße) interessiert.

Die Newtonsche Bewegungsgleichung liefert mit der Masse M , der Federsteifigkeit C und dem Dämpfungsbeiwert B die Differentialgleichung

$$M\ddot{y} + B\dot{y} + Cy = f , \quad (2.2)$$

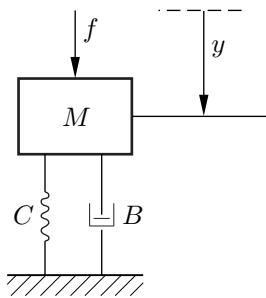


Bild 2-2: Einmassenschwinger

die nach der Auflösung nach \ddot{y}

$$\ddot{y} = \frac{1}{M} (-By - Cy + f) \quad (2.3)$$

ergibt. Der Einmassenschwinger besitzt also die Systemordnung zwei.

Tauchen in der Differentialgleichung zusätzlich noch andere Ableitungen wie beispielsweise nach dem Ort auf, so spricht man von einer partiellen Differentialgleichung, die zur Beschreibung sogenannter verteiltparametrischer Systeme herangezogen werden muss, für deren Behandlung auf entsprechende Spezialliteratur verwiesen sei [7, 11].

Zur vollständigen Beschreibung eines technischen Systems ist eine Differentialgleichung noch nicht ausreichend, da so die Verläufe der Ausgangsgrößen y im Allgemeinen noch nicht bestimmt werden können. Hierzu werden noch die *Anfangsbedingungen* der Differentialgleichung benötigt.

Anfangsbedingungen

Die Bedingungen

$$y^{n-1}(t_0) = {}_0y^{n-1}, \dots, \dot{y}(t_0) = {}_0\dot{y}, y(t_0) = {}_0y \quad (2.4)$$

heißen Anfangsbedingungen einer Differentialgleichung n -ter Ordnung zum Zeitpunkt t_0 . Zur Unterscheidung von Indizierungen wie y_i wird die 0 zur Kennzeichnung einer Anfangsbedingung auf die linke Seite geschrieben: ${}_0y$

Die Existenz der Lösung einer gewöhnlichen Differentialgleichung unter einer solchen Anfangsbedingung ist für die Zeitpunkte $t \geq t_0$ unter bestimmten mathematischen Bedingungen an die Funktion f und die Funktion u sichergestellt [43]. Diese Anforderungen sind für Differentialgleichungen und Eingangssignale, die technische Abläufe beschreiben, stets erfüllt. Für die Ermittlung der Lösung ist dabei der Verlauf von u oder y für $t < t_0$ unerheblich. Die für den zukünftigen Verlauf relevanten Informationen aus der Vergangenheit werden nämlich vollständig in der Anfangsbedingung kodiert. Für zeitinvariante Systeme lässt sich außerdem ohne Beschränkung der Allgemeinheit der Zeitpunkt t_0 auf $t_0 = 0$ setzen. Daher hat es sich in der Regelungstechnik eingebürgert, alle Signale nur für positive Zeiten $t \geq 0$ zu betrachten und die Signale für negative Zeiten zu Null zu setzen, ohne dies gesondert zu kennzeichnen. Dies vereinfacht die Schreibweise, erfordert aber zusätzliche Aufmerksamkeit, wenn Signale in einer Umgebung von $t = 0$ betrachtet werden.

Differentialgleichungen und die durch sie beschriebenen Systeme können in Abhängigkeit der Eigenschaften der Funktion f klassifiziert werden.

Zeitinvariante Systeme

Ist im Gegensatz zu Gl.(2.1) f nicht explizit von der Zeit abhängig, d. h.

$$y^{(n)}(t) = f\left(y^{(n-1)}(t), \dots, \ddot{y}(t), \dot{y}(t), y(t), u^{(m)}(t), \dots, u(t)\right) \quad , \quad (2.5)$$

so heißt die Differentialgleichung und das durch die Differentialgleichung beschriebene System *zeitinvariant*.

Der Einmassenschwinger ist beispielsweise zeitinvariant, da t in Gl.(2.3) nicht auftaucht. Zeitinvariante Systeme sind nicht explizit von der Zeit abhängig; das bedeutet aber nicht, dass sie von der Zeit unabhängig sind, da die Eingangsgröße $u(t)$ eine Funktion der Zeit ist. Allerdings gibt es keine Zeitabhängigkeit, die nicht durch die Ein- und Ausgangsgrößen erfasst wird.

Zeitinvariante Systeme haben die nützliche Eigenschaft, dass die Systemantwort $y(t)$ bei gleichbleibender Anfangsbedingung unabhängig davon ist, wann sich die Eingangsgröße $u(t)$ verändert. Das heißt, dass eine Verschiebung der Eingangsgröße auf der Zeitachse eine gleich große Verschiebung

der Ausgangsgröße bewirkt, also

$$u(t) \mapsto y(t) \Rightarrow u(t-T) \mapsto y(t-T) . \quad (2.6)$$

Physikalisch bedeutet dies, dass sich das Übertragungsverhalten des Systems selbst nicht über die Zeit ändert, auch wenn die Ein- und Ausgangssignale dies tun. Viele Systeme können in der Regelungstechnik als zumindest näherungsweise zeitinvariant angenommen werden, da entsprechende Systemveränderungen durch z. B. Verschleiß entsprechend langsam ablaufen und für die relevante Dynamik der Regelung keine Rolle spielen.

Lineare Systeme

Ist f eine lineare Funktion, d. h.

$$y^{(n)}(t) = a_{n-1}(t)y^{(n-1)} + \dots + a_2(t)\ddot{y}(t) + a_1(t)\dot{y}(t) + a_0(t)y(t) + b_m(t)u^{(m)}(t) + \dots + b_0(t)u(t) , \quad (2.7)$$

so heißt die Differentialgleichung und das durch die Differentialgleichung beschriebene System *linear*. Andernfalls heißt die Differentialgleichung und das durch die Differentialgleichung beschriebene System *nichtlinear*.

Der Einmassenschwinger ist linear, da f in Gl.(2.3) linear ist. Für lineare Systeme gelten das *Verstärkungsprinzip* und das *Überlagerungsprinzip*. Das Verstärkungsprinzip besagt, dass einer mit einem beliebigen konstanten Faktor c multiplizierten Eingangsgröße $c \cdot u(t)$ eine Ausgangsgröße $c \cdot y(t)$ zugeordnet wird, d. h.

$$u(t) \mapsto y(t) \Rightarrow c \cdot u(t) \mapsto c \cdot y(t) . \quad (2.8)$$

Das Überlagerungsprinzip behandelt den Fall, dass die Eingangsgröße aus mehreren Komponenten $u(t) = u_1(t) + u_2(t) + \dots$ besteht. Es besagt, dass die zugehörige Ausgangsgröße in gleicher Weise, nämlich als $y(t) = y_1(t) + y_2(t) + \dots$, gebildet werden kann:

$$u_i(t) \mapsto y_i(t) \Rightarrow \sum u_i(t) \mapsto \sum y_i(t) . \quad (2.9)$$

Tatsächlich sind diese beiden Eigenschaften gleichbedeutend mit der Linearität des Systems, d. h.: Gelten Verstärkungs- und Überlagerungsprinzip, so ist das System linear.

LTI-Systeme

Ist ein System linear und zeitinvariant, d. h. es kann durch eine lineare gewöhnliche Differentialgleichung mit konstanten Koeffizienten beschrieben werden

$$y^{(n)} + a_{n-1}y^{(n-1)} + \dots + a_0y = b_0u + \dots + \underbrace{b_m u^m}_{\neq 0}, \quad (2.10)$$

so nennt man das System auch LTI-System (aus dem Englischen von Linear Time Invariant).

Bei dem Einmassenschwinger in Gl.(2.2)

$$M\ddot{y} + B\dot{y} + Cy = f \quad (2.11)$$

handelt es sich um ein LTI-System, da nach einem Teilen durch $M \neq 0$ die Form Gl.(2.10) mit $f = u$ vorliegt.

Für die Behandlung von LTI-Systemen gibt es eine umfangreiche mathematische Theorie [43]. Allerdings werden praktisch auftretende Zusammenhänge zwischen Eingangsgrößen und Ausgangsgrößen oft nichtlinear sein. Das in Abschnitt 3.3 einzuführende Werkzeug der Linearisierung wird es aber ermöglichen, die nichtlinearen Zusammenhänge in einem gewissen Arbeitsbereich ohne unzulässige Fehler durch lineare Ausdrücke anzunähern, wodurch die Betrachtung linearer Systeme ihre Berechtigung erhält. Somit stellen LTI-Systeme die grundlegendste Systemklasse der Regelungstechnik dar.

Relativer Grad

Der relative Grad eines LTI-Systems ist $r = n - m$ mit n und m nach Gl.(2.10) und beschreibt damit die Differenz zwischen höchster auftretender Ableitung der Ausgangsgröße und höchster auftauchender Ableitung der Eingangsgröße.

Der relative Grad ist eine regelungstechnisch wichtige Kenngröße, da er einige Informationen über das Systemverhalten in sich trägt, was in Abschnitt 4.6 ausgeführt wird. Insbesondere lässt sich über den relativen Grad erkennen, ob ein System kausal ist. Für nichtlineare Systeme ist die Definition des relativen Grades etwas komplexer, siehe Abschnitt 17.1.

Kausale Systeme

Ist $r \geq 0$, d. h. die höchste auftretende Ableitung der Ausgangsgröße ist mindestens so groß wie die höchste auftretende Ableitung nach der Eingangsgröße, so heißt die Differentialgleichung und das durch die Differentialgleichung beschriebene LTI-System *kausal*. Andernfalls heißt es *akausal*.

Kausalität hängt mit der Beobachtung in Abschnitt 1.2 zusammen, dass die Elemente des Wirkungsplan gerichtete Operationen gemäß ihres Ursache-Wirkungs-Zusammenhangs darstellen. Beschreibt die Eingangsgröße die Ursache und die Ausgangsgröße die Wirkung, so ist klar, dass die Rollen von y und u in Gl.(2.1) nicht ohne Weiteres vertauscht werden dürfen, da dies einer physikalischen Umkehrung des Ursache-Wirkungs-Zusammenhangs entspräche. Dies wird durch die gegebene Definition abgedeckt: Denn gilt für ein kausales System Σ die Bedingung $n > m$, so wäre das System $\tilde{\Sigma}$, welches durch ein Vertauschen der Ein- und Ausgänge gemäß $\tilde{u} = y$ und $\tilde{y} = u$ entstünde, wegen $\tilde{n} = m < n = m$ akausal. Ein Vertauschen der Ein- und Ausgänge wäre nur im Fall $r = 0$ möglich – ein Fall der unter anderem bei rein algebraischen Zusammenhängen $y = f(u)$ ohne zeitliche Vorzugsrichtung auftritt.

Als Beispiel für Kausalität wird der Einmassenschwinger in Gl.(2.2)

$$M\ddot{y} + B\dot{y} + Cy = f \quad (2.12)$$

betrachtet. Dieser besitzt mit der Zuordnung von y als Ausgang und f als Eingang den relativen Grad $r = 2$ und ist kausal: Die Kraft f bewirkt eine Auslenkung y . Eine umgekehrte Zuordnung mit der Ausgangsgröße f und der Eingangsgröße y , sodass eine Auslenkung y eine Kraft f bewirkt, ist zwar mathematisch möglich, entspricht aber nicht der Ursache-Wirkungs-Richtung und führt auf ein akausales System.

Reale, physikalisch-technische Systeme sind kausal, weswegen eine Beschränkung auf diese Systemklasse sinnvoll ist. Dennoch kann es in einigen Fällen zielführend sein auch akausale Systeme zu betrachten, wofür noch Beispiele angeführt werden. Da Differentialgleichungen Kernbestandteil der Regelungstechnischen Systembeschreibung sind, lohnt sich der Blick auf bestimmte gebräuchliche Darstellungsformen abseits der Darstellung in Gl.(2.1) oder Gl.(2.10).

2.3 Darstellung von Differentialgleichungen im Zustandsraum

Jede kausale Differentialgleichung der Ordnung n kann in ein System von n Differentialgleichungen erster Ordnung überführt werden. Hierzu werden zusätzlich zu den Eingangs- und Ausgangsgrößen weitere Variablen benötigt, die üblicherweise als Zustandsvariablen mit dem Buchstaben x bezeichnet werden. Diese Form der Darstellung von Differentialgleichungen wird entsprechend Zustandsraumdarstellung genannt.

Zur Motivation der Zustandsraumdarstellung wird zunächst eine Differentialgleichung betrachtet, in welche nur die Eingangsgröße u , nicht aber deren Ableitung eingeht:

$$y^{(n)}(t) = f\left(y^{(n-1)}(t), \dots, \ddot{y}(t), \dot{y}(t), y(t), u(t), t\right) . \quad (2.13)$$

Die einfache Wahl der Zustandsgrößen

$$x_1 = y , \quad x_2 = \dot{y} , \quad x_3 = \ddot{y} , \quad \dots \quad x_n = y^{(n-1)} \quad (2.14)$$

ermöglicht eine Umformung von Gl.(2.13) zu

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ &\vdots \\ \dot{x}_{n-1} &= x_n \\ \dot{x}_n &= f(x_1, x_2, \dots, x_{n-1}, u, t) \\ y &= x_1 . \end{aligned}$$

(2.15)

Bei Gl.(2.15) handelt es sich um ein System von n Differentialgleichungen erster Ordnung. Diese lassen sich auch in der vektoriellen Differentialglei-

chung

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} x_2 \\ x_3 \\ \vdots \\ x_n \\ f(\mathbf{x}, u, t) \end{bmatrix} \quad y = \underbrace{\begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}}_{g(\mathbf{x}, u, t)} \cdot \mathbf{x} \quad (2.16)$$

zusammenfassen. Hierbei kennzeichnen Fettbuchstaben, dass es sich bei den Signalen und Funktionen um vektorwertige Größen handelt.

Man erkennt, dass das System von Differentialgleichungen in der Zustandsraumdarstellung so aufgebaut wird, dass die n Ableitungen \dot{x}_i der Zustandsgrößen x_i als Funktionen f dieser Zustandsgrößen und der Eingangsgröße u ausgedrückt werden, während die Ausgangsgröße y als Funktion g von \mathbf{x} und u dargestellt wird. Wichtig ist, dass bei der Zustandsraumdarstellung im Gegensatz zur allgemeinen Differentialgleichung keine Ableitungen in den Eingangsgrößen u auftauchen. Stattdessen dürfen Ableitungen nach der Zeit nur in Form des Vektors \mathbf{x} auftreten. Dies ist i. Allg. auch physikalisch sinnvoll, weil Ableitungen der Eingangsgröße in Differentialgleichungen fast nie durch physikalische Gegebenheiten sondern fast immer durch mathematische Umformungen entstehen. Die resultierende Darstellung heißt Zustandsraumdarstellung und lässt sich durch Einführung von Vektoren \mathbf{u} und \mathbf{y} auf Systeme mit mehreren Ein- und Ausgangsgrößen erweitern.

Zustandsraumdarstellung

Die Darstellung

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \\ \mathbf{y} &= \mathbf{g}(\mathbf{x}, \mathbf{u}, t) \end{aligned} \quad (2.17)$$

mit $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{u} \in \mathbb{R}^p$, $\mathbf{y} \in \mathbb{R}^q$ und entsprechenden vektorwertigen Funktionen $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R} \rightarrow \mathbb{R}^n$ und $\mathbf{g} : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R} \rightarrow \mathbb{R}^q$ heißt Zustandsraumdarstellung.

Im Falle von LTI-Systemen vereinfacht sich die Gl.(2.17) zu

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \\ \mathbf{y} &= \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}\end{aligned}\tag{2.18}$$

mit $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ als Matrizen passender Dimension. Die Matrizen heißen:

- \mathbf{A} : Systemmatrix
- \mathbf{B} : Eingangsmatrix
- \mathbf{C} : Ausgangsmatrix
- \mathbf{D} : Durchgangsmatrix

Die zugehörige Anfangsbedingung ist $\mathbf{x}(t = 0) = {}_0\mathbf{x}$. Bei Systemen mit einer einzigen Eingangsgröße ($p = 1$) wird die Eingangsmatrix \mathbf{B} zu einem Vektor \mathbf{b} ; entsprechendes gilt für die Ausgangsmatrix $\mathbf{C} \rightarrow \mathbf{c}^T$, wenn nur eine einzige Ausgangsgröße ($q = 1$) interessiert. Gelten beide Bedingungen, so wird die Durchgangsmatrix \mathbf{D} zu einem Skalar d .

Für die Zustandsraumdarstellung des Einmassenschwingers Gl.(2.2) kann man beispielhaft die Zustände $x_1 = y$ und $x_2 = \dot{y}$ einführen. Mit $u = f$ erhält man aus

$$M\ddot{y} + B\dot{y} + Cy = u \Rightarrow \dot{x}_2 = -\frac{B}{M}x_2 - \frac{C}{M}x_1 + \frac{1}{M}u\tag{2.19}$$

die lineare Zustandsraumdarstellung

$$\begin{aligned}\Rightarrow \dot{\mathbf{x}} &= \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 \\ -\frac{C}{M} & -\frac{B}{M} \end{bmatrix}}_{\mathbf{A}} \mathbf{x} + \underbrace{\begin{bmatrix} 0 \\ \frac{1}{M} \end{bmatrix}}_{\mathbf{b}} u \\ \mathbf{y} &= \underbrace{\begin{bmatrix} 1 & 0 \end{bmatrix}}_{\mathbf{c}^T} \mathbf{x} + \underbrace{0}_{d} u.\end{aligned}\tag{2.20}$$

Abhängig von der Anzahl der Ein- und Ausgangsgrößen haben sich die beiden folgenden Begrifflichkeiten etabliert.

SISO und MIMO

Ein System mit einer skalaren Eingangsgröße und einer skalaren Ausgangsgröße wird auch als SISO-System (Single Input Single Output) bezeichnet.

net. Systeme mit mehreren Ein- oder Ausgangsgrößen heißen MIMO-Systeme (Multiple Input Multiple Output).

Die Forderung, dass Ableitungen nicht in den Eingangsgrößen auftauchen, stellt insbesondere für den wichtigen Fall der LTI-Systeme keine Einschränkung dar, weil durch eine geschickte Wahl der Zustandsgrößen eine Ableitung der Eingangsgrößen vermieden werden kann. Hierzu stellt man zunächst fest, dass – Differenzierbarkeit entsprechend vorausgesetzt – die Ableitung \dot{u} als Grenzwert des Differenzenquotienten aufgefasst werden kann:

$$\dot{u} = \lim_{h \rightarrow 0} \frac{u(t+h) - u(t)}{h} = \lim_{h \rightarrow 0} \frac{u(t+h) - u(t)}{h} - \frac{u(t) - u(t)}{h}. \quad (2.21)$$

Aufgrund der Zeitinvarianz und der Linearität gilt dann

$$u \mapsto y \Rightarrow \frac{u(t+h)}{h} \mapsto \frac{y(t+h)}{h} \quad (2.22)$$

und damit im Grenzübergang auch

$$u \mapsto y \Rightarrow \dot{u} \mapsto \dot{y} \quad (2.23)$$

unter Voraussetzung der Differenzierbarkeit der Signale. Damit folgt dann aus $u \mapsto y$ beispielsweise auch $b_1 \dot{u} \mapsto b_1 \dot{y}$. Setzt man diesen Zusammenhang für alle Ableitungen und alle b_i an, so gewinnt man aus Gl.(2.23) die Ausgangsgröße zu

$$b_0 u + b_1 \dot{u} + \dots + b_n u^{(n)} \mapsto b_0 y + b_1 \dot{y} + \dots + b_n y^{(n)}. \quad (2.24)$$

Daher muss in Gl.(2.18) nur die Bildung von y modifiziert werden. Für $b_n = 0$ ergibt sich direkt die Zustandsraumdarstellung

$$\begin{aligned} \dot{\mathbf{x}} &= \underbrace{\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix}}_{\mathbf{A}} \mathbf{x} + \underbrace{\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}}_b u \\ y &= \underbrace{\begin{bmatrix} b_0 & b_1 & b_2 & \cdots & b_{n-1} \end{bmatrix}}_{c^T} \mathbf{x} \end{aligned} \quad (2.25)$$

Im Fall $b_n \neq 0$ muss für $y^{(n)} = \dot{x}_n$ die Zeitableitung durch Einsetzen der Differentialgleichung aufgelöst werden. Es ergibt sich abweichend

$$\mathbf{c}^T = [b_0 - b_1 a_0 \quad b_1 - b_2 a_1 \quad \cdots \quad b_{n-1} - b_n a_{n-1}] , \quad d = b_n. \quad (2.26)$$

Für MIMO-Systeme ist ein analoges Vorgehen möglich.

Regelungsnormalform

Die Darstellungsform von Differentialgleichungen im Zustandsraum gemäß Gl.(2.25) bzw. Gl.(2.26) heißt *Regelungsnormalform*.

Es ist zu betonen, dass die vorgestellte Wahl der Zustandsgrößen exemplarisch war. Die Wahl der Zustände, die auf die Regelungsnormalform führt, hat einige regelungstechnischen Vorteile, die später deutlich werden. Eine abweichende Wahl der Zustände ist aber ebenso möglich. Oft wird die Wahl der Zustandsgrößen durch die physikalische Wirklichkeit nahe gelegt. Wenn dagegen nur der Zusammenhang zwischen Eingangs- und Ausgangsgrößen vorgegeben ist und dieser Zusammenhang durch Zustandsgrößen ausgedrückt werden soll, gibt es viele zweckmäßige und auch weniger zweckmäßige Möglichkeiten der Definition dieser Zustandsgrößen. In jedem Fall sollten für eine Differentialgleichung n -ter Ordnung genau n Zustände gewählt werden. Nutzt man weniger Zustände, kann die Dynamik der Differentialgleichung nicht voll erfasst werden. Nutzt man mehr Zustände, so sind die zusätzlichen Zustände redundant, was einige unvorteilhafte Eigenschaften mit sich bringt. Hierauf wird in Abschnitt 2.7, auf die Umrechnung verschieden gewählter Zustandsgrößen ineinander in Abschnitt 3.6 eingegangen.

Die Beschreibung dynamischer Systeme im Zustandsraum ist vorteilhaft beim Lösen zahlreicher Aufgaben in der regelungstechnischen Theorie aber auch in der Anwendungspraxis. Diese Beschreibungsform erlaubt es ohne weiteres, Systeme mit mehreren Eingangs- und Ausgangsgrößen zu beschreiben und zu behandeln und sie ist die Grundlage vieler Verfahren zur digitalen Simulation dynamischer Systeme. Weil es leistungsfähige Software zum Bearbeiten von Matrizen gibt, ist die auf Matrizen aufbauende Beschreibung linearer Differentialgleichungen eine gute Basis für weitere rechnerunterstützte Verfahren zur Analyse dynamischer Systeme und zur Synthese von Regelungen.

2.4 Darstellung von Differentialgleichungen im Wirkungsplan

Eine andere zentrale Darstellungsform, der bei der Modellierung dynamischer Systeme eine besondere Bedeutung zukommt, ist der aus Abschnitt 1.2 bekannte Wirkungsplan. Hierzu spezifiziert man die Übertragungsböcke, indem eine Zeichnung mit der qualitativen Darstellung des Zusammenhangs zwischen Eingangs- und Ausgangsgröße eingetragen wird. Das ist

- bei nichtlinearen Elementen eine Kennlinie,
- bei linearen Elementen der zeitliche Verlauf der Ausgangsgröße nach einem Sprung der Eingangsgröße mit dem Wert eins.

Beispiele für beide Fälle sind in Tab. 2-1 gezeigt. Dabei wird stets von einem positiven Übertragungsverhalten ausgegangen ($K > 0$). Ein negatives Übertragungsverhalten wird außerhalb des Blocks durch Negation des Signals am Summenpunkt kenntlich gemacht. Der Fall linearer bzw. nichtlinearer Systeme kann graphisch leicht über die Positionierung des Koordinatensystems unterschieden werden.

Die gegebene Differentialgleichung $T\ddot{y} + y = Ku$ steht hier repräsentativ für eine beliebiges LTI-System. Für den Moment sind insbesondere die beiden untersten Zeitverläufe für lineare Systeme von Interesse. Das Proportional-Element – auch P-Element oder P-Glied genannt – folgt der „Differential“gleichung $y = Ku$, in welcher keine Ableitungen auftauchen. Daher ergibt sich der Systemausgang als das K -fache des Systemeingangs und damit bei einem Sprung der Eingangsgröße ebenfalls sprungförmig. Für den Integrator – auch I-Element oder I-Glied genannt – gilt hingegen $\dot{y} = Ku$, d. h. die Steigung der Ausgangsgröße ist proportional zur Eingangsgröße, wodurch sich der gezeigte Verlauf mit konstanter Steigung ergibt. Beide Elemente werden detailliert in Abschnitt 7.2 diskutiert.

Als Beispiel diene erneut der Einmassenschwinger in der Form aus Gl.(2.3)

$$\ddot{y} = \frac{1}{M} \underbrace{(-B\dot{y} - Cy + f)}_{\Sigma f} . \quad (2.27)$$

Dieser lässt sich sofort in den Wirkungsplan in Bild 2-3 übertragen, indem die niedrigeren Ableitungen bis hin zur Ausgangsgröße selbst durch

Bezeichnung	Symbol	Funktion
Nichtlineares System		$y = u^2$
Lineares System		$T\dot{y} + y = Ku$
P-Element		$y = Ku$
I-Element Integrator		$\dot{y} = Ku$

Tabelle 2-1: Lineare und Nichtlineare Systeme im Wirkungsplan

Integration gewonnen werden. Hierdurch stehen sie zum Abgriff und zur Rückführung in die Summenpunkte zur Verfügung.

Bei der Darstellung des Einmassenschwingers kommt man mit der Verwendung von P- und I-Elementen aus. Dies gilt insoweit allgemein, als dass alle kausalen LTI-Systeme sich im Wirkungsplan durch eine geeignete Verschaltung dieser beiden Element-Typen darstellen lassen. Eine Möglichkeit, eine solche Verschaltung zu ermitteln, besteht in der zuvor diskutierten Zustandsraumdarstellung als Regelungsnormalform in Gl.(2.25). Der differentielle Zusammenhang $\dot{x}_i = x_{i+1}$ für $i = 1, \dots, n-1$ wird dabei zu einer Integratorkette, die Beziehung für $x^{(n)}$ zu einer Summation von P-Elementen. Dies führt auf den in Bild 2-4 gezeigten Wirkungsplan.

Im Gegensatz zur kompakten Darstellung im Zustandsraum, in welchem die Systemdynamik in zwei Gleichungen $\dot{\mathbf{x}} = \dots$ und $\mathbf{y} = \dots$ kompakt gefasst wird, werden beim Wirkungsplan alle Einzelwirkungen der Zustandsgrößen separat dargestellt. Da normalerweise die Zustandsgrößen physikalisch motiviert sind, unterstützt der Wirkungsplan somit die Modellbildung kom-

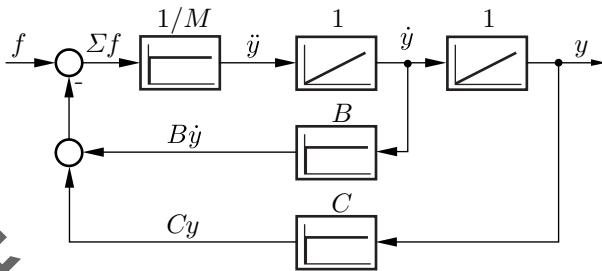


Bild 2-3: Wirkungsplan zum Einmassenschwinger Gl.(2.2)/Gl.(2.3)

plexer Systeme, indem eine Zerlegung in Teilsysteme vorgenommen und das daraus resultierende Wirkungsgefüge transparent gemacht wird. Insoweit verkörpert die Darstellungsform Wirkungsplan die grundsätzliche Be- trachtungsweise und auch das wesentliche Ziel des Faches Regelungstechnik, nämlich Hilfsmittel bereitzustellen, um dynamische technische Systeme mit komplexer Struktur analysieren, zielgerichtet beeinflussen und auch an deren Gestaltung mitwirken zu können.

2.5 Aufstellen von Differentialgleichungen

Die Regelungstechnische Modellierung technischer Systeme erfolgt immer in Hinblick auf eine zu erfüllende Regelungsaufgabe gemäß Bild 1-6. Der Versuch, direkt für das gesamte System eine Differentialgleichung aufzustellen, führt dabei bei komplexen Systemen oft nicht zum Ziel. Stattdessen bietet sich ein modulares Vorgehen an, welches durch den Wirkungsplan unterstützt werden kann. Hierfür kann – und zwar im Sinne einer damit empfohlenen top-down-Vorgehensweise – der folgende Leitfaden aufgestellt werden:

1. Klärung der Eingangs- und Ausgangsgrößen
2. Zerlegung in Teilsysteme
3. Übertragungsverhalten der Teilsysteme

Die einzelnen Schritte werden im Folgenden ausgeführt.

1. Klären der Ein -und Ausgangsgrößen

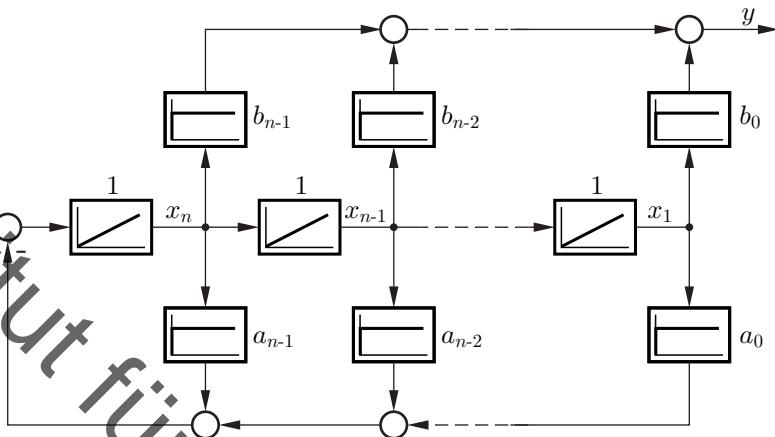


Bild 2-4: Wirkungsplan der Regelungsnormalform in Gl.(2.25)

Die Ein- und Ausgangsgrößen ergeben sich aus der Aufgabenstellung des zu modellierenden technischen Systems. Bereits die ersten Vorüberlegungen in Kap. 1 zeigten auf, dass z. B. eine (ungeregelte) Regelstrecke als Eingangsgrößen die Stell- und Störgröße und als Ausgangsgröße die Regelgröße besitzt. Folglich wird ein Regler als Eingangsgrößen die Regel- und Führungsgröße (bzw. die Regelabweichung) und als Ausgangsgröße die Stellgröße aufweisen. Schließlich kennt ein geregeltes System (der geschlossene Regelkreis) als Eingangsgrößen die Stör- und Führungsgröße und als Ausgangsgröße die Regelgröße (vgl. auch Bild 1-6).

2. Zerlegung in Teilsysteme

Bei der Zerlegung eines Gesamtsystems in Teilsysteme wird nach unmittelbaren Ursache-/Wirkungszusammenhängen gesucht, wobei empfohlen wird, ausgehend von der Ausgangsgröße des Gesamtsystems sukzessive rückwärts vorzugehen, bis schließlich nur noch die in Schritt 1 festgelegten Eingangsgrößen als solche auftreten. Das Vorgehen entgegen der Wirkungsrichtung der Größen bedeutet, ausgehend von einer Größe nach deren Ursachen zu fragen und diese Antworten festzuhalten. Dadurch kann man leichter sicherstellen, dass alle auf eine Größe wirkenden Einflüsse erfasst werden. Ein weiteres wichtiges Hilfsmittel zur Bildung von Teilsystemen besteht dar-

in, Speicher für Wärme, Energie oder Materie zu identifizieren und durch geeignete Grundgleichungen zu beschreiben. Verknüpfungen zwischen den Speichern beschreiben dann das Zusammenwirken der Teilsysteme.

Das Ziel von Schritt 2 besteht darin, ein erstes Wirkungsgefüge von Teilsystemen aufzustellen, aus denen ein Überblick über Struktur, Dynamik und Vorzeichen der Wirkzusammenhänge hervorgeht.

3. Übertragungsverhalten der Teilsysteme

Für viele der in Schritt 2 definierten Teilsysteme wird das abzubildende dynamische Übertragungsverhalten unmittelbar aus der technischen Ausführung des zu modellierenden technischen Systems zu entnehmen sein. Bei den übrigen, weniger trivialen Teilsystemen hilft oft eine Wiederholung von Schritt 2, um diese weiter zu unterteilen. Ansonsten werden die Differentialgleichungen zur Beschreibung dieser Teilsysteme aus den formalen Beschreibungen der betreffenden Fachdisziplinen hervorgehen. Hier sind insbesondere Energieerhaltungssätze, Gleichgewichtsbeziehungen von Kräften und Drehmomenten, Bewegungsgleichungen, Wärmeübergang und -speicherung, elektrische Netzwerke, Stromungsvorgänge und chemische Reaktionen zu nennen.

2.6 Beispiele für Modellbildung

2.6.1 Zerlegung in Teilsysteme

Vier Beispiele sollen das Vorgehen der Modellbildung illustrieren. Als erstes Beispiel soll ein Flüssigkeitsbehälter mit Zu- und Ablauf nach Bild 2-5 betrachtet werden, dessen Füllstand durch passende Wahl der Ventilstellung geregelt werden soll. Aufgrund der gegebenen Aufgabenstellung ist offenbar der Flüssigkeitsstand des Behälters die Ausgangsgröße, die daher mit y bezeichnet wird. Von außen auf den Flüssigkeitsstand wirkende Größen sind der Vordruck und die Ventilstellung. Der Umgebungsdruck p_U wird als konstant angenommen und muss daher nicht als dynamisches Signal modelliert werden. Laut Aufgabenstellung ist die Ventilstellung die Ausgangsgröße des Reglers und somit die Stellgröße u . Der Vordruck wird als Störgröße z aufgefasst, d. h. dass die Regelung das korrekte Einstellen eines gewünschten Füllstandes trotz möglicherweise schwankender Vordrücke sicherstellen soll.

Nach Klärung der Ein- und Ausgangsgrößen sollen Teilsysteme gebildet

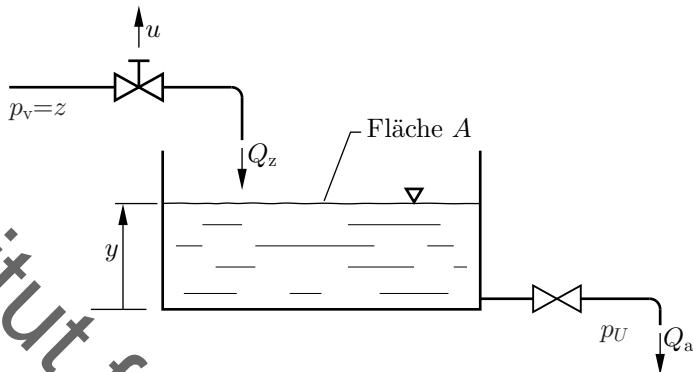


Bild 2-5: Flüssigkeitsbehälter mit Zu- und Ablauf

werden, indem ausgehend von der Ausgangsgröße nach Ursachen für die Änderung der jeweiligen Größe gesucht wird. Hierbei hilft die Identifikation des Flüssigkeitsbehälters als Speicher von Materie, dessen Inhalt sich in Abhängigkeit der Materieströme ändert. Für den Zusammenhang zwischen dem Flüssigkeitsstand und den Flüssigkeitsströmen gilt also, dass die Änderungsgeschwindigkeit des Flüssigkeitsstandes y proportional zu dem Flüssigkeitsstrom ist, der dem Behälter insgesamt zugeführt wird. D. h. das dynamische Verhalten des Teilsystems *Flüssigkeitsbehälter* mit den Massenströmen Q_z und Q_a als Eingangsgrößen und dem Flüssigkeitsstand y als Ausgangsgröße wird durch

$$\dot{y} = \frac{1}{A \cdot \rho} (Q_z - Q_a) \quad (2.28)$$

wiedergegeben, mit A als Oberfläche des Flüssigkeitsspiegels und ρ als Dichte der Flüssigkeit.

Da die Eingangsgrößen des Teilsystems Flüssigkeitsbehälter nicht den Eingangs- und Störgrößen u und z des Gesamtsystems entsprechen, muss Schritt 2 des modularen Vorgehens erneut durchgeführt werden. Die Flüssigkeitströme werden sich dabei durch entsprechende Durchflussgleichungen der Drosseln ergeben. Typische Modelle hierfür können der entsprechenden Fachliteratur entnommen werden, wobei einfache Modelle die Struktur

$$Q = K \cdot \alpha \cdot \text{sign}(\Delta p) \sqrt{|\Delta p|} \quad (2.29)$$

mit der Ventilstellung α , einer Konstanten K und der Druckdifferenz über der Drossel Δp aufweisen [50]. Die Vorzeichenfunktion $\text{sign}(\Delta p)$ sorgt dafür, dass für negative Druckdifferenzen die Richtung und damit das Vorzeichen des Flüssigkeitsstroms Q wechselt. Zur Bestimmung der genauen Gleichungen müssen nun gewissen Modellannahmen getroffen werden. Diese schränken im Allgemeinen den Gültigkeitsbereich von Modellen ein. Dennoch sind Modellannahmen für eine erfolgreiche Modellbildung essentiell, da häufig nur so ein hinreichendes Abstraktionsniveau erreicht werden kann. An dieser Stelle wird angenommen, dass sich der Füllstand im Behälter in bestimmten Grenzen bewegt: Der Zulauf liegt stets oberhalb des Füllstandes, während der Ablauf, welcher sich am Boden befinden soll, sich innerhalb des Füllstandes befindet. Der Druck an der Drossel des Ablaufes p_a entspricht dann dem hydrostatischen Druck zuzüglich des Umgebungsdrucks

$$p_a = \rho gy + p_U > p_U \quad (2.30)$$

mit dem konstanten Umgebungsdruck p_U , der als konstant angenommenen Dichte ρ und dem ebenfalls konstant angenommenen Ortsfaktor g . Der Vordruck z wird für eine sinnvolle technische Ausführung stets größer als der Umgebungsdruck sein. Somit ergibt sich für den Zulauf und Ablauf entsprechend

$$Q_z = K_z \cdot u \cdot \sqrt{z - p_U} \quad , \quad Q_a = K_a \cdot \alpha_a \cdot \sqrt{\rho gy} \quad . \quad (2.31)$$

Die Eingangsgrößen der Teilsysteme *Drossel-Zulauf* und *Drossel-Ablauf* sind z , u und y , da alle anderen auftretenden Größen konstant sind. Der Füllstand y ist durch die Beschreibung des Teilsystems *Flüssigkeitsbehälter* bereits bekannt; bei den anderen Größen handelt es sich um die in Schritt 1 definierten Eingangsgrößen. Einsetzen der Gleichungen für Q_z und Q_a führt nach Umformen auf

$$\dot{y} = \frac{1}{A \cdot \rho} (K_z \cdot u \cdot \sqrt{z - p_U} - K_a \cdot \alpha_a \cdot \sqrt{\rho gy}) \quad . \quad (2.32)$$

Somit liegt eine nichtlineare, zeitinvariante Differentialgleichung erster Ordnung vor und die Modellbildung ist abgeschlossen. Im resultierenden Wirkungsplan in Bild 2-6, der diese Differentialgleichung repräsentiert, ist die Zerlegung in die einzelnen Teilsysteme gut zu erkennen, wodurch dieser über

die Differentialgleichung hinaus zum Verständnis des modellierten Systems beiträgt. Zudem erkennt man, dass die Wirkung des Füllstandes auf den Abfluss strukturell einer Rückführung wie in Bild 1-6 entspricht. Hierbei handelt es sich um die Rückwirkung einer abhängigen Größe auf ihre Ursache. Dies stellt keine Verletzung der Voraussetzung der Rückwirkungsfreiheit dar, da diese Rückwirkung durch entsprechende Signalpfade explizit im Wirkungsplan dargestellt ist. Auch handelt es sich bei dieser Rückwirkung zwar um eine Rückführung, nicht aber um einen Regler. Stattdessen repräsentiert diese Rückführung den natürlichen Ausgleichsprozess, dass der Ablauf aus einem Tank mit der Füllhöhe des Tanks zunimmt. Des Weiteren ist zu betonen, dass die Anzahl der Speicher im System offenbar der Anzahl an Integratoren im Wirkungsplan und der Ordnung der Differentialgleichung entspricht.

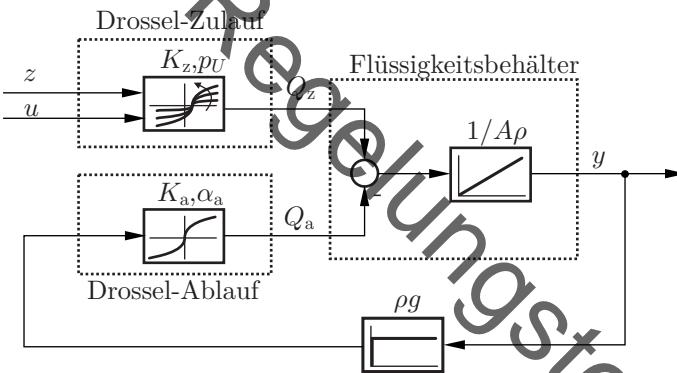


Bild 2-6: Wirkungsplan des Flüssigkeitsbehälters

Die hergeleitete Differentialgleichung lässt sich auf einfacherem Wege auch in eine Zustandsraumdarstellung überführen. Da das zu beschreibende System nur einen Speicher enthält und deshalb durch eine Differentialgleichung erster Ordnung vollständig beschrieben wird, wird nur eine einzige Zustandsgröße für die Zustandsraumdarstellung benötigt. Als physikalisch motivierte Wahl dieses Zustands bietet sich dabei der Füllstand y des einzigen Spei-

chers an. Somit ergibt sich direkt

$$\dot{x} = \underbrace{\frac{1}{A \cdot \rho} (K_z \cdot u_1 \sqrt{u_2 - p_U} - K_a \cdot \alpha_a \cdot \sqrt{\rho g x})}_{f(x, \mathbf{u})}, \quad y = \underbrace{x}_{g(x, \mathbf{u})}. \quad (2.33)$$

Die Eingangsgrößen u und z wurden hier zu einem zweireihigen Vektor $\mathbf{u} = [u \ z]^T$ zusammengefasst.

2.6.2 Rückwirkungen

Beim Aufstellen der Gleichungen für das Übertragungsverhalten eines Systems muss man darauf achten, dass man alle Zusammenhänge erfasst, besonders auch die Rückwirkungen einer abhängigen Größe auf ihre Ursache. Hierbei können sich selbst bei überschaubaren Systemen schnell komplexe Ausdrücke ergeben. Das folgende Beispiel (Bild 2-7), in dem der Druck in miteinander verbundenen gasgefüllten Behältern interessiert, soll dies veranschaulichen.

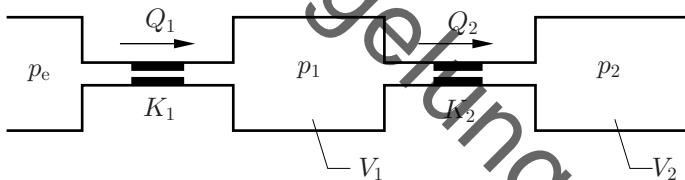


Bild 2-7: Reihenschaltung gasgefüllter Behälter

Gesucht ist eine Darstellung des Drucks p_2 abhängig vom Druck p_e in Form einer Differentialgleichung. Man erkennt, dass p_2 nur von p_1 abhängig ist; demgegenüber ist p_1 sowohl von p_e als auch von p_2 abhängig, da der Behälter 1 sowohl mit der Zuleitung als auch mit dem Behälter 2 verbunden ist. Der Zusammenhang zwischen p_1 und p_2 ist nicht rückwirkungsfrei, weil Änderungen von p_2 – hervorgerufen durch Änderungen von p_1 – auf p_1 zurückwirken.

Als Modellierungsannahme wird vorausgesetzt, dass die Behälter mit einem idealen Gas bei niedrigem Druck gefüllt sind und dass alle Zustandsänderungen bei konstanter Temperatur (isotherm) verlaufen. Dazu ist zu be-

merken, dass diese Voraussetzungen für die Beschreibung des dynamischen Verhaltens pneumatischer Geräte im Allgemeinen mit genügender Genauigkeit zutreffen.

Die Anordnung von Behältern in Bild 2-7 besteht aus zwei Drossel- und zwei Speichergliedern, die in Reihe geschaltet sind. Hiermit ist die Zerlegung in Teilsysteme bereits vorgenommen. Entsprechend des vorherigen Beispiels ergeben sich unter der Annahme $p_e \geq p_1 \geq p_2$ die Durchflussgleichungen

$$Q_2 = K_2 \cdot \sqrt{p_1 - p_2} \quad , \quad Q_1 = K_1 \cdot \sqrt{p_e - p_1} \quad . \quad (2.34)$$

Für die Speicher muss nun untersucht werden, wie sich die zu- und abströmenden Massen auf die in den Speichern herrschenden Drücke auswirken. Es gilt die ideale Gasgleichung

$$p \cdot V = m \cdot R \cdot T \quad , \quad (2.35)$$

darin ist p der Druck, V das Volumen, m die Masse und T die Temperatur des Gases sowie R ist die (spezielle) Gaskonstante. Ideale Gase ändern bei Drosselung ohne Wärmeaustausch ihre Temperatur nicht und die Volumina sind ebenfalls konstant. Durch Zusammenfassen der Konstanten erhält man

$$p = \underbrace{\frac{R \cdot T}{V}}_{K^*} m \quad , \quad (2.36)$$

d. h. der Druck ist zu der im Behälter eingeschlossenen Gasmasse proportional. Die in den Behältern eingeschlossenen Gasmassen werden durch die Zu- und Abflüsse verändert. Analog zum vorherigen Beispiel muss

$$\dot{m}_2 = \rho Q_2 \quad , \quad \dot{m}_1 = \rho(Q_1 - Q_2) \quad (2.37)$$

gelten, d. h. die zeitliche Ableitung des Inhaltes ist gleich der Differenz aus zufließendem und abfließendem Massenstrom. Durch Zusammenfassen von Gl.(2.36) und Gl.(2.37) und Einsetzen der Beziehung aus den Durchflussgleichungen Gl.(2.34) erhält man

$$\dot{p}_2 = K_2^* q_2 = K_2^* K_2 \cdot \sqrt{p_1 - p_2} \quad (2.38)$$

$$\dot{p}_1 = K_1^*(q_1 - q_2) = K_1^* (K_1 \cdot \sqrt{p_e - p_1} - K_2 \cdot \sqrt{p_1 - p_2}) \quad , \quad (2.39)$$

d. h. die Änderungsgeschwindigkeit der Behälterdrücke ist der Differenz aus aufließendem und abfließendem Massenstrom proportional und hängt weiter von der Druckverteilung in den Behältern und der Zuleitung ab.

Die beiden Gleichungen beschreiben die Änderungen der beiden Behälterdrücke. Um eine einzelne Differentialgleichung für den Druck p_2 zu erhalten, müssen die beiden Gleichungen zusammengefasst und die Variable p_1 eliminiert werden. Diese scheinbar einfache Operation gestaltet sich hier überraschend schwierig. Der einzige gangbare Weg ist es, die Gleichung, die p_1 in einfacherer (d. h. nicht differentieller Weise) enthält, nach p_1 aufzulösen und in die andere Gleichung einzusetzen. Es ergibt sich aus Gl.(2.38)

$$p_1 = p_2 + \left(\frac{\dot{p}_2}{K_2^* K_2} \right)^2 \Rightarrow \dot{p}_1 = \dot{p}_2 + \frac{2\dot{p}_2 \ddot{p}_2}{(K_2^* K_2)^2} \quad (2.40)$$

und eingesetzt in Gl.(2.39)

$$\frac{2}{(K_2^* K_2)^2} \dot{p}_2 \ddot{p}_2 + \left(1 + \frac{K_1^*}{K_2^*} \right) \dot{p}_2 = K_1^* K_1 \cdot \sqrt{p_e - p_2 - \left(\frac{\dot{p}_2}{K_2^* K_2} \right)^2} \quad (2.41)$$

Diese Gleichung ist trotz der Übersichtlichkeit des zu modellierenden Systems bereits von erheblicher Komplexität und die zugehörige Rechnung fehleranfällig. Eine Auflösung nach der höchsten Ableitung \ddot{p}_2 , um eine explizite Darstellung wie in Gl.(2.1) zu erhalten, benötigt Annahmen bezüglich der Ableitung \dot{p}_2 , welche nicht verschwinden darf. An dieser Stelle spielen Zustandsraumdarstellung und Wirkungsplan ihre Stärken aus, da sie Zwischengrößen wie p_1 zulassen und somit eine zugängliche Darstellung von Differentialgleichungen ermöglichen. So ergibt sich mit der Einführung des Zustandsvektors $\mathbf{x}^T = [p_1 \ p_2]$ sowie $y = p_2$ und $u = p_e$ die Zustandsraumdarstellung

$$\dot{\mathbf{x}} = \begin{bmatrix} K_1^* (K_1 \cdot \sqrt{u - x_1} - K_2 \cdot \sqrt{x_1 - x_2}) \\ K_2^* K_2 \cdot \sqrt{x_1 - x_2} \end{bmatrix} \quad , \quad y = x_2 \quad . \quad (2.42)$$

Eine Zusammenfassung mehrerer nichtlinearer Differentialgleichungen zu einer Gesamtgleichung ist im Allgemeinen wenig empfehlenswert. Dennoch

tritt es regelmäßig auf, dass Systeme modelliert werden müssen, die (wie die beiden Behälter des Beispiels) in Reihe geschaltet sind. Hier gibt es zwei Möglichkeiten der Abhilfe. So kann man auf ein Zusammenfassen der Gleichungen verzichten, indem man eine Darstellung im Zustandsraum oder Wirkungsplan wählt. Alternativ nähert man nichtlineare Gleichungen durch lineare Zusammenhänge an, wie dies in Abschnitt 3.3 diskutiert wird. Für lineare Differentialgleichung gestaltet sich das Zusammenfassen leichter als im nichtlinearen Fall, da dort Hilfsmittel zur Verfügung stehen, die diese Aufgabe erheblich erleichtern (siehe Kapitel 6).

2.6.3 Zusammenfassen von Teilsystemen im Wirkungsplan

Als vorletztes Beispiel sollen die Möglichkeiten der Zusammenfassung von Teilsystemen auf Ebene des Wirkungsplans anhand des Einmassenschwingers diskutiert werden. Dies ist auch als kleiner Vorausblick auf damit zusammenhängende Inhalte in Kapitel 7 gedacht und soll daher an dieser Stelle nicht erschöpfend behandelt werden.

Für den Einmassenschwinger konnte der Wirkungsplan in Bild 2-3 hergeleitet werden. Hierbei konnten neben der Ausgangsgrößen y auch deren Ableitungen direkt abgegriffen werden. Wenn die internen Größen des Wirkungsplans, hier die zeitlichen Ableitungen \dot{y} und \ddot{y} , nicht benötigt werden, um z. B. weitere Ausgangsgrößen zu bestimmen, kann die Darstellung im Wirkungsplan durch Charakterisierung des Übertragungsverhaltens von f auf y auch auf einen einzelnen Übertragungsblock kondensiert werden. Da es sich um ein lineares Element handelt, wird hierzu im Übertragungsblock der zeitliche Verlauf der Auslenkung y bei einer sprungförmigen Aufprägung einer Kraft f dargestellt. Im Falle eines im Vergleich zur Federsteifigkeit und Masse geringen Dämpfungsbeiwertes wird sich dabei ein Verlauf wie in Bild 2-8 ergeben, wobei Werkzeuge zur Bestimmung dieses Verlaufes Teil von Abschnitt 4.4 sind. Die Ermittlung und graphische Darstellung basiert auf der normierten Form einer Differentialgleichung zweiter Ordnung

$$\ddot{y} + 2D\omega_0\dot{y} + \omega_0^2y = K\omega_0^2f, \quad (2.43)$$

welche die statische Verstärkung K , den Dämpfungsgrad D und die Kennkreisfrequenz ω_0 enthält, die aus dem Koeffizientenvergleich mit Gl.(2.2) hervorgehen.

Die graphische Darstellung des Wirkungsplans wird hierdurch auf einen einzigen Übertragungsblock reduziert und die Information über das dynamische Übertragungsverhalten in die damit symbolisierte Differentialgleichung höherer Ordnung sowie in die zugehörigen Koeffizienten K , D und ω_0 verlagert.

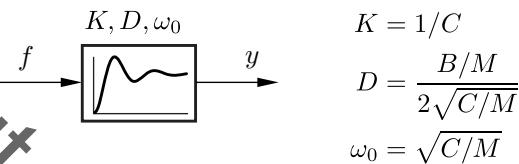


Bild 2-8: Wirkungsplan zur Differentialgleichung Gl.(2.43)

Bezüglich des damit dargestellten Übertragungsverhaltens von der Kraft f auf die Auslenkung y sind die beiden Wirkungspläne der Bilder 2-3 und 2-8 als äquivalent anzusehen – schließlich beschreiben sie die äquivalenten Differentialgleichungen Gl.(2.2) und Gl.(2.43). Ein Unterschied, auch im Nutzen für die Anwendung, ist jedoch darin zu sehen, dass ein ausführlicher Wirkungsplan wie in Bild 2-3 auch den Abgriff interner Größen ermöglicht, die ein zusammengefasster Wirkungsplan wie in Bild 2-8 im Gegensatz dazu nicht bereitstellt.

2.6.4 Modellierung von Regelungen

Die bisherigen drei Beispiele zeigen, wie mithilfe von Differentialgleichungen Regelstrecken beschrieben werden können. Um das anvisierte Vorgehen aus Bild 2-1 nutzen zu können, ist es jedoch notwendig, nicht nur Regelstrecken, sondern auch die zu entwerfenden Regler modellhaft abilden zu können. Dabei ist es zweckmäßig, Regelstrecken sowie Regler in einer einheitlichen Beschreibungssprache zu fassen. Dies schränkt die Menge der für Regelungsaufgaben verwendbaren Algorithmen ein, da nicht alle denkbaren Rechenvorschriften ein entsprechendes Differentialgleichungspendant in der Modellwelt besitzen müssen. Das folgende Beispiel soll aufzeigen, dass die Beschränkung auf solche Algorithmen, die sich über eine Differentialgleichung modellieren lassen, keine wesentliche Einschränkung der Menge der sinnvollen Algorithmen darstellt.

Betrachtet wird die Regelungsaufgabe, ein Fahrzeug auf einer Autobahn in einem konstanten Sollabstand zu einem Vorderfahrzeug fahren zu lassen. Der Sollabstand w ist die Führungsgröße, der Abstand y die Regelgröße und die Gaspedalstellung u die Stellgröße. Mögliche Störungen sind der Neigungswinkel der Straße oder Einflüsse wie Wind. Der Einfachheit halber werden nur ein Fahrstreifen und zwei Fahrzeuge betrachtet. Die beschriebene Regelungsaufgabe wird von zahllosen Menschen täglich erfolgreich gelöst. Das intuitive Vorgehen zur Bestimmung der Gaspedalstellung u besitzt dabei vorrangig die folgenden drei Anteile:

- 1) Wähle u abhängig vom aktuellen Abstand. Man wird dabei u verringern, wenn der Abstand zu klein ist und u vergrößern, wenn der Abstand zu groß ist.
- 2) Manchmal verändert sich der Abstand (z. B. durch ein Bremsmanöver des Vorderfahrzeugs), wodurch eine zusätzliche Anpassung von u notwendig ist. Wähle daher u auch abhängig von der aktuellen Abstandsänderung \dot{y} . Man wird dabei u verringern, wenn \dot{y} negativ ist und u vergrößern, wenn \dot{y} positiv ist.
- 3) Das beschriebene Vorgehen stellt bisher nicht sicher, dass sich der gewünschte Abstand langfristig einstellt. Passe u daher so lange an, bis $y = w$ gilt. Folglich gilt $\dot{u} = 0$ nur für $y = w$. Man wird dabei u verringern, wenn y langfristig zu klein ist und u vergrößern, wenn y langfristig zu groß ist.

Der erste Effekt lässt sich mathematisch durch $u_1 = f_1(y)$ fassen, der zweite durch $u_2 = f_2(\dot{y})$, wobei die Funktionen f_1 und f_2 dasselbe Vorzeichen wie ihr Funktionsargument besitzen. Der letzte Effekt führt auf die Beschreibung $\dot{u}_3 = f_3(y)$. Nimmt man an, dass sich die Gesamtentscheidung des Menschen aus einer linearen Überlagerung der drei Einzeleffekte ergibt, so ergibt sich nach Ableiten von u_1 und u_2 die mathematische Beschreibung

$$\dot{u} = \frac{\partial f_1}{\partial y} \dot{y} + \frac{\partial f_2}{\partial \dot{y}} \ddot{y} + f_3(y) = f(y, \dot{y}, \ddot{y}) \quad . \quad (2.44)$$

Diese Beschreibung des Regelverhalten eines Menschen in Gl.(2.44) führt offenbar auf eine gewöhnliche Differentialgleichung. Sie ist nichtlinear und zeitinvariant. Da u die Ausgangsgröße eines Reglers, y aber die Eingangsgröße eines Reglers ist, besitzt sie die Ordnung 1 und ist akausal. Die Akausalität kann dabei so gedeutet werden, dass das bisherige Modell des Regelverhal-

tens keine Reaktionszeit des Menschen beinhaltet und somit eine Idealisierung darstellt. Dies illustriert die zuvor getätigte Aussage, dass auch die Betrachtung akausaler Systeme in gewissen Fällen zielführend sein kann. Das Regelverhalten des Menschen entspricht dabei in seinen Grundzügen dem sogenannten PID-Regler, der in Abschnitt 7.2 genauer betrachtet wird.

Das gegebene Beispiel zeigt, dass eine Beschränkung auf Differentialgleichungen zur mathematischen Beschreibung keine wesentliche Einschränkung darstellt, da viele intuitive Regelgesetze sich als Differentialgleichung darstellen lassen. Dennoch gibt es natürlich mögliche Stellgesetze, für die eine solche Darstellung nicht möglich ist und die den Rahmen dieses Buches sprengen [27].

2.7 Das Gesetz der Sparsamkeit

Bei der Modellierung von dynamischen Systemen, die einer Eingangsgröße $u(t)$ eine Ausgangsgröße $y(t)$ zuordnen, ist es grundsätzlich möglich, verschiedene Differentialgleichungen aufzustellen, die das identische Übertragungsverhalten $u \mapsto y$ aufweisen. Hierzu gehört das unerhebliche Multiplizieren beider Gleichungsseiten der Differentialgleichung mit einer Konstanten, was strukturell an der Differentialgleichung als Modell des Systems nichts verändert. Es sind aber auch Modifikationen denkbar, die die Differentialgleichung erheblich verkomplizieren und die bei der Modellierung daher unterlassen werden sollten.

Als Beispiel wird die einfache Differentialgleichung $\dot{y} = u$ betrachtet, die die Ordnung eins besitzt. Differenziert man nun auf beiden Seiten, erhält man die neue Differentialgleichung $\ddot{y} = \dot{u}$, die Ordnung zwei besitzt. Betrachtet man das durch die beiden Differentialgleichungen modellierte Verhalten, so ist das Übertragungsverhalten $u \mapsto y$ genau identisch, sofern u differenzierbar ist.

Nimmt man die Ordnung einer Differentialgleichung als Indikator für die Komplexität des Modells, so ist das Modell $\ddot{y} = \dot{u}$ dem Modell $\dot{y} = u$ offenbar unterlegen. Auch die zusätzlichen Forderungen an Differenzierbarkeit des Eingangs sprechen nicht für das Modell der Ordnung zwei. Es ist daher erstrebenswert, solche Differentialgleichungen als Modelle zu verwenden, die eine möglichst geringe Ordnung aufweisen.

Minimale Realisierung

Eine Differentialgleichung, die mit der minimal möglichen Ordnung auskommt, um das Übertragungsverhalten $u \mapsto y$ zu beschreiben, wird *minimale Realisierung* genannt.

Diese Forderung an die Ordnung einer Differentialgleichung überträgt sich direkt auf die Darstellungsformen als Wirkungsplan und im Zustandsraum. Die minimale Realisierung einer Differentialgleichung der Ordnung n im Wirkungsplan kommt mit genau n Integratoren oder vergleichbaren Gliedern aus. Eine minimale Realisierung im Zustandsraum verwendet für eine solche Differentialgleichung genau n Zustände. Alle von dieser Forderung abweichenden Modelle genügen nicht dem Gesetz der Sparsamkeit (*lex parsimoniae*), das auch unter dem Schlagwort von „Ockhams¹ Rasiermesser“ Bekanntheit erlangt hat.

Gesetz der Sparsamkeit (*lex parsimoniae*)

Innerhalb einer Menge von wissenschaftlichen Erklärungen sollten solche Erklärungen bevorzugt werden, die mit weniger Variablen oder Elementen auskommen.

Die überflüssigerweise hinzugefügte Ordnung der Differentialgleichung, Integratoren im Wirkungsplan oder Zustände, sind wahlweise redundant oder leisten keinen Beitrag zum Übertragungsverhalten des Systems. In Bild 2-9 sind drei Wirkungspläne gezeigt, die allesamt die Differentialgleichung $y = u$ darstellen.

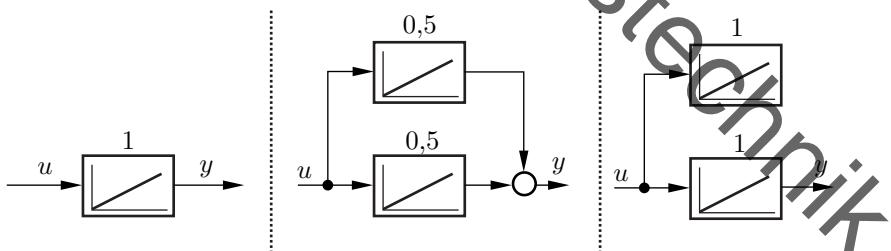


Bild 2-9: Drei Wirkungspläne zur Illustration einer minimalen Realisierung

¹Wilhelm von Ockham (1288?-1347), englischer Philosoph [60]

Während der Wirkungsplan links einen Integrator verwendet, enthält der in der Mitte einen redundanten Integrator und der Plan zur Rechten einen zusätzlichen Integrator, der keinen Einfluss auf das Verhalten von u nach y hat. Der linke Wirkungsplan ist der einzige, der dem Gesetz der Sparsamkeit genügt und eine minimale Realisierung darstellt.

Modelle, die keine minimale Realisierungen sind, besitzen nicht nur unnötige Zustände, sondern auch regelungstechnisch äußerst unvorteilhafte Eigenschaften, worauf an den passenden Stellen verwiesen wird. In allen weiteren Ausführungen wird daher stets davon ausgegangen, dass es sich bei der betrachteten Differentialgleichung um eine minimale Realisierung handelt.

2.8 Einheiten

Die in den vorangegangenen Abschnitten verwendeten Größen wie Masse oder Temperatur besitzen allesamt eine Einheit, die bisher geflissentlich unterschlagen wurde. Allen auftretenden Signalen muss eine physikalische Einheit zugeordnet werden. Dasselbe gilt auch für die meisten regelungstechnischen Kenngrößen wie die erwähnten Koeffizienten K , D und ω_0 . Da diese Koeffizienten Teil einer Differentialgleichung sind, welche Zeitableitungen enthält, sind die für die Koeffizienten verwendeten Einheiten oft ebenfalls Zeiteinheiten.

Zeitkonstanten und Frequenzen

In der einfachen Differentialgleichung $T\dot{y}(t) + y(t) = 0$ muss – wenn die Zeit in Sekunden sec gemessen wird und unabhängig von der Einheit von y – auch T die Einheit sec haben. Man bezeichnet T als *Zeitkonstante* des Systems.

Schreibt man stattdessen $\dot{y} + \omega y = 0$, so hat ω die Einheit sec^{-1} und ist daher eine Frequenz, die *Eckkreisfrequenz* genannt wird. In anderen Fällen ergeben sich Frequenzen, die als *Kennkreisfrequenz* (siehe auch Gl.(2.43)) oder *Eigenkreisfrequenz* bezeichnet werden.

In vielen Fällen spricht man nicht von den Koeffizienten einer Differentialgleichung, sondern von den Zeitkonstanten eines Systems. Die Kreisfrequenzen charakterisieren in einigen Fällen das Schwingungsverhalten oder frequenzabhängige Eigenschaften des Systems. In der Regelungstechnik ist

dabei die vom SI-Standard abweichende Notation sec für Sekunde üblich, da die Variable s durch die Laplace-Transformation anderweitig belegt ist. Man versucht, die Koeffizienten der Differentialgleichung so zusammenzufassen, dass sich Zeitkonstanten, Frequenzen oder dimensionslose Kennzahlen ergeben. Insbesondere bei der Verrechnung unterschiedlicher Signale wie y und u werden aber Koeffizienten mit Einheiten verbleiben, die diese beiden einheitenbehafteten Signale ins Verhältnis setzen.

Die genutzten Einheiten sind dabei so zu wählen, dass sie Differenzen zwischen Größen adäquat beschreiben – so sollte beispielsweise für Temperaturen die Einheit K anstelle von C° verwendet werden. Dies liegt daran, dass es die Aufgabe der Regelungstechnik ist, die Ausgangsgrößen eines Systems ihren Führungsgrößen auszugleichen. Daher wird in der Regelungstechnik vielfach nicht mit den Absolutwerten von Größen, sondern mit Abweichungsgrößen gearbeitet, die im einfachsten Fall als Differenz von Absolutwerten und Bezugswerten gebildet werden.

Notation von Absolut- und Abweichungsgrößen

Normalerweise werden alle Signale mit Kleinbuchstaben geschrieben. Treten in der Systembeschreibung sowohl Absolutgrößen als auch Abweichungsgrößen auf, so werden diese dadurch unterschieden, dass Absolutgrößen mit Großbuchstaben, Abweichungsgrößen hingegen mit Kleinbuchstaben geschrieben werden.

Auf Großbuchstaben für Signale wird (abseits von der normkonformen Darstellung physikalischer Größen bei der Modellbildung) nur dann zurückgegriffen, wenn zwischen Abweichungs- und Absolutgrößen unterschieden werden muss. Im Falle von (absoluten) Bezugswerten Y_0 und U_0 ergeben sich die Abweichungsgrößen entsprechend zu

$$y = Y - Y_0 \quad , \quad u = U - U_0 \quad (2.45)$$

In einigen Fällen kann es zweckmäßig sein, die Abweichungsgrößen zusätzlich auf konstante Bezugswerte, z. B. Maximalwerte, Minimalwerte o. A. zu normieren. Die Möglichkeit der Normierung wird dann genutzt, wenn das Mitführen von Dimensionen keine zusätzliche Klarheit oder Sicherheit vermittelt oder wenn relative Änderungen oder Prozentangaben mehr Einsicht über das Systemverhalten geben. Insbesondere im Falle von MIMO-Systemen ist eine Normierung anzuraten, damit die verschiedenen Ein- und

Ausgänge miteinander sinnvoll ins Verhältnis gesetzt werden (siehe auch Abschnitt 13.2.4). Auf die ausdrückliche Kennzeichnung normierter Größen wird dabei meist verzichtet, weil dies aus dem Zusammenhang hervorgeht.

Der Wechsel zu normierten Abweichungsgrößen fördert die Übertragbarkeit Regelungstechnischer Erkenntnisse zwischen einzelnen Fachdisziplinen, da von der konkreten Einheit und Anwendung abstrahiert wird und eine allgemeingültige Aussage abgeleitet werden kann. Eine Herausforderung ist hierbei, mit den einheitenlosen Darstellungen in Abweichungsgrößen umzugehen zu können ohne die physikalische Bedeutung der Koeffizienten aus dem Blick zu verlieren. So ist es für ein P-Element, das die Verstärkung $K = 1$ besitzt, von entscheidender Bedeutung, ob K die Einheit $\frac{V}{m}$ oder $\frac{V}{mm}$ besitzt, da das einen Unterschied in der tatsächlichen Verstärkung um den Faktor 1000 entspricht.

Ein prominentes Beispiel für ein Scheitern dieses Spagats zwischen Abstraktion und Anwendung ist dabei der Absturz der NASA-Sonde Mars Climate Orbiter (MCO) am 23. September 1999 [35]: Ziel der Mission war es, eine Sonde in den Orbit des Mars zu bringen um von dort aus das Marsklima zu untersuchen. Beim Einflug in den Orbit musste dabei ein durch den Solar-druck hervorgerufenes Drehmoment kompensiert werden. Hierbei erwartete die Regelung vom Messglied einen Wert im imperialen System lb_f , erhielt aber einen Wert in der SI-Einheit N. Hierdurch war die Reglerverstärkung um etwa den Faktor 4,45 höher als eigentlich vorgesehen, wodurch der Kurs überkorrigiert wurde. Die Sonde gelangte in die Marsatmosphäre und wurde durch die Reibungshitze zerstört.

3 Autonome Systeme

3.1 Arbeitspunkte und Ruhelagen

Bei der Bildung von Abweichungsgrößen müssen Bezugswerte definiert werden, zu denen die Absolutwerte ins Verhältnis gesetzt werden. Dabei liegt es nahe, die miteinander verträglichen Bezugswerte zu nehmen, welche sich bei einer erfolgreichen Umsetzung der gegebenen Regelungsaufgabe einstellen – die sogenannten *Arbeitspunkte*.

Arbeitspunkt

Ein Tupel $(\mathbf{u}_0(t), \mathbf{y}_0(t))$ heißt Arbeitspunkt eines Systems, wenn bei einer passenden Anfangsbedingung für die Lösung der Differentialgleichung des Systems gilt:

$$\mathbf{u}_0(t) \mapsto \mathbf{y}_0(t) . \quad (3.1)$$

Unter einem Arbeitspunkt soll also eine definierte Lösungstrajektorie des Systems verstanden werden. Bei einer Beschreibung im Zustandsraum muss entsprechend der Zustand $\mathbf{x}_0(t)$ dem Tupel hinzugefügt werden. Meistens besitzt ein Arbeitspunkt besonders günstige Eigenschaften und das Ziel der Regelung ist es, das System trotz abweichender Anfangsbedingungen und Störungen auf diesen Arbeitspunkt zu führen und dort zu halten. Für zeitinvariante System wird es dabei in vielen Fällen ausreichen, konstante Größen $(\mathbf{u}_0, \mathbf{y}_0)$ als Arbeitspunkt zu verwenden. Da Größen genau dann konstant bleiben, wenn keine zeitliche Änderung des Systems erfolgt, spricht man auch von *Ruhelagen*, *Gleichgewichtspunkten* oder auch *stationären* Arbeitspunkten des Systems.

Zur Bestimmung von Ruhelagen wird dabei zweckmäßig ausgehend von einem festen $\mathbf{u} = \mathbf{u}_0$ nach verträglichen Werten für \mathbf{x} gesucht, aus denen sich die zugehörigen Ausgänge \mathbf{y} ergeben. Setzt man zur Berechnung der Ruhelagen konstante Eingangsgrößen \mathbf{u}_0 in die zeitinvariante, nichtlineare Zustandsraumdarstellung ein, so erhält man mit

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u} = \mathbf{u}_0) \quad (3.2)$$

ein System, das keine Eingangsgröße mehr besitzt – ein sogenanntes autonomes System.

Autonome Systeme

Ein autonomes System folgt der Zustandsgleichung

$$\dot{\boldsymbol{x}} = \mathbf{f}(\boldsymbol{x}) \quad , \quad \boldsymbol{x}(t=0) = {}_0\boldsymbol{x} \quad . \quad (3.3)$$

Der Verlauf der Zustandsgrößen ist also nur vom Anfangszustand ${}_0\boldsymbol{x}$ abhängig.

Trotz der Einschränkung, keinen Eingang zu besitzen, ist die Klasse der autonomen Systeme vergleichsweise groß. So können sowohl Systeme mit konstanter Eingangsgröße (siehe Gl.(3.2)) als auch Systeme mit zeitinvarianten Stellgesetzen $\mathbf{h}: \mathbb{R}^n \rightarrow \mathbb{R}^p$

$$\dot{\boldsymbol{x}} = \mathbf{f}(\boldsymbol{x}, \boldsymbol{u} = \mathbf{h}(\boldsymbol{x})) = \tilde{\mathbf{f}}(\boldsymbol{x}) \quad (3.4)$$

als autonome Systeme aufgefasst werden, so dass die meisten praxisrelevanten Regelkreisstrukturen und insbesondere der Fall stationärer Arbeitspunkte erfasst werden.

Ruhelage

Der Punkt \boldsymbol{x}_0 heißt Ruhelage des autonomen Systems $\dot{\boldsymbol{x}} = \mathbf{f}(\boldsymbol{x}, \boldsymbol{u}_0)$, wenn die Zustandsgleichung für $\boldsymbol{x} = \boldsymbol{x}_0$ verschwindet

$$\dot{\boldsymbol{x}} = \mathbf{f}(\boldsymbol{x}_0, \boldsymbol{u}_0) = \mathbf{0} \quad . \quad (3.5)$$

Da die zeitliche Änderung $\dot{\boldsymbol{x}}$ des Zustands \boldsymbol{x} am Punkt \boldsymbol{x}_0 verschwindet, verweilt das System am gegebenen Punkt. Mathematisch führt die Bestimmung der Ruhelagen bei gegebenem (konstantem) Eingangssignal $\boldsymbol{u} = \boldsymbol{u}_0$ auf eine Nullstellenbestimmung der Abbildung \mathbf{f} . Im Falle eines linearen autonomen Systems

$$\dot{\boldsymbol{x}} = \mathbf{A}\boldsymbol{x} \stackrel{!}{=} \mathbf{0} \quad (3.6)$$

lassen sich die Ruhelagen besonders einfach gewinnen. Die Bestimmungs-gleichung besitzt in Abhängigkeit der Eigenwerte der Systemmatrix \mathbf{A} ent-weder eine Lösung oder ein Kontinuum an Lösungen. Während in jedem Fall eine Ruhelage $\boldsymbol{x}_0 = \mathbf{0}$ im Ursprung liegt (triviale Lösung), existiert für

den Fall eines Eigenwertes bei Null ein Kontinuum von miteinander verbundenen Ruhelagen, welches über den zugehörigen Eigenraum berechnet werden kann.

Die Nullstellen der Funktion f werden im Falle eines Systems mit konstanter Eingangsgröße Gl.(3.2) vom gewählten Wert für u_0 abhängig sein, wodurch jedes u_0 einen Arbeitspunkt mit abweichendem x_0 und y_0 definiert. Möchte man dasjenige u_0 auswählen, welches auf ein besonders wünschenswerten Wert wie $y_0 = w_0$ mit der konstanten Führungsgröße w_0 führt, so erfordert dies den Entwurf einer Steuerung, da ausgehend von einem gegebenen w ein passendes u berechnet wird. Als Beispiel wird dazu erneut der Flüssigkeitsbehälter in Bild 2-5 betrachtet, für den bereits die folgende Differentialgleichung aufgestellt wurde:

$$\dot{y} = \frac{1}{A \cdot \rho} (K_z \cdot u \sqrt{z - p_U} - K_a \cdot \alpha_a \cdot \sqrt{\rho g y}) \quad (3.7)$$

Eine erfolgreiche Regelung wird den Füllstand y der konstant angenommenen Führungsgröße $y_0 = w_0$ dauerhaft angeglichen haben. Da die Regelung y auf der gewünschten Führungsgröße hält, werden sich der Zu- und der Ablauf genau im Gleichgewicht befinden, sodass keine zeitliche Änderung von y auftritt, d. h. eine Ruhelage vorliegt. Um ein autonomes System zu erhalten, muss auch Störgröße $z = z_0 > p_U$ als konstant angenommen werden. Nun lässt sich die benötigte Stellgröße u_0 wie folgt berechnen:

$$\begin{aligned} \dot{y} &= \frac{1}{A \cdot \rho} (K_z \cdot u_0 \sqrt{z_0 - p_U} - K_a \cdot \alpha_a \cdot \sqrt{\rho g w_0}) = 0 \\ \Rightarrow u_0 &= \frac{K_a \cdot \alpha_a \cdot \sqrt{\rho g w_0}}{K_z \cdot \sqrt{z_0 - p_U}} \end{aligned} \quad (3.8)$$

Das Tupel (u_0, z_0, y_0) charakterisiert den wünschenswerten Zustand des Systems und ist somit ein stationärer Arbeitspunkt. Da Gl.(3.8) ausgehend von einer gegebenen Führungsgröße w_0 und einer gemessenen Störung z_0 die passende Stellgröße u_0 berechnet, handelt es sich dabei um eine (statische) Steuerung.

Lineare Systeme sind für ein festes u_0 typischerweise durch eine einzige Ruhelage gekennzeichnet. Im Gegensatz hierzu können nichtlineare Systeme mehrere isolierte Ruhelagen besitzen, in deren Umgebung das System

teilweise stark abweichendes Verhalten aufweist. Als Beispiel sei das reibungsbehaftete, mathematische Pendel betrachtet, welches als System mit zwei Zuständen modelliert werden kann. Wird der Zustand \mathbf{x} gemäß Bild 3-1 zu

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \theta \\ \dot{\theta} \end{bmatrix} \quad (3.9)$$

gewählt, erhält man mit den Methoden der Mechanik bei einer geschwindigkeitsabhängigen Reibung die Zustandsgleichung des nichtlinearen Zustandsraummodells Gl.(3.10)

$$\dot{\mathbf{x}} = \begin{bmatrix} x_2 \\ -\frac{g}{L} \sin(x_1) - \frac{\mu}{ML} x_2 \end{bmatrix} \quad (3.10)$$

mit der Masse der Last M , der Länge des (masselosen) Stabs L und dem Reibungskoeffizienten μ . Bereits aus der physikalischen Anschauung ist erkennbar, dass das System zwei Ruhelagen besitzt. Unter Vernachlässigung der Periodizität des Winkels θ liegen diese bei $\theta = \{0, \pi\}$ und $\dot{\theta} = 0$, so dass sich eine untere (hängendes Pendel) und eine obere (stehendes Pendel) Gleichgewichtslage ausbilden. Dies kann leicht durch Einsetzen der Werte in Gl.(3.10) überprüft werden.

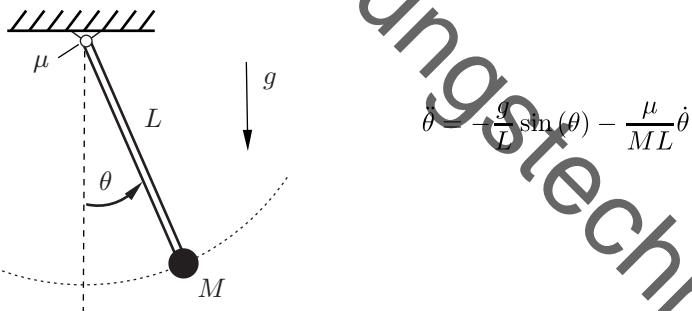


Bild 3-1: Das mathematische Pendel

Zwar handelt es sich sowohl bei der oberen als auch der unteren Gleichgewichtslage um Ruhelagen desselben Systems. Dennoch ist aus der Praxis bekannt, dass sich das Pendel in der Nähe dieser Ruhelagen vollkommen

anders verhält. Würde das Pendel in der oberen Ruhelage initialisiert und ist keinen Störungen unterlegen, verbleibt es in dieser Position und die untere Ruhelage wird nie erreicht. Aber offenbar verlässt das System bereits bei kleinsten Auslenkungen der Last die obere Ruhelage, ohne im Weiteren zu ihr zurückzukehren. Hingegen nimmt das unerregte Pendel die untere Ruhelage auch nach kleinen Auslenkungen oder Störungen wieder ein.

3.2 Stabilität

Die beschriebenen dynamische Eigenschaft des Pendels, bei kleinen Störungen in die Ruhelage zurückzukehren bzw. diese bei beliebig kleinen Störungen zu verlassen, wird in der Regelungstechnik unter dem Begriff der *Stabilität* einer Ruhelage gefasst. Anschaulich bedeutet dabei Stabilität, dass kleine Störungen nicht zu einem Verlassen einer Ruhelage führen, sondern ausgeglichen werden. Somit ist Stabilität eine Eigenschaft, die sich nicht auf ein gesamtes System, sondern auf eine einzelne Ruhelage bezieht und lokal das Systemverhalten nahe der Ruhelage charakterisiert.

Stabilität ist eine enorm wichtige Eigenschaft und nutzbare technische Prozesse sollten diese Eigenschaft in jedem Fall erfüllen. Die regelungstechnisch typische Situation ist, dass die Regelstrecke in eine bestimmte Ruhelage überführt und dort gehalten werden soll. Das Bestimmen einer geeigneten Ruhelage ist dabei ein Problem des Steuerungsentwurfs und wurde in Abschnitt 3.1 bereits angerissen. Die Aufgabe der Regelung ist es, diese Steuerung abzusichern und Abweichungen, die durch fehlerhafte Modellierung, den Anfangszustand oder Störungen verursacht werden, auszugleichen. Eine erfolgreiche Umsetzung dieser Absicherung wird dann durch das Kriterium der Stabilität beschrieben.

Asymptotische Stabilität

Eine Ruhelage x_0 heißt *asymptotisch stabil*, wenn alle Systemtrajektorien, welche in der Nähe der Ruhelage starten, in deren Nähe bleiben und letztlich gegen diese Ruhelage streben. Formal wird dieser Zusammenhang über zwei Grenzen δ und ϵ und eine Grenzwertbedingung definiert

Gelten für die Ruhelage \mathbf{x}_0 die beiden Bedingungen

$\forall \epsilon > 0 \exists \delta > 0$, so dass

$$\forall \mathbf{x}(t) \text{ mit } \|\mathbf{x}(t=0) - \mathbf{x}_0\| < \delta \text{ gilt: } \|\mathbf{x}(t) - \mathbf{x}_0\| < \epsilon \quad \forall t \geq 0 \quad (3.11)$$

sowie

$$\lim_{t \rightarrow \infty} \mathbf{x}(t) = \mathbf{x}_0 \quad , \quad (3.12)$$

so heißt die Ruhelage *asymptotisch stabil* oder kürzer auch *stabil*.

In der ersten Bedingung Gl.(3.11) beschreibt δ die Umgebung um die Ruhelage, der die Startpunkte $\mathbf{x}(t=0)$ der Systemtrajektorien entnommen werden. Die ϵ -Grenze beschreibt hingegen die Umgebung, in der die Systemtrajektorien verweilen. Die zweite Bedingung Gl.(3.12) stellt die Konvergenz des Systems in die Ruhelage sicher.

Asymptotische Stabilität ist die zentrale Zieleigenschaft der Regelungstechnik. Daher spricht man abkürzend oft auch einfach nur von *Stabilität* und immer wenn in diesem Buch über Stabilität geschrieben wird, ist asymptotische Stabilität gemeint. Systeme, die die Bedingung für Stabilität nicht erfüllen und somit nicht stabil sind, lassen sich in zwei Unterklassen aufteilen.

Grenzstabilität und Instabilität

Erfüllt eine Ruhelage nur Bedingung Gl.(3.11), nicht aber Gl.(3.12), so heißt diese Ruhelage *grenzstabil*. Erfüllt eine Ruhelage Bedingung Gl.(3.11) nicht, so heißt sie *instabil*.

Jede Ruhelage ist entweder stabil, grenzstabil oder instabil. Die verschiedenen Stabilitätsbegriffe sollen am Beispiel des Pendelsystems in Bild 3.1 verdeutlicht werden. Für $\mu \neq 0$ dissipiert das System aufgrund der Reibung fortlaufend Energie, sollten die Anfangsbedingungen nicht einer der Ruhelagen entsprechen. Die physikalische Anschauung zeigt, dass das Pendel dann stets gegen die untere Ruhelage konvergiert. Es handelt sich also um eine stabile Ruhelage. Die obere Ruhelage des Pendels ist instabil, da eine beliebig kleine Auslenkung des Pendels zu einem Verlassen der Ruhelage führt. Mathematisch kann also für beispielsweise den Wert $\epsilon = \frac{\pi}{2}$ kein Wert für

$\delta > 0$ gefunden werden, der Gl.(3.11) erfüllt. Für den Fall des reibungsfreien Pendels $\mu = 0$ führt das System eine Dauerschwingung mit konstanter Amplitude aus, solange der Anfangszustand außerhalb einer Ruhelage initialisiert wird. Der Abstand $\|\mathbf{x} - \mathbf{x}_0\|$ zum unteren Gleichgewichtspunkt \mathbf{x}_0 bleibt jedoch in Abhängigkeit des Anfangszustands beschränkt, so dass die untere Ruhelage nunmehr grenzstabil aber nicht mehr stabil ist.

Stabilität ist die Regelungstechnisch wichtigste Eigenschaft. Die Definition in Gl.(3.11) ist dabei für einen direkten mathematischen Nachweis der Stabilität nicht geeignet. Glücklicherweise gibt es ein mathematisches Resultat [17], das die Stabilitätsprüfung erheblich erleichtert. Da nämlich die Stabilität eine lokale Eigenschaft einer Ruhelage ist, kann man diese in vielen Fällen dadurch nachweisen, indem man das eigentlich zu untersuchende (nichtlineare) System durch ein lineares Ersatzsystem annähert – der sogenannten *Linearisierung*.

3.3 Linearisierung

Mit der Linearisierung lassen sich viele Eigenschaften eines Systems in der Umgebung eines Arbeitspunktes untersuchen. Dies hat den Vorteil, die Analyse nichtlinearer Differentialgleichungen auf die Analyse linearer Differentialgleichungen zurückzuführen und damit den mathematischen Untersuchungsmethoden für lineare Differentialgleichungen zugänglich zu machen. Dabei ist die Behandlung linearer Differentialgleichung vergleichsweise einfach, während die große Vielfalt an Funktionen f und damit an möglichen nichtlinearen Differentialgleichungen eine Schematisierung der Untersuchungsmethoden für nichtlineare Systeme erheblich erschwert.

Die Linearisierung wird als Näherung des eigentlich nichtlinearen Systems nur innerhalb bestimmter Bereiche der beteiligten Größen genügend genau sein. Weil aber das Ziel einer Regelung häufig darin besteht, Größen innerhalb enger Bereiche zu halten, sorgt in sehr vielen Fällen eine funktionierende Regelung dafür, dass linearisierte Beziehungen zur Beschreibung der interessierenden Zusammenhänge genügen. Obgleich die allermeisten technisch wichtigen Zusammenhänge nichtlinear sind, strebt man in der Regelungstechnik also ihre Linearisierung an. Zu deren Herleitung wird zunächst die (graphische) Linearisierung einer Funktion wiederholt, um dann den Weg zur Linearisierung einer Differentialgleichung zu beschreiben.

3.3.1 Linearisierung einer Funktion

Betrachtet wird eine nichtlineare Funktion in einer Variablen $Y = f(U)$ mit einem Arbeitspunkt A mit $(U_0, Y_0) = f(U_0)$, für die eine lineare Eratzbeschreibung $y = K \cdot u$ in der Umgebung des Arbeitspunktes gefunden werden soll. Da lineare Abbildungen stets durch den Ursprung des Koordinatensystems verlaufen, muss sinnvollerweise das Koordinatensystem der Linearisierung zunächst in den Arbeitspunkt verschoben werden (siehe Bild 3-2). Dies entspricht genau dem Wechsel von Absolut- zu Abweichungsgrößen, die ein neues Koordinatensystem definieren, dessen Ursprung im Arbeitspunkt liegt. Für eine möglichst gute Annäherung der nichtlinearen Funktion fordert man nun, dass die Linearisierung und die nichtlineare Funktion im Arbeitspunkt die gleiche Steigung aufweisen sollen. Graphisch kann somit die Linearisierung näherungsweise durch das Einzeichnen der Tangente im Arbeitspunkt ermittelt werden.

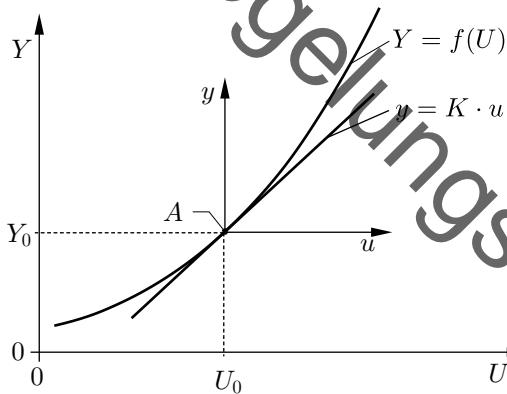


Bild 3-2: Graphische Linearisierung einer Funktion einer Variablen

Das analytische Pendant zum Einzeichnen der Tangente ist das Bestimmen der Ableitung der nichtlinearen Funktion f am Arbeitspunkt. Die Linearis-

sierung lautet also

$$y = \frac{\partial f}{\partial U} \Big|_{U_0} u \quad \text{mit } y = Y - Y_0 \quad \text{und } u = U - U_0 \quad . \quad (3.13)$$

Mathematisch entspricht somit die Linearisierung einer Taylorreihe¹, die nach dem linearen Glied abgebrochen wird.

Taylorreihe

Sei $f : \mathbb{R} \rightarrow \mathbb{R}$ eine glatte Funktion. Dann heißt die unendliche Reihe

$$T(U) = \sum_{n=0}^{\infty} \frac{1}{n!} \frac{\partial^n f}{\partial U^n} \Big|_{U_0} \cdot (U - U_0)^n \quad (3.14)$$

Taylorreihe von f im Punkt U_0 .

Das Verschieben des Koordinatensystems in den Arbeitspunkt sorgt dafür, dass die Abweichungsgrößen so gebildet werden, dass der nullte Term der Taylorreihe in der Linearisierung abgebildet wird. Durch die korrekte Wahl der Steigung der Linearisierung stimmt auch der erste Term der Taylorreihe überein. Daraus folgt, dass – entsprechende Differenzierbarkeit von f vorausgesetzt – die Linearisierung die bestmögliche lineare Approximation von f ist. Dabei meint „bestmöglich“, dass der Approximationsfehler zwischen nichtlinearer Funktion und Linearisierung quadratisch verschwindet. Bricht man die Taylorreihe nach dem ersten Glied ab, so erhält man

$$f(U) \approx f(U_0) + \frac{\partial f}{\partial U} \Big|_{U_0} \cdot (U - U_0) \quad . \quad (3.15)$$

Mit dem Wechsel auf die Abweichungsgrößen $u = U - U_0$ und $y = Y - Y_0 = f(U) - f(U_0)$ erhält man die Linearisierung

$$y \approx \frac{\partial f}{\partial U} \Big|_{U_0} \cdot u \quad . \quad (3.16)$$

Diese Definition der Linearisierung lässt sich über die Jacobi²-Matrix auf Funktionen mit mehreren Ein- und Ausgängen erweitern:

¹Brook Taylor (1685-1731), britischer Mathematiker [56]

²Carl Gustav Jacob Jacobi (1804-1851), preußischer Mathematiker [22]

Linearisierung einer Funktion

Sei $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ eine differenzierbare Funktion und $\mathbf{U}_0 \in \mathbb{R}^n$, $\mathbf{Y}_0 = f(\mathbf{U}_0) \in \mathbb{R}^m$ ein Arbeitspunkt. Mit dem Wechsel auf die Abweichungsgrößen $\mathbf{u} = \mathbf{U} - \mathbf{U}_0$ und $\mathbf{y} = \mathbf{Y} - \mathbf{Y}_0$ erhält man die *Linearisierung*

$$y = \left. \frac{\partial f}{\partial \mathbf{U}} \right|_{\mathbf{U}_0} \cdot \mathbf{u} \quad (3.17)$$

mit der Jacobi-Matrix \mathbf{J} , die alle partiellen Ableitungen enthält:

$$\mathbf{J} = \left. \frac{\partial f}{\partial \mathbf{U}} \right|_{\mathbf{U}_0} = \begin{bmatrix} \left. \frac{\partial f_1}{\partial U_1} \right|_{\mathbf{U}_0} & \left. \frac{\partial f_1}{\partial U_2} \right|_{\mathbf{U}_0} & \cdots & \left. \frac{\partial f_1}{\partial U_n} \right|_{\mathbf{U}_0} \\ \left. \frac{\partial f_2}{\partial U_1} \right|_{\mathbf{U}_0} & \left. \frac{\partial f_2}{\partial U_2} \right|_{\mathbf{U}_0} & \cdots & \left. \frac{\partial f_2}{\partial U_n} \right|_{\mathbf{U}_0} \\ \vdots & \vdots & \ddots & \vdots \\ \left. \frac{\partial f_m}{\partial U_1} \right|_{\mathbf{U}_0} & \left. \frac{\partial f_m}{\partial U_2} \right|_{\mathbf{U}_0} & \cdots & \left. \frac{\partial f_m}{\partial U_n} \right|_{\mathbf{U}_0} \end{bmatrix}. \quad (3.18)$$

Die Linearisierung stellt nur eine lokale Näherung der nichtlinearen Funktion dar. Für eindimensionale Ein- oder Ausgangsgrößen wird die Jacobi-Matrix zu einem Zeilen- oder Spaltenvektor. Für den Fall $n = m = 1$ ist sie ein Skalar, der identisch zu Gl.(3.16) ist. Als Beispiel soll die folgende nichtlineare Beziehung mit einer Ausgangsgröße Y und zwei Eingangsgrößen U und Z linearisiert werden.

$$Y = \frac{Z^2}{U} + Z + B, \quad B = \text{konst.} \quad (3.19)$$

Die Jacobi-Matrix ist ein Zeilenvektor. Mit Gl.(3.17) erhält man nach Ausmultiplizieren den linearen Ausdruck

$$y = -\frac{Z_0^2}{U_0^2} u + \frac{2Z_0}{U_0} z + z \quad (3.20)$$

mit U_0 und Z_0 als den Arbeitspunktwerten von U und Z . Aus diesem Beispiel sieht man, dass durch die Linearisierung ein nichtlinearer Zusammenhang in eine Addition linearer Zusammenhänge zerfällt. Lineare Terme

werden dabei linear übernommen. Entsprechend kann – wenn mit der Linearisierung in der Umgebung eines Arbeitspunktes gearbeitet wird – bei einer Darstellung eines Systems im Wirkungsplan jede nichtlineare Funktion durch eine passende Verwendung von P-Elementen und Summenpunkten dargestellt werden. Die Verstärkungsfaktoren der P-Elemente entsprechen dabei genau den partiellen Ableitungen im definierten Arbeitspunkt.

Das gezeigte Vorgehen zur Linearisierung von Funktionen ist auch für implizite Funktionen

$$0 = f(Y, U, \dots) \quad (3.21)$$

anwendbar. Im Falle von Gleichungssystemen mit Zwischengrößen H wie beispielsweise

$$\begin{aligned} H &= f(Y, U) \\ 0 &= g(U, Z, H) \end{aligned} \quad (3.22)$$

kann aufgrund der Kettenregel jede einzelne Gleichung für sich linearisiert, um diese anschließend ineinander einzusetzen und die Zwischengröße h zu eliminieren. Der Weg, zunächst die Zwischengröße H im Nichtlinearen zu eliminieren und dann zu linearisieren, führt mit einem deutlich höheren Rechenaufwand auf das gleiche Resultat.

3.3.2 Linearisierung einer Differentialgleichung

Die Linearisierung einer Funktion kann dazu genutzt werden, auch nichtlineare Differentialgleichungen zu linearisieren. Hierzu fasst man in der Form

$$Y^{(n)}(t) = f \left(Y^{(n-1)}(t), \dots, \ddot{Y}(t), \dot{Y}(t), Y(t), U^{(m)}(t), \dots, U(t) \right) \quad (3.23)$$

die $n + m + 1$ Argumente $Y^{(n-1)}, \dots, U$ der Funktion f als unabhängige Variablen auf und verfährt wie bei der Linearisierung einer Funktion. Das Ergebnis ist eine lineare Differentialgleichung. Diese Vorgehensweise soll an einem Beispiel erläutert werden.

Es wird angenommen, dass ein beliebiges System durch die folgende nichtlineare Differentialgleichung beschrieben wird:

$$\ddot{Y} = B \cdot \dot{Y}^2 + C \cdot Y \cdot U = f(Y, \dot{Y}, U) \quad , \quad B, C = \text{konst.} \quad (3.24)$$

Die Linearisierung von Gl.(3.24) erfolgt an einem geeigneten Arbeitspunkt A , der im Folgenden als gegeben angenommen wird. Es berechnet sich:

$$\ddot{y} = \frac{\partial f}{\partial Y} \Big|_A \cdot y + \frac{\partial f}{\partial \dot{Y}} \Big|_A \cdot \dot{y} + \frac{\partial f}{\partial U} \Big|_A \cdot u \quad (3.25)$$

wobei die Abweichungsgrößen zu

$$y = Y - Y_0 \quad , \quad \dot{y} = \dot{Y} - \dot{Y}_0 \quad , \quad \ddot{y} = \ddot{Y} - \ddot{Y}_0 \quad , \quad u = U - U_0 \quad (3.26)$$

gesetzt werden und die Größen \ddot{Y}_0 , \dot{Y}_0 , Y_0 und U_0 die entsprechenden Werte am Arbeitspunkt darstellen.

Das Berechnen der einzelnen Ableitungen, das Einsetzen derselben und das Sortieren nach Eingangs- und Ausgangsgrößen führt anschließend auf die gesuchte Linearisierung

$$\underbrace{1}_{a_2} \cdot \ddot{y} - \underbrace{-2 \cdot B \cdot \dot{Y}_0}_{a_1} \cdot \dot{y} - \underbrace{-C \cdot U_0 \cdot y}_{a_0} = \underbrace{C \cdot Y_0 \cdot u}_{b_0} . \quad (3.27)$$

Es ergibt sich eine lineare Differentialgleichung. Im Falle von konstanten Werten Y_0, \dots ist sie zeitinvariant und die Größen a_0 , a_1 , a_2 und b_0 sind die konstanten Koeffizienten der linearisierten Differentialgleichung.

Die Linearisierung eines zeitinvarianten Systems im Zustandsraum

$$\dot{\mathbf{X}} = \mathbf{f}(\mathbf{X}, \mathbf{U}) \quad , \quad \mathbf{Y} = \mathbf{g}(\mathbf{X}, \mathbf{U}) \quad (3.28)$$

in einem stationären Arbeitspunkt A führt auf die lineare Zustandsraumdarstellung Gl.(2.18) eines LTI-Systems. Die Matrizen \mathbf{A} , \mathbf{B} , \mathbf{C} und \mathbf{D} ergeben sich dabei als Jacobi-Matrizen von \mathbf{f} und \mathbf{g} nach \mathbf{X} bzw. \mathbf{U} :

$$\mathbf{A} = \frac{\partial \mathbf{f}}{\partial \mathbf{X}} \Big|_A \quad , \quad \mathbf{B} = \frac{\partial \mathbf{f}}{\partial \mathbf{U}} \Big|_A \quad , \quad \mathbf{C} = \frac{\partial \mathbf{g}}{\partial \mathbf{X}} \Big|_A \quad , \quad \mathbf{D} = \frac{\partial \mathbf{g}}{\partial \mathbf{U}} \Big|_A . \quad (3.29)$$

So ergibt sich bei der Linearisierung des mathematischen Pendels Gl.(3.10)

$$\dot{\mathbf{x}} = \begin{bmatrix} x_2 \\ -\frac{g}{L} \sin(x_1) - \frac{\mu}{ML} x_2 \end{bmatrix} \quad (3.30)$$

mit der Auslenkung des Pendels als Ausgangsgröße $y = x_1$ der allgemeine Ausdruck

$$\mathbf{A} = \frac{\partial \mathbf{f}}{\partial \mathbf{X}} \Big|_{\mathbf{A}} = \begin{bmatrix} 0 & 1 \\ -\frac{g}{L} \cos({}_0 X_1) & -\frac{\mu}{ML} \end{bmatrix}, \quad \mathbf{b} = \frac{\partial \mathbf{f}}{\partial \mathbf{U}} \Big|_{\mathbf{A}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad (3.31)$$

$$\mathbf{c}^T = \frac{\partial \mathbf{g}}{\partial \mathbf{X}} \Big|_{\mathbf{A}} = [1 \quad 0], \quad d = \frac{\partial \mathbf{g}}{\partial \mathbf{U}} \Big|_{\mathbf{A}} = 0.$$

Abhängig davon, ob die obere oder unterer Ruhelage betrachtet wird, ergibt sich bei der Linearisierung ein unterschiedliches Vorzeichen innerhalb der Systemmatrix:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ \pm \frac{g}{L} & -\frac{\mu}{ML} \end{bmatrix}. \quad (3.32)$$

3.3.3 Linearisierung im Kennlinienfeld

Die beschriebenen Methoden zur Linearisierung von Funktionen und Differentialgleichungen reichen in der präsentierten Form gelegentlich nicht aus, um ein lineares Ersatzsystem zu bestimmen. Das ist insbesondere dann der Fall, wenn der interessierende nichtlineare Zusammenhang nicht als Funktion, sondern als Kennlinienfeld gegeben ist. Kennlinienfelder sind in den Ingenieurwissenschaften zur Darstellung der Zusammenhänge zwischen einer abhängigen und zwei oder mehreren unabhängigen Größen verbreitet. Zur näherungsweisen Bestimmung der Ableitung der abhängigen Größe nach einer Größe, die Parameter von Kennlinien ist, benutzt man zweckmäßigerweise einen geeigneten Differenzenquotienten.

Die Linearisierung des in Bild 3-3 dargestellten Kennlinienfeldes liefert die Koeffizienten K_u und K_z der linearen Gleichung für Abweichungsgrößen

$$y = K_u \cdot u + K_z \cdot z. \quad (3.33)$$

Der Koeffizient K_u ergibt sich analog zur Linearisierung einer Funktion aus der Steigung der Kennlinie für $Z = Z_0$ im Arbeitspunkt. Um den Koeffizienten K_z nach dem gleichen Verfahren bestimmen zu können, benötigt man eine Kennlinie $Y = f(Z, U_0)$ mit $U = U_0$ als Parameter. Im rechten Teil von Bild 3-3 ist eine solche Kennlinie dargestellt. Man erkennt, dass der

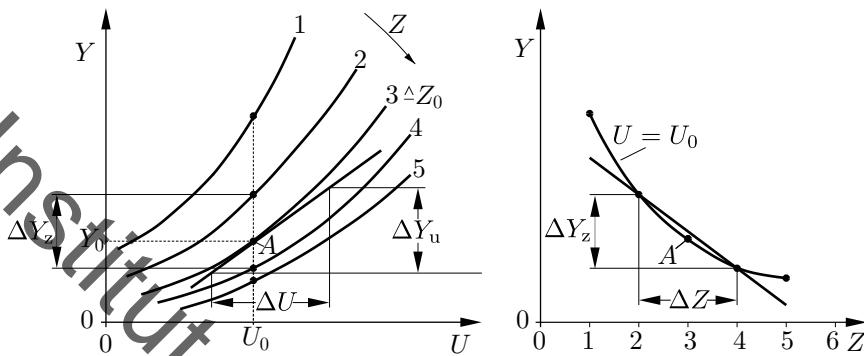


Bild 3-3: Linearisierung eines Kennlinienfeldes

Unterschied zwischen der Steigung der (nicht eingezeichneten) Tangente im Arbeitspunkt an diese Kennlinie und der Steigung der Sekante durch die dem Arbeitspunkt benachbarten Punkte $Y(U_0, Z = 2)$ und $Y(U_0, Z = 4)$ gering ist. Bei geeigneter Wahl der Sekante erhält man also nur kleine Fehler, wenn man auf die Tangente verzichtet. Dieser Schritt liegt nahe, weil die Steigung der Sekante ohne zusätzliche Zeichenarbeit aus dem ursprünglichen Kennlinienfeld (Bild 3-3, links) über

$$K_z = \frac{\partial Y}{\partial Z} \Big|_A \cong \frac{\Delta Y_z}{\Delta Z} \Big|_{U=U_0} \quad (3.34)$$

bestimbar ist. Bei dieser Vorgehensweise ist darauf zu achten, dass die Sekanten durch Punkte gelegt werden, die in etwa symmetrisch zum Arbeitspunkt liegen und nicht allzu weit von ihm entfernt sind.

Da das Vorzeichen von der Richtung wachsender Werte des Parameters Z abhängt, ist es wichtig, die Differenzen so zu bilden, dass die zu einem Punkt gehörenden Y - und Z -Werte mit gleichem Vorzeichen eingesetzt werden. Welche Werte dann mit positivem und welche mit negativem Vorzeichen versehen werden, ist gleichgültig, d.h.

$$K_z \cong \frac{Y(U_0, Z = 4) - Y(U_0, Z = 2)}{(Z = 4) - (Z = 2)} = \frac{Y(U_0, Z = 2) - Y(U_0, Z = 4)}{(Z = 2) - (Z = 4)} . \quad (3.35)$$

3.4 Charakteristisches Polynom

Mit der Linearisierung liegt ein mathematisches Werkzeug bereit, ein nicht-lineares System durch ein lineares Ersatzsystem zu beschreiben. Um hierüber die Stabilität des nichtlinearen autonomen Systems analysieren zu können, muss zunächst geklärt werden, wie Stabilität für ein autonomes LTI-System nachgewiesen werden kann. Zum Nachweis wird prinzipiell der Verlauf $x(t)$ benötigt. Um diesen zu erhalten, scheint es notwendig, die zugehörige Differentialgleichung zu lösen. Die Ermittlung dieser Lösung ist aber oft aufwendig, sodass man bestrebt ist, diese zu umgehen und dennoch über ihre Struktur Aussagen über ihre dynamischen Eigenschaften zu gewinnen. Hierfür ist die Kenntnis von Verfahren zur Lösung linearer Differentialgleichungen mit konstanten Koeffizienten notwendig.

Zunächst wird die autonome Differentialgleichung erster Ordnung

$$\dot{x} = \lambda x \quad x(t=0) = {}_0x \quad (3.36)$$

betrachtet. Es wird also eine Funktion $x(t)$ gesucht, die einmal abgeleitet sich selbst mit dem Faktor $\lambda \in \mathbb{R}$ multipliziert ergibt. Die Lösungsfunktion ist aufgrund der Kettenregel die Exponentialfunktion

$$x(t) = C \cdot e^{\lambda t} \Rightarrow \dot{x}(t) = C \cdot \lambda \cdot e^{\lambda t} \quad (3.37)$$

mit der Konstanten C . Diese kann über ein Einsetzen von $t = 0$ direkt zu $C = {}_0x$ bestimmt werden. Diese Lösung ist eindeutig und damit die einzige Lösung der Differentialgleichung Gl.(3.36).

Aus dieser Lösung der Differentialgleichung können direkt die Stabilitätseigenschaften des Systems aus Gl.(3.36) abgelesen werden. Die Ruhelage liegt bei $x = 0$. Die Forderung in Gl.(3.12) führt auf

$$\lim_{t \rightarrow \infty} x(t) = \lim_{t \rightarrow \infty} {}_0x \cdot e^{\lambda t} \stackrel{!}{=} 0 \quad . \quad (3.38)$$

Für ${}_0x \neq 0$ (also eine beliebige Auslenkung aus der Ruhelage) ist die Grenzwertbedingung erfüllt, wenn die Exponentialfunktion für steigende t abklingt. Das ist genau dann der Fall, wenn $\lambda < 0$ ist. Die Ruhelage ist in diesem Fall stabil. Für $\lambda > 0$ klingt die Exponentialfunktion auf und die Ruhelage ist instabil. Für $\lambda = 0$ ist die Lösung $x(t) = {}_0x$ konstant. Folglich

ist jedes $x_0 \in \mathbb{R}$ eine Ruhelage – es gibt derer also unendlich viele. Da alle Lösungen konstant bleiben, ist auch die Forderung Gl.(3.11) mit $\epsilon = \delta$ erfüllt. Alle Ruhelagen sind grenzstabil.

Die sich ergebende Exponentialfunktion für $\lambda = -\frac{1}{T} < 0$ ist in Bild 3-4 dargestellt. Zu den Eigenschaften dieser Funktion gehört, dass eine Tangente an einen beliebigen Punkt der Funktion eine Strecke von der Größe der Zeitkonstanten T auf der Asymptoten abschneidet.

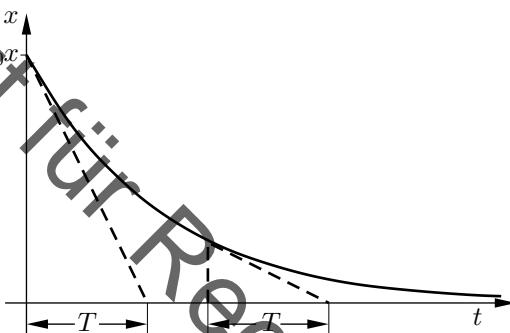


Bild 3-4: Lösung $x(t) = x_0 \cdot e^{-t/T}$

Ausgehend von der Differentialgleichung erster Ordnung können diese Erkenntnisse auf den allgemeinen Fall eines autonomen LTI-Systems n -ter Ordnung übertragen werden. Die allgemeine Form lautet

$$a_n y^{(n)} + \dots + a_1 \dot{y} + a_0 y = 0 \quad . \quad (3.39)$$

Setzt man hier analog zum Vorgehen bei einer Differentialgleichung erster Ordnung den Lösungskandidaten $C \cdot e^{\lambda t}$ an, so erhält man

$$a_n y^{(n)} + \dots + a_1 \dot{y} + a_0 y = a_n \lambda^n C \cdot e^{\lambda t} + \dots + a_1 \lambda C \cdot e^{\lambda t} + a_0 C \cdot e^{\lambda t} = 0. \quad (3.40)$$

Ausklammern des Lösungskandidaten führt auf

$$C \cdot e^{\lambda t} \cdot (a_n \lambda^n + \dots + a_2 \lambda^2 + a_1 \lambda + a_0) = 0 \quad . \quad (3.41)$$

Da der erste Term des Produkts für $C \neq 0$ nicht verschwindet, muss λ so gewählt werden, dass der Ausdruck innerhalb der Klammer verschwindet:

$$a_n \lambda^n + \dots + a_2 \lambda^2 + a_1 \lambda + a_0 \stackrel{!}{=} 0 \quad . \quad (3.42)$$

Also muss λ die Wurzel eines aus den Koeffizienten a_i gebildeten Polynoms sein, das auch *charakteristische Gleichung* oder *charakteristisches Polynom* genannt wird.

Charakteristisches Polynom

Jede Differentialgleichung eines autonomen Systems der Ordnung n

$$a_n y^{(n)} + \dots + a_1 \dot{y} + a_0 y = 0 \quad (3.43)$$

besitzt das folgende charakteristische Polynom der Ordnung n :

$$p(\lambda) = a_n \lambda^n + \dots + a_1 \lambda^1 + a_0 . \quad (3.44)$$

Zur Lösung der Differentialgleichung benötigt man also die Wurzeln des zugehörigen charakteristischen Polynoms. Ein solches Polynom n -ten Grades mit reellen Koeffizienten hat n Wurzeln $\lambda_1, \dots, \lambda_n$, die entweder reell oder paarweise konjugiert komplex sind. Jedes λ_i führt auf einen Lösungskandidaten der Differentialgleichung in der Form $e^{\lambda_i t}$. Aufgrund des Überlagerungsprinzips für lineare Systeme ist auch jede lineare Überlagerung ein Lösungskandidat.

Zunächst werden nur Systeme betrachtet, deren Wurzeln λ_i des charakteristischen Polynoms alle voneinander verschieden sind. Unter dieser Voraussetzung hat die Lösung der Differentialgleichung Gl.(3.39) die Form

$$y(t) = C_1 \cdot e^{\lambda_1 t} + C_2 \cdot e^{\lambda_2 t} + \dots + C_n \cdot e^{\lambda_n t} . \quad (3.45)$$

Die Konstanten C_1, \dots, C_n werden aus den Anfangsbedingungen bestimmt. Da für die Betrachtung der Stabilität beliebige Anfangsbedingungen in der Nähe der Ruhelage möglich sind, wird im Allgemeinen $C_i \neq 0$ gelten und alle λ_i werden das Systemverhalten für eine Anfangsbedingung in der Nähe der Ruhelage beeinflussen. Sind alle λ_i rein reell, so kann mit dem gleichen Argument wie für den Fall einer Differentialgleichung erster Ordnung gefolgert werden, dass alle $\lambda_i < 0$ sein müssen, damit das System stabil ist. Ist ein einziges $\lambda_i > 0$, so ist die Ruhelage instabil. Ist ein $\lambda_i = 0$, gibt es erneut unendlich viele Ruhelagen. Diese sind grenzstabil, wenn die übrigen λ_j negativ sind.

Dieses Resultat lässt sich auch auf komplex konjugierte Wurzeln $\sigma \pm j\omega$ mit

der imaginären Einheit j übertragen. Wegen der Rechenregel

$$e^{(\sigma+j\omega)t} + e^{(\sigma-j\omega)t} = e^{\sigma t} (e^{j\omega t} + e^{-j\omega t}) = 2e^{\sigma t} \cos(\omega t) \quad (3.46)$$

führen paarweise komplexe konjugierte Wurzeln zu Lösungsanteilen mit $\sin(\omega t)$ und $\cos(\omega t)$. Diese Systeme werden *schwingungsfähig* genannt.

Schwingungsfähige Systeme

Besitzt das charakteristische Polynom eines Systems Wurzeln mit nicht verschwindendem Imaginärteil, so heißt das System *schwingungsfähig*, da die homogene Lösung sinusförmige Anteile enthält.

Die Frequenz dieser Schwingungen heißt *Eigenkreisfrequenz*.

Eigenkreisfrequenz

Die Frequenz der Schwingung eines schwingungsfähigen Systems wird durch den Imaginärteil der komplexe konjugierten Wurzeln vorgegeben. Diese Frequenz wird *Eigenkreisfrequenz* genannt und oft mit ω_D bezeichnet.

Das Abklingverhalten der Schwingung wird durch den Realteil σ bestimmt. Hieraus folgt, dass der Realteil aller λ_i negativ sein muss, damit die Ruhelage in $y = 0$ stabil ist. Ist ein Realteil positiv, ist die Ruhelage instabil. Ist ein $\lambda_i = 0$, gibt es ein Kontinuum von Ruhelagen und damit unendlich viele. Diese sind grenzstabil, wenn die Realteile der übrigen λ_j negativ sind. Gibt es ein rein imaginäres Wurzelpaar bei $\pm j\omega$, so führt das System nicht abklingende Dauerschwingungen $\cos(\omega t)$ aus. Analog zum mathematischen Pendel ist die Amplitude dieser Schwingung konstant und die Ruhelage grenzstabil. Diese Überlegungen sind in Bild 3-5 visuell zusammengefasst, wo die Wurzeln durch Kreuze markiert sind.

Die vorherigen Ausführungen zur Stabilität gelten für den Fall einfacher Wurzeln, die alle voneinander verschieden sind. Im Falle mehrfacher Wurzeln besitzt der Ansatz in Gl.(3.45)

$$y(t) = C_1 \cdot e^{\lambda_1 t} + C_2 \cdot e^{\lambda_2 t} + \dots + C_n \cdot e^{\lambda_n t} \quad (3.47)$$

mehrere Lösungsanteile mit identischer e -Funktion. Hierdurch werden die $e^{\lambda_i t}$ nicht mehr mit konstanten Koeffizienten C_i , sondern mit Polynome

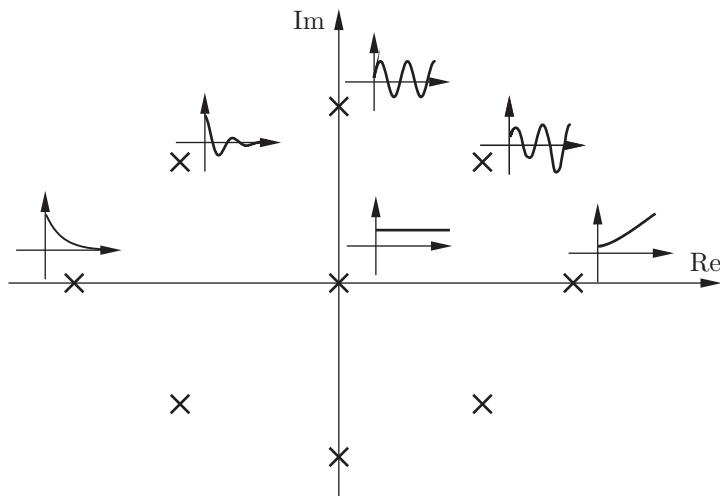


Bild 3-5: Wurzeln des charakteristischen Polynoms und zugehörige Lösungen der homogenen Differentialgleichungen

in der Zeit t gewichtet [43]. Liegt eine k -fache Wurzel vor, ergibt sich der zugehörige Lösungsanteil zu

$$e^{\lambda t} (C_1 + C_2 t + \dots + C_k t^{k-1}) \quad . \quad (3.48)$$

Falls die charakteristische Gleichung mehrfache Wurzeln aufweist, ändert sich also die Form der Lösung; sie enthält aber auch dann Terme $e^{\lambda t}$, die das Auf- und Abklingenverhalten festlegen. Allerdings fällt eine Exponentialfunktion mit negativem Exponenten schneller ab als jedes Polynom ansteigt. Daher bleiben die Aussagen zur Stabilität für negative und positive Realteile der λ_i unverändert. Für den Fall eines verschwindenden Realteils $\text{Re}(\lambda_i) = 0$ führen die zusätzlichen polynomialen Anteile jedoch zur Instabilität der Ruhelage. Somit kann die Stabilität einer Ruhelage für den Fall eines autonomen LTI-Systems vollständig auf die Realteile der Wurzeln des charakteristischen Polynoms zurückgeführt werden.

Stabilität von LTI-Systemen

Für ein autonomes LTI-System mit einer Ruhelage in $y = 0$ hängt die Stabilität der Ruhelage mit den Wurzeln λ_i des zugehörigen charakteristischen Polynoms wie folgt zusammen:

- Gilt für den Realteil aller Wurzeln $\operatorname{Re}(\lambda_i) < 0$, so ist die Ruhelage stabil.
- Gilt für den Realteil mindestens einer Wurzel $\operatorname{Re}(\lambda_i) > 0$, so ist die Ruhelage instabil.
- Gilt für den Realteil einer mehrfachen Wurzel $\operatorname{Re}(\lambda_i) = 0$, so ist die Ruhelage instabil.
- Ansonsten (d.h. einfache Wurzeln mit $\operatorname{Re}(\lambda_i) = 0$, alle anderen Wurzeln mit $\operatorname{Re}(\lambda_j) < 0$) ist die Ruhelage grenzstabil.

Für den Fall $\lambda = 0$ treten unendlich viele Ruhelagen auf, da die Lösung einen konstanten Lösungsanteil enthält. Alle diese Ruhelagen besitzen aber offenbar dieselbe Stabilitätseigenschaft (entweder grenzstabil oder instabil). Daher hat es sich eingebürgert, bei LTI-Systemen nicht von der Stabilität einer Ruhelage, sondern der Stabilität des Systems zu sprechen. Da die imaginäre Achse offenbar der Rand des Bereiches der stabilen Wurzeln ist, hat sich folgende Begrifflichkeit etabliert.

Stabilitätsrand

Die imaginäre Achse wird auch als *Stabilitätsrand* bezeichnet. Ein LTI-System befindet sich genau dann *am Stabilitätsrand*, wenn eine oder mehrere Wurzeln des charakteristischen Polynoms auf dem Stabilitätsrand liegen, während sich alle anderen Wurzeln im stabilen Bereich (negativer Realteil) befinden.

Im Unterschied zu Systemen mit Wurzeln mit positivem Realteil, deren Lösungen exponentiell wachsen, und ausschließlich negativem Realteil, deren Lösungen exponentiell abfallen, besitzen Systeme am Stabilitätsrand ein polynomielles Wachstum. Die Ordnung des Polynoms wird durch die Vielfachheit der Wurzeln auf dem Stabilitätsrand vorgegeben, wodurch das System entweder als grenzstabil oder instabil zu klassifizieren ist. Einfache Wurzeln führen zu einem „Wachstum“ in Form eines Polynoms der Ordnung 0. Die Verläufe bleiben also beschränkt ohne zu wachsen oder abzuklingen.

3.5 Linearisierungstheorem

Diese Auflistung zum Stabilitätsrand zeigt, dass der Fall mit strikt positivem Realteil und strikt negativem Realteil vergleichsweise einfach zu behandeln ist, während bei Wurzeln auf der imaginären Achse die Stabilitätsuntersuchung durch die Betrachtung der Vielfachheiten an Komplexität gewinnt und Systeme grenzstabil oder instabil sein können. Diese Komplexität finden sich auch im Linearisierungstheorem (Satz von Hartman³-Grobman) wieder, welches den Zusammenhang zwischen linearisiertem und nichtlinearem System beschreibt:

Linearisierungstheorem (Hartman-Grobman)

Gegeben ist ein nichtlineares System mit der Ruhelage \mathbf{X}_0 und seine Linearisierung in \mathbf{X}_0 , sodass das linearisierte System eine Ruhelage in $\mathbf{x}_0 = 0$ besitzt. Die Wurzeln des charakteristischen Polynoms der Linearisierung sind λ_i . Dann ähneln sich das Systemverhalten des linearisierten Systems und des nichtlinearen Systems in einer Umgebung der Ruhelage. Das bedeutet:

- Gilt für alle Wurzeln $\operatorname{Re}(\lambda_i) < 0$, so ist \mathbf{X}_0 eine stabile Ruhelage des nichtlinearen Systems.
- Gilt für mindestens eine Wurzel $\operatorname{Re}(\lambda_i) > 0$, so ist \mathbf{X}_0 eine instabile Ruhelage des nichtlinearen Systems.
- Ist das linearisierte System am Stabilitätsrand, ist keine Aussage über die Stabilität von \mathbf{X}_0 anhand der Linearisierung möglich. Diese könnte stabil, grenzstabil oder instabil sein.

Durch die Betrachtung der Linearisierung anstelle des nichtlinearen Systems wird ein kleiner Fehler gemacht. Anschaulich kann man sich vorstellen, dass dieser Fehler die Wurzeln des charakteristischen Polynoms minimal verschiebt. Sind die Realteile strikt positiv oder negativ, ändert sich das Vorzeichen nicht und die Stabilität oder Instabilität der Ruhelage bleibt erhalten. Ist der Realteil jedoch identisch Null, so kann dieser Linearisierungsfehler die Wurzel in den stabilen oder instabilen Bereich verschieben. Eine Aussage über die Stabilität ist auf diesem Wege nicht möglich und muss – sofern erforderlich – durch andere Methoden, wie sie im Kapitel 16

³Philip Hartman (1915-2015), amerikanischer Mathematiker [17]

vorgestellt werden, ermittelt werden. Daher wird man im Falle eines Reglerentwurfs den Regler immer so auslegen, dass das geregelte linearisierte System stabil und nicht nur grenzstabil ist, da nur dadurch auch die Stabilität der Ruhelage des nichtlinearen Systems sichergestellt werden kann.

Als zentraler Zusammenhang bleibt festzuhalten, dass sich die Stabilität des linearisierten Systems auf das nichtlineare System überträgt. Dieses Resultat liefert die mathematische Rechtfertigung für das regelungstechnische Vorgehen, Systeme zu linearisieren und das Systemverhalten in einer Umgebung eines Arbeitspunktes zu untersuchen. Für lineare Systeme lässt sich die Stabilität vergleichsweise einfach untersuchen und der Umweg über das linearisierte System stellt ein Hauptwerkzeug für die Stabilitätsanalyse nichtlinearer Systeme dar. Daher werden im Folgenden mit der Ausnahme weniger Kapitel ausschließlich LTI-Systeme untersucht, da alle anderen Fälle mindestens lokal auf diese Systemklasse zurückgeführt werden können.

Als Beispiel für die Stabilitätsuntersuchung über die Linearisierung wird das mathematische Pendel betrachtet. Die Linearisierung in Gl.(3.32) führt nach Umstellung als Differentialgleichung auf

$$\ddot{y} + \frac{\mu}{ML}\dot{y} \pm \frac{g}{L}y = 0 \quad , \quad (3.49)$$

wobei das positive Vorzeichen die untere, das negative Vorzeichen die obere Ruhelage beschreibt. Das charakteristische Polynom ist

$$\lambda^2 + \frac{\mu}{ML}\lambda \pm \frac{g}{L} . \quad (3.50)$$

Für die untere Ruhelage sind die Wurzeln

$$\lambda_{1,2} = -\frac{\mu}{2ML} \pm \sqrt{\underbrace{\frac{\mu^2}{4M^2L^2} - \frac{g}{L}}_{\text{Wurzelterm}}} . \quad (3.51)$$

Der Wurzelterm ist für große $\frac{g}{L}$ imaginär. Der verbliebene Realteil ergibt sich dann zu $-\frac{\mu}{2ML}$. Für kleine $\frac{g}{L}$ ist der Wurzelterm betragsmäßig kleiner als $-\frac{\mu}{2ML}$, sodass sich das Vorzeichen des Realteils nicht ändert. Für $\mu > 0$ ist der Realteil aller Wurzeln negativ und das linearisierte System ist stabil. Somit ist die untere Ruhelage des Pendels ebenfalls stabil. Für $\mu = 0$ wird der Realteil null und es liegt ein konjugiert komplexes Wurzelpaar mit

Vielfachheit 1 vor. Das linearisierte System ist am Stabilitätsrand und es kann nicht auf die Stabilität der unteren Ruhelage des nichtlinearen Systems geschlossen werden. Dass diese für den Fall $\mu = 0$ ebenfalls grenzstabil ist, kann aus der Untersuchung des linearisierten Systems nicht abgeleitet werden.

Für die obere Ruhelage ergeben sich aufgrund des anderen Vorzeichens die abweichenden Wurzeln

$$\lambda_{1,2} = -\frac{\mu}{2ML} \pm \sqrt{\underbrace{\frac{\mu^2}{4M^2L^2} + \frac{g}{L}}_{\text{Wurzelterm}}} . \quad (3.52)$$

Der Wurzelterm ist für alle positiven g und L betragsmäßig größer als $-\frac{\mu}{2ML}$. Somit besitzt das System für alle μ eine Wurzel mit positivem Realteil. Die obere Ruhelage ist also unabhängig von μ instabil.

3.6 Analyse im Zustandsraum

Der Terminus des *charakteristischen Polynoms* ist aus der Mathematik vor allem im Kontext der Bestimmung von Eigenwerten einer Matrix geläufig. Tatsächlich besteht ein direkter Zusammenhang zwischen den Eigenwerten der Systemmatrix \mathbf{A} und den Wurzeln λ_i des charakteristischen Polynoms der Differentialgleichung, sodass die Stabilitätseigenschaften statt über die Differentialgleichung ohne Umweg auch direkt im Zustandsraum ermittelt werden können. Zur Herleitung der entsprechenden Stabilitätsbedingungen im Zustandsraum wird von der Darstellung der Differentialgleichung in Regelungsnormalform

$$\dot{\mathbf{x}} = \underbrace{\begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & & 1 \\ -a_0 & -a_1 & -a_2 & \dots & -a_{n-1} \end{bmatrix}}_{\mathbf{A}} \cdot \mathbf{x} \quad (3.53)$$

ausgegangen, in welcher der führende Koeffizient a_n zu eins normiert wird. Möchte man die Eigenwerte der Systemmatrix \mathbf{A} ermitteln, so besteht eine

Möglichkeit darin, die Determinante der Matrix $\lambda\mathbf{I} - \mathbf{A}$ mit der Identitätsmatrix \mathbf{I} zu null zu setzen:

$$\det(\lambda\mathbf{I} - \mathbf{A}) = \det \begin{pmatrix} \lambda & -1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda & -1 \\ a_0 & a_1 & a_2 & \cdots & \lambda + a_{n-1} \end{pmatrix} \stackrel{!}{=} 0. \quad (3.54)$$

Dabei stehen die Koeffizienten des charakteristischen Polynoms in der letzten Zeile. Möchte man nun $\det(\lambda\mathbf{I} - \mathbf{A})$ berechnen, so kann dies über den Laplaceschen Entwicklungssatz geschehen. Entwickelt man dabei nach der letzten Zeile, so ergibt sich

$$\begin{aligned} \det(\lambda\mathbf{I} - \mathbf{A}) &= (-1)^n a_0 \det(\mathbf{H}_0) + (-1)^{n+1} a_1 \det(\mathbf{H}_1) + \cdots \\ &\quad + (-1)^{2n-1} (\lambda + a_{n-1}) \det(\mathbf{H}_{n-1}) \end{aligned} \quad (3.55)$$

mit den Untermatrizen \mathbf{H}_i . Diese haben durch das Streichen der letzten Zeile und $i+1$ -ten Spalte die spezielle Struktur

$$\mathbf{H}_i = \left\{ \begin{array}{c|cc|c} i \text{ Spalten} & & & n-1-i \text{ Spalten} \\ \hline \begin{array}{c} i \text{ Zeilen} \\ \hline n-1-i \text{ Zeilen} \end{array} & \begin{array}{cccc} \lambda & -1 & 0 & \cdots \\ 0 & \ddots & \ddots & \\ \vdots & & \ddots & -1 \\ \vdots & & & \lambda \end{array} & \begin{array}{ccc} \cdots & \cdots & 0 \\ \vdots & & \vdots \\ -1 & & \vdots \\ \lambda & \ddots & \vdots \\ \vdots & & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & \lambda & -1 \end{array} \end{array} \right.. \quad (3.56)$$

Die verbleibenden Untermatrizen bestehen also aus einer unteren und einer oberen Dreiecksmatrix, die blockweise zusammengesetzt sind. Daher lässt sich die Determinante – ähnlich zu Dreiecksmatrizen – direkt als das

Produkt der Diagonalelemente angeben. Hiermit ergibt sich für die Determinante

$$\begin{aligned}\det(\lambda\mathbf{I} - \mathbf{A}) &= (-1)^n a_0 \cdot (-1)^{n-1} + (-1)^{n-1} a_1 \cdot (-1)^{n-2} \lambda + \dots \\ &\quad (-1)^{n-2} a_2 \cdot (-1)^{n-3} \lambda^2 + \dots + (\lambda + (-1)^{2n-1} a_{n-1} \cdot \lambda^{n-1}) \\ &= \pm (a_0 + a_1 \lambda + a_2 \lambda^2 + \dots + a_{n-1} \lambda^{n-1} + \lambda^n)\end{aligned}\quad (3.57)$$

genau das charakteristische Polynom der Differentialgleichung. Das wechselnde Vorzeichen ist dabei irrelevant, da die Gleichung zu null gesetzt wird. Nun war aber die Formel $\det(\lambda\mathbf{I} - \mathbf{A}) = 0$ die Bestimmungsgleichung für die Eigenwerte λ der Matrix \mathbf{A} . Also sind die Wurzeln des charakteristischen Polynoms einer Differentialgleichung und die Eigenwerte der Systemmatrix \mathbf{A} der Zustandsraumdarstellung in Regelungsnormalform identisch. Dieses Resultat beschränkt sich allerdings nicht nur auf die Regelungsnormalform, sondern gilt für alle Zustandsraumdarstellungen, die sich in diese transformieren lassen.

Zustandstransformation

Unter Anwendung der linearen Zustandstransformation \mathbf{T}

$$\tilde{\mathbf{x}} = \mathbf{T}\mathbf{x} \quad (3.58)$$

ergeben sich die transformierten Zustandsgleichungen zu

$$\begin{aligned}\dot{\tilde{\mathbf{x}}} &= \underbrace{\mathbf{T} \cdot \mathbf{A} \cdot \mathbf{T}^{-1}}_{\tilde{\mathbf{A}}} \cdot \tilde{\mathbf{x}} + \underbrace{\mathbf{T} \cdot \mathbf{B}}_{\tilde{\mathbf{B}}} \cdot \mathbf{u} \\ \mathbf{y} &= \underbrace{\mathbf{C} \cdot \mathbf{T}^{-1}}_{\tilde{\mathbf{C}}} \cdot \tilde{\mathbf{x}} + \underbrace{\mathbf{D}}_{\tilde{\mathbf{D}}} \cdot \mathbf{u},\end{aligned}\quad (3.59)$$

wobei jede reguläre Transformationsmatrix \mathbf{T} auch invertierbar ist.

Nun ist aus der Mathematik bekannt, dass eine reguläre Zustandstransformation die Eigenwerte nicht verändert, d.h. die Eigenwerte von \mathbf{A} und $\tilde{\mathbf{A}}$ sind identisch. Also überträgt sich das vorherige Resultat auf jede Zustandsraumdarstellung, die sich über eine Zustandstransformation \mathbf{T} aus der Regelungsnormalform gewinnen lässt. Für alle minimalen linearen Zustandsraumdarstellungen lässt sich eine solche Matrix \mathbf{T} finden, weshalb es sinnvoll ist, nur solche Zustandsraumdarstellungen zu betrachten.

Stabilität im Zustandsraum

Gegeben ist die autonome Differentialgleichung eines LTI-Systems und die zugehörige Zustandsraumdarstellung $\dot{x} = \mathbf{A} \cdot x$. Dann ist das charakteristische Polynom

$$p(\lambda) = \det(\lambda \mathbf{I} - \mathbf{A}) \quad . \quad (3.60)$$

Die Wurzeln des charakteristischen Polynoms entsprechen damit genau den Eigenwerten von \mathbf{A} , die für Stabilität alle einen negativen Realteil aufweisen müssen.

Als Beispiel soll das autonome System

$$\dot{x} = \mathbf{A}x = \begin{bmatrix} 3 & -2 & -12 \\ 4 & -2 & -12 \\ 1,5 & -0,5 & -6 \end{bmatrix} x \quad (3.61)$$

auf Stabilität untersucht werden. Da es für die Bestimmung von Eigenwerten entsprechende Rechnerwerkzeuge gibt, kann in einfacher Weise die Stabilität geprüft werden. Der Aufruf einer passenden Software liefert für die Eigenwerte

$$\lambda_1 = -3 \quad , \quad \lambda_{2,3} = -1 \pm j \quad . \quad (3.62)$$

Da alle Realteile der Eigenwerte negativ sind, ist das System stabil. Zudem ist das System aufgrund der komplexwertigen Eigenwerte $\lambda_{2,3}$ schwingungsfähig. Ebenfalls mit Rechnerunterstützung gewinnt man leicht die Transformationsmatrix

$$\mathbf{T} = \begin{bmatrix} 0 & -1 & 2 \\ -1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad , \quad (3.63)$$

welche das System in die Regelungsnormalform

$$\mathbf{A}_{RNF} = \mathbf{T} \cdot \begin{bmatrix} 3 & -2 & -12 \\ 4 & -2 & -12 \\ 1,5 & -0,5 & -6 \end{bmatrix} \cdot \mathbf{T}^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -8 & -5 \end{bmatrix} \quad (3.64)$$

überführt. Das charakteristische Polynom steht hier in der untersten Zeile und ist genau

$$p(\lambda) = \lambda^3 + 5\lambda^2 + 8\lambda + 6 = (\lambda + 3) \cdot (\lambda + 1 + j) \cdot (\lambda + 1 - j) \quad (3.65)$$

mit den Nullstellen als zuvor numerisch berechneten Eigenwerten.

Die Stabilität und Schwingungsfähigkeit eines LTI-Übertragungssystems kann also auch direkt im Zustandsraum analysiert werden. Dies wird auch durch die Herleitung der allgemeinen Lösung der autonomen Differentialgleichung $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ im Zustandsraum deutlich. Da die Zustandsraumdarstellung eine (vektorielle) Differentialgleichung erster Ordnung ist, kann deren Lösung in Analogie zu Differentialgleichung erster Ordnung in Gl.(3.36)

$$\dot{x} = \lambda x \Rightarrow x(t) = e^{\lambda t} \quad (3.66)$$

gefunden werden. Die Lösung der Differentialgleichung erster Ordnung ergibt sich dabei über die Reihenentwicklung der Exponentialfunktion

$$e^{\lambda t} = \sum_{k=0}^{\infty} \frac{1}{k!} (\lambda t)^k = 1 + \frac{\lambda t}{1!} + \frac{(\lambda t)^2}{2!} + \dots \quad (3.67)$$

aus welcher durch Vertauschen von Ableitung und unendlicher Summe die Rechenregel

$$\begin{aligned} \frac{d}{dt} e^{\lambda t} &= \frac{d}{dt} \sum_{k=0}^{\infty} \frac{1}{k!} (\lambda t)^k = \sum_{k=0}^{\infty} \frac{1}{k!} \frac{d}{dt} (\lambda t)^k = \sum_{k=1}^{\infty} \frac{k}{k!} (\lambda t)^{k-1} \\ &= \lambda \sum_{k=1}^{\infty} \frac{1}{(k-1)!} (\lambda t)^{k-1} \stackrel{k-1 \rightarrow i}{=} \lambda \sum_{i=0}^{\infty} \frac{1}{i!} (\lambda t)^i = \lambda e^{\lambda t} \end{aligned} \quad (3.68)$$

folgt. Das Vertauschen von Ableitung und unendlicher Summe ist dabei aufgrund der Konvergenzeigenschaften der Reihe sichergestellt [43]. Die Definition der Exponentialfunktion als Reihe kann man nun direkt auch auf Matrizen übertragen:

Transitionsmatrix

Sei \mathbf{A} eine $(n \times n)$ -Matrix. Dann heißt die $(n \times n)$ -Matrix $e^{\mathbf{A}}$, die sich aus

der Reihe

$$e^{\mathbf{A}} = \sum_{k=0}^{\infty} \frac{1}{k!} (\mathbf{A})^k = \mathbf{I} + \frac{1}{1!} \mathbf{A} + \frac{1}{2!} \mathbf{A}^2 + \dots \quad (3.69)$$

ergibt, *Matrixexponential*. Die Matrix $e^{\mathbf{A}t}$ wird *Transitionsmatrix* genannt.

Hier ist $e^{\mathbf{A}}$ ebenfalls eine $(n \times n)$ -Matrix, weil Gl.(3.69) eine Summe von $(n \times n)$ -Matrizen darstellt. Außerdem kommutieren aufgrund der Reihendarstellungen \mathbf{A} und $e^{\mathbf{A}}$ miteinander. Die Reihe Gl.(3.69) konvergiert unter analogen Voraussetzungen und es gibt sich mit wortgleichem Beweis wie zuvor die Berechnungsvorschrift

$$\frac{d}{dt} e^{\mathbf{A}t} = \mathbf{A} \cdot e^{\mathbf{A}t} = e^{\mathbf{A}t} \cdot \mathbf{A} \quad . \quad (3.70)$$

Mit diesen Festlegungen führt die entsprechende Lösung der Zustandsgleichung auf

$$\mathbf{x}(t) = e^{\mathbf{A}t} \cdot \mathbf{x}_0 \quad . \quad (3.71)$$

Diese Lösung entspricht tatsächlich genau der Lösung in Gl.(3.45) beziehungsweise Gl.(3.48). Um dies einzusehen wird zuerst der Sonderfall untersucht, in welcher \mathbf{A} eine Diagonalmatrix ist, d.h. \mathbf{A} besitzt ausschließlich auf der Diagonalen Einträge ungleich null. Diese Diagonaleinträge sind dann gleichzeitig auch die Eigenwerte λ_i dieser Matrix, die zunächst als voneinander verschieden angenommen werden. Diagonalmatrizen lassen sich dadurch potenzieren, dass man ihre Diagonaleinträge potenziert. Somit folgt für die Reihe aus Gl.(3.69)

$$e^{\mathbf{A}t} = \sum_{k=0}^{\infty} \frac{t^k}{k!} \begin{bmatrix} (\lambda_1 t)^k & 0 & & \\ 0 & (\lambda_2 t)^k & \ddots & \\ & \ddots & \ddots & 0 \\ & & 0 & (\lambda_n t)^k \end{bmatrix} = \begin{bmatrix} e^{\lambda_1 t} & 0 & & \\ 0 & e^{\lambda_2 t} & \ddots & \\ & \ddots & \ddots & 0 \\ & & 0 & e^{\lambda_n t} \end{bmatrix} \quad . \quad (3.72)$$

Die allgemeine Lösung entspricht damit genau der Lösung in Gl.(3.45) als Zusammensetzung von e -Funktionen mit den Eigenwerten λ_i als Exponen-

ten. Dies gilt zunächst nur für den Fall von Diagonalmatrizen. Allerdings lassen sich alle diagonalisierbaren Matrizen durch eine passende Wahl der Zustandsgrößen x_i in eine Diagonalmatrix überführen. Mathematisch entspricht dies einer Zustandstransformation gemäß Gl.(3.58) mit einer Transformationsmatrix, die die Eigenvektoren von \mathbf{A} als Spalten enthält. Dabei sind insbesondere alle Matrizen, die nur einfache Eigenwerte haben, diagonalisierbar. Somit gilt die vorherige Argumentation nicht nur für Diagonalmatrizen mit einfachen Eigenwerten, sondern für jede Matrix mit einfachen Eigenwerten.

Diagonale Normalform

Gegeben ist die Differentialgleichung eines LTI-Systems wie in Gl.(2.10)

$$y^{(n)} + a_{n-1}y^{(n-1)} + \dots + a_0y = b_0u + \dots + b_{n-1}u^{(n-1)}. \quad (3.73)$$

Das charakteristische Polynom des autonomen Systems $u = 0$ besitze einfache Wurzeln λ_i . Dann kann der Zustand \mathbf{x} so gewählt werden, dass die Zustandsraumdarstellung die folgende Form besitzt:

$$\dot{\mathbf{x}} = \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & & 0 \\ 0 & 0 & \lambda_3 & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_n \end{bmatrix} \cdot \mathbf{x} + \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \cdot u$$

$$y = [r_1 \ r_2 \ r_3 \ \dots \ r_n] \cdot \mathbf{x} . \quad (3.74)$$

Aufgrund der Diagonalstruktur zerfällt das System in n parallele, ungekoppelte Übertragungskanäle, die vollständig separat voneinander betrachtet werden können. Jeder Übertragungskanal ist dabei einem Lösungsanteil $C \cdot e^{\lambda t}$ zugeordnet. Es gilt $r_i \neq 0$ für alle $i = 1, \dots, n$, da ansonsten ein Übertragungskanal keinen Beitrag zum Übertragungsverhalten leisten würde und keine minimale Realisierung vorläge. Diese besondere Struktur zeigt sich auch im zugehörigen Wirkungsplan in Bild 3-6.

Die Einträge der Matrizen in Diagonaler Normalform sind dabei im Falle konjugiert komplexer Polpaare ebenfalls komplex. Auch die Koeffizienten r_i und die Zustände x_i können in diesem Fall komplexe Werte annehmen. Das

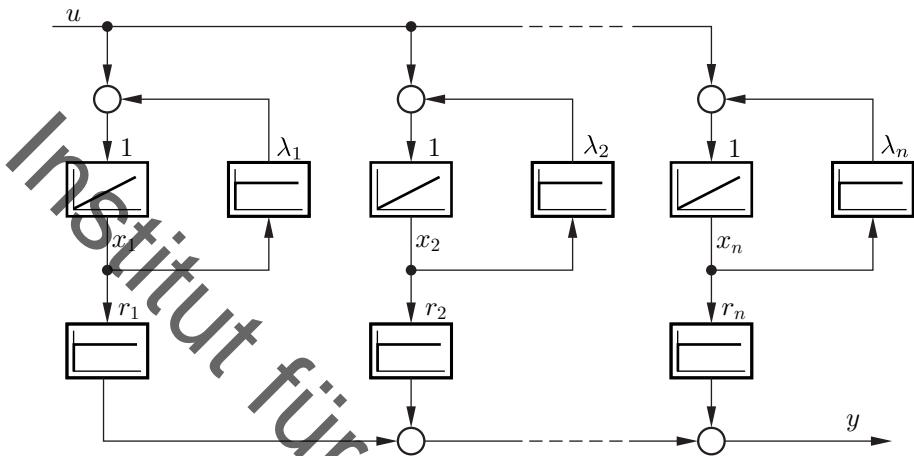


Bild 3-6: Diagonale Normalform mit einfachen und reellen Eigenwerten

ist im Allgemeinen unerwünscht und kann durch die Einführung passender 2×2 -Unterblöcke innerhalb der diagonalen Struktur vermieden werden. Hierzu ersetzt man die komplex konjugierten Eigenwerte $\lambda_i = \sigma + j\omega$ und $\lambda_j = \bar{\lambda}_i = \sigma - j\omega$ wie folgt:

$$\begin{bmatrix} \sigma + j\omega & 0 \\ 0 & \sigma - j\omega \end{bmatrix} \Rightarrow \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix} . \quad (3.75)$$

Man kann nachrechnen, dass beide Matrizen dieselben Eigenwerte haben. Hierdurch werden die beiden Zustände x_i und x_j zwar miteinander verkoppelt, aber bis auf diese 2×2 -Unterblöcke behält die Matrix ihre Diagonalstruktur. Dieses Vorgehen kann analog zum Ersetzen der komplexen Exponenten der e -Funktion durch cos- und sin-Funktionen in Gl.(3.46) gedeutet werden.

Ähnlich zu der Fallunterscheidung von einfachen und mehrfachen Wurzeln des charakteristischen Polynoms in Gl.(3.45) und Gl.(3.48) muss neben dem diskutierten Fall von Matrizen mit unterschiedlichen Eigenwerten auch der Fall von Matrizen mit mehrfachen Eigenwerten betrachtet werden. Matrizen, die trotz mehrfacher Eigenwerte diagonalisierbar sind, können ausgeschlossen werden, da dann in der Diagonalen Normalform redundante Übertragungskanäle vorliegen und es sich somit um keine minimale Realisierung

handeln würde. Matrizen, die aufgrund mehrfacher Eigenwerte nicht diagonalisierbar sind, lassen sich zwar nicht in die Form Gl.(3.74), zumindest aber in eine diagonalähnliche Form bringen. Diese sogenannte *Jordansche⁴ Normalform* ist für alle in der Regelungstechnik auftretenden Matrizen möglich. Hierzu ergänzt man die Diagonalmatrix im Falle mehrfacher Eigenwerte um Einsen auf der ersten Nebendiagonalen:

Jordansche Normalform

Gegeben ist die Differentialgleichung eines LTI-Systems wie in Gl.(2.10)

$$y^{(n)} + a_{n-1}y^{(n-1)} + \dots + a_0y = b_0u + \dots + b_{n-1}u^{(n-1)}. \quad (3.76)$$

Das charakteristische Polynom des autonomen Systems $u = 0$ besitze eine k -fache Wurzel bei λ_0 sowie einfache Wurzeln $\lambda_1, \dots, \lambda_{n-k}$. Dann kann der Zustand \mathbf{x} so gewählt werden, dass die Zustandsraumdarstellung die folgende Form besitzt.

$$\dot{\mathbf{x}} = \underbrace{\begin{bmatrix} \lambda_0 & 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & 0 & & & \vdots \\ \vdots & & \ddots & 1 & & & \vdots \\ \vdots & & & \lambda_0 & 0 & & \vdots \\ \vdots & & & & \lambda_1 & 0 & \vdots \\ \vdots & & & & & \ddots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 & \lambda_{n-k} \end{bmatrix}}_{\begin{matrix} k \\ n-k \end{matrix}} \cdot \mathbf{x} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \frac{1}{1} \\ \vdots \\ 1 \end{bmatrix} \cdot u \quad (3.77)$$

$$y = [r_1 \ r_2 \ r_3 \ \dots \ r_n] \cdot \mathbf{x} .$$

Für den Fall von einfachen Eigenwerten entspricht die Jordansche Normalform genau der Diagonalen Normalform mit den Eigenwerten auf der Diagonale. Ansonsten nimmt die Systemmatrix in Gl.(3.77) eine Blockdiagonalfom an. Im Falle verschiedener mehrfacher Eigenwerte ergibt sich die Jordansche Normalform entsprechend durch Hinzufügen von Einsen auf der ersten Nebendiagonalen für alle mehrfachen Eigenwerte.

⁴Camille Jordan (1838-1922), französischer Mathematiker [23]

Die Jordansche Normalform ist neben der Regelungnormalform in Gl.(2.25) die wichtigste Normalform der Regelungstechnik. Während bei der Regelungnormalform sich das charakteristische Polynom direkt in der letzten Zeile der Systemmatrix \mathbf{A} ablesen lässt, finden sich bei der Jordanschen Normalform direkt die gesuchten Eigenwerte auf der Diagonalen. Im Falle komplexer Eigenwerte kann analog zur Diagonalen Normalform vorgegangen werden, um die Einträge reellwertig zu halten. Die Jordansche Normalform ermöglicht zudem die Ermittlung der Lösung im Fall mehrfacher Eigenwerte wie in Gl.(3.48). Man kann dabei berechnen, dass die Einsen auf der Nebendiagonalen bei der Berechnung der Transitionsmatrix zu Polynomen in t führen, die ganz analog zu Gl.(3.48) gebildet werden. Somit ergeben sich bei der allgemeinen Lösung im Zustandsraum die gleichen Resultate wie bei der Lösung der Differentialgleichung.

Als abschließendes Beispiel für die Jordansche Normalform wird die Differentialgleichung

$$\ddot{y} = u \quad y(t=0) = {}_0y \quad , \quad \dot{y}(t=0) = {}_0\dot{y} \quad (3.78)$$

betrachtet, die auch Doppelintegrator genannt wird, da der Eingang zur Bestimmung des Ausgangs zweimal integriert wird. Für $u = 0$ ergibt sich ein autonomes System mit dem charakteristischen Polynom

$$p(\lambda) = \lambda^2 = 0 \quad \Rightarrow \quad \lambda_{1,2} = 0. \quad (3.79)$$

Das System besitzt also einen doppelten Eigenwert bei $\lambda = 0$ und ist somit instabil. Die allgemeine Lösung ist

$$y(t) = (C_1 + C_2 t) \cdot e^{0t} = C_1 + C_2 t = {}_0y + {}_0\dot{y} \cdot t \quad , \quad (3.80)$$

eine Rampe mit Steigung ${}_0\dot{y}$. Die Lösung divergiert wie zu erwarten selbst für kleinste ${}_0\dot{y}$. Eine Analyse im Zustandsraum führt auf das identische Ergebnis. Wegen des doppelten Eigenwertes kann nicht die Diagonale Normalform gewählt werden, sondern muss auf die Jordansche Normalform zurückgegriffen werden. Diese entspricht in diesem Beispiel zufälligerweise auch der Regelungnormalform mit $x_1 = y$ und $x_2 = \dot{y}$:

$$\dot{\mathbf{x}} = \underbrace{\begin{bmatrix} \lambda_1 = 0 & 1 \\ 0 & \lambda_2 = 0 \end{bmatrix}}_{\mathbf{A}} \mathbf{x} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u \quad , \quad y = [1 \quad 0] \mathbf{x} \quad . \quad (3.81)$$

Zur Bestimmung der Transitionsmatrix berechnet man

$$\mathbf{A}^2 = \mathbf{A} \cdot \mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \mathbf{0} \quad . \quad (3.82)$$

Somit ist $\mathbf{A}^k = \mathbf{0}$ für $k \geq 2$ und die unendliche Reihe der Transitionsmatrix besteht nur aus zwei Gliedern

$$\mathbf{x}(t) = e^{\mathbf{A}t} \mathbf{x}_0 = (\mathbf{I} + \mathbf{A}t) \mathbf{x}_0 = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \mathbf{x}_0 = \begin{bmatrix} {}_0y + {}_0\dot{y} \cdot t \\ {}_0\dot{y} \end{bmatrix} \quad (3.83)$$

mit der identischen Lösung wie in Gl.(3.80).

4 Verhalten bei allgemeiner Anregung

4.1 Homogene und partikuläre Lösung

Das vorherige Kapitel hatte durch die Betrachtung autonomer Systeme geklärt, wie sich die Anfangsbedingungen ${}_0\boldsymbol{x}$ auf die Lösung einer Differentialgleichung auswirken. Möchte man zusätzlich auch die Auswirkungen von Eingangsgrößen \boldsymbol{u} beurteilen, so kann man die Gesamtlösung aufgrund des Überlagerungsprinzips linearer Systeme aus zwei Anteilen zusammensetzen: den einen für die Auswirkungen der Anfangsbedingung und den anderen für die Auswirkungen von \boldsymbol{u} . Zur Darstellung wird dabei aufgrund ihrer Kompaktheit auf die Zustandsraumdarstellung zurückgegriffen.

Homogene Lösung

Genügt die Funktion $\boldsymbol{x}_h(t)$ der Gleichung

$$\dot{\boldsymbol{x}}_h = \mathbf{A} \cdot \boldsymbol{x}_h + \mathbf{B} \cdot \mathbf{0} \quad , \quad \boldsymbol{x}(t=0) = {}_0\boldsymbol{x} \quad , \quad (4.1)$$

so heißt $\boldsymbol{x}_h(t)$ *homogene Lösung* der Differentialgleichung

$$\dot{\boldsymbol{x}} = \mathbf{A} \cdot \boldsymbol{x} + \mathbf{B} \cdot \boldsymbol{u} \quad , \quad \boldsymbol{x}(t=0) = {}_0\boldsymbol{x} \quad . \quad (4.2)$$

Die homogene Lösung beschreibt die Auswirkungen der Anfangsbedingung ${}_0\boldsymbol{x}$ für $\boldsymbol{u} = \mathbf{0}$ und entspricht damit der Lösung einer autonomen Differentialgleichung.

Partikuläre Lösung

Genügt die Funktion $\boldsymbol{x}_p(t)$ der Gleichung

$$\dot{\boldsymbol{x}}_p = \mathbf{A} \cdot \boldsymbol{x}_p + \mathbf{B} \cdot \boldsymbol{u} \quad , \quad \boldsymbol{x}(t=0) = \mathbf{0} \quad , \quad (4.3)$$

so heißt $\boldsymbol{x}_p(t)$ *partikuläre Lösung* der Differentialgleichung

$$\dot{\boldsymbol{x}} = \mathbf{A} \cdot \boldsymbol{x} + \mathbf{B} \cdot \boldsymbol{u} \quad , \quad \boldsymbol{x}(t=0) = {}_0\boldsymbol{x} \quad . \quad (4.4)$$

Die partikuläre Lösung erfasst die Auswirkungen von \boldsymbol{u} , setzt aber die Anfangsbedingungen zu null. Eine einfache Rechnung zeigt, dass die Summe

$\dot{x}_h + x_p$ die Gesamtlösung der Differentialgleichung ergibt: Einsetzen von $x_h + x_p$ liefert

$$\dot{x}_h + \dot{x}_p = \mathbf{A} \cdot x_h + \mathbf{A} \cdot x_p + \mathbf{B} \cdot u = \mathbf{A} \cdot (x_h + x_p) + \mathbf{B} \cdot u \quad (4.5)$$

sowie für die Anfangsbedingung

$$x_h(t=0) + x_p(t=0) = {}_0x + {}_0\mathbf{0} = {}_0x \quad , \quad (4.6)$$

womit $x_h + x_p$ die Lösung der Differentialgleichung

$$\dot{x} = \mathbf{A} \cdot x + \mathbf{B} \cdot u \quad , \quad x(t=0) = {}_0x \quad (4.7)$$

ist. Die homogene Lösung ist aus den Betrachtungen zu autonomen Systemen bereits bekannt, womit die partikuläre Lösung als zu bestimmender Lösungsanteil verbleibt. Daher ist es bei Bedarf zulässig, für die Analyse von Systemen bei Anregung die Anfangsbedingung zu null zu setzen, ohne die gewonnenen Resultate einzuschränken. Außerdem kann ohne Beschränkung der Allgemeinheit die Auswirkung einer einzelnen Eingangsgröße $u(t)$ untersucht werden, da aufgrund des Überlagerungsprinzips sich bei mehreren Eingangsgrößen die Lösung erneut als Addition der Auswirkungen der einzelnen $u_i(t)$ ergibt.

Zur Bestimmung der partikulären Lösung wird zunächst von der einfachen Differentialgleichung analog zu Gl.(3.36)

$$\dot{x} = ax + bu \quad , \quad x(t=0) = 0 \quad (4.8)$$

ausgegangen. Im Gegensatz zu Gl.(3.36) wurde die Anfangsbedingung zu null gesetzt und hierfür der Term bu ergänzt. Aufbauend auf der homogenen Lösung

$$x_h(t) = Ce^{at} \quad (4.9)$$

lässt sich die partikuläre Lösung über die Variation der Konstanten mit dem Ansatz

$$x_p(t) = C(t)e^{at} \quad \Rightarrow \quad \dot{x}_p(t) = \dot{C}(t)e^{at} + C(t)ae^{at} \quad (4.10)$$

finden. Eingesetzt in die Differentialgleichung Gl.(4.8) ergibt sich

$$(\dot{C}(t) + C(t)a - aC(t))e^{at} = bu(t) \quad . \quad (4.11)$$

Die unbekannte Funktion $C(t)$ bestimmt man aus Gl.(4.11) mit

$$\dot{C}(t) = b e^{-at} u(t) \quad (4.12)$$

durch Integration über der Zeit von 0 bis t zu

$$C(t) = C_0 + b \int_{-0}^t e^{-a\tau} u(\tau) d\tau \quad (4.13)$$

mit der Integrationskonstanten

$$C_0 = C(t=0) = x(t=0) = 0 \quad . \quad (4.14)$$

Die Notation der unteren Integrationsgrenze -0 bedeutet hier, dass eine bei $t = 0$ möglicherweise auftretende Unstetigkeit in die Integration einbezogen wird. Solche Unstetigkeiten können bei unstetigen Eingangssignalen $u(t)$, die sich in $t = 0$ bereits ändern, auftreten.

Die Gesamtlösung der Differentialgleichung ergibt sich als Summe der homogenen und partikulären Lösung zu

$$x(t) = \underbrace{{}_0 x \cdot e^{at}}_{\text{homogen}} + \underbrace{\int_{-0}^t e^{a(t-\tau)} bu(\tau) d\tau}_{\text{partikulär}} \quad . \quad (4.15)$$

Die Lösung der Zustandsdifferentialgleichung

$$\dot{x} = \mathbf{A} x + \mathbf{B} u \quad (4.16)$$

lässt sich in analoger Weise durch Verwendung der Transitionsmatrix gewinnen, wobei sich

$$x(t) = e^{\mathbf{A}t} {}_0 x + \int_{-0}^t e^{\mathbf{A}(t-\tau)} \mathbf{B} u(\tau) d\tau \quad (4.17)$$

ergibt. Die allgemeine Lösung für die Ausgangsgröße y lässt sich mit

$$y = \mathbf{C} \cdot x + \mathbf{D} \cdot u \quad (4.18)$$

aus Gl.(4.17) gewinnen zu

$$\mathbf{y}(t) = \mathbf{C} \cdot e^{\mathbf{A}t} \cdot {}_0\mathbf{x} + \int_{-0}^t \mathbf{C} \cdot e^{\mathbf{A}(t-\tau)} \cdot \mathbf{B} \cdot \mathbf{u}(\tau) d\tau + \mathbf{D} \cdot \mathbf{u}(t) \quad . \quad (4.19)$$

4.2 Übergangsfunktion

Die homogene Lösung beschreibt die Auswirkungen von Anfangsbedingungen ${}_0\mathbf{x} \in \mathbb{R}^n$, die als Vektoren aufgefasst werden können und daher über n Freiheitsgrade verfügen. Die partikuläre Lösung hingegen enthält in der Bestimmungsgleichung die Eingangsgröße $\mathbf{u}(t)$ und damit eine Funktion als freie Variable. Hier sind potentiell unendlich viele und gänzlich verschiedene Eingangsfunktionen denkbar, wobei für einige $\mathbf{u}(t)$ die in der partikulären Lösung auftretenden Integrale schwer zu lösen sein werden. Daher ist es sinnvoll, sich auf einige wenige Eingangsfunktionen $\mathbf{u}(t)$ zu beschränken, deren Lösung vergleichsweise einfach zu ermitteln ist und über die man ein möglichst repräsentatives Bild der dynamischen Eigenschaften des Systems erhält. Aufgrund des Überlagerungsprinzips ist dabei die Betrachtung einzelner Eingangsgrößen $u(t)$ möglich. Folglich werden häufig die Systemantworten auf spezielle, genormte Eingangsfunktionen herangezogen – unabhängig davon, dass solche Verläufe u. U. nicht realisiert werden können. Die wichtigsten Standardfunktionen sind die *Sprungfunktion*, die bei den Ausführungen zum Wirkungsplan in Abschnitt 2.4 bereits aufgetreten ist, sowie die *Impulsfunktion* (Bild 4-1).

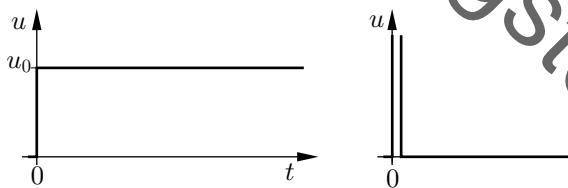


Bild 4-1: Sprung- und Impulsfunktion

Die Sprungfunktion ist dadurch gekennzeichnet, dass sie zum Zeitpunkt null um einen endlichen Wert springt. Üblicherweise wählt man beim Rechnen mit Abweichungsgrößen den Arbeitspunkt so, dass die Abweichungsgröße von null aus auf einen endlichen positiven Wert springt. Ferner setzt man

voraus, dass das betrachtete System vor dem Sprung in Ruhe war, da nur der partikuläre Lösungsanteil interessiert. Für die Impulsfunktion gilt Ähnliches mit der Ausnahme, dass die Impulsfunktion theoretisch die Amplitude unendlich und die Breite null hat. Die Fläche unter der Impulsfunktion ist endlich und ist ein Maß für die Intensität des Impulses. Diese Definition entspricht nicht den gängigen Vorstellungen von Funktionen, weswegen der Impuls auch als Pseudofunktion bezeichnet wird. Praktisch wird die Impulsfunktion durch Signalverläufe von genügend kurzer Dauer und realisierbarer Amplitude angenähert.

Sprungantwort, Impulsantwort, Übergangsfunktion, Gewichtsfunktion

Die sich als Folge einer sprung- beziehungsweise impulsförmigen Eingangsgröße ergebenden Verläufe der Ausgangsgröße werden *Sprungantwort* beziehungsweise *Impulsantwort* genannt. Durch Normieren auf die Sprunghöhe u_0 beziehungsweise die Impulsfläche $\int u dt$ erhält man die *Übergangsfunktion* $h(t)$ beziehungsweise die *Gewichtsfunktion* $g(t)$.

$$h(t) = \frac{y(t)}{u_0} = \frac{\text{Sprungantwort}}{\text{Sprunghöhe}} \quad (4.20)$$

$$g(t) = \frac{y(t)}{\int u dt} = \frac{\text{Impulsantwort}}{\text{Impulsfläche}} \quad (4.21)$$

Alternativ zur Normierung der Sprungantwort beziehungsweise Impulsantwort kann man aufgrund des Verstärkungsprinzips für lineare Systeme die Übergangsfunktion beziehungsweise Gewichtsfunktion auch dadurch erhalten, dass man die zugehörige Differentialgleichung für den (dimensionslosen) Einheitssprung

$$1(t) = \begin{cases} 1 & \text{für } t \geq 0 \\ 0 & \text{für } t < 0 \end{cases} \quad (4.22)$$

mit der Höhe eins beziehungsweise Einheitsimpuls $\delta(t)$ mit der Fläche eins löst.

Zur Beurteilung der Eigenschaften von Systemen wird hauptsächlich mit

der Übergangsfunktion gearbeitet werden, deren Piktogramme beispielsweise zur Darstellungen von Wirkungsplänen genutzt werden. Die Gewichtsfunktion wird vorwiegend für mehr theoretische Überlegungen eingesetzt. Als Beispiel für den praktischen Einsatz einer Sprungantwort wird angenommen, dass ein Autofahrer das Beschleunigungsverhalten eines Fahrzeugs untersuchen möchte. Eine Möglichkeit besteht darin, das Gaspedal schlagartig durchzudrücken und das Beschleunigungsverhalten des Fahrzeugs zu beobachten. Die plötzliche Änderung des Gaspedals als Eingangsgröße entspricht näherungsweise einem Sprung und das Beschleunigungsverhalten über der Zeit ist dann die Sprungantwort. Soll hingegen herausgefunden werden, welche Eigenschwingungen ein gefertigtes Bauteil besitzt, so besteht ein gängiger Ansatz darin, das Werkstück über einen speziellen Hammer mit Stößen zu versehen und die entstehenden Schwingungen messtechnisch zu erfassen. Die schlagartige Anregung durch den Hammer entspricht strukturell einem Impuls und das Schwingverhalten des Bauteils somit einer Impulsantwort.

Mit Überlegungen analog zu Gl.(2.23) kann man Übergangs- und Gewichtsfunktion direkt miteinander ins Verhältnis setzen. Da nämlich der Einheitsprung durch Integration des Einheitsimpulses über der Zeit entsteht

$$1(t) = \int_{-0}^t \delta(\tau) d\tau , \quad (4.23)$$

gilt für lineare Systeme, dass auch die Übergangsfunktion $h(t)$ durch Integration der Gewichtsfunktion bestimmbar ist:

$$h(t) = \int_{-0}^t g(\tau) d\tau . \quad (4.24)$$

Die Umkehrung

$$g(t) = \frac{dh(t)}{dt} \quad (4.25)$$

gilt mit Einschränkungen hinsichtlich der Differenzierbarkeit der Übergangsfunktion.

Aufgrund der einfachen Form des Eingangssignals $u(t)$ für die Übergangsfunktion, lässt sich das in Gl.(4.19) auftretende Integral rasch lösen. Es ergibt sich im SISO-Fall mit $d = 0$, ${}_0\mathbf{x} = 0$ und $u(t) = 1(t)$ mithilfe der Substitution $\vartheta = t - \tau$

$$h(t) = \int_0^t \mathbf{c}^T \cdot e^{\mathbf{A}(t-\tau)} \cdot \mathbf{b} \cdot 1(\tau) d\tau = \int_{-0}^t \mathbf{c}^T \cdot e^{\mathbf{A}\vartheta} \cdot \mathbf{b} d\vartheta \quad . \quad (4.26)$$

Mit Gl.(4.26) lässt sich auch direkt die Gewichtsfunktion bestimmen, da für diese als Ableitung der Übergangsfunktion

$$g(t) = \dot{h}(t) = \mathbf{c}^T \cdot e^{\mathbf{A}t} \cdot \mathbf{b} \quad (4.27)$$

folgt, da Integration und Differentiation sich gegenseitig aufheben.

4.3 Faltung

Bei genauer Betrachtung von Gl.(4.27) stellt man fest, dass $g(t)$ identisch mit einem Teilausdruck der allgemeinen Lösung in Gl.(4.19) ist. Die partikuläre Lösung lässt sich folglich zu

$$y_p(t) = \int_{-0}^t g(t - \tau) \cdot u(\tau) d\tau \quad (4.28)$$

umschreiben. Der Zusammenhang in Gl.(4.28) wird auch *Faltung* oder *Faltungintegral* genannt.

Faltung

Seien $a(t)$ und $b(t)$ zwei Signale mit $a(t) = b(t) = 0$ für $t < 0$. Dann ist die *Faltung* von $a(t)$ und $b(t)$ definiert als

$$a(t) * b(t) = \int_{-0}^t a(t - \tau) \cdot b(\tau) d\tau = b(t) * a(t) \quad . \quad (4.29)$$

Die Faltung ist kommutativ, was man mit derselben Substitution wie in Gl.(4.26) nachweisen kann. Somit besitzt die Faltung ähnliche Eigenschaften wie die Multiplikation von Zahlen.

ten wie ein Produkt. Das hat zu der Schreibweise $y_p(t) = g(t) * u(t) = u(t) * g(t)$ geführt.

Mithilfe der Faltung lässt sich die partikuläre Lösung eines linearen Systems unter ausschließlich der Zuhilfenahme der Gewichtsfunktion $g(t)$ für potentiell beliebige Eingangsfunktionen $u(t)$ berechnen. Dabei bietet die Faltung neben einem analytischen Zugang, der bei komplizierten $u(t)$ oft an seine Grenzen stößt, auch einen graphischen Ansatz.

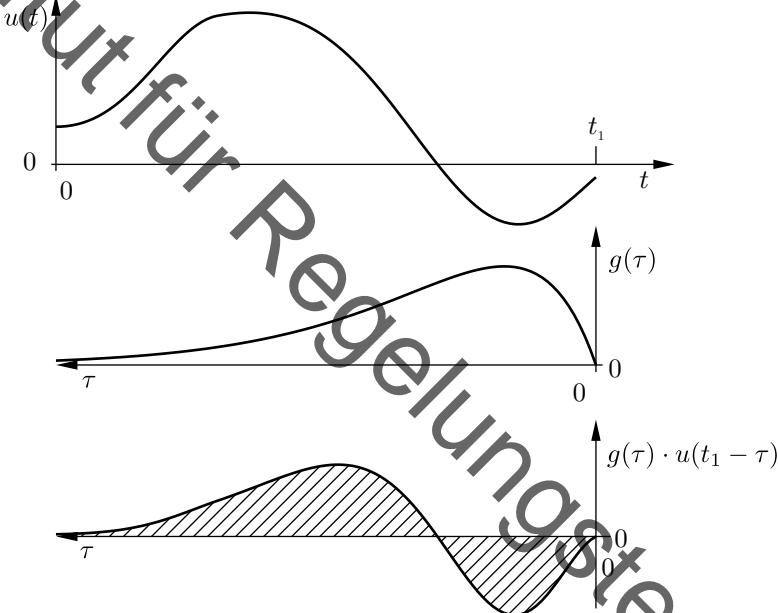


Bild 4-2: Faltung

In Bild 4-2 ist hierzu die Rechenvorschrift der Faltung visualisiert. Der Wert der Ausgangsgröße zu einem bestimmten Zeitpunkt $t = t_1$ wird dadurch bestimmt, dass die Werte der Eingangsgröße $u(t)$ zu allen diesem Zeitpunkt vorausgehenden Zeitpunkten mit der Gewichtsfunktion $g(t)$ multipliziert und somit gewichtet werden, woher auch die Bezeichnung „Gewichtsfunktion“ stammt. Dieses Ergebnis wird anschließend integriert. Der Wert der Ausgangsgröße zum Zeitpunkt t_1 wird daher durch die schraffierte Fläche

unter der Kurve $g(\tau) \cdot u(t_1 - \tau)$ dargestellt. Diese Berechnungsmethode gestaltet sich insgesamt sehr aufwändig, da beide Signale schrittweise gegeneinander verschoben werden und die sich ergebenen Flächen ermittelt werden müssen. Die Tatsache, dass die beiden Signale wegen der Funktionsargumente $g(\tau)$ und $u(t - \tau)$ gegenläufige Zeitachsen haben und somit eines der Signale „umgeklappt“ werden muss, hat dabei zu der Bezeichnung als „Faltung“ geführt. Die Berechnung lässt sich durch Rechnerunterstützung zwar gut systematisieren. Dennoch möchte man es üblicherweise vermeiden, die komplizierte Faltung zweier Signale $g(t) * u(t)$ ausrechnen zu müssen.

Auch ohne die explizite Berechnung der Faltung können über ihre Berechnungsvorschrift wichtige Aussagen über das dynamische Verhalten linearer Systeme gewonnen werden – insbesondere auch zur Stabilität. Stellt man die Berechnungsformel für $g(t)$ in Gl.(4.27) der homogenen Lösung gegenüber

$$g(t) = \dot{h}(t) = \mathbf{c}^T \cdot e^{\mathbf{A}t} \cdot \mathbf{b} \quad \text{und} \quad y_h(t) = \mathbf{c}^T \cdot e^{\mathbf{A}t} \cdot {}_0\mathbf{x} \quad (4.30)$$

so erkennt man direkt, dass beide Lösungen identisch aufgebaut sind. Wenn die homogene Lösung nur Terme mit abklingenden Exponentialfunktionen aufweist, so wird auch $g(t)$ dieses Verhalten zeigen und exponentiell abklingen. Für ein stabiles System läuft die Gewichtsfunktion also gegen null. Das Abklingverhalten von $g(t)$ ist wegen der negativen Exponenten der e -Funktion dabei so schnell, dass auch die Fläche unter $g(t)$ beschränkt bleibt. Die Übergangsfunktion konvergiert dann nach Gl.(4.24) gegen einen endlichen Wert:

$$h(t) = \int_{-0}^t g(\tau) d\tau \xrightarrow{t \rightarrow \infty} h_\infty < \infty \quad (4.31)$$

Ein stabiles System reagiert also auf den beschränkten Einheitssprung mit einer gleichfalls beschränkten Ausgangsgröße. Diese Aussage lässt sich durch einige Abschätzungen auf beliebige beschränkte Eingangsgrößen erweitern, wobei es ausreicht, den partikulären Lösungsanteil zu betrachten, da der homogene gegen null läuft. Ausgangspunkt ist die Faltung. Mit den Beträgen von Eingangs- und Ausgangsgröße sowie der Gewichtsfunktion wird aus

dieser die Ungleichung

$$|y(t)| = \left| \int_0^t u(t-\tau) \cdot g(\tau) d\tau \right| \leq \int_0^t |u(t-\tau)| \cdot |g(\tau)| d\tau \quad . \quad (4.32)$$

Wenn die Eingangsgröße beschränkt ist

$$|u(t)| \leq M \quad , \quad M < \infty \quad (4.33)$$

wird daraus

$$|y(t)| \leq M \int_0^t |g(t-\tau)| d\tau = M \cdot \int_0^t |g(\vartheta)| d\vartheta \quad . \quad (4.34)$$

Die Ausgangsgröße ist also genau dann beschränkt, wenn die Gewichtsfunktion absolut integrierbar ist. Aufgrund des Überlagerungsprinzips gelten identische Ausführungen für den Mehrgrößenfall mit den vektorwertigen Impulsantworten $\mathbf{g}_i(t)$ für die einzelnen Eingänge $u_i(t)$. Die Eigenschaft, dass stabile Systeme beschränkte Eingangsgrößen auf beschränkte Ausgangsgrößen abbilden, hat zu der folgenden Definition geführt.

BIBO-Stabilität

Ein lineares System $u \mapsto y$ heißt übertragungsstabil oder auch BIBO-stabil (von Englisch Bounded Input Bounded Output), wenn für jedes beschränkte Eingangssignal

$$|u(t)| \leq M_u < \infty \quad (4.35)$$

auch das Ausgangssignal beschränkt ist:

$$|y(t)| \leq M_y < \infty \quad . \quad (4.36)$$

Ein lineares System ist genau dann BIBO-stabil, wenn die Impulsantwort $g(t)$ absolut integrierbar ist.

Zu beachten ist, dass diese Definition im Gegensatz zu Gl.(3.11) nur für lineare Systeme gültig ist. Zudem kann es passieren, dass in den Signalen

$u(t)$ oder $y(t)$ Impulse $\delta(t)$ auftreten. Das ist für $y(t)$ insbesondere bei akausalen Systemen der Fall. Da diese Impulse eine endliche Fläche unter sich besitzen, werden diese im Rahmen der BIBO-Stabilität ebenfalls als beschränkt angesehen.

Offenbar ist jedes lineare stabile System auch BIBO-stabil, da die Impulsantwort für stabile Systeme absolut integrierbar ist. Doch gilt auch die Umkehrung? In Gl.(4.30) übernimmt der Eingangsvektor \mathbf{b} die Rolle der Anfangsbedingung ${}_0\mathbf{x}$. Somit ist nicht klar, ob durch \mathbf{b} auch alle Lösungskomponenten des Systems angeregt werden, da die Definition für Stabilität Konvergenz für alle möglichen Anfangsbedingungen fordert.

Es zeigt sich, dass die Umkehrung gilt, sofern es sich bei dem System um eine minimale Realisierung handelt. Hierzu wird ein BIBO-stabiles System betrachtet, das also ein absolut integrierbares $g(t)$ besitzt. Aus Gründen der Übersichtlichkeit wird angenommen, dass alle Eigenwerte einfach sind. Da das System eine minimale Realisierung ist, kann ohne Beschränkung der Allgemeinheit die Diagonale Normalform zur Darstellung im Zustandsraum genutzt werden:

$$g(t) = \mathbf{c}^T \cdot e^{\mathbf{A}t} \cdot \mathbf{b} = \underbrace{\begin{bmatrix} r_1 & \dots & r_n \end{bmatrix}}_{r_i \neq 0} \cdot \begin{bmatrix} e^{\lambda_1 t} & & \\ & \ddots & \\ & & e^{\lambda_n t} \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \quad (4.37)$$

Hier wird wegen $\mathbf{b} = [1 \ \dots \ 1]^T$ jeder Übertragungskanal $e^{\lambda_i t}$ angeregt und jeder angeregte Kanal wegen $r_i \neq 0$ in die Ausgangsgröße überführt. Da alle Eigenwerte verschieden sind, können sich diese Kanäle nicht gegenseitig kompensieren. Also ist jeder Übertragungskanal in $y(t)$ sichtbar. Da $g(t)$ absolut integrierbar ist, müssen alle λ_i negativen Realteil besitzen. Eine analoge Betrachtung ist auch für den allgemeinen Fall mithilfe der Jordanschen Normalform möglich.

BIBO-Stabilität und Stabilität

Für jede minimale Realisierung sind BIBO-Stabilität und (asymptotische) Stabilität gleichbedeutend. Alle stabilen Systeme reagieren auf beschränkte Eingangsgrößen mit beschränkten Ausgangsgrößen. Jedes System, das eine minimale Realisierung ist und auf beschränkte Eingangsgrößen mit beschränkten Ausgangsgrößen antwortet, ist stabil.

Handelt es sich nicht um eine minimale Realisierung, kann es sein, dass $g(t)$ absolut integrierbar ist, es aber Lösungsanteile $e^{\lambda t}$ gibt, die zwar instabil sind, jedoch keinen Beitrag zu $g(t)$ leisten (vgl. Bild 2-9). Diese Beobachtung ist ein weiteres Argument, das Gesetz der Sparsamkeit zu befolgen. Da stets von minimalen Realisierungen ausgegangen wird, werden die Begriffe „stabil“ und „BIBO-stabil“ gleichwertig verwendet.

Die BIBO-Stabilität ermöglicht neben der Prüfung der Wurzeln des charakteristischen Polynoms einen weiteren Zugang zur Stabilitätsprüfung, der auch experimentell ablaufen kann. Aufgrund der Faltung ist nämlich ein System genau dann stabil, wenn die Impulsantwort absolut integrierbar ist. Es müssen folglich nicht alle möglichen Eingangssignale $u(t)$ untersucht werden, sondern es reicht die Prüfung genau eines Eingangssignales aus – des Impulses. Da dieser experimentell schwierig umzusetzen ist, wird meist für eine experimentelle Untersuchung die Sprungantwort aufgezeichnet. Wegen Gl.(4.31) ist Stabilität gleichbedeutend damit, dass die Sprungantwort gegen einen endlichen Wert konvergiert. Als Stabilitätstest regt man folglich das System mit einem realen Sprung der Einheitsgröße an und zeichnet den Verlauf der Ausgangsgröße auf. Konvergiert dieser gegen einen stationären Endwert, so ist das System stabil. Divergiert das System oder stellen sich Dauerschwingungen ein, so ist das System instabil oder grenzstabil.

Statische Verstärkung

Für ein stabiles System wird der existente stationäre Endwert K der Übergangsfunktion

$$\lim_{t \rightarrow \infty} h(t) = K \quad (4.38)$$

als *statische Verstärkung* (englisch: „steady-state gain“) bezeichnet. Gemäß der Definition der Übergangsfunktion ist die Einheit der statischen Verstärkung $[y]/[u]$. Alternative Bezeichnungen sind statischer Übertragungsfaktor, statischer Übertragungsbeiwert, statischer Verstärkungsfaktor oder diese Bezeichnungen in Kombination mit dem Adjektiv „stationär“ anstelle von „statisch“.

Wichtig ist hier zu betonen, dass die Beschränktheit der Sprungantwort $h(t)$ nicht ausreichend ist, sondern zusätzlich auch die Konvergenz gegen einen endlichen Endwert gefordert werden muss. So würde beispielsweise

für $h(t) = \sin(t)$, was beschränkt aber nicht konvergent ist, entsprechend $g(t) = \dot{h}(t) = \cos(t)$ folgen. Die Fläche unter der Kurve $\cos(t)$ ist aber unbeschränkt.

4.4 Laplace-Transformation

Die Faltung und die Zerlegung in einen homogenen und einen partikulären Lösungsanteil ermöglicht die Lösung einer linearen Differentialgleichung mit konstanten Koeffizienten. Abseits von Sonderfällen wie Sprüngen oder Impulsen ist das Ausrechnen der Faltung aber mit erheblichen Aufwand verbunden und sollte nach Möglichkeit vermieden werden. Zur Berechnung der Systemantwort gibt es eine viel effizientere Möglichkeit als die Faltung – die Laplace¹-Transformation.

4.4.1 Laplace-Transformation von Zeitfunktionen

Gegeben sei eine zeitabhängige Funktion $f(t)$, welche für negative Werte des Arguments t null ist. Die Laplace-Transformation ordnet einer solchen Funktion $f(t)$ im Zeitbereich eine andere Funktion $F(s)$ im sogenannten Bildbereich umkehrbar eindeutig zu. Durch die besondere Form der Transformation wird erreicht, dass die Operationen Differentiation und Integration von Zeitfunktionen in algebraische Operationen mit den zugehörigen Bildfunktionen übergehen, sodass die Abbilder von Differentialgleichungen algebraische Gleichungen ergeben, die einfacher als die Originalgleichungen umgeformt und miteinander verknüpft werden können. Da sehr viele Aufgaben ohne allzu tief gehende Kenntnis der Theorie der Laplace-Transformation mit Hilfe so genannter Korrespondenztabellen gelöst werden können, soll hier nur ein kurzer Abriss der Verfahrensweise gegeben werden. Für weitergehende Fragen sei auf die einschlägige Literatur verwiesen [14].

Laplace-Transformation

Gegeben sei eine Funktion $f(t)$ mit $f(t) = 0$ für $t < 0$ und $t \in \mathbb{R}$. Die Laplace-Transformation $\mathcal{L}\{f(t)\}$ ordnet dieser Funktion $f(t)$ im Zeitbereich eine neue Funktion $F(s)$ mit der komplexen Variable $s = \sigma + j\omega \in \mathbb{C}$

¹Pierre-Simon Laplace (1749-1827), französischer Physiker [34]

im Bildbereich über die folgende Formel zu:

$$F(s) = \int_{-0}^{\infty} f(t) \cdot e^{-st} dt = \mathcal{L}\{f(t)\} . \quad (4.39)$$

Die Laplace-Transformation ist eindeutig umkehrbar mit der inversen Laplace-Transformation

$$f(t) = \begin{cases} \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} F(s) \cdot e^{st} ds & \text{für } t \geq 0 \\ 0 & \text{für } t < 0 \end{cases} = \mathcal{L}^{-1}\{F(s)\}. \quad (4.40)$$

Darin ist α eine positive Konstante, die wie σ so groß zu wählen ist, dass die Integrale in Gl.(4.39) und Gl.(4.40) konvergieren. Abkürzend schreibt man für die Verknüpfung von der Funktion $F(s)$ im Bildbereich mit der Funktion $f(t)$ im Zeitbereich

$$F(s) \bullet \circ f(t) \quad \text{beziehungsweise} \quad f(t) \circ \bullet F(s) . \quad (4.41)$$

Aufgrund des Integrationsanfangs bei $t = -0$ können nur Signale transformiert werden, die für negative Zeiten null sind, da die Transformation sonst nicht umkehrbar eindeutig ist. Durch passende Wahl von α und σ kann die Konvergenz der Integrale für alle relevanten Funktionen sichergestellt werden.

Auch wenn die Integrale in Gl.(4.39) auf den ersten Blick abschreckend wirken, so kann die Bildfunktion für viele Zeitfunktionen rasch ermittelt werden. Als Bildfunktion des Einheitssprungs $f(t) = 1(t)$ erhält man

$$F(s) = \int_{-0}^{\infty} 1(t) \cdot e^{-st} dt = \left[-\frac{1}{s} \cdot e^{-st} \right]_{-0}^{\infty} = -\frac{1}{s}(0 - 1) = \frac{1}{s} , \quad (4.42)$$

sofern $\operatorname{Re}(s) = \sigma > 0$. Durch diese Bedingung an σ wird der sogenannte Konvergenzbereich der Transformation beschrieben – der Bereich der unabhängigen Variablen s , in dem Gl.(4.39) und Gl.(4.40) gelten.

Die Bildfunktion des Einheitsimpuls $\delta(t)$ kann mithilfe der sogenannten „Ausblendeigenschaft“ hergeleitet werden [14]. Diese besagt, dass

$$\int_{-\infty}^{\infty} f(t) \cdot \delta(t - t_0) dt = f(t_0) \quad (4.43)$$

gilt, d.h. das Integral des Produktes einer (stetigen) Funktion $f(t)$ mit einer (um t_0 zeitverschobenen) δ -Funktion ergibt den Funktionswert an der Stelle, an der die δ -Funktion von null verschieden ist. Dies führt zu der Korrespondenz

$$\mathcal{L}\{\delta(t - t_0)\} = \int_{-0}^{\infty} \delta(t - t_0) e^{-st} dt = e^{-st_0} \quad \text{für } t_0 \geq 0 \quad (4.44)$$

und mit $t_0 = 0$ zu

$$\mathcal{L}\{\delta(t)\} = 1 \quad . \quad (4.45)$$

In ganz ähnlicher Weise lässt sich die Exponentialfunktion $f(t) = e^{\lambda t}$ in die Bildfunktion

$$F(s) = \int_{-0}^{\infty} e^{\lambda t} \cdot e^{-st} dt = -\frac{1}{s - \lambda} \left[e^{-t(s - \lambda)} \right]_{-0}^{\infty} = \frac{1}{s - \lambda} \quad (4.46)$$

überführen, sofern $\operatorname{Re}(s) = \sigma > \operatorname{Re}(\lambda)$ gilt.

Für die praktische Anwendung der Laplace-Transformation müssen die jeweiligen Bildfunktionen nicht ausgerechnet werden. Stattdessen greift man auf Korrespondenztafeln zurück, in denen zahlreiche häufiger vorkommende Zeit- und Bildfunktionen aufgeführt sind. Die Tab. 4-1 ist ein Beispiel für eine solche Tafel.

Die Laplace-Transformation lässt sich ohne Weiteres auch auf vektorwertige Funktionen $\mathbf{f}(t)$ anwenden, indem alle Einträge des Vektors komponentenweise transformiert werden.

$F(s)$	$f(t)$ für $t > 0$	$f(t) = 0$ für $t \leq 0$
$\frac{1}{(s - \lambda)^n}$	$\frac{1}{(n-1)!} t^{n-1} e^{\lambda t}$	$n = 1, 2, 3\dots$
$\frac{1}{s}$	$\delta(t)$	
$\frac{1}{s^2}$	$1(t)$	
$\frac{1}{1+sT}$	t	
$\frac{\omega_0^2}{s^2 + 2D\omega_0 s + \omega_0^2}$	$\frac{1}{T} e^{-t/T}$	
	$\frac{\omega_0}{\sqrt{1-D^2}} e^{-D\omega_0 t} \sin(\omega_D t)$	$ D < 1$
	$\omega_0^2 t e^{-D\omega_0 t}$	$\omega_D = \sqrt{1-D^2}\omega_0 \quad D = 1$
$\frac{1}{(1+sT_1)(1+sT_2)}$	$\frac{1}{T_1 - T_2} (e^{-t/T_1} - e^{-t/T_2})$	$T_1 \neq T_2$
$\frac{s}{1+sT}$	$\frac{1}{T} \left(\delta(t) - \frac{1}{T} e^{-t/T} \right)$	
$\frac{s}{(1+sT_1)(1+sT_2)}$	$\frac{1}{T_1 T_2 (T_1 - T_2)} (T_1 e^{-t/T_2} - T_2 e^{-t/T_1})$	$T_1 \neq T_2$
$\frac{s\omega_0^2}{s^2 + 2D\omega_0 s + \omega_0^2}$	$\omega_0^2 e^{-D\omega_0 t} \left(\cos(\omega_D t) - \frac{D}{\sqrt{1-D^2}} \sin(\omega_D t) \right)$	$ D < 1$
	$\omega_D = \sqrt{1-D^2}\omega_0$	
$\frac{1}{s(1+sT)}$	$1 - e^{-t/T}$	
$\frac{1}{s(1+sT_1)(1+sT_2)}$	$1 - \frac{1}{T_1 - T_2} (T_1 e^{-t/T_1} - T_2 e^{-t/T_2})$	$T_1 \neq T_2$
$\frac{\omega_0^2}{s(s^2 + 2D\omega_0 s + \omega_0^2)}$	$1 - e^{-D\omega_0 t} \left(\cos(\omega_D t) + \frac{D}{\sqrt{1-D^2}} \sin(\omega_D t) \right)$	$ D < 1$
	$\omega_D = \sqrt{1-D^2}\omega_0$	

Tabelle 4-1: Korrespondenztafel $F(s) \bullet \circ f(t)$

4.4.2 Laplace-Transformation von Operationen

Im Zeitbereich ist es möglich, Funktionen beispielsweise zu addieren oder abzuleiten. Zu diesen Operationen im Zeitbereich gehören entsprechende Operationen mit den zugehörigen Bildfunktionen im Bildbereich. Aus Gl.(4.39) kann man sehen, dass wegen

$$\begin{aligned} \mathcal{L}\{a_1 \cdot f_1(t) + a_2 \cdot f_2(t)\} &= \int_{-0}^{\infty} (a_1 \cdot f_1(t) + a_2 \cdot f_2(t)) e^{-st} dt \\ &= a_1 \int_{-0}^{\infty} f_1(t) e^{-st} dt + a_2 \int_{-0}^{\infty} f_2(t) e^{-st} dt \\ &= \hat{a}_1 \cdot F_1(s) + a_2 \cdot F_2(s) \end{aligned} \quad (4.47)$$

die Laplace-Transformation linear im Eingangsargument $f(t)$ ist.

Linearität der Laplace-Transformation

Die Laplace-Transformation ist linear und es gilt

$$a_1 \cdot f_1(t) + a_2 \cdot f_2(t) \circ\bullet a_1 \cdot F_1(s) + a_2 \cdot F_2(s) , \quad (4.48)$$

d. h. die Summation von Funktionen und die Multiplikation mit Konstanten bleibt im Bildbereich bestehen.

Die Differentiation ist eine wichtige Operation, wenn man Differentialgleichungen lösen will. Die Regeln zur partiellen Integration

$$\int_{-0}^{\infty} \left(u \cdot \frac{dv}{dt} \right) dt = [u \cdot v]_{-0}^{\infty} - \int_{-0}^{\infty} \left(v \cdot \frac{du}{dt} \right) dt \quad (4.49)$$

werden benutzt, um für die Ableitung $\dot{f}(t) = \frac{df(t)}{dt}$ einer Funktion $f(t)$ die Laplace-Transformation zu bestimmen. Mit der Definition der Laplace-

Transformation in Gl.(4.39) und Gl.(4.49) erhält man

$$\begin{aligned}\mathcal{L}\{\dot{f}(t)\} &= \int_{-0}^{\infty} \frac{df(t)}{dt} e^{-st} dt = \int_{-0}^{\infty} e^{-st} \frac{df(t)}{dt} dt \\ &= [e^{-st} \cdot f(t)]_{-0}^{\infty} - \int_{-0}^{\infty} f(t) \cdot (-s \cdot e^{-st}) dt .\end{aligned}\quad (4.50)$$

Wählt man σ hinreichend groß, so wird e^{-st} für transformierbare $f(t)$ schneller abfallen als $f(t)$ steigen kann und es ergibt sich

$$\begin{aligned}\mathcal{L}\{\dot{f}(t)\} &= \underbrace{\lim_{t \rightarrow \infty} e^{-st} \cdot f(t)}_{\rightarrow 0} - e^0 \cdot f(-0) + s \cdot \int_{-0}^{\infty} f(t) e^{-st} dt \\ &= -f(-0) + s \cdot \mathcal{L}\{f(t)\} .\end{aligned}\quad (4.51)$$

Laplace-Transformation von Ableitungen

Durch Anwendung der Laplace-Transformation wird aus der Differentiation im Zeitbereich eine Multiplikation mit der unabhängigen Variablen s im Bildbereich

$$\mathcal{L}\{\dot{f}(t)\} = -f(-0) + s \cdot \mathcal{L}\{f(t)\} \quad (4.52)$$

mit $f(-0)$ als linksseitiger Grenzwert der Funktion im Nullpunkt.

Damit wird durch Laplace-Transformation aus der unter Umständen schwierigen Differentiation im Zeitbereich eine einfache algebraische Multiplikation im Bildbereich. In vielen Fällen – wie wenn ausschließlich die partikuläre Lösung betrachtet wird – können die Anfangsbedingungen wie $f(-0)$ zu null angenommen werden.

Für höhere Ableitungen gewinnt man auf ähnlichem Wege

$$\begin{aligned}\mathcal{L}\{\ddot{f}(t)\} &= s^2 \mathcal{L}\{f(t)\} - s \cdot f(-0) - \dot{f}(-0) \\ \mathcal{L}\{f^{(n)}(t)\} &= s^n \mathcal{L}\{f(t)\} - \sum_{k=1}^n s^{n-k} f^{(k-1)}(-0) .\end{aligned}\quad (4.53)$$

Operation	Zeitbereich: $f(t)$	Bildbereich: $F(s)$
Multiplikation mit Konstante	$c \cdot f(t)$	$c \cdot F(s)$
Summenbildung	$f_1(t) + f_2(t) + \dots$	$F_1(s) + F_2(s) + \dots$
Verschiebung	$f(t - T_t)$	$F(s) \cdot e^{-sT_t}$
	$\dot{f}(t)$	$s \cdot F(s) - f(-0)$
	$\ddot{f}(t)$	$s^2 \cdot F(s) - sf(-0) - \dot{f}(-0)$
Differentiation	$f^{(n)}(t)$	$s^n \cdot F(s) - \sum_{k=1}^n s^{n-k} f^{(k-1)}(-0)$
		$f(-0)$ ist der Grenzwert von $f(t)$, der sich ergibt, wenn t von negativen Werten aus gegen null geht.
Integration	$\int_0^t f(\tau) d\tau$	$\frac{1}{s} F(s)$
Anfangswert	$\lim_{t \rightarrow 0^+} f(t)$	$\lim_{s \rightarrow \infty} s \cdot F(s)$
Endwert	$\lim_{t \rightarrow \infty} f(t)$	$\lim_{s \rightarrow 0} s \cdot F(s)$

Tabelle 4-2: Operationen im Zeit- und Bildbereich

Weitere nützliche Korrespondenzen, wie die zur Integration und zur Bestimmung von Grenzwerten, sind ohne Beweis in Tab. 4-2 aufgeführt.

4.4.3 Bestimmung des Zeitverlaufes linearer Systeme

Mit Hilfe der Korrespondenzen in Tab. 4-2 kann man eine gewöhnliche Differentialgleichung oder ein System miteinander gekoppelter Differentialgleichungen einschließlich ihrer Anfangsbedingungen in den Bildbereich transformieren. Das soll im SISO-Fall am Beispiel der Lösung einer Differentialgleichung erster Ordnung dargestellt werden. Der MIMO-Fall lässt sich aufgrund des Überlagerungsprinzips und der Linearität der Laplace-Transformation identisch behandeln, indem das gezeigte Verfahren auf die einzelnen Eingangsgrößen $u_i(t)$ angewendet wird und die Lösungen überlagert werden.

Die Differentialgleichung $T\dot{y} + y = K \cdot u$ wird zur Lösung in den Bildbereich der Laplace-Transformation übertragen, indem beide Seiten der Gleichung transformiert werden:

$$\mathcal{L}\{T\dot{y} + y\} = \mathcal{L}\{K \cdot u\} \quad (4.54)$$

Mit den Korrespondenzen in Tab. 4-2 findet man

$$T \cdot (s \cdot Y(s) - y(-0)) + Y(s) = K \cdot U(s), \quad (4.55)$$

und das ist eine algebraische Gleichung, die das Abbild der Eingangsgröße $U(s)$ mit dem der Ausgangsgröße $Y(s)$ verbindet. Der Term $y(-0)$ aus der Anfangsbedingung ist eine Konstante, die aus der Vorgeschichte der Größe $y(t)$ bekannt ist. Die Gleichung liefert nach der Ausgangsgröße aufgelöst

$$Y(s) = \frac{T \cdot y(-0)}{1 + sT} + \frac{K}{1 + sT} \cdot U(s) \quad . \quad (4.56)$$

Man findet hier die Struktur der homogenen und partikulären Lösung wieder, da der erste Summand nur durch die Anfangsbedingung, der zweite nur durch die Eingangsgröße bestimmt wird.

Es soll jetzt zunächst die homogene Lösung betrachtet werden. Man kann leicht nachrechnen, dass für $u(t) = 0$ die zugehörige Bildfunktion $U(s) = 0$

ist. Als Anfangsbedingung sei $y(-0) = {}_0y$ gegeben. Damit ergibt sich die Bildfunktion der Lösung zu

$$\text{Institut für Regelungstechnik} \quad Y(s) = \frac{1}{1+sT} \cdot T \cdot {}_0y \quad . \quad (4.57)$$

Der Korrespondenztafel entnimmt man die zugehörige Zeitfunktion

$$y(t) = \frac{1}{T} e^{-t/T} \cdot T \cdot {}_0y = {}_0y \cdot e^{-t/T} \quad , \quad (4.58)$$

die aus Bild 3-4 bereits bekannt ist.

Auf analogem Wege lässt sich beispielsweise auch die Übergangsfunktion ermitteln. Der Korrespondenztafel Tab. 4-1 entnimmt man

$$u(t) = 1 \quad \circ \bullet \quad U(s) = \frac{1}{s} \quad (4.59)$$

setzt dies und die Anfangsbedingung $y(0) = 0$ in die transformierte Differentialgleichung ein und erhält als Bildfunktion der Lösung

$$Y(s) = \frac{K}{s \cdot (1+sT)} \quad . \quad (4.60)$$

Obgleich diese Funktion in der Tab. 4-1 enthalten ist, soll hieran ein Weg gezeigt werden, auf dem kompliziertere Ausdrücke mit Nennern höheren Grades der Auswertung mit Korrespondenztafeln zugänglich gemacht werden können.

Partialbruchzerlegung, Residuum

Gegeben ist eine komplexwertige Funktion $F(s) : \mathbb{C} \rightarrow \mathbb{C}$ mit n verschiedenen, einfachen Polstellen $\lambda_i \in \mathbb{C}$. Dann lässt sich $F(s)$ in ein Polynom $P(s)$ (ohne Polstellen) und eine Summe von n Brüchen mit je einer Polstelle zerlegen:

$$F(s) = P(s) + \sum_{i=1}^n \frac{r_i}{s - \lambda_i} \quad . \quad (4.61)$$

Hierbei heißt r_i das *Residuum* der Polstelle λ_i . Im Falle mehrfacher Pol-

stellen muss der zugehörige Bruch zu

$$\frac{r_i^0 + r_i^1 s + \dots + r_i^{p-1} s^{p-1}}{(s - \lambda_i)^p} \quad (4.62)$$

erweitert werden mit p als der Vielfachheit der Polstelle.

Mit der Partialbruchzerlegung wird ein gegebener Ausdruck in eine Summe einfacherer Ausdrücke zerlegt, die einzeln transformiert werden können. Im vorliegenden Fall hat $Y(s)$ die Polstellen

$$\lambda_1 = 0 \quad , \quad \lambda_2 = \frac{1}{T} \quad . \quad (4.63)$$

Damit lautet die Partialbruchzerlegung

$$Y(s) = \frac{r_1}{s - \lambda_1} + \frac{r_2}{s - \lambda_2} = \frac{r_1}{s} + \frac{r_2}{s + 1/T} = \frac{r_1}{s} + \frac{r_2 T}{1 + sT} \quad . \quad (4.64)$$

Durch Koeffizientenvergleich findet man die Residuen

$$r_1 = K \quad , \quad r_2 T = -KT \quad (4.65)$$

und damit als Partialbruchzerlegung der Lösungsfunktion

$$Y(s) = \frac{K}{s} - \frac{KT}{1 + sT} \quad . \quad (4.66)$$

Aufgrund der Linearität der Laplace-Transformation kann man die obige Summe gliedweise in den Zeitbereich zurückübersetzen. Man gewinnt

$$y(t) = h(t) = K - K \cdot e^{-t/T} = K \cdot (1 - e^{-t/T}) \quad (4.67)$$

mit dem in Bild 4-3 gezeigten Verlauf. Da die Übergangsfunktion gegen einen endlichen Endwert strebt, ist das System stabil, wobei K genau der statischen Verstärkung entspricht. Der resultierende Verlauf entspricht dabei dem Piktogramm in Tab. 2-1.

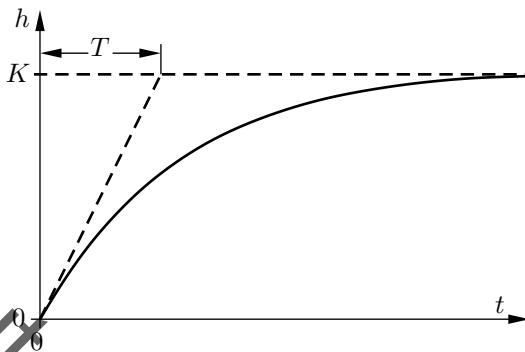


Bild 4-3: Übergangsfunktion eines Systems erster Ordnung

4.5 Übertragungsfunktion

Im vorherigen Abschnitt war die Bildfunktion $Y(s)$ der Lösung einer Differentialgleichung mit der Bildfunktion der Eingangsgröße $U(s)$ multiplikativ über einen gebrochen rationalen Ausdruck in s verknüpft. Diese Aussage ist allgemeingültig. Aus der Differentialgleichung

$$a_n y^{(n)} + \dots + a_1 \dot{y} + a_0 y = b_0 u + b_1 \dot{u} + \dots + b_m u^{(m)} \quad (4.68)$$

erhält man durch Laplace-Transformation beider Seiten bei verschwindenden Anfangsbedingungen

$$\begin{aligned} a_n s^n Y(s) + \dots + a_1 s Y(s) + a_0 Y(s) &= \\ b_0 U(s) + b_1 s U(s) + \dots + b_m s^m U(s) \end{aligned} \quad (4.69)$$

und daraus durch Zusammenfassen

$$Y(s)(a_n s^n + \dots + a_1 s + a_0) = U(s)(b_0 + b_1 s + \dots + b_m s^m). \quad (4.70)$$

Daraus lässt sich ein Quotient bilden, der als *Übertragungsfunktion* bezeichnet wird, da er beschreibt, wie die Größe $U(s)$ in die Größe $Y(s)$ umgewandelt wird, d.h. wie eine Größe vom Eingang des durch die Funktion beschriebenen Übertragungsgliedes zum Ausgang übertragen wird.

Übertragungsfunktion

Gegeben sei die Differentialgleichung eines LTI-Systems mit verschwindenden Anfangsbedingungen

$$a_n y^{(n)} + \dots + a_1 \dot{y} + a_0 y = b_0 u + b_1 \dot{u} + \dots + b_m u^{(m)} . \quad (4.71)$$

Durch Laplace-Transformation gewinnt man die Übertragungsfunktion $G(s)$ des Systems zu

$$G(s) = \frac{Y(s)}{U(s)} = \frac{b_m s^m + \dots + b_1 s + b_0}{a_n s^n + \dots + a_1 s + a_0} . \quad (4.72)$$

Darin sind $Y(s)$ und $U(s)$ die Bildfunktionen der entsprechenden Größen und $G(s)$ eine Funktion, die ausschließlich von der Differentialgleichung bestimmt wird.

Man erkennt, dass die Übertragungsfunktion eine gebrochen rationale Funktion der Variablen s ist und alle Koeffizienten der Differentialgleichung enthält. Sie beschreibt daher den Zusammenhang zwischen Eingangs- und Ausgangsgröße genauso gut wie die Differentialgleichung.

Ein analoges Vorgehen ist im Mehrgrößenfall durch die komponentenweise Betrachtung aller Kombinationen von Ein- und Ausgangsgrößen möglich, was auf die $n \times m$ -Übertragungsmatrix

$$\mathbf{G}(s) = \begin{bmatrix} \frac{Y_1(s)}{U_1(s)} & \dots & \frac{Y_1(s)}{U_m(s)} \\ \vdots & & \vdots \\ \frac{Y_n(s)}{U_1(s)} & \dots & \frac{Y_n(s)}{U_m(s)} \end{bmatrix} \quad (4.73)$$

führt. Dies lässt sich auch durch Anwendung der Laplace-Transformation auf die lineare Zustandsraumdarstellung plausibilisieren. Die Zustandsgleichungen im Bildbereich der Laplace-Transformation lauten für verschwindende Anfangsbedingungen

$$\begin{aligned} s \cdot \mathbf{X}(s) &= \mathbf{A} \cdot \mathbf{X}(s) + \mathbf{B} \cdot \mathbf{U}(s) \\ \mathbf{Y}(s) &= \mathbf{C} \cdot \mathbf{X}(s) + \mathbf{D} \cdot \mathbf{U}(s) . \end{aligned} \quad (4.74)$$

Die Koeffizientenmatrizen \mathbf{A} , \mathbf{B} , \mathbf{C} , \mathbf{D} werden durch die Transformation ebenso wenig verändert wie die Koeffizienten einer einfachen Differential-

gleichung. Hieraus gewinnt man nach den Regeln der Matrizenrechnung

$$\mathbf{Y}(s) = \underbrace{\left(\mathbf{C} \cdot (s \cdot \mathbf{I} - \mathbf{A})^{-1} \cdot \mathbf{B} + \mathbf{D} \right)}_{\mathbf{G}(s)} \cdot \mathbf{U}(s) \quad (4.75)$$

eine kompakte Umrechnungsformel vom Zustandsraum in den Bildbereich.

Insgesamt erhält man im Fall verschwindender Anfangsbedingungen immer eine Lösung im Bildbereich von der Form $\mathbf{Y}(s) = \mathbf{G}(s) \cdot \mathbf{U}(s)$ und somit einen geschlossenen Lösungsausdruck für die partikuläre Lösung, welcher aufgrund der Eindeutigkeit der Lösung identisch zur Faltung sein muss.

Laplace-Transformation der Faltung

Es gilt

$$Y(s) = G(s) \cdot U(s) \quad \bullet \circ \quad y(t) = g(t) * u(t) \quad . \quad (4.76)$$

Also wird durch die Laplace-Transformation die Operation der Faltung $f(t) * g(t)$ im Zeitbereich in eine Multiplikation $F(s) \cdot G(s)$ im Bildbereich überführt.

Für die Laplace-Transformierte der Gewichtsfunktion $g(t)$ ergibt sich dabei

$$u(t) = \delta(t) \quad \circ \bullet \quad 1 = U(s) \Rightarrow \mathcal{L}\{g(t)\} = G(s) \cdot 1 = G(s) \quad , \quad (4.77)$$

und das ist genau die Übertragungsfunktion, was die Buchstabenvergabe g und G rechtfertigt. Dieser Zusammenhang ist die Begründung dafür, dass das Lösen von Differentialgleichungen im Bildbereich wesentlich einfacher als im Zeitbereich ist, da die komplexe Rechenoperation der Faltung in eine simple Multiplikation überführt wird.

Die Stabilität eines LTI-Systems lässt sich nicht nur auf Ebene der Differentialgleichungen oder im Zustandsraum, sondern auch über die Übertragungsfunktion überprüfen. Offenbar ist der Nenner der Übertragungsfunktion in Gl.(4.72) genau das charakteristische Polynom der Differentialgleichung. Die Nullstellen des Nenners – d. h. alle Polstellen von $G(s)$ – sind also genau die Eigenwerte von \mathbf{A} und Wurzeln des charakteristischen Polynoms.

Diese Aussage gilt streng genommen nur dann, wenn es innerhalb von $G(s)$ nicht zu einer Pol-Nullstellen-Kürzung zwischen Zähler- und Nennerpolynom kommt. In einem solchen Fall könnte man Zähler und Nenner kürzen und das Nennerpolynom wäre in Wirklichkeit um eine Ordnung geringer bei identischem Übertragungsverhalten. Eine Pol-Nullstellen-Kürzung innerhalb von $G(s)$ kann also nur dann auftreten, wenn es sich nicht um eine minimale Realisierung handelt. Daher kann eine solche Kürzung hier ausgeschlossen werden.

Stabilität im Bildbereich

Gegeben ist die Differentialgleichung eines LTI-Systems und die zugehörige Übertragungsfunktion $G(s)$. Dann ist das charakteristische Polynom das Nennerpolynom von $G(s)$

$$p(s) = a_n s^n + \dots + a_1 s + a_0 \quad . \quad (4.78)$$

Die Wurzeln des charakteristischen Polynoms entsprechen damit genau den Polstellen von $G(s)$, die für Stabilität alle einen negativen Realteil aufweisen müssen.

Die Polstellen von $G(s)$ sind also ein entscheidender Indikator für das Systemverhalten. Eine häufig genutzte graphische Darstellungsweise für gebrochen rationale Übertragungsfunktionen setzt hier an, indem die Polstellen und Nullstellen von $G(s)$ in ein gemeinsames Diagramm überführt werden. Nach dem Fundamentalsatz der Algebra kann jedes Polynom durch seine Nullstellen λ_i und den Koeffizienten a_n der höchsten Potenz der Variablen s ausgedrückt werden:

$$a_n s^n + \dots + a_1 s + a_0 = a_n \cdot (s - \lambda_1) \cdot (s - \lambda_2) \dots (s - \lambda_n) \quad . \quad (4.79)$$

Daher lässt sich eine gebrochen rationale Übertragungsfunktion in die Form

$$G(s) = K \cdot \frac{(s - \eta_1) \cdot (s - \eta_2) \dots (s - \eta_m)}{(s - \lambda_1) \cdot (s - \lambda_2) \dots (s - \lambda_n)} \quad (4.80)$$

mit den Nullstellen η_i und Polstellen $\lambda_i \in \mathbb{C}$ überführen.

Pol-Nullstellen-Diagramm

Die graphische Darstellung einer gebrochen rationalen Übertragungsfunktion $G(s)$ in Form ihrer in der komplexen s -Ebene mit Kreuzen markierten Polstellen und mit Kreisen markierten Nullstellen beschreibt die Übertragungsfunktion bis auf den Vorfaktor K eindeutig.

Differentialgleichung Übertragungsfunktion	Pol-Nullstellen-Diagramm	Interpretation
$y = K \cdot u$ $G(s) = K$		P-Glied nicht darstellbar
$\dot{y} = K_I \cdot u$ $G(s) = \frac{K_I}{s}$		I-Glied grenzstabil
$-T\dot{y} + y = K \cdot u$ $G(s) = \frac{K}{-Ts + 1}$		instabiles System relativer Grad 1
$\ddot{y} + 2\dot{y} + 2y = u - \dot{u}$ $G(s) = \frac{-s + 1}{s^2 + 2s + 2}$		stabiles System schwingungsfähig relativer Grad 1

Tabelle 4-3: Beispiele für Pol-Nullstellen-Diagramme

Beispiele für Pol-Nullstellen-Diagramme sind in Tab. 4-3 ausgeführt. Da Pol- und Nullstellen konstante komplexe Werte sind, ist die Lage der sie bezeichnenden Symbole keine Funktion irgendeiner unabhängigen Variablen und es ergibt sich für jede Übertragungsfunktion genau ein Pol-Nullstellen-Diagramm. Man erkennt in Tab. 4-3, dass die Übertragungsfunktion des proportional wirkenden Gliedes nur aus dem nicht durch Pol- und Nullstellen darstellbaren Übertragungsfaktor besteht. Für eine vollständige Darstellung eines Systems muss dieser ergänzend zum Pol-Nullstellen-Diagramm angegeben werden. Aus dem Pol-Nullstellen-Diagramm lassen sich wesentliche dynamische Eigenschaften direkt ablesen. Für Stabilität müssen alle durch Kreuze markierten Polstellen in der linken offenen komplexen Halbebene liegen. Das System ist schwingungsfähig, sobald nicht alle durch Kreuze markierten Polstellen auf der reellen Achse liegen. Der relative Grad und damit die Kausalität lässt sich direkt über die Differenz von Polstellen und Nullstellen ablesen. Letzteres ist ein Vorteil der Übertragungsfunktion gegenüber beispielsweise der Darstellung im Zustandsraum.

4.6 Grenzwertsätze

Neben der Lösung von Differentialgleichungen für definierte Eingangsgrößen $u(t)$ liefert die Laplace-Transformation über die Grenzwertsätze in Tab. 4-2 ein sehr handliches Hilfsmittel. Die Formeln

$$\begin{aligned}\lim_{t \rightarrow +0} f(t) &= \lim_{s \rightarrow \infty} sF(s) \\ \lim_{t \rightarrow \infty} f(t) &= \lim_{s \rightarrow 0} sF(s)\end{aligned}\tag{4.81}$$

sind dabei nur anwendbar, wenn die zugehörigen Grenzwerte auch existieren (d. h. insbesondere endlich sind). Dies vorausgesetzt erhält man mit Gl. (4.81) und $H(s) = 1/s \cdot G(s)$ den Zusammenhang zwischen den Grenzwerten von Übertragungsfunktion $G(s)$ und Übergangsfunktion $h(t)$ zu

$$\begin{aligned}\lim_{t \rightarrow \infty} h(t) &= \lim_{s \rightarrow 0} sH(s) = \lim_{s \rightarrow 0} G(s) \quad , \\ \lim_{t \rightarrow +0} h(t) &= \lim_{s \rightarrow \infty} sH(s) = \lim_{s \rightarrow \infty} G(s) \quad .\end{aligned}\tag{4.82}$$

Hierüber lässt sich der Anfangswert und für stabile Systeme der Endwert der Übergangsfunktion leicht bestimmen.

Berechnung der statischen Verstärkung

Aus den Grenzwertsätzen erhält man für die statische Verstärkung

$$K = \lim_{t \rightarrow \infty} h(t) = \lim_{s \rightarrow 0} G(s) = G(0) \quad , \quad (4.83)$$

d.h. für stabile Systeme entspricht die statische Verstärkung dem Wert der Übertragungsfunktion für $s = 0$.

Für die Berechnung der Anfangswerte $t \rightarrow +0$ muss der Grenzwert für $s \rightarrow \infty$ gebildet werden. Betrachtet man die Übergangsfunktion, so ergibt sich für ein System mit n Polstellen und m Nullstellen

$$\lim_{t \rightarrow +0} h(t) = \lim_{s \rightarrow \infty} \frac{b_0 + \dots + b_m s^m}{a_0 + \dots + a_n s^n} = \begin{cases} 0 & \text{für } n > m \\ \frac{b_m}{a_n} & \text{für } n = m \\ \infty & \text{für } n < m \end{cases} \quad (4.84)$$

Der Anfangswert der Übergangsfunktion hängt also maßgeblich vom relativen Grad ab. Während sich für aukausale Systeme ein nicht endlicher Anfangswert ergibt, ist für Systeme mit relativem Grad von 1 oder höher der Anfangswert stets null. Dies gilt unabhängig von der Stabilität des betrachteten Systems. Für Systeme mit $n = m$ und relativem Grad 0 reagiert die Ausgangsgröße bei einem Sprung der Eingangsgröße auch direkt mit einem Sprung, der um den Faktor b_m/a_n verstärkt ist.

Sprungfähigkeit, Durchgriff

Ein System heißt sprungfähig, wenn bei einem Sprung der Eingangsgröße u in $t = t_0$ auch die Ausgangsgröße y einen Sprung in $t = t_0$ aufweist. Ein System ist genau dann sprungfähig, wenn sein relativer Grad null ist.

Alternativ spricht man davon, dass das System einen *Durchgriff* besitzt, da es eine direkte, unverzögerte Wirkung von u auf y gibt.

Sprungfähigkeit entspricht im Zustandsraum der Eigenschaft, dass die Durchgangsmatrix $\mathbf{D} \neq 0$ nicht verschwindet. Die meisten Regelstrecken werden einen relativen Grad von eins oder größer besitzen und nicht sprungfähig sein. In Reglern, die schnell auf Abweichungen in der Regelgröße und Störungen reagieren müssen, ist ein Durchgriff hingegen häufig anzutreffen.

Die Argumentation bezüglich des Anfangswertes der Übergangsfunktion lässt sich auch auf deren Ableitungen übertragen. Wegen $df/dt \circ \bullet sF(s)$ erhält man für die Anfangssteigung der Übergangsfunktion

$$\lim_{t \rightarrow +0} g(t) = \lim_{s \rightarrow \infty} \frac{b_0 s + \dots + b_m s^{m+1}}{a_0 + \dots + a_n s^n} = \begin{cases} 0 & \text{für } n > m + 1 \\ \frac{b_m}{a_n} & \text{für } n = m + 1 \\ \infty & \text{für } n < m + 1 \end{cases}. \quad (4.85)$$

Die erste Ableitung der Übergangsfunktion springt also auf einen endlichen Wert in $t = 0$, wenn das System einen relativen Grad von eins besitzt. Dies lässt sich auf beliebige Ableitungen fortsetzen:

Relativer Grad und Anfangswert

Die r -te Zeitableitung der Übergangsfunktion eines Systems mit relativem Grad r wird sich in $t = 0$ sprungförmig verändern.

Als Beispiel für diesen Zusammenhang sei auf Bild 4-3 verwiesen. Für dieses System mit relativem Grad $r = 1$ weist die Übergangsfunktion in $t = 0$ einen Knick auf. Obgleich die Funktion $h(t)$ selbst stetig ist, liegt in $\dot{h}(t)$ für $t = 0$ ein Sprung des Funktionswertes vor, der sich in dem Knick bei $t = 0$ zeigt.

Wichtig ist es, vor Anwendung der Grenzwertsätze stets deren Anwendbarkeit zu prüfen, da ein voreiliges Anwenden ohne vorangegangene Prüfung der Voraussetzung zu Fehlschlüssen führen kann. Beispielsweise ergibt die Anwendung der Grenzwertsätze für

$$G(s) = \frac{K}{1 + Ts} \Rightarrow \lim_{s \rightarrow 0} = K \quad (4.86)$$

unabhängig von T . Allerdings existiert der statische Endwert $\lim_{t \rightarrow \infty} h(t)$ nur für positive $T > 0$. Für $T < 0$ führt die Anwendung der Grenzwertsätze ohne Prüfung der Voraussetzungen in die Irre. In der praktischen Anwendung auf regelungstechnische Probleme kann man fast immer annehmen, dass $\lim h(t)$ für $t \rightarrow +0$ existiert, wenn $\lim G(s)$ für $s \rightarrow \infty$ existiert. Die Existenz von $\lim h(t)$ für $t \rightarrow \infty$ ist jedoch nur gesichert, wenn $G(s)$ ein stabiles Übertragungssystem beschreibt.

5 Verhalten bei sinusförmiger Anregung

5.1 Frequenzgang

Mit den im vorangegangenen Kapitel 4 vorgestellten Methoden lässt sich die Antwort eines LTI-Systems auf jede transformierbare Eingangsfunktion ermitteln. Eine besondere Bedeutung kommt dabei sinusförmigen Eingangssignalen wie

$$u(t) = U \cdot \cos(\omega t) \quad (5.1)$$

mit der Amplitude U und der Frequenz ω zu. Will man die partikuläre Lösung $y_p(t)$ eines LTI-Systems auf dieses Eingangssignal mithilfe der Laplace-Transformation berechnen, so führt dies mit

$$U(s) = U \frac{s}{s^2 + \omega^2} \quad (5.2)$$

auf

$$Y_p(s) = G(s) \cdot U(s) = U \frac{b_0 + b_1 s + \dots + b_m s^m}{a_0 + a_1 s + \dots + a_n s^n} \cdot \frac{s}{s^2 + \omega^2} \quad . \quad (5.3)$$

Mit dem Lösungsansatz der Partialbruchzerlegung ergibt sich dann strukturell

$$Y_p(s) = U \underbrace{\frac{Z(s)}{a_0 + a_1 s + \dots + a_n s^n}}_{Y_1(s)} + U \underbrace{\frac{r_1 \cdot s + r_2}{s^2 + \omega^2}}_{Y_2(s)} \quad (5.4)$$

mit den zu bestimmenden Residuen r_1 und r_2 sowie dem Zähler $Z(s)$. Der erste Term der Partialbruchzerlegung $Y_1(s)$ enthält im Nenner das charakteristische Polynom. Der zweite Term $Y_2(s)$ enthält im Nenner $s^2 + \omega^2$ und entspricht damit erneut einem sinusförmigen Signal mit Frequenz ω , dessen Amplitude und Phasenlage sich durch die Residuen jedoch verändert hat. Damit wird für stabile Systeme

$$\lim_{t \rightarrow \infty} y_p(t) = \lim_{t \rightarrow \infty} y_1(t) + \lim_{t \rightarrow \infty} y_2(t) = 0 + Y \cdot \cos(\omega t + \varphi) \quad (5.5)$$

gelten. Die Situation, wenn alle Anteile von $y(t)$, die nicht der Frequenz der sinusförmigen Anregung entsprechen, abgeklungen sind, wird *eingeschwungener Zustand* genannt.

Eingeschwungener Zustand

Gegeben ist ein stabiles LTI-System, das mit einem sinusförmigen Eingangssignal mit Frequenz ω angeregt wird. Dann wird die Ausgangsgröße des LTI-Systems für $t \rightarrow \infty$ gegen eine sinusförmige Schwingung konvergieren, die im Vergleich zur Eingangsschwingung die gleiche Frequenz ω , aber eine veränderte Amplitude und Phasenlage aufweisen wird. Dieser Grenzwert ist der *eingeschwungene Zustand*.

LTI-Systeme bilden im eingeschwungenen Zustand also sinusförmige Eingangssignale auf sinusförmige Ausgangssignale ab, wenn sie stabil sind. Von großem Interesse ist dabei die Phasenverschiebung und die Amplitudenverstärkung, die zwischen Ein- und Ausgangssignal auftritt. So kann eine hohe Amplitudenverstärkung auf eine Resonanzfrequenz hindeuten und die Phasenlage wird später für die Stabilität des geschlossenen Regelkreises relevant werden (siehe Abschnitt 9.3). Sowohl die Amplitudenveränderung als auch die Phasenverschiebung hängen von den Koeffizienten der Partialbruchzerlegung ab und werden daher auch von der Frequenz ω des Eingangssignals abhängen. Eine Bestimmung der frequenzabhängigen Verstärkung Y und von $\varphi(\omega)$ über die Laplace-Transformation ist dabei aufwändig. Alternativ kann man versuchen, diese durch Einsetzen in die Differentialgleichung zu gewinnen. Schon bei einer Differentialgleichung erster Ordnung $T\dot{y} + y = Ku$ erhält man für $u(t) = U \cdot \cos(\omega t)$ und dem Ansatz $y(t) = Y \cdot \cos(\omega t + \varphi)$

$$-Y \cdot T \cdot \omega \sin(\omega t + \varphi) + Y \cdot \cos(\omega t + \varphi) = U \cdot \cos(\omega t) \quad (5.6)$$

und damit eine Gleichung, deren Lösung nach φ und Y Probleme bereitet. Die Schwierigkeit besteht darin, dass sowohl sin- als auch cos-Ausdrücke erscheinen, die miteinander verrechnet werden müssen, und dass die Addition sinusförmiger Signale über die Additionstheoreme umständlich ist.

Einen wesentlich schnelleren Weg zum Ziel bietet die Zeigerdarstellung für sinusförmige Signale.

Zeigerdarstellung für sinusförmige Signale

Ein sinusförmiges Signal $f(t) = \hat{f} \cdot \cos(\omega t + \varphi_0)$ lässt sich durch einen Zeiger \underline{f} in der komplexen Ebene darstellen. Dieser Zeiger hat die Länge \hat{f} und den Startwinkel φ_0 und rotiert mit der Frequenz ω mathematisch positiv (d. h. gegen den Uhrzeigersinn) um den Ursprung. In komplexer Schreibweise ergibt sich für den Zeiger

$$f(t) = \hat{f} \cdot e^{j\omega t + \varphi_0} = \hat{f} \cdot (\cos(\omega t + \varphi_0) + j \sin(\omega t + \varphi_0)) \quad (5.7)$$

mit der Eulerschen Darstellung für komplexe Zahlen. Der aktuelle Funktionswert ergibt sich dann aus dem Realteil des Zeigers für den gesuchten Zeitpunkt

$$f(t) = \operatorname{Re}(\underline{f}(t)) = \hat{f} \cdot \cos(\omega t + \varphi_0) \quad . \quad (5.8)$$

Die Darstellung einer sinusförmigen Schwingung als Zeiger hat den Vorteil, dass sin und cos in einer einheitlichen Darstellung gefasst werden. Hierdurch kann die Summe von zwei sinusförmigen Schwingungen gleicher Frequenz wie $\sin(\omega t) + \cos(\omega t)$ durch Vektoraddition direkt berechnet werden. Dazu addiert man beide Zeiger und erhält unter der Beachtung von $\sin(\omega t) = \cos(\omega t - \frac{\pi}{2})$ nach Bild 5-1 sofort als Summe $\sqrt{2} \cos(\omega t - \frac{\pi}{4})$.

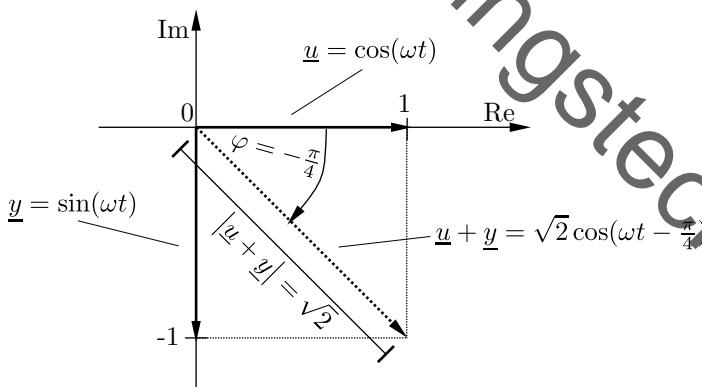


Bild 5-1: Addition zweier sinusförmiger Signale über Zeiger

Grad und Bogenmaß

Winkel werden nach Bedarf in Grad $0^\circ \dots 360^\circ$ oder Bogenmaß (Radian) $0 \dots 2\pi$ angegeben. Werden Winkel mit (Kreis-)Frequenzen ω verknüpft, ist in den vielen Fällen das Bogenmaß die zielführendere Einheit. Die Umrechnung zwischen einem Winkel in Radian φ_π und einem Winkel in Grad φ_{Grad} kann durch

$$\varphi_\pi = \varphi_{\text{Grad}} \cdot \frac{\pi}{180^\circ} \quad (5.9)$$

erfolgen.

Zeiger können rotiert und skaliert werden, indem man diese mit einer komplexen Zahl multipliziert. Dabei führt die Multiplikation mit $r \cdot e^{j\varphi}$ auf eine Skalierung mit r und eine Rotation um φ . Der Wechsel von der Darstellung als $\sin(\omega t)$ oder $\cos(\omega t)$ auf eine einheitliche Form als $e^{j\omega t}$ ermöglicht zudem das einfache Ableiten der Schwingungen. Man erhält

$$\underline{u}(t) = U \cdot e^{j\omega t + \varphi_0} \Rightarrow \dot{\underline{u}}(t) = (j\omega) \cdot U \cdot e^{j\omega t + \varphi_0} = (j\omega) \cdot \underline{u} \quad . \quad (5.10)$$

Man erkennt, dass die Ableitung durch Multiplikation mit $(j\omega)$ gebildet wird. Das entspricht einer Rotation des Zeigers um 90° und Streckung um den Betrag ω . Der Wechsel auf einen Zeiger und damit eine komplexe Zahl führt aufgrund der Linearität des Systems dabei zu keinen zusätzlichen Schwierigkeiten, da durch das Überlagerungsprinzip der imaginäre Anteil des Eingangszeigers auf einen rein imaginären Ausgangszeiger abgebildet wird: $u \mapsto y \Rightarrow ju \mapsto jy$. Normalerweise interessiert nur die von der eigentlichen Schwingung (und damit dem Realteil des Eingangszeigers) hervorgerufene Lösung. Daher reicht es aus, den Realteil des Ausgangszeigers zu betrachten. Einsetzen der Zeiger \underline{u} und \underline{y} in die Differentialgleichung ergibt

$$a_n(j\omega)^n \underline{y} + \dots + a_1(j\omega) \underline{y} + a_0 \underline{y} = b_m(j\omega)^m \underline{u} + \dots + b_1(j\omega) \underline{u} + b_0 \underline{u} \quad (5.11)$$

Hier kann man die Zeiger \underline{u} und \underline{y} ausklammern und erhält den sogenannten *Frequenzgang*:

Frequenzgang

Gegeben ist ein stabiles LTI-System, das mit einem sinusförmigen Eingangssignal $u(t)$ angeregt wird und hierauf im eingeschwungenen Zustand mit $y(t)$ reagiert. Der Frequenzgang ist im eingeschwungenen Zustand definiert als das Verhältnis

$$\text{Frequenzgang} = \frac{\text{Zeiger der sinusförmigen Ausgangsgröße}}{\text{Zeiger der sinusförmigen Eingangsgröße}} \quad (5.12)$$

$$G(j\omega) = \frac{y}{u} .$$

und berechnet sich für die Differentialgleichung

$$a_n y^n + \dots + a_0 y = b_0 u + \dots + b_m u^m$$

zu

$$G(j\omega) = \frac{b_0 + b_1(j\omega) + \dots + b_m(j\omega)^m}{a_0 + a_1(j\omega) + \dots + a_n(j\omega)^n} . \quad (5.13)$$

Der Frequenzgang ist demnach ein komplexer Übertragungsfaktor, der nur von der Frequenz abhängt und der die Zeiger der (sinusförmigen) Ein- und Ausgangsgrößen multiplikativ verknüpft. Folglich ergibt sich der Zeiger der Ausgangsgröße (und damit der eingeschwungene Zustand der Ausgangsgröße) aus dem Zeiger der Eingangsgröße, welcher durch die Multiplikation mit der komplexen Zahl $G(j\omega)$ um den Faktor $|G(j\omega)|$ skaliert und den Winkel $\angle G(j\omega)$ rotiert wurde. Der Betrag des Frequenzgangs beschreibt also die Amplitudenverstärkung und der Winkel des Frequenzgangs die Phasenverschiebung zwischen Ein- und Ausgangssignal. Dabei hat es sich eingebürgert, den Frequenzgang als Funktion von $j\omega$ zu schreiben, obgleich er eine (komplexe) Funktion der reellen Frequenz ω ist.

Übertragung sinusförmiger Signale

Für ein stabiles LTI-System mit dem Frequenzgang $G(j\omega)$ gilt

$$u(t) = \sin(\omega t) \mapsto y(t) = |G(j\omega)| \cdot \sin(\omega t + \angle(G(j\omega))) . \quad (5.14)$$

Aufgrund der großen Bedeutung von Betrag und Phase wird der Betrag

des Frequenzgangs $|G(j\omega)|$ auch als *Amplitudengang* und der Phasenwinkel $\angle(G(j\omega)) = \varphi(\omega)$ als *Phasengang* bezeichnet.

Die Bestimmungsgleichung Gl.(5.13) weist einige auffällige Gemeinsamkeiten mit der Definition der Übertragungsfunktion Gl.(4.72) auf:

$$G(j\omega) = \frac{b_0 + \dots + b_m(j\omega)^m}{a_0 + \dots + a_n(j\omega)^n} \Leftrightarrow G(s) = \frac{b_m s^m + \dots + b_0}{a_n s^n + \dots + a_0}$$

Abschließen der unterschiedlichen Reihenfolge der Koeffizienten, entspricht der Frequenzgang der Übertragungsfunktion für $s = j\omega$, d. h. für $\sigma = 0$ in der Laplace-Transformation. Aus der Übertragungsfunktion $G(s)$ kann also durch einen recht einfachen formalen Schritt der Frequenzgang gewonnen werden. Man muss nur die komplexe Variable $s = \sigma + j\omega$ ersetzen durch die imaginäre Variable $j\omega$.

Der Fakt, dass sich über den Frequenzgang streng genommen nur stabile Systeme beschreiben lassen, da nur diese einen eingeschwungenen Zustand besitzen, spiegelt sich in der Wahl $\sigma = 0$ und dem daraus resultierenden Konvergenzbereich der Laplace-Transformation wider. Der Grenzwertsatz für den Endwert $t \rightarrow \infty$ ergibt somit im Frequenzbereich $\omega \rightarrow 0$. Folglich wird der statische Fall mit einer konstant bleibenden Anregung ($t \rightarrow \infty$) mit einer sinusförmigen Anregung mit Frequenz $\omega = 0$ in Verbindung gebracht, was auch dem intuitiven physikalischen Verständnis entspricht.

Als Beispiel für das Arbeiten mit Frequenzgängen soll die Bewegung eines gefederten Einachsfahrzeuges auf welliger Fahrbahn nach Bild 5-2 untersucht werden. Die (masselosen) Räder folgen der Fahrbahn, sodass

$$u(t) = U_0 \cdot \sin\left(2\pi \cdot \frac{s(t)}{S_0}\right) = U_0 \cdot \sin\left(2\pi \cdot \frac{vt}{S_0}\right) = U_0 \cdot \sin(\omega t) \quad (5.15)$$

gilt. Dabei hängt die Kreisfrequenz $\omega = 2\pi \cdot v/S_0$ von der Fahrgeschwindigkeit v und der Länge S_0 der Fahrbahnwellen ab. Die Bewegung des Wagenkastens wird mit dem Reibbeiwert B durch die Differentialgleichung

$$M\ddot{y} + B\dot{y} + Cy = Cu \quad (5.16)$$

beschrieben (vgl. auch Gl.(2.2)). Der zugehörige Frequenzgang ist

$$G(j\omega) = \frac{\underline{y}}{\underline{u}} = \frac{C}{M(j\omega)^2 + B(j\omega) + C} \quad . \quad (5.17)$$

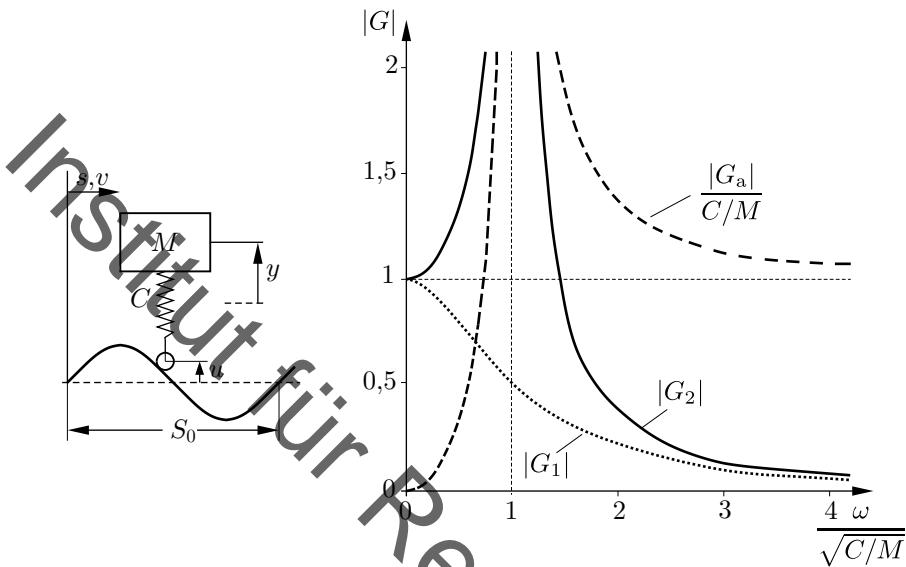


Bild 5-2: Modell eines Wagens auf welliger Fahrbahn und zugehörige Amplitudengänge

Es lässt sich leicht nachrechnen, dass der Wagenkasten für positive Konstanten M , B und C stabil ist, wodurch der eingeschwungene Zustand wohldefiniert ist. In diesem wird der Wagenkasten eine Bewegung

$$y(t) = Y_0 \sin(\omega t + \varphi)$$

ausführen. Wenn man sich nur für die Amplitude Y_0 interessiert, muss nach Gl.(5.14) der Amplitudengang für die Anregungsfrequenz ω ausgewertet werden. Diesen Betrag eines komplexwertigen Bruchs erhält man aus den Regeln für das Rechnen mit komplexen Zahlen zu

$$|G(j\omega)| = \frac{|C|}{|C - M\omega^2 + j\omega B|} = \frac{C}{\sqrt{(C - M\omega^2)^2 + \omega^2 B^2}}. \quad (5.18)$$

Dieser Ausdruck beschreibt das Verhältnis der Amplituden von Rad- und Wagenkastenbewegung abhängig von ω , was in der Mechanik auch Vergrößerungsfunktion genannt wird.

Der Verlauf von $|G(j\omega)|$ über ω wird dabei wesentlich von den Parametern C , M und B abhängen, wobei insbesondere dem Verhältnis der Reibung B zu den anderen Parametern eine Schlüsselrolle zukommt. Dies wird ausführlicher in Kapitel 7 diskutiert, während hier nur zwei Grenzfälle vorgestellt werden sollen. So wird beispielsweise für eine große Reibung $B^2 = 4 \cdot M \cdot C$ der Betrag zu

$$|G_1(j\omega)| = \frac{C}{C + M\omega^2} ,$$

welcher ein Maximum für $\omega = 0$ mit $|G_1(j0) = 1|$ aufweist und danach kontinuierlich abfällt. Für $C = M$ würden nur sehr langsame Änderungen der Fahrbahn nahezu unverändert auf den Wagenkasten übertragen und die Amplitude des Wagenkastens wäre im eingeschwungenen Zustand stets geringer als die Amplitude der Fahrbahn.

Betrachtet man hingegen eine sehr geringe Reibung, so erhält man im Grenzfall $B \rightarrow 0$ den Betrag

$$|G_2(j\omega)| = \frac{C}{C - M\omega^2} .$$

Dieser weist bei der Frequenz $\sqrt{C/M}$ des Systems eine Polstelle auf, was einer Resonanz entspricht. Man sieht aber auch, dass selbst hier das Phänomen auftritt, dass bei hoher Fahrgeschwindigkeit, d. h. bei Frequenzen, die wesentlich größer sind als $\sqrt{C/M}$, die Amplituden der Wagenkastenbewegung wesentlich kleiner werden können als die der Fahrbahn (Bild 5-2). Dies liegt an der Trägheit der Masse, aufgrund derer der Wagenkasten den sehr schnellen Anregungen durch die Fahrbahn nicht mehr folgen kann. Hier ist zu beachten, dass im Grenzfall $B = 0$ der Wagenkasten nicht stabil ist und daher kein eingeschwungener Zustand existiert. Diese Darstellung ist daher als repräsentativ für sehr kleine Werte für B zu verstehen.

Häufig ist neben dem Weg, den der Wagenkasten beschreibt, auch die Beschleunigung von Interesse, weil sie ein Maß für die Belastung etwa der Insassen eines Fahrzeugs darstellt. Die Amplitude A_0 der Beschleunigung $a = \ddot{y}$ lässt sich mit Hilfe des Frequenzganges leicht bestimmen, weil der zugehörige Zeiger

$$\underline{a} = (j\omega)^2 \cdot \underline{y}$$

ist und damit die Amplitude der Beschleunigung über

$$|G_a(j\omega)| = |(j\omega)^2 \cdot G(j\omega)|$$

mit der Amplitude der Fahrbahnwellen verknüpft ist.

Aus Bild 5-2 erkennt man u. a., dass bei Fahrgeschwindigkeiten oberhalb von $\sqrt{C/M}$ zwar die Amplitude des Wagenkastenweges mit wachsender Frequenz stark abnimmt, die Amplitude der Beschleunigung hingegen einem Grenzwert zustrebt, sodass z. B. eine hohe Fahrgeschwindigkeit nicht als geeignetes Mittel zum schonenden Personentransport erscheint.

5.2 Ortskurve

Wie fast alle Funktionen können auch Frequenzgänge, die Funktionen der Frequenz sind, entweder als analytische Ausdrücke oder als Wertetabellen oder graphisch dargestellt werden. Aus den zahlreichen Möglichkeiten der graphischen Darstellung werden die Darstellung als *Ortskurve* und die logarithmische Darstellung im sogenannten *Bode¹-Diagramm* (siehe Abschnitt 5.3) für Regelungstechnische Zwecke am häufigsten benutzt.

Ortskurven von Frequenzgängen

Für jedes feste ω ist der Frequenzgang $G(j\omega)$ eine feste komplexe Zahl und damit ein Punkt in der komplexen Ebene. Verbindet man für ein definiertes Frequenzintervall all diese Punkte, so erhält man einen Linienzug – die Ortskurve des Frequenzgangs.

Ein Beispiel für eine Ortskurve zeigt Bild 5-3.

Die verbundenen Punkte können auch als Endpunkte von Zeigern aufgefasst werden, die den Frequenzgang repräsentieren oder als Zeiger der Ausgangsgröße y für den Fall, dass die Eingangsgröße durch den Einheitszeiger $u = e^{j\omega t}$ beschrieben wird; in diesem Fall ist $y = G(j\omega)$. Üblicherweise werden Ortskurven von Frequenzgängen als Linienzüge mit einer Frequenzparametrierung, mindestens aber einer Angabe über die Richtung wachsender Frequenz, dargestellt. Aus der Ortskurve lässt sich die bei einer bestimmten Frequenz auftretende Amplitudenverstärkung und Phasenverschiebung

¹Hendrik Wade Bode (1905-1982), amerikanischer Elektrotechniker [3]

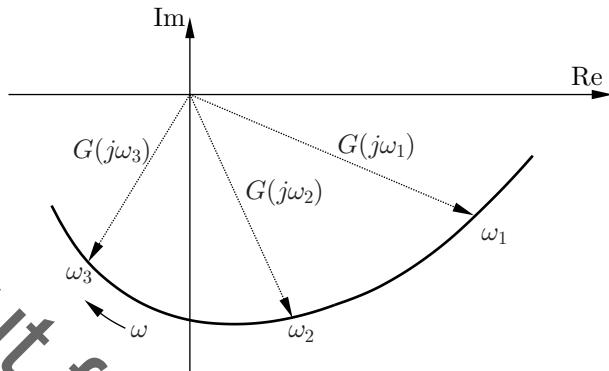


Bild 5-3: Ortskurve eines Frequenzgangs

leicht über den Betrag und Winkel des zugehörigen Zeigers ermitteln. Einige Beispiele sind in Tab. 5-1 zusammengestellt. Das vierte Beispiel wird aus Platzgründen in dimensionsloser Form angegeben. Man erkennt, dass die Ortskurve des Frequenzganges für ein P-Glied zu einem Punkt entartet und dass die zu Integriereru gehörenden Ortskurven die negative imaginäre Achse bedecken. Der Frequenzgang eines Systems erster Ordnung $T\dot{y} + y = Ku$ ist weniger einfach darzustellen. Die Mathematik zeigt, dass alle komplexen Funktionen

$$A(\omega) = \frac{j\omega \cdot a + b}{j\omega \cdot c + d} \quad (5.19)$$

für beliebige reelle Werte der Koeffizienten a, b, c, d Kreise in der komplexen Ebene beschreiben, deren Mittelpunkte auf der reellen Achse liegen. Durch diese Aussage und die Grenzwerte des Frequenzganges für große beziehungsweise kleine Werte der Frequenz wird der Verlauf der Ortskurve im vorliegenden Fall festgelegt. Durch Erweitern mit dem Konjugiert-Komplexen des Nenners des Frequenzganges des Systems erster Ordnung

$$G(j\omega) = K \cdot \frac{1}{1 + j\omega T} \cdot \frac{1 - j\omega T}{1 - j\omega T} = K \cdot \frac{1 - j\omega T}{1 + \omega^2 T^2} \quad (5.20)$$

erhält man einen Ausdruck mit reelem Nenner, der leicht in einen Real- und einen Imaginärteil zu zerlegen ist. Diese können falls erforderlich für

Differentialgleichung Frequenzgang	Ortskurve	Interpretation
$y = K \cdot u$ $G(j\omega) = K$		P-Glied alle ω gleich übertragen keine Phasenverschiebung
$\dot{y} = K_I \cdot u$ $G(j\omega) = \frac{K_I}{j\omega}$		I-Glied Amplitudenabsenkung für hohe und -verstärkung für niedrige ω Phasenverschiebung stets -90°
$T\dot{y} + y = K \cdot u$ $G(j\omega) = \frac{K}{Tj\omega + 1}$		Sinkender Amplitudengang für steigende ω Phasenverschiebung zwischen 0° und -90°
dimensionslos: $\ddot{y} + 2\dot{y} + 2y = u - \dot{u}$ $G(j\omega) = \frac{-j\omega + 1}{(j\omega)^2 + 2(j\omega) + 2}$		Halbierte Amplitude für $\omega=2$ bei einer Phasenverschiebung -180° Phasenverschiebung zwischen 0° und -270°

Tabelle 5-1: Beispiel für Frequenzgänge und ihre Ortskurven

eine genügende Zahl von Frequenzwerten ausgerechnet werden. Man erhält

$$G(j\omega) = \frac{K}{1 + \omega^2 T^2} - j \frac{K\omega T}{1 + \omega^2 T^2} \quad (5.21)$$

und erkennt, dass der Imaginärteil des Frequenzganges für alle positiven Frequenzen negativ ist. Der Startwert für $\omega = 0$ entspricht bei stabilen Systemen gemäß der Grenzwertsätze der statischen Verstärkung des Systems. Man erkennt ferner (Tab. 5-1), dass die Ortskurve des Frequenzganges in Richtung wachsender ω -Werte im Uhrzeigersinn durchlaufen wird. Diese Eigenschaft ist für viele Systeme typisch, da aufgrund der kausalen Systemen innenwohnenden Trägheit Ausgangssignale im eingeschwungenen Zustand den Eingangssignalen meist verzögert hinterherlaufen. Dies führt oft auf eine negative Phasenverschiebung und somit auf eine im Uhrzeigersinn verlaufende Ortskurve. Dieses Phänomen wird ausführlicher in Rahmen von Tab. 6-2 diskutiert.

Die meisten technisch interessanten Frequenzgänge werden durch Ortskurven dargestellt, die nur durch punktweise Auswertung von Gleichungen für Real- und Imaginärteil analog zu Gl.(5.21) zu bestimmen sind. In solchen Fällen wird die Ortskurve mit Rechnerunterstützung gezeichnet, wie bei dem vierten Beispiel in Tab. 5-1 geschehen. Die graphische Darstellung als Ortskurve ist ein nennenswerter Vorteil des Frequenzgangs gegenüber der Übertragungsfunktion. Dies wird dadurch möglich, dass der Frequenzgang Funktion einer einzigen reellen Variablen ist, während Übertragungsfunktionen von der komplexen Variablen $s = \sigma + j\omega$ abhängen.

5.3 Bode-Diagramm

Neben der Ortskurvendarstellung wird eine logarithmische Darstellung von Frequenzgängen häufig benutzt.

Bode-Diagramm

Im sogenannten Bode-Diagramm werden Betrag und Phasenwinkel des Frequenzganges als Funktionen der positiv angenommenen Frequenz separat voneinander dargestellt. Dabei sind die Frequenzachsen und die Betragsachse logarithmisch geteilt, der Phasenwinkel wird linear aufgetragen.

Ein Beispiel ist in Bild 5-4 gezeigt. Die Darstellung von Frequenzgängen im Bode-Diagramm hat gegenüber der Ortskurvendarstellung einige Vorteile. Zum einen ermöglicht die kontinuierliche Frequenzachse des Bode-Diagramms im Gegensatz zu den diskreten und oft rudimentären Frequenz-

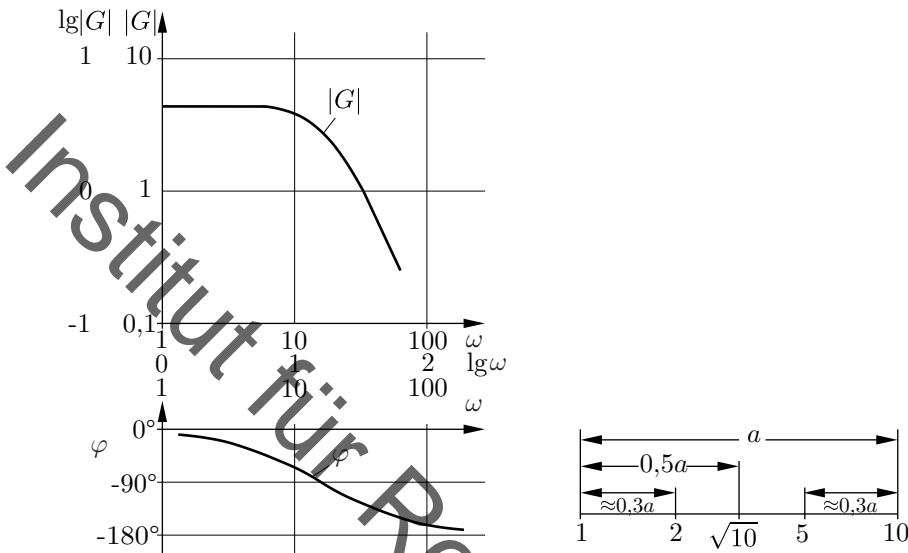


Bild 5-4: Bode-Diagramm und logarithmische Teilung

angaben der Ortskurve ein passgenaues Ablesen der gesuchten Frequenz. Zum anderen lassen sich für sehr viele Frequenzgänge entsprechende Konstruktionsvorschriften angeben, die ohne viel Rechenarbeit zu hinreichend genauen Darstellungen im Bode-Diagramm führen. Insbesondere lässt sich die häufig vorkommende Multiplikation von Frequenzgängen im Bode-Diagramm recht einfach durchführen: Da der Frequenzgang durch Betrag und Phasenwinkel

$$G(j\omega) = |G| \cdot e^{j\varphi} \quad (5.22)$$

dargestellt wird, erhält man als Produkt zweier Frequenzgänge G_1 und G_2

$$G = G_1 \cdot G_2 = |G_1| \cdot e^{j\varphi_1} \cdot |G_2| \cdot e^{j\varphi_2} = |G_1| \cdot |G_2| \cdot e^{j(\varphi_1 + \varphi_2)} \quad (5.23)$$

und somit den Betrag zu

$$|G| = |G_1| \cdot |G_2| \quad (5.24)$$

und wegen der logarithmischen Teilung der $|G|$ -Achse

$$\lg |G| = \lg |G_1| + \lg |G_2| \quad . \quad (5.25)$$

Für die Winkel gilt nach Gl.(5.23)

$$\varphi = \varphi_1 + \varphi_2 \quad . \quad (5.26)$$

Man erkennt, dass durch die gewählte Darstellung die Multiplikation in eine graphische Addition übergeführt wird. Für die praktische Handhabung logarithmischer Darstellungen ist die Kenntnis der ungefähren Aufteilung der logarithmischen Einheit nach Bild 5-4 hilfreich.

Frequenzgang	Bode-Diagramm
$G(j\omega) = K$ $ G = K$ $\varphi = 0^\circ$	
$G(j\omega) = \frac{K_I}{j\omega}$ $ G = \frac{K_I}{\omega}$ $\varphi = -90^\circ$	
$G(j\omega) = \frac{K}{Tj\omega + 1}$ $ G = \frac{K}{\sqrt{1 + (\omega T)^2}}$ $\varphi = \arctan(-\omega T)$	

Tabelle 5-2: Beispiele für Frequenzgänge und Bode-Diagramme

Tab. 5-2 zeigt einige Beispiele für die Darstellung von Frequenzgängen im Bode-Diagramm. Man erkennt, dass der Amplitudengang eines I-Elements einer Geraden mit der Steigung -1 entspricht, sofern gleiche Maßstäbe für Betrag und Frequenz vorliegen. Die Linie $|G| = 1$ wird bei der Frequenz K_1 geschnitten. Der Phasengang ist konstant bei -90° .

Der Frequenzgang der Differentialgleichung erster Ordnung $T\dot{y} + y = Ku$ lässt sich für Frequenzen weit unterhalb beziehungsweise oberhalb der Eckkreisfrequenz

$$\omega_E = \frac{1}{T} \quad (5.27)$$

durch Asymptoten annähern. Die Asymptoten des Amplitudenganges sind für kleine Werte der Frequenz

$$\lg |G(\omega \ll 1/T)| \simeq \lg K \quad (5.28)$$

und für große Frequenzwerte

$$\lg |G(\omega \gg 1/T)| \simeq \lg K - \lg(\omega T) = \lg K - \lg T - \lg \omega \quad , \quad (5.29)$$

das ist eine Gerade mit der Steigung -1 durch den Punkt $\omega = 1/T$, $|G| = K$. Für die Eckkreisfrequenz selbst ist der Betrag

$$\lg(|G(\omega = 1/T)|) = \lg K - 0,5 \cdot \lg 2 \simeq \lg K - 0,15 \quad . \quad (5.30)$$

Zur Konstruktion des Phasenganges dienen die Grenzwerte null und -90° für kleine beziehungsweise große Frequenzwerte und für die Eckkreisfrequenz

$$\varphi(\omega = 1/T) = -45^\circ \quad . \quad (5.31)$$

Die Kurve ist punktsymmetrisch zum Punkt $(\omega_E, -45^\circ)$; die in diesem Punkt angelegte Wendetangente schneidet auf den Asymptoten näherungsweise Abschnitte der Länge 0,7 logarithmische Einheiten ab. Ein genauere Betrachtung findet sich in Abschnitt 7.3 in Tab. 7-4.

Das Bode-Diagramm wird oft dazu benutzt, das Produkt verschiedener Frequenzgänge graphisch zu bestimmen, weil die Multiplikation durch die gewählte Darstellungsform leicht auszuführen ist. Am Beispiel der Multiplikation der Frequenzgänge eines Integrators G_1 und eines Systems erster

Ordnung G_2 soll gezeigt werden, wie eine solche Multiplikation im Bode-Diagramm durchgeführt wird (Bild 5-5).

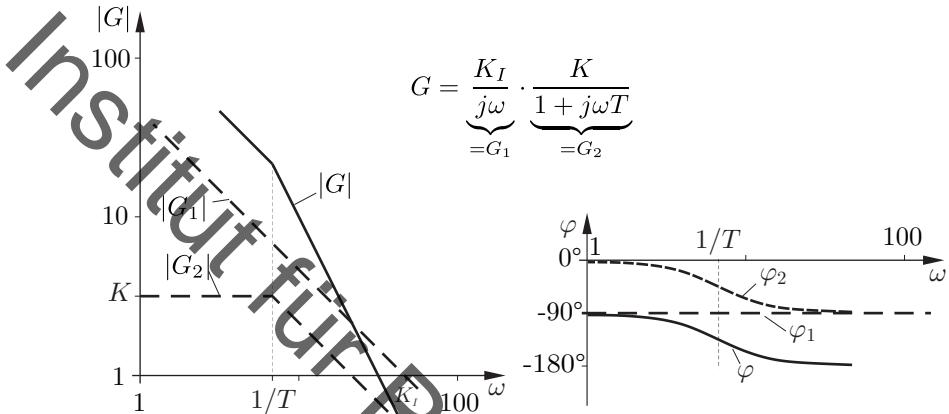


Bild 5-5: Multiplikation von Frequenzgängen

Man erkennt in Bild 5-5, dass die resultierenden Verläufe $|G|$ und φ durch graphische Addition der mit unterbrochenen Linien dargestellten Verläufe von $|G_1|$ und $|G_2|$ beziehungsweise φ_1 und φ_2 entstehen. Dabei ist zu beachten, dass in der Darstellung des Amplitudenganges die Linie $|G| = 1$ Bezugslinie ist, sodass bei der graphischen Addition Abstände zu Punkten oberhalb dieser Linie positiv und solche zu Punkten unterhalb dieser Bezugslinie negativ zu werten sind. Für den Phasengang gilt das Entsprechende für die Nulllinie. Diese Verfahrensweise wird dadurch noch erleichtert, dass für viele regelungstechnischen Fragestellungen eine Darstellung des Amplitudenganges durch Geraden – die Asymptoten – genügend genau ist. Eine Asymptotendarstellung des Phasenganges ist dagegen nur in Sonderfällen ausreichend.

5.4 Fourier-Transformation

In den vorangegangen Abschnitten wurde der Frequenzgang über das Verhältnis der Zeiger im eingeschwungenen Zustand definiert. Hierbei wurde festgestellt, dass der Frequenzgang der Übertragungsfunktion mit $s = j\omega$

entspricht. Setzt man diese Variable in das Integral der Laplace-Transformation in Gl.(4.39) ein, so erhält man die Fourier²-Transformation:

Fourier-Transformation

Gegeben sei eine absolut integrierbare Funktion $f(t)$ mit $f(t) = 0$ für $t < 0$ und $t \in \mathbb{R}$. Die Fourier-Transformation $\mathcal{F}\{f\}$ ordnet dieser Funktion $f(t)$ im Zeitbereich eine neue Funktion $F(j\omega)$ im Frequenzbereich über die folgende Formel zu

$$F(j\omega) = \int_{-\infty}^{\infty} f(t) \cdot e^{-j\omega t} dt = \mathcal{F}\{f(t)\} \quad . \quad (5.32)$$

Die Fourier-Transformation ist eindeutig umkehrbar mit der inversen Fourier-Transformation

$$f(t) = \begin{cases} \frac{1}{2\pi j} \int_{-\infty}^{+\infty} F(j\omega) \cdot e^{j\omega t} d\omega & \text{für } t \geq 0 \\ 0 & \text{für } t < 0 \end{cases} = \mathcal{F}^{-1}\{F(j\omega)\} \quad (5.33)$$

Der Frequenzgang lässt sich dann analog zur Übertragungsfunktion auch als Quotient der Fourier-Transformierten von Ein- und Ausgangssignal deuten. Das Integral der Fourier-Transformation konvergiert nur im Fall von Signalen, die absolut integrierbar sind. Folglich sind als Systeme auch nur stabile Systeme fouriertransformierbar, da nur solche Systeme absolut integrierbare Impulsantworten besitzen. Die Laplace-Transformation kann in diesem Sinne als Erweiterung der Fourier-Transformation auf instabile Systeme und deren Signale verstanden werden.

Frequenzgang für instabile Systeme

Obgleich der eingeschwungene Zustand für Systeme, die nicht stabil sind, nicht definiert ist, so kann diesen Systemen aufgrund der Ähnlichkeit von Übertragungsfunktion und Frequenzgang ein Frequenzgang zugeordnet werden. Dabei entspricht der Frequenzgang einer Spezialisierung der Übertragungsfunktion auf die imaginäre Achse mit $G(s = j\omega)$. Die Inter-

²Jean Baptiste Joseph Fourier (1768-1830), französischer Mathematiker [13]

interpretation des Frequenzgangs über den eingeschwungenen Zustand geht dabei allerdings verloren.

Die Fouriertransformation bietet noch eine weitere Interpretationsebene, da sie angibt, welche Frequenzanteile in den transformierten Signalen wie stark vertreten sind. Dies sieht man am leichtesten durch die Betrachtung der inversen Fourier-Transformation für $t > 0$ ein. Diese hat strukturell sehr große Ähnlichkeit zur aus der Mathematik bekannten Fourier-Reihe:

$$f(t) = \frac{1}{2\pi j} \int_{-j\infty}^{+j\infty} F(j\omega) \cdot e^{j\omega t} d\omega \quad \Leftrightarrow \quad f(t) = \sum_{k=-\infty}^{\infty} c_k \cdot e^{j\omega kt} \quad (5.34)$$

Die Fourierreihe ist dabei nur für periodische Signale definiert, was sich darin äußert, dass zur Darstellung der Funktion eine diskrete Menge von Schwingungen $e^{j\omega kt}$, die alle Vielfache der Grundschwingung des Signals sind, ausreicht. Im Falle nicht-periodischer Signale, wie sie wegen der Forderung der Kausalität $f(t) = 0$ für $t < 0$ in der Regelungstechnik stets vorliegen, reicht eine solche diskrete Menge an Frequenzen nicht mehr aus. An ihre Stelle tritt eine kontinuierliche Menge an Frequenzen, wie sie in der Fouriertransformation abgebildet werden. Ebenso wie c_k das Gewicht des k -ten Fourierreihenkoeffizienten beschreibt, ist $F(j\omega)$ das Gewicht des Frequenzanteils mit der Frequenz ω .

Frequenzspektrum

Gegeben sei ein absolut integrierbares Signal $f(t)$. Dann lässt sich dieses Signal gemäß

$$f(t) = \frac{1}{2\pi j} \int_{-j\infty}^{+j\infty} F(j\omega) \cdot e^{j\omega t} d\omega$$

in seine unterschiedlichen Signalanteile zerlegen. $F(j\omega)$ heißt *Frequenzspektrum* (oder kurz: Spektrum) des Signals und gibt an, wie stark die zu der jeweiligen Frequenz ω gehörige Schwingung $e^{j\omega t}$ in das Gesamtsignal eingeht.

Die Forderung nach Integrierbarkeit des Signals $f(t)$ kann hier als Bedingung verstanden werden, dass die Signalenergie von $f(t)$ endlich sein muss.

Das Frequenzspektrum eines Signals lässt sich direkt über die Fourier- bzw. Laplace-Transformation mit $s = j\omega$ berechnen. Die Anwendung des Spektrums ist vielfältig: In der Akustik zeigt das Frequenzspektrum an, welche Frequenzen wie stark in einem akustischen Signal vertreten sind. Die konkrete Zusammensetzung ist dabei beispielsweise für jedes Musikinstrument unterschiedlich, was im menschlichen Gehör in unterschiedliche Tonfarbungen trotz gleicher Tonhöhe umgesetzt wird. In der Optik beschreibt das Farbspektrum, welche Spektralfarbe wie stark in einem optischen Signal vertreten sind, was beispielsweise Schlüsse auf die chemische Zusammensetzung zulässt.

Wendet man die Interpretation als Frequenzspektrum auf die in Tab. 5-2 bereits berechneten Frequenzgänge an, so beschreiben die dort gezeigten Verläufe die Frequenzspektren der Gewichtsfunktionen der zugehörigen Systeme. Im Falle des P-Elements ist die Gewichtsfunktion ein Impuls mit einem konstanten Amplitudengang. Das zeigt, dass der Impuls alle Frequenzen mit gleicher Gewichtung enthält. Daher eignet er sich in besonderer Weise zur Systemidentifikation, da er die Systemdynamik über alle Frequenzen gleichermaßen anregt. Beim Integrator mit der Sprungfunktion als Gewichtsfunktion ergibt sich zunächst scheinbar ein Spektrum, das niedrige Frequenzen übermäßig stark enthält und in dem hohe Frequenzen eine untergeordnete Rolle spielen. Allerdings ist die Sprungfunktion nicht absolut integrierbar, da die Fläche unter ihr nicht endlich ist; daher ist die Fouriertransformation und das Frequenzspektrum hier nicht definiert. Die Exponentialfunktion $e^{-t/T}$ aus Bild 4-3 ist die Gewichtsfunktion des Systems erster Ordnung und enthält daher gemäß Tab. 5-2 vorwiegend Signalelemente mit Frequenzen $\omega < \omega_E = 1/T$.

5.5 Filter

Die für die Exponentialfunktion beschriebene Eigenschaft, dass der Amplitudengang bis zur Eckkreisfrequenz $\omega_E = 1/T$ nahezu konstant verläuft, danach aber mit Steigung -1 abfällt, lässt sich zur Auslegung von Filtern nutzen.

Filter, das

Ein Filter ist ein System, dass ein Eingangssignal abhängig von seiner Frequenz in Amplitude und Phaselage verändert.

Filter sind spezielle dynamische Systeme. Bei einem Filter haben Eingangsgröße (das Eingangssignal) und Ausgangsgröße (das veränderte Eingangssignal) dieselbe physikalische Einheit. Filter werden genutzt, um beispielsweise in einem verrauschten Messsignal das Rauschen zu verringern oder eine konstante Verschiebung im Messsignal, welche aus einer ungenauen Kalibrierung herröhrt, zu entfernen. Aus den klassischen Anwendungsfällen für Filter haben sich dabei vor allem die folgenden drei Klassen von Filtern herauskristallisiert, wobei die Bezeichnungen nicht nur auf Filter, sondern generell auf Systeme angewandt werden.

Tiefpass, Hochpass, Bandpass

Ein System oder Filter wird *Tiefpass*(filter) genannt, wenn sinusförmige Eingangssignale mit Frequenzen unterhalb der Grenzfrequenz $\omega \ll \omega_g$ im eingeschwungenen Zustand in ihrer Amplitude nahezu unverändert bleiben, während für Frequenzen oberhalb der Grenzfrequenz $\omega \gg \omega_g$ ihre Amplitude verringert wird.

Ein System oder Filter wird *Hochpass*(filter) genannt, wenn sinusförmige Eingangssignale mit Frequenzen oberhalb der Grenzfrequenz $\omega \gg \omega_g$ im eingeschwungenen Zustand in ihrer Amplitude nahezu unverändert bleiben, während für Frequenzen unterhalb der Grenzfrequenz $\omega \gg \omega_g$ ihre Amplitude verringert wird.

Die Kombination aus einem Tiefpass und einem Hochpass heißt *Bandpass*(filter), für den nur Eingangssignale mit Frequenzen im Bereich $\omega_1^* \ll \omega \ll \omega_2^*$ nahezu unverändert bleiben.

Die Bezeichnung als „Tiefpass“ röhrt daher, dass das System tiefe (also niedrige) Frequenzen ungeschwächt passieren lässt, während die hohen Frequenzen herausgefiltert werden. Die Bezeichnungen als „Hochpass“ und „Bandpass“ ergeben sich analog. Die englische Bezeichnung der Grenzfrequenz als „cut-off frequency“ unterstreicht dabei das beschriebene Verhalten. Aufgrund des Überlagerungsprinzips gelten die Aussagen nicht nur für sinusförmige Signale, sondern auch für jedes Signal, dass sich als Linear-kombination sinusförmiger Signale darstellen lässt. Eine solche Darstellung ist über die Fourier-Transformation für alle absolut integrierbaren Signale möglich. Streng genommen gelten die Zusammenhänge nur im eingeschwungenen Zustand. In vielen Fällen wird sich aber eine entsprechende Abschwächung der Amplituden der Signalanteile bereits nach recht kur-

er Zeit einstellen, sodass die Filterung auch auf kurzen Zeitskalen sinnvoll betrieben werden kann.

Wird ein LTI-System $u \mapsto y$ mit dem Frequenzgang $G(j\omega)$ als Filter eingesetzt, so ergibt sich wie zuvor $Y(j\omega) = G(j\omega) \cdot U(j\omega)$. Folglich ergibt die Fourier-Transformation (und damit das Frequenzspektrum) des gefilterten Ausgangssignals sich aus der Multiplikation des Frequenzgangs mit dem Frequenzspektrum des Eingangssignals. Daher kann man dem Amplitudengang direkt entnehmen, welche Frequenzen durch das System herausgefiltert werden, welche unverändert bleiben und welche ggf. verstärkt werden. Die generischen Verläufe des Amplitudengangs von Tiefpass, Hochpass und Bandpass im Bode-Diagramm ergeben sich daher wie in Bild 5-6 gezeigt.

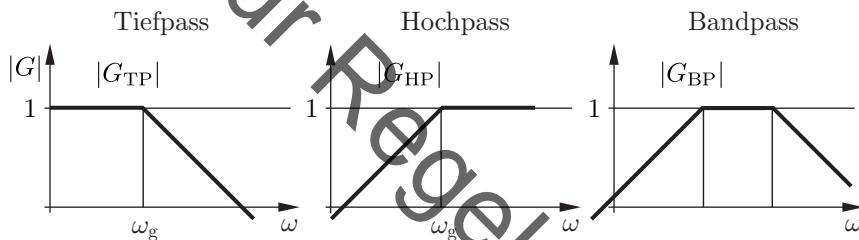
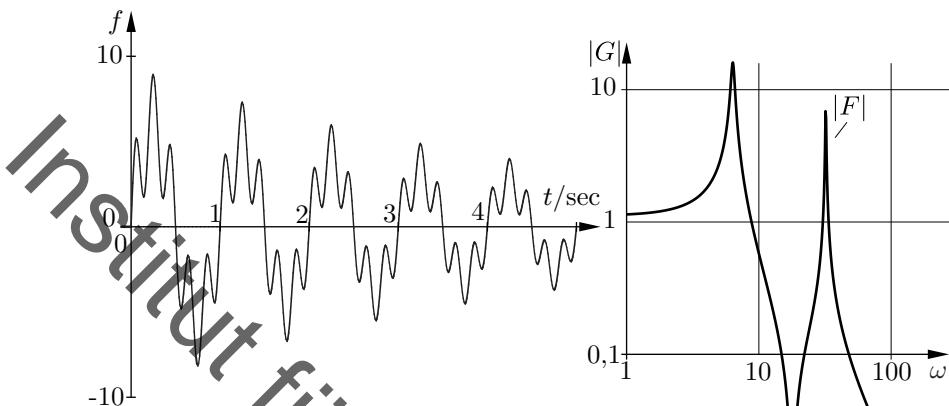


Bild 5-6: Generische Amplitudengänge der verschiedenen Filtertypen

Als Beispiel für die Verwendung von Frequenzspektren und Filtern wird das in Bild 5-7 gezeigte Signal $f(t)$ betrachtet. Aus dem Zeitverlauf kann man erkennen, dass $f(t)$ für $t > 0$ vor allem einen niederfrequenten mit Periodendauer $T \approx 1$ sec und einen hochfrequenten Signalanteil mit $T \approx 0,2$ sec besitzt. Diese durch scharfes Hinsehen extrahierten Informationen findet man rechnerisch auch im zugehörigen Frequenzspektrum $F(j\omega)$ wieder, welches durch Fouriertransformation berechnet werden kann und in Bild 5-7 im Amplitudengang gezeigt ist. Stellt sich nun die Aufgabe, aus dem Signal den hochfrequenten Signalanteil zu entfernen, so muss ein Tiefpassfilter mit der Grenzfrequenz $2\pi \text{ sec}^{-1} < \omega_g < 10\pi \text{ sec}^{-1}$ gewählt werden. Die Amplitudengänge eines beispielhaften Tiefpassfilters $G(j\omega)$ und des gefilterten Spektrums $\tilde{F}(j\omega)$ sind zusammen mit dem gefilterten Signal $\tilde{f}(t)$ in Bild 5-8 gezeigt. Es ist klar zu sehen, dass das Filter die gegebene Aufgabe erfüllen kann und der Frequenzanteil bei $10\pi \text{ sec}^{-1}$ stark gedämpft wird. Allerdings

Bild 5-7: Signal $f(t)$ und Amplitudengang des Frequenzspektrum $F(j\omega)$

tritt auch eine negative Phasenverschiebung im gefilterten Signal auf. Gefilterte Signale sehen visuell ansprechender aus. Es ist daher verlockend, Messsignale stark zu filtern. Gleichzeitig kann die Phasenverschiebung aber zur Instabilität geschlossener Regelkreise führen, weshalb hier Vorsicht geboten ist.

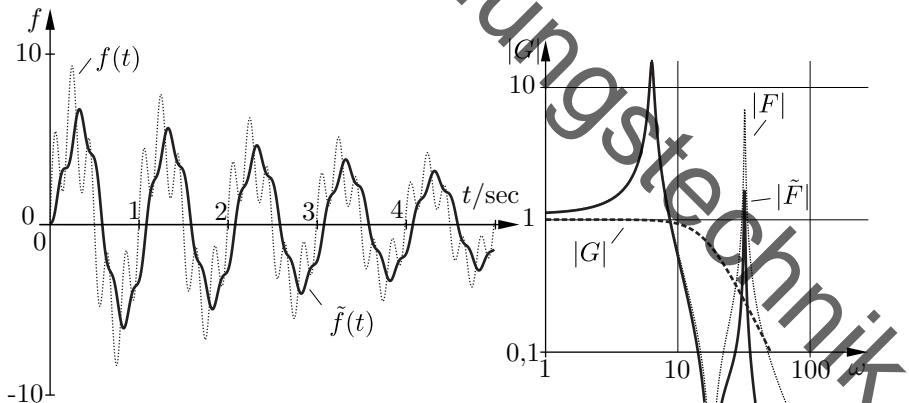


Bild 5-8: Original und Filterung im Zeit- und Frequenzbereich

6 Verschaltungen von Systemen

In den vorangehenden Kapiteln sind wiederholt einzelne Regelkreisglieder und ihre dynamischen Eigenschaften als Beispiele behandelt worden. Dies hatte vorrangig das Ziel, die neu eingeführten Beschreibungsmittel für dynamische Systeme zu erläutern.

Nachfolgend soll eine systematische Vorgehensweise präsentiert werden, die für ein allgemeines LTI-System die unterschiedlichen Beschreibungsformen effizient nutzt. Ausgangspunkt ist dabei die Idee, ein kompliziertes LTI-System als Verschaltung einfacher Teilsysteme aufzufassen. Diese einfachen Teilsysteme bilden dann eine überschaubare Menge von Bausteinen, aus denen ein allgemeines LTI-System durch eine geeignete Verschaltung hervorgeht.

Auf dieselbe Art und Weise lassen sich komplexe Systeme, die aus vielen unübersichtlich zusammengesetzten Teilsystemen bestehen, zu einer kompakteren Beschreibung in Form eines einziges Systems zusammenfassen.

Hierfür ist es notwendig, die verschiedenen Verschaltungsmöglichkeiten von Systemen und die zugehörigen mathematischen Operationen zu verstehen.

6.1 Zusammenfassen von Teilsystemen

Viele regelungstechnische Zwecke erfordern das Modell einer Anordnung von signalübertragenden Gliedern. Dieses Modell kann man dadurch erhalten, dass man die aus einzelnen Teilsystemen bestehende Schaltung als ein einziges Übertragungsglied mit einem Eingang und einem Ausgang interpretiert. Dabei ist es ratsam, die einzelnen Teilsysteme im Bildbereich zu beschreiben, da auf diese Weise die Anordnung durch algebraische Beziehungen beschrieben wird.

In Tab. 6-1 sind die drei wichtigsten aus zwei Übertragungsgliedern bestehenden Schaltungen abgeleitet und zusammengestellt. Hierbei kann G sowohl ein Frequenzgang als auch eine Übertragungsfunktion mit der Zuordnung $u \mapsto y = G \cdot u$ sein. Streng genommen sind u und y hier als Laplace-Transformierte als $U(s)$ und $Y(s)$ zu setzen. Es hat sich aber eingebürgert, auch einfach $y = G \cdot u$ zu schreiben, obgleich mathematisch eigentlich eine Faltung und keine Multiplikation verwendet wird.

Frequenzgang Übertragungsfunktion	Wirkungsplan
$\begin{aligned}y &= y_1 + y_2 \\&= (G_1 + G_2) \cdot u\end{aligned}$	<p>Parallelschaltung</p>
$\begin{aligned}y &= G_1 v \\&= G_1 \cdot (G_2 \cdot u) \\&= G_1 \cdot G_2 \cdot u\end{aligned}$	<p>Reihenschaltung</p>
$\begin{aligned}y &= G_v \cdot (u \mp v) \\&= G_v \cdot (u \mp G_r \cdot y) \\&\Rightarrow y = \frac{G_v}{1 \pm G_v \cdot G_r} u\end{aligned}$	<p>Rückkopplung</p>

Tabelle 6-1: Schaltungen mit zwei Übertragungsgliedern

Die Rechenvorschriften für die Verschaltungen aus Tab. 6-1 lassen sich auf Basis der Signale und der Definition des Wirkungsplans herleiten.

Parallelschaltung, Reihenschaltung und Rückführung

Im Falle der Parallelschaltung ergibt sich

$$y = y_1 \pm y_2 = G_1 \cdot u \pm G_2 \cdot u = \underbrace{(G_1 \pm G_2)}_{=G} \cdot u \quad , \quad (6.1)$$

für die Reihenschaltung

$$y = G_1 \cdot v = G_1 \cdot (G_2 \cdot u) = \underbrace{G_1 \cdot G_2}_{=G} \cdot u \quad (6.2)$$

und für die Rückführung

$$\begin{aligned} y &= G_v \cdot (u \mp v) = G_v \cdot (u \mp G_r \cdot y) = G_v \cdot u \mp G_v \cdot G_r \cdot y \\ \Leftrightarrow y \pm G_v \cdot G_r \cdot y &= G_v \cdot u \Leftrightarrow y = \underbrace{\frac{G_v}{1 \pm G_v \cdot G_r}}_{=G} u \quad . \end{aligned} \quad (6.3)$$

Der Nenner in der Gleichung für die Rückkopplung wird zu „ $1 + G_v \cdot G_r$ “, wenn der Regelkreis über ein „-“ geschlossen ist. Andernfalls wird er zu „ $1 - G_v \cdot G_r$ “. Der Vorwärtszweig G_v entspricht der direkten Wirkung der Eingangs- auf die Ausgangsgröße ohne Rückführung.

Aus den Verschaltungen kann man direkt ermitteln, inwiefern sich die Stabilität der Teilsysteme auf die Stabilität des Gesamtsystems überträgt. Hierbei muss angenommen werden, dass es sich nicht nur bei den einzelnen G_i , sondern auch bei dem Gesamtsystem G um eine minimale Realisierung handelt – es treten also keine Pol-Nullstellen-Kürzungen auf. Im Falle einer Parallelschaltung ergibt sich mit dem Zähler Z_i und dem Nenner N_i , mit $i = 1, 2$,

$$G_1 \pm G_2 = \frac{Z_1}{N_1} \pm \frac{Z_2}{N_2} = \frac{Z_1 \cdot N_2 \pm Z_2 \cdot N_1}{N_1 N_2} \quad . \quad (6.4)$$

Die Pole des Gesamtsystems sind also die Polstellen von G_1 und die Polstellen von G_2 . Folglich ist die Parallelschaltung genau dann stabil, wenn G_1 und G_2 stabil sind. Dies ist auch im Bezug auf die BIBO-Stabilität plausibel, da y dann beschränkt sein wird, wenn sowohl y_1 als auch y_2 beschränkt sind. Eine analoge Rechnung liefert das identische Ergebnis für die Reihenschaltung.

Für den durch die Rückführung entstehenden geschlossenen Regelkreis ergibt sich ein abweichendes Ergebnis, da die durch den Ausdruck $1 \pm G_v \cdot G_r = 0$ bestimmbaren Polstellen des rückgekoppelten Systems sich nicht in ähnlicher Weise aus den Polen und Nullstellen von G_v und G_r ableSEN lassen. Da dem Produkt von Vorwärts- und Rückwärtszweig $G_v \cdot G_r$ hier eine entscheidende Bedeutung zukommt, erhält es eine eigene Bezeichnung:

Aufgeschnittener Regelkreis

Der aufgeschnittene Regelkreis G_0 ist die negative Wirkung der Ausgangsgröße auf sich selbst. Der Name röhrt daher, dass man gedanklich den Regelkreis am Anfang der Rückführung aufschneidet (siehe Bild 6-1). Also beschreibt G_0 das Übertragungsverhalten von y_e auf y_a . Es entspricht bei einer einfachen Rückführung der Reihenschaltung des Vorwärts- und Rückwärtszweiges.

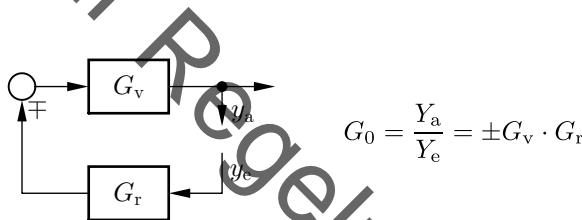


Bild 6-1: Aufgeschnittener Regelkreis

Stabilität von Verschaltungen

Eine Parallel- oder Reihenschaltung von Teilsystemen ist – unter Ausschluss von Pol-Nullstellen-Kürzungen – genau dann stabil, wenn die einzelnen Teilsysteme stabil sind.

Der durch die Rückführung entstehende geschlossene Regelkreis kann andere Stabilitätseigenschaften als der aufgeschnittene Regelkreis besitzen. Dabei kann sowohl

- der aufgeschnittene Regelkreis stabil, der geschlossene Regelkreis jedoch instabil
- der aufgeschnittene Regelkreis instabil, der geschlossene Regelkreis jedoch stabil

sein.

Die unterschiedlichen Stabilitätseigenschaften zwischen aufgeschnittenem und geschlossenem Regelkreis sind ein zentraler Vorteil der Regelung gegenüber einer Steuerung. Betrachtet man die Wirkungspläne für Steuerung und Regelung genauer, so erkennt man in Bild 1-7, dass eine Steuerung einer Reihenschaltung ohne Rückführung entspricht. Somit wird eine Steuerung niemals ein instabiles System stabilisieren können. Durch eine Regelung wie in Bild 1-6 entsteht ein geschlossener Regelkreis, der auch bei einer instabilen Regelstrecke stabil sein kann und die Regelstrecke somit stabilisiert.

Allerdings kann ein ungeschickt gewählter Regler eine bereits stabile Regelstrecke auch destabilisieren, selbst wenn der Regler ebenfalls stabil ist. Diese Gefahr ist bei einer Steuerung nicht gegeben, sofern die Steuerung selbst stabil ist.

Somit erfordert der Entwurf einer Regelung zusätzliche Aufmerksamkeit und Vorsicht im Vergleich zum Entwurf einer Steuerung. Deswegen ist es oft notwendig, die Übertragungsfunktionen des geschlossenen Regelkreises zu berechnen, wobei man grundlegend zwischen zwei verschiedenen Übertragungsfunktionen unterscheidet:

Führungs- und Störübertragungsfunktion

In dem einfachen Regelkreis in Bild 6-2 wird die Übertragungsfunktion von der Stör- auf die Regelgröße

$$S(s) = \frac{Y(s)}{Z(s)} = \frac{1}{1 + G_0(s)} = \frac{N_S(s)N_R(s)}{Z_S(s)Z_R(s) + N_S(s)N_R(s)} \quad (6.5)$$

als Störübertragungsfunktion oder auch *Sensitivität* bezeichnet.

Die Übertragungsfunktion von der Führungs- auf die Regelgröße

$$T(s) = \frac{Y(s)}{W(s)} = \frac{G_0(s)}{1 + G_0(s)} = \frac{Z_S(s)Z_R(s)}{Z_S(s)Z_R(s) + N_S(s)N_R(s)} \quad (6.6)$$

wird als Führungsübertragungsfunktion oder auch *komplementäre Sensitivität* bezeichnet.

Die Bezeichnung als Sensitivitätsfunktion röhrt dabei historisch daher, dass die Sensitivität das Verhältnis der relativen Änderung des Führungsverhaltens zur relativen Änderung der Regelstreckendynamik beschreibt.

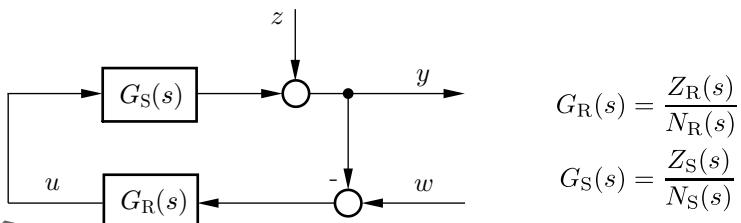


Bild 6-2: Standard-Regelkreis

Die komplementäre Sensitivität $T(s)$ wird deshalb als komplementär bezeichnet, weil sie die Sensitivitätsfunktion $S(s)$ zu eins ergänzt:

$$S(s) + T(s) = 1. \quad (6.7)$$

Offenbar unterscheiden sich beide Übertragungsfunktionen des geschlossenen Regelkreises nur in ihren Nullstellen. Daher ist es z. B. für die Beurteilung der Stabilität irrelevant, welche der Übertragungsfunktionen des geschlossenen Regelkreises zu deren Überprüfung herangezogen werden.

Die vorgestellten Rechenregeln dienen neben der überschlägigen Beurteilung der Stabilität von Verschaltungen vorwiegend der Bestimmung der Übertragungsfunktion oder des Frequenzgangs komplizierter Schaltungen wie z. B. der in Bild 6-3, indem sie schrittweise Teile der Schaltung zusammenfassen.

Im Falle mehrerer verschalteter Rückführungen empfiehlt es sich oft – ähnlich wie in den Herleitungen in Tab. 6-1 – ausgewählte in der Schaltung auftretende Größen zu benennen und zunächst Übertragungsfunktionen zu diesen Zwischengrößen zu bestimmen. Die zusätzlich eingeführten Zwischengrößen können dann in einem zweiten Schritt durch die algebraischen Beziehungen der Übertragungsfunktion ersetzt werden.

Dies soll am Beispiel von Bild 6-3 illustriert werden. Da zwei verschaltete Rückführungen vorliegen, wird die Eingangsgröße in die innere Rückführung als Zwischengröße v definiert. Nun lassen sich G_1 und G_2 sowie G_3 und G_5 als Reihen- bzw. Parallelschaltung zusammenfassen. Somit erhält man mit

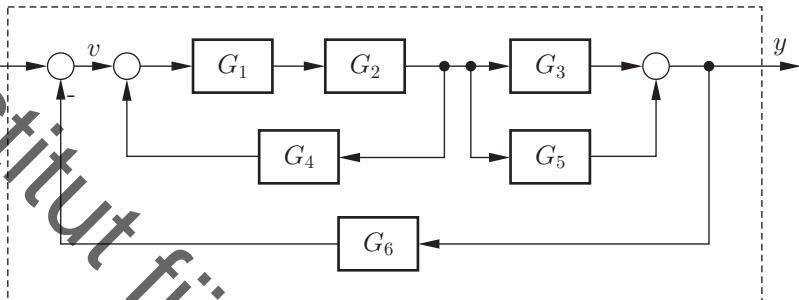


Bild 6-3: Schaltung mit mehreren Übertragungsgliedern

der Rechenregel für Rückführungen sowie einer weiteren Reihenschaltung

$$y = \underbrace{\frac{G_1 G_2}{1 - G_1 G_2 G_4} \cdot (G_3 + G_5) \cdot v}_{=\tilde{G}} . \quad (6.8)$$

Es verbleibt eine letzte Rückführung, die bereits in Standardform vorliegt, womit sich die Gesamtübertragungsfunktion

$$G = \frac{Y}{U} = \frac{\tilde{G}}{1 + \tilde{G} \cdot G_6} = \frac{G_1 G_2 (G_3 + G_5)}{1 - G_1 G_2 G_4 + G_1 G_2 G_6 (G_3 + G_5)} \quad (6.9)$$

ergibt.

Das Zusammenfassen von Teilsystemen im Bildbereich ist in den allermeisten Fällen schneller und weniger fehleranfällig als ein entsprechendes Zusammenfassen im Zeitbereich, also auf Ebene der Differentialgleichungen. Dabei wird der Vorteil ausgenutzt, dass die Anordnung im Bildbereich durch algebraische Beziehungen beschrieben wird. Wird eine Differentialgleichung als Ergebnis gesucht, so kann man diese direkt aus der Übertragungsfunktion ablesen.

Auch Anordnungen mit mehreren Eingangsgrößen lassen sich auf diese Weise behandeln, wie das nächste Beispiel zeigen soll. Gesucht ist die gesamte

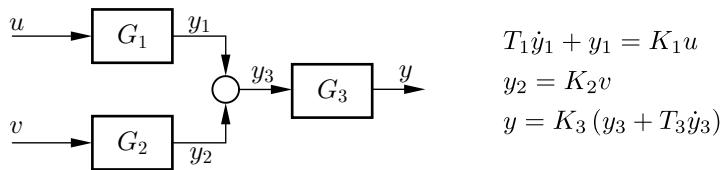


Bild 6-4: Parallelschaltung mit zwei Eingängen

Differentialgleichung mit u und v als Eingangs- und y als Ausgangsgrößen mit den gegebenen Differentialgleichungen von G_1 , G_2 und G_3 . Zur Lösung dieser Aufgabe ist ein Umweg über den Bildbereich anzuraten. Aus den zu Bild 6-4 gehörenden Differentialgleichungen wird dabei im Bildbereich

$$\begin{aligned}
 Y_1(s) &= G_1(s) \cdot U(s) = \frac{K_1}{1+sT_1} U(s) \\
 Y_2(s) &= G_2(s) \cdot V(s) = K_2 \cdot V(s) \\
 Y(s) &= G_3(s) \cdot Y_3(s) = K_3(1+sT_3) \cdot Y_3(s) \quad .
 \end{aligned} \tag{6.10}$$

Zusammengefasst erhält man daraus

$$Y(s) = \underbrace{K_3(1+sT_3)}_{G_3} \left[\underbrace{\frac{K_1}{1+sT_1}}_{G_1} \cdot U(s) + \underbrace{K_2}_{G_2} \cdot V(s) \right] . \tag{6.11}$$

Multipliziert man auf beiden Seiten mit dem Nenner, so erhält man eine Form, aus der man die Differentialgleichung rasch bestimmen kann:

$$T_1 \dot{y} + y = K_1 K_3 u + K_1 K_3 T_3 \dot{u} + K_2 K_3 v + K_2 K_3 (T_1 + T_3) \dot{v} + K_2 K_3 T_1 T_3 \ddot{v} . \tag{6.12}$$

Dieses Ergebnis kann man auch durch Einsetzen der Zusammenhänge direkt in die Differentialgleichungen im Zeitbereich erhalten. Insbesondere bei Rückführungen fallen die dabei notwendigen Rechnungen aber komplizierter aus.

6.2 Zerlegung in einfache Elemente

Anstatt eine Verschaltung von Teilsystemen zu einem Gesamtsystem zusammenzufassen, ist es ebenso möglich, ein vorliegendes (Gesamt-)System in eine Verschaltung von Teilsystemen zu zerlegen. Eine geschickte Wahl dieser Teilsysteme erlaubt dabei erhellende Blicke in das Verhalten des Gesamtsystems.

Hierzu wird die in Pol- und Nullstellen zerlegte Übertragungsfunktion eines SISO-Systems wie in Gl.(4.80) betrachtet:

$$G(s) = K \cdot \frac{(s - \eta_1) \cdot (s - \eta_2) \dots (s - \eta_m)}{(s - \lambda_1) \cdot (s - \lambda_2) \dots (s - \lambda_n)} . \quad (6.13)$$

Offenbar lässt sich $G(s)$ als eine Multiplikation von einzelnen Übertragungsfunktionen auffassen, die allesamt nur eine Pol- oder nur eine Nullstelle aufweisen. Im Allgemeinen gilt zwar η_i und $\lambda_i \in \mathbb{C}$, hier wird aber zunächst nur der Fall reeller Null- und Polstellen diskutiert.

Dabei entspricht $G(s)$ in seiner multiplikativen Form einer Reihenschaltung von $m + n$ Elementen, von denen n Glieder je eine Polstelle und m Glieder je genau eine Nullstelle besitzen. In diesem Sinne wäre beispielsweise das System in Bild 5-5 die Reihenschaltung eines Integrators mit einem System erster Ordnung.

Diese Zerlegung als Reihenschaltung einfacher Elemente ist insbesondere für Betrachtungen im Bildbereich zweckmäßig, da mit den im Abschnitt 5.1 vorgestellten Verfahren diese Reihenschaltungen (d.h. Multiplikationen) in besonders einfacher Weise graphisch dargestellt werden können. Hierzu müssen nur die Bode-Diagramme der Einzelemente bekannt sein, um direkt das Bode-Diagramm des Gesamtsystems abzuleiten. Analoges gilt für eine überschlagsmäßige Darstellung der Ortskurve. Hierdurch kann man sich in seiner Kenntnis auf einfache Kernsysteme konzentrieren und kann komplexere Systeme auf diese zurückführen.

Zur Ableitung der Bode-Diagramme und Ortskurven dieser Kernsysteme, die zunächst aus einer reellwertigen Pol- oder Nullstelle bestehen, wird mit einer einfachen stabilen Polstelle begonnen. Der entstehende Frequenzgang ist von der Struktur $G_+(j\omega) = 1/(j\omega + \lambda)$ mit $\lambda > 0$ und bereits in Tab. 5-1 beziehungsweise Tab. 5-2 aufgeführt. Dort findet sich ebenfalls der Fall $\lambda = 0$, der einem Integrator entspricht. Aus diesen beiden Fällen lassen sich

durch mathematische Operationen auch der Fall negativer $\lambda < 0$ und der Fall einfacher reeller Nullstellen ableiten. Für instabile Polstellen ergibt sich aus der Gegenüberstellung

$$G_+(j\omega) = \frac{1}{j\omega + \lambda} = \frac{-j\omega + \lambda}{\omega^2 + \lambda^2} \quad \text{bzw.} \quad G_-(j\omega) = \frac{1}{-j\omega + \lambda} = \frac{j\omega + \lambda}{\omega^2 + \lambda^2}, \quad (6.14)$$

dass $G_-(j\omega)$ genau $G_+(j\omega)$ entspricht, bis auf das umgekehrte Vorzeichen des Imaginärteils. Das führt zu einer Spiegelung der Ortskurve an der reellen Achse, sodass der Phasenwinkel von 0° nach 90° läuft. Der Amplitudengang bleibt jedoch identisch.

Systeme mit einer reellwertigen Nullstelle lassen sich auf ähnliche Weise ableiten: Wegen

$$G_N(j\omega) = (j\omega + \eta) = \left(\frac{1}{j\omega + \eta} \right)^{-1} = (G_P(j\omega))^{-1} \quad (6.15)$$

entspricht ein System mit nur einer Nullstelle $G_N(j\omega)$ genau der Inversion eines Systems $G_P(j\omega)$ mit nur einer Polstelle an identischer Position. Die Inversion einer komplexen Zahl ergibt sich zu

$$G = r \cdot e^{j\varphi} \Rightarrow G^{-1} = \frac{1}{r} \cdot e^{-j\varphi} \quad . \quad (6.16)$$

Folglich wird der Phasenwinkel mit umgekehrten Vorzeichen versehen (d. h. an der 0° -Achse gespiegelt) und der Betrag invertiert. Eine Inversion des Betrages führt im Bode-Diagramm wegen

$$\log(|G|^{-1}) = -\log|G| \quad (6.17)$$

auf einen Vorzeichenwechsel des Betrags, der folglich an der 10^0 -Achse gespiegelt werden muss – siehe auch Bild 6-5.

All diese Überlegungen werden in Tab. 6-2 gebündelt. Hier finden sich die Änderung des Phasengangs sowie der Asymptoten des Amplitudengangs für einzelne Pol- oder Nullstellen abhängig von deren Vorzeichen. Dort ist auch der Fall komplex konjugierter Pol- und Nullstellen aufgenommen, der abweichend behandelt werden muss: Auch dort führen Nullstellen im Vergleich zu Polstellen zu einer Inversion des Betrages und Spiegelung der

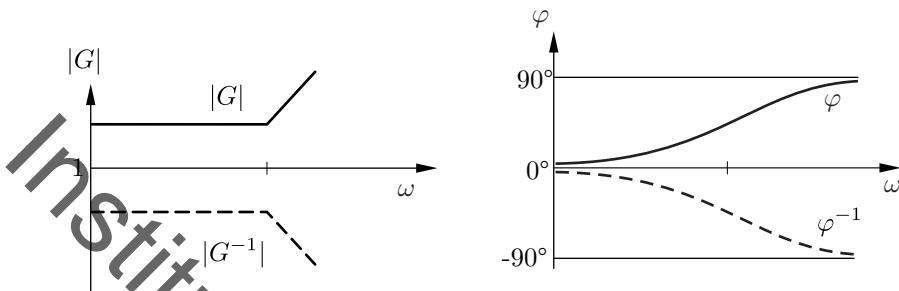


Bild 6-5: Inversion eines Frequenzgangs

Phase, während ein Vorzeichenwechsel im Realteil ein Spiegeln des Phasengangs bewirkt. Das im nächsten Kapitel in Abschnitt 7.3 vorgestellte PT₂-Element kann dabei prototypisch für den Fall komplex konjugierter Polstellen verwendet werden. Die genaue Herleitung der Zusammenhänge findet sich ebenfalls dort.

Zum Umgang mit Tab. 6-2 sei angemerkt, dass entweder eine einzelne reelle Polstelle bzw. Nullstelle bei s oder ein komplex konjugiertes Paar bei s und \bar{s} betrachtet wird. Die Eckkreisfrequenz entspricht dem Betrag von s . Im Amplitudengang sind nur die Asymptoten aufgeführt. Die Phase zeigt die Phasenänderung an, da die Startphase auch von K abhängt. Zudem zeigt im komplex konjugierten Fall auch die Phase nur den prinzipiellen Verlauf, da die genaue Krümmung des Phasengangs von der Dämpfung abhängig ist (siehe Abschnitt 7.3 und Tab. 7-4).

An Tab. 6-2 erkennt man, dass Polstellen zu einem Abfall des Betrages führen, während dieser bei Nullstellen steigt. Da kausale Systeme mindestens so viele Pole wie Nullstellen besitzen, werden die Ortskurven kausaler Systeme für $\omega \rightarrow \infty$ gegen einen endlichen reellen Wert, für einen relationalen Grad von eins oder größer sogar gegen den Ursprung streben. Dabei wird im Falle stabiler Systeme die Ortskurve typischerweise im Uhrzeigersinn durchlaufen, da stabile Polstellen zu einem Phasenabfall führen. Ein mathematisch positiver Verlauf der Ortskurve deutet hingegen auf Nullstellen oder instabile Polstellen hin.

Bei diesen Überlegungen wurde der Vorfaktor K der Übertragungsfunktion

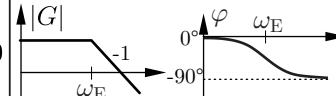
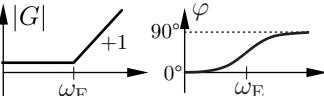
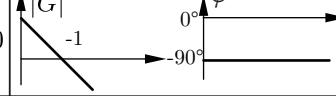
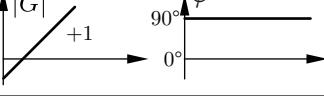
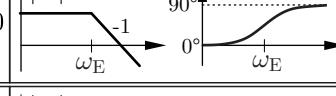
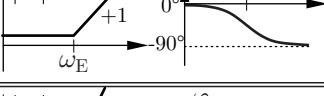
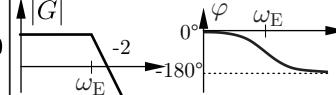
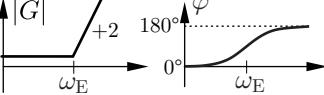
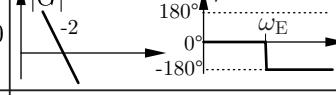
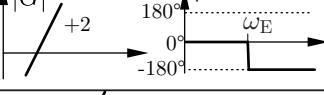
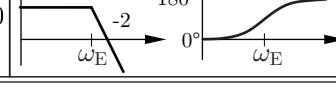
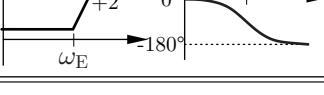
$\omega_E = s $	Polstelle	Nullstelle
reell	$\text{Re}(s) < 0$ 	$\text{Re}(s) < 0$ 
	$\text{Re}(s) = 0$ 	$\text{Re}(s) = 0$ 
	$\text{Re}(s) > 0$ 	$\text{Re}(s) > 0$ 
komplex konjugiert	$\text{Re}(s) < 0$ 	$\text{Re}(s) < 0$ 
	$\text{Re}(s) = 0$ 	$\text{Re}(s) = 0$ 
	$\text{Re}(s) > 0$ 	$\text{Re}(s) > 0$ 

Tabelle 6-2: Bode-Diagramme in Abhängigkeit der Pol- und Nullstellen

$$G(s) = K \cdot \frac{(s - \eta_1) \cdot (s - \eta_2) \dots (s - \eta_m)}{(s - \lambda_1) \cdot (s - \lambda_2) \dots (s - \lambda_n)} \quad (6.18)$$

nicht betrachtet. Dieser führt abhängig von seinem Betrag zu einer Skalierung der Ortskurve um den Faktor $|K|$ beziehungsweise eine vertikale Verschiebung des Amplitudengangs im Bode-Diagramm. Das Vorzeichen von K hat keinen Einfluss auf den Betrag, wohl aber auf den Phasenwinkel. Wegen

$$-G = -r \cdot e^{j\varphi} = (-1) \cdot r \cdot e^{j\varphi} = (e^{\pm j\pi}) \cdot r \cdot e^{j\varphi} = r \cdot e^{j(\varphi \pm \pi)} \quad (6.19)$$

führt die Multiplikation einer komplexen Zahl mit -1 zu einer Drehung um $\pm\pi \equiv \pm180^\circ$. Das Vorzeichen der Drehung ist mathematisch wegen der 360° -Periodizität irrelevant. Im Bode-Diagramm muss folglich der Frequenzgang um 180° nach oben oder unten verschoben werden. In der Ortskurve muss die gesamte Ortskurve um 180° gedreht werden.

Die Zerlegung in eine Reihenschaltung hat Vorteile bei einer Analyse im Bildbereich. Im Zeitbereich hingegen ist diese Art der Zerlegung für eine Analyse wenig vielversprechend, da eine Multiplikation im Bildbereich einer Faltung im Zeitbereich entspricht. Dort bietet es sich an, abweichend eine Zerlegung in eine Summe vorzunehmen. Das kann man über eine Partialbruchzerlegung der Übertragungsfunktion $G(s)$ erreichen.

Für ein $G(s)$ wie in Gl.(4.80)

$$G(s) = K \frac{\prod_{i=1}^m (s - \eta_i)}{\prod_{i=1}^n (s - \lambda_i)} \quad (6.20)$$

ergibt sich für den aus Übersichtlichkeitsgründen gewählten Fall einfacher reeller Polstellen und $m < n$

$$G(s) = K \sum_{k=1}^n \frac{r_k}{s - \lambda_k} , \quad (6.21)$$

mit den Polstellen λ_k und den Residuen r_k , und damit eine Parallelschaltung von n Systemen erster Ordnung.

Aufgrund der Linearität des Systems ergibt sich dann auch die Systemantwort $y(t)$ auf ein Eingangssignal $u(t)$ aus der linearen Überlagerung der Systemantworten $y_k(t)$ der einzelnen $r_k/(s - \lambda_k)$ auf $u(t)$. Diese lineare Überlagerung ist im Zeitbereich einfach darzustellen, wofür das nächste Kapitel zahlreiche Beispiele liefert.

Tatsächlich entspricht diese Zerlegung des Systems in parallele Übertragungskanäle, die linear überlagert werden können, genau der Diagonalen Normalform in Bild 3-6. Folglich kann diese Normalform auch über die Partialbruchzerlegung gewonnen werden und umgekehrt.

Die Jordansche Normalform für mehrfache Eigenwerte entspricht genau den abweichenden Ansätzen für eine Partialbruchzerlegung bei mehrfachen Polstellen.

Auch hier sind die diversen Darstellungsmöglichkeiten für dynamische Systeme also eng miteinander verwoben und die Diagonalisierung einer Matrix kann als analoges Problem zur Bestimmung der Residuen der Partialbruchzerlegung gedeutet werden.

Diese Darstellungsform als Summe kann dazu genutzt werden, diejenigen Übertragungskanäle zu identifizieren, welche den dominanten Beitrag zum dynamischen Verhalten eines Systems leisten. Hierfür wird zunächst ein Signal $f(t)$ betrachtet, dessen Laplace-Transformierte $F(s)$ ausschließlich einfache Polstellen λ_i besitzt:

$$F(s) = K \sum_{k=1}^n \frac{r_k}{s - \lambda_k} \quad \bullet \circ \quad f(t) = K \sum_{k=1}^n r_k e^{\lambda_k t} . \quad (6.22)$$

Welcher der n summierten Anteile wird $f(t)$ maßgeblich bestimmen, d. h. betragmäßig am größten sein?

Für große t wird das der Anteil sein, dessen λ_k den größten Realteil hat, da er am schnellsten wächst bzw. am langsamsten abklingt und somit für große t alle anderen betragmäßig übersteigt. Für kleine t wird dies allerdings wegen $e^{\lambda_k 0} = 1$ derjenige Anteil sein, dessen Residuum r_k den größten Betrag aufweist.

Dominante Signalanteile im Zeitbereich

Eine Funktion $f(t)$ mit

$$f(t) = K \sum_{k=1}^n r_k e^{\lambda_k t} \quad (6.23)$$

wird in ihrem Verhalten für große t maßgeblich durch die λ_k bestimmt, deren Realteile deutlich größer als die Realteile der anderen λ_i sind. Für kleine t werden dies hingegen die λ_k sein, deren Residuen r_k betragmäßig deutlich größer als die Residuen der anderen λ_i sind. Diese λ_k werden *dominante Signalanteile* von $f(t)$ genannt.

Zur Bestimmung der dominanten Signalanteile werden also neben den Realteilen der Polstellen auch die Residuen der Partialbruchzerlegung benötigt. In der Regelungstechnik ist als Signal dabei vor allem die Übergangsfunktion $h(t)$ von Interesse. Deren dominanten Signalanteile können dabei durch

die Pole- und Nullstellen der Übertragungsfunktion $G(s)$ abgeschätzt werden. Die zugehörige Laplace-Transformation $H(s)$ der Übergangsfunktion ergibt sich zu

$$H(s) = G(s) \frac{1}{s} = K \sum_{k=0}^n \frac{r_k}{s - \lambda_k} \quad \bullet \circ \quad y(t) = K \sum_{k=0}^n r_k e^{\lambda_k t} \quad (6.24)$$

mit $\lambda_0 = 0$ aus der Polstelle bei $s = 0$, die durch den Einheitssprung zusätzlich in $H(s)$ miteingebracht wird. Für die Bewertung der Dominanz werden nun die Residuen benötigt, für die eine geschlossene Berechnungsformel angegeben werden kann.

Berechnung der Residuen

Gegeben sei eine Funktion $F(s)$ in der Form

$$F(s) = K \frac{\prod_{i=1}^m (s - \eta_i)}{\prod_{i=1}^n (s - \lambda_i)} \quad (6.25)$$

mit einfachen Polstellen λ_i . Dann berechnen sich die Residuen der Partialbruchzerlegung zu

$$r_k = \lim_{s \rightarrow \lambda_k} (s - \lambda_k) F(s) = \frac{\prod_{\substack{i=1 \\ i \neq k}}^m (\lambda_k - \eta_i)}{\prod_{\substack{i=1 \\ i \neq k}}^n (\lambda_k - \lambda_i)} . \quad (6.26)$$

Für den allgemeinen Fall mehrfacher Polstellen ergeben sich analoge Formeln. Wendet man diese Formel auf $H(s)$ an, so erhält man zunächst für das Gewicht r_0 wegen

$$r_0 = \lim_{s \rightarrow 0} (s - 0) H(s) = \lim_{s \rightarrow 0} s \cdot \frac{1}{s} G(s) = G(0) \quad (6.27)$$

genau die statische Verstärkung $r_0 = K$ des Systems. Das ist einsichtig, da das Gewicht r_0 genau den konstanten Anteil e^0 der Lösungsfunktion $y(t)$ beschreibt.

Für die verbleibenden Gewichte r_k ermöglicht es die Berechnungsvorschrift Gl.(6.26), aus der Lage der Pol- und Nullstellen zueinander abzuschätzen, welches Residuum relativ zu den anderen betragsmäßig größer ausfallen wird: Aus Gl.(6.26) ist zu erkennen, dass Nullstellen η_i in der Nähe einer Polstelle λ_k das Residuum im Betrag verringern – im Grenzfall $\eta_i = \lambda_k$ auf $r_k = 0$. Treten diese Pol- und Nullstellen in $G(s)$ auf, so werden sie einen geringen Einfluss auf die Systemdynamik haben. Nullstellen in der Nähe des Koordinatenursprungs werden hingegen das Gewicht, welches zum durch den Sprung gehörigen konstanten Lösungsanteil gehört, stark verringern und somit einen entscheidenden Einfluss besitzen.

Analog werden λ_i , die nahe an $\lambda_0 = 0$ liegen, aufgrund von Gl.(6.26) ein relativ großes Residuum aufweisen und einen dominanten Einfluss ausüben. Die Vermutung, dass weit von λ_k entfernte Nullstellen η_i den Anteil der zugehörigen Exponentialfunktion erhöhen, hält einer vertieften Analyse hingegen nicht stand. Solche Nullstellen mit z. B. $\eta_i \rightarrow \infty$ liegen nämlich von allen λ_k gleich weit entfernt, sodass sich die relative Gewichtung der Residuen zueinander nicht verändert.

Diese Überlegungen lassen sich wie folgt zusammenfassen:

Dominante Pol- und Nullstellen

Die Residuen r_k werden in ihrem Betrag normalerweise durch die Pol- bzw. Nullstellen festgelegt,

- die sich deutlich näher am Koordinatenursprung als die anderen Pol- und Nullstellen befinden und
- in deren Nähe sich keine Nullstelle (für Polstellen) bzw. Polstelle (für Nullstellen) befindet.

Diese Pol- und Nullstellen werden *dominante Pol- und Nullstellen* genannt. Zusätzlich bezeichnet man Polstellen, die einen deutlich größeren Realteil als andere Polstellen aufweisen, ebenfalls als dominant.

Bei den angegebenen Regeln handelt es sich um Faustregeln, die in den meisten Fällen korrekte Aussagen liefern. Im Einzelfall müssen die Residuen berechnet und verglichen werden.

Trotz der Bedeutung der Polstellen einer Übertragungsfunktion für die dynamischen Eigenschaften des Systems darf die Wirkung von Nullstellen nicht vernachlässigt werden, da diese die Residuen empfindlich beeinflussen

und somit das Verhalten für kleine Zeiten mitbestimmen.

Die Dominanz von Pol- und Nullstellen mit geringem Betrag ist auch nach Tab. 6-2 logisch, da solche Pole und Nullstellen die niedrigsten Eckkreisfrequenzen in den Frequenzgang einbringen.

Im Sinne der Vereinfachung von Modellen kann die Analyse der dominanten Polstellen dazu genutzt werden, unwesentliche Modellanteile zu entfernen. Beispiele für dominante Pol- und Nullstellen liefert Kapitel 7.

6.3 Zerlegung nicht-minimalphasiger Systeme

Die vorgestellte Zerlegung von Systemen ist unabhängig davon anwendbar, ob die Pole bzw. Nullstellen in der rechten oder linken komplexen Halbebenen liegen. Das hat zur Folge, dass aus einem gegebenen Amplitudengang nicht eindeutig auf den zugehörigen Phasengang geschlossen werden kann. So wäre nämlich zu einem Amplitudengang, der zu einer einfachen Polstelle in $|s| = 1/T$ gehört, sowohl eine stabile Polstelle bei $s = -1/T$ als auch eine instabile Polstelle bei $s = 1/T$ passend. Zweitens besäße aber eine steigende Phase, während zur stabilen Polstelle eine fallende Phase gehört. Möchte man erreichen, dass man aus dem Amplitudengang direkt auf den zugehörigen Phasengang schließen kann, so muss Tab. 6-2 dergestalt eingeschränkt werden, dass diese Doppeldeutigkeiten ausgeschlossen werden können. Dies motiviert die folgende Definition:

Minimalphasigkeit

Ein lineares System in der Form wie in Gl.(4.80)

$$G(s) = K \cdot \frac{(s - \eta_1) \cdot (s - \eta_2) \dots (s - \eta_m)}{(s - \lambda_1) \cdot (s - \lambda_2) \dots (s - \lambda_n)} \quad (6.28)$$

heißt *minimalphasig*, wenn $K > 0$ gilt und alle Pol- und Nullstellen nicht-positive Realteile aufweisen.

Diese Definition wird in der regelungstechnischen Literatur nicht einheitlich gehandhabt. Nach der hier gegebenen Definition müssen minimalphasige Systeme nicht zwingend stabil sein, da auch Polstellen auf der imaginären Achse zugelassen sind. In allen anderen Fällen entspricht K der statischen Verstärkung, welche für minimalphasige Systeme positiv sein muss.

Der Name „minimalphasig“ röhrt daher, dass bei einem solchen System sich von einem gegebenen Betragsverlauf direkt auf den Phasenverlauf schließen lässt. Sind nämlich Pole und Nullstellen mit positivem Realteil ausgeschlossen, müssen in Tab. 6-2 nur die Zeilen mit $\text{Re}(s) < 0$ und $\text{Re}(s) = 0$ betrachtet werden. Folglich entspricht eine Steigung von „-2“, „-1“, „0“, „1“ usw. der Asymptoten des Amplitudenganges im Bode-Diagramm stets einem Phasenverlauf mit Asymptoten von -180° , -90° , 0° , $+90^\circ$. Dadurch, dass zusätzlich auch das Vorzeichen von K bekannt ist, kann also der Phasengang zu einem gegebenen Amplitudengang sofort ermittelt werden, wenn das System minimalphasig ist.

Die Umkehrung gilt nur, wenn der exakte Wert von K bekannt ist, da ansonsten der Amplitudengang bei gleichbleibendem Phasengang nach oben oder unten verschoben werden könnte.

Übertragungsglieder, die die Bedingungen für Minimalphasigkeit nicht erfüllen, heißen nicht-minimalphasig. Diese Bezeichnung wird man dabei vor allem für Systeme verwenden, die stabil sind aber Nullstellen mit positivem Realteil besitzen, da die Systeme ansonsten als „instabil“ schon hinreichend charakterisiert sind.

In der Anwendung werden auch nicht-minimalphasige Systeme auftreten. Diese kann man jedoch durch ein geschicktes Ergänzen von Pol- und Nullstellen in ein minimalphasiges System und ein sogenanntes *Allpass*-Glied aufteilen.

Allpass

Ein System oder Filter wird *Allpass(filter)* genannt, wenn sein Amplitudengang für alle Kreisfrequenzwerte konstant ist.

Damit werden stabile Allpassfilter sinusförmige Eingangssignale im eingeschwungenen Zustand abhängig von ihrer Frequenz in ihrer Phaselage verändern, in ihrer Amplitude aber unabhängig von der Frequenz skalieren. Das P-Element $G(s) = K$ wird man dabei üblicherweise nicht als Allpass bezeichnen, da das Eingangssignal in der Phase gar nicht verändert wird.

Aus Tab. 6-2 kann ohne größeren Aufwand darauf geschlossen werden, welche Pol-Nullstellen-Konfigurationen bei einem Allpass vorliegen müssen. Da einzelne Pole und Nullstellen den Amplitudengang um je -1 und $+1$ in den Asymptoten verändern, ist der geforderte konstante Verlauf nur bei einer spiegelbildlichen Anordnung von Pol- und Nullstellen in rechter und lin-

ker komplexer s -Halbebene möglich. Damit ist gemeint, dass ein Allpass, wenn er eine Pol- oder Nullstelle in $s = a + bj$ aufweist, auch eine Null- oder Polstelle in $s = -a + bj$ besitzen wird. Hierdurch wird jede durch eine Pol- oder Nullstelle möglicherweise hervorgerufene Änderung im Amplitudengang durch ein entsprechendes Pendant mit genau gegenläufigem Amplitudengang aber gleicher Eckkreisfrequenz ω_E ausgeglichen und der Betrag bleibt konstant.

Zerlegung nicht-minimalphasiger Systeme

Jedes nicht-minimalphasige System kann in die Reihenschaltung eines minimalphasigen Systems und eines Allpasses zerlegt werden.

Hierzu schlägt man alle Pole und Nullstellen mit $\operatorname{Re}(s) > 0$ des nicht-minimalphasigen Systems dem Allpass zu. Damit dieser einen konstanten Amplitudengang aufweist, ergänzt man diesen um eine spiegelbildliche Anordnung von Pol und Nullstellen, die dann alle negative Realteile aufweisen. Diese ergänzten Pol- und Nullstellen kompensiert man dann durch eine passende Pol-Nullstellen-Kürzung. Folglich entsteht bei dieser Zerlegung keine minimale Realisierung.

Als Beispiel dient das Übertragungsglied mit der Übertragungsfunktion

$$G(s) = \frac{1 - sT_A}{1 + sT} , \quad (6.29)$$

die sich erweitern und umstellen lässt zu

$$G(s) = \frac{1 - sT_A}{1 + sT} \frac{1 + sT_A}{1 + sT_A} = \underbrace{\frac{1 + sT_A}{1 + sT}}_{\text{minimalphasig}} \underbrace{\frac{1 - sT_A}{1 + sT_A}}_{\text{Allpass}} \quad (6.30)$$

und damit die Reihenschaltung eines (phasenminimalen) Gliedes und eines Allpasses beschreibt. Bild 6-6 zeigt beispielhaft Pole und Nullstellen sowie die Frequenzgänge im Bode-Diagramm.

Der Vorteil dieser Zerlegung ist, dass nicht-minimalphasige Anteile in einem System schnell sichtbar gemacht werden können. Dies ist für den Reglerentwurf hilfreich, da nicht-minimalphasige Systeme im Allgemeinen schwieriger als minimalphasige Systeme zu regeln sind. Ist das System aufgrund instabiler Polstellen nicht-minimalphasig, so leuchtet dies direkt ein, da offenbar zunächst ein stabilisierender Regler gefunden werden muss. Besitzt

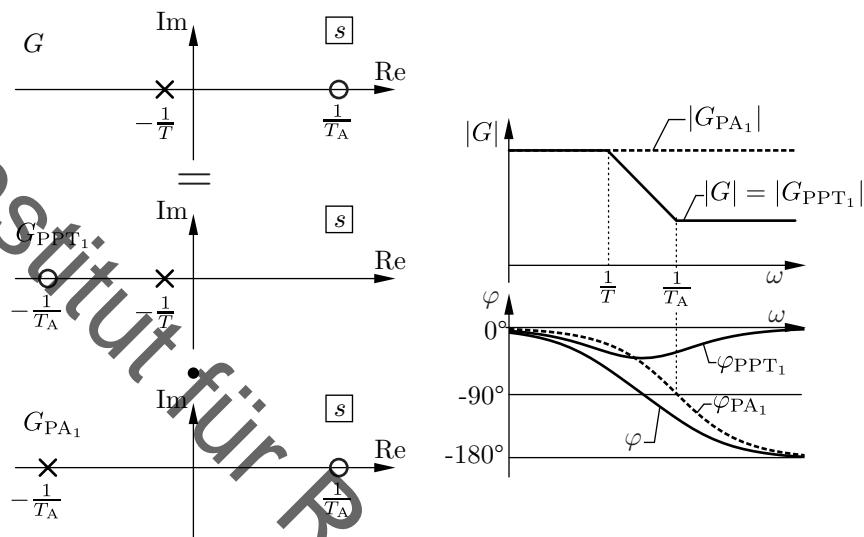


Bild 6-6: Zerlegung eines Nichtminimalphasengliedes

das System eine nicht-minimalphasige Nullstelle, so wird der geschlossene Regelkreis für hohe Reglerverstärkungen instabil werden. Dies wird in Abschnitt 7.5.1 genauer ausgeführt.

7 Typische Übertragungsglieder

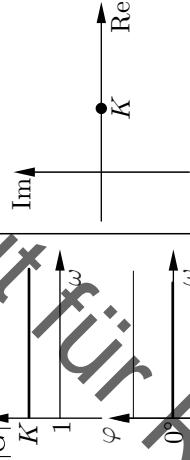
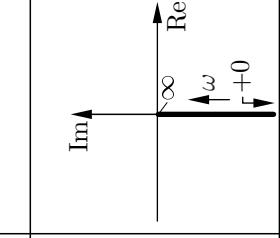
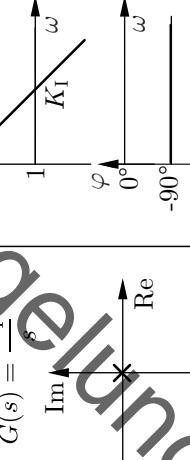
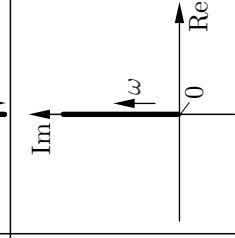
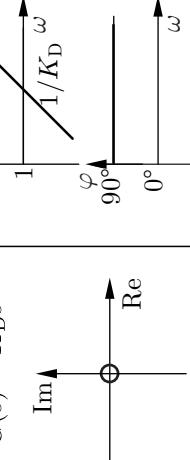
7.1 Übersicht

Da komplexe Systeme durch eine Zerlegung auf einfache Teilsysteme zurückgeführt werden können, lohnt sich ein intensiver Blick auf typische Übertragungsglieder, die Regelungstechnisch wahlweise als Regler, Messglied oder als Regelstrecke gehäuft anzutreffen sind. Diese werden zunächst tabellarisch aufgeführt und anschließend in entsprechenden Unterabschnitten ausführlich diskutiert. Die Elemente stellen dabei auch Beispiele für die Zerlegung in beziehungsweise das Zusammenfassen von Teilsystemen dar.

Mit den Tabellen 7-1 bis 7-3 soll besonders betont werden, dass die analytischen Ausdrücke für Differentialgleichung und Übertragungsfunktion, die Übergangsfunktion, die Pole und Nullstellen der Übertragungsfunktion, das Bode-Diagramm und die Ortskurve des Frequenzgangs prinzipiell gleichberechtigte Beschreibungsmittel für SISO-LTI-Systeme sind. Dies bedeutet auch, dass das dynamische Verhalten ein und desselben Regelkreisgliedes durch unterschiedliche Mittel dargestellt und dass für unterschiedliche Anwendungsfälle die jeweils geeignete Darstellungsform ausgewählt werden kann.

Um den Umfang der Tabellen zu begrenzen, sind nur häufig vorkommende Übertragungsglieder berücksichtigt worden. Die Graphiken für Übergangsfunktion, Ortskurve und Bode-Diagramm sollen den prinzipiellen Verlauf dieser Funktionen veranschaulichen.

In den Tabellen 7-1 bis 7-3 werden Übertragungsglieder mit Buchstaben oder Buchstabengruppen bezeichnet, die das dynamische Verhalten kennzeichnen sollen. Dabei werden Buchstabengruppen ohne Rücksicht darauf gebildet, ob das zu kennzeichnende Verhalten durch Reihen- oder durch Parallelschaltung der durch die Einzelbuchstaben bezeichneten Grundglieder erzeugt wird, weil Verwechslungen kaum möglich sind.

Bez.	Differentialgleichung Übergangsfunktion	Übertragungsfunktion Pol-Nullstellen-Diagramm	Bode-Diagramm	Ortskurve
P	$y = Ku$	$G(s) = K$		
I	$y = K_1 \int u dt$	$G(s) = \frac{K_1}{s}$		
D	$y = K_D \dot{u}$	$G(s) = K_D s$		

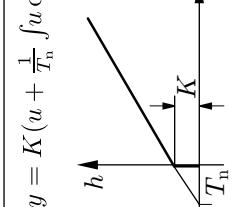
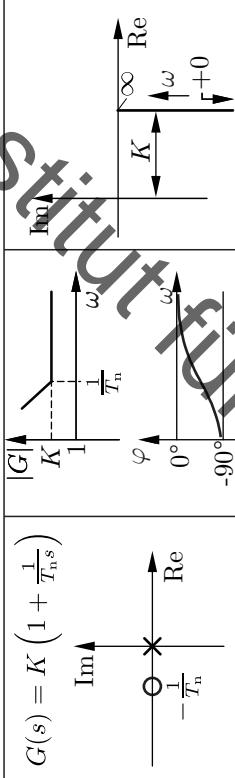
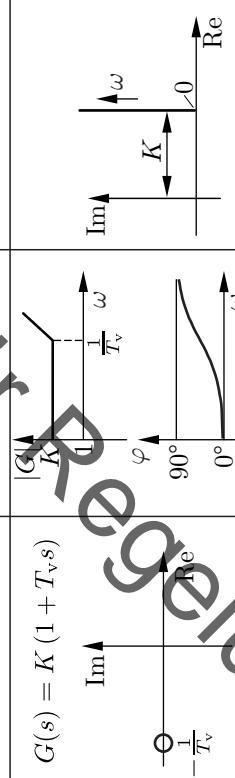
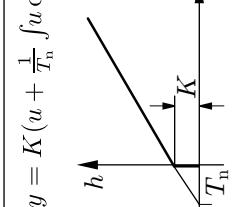
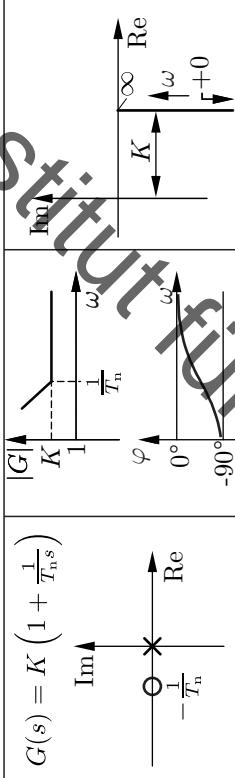
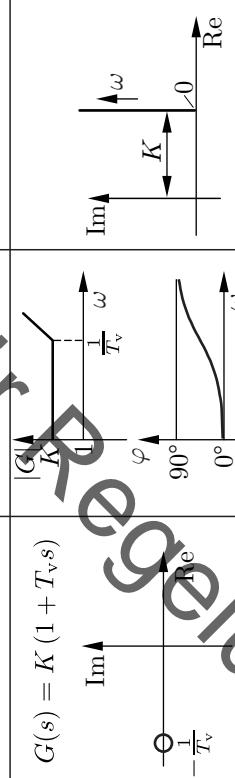
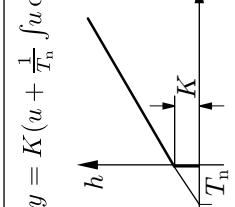
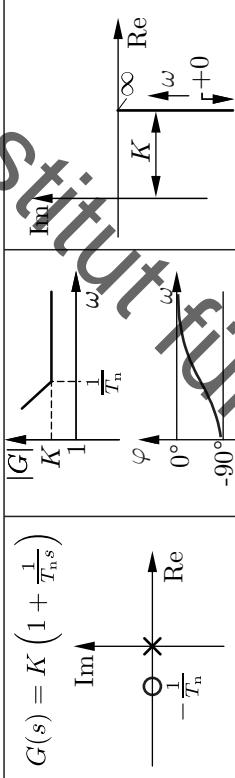
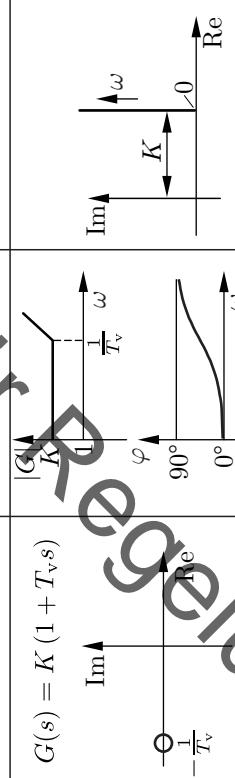
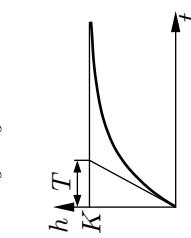
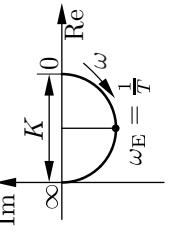
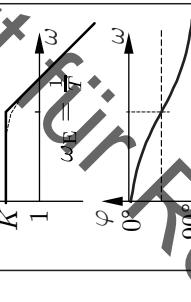
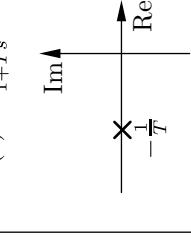
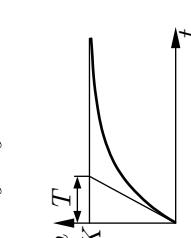
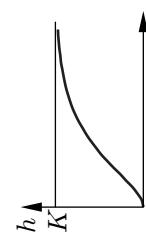
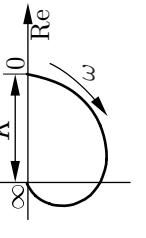
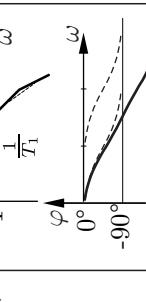
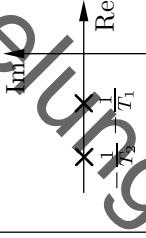
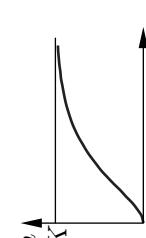
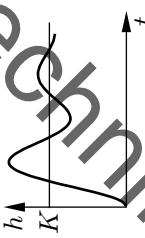
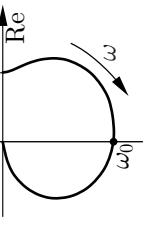
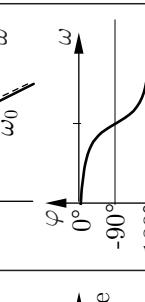
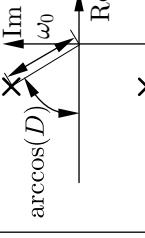
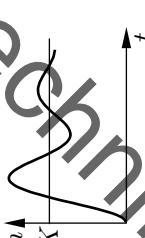
$y = K(u + \frac{1}{T_n} \int u dt)$ 	$G(s) = K \left(1 + \frac{1}{T_n s} \right)$ 	$ G $ 
$y = K(u + T_v \dot{u})$ 	$G(s) = K (1 + T_v s)$ 	$ G $ 
$y = K(u + \frac{1}{T_n} \int u dt + T_v \dot{u})$ 	$G(s) = K \left(1 + \frac{1}{T_n s} + T_v s \right)$ 	$ G $ 

Tabelle 7-1: Regelkreisglieder

Bez.	Differentialgleichung Übergangsfunktion	Übertragungsfunktion Pol-Nullstellen-Diagramm	Bode-Diagramm	Ortskurve
PT ₁	$T\dot{y} + y = Ku$ 	$G(s) = \frac{K}{1+Ts}$ 	 	
PT ₂ (D ≥ 1)	$T_1 T_2 \ddot{y} + (T_1 + T_2)\dot{y} + y = Ku$ 	$G(s) = \frac{K}{T_1 T_2 s^2 + (T_1 + T_2)s + 1}$ 	 	
PT ₂ (D < 1)	$\ddot{y} + 2D\omega_0\dot{y} + \omega_0^2 y = K\omega_0^2 u$ 	$G(s) = \frac{K\omega_0^2}{s^2 + 2D\omega_0 s + \omega_0^2}$ 	 	

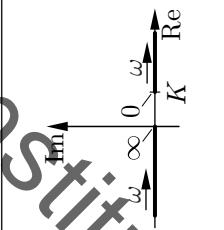
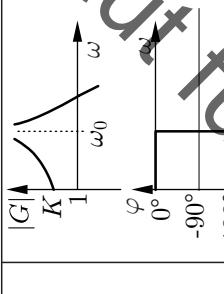
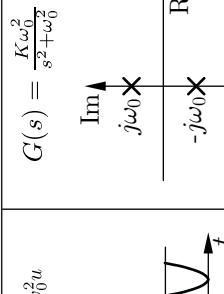
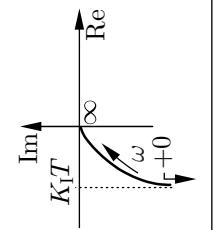
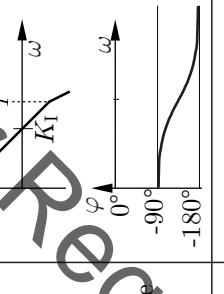
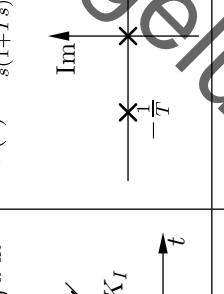
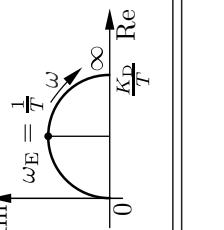
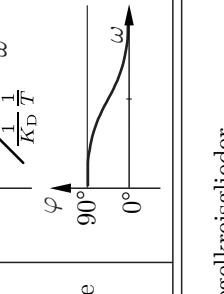
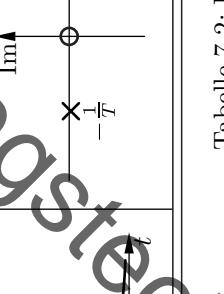
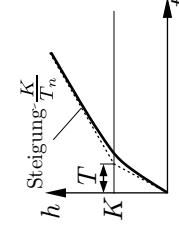
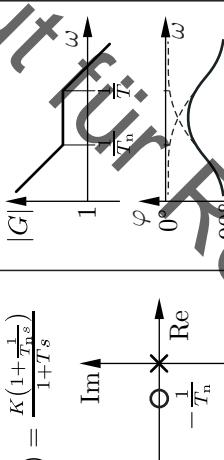
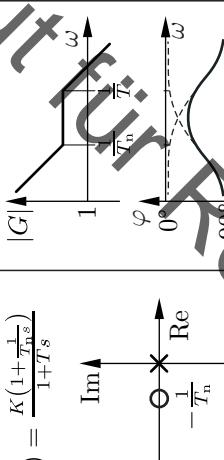
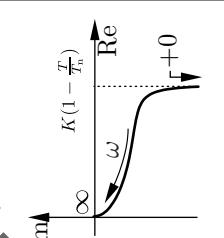
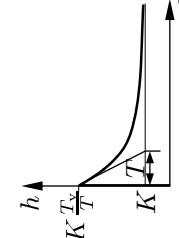
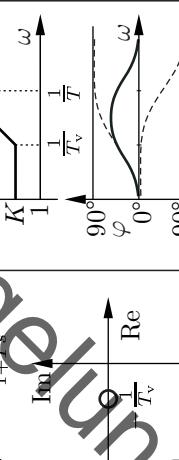
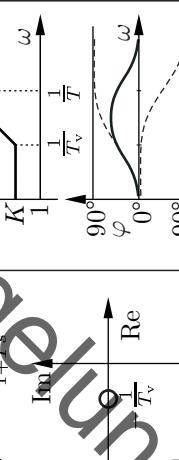
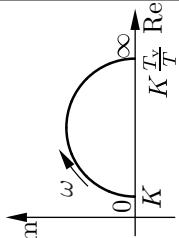
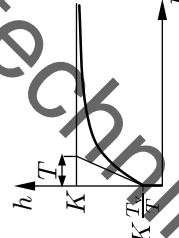
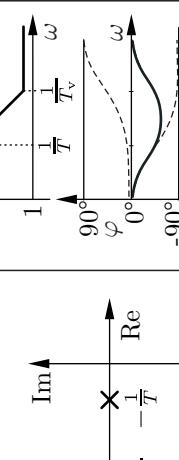
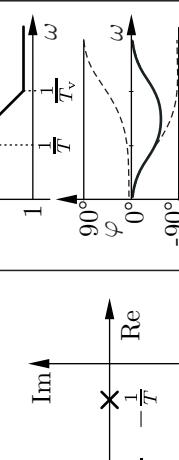
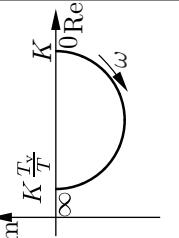
PT_2 ($D = 0$)	$\ddot{y} + \omega_0^2 y = K \omega_0^2 u$	$G(s) = \frac{K \omega_0^2}{s^2 + \omega_0^2}$			
IT_1	$T \dot{y} + y = K_1 \int u \, dt$	$G(s) = \frac{K_1}{s(1+Ts)}$			
DT_1	$T \dot{y} + y = K_D \dot{u}$	$G(s) = \frac{K_D s}{1+Ts}$			

Tabelle 7-2: Regelkreisglieder

Bez.	Differentialgleichung Übergangsfunktion	Übertragungsfunktion Pol-Nullstellen-Diagramm	Bode-Diagramm	Ortskurve
PIT ₁	$T\dot{y} + y = K(u + \frac{1}{T_n} f_0 t)$ 	$G(s) = \frac{K(1 + \frac{1}{T_n s})}{1 + T_n s}$ 		
PDT ₁	$T\dot{y} + y = K(u + T_v \dot{u})$ 	$G(s) = K \frac{1 + T_v s}{1 + T_s s}$ 		
PPT ₁	$T\dot{y} + y = K(u + \frac{T_v}{T} \dot{u})$ 	$G(s) = K \frac{1 + T_v s}{1 + T s}$ 		

$y(t) = Ku(t - T_t)$	$G(s) = Ke^{-sT_t}$	
PT_t		
$T\dot{y}(t) + y(t) = Ku(t - T_t)$	$G(s) = \frac{K}{1+Ts} e^{-sT_t}$	
PT_1T_t		
$T\dot{y} + y = K(u - Tu)$	$G(s) = \frac{K(1-Ts)}{1+Ts}$	
PA_1		

Tabelle 7-3: Regalkreisglieder

7.2 Grundlegende Reglertypen

7.2.1 P-Element

Übertragungsglieder mit proportionalem (P) oder integrierendem (I) Verhalten sind in den vorherigen Kapiteln bereits angeklungen. Sie sind in Regelstrecken, Mess- und Stellgeräten sowie in Reglern anzutreffen.

Beim P-Element mit der Gleichung $y = K \cdot u$ ergibt sich ein rein statischer Zusammenhang zwischen Ein- und Ausgangsgröße, wobei K der statischen Verstärkung entspricht. Ein P-Regler bezieht ausschließlich die aktuelle Regelabweichung in die Berechnung der Stellgröße ein. In diesem Sinne ist der P-Regler nicht dynamisch, da in die Regelentscheidung nur der aktuelle Zeitpunkt, nicht aber die Vergangenheit einfließt.

7.2.2 I-Element

Das I-Element folgt der Gleichung $\dot{y} = Ku$ bzw. $y = K \int_0^t u(\tau) d\tau$. In der zweiten Form ist die Bezeichnung als integrierendes Element leichter einzusehen, da die Ausgangsgröße sich aus der Integration der Eingangsgröße ergibt. Ein I-Regler verwendet die gesamte Vergangenheit der Regelabweichung, um aus deren Integration den aktuellen Stelleingriff zu berechnen. Dies hat zur Folge, dass der I-Regler bei gleicher Verstärkung wie ein P-Regler auf eine z. B. durch eine Störung plötzlich auftretende Regelabweichung langsamer reagiert, da diese erst über eine gewisse Zeitspanne integriert werden muss, um einen Beitrag zur Stellgröße zu leisten.

Das verzögerte Verhalten bei plötzlichen Störungen ist ein Nachteil des I-Reglers, der potentiell sogar Systeme destabilisieren kann. Dies passiert, wenn die Stelleingriffe so stark verzögert auf eine Strecke einwirken, dass die Situation sich bereits wieder geändert hat und die Stelleingriffe inzwischen unpassend geworden sind – hierauf wird in Kapitel 9 eingegangen. Diesen Nachteil kompensiert der I-Regler durch die Eigenschaft, dass er – sofern der geschlossene Regelkreis stabil ist – keine *bleibende Regelabweichung* zulässt.

Stationäre Genauigkeit

Ein stabiles System $u \mapsto y$ heißt *stationär genau*, wenn für jedes konver-

gente Eingangssignal $u(t)$ gilt:

$$\lim_{t \rightarrow \infty} u(t) = \lim_{t \rightarrow \infty} y(t) , \quad (7.1)$$

d. h., dass der Grenzwert von Ein- und Ausgangsgröße identisch ist.

Bleibende Regelabweichung

Für einen stabilen geschlossenen Regelkreis mit der Führungsübertragungsfunktion $w \mapsto y$ ergibt sich für jedes konvergente $w(t)$ ein Grenzwert

$$\lim_{t \rightarrow \infty} (w(t) - y(t)) \doteq \lim_{t \rightarrow \infty} e(t) = e_\infty , \quad (7.2)$$

der als *bleibende Regelabweichung* bezeichnet wird. Ist das System $w \mapsto y$ stationär genau, so gilt $e_\infty = 0$ und man spricht davon, dass *keine bleibende Regelabweichung* vorliegt.

Keine bleibende Regelabweichung zu besitzen ist eine sehr wünschenswerte Eigenschaft eines geschlossenen Regelkreises. Praktisch bedeutet dies, dass die Anforderung $w(t)$ an den Regelkreis nach einiger Zeit ohne einen Fehler in der Regelgröße $y(t)$ umgesetzt werden kann. Also wird die Anforderung $w(t)$ vollständig erfüllt, solange sie einen Grenzwert besitzt und sich nicht ständig ändert.

Würde z.B. der Fahrer eines Fahrzeuges mit einem stationär genau arbeitenden Tempomaten eine neue Wunschgeschwindigkeit vorgeben, so würde der Tempomat diese Wunschgeschwindigkeit nach einiger Zeit exakt einstellen.

Stationäre Genauigkeit ist beileibe nicht die einzige Anforderung an einen Regelkreis, da der Weg hin zur Erreichung dieses Ziels ebenfalls von entscheidender Bedeutung ist. Nichtsdestoweniger ist eine verschwindende bleibende Regelabweichung eine zentrale Anforderung an viele Regelungssysteme.

Wird ein I-Regler $e \mapsto u$ in einem stabilen Regelkreis eingesetzt, so besitzt die Stellgröße u des I-Reglers einen Grenzwert, d. h. es gilt $\dot{u} = 0$ für $t \rightarrow \infty$. Daraus folgt

$$0 = \dot{u}_\infty = K e_\infty \Rightarrow e_\infty = 0 \quad \text{für } K \neq 0. \quad (7.3)$$

Ein P-Regler $u = K \cdot e$ wird diese Bedingung nicht erfüllen, da dort die Forderung $\dot{u} = 0$ für $t \rightarrow \infty$ nur auf

$$0 = \lim_{t \rightarrow \infty} \dot{u} = \lim_{t \rightarrow \infty} K \cdot \dot{e} \Rightarrow \dot{e} = 0 \Rightarrow e_\infty = \text{beliebig} \quad (7.4)$$

führt. Ein I-Regler lässt also im Gegensatz zu einem P-Regler keine bleibende Regelabweichung zu.

Diese Argumentation lässt sich mithilfe der Grenzwertsätze in Abschnitt 4.6 formalisieren, wobei entscheidend ist, ob der aufgeschnittene Regelkreis einen *integrierenden Anteil* aufweist.

Integrierender Anteil und integrierendes Verhalten

Ein LTI-System besitzt genau dann einen integrierenden Anteil, wenn es eine Polstelle in $s = 0$ besitzt.

Haben zusätzlich alle weiteren Polstellen des Systems einen negativen Realteil, so gilt

$$\lim_{t \rightarrow \infty} g(t) = \text{konstant} \neq 0 \quad (7.5)$$

und das System besitzt *integrierendes Verhalten*.

Bei dieser Definition wird erneut vorausgesetzt, dass es sich um eine minimale Realisierung handelt. Integrierendes Verhalten bedeutet, dass sich das System für $t \rightarrow \infty$ dem Verhalten eines Integrators annähert, da alle dynamischen Anteile außer dem integrierenden Anteil mit $e^{-\lambda t}$ abklingen. Wegen Gl.(7.5) konvergiert die Sprungantwort eines Systems mit integrierendem Verhaltens gegen eine Rampe mit konstanter Steigung, wie sie aus der Sprungantwort des Integrators bekannt ist. Ein IT₁ besitzt beispielsweise integrierendes Verhalten.

Wendet man die Grenzwertsätze an, um die bleibende Regelabweichung zu berechnen, so erhält man allgemein für einen stabilen geschlossenen Regelkreis und die Führungsübertragungsfunktion $w \mapsto y$ wie in Gl.(6.6) den Zusammenhang

$$y_\infty = \lim_{s \rightarrow 0} sW(s)T(s) = \lim_{s \rightarrow 0} sW(s) \frac{G_0(s)}{1 + G_0(s)} \quad . \quad (7.6)$$

Das Eingangssignal $w(t)$ ist nach Voraussetzung konvergent. Bezeichnet man den Endwert mit w_∞ , so führt die Partialbruchzerlegung von $W(s)$

auf

$$W(s) = \frac{w_\infty}{s} + \sum_{i=1}^n \frac{r_i}{s - s_i} \quad \Rightarrow \quad \lim_{s \rightarrow 0} sW(s) = w_\infty \quad . \quad (7.7)$$

Folglich berechnet sich weiter

$$y_\infty = w_\infty \cdot \lim_{s \rightarrow 0} \frac{G_0(s)}{1 + G_0(s)} \quad . \quad (7.8)$$

Die relative bleibende Regelabweichung hängt also ausschließlich von der Verstärkung von $G_0(s)$ für $s \rightarrow 0$ ab. Die Regelabweichung wird dabei umso geringer, je größer die statische Verstärkung des aufgeschnittenen Regelkreises wird. Die Forderung $y_\infty = w_\infty$ lässt sich dabei nur für $|G_0(s)| \rightarrow \infty$ für $s \rightarrow 0$ erfüllen. Das ist genau dann der Fall, wenn $G_0(s)$ eine Polstelle in $s = 0$ besitzt.

Stationäre Genauigkeit im geschlossenen Regelkreis

Die Führungsübertragungsfunktion des geschlossenen Standard-Regelkreises ist genau dann stationär genau, wenn

- 1) der geschlossene Regelkreis stabil ist und
- 2) der aufgeschnittene Regelkreis einen integrierenden Anteil besitzt.

Hieraus folgt, dass unter Voraussetzung der Stabilität ein I-Regler stets stationäre Genauigkeit sicherstellt, sofern die Regelstrecke keine Nullstelle in $s = 0$ besitzt. Für Strukturen abseits des Standard-Regelkreises wie in Abschnitt 12.2 kann eine stationäre Genauigkeit im Führungsverhalten auch ohne I-Anteil im Regler realisiert werden.

Besitzt die Regelstrecke einen integrierenden Anteil, so reicht auch ein P-Regler aus, um eine bleibende Regelabweichung zu verhindern. Außerdem folgt aus $|G_0(0)| = \infty$ auch direkt, dass für die Störübertragungsfunktion nach Gl.(6.5) gilt:

$$\lim_{s \rightarrow 0} S(s) = \lim_{s \rightarrow 0} \frac{1}{1 + G_0(s)} \stackrel{|G_0| \rightarrow \infty}{=} 0 \quad . \quad (7.9)$$

Folglich ist der Grenzwert der Regelgröße y bei einem Sprung der Störgröße z identisch null. Konstante Störgrößen werden also vollständig ausgeregelt.

Vollständige Ausregelung von Störungen

Die Führungsübertragungsfunktion $T(s)$ eines Regelkreises ist genau dann stationär genau, wenn konstante Störgrößen z nach Bild 6-2 vollständig ausgeregelt werden.

Abseits des unterschiedlichen dynamischen Verhaltens und verschiedener Eigenschaften für den Grenzwert $t \rightarrow \infty$ unterscheiden sich P-Element und Integrator auch in ihren Darstellungsformen wie Frequenzgang und Pol-Nullstellen-Diagramm erheblich. Beide Elemente sind dabei aus Kapitel 5 in Form der Beispiele in Tab. 5-1 und Tab. 5-2 bereits bekannt, sodass die resultierenden Verläufe hier nicht nochmals diskutiert werden sollen.

7.2.3 D-Element

Der Differenzierer – auch D-Glied oder D-Element genannt – folgt der Differentialgleichung $y = K \cdot \dot{u}$ mit Eingang u und Ausgang y . Er entspricht damit einem Integrator mit umgekehrter Ursache-Wirkungs-Richtung, sodass sich die Ausgangsgröße nicht als Integral der Eingangsgröße, sondern als deren Ableitung ergibt.

Aus der Umkehrung der Ursache-Wirkungs-Richtung ergibt sich direkt, dass strukturell

$$G_D(s) = (G_I(s))^{-1} \quad (7.10)$$

gelten muss, d. h. die Übertragungsfunktion des Differenzierers ist die inverse Übertragungsfunktion des Integrators. Entsprechend ergibt sich mit den in Kapitel 6 vorgestellten Rechenregeln die Darstellung des Differenzierers im Bode-Diagramm und in der Ortskurve über eine passende Spiegelung des Integrators.

Der Differenzierer ist acausal und daher in Regelstrecken nur in Sonderfällen anzutreffen. Auch als Regler kann er in seiner Reinform nicht eingesetzt werden, wird aber dennoch gelegentlich für Auslegungszwecke genutzt. Hierbei geht man so vor, dass man den D-Regler zunächst in seiner acausalen Reinform auslegt, um diese dann in einem zweiten Schritt zu kausalisieren – siehe auch Abschnitt 7.4.

Ein Vorteil des D-Reglers ist die enorm schnelle Reaktion auf rasche Änderungen in der Regelabweichung, da diese entsprechend hohe Ableitungen \dot{e}

und damit große Stelleingriffe $u = K \cdot \dot{e}$ bedeuten.

Nachteilig ist, dass im stationären Fall $\dot{e} = 0$ und damit auch $u = 0$ gilt und somit ohne das Vorliegen einer Änderung der Regler keine Eingriffe tätig. Somit folgt für lineare Systeme in minimaler Realisierung auch $u = 0 \Rightarrow y = 0$ und damit $e_\infty = w_\infty$. Das ist insbesondere bei Arbeitspunktwechseln nachteilig und resultiert in einer hohen bleibenden Regelabweichung.

Ein weiterer praktischer Nachteil besteht darin, dass durch die Ableitung auch evtl. vorhandenes Rauschen im Fehlersignal stark verstärkt wird.

Man spricht analog zum integrierenden Anteil auch von einem *differenzierenden Anteil*:

Differenzierender Anteil und differenzierendes Verhalten

Ein LTI-System besitzt genau dann einen differenzierenden Anteil, wenn es eine Nullstelle in $s = 0$ besitzt.

Haben zusätzlich alle Polstellen des Systems einen negativen Realteil, so gilt

$$\lim_{t \rightarrow \infty} h(t) = 0 \quad (7.11)$$

und das System besitzt *differenzierendes Verhalten*.

Offenbar kann ein System nicht gleichzeitig einen differenzierenden und einen integrierenden Anteil besitzen bzw. differenzierendes und integrierendes Verhalten aufweisen. In solchen Fällen käme es zu einer Pol-Nullstellen-Kürzung und beide Anteile höben sich auf.

Ein System mit differenzierendem Verhalten nähert sich mit der Zeit dem Verhalten eines Differenzierers an. Dies äußert sich darin, dass dieser nur bei Änderungen der Eingangsgröße ein nicht verschwindendes Ausgangssignal erzeugt und somit im Falle einer konstanten Anregung, wie bei einem Sprung, gegen null läuft. Die statische Verstärkung beträgt also null. Weil diesses Verhaltens spricht man gelegentlich statt von differenzierendem Verhalten auch von *nachgebendem Verhalten* mit synonymer Bedeutung.

Ist im einfachen Standard-Regelkreis die Führungsübertragungsfunktion stationär genau, dann besitzt die Störübertragungsfunktion differenzierendes Verhalten.

P- und D-Element sind stabil, während das I-Element grenzstabil ist.

7.2.4 PI, PD und PID

Übertragungsglieder mit PI-, PD- und PID-Verhalten treten hauptsächlich als Regler auf. Auch umgekehrt sind die meisten praktisch eingesetzten linearen Regler vom so genannten PID-Typ, d. h. sie sind PID-Regler oder lassen sich als Vereinfachungen des PID-Reglers (wie PI- oder PD-Regler) auffassen.

Das PD-Glied entsteht durch Parallelschalten eines proportional und eines differenzierend wirkenden Gliedes nach Bild 7-1. Als PD-Regler (Proportional - Differential - Regler) eingesetzt, nutzt es neben der Regelabweichung auch noch deren Änderungsgeschwindigkeit zum Bilden der Stellgröße aus. Die Differentialgleichung dieser Parallelschaltung

$$y = K_R \cdot u + K_D \cdot \dot{u} \quad (7.12)$$

wird üblicherweise in der Form

$$y = K_R(u + T_v \cdot \dot{u}) \quad (7.13)$$

mit der Zeitkonstanten $T_v = K_D/K_R$ geschrieben, die auch als Vorhaltzeit oder Vorhaltezeit bezeichnet wird. K_R entspricht dann der statischen Verstärkung.

Die Übergangsfunktion des PD-Elements (Tab. 7-1) entsteht durch Addition der Übergangsfunktionen des P- und des D-Elementes.

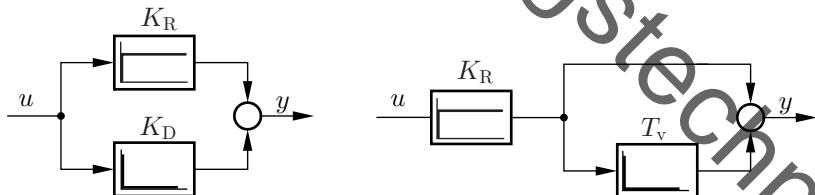


Bild 7-1: PD-Regler, Wirkungsplan

Der Frequenzgang

$$G(j\omega) = K_R(1 + j\omega T_v) \quad (7.14)$$

hat einen konstanten Realteil K_R und einen kreisfrequenzabhängigen Imaginärteil $\omega K_R T_v$. Seine Ortskurve ist eine Parallelle zur positiv-imaginären Achse.

Die Übertragungsfunktion des PD-Reglers

$$G(s) = K_R(1 + sT_v) \quad (7.15)$$

wird durch die Vorhaltzeit T_v und die zugehörige einzige Nullstelle $\eta = -1/T_v$ gekennzeichnet. Die Darstellung im Bode-Diagramm kann direkt aus Tab. 6-2 abgeleitet werden. Alternativ sieht man, dass die Übertragungsfunktion des PD-Reglers genau der inversen Übertragungsfunktion des Systems erster Ordnung entspricht, für welches Ortskurve und Bode-Diagramm in Tab. 5-1 und Tab. 5-2 bereits bekannt sind. Für den Amplitudengang

$$|G(j\omega)| = K_R \sqrt{1 + \omega^2 T_v^2} \quad (7.16)$$

mit den Asymptoten

$$\begin{aligned} \lg |G(\omega \ll \omega_E)| &\simeq \lg K_R \\ \lg |G(\omega \gg \omega_E)| &\simeq \lg K_R + \lg(\omega T_v) = \lg K_R + \lg T_v + \lg \omega \end{aligned} ; \quad (7.17)$$

ergibt sich dabei direkt der in Tab. 7-1 wiedergegebene Verlauf. Analog zu den Ausführungen in Tab. 5-2 oder Tab. 6-2 tritt die größte Abweichung des Betrages von den Asymptoten bei der Eckkreisfrequenz auf:

$$\lg |G(\omega = \omega_E)| \approx \lg K_R + 0,15 . \quad (7.18)$$

Der Phasengang

$$\varphi(\omega) = \arctan(\omega T_v) \quad (7.19)$$

wird entsprechend durch die Grenzwerte

$$\varphi(\omega \ll \omega_E) = 0^\circ , \quad \varphi(\omega \gg \omega_E) = 90^\circ \quad (7.20)$$

und den Wert bei der Eckkreisfrequenz

$$\varphi(\omega = \omega_E) = 45^\circ \quad (7.21)$$

bestimmt. Genauso wie der D-Regler ist auch der PD-Regler aksausal und muss nach seinem Entwurf kausalisiert werden. PD-Regler werden z. B. gern

für Positionsregelungen in der Robotik oder der Produktionstechnik eingesetzt.

Der PI-Regler kann durch Parallelschaltung eines P-Elements und eines integrierenden Gliedes nach Bild 7-2 aufgebaut werden. Dieser Regler vereinigt in gewissem Umfang die positiven Eigenschaften des P-Reglers – schnelle Stellgrößenbeeinflussung bei Regelabweichung – mit denen des I-Reglers, nämlich keine bleibende Regelabweichung zuzulassen. Er besitzt genau wie der I-Regler integrierendes Verhalten. Die Gleichung der Parallelschaltung

$$y = K_R \cdot u + K_I \cdot \int_0^t u \, d\tau \quad (7.22)$$

wird üblicherweise umgeformt zu

$$y = K_R(u + \frac{1}{T_n} \cdot \int_0^t u \, d\tau) \quad (7.23)$$

mit der Zeitkonstanten $T_n = K_R / K_I$, die auch Nachstellzeit genannt wird. Dabei ist zu beachten, dass eine verringerte Nachstellzeit einer stärkeren Wirksamkeit des integrierenden Anteils des Reglers entspricht.

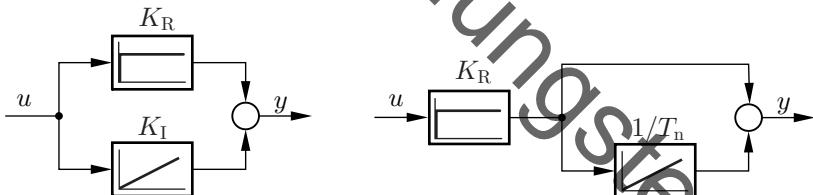


Bild 7-2: PI-Regler, Wirkungsplan

Aus der Differentialgleichung ist zu erkennen, dass die Übergangsfunktion durch Addition der Übergangsfunktionen des P- und des I-Gliedes entsteht (Tab. 7-1). Da der Integration im Zeitbereich die Division durch $j\omega$ im Frequenzbereich entspricht, ergibt sich der Frequenzgang des PI-Reglers zu

$$G(j\omega) = K_R \left(1 + \frac{1}{j\omega T_n} \right) \quad , \quad (7.24)$$

die Übertragungsfunktion wird zweckmäßigerverweise als

$$G(s) = K_R \cdot \frac{1 + sT_n}{sT_n} \quad (7.25)$$

geschrieben. Dadurch ist unmittelbar zu erkennen, dass diese Funktion einen Pol $\lambda = 0$ und eine Nullstelle $\eta = -1/T_n$ aufweist (Tab. 7-1) und somit direkt mit Tab. 6-2 gearbeitet werden kann.

Durch graphische Addition der zugehörigen Frequenzgänge im Bode-Diagramm entsteht das in Tab. 7-1 dargestellte Bild. Der Frequenzgang weist einen kreisfrequenzunabhängigen Realteil K_R und einen Imaginärteil $-K_I/\omega$ auf, der dem des integrierenden Gliedes entspricht. Die Ortskurve ist daher eine Parallele zur negativ-imaginären Achse im Abstand K_R .

PI-Regler werden zum Beispiel für Kraftregelungen in der Robotik verwendet, wo differenzierende Anteile u. a. aufgrund des Rauschens auf Kraftsignale meist nicht zielführend sind [58].

Der PID-Regler kann entsprechend Bild 7-3 als Kombination der beiden zuvor behandelten Reglertypen PD- und PI-Regler aufgefasst werden. Die Parallelschaltung von P-, I- und D-Glied ergibt den aufwendigsten unter den Standardreglern. Er wird durch die Gleichung

$$y = K_R u + K_I \int_0^t u \, d\tau + K_D \dot{u} \quad (7.26)$$

und umgeformt durch

$$y = K_R \left(u + \frac{1}{T_n} \int_0^t u \, d\tau + T_v \dot{u} \right) \quad (7.27)$$

mit

$$T_n = \frac{K_R}{K_I} \quad , \quad T_v = \frac{K_D}{K_R} \quad (7.28)$$

beschrieben. Die Übergangsfunktion ergibt sich durch Addition der Übergangsfunktionen der drei parallel geschalteten Glieder.

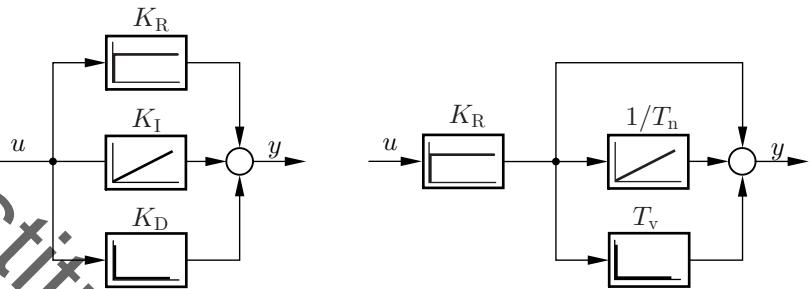


Bild 7-3: PID-Regler, Wirkungsplan

Der Frequenzgang

$$G(j\omega) = K_R \left(1 + \frac{1}{j\omega T_n} + j\omega T_v \right) \quad (7.29)$$

weist einen kreisfrequenzunabhängigen Realteil K_R und einen kreisfrequenz-abhängigen Imaginärteil

$$\text{Im}(G(j\omega)) = K_R \left(\omega T_v - \frac{1}{\omega T_n} \right) \quad (7.30)$$

auf. Die Ortskurve ist demzufolge eine Parallele zur imaginären Achse (Tab. 7-1).

Zur Darstellung im Bode-Diagramm bringt man den Frequenzgang zweckmäßigerweise auf den gemeinsamen Nenner

$$G(j\omega) = K_R \frac{1 + j\omega T_n + (j\omega)^2 T_n T_v}{j\omega T_n} \quad . \quad (7.31)$$

Hieran sieht man, dass im Gegensatz zu PD- und PI-Regler die Zeithorizontanten T_n und T_v leider nicht mit den Nullstellen korrespondieren. Wenn jedoch T_v wesentlich kleiner als T_n ist, was häufig aber nicht immer der Fall ist, kann man näherungsweise

$$G(j\omega) \simeq K_R \frac{(1 + j\omega T_n)(1 + j\omega T_v)}{j\omega T_n} \quad (7.32)$$

ansetzen und den PID-Regler als Reihenschaltung von einem PI- und einem PD-Regler auffassen. Daraus resultieren die Eckkreisfrequenzen $\omega_{E1} = 1/T_n$, $\omega_{E2} = 1/T_v$, der trogförmige Amplitudengang und der von -90° bis $+90^\circ$ verlaufende Phasengang. Falls die Voraussetzung $T_v \ll T_n$ nicht erfüllt ist, ändern sich die Werte für die (im Allg. komplexen) Nullstellen der Übertragungsfunktion bzw. die Eckkreisfrequenzen des Frequenzgangs; die grundsätzlichen Eigenschaften des PID-Reglers bleiben dagegen erhalten.

7.3 Verzögerungsglieder

7.3.1 PT₁

Als Beispiel in vorangegangenen Abschnitten wurde oft die Differentialgleichung erster Ordnung

$$T\ddot{y} + y = Ku \quad (7.33)$$

genutzt. Charakteristisch für den Zeitverlauf ist die verzögerte Antwort des Ausgangs auf einen Sprung der Eingangsgröße (siehe Bild 4-3). Daher hat sich die Bezeichnung als Verzögerungsglied für Systeme mit vergleichbarem Verhalten eingebürgert.

Verzögerungsglieder

Ein LTI-System nach Gl.(2.10), das durch eine stabile Differentialgleichung beschrieben wird, welche keine Ableitungen der Eingangsgrößen enthält, heißt *Verzögerungsglied*.

Im Bildbereich zeichnen sich Verzögerungsglieder dadurch aus, dass der Zähler der Übertragungsfunktion bzw. des Frequenzganges nur aus einer Konstanten besteht. Somit besitzen Verzögerungsglieder nur Polstellen und keine Nullstellen. Die Anzahl der Polstellen n gibt die Ordnung des Systems wieder und man spricht entsprechend von einem Verzögerungsglied n -ter Ordnung. Als Akronyme haben sich im Deutschen die Kurznotationen PT₁, PT₂ oder PT_n etabliert. Für das PT₁ ergeben sich alle Verläufe wortgleich zu den bisherigen Ausführungen zu einem System erster Ordnung.

In bestimmten anwendungsorientierten Fachdisziplinen wie der Servohydraulik spricht man anstelle von Verzögerungsgliedern auch von Systemen *mit Ausgleich* [50]. Hierdurch werden Verzögerungsglieder Systemen mit integrierendem Verhalten gegenübergestellt, die als Systeme ohne Ausgleich

bezeichnet werden. Diese Vokabel wird einsichtig, wenn man sich bewusst macht, dass es sich bei einem stabilen System erster Ordnung um einen rückgekoppelten Integrator handelt (siehe Bild 7-4), was man leicht nachrechnen kann.

Eine solche Rückkopplung tritt beispielsweise bei einem Tank auf, der neben dem Zufluss u auch einen zum Füllstand y proportionalen Abfluss v besitzt. Ein anderes Beispiel wäre ein Wärmespeicher, der proportional zu seiner Temperatur y Wärme an die Umgebung abgibt. Diese Verluste v sorgen für einen Ausgleich innerhalb des Systems, wodurch der Integrator gesättigt wird und das System stabil ist. Fehlt dieser Ausgleich, liegt integrierendes Verhalten vor.

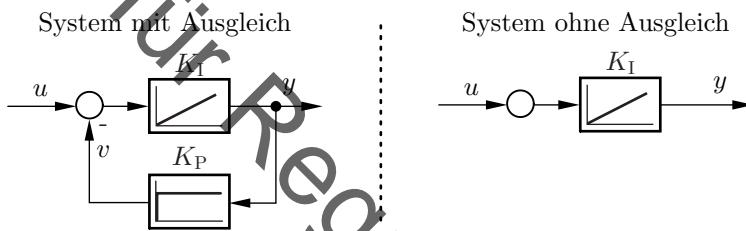


Bild 7-4: Wirkungspläne von Systemen mit und ohne Ausgleich

Verzögerungsglieder treten vorzugsweise bei der Beschreibung von Regelstrecken auf. Aber auch Mess- und Stellgeräte werden häufig durch PT₁- oder PT₂-Verhalten gekennzeichnet. Außerdem werden sie in der Digitaltechnik zum Filtern von verrauschten Messungen verwendet, da alle Verzögerungsglieder Tiefpassfilter sind.

Die Eignung als Tiefpassfilter sieht man beispielsweise an der asymptotischen Darstellung des PT₁ im Bode-Diagramm, wo der Amplitudengang zunächst konstant bleibt, um dann mit der Steigung -1 abzufallen. Hieraus ergibt sich auch die englische Bezeichnung des PT₁, das meistens als „first-order-lowpass“, also Tiefpass erster Ordnung, bezeichnet wird.

Die Grenzfrequenz (siehe Abschnitt 5.5) ω_g ist identisch zur Eckkreisfrequenz ω_E . Die Tatsache, dass dort genau

$$|G(j\omega_g)| = \frac{\sqrt{2}}{2} \quad \text{bzw.} \quad |G(j\omega_g)|^2 = \frac{1}{2} \quad (7.34)$$

gilt, hat zu einer alternativen Definition der Grenzfrequenz geführt:

Grenzfrequenz und Bandbreite

Die Frequenz ω , für die der Frequenzgang im Amplitudengang den $\sqrt{2}/2$ -fachen Wert seines Maximums annimmt, heißt Grenzfrequenz ω_g .

Die Frequenz ω , für die das Frequenzspektrum im Amplitudengang den $\sqrt{2}/2$ -fachen Wert seines Maximums annimmt, heißt Bandbreite ω_g .

Beide Definitionen sind genau analog – es haben sich lediglich unterschiedliche Fachwörter etabliert, je nachdem ob man von (prinzipiell gleichwertigen) Frequenzgängen oder Frequenzspektren spricht. Hintergrund ist, dass $|G(j\omega)|^2$ oft ein Maß für die Leistung eines Signals ist. Die Bandbreite gibt dann an, für welche Frequenz der halbe maximale Leistungswert erreicht wird.

7.3.2 PT₂

Das PT₂-Glied, Verzögerungsglied zweiter Ordnung, (Tab. 7-2) unterscheidet sich in mehreren Punkten vom PT₁-Glied, sodass es hier gesondert betrachtet werden soll.

Das PT₂-Glied kann als Prototyp komplex konjugierter Polstellen verstanden werden (vgl. Abschnitt 6.2). Insbesondere können alle Verzögerungsglieder höherer als zweiter Ordnung als Reihenschaltung von PT₁- und PT₂-Gliedern aufgefasst werden.

Um die allgemeine Differentialgleichung des PT₂-Gliedes

$$a_2 \ddot{y} + a_1 \dot{y} + a_0 y = b_0 u \quad (7.35)$$

zu lösen, sind die Nullstellen $\lambda_{1,2}$ des charakteristischen Polynoms

$$\lambda_{1,2} = -\frac{a_1}{2a_2} \pm \sqrt{\frac{a_1^2}{4a_2^2} - \frac{a_0}{a_2}} \quad (7.36)$$

zu bestimmen.

Abhängig vom Vorzeichen des Ausdruckes unter dem Wurzelzeichen in Gl.(7.36) wird das PT₂ dabei rein reelle oder komplexe Eigenwerte besitzen.

Entsprechend ergibt sich für die homogene Lösung der Differentialgleichung entweder

$$y_h = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t} \quad (\lambda_1 \neq \lambda_2 \text{ reell}) \quad , \quad y_h = C t e^{\lambda t} \quad (\lambda_1 = \lambda_2) \quad (7.37)$$

oder

$$y_h = e^{\alpha t} (A \cos \omega t + B \sin \omega t) \quad (\lambda_{1,2} = \alpha \pm j\omega, \text{ komplex}) . \quad (7.38)$$

Für reelle $\lambda_{1,2}$ erhält man eine aperiodisch verlaufende Zeitfunktion, während das Paar konjugiert komplexer Polstellen $\lambda_{1,2}$ zu einem schwingenden Verlauf der Zeitfunktion führt, der im Fall negativen Realteils α abklingt. Durch eine andere Schreibweise der Differentialgleichung, die in Gl.(2.43) bereits vorgestellt worden ist, kann dieser Zusammenhang noch deutlicher gemacht werden:

Dämpfungsgrad und Kennkreisfrequenz

Die Differentialgleichung eines PT₂ kann in der Form

$$\ddot{y} + 2D\omega_0 \dot{y} + \omega_0^2 y = K\omega_0^2 u \quad (7.39)$$

geschrieben werden, mit der statischen Verstärkung K , D als (dimensionslosem) Dämpfungsgrad und ω_0 als Kennkreisfrequenz.

Mit Gl.(7.39) ergeben sich die Wurzeln des charakteristischen Polynoms zu

$$\lambda_{1,2} = -\omega_0(D \pm \sqrt{D^2 - 1}) \quad (7.40)$$

und man unterscheidet je nach Größe des Dämpfungsgrades

$D > 1$	aperiodische Lösung	$\lambda_{1,2}$ reell
$D = 1$	aperiodischer Grenzfall	$\lambda_1 = \lambda_2$, reell
$0 < D < 1$	gedämpfte Schwingung	$\lambda_{1,2}$ konjugiert komplex
$D = 0$	ungedämpft, grenzstabil	$\lambda_{1,2}$ konjugiert komplex
$D < 0$	instabil	

Bild 7-5 zeigt die Lage der Polstellen der zugehörigen Übertragungsfunktion. Man erkennt, dass die Polstellen mit abnehmendem Dämpfungsgrad zunächst auf der reellen Achse aufeinander zu wandern. Für $D = 1$ bilden

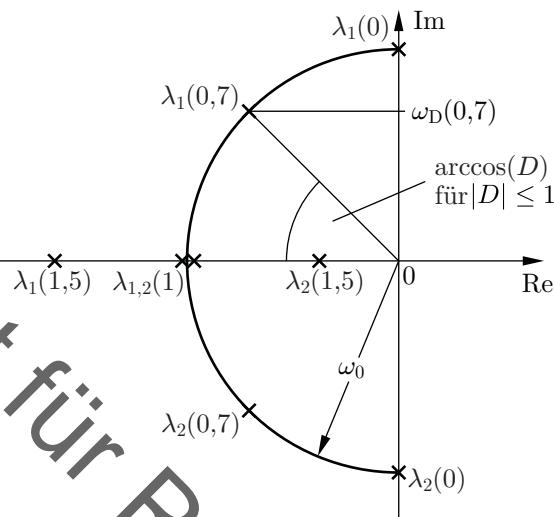


Bild 7-5: Polstellen eines PT_2 -Gliedes für unterschiedliche D und konstantes ω_0

sie dann eine doppelte Polstelle. Für einen weiter abnehmenden Dämpfungsgrad (bei $\omega_0 = \text{konst}$) beschreiben die Polstellen Kreisbögen um den Koordinatenursprung in der komplexen Ebene, wobei sie für $D = 0$ auf der imaginären Achse liegen.

Für $D \geq 1$ ist $\lambda_{1,2}$ reell und man kann gemäß der Ausführungen in Abschnitt 6.2 das PT_2 als System zweiter Ordnung in die Reihenschaltung zweier Systeme erster Ordnung zerlegen.

Konkret bringt man mit der Wahl

$$\omega_0^2 = \frac{1}{T_1 T_2} \quad , \quad D = \frac{1}{2} \frac{T_1 + T_2}{\sqrt{T_1 T_2}}$$

die Differentialgleichung auf die Form

$$T_1 T_2 \ddot{y} + (T_1 + T_2) \dot{y} + y = K \cdot u \quad (7.42)$$

mit reellen Zeitkonstanten T_1 und T_2 .

Eine solche Differentialgleichung bzw. die zugehörige Übertragungsfunktion

$$G(s) = \frac{K}{T_1 T_2 s^2 + (T_1 + T_2)s + 1} = K \cdot \frac{1}{1 + sT_1} \cdot \frac{1}{1 + sT_2} \quad (7.43)$$

beschreibt das dynamische Verhalten einer Reihenschaltung von zwei Verzögerungsgliedern erster Ordnung mit den Zeitkonstanten T_1 und T_2 .

Eine solche Reihenschaltung hat für beliebige positive reelle Zeitkonstanten stets eine aperiodisch verlaufende Übergangsfunktion, weil mit Gl.(7.41) $D \geq 1$ ist. Ist eine der beiden Zeitkonstanten T_1, T_2 wesentlich kleiner als die andere, so wird die größere Zeitkonstante die dominante Polstelle sein und das PT₂ sich fast identisch zu einem PT₁ verhalten.

Die Darstellung des Frequenzganges

$$G(j\omega) = \frac{K}{T_1 T_2 (j\omega)^2 + (T_1 + T_2)j\omega + 1} = K \cdot \frac{1}{1 + j\omega T_1} \cdot \frac{1}{1 + j\omega T_2} \quad (7.44)$$

im Bode-Diagramm gewinnt man, indem man z. B. die Teilstrecken entsprechend den dafür geltenden Regeln graphisch multipliziert.

Zusätzliche Werte können Tab. 7-4 entnommen werden. Man erkennt, dass der Phasengang für große Werte der Kreisfrequenz gegen -180° und die Steigung des Amplitudengangs gegen -2 gehen. Die Ortskurve des Frequenzganges durchläuft entsprechend dem Winkelverlauf im Bode-Diagramm den 4. und 3. Quadranten. Da für große Werte der Kreisfrequenz

$$G(\omega \rightarrow \infty) = \frac{K}{T_1 T_2} \cdot \frac{1}{(j\omega)^2} = -\frac{K}{T_1 T_2} \cdot \frac{1}{\omega^2} \quad (7.45)$$

gilt, läuft die Ortskurve mit der negativen reellen Achse als Asymptote in den Nullpunkt.

Für $D = 1$ wird in Gl.(7.41) $T_1 = T_2$. Die beiden Pole der Übertragungsfunktion Gl.(7.43) bei $-1/T_1$ und $-1/T_2$ fallen zu einem reellen Doppelpol zusammen. $D = 1$ ist der kleinste Wert des Dämpfungsgrades, für den eine aperiodische Lösung existiert und z. B. die Übergangsfunktion nicht über ihren Endwert hinaus überschwingt (Bild 7-6). Er wird daher auch aperiodischer Grenzfall genannt.

Aperiodischer Grenzfall

Besitzt ein System nur stabile reelle Polstellen, von denen mindestens eine mehrfach ist, so spricht man von einem aperiodischen Grenzfall, da eine geeignete minimale Änderung der Systemparameter zu einem schwingungsfähigen System führt.

Überschwingen, Unterschwingen

Nimmt die Sprungantwort eines stabilen Systems sowohl Werte kleiner als auch größer des statischen Endwertes $t \rightarrow \infty$ an, so spricht man von *Überschwingen*.

Nimmt die Sprungantwort eines stabilen Systems sowohl Werte kleiner als auch größer des Startwertes $t \rightarrow -0$ an, so spricht man von *Unterschwingen*.

Schwingungsfähigkeit und Überschwingen haben nur bedingt miteinander zu tun. Nicht alle schwingungsfähigen Systeme schwingen über und nicht alle Systeme, die überschwingen, sind schwingungsfähig.

Man erkennt aus Tab. 7-2, dass für $D = 0$ die Pole auf der imaginären Achse liegen. In diesem Fall führt das Übertragungsglied bei Anregung mit einem Sprung ungedämpfte Schwingungen mit der Kreisfrequenz ω_0 aus. Das System ist am Stabilitätsrand.

Für $1 > D > 0$ ist das Verzögerungsglied zweiter Ordnung nicht als Reihenschaltung einfacher Systeme mit reellen Koeffizienten darstellbar. Dies liegt daran, dass die Polstellen komplexe Werte annehmen und die Zerlegung in reelle Polstellen scheitert. Das System ist schwingungsfähig, weswegen die Übergangsfunktion damit über ihren Endwert hinaus schwingt (siehe Bild 7-6). Die Übertragungsfunktion weist ein konjugiert komplexes Polpaar auf (Bild 7-5), dessen Lage durch die Kennkreisfrequenz ω_0 und den Dämpfungsgrad D bestimmt wird und dessen Imaginärteil gleich der Eigenkreisfrequenz ω_D ist – also der Frequenz mit der ein System nach Anregung mit einem Sprung schwingt.

Für das PT_2 -System ergibt sich die Eigenkreisfrequenz aus dem Dämpfungsgrad D und der Kennkreisfrequenz ω_0 zu

$$\omega_D = \omega_0 \sqrt{1 - D^2} < \omega_0 \quad . \quad (7.46)$$

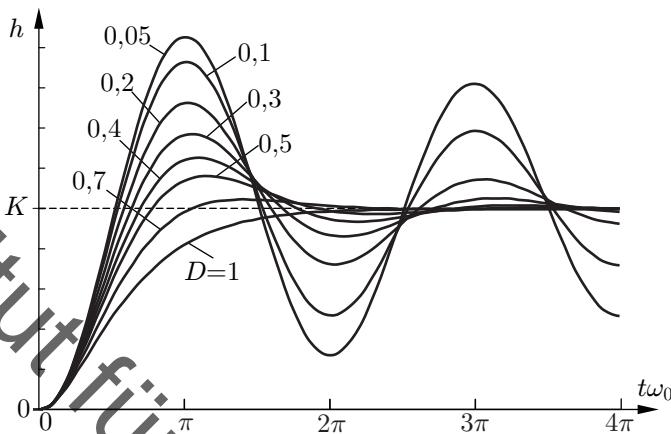


Bild 7-6: Übergangsfunktion für PT_2 -Glieder mit $0,05 \leq D \leq 1$

Der Frequenzgang

$$G(j\omega) = \frac{K\omega_0^2}{(j\omega)^2 + 2D\omega_0 j\omega + \omega_0^2} \quad (7.47)$$

wird für $1 \geq D \geq 0$ im Bode-Diagramm mit nur einer Eckkreisfrequenz $\omega_E = \omega_0$ dargestellt.

Die Asymptoten des Amplitudengangs haben die Steigungen 0 und -2 ; der tatsächliche Amplitudengang weicht u. U. erheblich vom Verlauf der Asymptoten ab, wie Bild 7-7 zeigt.

Der Phasengang hat die Asymptoten $\varphi_{\text{Asymp.}} = 0^\circ$ für kleine und $\varphi_{\text{Asymp.}} = -180^\circ$ für große Kreisfrequenzwerte.

Tab. 7-4 enthält Korrekturwerte, nämlich die Abweichungen des Betrags und des Phasenwinkels von den Asymptoten. Hiermit können Amplituden- und Phasengang mit ausreichender Genauigkeit dargestellt werden. Es hat sich als zweckmäßig erwiesen, diese Korrekturwerte als Funktion von ω/ω_E anzugeben, u. a. weil die Korrekturwerte für einen Argumentwert ω_1/ω_E und für dessen Reziprokwert $\omega_2/\omega_E = \omega_E/\omega_1$ gleich groß sind. Daher sind in Tab. 7-4 nur Werte für $0 < \omega/\omega_E \leq 1$ angegeben.

Die aufgeführten Werte gelten dabei gemäß Tab. 6-2 auch für instabile Pole

$\frac{\omega}{\omega_E}$ bzw. $\frac{\omega_E}{\omega}$	$\lg(G) - \lg(\text{Asymptote})$				$ \varphi - \varphi_{\text{Asymptote}} $			
	0,1	0,5	0,8	1	0,1	0,5	0,8	1
PT ₁ bzw. mit anderem Vorzeichen für doppelte Pol- und Nullstellen	-0,002	-0,048	-0,107	-0,151	5,7	26,6	38,7	45,0
	-0,004	-0,097	-0,215	-0,301	11,4	53,1	77,3	90,0
	0,707	0,000	-0,013	-0,075	-0,151	8,1	43,4	72,3
	0,5	0,002	0,045	0,057	0,000	5,8	33,7	65,8
	0,4	0,003	0,071	0,134	0,097	4,6	28,1	60,6
	0,3	0,004	0,093	0,222	0,222	3,5	21,8	53,1
	0,2	0,004	0,110	0,317	0,398	2,3	14,9	41,6
	0,1	0,004	0,121	0,405	0,699	1,2	7,6	24,0
	0,05	0,004	0,124	0,433	1,000	0,6	3,8	12,5
								90,0

Tabelle 7-4: Korrekturwerte für Betrag (in log. Einheiten) und Phasenwinkel (in Winkelgraden)

oder Nullstellen, indem die Korrekturwerte mit den passenden Asymptoten und ggf. geändertem Vorzeichen verrechnet werden.

Man erkennt, dass für $D > \sqrt{2}/2$ der Amplitudengang für alle Kreisfrequenzen kleiner ist als der durch die Asymptoten angegebene Wert und für $\omega \rightarrow 0$ gegen die Asymptoten strebt. Daher liegt das Maximum des Amplitudengangs bei $\omega = 0$ und fällt danach kontinuierlich ab. Dieses Tiefpassverhalten ist für viele Systeme typisch. Für kleinere Werte des Dämpfungsgrades ist dieses Verhalten nicht mehr gegeben.

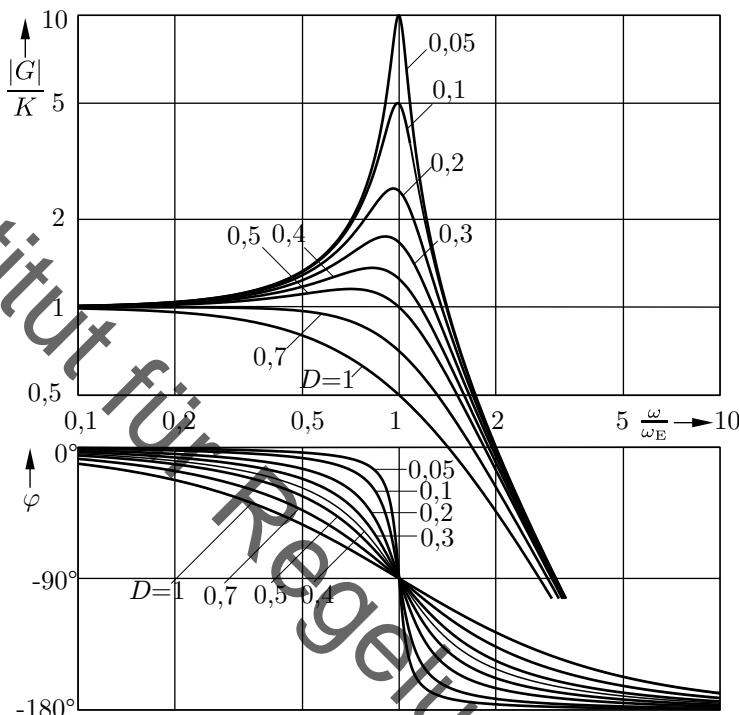


Bild 7-7: Bode-Diagramm für PT_2 -Glieder mit $0,05 \leq D \leq 1$

Resonanzüberhöhung

Nimmt der Amplitudengang eines Systems sein Maximum nicht bei $\omega = 0$ an, so spricht man von einer *Resonanzüberhöhung*. Die zugehörige Frequenz, bei der das Maximum erreicht wird, heißt *Resonanzfrequenz* ω_{res} .

Ein System kann auch mehrere Resonanzfrequenzen besitzen, wenn es aus mehreren Teilsystemen mit Resonanzüberhöhung zusammengesetzt ist. Für $D < \sqrt{2}/2$ zeigt der Amplitudengang des PT_2 eine Resonanzüberhöhung, die für $D \rightarrow 0$ über alle Grenzen geht. Die Resonanzfrequenz lässt sich dabei für das PT_2 -System zu

$$\omega_{\text{res}} = \omega_0 \sqrt{1 - 2D^2} < \omega_D < \omega_0 \quad (7.48)$$

berechnen und ist ein wenig kleiner als die Eckkreisfrequenz ω_0 oder die Eigenkreisfrequenz ω_D .

Der Wert $D = \sqrt{2}/2 \approx 0,707$ wird häufig bei Messgeräten oder Tiefpassfiltern zweiter Ordnung angestrebt. Wie Bild 7-6 zeigt, schwingt die Übergangsfunktion eines solchen Gliedes über ihren Endwert hinaus, obgleich der Amplitudengang keine Resonanzüberhöhung aufweist. Die Grenzfrequenz ergibt sich in diesem Fall wie beim PT₁ zu $\omega_g = \omega_0$.

In der Ortskurve des Frequenzganges von PT₂-Gliedern mit $1 \geq D \geq 0$ wird die Resonanzüberhöhung aber $D < \sqrt{2}/2$ in Form einer Ausbuchtung sichtbar, wie aus Bild 7-8 zu ersehen ist.

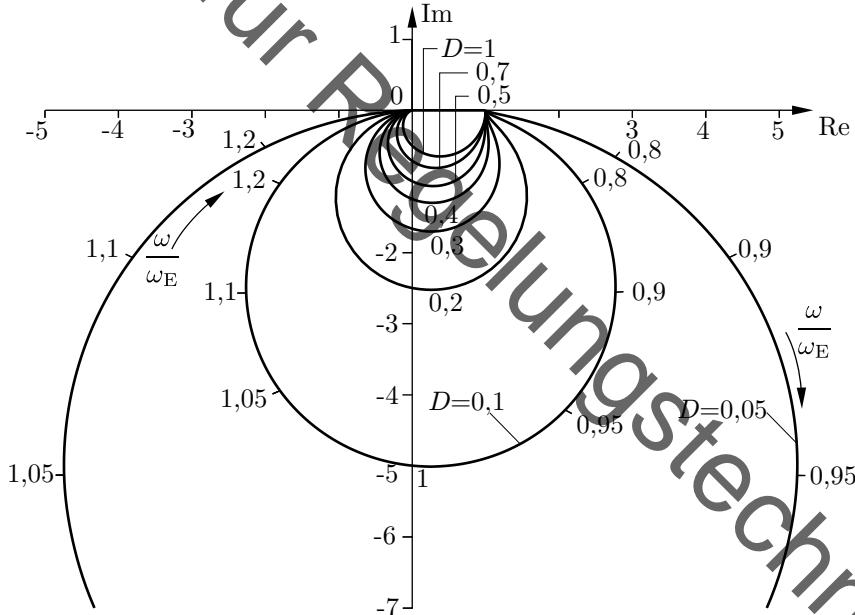


Bild 7-8: Frequenzgangortskurven für PT₂-Glieder mit $0,05 \leq D \leq 1$

Das PT₂ kann prototypisch für ein System mit zwei Polstellen gedeutet werden.

7.3.3 PT_n

Verzögerungsglieder n -ter Ordnung (PT_n) lassen sich als Reihenschaltung von Verzögerungsgliedern erster und zweiter Ordnung auffassen: Hierbei können die PT_1 -Teilsysteme für die Dynamik der reellen Polstellen genutzt werden, während die PT_2 -Anteile mit komplex-konjugierten Polstellen die schwingungsfähigen Dynamiken abdecken.

Für ein PT_n hat die Differentialgleichung die allgemeine Form

$$a_n y^{(n)} + \dots + a_1 \dot{y} + a_0 y = b_0 u \quad , \quad (7.49)$$

d. h. auf der rechten Seite stehen entsprechend der Definition eines Verzögerungsgliedes keine Ableitungen der Eingangsgröße. Demzufolge weist die Übertragungsfunktion nur Pole und keine Nullstellen auf.

Der Frequenzgang

$$G(j\omega) = \frac{b_0}{a_n(j\omega)^n + \dots + a_1 j\omega + a_0} \quad (7.50)$$

geht für große Werte der Kreisfrequenz gegen

$$G(\omega \rightarrow \infty) = \frac{b_0}{a_n} \cdot \frac{1}{(j\omega)^n} \quad . \quad (7.51)$$

Das bedeutet, dass der Phasenwinkel gegen $-n \cdot 90^\circ$ geht, die Ortskurve n Quadranten durchläuft und die Steigung des Amplitudengangs im Bode-Diagramm gegen $-n$ strebt. Diese Resultate ergeben sich auch direkt aus der Addition der Teilsysteme erster oder zweiter Ordnung im Bode-Diagramm, in die das PT_n zerlegt werden kann.

Die Übergangsfunktionen aller Verzögerungsglieder von höherer als erster Ordnung haben die Gemeinsamkeit, dass die Tangente im Zeitnullpunkt waagerecht verläuft; lediglich die Übergangsfunktion des Verzögerungsgliedes erster Ordnung (PT_1) hat im Zeitnullpunkt eine von null verschiedene Steigung. Dies ergibt sich direkt aus den Zusammenhängen zwischen Anfangssteigung und relativem Grad aus Abschnitt 4.6.

7.4 Kombinationen

Aus den Verzögerungsgliedern und Systemen vom PID-Typ lassen sich durch Reihen- und Parallelschaltung eine Vielzahl von Elementen erster

oder zweiter Ordnung gewinnen, von denen einige Systeme häufiger als Regler oder Regelstrecken anzutreffen sind.

7.4.1 IT_1

IT_1 -Glieder entstehen durch Reihenschaltung eines integrierenden (I) und eines Verzögerungsgliedes erster Ordnung (PT_1). Manche Regelstrecken und integrierend wirkende Stellantriebe mit Verzögerung weisen IT_1 -Verhalten auf.

Die Beschreibungen in Tab. 7-2 lassen sich ohne besondere Schwierigkeiten aus der Reihenschaltung von I- und PT_1 -Glied gewinnen. Zu bemerken ist allenfalls, dass man zeigen kann, dass die Asymptote der Übergangsfunktion die Zeitachse im Punkte T schneidet mit T als Zeitkonstante (Tab. 7-2). Die Existenz der Asymptote als Gerade zeigt, dass das IT_1 -Element integrierendes Verhalten besitzt.

Der integrierende Anteil wird im Frequenzgang durch den Faktor $1/(j\omega)$ und in der Übertragungsfunktion durch einen Pol im Koordinatenursprung dargestellt. Wegen dieses Faktors $1/(j\omega)$ geht der Phasenwinkel für kleine Werte ω gegen -90° , die Steigung des Amplitudengangs im Bode-Diagramm gegen -1 und der Betrag selbst über alle Grenzen. Zudem strebt der Realteil des Frequenzganges

$$G(j\omega) = \frac{K_I}{j\omega(1 + j\omega T)} \quad (7.52)$$

für kleine Werte ω gegen den endlichen Wert $-K_IT$. Außerdem entspricht die Sprungantwort des IT_1 strukturell der *Rampenantwort* eines Übertragungssystems mit PT_1 -Verhalten.

Rampenantwort

Analog zur Sprung- und Impulsantwort bezeichnet man den Verlauf der Ausgangsgröße bei Anregung mit einer Rampe $u(t) = K \cdot t$ für $t > 0$ als *Rampenantwort*.

Die Übereinstimmung von Sprungantwort des IT_1 und Rampenantwort des

PT_1 sieht man am leichtesten im Bildbereich. Hier ergibt sich

$$\mathcal{L}\{h_{IT_1}(t)\} = \frac{1}{s} \cdot \frac{K_I}{s(1+sT)} = \underbrace{\frac{1}{s^2}}_{\mathcal{L}\{t\}} \cdot \underbrace{\frac{K_I}{1+sT}}_{PT_1} . \quad (7.53)$$

Anschaulich kann aufgrund der Kommutativität der Reihenschaltung die im IT_1 intern vorgenommen Integration in die Zeitintegration des Eingangssignals verschoben werden. Dies entspricht strukturell dem bekannten Zusammenhang zwischen $g(t)$ und $h(t)$. Folglich ist auch die Gewichtsfunktion des IT_1 mit der Übergangsfunktion des PT_1 identisch.

Die Rampenantwort ist vor allem bei der Folgeregelung wichtig. Die normierte Rampenantwort hat allerdings keinen speziellen Namen.

Festwertregelung und Folgeregelung

Regelungen, die vorwiegend Störungen unterdrücken sollen und bei konstanten Führungsgrößen arbeiten, werden *Festwertregelung* genannt. Regelungen, die in erster Linie die Regelgröße einer sich ändernden Führungsgröße nachführen sollen, werden *Folgeregelung* genannt.

Möchte man die Qualität einer Folgeregelung für $t \rightarrow \infty$ untersuchen, so wird für alle stationär genau arbeitenden Regelkreise die Sprungantwort wenig Aufschluss geben, da die bleibende Regelabweichung für all diese Regelkreise null und somit identisch ist.

Betrachtet man aber die Antwort eines solchen Regelkreises auf eine Rampe, so wird sich – ganz analog zum Verhalten des IT_1 -Glieds im Vergleich zum I-Glied – ein sogenannter *Schleppfehler* (manchmal auch *Geschwindigkeitsfehler* genannt) einstellen (siehe Bild 7-9). Er beschreibt als Kennwert den Fehler, der sich nach Abklingen des Anlaufvorganges einstellt.

Der Geschwindigkeitsfehler ist genau dann endlich, wenn die Führungsgröße und die Regelgröße für große Werte der Zeit t parallel zueinander, d. h. mit gleicher Steigung verlaufen. Da die Steigung der Rampenantwort mit dem Wert der Sprungantwort identisch ist, ist diese Bedingung für alle stationär genauen Regelkreise erfüllt. Der Schleppfehler wird dann zu einem konstanten Wert, der es ermöglicht, das Folgeverhalten von stationär genauen Regelkreisen zueinander ins Verhältnis zu setzen.

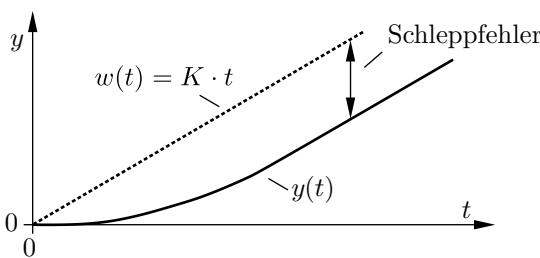


Bild 7-9: Rampenantwort eines einfachen Folgesystems

7.4.2 DT₁

DT₁ bezeichnet das dynamische Verhalten eines Differenzierers mit Verzögerung erster Ordnung, der als Reihenschaltung eines D- und eines PT₁-Gliedes aufgefasst werden kann.

Ein D-Glied ohne Verzögerung ist nicht kausal und damit technisch nicht realisierbar. Das sieht man daran, dass das D-Glied mehr Null- als Polstellen besitzt. Nach Tab. 6-2 wird für aukausale Systeme folglich der Betrag des Frequenzgangs für große Werte der Kreisfrequenz über alle Grenzen wachsen. Dies ist eine weitere physikalische Erklärung für die Unmöglichkeit, aukausale Systeme technisch zu realisieren, da diese unendlich hohe Frequenzen unendlich stark verstärken müssten.

In diesem Sinne kausalisiert man D-Glieder durch das Hinzufügen zusätzlicher Verzögerungsglieder. Daher hat sich als alternative Bezeichnung für das DT₁-Element auch *realer Differenzierer* eingebürgert.

Die zusätzlichen Verzögerungsglieder können je nach Größe ihrer Zeitkonstanten und dem Anwendungsfall möglicherweise vernachlässigt werden. Da die Nullstelle des Differenzierers bei $s = 0$ sehr dominant ist, werden kleine Zeitkonstanten fast keine Auswirkungen auf die Dynamik haben.

Die Differentialgleichung des DT₁-Gliedes ist

$$T\dot{y} + y = K_D \dot{u} \quad (7.54)$$

und sein Frequenzgang

$$G(j\omega) = \frac{j\omega K_D}{1 + j\omega T} \quad . \quad (7.55)$$

Die Übergangsfunktion (Tab. 7-2) bleibt für alle Werte der Zeit endlich und geht gegen null für große Werte der Zeit, was zeigt, dass der reale Differenzierer differenzierendes Verhalten besitzt.

Die Ortskurve des Frequenzganges erhält man aus den Werten für $\omega = 0$ und $\omega \rightarrow \infty$, aus der Struktur der Gl.(7.55), die auf einen Kreis schließen lässt und aus der Tatsache, dass die Ortskurve im Uhrzeigersinn durchlaufen wird (Tab. 7-2).

Die Darstellung im Bode-Diagramm und die Pol-Nullstellendarstellung der Übertragungsfunktion erhält man dadurch, dass man von einer Zerlegung gemäß Tab. 6-2 in ein Teilsystem mit einer Nullstelle bei $s = 0$ und einem Pol bei $-1/T$ (oder gleichbedeutend einer Reihenschaltung mit einem D- und einem PT_1 -Glied) ausgeht.

Bemerkenswert ist, dass auch die Subtraktion eines P- und eines PT_1 -Gliedes mit gleichem Übertragungsfaktor gemäß Bild 7-10 DT₁-Verhalten aufweist, obgleich in beiden Teilsystemen keine Differentiation vorgenommen wird. Dies verifiziert die Rechnung

$$G(s) = \frac{K_D}{T} - \frac{K_D}{T} \frac{1}{1+sT} = \frac{\frac{K_D}{T}(1+sT) - 1}{1+sT} = \frac{K_D \cdot s}{1+sT} \quad . \quad (7.56)$$

Folglich entspricht die Sprungantwort des DT₁ der eines P-Elementes, von der die Sprungantwort eines PT_1 mit identischem Übertragungsfaktor abgezogen wird. Hierdurch läuft die Sprungantwort des DT₁ für $t \rightarrow \infty$ gegen den Endwert null.

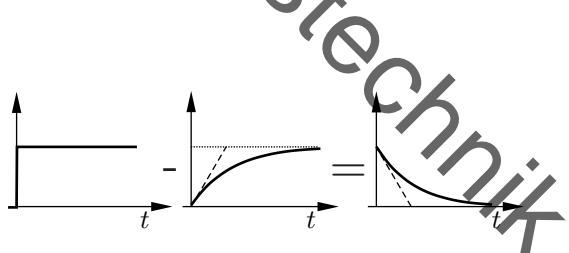
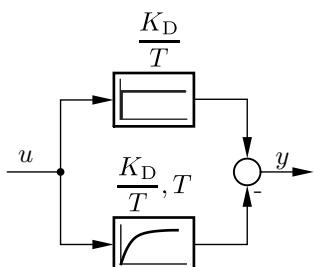


Bild 7-10: DT₁-Glied durch Parallelschaltung von P und PT_1

7.4.3 PIT₁

PIT₁-Verhalten weist ein PI-Regler mit Verzögerung erster Ordnung auf. Technisch ist dies beispielsweise dann gegeben, wenn ein PI-Regler mit einem Tiefpassfilter erster Ordnung in Reihe geschaltet wird.

Aus der Reihenschaltung eines PI- und eines PT₁-Gliedes lassen sich die Darstellungen in Tab. 7-3 mit geringem Aufwand herleiten. Da das genaue Aussehen aller zugehöriger Graphen davon abhängt, ob $-1/T_n$ oder $-1/T$ eine dominante Pol- bzw. Nullstelle ist, können die Verläufe in Tab. 7-3 nur als Orientierungshilfe dienen.

7.4.4 PPT₁ und PDT₁

Glieder mit der Differentialgleichung

$$T\ddot{y} + y = K(u + T_v \dot{u}) \quad (7.57)$$

mit genau einer Pol- und einer Nullstelle besitzen abhängig vom Verhältnis der Zeitkonstanten T und T_v unterschiedliche dynamische Eigenschaften und Bezeichnungen. Sie verhalten sich sehr unterschiedlich, je nachdem ob die Zeitkonstante T größer oder kleiner als T_v ist, da entweder die Polstelle in $-1/T$ oder die Nullstelle in $-1/T_v$ dominant ist. Im Sonderfall $T = T_v$ erhält man das Übertragungsverhalten eines P-Gliedes mit Verstärkung K .

Zur besseren Analyse der Verläufe im Zeitbereich zerlegt man das System zweckmäßig in eine Parallelschaltung

$$G(s) = K \cdot \frac{1 + T_v s}{1 + T s} = \underbrace{K \frac{T_v}{T}}_{\text{P-Glied}} + \underbrace{\frac{K(1 - \frac{T_v}{T})}{1 + T s}}_{\text{PT}_1\text{-Glied}} . \quad (7.58)$$

Für $T_v < T$ handelt es sich um eine Addition von einem P- und einem PT₁-Element, weswegen das Element als PPT₁ bezeichnet wird. Die Übergangsfunktion ergibt als Summe der Übergangsfunktionen von P- und PT₁-Glied und hat daher einen vom Startwert des P-Elements KT_v/T aus ansteigenden Verlauf, der im statischen Übertragungsfaktor K endet. In diesem Fall liegt die Polstelle $-1/T$ näher am Ursprung und ist dominant. Das Übertragungsverhalten ähnelt daher strukturell eher einem PT₁-Element.

Für $T < T_v$ ergibt sich hingegen eine Subtraktion eines P- und eines PT_1 -Elements. Die Übergangsfunktion wird daher ausgehend vom Startwert des P-Elements KT_v/T abfallen und dem nun kleineren Endwert K entgegenstreiben. Der Verlauf im Zeitbereich hat daher wenig Gemeinsamkeiten mit dem des PPT_1 -Gliedes. Da aus dem DT_1 bereits bekannt ist, dass sich dieses aus einer Subtraktion von P und PT_1 mit identischem Übertragungsfaktor ergibt, fasst man den Fall $T < T_v$ als eine Parallelschaltung von einem P-Element mit einem DT_1 -Element auf und vergibt den Namen PDT_1 . Dies passt auch dazu, dass die Nullstelle $-1/T_v$ hier näher am Ursprung liegt und somit dominant ist. Diese Dominanz schlägt sich darin nieder, dass die Sprungantwort eher einem PD-Element, als einem PT_1 gleicht.

Tatsächlich entsprechen beide Elemente einer Reihenschaltung von einem PD- und einem PT_1 -Element. Sie treten daher als kausalierte PD-Regler auf, indem einem PD-Regler eine Verzögerung erster Ordnung hinzugefügt wird. Hierbei wird man das Verhalten des idealen PD-Reglers imitieren wollen, weswegen $T < T_v$ gewählt werden muss, da nur so die Nullstelle dominant ist und der Verlauf dem eines PD-Reglers ähnelt.

Da das genaue Aussehen aller zugehöriger Graphen maßgeblich vom Verhältnis T_v/T abhängt, können die Verläufe in Tab. 7-3 nur als Orientierungshilfe dienen. Der Phasengang muss beispielsweise durch korrekte Addition der Phasengänge der Null- und der Polstelle wie in Tab. 6-2 bestimmt werden. Der Frequenzgang beider Elemente entspricht in seiner Form Gl.(5.19), weswegen die Ortskurve die Form eines Kreises hat.

Je nach Verhältnis der Zeitkonstanten T und T_v wird der Einfluss für niedrige Frequenzen wahlweise von der Nullstelle oder der Polstelle vorgegeben. Aus der Addition der Verläufe der Pol- und Nullstelle gemäß Tab. 6-2 ergeben sich dabei direkt die gezeigten Verläufe des Frequenzganges im Bode-Diagramm. Das PDT_1 , in welchem für niedrige Frequenzen die Nullstelle dominiert, besitzt dabei eine positive Phase und einen ansteigenden Amplitudengang. Genau umgekehrt verhält es sich bei PPT_1 , welches sich strukturell aus der Inversion des PDT_1 ergibt.

In der Regelungstechnik werden neben den bereits erwähnten Einsatzzwecken beide Elemente als sogenannte *Kompensationsglieder* für den Reglerentwurf verwendet. Ein später in Abschnitt 11.1 vorgestelltes Reglerentwurfsverfahren basiert nämlich auf der Idee, dem Frequenzgang des aufgeschnittenen Regelkreises $G_0(j\omega)$ eine bestimmte Form zu verleihen. Hier

kann man PDT₁ und PPT₁ nutzen, um gezielte, punktuelle Eingriffe in den Frequenzgang vorzunehmen, da beide Glieder außerhalb des Bereiches zwischen den beiden Zeitkonstanten weitestgehend konstante Verläufe in Betrag und Phase aufweisen. Hierbei wird man das PDT₁ als phasenanhebendes Glied verwenden, um die Phase in einem kleinen Frequenzbereich gezielt zu erhöhen. Das PPT₁ hingegen wird eingesetzt, um die Amplitude abzusenken. Aus dieser Verwendung ergeben sich auch die englischen Bezeichnungen als „lead“- (PDT₁) und „lag“-Element (PPT₁).

7.5 Nicht-minimalphasige Systeme

7.5.1 PA₁

Die bisher in diesem Kapitel diskutierten Elemente besitzen allesamt Pole und Nullstellen, deren Realteil negativ oder null war. Liegen Polstellen in der rechten Halbebene, so ist das System instabil. Aber auch Nullstellen in der rechten Halbebene verändern das dynamische Verhalten maßgeblich.

Tab. 7-3 zeigt die Beschreibung eines stabilen Allpasses erster Ordnung (PA₁). In der Differentialgleichung fällt das negative Vorzeichen vor einem Term der rechten Seite auf. Dies führt zu entsprechenden negativen Vorzeichen im Zähler von Übertragungsfunktion und Frequenzgang und somit einer Nullstelle, welche sich in der rechten offenen komplexen Halbebene befindet.

Der Betrag des Frequenzgangs ist konstant (weil die Beträge von Zähler und Nenner gleich sind) während der Phasengang wegen des negativen Vorzeichens im Zähler für große ω nach -180° strebt. Bei Allpässen höherer Ordnung (PA_n) geht der Phasenwinkel des Frequenzganges für große ω gegen $-180^\circ \cdot n$.

In der Anwendungspraxis sind Übertragungsglieder mit reinem Allpassverhalten nur selten anzutreffen. Allerdings besitzen zahlreiche Regelstrecken einen Allpassanteil in Form der in Abschnitt 6.3 vorgestellten Zerlegung. Dabei kann anhand der Übergangsfunktion des PA₁-Gliedes in Tab. 7-3 gut veranschaulicht werden, warum dies die Regelungsaufgabe erheblich erschwert.

Die Übergangsfunktion des PA₁-Elements in Tab. 7-3 wechselt zunächst in den negativen Wertebereich, besitzt aber eine positive statische Verstär-

kung. Obwohl eine Erhöhung der Eingangsgröße langfristig eine Erhöhung der Ausgangsgröße zur Folge hat, so scheint kurzfristig diese eine Absenkung der Ausgangsgröße zu bewirken. Somit wird ein Stelleingriff des Reglers die Regelabweichung zunächst erhöhen, was ein weiteres Wachsen des Stelleingriffs zur Folge hat. Für eine geringe Reglerverstärkung wird diese Anpassung des Stelleingriffs langsamer als die Dynamik des Allpasses ablaufen, sodass das korrekte Vorzeichen der Regelstrecke rechtzeitig in der Regelabweichung bemerkbar wird. Ist die Reglerverstärkung jedoch hoch, so erfolgt die Anpassung der Stellgröße schneller als die Dynamik des Allpasses und die Stellgröße wächst über alle Grenzen.

Die Tatsache, dass die Übergangsfunktion für kleine Zeiten ein anderes Vorzeichen als die statische Verstärkung besitzt, lässt sich mithilfe der Grenzwertsätze für alle Allpässe ungerader Ordnung (also eine ungerade Anzahl nicht-minimalphasiger Nullstellen) nachweisen. Beispielhaft wird ein stabiles System mit relativem Grad eins und einer nicht-minimalphasigen Nullstelle betrachtet. Die Grenzwertsätze für die Startsteigung und den Endwert liefern

$$\lim_{t \rightarrow 0} \dot{h}(t) = \lim_{s \rightarrow \infty} sG(s) = \lim_{s \rightarrow \infty} sK \frac{(s - \eta_1) \dots (s - \eta_{n-1})}{(s - \lambda_1) \dots (s - \lambda_n)} = K \quad (7.59)$$

sowie

$$\lim_{t \rightarrow \infty} h(t) = \lim_{s \rightarrow 0} G(s) = \lim_{s \rightarrow 0} K \frac{(s - \eta_1) \dots (s - \eta_{n-1})}{(s - \lambda_1) \dots (s - \lambda_n)} = K \underbrace{\frac{\prod\limits_{i=1}^{n-1} -\eta_i}{\prod\limits_{i=1}^n -\lambda_i}}_{<0}. \quad (7.60)$$

Somit besitzen Endwert und Anfangssteigung ein unterschiedliches Vorzeichen und die Übergangsfunktion läuft zunächst in die falsche Richtung (Unterschwingen).

Diese Eigenschaft lässt sich auch dadurch plausibilisieren, dass man einen Allpass in eine Parallelschaltung zerlegt. Dies kann beispielsweise durch

$$G(s) = K \cdot \frac{1 - sT}{1 + sT} = \underbrace{\frac{K}{1 + sT}}_{\text{PT}_1} - \underbrace{\frac{KTs}{1 + sT}}_{\text{DT}_1} \quad (7.61)$$

erfolgen.

Diese Parallelschaltung ist in Bild 7-11 gezeigt. Offenbar gibt es zwei parallele Übertragungskanäle mit unterschiedlichem Vorzeichen. Der negative Übertragungskanal (DT_1) besitzt dabei aufgrund seiner Sprungfähigkeit einen hohen Anfangsausschlag, aber auch wegen seines differenzierenden Verhaltens einen statischen Endwert von null. Der positive Übertragungskanal (PT_1) wächst hingegen langsamer (relativer Grad von eins) und besitzt den statischen Endwert K . Folglich wird für kleine Zeiten der negative Übertragungskanal, für große Zeiten der positive Übertragungskanal dominieren. Dieses Phänomen veranschaulicht die schlechte Regelbarkeit und das wechselnde Vorzeichen innerhalb der Sprungantwort.

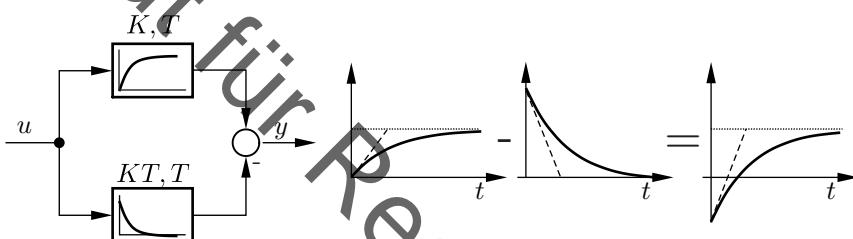


Bild 7-11: Parallelschaltung mit Allpassverhalten

Nicht-minimalphasige Nullstellen treten in einem technischen System stets dann auf, wenn es wie in Bild 7-11 zwei Wirkpfade unterschiedlichen Vorzeichens und unterschiedlich schneller Dynamik besitzt. Außerdem entstehen bei der Stabilisierung instabiler Systeme durch Regelungen ebenfalls nicht-minimalphasige Nullstellen.

Zur Illustration wird die Übertragungsfunktion von der Führungsgröße w auf die Stellgröße u in Bild 6-2 betrachtet. Diese berechnet sich zu

$$G_u(s) = \frac{U(s)}{W(s)} = \frac{G_R(s)}{1 + G_0(s)} = \frac{N_S(s)Z_R(s)}{Z_S(s)Z_R(s) + N_S(s)N_R(s)} . \quad (7.62)$$

Folglich sind wegen des Ausdrucks $N_S(s)$ im Zähler alle Polstellen der Regelstrecke auch Nullstellen dieser Stellgrößenübertragungsfunktion. Diese besitzt also für eine instabile Regelstrecke nicht-minimalphasige Nullstellen, welche man in ihren Auswirkungen auf die zeitliche Dynamik gut beobachten kann.

Als Beispiel hierfür wird in Bild 7-12 ein inverses Pendel ähnlich zu Bild

3-1 betrachtet, das sich in der oberen Ruhelage befindet und dessen unterer Aufpunkt durch einen Wagen verschoben werden kann. Aus Kapitel 2 ist bekannt, dass das Pendel in seiner oberen Ruhelage instabil ist.

Es wird angenommen, dass ein stabilisierender Regler (z. B. ein Mensch) den Wagen passend bewegt, um das Pendel in eine neue gewünschte Position zu verfahren und dabei die obere instabile Ruhelage beizubehalten. Wie wird man den Wagen anfangs verschieben müssen, um das Pendel insgesamt zu einem neuen Zielort nach rechts zu bewegen? Die einzige Möglichkeit besteht darin, dass Pendel geeignet zu neigen und den Wagen zunächst vom geplanten Zielort weg nach links zu bewegen. Folglich wirkt sich eine Änderung der Führungsgröße (Zielort) auf die Stellgröße (Wagenposition) zunächst mit dem genau umgekehrten Vorzeichen aus. Dies ist genau der beschriebene Effekt der nicht-minimalphasigen Nullstelle der Übertragungsfunktion von $w(t)$ auf $u(t)$. Ähnliche Effekte sind beispielsweise auch beim Lenken eines Fahrrades zu beobachten [38].

7.5.2 \mathbf{PT}_t , $\mathbf{PT}_1\mathbf{T}_t$

Alle bisherigen Übertragungsglieder konnten durch lineare gewöhnliche Differentialgleichungen mit konstanten Koeffizienten beschrieben werden. Diese Modellform reicht dabei zur Beschreibung nahezu aller häufig vorkommenden Systeme aus. Hiervon gibt es nur eine nennenswerte Ausnahme: Die sogenannten Totzeit-Glieder. Diese werden durch die Gleichung

$$y(t) = K \cdot u(t - T_t) \quad (7.63)$$

beschrieben. Diese sagt aus, dass die Ausgangsgröße gleich ist dem Wert der Eingangsgröße zu einem um die Totzeit T_t früher gelegenen Zeitpunkt, multipliziert mit dem Verstärkungsfaktor K .

Die Übergangsfunktion ergibt sich entsprechend als ein Sprung auf den Wert K zum Zeitpunkt T_t wie in Tab. 7-3 dargestellt.

Derartige Zeitverläufe treten bei allen Transportprozessen auf und das Totzeit-Glied kann prototypisch als Transportband mit konstanter Geschwindigkeit gedeutet werden – siehe auch Bild 7-13. Hier stellen sich am Ende des Bandes (d. h. als Ausgangsgröße) genau die Werte ein, die man an den Anfang des Bandes (Eingangsgröße) vor T_t Sekunden gelegt hat. Diese Form des Transportes tritt bei nahezu jeder Form der Signalübermittlung auf und

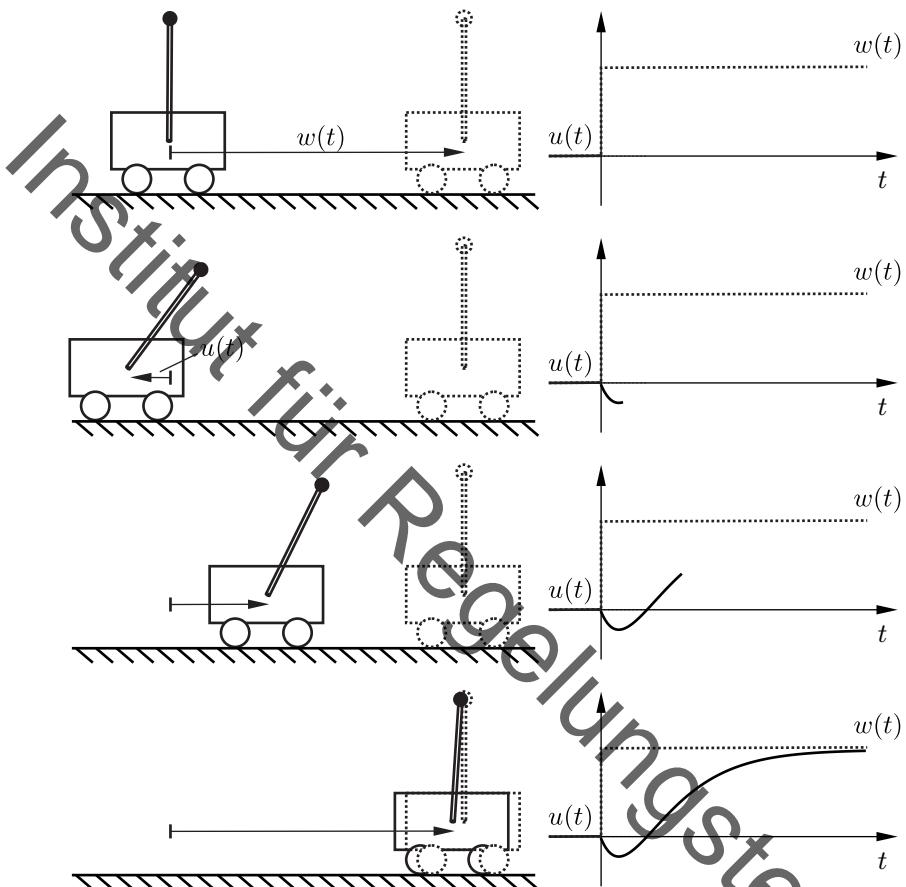


Bild 7-12: Zeitverlauf bei Änderung der Führungsgröße eines geregelten inversen Pendels

wird daher (bei einer ggf. sehr kleinen Totzeit) Teil jeder vollständigen Streckenbeschreibung sein.

Die Beschreibung in Gl.(7.63) unterscheidet sich von den bisher behandelten Differentialgleichungen dadurch, dass keine Ableitungen von Ein- oder Ausgangsgrößen auftreten und dass die Argumente der Zeitfunktionen auf

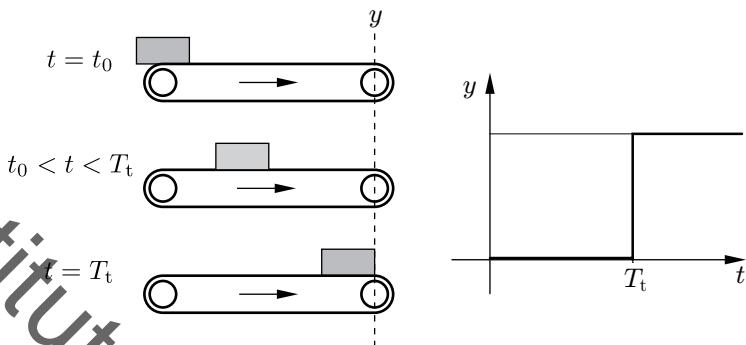


Bild 7.13: Transportprozess als Totzeit-Glied

der rechten und der linken Gleichungsseite voneinander verschieden sind. Mathematisch entspricht dabei der Transportprozess einer partiellen statt einer gewöhnlichen Differentialgleichung, die im Zeitbereich in der Form Gl.(7.63) gefasst werden kann [11].

Diese Eigenschaften finden sich in der speziellen Form des zugehörigen Frequenzganges wieder. Analog zu der Vorgehensweise in Abschnitt 5.1 wird Gl.(7.63) über die Zeigerdarstellung untersucht. Einsetzen der Größen führt unter Beachtung der unterschiedlichen Argumente zu

$$Ye^{j\omega t} = K \cdot U \cdot e^{j\omega(t-T_t)} = K \cdot U \cdot e^{j\omega t} \cdot e^{-j\omega T_t} . \quad (7.64)$$

Daraus ergibt sich der gesuchte Frequenzgang als Quotient direkt zu

$$G(j\omega) = \frac{y}{u} = K \cdot e^{-j\omega T_t} . \quad (7.65)$$

Der gewonnene Frequenzgang ist also im Gegensatz zu den bisherigen Frequenzgängen keine rationale sondern eine transzendente Funktion der Kreisfrequenz ω . Daher unterscheiden sich auch die graphischen Darstellung in Ortskurve und Bode-Diagramm stark von den Darstellungen der bisher behandelten Elemente. Die zugehörige Ortskurve ist ein Kreis mit dem Radius K um den Ursprung des Koordinatensystems, der mit wachsender Kreisfrequenz immer wieder durchlaufen wird und dessen Parametrierung daher mehrdeutig ist (Tab. 7-3). Der Phasenwinkel in Radian

$$\varphi = -\omega T_t \quad (7.66)$$

geht für $\omega \rightarrow \infty$ gegen $-\infty$.

Im Bode-Diagramm wird der Frequenzgang durch einen konstanten Betrag dargestellt. Somit handelt es sich beim Totzeit-Element um einen Allpass. Der Phasenwinkel nimmt mit der Kreisfrequenz linear ab. Wegen der logarithmischen Teilung der Kreisfrequenzachse erscheint der Phasengang jedoch nicht als Gerade, sondern als Exponentialfunktion und ist nach unten gekrümmt. Als charakteristischer Wert kann für $\omega = 1/T_t$ der Phasenwinkel $-1 = -180^\circ/\pi \approx -57^\circ$ im Bogenmaß bzw. Grad angegeben werden.

Die Übertragungsfunktion ist nicht rational und daher nicht durch endlich viele Pol- und Nullstellen darstellbar. Die entsprechenden Felder in Tab. 7-3 sind leer. Insofern kann auch ein Stabilitätstest über die Wurzeln des charakteristischen Polynoms für Totzeit-Glieder nicht durchgeführt werden. Es ist aber ersichtlich, dass die Übergangsfunktion konvergent und das Totzeit-Glied somit stabil ist.

Der kontinuierliche Abfall der Phase des Totzeit-Gliedes und das wiederholte Durchlaufen eines Kreises durch die Ortskurve sorgen dafür, dass e^{-sT_t} für $s \rightarrow \infty$ keinen Grenzwert besitzt. Dies macht die Anwendung der Grenzwertsätze für $t \rightarrow 0$ für Totzeitglieder unmöglich und eine regelwidrige Anwendung kann zu Fehlschlüssen führen.

Die Eigenschaften einer Reihenschaltung aus Verzögerungs- und Totzeitglied, des PT_1T_t -Gliedes, ergeben sich durch Multiplikation der Übertragungsfunktionen oder Frequenzgänge in einfacher Weise (Tab. 7-3).

8 Identifikation linearer Regelkreisglieder

8.1 Allgemeines

Um Systeme mit den bisher beschriebenen Werkzeugen analysieren zu können, benötigt man ein Modell in Form einer Differentialgleichung oder einer äquivalenten Darstellung.

Kapitel 2 erwähnte bereits zwei Wege zum Herleiten solcher Modelle: Zum einen kann aus physikalischen Zusammenhängen und Grundgleichungen ein theoretisches „white-box-Modell“ abgeleitet werden. Dieses Vorgehen wurde vorrangig in Kapitel 2 beim Aufstellen von Differentialgleichungen im Wirkungsplan verfolgt. Zum anderen kann ein Modell durch das Aufzeichnen von Messdaten im Experiment bestimmt werden, was als *Systemidentifikation* (oder kurz „Identifikation“) bezeichnet wird.

Systemidentifikation

Bei der *Systemidentifikation* wird aus der Messung von Ein- und Ausgangssignalen eines Systems und unter möglichem Einsatz von Vorkenntnissen ein mathematisches Modell des Signalübertragungsverhaltens des Systems bestimmt.

Eine theoretische Modellbildung ist nur dann möglich, wenn alle Gesetzmäßigkeiten des Prozesses hinreichend genau bekannt sind. Oft entziehen sich technische Prozesse – zumindest in Teilen – jedoch einer handhabbaren Beschreibung. Diese Teilsysteme stehen einer theoretischen Modellbildung dann nicht offen.

Selbst wenn eine rigorose Beschreibung des Systems über physikalische Gleichungen möglich ist, so werden in den allermeisten Fällen zumindest einzelne Parameter (wie z. B. eine Federsteifigkeit oder ein Reibbeiwert) eines von der Struktur her bekannten Modells nicht genau bestimmbar sein.

In diesen Fällen muss auf eine Systemidentifikation zurückgegriffen werden, die sich unter Umständen nur auf die einer theoretischen Modellbildung nicht zugänglichen Teilsysteme oder Parameter bezieht. Stützt sich die Modellbildung ausschließlich auf Messdaten, ohne Verwendung physikalischer Gesetze, so spricht man von einem „Black-Box“-Modell. Zwischenformen werden auch „Grey-Box“-Modelle genannt (vgl. Kapitel 2).

Das Thema der Systemidentifikation im regelungstechnischen Kontext füllt diverse Bücher, die sich erschöpfend mit den diversen Ansätzen und ihren Vor- und Nachteilen auseinandersetzen [20]. An dieser Stelle soll nur ein grober Abriss über das grundsätzliche Vorgehen gegeben werden und anschließend einige wenige exemplarische Verfahren skizziert werden.

Der grundsätzliche Ablauf einer Systemidentifikation ist in Bild 8-1 für den SISO-Fall gezeigt: Das zu identifizierende System wird mit Eingangssignalen $u(t)$ angeregt. Hieraus ergeben sich – überlagert durch Messrauschen $n(t)$ und Störungen $z(t)$ – die verfälschten gemessenen Systemausgänge $y_m(t)$. Die Aufgabe der Identifikation ist es, aus der Messung des Eingangssignals $u(t)$ und des gestörten Ausgangssignals $y_m(t)$ ein Modell des Prozesses derart zu ermitteln, dass ein zwischen echtem Systemausgang $y(t)$ und Modellausgang $\hat{y}(t)$ gebildetes Fehlersignal $e(t)$ möglichst klein wird.

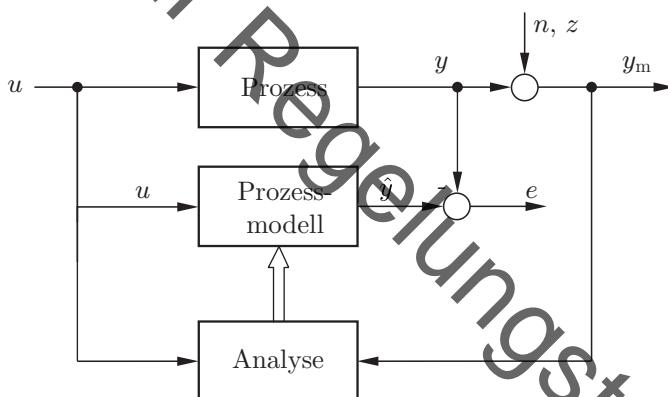


Bild 8-1: Wirkungsplan für den Aufbau einer Systemidentifikation

Rauschen $n(t)$ und Störungen $z(t)$ beeinflussen die Messung unweigerlich. Da das Modell den Einfluss der Eingänge auf die Ausgänge beschreiben soll, müssen die Signale in $y_m(t)$, die nicht $u(t)$ sondern $n(t)$ oder $z(t)$ zuzuordnen sind, durch entsprechende Maßnahmen von den anderen separierbar sein.

Daher sind neben einer Messwertaufbereitung (Filterung) auch die Wahl passender Testsignale $u(t)$, die das System hinreichend anregen, sowie wiederholte und lange Messungen Bestandteil einer erfolgreichen Identifikation.

Da diese Voraussetzungen im normalen Betrieb nicht gegeben sein werden, wird die Systemidentifikation meist im Rahmen einer eigenen Messkampagne mit speziellen Testsignalen wie Sprungfunktionen umgesetzt.

Auch wird das beschriebene Vorgehen nur im Falle stabiler Systeme von Erfolg gekrönt sein, weshalb instabile Systeme vor der Identifikation erst stabilisiert werden müssen. Eine weitere Annahme dieses Vorgehens ist, dass ein eindeutiger Zusammenhang zwischen Ein- und Ausgangssignalen besteht. Diese Voraussetzung ist insbesondere für die zentrale Klasse der LTI-Systeme gegeben, auf die sich alle weiteren Ausführungen beziehen.

Die Kenntnisse über das zu identifizierende System werden beim beschriebenen Vorgehen lediglich aus den Ein- und Ausgangssignalen und nicht etwa aus den Systemzuständen gewonnen. Daher können Anteile, die sich nicht im Ein-Ausgangs-Verhalten wiederfinden, hierüber nicht bestimmt werden. Somit muss auch bei der Systemidentifikation angenommen werden, dass das zu bestimmende Modell eine minimale Realisierung ist.

Abseits von dieser Vorgabe an das zu bestimmende Modell können weitere Vorgaben an die Modellklasse gemacht werden. So ist es z.B. möglich, die Modellform gänzlich frei zu belassen, oder aber – beispielsweise aus Vorwissen über die Struktur oder Ordnung des Systems – die Modellstruktur bereits vorzugeben und nur die Parameter des Modells durch die Identifikation bestimmen zu lassen. Der zweite Weg ist dann empfehlenswert, wenn beispielsweise aus einer theoretischen Modellbildung die Modellstruktur bekannt ist, nicht aber spezielle Werte der Koeffizienten der Differentialgleichung wie die Wärmeübergangskoeffizienten oder andere schwer bestimmbarer Parameter. Um beide Ansätze voneinander abzugrenzen, spricht man von *parametrischer* und *nicht-parametrischer* Systemidentifikation.

Parametrische und nicht-parametrische Identifikation

Werden bei der Systemidentifikation konkrete Vorgaben an die Struktur des zu identifizierenden Modells gemacht, sodass nicht mehr die Modellform, sondern nur die Koeffizienten einer vorgegebenen Modellform bestimmt werden müssen, so spricht man von einer *parametrischen* Systemidentifikation.

Werden hingegen keine Vorgaben hinsichtlich der Modellstruktur gemacht, wird das Verfahren als *nicht-parametrisch* bezeichnet.

Im Folgenden werden für beide Ansätze Vertreter vorgestellt, wobei nicht-parametrische Verfahren im Frequenzbereich und parametrische im Zeitbereich diskutiert werden.

8.2 Nicht-parametrische Identifikation

Nicht-parametrische Identifikationsverfahren werden bevorzugt im Frequenzbereich durchgeführt. Die Definition des Frequenzgangs über den eingeschwungenen Zustand ermöglicht es nämlich, für stabile Systeme eine Modellbeschreibung ohne Vorgabe einer Modellstruktur messtechnisch zu erfassen.

Es ist bereits bekannt, dass ein stabiles LTI-System auf ein sinusförmiges Eingangssignal nach Abklingen aller Einschwingvorgänge mit einer harmonischen Ausgangsschwingung mit identischer Frequenz antwortet:

$$u(t) = U \cos(\omega t) \Rightarrow y(t) = Y \cos(\omega t + \varphi). \quad (8.1)$$

Die Amplitude Y sowie die Phasenverschiebung φ können dem Frequenzgang des Systems über Betrag und Phasenwinkel direkt zugeordnet werden. In einigen Messaufbauten wird nicht der Phasenwinkel, sondern die Phasenverschiebung gemessen (siehe Bild 8-2).

$$\varphi(\omega) = \varphi_y(\omega) - \varphi_u(\omega) = -\frac{t_\varphi}{T} \cdot 360^\circ. \quad (8.2)$$

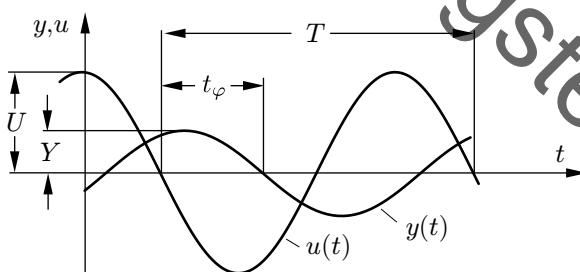


Bild 8-2: Harmonische Eingangs- und Ausgangsgröße

Dabei ist zu beachten, dass t_φ der zeitliche Abstand eines Nulldurchgangs der Eingangsgröße vom entsprechenden gleichsinnigen Nulldurchgang

der Ausgangsgröße ist. Bei den meisten stabilen Systemen wird ein negativer Phasengang vorliegen, sodass die Ausgangsgröße der Eingangsgröße nachfolgt und der Phasenwinkel einen negativen Wert erhält. Es gilt somit $t_\varphi > 0$ für $\varphi(\omega) < 0$ und umgekehrt.

Für die vollständige Beschreibung eines Übertragungssystems muss dessen Frequenzgang bei zahlreichen Frequenzwerten ermittelt werden. Die Ergebnisse derartiger Messungen können als Wertetabelle oder graphisch dargestellt werden. In der Praxis wird oft innerhalb einer einzigen Messung mit einem entsprechenden Eingangssignal eine Reihe von Frequenzen durchlaufen. Hierbei muss darauf geachtet werden, dass jede Frequenz genügend lang durchlaufen wird, um den eingeschwungenen Zustand zu erreichen. Zudem sollte die Amplitude der Anregung meist mit wachsender Frequenz sinken, um das physikalische System nicht übermäßig zu belasten.

Die Frequenzwerte können dabei an die regelungstechnischen Bedürfnisse und andere Randbedingungen angepasst werden. So empfiehlt es sich aus theoretischer Sicht, im Bereich kritischer Frequenzen wie Resonanz- oder Eigenkreisfrequenzen die Abstände der Anregungsfrequenzen zu verkleinern, um in diesen besonders relevanten Bereichen ein hinreichend gutes Modell zu erhalten. Der Fall von Resonanz erfordert jedoch eine praktische Abwägung zwischen Modellgüte und Systembelastung durch den Vorgang der Identifikation selbst.

Aufgrund der Überlagerung durch Störsignale $n(t)$ und $z(t)$ empfiehlt es sich, die gesuchten Größen anhand mehrerer Schwingungen mit ausreichend langer Messzeit zu bestimmen und die gewonnenen Werte zu mitteln. Solange die Störsignale zum Eingangssignal unkorreliert sind, werden sich so die entstandenen Einflüsse der Störsignale herausheben. Dieses Vorgehen kann auch bereits im Messaufbau durch Verwendung sogenannter Korrelationsverfahren vorgesehen werden [57].

Im Zeitbereich ist prinzipiell ein analoges Vorgehen zum Frequenzbereich möglich. Hierzu kann man durch eine sprungförmige Änderung der Eingangsgröße die Übergangsfunktion des Systems in graphischer Form oder als Wertetabelle messen. Selbiges ist auch für die Gewichtsfunktion möglich. Auf diese Weise erhält man ein graphisches Modell im Zeitbereich. Dieses kann man jedoch – im Gegensatz zu einem graphischen Modell im Frequenzbereich, siehe Abschnitt 11.1 – nur eingeschränkt für den Reglerentwurf nutzen. Im Zeitbereich arbeiten nämlich die meisten leistungsfähigen Ver-

fahren mit einem Modell in Form einer Differentialgleichung. Möchte man die Menge der anwendbaren Reglerentwurfsverfahren möglichst groß halten, sollte das Ergebnis einer Identifikation im Zeitbereich daher nicht graphisch sondern als Differentialgleichung vorliegen. Bei einer Identifikation im Zeitbereich werden daher vorrangig parametrische Verfahren verwendet.

8.3 Parametrische Identifikation

8.3.1 Überanpassung

Für eine parametrische Identifikation muss eine Modellstruktur vorgegeben werden. Diese Struktur kann aus Kenntnis der physikalischen Gesetzmäßigkeiten oder auch aus Messungen gewonnen werden. Um eine Struktur für ein Modell im Zeitbereich aus Messungen zu gewinnen, geht man so vor, dass man zunächst einige Sprungantworten aufzeichnet und aus dem Verlauf der Ausgangsgröße abschätzt, welche Modellformen geeignet sind. Von besonderem Interesse ist dabei, ob die Verläufe sich durch starke Verzögerungen, Totzeiten, Überschwingen, Schwingungsfähigkeit oder nicht-minimalphasiges Verhalten auszeichnen.

Entsprechend der beobachteten Eigenschaften wird eine passende Differentialgleichung z.B. als Reihenschaltung entsprechend passender Elemente gewählt, in welcher Parameter wie Kennkreisfrequenz, Dämpfung, Verstärkung und Zeitkonstanten unbekannte Parameter sind, die im Rahmen der Identifikation zu bestimmen sind.

Bei der Identifikation ist es wichtig, nicht zu viele unbekannte Parameter anzusetzen. Zum einen werden einige wenige Teilsysteme und damit Parameter den wesentlichsten Beitrag zum Ein-Ausgangsverhalten leisten. Hierdurch werden Parameter, die nicht zu den dominanten Pol- oder Nullstellen gehören, normalerweise nur sehr ungenau zu bestimmen sein. Zum anderen droht bei Auswahl zu vieler Parameter das Phänomen der *Überanpassung*.

Überanpassung

Unter *Überanpassung* (englisch: overfitting) versteht man, dass ein identifiziertes oder anderweitig gelerntes Modell die zur Erstellung des Modells verwendeten Daten zu genau wiedergibt, d.h. auch eigentlich nicht modellrelevante Informationen und Artefakte in diesem speziellen Daten-

satz (wie Rauschen) abbildet. Dies kann zur Folge haben, dass das Modell daran scheitert, zusätzliche Datenpunkte abseits der zum Bestimmen des Modells verwendeten adäquat wiederzugeben.

Die Auswahl einiger weniger charakteristischer Parameter ist folglich wesentlicher Bestandteil einer erfolgreichen Identifikation. Überanpassung tritt bei jeder Form der parametrischen Identifikation und sowohl im Zeit- wie auch Frequenzbereich auf. Die gesuchten Parameter können näherungsweise graphisch oder aber über numerische Verfahren ermittelt werden, wie in den folgenden Abschnitten erläutert wird.

8.3.2 Graphische Parameteridentifikation

Bei der graphischen Identifikation bestimmt man ausgehend von dem gewählten (einfachen) Modell des Prozesses die Parameter mithilfe der Tabellen Tab. 7-1 bis 7-3. Die statische Verstärkung liest man am statischen Endwert der Übergangsfunktion ab. Zeitkonstanten und Totzeiten können durch Tangenten und den Zeitverzug bestimmt werden. Kennkreisfrequenz und Dämpfung werden für $D < 1$ meist abweichend über Real- und Imaginärteil des komplex konjugierten Polstellenpaars ermittelt, da sich ω_0 im Zeitbereich nicht so gut ablesen lässt wie ω_D . Hierzu bestimmt man ω_D über die Nulldurchgänge der entstehenden Dauerschwingung und anschließend den Realteil über das Abklingverhalten der zugehörigen Hüllkurve.

An dieser Beschreibung sieht man bereits, dass das Bestimmen dieser Größen für schwingungsfähige Systeme eine gewisse Komplexität besitzt und die ermittelten Parameter recht ungenau sein werden. Einen einfachen Zugang bietet das Verfahren jedoch für Systeme vom Typ des PT_1T_t , wo eine schnelle Kennwertermittlung möglich ist. Glücklicherweise lässt sich mit dem PT_1T_t eine Vielzahl von typischen Verzögerungen höherer Ordnung, wie sie beispielsweise in vielen Antrieben vorliegen, mit hinreichender Genauigkeit beschreiben.

T_u-T_g -Modell

Wird ein Verzögerungselement höherer Ordnung ersatzweise über ein PT_1T_t -System beschrieben, so spricht man von einem T_u-T_g -Modell.[45]

Ein klassisches Vorgehen zum Ableiten dieses Ersatzmodell aus einer Sprun-

gantwort besteht darin, die Zeitkonstante T_g , welche Ausgleichszeit genannt wird und der Zeitkonstanten T des PT_1 entspricht, über die Steigung der Wendetangente der Sprungantwort zu bestimmen – siehe gestrichelt in Bild 8-3. Die Verzugszeit T_u , welche analog zur Totzeit zu deuten ist, ergibt sich dann als die Zeit, welche die angelegte Tangente an der Zeitachse abschneidet. Die statische Verstärkung K kann über den statischen Endwert abgelesen werden, wobei auf eine Normierung zu achten ist, wenn mit einer Sprunghöhe ungleich eins gearbeitet wurde.

Die abweichenden Benennungen als T_u und T_g anstelle der üblichen Bezeichner T und T_t haben sich eingebürgert, um den Charakter als Ersatzmodell zu unterstreichen. Das tatsächliche System kann also sogar totzeitfrei sein. Aus Bild 8-3 ist zu erkennen, dass das gepunktet eingezeichnete Ersatzmodell eine ungefähre Näherung des Originalsystems ist. Dabei wird vorausgesetzt, dass das zu identifizierende System auch ungefähr den Verlauf eines $PT_1 T_t$ aufweist.

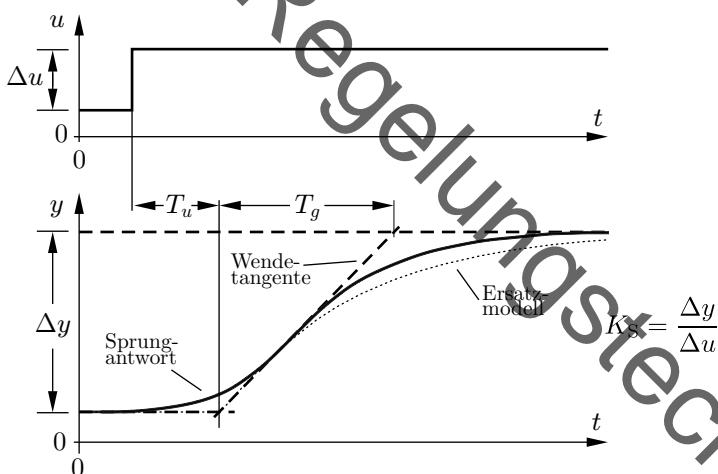


Bild 8-3: Sprungantwort und Kennwerte des T_u - T_g -Modells

Aus der Herleitung des identifizierten Modells ist klar, dass kein Qualitätsmaß für die Übereinstimmung von Ersatzmodell und Originalsystem angegeben werden kann. Das Näherungsverfahren über die Tangente liefert keine Garantien dafür, die beste mögliche Näherung zu sein und ist anfällig

lig für Ungenauigkeiten in der Konstruktion. Im Beispiel in Bild 8-3 liegt trotz exakter Ausführung das Ersatzmodell stets unterhalb der originalen Sprungantwort.

Diese Gründe haben dafür gesorgt, dass aufbauend auf den gleichen Grundideen die historischen Ansätze der graphischen Näherung durch Verfahren der Optimierung abgelöst wurden, bei welchen (mit Rechnerunterstützung) die graphische Näherung analytisch bestimmt wird.

8.3.3 Methode der kleinsten Fehlerquadrate

Zunächst wird angenommen, dass die Eingangsgröße und die gemessene Ausgangsgröße als Daten in ihrem gesamten Verlauf verfügbar sind. Um diese von üblichen Ein- und Ausgängen zu unterscheiden, werden diese gemessenen (und damit ggf. fehlerbehafteten Größen) mit $\tilde{u}(t)$ und $\tilde{y}(t)$ bezeichnet. Da diese Verläufe bekannt sind, können deren Zeitableitung $\dot{\tilde{u}}(t)$, $\ddot{\tilde{y}}(t)$ und höhere Ableitungen prinzipiell berechnet werden.

Als Modell wird die gewöhnliche Differentialgleichung

$$b_0 u(t) + b_1 \dot{u}(t) + \dots + b_m u^{(m)} = a_0 y(t) + a_1 \dot{y}(t) + \dots + y^{(n)} \quad (8.3)$$

angesetzt, wobei der führende Koeffizient a_n zu eins normiert wurde. Hierbei sind alle oder einige der Koeffizienten die im Rahmen der Identifikation zu bestimmenden Parameter des Systemmodells.

Passen die Messdaten für Ein- und Ausgangsgröße perfekt zu dem gewählten Modell (mit den optimalen Parametern), so wird beim Einsetzen dieser Verläufe in Gl.(8.3) auf beiden Seiten das gleiche Ergebnis stehen:

$$b_0 \tilde{u}(t) + b_1 \dot{\tilde{u}}(t) + \dots + b_m \tilde{u}^{(m)} \stackrel{!}{=} a_0 \tilde{y}(t) + a_1 \dot{\tilde{y}}(t) + \dots + \tilde{y}^{(n)} . \quad (8.4)$$

Da die Messdaten fehlerbehaftet sind, weil Messrauschen und Störungen ebenfalls auf die Ausgänge wirken, wird diese Gleichung nicht mit dem Gleichheitszeichen erfüllt sein. Es kann daher nicht erwartet werden, Parameter a_i und b_i zu finden, die Gl.(8.4) exakt erfüllen. Für eine Lösung von

Gl.(8.4) wechselt man zunächst auf eine vektorwertige Notation:

$$\underbrace{[\tilde{u}(t) \quad \dots \quad \tilde{u}^{(m)}(t) \quad -\tilde{y}(t) \quad \dots \quad -\tilde{y}^{(n)}(t)]}_{\mathbf{m}(t), \text{Messdaten}} \cdot \begin{bmatrix} b_0 \\ \vdots \\ b_m \\ a_0 \\ \vdots \\ a_{n-1} \end{bmatrix} \stackrel{!}{=} \tilde{y}^{(n)} \quad . \quad (8.5)$$

Hierbei enthält der Vektor $\mathbf{m}(t)$ Messdaten als Funktion der Zeit und der Vektor $\boldsymbol{\vartheta} \in \mathbb{R}^{m+n+1}$ die zu bestimmenden Parameter. Auch dieser Ausdruck wird nicht zu Null verschwinden – vor allem nicht für alle t .

Beschränkt man sich bei der Lösung von Gl.(8.5) auf lediglich p Zeitpunkte t_i statt auf das komplette Zeitkontinuum t , so wird aus den Zeitfunktionen der Messdaten die Messmatrix \mathbf{M} sowie der Vektor \mathbf{b} :

$$\mathbf{M} = \begin{bmatrix} \tilde{u}(t_1) & \dots & \tilde{u}^{(m)}(t_1) & -\tilde{y}(t_1) & \dots & -\tilde{y}^{(n-1)}(t_1) \\ \vdots & & \vdots & \vdots & & \vdots \\ \tilde{u}(t_p) & \dots & \tilde{u}^{(m)}(t_p) & -\tilde{y}(t_p) & \dots & -\tilde{y}^{(n-1)}(t_p) \end{bmatrix} \quad (8.6)$$

$$\mathbf{b} = \begin{bmatrix} \tilde{y}^{(n)}(t_1) \\ \vdots \\ \tilde{y}^{(n)}(t_p) \end{bmatrix} .$$

Hiermit erhält man aus Gl.(8.5) das lineare Gleichungssystem

$$\mathbf{M} \cdot \boldsymbol{\vartheta} = \mathbf{b} \quad . \quad (8.7)$$

Dieses besitzt für invertierbare Matrizen \mathbf{M} die eindeutige Lösung $\boldsymbol{\vartheta} = \mathbf{M}^{-1}\mathbf{b}$, die man zur Bestimmung der gesuchten Parameter verwendet könnte. Dies ist allerdings nicht zielführend, weil die Parameter die vollständigen Messdaten \tilde{y} inklusive Störungen und Messrauschen erklären müssten. Stattdessen müssen wesentlich mehr Zeitpunkte t_i als zu bestimmende Parameter verwendet werden: $p >> m + n + 1$. Hierdurch ist \mathbf{M} nicht quadratisch und somit nicht invertierbar, wodurch Gl.(8.7) nicht mit beiderseitiger

Gleichheit gelöst werden kann. Daher wird man ersatzweise fordern, dass der entstehende Fehler beim Lösen des überbestimmten Gleichungssystems klein sein wird. Eine gängige Wahl zur Bestimmung, was unter „klein“ verstanden werden soll, ist der Fehler der kleinsten Quadrate (least-square), da dieser eine analytisch geschlossene Lösung der Minimierung erlaubt.

Methode der kleinsten Fehlerquadrate (least-square)

Das Optimierungsproblem zum Finden des optimalen ϑ^* unter der Verwendung der klassischen euklidischen 2-Norm

$$\vartheta^* = \arg \min \| \mathbf{M} \cdot \vartheta - \mathbf{b} \|_2^2, \quad \|\mathbf{x}\|_2 = \sqrt{x_1^2 + \dots + x_n^2}, \quad x \in \mathbb{R}^n, \quad (8.8)$$

heißt *least-square*.

Da hier die einzelnen Spalten der Vektoren den Zeitpunkten t_i entsprechen, bedeutet dies, dass die Parameter ϑ gesucht werden, die den quadratischen Fehler im zeitlichen Mittel minimieren.

Das Optimierungsproblem in Gl.(8.8) besitzt eine geschlossene Lösung, die sich durch eine verallgemeinerte Inversenbildung formulieren lässt:

Pseudoinverse

Gegeben ist das überbestimmte lineare Ausgleichsproblem $\mathbf{M}\vartheta = \mathbf{b}$ mit der $p \times q$ -Matrix \mathbf{M} und $p > q$. Dann lautet ist die *Pseudoinverse* [8]

$$\mathbf{M}^\dagger = (\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T. \quad (8.9)$$

Man kann zeigen, dass die Matrix \mathbf{M}^\dagger im Falle invertierbarer Matrizen \mathbf{M} genau der Inversen entspricht. Die Pseudoinverse existiert dagegen, sobald \mathbf{M} vollen Rang hat und damit auch für nicht quadratische Matrizen.

Mit der Pseudoinversen kann man zeigen, dass die Lösung des Optimierungsproblems in Gl.(8.8) genau

$$\vartheta = \mathbf{M}^\dagger \cdot \mathbf{b} \quad (8.10)$$

ist. Hiermit lassen sich die Parameter in ϑ schnell bestimmen, sofern \mathbf{M}^\dagger existiert. Dabei muss die Inverse nicht explizit berechnet werden, da nur das Produkt $\mathbf{M}^\dagger \mathbf{b}$ gesucht ist, welches numerisch weitaus günstiger als die vollständige Inverse ermittelt werden kann.

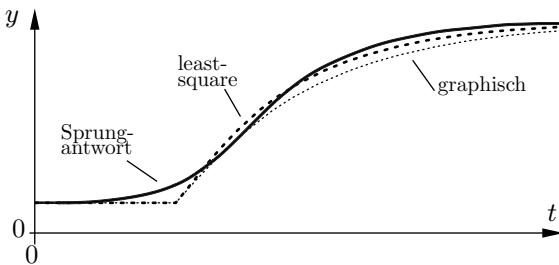
Um in \mathbf{M} einen vollen Rang sicherzustellen, muss die Messmatrix genügend linear unabhängige Spalten aufweisen. Regelungstechnisch kann diese mathematische Bedingung so interpretiert werden, dass die Messdaten eine hinreichende Anregung des zu identifizierenden Prozesses enthalten müssen. Ist dies nicht der Fall (z. B. wenn die Messdaten nur konstante Verläufe enthalten), so ist es nicht möglich, hieraus eindeutig auf die Parameter des Modells zu schließen. In allen anderen Fällen führt Gl.(8.10) auf die gesuchten Modellparameter, wobei eine schwache Anregung zu einer schlechten Konditionszahl von \mathbf{M} und somit zu numerischen Problemen führen kann.

Bei der Implementierung des oben beschriebenen Ansatzes muss noch beachtet werden, dass sowohl die Messmatrix als auch der Vektor \mathbf{b} Ableitungen enthalten. Die Ableitungsbildung ist aber als idealer Differenzierer acausal. Durch die konstante Steigung von +1 im Bode-Diagramm werden hochfrequente Anteile innerhalb der Signale (wie sie durch Messrauschen hervorgerufen werden) verstärkt. Um dies zu vermeiden, bietet es sich an, beide Seiten der Ausgangsgleichung Gl.(8.4) n -mal zu integrieren. Das hochfrequente Messrauschen wird so abgeschwächt und beeinflusst die Parameteridentifikation nur noch nachrangig. Allerdings müssen konstante Messfehler, die durch die Integration verstärkt werden, zuvor aus den Messdaten entfernt werden. Die weitere Herleitung der Parameteridentifikation erfolgt dann wortgleich.

Der vorgestellte Algorithmus wurde auf das in Bild 8-3 gezeigte Beispiel angewendet. Die Totzeit wurde dabei aus der graphischen Identifikation übernommen und nur der Parameter $\vartheta = T$ durch die numerische Optimierung bestimmt. Das Resultat in Bild 8-4 zeigt klar, dass ein kleineres T durch die Optimierung vorgeschlagen wird. Hierdurch kann der Effekt, dass das Ersatzmodell stets unterhalb der Messdaten lag, vermieden werden. Der Fehler zwischen Ersatzmodell und Messdaten kann durch die Vorgabe eines anderen parametrischen Modells mit mehr unbekannten Parametern noch weiter verringert werden. Für den Reglerentwurf reicht eine Näherung wie in Bild 8-4 aber in aller Regel aus.

Das beschriebene Identifikationsverfahren wird beim praktischen Einsatz noch um zwei Modifikationen ergänzt. Bisher nimmt das Verfahren nämlich beliebig große Parameter ϑ in Kauf, um die Messdaten zu erklären. Auch eine initiale Schätzung für ϑ kann nicht eingebracht werden.

Die einfachste Lösung dieses Sachverhaltes ist es, als erste Modifikation den

Bild 8-4: Identifikation eines T_u - T_g -Modells über least-square

quadratischen Fehler in Gl.(8.8) um den zusätzlichen Term

$$\|\mathbf{M} \cdot \boldsymbol{\vartheta} - \mathbf{b}\|_2^2 + \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2^2 \quad (8.11)$$

zu ergänzen. Mit der initialen Schätzung $\boldsymbol{\vartheta}_0$ wird so die Summe aus der Abweichung der Messdaten und der Abweichung von der Startschätzung bestraft. Die Lösung des Optimierungsproblems ändert sich dabei strukturell nicht, da die Summe der quadratischen Abweichungen vektorwertig als

$$\begin{bmatrix} \mathbf{M} \\ \mathbf{I} \end{bmatrix} \cdot \boldsymbol{\vartheta} - \begin{bmatrix} \mathbf{b} \\ \boldsymbol{\vartheta}_0 \end{bmatrix} \quad (8.12)$$

mit der Identitätsmatrix \mathbf{I} zusammengefasst werden kann. Somit kann auch dieses Optimierungsproblem mit der Pseudoinversen gelöst werden.

Würde man nun aber Gl.(8.11) direkt anwenden, würden Startschätzung und Messabweichung in der Optimierung genau gleich gewichtet. Das bedeutet, dass man nicht in die Identifikation miteinfließen lassen kann, wie sicher man sich mit seiner Startschätzung oder aber auch den aufgezeichneten Messwerten ist.

Daher wird als zweite Modifikation das least-square mit einer Gewichtung versehen, indem in der Optimierung die gewichtete 2-Norm

$$\|\mathbf{x}\|_{\mathbf{W}}^2 = \mathbf{x}^T \mathbf{W} \mathbf{x} \quad (\text{falls } \mathbf{W} \text{ diagonal: } = W_{11}x_1^2 + \dots + W_{nn}x_n^2) \quad (8.13)$$

mit der Gewichtungsmatrix \mathbf{W} verwendet wird. Für sinnvolle Gewichtungen wird die Gewichtungsmatrix dabei immer *positiv definit* sein.

Positive Definitheit

Eine quadratische reelle Matrix \mathbf{W} ist genau dann positiv definit, wenn

$$\mathbf{x}^T \mathbf{W} \mathbf{x} > 0 \quad \forall \mathbf{x} \neq \mathbf{0} \quad (8.14)$$

gilt. Für eine symmetrische Matrix bedeutet dies, dass alle (reellen) Eigenwerte $\lambda > 0$ strikt positiv sind. Eine Diagonalmatrix ist genau dann positiv definit, wenn alle Diagonaleinträge strikt positiv sind.

Auch für das gewichtete Ausgleichsproblems kann die Lösung über die (nun gewichtete) Pseudoinverse mit

$$\mathbf{M}_\mathbf{W}^\dagger = (\mathbf{M}^T \mathbf{W} \mathbf{M})^{-1} \mathbf{M}^T \mathbf{W} \quad (8.15)$$

bestimmt werden. Die Lösung des Identifikationsproblems hat sich also nur unwesentlich erschwert, die Flexibilität aber enorm gesteigert.

Neben der Abwägung zwischen Messdaten und Startschätzung ermöglicht es die Gewichtung \mathbf{W} auch zwischen unterschiedlichen Einträgen in \mathbf{M} zu unterscheiden. Hierbei reicht es in den allermeisten Fällen aus, \mathbf{W} als Diagonalmatrix anzusetzen.

Wahl der Gewichtungsmatrix

Wird die Gewichtungsmatrix \mathbf{W} als Diagonalmatrix gewählt, so bedeuten große Einträge W_{ii} , dass die Optimierung der i -ten Zeile in \mathbf{M} ein großes Gewicht beimisst und diese vorrangig erfüllen möchte. Niedrige Einträge W_{jj} bedeuten umgekehrt, dass die Optimierungen Abweichungen in der j -ten Zeile eher tolerieren wird.

Entscheidend ist dabei nur die relative Gewichtung der Diagonaleinträge zueinander. Ist beispielsweise bekannt, dass der Messung zum i ten Zeitpunkt weniger vertraut werden kann, so wird man den Wert W_{ii} absenken. Ist die Startschätzung für j -te Eintrag von $\boldsymbol{\vartheta}$ recht genau, so wird man W_{kk} mit $k = p + j$ entsprechend erhöhen.

Diese beschriebene Identifikation ist auch für eine große Klasse von nichtlinearen Systemen anwendbar, solange diese linear in den zu identifizierenden Parametern sind. Dies lässt sich oft durch die Definition von Ersatzparametern erreichen [39].

9 Stabilitätsprüfung

9.1 Problemstellung

Da eine technisch brauchbare Regelung unbedingt stabil sein muss, sind Stabilitätsuntersuchungen schon sehr lange fester Bestandteil der Regelungstechnik. In Kapitel 3 wurde gezeigt, dass in den allermeisten Fällen die Stabilität der Ruhelage eines nichtlinearen Systems auf die Stabilität eines LTI-Systems (nämlich der Linearisierung in dieser Ruhelage) zurückgeführt werden kann.

Ein LTI-System ist genau dann stabil, wenn eines der folgenden gleichbedeutenden Kriterien erfüllt ist:

Stabilitätskriterien

Ein LTI-System ist genau dann stabil, wenn die

- Wurzeln des charakteristischen Polynoms alle negativen Realteil aufweisen.
- Eigenwerte der Systemmatrix A alle negativen Realteil aufweisen.
- Polstellen der Übertragungsfunktion alle negativen Realteil aufweisen.
- Übergangsfunktion gegen einen endlichen Wert konvergiert.
- Gewichtsfunktion gegen null konvergiert und für LTI-Systeme somit absolut integrierbar ist.

Das zu untersuchende System ist dabei meist nicht die Regelstrecke, sondern der geschlossene Regelkreis. Dabei ist es unerheblich, welche Übertragungsfunktion am geschlossenen Regelkreis (Führungsübertragungsfunktion, Störübertragungsfunktion oder andere) untersucht wird, da diese alle denselben Nenner $1 + G_0(s)$ und somit identische Stabilitätseigenschaften besitzen (für Ausnahmen siehe Abschnitt 9.4). Dies kann auch so gedeutet werden, dass die Stabilität eines linearen Übertragungssystems nicht von der Eingangsgröße abhängt, und damit auch nicht davon, ob das Stör- oder das Führungsverhalten untersucht wird.

Der geschlossene Regelkreis hängt von den Einstellparametern des Reglers ab. Meistens soll die Stabilität nicht für fix vorgegebene Parameter des

Reglers untersucht werden, sondern es wird der Bereich gesucht, in dem sich die Parameter bewegen dürfen, damit der geschlossene Regelkreis stabil arbeitet.

Diese Aufgabe lässt sich kaum simulativ oder experimentell lösen, da der geschlossene Regelkreises für potentiell unendlich viele Parameterkombinationen getestet werden müsste. Stattdessen sucht man für den geschlossenen Regelkreis die Wurzeln des charakteristischen Polynoms in Abhängigkeit der Parameter.

Für Regelkreise, die durch Differentialgleichungen höherer Ordnung beschrieben werden, müssten also die Nullstellen eines Polynoms höheren Grades analytisch ermittelt werden. Hierfür fehlen für Ordnung drei oder größer geschlossene Lösungsformeln und Ansätze wie Polynomdivision müssen wegen der Parameterabhängigkeit der Koeffizienten verworfen werden. Auch für Systeme mit Ordnung zwei müssen oft Fallunterscheidungen bezüglich der Vorzeichen der auftretenden Wurzelterme getroffen werden, die die Berechnung aller stabilisierender Reglerparameter erschweren.

Tatsächlich wird aber für die Bestimmung der Stabilitätseigenschaften gar nicht die genaue Position der Wurzeln benötigt, sondern es reicht die Kenntnis des Vorzeichens der Wurzeln aus. Für dieses neue Problem, nicht die Wurzeln selbst, sondern das Vorzeichen des Realteils zu ermitteln, gibt es wesentlich einfachere Berechnungsverfahren. Die beiden prominentesten Ansätze – die algebraischen Stabilitätskriterien und das Nyquist-Kriterium – werden im folgenden vorgestellt.

9.2 Algebraische Stabilitätskriterien

9.2.1 Grundidee

Die sogenannten algebraischen Stabilitätskriterien umgehen nicht nur die explizite Lösung der Differentialgleichung, sondern auch die des charakteristischen Polynoms. Ihre Grundidee soll kurz skizziert werden.

Betrachtet wird ein charakteristisches Polynom in der Form

$$p(\lambda) = a_n \lambda^n + \dots + a_1 \lambda + a_0 \quad . \tag{9.1}$$

Zur Illustration der Grundidee wird angenommen, dass alle λ rein reell sind. Wären alle Koeffizienten streng positiv $a_i > 0$ und $\lambda \geq 0$ eine Nullstelle, so

würde gelten:

$$\underbrace{a_n}_{>0} \underbrace{\lambda^n}_{\geq 0} + \dots + \underbrace{a_1}_{>0} \underbrace{\lambda}_{\geq 0} + \underbrace{a_0}_{>0} > 0 \quad (9.2)$$

Also kann λ keine Nullstelle sein, da $p(\lambda) > 0 \neq 0$ gilt. Folglich handelt es sich um einen Widerspruch und alle Nullstellen müssen negativ sein. Diese Argumentation ist möglich, falls alle Koeffizienten negativ sind.

Die Bedingungen, dass alle a_i dasselbe Vorzeichen haben, ist wesentlich einfacher zu prüfen, als die Nullstellen eines Polynoms zu bestimmen. Der Stabilitätstest konnte somit weg von den konkreten Wurzeln hin zu Vorzeichen der Koeffizienten des charakteristischen Polynoms verlagert werden.

Die obige Beweisskizze befasst sich nur mit reellen Wurzeln. Für den allgemeinen komplexen Fall ergeben sich aufwändigere Bedingungen, die prüfen, ob alle Nullstellen eines Polynoms (hier des charakteristischen Polynoms der Differentialgleichung eines Übertragungssystems) einen negativen Realteil haben. Dazu werden bestimmte Schemata und die Grundrechenarten benutzt, insbesondere in den Kriterien nach Routh und Hurwitz.

9.2.2 Stabilitätskriterien nach Routh und Hurwitz

Bereits 1877 wurde von Routh¹ und 1895 von Hurwitz² jeweils ein Kriterium zur Prüfung der Nullstellen von Polynomen angegeben, die noch heute zur Stabilitätsprüfung benutzt werden. Obgleich es mittlerweile effizientere Verfahren wie das Parameterraumverfahren [1] gibt, ist der Bekanntheitsgrad der beiden traditionellen Kriterien wesentlich höher.

Sowohl Routh als auch Hurwitz prüfen die Koeffizienten a_i des charakteristischen Polynoms. Diese Koeffizienten findet man bekanntlich in der homogenen Differentialgleichung oder im Nenner der Übertragungsfunktion bzw. des Frequenzgangs (sofern diese als Quotienten zweier Polynome ohne Doppelbrüche dargestellt sind) oder in der Determinante von $\lambda I - A$ des zu prüfenden Übertragungssystems.

Beiden Kriterien gemeinsam ist die sogenannte 1. Bedingung, die für reelle Nullstellen zu der Ausführung in Abschnitt 9.2.1 identisch ist.

¹Edward Routh (1831-1907), englischer Mathematiker [48]

²Adolf Hurwitz (1859-1919), deutscher Mathematiker [19]

1. Bedingung

Ein System der Ordnung n mit dem charakteristischen Polynom $a_0 + a_1\lambda + \dots + a_n\lambda^n$ ist nur dann stabil, wenn alle Koeffizienten a_i vorhanden sind und dasselbe Vorzeichen besitzen.

Um die folgenden Bedingungen kompakter zu fassen, einigt man sich bei der ersten Bedingung darauf, dass alle Koeffizienten a_i vorhanden und echt positiv sein müssen. Falls alle Koeffizienten negativ sind, kann die Bedingung dann erfüllt werden, indem die Differentialgleichung mit -1 multipliziert wird.

Die formulierte erste Bedingung ist eine notwendige Bedingung. Ist sie verletzt, so ist das System in jedem Fall nicht stabil. Das bedeutet aber nicht, dass ein Erfüllen der ersten Bedingung die Stabilität des Systems sicherstellt. Diese Bedingung ist also im Allgemeinen nicht hinreichend.

Um ein hinreichendes Kriterium zu erhalten, muss eine nach Hurwitz bzw. Routh unterschiedlich formulierte, zweite Bedingung noch zusätzlich erfüllt sein:

2. Bedingung nach Hurwitz

Die Determinante $\det(H)$ der Hurwitz-Matrix H in Gl.(9.3) sowie die nach Gl.(9.3) gebildeten Unterterminanten sind sämtlich größer als null.

$$\left[\begin{array}{ccccccc} a_{n-1} & a_{n-3} & a_{n-5} & \cdots & \cdots & \cdots & 0 \\ a_n & a_{n-2} & a_{n-4} & \cdots & \cdots & \cdots & 0 \\ 0 & a_{n-1} & a_{n-3} & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & & \\ \vdots & \vdots & \vdots & & a_3 & a_1 & 0 \\ \vdots & \vdots & \vdots & & a_4 & a_2 & a_0 \\ 0 & 0 & 0 & & a_5 & a_3 & a_1 \end{array} \right] = H \quad (9.3)$$

Das Bildungsgesetz für die Hurwitz-Matrix Gl.(9.3) lässt sich so beschreiben, dass auf der Hauptdiagonalen die Koeffizienten a_{n-1}, \dots, a_1 in ihrer

natürlichen Reihenfolge stehen und die Spalten so aufgefüllt werden, dass die Koeffizienten mit von oben nach unten zunehmenden Indizes angeordnet sind. Fehlende Koeffizienten werden durch Nullen dargestellt. Die Definition der Unterdeterminanten ergibt sich aus dem Schema.

Die Kombination von erster und zweiter Bedingung nach Hurwitz ist ein notwendiges und hinreichendes Kriterium für Stabilität eines LTI-Systems. Die entstehenden Determinanten und Unterdeterminanten kann man analytisch in Abhängigkeit der unbekannten Parameter auswerten. Dabei wird man bei der Anwendung des Hurwitz-Kriteriums die Determinanten normalerweise nicht von Hand nach Gl.(9.3) berechnen, sondern auf Tabellenwerke zurückgreifen, wo die entsprechenden Ausdrücke bereits ausmultipliziert sind.

Bis zur Ordnung $n = 3$ sind die Ausdrücke der Hurwitz-Determinanten übersichtlich. Für höhere Ordnung werden sie hingegen recht langlich, sodass sich dort ein anderes Kriterium eher anbietet.

Für das Kriterium nach Routh lautet die zweite Bedingung:

2. Bedingung nach Routh

Die Routhschen Probefunktionen R_i sind sämtlich größer als null.

Die Probefunktionen werden durch das Rechenschema Gl.(9.5) (oder ähnliche Schemata) ermittelt. Dazu werden in zwei Zeilen die Koeffizienten der Differentialgleichung angeschrieben und dann aus jeweils zwei Zeilen eine dritte gebildet. Dieses Verfahren ist so lange fortzusetzen, bis man zwei Zeilen mit jeweils nur einem Element erhält. Die Elemente der ersten Spalte dieses Schemas sind die Routhschen Probefunktionen R_n, \dots, R_0 (Gl.(9.5)).

a_n	a_{n-2}	a_{n-4}
a_{n-1}	a_{n-3}	a_{n-5}
$a_{n-2} - \frac{a_n}{a_{n-1}} a_{n-3}$	$a_{n-4} - \frac{a_n}{a_{n-1}} a_{n-5}$	$a_{n-6} - \frac{a_n}{a_{n-1}} a_{n-7}$
$\underbrace{a'_{n-2}}$	$\underbrace{a'_{n-4}}$	$\underbrace{a'_{n-6}}$
$a_{n-3} - \frac{a_{n-1}}{a'_{n-2}} a'_{n-4}$	$a_{n-5} - \frac{a_{n-1}}{a'_{n-2}} a'_{n-6}$	

(9.4)

$$R_n = a_n , \quad R_{n-1} = a_{n-1} , \quad R_{n-2} = a'_{n-2} , \quad \dots \quad (9.5)$$

Ein Vorteil des Routh-Kriteriums gegenüber dem Hurwitz-Kriterium ist, dass es neben der Aussage über Stabilität auch die Zahl der Nullstellen des charakteristischen Polynoms mit positivem Realteil liefert: Diese entspricht nämlich der Zahl an Vorzeichenwechseln in der Folge der Probefunktionen.

Ein weiterer Vergleich beider Kriterien ergibt, dass das Kriterium nach Hurwitz zwar eleganter wirkt, wegen der zahlreichen Determinantenbestimmungen aber umständlicher zu handhaben ist als das nach Routh. Für die rechnerische Auswertung, insbesondere bei Systemen von höherer Ordnung, wird allgemein das Routh-Kriterium bevorzugt.

Bei der Anwendung beider Kriterien auf Systeme niedriger Ordnung erhält man Sonderfälle, die leicht überschaubar sind: Bei der Prüfung von Differentialgleichungen erster und zweiter Ordnung ist Stabilität bereits gesichert, wenn die erste der genannten Bedingungen erfüllt ist. Aus den Schemata für die zweite Bedingung erkennt man für die Differentialgleichung zweiter Ordnung, dass die Hurwitzdeterminante $\det(H) = a_1 a_2$ und die Routhschen Probefunktionen $R_n = a_2$, $R_{n-1} = a_1$, $R_{n-2} = a_0$ sind. Diese Ausdrücke sind positiv, wenn die erste Bedingung erfüllt ist.

Demnach sind Systeme, die durch Differentialgleichungen erster oder zweiter Ordnung beschrieben werden, für alle positiven Werte ihrer Koeffizienten stabil. Das entspricht genau den Überlegungen von Abschnitt 7.3, wo gezeigt wurde, dass PT₂-Glieder genau dann stabil sind, wenn für positive ω_0 die Dämpfung D ebenfalls positiv ist. Wegen des Nennerpolynoms $s^2 + 2D\omega_0 s + \omega_0^2$ sind in diesem Fall alle Koeffizienten vorhanden und positiv.

Für die Differentialgleichung dritter Ordnung ist der einzige Ausdruck, der im Rahmen der zweiten Bedingung negativ ausfallen kann, obgleich die erste Bedingung erfüllt ist, die Hurwitzdeterminante

$$\det(H) = \begin{vmatrix} a_2 & a_0 \\ a_3 & a_1 \end{vmatrix} = a_1 a_2 - a_0 a_3 \quad (9.6)$$

oder die Probefunktion

$$R_1 = a_1 - \frac{a_3}{a_2} \cdot a_0 . \quad (9.7)$$

Beide lassen sich in die Bedingung für Stabilität

$$a_1 a_2 > a_0 a_3 \quad (9.8)$$

umformen, die man sich dadurch einprägen kann, dass das Produkt der inneren Koeffizienten größer sein muss als das der äußeren Koeffizienten der Differentialgleichung.

Stabilitätskriterien für Systeme bis Ordnung $n = 3$

Ein System der Ordnung $n = 1$ oder $n = 2$ ist genau dann stabil, wenn alle Koeffizienten a_i vorhanden und positiv sind.

Ein System der Ordnung $n = 3$ ist genau dann stabil, wenn a_0, a_1, a_2 und a_3 vorhanden und positiv sind sowie $a_1 a_2 > a_0 a_3$ gilt.

Auch hier kann im Fall, dass alle Koeffizienten negativ sind, die Differentialgleichung natürlich mit -1 multipliziert werden.

Bei der Anwendung der Kriterien ist zu beachten, dass Koeffizienten a_i , die null sind, als nicht vorhanden gelten und daher die erste Bedingung nicht erfüllt ist. Damit erledigt sich auch die Frage, ob Null eine positive oder negative Zahl ist. Allerdings lassen sich diese Fälle verschwindender Koeffizienten mit dem folgenden mathematischen Satz noch ausführlicher deuten:

Stetigkeit der Eigenwerte und Nullstellen

Die Eigenwerte einer Matrix sind stetig von den Einträgen der Matrix abhängig.

Die Nullstellen eines Polynoms $s^n + a_{n-1}s^{n-1} + \dots + a_0$ sind stetig von den Koeffizienten des Polynoms a_i abhängig.

Das bedeutet, dass Eigenwerte nicht von der linken in die rechte komplexe Halbebene springen, wenn man die Einträge der Matrix nur wenig ändert. Eigenwerte können nur dadurch von der linken in die rechte komplexe Halbebene gelangen, indem sie die imaginäre Achse kreuzen.

Für die Stabilität bedeutet dies, dass Systeme nicht plötzlich instabil werden, wenn sich die Koeffizienten der Systemmatrix nur leicht ändern. Stattdessen werden diese Systeme Eigenwerte mit stetig wachsendem Realteil aufweisen. Folglich werden die homogenen Lösungen zunächst immer langsamer gegen Null abfallen, um dann (bei Eigenwerten am Stabilitätsrand) Dauerschwingungen durchzuführen oder polynomiell zu wachsen und schließlich exponentiell ins Unendliche zu laufen.

Was bedeutet das für das Hurwitz-Kriterium? Die zu prüfenden Unterterminanten sind stetig von den Koeffizienten abhängig, da diese nur miteinander multipliziert und addiert werden. Folglich sind alle zu prüfenden Bedingungen wie beispielsweise $a_i > 0$ und $a_1 a_2 > a_0 a_3$ stetig von den Koeffizienten abhängig. Sind alle Bedingungen erfüllt, ist das System stabil. Verschwindet eine der Bedingungen zu null, ist dies der Grenzfall, wo das System gerade nicht mehr stabil ist. Folglich muss wegen der Stetigkeit der Eigenwerte das System am Stabilitätsrand sein.

Hurwitz-Kriterium und Stabilitätsrand

Sind alle Koeffizienten und alle Unterterminanten entweder positiv oder identisch null, so befindet sich das System am Stabilitätsrand.

Für das Routh-Kriterium gibt es vergleichbare Aussagen, die aber wegen der auftretenden Brüche, welche für verschwindende Nenner unstetig sind, und den daraus resultierenden Fallunterscheidungen recht unübersichtlich sind. Für Sonderfälle sei auf Abschnitt 9.4 verwiesen.

9.2.3 Beispiele

Die Stabilitätskriterien nach Routh und Hurwitz sollen zur Illustration auf einige Beispiele angewandt werden, die im folgenden Abschnitt wieder aufgegriffen werden. Betrachtet wird ein Standardregelkreis wie in Bild 6-2, der für gegebene Übertragungsfunktionen für Regelstrecke G_S und Regler G_R auf Stabilität untersucht werden soll. Dabei ist es unerheblich, welche der möglichen Übertragungsfunktionen des geschlossenen Kreises zur Prüfung herangezogen wird, da alle denselben Nenner besitzen. Dieser Nenner berechnet sich gemäß Gl.(6.6) oder Gl.(6.5) zu

$$p(s) = Z_0(s) + N_0(s) = Z_S(s)Z_R(s) + N_S(s)N_R(s) \quad , \quad (9.9)$$

wofür man die Regel „Zählerprodukt plus Nennerprodukt“ als Gedankenstütze nutzen kann.

Hervorzuheben ist, dass über Gl.(9.9) die Übertragungsfunktion des geschlossenen Regelkreises gar nicht vollständig aufgestellt werden muss. Das liegt daran, dass es ausreicht, für Stabilität den Nenner zu betrachten.

Als erstes Beispiel wird ein PT₁ mit einem P-Regler betrachtet:

$$G_S(s) = \frac{K_S}{Ts + 1} , \quad G_R(s) = K_R , \quad K_S, K_R, T > 0. \quad (9.10)$$

Das charakteristische Polynom des geschlossenen Regelkreises ist

$$p(s) = (Ts + 1) + K_S K_R = \underbrace{Ts}_{a_1} + \underbrace{K_S K_R + 1}_{a_0}. \quad (9.11)$$

Zunächst wird die Stabilität über die Bestimmung der Eigenwerte untersucht. Es berechnet sich

$$a_1 \dot{x} + a_0 x = 0 \Rightarrow \lambda = -\frac{a_0}{a_1} = -\frac{K_S K_R + 1}{T}. \quad (9.12)$$

Dieser Eigenwert ist negativ, sofern $K_R > -1/K_S$ gilt. Beispielsweise werden alle positiven Reglerverstärkungen K_R auf einen stabilen geschlossenen Regelkreis führen. Auch das Weglassen des Reglers (d. h. $K_R = 0$) führt auf einen stabilen geschlossenen Regelkreis, da die Regelstrecke selbst bereits stabil war.

Das identische Ergebnis erhält man über das Hurwitz oder Routh-Kriterium: Die erste Bedingung für Stabilität nach Routh und Hurwitz ist erfüllt, wenn alle Koeffizienten dasselbe Vorzeichen besitzen. Da $a_1 = T > 0$ ist, muss auch $a_0 = K_S K_R + 1 > 0$ sein und man erhält den identischen Stabilitätsbereich. Eine zweite Bedingung muss nicht geprüft werden muss, weil das System von erster Ordnung ist.

Wenn man in Wirkungsplan 6-2 das negative Vorzeichen an dem Summenpunkt, durch den y von w subtrahiert wird, durch ein positives ersetzt, wird die Rückführung (oder Rückkopplung) zu einer Mitkopplung. Im charakteristischen Polynom ist das äquivalent zu einem Vorzeichenwechsel $G_R = -K_R$ innerhalb des Reglers. Beim Auflösen der neuen Ungleichung $-K_S K_R + 1 > 0$ nach K_R ist zu beachten, dass bei einer Multiplikation mit einer negativen Zahl das Ungleichheitszeichen gedreht werden muss. Dadurch erhält man $K_R < 1/K_S$ als Stabilitätsbereich. Für größere K_R (also eine starke Mitkopplung) wird der Regelkreis also instabil.

Bei diesem ersten Beispiel brachte die Anwendung von Hurwitz oder Routh keinerlei Vorteil gegenüber einer direkten Berechnung der Eigenwerte. Stattdessen diente es eher der Plausibilisierung, dass beide Ansätze die gleichen Resultate liefern.

Für das zweite Beispiel mit einer PT_3 -Regelstrecke und einem P-Regler

$$G_S(s) = \frac{K_S}{(Ts + 1)^3} , \quad G_R(s) = K_R , \quad K_S, K_R, T > 0. \quad (9.13)$$

können die neuen Verfahren ihre Stärken aber bereits ausspielen. Das charakteristische Polynom ist mit Gl.(9.9)

$$p(s) = (Ts + 1)^3 + K_S K_R = \underbrace{T^3 s^3}_{a_3} + \underbrace{3T^2 s^2}_{a_2} + \underbrace{3T s + 1}_{a_1} + \underbrace{K_S K_R}_{a_0}. \quad (9.14)$$

Eine Berechnung der Polstellen als analytischer Ausdruck in Abhängigkeit von T , K_S und K_R gestaltet sich schwierig. Das Hurwitz-Kriterium liefert hingegen direkt, dass alle Koeffizienten vorhanden und positiv sind, womit die erste Bedingung für Stabilität erfüllt ist. Die zweite Bedingung in der für Systeme dritter Ordnung geltenden verkürzten Form ergibt

$$3T \cdot 3T^2 > (1 + K_S K_R) \cdot T^3 \quad (9.15)$$

und daraus gewinnt man als Bedingung für Stabilität $8 > K_S K_R$. Die Störübergangsfunktion des Regelkreises für $K_S K_R = 8$ ist in Bild 9-1 wiedergegeben. Man erkennt, dass der Regelkreis aufgrund der ausgeführten Dauerschwingungen grenzstabil ist und sich daher am Stabilitätsrand befindet.

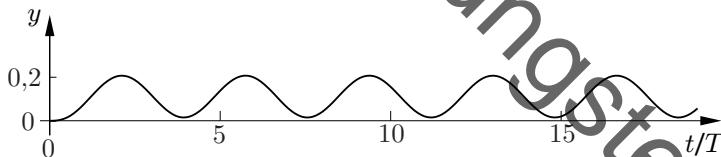


Bild 9-1: Störübergangsfunktion für $K_S K_R = 8$ am Stabilitätsrand.

Als letztes Beispiel wird die dimensionslose Regelstrecke G_S mit

$$G_S(s) = \frac{1}{2s^2 - s - 1} \quad (9.16)$$

untersucht.

Das Hurwitz- und das Routh-Kriterium lassen sich selbstverständlich auch für die Prüfung der Stabilität von G_S einsetzen. Dort sieht man, dass die

Koeffizienten 2, -1 und -1 unterschiedliche Vorzeichen haben. Die Regelstrecke ist folglich instabil.

Um die Strecke zu stabilisieren, soll ein Regler eingesetzt werden. Damit dies gelingen kann, muss der Regler sowohl einen Einfluss auf a_1 als auch a_0 haben, da nur auf diese Weise die erste Bedingung erfüllbar ist. Daher kommt ein P-Regler, der im geschlossenen Regelkreis nur a_0 verändern kann, nicht in Frage. Die Stabilitätskriterien können also auch zur Abschätzung der passenden Reglerstruktur genutzt werden, wofür oftmals ein schneller Blick ausreicht.

Mit dem PD-Regler $G_R(s) = K_R(s + 1)$ erhält man im geschlossenen Regelkreis

$$p(s) = \underbrace{2}_{a_2} s^2 + \underbrace{(K_R - 1)}_{a_1} s + \underbrace{K_R - 1}_{a_0} . \quad (9.17)$$

Es muss also $K_R > 1$ für Stabilität gelten. Für $K_R = 1$ ist der geschlossene Regelkreis am Stabilitätsrand. Setzt man $K_R = 1$ ein und berechnet $T(s)$ nach Gl.(6.6), so erhält man

$$T(s) = \frac{G_0}{1 + G_0} = \frac{s + 1}{2s^2}, \quad (9.18)$$

und damit den Nenner eines Doppelintegrators. In Gl.(3.80) wurde bereits berechnet, dass die Lösung von der Form $y(t) = C_2 t + C_1$ für $t \rightarrow \infty$ über alle Grenzen wächst. Das zeigt, dass Systeme am Stabilitätsrand (d. h. mit Polen auf der imaginären Achse) sowohl grenzstabil als auch instabil sein können.

In vielen Fällen ergeben sich die Stabilitätsbereiche nicht als unendlich große Intervalle $K_R > \dots$, sondern nur eine beschränkte Parametermenge $\dots > K_R > \dots$ führt auf einen stabilen geschlossenen Regelkreis.

9.3 Nyquist-Kriterium

9.3.1 Vollständiges Nyquist-Kriterium

Die algebraischen Stabilitätskriterien besitzen gewisse Nachteile, die ihre Anwendung in einigen Fällen erschweren. Sie gehen nämlich davon aus, dass

ein charakteristisches Polynom als analytischer Ausdruck vorliegt, über welchen die Stabilität untersucht und nachgewiesen werden kann. Vielfach sind solche Ausdrücke aber nicht vorhanden, entweder weil für ein oder mehrere Glieder des Regelkreises nur Messergebnisse zur Beschreibung ihres dynamischen Verhaltens vorliegen, oder weil zur Vereinfachung eine graphische Darstellung, z. B. im Bode-Diagramm, benutzt wird.

Auch in Fällen, wo eine analytische Beschreibung vorliegt, muss es sich hierbei nicht um ein charakteristisches Polynom handeln, bspw. wenn Totzeiten im Regelkreis vorhanden sind. So erhält man für einen totzeitbehafteten aufgeschnittenen Regelkreis (PT_1T_t)

$$G_0 = \frac{K}{Ts + 1} e^{-sT_t} \Leftrightarrow S(s) = \frac{1}{1 + G_0(s)} = \frac{Ts + 1}{Ts + 1 + Ke^{-sT_t}} \quad (9.19)$$

einen Nenner, der kein Polynom in s ist, sondern nicht-rationale Anteile in Form der e -Funktion enthält. Der aufgeschnittene Regelkreis lässt sich zwar mit algebraischen Stabilitätskriterien auf Stabilität untersuchen: Hierzu nutzt man die Eigenschaft, dass die Reihenschaltung stabiler Elemente stabil ist, weshalb es ausreicht, den totzeitfreien Anteil in G_0 zu betrachten. Für den geschlossenen Regelkreis ist dieser Trick jedoch nicht möglich.

Ein weiterer Nachteil der algebraischen Stabilitätskriterien ist, dass diese abseits der reinen Information, ob ein System stabil ist oder nicht, keine Anhaltspunkte liefern, wie dynamische Eigenschaften wie Dämpfung oder Zeitkonstanten im Regelkreis ausgeprägt sind. Daher eignen sie sich weniger zur konkreten Festlegung von Reglerparametern, sondern nur zu deren Eingrenzung auf ein Gebiet möglicher Kandidaten.

Das im Folgenden zu behandelnde Kriterium ist nach Harry Nyquist³ benannt, auf den auch die englische Bezeichnung für „Ortskurve“ (Nyquist plot) zurückgeht. Dieses Kriterium benutzt den Frequenzgang des aufgeschnittenen Regelkreises in seiner graphischen Darstellung als Ortskurve oder im Bode-Diagramm. Weil dieser Frequenzgang oft einfach anzugeben ist und auch durch Messungen (siehe 8) ohne analytische Modellbildung verfügbar ist, wird das Nyquist-Kriterium oft benutzt. Zudem besitzt auch das Totzeit-Glied eine geschlossene Darstellung im Frequenzbereich und in der Ortskurve. Schließlich liefert der Verlauf des Frequenzganges des aufgeschnittenen Regelkreises zusätzlich zur Aussage über die Stabi-

³Harry Nyquist (1889-1976), schwedisch-amerikanischer Ingenieur [41]

lität des geschlossenen Regelkreises auch noch Aussagen über die Stabilitätsgüte und Hinweise zu deren Verbesserung. Hierdurch eignet sich das Nyquist-Kriterium auch für eine Reglerentwurfsmethodik, die in Abschnitt 11.1 ausführlich diskutiert wird.

Für das Nyquist-Kriterium gibt es verschiedene Möglichkeiten der Herleitung, wovon die hier präsentierte ohne zusätzliche mathematische Kenntnisse direkt auf die von Pol- und Nullstellen hervorgerufene Phasenänderung zurückgreift. Dabei wird zunächst der Fall von Pol- oder Nullstellen auf der imaginären Achse ausgeklammert.

Aus Tab. 6-2 ist dabei bekannt, dass für positive Frequenzen $0 < \omega < \infty$ folgendes gilt:

- stabile Polstellen ($\operatorname{Re}(\lambda) < 0$) senken die Phase um $\Delta\varphi = -90^\circ$.
- instabile Polstellen ($\operatorname{Re}(\lambda) > 0$) heben die Phase um $\Delta\varphi = +90^\circ$.
- „stabile“ Nullstellen ($\operatorname{Re}(\eta) < 0$) heben die Phase um $\Delta\varphi = +90^\circ$.
- „instabile“ Nullstellen ($\operatorname{Re}(\eta) > 0$) senken die Phase um $\Delta\varphi = -90^\circ$.

Die Bezeichnungen der Nullstellen als „stabil“ wird dabei in Anführungszeichen gesetzt, da diese keinen Einfluss auf die Stabilität haben und das Adjektiv lediglich auf die identische Bedingung an den Realteil verweisen soll.

Betrachtet man die Ortskurve $G(j\omega)$ oder den Phasengang eines auf Stabilität zu untersuchenden Systems, kann man folglich aus der Gesamtänderung der Phase ablesen, welche Kombinationen von stabilen und instabilen Pol- bzw. Nullstellen möglich sind. Bezeichnet man

- die Anzahl der stabilen Polstellen mit p_-
- die Anzahl der instabilen Polstellen mit p_+
- die Anzahl der „stablen“ Nullstellen mit n_-
- die Anzahl der „instabilen“ Nullstellen mit n_+

so gilt für die Gesamtphasenänderung $\Delta\varphi$ in $0 < \omega < \infty$

$$\Delta\varphi(G) = 90^\circ [p_+(G) - p_i(G) + n_-(G) - n_+(G)] \quad . \quad (9.20)$$

Als Beispiel zeigt Bild 9-2 den Phasengang eines kausalen Systems zweiter Ordnung. Da jede Pol- oder Nullstelle einen Phasenabfall von $\pm 90^\circ$

erzeugt, führt eine gerade Anzahl von Pol- und Nullstellen zu einem Phasenwinkel für $\omega \rightarrow \infty$, der ein ganzzahliges Vielfaches von 180° ist. Das ist hier nicht der Fall, weshalb das System neben den zwei Polstellen (da es zweiter Ordnung ist) genau eine Nullstelle besitzt. Da die Phase um insgesamt $\Delta\varphi = -90^\circ$ abfällt, gibt es hierfür zwei mögliche Kombinationen.

- Das System hat zwei stabile Polstellen (-180°) und eine „stabile“ Nullstelle ($+90^\circ$).
- Das System hat eine stabile Polstelle (-90°), eine instabile Polstelle ($+90^\circ$) und eine „instabile“ Nullstelle (-90°).

Andere Kombinationen sind nicht möglich.

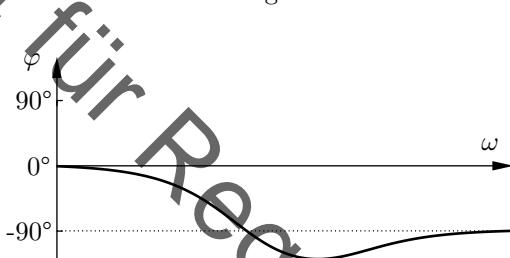


Bild 9-2: Phasengang eines Systems zweiter Ordnung

Mit wenigen Zusatzinformationen kann man aus der Menge der Kombinationen, die laut des Phasengangs möglich sind, die tatsächlich vorliegende Kombination bestimmen. Das kann über die Betrachtung der Betragssänderungen geschehen oder aber dadurch, dass einige Positionen der Null- oder Polstellen bereits bekannt sind. Dieser zweite Fall liegt vor, wenn ein geschlossener Regelkreis auf Stabilität geprüft werden soll und die Stabilität des aufgeschnittenen Regelkreises bekannt ist.

Die Ortskurve kann deshalb helfen, für eine Regelstrecke mit bekannten Stabilitätseigenschaften die passenden Parameter eines Reglers zu finden. Diese Frage stellt sich sehr häufig in regelungstechnischen Anwendungen. Hierzu wird die Störübertragungsfunktion $S(s) = 1/(1 + G_0(s))$ betrachtet. Für diese gilt:

- Die Nullstellen von $S(s)$ entsprechen den Polstellen von $1 + G_0(s)$.
- Die Polstellen von $S(s)$ entsprechen den Nullstellen von $1 + G_0(s)$.

Um die Pol- und Nullstellen von $S(s)$ zu bestimmen, reicht es also aus, $1 + G_0$ zu betrachten. Die Polstellen von $1 + G_0$ entsprechen dabei genau den Polstellen von G_0 . Für eine Polstelle λ gilt nämlich

$$1 + G_0(\lambda) \rightarrow \infty \Leftrightarrow G_0(\lambda) \rightarrow \infty . \quad (9.21)$$

Sind die Stabilitätseigenschaften von G_0 bekannt, so ist folglich auch bekannt, ob die Nullstellen von $S(s)$ positiven oder negativen Realteil haben.

Somit gilt:

$$\begin{aligned} n_+(S) &= p_+(1 + G_0) = p_+(G_0) \Rightarrow \text{bekannt} \\ n_-(S) &= p_-(1 + G_0) = p_-(G_0) \Rightarrow \text{bekannt} \\ p_+(S) &= n_+(1 + G_0) \Rightarrow \text{unbekannt} \\ p_-(S) &= n_-(1 + G_0) \Rightarrow \text{unbekannt} \end{aligned} \quad (9.22)$$

und man kann Gl.(9.20) umstellen nach

$$\begin{aligned} \frac{1}{90^\circ} \cdot \underbrace{\Delta\varphi(1 + G_0)}_{\text{Phasenänderung von } 1 + G_0} \\ &= (p_+(1 + G_0) - p_-(1 + G_0)) + (n_-(1 + G_0) - n_+(1 + G_0)) \quad (9.23) \\ &= \underbrace{(p_+(G_0) - p_-(G_0))}_{\text{Pole von } G_0; \text{ bekannt}} + \underbrace{(p_-(S) - p_+(S))}_{\text{Pole von } S; \text{ unbekannt}} . \end{aligned}$$

Kennt man durch bspw. eine Skizze von $1 + G_0$ die Phasenänderung von $1 + G_0$, so kann man mit Gl.(9.23) die Differenz $p_-(S) - p_+(S)$ der stabilen und instabilen Pole des geschlossenen Regelkreises ausrechnen. Gesucht ist aber nicht diese Differenz, sondern die Anzahl der instabilen Polstellen $p_+(S) = n_+(1 + G_0)$, die möglichst null sein soll. Dies ist hier ein notwendiges und hinreichendes Stabilitätskriterium, da der Fall von Polen auf der imaginären Achse zunächst ausgeschlossen wurde.

Die Zahl $n_+(1 + G_0)$ lässt sich mit einer geringfügigen Zusatzannahme finden. Hierzu wird nochmals $S = 1/(1 + G_0) = N_0/(Z_0 + N_0)$ betrachtet. Der Nenner des geschlossenen Regelkreises ist also $Z_0 + N_0$. Wie viele Polstellen hat der geschlossene Regelkreis insgesamt? Ist G_0 kausal (wovon ausgegangen werden kann), so ist die Ordnung von Z_0 (Anzahl der Nullstellen) nicht größer als die Ordnung von N_0 (Anzahl der Polstellen).

Anzahl der Polstellen des geschlossenen Regelkreises

Ist der aufgeschnittene Regelkreis kausal, so besitzt der geschlossene Regelkreis genauso viele Polstellen wie der aufgeschnittene Regelkreis.

Auf Sonderfälle für sprungfähige Systeme wird in Abschnitt 9.4 eingegangen.

Mithilfe dieser Erkenntnis ist klar, dass für die Gesamtanzahl der Polstellen von S gelten muss:

$$\underbrace{n_+(S) + n_-(S)}_{\text{Anzahl Nullstellen } S} = \underbrace{p_+(S) + p_-(S)}_{\text{Anzahl Pole } S} = \underbrace{n_+(1 + G_0) + n_-(1 + G_0)}_{\text{Anzahl Nullstellen } 1+G_0}. \quad (9.24)$$

Setzt man Gl.(9.24) in Gl.(9.23) ein, so gewinnt man

$$\begin{aligned} \frac{\Delta\varphi(1 + G_0)}{90^\circ} &= 2p_+(G_0) - 2n_+(1 + G_0) \\ \Rightarrow \frac{\Delta\varphi(1 + G_0)}{180^\circ} &= p_+(G_0) - n_+(1 + G_0) = p_+(G_0) - p_+(S) \end{aligned} \quad . \quad (9.25)$$

Mit Gl.(9.25) hat man eine erste vollständig anwendbare Formulierung des Nyquist-Kriteriums, um die Stabilität des geschlossenen Regelkreises zu prüfen. Hierzu geht man wie folgt vor:

- 1) Messung (Identifikation) oder Bestimmung (Modellbildung) des Frequenzgangs der Regelstrecke $G_S(j\omega)$
- 2) Festlegen eines Reglers $G_R(j\omega)$
- 3) Bestimmung der Anzahl der instabilen Polstellen $p_+(G_0)$ des aufgeschnittenen Regelkreises $G_0 = G_S G_R$
- 4) Zeichnen der Ortskurve $1 + G_0(j\omega)$ und Bestimmung der Gesamtphasenänderung
- 5) Ausrechnen der Anzahl der instabilen Postellen $p_+(S) = n_+(1 + G_0)$ des geschlossenen Regelkreises mit Gl.(9.25) (Pole am Stabilitätsrand zunächst ausgeschlossen)

Für Stabilität des geschlossenen Kreises muss $p_+(S) = n_+(1 + G_0) = 0$ gelten. In Schritt 3) wird man die Stabilität von G_R kennen und meist einen stabilen Regler oder Regler mit integrierendem Verhalten wählen. In

vielen Fällen wird auch die Regelstrecke stabil sein und somit $p_+(G_0) = 0$ gelten. In anderen Fällen ergibt sich die Übertragungsfunktion G_0 meist aus einer Reihenschaltung einfacher Glieder, deren Pole und Nullstellen sich entsprechend den Ausführungen in Abschnitt 6.2 überlagern. Hierdurch kann man die Zahl $p_+(G_0)$ der Polstellen in der rechten offenen s -Halbebene ohne großen Aufwand ermitteln.

In Schritt 4) reicht, da nur die Gesamtphasenänderung von Interesse ist, in vielen Fällen eine Skizze der Ortskurve aus. Nachteilig gestaltet sich dabei, dass nicht die Ortskurve $G_0(j\omega)$, sondern $1 + G_0(j\omega)$ benötigt wird. Dies erschwert insbesondere die Anwendung im Bode-Diagramm erheblich, da dort Additionen graphisch nur schwer durchgeführt werden können. Hier behilft man sich wie folgt: Die Gesamtphasenänderung $\Delta\varphi(1 + G_0)/180^\circ$ kann als Anzahl an Umdrehungen der Ortskurve $1 + G_0$ um den Ursprung gedeutet werden. Eine Phasenänderung von $\Delta\varphi = 180^\circ$ entspricht dabei einer halben Umdrehung in mathematisch positiver Richtung – also gegen den Uhrzeigersinn. Da die Ortskurve $1 + G_0$ der Ortskurve von G_0 entspricht, die auf der reellen Achse nur um 1 nach rechts verschoben wurde, gilt:

Umdrehungen von $1 + G_0$ und G_0

Die Anzahl der Umdrehungen von $1 + G_0$ um den Ursprung ist gleich der Anzahl der Umdrehungen von G_0 um den Punkt -1 .

Dabei wird -1 auch als „kritischer Punkt“ bezeichnet.

Es hat sich eingebürgert, Umdrehungen nicht mathematisch positiv, sondern im Uhrzeigersinn zu zählen. Eine Umdrehung von $m = 1$ entspricht dann einer Phasenänderung von $\Delta\varphi = -360^\circ$.

Als weitere Modifikation, die später die Anwendung für Systeme mit Polen auf der imaginären Achse erleichtern wird, betrachtet man nicht die Phasenänderung im Frequenzintervall $0 < \omega < \infty$, sondern in dem Intervall $-\infty < \omega < \infty$. Eine negative Frequenz kann dabei als Vorzeichenwechsel im Phasengang und wegen

$$\cos(-\omega t) + j \sin(-\omega t) = \cos(\omega t) - j \sin(\omega t) \quad (9.26)$$

als eine Spiegelung der Ortskurve an der reellen Achse aufgefasst werden. Somit wird durch die Verdopplung des betrachteten Frequenzintervalls auf

$\infty < 0 < \infty$ auch der Phasenabfall und die Anzahl der Umdrehungen verdoppelt.

Mit diesen Konventionen folgt aus Gl.(9.25) das sogenannten *vollständige Nyquist-Kriterium*:

Vollständiges Nyquist-Kriterium

Gegeben sei ein aufgeschnittener kausaler Regelkreis G_0 und ein zugehöriger geschlossener Regelkreis G . Für die drei Variablen

- $p = p_+(G_0)$: die Anzahl der Pole des aufgeschnittenen Regelkreises G_0 in der rechten offenen s -Halbebene.
- $n = n_+(1+G_0)$: die Anzahl der Pole des geschlossenen Regelkreises G in der rechten offenen s -Halbebene.
- m : Die Anzahl der Umdrehungen der Ortskurve von $G_0(j\omega)$ von $-\infty < \omega < \infty$ um den kritischen Punkt -1 im Uhrzeigersinn.

gilt der folgende Zusammenhang:

$$m = n - p \quad . \quad (9.27)$$

Mit diesem Kriterium lässt sich bei zwei gegebenen Variablen die letzte unbekannte Variable einfach bestimmen. Der Regelfall ist dabei, dass p (aus einer Stabilitätsbetrachtung von G_0) und m (aus einer Betrachtung der Ortskurve von G_0) bekannt sind und n bestimmt werden soll.

Damit der geschlossene Regelkreis stabil ist, muss $n = 0$ gelten. Hierzu muss $m = -p$ sein:

Stabilitätsbedingung nach Nyquist

Der geschlossene Regelkreis ist genau dann stabil, wenn die Ortskurve des Frequenzganges des kausalen aufgeschnittenen Regelkreises $G_0(j\omega)$ beim Durchlaufen der Frequenzwerte von $-\infty$ über null bis $+\infty$ den Punkt -1 genau p Mal im mathematisch positiven Sinn (d. h. gegen den Uhrzeigersinn) umfährt.

Das Nyquist-Kriterium gilt im Gegensatz zu den algebraischen Kriterien ohne weitere Einschränkung auch für Regelkreise mit Totzeitgliedern. Der (relativ aufwendige) Nachweis soll hier nicht geführt werden, sondern es wird auf die Spezialliteratur verwiesen [7].

9.3.2 Beispiele

Als erstes Beispiel soll ein Standard-Regelkreis mit einer PT_3 -Regelstrecke und einem P-Regler herangezogen werden, welches bereits beim Hurwitz-Kriterium verwendet wurde.

$$G_S(s) = \frac{K_S}{(Ts + 1)^3} \quad , \quad G_R(s) = K_R \quad , \quad K_S, K_R, T > 0. \quad (9.28)$$

Mit dem Hurwitz-Kriterium wurde dabei bestimmt, dass für Stabilität $K_SK_R < 8$ gelten muss. Möchte man stattdessen das Nyquist-Kriterium nutzen, so muss als erstes der Frequenzgang des aufgeschnittenen Regelkreises bestimmt werden, der sich direkt als

$$G_0(j\omega) = \frac{K_SK_R}{(1 + j\omega T)^3} \quad (9.29)$$

ergibt.

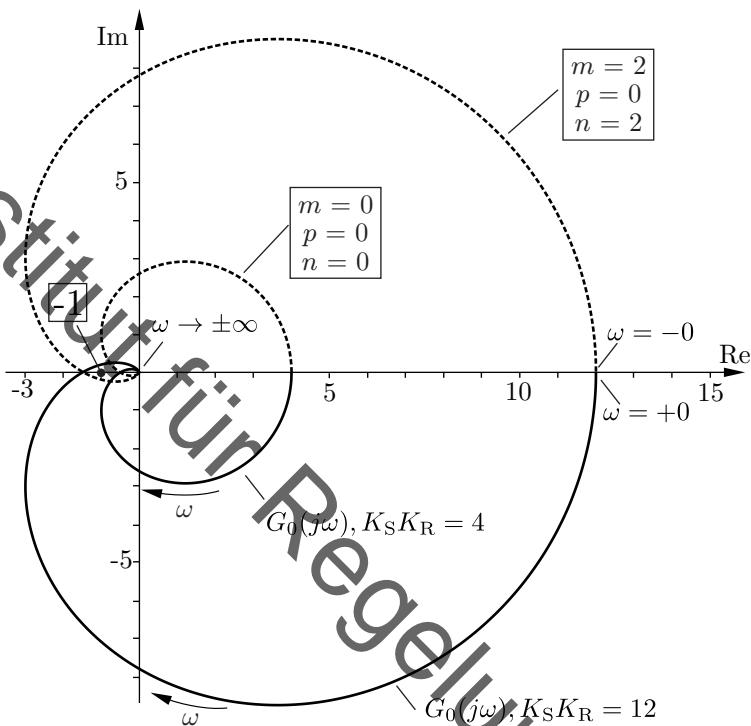
Der aufgeschnittene Regelkreis ist in jedem Fall stabil, die Zahl p der Polstellen seiner Übertragungsfunktion in der rechten offenen s -Halbebene ist daher null.

Die zugehörige Ortskurve ist in Bild 9-3 für zwei Werte von K_SK_R aufgetragen. Die Ortskurve kann dabei für $0 < \omega < \infty$ direkt aus den Tabellen bestimmt werden. Der Verlauf für $-\infty < \omega < 0$ ergibt sich dann durch Spiegelung an der reellen Achse und ist gestrichelt gezeichnet.

Aus Bild 9-3 ist zu ersehen, dass für $K_SK_R = 4$ die Ortskurve G_0 den kritischen Punkt -1 nicht umläuft, weil ein von diesem Punkt an die Ortskurve gezogener Zeiger keine Drehung ausführt, wenn seine Spitze die Ortskurve einmal abfährt. Da die Zahl m der Umläufe null ist und der aufgeschnittene Regelkreis stabil ist ($p = 0$), ist auch der geschlossene Regelkreis stabil.

Auf gleiche Weise stellt man fest, dass die Ortskurve für $K_SK_R = 12$ den Punkt -1 zweimal umläuft, der Nenner der Übertragungsfunktion des geschlossenen Regelkreises daher zwei Nullstellen in der rechten offenen s -Halbebene haben muss und dieser Regelkreis instabil sein wird, was mit den im vorangehenden Abschnitt gewonnenen Aussagen übereinstimmt.

Bei genauerer Betrachtung von Bild 9-3 ist zu erkennen, dass die Ortskurve für $K_SK_R = 12$ genau der Ortskurve für $K_SK_R = 4$ entspricht, nur dass diese um den Faktor 3 gestreckt wurde.

Bild 9-3: Ortskurven eines PT_3 mit verschiedenen Verstärkungen

Skalierung von Ortskurven

Die Ortskurve von $KG_0(s)$ entspricht der Ortskurve von G_0 um den Faktor K gestreckt bzw. gestaucht. Die Umdrehungen von KG_0 um den kritischen Punkt -1 entsprechen daher den Umdrehungen von G_0 um den Punkt $-1/K$.

Dieser Zusammenhang ergibt sich auch sofort aus den Rechenregeln für Reihenschaltungen und den Konstruktionsregeln für Ortskurven.

Da die Ortskurven die negative reelle Achse im Punkt $-0,5$ bzw. $-1,5$ schneiden, wird die (nicht dargestellte) Ortskurve für $K_S K_R = 8$ genau durch den kritischen Punkt -1 verlaufen. Mit einer Argumentation über

Stetigkeit analog zum Stabilitätsrand des Hurwitz-Kriteriums ist klar, dass für diesen Fall, wo die Ortskurve des aufgeschnittenen Regelkreises den Punkt -1 durchläuft (und daher m nicht wohldefiniert ist) der geschlossene Regelkreis Pole auf der imaginären Achse aufweisen muss. Es lässt sich sogar direkt die exakte Position dieser Polstellen berechnen. Für eine Polstelle λ des geschlossenen Regelkreises muss nämlich gelten:

$$|G(s = \lambda)| = \frac{1}{|1 + G_0(s = \lambda)|} \rightarrow \infty \Leftrightarrow G_0(s = \lambda) = -1 . \quad (9.30)$$

Also liegen die Polstellen am Stabilitätsrand genau bei $\pm\omega_\pi$, wobei ω_π hier genau die Frequenz ist, für welche die Ortskurve den Punkt -1 durchläuft.

Da es sich bei Gl.(9.30) um eine genau-dann-wenn-Beziehung handelt, ist dies auch die einzige Möglichkeit, wie Pole des geschlossenen Regelkreises auf den Stabilitätsrand gelangen können. Das ist eine wichtige Erkenntnis, da die Forderung $n_+ = 0$ für den allgemeinen Fall keine Stabilität sicherstellt, da n_+ nur die Pole von G in der rechten offenen s -Halbebene zählt und somit G durchaus Pole auf der imaginären Achsen hätte haben können. Dieser Fall kann aber ausgeschlossen werden, da dann die Ortskurve den Punkt -1 durchlaufen würde. Somit gelten alle vorangegangenen Aussagen auch für Pole des geschlossenen Regelkreises am Stabilitätsrand.

Nyquist-Kriterium und Stabilitätsrand

Der geschlossene Regelkreis besitzt genau dann Pole auf der imaginären Achse, wenn die Ortskurve des aufgeschnittenen Regelkreises G_0 den Punkt -1 genau durchläuft. Die Polstellen am Stabilitätsrand bei $\pm j\omega_\pi$ besitzen dabei einen Imaginärteil, der genau der Frequenz ω_π entspricht, für welchen der kritische Punkt -1 durchlaufen wird ($G(j\omega_\pi) = -1$).

Das Nyquist-Kriterium soll auf ein weiteres Beispiel mit einer instabilen Regelstrecke angewandt werden:

$$G_S(s) = \frac{-1}{1 - sT} , \quad G_R(s) = K_R , \quad G_0(s) = \frac{-K_R}{1 - sT} . \quad (9.31)$$

Man sieht direkt, dass G_0 einen Pol bei $s = 1/T$, also in der rechten offenen s -Halbebene hat. Damit ist $p = 1$ und der aufgeschnittene Regelkreis also instabil.

Bild 9-4 zeigt die zugehörigen Ortskurven von $G_0(j\omega)$ für $K_R = 0,75$ und $K_R = 1,5$. Für $K_R = 0,75$ ergibt Bild 9-4 null Umläufe um den Punkt -1 , der Regelkreis ist demnach instabil. Für $K_R = 1,5$ erhält man einen Umlauf im mathematisch positiven Sinn – der Regelkreis ist stabil. Da der Anfangswert der Ortskurven $G_0(j0) = -K_R$ ist, wird der geschlossene Regelkreis für $K_R < 1$ instabil und für $K_R > 1$ stabil sein. Für $K_R = 1$ durchläuft die Ortskurve für $\omega = \omega_\pi = 0 \text{ sec}^{-1}$ den Punkt -1 .

Nach den vorherigen Ausführungen wird für $K_R = 1$ das System also am Stabilitätsrand mit einem Pol im Ursprung (d. h. grenzstabil) sein. Berechnet man zur Probe für $K_R = 1$ den geschlossenen Regelkreis, so gewinnt man

$$G = \frac{1}{1 + G_0} = \frac{1}{1 + \frac{-1}{1 - sT}} = \frac{1}{-sT} \quad (9.32)$$

mit identischer Aussage.

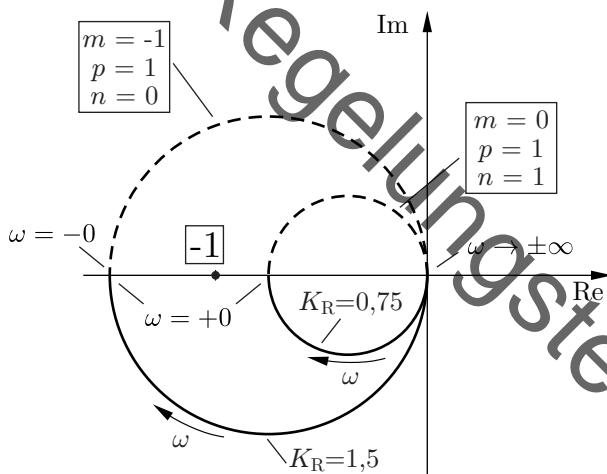


Bild 9-4: Ortskurven zu $G_0 = \frac{-K_R}{1 - j\omega T}$

In einigen komplexeren Fällen als den gezeigten Beispielen fällt das korrekte Ermitteln der Umdrehungen m gelegentlich schwer. Manchem hilft dabei das Verständnis, dass lediglich die Phasenänderung eines Zeigers, der vom

Punkt -1 auf die Ortskurve gerichtet ist, gesucht ist. In solchen Fällen bietet die Vorstellung der Ortskurve als „Seil“ zusätzliche Anschauung beim Zählen der Umdrehungen: Hierzu wird die Ortskurve als Seil aufgefasst, dessen beide Enden (z. B. in $\omega = 0$) zusammengeknotet sind, sodass sich ein geschlossenes Seil wie ein Lasso ergibt. Weiterhin versteht man die komplexe Ebene als eine zweidimensionale Ebene ohne Einschränkungen für das Bewegen des Seils auf dieser Ebene. Der kritische Punkt -1 hingegen ist ein im Boden gerammter Pflock, über welchen das Seil nicht hinwegbewegt werden darf. Die gesuchte Umdrehungszahl m entspricht dann der Anzahl der tatsächlichen Wicklungen des Seils um den Pflock. Kann das Seil gänzlich vom Pflock entfernt werden, muss $m = 0$ gelten.

9.3.3 Anwendung bei Polen am Stabilitätsrand

Die Herleitung des Nyquist-Kriteriums hatte bisher ausgeschlossen, dass sich Pole oder Nullstellen auf der imaginären Achse befinden. Der Fall der Nullstellen (d. h. der Polstellen des geschlossenen Regelkreises) konnte über das Durchlaufen der -1 behandelt werden. Der verbliebene Fall von Polstellen des aufgeschnittenen Regelkreises auf der imaginären Achse bedarf aber noch einer zusätzlichen Argumentation.

Zunächst wird die Ortskurve eines einfachen Integrators $G(j\omega) = 1/(j\omega)$ im Intervall $-\infty < \omega < \infty$ betrachtet. Während für positive Frequenzen die Ortskurve die negative imaginäre Achse bedeckt, verläuft sie für negative Frequenzen aufgrund der Spiegelung auf der positiven imaginären Achse. Wird der Phasengang für alle Frequenzen aufgetragen, so beträgt dieser

$$\varphi(\omega) = \begin{cases} +90^\circ & \text{für } \omega < 0 \\ \text{nicht definiert} & \text{für } \omega = 0 \\ -90^\circ & \text{für } \omega > 0 \end{cases}. \quad (9.33)$$

Obgleich der Wert des Phasengangs für $\omega = 0$ nicht definiert ist und somit eine Unstetigkeit vorliegt, scheint es zunächst so, als könnte man dennoch eine Phasenänderung von $\Delta\varphi = -180^\circ$ ablesen. Leider ist der Phasenwinkel aber bis auf $\pm 360^\circ$ nicht definiert. Daher wäre die Zuordnung der Phasen

über

$$\varphi(\omega) = \begin{cases} +90^\circ & \text{für } \omega < 0 \\ \text{nicht definiert} & \text{für } \omega = 0 \\ +270^\circ & \text{für } \omega > 0 \end{cases} \quad (9.34)$$

ebenso korrekt, was auf eine Phasenänderung von $\Delta\varphi = +180^\circ$ führt. Da die Phase im Punkt $\omega = 0$ nicht definiert ist, kann nicht zwischen den beiden Optionen unterschieden werden. Dasselbe gilt wortgleich für den Fall eines komplex konjugierten Polstellenpaars auf der imaginären Achse. Der gezeigte Verlauf in Tab. 6-2 ist insofern exemplarisch für die Interpretation einer fallenden Phase.

Welche Auswirkungen hat diese Beobachtung auf das Nyquist-Kriterium für aufgeschnittene Regelkreise, die Polstellen auf der imaginären Achse haben? Lässt man die Herleitung Revue passieren, so erkennt man, dass p_+ alle Pole mit steigender Phase, p_- hingegen alle Pole mit sinkender Phase umfasst. Alle weiteren Eigenschaften sind nicht von Belang. Daher gibt es zwei Möglichkeiten zum Umgang mit Polen am Stabilitätsrand.

- a) Pole am Stabilitätsrand werden als Pole mit fallender Phase aufgefasst. Sie gelten dann zwecks Anwendung des Nyquist-Kriteriums als stabil und tragen nicht zu $p = p_+$ bei. Die Ortskurve muss bei der Unstetigkeitsstelle mit fallender Phase geschlossen werden.
- b) Pole am Stabilitätsrand werden als Pole mit steigender Phase aufgefasst. Sie gelten dann zwecks Anwendung des Nyquist-Kriteriums als instabil und tragen voll zu $p = p_+$ bei. Die Ortskurve muss bei der Unstetigkeitsstelle mit steigender Phase geschlossen werden.

Beide Optionen sind valide Möglichkeiten. Üblicherweise entscheidet man sich für Option a), da so das bereits formulierte Nyquist-Kriterium nicht verändert werden muss. Dieses gilt dann ohne Einschränkungen auch für den allgemeinen Fall mit Pol- oder Nullstellen auf der imaginären Achse. Ansonsten muss p die Anzahl der Pole in der abgeschlossenen rechten s -Halbebene bezeichnen.

Das Schließen der Ortskurve wird für ein IT_1 -Glied beispielhaft ausgeführt. Zunächst wird gemäß Option a) der Pol in $s = 0$ als „stabil“ (d. h. mit fallender Phase) betrachtet und die Ortskurve daher mit einem Phasenabfall von -180° geschlossen. Hierdurch ergibt sich ein Halbkreis im Unendlichen –

siehe Bild 9-5 auf der linken Seite. Im Falle mehrfacher Polstellen in $s = 0$ würde für jede Polstelle ein Phasenabfall von -180° benötigt. Es ergibt sich mit der geschlossenen Ortskurve keine Umdrehung ($m = 0$), womit aufgrund der „Stabilität“ (der Pol am Stabilitätsrand zählt nicht zu p) des aufgeschnittenen Regelkreises auch der geschlossene Regelkreis stabil ist.

Die Alternative b) führt auf $m = -1$, da der Halbkreis mit steigender Phase geschlossen wird, und $p = 1$ (da der Integrator als instabil gilt) und damit auf $n = 0$ und ein identisches Ergebnis. Im folgenden wird ausschließlich Fall a) betrachtet.

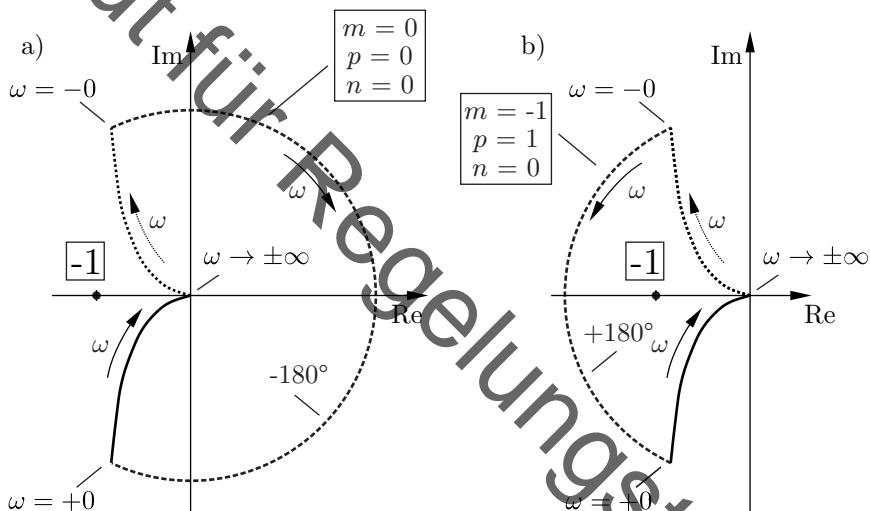


Bild 9-5: Schließen der Ortskurve für einen Regelkreis mit IT_1 -Element

9.3.4 Vereinfachtes Nyquist-Kriterium

Das vollständige Nyquist-Kriterium ermöglicht eine schnelle Stabilitätsprüfung anhand der Ortskurve von G_0 . Möchte man dies zum Reglerentwurf nutzen, so kann man versuchen, G_R so anzupassen, dass $G_0 = G_S \cdot G_R$ die entsprechenden Umläufe um den kritischen Punkt -1 aufweist.

Leider ist die Veränderung von G_0 bei einer Modifikation von G_R (außer

in den einfachen Fällen der Änderung der statischen Verstärkung des Reglers) in der Ortskurve nur schwer abschätzbar. Sehr viel einfacher wäre dies im Bode-Diagramm, wo sich G_0 aus der simplen Addition von G_S und G_R ergibt. In der Darstellung im Bode-Diagramm können die Anzahl der Umdrehungen um den Punkt -1 allerdings nicht so einfach wie in der Ortskurve abgelesen werden.

Das vereinfachte Nyquist-Kriterium bietet unter zwei Zusatzannahmen eine Möglichkeit, das Nyquist-Kriterium im Bode-Diagramm mit überschaubarem Aufwand anwenden zu können.

Als erste Annahme wird getroffen, dass $p = 0$ gilt. Bei einer großen Anzahl von Anwendungsfällen ist der aufgeschnittene Regelkreis stabil oder hat integrierendes Verhalten ($G_0(s)$ hat einen einfachen Pol bei $s = 0$, alle anderen Pole liegen in den linken offenen s -Halbebene), sodass diese Anforderung erfüllt werden kann.

Gilt $p = 0$, so darf die Ortskurve von $G_0(j\omega)$ den Punkt -1 nicht umlaufen, wenn der geschlossene Regelkreis stabil sein soll. Wie kann so eine unerwünschte Umdrehung entstehen? Sie ist nur möglich, wenn die Ortskurve von G_0 auf der reellen Achse sowohl durch Punkte rechts als auch links der -1 verläuft. Befinden sich die Punkte, an denen die Ortskurve auf der reellen Achse liegt, ausschließlich auf einer Seite der -1 , so liegt in jedem Fall keine Umdrehung vor. Es reicht dabei aus, das Frequenzintervall $-0 \leq \omega \leq \infty$ zu betrachten, da die Spiegelung der Ortskurve hier keine zusätzlichen Umdrehungen erzeugt, solange das korrekte Schließen der Ortskurve (für den Fall von Polen bei $s = 0$) mitbetrachtet wird. Die allermeisten Ortskurven (nämlich die von Systemen ohne Durchgriff) werden für $\omega \rightarrow \infty$ in den Ursprung laufen, welcher rechts der -1 liegt. Daher liegt es nahe, als Stabilitätsbedingung zu fordern, dass die Ortskurve von G_0 im Intervall $0 \leq \omega \leq \infty$ die reelle Achse ausschließlich rechts der -1 schneiden soll.

Stabilität und Schnitt mit der reellen Achse

Gilt für den aufgeschnittenen Regelkreis $p = 0$ und schneidet die Ortskurve von G_0 im Intervall $0 \leq \omega \leq \infty$ die reelle Achse ausschließlich rechts der -1 , so ist der geschlossene Regelkreis stabil.

Befindet sich die Ortskurve für $\omega = 0$ oder $\omega = \infty$ auf der reellen Achse, so gilt dies ebenfalls als Schnittpunkt, da die vollständige Ortskurve das

Intervall $-\infty < \omega < \infty$ umfasst.

Diese Bedingung lässt sich relativ einfach auch anhand des Bode-Diagramms prüfen. Ein Schnitt der Ortskurve an der positiven reellen Achse (das ist ein Phasenwinkel von 0°), ist unproblematisch. Kritisch wäre ein Punkt auf der negativen reellen Achse (das ist ein Phasenwinkel von -180°) mit einem Betrag größer als eins, da dann dieser links der -1 liegen würde. Folglich kommt den Werten, wo der Phasengang des aufgeschnittenen Kreises -180° annimmt, eine zentrale Bedeutung für die Stabilität des geschlossenen Kreises zu.

Durchtrittsfrequenz ω_π

Die Kreisfrequenzen, für die der Phasengang des aufgeschnittenen Regelkreises die Phasenwinkel $-180^\circ \pm n \cdot 360^\circ$ (in Grad) beziehungsweise $-\pi \pm 2\pi n$ (in Bogenmaß) annimmt, heißen Durchtrittsfrequenzen ω_π .

Das Hinzufügen der Vielfachen von 360° ist aufgrund der Periodizität der Winkelfunktionen nötig. Auch ein Wert von $\omega_\pi = \infty$ ist zugelassen, wenn die Ortskurve gegen die negative reelle Achse konvergiert. Ebenso kann es sein, dass ein Phasengang keine einzige Kreisfrequenz ω_π besitzt und die Menge aller ω_π leer ist. Ist der aufgeschnittene Regelkreis am Stabilitätsrand, müssen die ω_π , welche möglicherweise durch das Vervollständigen der Ortskurve entstehen, mitberücksichtigt werden. In vielen Fällen wird es genau eine Frequenz ω_π geben, die sich im Bode-Diagramm aus dem Phasengang ebenso wie der Betrag an dieser Stelle sehr einfach ermitteln lässt. Es bietet sich nun die folgende Stabilitätsbedingung an, die sich auch im Bode-Diagramm leicht prüfen lässt:

Hinreichende Stabilitätsbedingung

Gilt für den aufgeschnittenen Regelkreis $p = 0$ und gilt $|G_0(j\omega_\pi)| < 1$ für alle ω_π , so ist der geschlossene Regelkreis stabil.

Der Nachteil an dieser Bedingung ist, dass ihre Erfüllung zwar Stabilität impliziert, aber ihre Verletzung nicht unbedingt Instabilität bedeutet. Es ist also bisher nur ein hinreichendes und kein notwendiges Kriterium.

Bild 9-6 zeigt die Ortskurve des Frequenzgangs eines aufgeschnittenen Regelkreises G_0 , der die Vorbedingungen $p = 0$ erfüllt. Die negative reelle Achse wird dreimal geschnitten (einmal im Unendlichen), sodass es drei ver-

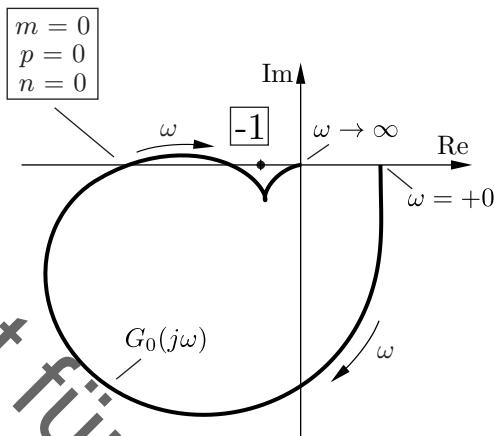


Bild 9-6: $G_0(j\omega)$ mit mehrfachen Schnittpunkten mit der negativen reellen Achse

schiedene ω_π gibt. Die Bedingung $|G_0(j\omega_\pi)| < 1$ ist für die beiden Schnittpunkte nicht erfüllt. Dennoch erkennt man, dass der geschlossene Regelkreis stabil sein wird, weil der Punkt -1 von der Ortskurve nicht umschlossen wird.

Regelungssysteme mit Frequenzgangortskurven, die die reelle Achse mehrfach schneiden, wie die in Bild 9-6, kommen recht selten vor. Dennoch sorgen die mehrfachen Schnittpunkte links der -1 dafür, dass der geschlossene Regelkreis stabil ist, obgleich die formulierte Bedingung fälschlicherweise auf eine mögliche Instabilität hindeutet.

Zur Reglerauslegung und Bestimmung der Stabilitätsbereiche der Reglerparameter ist ein notwendiges und hinreichendes Stabilitätskriterium essentiell. Um ein solches zu erhalten, müssen Fälle wie in Bild 9-6 ausgeschlossen werden. Dies erreicht man mit der Forderung, dass die Ortskurve für steigende $\omega \geq 0$ stets vom dritten Quadranten aus kommend die negative reelle Achse schneidet, nicht aber vom zweiten Quadranten aus. Für das Bode-Diagramm bedeutet dies, dass der Phasengang die $-180^\circ \pm n \cdot 360^\circ$ -Linien nur mit negativer Steigung $d\varphi/d\omega \leq -90^\circ$ schneiden darf. Hierbei handelt es sich (neben der Bedingung $p = 0$) um die zweite Bedingung für die Anwendbarkeit des vereinfachten Nyquist-Kriteriums:

Anwendbarkeit des Vereinfachten Nyquist-Kriteriums

Gilt für den aufgeschnittenen Regelkreis $p = 0$ und schneidet der Phasengang die $-180^\circ \pm n \cdot 360^\circ$ -Linien mit negativer Steigung, so ist das vereinfachte Nyquist-Kriterium *anwendbar*.

Aussage des Vereinfachten Nyquist-Kriterium

Der geschlossene Regelkreis ist genau dann stabil, wenn $|G_0(j\omega_\pi)| < 1$ für alle ω_π gilt.

Für den Fall $|G_0(j\omega_\pi)| = 1$ ist der geschlossene Regelkreis am Stabilitätsrand. Es handelt sich um ein hinreichendes und notwendiges Kriterium, sodass mit diesem (Anwendbarkeit vorausgesetzt) genau die Reglerparameter ausgerechnet werden können, für die der geschlossene Regelkreis stabil ist.

Das so gewonnene vereinfachte Nyquist-Kriterium soll durch ein Gedankenexperiment zusätzlich erläutert werden. Dazu soll der in Bild 9-7 dargestellte aufgeschnittene Regelkreis dienen. Der aufgeschnittene Regelkreis $G_0(j\omega)$ möge stabil und so beschaffen sein, dass das Nyquist-Kriterium in seiner zuletzt angegebenen, vereinfachten Fassung anwendbar ist.

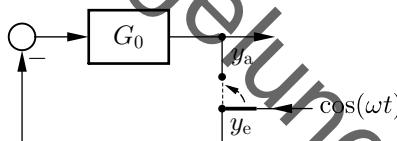


Bild 9-7: Aufgeschnittener Regelkreis mit Schalter

Der Regelkreis wird, wie in Bild 9-7 dargestellt, aufgeschnitten und mit

$$y_e(t) = \cos(\omega t) \quad (9.35)$$

als Eingangsgröße des Reglers erregt. Im eingeschwungenen Zustand wird die Ausgangsgröße der Regelstrecke zu

$$y_a(t) = -|G_0(j\omega)| \cdot \cos(\omega t + \varphi(\omega)) \quad (9.36)$$

und sie ist damit in Amplitude und Phasenlage von der Frequenz der Eingangsgröße abhängig. Diese Frequenz wird nun so gewählt, dass der Phasenwinkel $\varphi = -\pi$ das Minuszeichen vor $|G_0|$ kompensiert. Es gilt für die

so gewählte Frequenz $\omega = \omega_\pi$ wegen $-\cos(\omega) = \cos(\omega + \pi)$:

$$\begin{aligned} y_a(t) &= -|G_0(j\omega)| \cdot \cos(\omega t + \varphi(\omega)) \\ &= |G_0(j\omega_\pi)| \cdot \cos(\underbrace{\omega t + \varphi(\omega_\pi)}_{=-\pi} + \pi) = |G_0(j\omega_\pi)| \cdot y_e(t) \quad . \end{aligned} \quad (9.37)$$

Die Amplitude der Ausgangsgröße kann nun größer, gleich oder kleiner sein als die der Eingangsgröße. Wenn sie (zufällig) wegen $|G_0(j\omega_\pi)| = 1$ gleich der Amplitude der Eingangsgröße sein sollte, dann ist $y_e(t) = y_a(t)$.

Man kann also mit dem in Bild 9-7 dargestellten Schalter die Ausgangsgröße der Regelstrecke dem Eingang des Reglers zuführen, ohne an den Größen selbst etwas zu ändern. Nach dem Umschalten hat man einen geschlossenen Regelkreis vor sich, der Dauerschwingungen ausführt und sich daher am Stabilitätsrand befindet.

Wenn bei der schon definierten Frequenz ω_π hingegen $|G_0(j\omega_\pi)| > 1$ sein sollte, so würde nach Umschalten des Schalters in Bild 9-7 der Regler eine größere Eingangsgröße erhalten als vorher. Diese größere Eingangsgröße hätte (neben Einschwingvorgängen) eine größere Ausgangsgröße der Regelstrecke zur Folge, welche als Eingangsgröße des Reglers zu abermaliger Vergrößerung der Ausgangsgröße führen würde. Ein solcher geschlossener Regelkreis, der aufklingende Schwingungen ausführt, ist instabil.

Im entgegengesetzten Fall gilt bei $\omega = \omega_\pi$ die Bedingung $|G_0(j\omega_\pi)| < 1$. Dann würde nach Umschalten des Schalters der Regler eine kleinere Eingangsgröße als vorher erhalten. Die durch das äußere Signal aufgeprägte Schwingung würde verlöschen und der geschlossene Regelkreis wäre stabil.

Aus den Ergebnissen des Gedankenexperiments kann man folgern, dass ein geschlossener Regelkreis stabil ist, wenn der Betrag des Frequenzgangs des aufgeschnittenen Regelkreises bei der Frequenz (oder den Frequenzen) ω_π kleiner ist als eins; dabei ist ω_π dadurch festgelegt, dass für diese Frequenz(en) der Phasenwinkel des Frequenzgangs des aufgeschnittenen Regelkreises $\varphi_0(\omega_\pi) = -(2n+1)\pi$ ist. Dies lässt sich zu der Vorschrift umformen, dass die Ortskurve von $G_0(j\omega)$ die negativ-reelle Achse nur rechts vom Punkt -1 schneiden darf. Bedauerlicherweise führt der betrachtete Gedankenversuch in wenigen Fällen zu falschen Schlüssen. Einer dieser Fälle ist in Bild 9-6 dargestellt; die Ortskurve von G_0 schneidet die reelle Achse links der -1 , dennoch ist der zugehörige geschlossene Regelkreis stabil.

Für die praktische Behandlung linearer Regelungssysteme wird i. Allg. die Darstellung von Frequenzgängen im Bode-Diagramm gegenüber der Ortskurvendarstellung bevorzugt, weil sie mit sehr viel geringerem Aufwand (z. B. bei der Multiplikation) verbunden ist. Insbesondere das vereinfachte Nyquist-Kriterium lässt sich dabei in einfacher Weise im Bode-Diagramm anwenden.

Wie in Bild 9-8 beispielhaft gezeigt wird, kann man zur Stabilitätsprüfung mit dem vereinfachten Nyquist-Kriterium aus dem Phasengang bei $\varphi_0(\omega_\pi) = -\pi$ die Frequenz ω_π bestimmen. Man muss dann prüfen, ob der Betrag des Frequenzgangs $|G_0(j\omega_\pi)|$ kleiner als eins ist.

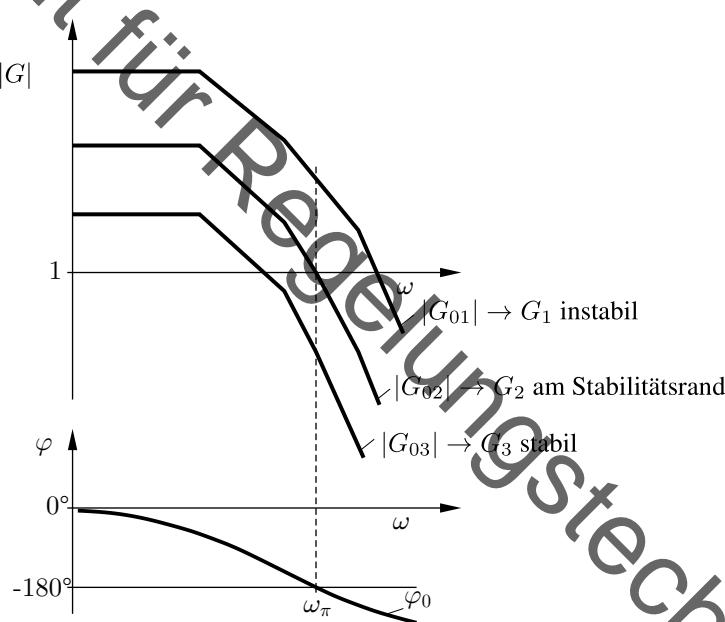


Bild 9-8: Nyquist-Kriterium im Bode-Diagramm

Prinzipiell lässt sich mit der Darstellung von $G_0(j\omega)$ im Bode-Diagramm auch die Stabilität über das vollständige Nyquist-Kriterium prüfen. Sofern der Phasengang die -180° -Linie mehrfach schneidet oder aus anderen Gründen schwierig zu interpretieren ist, so führt eine anhand des Bode-Diagramms angefertigte Skizze der Ortskurve zu den gesuchten Antworten.

Wichtig ist vor Anwendung des vereinfachten Nyquist-Kriteriums seine Anwendbarkeit zu prüfen. Trotz der Übersichtlichkeit der Bedingungen unterlaufen hierbei insbesondere im Bode-Diagramm beim Prüfen der Phasenbe dingung typische Fehler. Hierfür sind in Bild 9-9 zwei Phasengänge gezeigt, die zu offenen Regelkreisen mit $p = 0$ gehören sollen.

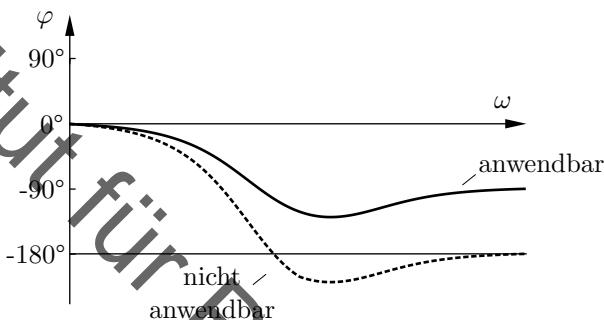


Bild 9-9: Zwei Phasengänge und die Anwendbarkeit des vereinfachten Nyquist-Kriteriums

Der durchgezogene Phasengang entspricht dem Phasengang aus 9-2 und erreicht die -180° -Linie gar nicht. Somit liegt auch nicht der Fall vor, dass die negative reelle Achse vom zweiten Quadranten aus erreicht wird. Das vereinfachte Nyquist-Kriterium ist also anwendbar. Da ω_π nicht definiert ist, ist die Bedingung $|G(j\omega_\pi)| < 1$ für alle ω_π automatisch erfüllt.

Systeme mit $\varphi(\omega) \neq -180^\circ$

Gilt für den aufgeschnittenen Regelkreis $p = 0$ und $\varphi(\omega) \neq -180^\circ \pm n \cdot 360^\circ \forall \omega$, so ist der geschlossene Regelkreis (unabhängig vom Betragsverlauf) stabil.

Der gestrichelte Phasengang schneidet die -180° -Linie zwar zunächst mit negativer Steigung, schneidet diese dann aber mit einer positiven Steigung für $\omega \rightarrow \infty$. Dieser Grenzwert muss bei der Anwendung des vereinfachten Nyquist-Kriteriums mitberücksichtigt werden. Er entspricht hier einem Endwert der Ortskurve auf der negativen reellen Achse aus dem zweiten Quadranten kommend, weshalb das Kriterium nicht anwendbar ist.

Im Falle von nicht sprungfähigem G_0 wird der zugehörige Wert der Ortskur-

ve allerdings im Ursprung liegen, sodass Situationen wie in Bild 9-6 nicht auftreten können. Dies muss separat geprüft werden.

Bei der beliebten Anwendung des vereinfachten Nyquist-Kriteriums im Bode-Diagramm ist insbesondere bei aufgeschnittenen Regelkreisen mit integrierendem Verhalten Vorsicht geboten. Das liegt daran, dass der Punkt $\omega = 0$, in welchem die Ortskurve geschlossen werden muss, im Bode-Diagramm aufgrund der logarithmischen Skalierung der Frequenzachse nicht dargestellt wird. Hierzu sind in Bild 9-10 die zwei Phasengänge der Frequenzgänge $1/(j\omega)$ und $-1/(j\omega)$ einmal in logarithmischer und einmal in linearer Skalierung der Frequenzachse gezeigt.

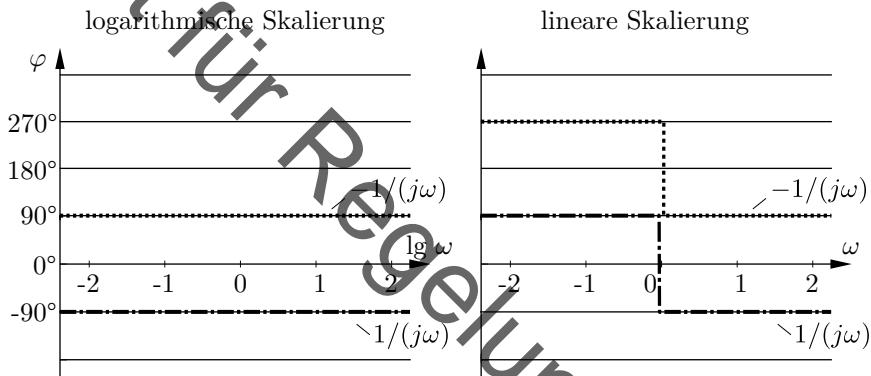


Bild 9-10: Phasengänge von $1/(j\omega)$ und $-1/(j\omega)$ mit linearer und logarithmischer Skalierung der Frequenzachse

Beide Phasengänge scheinen in der logarithmischen Skalierung die $-180^\circ + n \cdot 360^\circ$ -Linien gar nicht zu schneiden. Da integrierendes Verhalten vorliegt, muss die Ortskurve aber noch geschlossen werden. Um die Bedingung $p = 0$ zu erfüllen, muss die Ortskurve in $\omega = 0$ mit einer um -180° fallenden Phase geschlossen werden.

Folglich verläuft die Phase im Fall $1/(j\omega)$ von $+90^\circ$ nach -90° – es entsteht kein Schnittpunkt mit der $+180^\circ$ -Linie und der geschlossene Regelkreis ist stabil. Für den Fall $-1/(j\omega)$ verläuft die Phase in $\omega = 0$ jedoch von $+270^\circ$ nach $+90^\circ$ wie in der linearen Skalierung zu sehen. Es liegt also ein Schnittpunkt für $\omega_\pi = 0$ vor, der mit einem Betrag $|G(j0)| = \infty > 1$ korrespondiert, weswegen der geschlossene Regelkreis instabil ist.

Vereinfachtes Nyquist-Kriterium für Systeme mit integrierendem Verhalten

Gegeben ist ein aufgeschnittener Regelkreis mit integrierendem Verhalten, für den das vereinfachte Nyquist-Kriterium anwendbar ist. Besitzt dieser einer Startphase von $\varphi(0) = -90^\circ$, so entstehen durch das Schließen der Ortskurve keine zusätzlichen Schnittpunkte mit der -180° -Linie. Beträgt die Startphase jedoch $\varphi(0) = +90^\circ$, so ist der geschlossene Regelkreis in jedem Fall instabil.

Eine weitere Schlussfolgerung, die sich direkt aus den Überlegungen, die zur Herleitung des vereinfachten Nyquist-Kriteriums führten, ergibt, ist der Satz der kleinen Verstärkungen (englisch: „small gain theorem“). Grundidee ist, dass ein $G_0(j\omega)$, dessen Ortskurve sich stets im Einheitskreis befindet, unmöglich eine Umdrehung um den Punkt -1 besitzen kann.

Satz der kleinen Verstärkungen

Gegeben ist ein stabiler aufgeschnittener Regelkreis. Gilt

$$|G_0(j\omega)| < 1 \quad \forall \omega , \quad (9.38)$$

so ist der geschlossene Regelkreis (unabhängig vom Phasenverlauf) stabil.

Dieser Satz ist für die Einstellungen von Reglern von gewissem Interesse. Ist die Regelstrecke G_S stabil, so wird ein stabiler Regler, der eine sehr kleine statische Verstärkung besitzt, den Regelkreis nicht destabilisieren können. Das hat zu dem häufigen Vorgehen geführt, bei der Reglereinstellung für stabile Strecken mit sehr kleinen Verstärkungen des Reglers zu beginnen und diese dann sukzessive zu erhöhen.

Für den Stabilitätsnachweis eines praktisch eingesetzten Reglers ist der Satz der kleinen Verstärkungen aber meist unbrauchbar, da für ein gutes Reglerverhalten weitaus größere Verstärkungen notwendig sein werden (siehe Abschnitt 10.3).

9.3.5 Amplituden- und Phasenreserve

Im Gegensatz zum Hurwitz-Kriterium kann man mithilfe des Nyquist-Kriteriums interpretierbare Angaben über die Stabilitätsgüte des geschlos-

senen Regelkreises gewinnen. Hierzu betrachtet man den Verlauf der Ortskurve des Frequenzganges des aufgeschnittenen Regelkreises in der Umgebung des kritischen Punktes -1 . Es leuchtet ein, dass Regelkreise, für die die Ortskurve des Frequenzganges des aufgeschnittenen Regelkreises sehr nahe am kritischen Punkt verläuft, nahe am Stabilitätsrand sind. Solche Regelkreise erscheinen schlecht gedämpft und können durch geringfügige Veränderungen im dynamischen Verhalten ihrer Glieder instabil werden.

Im Gegensatz dazu arbeiten Regelkreise, deren Ortskurven des Frequenzganges des aufgeschnittenen Kreises mit großem Abstand vom kritischen Punkt verlaufen, meist sehr träge. Dieser Verlauf der Ortskurve wird nämlich oft durch einen sehr kleinen Übertragungsfaktor des Reglers erreicht. Zwischen den skizzierten Extremen liegt ein mittlerer technisch brauchbarer Bereich, der durch Entwurfsregeln (siehe Abschnitt 11.1) umschrieben wird.

Um die vage Umschreibung des „Abstands vom Punkt -1 “ mathematisch zu fassen, haben sich zwei Faktoren etabliert, die neben einer leichten Bestimmbarkeit auch eine starke technische Interpretation mit sich bringen.

Amplituden- und Phasenreserve; Durchtrittsfrequenz ω_d

Die *Amplitudenreserve* ist die Zahl, mit der die Ortskurve des aufgeschnittenen Regelkreises multipliziert werden muss, so dass diese genau durch den kritischen Punkt -1 verläuft:

$$A_R = \frac{1}{|G_0(j\omega_\pi)|} . \quad (9.39)$$

Gibt es verschiedene Amplitudenreserven, so ist die kleinste Amplitudenreserve die entscheidende Größe.

Die *Phasenreserve* ist der Winkel, um den die Ortskurve des aufgeschnittenen Regelkreises gedreht werden muss, so dass diese genau durch den kritischen Punkt -1 verläuft:

$$\alpha_R = \varphi_0(\omega_d) - (180^\circ \pm n \cdot 360^\circ) . \quad (9.40)$$

mit der Durchtrittsfrequenz ω_d definiert durch

$$|G_0(j\omega_d)| = 1 . \quad (9.41)$$

Gibt es verschiedene Phasenreserven, so ist die betragsmäßig kleinste Phasenreserve entscheidend.

Beide Definitionen sind graphisch in Bild 9-11 gezeigt.

Die Durchtrittsfrequenz ω_d ist nicht mit der Eigenkreisfrequenz ω_D zu verwechseln. Sowohl Amplituden- als auch Phasenreserve werden üblicherweise nur im Kontext des vereinfachten Nyquist-Kriteriums genutzt und lassen sich leicht im Bode-Diagramm bestimmen (siehe Bild 9-12).

Für einen Regelkreis am Stabilitätsrand ist offenbar

$$A_R = 1 \quad , \quad \alpha_R = 0 \quad . \quad (9.42)$$

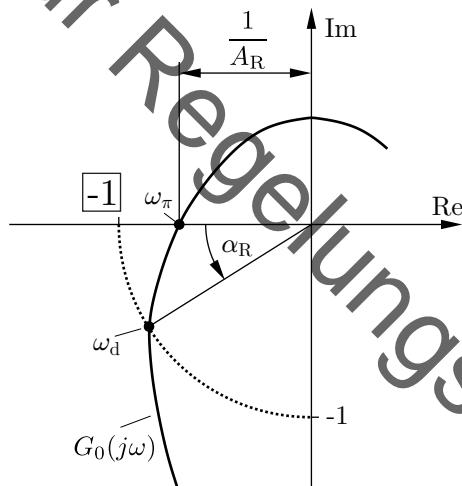


Bild 9-11: Amplituden- und Phasenreserve

Mit Amplituden- und Phasenreserve wird der Abstand vom kritischen Punkt -1 in zwei Anteile aufgeteilt: Zum einen eine reine Streckung oder Stauchung der Ortskurve, zum anderen eine reine Drehung der Ortskurve.

Ortskurven werden genau dadurch gestreckt, indem die statische Verstärkung des Systems verändert wird. Folglich beschreibt (ausgehend von einem

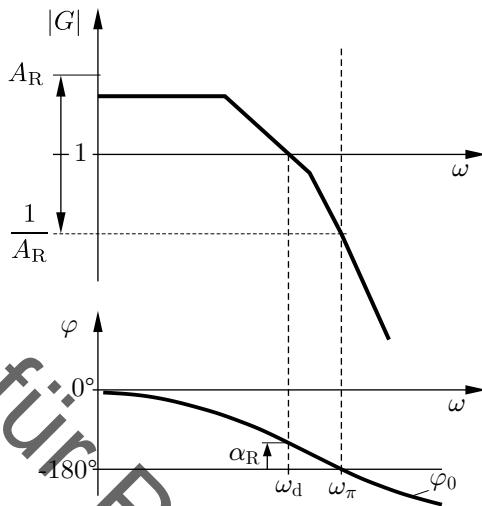


Bild 9-12: Amplituden- und Phasenreserve im Bode-Diagramm

stabilen geschlossenen Regelkreis) die Amplitudenreserve den Faktor, um welchen die statische Verstärkung des aufgeschnittenen Regelkreises verändert werden darf, sodass sich das System am Stabilitätsrand befindet. Daher kann die Amplitudenreserve als maximal zulässige zusätzliche Verstärkung im System verstanden werden. Diese Verstärkung kann dabei von einem Regler herrühren, wo sie ein Maß dafür ist, wie nah die Reglerverstärkung an ihrem zulässigen Maximum ist. Alternativ entstammt sie einer ungenauen Modellierung der Regelstrecke und beschreibt damit, wie genau das statische Verhalten der Regelstrecke mindestens bekannt sein muss.

Die Phasenreserve beschreibt die maximal zulässige Drehung der Ortskurve. Hiermit ist sie ein Maß dafür, wie groß eine zusätzlich im aufgeschnittenen Regelkreis befindliche Totzeit sein darf, ohne die Stabilitätseigenschaften des geschlossenen Regelkreises zu verändern. Um dies einzusehen rekapituliert man, dass das Totzeit-Element einen konstanten Amplitudengang und eine fallende Phase besitzt. Ein mit G_0 zusätzlich in Reihe geschaltetes Totzeit-Glied wird also die Amplitude des aufgeschnittenen Regelkreises nicht verändern, womit die Durchtrittsfrequenz ω_d gleich bleibt. Die durch die zusätzliche Totzeit T_t hervorgerufene Phasenänderung zur Frequenz ω_d

beträgt genau $-\omega_d T_t$. Für einen Wert der Totzeit von genau

$$-\omega_d T_t = -\alpha_R \quad \Rightarrow \quad T_t = \frac{\alpha_R}{\omega_d} \quad (9.43)$$

wird der Punkt, an welchem die Ortskurve von G_0 den Einheitskreis erreicht, in den kritischen Punkt -1 gedreht, sodass das System am Stabilitätsrand ist. Somit sind Amplituden- und Phasenreserve ein Maß für die Robustheit des Regelkreises.

Robustheit

Ein Regelkreis heißt *robust*, wenn eine zufriedenstellende Funktion trotz Modellunsicherheiten sichergestellt ist.

Robustheit ist eine wichtige Anforderung an Regelkreise, da die zum Reglerentwurf genutzten Modelle Vereinfachungen, Linearisierungen oder identifizierte Anteile enthalten sowie sich über die Zeit (langsam) ändern können. In diesem Sinne kann das Linearisierungstheorem aus Abschnitt 3.5 so gedeutet werden, dass Systeme, die grenzstabil sind, keinerlei Robustheit besitzen, da ein beliebig kleiner Linearisierungsfehler die Stabilitätseigenschaften verändern kann.

Die Robustheit linearer Systeme kann mithilfe von Amplituden- und Phasenreserve quantifiziert werden, wobei größere Reserven auf einen robusteren Regelkreis schließen lassen. Den Verfahren der robusten Regelung widmet sich dabei umfangreiche Spezialliteratur, deren Methoden den Umfang eines Einführungswerkes sprengen [1, 53].

Aus den Ausführungen zur Phasenreserve sieht man zusätzlich, dass Regelkreise, die die Bedingungen des Satzes für kleine Verstärkungen nicht erfüllen, durch eine hinreichend große Totzeit im System destabilisiert werden können. Sofern nämlich $|G_0(j\omega)| > 1$ für mindestens ein $\omega \neq 0$ gilt, kann mit einer entsprechend großen Totzeit dieser Punkt der Ortskurve so gedreht werden, dass dieser links der -1 zum Liegen kommt und hierdurch eine Umdrehung im Uhrzeigersinn entsteht ($m > 0$). Der geschlossene Regelkreis ist dann (unabhängig von der Stabilität des aufgeschnittenen Regelkreises) instabil, da $0 < m = n - p$ nur für $n > 0$ erfüllt werden kann.

Einfluss von Totzeiten auf Stabilität

Jeder aufgeschnittene Regelkreis $G_0(j\omega)$, der mindestens eine Frequenz $\omega \neq 0$ mit $|G_0(j\omega)| > 1$ besitzt, führt für eine hinreichend große Totzeit in G_0 auf einen instabilen geschlossenen Regelkreis.

Totzeiten verschlechtern also die Robustheitseigenschaften und verursachen Stabilitätsprobleme im geschlossenen Regelkreis. Beim Aufbau eines Regelungssystems ist also eine geringe Totzeit wünschenswert. Das kann schnelle Taktraten des (digital umgesetzten) Reglers und eine räumliche Nähe des Reglers zur Regelstrecke für niedrige Latenzen notwendig machen. Anschaulich reicht es für eine erfolgreiche Regelung nicht aus, die Informationen zur aktuellen Regelgröße y irgendwann zu erhalten. Stattdessen müssen diese Informationen rechtzeitig vorliegen, da nur so der Regler geeignet darauf reagieren kann. Veraltete Informationen (d. h. eine Regelung mit zu hoher Totzeit) können also schlimmere Auswirkungen haben als gar keine Informationen (d. h. keine Regelung).

9.4 Sonderfälle

Bei der Herleitung der Stabilitätskriterien nach Hurwitz/Routh sowie nach Nyquist wurden bestimmte Sonderfälle durch unauffällige mathematische Details ausgeschlossen. Diese Sonderfälle sind für den praktischen Gebrauch von nachrangiger Relevanz, sollen aber der Vollständigkeit halber hier diskutiert werden, da man ansonsten in diesen Situationen zu falschen Schlüssen gelangen kann. Konkret geht es um Pol-Nullstellen-Kürzungen sowie um sich unstetig verändernde Polstellen.

9.4.1 Pol-Nullstellen-Kürzungen

Pol-Nullstellen-Kürzungen treten genau dann in einem System auf, wenn dieses keine minimale Realisierung ist. Nur für minimale Realisierungen ist die BIBO-Stabilität, die über das Ein-Ausgangsverhalten wie z. B. Übertragungsfunktionen untersucht wird, mit der Stabilität im Zustandsraum identisch. Im Falle von Pol-Nullstellen-Kürzungen kann es passieren, dass zwar das Ein-Ausgangsverhalten BIBO-stabil ist, aber interne Größen, die nicht im Systemausgang sichtbar sind, divergieren. Daher wurde bei den entsprechenden Stabilitätsbetrachtungen stets der Fall von Pol-Nullstellen-

Kürzungen ausgeschlossen.

Die Annahme, dass es sich bei der Beschreibung einer Regelstrecke oder eines Reglers um eine minimale Realisierung handelt, ist keine wesentliche Einschränkung. Setzt man aber Regelstrecke G_S und Regler G_R zusammen, so kann es durchaus passieren, dass es in $G_0 = G_S \cdot G_R$ zu einer Pol-Nullstellen-Kürzung kommt. Als Beispiel kann der bereits diskutierte Fall herausgegriffen werden, dass die Regelstrecke differenzierendes Verhalten (Nullstelle in $s = 0$), der Regler hingegen integrierendes Verhalten (Pol bei $s = 0$) aufweist.

Welche Auswirkungen haben solche Pol-Nullstellen-Kürzungen auf die Stabilitätsprüfung? Hier wird repräsentativ der Fall herausgegriffen, dass ein solcher Anteil $A(s)$ im Nenner der Übertragungsfunktion der Regelstrecke und im Zähler der Übertragungsfunktion des Reglers auftritt, sodass

$$G_S(s) = \frac{Z_S(s)}{N_S(s) \cdot A(s)}, \quad G_R(s) = \frac{Z_R(s) \cdot A(s)}{N_R(s)} \quad (9.44)$$

gilt. Berechnet man den aufgeschnittenen Regelkreis G_0 und kürzt diesen, so erhält man

$$G_0(s) = G_S(s) \cdot G_R(s) = \frac{Z_S(s) \cdot Z_R(s) \cdot A(s)}{N_S(s) \cdot N_R(s) \cdot A(s)} = \frac{Z_S(s) \cdot Z_R(s)}{N_S(s) \cdot N_R(s)} \quad (9.45)$$

Damit führt die Berechnung $1 + G_0 = 0$ auf den bekannten Ausdruck

$$Z_S(s) \cdot Z_R(s) + N_S(s) \cdot N_R(s) = 0, \quad (9.46)$$

worin $A(s)$ nicht mehr auftaucht.

Es scheint so, als hätte der gekürzte Anteil $A(s)$ keinen Einfluss auf die Stabilität. Berechnet man allerdings die Übertragungsfunktion von der Störung z am Eingang der Regelstrecke auf die Ausgangsgröße y , so gewinnt man

$$\begin{aligned} \frac{Y(s)}{Z(s)} &= \frac{G_S(s)}{1 + G_0(s)} = \frac{\frac{Z_S(s)}{N_S(s) \cdot A(s)}}{\frac{Z_S(s) \cdot Z_R(s) + N_S(s) \cdot N_R(s)}{N_S(s) \cdot N_R(s)}} \\ &= \frac{Z_S(s) \cdot N_R(s)}{A(s) \cdot (Z_S(s) \cdot Z_R(s) + N_S(s) \cdot N_R(s))}. \end{aligned} \quad (9.47)$$

Hier kürzt sich der Term $A(s)$ nicht heraus und gelangt über den Nenner von $G_S(s)$ in den Nenner der Gesamtübertragungsfunktion. Dies Berechnung in Gl.(9.47) scheint im Widerspruch zu Gl.(9.46) zu stehen.

Die Folge ist, dass die am Anfang dieses Kapitels getroffene Aussage, dass es unerheblich sei, welche Übertragungsfunktion am geschlossenen Regelkreis untersucht würde, Einschränkungen unterliegt. Das Argument, dass alle Übertragungsfunktionen denselben Nenner $1 + G_0(s)$ besitzen, gilt nämlich nur für den Fall einer minimalen Realisierung, wo es zu keinen Pol-Nullstellen-Kürzungen kommt. Der Unterschied zwischen den verschiedenen Nennern ist dabei die gekürzten Pol- und Nullstellen $A(s)$.

Solang man sicher sein kann, dass $A(s)$ keine stabilitätsgefährdenden Nullstellen mit nichtnegativem Realteil hat, führt eine Stabilitätsprüfung, die einen gekürzten aufgeschnittenen Regelkreis G_0 nutzt, dennoch zu richtigen Aussagen.

Vorsicht ist allerdings dann geboten, wenn $A(s)$ eine oder mehrere Nullstellen mit nichtnegativem Realteil hat. Aus Gl.(9.47) ist zu erkennen, dass der zugehörige Regelkreis dann auch mit beliebiger Reglerübertragungsfunktion nicht stabil arbeiten kann.

Kürzungen von Pol- und Nullstellen

Die Pole und Nullstellen einer Übertragungsfunktion dürfen nur dann gegeneinander gekürzt werden, wenn sie in der linken offenen Halbebene liegen.

Eine zentrale Schlussfolgerung hieraus ist, dass es nicht möglich ist, eine instabile Strecke durch eine Reihenschaltung mit einer nicht-minimalphasigen Steuerung, welche mit ihren Nullstellen die instabilen Polstellen der Strecke herauskürzt, zu stabilisieren. Wäre dies möglich, wäre das Skript an dieser Stelle zu Ende und die Regelungstechnik ein im Kern gelöstes Problem.

Dass es nicht möglich ist, instabile Polstellen zu kürzen, wird durch ein zweites weniger mathematisches Argument gestützt. Hierzu muss man sich bewusst machen, dass es sich bei den untersuchten Systemen stets um Modelle realer Prozesse handelt. Modelle geben immer nur einen Teil des realen Systemverhaltens wieder und stellen eine Näherung des echten Systemverhaltens dar.

Beispielhaft wird das Streckenmodell vielleicht $G_S(s) = 1/(-s + 2)$ betrachtet, welches mathematisch scheinbar durch eine Pol-Nullstellen-Kürzung mit dem Regler $G_R(s) = (-s + 2)$ gekürzt werden kann. Tatsächlich wird die reale Regelstrecke sich anders verhalten: Selbst bei genauerster Modellierung muss davon ausgegangen werden, dass durch kleinste Fehler, nicht

exakt bestimmte Parameter oder Verschleiß sich die reale Regelstrecke wie $G(s) = 1/(-s+2+\epsilon)$ verhält. Sie beinhaltet also einen (möglicherweise sehr kleinen) Modellfehler. Bereits für beliebige $\epsilon \neq 0$ kommt es zu keiner Pol-Nullstellen-Kürzung und die Reihenschaltung besitzt neben der bereits existenten instabilen Polstelle nun auch eine ungünstige nicht-minimalphasige Nullstelle durch den Regler.

Wird die Kürzung hingegen im stabilen Bereich vorgenommen, entsteht aufgrund des Modellfehlers aus einem aus der Kürzung hervorgehendem P-Element lediglich ein PDT₁ oder PPT₁ mit sehr nah aneinanderliegenden Zeitkonstanten. Eine Betrachtung von Tab. 7-2 zeigt, dass diese Elemente kaum Auswirkungen auf Phasen- und Amplitudengang haben, sodass eine Kürzung als Modellvereinfachung gerechtfertigt erscheint.

9.4.2 Unstetige Polstellen

Ein anders gelagerter Sonderfall kann im Falle sprungfähiger Systeme entstehen und zu Polstellen führen, die die linke komplexe Halbebene nicht stetig über den Stabilitätsrand, sondern unstetig im Unendlichen verlassen. Da die meisten Regelstrecken und Regelkreise nicht sprungfähig sind, sondern einen relativen Grad von eins oder höher besitzen, tritt dieser Fall praktisch sehr selten auf.

Schließt man den Regelkreis mit einem sprungfähigen G_0 , so erhält man für das charakteristische Polynom $p(s)$ des geschlossenen Regelkreises

$$G_0(s) = \frac{b_0 + \dots + b_n s^n}{a_0 + \dots + a_n s^n} \Rightarrow p(s) = (a_0 + b_0) + \dots + (a_n + b_n) s^n \stackrel{!}{=} 0. \quad (9.48)$$

Für nicht-sprungfähige Systeme der Ordnung n ist $b_n = 0$ und somit der führende Koeffizient des aufgeschnittenen und geschlossenen Regelkreises gleich. Daher besitzt auch der geschlossene Regelkreis Ordnung n und somit n Polstellen. Falls aber für ein sprungfähiges System (zufälligerweise) $a_n = -b_n$ gelten sollte, so besitzt der geschlossene Regelkreis eine geringere Ordnung als der aufgeschnittene Regelkreis. Für diese spezielle Konstellation ist ein Pol des Regelkreises „verschwunden“. Der Fall $a_n = -b_n$ bzw. $b_n/a_n = -1$ entspricht wegen

$$\lim_{\omega \rightarrow \infty} G_0(j\omega) = \frac{b_n}{a_n} \stackrel{!}{=} -1 \quad (9.49)$$

genau dem Fall, dass der Grenzwert der Ortskurve von G_0 für $\omega \rightarrow \infty$ im kritischen Punkt -1 liegt.

Aufgrund der früheren Ausführungen zum Stabilitätsrand entspricht dies $\omega_\pi = \infty$ – also einem „Sprung“ der Polstellen vom stabilen in den instabilen Bereich im Unendlichen. Im Moment des „Sprungs“ geht dabei kurzzeitig eine Polstelle „verloren“ (Singularität). Die Ausführungen zum Stabilitätsrand und zur Stetigkeit der Polstellen gelten folglich nur für den Fall, dass der höchste Koeffizient des charakteristischen Polynoms nicht verschwindet.

Ein kurzes Beispiel mit einem Allpass soll dieses (vorwiegend mathematische) Phänomen erläutern. Für einen Allpass 1. Ordnung als Regelstrecke ergibt sich im geschlossenen Regelkreis

$$G_S = \frac{K(1 - sT)}{1 + sT} \quad K, T > 0 \quad \Rightarrow \quad p(s) = 1 + K + s(T - KT) . \quad (9.50)$$

Das Hurwitz-Kriterium liefert als Stabilitätsbereich $-1 < 0 < K < 1$. Für $K = 1$ verschwindet dabei der höchste Koeffizient $a_1 = 0$. Das Nyquist-Kriterium mit der Ortskurve aus Tab. 7-3, welche einen Halbkreis mit Radius K beschreibt, ergibt, dass für $K < 1$ die Ortskurve von G_S nur zur Rechten der -1 liegt, wodurch der geschlossene Kreis stabil ist. Für $K = 1$ endet die Ortskurve mit $\omega \rightarrow \infty$ im kritischen Punkt -1 und für $K > 1$ entsteht eine Umdrehung $m = 1$ und somit $n = 1$ und ein instabiler Regelkreis.

Trotzdem ist der Regelkreis für $K = 1$ nicht am Stabilitätsrand. Für die Polstelle des geschlossenen Regelkreises berechnet sich nämlich

$$\lambda = (1 + K)/(T - KT) . \quad (9.51)$$

Dieser Ausdruck ist für $-1 < K < 1$ negativ und ansonsten positiv, für den Fall $K = 1$ allerdings gar nicht definiert.

Für $K = 1$ ist das charakteristische Polynom $p(s) = 2$ ohne eine Polstelle. Die Polstelle λ läuft für steigende K auf der reellen Achse nach $-\infty$ um dann für $K = 1$ zu verschwinden und für $K > 1$ bei $+\infty$ im instabilen Bereich erneut zu erscheinen.

Der Sonderfall $a_n = 0$ (Hurwitz/Routh) bzw. $\lim_{\omega \rightarrow \infty} G(j\omega) = -1$ wird in der Praxis selten vorkommen. Wenn er auftritt, kann er bei falscher Behandlung zu Fehlschlüssen führen, da das System in diesem Fall keine Pole auf der imaginären Achse besitzt.

10 Einführung in den Reglerentwurf

10.1 Ziele und Lösungsansätze

10.1.1 Motivation

In Kapitel 1 wurde gezeigt, dass die Verbindung von Regelstrecke und Regler zu einem Gebilde mit in sich geschlossenem Wirkungsablauf, dem Regelkreis, führt. Aufbauend auf den Prinzipien der Modellbildung und linearen Ersatzsystemen in Kapitel 2 wurde in Kapitel 4 gezeigt, wie die Null- und vor allem Polstellen eines Systems dessen dynamische Eigenschaften bestimmen. Weiter wurde in Kapitel 6 hergeleitet, dass der geschlossene Regelkreis andere Polstellen als die Regelstrecke besitzt und damit auch abweichende dynamische Eigenschaften aufweisen wird. Wie sich die zentrale Eigenschaft der Stabilität nachweisen lässt, wurde in Kapitel 9 behandelt.

Zusammenfassend entsteht durch den Einsatz einer Regelung also ein neues System, welches andere dynamische Eigenschaften als die Regelstrecke besitzt. Wenn man, wie es häufig der Fall ist, davon ausgehen muss, dass die Regelstrecke hinsichtlich ihrer dynamischen Eigenschaften vorgegeben ist und nicht verändert werden kann, so entsteht für die Regelungstechnik die Aufgabe, eine Regeleinrichtung zu konzipieren, die zusammen mit der Regelstrecke ein neues System mit den gewünschten Eigenschaften ergibt.

Ziel des Reglerentwurfs

Gegeben ist ein dynamisches System (Regelstrecke) mit unzureichenden dynamischen Eigenschaften. Ziel des Reglerentwurfs ist die Schaffung eines anderen dynamischen Systems (Regler), welches zusammen mit der Regelstrecke im geschlossenen Wirkungsablauf neue, gewünschte Eigenschaften aufweist.

Von einem Regelungssystem erwartet man, dass es stabil ist und die Regelgröße bei Störungen wenig und nur für kurze Zeit von der Führungsgröße abweicht sowie Änderungen der Führungsgröße mit geringen Abweichungen folgt. Ob der Regler diese Ziele erreicht, hängt von der richtigen Wahl sowohl des Reglertyps als auch der Reglerparameter ab, wie das folgende Beispiel erläutern soll.

Als Regelungsaufgabe (siehe Bild 10-1) soll die Ausgangsgröße einer Regelstrecke mit Übertragungsfunktion

$$G_S = \frac{1}{(1 + sT)^3} \quad (10.1)$$

trotz einwirkender Störungen konstant gehalten werden. Die Störgröße z ändert sich sprungförmig.

Das Streckenmodell entspricht dabei einem Verzögerungsglied dritter Ordnung (PT_3), dessen statische Verstärkung durch geeignete Normierung zu eins gesetzt wurde. Als Regler G_R werden P-, I-, PI- und PID-Regler in Betracht gezogen. Bild 10-2 zeigt die Störübergangsfunktionen des Regel-

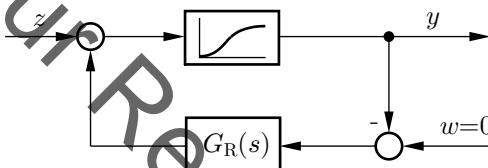


Bild 10-1: Regelkreis mit verzögernder Regelstrecke

kreises für verschiedene Regler mit verschiedener Reglerverstärkung. Dabei sind zuoberst die Verläufe für einen P-Regler $G_R(s) = K_R$ dargestellt.

Man erkennt, dass mit wachsendem K_R die Störung schneller und mit geringerem bleibendem Fehler ausgeglichen wird, aber auch, dass der Regelkreis für $K_R = 8$ Dauerschwingungen und für noch größere K_R aufklingende Schwingungen ausführt.

Dieses Resultat entspricht denen aus der Stabilitätsprüfung in Kapitel 9. Obgleich das statische Verhalten für $t \rightarrow \infty$ bei Erhöhung von K_R verbessert wird, da der Fehler sinkt, verschlechtert sich das dynamische Verhalten durch die auftretenden Schwingungen.

Es ist bereits bekannt, dass bleibende Regelabweichungen mit I-Reglern vermieden werden können. Bild 10-2 zeigt in der mittleren Zeile, wie sich der Regelkreis mit einem solchen Regler $G_R(s) = K_I/s$ verhält. Man erkennt, dass die Regelvorgänge wesentlich langsamer ablaufen als bei der Regelung mit P-Regler und dass auch der Regelkreis mit I-Regler bei größeren Werten von K_I zu Schwingungen neigt.

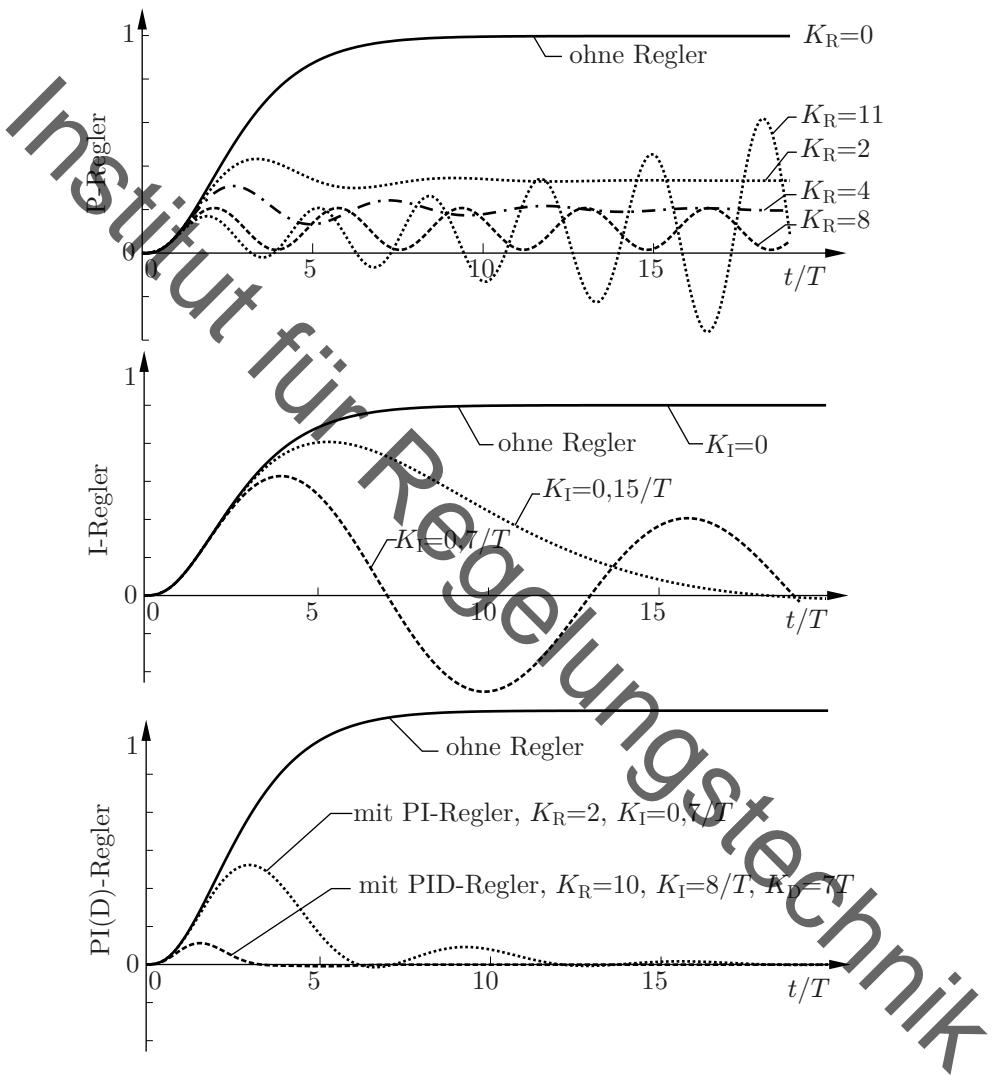


Bild 10-2: Störübergangsfunktionen des Regelkreises nach Bild 10-1 für P-, I-, PI- und PID-Regler

Dass der PI-Regler in einem gewissen Maße die positiven Eigenschaften des P- und des I-Reglers vereinigt, zeigt Bild 10-2 in der untersten Zeile mit der Störübergangsfunktion des Regelkreises mit einem PI-Regler mit $K_R = 2$ und $K_I = 0,7/T$ entsprechend einer Nachstellzeit $T_n = 2,9T$. Es tritt keine bleibende Regelabweichung auf, aber der Regelkreis reagiert im Vergleich zu einem reinen I-Regler zu Beginn wesentlich schneller. Eine weitere wesentliche Verbesserung lässt sich, wie zu sehen ist, mit einem PID-Regler erzielen. Durch den zusätzlichen differenzierenden Anteil im Regler können dabei wesentlich größere K_R -Werte eingestellt werden können, ohne dass die Regelgröße übermäßig schwingt.

10.1.2 Gütemaße und Kennwerte

Das Regelverhalten wurde im obigen Beispiel auf Basis der Zeitverläufe des geschlossenen Regelkreises bewertet. Die Bewertung erfolgte dabei aufgrund recht vager sprachlicher Beschreibungen wie „übermäßig schwingt“ oder „reagiert wesentlich schneller“. Um diese Bewertungen exakter zu fassen und objektiv vergleichbar zu machen, werden zur Beurteilung der Güte einer Regelung geeignet definierte Gütemaße und Kennwerte genutzt. Gütemaße können einmal als Leitlinien für die Entwurfsarbeit, zum anderen aber auch zur Definition von Forderungen, als Grundlage für Abnahmeverweise, Gewährleistungen, Regressansprüche und vieles andere mehr benutzt werden.

Sehr einfache Kennwerte können aus den Sprungantworten des geschlossenen Regelkreises (Bild 10-3) abgeleitet werden.

Kennwerte für die Güte einer Regelung

Die *Anschwingzeit* T_{an} ist die Zeit, die beginnt, wenn die Regelgröße eine vereinbarte Einschwingtoleranz verlässt, und die endet, wenn sie in diesen Bereich zum ersten Mal wieder eintritt.

Die *Einschwingzeit* T_{ein} ist die Zeit, die beginnt, wenn die Regelgröße eine vereinbarte Einschwingtoleranz verlässt, und die endet, wenn sie in diesen Bereich wieder eintritt und dauerhaft darin verbleibt.

Die *bleibende Regelabweichung* e_∞ ist (vgl. Gl.(7.2)) die nach Abklingen von Einschwingvorgängen verbleibende Differenz von Führungssgröße und

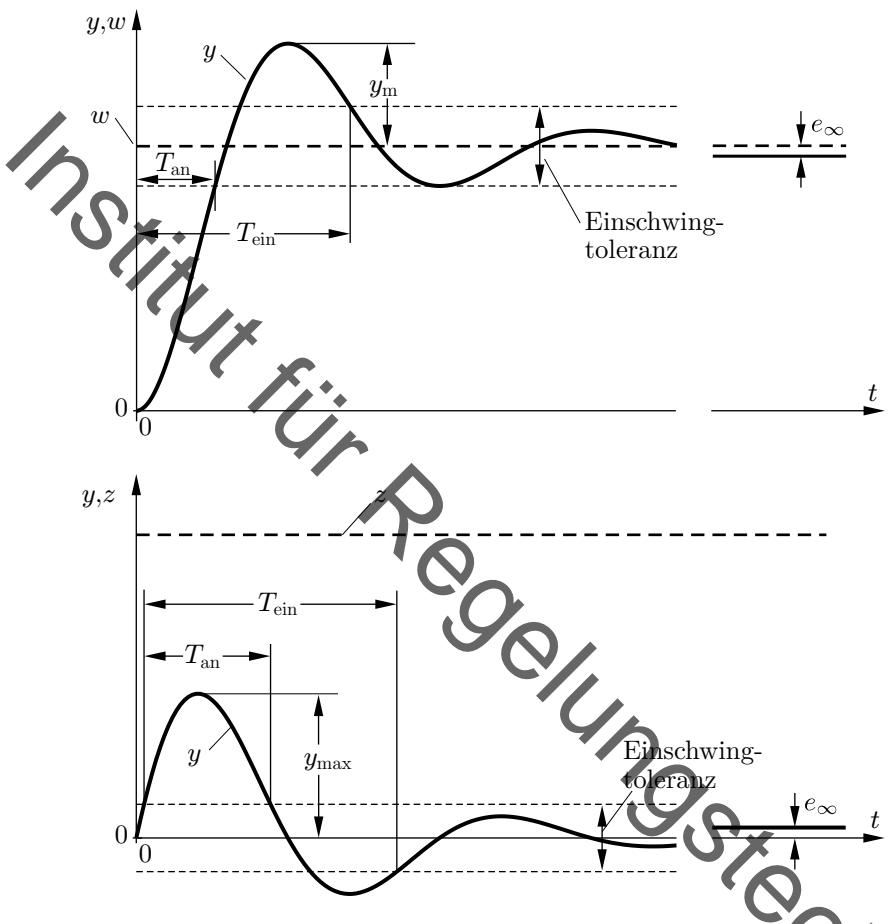


Bild 10-3: Sprungantworten bei Führung (oben) und Störung (unten)

Regelgröße.

Die *Überschwingweite* y_m ist der größte Wert, um den die Regelgröße bei Sprüngen der Führungsgröße über ihre Führungsgröße hinaus über- schwingt, wobei oft mit einer auf diese Führungsgröße bezogenen Über- schwingweite gearbeitet wird.

Bei Antworten auf Störgrößensprünge wird statt der Überschwingweite die maximale Regelabweichung y_{\max} benutzt.

10.1.3 Ansätze des Reglerentwurfs

Eine einleuchtende Forderung für einen Reglerentwurf wäre T_{an} , T_{ein} , y_{\max} und e_{∞} zu minimieren. Hierzu gibt es grundsätzlich vor allem zwei Ansätze: Der erste Ansatz nutzt die Zusammenhänge zwischen den Regelzielen im Zeitbereich wie in Bild 10-3 und bestimmten Beschreibungsformen von LTI-Systemen. So lassen sich Überschwingen aber auch die Einschwingzeit mit der Lage der Pol- und Nullstellen in der komplexen Ebene assoziieren und der Reglerentwurf die Polstellen des geschlossenen Regelkreises in entsprechende erwünschte Bereiche der s -Ebene platzieren. Der zweite Ansatz nutzt den Frequenzbereich, um Anforderungen an den geschlossenen Regelkreis indirekt zu adressieren, indem er diese zu Anforderungen an den aufgeschnittenen Regelkreis umformuliert. Diese beiden Entwurfsmethoden werden in Kapitel 11 anhand der bekanntesten Vertreter vorgestellt.

Entwurf am offenen oder geschlossenen Regelkreis

Die Regelziele für den geschlossenen Regelkreis können wahlweise direkt am geschlossenen Regelkreis untersucht werden oder aber auf Regelziele für den aufgeschnittenen Regelkreis übertragen werden.

Neben dieser Unterscheidung des Vorgehens beim Reglerentwurf stehen noch zwei grundsätzlich unterschiedliche Reglerstrukturen für den Reglerentwurf zur Verfügung.

Die bisher vorgestellten PID-Regler sind dabei vom Typ der Ausgangsrückführung. Darunter ist zu verstehen, dass der Regler als Eingangssignal den Ausgang der Regelstrecke \mathbf{y} (bzw. die gemessene Ausgangsgröße oder die Regelabweichung) erhält. Der Regler berechnet dann auf Basis der Ausgangsgröße \mathbf{y} und des vorgegebenen Sollwerts \mathbf{w} für die Ausgangsgröße den notwendigen Stelleingriff \mathbf{u} .

Im Ausgangssignal sind (eine minimale Realisierung vorausgesetzt) alle dynamischen Informationen des Systems enthalten. Diese lassen sich aber einfacher durch den Zustandsvektor \mathbf{x} beschreiben. Dieser enthält neben den Ausgangsgrößen \mathbf{y} auch Informationen über die Ableitung der Ausgangsgrößen. So ergibt sich bei einem reinen Verzögerungsglied zweiter Ordnung

im Zustandsraum

$$\dot{\boldsymbol{x}} = \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & -2D\omega_0 \end{bmatrix} \boldsymbol{x} + \begin{bmatrix} 0 \\ K\omega_0^2 \end{bmatrix} u \quad , \quad y = x_1 \quad , \quad (10.2)$$

womit $y = x_1$ aber auch $\dot{y} = x_2$ gilt. Folglich enthält der Zustandsvektor \boldsymbol{x} nicht nur die Ausgangsgröße, sondern auch deren Ableitung.

Die Reglerentwürfe in Bild 10-2 haben gezeigt, dass differenzierende Anteile wie bei einem PD- oder PID-Regler das Regelungsergebnis deutlich verbessern können. Daher verspricht eine Regelung, die die Ableitungen in Form des Zustands zur Regelung verwenden, ebenfalls eine verbesserte Regelgüte.

Ausgangsrückführung oder Zustandsrückführung

Wird bei einer Regelung die Ausgangsgröße y zurückgeführt, so spricht man von einer *Ausgangsrückführung*.

Wird bei einer Regelung die Zustandsgröße \boldsymbol{x} zurückgeführt, so spricht man von einer *Zustandsrückführung* oder einem *Zustandsregler*.

Beide Reglerstrukturen sind in Bild 10-4 im linearen Zustandsraum gegenübergestellt. In der in diesem Abschnitt behandelten Form benötigt dabei der Zustandsregler keinen Sollwert. Stattdessen wird implizit eine Festwertregelung angenommen, die versucht, den Zustand nach $\mathbf{0}$ zu überführen. Zustandsregler können mit den Methoden aus Abschnitt 12.2 auch für Folgeregelungen dienen.

Rein physikalisch lässt sich die Zustandsgröße eigentlich nicht zurückführen, da die Zustände im Allgemeinen nicht messbar sind und nur die tatsächlichen Ausgangsgrößen der Regelstrecke deshalb für eine Regelung genutzt werden können. Es ist jedoch möglich, Zustände zu beobachten oder zu schätzen, wie in Abschnitt 11.4 noch erläutert wird. Deshalb kann für den Reglerentwurf zunächst theoretisch angenommen werden, dass die Zustände zur Verfügung stehen.

Keiner der erwähnten Ansätze kann alle formulierten Kenngrößen zur Beurteilung der Regelgüte beliebig klein werden lassen. Vielfach hat die Verringerung der einen eine Vergrößerung der anderen zur Folge. Dies illustriert Bild 10-2 am Beispiel des I-Reglers: Eine Erhöhung der Reglerverstärkung führt hier zu einer geringeren Anschwingzeit und einer geringeren maximalen Regelabweichung, erhöht aber die Einschwingzeit.

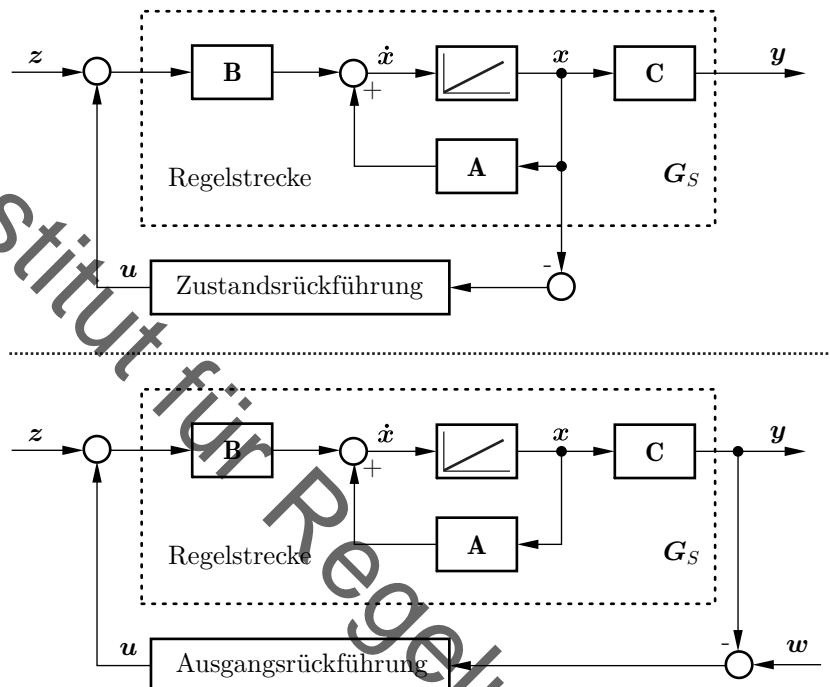


Bild 10-4: Wirkungsplan für ein lineares System mit Zustandsrückführung und Ausgangsrückführung

Meist werden die an ein Regelungssystem zu stellenden Forderungen also nur unvollständig erfüllt werden können. Man wird daher beim Entwurf solcher Systeme versuchen, die als besonders wichtig angesehenen Forderungen so weit wie möglich zu erfüllen und in Kauf nehmen, dass als weniger wichtig eingestufte unberücksichtigt bleiben.

So ist es vielfach sinnvoll, zwischen Festwertregelungen und Folgeregelungen zu unterscheiden und damit bereits beim Entwurf Prioritäten festzulegen. Beide Ziele lassen sich aber auch gleichzeitig durch einen komplexeren Aufbau des Regelungssystems wie einer Kombination von Regelung und Steuerung umsetzen (siehe Abschnitt 12.2). Oft nutzt man auch besser geeignete Kennwerte wie die sogenannten Integralkriterien, die die Güte einer Rege-

lung mit einem einzigen Zahlenwert erfassen. Eine Regelung gilt dann als optimal im Sinne eines dieser Gütemaße, wenn sie so ausgelegt bzw. eingestellt ist, dass das betreffende Gütemaß einen minimalen Wert annimmt. Auf diesen Ideen beruhen die Ansätze der optimale Regelung, die in Kapitel 18 vorgestellt werden.

10.2 Statischer Reglerentwurf

Bevor auf die verschiedenen Ansätze zum Reglerentwurf in den folgenden Kapiteln und Abschnitten eingegangen wird, lohnt es sich, einen Schritt zurückzutreten und zunächst auf eine besonders einfache Form der Reglerauslegung einzugehen. Diese besteht darin, von den aufgeführten Kennwerten lediglich die bleibende Regelabweichung zu adressieren. Dieser statische Reglerentwurf lässt nicht den transienten Übergang, sondern nur das Verhalten für $t \rightarrow \infty$ in die Bewertung der Regelgüte einfließen.

Diese Form der Reglerauslegung ist für praktische Zwecke von untergeordneter Bedeutung, besitzt aber einen didaktischen Wert, da wichtige Aussagen über das Vorzeichen einer erfolgreichen Regelung hierüber getroffen werden können.

Für die Betrachtung des statischen Verhaltens $t \rightarrow \infty$ wird davon ausgegangen, dass von allen betrachteten linearen Teilsystemen die Grenzwerte $t \rightarrow \infty$ beziehungsweise $s \rightarrow 0$ der Ein- und Ausgangssignale existieren und die Grenzwertsätze somit anwendbar sind. Im statischen Fall $t \rightarrow \infty$ können dann alle Signale und Systeme durch die zugehörigen statischen Endwerte ersetzt werden.

Bei Betrachtung des Störverhaltens eines Standard-Regelkreises wie in Bild 10-5 werden aus den allgemeinen Strecken- und Reglerübertragungsfunktionen $G_S(s)$ und $G_R(s)$ die statischen Verstärkungen K_S und K_R ; die Signale $z_1(t)$, $z_2(t)$, und $y(t)$ werden durch die Endwerte $z_{1,\infty}$, $z_{2,\infty}$, und y_∞ ersetzt. Die Betrachtung von Störungen z_1 und z_2 ermöglicht dabei die Behandlung des allgemeinen Falls mit Eingangsstörungen z_1 und Ausgangsstörungen z_2 .

Um die durch Einsatz einer Regelung erzielte Verbesserung im statischen Verhalten eines Systems zu quantifizieren, ist es sinnvoll, die Auswirkungen von Störungen auf die Regelgröße im Fall ohne Regler und im Fall mit Regler zueinander ins Verhältnis zu setzen. Wird kein Regler eingesetzt ($K_R = 0$), so erhält man $y_\infty^{0R} = z_{2,\infty} + K_S z_{1,\infty}$. Über die Rechenregeln

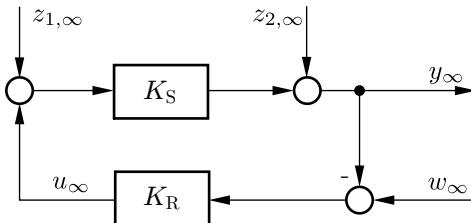


Bild 10-5: Standard-Regelkreis in einer statischer Betrachtungsweise

für Rückkopplungen erhält man bei Einsatz eines Reglers mit $w_\infty = 0$ den Zusammenhang

$$y_\infty^{\text{mR}} = \frac{z_{2,\infty} + K_S z_{1,\infty}}{1 + K_S K_R}. \quad (10.3)$$

Das Verhältnis der Auswirkung der Störung auf die Ausgangsgröße ohne Regler y_∞^{R} zur Auswirkung mit Regler y_∞^{mR} ist dann ein (prozentualer) Wert, der *Regelfaktor* genannt wird.

Regelfaktor

Der *Regelfaktor* charakterisiert die statische Leistungsfähigkeit eines Reglers gegenüber Störungen und ist als das Verhältnis

$$R = \frac{y_\infty^{\text{mR}}}{y_\infty^{\text{R}}} = \frac{1}{1 + K_S K_R} = \frac{1}{1 + G_0(s=0)} \quad (10.4)$$

definiert. Sein konkreter Wert ist im Fall stabiler linearer Systeme unabhängig vom statischen Endwert der Störungen.

Der Regelfaktor ist bei sinnvollen Regelungen kleiner als eins. Für technisch sinnvolle Regelungen wird fast immer ein Regelfaktor gefordert, der kleiner ist als etwa 0,2, d. h. eine Regelung sollte die Wirkung von Störgrößen auf etwa 20% des Wertes ohne Regelung vermindern können, um den notwendigen Aufwand zu rechtfertigen.

Um einen Regelfaktor $R < 1$ zu erreichen, muss offenbar die statische Verstärkung K_R des Reglers dasselbe Vorzeichen wie die statische Verstärkung K_S der Regelstrecke aufweisen.

Vorzeichen einer Regelung

Bei einer stabilen Regelstrecke muss für eine sinnvolle Regelung das Vorzeichen der statischen Verstärkung des Reglers so gewählt werden, dass der Regler als Rückkopplung arbeitet und Regelabweichungen mit insgesamt negativem Vorzeichen auf die Regelgrößen zurückgeführt werden.

Besitzen K_S und K_R unterschiedliche Vorzeichen, so heben diese das im Regelkreis vorgesehene Minuszeichen der Rückführung auf und es entsteht eine Mitkopplung, bei der Regelabweichungen positiv auf die Regelgrößen zurückgeführt werden. Dies führt bei stabilen Regelstrecken in jedem Fall zu einer Verschlechterung des Systemverhaltens durch die Regelung und kann diese auch leicht destabilisieren.

Die erste Bedingung nach Hurwitz/Routh bestätigt dies: Sie fordert, dass alle Vorzeichen des charakteristischen Polynoms gleich sein müssen. Besitzen K_S und K_R unterschiedliche Vorzeichen, führt dies in den allermeisten Fällen für hinreichend große K_R zu einer Verletzung dieser Bedingung, wie in Abschnitt 9.2.2 bereits demonstriert wurde. Eine Mitkopplung ist unter Umständen sinnvoll, wenn die Regelstrecke instabil ist, für stabile Regelstrecke aber grundsätzlich unsinnig.

Abseits der passenden Wahl des Vorzeichens kann man aus Gl.(10.4) noch weitere Forderungen an die statische Reglerverstärkung ableiten: Offenbar wird der Regelfaktor umso kleiner, je größer K_R wird. Das bedeutet, dass aggressive Regler mit einer hohen Reglerverstärkung eine bessere Leistungsfähigkeit im Regelkreis erzielen können und K_R aus einer statischen Perspektive heraus möglichst hoch gewählt werden sollte. Um einen optimalen Regelfaktor von $R = 0$ zu erreichen, muss dabei $G_0(s \rightarrow 0) = \infty$ gelten, d. h. der aufgeschnittene Regelkreis muss einen integrierenden Anteil besitzen. Dieses Resultat ist aus den Ausführungen zur stationären Genauigkeit bekannt, da stationär genaue Regelkreise konstante Störgrößen vollständig ausregeln.

Ein auf dem Regelfaktor basierender Reglerentwurf berechnet ausgehend von einem geforderten maximalen Regelfaktor die notwendige minimale statische Verstärkung des Reglers. Teilweise können bei der statischen Betrachtungsweise auch nichtlineare Abhängigkeiten berücksichtigt werden, wenn diese beispielsweise in Form einer Kennlinie oder eines Kennfeldes gegeben sind.

10.3 Abwägungen bei der Reglerverstärkung

10.3.1 Vorteile hoher Verstärkungen

Der Vergleich des statischen Verhaltens von Reglern mit und ohne integrierendem Verhalten fällt offensichtlich zugunsten der Regler mit I-Anteil aus. Das liegt daran, dass die Forderung nach möglichst kleiner bleibender Regelabweichung nur mit sehr großen (statischen) Übertragungsfaktoren des Reglers zu erfüllen ist, welche mit integrierenden Reglern leicht zu realisieren sind. Das ist auch in Bild 10-2 klar zu erkennen.

Prinzipiell ähnliche Ergebnisse erhält man, wenn man möglichst kleine dynamische Abweichungen fordert. So berechnet sich die Übertragungsfunktion von der Führungsgröße auf die Regelabweichung zu

$$\begin{aligned} E(s) &= W(s) - Y(s) = W(s) - \frac{G_0(s)}{1 + G_0(s)} W(s) = \frac{1}{1 + G_0(s)} W(s) \\ \Rightarrow \quad \frac{E(s)}{W(s)} &= \frac{1}{1 + G_0(s)} = S(s) \quad . \end{aligned} \tag{10.5}$$

Gilt nun beispielsweise für die Führungsgröße $w(t) = \cos(\omega t)$, d. h. die Führungsgröße ändert sich sinusförmig und ihr soll möglichst gut gefolgt werden, dann ist der entstehende Fehler für stabile Regelkreise im eingeschwungenen Zustand

$$e(t) = \frac{1}{|1 + G_0(j\omega)|} \cos(\omega t + \varphi(\omega)) \quad . \tag{10.6}$$

Die Amplitude des Folgefehlers ist also – ganz ähnlich wie der Regelfaktor – von der Verstärkung $|1 + G_0(j\omega)|$ abhängig, welche folglich möglichst groß sein sollte. Dies kann nur durch eine möglichst große Verstärkung von $|G_0(j\omega)|$ erreicht werden. Auf den ersten Blick erscheint es also auch im dynamischen Fall so, als wäre eine möglichst hohe Reglerverstärkung die Lösung aller regelungstechnischer Probleme. Der Vergrößerung der Reglerverstärkung sind aber aus verschiedenen Gründen Grenzen gesetzt, die im Folgenden ausgeführt werden.

Wahl der Reglerverstärkung

Die Kunst des Reglerentwurfs besteht darin, die Verstärkung des Reglers dergestalt zu wählen, dass diese in den für die Regelgüte relevanten Bereichen möglichst hoch ist, ohne die Grenzen des systemtheoretisch Erlaubten zu überschreiten.

10.3.2 Nachteile hoher Verstärkungen

Zunächst führt eine große Reglerverstärkung wegen $U(s) = G_R(s)E(s)$ auch zu großen Werten der Stellgröße. Eine Verdopplung der Reglerverstärkung $|G_R(s)|$ bewirkt dabei bei gleichgroßer Regelabweichung eine Verdopplung der Stellgröße. Dasselbe gilt auch für Zustandsrückführungen.

Sehr große Stelleingriffe sind aber normalerweise nur im idealisierten Fall linearer Systeme möglich, wo potentiell unendlich große Stellgrößen denkbar sind. Praktisch gelten diese linearen Systembeschreibungen nur in der Umgebung eines Arbeitspunktes und das tatsächliche System wird immer eine *Stellgrößenbeschränkung* aufweisen.

Stellgrößenbeschränkungen

Aufgrund gerätetechnischer Einschränkungen wie der limitierten Energie, die eine Regeleinrichtung aufbringen kann, ist die Stellgröße in allen praktischen Anwendungen auf den sogenannten Stellbereich $\underline{u} \leq u(t) \leq \bar{u}$ beschränkt und kann die zugehörigen Maximal- und Minimalwerte nicht über- bzw. unterschreiten.

Diese Beschränkung der aufprägbaren Stellgröße $u(t)$ sorgen dafür, dass bei einem Regler, der Stellwerte außerhalb des Stellbereiches anfordert, statt der geforderten Stellwerte die entsprechenden Maximal- und Minimalwerte an die Regelstrecke ausgegeben werden. Dies nennt sich Saturierung.

Dadurch verändern sich die Eigenschaften des Regelkreises, der ein nichtlineares Verhalten erhält. Die in der linearen Analyse scheinbar erreichbare Regelgüte wird nicht erreicht. Stattdessen ergeben sich eine Vielzahl zusätzlicher Probleme, auf deren Analyse unter anderem in Abschnitt 16.5 eingegangen wird. An dieser Stelle sei insbesondere auf die Problematik des Integrator-Windup verwiesen, welches auftritt, wenn Regler mit integrierendem Anteil in ihre Stellgrößenbeschränkungen laufen [2].

Eng verwoben mit den Stellgrößenbeschränkungen sind energetische und wirtschaftliche Überlegungen. Eine gute Umsetzung der Regelungsaufgabe spart üblicherweise Kosten, da weniger Verluste entstehen oder ein System näher an seinem optimalen Betriebspunkt gehalten werden kann. Es werden aber dennoch Kosten erzeugt, in Form von Aufwendungen für die Stelleingriffe $u(t)$.

Energieverbrauch des Aktors

Typischerweise führen große und sich schnell ändernde Stellgrößenänderungen zu einem höheren Energieverbrauch und Verschleiß des Aktors. Daher sollte – auch innerhalb der Stellgrößenbeschränkungen – auf einen maßvollen Gebrauch der Stellgröße geachtet werden.

Ein weiteres zentrales Problem mit sehr hohen Reglerverstärkungen ist das bereits mehrfach angeklungene Thema der Stabilität. Das Nyquist-Kriterium zeigte, dass hohe Reglerverstärkungen bei der Frequenz ω_π eine ursprünglich stabile Regelsstrecke destabilisieren. Somit beschränkt ω_π den Frequenzbereich, in welchem hohe Reglerverstärkungen möglich sind. Folglich kann nach Gl.(10.6) einer Führungsgröße mit Frequenzanteilen größer als ω_π nicht mit einer kleinen Abweichung gefolgt werden. Dies kann durch die Bandbreite des geschlossenen Regelkreises ausgedrückt werden, die beschreibt, in welchem Frequenzbereich sich die wesentlichen Signalanteile von $e(t)$ befinden.

Ein Maß für die Bandbreite ω_g des geschlossenen Regelkreises kann mit der sehr groben Näherung

$$\frac{1}{|1 + G_0(j\omega_g)|} \approx \frac{1}{\sqrt{1^2 + (G_0(j\omega_g))^2}} \stackrel{!}{=} \frac{1}{\sqrt{2}} \quad (10.7)$$

$$\Rightarrow |G_0(j\omega_g)| = 1 \quad \Rightarrow \quad \omega_g \approx \omega_d$$

gegeben werden.

Die Näherung beruht auf der Annahme, dass der Phasenwinkel $\varphi(j\omega_d)$ sich nahe der $\pm 90^\circ$ -Linie befindet. Diese Annahme ist oft näherungsweise erfüllt, wenn der Amplitudengang die 10^0 -Linie mit einer Steigung von -1 schneidet, was – wie in Abschnitt 11.1 gezeigt wird – ohnehin wünschenswert ist.

Aus Gl.(10.7) folgt, dass ein typischer Regelkreis nur bis zu der Durchtrittsfrequenz ω_d Störungen erfolgreich unterdrücken kann. Ein leistungsfähiger

Reglerentwurf muss daher versuchen, ω_d zu erhöhen, beispielsweise durch das Erhöhen der Reglerverstärkung. Die stabilitätsbedingte Beschränkung von ω_d nach oben durch ω_π setzt der Reglerverstärkung aber immer Grenzen. Daher läuft der Entwurf einer Regelung vielfach auf einen Kompromiss hinaus zwischen günstigem Stör- und Führungsverhalten, das mit großen Übertragungsbeiwerten in einem großen Frequenzbereich erreicht wird, und der Forderung nach Stabilität, der i. Allg. durch kleinere Übertragungsfaktoren Rechnung getragen werden kann.

Frequenzabhängige Reglerverstärkung

Leistungsfähige Regler besitzen in einem möglichst großen bei $\omega = 0$ beginnenden Frequenzbereich eine große Verstärkung. Aus Stabilitätsgründen ist dieser Frequenzbereich aber nach oben hin beschränkt, z. B. durch ω_π .

Nicht-minimalphasige Nullstellen verstärken diese Beschränkung der Reglerverstärkung zusätzlich. In Abschnitt 7.5.1 wurde bereits argumentiert, wieso Systeme mit Allpassanteil für große Reglerverstärkungen instabil werden. Dies lässt sich auch durch eine Rechnung untermauern: Gilt für die statische Reglerverstärkung $K_R \rightarrow \infty$, so ist auch $1 \ll |G_0|$. Damit ist der Nenner des geschlossenen Regelkreis $1 + G_0(s) \approx G_0(s)$ und die Polstellen des geschlossenen Regelkreises sind die Nullstellen des aufgeschnittenen Regelkreises.

Polstellen für hohe Reglerverstärkungen

Für eine nach unendlich hin wachsende Verstärkung im aufgeschnittenen Regelkreis $|G_0| \rightarrow \infty$ wandern die Pole des geschlossenen Regelkreises in die Nullstellen des aufgeschnittenen Regelkreises.

Besitzt der aufgeschnittene Regelkreis nicht-minimalphasige Nullstellen, ist der geschlossene Regelkreis daher für hinreichend große Verstärkungen in jedem Fall instabil.

Ein weiteres Problem mit hohen Reglerverstärkungen ist die fehlende Robustheit, d. h. dass die scheinbar erreichbare Regelgüte bei kleinen Abweichungen zwischen Streckenmodell und echter Regelstrecke nicht erhalten bleibt. Dies wurde anhand des Beispiels einer unberücksichtigten Totzeit bereits in Abschnitt 9.3 diskutiert, betrifft aber nicht nur Totzeiten, sondern sämtliche vernachlässigte Dynamik. Jede Modellbildung unterschlägt

Dynamik – durch die Approximation über dominante Pol- und Nullstellen, durch die Linearisierung oder strukturell bei jeder vereinfachenden Modellierungsannahme.

Zur Veranschaulichung mangelnder Robustheit wird die Geschwindigkeitsregelung eines Fahrzeuges betrachtet. Als Stellgröße wird die von der Straße auf das Fahrzeug ausgeübte resultierende Kraft (Vortriebskraft) angenommen. Ein einfaches Modell der Regelstrecke wäre dann

$$G_S(s) = \frac{1}{ms} , \quad (10.8)$$

mit der Fahrzeugmasse m .

Offenbar führt ein einfacher P-Regler $G_R(s) = K_R$ im geschlossenen Regelkreis auf die Führungsübertragungsfunktion

$$T(s) = \frac{G_S G_R}{1 + G_S G_R} = \frac{K_R}{ms + K_R} \quad (10.9)$$

und damit auf ein PT_1 mit der Zeitkonstante m/K_R .

Folglich führen sehr hohe Reglerverstärkungen auf eine schnelle Zeitkonstante im System und damit auf ein besseres Regelverhalten. Stabilität ist für alle K_R gegeben, sodass ein sehr hoher Wert für K_R zunächst sinnvoll erscheint.

Im realen Fahrzeug wird aber nicht (wie bisher vereinfacht angenommen) direkt die Vortriebskraft eingestellt werden können, sondern stattdessen diese Kraft nur mit einer (möglicherweise sehr kleinen) Verzögerung im Antriebsstrang auf die Regelstrecke aufgebracht werden. Nimmt man für diese Verzögerung eine PT_1 -Dynamik an, so erhält man

$$G_S(s) = \frac{1}{ms} \cdot \frac{1}{Ts + 1} \quad (10.10)$$

und als Führungsübertragungsfunktion des geschlossenen Regelkreises

$$T(s) = \frac{K_R}{mTs^2 + ms + K_R} . \quad (10.11)$$

Da T sehr klein ist, spielt der führende Koeffizient mT eigentlich eine untergeordnete Rolle und es ist zu erwarten, dass diese sehr schnelle Dynamik

auch im Sinne einer Approximation durch die dominante Polstelle von G_S in $s = 0$ vernachlässigt werden kann. Durch Vergleich mit einem schwingungsfähigen PT₂ (Tab. 7-2) bestimmen sich aber Kennkreisfrequenz ω_0 und Dämpfungsgrad D zu

$$\omega_0 = \sqrt{\frac{K_R}{mT}}, \quad D = \frac{\sqrt{m}}{2\sqrt{K_R T}} \quad , \quad (10.12)$$

d. h. mit wachsendem Übertragungsfaktor K_R nimmt die Kennkreisfrequenz zu, was erwünscht ist, aber es nimmt damit auch der Dämpfungsgrad ab, und das kann unerwünscht sein. Werte des Dämpfungsgrades unter 0,5 sind etwa in den wenigsten Fällen erwünscht, weil die Ausgleichsvorgänge sehr lange Zeit dauern und es zu starkem Überschwingen kommt.

Dieses Phänomen kann im Frequenzbereich gedeutet werden: das PT₁ besitzt für $\omega \ll 1/T$ einen nahezu konstanten Amplituden- und Phasengang. Daher entspricht es in diesem Frequenzbereich strukturell einem P-Element mit Verstärkung eins und kann für die Modellierung ignoriert werden. Für Frequenzen $\omega > 1/T$ hingegen hat das PT₁ sowohl auf die Amplitude als auch auf die Phase einen erheblichen Einfluss.

Arbeitet der Regelkreis mit einer kleinen Bandbreite, d. h. $\omega_d \ll 1/T$, so tritt dieser Einfluss des PT₁ in einem Frequenzbereich auf, in dem der Betrag von G_0 sehr klein ist und wird daher kaum angeregt. Ist die Bandbreite hingegen groß, so wird der Einfluss des PT₁ kritisch, da er die Phase und Phasenreserve aber auch die Durchtrittsfrequenz ω_d verändert. Das ist in Bild 10-6 visualisiert. Bei kleinen Verstärkungen (links) verhält sich das IT₁ im Wesentlichen wie ein I-Element und die zusätzliche Verzögerung kann vernachlässigt werden. Für hohe Verstärkungen (rechts) hat die Zeitkonstante T hingegen einen entscheidenden Einfluss.

Vernachlässigbare Dynamik

In einem Regelkreis haben alle Anteile mit Frequenzen $\omega < \omega_d$ mit der Durchtrittsfrequenz ω_d einen wesentlichen Einfluss auf die Dynamik. Nur dynamische Anteile mit $\omega \gg \omega_d$ sind vernachlässigbar.

Durch hohe Reglerverstärkungen und steigende ω_d verliert der Regelkreis also Robustheitseigenschaften, da ein immer größerer Frequenzbereich in seinen dynamischen Auswirkungen nicht vernachlässigt werden kann.

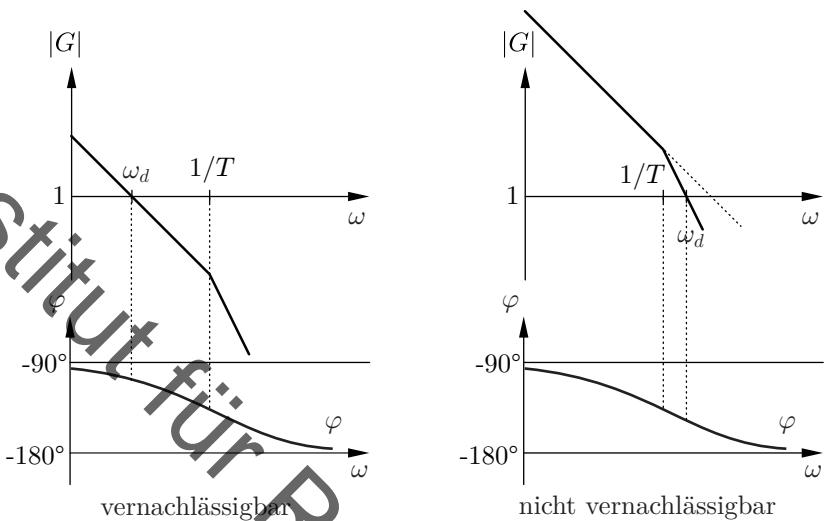


Bild 10-6: IT_1 mit vernachlässigbarer und nicht vernachlässigbarer Zeitkonstante abhängig von der Reglerverstärkung

Mit der Forderung nach Robustheit verwoben ist ein letzter negativer Effekt, den hohe Reglerverstärkungen mit sich bringen. Hierzu wird der Regelkreis mit Messrauschen n in Bild 10-7 betrachtet.

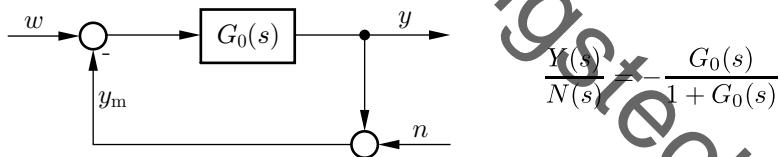


Bild 10-7: Regelkreis mit Messrauschen

Idealerweise soll sich das Messrauschen nicht auf die tatsächliche Regelgröße auswirken. Wird aber eine hohe Reglerverstärkung $|G_0| \gg 1$ gewählt, so gilt mit $w = 0$

$$|Y(s)| = \left| -\frac{G_0(s)}{1 + G_0(s)} \right| \cdot |N(s)| \stackrel{|G_0| \gg 1}{\approx} |N(s)| , \quad (10.13)$$

d. h., dass das Rauschen nicht gedämpft wird, sondern sich mit seiner voll-

ständigen Amplitude in der Regelgröße wiederfindet.

Soll das Messrauschen unterdrückt werden, so muss stattdessen eine möglichst niedrige Reglerverstärkung gewählt werden, weil mit $|G_0| \ll 1$ folgt:

$$|Y(s)| = \left| -\frac{G_0(s)}{1 + G_0(s)} \right| \cdot |N(s)| \stackrel{|G_0| \ll 1}{\approx} 0 . \quad (10.14)$$

Die Forderung zur Unterdrückung von Messrauschen ist also genau komplementär zur Forderung einer geringen Regelabweichung. Die beide Ziele „Rauschunterdrückung“ und „Störgrößenunterdrückung“ können folglich nicht gleichzeitig erreicht werden. Dies spiegelt sich auch darin wider, dass das Störverhalten durch die Sensitivität $S(s)$, das Rauschverhalten aber durch die komplementäre Sensitivität $T(s)$ vorgegeben wird.

Fundamentales Dilemma der Regelungstechnik

Wegen Gl.(6.7)

$$S(s) + T(s) = \frac{1}{1 + G_0} + \frac{G_0}{1 + G_0} = 1 \quad (10.15)$$

können Rausch- und Störgrößenunterdrückung nicht unabhängig voneinander eingestellt werden.

10.4 Einstellregeln

Wie der vorherige Abschnitt 10.3 zeigte, ist viel Geschick dafür nötig, die Reglerverstärkung in den passenden Gebieten groß zu wählen, ohne sie in anderen Gebieten wegen Stabilität, Robustheit und Messrauschen zu stark anzuheben. Diese Aufgabe ist ein kreativer Entwurfsprozess mit vielen Freiheiten in der Wahl passender Regler und Reglerstrukturen und profitiert neben dem Wissen um die systemtheoretischen Zusammenhänge auch von Erfahrung bei der Einstellung von Reglern.

Die Erfahrung hilft dabei vor allem auch deswegen weiter, da viele Regelstrecken – auch in unterschiedlichen Fachbereichen von der Verfahrenstechnik über Elektrotechnik und Produktionstechnik bis hin zur Robotik – sich über strukturell ähnliche Systemmodelle wie Verzögerungsglieder höherer Ordnung beschreiben lassen.

Sogenannte Einstellregeln haben zum Ziel, diese Erfahrungswerte in Kombination mit minimalen Kenntnissen über das zu regelnde Objekt für eine initiale Reglereinstellung nutzbar zu machen.

Die hieraus entstehenden Reglerparametrierungen basieren auf den Ausführungen zur passenden Verstärkungswahl, sind aber dennoch nur als Ausgangspunkt für weitere Anpassungen der Reglerparameter zu verstehen oder dann zu nutzen wenn die betreffende Regelung von so untergeordneter Bedeutung ist, dass der Entwurfsaufwand auf ein Minimum zu beschränken ist.

10.4.1 Einstellung mittels T_u - T_g -Ersatzmodell

Auf Arbeiten von Chien¹, Hrones und Reswick (1952) beruhen Empfehlungen, die von einem T_u - T_g -Modell für die Übergangsfunktion der Regelstrecke ausgehen. Folglich sind diese Empfehlungen auch nur anwendbar, wenn ein T_u - T_g -Ersatzmodell das dynamische und statische Verhalten der Regelstrecke ausreichend genau beschreibt. Sie sind daher insbesondere für nicht-minimalphasige Systeme oder ausgeprägt schwingungsfähige Regelstrecken nicht geeignet.

Auf den ermittelten Kennwerten des T_u - T_g -Modells aus Bild 8-3 basieren Empfehlungen für die Einstellung von Reglern für günstiges Verhalten bei Stör- bzw. Führungsgrößenänderung, wobei Störungen als Störungen am Eingang der Regelstrecke verstanden werden. So werden für $T_u/T_g < 1/3$ als günstige bzw. brauchbare Einstellwerte die in Tab. 10-1 wiedergegebenen Parameterwerte empfohlen. Die Beschränkung auf $T_u/T_g < 1/3$ besagt, dass die mit diesen Formeln ermittelten Werte sich weniger gut für Regelstrecken mit ausgeprägtem Totzeitverhalten eignen.

Wenn man die soeben eingeführten Einstellregeln auf den Regelkreis mit PT₃ nach Bild 10-1 anwendet, so erhält man für die Regelstrecke einen Wert $T_g/T_u \approx 5$. Nach Tab. 10-1 sollte ein P-Regler für einen Regelverlauf mit 20 % Überschwingen demnach auf $K_R \approx 3,5$ eingestellt werden. Die in Bild 10-2 dargestellte Störübergangsfunktion für einen Wert von $K_R = 4$ lässt erwarten, dass diese Einstellung zu brauchbaren Ergebnissen führen wird.

¹Kun Li Chien, amerikanischer Regelungstechniker [6]

Regler	Aperiodischer Regelverlauf		Regelverlauf mit 20% Überschwingen	
	Störung	Führung	Störung	Führung
P K_R	$\frac{0,3 T_g}{K_S T_u}$	$\frac{0,3 T_g}{K_S T_u}$	$\frac{0,7 T_g}{K_S T_u}$	$\frac{0,7 T_g}{K_S T_u}$
	$\frac{0,6 T_g}{K_S T_u}$	$\frac{0,35 T_g}{K_S T_u}$	$\frac{0,7 T_g}{K_S T_u}$	$\frac{0,6 T_g}{K_S T_u}$
	$4 T_u$	$1,2 T_g$	$2,3 T_u$	$1 T_g$
PI K_R T_n	$\frac{0,95 T_g}{K_S T_u}$	$\frac{0,6 T_g}{K_S T_u}$	$\frac{1,2 T_g}{K_S T_u}$	$\frac{0,95 T_g}{K_S T_u}$
	$2,4 T_u$	$1 T_g$	$2 T_u$	$1,35 T_g$
	$0,42 T_u$	$0,5 T_u$	$0,42 T_u$	$0,47 T_u$
PID T_n T_v				

Tabelle 10-1: Einstellwerte für Reglereinstellung nach Sprungantwort der Regelstrecke

10.4.2 Einstellung mittels Schwingversuch

Ein anderer Satz von Empfehlungen basiert auf Arbeiten von Ziegler² und Nichols und hat die Amplitudenreserve und das vereinfachte Nyquist-Kriterium als sein theoretisches Fundament. Die Empfehlungen sind dabei so formuliert, dass das Aufstellen eines Systemmodells nicht notwendig ist, sondern dass die notwendigen Kenntnisse über die Regelstrecke, die die Anforderungen des vereinfachten Nyquist-Kriteriums erfüllen muss, durch einen Versuch beschafft werden können.

Der Regler wird dazu zunächst als P-Regler mit sehr kleiner Verstärkung $K_R \rightarrow 0$ betrieben und der Regelkreis geschlossen. Ausgehend von einem

²John G. Ziegler (1909-1997), amerikanischer Regelungstechniker [64]

stabilen Betrieb der Regelung wird die Verstärkung des Reglers K_R so weit vergrößert, bis der geschlossene Regelkreis den Stabilitätsrand erreicht und Dauerschwingungen ausführt. Von dem dabei erreichten Übertragungsfaktor $K_{R\text{krit}}$ und der Periodendauer der sich ergebenden Schwingung T_{krit} wird indirekt auf die Amplitudenreserve und die Kreisfrequenz ω_π und von diesen nach Tab. 10-2 auf empfehlenswerte Reglereinstellungen geschlossen.

Das Verfahren eignet sich gut für stabile Regelstrecken, die für Versuche zugänglich sind, aber in ihrer Modellierung z. B. wegen unübersichtlichen Mess- und Stellgeräteketten zu aufwendig sind.

Sollte doch ein (vereinfachtes) Modell des Systems vorliegen, können die Einstellregeln dennoch nützlich sein. In dem Fall lassen sich $K_{R\text{krit}}$ und T_{krit} aus einer Analyse des Frequenzganges ohne eigentlichen Betriebsversuch gewinnen. Hierbei lässt sich K_R über die Amplitudenreserve und T_{krit} über die Kreisfrequenz ω_π bestimmen, da ω_π genau der Eigenkreisfrequenz der Pole am Stabilitätsrand entspricht. Bezeichnet G_0 den aufgeschnittenen Regelkreis ohne P-Regler, so erhält man

$$K_{R\text{krit}} = \frac{1}{|G_0(j\omega_\pi)|} \quad , \quad T_{\text{krit}} = \frac{2\pi}{\omega_\pi} . \quad (10.16)$$

Auf den Regelkreis mit PT_3 -Regelstrecke nach Bild 10-1 mit P-Regler angewandt, erhält man aus Bild 10-2 ein $K_{R\text{krit}} = 8$ und mit Tab. 10-2 die Empfehlung $K_R = 4$. Diese Einstellung entspricht dabei aufgrund des Aufbaus des Schwingversuchs genau der Einstellung einer Amplitudenreserve von $A_R = 2$.

Regler	K_R	T_n	T_v
P	$0,5 \cdot K_{R\text{krit}}$	-	-
PI	$0,45 \cdot K_{R\text{krit}}$	$0,85 \cdot T_{\text{krit}}$	-
PID	$0,6 \cdot K_{R\text{krit}}$	$0,5 \cdot T_{\text{krit}}$	$0,12 \cdot T_{\text{krit}}$

Tabelle 10-2: Einstellwerte nach einem Schwingversuch

11 Grundlegende modellbasierte Reglerentwurfsverfahren

Die im vorherigen Kapitel vorgestellte Einstellregel nach Ziegler und Nichols basiert auf der Idee, die Reglereinstellung über Anforderungen an die Amplituden- und Phasenreserve zu bestimmen, hierbei aber mit minimalen Informationen auszukommen. Lässt man diese Einschränkung der minimalen Informationen fallen und geht stattdessen davon aus, dass ein vollständiges Modell der Regelstrecke gegeben ist, so ermöglicht diese Modellbeschreibung einen systematischen und leistungsstarken Reglerentwurf.

Die hier vorgestellten Verfahren in Abschnitt 11.1 und Abschnitt 11.2 verfolgen dabei den Ansatz, den Entwurf über den aufgeschnittenen Regelkreis vorzunehmen. Die nachfolgenden Abschnitte 11.3 und 11.4 behandeln hingegen Verfahren, die den Entwurf am geschlossenen Regelkreis vollziehen.

11.1 Frequenzkennlinienverfahren

11.1.1 Grundidee

Der Reglerentwurf nach dem Frequenzkennlinienverfahren basiert auf einem Modell der Regelstrecke in Form des Frequenzgangs, wobei sich für die Durchführung eine Darstellung im Bode-Diagramm in besonderer Weise eignet. Zudem setzt es für einen handhabbaren Entwurf die Anwendbarkeit des vereinfachten Nyquist-Kriteriums zur sinnvollen Definition von Amplituden- und Phasenreserve voraus und basiert auf den folgenden drei Grundüberlegungen aus Abschnitt 10.3:

- Um Störungen gut zu unterdrücken und auf Sollwertänderungen schnell zu reagieren, sollte $|S| = \frac{1}{|1+G_0|}$ möglichst klein und $|G_0|$ daher möglichst groß sein.
- Für Stabilität und gute Dämpfungseigenschaften sollten Amplituden- und Phasenreserve hinreichend groß sein.
- Um unmodellierte Dynamik nicht anzuregen und Messrauschen nicht zu verstärken, sollte $|T| = \frac{|G_0|}{|1+G_0|}$ möglichst klein und $|G_0|$ daher möglichst klein sein.

Alle drei Anforderungen sind in sich widersprüchlich, da sie unterschiedliche Anforderungen an die Verstärkung des aufgeschnittenen Regelkreises formulieren. Vorteilhaft ist aber, dass – obgleich der Reglerentwurf ein gewünschtes Verhalten im geschlossenen Regelkreis erreichen soll – alle formulierten Bedingungen Anforderungen an den Frequenzgang des aufgeschnittenen Regelkreises stellen.

Hierdurch wird es möglich, den Reglerentwurf vollständig auf gewünschten Eigenschaften des aufgeschnittenen Regelkreises zurückzuführen. Das ist hilfreich, da der Zusammenhang zwischen dem zu entwerfenden Regler G_R und dem aufgeschnittenen Regelkreis durch $G_0 = G_S G_R$ sehr klar ist und insbesondere im Bode-Diagramm durch eine Addition leicht umgesetzt werden kann. Der Zusammenhang zwischen Regler G_R und dem geschlossenen Regelkreis hingegen ist durch den Bruch $1/(1 + G_0)$ sehr nichtlinear und weniger leicht zu handhaben.

Das Frequenzkennlinienverfahren überführt die drei aufgeführten Grundüberlegungen in Wunschbereiche, die der Frequenzgang von G_0 durchqueren soll. Da die Ziele in sich widersprüchlich sind, muss dabei das Frequenzband in Bereiche aufgeteilt werden, in denen jeweils eines der Ziele priorisiert wird. Hierzu wird man fordern (siehe Bild 11-1):

- im Bereich niedriger Frequenzen eine hohe Verstärkung $|G_0|$, um zumindest niederfrequente Störungen auszuregeln und die statischen Fehler klein zu halten.
- im Bereich hoher Frequenzen eine niedrige Verstärkung $|G_0|$, da unmodellierte Dynamik, Linearisierungsfehler und Messrauschen tendenziell hochfrequent sind.
- im dazwischenliegenden Übergangsbereich eine gute Amplituden- und Phasenreserve, wobei ω_d und ω_π möglichst groß sein sollten, um eine hohe Bandbreite im Regelkreis zu erzielen.

Typischerweise erfüllt der Frequenzgang der Regelstrecke G_S die formulierten Anforderungen nicht oder nur teilweise. In diesem Fall soll der Regler G_R den Frequenzgang G_0 des aufgeschnittenen Regelkreises ausgehend von G_S so verformen, dass G_0 möglichst viele Anforderungen erfüllt. Da dem aufgeschnittenen Regelkreis eine Wunschform verliehen werden soll, wird dieses Entwurfsverfahren in der englischen Literatur auch als „Open Loop Shaping“ bezeichnet.

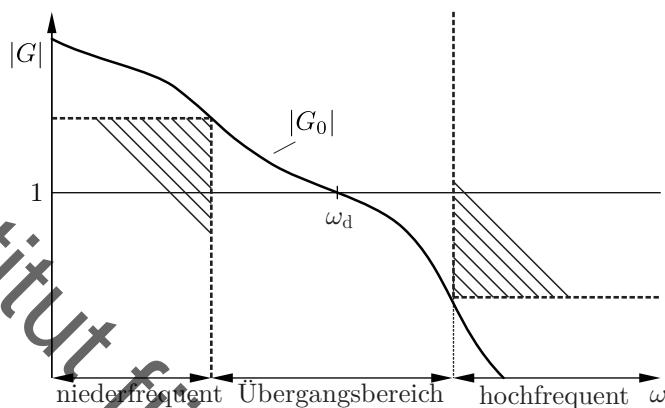


Bild 11-1: Unterteilung des Frequenzbereiches in Gebiete hoher und niedriger Verstärkung

Der Entwurf selbst ist komplex und erfolgt mehrstufig, indem der Regler G_R immer weiter um zusätzliche Pol- und Nullstellen ergänzt wird, um auf Basis der graphischen Addition im Bode-Diagramm eine gewünschte Kontur von G_0 zu erreichen.

Die genauen Anforderungen müssen auf das jeweils vorliegende System abgestimmt werden, da sinnvolle Frequenzbereiche von den Eigenschaften der Regelstrecke und den Anforderungen an das Regelungssystem und damit vom konkreten Anwendungsfall abhängen.

Im Folgenden werden Maßnahmen erläutert, um in jedem der drei beschriebenen Frequenzbereiche jeweils einzeln die Regelziele zu erfüllen. Die geeignete Kombination dieser Maßnahmen über alle drei Bereiche hinweg führt dann auf den Gesamtregler.

11.1.2 Hohe Verstärkung bei niedrigen Frequenzen

Um bei niedrigen Frequenzen eine hohe Reglerverstärkung zu erzielen, ist es zweckmäßig, dass der aufgeschnittene Regelkreis integrierendes Verhalten besitzt, d. h. neben stabilen Polen genau einen Pol in $s = 0$ besitzt. Hierdurch können hohe Verstärkungen für kleine Frequenzen und wegen

$|G_0(j\omega)| \rightarrow \infty$ für $\omega \rightarrow 0$ auch stationäre Genauigkeit sichergestellt werden.

Ein mehrfacher Pol in $s = 0$ ist dabei zu vermeiden, da die Phase bei einem n -fachen Pol in $s = 0$ für $\omega = 0$ bei $-90^\circ \cdot n$ startet, wodurch es für $n \geq 2$ zu Stabilitätsproblemen kommen kann.

Daher wird man bei Regelstrecken mit Ausgleich einen Regler mit integrierendem Verhalten vorsehen, bei Regelstrecken ohne Ausgleich jedoch auf einen integrierenden Anteil im Regler verzichten.

Die Festlegung des Bereiches, in welchem Störungen unterdrückt werden sollen, ergibt sich üblicherweise aus der Spezifikation des geschlossenen Regelkreises. Wird dabei gefordert, dass Ausgangsstörungen mit einer maximalen Frequenz von ω_g mit maximal $p\%$ ihrer Amplitude auf den Ausgang wirken sollen, so muss gelten:

$$|S(j\omega)| = \frac{1}{|1 + G_0(j\omega)|} \leq p\% \quad \forall \omega \leq \omega_g \quad . \quad (11.1)$$

Aus der umgekehrten Dreiecksungleichung lässt sich eine hinreichende Bedingung zur Erfüllung der gegebenen Anforderung herleiten:

$$|G_0(j\omega)| \geq 1 + \frac{1}{p\%} \quad \Rightarrow \quad \frac{1}{|1 + G_0(j\omega)|} \leq p\%. \quad (11.2)$$

Als Beispiel zum Reglerentwurf im niederfrequenten Bereich wird ein PT₁ als Regelstrecke betrachtet und gefordert, dass Störungen mit einer maximalen Bandbreite von $\omega_g = 2 \text{ sec}^{-1}$ auf mindestens 20% gedämpft werden sollen und der Regelkreis stationär genau arbeiten soll.

Als Regler wird ein I -Regler angestrebt, der stationäre Genauigkeit sicherstellt. Der geschlossene Regelkreis ist dabei stabil, da $\varphi(\omega)$ nur im Unendlichen die -180° -Linie berührt, wo $|G_0(\infty)| = 0$ gilt.

Über Gl.(11.2) erhält man mit $p\% = 0,2$ eingesetzt die Bedingung, dass der aufgeschnittene Regelkreis eine Verstärkung von $|G_0(j\omega_g)| \geq 6$ aufweisen muss. Diese Ungleichung lässt sich als Sperrfläche für den Amplitudengang $|G_0|$ auffassen, wie sie in Bild 11-2 gezeigt ist. Die markierte Ecke der Sperrfläche stellt dabei die Mindestanforderungen an $|G_0|$ dar, bei der die Ungleichung $G_0 \geq 6$ zu einer Gleichung wird.

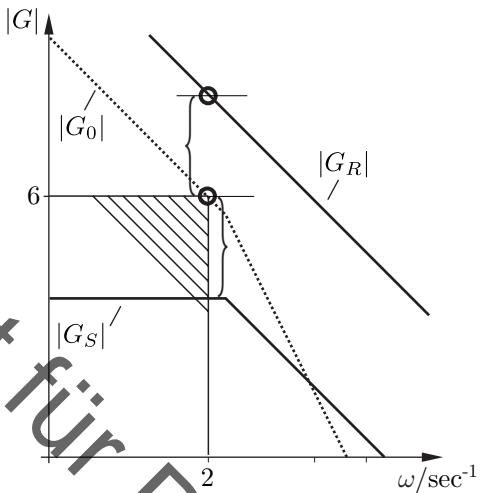


Bild 11-2: Auslegung eines I-Reglers für den niederfrequenten Bereich

Um eine übermäßig hohe Reglerverstärkung mit den in Abschnitt 10.3 erläuterten Problemen zu vermeiden, bietet es sich an, den Amplitudengang des offenen Regelkreises $|G_0|$ genau durch die markierte Ecke laufen zu lassen. Durch graphische Addition kann die geforderten Reglerverstärkung an dieser Frequenz ermittelt werden (siehe Bild 11-2).

Bei dieser Auslegung handelt es sich selbstredend nicht um die einzige mögliche. Tatsächlich erfüllt auch jede andere Regelübertragungsfunktion die Anforderungen, solange der entsprechende Amplitudengang von $|G_0|$ jenseits der Sperrflächen verläuft, G_R einen integrierenden Anteil hat und der geschlossene Regelkreis stabil ist.

11.1.3 Übergangsbereich

Der Übergangsbereich ist beim Reglerentwurf über das Frequenzkennlinienverfahren der komplexeste. Zur genaueren Erläuterung des Entwurfsverfahrens wird daher bei der Auslegung eines einfachen P-Reglers begonnen, welcher für den Übergangsbereich ausgelegt werden soll.

Bild 11-3 zeigt den Frequenzgang eines aufgeschnittenen Regelkreises mit einer Regelstrecke mit Verzögerung dritter Ordnung. Der Übertragungsfaktor K_R des P-Reglers ist so zu wählen, dass $A_{R\max} \geq A_R \geq A_{R\min}$ und $\alpha_{R\max} \geq \alpha_R \geq \alpha_{R\min}$ gelten.

Empfohlene Bereiche für Amplituden- und Phasenreserve

Für Festwertregelungen wird als allgemeine Entwurfsregel

$$1,5 < A_R < 3,0 \quad ; \quad 20^\circ < \alpha_R < 70^\circ \quad (11.3)$$

empfohlen; für Folgeregelungen lauten die empfohlenen Bereiche

$$4 < A_R < 10 \quad ; \quad 40^\circ < \alpha_R < 60^\circ \quad . \quad (11.4)$$

Man sieht, dass die für gutes Führungsverhalten empfohlenen Werte durch eine schwächere Reglereinstellung erreicht werden als die für gutes Störverhalten vorgeschlagenen. Dies liegt u. a. daran, dass Änderungen der Führungsgröße unmittelbar auf den Regler wirken, während Eingangsstörungen durch die Regelstrecke gedämpft werden. Daraus folgt auch, dass ein Regelkreis, der eine am Ausgang der Regelstrecke wirkende Störgröße gut unterdrücken soll (vergleiche auch Bild 10-5), nach den Empfehlungen für gutes Führungsverhalten ausgelegt werden sollte.

Zur Bestimmung günstiger Werte für K_R wird zunächst der Frequenzgang $G'_0(j\omega) = G_0(j\omega)/K_R$ des aufgeschnittenen Regelkreises eingetragen. Dies entspricht der Verwendung eines Reglerübertragungsfaktors von $K_R = 1$. Veränderungen von K_R (ohne Vorzeichenwechsel) werden dabei den Phasengang nicht beeinflussen ($\varphi_0 = \varphi'_0$) und lediglich den Amplitudengang parallel nach oben und unten verschieben. Folglich kann man aus φ_0 die Frequenz ω_π ermitteln, bei der der endgültige Betrag von $G_0(j\omega)$ gleich dem Reziprokwert der Amplitudenreserve sein soll. Da zwei Grenzen für die Amplitudenreserve vorgegeben sind, erhält man einen Bereich von $1/A_{R\min}$ bis $1/A_{R\max}$, innerhalb dessen der Amplitudengang von $G_0(j\omega)$ die Linie ω_π schneiden muss.

Aus den Angaben über die Phasenreserve erhält man zwei Frequenzen ω_{d1} und ω_{d2} , die einen Bereich kennzeichnen, innerhalb dessen $|G_0(j\omega)|$ die 10^0 -Linie schneiden muss. Man erkennt, dass sowohl $G_{01}(j\omega)$ als auch $G_{02}(j\omega)$ und alle zwischen diesen beiden Kurven verlaufenden Frequenzgänge die vorgegebenen Grenzen für Amplituden- und Phasenreserve einhalten.

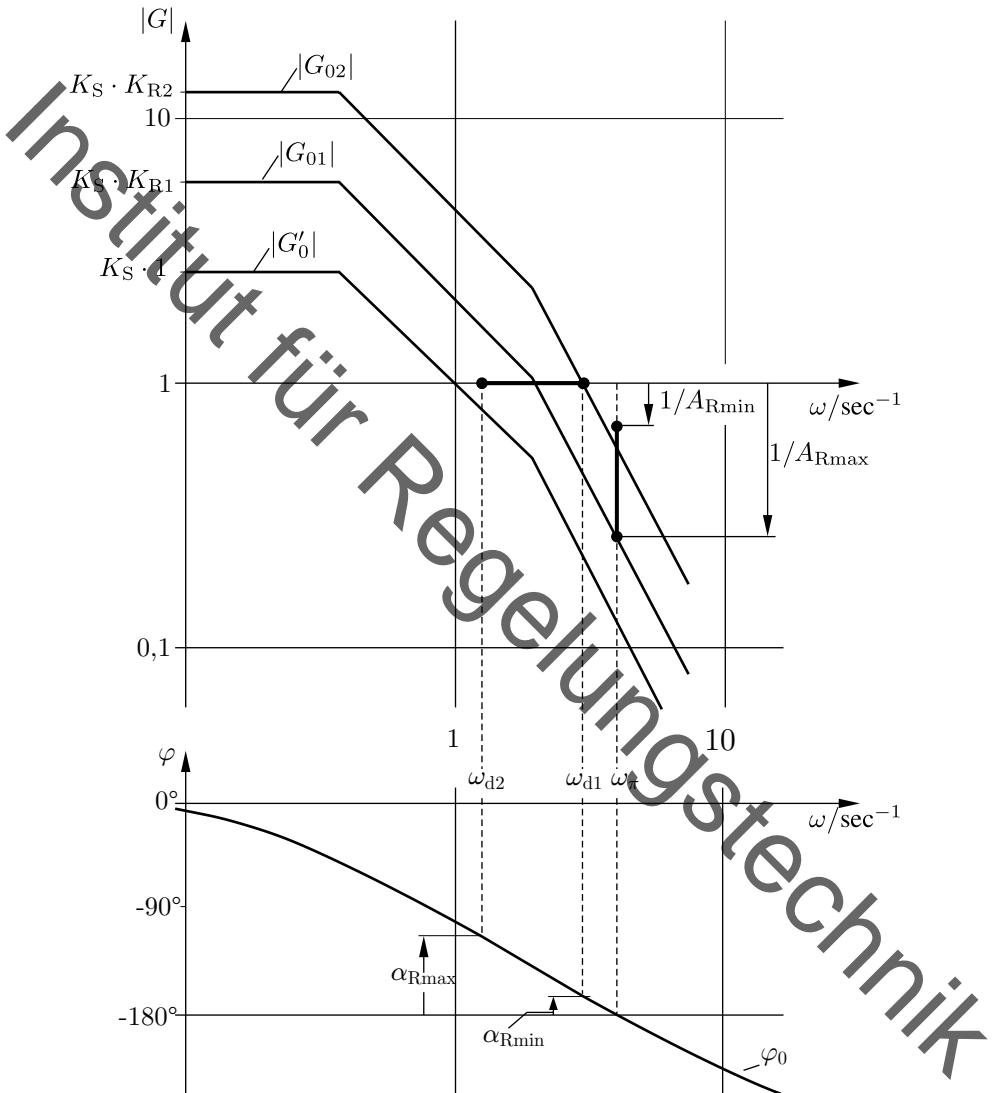


Bild 11-3: Auslegung eines P-Reglers für den Übergangsbereich

In diesem Beispiel ist die obere Grenze von K_R durch $\varphi_{R,\min}$ und die untere Grenze durch $1/A_{R,\max}$ bestimmt. Damit erhält man $K_{R1} \leq K_R \leq K_{R2}$ als Einstellvorschrift für den Übergangsbereich.

Aus dem Beispiel in Bild 11-3 sieht man weiterhin, dass eine hohe Phasenreserve für solche ω_d erzielt werden kann, bei denen die Steigung von $|G_0|$ genau -1 beträgt (siehe ω_{d2}). Für größere ω_d (siehe ω_{d1}), bei denen die Steigung bereits -2 beträgt, fällt die Phasenreserve dabei wesentlich geringer aus.

Das Beispiel ist kein Sonderfall, sondern zeigt ein typisches Phänomen stabiler Regelstrecken: Jeder Steigungsabfall im Amplitudengang um -1 ist mit einer stabilen Polstelle assoziiert, die die Phase um -90° asymptotisch fallen lässt. Folglich ist für minimalphasige aufgeschnittene Regelkreise zu erwarten, dass sich bei einer Steigung von -1 die Phase grob abgeschätzt bei -90° bewegen wird, während bei einer Steigung von -2 eine Phase im Bereich von -180° zu erwarten ist.

Dies ist natürlich nur eine Faustformel, da Pol- und Nullstellen die Phase in einem größeren Frequenzbereich beeinflussen, weswegen das Kurvenstück mit der passenden Steigung sich in einem hinreichend großen Bereich befinden sollte und bspw. der Knick auf -2 nicht unmittelbar nach der Durchtrittsfrequenz ω_d erfolgen sollte.

Daher kann man als Erfahrungsregel für G_0 ableiten:

- der Amplitudengang soll die 10^0 -Linie mit der Steigung -1 schneiden, weil zu dieser Steigung bei Phasenminimumsystemen ein Phasenwinkel von -90° gehört,
- das Kurvenstück mit der Steigung -1 sollte sich mindestens von $|G_0| = 2$ bis $|G_0| = 0,4$ erstrecken.
- für die Durchtrittsfrequenz ω_d gilt häufig näherungsweise der Zusammenhang $\omega_d T_{\text{an}} \approx 1,5$.

Diese Anforderung wird sich durch die Anpassung der Verstärkung K_R des P-Reglers normalerweise nicht zur Zufriedenheit umsetzen lassen, da der P-Regler die Steigung des Amplitudengangs nicht ändern kann. Sie werden stattdessen häufig dadurch erfüllt, dass man den Regler um in Reihe geschaltete Kompensationsglieder erweitert. Der Regler G_R bildet zusammen mit dem Kompensationsglied G_K einen modifizierten Regler $G_R G_K$.

Als Kompensationsglieder werden sehr häufig Übertragungsglieder mit dem Frequenzgang

$$G_K(j\omega) = \frac{1 + j\omega T_1}{1 + j\omega T_2} \quad (11.5)$$

benutzt, die in Tab. 7-3 als PDT₁-Glied (lead) mit $T_1 > T_2$ und als PPT₁-Glied (lag) mit $T_1 < T_2$ aufgeführt sind. Üblicherweise werden Glieder eingesetzt, bei denen sich die Zeitkonstanten T_1 und T_2 um maximal eine Dekade unterscheiden.

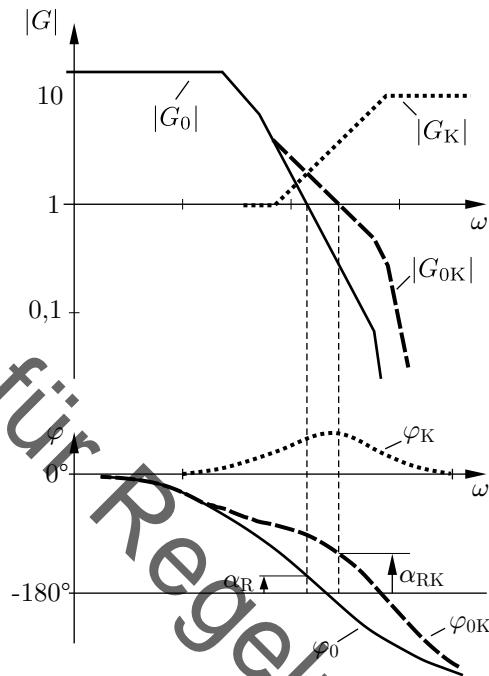
Zur Erreichung des Ziels einer guten Phasenreserve durch eine Steigung von -1 gibt es dabei grundsätzlich zwei Ansätze. Zum einen kann versucht werden, eine zu kleine Steigung von bspw. -2 im gewünschten Bereich von ω_d durch ein PDT₁-Element auf -1 anzuheben. Da das PDT₁-Element gleichzeitig die Phase anhebt, ist dabei mit einer Verbesserung der Phasenreserve zu rechnen. Zusätzlich erhöht sich der Betrag von G_0 , wodurch ω_d vergrößert wird, was die Leistungsfähigkeit des Regelkreises weiter erhöht, da nach der gegebenen Faustformel hierdurch die Anschwingzeit sinkt. Folglich wird der Regelkreis schneller ansprechen und dennoch besser gedämpft erscheinen.

Bild 11-4 illustriert diesen ersten Ansatz für eine Regelstrecke mit Verzögerung vierter Ordnung, wobei $G_{0K} = G_0 G_K$ den aufgeschnittenen Regelkreis mit Kompensationsglied bezeichnet.

In einigen Fällen ist diese Anhebung der Steigung nicht anzuraten, weil beispielsweise durch vorhandene Totzeiten im System die Phase so stark abfällt, dass eine Verschiebung von ω_d hin zu größeren Frequenzen zur Instabilität führt. Da bei stabilen Systemen typischerweise die Phase und der Betrag mit steigendem ω abfallen, ist hier der Weg zu einer guten Phasenreserve, den Amplitudengang vorzeitig abzusenken, damit die Durchtrittsfrequenz ω_d noch in einem Bereich mit einer größeren Phase liegt.

In diesem Fall wird ein PPT₁-Element genutzt, um ω_d hin zu niedrigeren Frequenzen in einen günstigeren Bereich zu verschieben. Dieser zweite mögliche Ansatz ist für einen stark totzeitbehafteten Integrator in Bild 11-5 gezeigt.

Dieser Reglerentwurf zielt auf eine Erhöhung der Robustheit und weniger auf ein – aufgrund der schlechten Eigenschaften der Regelstrecke nicht anzuratendes – Ausreizen der Regelgüte ab. Dies zeigen auch die in Bild 11-5

Bild 11-4: Korrektur des Frequenzganges mit einem PDT₁

gezeigten Sprungantworten des geschlossenen Regelkreises. Die Anschwingzeit des Regelkreises ohne Kompensationsglied $h(t)$ ist offenbar geringer (da ω_d größer ist), führt aber wegen der geringen Phasenreserve zu einer schlechten Dämpfung. Mit Kompensationsglied zeigt $h_k(t)$ eine größere Anschwingzeit, aber einen insgesamt besseren Verlauf.

11.1.4 Niedrige Verstärkung bei hohen Frequenzen

Aufgrund der Regelstrecken innenwohnenden Trägheit besitzen diese üblicherweise einen relativen Grad von $r = 1$ oder größer, weswegen der Amplitudengang von G_S für große Frequenzen gegen null abfällt. Oft erfolgt dieser Abfall allerdings nicht schnell genug – insbesondere dann, wenn Reg-

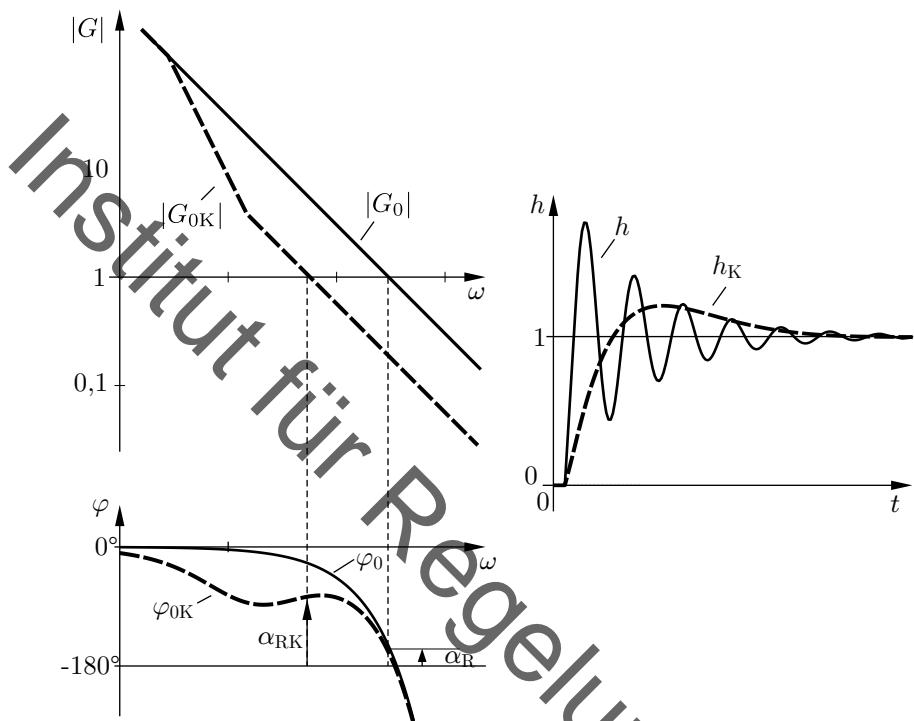


Bild 11-5: Korrektur des Frequenzganges mit einem PPT_1 und resultierende Übergangsfunktionen des geschlossenen Kreises

ler mit einer hohen Reglerverstärkung eingesetzt werden. Insbesondere gilt dies für akausale Regler wie PD-Regler, die den Amplitudengang G_0 für steigende Frequenzen zunehmend anheben.

Um in solchen Fällen einen zusätzlichen Abfall im Amplitudengang für hohe Frequenzen zu erreichen, schaltet man üblicherweise zusätzliche Tiefpassfilter passender Ordnung mit dem Regler in Reihe. Meist setzt man hierfür Verzögerungsglieder PT_n oder PPT_1 -Glieder ein, deren Zeitkonstanten so gewählt werden müssen, dass sie hinreichend weit vom Übergangsbereich und ω_d entfernt liegen, um die in vorherigen Schritten eingestellte Phasenreserve nicht negativ zu beeinflussen, aber dennoch einen hinreichend

großen Amplitudenabfall im hochfrequenten Bereich zu erzielen.

Um den notwendigen Amplitudenabfall quantifizieren zu können, bedient man sich einer ähnlichen Abschätzung wie im niederfrequenten Bereich. Mit der Anforderung

$$|T(j\omega)| = \frac{|G_0(j\omega)|}{|1 + G_0(j\omega)|} \leq p\% \quad \forall \omega \leq \omega_g \quad (11.6)$$

ergibt sich mithilfe der Dreiecksungleichung die Abschätzung

$$|G_0(j\omega)| \leq \frac{p\%}{1 + p\%} \Rightarrow \frac{|G_0(j\omega)|}{|1 + G_0(j\omega)|} \leq p\% . \quad (11.7)$$

Das Verbot einer Verstärkung von Messrauschen würde beispielsweise nach Gl.(11.7) auf die Forderung $|G_0| \leq 0,5$ führen. Das passende Filter kann dabei mit den entsprechenden Sperrbereichen analog zum Vorgehen für den niederfrequenten Bereich entworfen werden.

11.2 Betragkriterium und Symmetrisches Kriterium

Das vorgestellte Frequenzkennlinienverfahren verfolgt das Ziel, die Störübertragungsfunktion S zu null werden zu lassen, indem die Verstärkung $|G_0|$ in den passenden Bereichen groß gewählt wird. Dabei wird beim Entwurf weniger rechnerisch als graphisch vorgegangen.

Das Betragkriterium und das Symmetrische Kriterium sind zwei Methoden zur Reglerauslegung, die ebenfalls mit Hilfe des Frequenzgangs des aufgeschnittenen Regelkreises das Verhalten des geschlossenen Regelkreises optimieren. Dabei gehen sie jedoch primär technisch vor. Dies ermöglicht geschlossene Berechnungen, bedingt aber für die Handhabbarkeit der Gleichungen einige Voraussetzungen.

Die Verfahren erfreuen sich insbesondere in der Antriebsregelung großer Beliebtheit, wobei man sie dort vorwiegend unter den Namen „Betragsoptimum“ und „Symmetrisches Optimum“ findet. Da bei beiden Verfahren streng genommen keine Optimierung vorgenommen wird, werden sie hier abweichend als „Kriterien“ bezeichnet, um die Bezeichnung der Optimierung den Verfahren in Kapitel 18 und 19 vorzubehalten [45].

Bei beiden Verfahren wird die Regelstrecke als reines Verzögerungselement angenommen, d. h. die Übertragungsfunktion $G_S(s)$ ist stabil und besitzt keine Nullstellen. Abweichend zum bisherigen Entwurf wird statt der Störübertragungsfunktion S , welche zu Null gebracht werden soll, die Führungsübertragungsfunktion T betrachtet, die gemäß $S + T = 1$ folglich eins entsprechen soll. Die beiden Verfahren suchen eine Reglereinstellung, die die daraus abgeleitete Forderung an den Betrag von $T(j\omega)$

$$|T(j\omega)| = \left| \frac{G_0(j\omega)}{1 + G_0(j\omega)} \right| \stackrel{!}{=} 1 \quad (11.8)$$

möglichst gut erfüllt.

Beim Betragkriterium deutet schon der Name darauf hin, dass obige Idealbedingung angenähert werden soll. Dazu wird die Bedingung $|T| = 1$ quadratiert und invertiert

$$\left| \frac{1}{T(j\omega)} \right|^2 = \frac{1}{T(j\omega)} \frac{1}{T(-j\omega)} \stackrel{!}{=} 1 \quad , \quad (11.9)$$

sodass sich für den Regelkreis die Forderung

$$1 \stackrel{!}{=} \left(1 + \frac{1}{G_0(j\omega)}\right) \left(1 + \frac{1}{G_0(-j\omega)}\right) \quad (11.10)$$

und mit $G_0(j\omega) = \frac{Z_0(j\omega)}{N_0(j\omega)}$

$$\begin{aligned} 1 &\stackrel{!}{=} 1 + \frac{N_0(j\omega)}{Z_0(j\omega)} + \frac{N_0(-j\omega)}{Z_0(-j\omega)} + \frac{N_0(j\omega)N_0(-j\omega)}{Z_0(j\omega)Z_0(-j\omega)} \\ &= 1 + \frac{Z_0(-j\omega)N_0(j\omega) + Z_0(j\omega)N_0(-j\omega) + N_0(j\omega)N_0(-j\omega)}{Z_0(j\omega)Z_0(-j\omega)} \end{aligned} \quad (11.11)$$

ergibt.

Die Forderung aus Gl.(11.9) wird nur erfüllt, wenn

$$P(\omega) = Z_0(-j\omega)N_0(j\omega) + Z_0(j\omega)N_0(-j\omega) + N_0(j\omega)N_0(-j\omega) = 0 \quad (11.12)$$

gilt. Das Polynom $P(\omega)$ enthält nur gerade Potenzen von ω und hat reelle Koeffizienten

$$P(\omega) = p_2\omega^2 + p_4\omega^4 + p_6\omega^6 + \dots \quad (11.13)$$

Um die Bedingung $|T(j\omega)| = 1$ exakt zu erfüllen, müssten alle Koeffizienten des Polynoms zu null gesetzt werden. Da die Parameter des Reglers in den Koeffizienten enthalten sind, ergeben sich somit Bedingungen für die Reglereinstellung. Allerdings wird man meist nur wenige Koeffizienten verschwinden lassen können.

Betragskriterium

Das Betragskriterium versucht, Gl.(11.13) dadurch gut zu erfüllen, dass von der niedrigsten Potenz von ω zu höheren fortschreitend möglichst viele Glieder in Gl.(11.13) zu null gesetzt werden.

Man erreicht damit, dass $|T(j\omega)| = 1$ in einer möglichst großen Umgebung von $\omega = 0$ erfüllt wird.

Wie viele Koeffizienten man zu null setzen kann, hängt von der Ordnung des Reglers ab. Mit einem PI-Regler kann man z. B. die ersten beiden und mit einem PID-Regler die ersten drei Glieder verschwinden lassen.

Weil vorausgesetzt wurde, dass die Strecke reines Verzögerungsverhalten hat, lässt sich ihr Frequenzgang in der Form

$$G_S(j\omega) = \frac{1}{a_0 + a_1 j\omega + a_2 (j\omega)^2 + \dots} \quad (11.14)$$

angeben.

Setzt man für den Regler einen idealisierten (akausalen) PID-Regler an mit

$$G_R(j\omega) = \frac{r_0 + r_1 j\omega + r_2 (j\omega)^2}{j\omega} , \quad (11.15)$$

so ergeben sich für den Zähler Z_0 und den Nenner N_0 des Frequenzgangs des aufgeschnittenen Regelkreises die folgenden Polynome:

$$\begin{aligned} Z_0(j\omega) &= r_0 + r_1 j\omega + r_2 (j\omega)^2 \\ N_0(j\omega) &= a_0 j\omega + a_1 (j\omega)^2 + a_2 (j\omega)^3 \dots \end{aligned} \quad (11.16)$$

Eingesetzt in Gl.(11.12) erhält man daraus das Polynom

$$\begin{aligned} P(\omega) &= (a_0^2 - 2r_0 a_1 + 2r_1 a_0) \omega^2 + \\ &+ (-2a_0 a_2 + a_1^2 + 2r_0 a_3 - 2r_1 a_2 + 2r_2 a_1) \omega^4 + \\ &+ (2a_0 a_4 - 2a_1 a_3 + a_2^2 - 2r_0 a_5 + 2r_1 a_4 - 2r_2 a_3) \omega^6 \\ &+ \dots = 0 . \end{aligned} \quad (11.17)$$

Für die Auslegung eines PI-Reglers ($r_2 = 0$) würde man die Koeffizienten der Potenzen ω^2 und ω^4 zu null setzen und damit r_0 und r_1 zu

$$r_0 = \frac{a_0(a_1^2 - a_0 a_2)}{2(a_1 a_2 - a_0 a_3)} , \quad r_1 = \frac{a_1(a_1^2 - a_0 a_2)}{2(a_1 a_2 - a_0 a_3)} - \frac{a_0}{2} \quad (11.18)$$

berechnen.

Die Methode des Symmetrischen Kriteriums ist der des Betragskriteriums sehr ähnlich. Es wird jedoch zusätzlich angenommen, dass die Zeitkonstanten der Strecke in eine Gruppe großer Zeitkonstanten T_1, \dots, T_n und eine Gruppe kleiner Zeitkonstanten τ_1, \dots, τ_m aufgespalten werden können, wobei $T_1, \dots, T_n \gg \sum_{i=1}^m \tau_i$ gelten soll. Der Frequenzgang der Strecke lautet

also

$$G_S(j\omega) = \frac{K_S}{\prod_{i=1}^n (1 + T_i j\omega) \prod_{k=1}^m (1 + \tau_k j\omega)} \quad (11.19)$$

und der Frequenzgang des Reglers wird in der Form

$$G_R(j\omega) = K_R \frac{(1 + T_R j\omega)^n}{T_R j\omega} \quad (11.20)$$

angesetzt.

Ist der Regler ein PI-Regler, also $n = 1$, wird man nur die größte Zeitkonstante der Strecke berücksichtigen können, entsprechend bei einem PID-Regler ($n = 2$), die zwei größten Zeitkonstanten.

Symmetrisches Kriterium

Beim Symmetrischen Kriterium geht man davon aus, dass der Amplitudengang eines richtig ausgelegten aufgeschnittenen Regelkreises eine Durchtrittsfrequenz ω_d hat, die zwischen den niedrigen Eckfrequenzen $1/T_i$ der großen Zeitkonstanten und den hohen Eckfrequenzen $1/\tau_k$ der kleinen Zeitkonstanten liegt.

Der Frequenzgang $G_0(j\omega)$ hat dann eine gewisse Symmetrie zur Durchtrittsfrequenz, was diesem Verfahren den Namen verliehen hat.

Für das dynamische Verhalten des Regelkreises ist vorrangig der Bereich des Frequenzgangs um die Durchtrittsfrequenz maßgebend. Da die Eckfrequenzen $1/T_i$ der großen Zeitkonstanten genügend weit links von ω_d liegen sollen, kann man folgende Näherung zulassen.

$$\frac{1}{1 + T_i j\omega} \approx \frac{1}{T_i j\omega} \quad \text{für } \omega \gg \frac{1}{T_i} \quad (11.21)$$

Bei den kleinen Zeitkonstanten, die rechts von ω_d liegen, wird es ausreichen, sie durch eine Summenzeitkonstante

$$T_\Sigma = \sum_{i=1}^m \tau_i \quad (11.22)$$

zu ersetzen.

Mit diesen Näherungen folgt dann für den Frequenzgang des aufgeschnittenen Regelkreises

$$G_0(j\omega) = \frac{K_S}{\prod_{i=1}^n T_i(j\omega)^n (1 + T_\Sigma j\omega)} \frac{K_R (1 + T_R j\omega)^n}{T_R j\omega} . \quad (11.23)$$

Durch die Abschätzung

$$\left(\frac{1 + T_R j\omega}{T_R j\omega} \right)^n = \left(1 + \frac{1}{T_R j\omega} \right)^n \approx 1 + n \frac{1}{T_R j\omega} \quad (11.24)$$

vereinfacht sich der Frequenzgang weiter zu

$$G_0(j\omega) = \frac{(1 + \frac{T_R}{n} j\omega)}{T_* (1 + T_\Sigma j\omega) (j\omega)^2} = \frac{Z_0(j\omega)}{N_0(j\omega)} \text{ mit } T_* = \frac{\prod_{i=1}^n T_i}{n K_R K_S T_R^{n-2}} . \quad (11.25)$$

Für den Betragsverlauf des Frequenzgangs des geschlossenen Regelkreises

$$T(j\omega) = \frac{G_0(j\omega)}{1 + G_0(j\omega)} = \frac{Z_0(j\omega)}{N_0(j\omega) + Z_0(j\omega)} \quad (11.26)$$

wird ähnlich dem Betragskriterium gefordert, dass in einer möglichst großen Umgebung von $\omega = 0$ dann $|T(j\omega)| = 1$ sein soll. Deshalb verlangt man $|N_0(j\omega) + Z_0(j\omega)| = 1$ und versucht dieses anzunähern, indem man in dem Polynom $P(\omega)$

$$P(\omega) = (N_0(j\omega) + Z_0(j\omega)) (N_0(-j\omega) + Z_0(-j\omega))^{-1} \quad (11.27)$$

nach Einsetzen und Ausmultiplizieren

$$P(\omega) = 1 + \left(\frac{T_R^2}{n^2} - 2T_* \right) \omega^2 + \left(-2 \frac{T_\Sigma T_R T_*}{n} + T_*^2 \right) \omega^4 + T_*^2 T_R^2 \omega^6 \quad (11.28)$$

möglichst viele Koeffizienten der ersten ω -Potenzen durch geeignete Wahl der Reglerparameter zu null setzt.

Möchte man z. B. einen PI-Regler verwenden, so ist $n = 1$ und man hat zwei freie Reglerparameter, mit denen die Koeffizienten von ω^2 und ω^4 zu null gesetzt werden können:

$$\begin{aligned} T_R^2 - 2T_* &= 0 \\ -2T_\Sigma T_R T_* + T_*^2 &= 0 \quad \text{mit} \quad T_* = \frac{T_1 T_R}{K_R K_S} \end{aligned} . \quad (11.29)$$

Für die Reglerparameter folgen daraus die Einstellwerte

$$K_R = \frac{T_1}{2T_\Sigma K_S} \quad \text{und} \quad T_R = 4T_\Sigma \quad . \quad (11.30)$$

Als Beispiel zum Vergleich beider Verfahren ist für eine Verzögerungsstrecke dritter Ordnung mit einer großen und zwei kleinen Zeitkonstanten

$$G_S(j\omega) = \frac{K_S}{(1 + T_1 j\omega)(1 + T_2 j\omega)(1 + T_3 j\omega)} \quad (11.31)$$

$K_S = 1$, $T_1 = 10 \text{ sec}$, $T_2 = 1 \text{ sec}$, $T_3 = 0,1 \text{ sec}$

ein PI-Regler nach den Einstellregeln des Betragkriteriums und des Symmetrischen Kriteriums ausgelegt worden.

Bild 11-6 zeigt die Sprungantworten der geschlossenen Regelkreise bei einem Sprung der Führungsgröße. Man erkennt, dass in beiden Fällen die Regelgröße überschwingt und der Einschwingvorgang beim Symmetrischen Kriterium schwächer gedämpft ist als beim Betragskriterium.

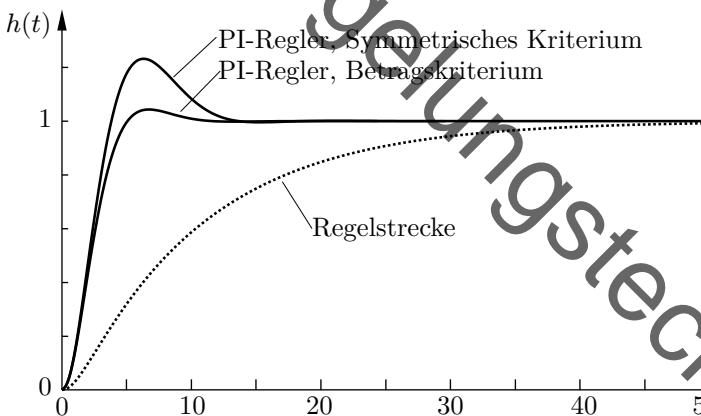


Bild 11-6: Führungsübergangsfunktion des Regelkreises mit PI-Regler

11.3 Polvorgabe

11.3.1 Polvorgabe für Ausgangsrückführungen

Die bisher vorgestellten Reglerentwurfsverfahren adressieren den geschlossenen Regelkreis, für welche eigentlich die Regelziele definiert sind, nur indirekt über Approximationen oder Abschätzungen an den aufgeschnittenen Regelkreis. Durch die Abschätzungen geht dabei stets Genauigkeit verloren und die Leistungsfähigkeit des Reglers kann ggf. nicht voll ausgeschöpft werden. Zusätzlich liefern Daumenregeln – wie die für Amplituden- und Phasenreserven gegebenen – nur eine grobe Orientierung für das erwartete dynamische Verhalten im geschlossenen Regelkreis.

Diese Problematik ist bei Verfahren, die direkt am geschlossenen Regelkreis arbeiten, weniger stark ausgeprägt. Da der Zusammenhang der einzustellenden Reglerparameter mit der Systemdynamik des geschlossenen Regelkreises sehr komplex ausfällt, ist es unklug, mit der analytischen Lösung des geschlossenen Regelkreises zu arbeiten. Stattdessen behilft man sich auch hier mit einer Näherung für das vermutete Verhalten im geschlossenen Regelkreis – die Polstellen bzw. Eigenwerte des geschlossenen Regelkreises.

Auch wenn diese Näherung nicht die vollständige Dynamik abbilden, da diese auch von – bei diesem Ansatz ausgeschlossenen – Totzeiten, den Residuen und Nullstellen abhängig ist, so bestimmen die Polstellen doch die wesentlichen Eigenschaften wie Schwingungsfähigkeit, Stabilität und die Schnelligkeit, mit der Einschwingvorgänge abklingen.

Bild 11-7 zeigt das Zielgebiet, in welchem die Polstellen des geschlossenen Regelkreises vorzugsweise liegen sollten. Die Pfeile zeigen dabei die Richtung besserer Eigenschaften an.

Zielgebiet der Pole des geschlossenen Regelkreises

Die Pole des geschlossenen Regelkreises sollten ...

- für schnelles Einschwingen möglichst geringen Realteil α
- für geringes Überschwingen möglichst hohe Dämpfung D
- für geringe Schwingungsfrequenzen möglichst kleinen Imaginärteil ω_D
- für schnelles Ansprechverhalten einen möglichst großen Betrag ω_0 aufweisen.

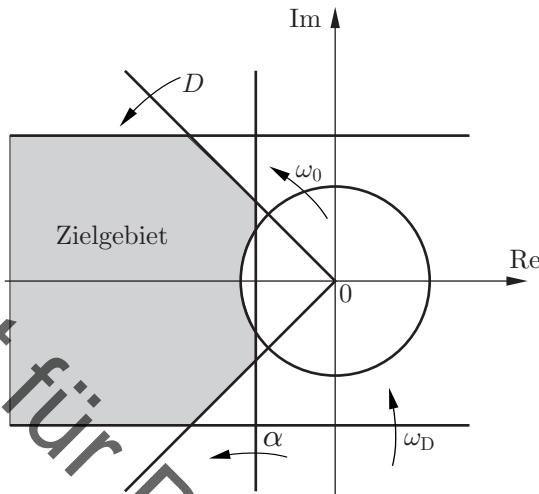


Bild 11-7: Zielgebiet der Pole des geschlossenen Regelkreises

Die in Bild 10-3 aufgeführten Spezifikationen im Zeitbereich können dabei überschlägig wie folgt mit den Positionen der Polstellen assoziiert werden.

- Einschwingzeit: Realteil α und Dämpfung D
- Anschwingzeit: Realteil α und Betrag ω_0
- Überschwingweite: Dämpfung D und Imaginärteil ω_D

Die stationäre Genauigkeit kann an dieser Stelle nicht abgelesen werden und muss auf anderem Wege (wie durch eine Vorsteuerung, siehe Abschnitt 12.2) sichergestellt werden.

Insgesamt scheint es nach Bild 11-7 vorteilhaft, die Pole möglichst weit nach links in der komplexen Halbebene zu verschieben. Hierbei ist aber zu beachten, dass dies in vielen Fällen nur mit sehr großen Reglerverstärkungen umsetzbar sein wird, deren Nachteile in Abschnitt 10.3 diskutiert worden sind. Außerdem wird es von der angesetzten Reglerstruktur abhängig sein, wie viele Pole des geschlossenen Regelkreises sich wohin verschieben lassen. Hierbei kommt insbesondere den dominanten Polstellen eine Schlüsselrolle zu, da diese für die Dynamik normalerweise ausschlaggebend sind.

Der Reglerentwurf am geschlossenen Regelkreis hat zum Ziel, die Reglerparameter so zu wählen, dass alle Pole des geschlossenen Regelkreises ein vorher definiertes Zielgebiet erreichen. Eine naheliegende Lösung hierfür ist es, die Pole des geschlossenen Regelkreises in Abhängigkeit der Parameter auszurechnen und die entstehenden Gleichungen nach den Reglerparametern umzustellen. Diese Rechnung gestaltet sich allerdings – ganz analog zu Stabilitätsprüfung über das Ausrechnen der Polstellen – insgesamt als schwierig, da zu viele symbolische Variablen in den Rechnungen erscheinen.

Eine Alternative besteht darin, bestimmte Polpositionen innerhalb des Zielgebietes als Sollpositionen der Pole des geschlossenen Kreises zu definieren und die Reglerparameter über einen Koeffizientenvergleich zu lösen. Dieses Verfahren führt zu einfacheren Rechnungen, die abhängig von den gewünschten Polstellen und dem angesetzten Reglertyp aber manchmal auch gar keine Lösung liefern.

Diese Problematik soll an einem Beispiel deutlich gemacht werden. Betrachtet wird eine PT_1 -Regelstrecke mit einem PI-Regler in der dimensionslosen Form

$$G_S(s) = \frac{1}{s+1} \quad , \quad G_R(s) = K \frac{T_n s + 1}{T_n s} . \quad (11.32)$$

Der Regler soll so gewählt werden, dass die Einschwingzeit der Regelstrecke halbiert wird und ein aperiodischer Grenzfall vorliegt. Folglich muss für die Dämpfung $D = 1$ und für den Realteil der Polstellen $\alpha = -2$ gelten, da die Regelstrecke einen Pol bei -1 aufweist. Der geschlossene Regelkreis besitzt das charakteristische Polynom

$$p(s) = T_n s^2 + (T_n + K T_n) s + K \quad (11.33)$$

mit zwei Wurzeln.

Aus den Anforderungen ergibt sich, dass beide Polstellen genau bei -2 liegen müssen. Durch Koeffizientenvergleich gewinnt man nach Teilen durch T_n

$$s^2 + \frac{(T_n + K T_n)}{T_n} s + \frac{K}{T_n} \stackrel{!}{=} (s+2)^2 = s^2 + 4s + 4 \Leftrightarrow K = 3, T_n = \frac{3}{4} \quad (11.34)$$

und damit die gesuchte Parametrierung des Reglers.

Versucht man die gleiche Rechnung bei einer Regelstrecke $G_S = 1/(s+1)^2$, so erhält man abweichend

$$s^3 + 2s^2 + \frac{(T_n + KT_n)}{T_n}s + \frac{K}{T_n} \stackrel{!}{=} (s+2)^3 = s^3 + 6s^2 + 12s + 8 \quad (11.35)$$

ohne eine gültige Lösung, da der Koeffizient a_2 für alle T_n und K nicht übereinstimmt.

In dem vergleichsweise einfach konstruierten Beispiel gibt es bereits erste Probleme beim Lösen der entstehenden Gleichungssysteme. Der Grund hierfür ist, dass mit dem gewählten PI-Regler als Ausgangsrückführung eine Reglerstruktur verfolgt wurde, die bei einer beliebigen Vorgabe für die Positionen der Pole des geschlossenen Regelkreises keine Lösung des Gleichungssystems garantieren kann.

Eine solche Garantie ist aber möglich, wenn man anstelle einer Ausgangsrückführung eine Zustandsrückführung ansetzt.

11.3.2 Polvorgabe für Zustandsrückführungen

Hierzu liegt eine LTI-Regelstrecke im Zustandsraum vor, wobei nur die Zustandsgleichung und nicht die Ausgangsgleichung betrachtet wird:

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu} \quad . \quad (11.36)$$

Als Zustandsregler reicht es für eine lineare Regelstrecke in jedem Fall aus, einen linearen Regler

$$\mathbf{u} = -\mathbf{Kx} \quad (11.37)$$

mit der Matrix \mathbf{K} als sogenannter *Rückführmatrix* anzusetzen, wobei das Minuszeichen das Vorhandensein einer Rückkopplung andeuten soll, auch wenn in \mathbf{K} negative Einträge zugelassen sind. Für Eingrößensysteme verwendet man einen Rückführvektor \mathbf{k} .

$$\mathbf{u} = \mathbf{x}(k_1x_1 + k_2x_2 + \dots k_nx_n) = -\mathbf{k}^T\mathbf{x} \quad (11.38)$$

Setzt man $\mathbf{u} = -\mathbf{Kx}$ in das Zustandsraummodell ein, so gewinnt man für den geschlossenen Regelkreis

$$\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{BK})\mathbf{x} \quad (11.39)$$

und damit ein autonomes System mit der neuen Systemmatrix \mathbf{A}_K als

$$\mathbf{A}_K = \mathbf{A} - \mathbf{B}\mathbf{K} . \quad (11.40)$$

Die Eigenwerte dieser Matrix sind dann die Polstellen des geschlossenen Regelkreises.

Zustandsrückführung über Polvorgabe

Werden die Eigenwerte der Matrix \mathbf{A}_K vorgegeben um damit aus den bekannten Matrizen \mathbf{A} und \mathbf{B} die Rückführmatrix \mathbf{K} zu bestimmen, so spricht man von dem Reglerentwurfsverfahren der *Polvorgabe*.

Der Zustandsregler besitzt bereits im SISO-Fall genügend Freiheitsgrade (nämlich die wählbaren Einträge von \mathbf{k}), um die Lösbarkeit der Gleichungssysteme sicherzustellen. Dies wird unter der Annahme gezeigt, dass die Regelstrecke in Regelungsnormalform (siehe Gl.(2.25)) beschrieben ist. Ist dies nicht der Fall, so lassen sich die Zustandsgrößen der Regelungsnormalform in der in Abschnitt 3.6 beschriebenen Weise durch Transformation aus den ursprünglichen Zustandsgrößen gewinnen.

Die Systemmatrix \mathbf{A}_K des Übertragungssystems mit Rückführung ergibt sich nach Gl.(11.40) zu

$$\mathbf{A}_K = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & & 1 \\ -a_0 - k_1 & -a_1 - k_2 & -a_2 - k_3 & \dots & -a_{n-1} - k_n \end{bmatrix} \quad (11.41)$$

und hat wiederum Regelungsnormalform.

Eine nützliche Eigenschaft der Systemmatrix in Regelungsnormalform ist, dass die letzte Zeile die Koeffizienten des charakteristischen Polynoms enthält. Damit kann unmittelbar

$$\det(\lambda\mathbf{I} - \mathbf{A}_K) = \lambda^n + (a_{n-1} + k_n)\lambda^{n-1} + \dots + (a_0 + k_1) \quad (11.42)$$

angeschrieben werden.

Wenn als Polstellen Werte $\lambda_1, \dots, \lambda_n$ vorgegeben sind, bedeutet dies, dass das charakteristische Polynom

$$(\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n) = \lambda^n + p_{n-1}\lambda^{n-1} + \dots + p_0 \quad (11.43)$$

lauten soll.

Mit den so bestimmten Koeffizienten p_i ergibt sich durch Koeffizientenvergleich der Rückführvektor zu

$$\mathbf{k}^T = [p_0 - a_0 \quad p_1 - a_1 \quad \cdots \quad p_{n-1} - a_{n-1}] \quad . \quad (11.44)$$

Folglich ist die Polvorgabe für Systeme in Regelungsnormalform stets und auch besonders einfach zu lösen.

11.3.3 Steuerbarkeit

Liegen (MIMO-)Zustandsraumdarstellungen als Paar (\mathbf{A}, \mathbf{B}) nicht in Regelungsnormalform vor, muss das System gewisse Eigenschaften erfüllen, damit die bei der Polvorgabe entstehenden Gleichungssysteme unter allem Umständen lösbar sind. Ausschaulich kann man sich vorstellen, dass das System es zulassen muss, dass seine Dynamik durch eine passende Wahl von \mathbf{u} beliebig verändert werden kann. Diese Eigenschaft wird *Steuerbarkeit* genannt.

Steuerbarkeit

Ein System $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}$ ist genau dann steuerbar, wenn es aus jedem beliebigen Anfangszustand \mathbf{x}_0 heraus durch eine geeignete Eingangsgröße, den Steuervektor $\mathbf{u}(t)$, in endlicher Zeit t_e in einen beliebigen Zielzustand $\mathbf{x}(t_e) = \mathbf{x}_e$ überführt werden kann.

Man sagt dann auch, dass das Paar (\mathbf{A}, \mathbf{B}) steuerbar ist.

Steuerbarkeit klingt zunächst nach einer sehr mächtigen Eigenschaft. Tatsächlich kann man zeigen, dass jede minimale Realisierung steuerbar ist. Folglich ist Steuerbarkeit eine Eigenschaft, die durch eine passende Modellierung des Systems immer sichergestellt werden kann.

Ist eine Regelstrecke nicht steuerbar, so gibt es analog zu Bild 2-9 parallele Übertragungskanäle, deren Dynamik nicht unabhängig voneinander eingestellt werden kann. Diese Eigenwerte können dann nicht durch eine (Zustands)-Regelung beliebig verschoben werden. In allen anderen Fällen kann dem System im geschlossenen Regelkreis eine gewünschte Dynamik in Form seiner Polstellen aufgeprägt werden.

Lösbarkeit der Polvorgabe

Die Polvorgabe über eine Zustandsrückführung ist genau dann für beliebige Eigenwerte von \mathbf{A}_K lösbar, wenn das Paar (\mathbf{A}, \mathbf{B}) steuerbar ist.

Steuerbarkeit ist also eine notwendige und hinreichende Voraussetzung zur Umsetzung einer Zustandsrückführung mit Polvorgabe. Dies stellt in den meisten Fällen aber keine praktische Einschränkung dar, da nicht steuerbare Systeme durch eine entsprechende Kürzung der Pole und Nullstellen in eine minimalen Realisierung mit identischem Ein-Ausgangsverhalten überführt werden können, die dann steuerbar ist. Systeme in Regelungsnormalform sind offenbar stets steuerbar.

Die Definition der Steuerbarkeit liefert ein mathematisches Kriterium, um Steuerbarkeit auf einfache Weise zu prüfen: Hierzu betrachtet man die allgemeine Lösung $\mathbf{x}(t)$ im Zustandsraum zum Zeitpunkt t_e :

$$\mathbf{x}(t_e) = \underbrace{e^{\mathbf{A} t_e} \mathbf{x}_0}_{\text{festgelegt}} + \int_0^{t_e} e^{\mathbf{A}(t_e - \tau)} \mathbf{B} u(\tau) d\tau \stackrel{!}{=} \mathbf{x}_e \quad (11.45)$$

Der erste Term der homogenen Lösung ist durch den Zeitpunkt t_e und den beliebigen Anfangszustands \mathbf{x}_0 festgelegt. Mit der Reihenentwicklung des Matrixexponentials ergibt sich für den Integranden der partikulären Lösung

$$e^{\mathbf{A}(t_e - \tau)} \mathbf{B} u(\tau) = \left(\mathbf{B} + \frac{(t_e - \tau)}{1!} \mathbf{A} \mathbf{B} + \frac{(t_e - \tau)^2}{2!} \mathbf{A}^2 \mathbf{B} + \dots \right) u(\tau). \quad (11.46)$$

Um den Vektor $\mathbf{x}_e \in \mathbb{R}^n$ mit n Freiheitsgraden einstellen zu können, benötigt man auch n Freiheitsgrade innerhalb der partikulären Lösung. Da diese nach Gl.(11.46) eine Linearkombination der Spalten von \mathbf{B} , $\mathbf{A} \mathbf{B}$, $\mathbf{A}^2 \mathbf{B}$ usw. ist, müssen sich unter diesen Spalten genügend (nämlich n) linear unabhängige befinden.

Nach dem Satz nach Cayley¹-Hamilton können \mathbf{A}^n und alle höheren Potenzen einer Matrix \mathbf{A} als Linearkombination von niedrigen Potenzen geschrie-

¹Arthur Cayley (1821-1895), englischer Mathematiker [5]

ben werden, sodass die Spalten ab $\mathbf{A}^n \mathbf{B}$ keine linear unabhängigen Informationen erbringen. Daher reicht es aus, die Potenzen bis \mathbf{A}^{n-1} zu betrachten. Dies mündet in dem folgenden Test auf Steuerbarkeit nach Kalman²:

Steuerbarkeitskriterium von Kalman

Das Paar (\mathbf{A}, \mathbf{B}) ist genau dann steuerbar, wenn die sogenannte Steuerbarkeitsmatrix

$$\mathbf{Q}_S = [\mathbf{B} \quad \mathbf{AB} \quad \mathbf{A}^2\mathbf{B} \quad \dots \quad \mathbf{A}^{n-1}\mathbf{B}] \quad (11.47)$$

vollen Rang besitzt.

Für SISO-Systeme ist \mathbf{Q}_S quadratisch, sodass diese Bedingung dadurch geprüft werden kann, dass $\det(\mathbf{Q}_S) \neq 0$ für Steuerbarkeit gelten muss. Man kann zudem aus Gl.(11.47) sehen, dass sich die Steuerbarkeit unter einer regulären Zustandstransformation nicht verändert.

Steuerbarkeit lässt sich also leicht überprüfen und stellt meist keine Einschränkung für den Reglerentwurf dar. Somit kann über eine Polvorgabe mit Zustandsrückführung nahezu jedem System (und unabhängig von dessen Stabilität) eine gewünschte Dynamik aufgeprägt werden. Auf diese Weise erscheint dieser Reglerentwurf außerordentlich mächtig zu sein.

Der Entwurf leidet aber strukturell ebenfalls unter den in Abschnitt 10.3 geschilderten Problemen. Auch hier werden Abweichungen in der Streckenbeschreibung, Stellgrößenbeschränkungen etc. die Reglerverstärkung nach oben limitieren. Die verwendete Reglerverstärkung kann dabei indirekt aus Gl.(11.44) ermittelt werden. Man sieht aus

$$\mathbf{k}^T = [p_0 - a_0 \quad p_1 - a_1 \quad \dots \quad p_{n-1} - a_{n-1}] \quad , \quad (11.48)$$

dass die Einträge von \mathbf{k} größer werden, wenn die gewünschten Koeffizienten p_i und die ungeregelten Koeffizienten a_i des charakteristischen Polynoms weit auseinanderliegen. Folglich werden die Einträge des Rückführvektors \mathbf{k} umso größer, je stärker das charakteristische Polynom (und damit die Eigenwerte) verändert werden sollen.

Eine deutliche Verschiebung der Eigenwerte weit in die linke Halbebene wird daher nur mit sehr großen Stellaufländen zu realisieren sein, welche

²Rudolf Kálmán (1930-2016), ungarisch-amerikanischer Ingenieur [25]

zu vermeiden sind. Daher wird man das Zielgebiet in 11-7 abhängig von den Streckenpolen in Betrag und Realteil beschränken.

Als abschließendes Beispiel für die Polvorgabe soll ein Regler für die obere Ruhelage des mathematischen Pendels (d. h. das stehende Pendel) entworfen werden. Mit dem Winkel ϑ und der Winkelgeschwindigkeit $\dot{\vartheta}$ als Zuständen ergibt sich

$$\dot{x} = \begin{bmatrix} \dot{\vartheta} \\ \ddot{\vartheta} \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 \\ \frac{g}{L} & -\frac{\mu}{ML} \end{bmatrix}}_{A} + \underbrace{\begin{bmatrix} 0 \\ \frac{1}{ML^2} \end{bmatrix}}_b u , \quad (11.49)$$

wobei eine Stellgröße u in Form eines im Gelenk angreifenden Momentes hinzugefügt wird, damit eine Regelung sinnvoll möglich ist.

Da die Regelstrecke ein instabiles komplexes Polstellenpaar besitzt, lässt sich kein Regler nach dem Frequenzkennlinienverfahren entwerfen. Eine Zustandsregelung über Polvorgabe ist jedoch möglich, da das System steuerbar ist.

Als Regelziel wird formuliert, dass der geschlossene Regelkreis für eine Dämpfung von $D = \sqrt{2}/2$ ausgelegt werden soll, um Resonanzüberhöhungen zu verhindern. Einschwingvorgänge sollen dabei mit einer Zeitkonstanten von $T = 1$ sec abklingen.

Hieraus ergibt sich als Sollpolynom des geschlossenen Regelkreises

$$p(\lambda) = (\lambda + 1 - j)(\lambda + 1 + j) = \lambda^2 + 2\lambda + 2 . \quad (11.50)$$

Durch Koeffizientenvergleich erhält man über die Regelungsnormalform direkt

$$k^T = [gL + 2ML^2 \quad 2ML^2 - \mu L] . \quad (11.51)$$

Hieraus sieht man, dass die Reglerverstärkung für große μ geringer wird, da die Reibung auf natürlichem Wege zu einer Dämpfung des Systems beiträgt. Des Weiteren werden für große Massen und Länge größere Stelleingriffe notwendig sein, da die Regelstrecke eine größere Trägheit besitzt.

Formt man den entworfenen Zustandsregler in eine Ausgangsrückführung

um, so erhält man mit der Ausgangsgleichung $y = \vartheta$ den Zusammenhang

$$\begin{aligned} u &= -\mathbf{k}^T \mathbf{x} = -(k_1 \vartheta + k_2 \dot{\vartheta}) = -(k_1 y + k_2 \dot{y}) \\ \Rightarrow G_R(s) &= \frac{U(s)}{Y(s)} = -(k_1 + k_2 s), \end{aligned} \quad (11.52)$$

also einen PD-Regler.

Dieses Resultat gilt für jedes System zweiter Ordnung in Regelungsnormalform. Durch entsprechende Zustandstransformationen erhält man allgemein den folgenden Zusammenhang:

Zustandsregler als PD_n -Regler

Ein Zustandsregler für ein SISO-System n -ter Ordnung entspricht strukturell einem Regler mit

$$G_R(s) = k_0 + k_1 s + \dots + k_{n-1} s^{n-1} \quad (11.53)$$

und damit einem akausalen PD_{n-1} -Regler.

11.4 Beobachterentwurf

11.4.1 Zustandsschätzung

Die im vorherigen Abschnitt entworfene Zustandsrückführung setzt voraus, dass die Zustandsgrößen in der physikalischen Wirklichkeit messbar sind. Das ist häufig nicht der Fall.

Auch lässt sich in aller Regel aus den Ausgangsgrößen y nicht ohne weitere Informationen auf den zugehörigen Zustand \mathbf{x} schließen. So führt beispielsweise die Ausgangsgleichung $y = [1 \ 1] \mathbf{x}$ für \mathbf{x} mit den Einträgen

$$x_1 = a + b \quad , \quad x_2 = -a \quad \Rightarrow \quad y = x_1 + x_2 = a + b - a = b \quad (11.54)$$

für beliebige a auf denselben Ausgang $y = b$.

Uneindeutiger Zusammenhang zwischen \mathbf{x} und y

Solange die Ausgangsmatrix \mathbf{C} nicht einen Rang von n hat, kann der Zustand $\mathbf{x} \in \mathbb{R}^n$ nicht direkt aus den Ausgangsgrößen \mathbf{y} ermittelt werden.

In den allermeisten Fällen wird \mathbf{C} diesen Rang nicht besitzen, da es weniger Messgrößen \mathbf{y} als Zustände \mathbf{x} gibt. Diese Tatsache ist der Grund für die resultierenden acausalen Stellgesetze in Gl.(11.53), wenn man den Zustandsregler in eine Ausgangsrückführung umformt.

Folglich muss man versuchen, aus den Eingangs- und den Ausgangsgrößen mit Kenntnis der Eigenschaften des Systems auf die Zustandsgrößen zu schließen und hierdurch die acausalen Regler zu kausalisieren. Eine Einrichtung, die dies leistet, wird (Zustands-)Beobachter genannt. Ein solcher Beobachter kann außer zur Regelung auch für Überwachungsaufgaben nützlich sein.

Bild 11-8 zeigt den strukturellen Aufbau eines Zustandsbeobachters, der den Zustand \mathbf{x} des Systems in der Form des geschätzten Zustandes $\hat{\mathbf{x}}$ wiedergibt, und diesen an einen Zustandsregler übergibt. Dieser nutzt dann ersatzweise den geschätzten Zustand $\hat{\mathbf{x}}$ anstelle des nicht verfügbaren Zustandes \mathbf{x} . Somit ergibt die Kombination von Zustandsrückführung und Zustandsbeobachter insgesamt eine Ausgangsrückführung.

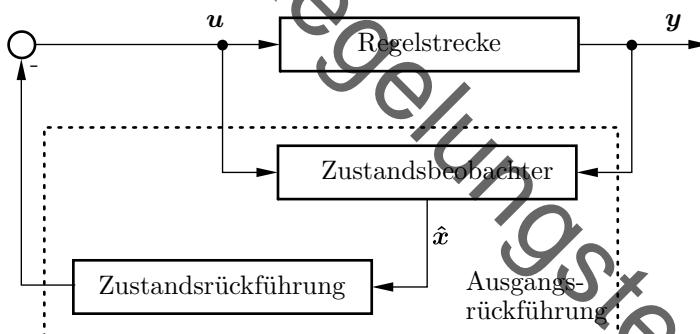


Bild 11-8: Wirkungsplan für ein System mit Zustandsrückführung und Zustandsbeobachter

Da für die Auslegung des Zustandsreglers ein Modell des Systems vorliegt, bietet es sich an, dieses auch innerhalb des Beobachters zum Schätzen der Zustände zu nutzen. Sofern das Modell den realen Prozess einigermaßen genau abbildet, antwortet es auf die Eingangssignale $u(t)$ der Regelstrecke mit einer sehr ähnlichen Ausgangsgröße $y(t)$. Daher sind die Zustände, die sich bei einer Simulation des Systemmodells bei identischen Eingangssignalen

$u(t)$ ergeben, eine gute Schätzung für die Zustände, die in der Regelstrecke vorliegen.

Ein erster Ansatz für einen Zustandsbeobachter ist daher, in diesem eine Kopie des Modells der Regelstrecke zu verwenden und parallel zur Regelstrecke mit zu simulieren. Diese Struktur ist für ein lineares SISO-Streckenmodell in Bild 11-9 gezeigt.

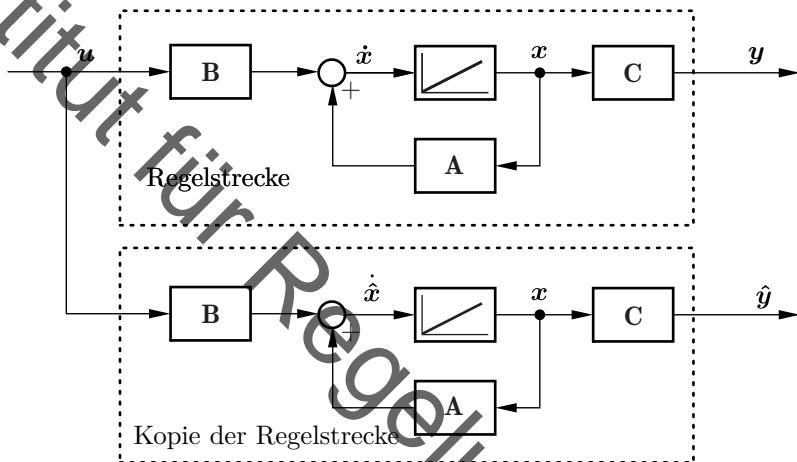


Bild 11-9: Zustandsbeobachter als Kopie des Regelstreckenmodells

Dieser Ansatz hat jedoch einen gewichtigen Nachteil, der die Struktur in Bild 11-9 ohne zusätzliche Maßnahmen für einen praktischen Einsatz unbrauchbar macht. Um diesen herauszuarbeiten, wird angenommen, dass die verwendete Kopie der echten Regelstrecke exakt entspricht. Hierbei handelt es sich um eine Idealvorstellung, da bei der Modellbildung immer Vereinfachungen vorgenommen werden. Unter dieser Voraussetzung sollte der geschätzte Zustand \hat{x} und der echte Zustand x möglichst genau übereinstimmen.

Schätzfehler

Gegeben sei eine Regelstrecke mit Zustand $x(t)$ und ein Beobachter mit geschätztem Zustand $\hat{x}(t)$. Der *Schätzfehler* $e(t)$ ist dann die Abweichung

dieser beiden Zustände

$$\mathbf{e}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t) \quad (11.55)$$

und sollte für beliebige $\mathbf{u}(t)$ und beliebige Anfangswerte ${}_0\mathbf{x}$ und ${}_0\hat{\mathbf{x}}$ möglichst schnell verschwinden: $\lim_{t \rightarrow \infty} \mathbf{e}(t) = \mathbf{0}$.

Berechnet man den Schätzfehler für den Aufbau in Bild 11-9, so gewinnt man mit der Gleichung für die Regelstrecke

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \quad \mathbf{x}(0) = {}_0\mathbf{x} \quad (11.56)$$

und mit der Gleichung für den Beobachter

$$\dot{\hat{\mathbf{x}}} = \mathbf{A}\hat{\mathbf{x}} + \mathbf{B}\mathbf{u}, \quad \hat{\mathbf{x}}(0) = {}_0\hat{\mathbf{x}} \quad (11.57)$$

die Gleichung für den Schätzfehler als

$$\begin{aligned} \dot{\mathbf{e}} &= \dot{\mathbf{x}} - \dot{\hat{\mathbf{x}}} \\ &= \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} - (\mathbf{A}\hat{\mathbf{x}} + \mathbf{B}\mathbf{u}) \\ &= \mathbf{A}(\mathbf{x} - \hat{\mathbf{x}}) \\ &= \mathbf{A}\mathbf{e} \end{aligned} \quad (11.58)$$

mit der Anfangsbedingung

$${}_0\mathbf{e} = {}_0\mathbf{x} - {}_0\hat{\mathbf{x}} \quad . \quad (11.59)$$

Dabei ist Gl.(11.58) strukturell identisch zu den in Kapitel 3 behandelten Systemen, sodass mit den dort vorgestellten Methoden die vollständige Lösung direkt als

$$\mathbf{e}(t) = e^{\mathbf{A}t} {}_0\mathbf{e} \quad (11.60)$$

angeschrieben werden kann.

Hieraus sieht man, dass der Schätzfehler nur dann gegen Null konvergiert, wenn die Systemdynamik \mathbf{A} stabil ist. Zusätzlich lässt sich die Geschwindigkeit der Konvergenz des Schätzfehlers offenbar nicht einstellen. Stattdessen erfolgt diese mit denselben Zeitkonstanten wie das beobachtete System und

damit stark verzögert. Dies wird für eine Regelung nicht ausreichend sein, da durch eine Regelung ein System in seinen dynamischen Eigenschaften verbessert werden soll. Dies wird auf Basis von veralteten Informationen selten gelingen.

Aufgrund der fehlenden Einstellbarkeit der Konvergenzgeschwindigkeit des Schätzfehlers ist der Beobachteraufbau aus Bild 11-9 also nicht geeignet. Dieser Nachteil röhrt daher, dass der gezeigte Aufbau strukturell einer Steuerung entspricht, da der Schätzfehler selbst nicht zurückgeführt wird. Die regelungstechnische Lösung besteht folglich darin, eine Rückführung vorzusehen, welche den geschätzten Zustand korrigiert.

11.4.2 Luenberger-Beobachter

Um den Beobachter zu vervollständigen, vergleicht man den durch den obigen Schätzer vorhergesagten Ausgang $\hat{\mathbf{y}} = \mathbf{C}\hat{\mathbf{x}}$ als „Regelgröße“ mit dem tatsächlichen Ausgang \mathbf{y} als „Führungsgröße“. Ergänzt man den Schätzer, also die Kopie der Regelstrecke, um eine lineare Rückführung mit Beobachtermatrix \mathbf{L} , so erhält man die in Bild 11-10 gezeigte Struktur des *Luenberger³-Beobachters*.

Luenberger-Beobachter

Gegeben sei eine lineare Regelstrecke mit Zustand $\mathbf{x}(t) \in \mathbb{R}^n$ und q Ausgangsgrößen. Ein Beobachter ist vom Typ des Luenberger-Beobachters, wenn er aus einer Kopie des Regelstreckenmodells mit einer zusätzlichen linearen Rückführung mit Beobachtermatrix $\mathbf{L} \in \mathbb{R}^{n \times q}$ besteht (siehe Bild 11-10).

Im Falle von nur einer Ausgangsgröße wird die Beobachtermatrix \mathbf{L} zu einem Spaltenvektor \mathbf{l} . Der Luenberger-Beobachter ist die grundlegendste aller Beobachterstrukturen und nahezu alle relevanten Beobachter können sich in seine oder eine sehr ähnliche Form gebracht werden, sodass er exemplarisch für einen linearen Beobachter schlechthin genommen werden kann. Berechnet man für den Luenberger-Beobachter die Dynamik des Schätzfeh-

³David Luenberger (*1937), amerikanischer Elektrotechniker [28]

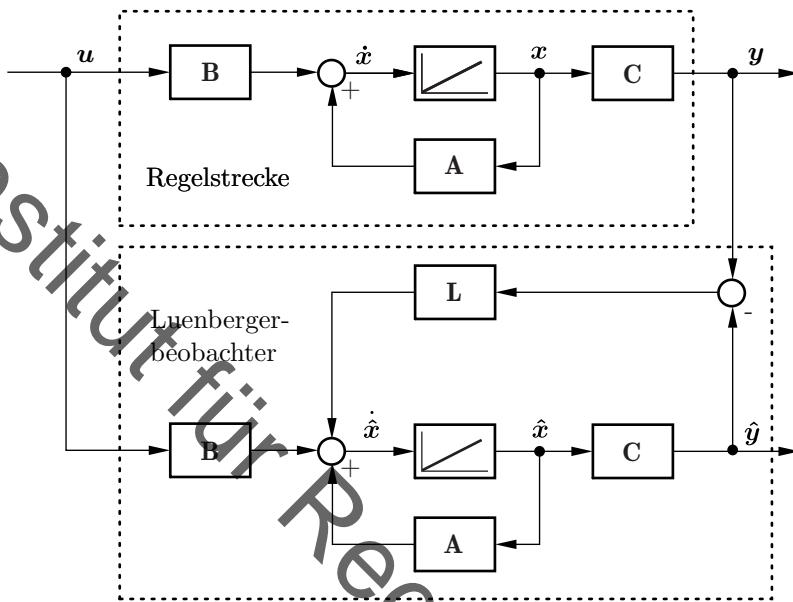


Bild 11-10: Luenberger-Beobachter als Kopie des Regelstreckenmodells mit zusätzlicher linearer Rückführung

lers, so erhält man wegen

$$\begin{aligned}\dot{\hat{x}} &= \mathbf{A}\hat{x} + \mathbf{B}u + \mathbf{L}(y - \hat{y}) \\ &= \mathbf{A}\hat{x} + \mathbf{B}u + \mathbf{L}(\mathbf{C}x + \mathbf{D}u - (\mathbf{C}\hat{x} + \mathbf{D}u)) \\ &= \mathbf{A}\hat{x} + \mathbf{B}u + \mathbf{L}\mathbf{C}e\end{aligned}\tag{11.61}$$

die Gleichung

$$\begin{aligned}\dot{e} &= \dot{x} - \dot{\hat{x}} \\ &= \mathbf{A}x + \mathbf{B}u - (\mathbf{A}\hat{x} + \mathbf{B}u + \mathbf{L}\mathbf{C}e) \\ &= (\mathbf{A} - \mathbf{L}\mathbf{C})e.\end{aligned}\tag{11.62}$$

Durch die zusätzliche Rückführung \mathbf{L} lassen sich die Eigenwerte des Schätzfehlers wegen Gl.(11.62) verschieben. Dabei ist Gl.(11.62) insbesondere un-

abhängig von den Eingangsgrößen \mathbf{u} . Die gewonnene Gleichung ähnelt dabei strukturell sehr der Gleichung, die beim Einsatz eines Zustandsregler entsteht:

$$\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{B}\mathbf{K})\mathbf{x} \Leftrightarrow \dot{\mathbf{e}} = (\mathbf{A} - \mathbf{LC})\mathbf{e} . \quad (11.63)$$

Der einzige Unterschied ist, dass die einzustellende Matrix \mathbf{K} bzw. \mathbf{L} beim Zustandsregler rechts, beim Beobachter links des Matrixprodukts mit \mathbf{B} bzw. \mathbf{C} steht. Da die Matrizenmultiplikation nicht kommutativ ist, lässt sich diese Reihenfolge nicht vertauschen und es ergeben sich strukturell andere Berechnungen.

Allerdings lässt sich mit einem einfachen Rechentrick der Entwurf eines Beobachters vollständig auf den Entwurf eines Zustandsreglers zurückführen, wie im Folgenden erläutert wird.

11.4.3 Beobachtbarkeit und Dualität

Hierzu bemerkt man, dass für die Vorgabe der gewünschten Dynamik an \mathbf{e} nur die Eigenwerte von $\mathbf{A} - \mathbf{LC}$ entscheidend sind. Da sich die Eigenwerte λ beim Transponieren einer Matrix nicht verändern, ergibt sich

$$\lambda(\mathbf{A} - \mathbf{LC}) = \lambda((\mathbf{A} - \mathbf{LC})^T) = \lambda(\mathbf{A}^T - \mathbf{C}^T \mathbf{L}^T) . \quad (11.64)$$

Duales Problem

Ein Luenberger-Beobachter \mathbf{L} kann für ein lineares System Σ mit den Zustandsraummatrizen $\mathbf{A}, \mathbf{B}, \mathbf{C}$ und \mathbf{D} genau dadurch entworfen werden, indem für das duale System $\tilde{\Sigma}$ ein Zustandsregler \mathbf{K} entworfen wird mit der Zuordnung

$$\tilde{\mathbf{A}} \equiv \mathbf{A}^T , \quad \tilde{\mathbf{B}} \equiv \mathbf{C}^T , \quad \mathbf{K} \equiv \mathbf{L}^T . \quad (11.65)$$

Das bedeutet, dass der Entwurf von Zustandsregler und Zustandsbeobachter genau äquivalent abläuft. Allgemein können alle Verfahren der Zustandsregelung in angepasster Form auch zur Zustandsbeobachtung genutzt werden, indem der Reglerentwurf für das duale System mit Systemmatrix \mathbf{A}^T und Eingangsmatrix \mathbf{C}^T durchgeführt wird. Ein naheliegendes Entwurfsverfahren für Zustandsbeobachter ist daher – ganz analog zum Vorgehen beim

Entwurf eines Zustandsreglers – eine Polvorgabe. Die Art der Berechnung kann dabei analog zur Polvorgabe für Zustandsregler erfolgen. Aufgrund der Transponierung der Matrizen bietet dabei die Regelungsnormalform keine Vorteile bei der Berechnung der Beobachterverstärkung \mathbf{L} mehr. Stattdessen ist es vorteilhaft, wenn das Paar $(\mathbf{A}^T, \mathbf{C}^T)$ in Regelungsnormalform vorliegt.

Beobachtungsnormalform

Ein System Σ mit den Zustandsraummatrizen \mathbf{A} , \mathbf{B} , \mathbf{C} und \mathbf{D} liegt genau dann in Beobachtungsnormalform vor, wenn das Paar $(\mathbf{A}^T, \mathbf{C}^T)$ in Regelungsnormalform vorliegt.

Die Beobachtungsnormalform hat im SISO-Fall mit $d = 0$ die Struktur

$$\dot{\mathbf{x}} = \underbrace{\begin{bmatrix} 0 & 0 & 0 & \cdots & -a_0 \\ 1 & 0 & 0 & & -a_1 \\ 0 & \ddots & & & \vdots \\ \vdots & 0 & 1 & 0 & -a_{n-2} \\ 0 & \cdots & 0 & 1 & -a_{n-1} \end{bmatrix}}_{\mathbf{A}} \mathbf{x} + \underbrace{\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n-2} \\ b_{n-1} \end{bmatrix}}_b u$$

$$y = \underbrace{\begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}}_{c^T} \mathbf{x} . \quad (11.66)$$

Der Beobachterentwurf ist offenbar für Systeme in Beobachtungsnormalform stets und einfach durchführbar. Führt man eine Polvorgabe für das duale System durch, so werden die entstehenden Gleichungssysteme unter allen Umständen genau dann lösbar sein, wenn das System *beobachtbar* ist.

Beobachtbarkeit

Ein System $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}$ und $y = \mathbf{Cx} + \mathbf{Du}$ ist genau dann beobachtbar, wenn man bei bekannter Eingangsgröße $\mathbf{u}(t)$ aus der Messung von $y(t)$ in endlicher Zeit t_e den beliebigen Anfangszustand ${}_0\mathbf{x} = \mathbf{x}(0)$ eindeutig bestimmen kann.

Man sagt dann auch, dass das Paar (\mathbf{A}, \mathbf{C}) beobachtbar ist.

Für Beobachtbarkeit gilt analog zur Steuerbarkeit, dass jede minimale Realisierung beobachtbar ist und somit auch diese Eigenschaft bei einer geeig-

neten Modellierung gegeben sein wird. Ist eine Regelstrecke nicht beobachtbar, so gibt es parallele Übertragungskanäle, deren Dynamik nicht voneinander unterschieden werden kann. Bild 11-11 zeigt verschiedene Beispiele von Systemen, die nicht steuerbar bzw. nicht beobachtbar sind.

Alle nicht steuerbaren oder nicht beobachtbaren Systeme sind keine minimalen Realisierungen. In diesem Sinne kann man eine minimale Realisierung auch wie folgt deuten:

Steuerbarkeit, Beobachbarkeit und minimale Realisierung

Durch die minimale Realisierung werden Zustände, die nicht steuerbar oder nicht beobachtbar sind, entfernt. Das ist möglich, da diese Zustände zwar für die Beschreibung der Eigenbewegung des Systems im Zustandsraummodell notwendig sein können, hingegen in der Beschreibung des Eingangs-Ausgangs-Verhaltens abkömmlig sind.

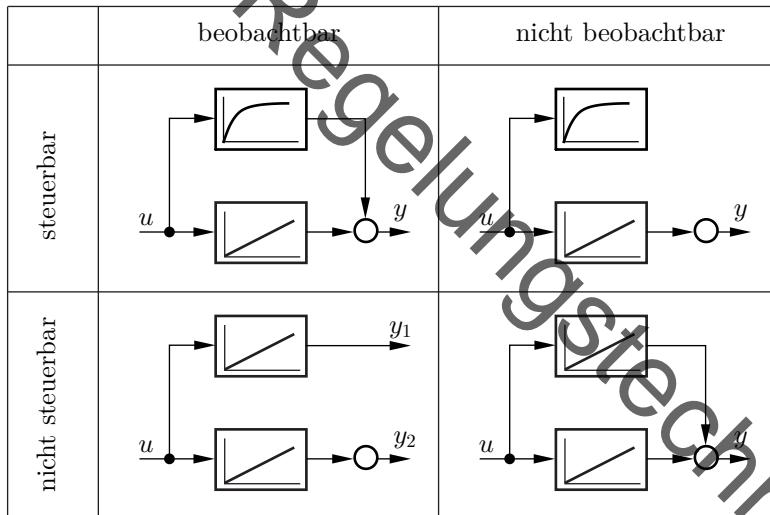


Bild 11-11: Beispiele für nicht steuerbare und nicht beobachtbare Systeme

Aufgrund der Äquivalenz von Zustandsregler und Zustandsrückführung ergeben sich nahezu wortgleiche Beweise und Überprüfungsformeln für Beobachtbarkeit. Ist das Paar (A, C) beobachtbar, so kann die Matrix L des

Luenberger-Beobachters so gewählt werden, dass $\mathbf{A} - \mathbf{LC}$ beliebige Eigenwerte besitzt. Weiter gilt das folgende Kriterium:

Beobachtbarkeitkriterium von Kalman

Das Paar (\mathbf{A}, \mathbf{C}) ist genau dann beobachtbar, wenn die sogenannte Beobachtbarkeitsmatrix

$$\mathbf{Q}_B = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \\ \vdots \\ \mathbf{CA}^{n-1} \end{bmatrix} \quad (11.67)$$

vollen Rang besitzt.

Alle weiteren Ausführungen erfolgen ebenfalls analog.

Der Entwurf eines Zustandsreglers und der Entwurf eines Zustandsbeobachters sind also für sich genommen handhabbare Rechnungen, die sich mit überschaubarem Aufwand durchführen lassen. Es ist aber an dieser Stelle noch unklar, inwiefern sich beide Entwürfe im Rahmen der Struktur in Bild 11-8 gegenseitig beeinflussen. Um dies zu untersuchen, ist es notwendig, das Gesamtsystem aus Bild 11-8 mit Zustandsregler, Regelstrecke und Zustandsbeobachter aufzustellen. Der Gesamtzustand ist dabei der Vektor $[\mathbf{x} \quad \mathbf{e}]^T$, um die Zustände der Regelstrecke \mathbf{x} aber auch den Schätzfehler \mathbf{e} zu erfassen. Die zusätzlich durch den Beobachter eingebrachten n Zustände \mathbf{e} entsprechen dabei $n-1$ zusätzlichen Polstellen des Reglers (Zustandsrückführung und Zustandsbeobachter), wodurch der Beobachter den aukausalen PD^{n-1} Zustandsregler kausalisiert.

Fügt man die Gleichungen

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu} \quad , \quad \mathbf{u} = -\mathbf{K}\hat{\mathbf{x}} \quad , \quad \dot{\hat{\mathbf{x}}} = \mathbf{A}\hat{\mathbf{x}} + \mathbf{Bu} + \mathbf{LC}(\mathbf{x} - \hat{\mathbf{x}}) \quad (11.68)$$

zusammen, so gewinnt man

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\mathbf{e}} \end{bmatrix} = \begin{bmatrix} \mathbf{A} - \mathbf{BK} & \mathbf{BK} \\ \mathbf{0} & \mathbf{A} - \mathbf{LC} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{e} \end{bmatrix} \quad . \quad (11.69)$$

Die Systemmatrix des Gesamtsystems ist also eine Blockdreiecksmatrix. Deinen Eigenwerten setzen sich aus den Eigenwerten der Diagonal-Blöcke – also

$\mathbf{A} - \mathbf{BK}$ und $\mathbf{A} - \mathbf{LC}$ – zusammen. Folglich beeinflusst der Beobachter die Eigenwerte des Zustandsrückgeführten Systems nicht und der Zustandsregler nicht die Eigenwerte des Schätzfehlers, was *Separationstheorem* genannt wird.

Separationstheorem

Die Eigenwerte eines Regelkreises, in dem eine Zustandsrückführung zusammen mit einem Beobachter eingesetzt wird, setzen sich zusammen aus der Vereinigung der Eigenwerte von

- $\mathbf{A} - \mathbf{BK}$, d. h. des Regelkreises mit Zustandsrückführung und ohne Beobachter
- $\mathbf{A} - \mathbf{LC}$, d. h. des Regelkreises mit Beobachter und ohne Zustandsrückführung

Man kann also die Zustandsrückführung und den Beobachter in getrennten Entwurfsschritten abhandeln.

Beim Entwurf eines Zustandsreglers inklusive Zustandsbeobachter geht man folglich zweischrittig vor. Zunächst legt man auf Basis der Anforderungen an den geschlossenen Regelkreis die Zustandsrückführung in Form der Matrix \mathbf{K} fest. Anschließend wählt man die Matrix \mathbf{L} des Zustandsbeobachters.

Damit der Beobachter die Leistungsfähigkeit des geschlossenen Regelkreises möglichst wenig beeinflusst, ist es notwendig, die Dynamik von $\mathbf{A} - \mathbf{LC}$ wesentlich schneller als die Dynamik von $\mathbf{A} - \mathbf{BK}$ zu wählen. Ansonsten wird der langsamer als der Regler agierende Beobachter die im ersten Entwurfsschritt erreichten Regelziele gefährden. Im Sinne der dominanten Polstellen müssen also die Eigenwerte des Schätzfehlers mit deutlich niedrigerem Realteil und deutlich größerem Betrag als die Eigenwerte des geschlossenen Regelkreises gewählt werden.

Auf den ersten Blick scheint es also angezeigt, die Eigenwerte der Matrix $\mathbf{A} - \mathbf{LC}$ möglichst weit in die linke komplexe Halbebene zu verschieben, da so der Beobachterfehler schneller konvergiert und der Einfluss des Beobachters auf den Zustandsregler immer geringer wird. Eine starke Verschiebung der Eigenwerte wird mit den gleichen Argumenten wie im Fall einer Zustandsrückführung über Polvorgabe dabei zu einer großen Beobachterverstärkung \mathbf{L} führen. Dies scheint zunächst unproblematisch, da das Argument, dass große Reglerverstärkungen wegen der großen Stellamplituden zu

vermeiden seien, hier keine Rolle spielt, da der Beobachter als „Stellgröße“ nur eine Korrektur des geschätzten Zustandes vornimmt, welche prinzipiell beliebige Werte annehmen kann.

Leider wird aber beim Beobachterentwurf die Beobachterverstärkung durch das Messrauschen nach oben beschränkt. Um das zu zeigen, werden die bisherigen Gleichungen um ein Messrauschen $n(t)$ in der Ausgangsgleichung

$$y = Cx + n \quad (11.70)$$

ergänzt. Hieraus ergibt sich für den Schätzfehler

$$\dot{e} = (\mathbf{A} - \mathbf{LC})e - \mathbf{Ln} \quad (11.71)$$

mit einem zusätzlichen partikulären Lösungsanteil hervorgerufen durch n .

Liegen die Eigenwerte des Schätzfehlers sehr weit von den Streckeneigenwerten entfernt und ist folglich eine große Beobachterverstärkung \mathbf{L} notwendig, so beeinflusst das Messrauschen den Beobachtungsfehler sehr stark. Wird eine geringe Verstärkung \mathbf{L} gewählt, so bleibt der Einfluss des Messrauschens gering; allerdings können dann die Eigenwerte möglicherweise nicht weit genug links in die komplexe Ebene verschoben werden, sodass die Regelung negativ beeinflusst wird.

Der Entwurf von Beobachter und Zustandsregler erfordert also eine enge Abstimmung beider Entwürfe, obgleich sie prinzipiell unabhängig voneinander sind. Als Richtwert hat sich dabei folgende Faustformel etabliert:

Eigenwerte des Schätzfehlers

Die Eigenwerte des Schätzfehlers sollten um den Faktor 3 bis 10 weiter links als die dominanten Pole des geschlossenen Regelkreises gewählt werden, wobei geringere Faktoren bei stärkerem Messrauschen anzuraten sind.

Der Zusammenhang zwischen dem Beobachterfehler und dem Messrauschen lässt es auch zu, Beobachter als Filter zu interpretieren. In der Elektrotechnik ist es sogar üblich, Beobachter allgemein als Filter zu bezeichnen. Auch das in Kapitel 15 diskutierte Kalmanfilter ist ein Beobachter. Bei dieser Deutung des Beobachters begreift man diesen als ein System mit Eingang

y und Ausgang \hat{y} . Die Übertragungsfunktion zwischen diesen Größen berechnet sich dann zu

$$G(s) = \frac{\hat{Y}(s)}{Y(s)} = \mathbf{c}^T (s\mathbf{I} - \mathbf{A} + \mathbf{l}\mathbf{c}^T)^{-1} \mathbf{l} \quad (11.72)$$

und beschreibt, welchen Einfluss der gemessene, verrauschte Ausgang y auf den geschätzten Ausgang \hat{y} hat.

Offenbar handelt es sich bei $G(s)$ um eine ganz normale Übertragungsfunktion mit Polstellen λ , die mit den Eigenwerten von $\mathbf{A} - \mathbf{l}\mathbf{c}^T$ identisch sind. Als Filter wird $G(s)$ folglich Signalanteile, die größer als die Grenzfrequenz von $G(s)$ bei $-1/\lambda$ sind, herausfiltern.

Beobachter als Filter

Man kann Beobachter als modellbasierte Filter einsetzen, die eine fehlerbehaftete Messung auf Basis eines vorhergesagten Verlauf korrigieren.

Das Ergebnis ist somit eine fehlerbereinigte Schätzung \hat{y} . Eine große Verstärkung \mathbf{l} verringert das Abdämpfen des Messrauschens. Eine zu geringe Verstärkung \mathbf{l} wird aber zu einem großen Vertrauen in die Modellprädiktion führen, weswegen Modellabweichungen, die in der Praxis immer gegeben sind, dazu führen, dass das gefilterte \hat{y} nur mit einem großen Phasenverzug der Messung y folgen wird – vergleiche auch Abschnitt 5.5.

11.4.4 Beispiel

Der Gesamtentwurf von Zustandsregler und Zustandsbeobachter soll an einem Beispiel veranschaulicht und bewertet werden, wobei der Zustandsregler für das inverse Pendel erneut aufgegriffen wird. Die betrachteten Gleichungen sind

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{\vartheta} \\ \ddot{\vartheta} \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 \\ \frac{g}{L} & -\frac{\mu}{ML} \end{bmatrix}}_{\mathbf{A}} + \underbrace{\begin{bmatrix} 0 \\ \frac{1}{ML^2} \end{bmatrix}}_b u \quad (11.73)$$

$$p(\lambda) = (\lambda + 1 - j)(\lambda + 1 + j) = \lambda^2 + 2\lambda + 2$$

$$\mathbf{k}^T = [gL M + 2ML^2 \quad 2ML^2 - \mu L] \quad .$$

Die Polstellen des geschlossenen Regelkreises liegen bei $-1 \pm j$. Somit handelt es sich um ein dominantes komplex konjugiertes Polstellenpaar und die Eigenwerte des Schätzfehlers sollten in Betrag und Realteil wesentlich größer als dieses Polpaar sein.

Unter der Annahme von typischem Rauschen und der Forderung nach einem nicht schwingungsfähigen Schätzfehler bietet sich eine Wahl der Beobachterpole in der Größenordnung von $-7\sqrt{2} \approx -10$ an. Es berechnet sich

$$\mathbf{A} - \mathbf{l}\mathbf{c}^T = \begin{bmatrix} 0 & 1 \\ \frac{g}{L} & -\frac{\mu}{ML} \end{bmatrix} - \begin{bmatrix} l_1 \\ l_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} = \begin{bmatrix} -l_1 & 1 \\ \frac{g}{L} - l_2 & -\frac{\mu}{ML} \end{bmatrix} \quad (11.74)$$

und damit für die Eigenwerte

$$\det(\lambda\mathbf{I} - \mathbf{A} + \mathbf{l}\mathbf{c}^T) = \lambda^2 + \left(l_1 + \frac{\mu}{ML}\right)\lambda + \frac{\mu l_1}{ML} + l_2 - \frac{g}{L} \stackrel{!}{=} \lambda^2 + 20\lambda + 100, \quad (11.75)$$

woraus über Koeffizientenvergleich die Beobachterverstärkung

$$\mathbf{l}^T = [l_1 \ l_2] = \left[20 - \frac{\mu}{ML} \quad 100 + \frac{g}{L} - \frac{20\mu}{ML} + \frac{\mu^2}{M^2 L^2} \right] \quad (11.76)$$

ermittelt werden kann.

Offenbar ist die Verstärkung l_2 wesentlich größer als die Verstärkung l_1 . Das ist logisch, da der erste Zustand ϑ gemessen werden kann und daher geringerer Korrekturen bedarf.

Im geschlossenen Regelkreis ergeben sich bei einer Startauslenkung von $\vartheta(0) = \pi/4$ die in Bild 11-12 gezeigten Verläufe für die Ausgangsgröße und Zustände. Dabei ist in Bild 11-12 links die Ausgangsgröße mit dem Zustandsregler ohne Beobachter y_{oB} gestrichelt, inklusive Beobachter y_{mB} durchgezogen und die verrauschte Messung $y_{m,mB}$ gepunktet gezeigt. Der Beobachter arbeitet mit der Startschätzung $\hat{\vartheta}(0) = 0$.

Es ist gut zu erkennen, dass sich die dynamischen Verläufe y_{oB} und y_{mB} relativ ähnlich sind und eine vergleichbare Regelgüte erreichen. Die Konvergenz des Schätzfehlers ist in Bild 11-12 auf der rechten Seite zu sehen, wo gestrichelt die tatsächlichen und durchgezogen die geschätzten Verläufe dargestellt sind. Man sieht, dass der geschätzte Zustand $\hat{\vartheta}$ trotz der großen

Abweichung bei der Startschätzung schnell gegen den tatsächlichen Zustand ϑ konvergiert. Auch ein Einfluss von Messrauschen ist kaum zu sehen, was die filternde Wirkung des Beobachter unterstreicht. Hingegen braucht die Ableitung von $\hat{\vartheta}$ länger um zu konvergieren und auch das Rauschen ist dort deutlicher zu sehen. Das liegt daran, dass l_2 wesentlich größer als l_1 ist. Diese Grenzfrequenz des Filters ist daher größer, weswegen weniger Signaleanteile des Rauschens gedämpft werden. Dennoch ist der Einfluss von Rauschen nur unwesentlich in der Regelgröße zu erkennen.

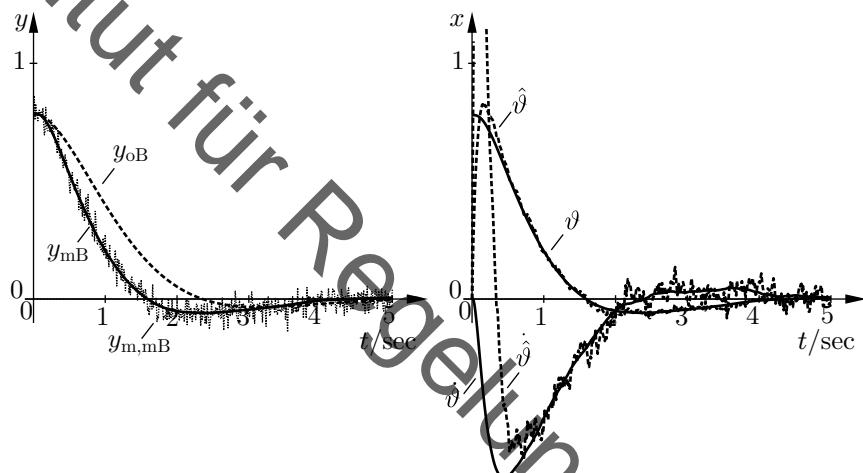


Bild 11-12: Verläufe des inversen Pendels im geschlossenen Regelkreis mit und ohne Beobachter

11.5 Wurzelortskurven

11.5.1 Grundidee

Bei der Polvorgabe wurde das Erreichen des durch die Regelziele beschriebenen Zielgebiets innerhalb der komplexen Ebene dadurch sichergestellt, dass Eigenwerte innerhalb des Zielgebiets ausgewählt wurden. Dieses Vorgehen nimmt durch die Vorgabe der Eigenwerte unnötige Beschränkungen in Kauf. So könnte es sein, dass bei leicht anders vorgegebenen Eigenwerten eine simplere Reglerstruktur möglich oder das Regelziel anderweitig einfacher zu erreichen wäre.

An dieser Stelle setzt das Wurzelortskurven-Verfahren an, welches als halbgraphisches Verfahren versucht, auf Vorgabe von konkreten Positionen innerhalb der Zielgebiete zu verzichten und stattdessen geeignete Regler aus einer näherungsweisen Bestimmung der Polstellen des geschlossenen Regelkreises für verschiedene Reglerverstärkungen abzuleiten.

Hierzu wird erneut die Störübertragungsfunktion $S(s) = 1/(1 + G_0(s))$ betrachtet. Die für Aussagen zur Stabilität oder Instabilität wichtigen Polstellen der Übertragungsfunktion sind die Nullstellen der charakteristischen Gleichung

$$1 + G_0(s) = 0 \quad (11.77)$$

und damit die Lösungen der Gleichung

$$G_0(s) = -1 \quad . \quad (11.78)$$

Wann nimmt aber eine komplexwertige Funktion $G_0(s)$ den Wert -1 an? Hierzu zerlegt man Gl.(11.78) in zwei reelle Gleichungen, nämlich in eine für den Betrag und eine für den Phasenwinkel.

$$|G_0(s)| = -1 \quad , \quad \varphi_0(s) = -180^\circ \pm n \cdot 360^\circ \quad , \quad (11.79)$$

die beide für ein s erfüllt sein müssen.

Der Trick des Wurzelortskurven-Verfahrens besteht nun darin, beide Gleichungen separat voneinander zu behandeln. Hierzu schreibt man – Totzeiten können erst einmal nicht berücksichtigt werden – die Übertragungsfunktion

$G_0(s)$ in der Form von Gl.(4.80) als Produkt der Pol- und Nullstellen:

$$G_0(s) = K \underbrace{\frac{(s - \eta_1)(s - \eta_2) \dots (s - \eta_m)}{(s - \lambda_1)(s - \lambda_2) \dots (s - \lambda_n)}}_{G'_0(s)} . \quad (11.80)$$

Betrags- und Winkelbedingung

Mit der Zerlegung in Gl.(11.80) zerfällt die Bedingung $G_0(s) = -1$ in die *Betragsbedingung*

$$|K||G'_0(s)| = 1 \quad (11.81)$$

und die *Winkelbedingung*

$$\varphi_0(s) = \begin{cases} \pi + q \cdot 2\pi & \text{für } K > 0 \\ 0 + q \cdot 2\pi & \text{für } K < 0 \end{cases} \quad q = 0, \pm 1, \pm 2, \dots . \quad (11.82)$$

Man erkennt, dass der Faktor K in der Betragsbedingung lediglich als Vorfaktor und in der Winkelbedingung nur mit seinem Vorzeichen auftritt. Dies kann man dazu nutzen, Aussagen über die Eigenschaften des geschlossenen Regelkreises für unterschiedliche Werte des Faktors K der Übertragungsfunktion des aufgeschnittenen Kreises zu gewinnen.

Hierzu stellt man zunächst fest, dass die Polstellen des geschlossenen Regelkreises diejenigen Werte von s sind, die sowohl die Betrags- als auch die Winkelbedingung erfüllen. Aber findet man ein s^* , welches die Winkelbedingung erfüllt, so findet man durch ein Auflösen der Betragsbedingung nach K gemäß

$$K = \pm \frac{1}{|G'_0(s^*)|} \quad (11.83)$$

stets ein K , für welches auch die Betragsbedingung erfüllt ist. Das Vorzeichen von K ist dabei durch die Winkelbedingung bereits festgelegt. Daraus folgt:

Wurzelortskurve

Es gibt für jedes s^* , welches die Winkelbedingung erfüllt, ein K , sodass s^* eine Polstelle des geschlossenen Regelkreis ist.

Verbindet man alle s^* , welche die Winkelbedingung für $K > 0$ oder $K < 0$ erfüllen, zu einem Linienzug, so erhält man die *Wurzelortskurve*.

Folglich zeigen die Wurzelortskurven die Positionen der Pole des geschlossenen Regelkreises für verschiedene Wert für K . Daher ist – im Gegensatz zur Ortskurve mit Laufparameter ω – die Verstärkung K der Laufparameter der Wurzelortskurve. Da der Faktor K in der Winkelbedingung nicht erscheint, sind die Wurzelortskurven ansonsten von ihm unabhängig.

Die Wurzelortskurven eignen sich gut zur Reglerauslegung, weil sie sich mit relativ überschaubaren Aufwand skizzieren lassen. Hierzu ist zunächst festzuhalten, dass prinzipiell lediglich die als bekannt vorausgesetzten Null- und Polstellen der Übertragungsfunktion $G_0(s)$ für die Konstruktion der Wurzelortskurve benötigt werden.

Für die Konstruktion von Wurzelortskurven gibt es Regeln, die auf Evans⁴ zurückgehen. Hierzu schreibt man Gl.(11.80) komplett in Betrag und Phasenwinkel aus:

$$\begin{aligned} G'_0(s) &= \frac{(s - \eta_1)(s - \eta_2) \cdot \dots \cdot (s - \eta_m)}{(s - \lambda_1)(s - \lambda_2) \cdot \dots \cdot (s - \lambda_n)} \\ &= \frac{|s - \eta_1|e^{j\psi_1}|s - \eta_2|e^{j\psi_2} \cdot \dots \cdot |s - \eta_m|e^{j\psi_m}}{|s - \lambda_1|e^{j\varphi_1}|s - \lambda_2|e^{j\varphi_2} \cdot \dots \cdot |s - \lambda_n|e^{j\varphi_n}} \\ &= \frac{\prod_{i=1}^m |s - \eta_i|}{\prod_{i=1}^n |s - \lambda_i|} e^{j(\sum_{i=1}^m \psi_i - \sum_{i=1}^n \varphi_i)} . \end{aligned} \quad (11.84)$$

Dabei ist mit ψ_i bzw. φ_i der zu der komplexen Zahl $s - \eta_i$ bzw. $s - \lambda_i$ gehörende Winkel bezeichnet.

Damit lautet die Betragsbedingung

$$|K||G'_0(s)| = |K| \frac{\prod_{i=1}^m |s - \eta_i|}{\prod_{i=1}^n |s - \lambda_i|} \stackrel{!}{=} 1 \quad (11.85)$$

⁴Walter Richard Evans (1920-1999), amerikanischer Regelungstechniker [12]

und die Winkelbedingung

$$\varphi_0(s) = \sum_{i=1}^m \varphi_{Ni} - \sum_{i=1}^n \varphi_{Pi} \stackrel{!}{=} \begin{cases} \pi + q \cdot 2\pi & \text{für } K > 0 \\ 0 + q \cdot 2\pi & \text{für } K < 0 \end{cases} \quad (11.86)$$

$$q = 0, \pm 1, \pm 2, \dots .$$

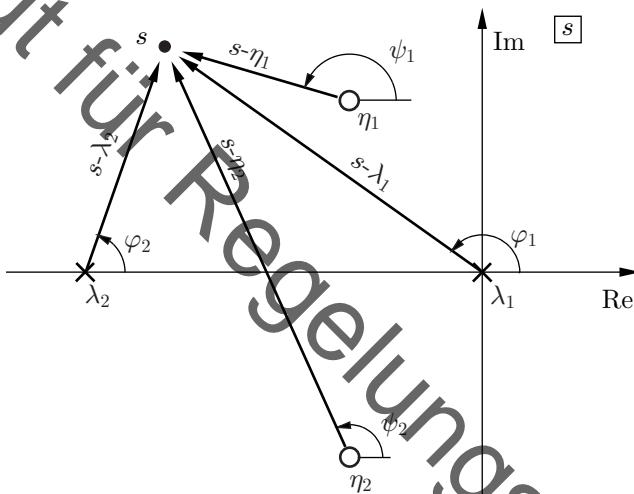


Bild 11-13: Prüfen der Winkelbedingung in der s -Ebene

Wenn man für jeden Punkt der komplexen s -Ebene die Winkelsumme φ_0 aufstellt und gemäß der Winkelbedingung prüft, kann man grundsätzlich die Punkte ermitteln, die zu Wurzelortskurven gehören. Man kann dies graphisch durchführen (Bild 11-13), indem man die als bekannt vorausgesetzten Nullstellen (\circ) und Polstellen (\times) in eine komplexe Ebene einträgt. Die Ausdrücke $s - \eta_i$ bzw. $s - \lambda_i$ sind komplexe Zahlen, die durch Vektoren dargestellt werden, die von der Null- bzw. Polstelle zu dem jeweils betrachteten Punkt s weisen. Die Winkel φ_i und ψ_i müssen ausgemessen

und zur Summe φ_0 nach Gl.(11.86) addiert werden. Der Punkt s in Bild 11-13 erfüllt die Winkelbedingung offenbar nicht.

11.5.2 Konstruktionsregeln

Anstelle dieses recht mühsamen graphischen Verfahrens, die Winkelsumme Gl.(11.86) für viele Punkte der s -Ebene probeweise zu ermitteln, benutzt man allgemein von Evans formulierte Regeln zur Konstruktion von Wurzelortskurven (WOKn). Diese Konstruktionsregeln sind exakt und ermöglichen eine Darstellung der Wurzelortskurve in einem relativ großen K -Bereich. Dieser Bereich wird dann durch eine skizzenhafte Verbindung zu $K > 0$ oder $K < 0$ erweitert.

Diese Kombination von exakter Konstruktion und skizzenhafter Vervollständigung ermöglicht – auch weil viele Konstruktionsregeln sehr einfach sind – eine schnelle überschlagige Bestimmung der Wurzelortskurve. Dies eignet sich insbesondere für das Abschätzen der richtigen Reglerstruktur. Zur konkreten Reglerauslegung wird üblicherweise die exakte Wurzelortskurve mittels Rechnerunterstützung bestimmt.

Zu den wichtigsten Konstruktionsregeln zählen die folgenden.

1. Die WOKn beginnen (mit $K = 0$) in den Polen (\times) und enden (mit $K \rightarrow \infty$) in den Nullstellen (\circ) von $G_0(s)$ oder im Unendlichen.
2. Für ein $G_0(s)$ mit relativem Grad r enden r Äste der WOK im Unendlichen.
3. Jeder Punkt auf der reellen Achse, auf dessen rechter Seite die Summe von Polen und Nullstellen im Fall $K > 0$ eine ungerade, im Fall $K < 0$ eine gerade Zahl ergibt, ist Wurzelort.
4. WOKn verlaufen symmetrisch zur reellen Achse und sind nur von der relativen Lage der Pole und Nullstellen abhängig.
5. Die r nach ∞ laufenden Äste haben Asymptoten, die sich im Wurzelschwerpunkt

$$s_W = \frac{\sum_{i=1}^n \lambda_i - \sum_{i=1}^m \eta_i}{r} \quad (11.87)$$

auf der reellen Achse schneiden und unter den Winkeln

$$\alpha_i = \begin{cases} \frac{2i-1}{r} 180^\circ & \text{für } K > 0 \\ \frac{2i}{r} 180^\circ & \text{für } K < 0 \end{cases} \quad i = 1, 2, \dots, r \quad (11.88)$$

verlaufen.

6. WOKn können sich schneiden, was mehrfachen Polstellen des geschlossenen Regelkreises entspricht. Die Schnittpunkte werden Verzweigungspunkte s_A genannt. Die notwendige Bedingung für die Verzweigungspunkte lautet

$$\sum_{i=1}^n \frac{1}{s_A - \lambda_i} = \sum_{i=1}^m \frac{1}{s_A - \eta_i} \quad . \quad (11.89)$$

Wenn sich nur zwei WOK-Aste schneiden, so geschieht das unter rechtem Winkel.

7. Der Winkel, unter dem die WOK aus einer einfachen Polstelle austritt oder in eine einfache Nullstelle einmündet, lässt sich durch Anwenden der Winkelbedingung auf einen Punkt, der der betrachteten Pol- oder Nullstelle beliebig nahe liegt, gewinnen. Für den Austrittswinkel an einem einfachen Pol λ_1 gilt

$$\varphi_1 = \sum_{i=1}^m \psi_i - \sum_{i=2}^n \varphi_i - \begin{cases} \pi + q2\pi & \text{für } K > 0 \\ 0 + q2\pi & \text{für } K < 0 \end{cases} \quad (11.90)$$

$$q = 0, \pm 1, \dots$$

und für den Eintrittswinkel an einer einfachen Nullstelle η_1

$$\psi_1 = \sum_{i=1}^n \varphi_i - \sum_{i=2}^m \psi_i + \begin{cases} \pi + q2\pi & \text{für } K > 0 \\ 0 + q2\pi & \text{für } K < 0 \end{cases} \quad (11.91)$$

$$q = 0, \pm 1, \dots .$$

8. Der an einem Punkt s der WOK anzutragende Wert K ergibt sich aus der Betragsbedingung zu

$$|K| = \frac{\prod_{i=1}^n |s - \lambda_i|}{\prod_{i=1}^m |s - \eta_i|} . \quad (11.92)$$

Aus der Gl.(11.88) für die Winkel α_i der Asymptoten in Regel 5 erhält man für $r = 1, \dots, 4$ und $K > 0$ die in Bild 11-14 dargestellten Anordnungen der Asymptoten.

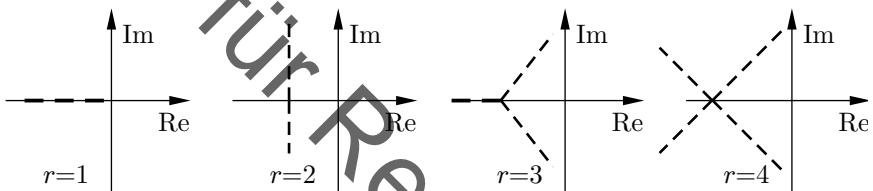


Bild 11-14: Asymptotenanordnungen für $K > 0$

Zur Ergänzung dieser konstruierten Teile der WOK zu einer Gesamtskizze eignet sich vor allem folgende Analogie:

Skizzen von Wurzelortskurven

Die WOKn können als Stromfäden in einer ebenen Potentialströmung mit den Polen als Quellen und den Nullstellen als Senken aufgefasst werden oder als Bahnkurven geladener Teilchen in einem elektrostatischen Feld, in dem die Teilchen von den Nullstellen angezogen und von den Polstellen abgestoßen werden.

Bild 11-15 zeigt als Beispiel die Wurzelortskurven für einen Regelkreis mit einem reellen und einem Paar konjugiert komplexer Pole. Eine Aussage zur Stabilität von Regelkreisen lässt sich wie folgt formulieren:

Stabilität mittels Wurzelortskurven

Der geschlossene Regelkreis für ein vorgegebenes K ist genau dann stabil, wenn alle zu K gehörenden Punkte der WOKn in der linken offenen s -Halbebene liegen.

Etwaige kritische Werte des Faktors K ergeben sich also an den Stellen, an denen WOKn die imaginäre Achse schneiden. Regelkreise, für die WOKn nur in der linken offenen s -Halbebene verlaufen, sind für alle Werte $0 < K < \infty$ stabil.

Wenn der Regelkreis außer stabil auch noch gut gedämpft und nicht zu träge arbeiten soll, so sollte man den Faktor K so wählen, dass die zugehörigen Punkte der WOKn in die entsprechenden Zielgebiete wie in Bild 11-7 wandern und nicht zu nahe an der imaginären Achse liegen.

In Bild 11-15 ist beispielhaft zu erkennen, dass mit wachsendem Übertragungsfaktor K der Regelkreis immer schlechter gedämpft und dann auch instabil wird.

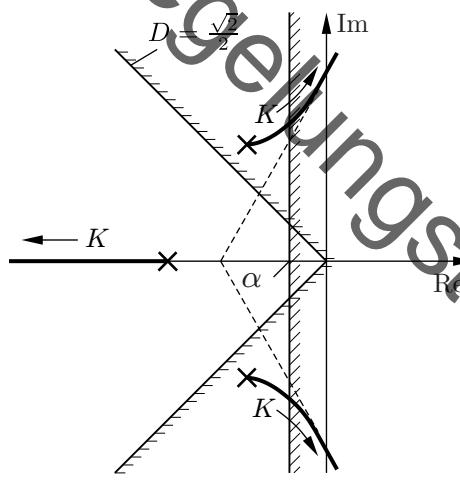


Bild 11-15: Wurzelortskurven für drei Polstellen, $K > 0$

Das Wurzelortskurvenverfahren ist ein sehr wirksames Mittel, um die „Wan-

derung“ von Polstellen unter dem Einfluss von Parametervariationen eines Regelkreises darzustellen. Man sollte dabei nicht vergessen, dass die Dynamik eines Übertragungssystems nicht nur von den Polstellen sondern auch von den Nullstellen seiner Übertragungsfunktion abhängt.

Bei der Auswertung von WOKn ist zu beachten, dass das zugehörige Koordinatensystem mit der Dimension $(\text{Zeit})^{-1}$ entsprechend der Dimension der Variablen s versehen ist und dass der Faktor K gemäß Gl.(11.80) definiert ist und dadurch die Dimension $(\text{Zeit})^{m-n}$ hat. Ferner ist das Vorzeichen von K nicht nur von den Vorzeichen an den Summenpunkten im Wirkungsplan des Regelkreises, sondern auch von evtl. vorhandenen Polen und Nullstellen in der rechten s -Halbebene abhängig. Daher ist dringend zu empfehlen, die Übertragungsfunktion des aufgeschnittenen Regelkreises auf die Form der Gl.(11.80) zu bringen, um Vorzeichen, Dimension und Zahlenwerte von K richtig verarbeiten zu können.

11.5.3 Beispiel

Die Handhabung des Wurzelortskurven-Verfahrens soll abschließend am Beispiel der Reglerentwurfs für ein instabiles System erster Ordnung illustriert werden. Der geschlossene Regelkreis soll stationär genau arbeiten und eine Dämpfung von mindestens $D = \sqrt{2}/2$ aufweisen.

Aufgrund der Forderung nach stationärer Genauigkeit wird zunächst untersucht, ob ein I-Regler (als einfachst möglicher Regler mit integrierendem Anteil) hierfür geeignet ist. Hiermit ergibt sich der aufgeschnittene Regelkreis zu

$$G_0(s) = \frac{K_I}{s} \frac{-1}{1 - sT} \quad . \quad (11.93)$$

Wenn die Übertragungsfunktion in die Form der Gl.(11.80) gebracht wird, erhält man

$$G_0 = - \underbrace{\frac{K_I}{(-T)}}_K \frac{1}{(s - 0)(s - \frac{1}{T})} \quad . \quad (11.94)$$

Zur Ermittlung der Wurzelortskurve trägt man nun zunächst die beiden Polstelle der Übertragungsfunktion des aufgeschnittenen Regelkreises ein.

Nach Regel 3 ist die reelle Achse von $+1/T$ bis 0 Teil der WOK, da K positiv ist. Da der relative Grad zwei beträgt, laufen die WOK-Äste nach Regeln 2 ins Unendliche entsprechend der Asymptoten nach Regel 5. Der Asymptotenschwerpunkt s_W und die Verzweigungspunkte s_A ergeben sich zu

$$s_W = \frac{0 + 1/T}{2} = \frac{1}{2T} \quad , \quad \frac{1}{s_A} + \frac{1}{s_A - 1/T} = 0 \Rightarrow s_A = \frac{1}{2T} \quad (11.95)$$

und sind damit in diesem Sonderfall identisch. Auf Basis dieser Berechnung ergibt sich die in Bild 11-16 gezeigte Wurzelortskurve.

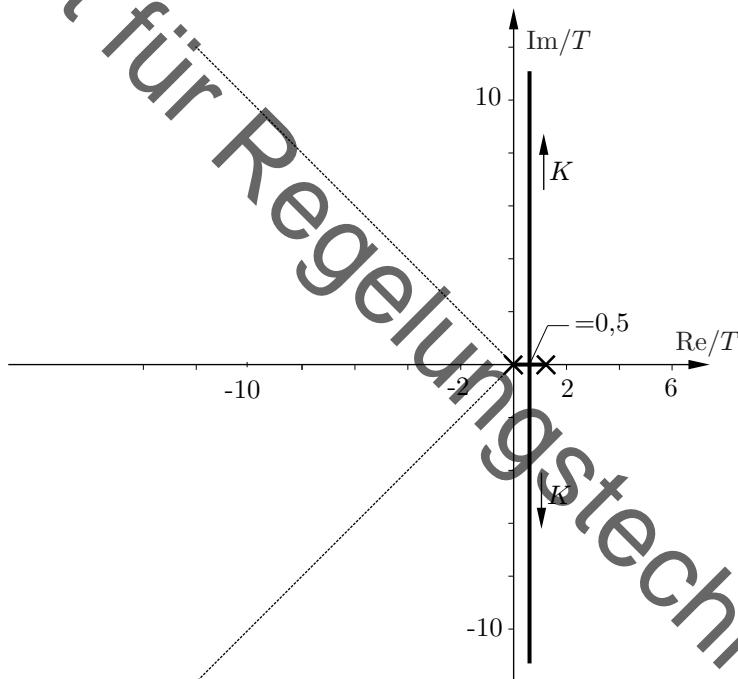


Bild 11-16: Wurzelortskurve zu $G_0 = \frac{K_I}{s} \frac{-1}{1 - sT}$

Offenbar verbleiben beide Äste der Wurzelortskurve für alle K in der rechten offenen Halbebene. Der I-Regler ist folglich keine geeignete Reglerstruk-

tur, da er die instabile Regelstrecke für keine Parametereinstellung stabilisieren kann.

Um die Äste der Wurzelortskurve in die linke Halbebene zu bewegen, bietet es sich – in Analogie zur ebenen Potentialströmung – an, mit dem Regler eine zusätzliche Nullstelle in die linke Halbebene einzuführen, da Nullstellen „anziehend“ wirken. Dies entspricht der Reglerstruktur eines PI-Reglers. Für die Wahl der Nachstellzeit $T_n = 0,1T$ ergibt sich in der Form Gl.(11.80)

$$G_0 = -\underbrace{\frac{K_1 0,1T}{(-T)}}_K \frac{s + \frac{1}{0,1T}}{(s - 0)(s - \frac{1}{T})} . \quad (11.96)$$

In Bild 11-17 sind erneut zunächst die zwei Pol- und die eine Nullstelle des aufgeschnittenen Regelkreises eingetragen worden. Nach Regel 3 ist nun die reelle Achse von $+1/T$ bis 0 und links von $-10T$ Wurzelort. Damit ist auch schon der eine nach ∞ verlaufende WOK-Ast nach Regel 2 und dessen Asymptote nach Regel 5 bestimmt worden.

Bereits jetzt kann man erkennen, dass nun der geschlossene Regelkreis für kleine Werte von K und damit für kleine K_1 instabil und für genügend große Werte stabil sein wird.

Da nach Regel 1 die WOKn in den Polen beginnen und in den Nullstellen enden sollen, fehlt noch ein Verbindungsstück zwischen den Teil-WOKn auf der reellen Achse. Dieses Stück muss außerhalb der reellen Achse liegen. Also ist ein Verzweigungspunkt nach Regel 6 zu bestimmen, wobei Gl.(11.89) liefert:

$$s_A = 10 \frac{-1 \pm \sqrt{1,1}}{T} \approx \frac{0,5}{T} \text{ und } \frac{-20,5}{T}. \quad (11.97)$$

Die in den Polstellen startenden Äste laufen in den Verzweigungspunkt bei $0,5/T$, wo sie im rechten Winkel die reelle Achse verlassen und in den Verzweigungspunkt bei $-20,5/T$ laufen. Dort läuft der eine Ast nach $-\infty$ und der andere in die Nullstelle. Tatsächlich kann man zweigen, dass in diesem konkreten Fall die Bahn zwischen den Verzweigungspunkten einem Kreis um die Nullstelle entspricht.

Aus der Zeichnung der WOK in Bild 11-17 sieht man, dass beide Äste

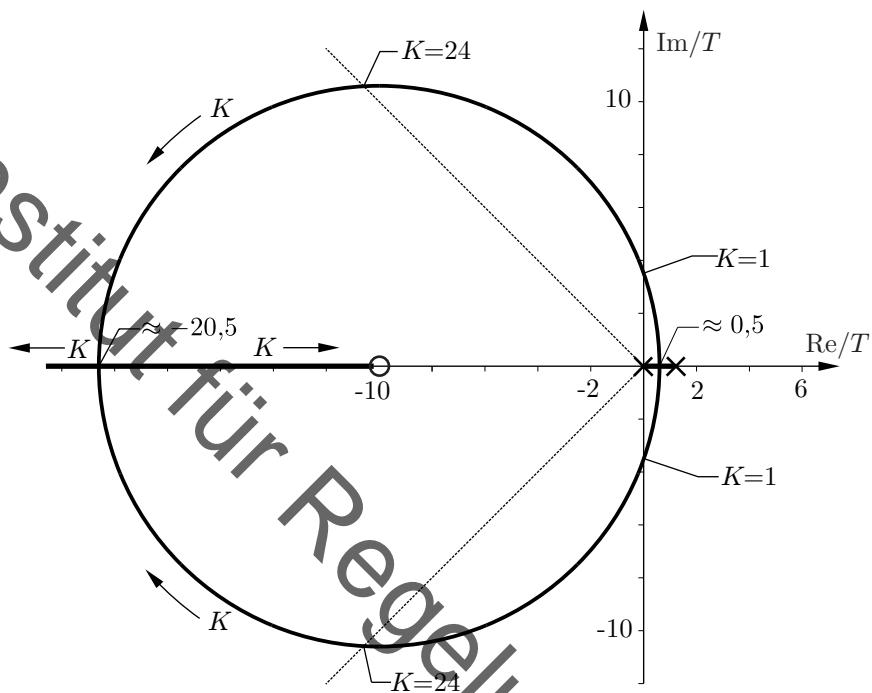


Bild 11-17: Wurzelortskurve zu $G_0 = \frac{K_I 1 + s 0,1T}{s 1 - sT}$

der WOK in den Zielbereich mit einer Dämpfung $D > \sqrt{2}/2$ laufen. Ein PI-Regler ist also zu Erreichung der formulierten Anforderungen geeignet.

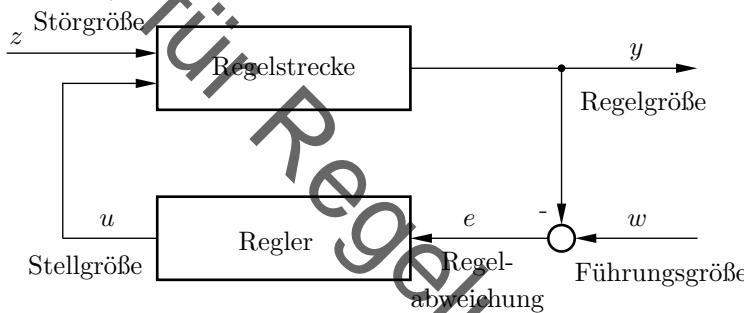
Über die Wurzelortskurve lässt sich auch die konkrete Reglerparametrierung bestimmen. Hierzu wird der Schnittpunkt der WOK mit der ersten Winkelhalbierenden bestimmt und dort nach Regel 8 die Betragsbedingung ausgewertet, wobei man $K_{\text{krit}} = 24 \text{ sec}^{-1}$ erhält. Die passende Verstärkung K_I gewinnt man dann über Gl.(11.96) zu

$$K_{I\text{krit}} = 10K_{\text{krit}} = 240 \text{ sec}^{-1} \quad . \quad (11.98)$$

12 Vermaschte Regelkreise

12.1 Erweiterung des Einfachregelkreises

Die in Bild 1-5 gezeigte und hier nochmals wiedergegebene Struktur des Einfachregelkreises lässt sich unter bestimmten Umständen durch zusätzlich einzuführende Wirkungslinien und Übertragungsglieder hinsichtlich der Erfüllung der Regelziele wesentlich verbessern.

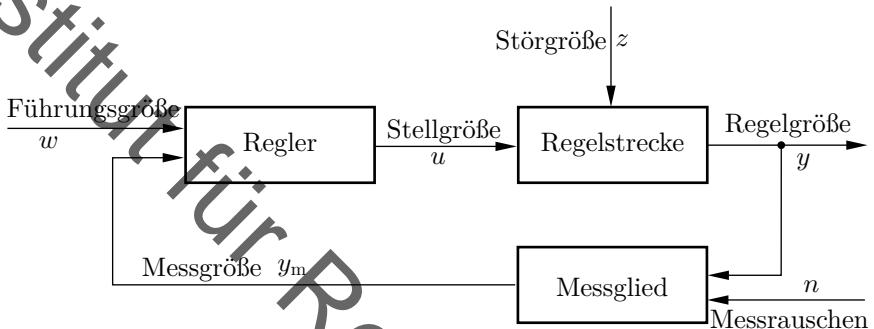


Im Folgenden sollen einige solche strukturellen Maßnahmen erläutert werden, die insbesondere besseres Führungsverhalten und bessere Störunterdrückung erzielen.

Diese Maßnahmen sind i. Allg. nur unter bestimmten Voraussetzungen anwendbar und bedingen oft einen zusätzlichen Geräteaufwand, der die Regelungseinrichtung verteuert. Andererseits sind sehr viele technische Prozesse mit einer einschleifigen Regelung überhaupt nicht oder nicht mit der erforderlichen Genauigkeit zu regeln. Die vorzustellenden Maßnahmen unterscheiden sich dabei maßgeblich darin, ob für ihre Umsetzung zusätzlichen Sensoren und/oder zusätzliche Aktoren benötigt werden. Neben den vorgestellten Verfahren gibt noch zahlreiche andere, die z. T. Mischformen der erwähnten darstellen. Die technischen Gegebenheiten sind so vielfältig, dass die folgende Darstellung nur als Orientierungshilfe aufgefasst werden kann.

12.2 Vorsteuerung

Die Vorsteuerung ist die einfachste Form, den Einfachregelkreis ohne zusätzliche Mess- oder Stellgrößen wesentlich zu verbessern. Strukturell entspricht diese dem Wechsel vom einfachen Standard-Regelkreis auf eine Form wie in Bild 1-6, die hier nochmals dargestellt ist (vgl. Abschnitt 1.2).



Der Regler in Bild 1-6 verarbeitet nicht nur die Regelabweichung $e = w - y$, sondern die beiden Größen w und y separat. Das hat einen entscheidenden Vorteil: Im Einfachregelkreis ergibt sich nämlich das Stör- und Führungsübertragungsverhalten wie in Gl.(6.7) (siehe Kapitel 6) zu

$$\frac{Y(s)}{W(s)} = T(s) = \frac{G_0(s)}{1 + G_0(s)} \quad , \quad \frac{Y(s)}{Z(s)} = S(s) = \frac{1}{1 + G_0(s)} \quad . \quad (12.1)$$

Wegen $S(s) + T(s) = 1$ können für diese Regelkreisstruktur Stör- und Führungsübertragungsverhalten nicht separat voneinander eingestellt werden, sodass man sich zwischen einer Festwertregelung und einer Folgeregelung entscheiden muss. Werden hingegen y und w als separate Eingangssignale verarbeitet, so gilt diese Einschränkung nicht mehr. Bei der Vorsteuerung wird dies so genutzt, dass parallel zur bestehenden Regelung G_R eine Steuerung G_V verwendet wird, wie es in Bild 12-1 gezeigt ist.

Vorsteuerung

Eine Vorsteuerung ist eine Steuerung, die zu einer bestehenden Regelung parallel geschaltet wird, um das Führungsverhalten zu verbessern.

Die Bezeichnung als Vorsteuerung dient dabei als Abgrenzung von einer

reinen Steuerung, die nicht in Kombination mit einer Regelung betrieben wird.

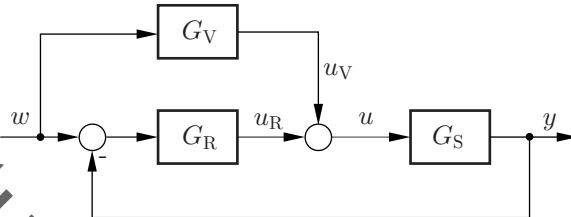


Bild 12-1: Vorsteuerung

Berechnet man für diese Regelkreisstruktur das Stör- und Führungsübertragungsverhalten, so gewinnt man

$$\frac{Y(s)}{W(s)} = \frac{(G_R(s) + G_V(s)) \cdot G_S(s)}{1 + G_0(s)} , \quad \frac{Y(s)}{Z(s)} = \frac{1}{1 + G_0(s)} , \quad (12.2)$$

mit einem veränderten Führungsverhalten und gleichbleibendem Störverhalten.

Der aufgeschnittene Regelkreis $G_0 = G_R \cdot G_S$ wird dabei durch die Vorsteuerung nicht verändert, da diese als Steuerung nicht in einem geschlossenen Wirkungsablauf steht. Folglich beeinflusst die Vorsteuerung (sofern sie selbst stabil ist) die Stabilität des Systems in keiner Weise. Durch die Vorsteuerung G_V können die dynamischen Fehler der Folgeregelung also ohne nachteilige Auswirkungen auf die Stabilität des Systems verminder werden.

Folglich wird man bei Verwendung einer Vorsteuerung zunächst den Regler auf gutes Störverhalten (Festwertregelung) hin auslegen und das Führungsverhalten in einem zweiten Schritt durch Auslegung der Vorsteuerung festlegen (Folgeregelung). Das ist insbesondere für Zustandsregler wichtig, die in der in Abschnitt 11.3 vorgestellten Form nur für eine Festwertregelung geeignet sind. Der Entwurf der Vorsteuerung kann dabei nach den gleichen Grundprinzipien wie der Entwurf einer Steuerung erfolgen.

Für das Folgeverhalten wäre es besonders wünschenswert, wenn die Regelgröße y der Führungsgröße w jederzeit entspräche. Aus der Forderung $y = w$ wird im Laplace-Bereich dabei $Y(s) = W(s)$, weshalb die Führungs-

übertragungsfunktion idealerweise eins sein sollte. Stellt man dies nach der (Vor-)Steuerung $G_V(s)$ um, so erhält man

$$\begin{aligned} \frac{Y(s)}{W(s)} &= \frac{(G_R(s) + G_V(s)) \cdot G_S(s)}{1 + G_0(s)} \stackrel{!}{=} 1 \\ \Leftrightarrow G_0(s) + G_V(s)G_S(s) &\stackrel{!}{=} 1 + G_0(s) \\ \Leftrightarrow G_V(s) &\stackrel{!}{=} \frac{1}{G_S(s)} \end{aligned} \quad (12.3)$$

als Formel für die perfekte Steuerung.

Perfekte Steuerung

Die Steuerung, die ein perfektes Führungsverhalten umsetzt, führt auf eine Invertierung der Regelstrecke.

Diese perfekte Steuerung lässt sich leider aus zwei Gründen im Allgemeinen nicht erreichen. Erstens ist $G_V(s)$ nicht notwendigerweise stabil. Die Polstellen von $G_V(s) = 1/G_S(s)$ sind nämlich die Nullstellen von $G(s)$. Falls $G(s)$ nicht-minimalphasige Nullstellen besitzt, so ist die Steuerung und damit der gesamte Regelkreis instabil. Zweitens ist für $G_V = 1/G_S$ der relative Grad der Steuerung der negative relative Grad der Regelstrecke. Daher ist nur für sprungfähige Systeme mit $r = 0$ die entstehende Steuerung kausal und ansonsten aksual.

Beide Probleme können durch zwei Modifikationen des Steuerungsentwurf behoben werden. Bezüglich der Stabilität wird die Regelstrecke $G_S(s)$ wie in Bild 6-6 in einen Allpass-Anteil $G_{AP}(s)$ und einen minimalphasigen Anteil $G_{MP}(s)$ aufgeteilt. Für den Steuerungsentwurf invertiert man dann nur den minimalphasigen Anteil:

$$G_S(s) = K \cdot G_{AP}(s) \cdot G_{MP}(s) \Rightarrow G_V(s) = \frac{1}{K \cdot G_{MP}(s)} \quad . \quad (12.4)$$

Die Führungsübertragungsfunktion wird so zu

$$\frac{Y(s)}{W(s)} = G_{AP}(s). \quad (12.5)$$

Sie ist also nicht ideal eins, aber im Betrag eins und zeigt nur Abweichungen vom idealen Verhalten im Phasengang. Diese Abweichung kann man oft

tolerieren. Man kann sogar zeigen, dass dieses Vorgehen optimal im Sinne des quadratischen Regelfehlers ist [29] (vgl. Kapitel 18).

Für die Problematik der Kausalität gibt es zwei Lösungsmöglichkeiten. Der Standardweg ist, die Steuerung – genau wie bei einem PD-Regler und einem realen PD-Regler – durch Hinzufügen eines Tiefpassfilters zu kausalieren. Oft setzt man hierfür ein reines Verzögerungselement r -ter Ordnung an und erhält

$$G_V(s) = \frac{1}{G_S(s)} \cdot \frac{1}{(sT + 1)^r} \quad . \quad (12.6)$$

Die Führungsübertragungsfunktion wird so zu

$$\frac{Y(s)}{W(s)} = G_{AP}(s) \cdot \frac{1}{(Ts + 1)^r} \quad . \quad (12.7)$$

Wählt man T so klein, dass $1/T$ den Frequenzbereich der zu erwartenden Führungsgrößenänderungen abdeckt, so sind keine nennenswerten Unterschiede im Führungsverhalten zu erwarten. Allerdings werden geringe T zu größeren Stellgrößenausschlägen führen, die insbesondere im Falle von Stellgrößenbeschränkungen problematisch sein können. Bei einer Vorsteuerung kann dies auch die Regelung behindern, wenn der Aktor durch die Vorsteuerung in die Stellgrößenbeschränkung gerät.

Modellbasierter Steuerungsentwurf

Der modellbasierte Steuerungsentwurf basiert auf einer Invertierung der Regelstrecke, wobei nur der minimalphasige Teil der Regelstrecke invertiert wird und die Steuerung anschließend durch ein Tiefpassfilter kausalisiert wird.

Wendet man diese Methodik des Steuerungsentwurfs beispielhaft auf die dimensionslose Strecke

$$G_S(s) = \frac{-s + 1}{(s + 2)(s + 6)} \quad . \quad (12.8)$$

an, so zerlegt man diese in Allpass- und Minimalphasenanteil

$$G_S(s) = \underbrace{\frac{-s + 1}{s + 1}}_{G_{AP}} \cdot \underbrace{\frac{s + 1}{(s + 2)(s + 6)}}_{G_{MP}} \quad . \quad (12.9)$$

Für die Kausalisierung wählt man T wesentlich schneller als die in der Regelstrecke vorhandenen Zeitkonstanten z. B. $T = 0,1$ und so

$$G_V(s) = \frac{(s+2)(s+6)}{(s+1)} \cdot \frac{1}{(0,1s+1)} . \quad (12.10)$$

Eine zweite Möglichkeit mit der Akausalität umzugehen, ergibt sich bei Folgeregelungen, die im Voraus bekannte Führungsgrößenverläufe verarbeiten. Dies ist beispielsweise bei Vorschubeinrichtungen an numerisch gesteuerten Werkzeugmaschinen der Fall. Da Akausalität bedeutet, dass Ableitungen benötigt werden, können diese in solchen Fällen analytisch im Voraus bestimmt werden. Das führt dazu, dass der Vorsteuerung außer dem Verlauf der Führungsgröße selbst noch die Verläufe ihrer Ableitungen nach der Zeit vorgegeben werden. Die Vorsteuerung muss in diesem Fall nur die vorgegebenen Ableitungen gewichten, entsprechend der Regelstrecke.

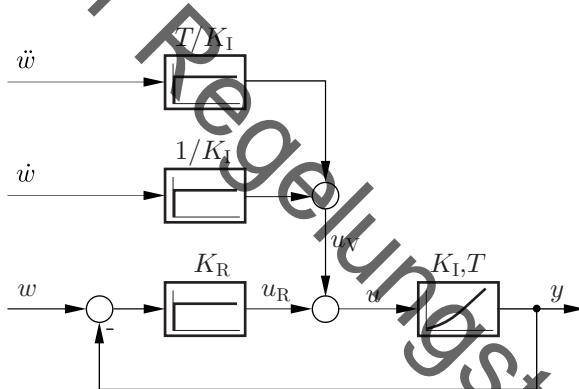


Bild 12-2: Aufschaltung der Ableitungen der Führungsgröße

Als Beispiel hierfür möge ein numerisch gesteuerter Vorschubantrieb (Bild 12-2) dienen, dessen dynamische Eigenschaften einem IT_1 -Glied entsprechen. Für die Vorsteuerung gilt daher

$$G_V(s) = \frac{1}{G_S(s)} = \frac{1}{\frac{K_I}{s(1+sT)}} = \frac{1}{K_I} s(1 + sT) \quad (12.11)$$

bzw. im Zeitbereich

$$u_V = \frac{1}{K_I} (\dot{w} + T\ddot{w}) . \quad (12.12)$$

In manchen Fällen ist der dargestellte modellbasierte Vorsteuerungsentwurf nicht möglich, da keine detaillierten Informationen über das Modell der Regelstrecke vorliegen. In derartigen Situation kann die sehr einfache statische Vorsteuerung

$$G_V(s) = \frac{1}{G_S(0)} \quad (12.13)$$

das Verhalten im Regelkreis dennoch deutlich verbessern.

Mit dieser Vorsteuerung ist die Führungsübertragungsfunktion des Regelkreises automatisch auch stationär genau. Dies gilt aber nicht bei konstanten Störungen, weil diese in der Steuerung nicht berücksichtigt werden. Oft wird die Verstärkung von $G_V(s)$ im Vergleich zu Gl.(12.13) noch etwas abgesenkt – insbesondere dann, wenn der Regler integrierendes Verhalten besitzt – da es sonst zu Überschwingen im Führungsverhalten kommt.

12.3 Führungsgrößenfilter

Als Alternative zu der in Bild 12-1 dargestellten Struktur der Vorsteuerung findet man in der Literatur das gleichwertige Konzept der Führungsgrößenfilter oder Vorfilter nach Bild 12-3.

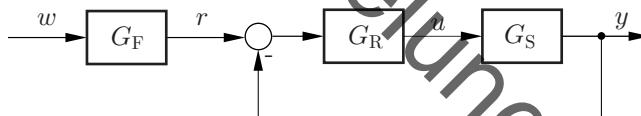


Bild 12-3: Führungsgrößenfilter

Man kann leicht nachrechnen, dass für nicht verschwindende Regler G_R Vorsteuerung und Führungsgrößenfilter äquivalente Formen sind, die sich über

$$\frac{G_V}{G_R} + 1 = G_F \quad (12.14)$$

ineinander umrechnen lassen, wobei G_F ggf. für eine praktischen Einsatz noch kausalisiert werden muss.

Somit gelten alle Ausführungen für Vorsteuerungen auch für Führungsgrößenfilter und umgekehrt. Nachteilig am Führungsgrößenfilter ist, dass dieses – im Gegensatz zur Vorsteuerung – nicht unabhängig vom verwendeten

Regler ist, sodass der Entwurf von Führungsgrößenfilter und Regler nicht losgelöst voneinander erfolgen kann.

Dennoch behauptet sich das Führungsgrößenfilter als Konzept neben der Vorsteuerung. Die Gründe hierfür sind oft praktischer Natur: Vor allem im Prototypenbau werden meist verschiedene Geräte mit jeweils verschiedenen Regelungsaufgaben zusammengeschaltet. Bei einer Vorsteuerung sind dann zwei Geräte notwendig, die beide auf die Stellgröße u zugreifen können müssen. Hierfür sind viele Geräte von der Stange allerdings nicht geeignet. Es ist gerätetechnisch dann wesentlich effizienter, einen gewöhnlichen Regler G_R zu nutzen, der die volle Kontrolle über die Stellgröße u besitzt, und die Informationen aus der Vorsteuerung in der Referenz r zu kodieren. Diese interpretiert der Regler immer noch als Führungsgröße, obwohl das Signal mehr Informationen beinhaltet. Ein solches Vorgehen wird beispielsweise bei Servokontrollern in der Robotik eingesetzt.

Dieses Konzept lässt sich wie folgt abstrahieren.

Führungsgrößenfilter

Ein Führungsgrößenfilter versucht das Verhalten eines Regelkreises dadurch zu verbessern, indem die eigentlich angeforderte Führungsgröße w zu einer Referenzgröße r verändert wird, welche zusätzliche Informationen über ein geeignetes Erreichen der Führungsgröße w im Bezug auf die Systemdynamik des Regelkreises enthält.

Diese Idee soll an einem weiteren Beispiel erläutert werden. Eine stabile PT₁-Regelstrecke wird mit einem I-Regler geregelt. Als Sollwertänderung wird gefordert, dass sich die Regelgröße sprungförmig in $t = 0$ von null auf eins verändern soll. Dieses Szenario ist in Bild 12-4 im durchgezogenen Verlauf y_{oF} gezeigt.

Es ist klar zu erkennen, dass der Regelkreis recht träge reagiert und es zu starkem Überschwingen kommt (vgl. auch Bild 10-2). Für dieses Führungsverhalten gibt es im Rahmen der Idee des Führungsgrößenfilters zwei Arten der Abhilfe.

Das träge Verhalten röhrt daher, dass der I-Regler anfangs eher schwach reagiert und erst aktiv wird, nachdem die Regelabweichung integriert ist. Der Regelkreis kann somit schneller gemacht werden, indem dem Regler für $t = 0$ eine viel größere Referenz $r(0) > w(0)$ übergeben wird, die die Rege-

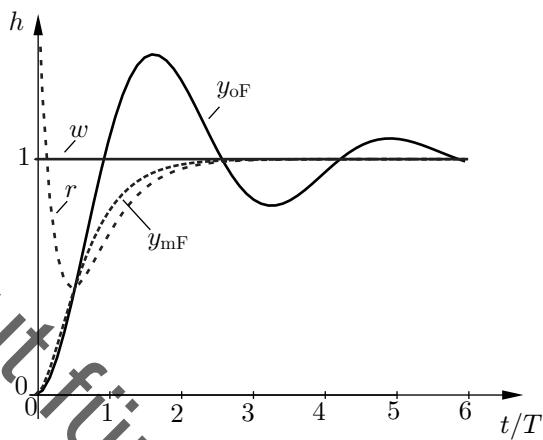


Bild 12-4: PT₁-Regelstrecke mit I-Regler mit (gestrichelt) und ohne (durchgezogen) Führungsgrößenfilter

Abweichung zu Beginn stark erhöht, sodass der I-Regler schneller größere Stelleingriffe erreicht.

Das Überschwingen kann anschaulich so gedeutet werden: Der geschlossene Regelkreis besitzt einen relativen Grad von zwei (eine Polstelle des Reglers, eine Polstelle der Regelstrecke, keine Nullstellen). Folglich ist der geschlossene Regelkreis nicht sprungfähig, sondern friestet die zweite Ableitung kann sich bei einem Sprungeingang schlagartig verändern. Insofern ist die geforderte Führungsgröße nicht umsetzbar. Der Regler reagiert zwar, dennoch verbleibt wegen der Unmöglichkeit, der Führungsgröße zu folgen, gerade zu Beginn eine erhebliche Abweichung. Diese wird im I-Anteil des Reglers aufintegriert, was dann zu einer langfristig zu großen Stellgröße und Überschwingen führt.

Folglich ist es sinnvoll, die Referenz im transienten Bereich der Sprungantwort abzusenken $r(t) < w(t)$, um zu verhindern, dass der Integrator zu große Stellgrößen aufprägt.

Das beschriebene Verhalten ist in Bild 12-4 in den gestrichelten Verläufen r und y_{mF} gezeigt. Es ist deutlich zu erkennen, wie das Führungsgrößenfilter G_F eine Referenz $r(t)$ erzeugt, die genau das geforderte Verhalten aufweist.

Da der geschlossene Regelkreis stationär genau arbeitet, gilt hier zudem

$$\lim_{t \rightarrow \infty} r(t) = \lim_{t \rightarrow \infty} w(t) \quad , \quad (12.15)$$

d. h. die Führungsgröße muss im stationären Fall nicht korrigiert werden. In anderen Fällen kann das Führungsgrößenfilter so auch die bleibende Regelabweichung verkleinern.

Das Grundkonzept des Führungsgrößenfilters, in einer Referenz Information darüber zu kodieren, wie einer Führungsgröße effektiv gefolgt werden kann, wird insbesondere bei sehr komplexen Folgeregelungen zu einer enormen Leistungssteigerungen der Regelung beitragen, wo dieses Modul meist *Trajektoriengenerator* genannt wird.

So betrachte man ein autonomes Fahrzeug, dass als Führungsgröße von einem Punkt A zu einem Punkt B fahren soll. Diese Aufgabe ist für einen Regler sehr schwer zu lösen, wenn sich beispielsweise Hindernisse zwischen A und B befinden. Schaltet man hier jedoch ein zusätzliches Modul zwischen, das – mit einem stark vereinfachten Bewegungsmodell des Fahrzeugs – eine gültige, fahrbare Trajektorie zwischen A und B ermittelt, so muss der Regler nur noch dieser Trajektorie folgen. Diese Regelkreisstruktur mit Folgeregelung und Trajektoriengenerator entspricht strukturell genau einem Führungsgrößenfilter.

12.4 Störgrößenaufschaltung

Die Vorsteuerung verbessert im Wesentlichen das Führungsverhalten des Regelkreises. Besonders vorteilhaft ist dabei der relativ kompakte Entwurf der Steuerung, welche zudem nicht stabilitätsgefährdend ist.

Ein wesentliches Ziel jeder Regelung ist jedoch auch die Minimierung des Einflusses von Störgrößen. Hierzu wird im Standard-Regelkreis die Auswirkung der Störgröße durch das Messen der Regelgröße bestimmt und daraus entsprechende Maßnahmen abgeleitet, welche der Störung entgegenwirken sollen. Somit erfährt der Regler von den Störungen erst, wenn diese sich bereits in der Ausgangsgröße bemerkbar machen, also stark verzögert.

Eine entscheidende Verbesserung des Störverhaltens lässt sich folglich dann erzielen, wenn auf anderem Wege Informationen über die Störgröße vorliegen – beispielsweise, indem die Störgröße selbst oder eine die Störgröße maßgeblich hervorrufende Größe selbst messbar ist. Dieser Sachverhalt ist

in Bild 12-5 gezeigt. Die Ausgangsstörung z wird über das Störmodell G_D durch die Störgröße d hervorgerufen, welche messbar ist. Die Idee der Störgrößenaufschaltung ist es nun, diese Messung d im Rahmen einer Steuerung zur Kompensation der Störungen zu nutzen.

Störgrößenaufschaltung

Eine Störgrößenaufschaltung ist eine Steuerung, die ihre Stellgröße auf Basis der Messung der Störgröße berechnet und diese mit der von einem Regler erzeugten Stellgröße überlagert, um das Störübertragungsverhalten zu verbessern.

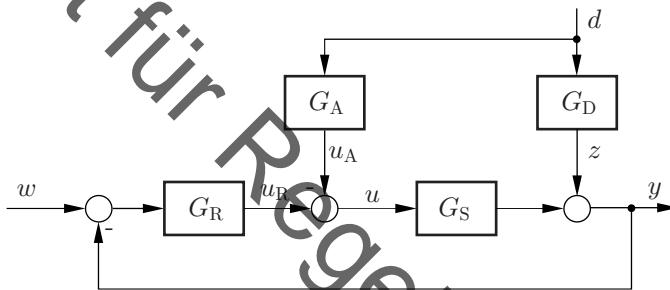


Bild 12-5: Störgrößenaufschaltung

Durch die Auslegung der Störgrößenaufschaltung als Steuerung können die gleichen Vorteile wie bei der Vorsteuerung geltend gemacht werden: Sie gefährdet weder die Stabilität noch und verändert sie das Führungsübertragungsverhalten.

Analog zum Entwurf der Vorsteuerung soll idealerweise die Wirkung der Störung genau kompensiert werden, sodass die Regelgröße durch diese Störung nicht beeinflusst wird. Berechnet man das hierfür notwendige Übertragungsverhalten der Störgrößenaufschaltung G_A , so gewinnt man

$$\frac{Y(s)}{Z(s)} \stackrel{!}{=} 0 \quad \Rightarrow \quad G_A(s) \stackrel{!}{=} \frac{G_D(s)}{G_S(s)} . \quad (12.16)$$

Diese Gleichung ähnelt sehr der Gleichung für den Entwurf von Vorsteuerungen $G_V = 1/G_S$. Daher lassen sich die entsprechenden Entwurfsprinzipien für Steuerungen über eine Zerlegung in Allpass- und Minimalpha-

senanteil sowie die Nutzung eines Tiefpassfilters hier ebenfalls anwenden. Nur durch das Auftreten der Übertragungsfunktion $G_D(s)$ ergeben sich ggf. Abweichungen. So kann beispielsweise für eine Störung am Eingang der Regelstrecke $G_D(s) = G_S(s)$ die ideale Störgrößenaufschaltung direkt zu $G_A(s) = 1$ angegeben werden – auch wenn die Regelstrecke nicht-minimalphasig ist oder sogar eine Totzeit besitzt. In den meisten Fällen erschwert $G_D(s)$ allerdings den Entwurf eher, da Modelle für das Störübertragungsverhalten oft nicht oder nur sehr ungenau bekannt sind. Hier behilft man sich – erneut analog zum Entwurf der Vorsteuerung – mit einer statischen Betrachtungsweise und nutzt $G_D(0)$ für eine zumindest statische Kompensation.

Störgrößenaufschaltungen sind dann zweckmäßig, wenn der zu regelnde Prozess durch wenige gut messbare Störgrößen beeinflusst wird. Ein weit verbreitetes Anwendungsbeispiel für Störgrößenaufschaltungen ist die Veränderung der Vorlauftemperatur in Zentralheizungsanlagen als Funktion von Änderungen der Außentemperatur. Bei Raumtemperaturregelungen ist die Außentemperatur eine der wichtigsten Störgrößen (siehe auch Abschnitt 1.1). Wenn durch geeignete Steuerung der Temperatur des Heizungsvorlaufs die Wirkung von Außentemperaturschwankungen auf die Raumtemperatur ganz oder teilweise ausgeglichen wird, so kann der Raumtemperaturregler i. Allg. einfacher und damit billiger sein und seine Aufgabe dennoch besser erfüllen als ein Regler in einem einfachen Regelkreis.

In manchen Fällen kann es vorkommen, dass eine Störgrößenaufschaltung trotz messbarer Störgrößen nicht in Frage kommt. Der häufigste Grund hierfür ist ein zu ungenaues Störmodell G_D . Wenn der Einfluss dieser Störgröße erheblich ist, so kann es sinnvoll sein, nach Modifikationen des technischen Prozesses (wie zusätzlichen Aktoren) zu suchen, die diese Störgrößen direkt beeinflussen können. Hierdurch wird diese Störgröße dann zur Regelgröße eines vorgeschalteten Reglers (Vorregelung) und man kann unabhängig von G_D ein gutes Störverhalten im Hauptregelkreis erzielen.

12.5 Kaskadenregelung

Die zentrale Voraussetzung für die Anwendung der Störgrößenaufschaltung ist, dass die Störgröße messbar sein muss. Dies ist jedoch häufig nicht der Fall. Die Grundidee, eintretende Störungen frühzeitig zu detektieren, kann jedoch auch bei nicht messbaren Störgrößen verfolgt werden, wenn statt-

dessen Zwischengrößen des Prozesses messbar sind.

Hierzu wird angenommen, dass der zu regelnde Prozess als eine Reihenschaltung von zwei Teilprozessen dargestellt werden kann (siehe Bild 12-6). Die Störungen z_1 und z_2 wirken auf beide Teilprozesse. Die Störgröße z_2

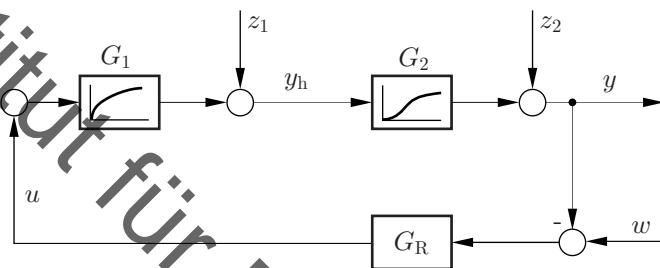


Bild 12-6: Reihenschaltung in zwei Teilprozesse

kann in der Regelgröße y direkt erfasst werden. Für die Störgröße z_1 ist dies leider nicht der Fall, wenn G_2 beispielsweise stark verzögernd ist.

Solange der erste Teilprozess G_1 nicht wesentlich langsamer als der zweite Teilprozess G_2 ist, wird die Störung z_1 in der Zwischengröße y_h wesentlich früher als in der eigentlichen Regelgröße y erkennbar sein. Daher ist zu erwarten, dass eine Nutzung der Informationen y_h das Störverhalten des Regelkreises verbessert.

Dies erreicht man dadurch, dass man einen zusätzlichen „inneren“ Regelkreis bildet, der die als messbar angenommene Zwischengröße y_h als Regelgröße einstellt. Da man aber nicht voraussetzen kann, dass für den zweiten Regelkreis ein zweiter Aktor zur Verfügung steht, muss man die Stellsignale des inneren und äußeren Regelkreises geschickt überlagern, um mit der einzigen Stellgröße u auszukommen.

Hierfür ist auch eine additive Überlagerung wie bei der Störgrößenaufschaltung denkbar. Es hat sich jedoch als besonders zweckmäßig herausgestellt, die beiden Regler so miteinander zu verschalten, dass der äußere Hauptregelkreis die Führungsgröße des unterlagerten inneren Regelkreises einstellt (siehe Bild 12-7).

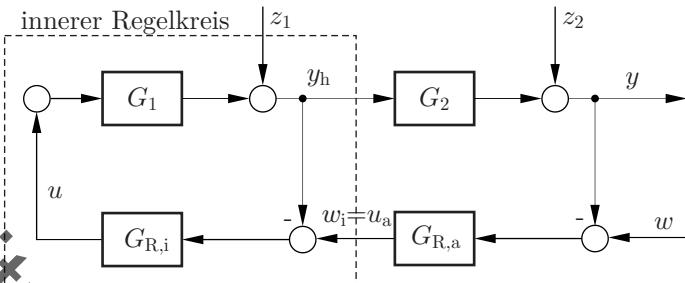


Bild 12-7: Kaskadenregelung

Kaskadenregelung

Eine Kaskadenregelung ist eine Regelung, bei der zwei oder mehr Regler mit unterschiedlichen Regelgrößen dergestalt miteinander verschaltet werden, dass die Stellgröße des sogenannten äußeren Regelkreises der Führungsgröße des sogenannten inneren Regelkreises entspricht.

Die Kaskadenregelung ist eine der am häufigsten genutzten Regelungsstrukturen und wurde bereits in den Beispielen in Abschnitt 1.3 vorgestellt. Eine kaskadierte Regelung ist besonders dann vorteilhaft, wenn eine an sich komplizierte Regelstrecke aus einer Reihenschaltung verschiedener einfacher Elemente mit eigenen Störungen oder Unsicherheiten aufgebaut ist, deren Ausgänge jeweils gemessen werden können. Die Kaskadenregelung unterteilt eine auf den ersten Blick komplexe Regelungsaufgabe durch die Einführung zusätzlicher Messgrößen in eine Folge einfacher Regelungsaufgaben.

Störungen können in den jeweils unterlagerten Regelkreisen leichter kompensiert werden, da der Signalweg der Störung durch die Regelstrecke stark verkürzt wird. Auch können Nichtlinearitäten und andere Unsicherheiten in einem Teil der Regelstrecke durch die unterlagerten Regler ausgeglichen werden.

Dennoch benötigt die Kaskadenregelung insgesamt nur ein einziges Stellglied, da alle weiteren inneren Stellgrößen die Führungsgrößen der unterlagerten Regelkreise sind. Konzeptionell kann die Kaskadenregelung dabei so gedeutet werden, dass ein Regelstrecke mit Eingang u und störbehaftetem

Ausgang y_h durch die unterlagerte Regelung zu einem Aktor wird, der ein angefordertes w_i – das ist der Sollwert für y_h – in y_h umsetzt. Hierdurch kann die vormalige Regelgröße y_h in Form der Sollvorgabe w_i als Stellgröße u_a für überlagerte Regelungsaufgaben genutzt werden.

Durch die schnellere Dynamik der jeweiligen Teilprozesse im Vergleich zum Gesamtprozess ist es zusätzlich meist möglich, in den unterlagerten Regelkreisen größere Reglerverstärkungen zuzulassen, als dies aufgrund der Trägheit des Hauptregelkreis möglich wäre. Dennoch ist zu beachten, dass die unterlagerten Regler das Gesamtsystem auch destabilisieren können.

Bei der Dimensionierung derartiger Regelkreise geht man daher zweckmäßigerweise von innen nach außen vor, d. h. man legt zuerst den Regler $G_{R,i}$ aus und fasst dann den mit diesem Regler gebildeten Unterregelkreis als Regelstrecke auf, an die der nächste Regler $G_{R,a}$, anzupassen ist.

Die Kaskadenregelung soll an einem Beispiel illustriert werden. Hierzu wird die Regelung eines Füllstands y eines Behälters mit dem Wirkungsplan in 12-8 betrachtet. Als Stellgröße dient dabei die Spannung u einer Pumpe, durch die abhängig von ihrer Drehzahl n und des Drucks p in der Zuleitung einen Volumenstrom q_{zu} gefördert wird. Aus dem Behälter fließt das Medium abhängig vom Füllstand mit dem Volumenstrom q_{ab} ab. Auf die Pumpe wirken Störungen in Form von Verlustleistungen z und auch der Druck p ist als Störgröße aufzufassen.

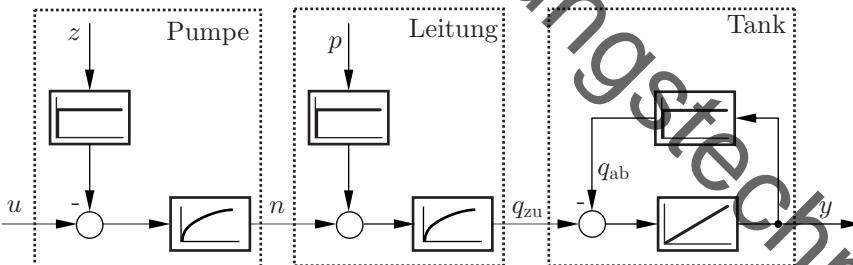


Bild 12-8: Wirkungsplan eines Tanks mit Zuleitung und Pumpe

Die Regelung des Füllstandes gestaltet sich ohne weitere Maßnahmen schwierig. Das liegt daran, dass Verlustleistungen innerhalb der Pumpe erst sehr viel später bei einer Abweichung des Füllstandes bemerkt und korrigiert werden.

Im Sinne einer Kaskadenregelung verfolgt man als erstes das Ziel, einen Drehzahlregler für die Pumpe zu entwerfen und diese Störungen so zu unterdrücken – siehe Bild 12-9. Die Schnittstelle der Spannung u verschwindet dabei nach außen hin und wird durch die Vorgabe einer Solldrehzahl n_{soll} ersetzt. Aus der zu regelnden Pumpe ist eine neue Aktor geworden, über den eine gewünschte Solldrehzahl direkt eingestellt werden kann.

Als zweites wendet man sich den störenden Druckschwankungen in der Zuleitung zu. Hier sieht man einen Volumenstromregler vor, der eine gewünschten Volumenstrom q_{soll} unabhängig von den Druckschwankungen präzise einstellt. Aus der Regelgröße q_{zu} wird somit eine neue Stellgröße in Form des Sollwertes q_{soll} . Diese kann dann für die eigentliche Hauptregelungsaufgabe der Füllstandsregelung genutzt werden.

Der sich so ergänzende kaskadierte Regelkreis ist in Bild 12-9 gezeigt. Hierbei sind für den Drehzahlregler und den Volumenstromregler jeweils ein PI-Regler vorsehen, damit die angeforderten Sollgrößen auch ohne Abweichung eingestellt werden können. Für den Füllstandsregler ist dies nicht notwendig, da die Regelstrecke bereits integrierendes Verhalten besitzt, weshalb ein P-Regler ausreicht.

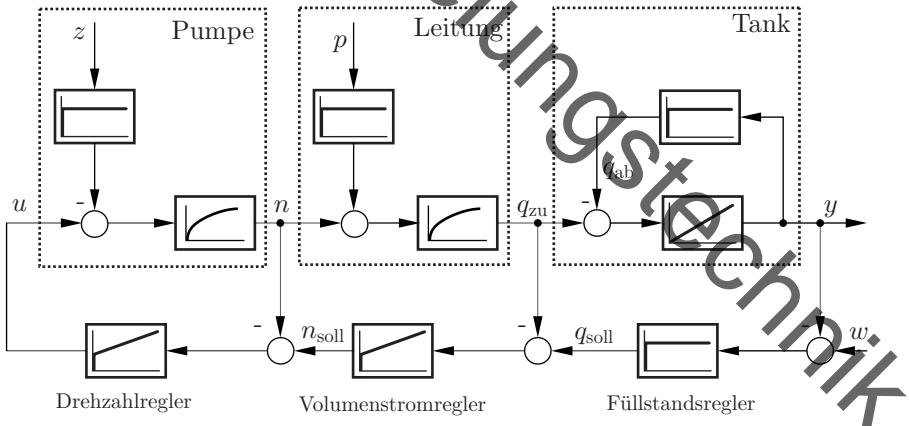


Bild 12-9: Füllstandsregelung als Kaskadenregelung

12.6 Hilfsstellgröße

Bei der Kaskadenregelung liegen zur Lösung der Regelungsaufgabe mehrere Messgrößen, aber nur eine Stellgröße vor. In anderen Fällen kann es vorkommen, dass nur eine einzige Messgröße, dafür aber mehrere Stellgrößen vorliegen. Die zusätzlich verfügbaren Stellgrößen werden *Hilfsstellgrößen* genannt und kann dazu genutzt werden, in einem unterlagerten Regelkreis mit Hilfsregler G_H das dynamische Verhalten zu verbessern – siehe Bild 12-10.

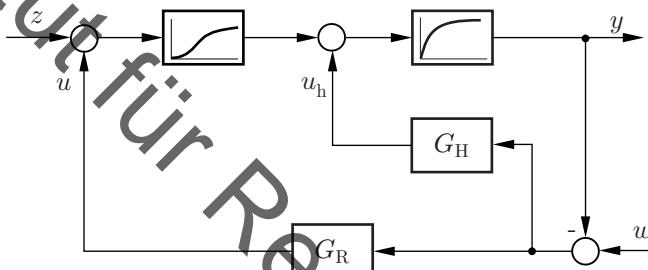


Bild 12-10: Aufschaltung einer Hilfsstellgröße

Der durch die Hilfsstellgröße gebildete Unterregelkreis kann i. Allg. wesentlich günstigere dynamische Eigenschaften als der Hauptregelkreis haben, weil die zugehörige Teilregelstrecke von niedrigerer Ordnung als die zum Hauptregelkreis gehörende Regelstrecke ist. Hierdurch kann der Hilfsregler G_H entsprechend schneller eingestellt werden.

Weil der Unterregelkreis die gleiche Führungsgröße und die gleiche Regelgröße wie der Hauptregelkreis verarbeitet und dies unter günstigeren Bedingungen für gute Dynamik tut, könnte man meinen, der Hauptregelkreis sei überflüssig. In der technischen Wirklichkeit ist dieser Schluss meist falsch, weil in vielen Fällen der Einsatz der Hilfsstellgröße mit hohen Kosten verbunden ist oder anderweitig als alleinige Stellgröße unwirtschaftlich ist. Aufgrund derartiger Einschränkungen werden als Hilfsregler meist solche mit differenzierendem Verhalten eingesetzt, damit die Hauptlast der Regelung nach wie vor beim Hauptregelkreis liegt.

Ein Anwendungsbeispiel für Regelungen mit Hilfsstellgrößen zeigt Bild 12-11. Die Temperatur des von einem Dampferzeuger mit Überhitzer abge-

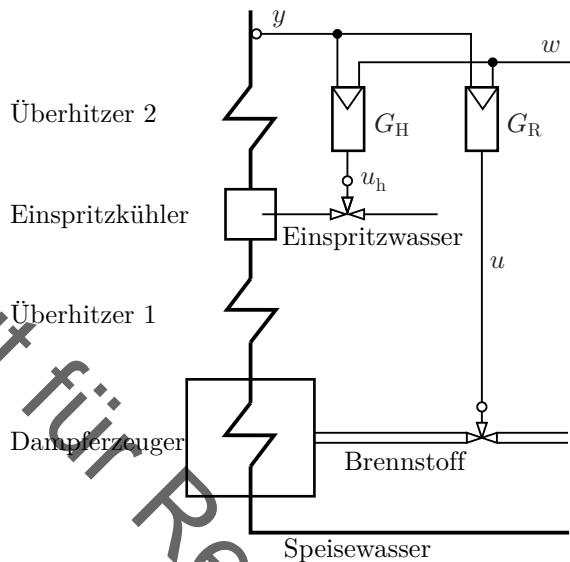


Bild 12-11: Dampftemperaturregelung mit Einspritzwasserstrom als Hilfsstellgröße

gebenen Dampfes wird durch die Brennstoffzufuhr beeinflusst. Um jedoch kurzzeitige Schwankungen der Temperatur besser ausgleichen zu können, werden nach einzelnen Überhitzern oder Überhitzergruppen sog. Einspritzkühler vorgesehen, in denen der Dampf durch eingespritztes Wasser gekühlt wird. Der Einspritzwasserstrom ist hier Hilfsstellgröße; er wirkt erheblich schneller auf die Dampftemperatur als eine Brennstoffstromänderung. Durch die Dampfkühlung wird allerdings der Gesamtwirkungsgrad des Dampferzeugers verschlechtert, sodass man i. Allg. anstrebt, möglichst wenig Einspritzwasser zu verwenden.

13 Mehrgrößenregelung

13.1 Zentrale vs. dezentrale Regelung

Die im vorangegangenen Kapitel vorgestellten Verfahren der Kaskadenregelung und der Hilfsstellgröße unterscheiden sich vom Standard-Regelkreis auch dadurch, dass nicht eine einzige Regelgröße bzw. Stellgröße vorlag, sondern mehrere genutzt wurden. So wurden bei der Kaskadenregelung bei einer einzigen Stellgröße mehrere Größen eingeregelt. Bei der Hilfsstellgröße wurden genau anders herum bei einer einzigen Regelgröße mehrere Stellgrößen genutzt. Somit handelte es sich bei den Regelstrecken (aber auch den vermaschten Regelungen selbst) nicht mehr um SISO, sondern um MIMO-Systeme.

Die Situation, dass ein zu regelnder MIMO-Prozess n Regelgrößen besitzt, die mit m Stellgrößen geregelt werden sollen, wird unter dem Schlagwort der Mehrgrößenregelung diskutiert. Das Verhältnis von n und m wird dabei maßgeblich bestimmen, ob es genügend Stellgrößen gibt, um alle n Regelgrößen unabhängig voneinander einzustellen zu können. Hierzu fasst man die Regelstrecke (siehe Gl.(4.73)) $\mathbf{Y}(s) = \mathbf{G}(s)\mathbf{U}(s)$ als $n \times m$ -Übertragungsmatrix auf. Sollen alle n Regelgrößen zumindest im stationären Fall korrekt eingestellt werden, so muss gelten:

$$\mathbf{y}_\infty = \mathbf{G}(0)\mathbf{u}_\infty \quad . \quad (13.1)$$

Dieses Gleichungssystem lässt sich für invertierbare Matrizen eindeutig lösen. Hierzu muss \mathbf{G} quadratisch sein, das System also genauso viele Stell- wie Regelgrößen besitzen. In diesem Fall gibt es eine eindeutige Kombination \mathbf{u}_∞ an Stellgrößen, die die geforderten Regelgrößen \mathbf{y}_∞ einstellen. Für $n > m$ gibt es mehr Regel- als Stellgrößen und Gl.(13.1) wird meist keine Lösung besitzen. Für $n < m$ gibt es potentiell unendlich viele Lösungen und daher zahlreiche Kombinationen \mathbf{u}_∞ , die auf das gleiche \mathbf{y}_∞ führen.

Unteraktuiert, Vollaktuiert, Überaktuiert

Gegeben ist eine Regelstrecke mit n Regelgrößen und m Stellgrößen.

- Für $n > m$ heißt die Strecke unteraktuiert.

- Für $n = m$ heißt die Strecke vollaktuiert.
- Für $n < m$ heißt die Strecke überaktuiert.

In den allermeisten Fällen wird es sich um vollaktuierte Regelstrecken handeln, sodass man sich oft auf diesen Fall beschränkt. Es gibt zwei grundsätzlich unterschiedliche Aufbauten, um solche Mehrgrößensysteme zu regeln: die *zentrale* und *dezentrale* Regelung.

Zentrale und Dezentrale Regelung

Wird die Regelungsaufgabe, die n Regelgrößen eines MIMO-Systems einzustellen, auf n verschiedene SISO-Regler aufgeteilt, so spricht man von einer *dezentralen* Regelung.

Wird die Regelungsaufgabe hingegen von einer einzigen Regelung übernommen, so spricht man von einer *zentralen* Regelung.

Das Konzept der zentralen Regelung hat den Nachteil, dass dieselbe Recheneinheit alle Ein- und Ausgangssignale verwalten muss. Bei örtlich verteilten Mehrgrößensystemen wie verfahrenstechnischen Anlagen kann die Kommunikation all dieser Signale zu zusätzlichen Totzeiten bei der Bereitstellung und Übermittlung der notwendigen Signale führen. Auch gefährdet die Überlastung oder der Ausfall einer einzigen Recheneinheit die Stabilität sämtlicher Ausgangsgrößen. Daher ist es unter Umständen erstrebenswert, statt einer zentralen Regelung mehrere dezentral arbeitende Regler einzusetzen.

Bei der dezentralen Regelung von MIMO-Systemen werden die jeweiligen Regelgrößen von separaten und voneinander unabhängigen Reglern geregelt und jede Stellgröße über genau eine einzige Regelabweichung ermittelt. So mit können Kommunikationswege kurz gehalten und die Funktionalität auf mehrere Steuergeräte verteilt werden. Als zusätzlichen Vorteil reduziert die dezentrale Regelung das Problem der Mehrgrößenregelung auf das bereits bekannte Probleme einer SISO-Regelung.

Die so entstehenden Einzelregelkreise sind aber häufig durch die Regelstrecke miteinander gekoppelt, d. h. eine Führungsgrößenänderung in einem der Regelkreise wirkt auf einen oder mehrere andere wie eine Störung. Die Reaktion der Regelkreise auf die Störungen beeinflusst benachbarte Regelkreise so, dass sich zwischen den Regelkreisen geschlossene Wirkungsabläufe einstellen können, die das Gesamtsystem instabil werden lassen, obgleich jeder

Teilregelkreis für sich betrachtet stabil ist – ein entscheidender Nachteil der dezentralen Regelung.

Dies soll an einem Beispiel erläutert werden: Es wird die gleichzeitige Regelung von Mischungstemperatur und Durchfluss in einer Flüssigkeitsmischstation (z. B. in einer Dusche) betrachtet, die in Bild 13-1 schematisch wiedergegeben ist.

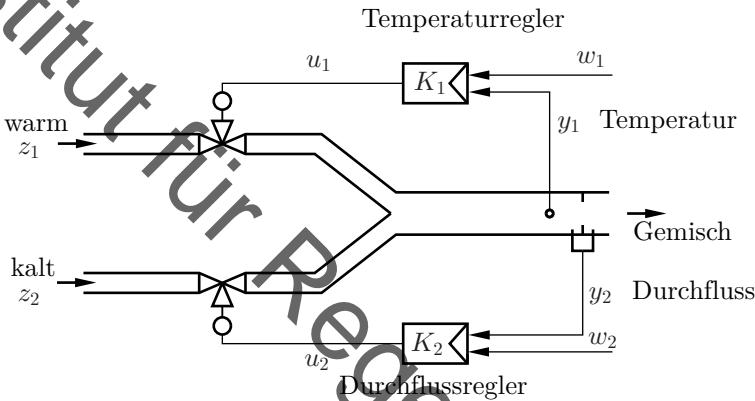


Bild 13-1: Regelung von Temperatur und Durchfluss bei Mischung

Die Temperatur der Mischung soll über den Temperatur-Regler K_1 durch Verstellen des Ventils für den Zulauf warmer Flüssigkeit eingestellt werden. Der Durchfluss der Mischung wird über den Durchfluss-Regler K_2 durch Verstellen des Ventils für den Zulauf kalter Flüssigkeit geregelt. Hierdurch sollen Störungen, die durch Schwankungen der Drücke z_1 und z_2 hervorgerufen werden, ausgeglichen werden.

Mit diesem Aufbau der Regelung werden sich beide Regelkreise gegenseitig beeinflussen. So wird eine Öffnung des Warmwasserventils u_1 durch den Regler K_1 eine (geplante) Erhöhung der Temperatur y_1 zur Folge haben – dies wird auch als *direkte Wirkung* bezeichnet. Zeitgleich wird aber ein Öffnen des Warmwasserventils u_1 auch den Durchfluss y_2 erhöhen. Dieser (ungeplanten) Erhöhung wird Regler K_2 durch ein Schließen des Kaltwasserventils entgegenwirken. Dieser kennt die eigentliche Ursache der Durchflusserhöhung nämlich nicht und wird diese als Störung auffassen. Die Schließung des Kaltwasserventils u_2 durch K_2 führt aber auf einen zusätzlichen

Anstieg der Temperatur y_1 . Dieser zweite Wirkungspfad über den Regler K_2 wird auch als *indirekte Wirkung* bezeichnet.

Direkte und indirekte Wirkung

Bei einem dezentralen Mehrgrößensystem beschreibt die *direkte Wirkung* die Auswirkungen, die eine Stellgrößenänderung u_i direkt auf die zugehörigen Regelgröße y_j hat, wobei alle anderen Regler K_k zu null gesetzt wurden. Es wird folglich jeder Teilregelkreis separat für sich betrachtet. Die *indirekte Wirkung* beschreibt die Auswirkungen, die eine Stellgrößenänderung u_i auf die nicht zugehörigen Regelgrößen $y_{k \neq j}$ und über die vorhandenen anderen Regler K_k und deren Stellgrößen u_k zusätzlich auf y_j hat.

Der Unterschied zwischen direkter und indirekter Wirkung für den Zweigrößenfall ist in Bild 13-2 dargestellt.

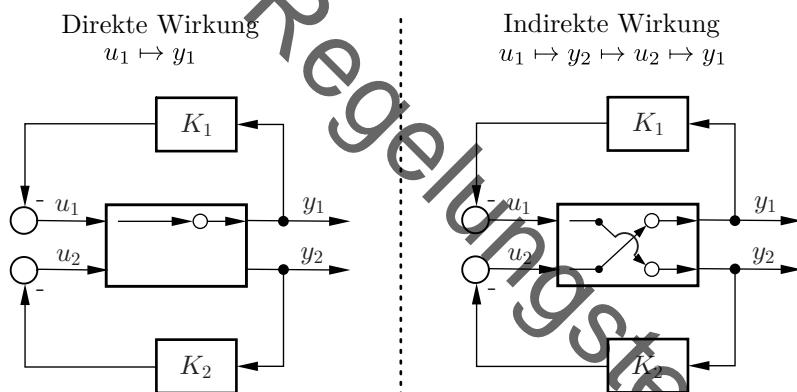


Bild 13-2: Direkte und indirekte Wirkung bei einem Zweigrößenystem

Die indirekte Wirkung sorgt dafür, dass die vom Regler vorgesehene Wirkung von der tatsächlichen Gesamtwirkung abweicht. Im vorliegenden Beispiel führt dadurch das Öffnen des Warmwasserventils zu einer weitaus größeren Temperaturerhöhung als beabsichtigt. Hierdurch wird der Regelkreis schlechter gedämpft erscheinen und kann sogar instabil werden.

Nachteilig am dezentralen Reglerentwurf ist, dass die indirekten Wirkungen nicht explizit adressiert werden können. Abhängig von deren Stärke muss

entschieden werden, ob ein dezentraler Reglerentwurf möglich ist.

Ist eine dezentrale Regelung möglich, so lassen sich mit entsprechenden Modifikationen die für SISO-Systeme bekannten Entwurfsmethoden anwenden. Dabei muss für jedes zusammengehörige Paar von Eingangs- und Ausgangssignal ein Regler entworfen werden. Um besonders geeignete Paare zu identifizieren und bewerten zu können, in welchen Fällen eine dezentrale Regelung eine zentrale Regelung ersetzen kann, ist es notwendig, MIMO-Systeme auf einige wichtige Eigenschaften hin zu untersuchen.

13.2 Eigenschaften von Mehrgrößensystemen

13.2.1 Verschaltungen von Mehrgrößensystemen

Lineare Mehrgrößensysteme lassen sich sowohl im Zustandsraum als auch im Bildbereich beschreiben, wobei die zugehörigen Beschreibungsformen bereits vorgestellt wurden. Schreibt man ein System mit n Regelgrößen und m Stellgrößen wie in Gl.(4.73) als Übertragungsmatrix

$$\begin{bmatrix} Y_1(s) \\ \vdots \\ Y_n(s) \end{bmatrix} = \underbrace{\begin{bmatrix} G_{11}(s) & \dots & G_{1m}(s) \\ \vdots & \ddots & \vdots \\ G_{n1}(s) & \dots & G_{nm}(s) \end{bmatrix}}_{\mathbf{G}(s)} \begin{bmatrix} U_1(s) \\ \vdots \\ U_m(s) \end{bmatrix}, \quad (13.2)$$

so beschreiben die Einzelübertragungsfunktionen $G_{ij}(s) = \frac{Y_i(s)}{U_j(s)}$ die direkten Wirkungen von u_j auf y_i .

Die Übertragungsmatrizen lassen sich ebenso wie Übertragungsfunktionen verknüpfen und daraus Parallel- und Reihenschaltungen sowie Rückführungen bilden. In Tab. 13-1 sind die drei wichtigsten aus zwei Übertragungsgliedern bestehenden Schaltungen abgeleitet und zusammengestellt.

Die Herleitung der Rechenvorschriften für die Verschaltungen erfolgt analog zum SISO-Fall. Hierbei ist nur gesondert darauf zu achten, dass es sich bei MIMO-Systemen nicht mehr um Übertragungsfunktionen, sondern um Übertragungsmatrizen handelt. Daher ist bei der Verschaltung die Reihenfolge der Übertragungsglieder wichtig. Die Reihenfolge der Multiplikation der Teilsysteme kann nämlich wegen der fehlenden Kommutativität der Matrizenmultiplikation nicht mehr unterschlagen werden. Zusätzlich müs-

Übertragungsmatrix	Wirkungsplan
$\begin{aligned} \mathbf{y} &= \mathbf{y}_1 + \mathbf{y}_2 \\ &= (\mathbf{G}_1 + \mathbf{G}_2) \cdot \mathbf{u} \end{aligned}$	<p>Parallelschaltung</p>
$\begin{aligned} \mathbf{y} &= \mathbf{G}_1 \cdot \mathbf{v} \\ &= \mathbf{G}_1 \cdot (\mathbf{G}_2 \cdot \mathbf{u}) \\ &= (\mathbf{G}_1 \cdot \mathbf{G}_2) \cdot \mathbf{u} \\ &\neq (\mathbf{G}_2 \cdot \mathbf{G}_1) \cdot \mathbf{u} \end{aligned}$	<p>Reihenschaltung</p>
$\begin{aligned} \mathbf{y} &= \mathbf{G}_v \cdot (\mathbf{u} \mp \mathbf{v}) \\ &= \mathbf{G}_v \cdot (\mathbf{u} \mp \mathbf{G}_r \cdot \mathbf{y}) \\ \Rightarrow \mathbf{y} &= (\mathbf{I} \pm \mathbf{G}_v \cdot \mathbf{G}_r)^{-1} \mathbf{G}_v \mathbf{u} \end{aligned}$	<p>Rückkopplung</p>

Tabelle 13-1: Schaltungen von Mehrgrößensystemen

sen die Dimensionen der Signale und Systeme zueinander passen.

So ist bei einer Reihenschaltung $\mathbf{G} = \mathbf{G}_1 \mathbf{G}_2$ von zwei Teilsystemen zu beachten, dass die Teilsysteme \mathbf{G}_1 und \mathbf{G}_2 entgegen der Signalflussrichtung angeordnet sind: Das Eingangssignal \mathbf{u} wird zuerst von Teilsystem \mathbf{G}_2 und danach von \mathbf{G}_1 verarbeitet. Da Vektoren (wie \mathbf{u}) immer von rechts an Matrizen multipliziert werden, ergibt sich die hergeleitete Reihenfolge.

Diese Reihenfolge muss auch bei Rückführungen beachtet werden. Dabei gilt unter Beachtung der Multiplikationsreihenfolge

$$\mathbf{G}(s) = (\mathbf{I} + \mathbf{G}_v(s)\mathbf{G}_r(s))^{-1} \mathbf{G}_v(s) = \mathbf{G}_v(s) (\mathbf{I} + \mathbf{G}_r(s)\mathbf{G}_v(s))^{-1}, \quad (13.3)$$

wobei die zweite Variante mit rechtsseitiger Inversen über Ausmultiplizieren nachgewiesen werden kann.

Aufgrund der fehlenden Kommutativität und den zwei Varianten in Gl.(13.3) ist es beim Aufstellen des aufgeschnittenen Regelkreises von Bedeutung, an welcher Stelle der Regelkreis geöffnet wird. Geschieht dies wie in 6-1 nach dem Ausgangsgrößenvektor \mathbf{y} , so erhält man $\mathbf{G}_0(s) = \mathbf{G}_v(s)\mathbf{G}_r(s)$, da \mathbf{y} zunächst durch den Rückwärtszweig verarbeitet wird. Schneidet man hingegen bei v , so ergibt sich $\mathbf{G}_0(s) = \mathbf{G}_r(s)\mathbf{G}_v(s)$ und damit die umgekehrte Reihenfolge.

Es sollte aber (analog dazu, dass es egal ist, ob die Führungs- oder Störübertragungsfunktion betrachtet wird) für die Stabilität keinen Unterschied machen, wo der Regelkreis aufgeschnitten wird. Tatsächlich kann man zeigen, dass beide Varianten für sämtliche Stabilitätsuntersuchungen auf dieselben Ergebnisse führen, solange Pol-Nullstellen-Kürzungen ausgeschlossen sind. Aus Gründen der Vergleichbarkeit und der Konvention einigt man sich auf die Reihenfolge, die sich beim Schnitt im Rückführzweig ergibt.

Aufgeschnittener Regelkreis und Rückführdifferenzmatrix

Der aufgeschnittene Regelkreis im Mehrgrößenfall ist

$$\mathbf{G}_0(s) = \mathbf{G}_v(s)\mathbf{G}_r(s) \quad . \quad (13.4)$$

Zudem wird die Matrix $\mathbf{F}(s) = \mathbf{I} + \mathbf{G}_0(s)$ *Rückführdifferenzmatrix* genannt.

13.2.2 Querkopplungen

Mit den Verschaltungsregeln lässt sich beschreiben, welche der Elemente $G_{ij}(s)$ der Übertragungsmatrix \mathbf{G} die Größe der indirekten Wirkung und der direkten Wirkung festlegen. Hierzu wird zunächst festgestellt, dass auch

der Regler $\mathbf{K}(s)$ sich als $m \times n$ - Übertragungsmatrix

$$\mathbf{K}(s) = \begin{bmatrix} K_{11}(s) & \dots & K_{1n}(s) \\ \vdots & \ddots & \vdots \\ K_{m1}(s) & \dots & K_{mn}(s) \end{bmatrix} \quad (13.5)$$

darstellen lässt. Die Dimensionen n und m sind im Vergleich zur Regelstrecke vertauscht, da die Regelgrößen die Eingänge von $\mathbf{K}(s)$ sind.

Im allgemeinen Fall einer zentralen Regelung ist $\mathbf{K}(s)$ potentiell voll besetzt, da es einen einzigen Regler gibt, der alle verfügbaren Regelgrößen aufnimmt, verarbeitet und alle Stellgrößen berechnet.

Im Falle einer dezentralen Regelung gibt es genauso viele Stell- wie Regelgrößen und $\mathbf{K}(s)$ ist daher quadratisch. Zudem wirkt jede Regelgröße nur auf eine einzige Stellgröße. Folglich besitzt im dezentralen Fall $\mathbf{K}(s)$ in jeder Spalte und jeder Zeile nur ein Element, dass von null verschieden ist.

Am Beispiel für $n = 3$ mit

$$\begin{aligned} \mathbf{K}(s) &= \begin{bmatrix} 0 & K_2(s) & 0 \\ 0 & 0 & K_3(s) \\ K_1(s) & 0 & 0 \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}}_{\mathbf{P}} \cdot \begin{bmatrix} K_1(s) & 0 & 0 \\ 0 & K_2(s) & 0 \\ 0 & 0 & K_3(s) \end{bmatrix} \end{aligned} \quad (13.6)$$

sieht man, dass die allgemeine dezentrale Reglermatrix $\mathbf{K}(s)$ aufgrund dieser Struktur in das Produkt einer Permutationsmatrix \mathbf{P} und einer Diagonalmatrix zerlegt werden kann. Dabei enthält die Permutationsmatrix die Informationen, welche Regelgröße welcher Stellgröße zugeordnet ist, während die Diagonalmatrix die dynamischen Reglerübertragungsfunktionen enthält.

Sortiert man den aufgeschnittenen Regelkreis als

$$\mathbf{G}_0(s) = \mathbf{G}(s) \cdot \mathbf{K}(s) = \underbrace{\mathbf{G}(s)\mathbf{P}}_{\mathbf{G}_P(s)} \cdot \begin{bmatrix} K_1(s) & & \\ & \ddots & \\ & & K_n(s) \end{bmatrix}, \quad (13.7)$$

so entspricht $\mathbf{G}_P(s)$ einer Regelstrecke, deren Stell- und Regelgrößen der-gestalt umsortiert wurden, dass die dezentrale Regelung mit Stellgröße u_i auch die Regelgröße y_i mit gleichem Index regelt.

Hauptregelkreise und Querkopplung

Es sei $\mathbf{G}(s)$ eine Regelstrecke mit umsortierten Ein- und Ausgängen, so dass die dezentrale Regelung $\mathbf{K}(s)$ eine Diagonalmatrix ist. Dann heißen die Diagonalelemente $G_{ii}(s)$ von $\mathbf{G}(s)$ *Hauptregelkreise*. Diese beschrei-ben die direkte Wirkung. Alle anderen Elemente $G_{ij}(s)$ mit $i \neq j$ heißen *Querkopplungen*.

Die Querkopplungen beschreiben, wie sich Stellgrößen u_i auf andere Regel-größen y_j auswirken. Diese Auswirkungen sind in der dezentralen Regelung nicht vorgesehen und führen durch entsprechende Verschaltungen zu indi-rekten Wirkungen.

Mehrgrößensysteme, bei denen die Querkopplung nur sehr schwach ausge-prägt ist, können durch Ausnutzung von Nulleinträgen in \mathbf{G} in mehrere einschleifige Regelkreise zerlegt werden, für die ein Entwurf der bereits be-schriebenen dezentralen Regelungen direkt vorgenommen werden kann.

Kann die Querkopplung nicht vernachlässigt werden, so kann das System oft trotzdem durch mehrere Teilregler geregelt werden, wobei beim Regler-entwurf die Querkopplungen berücksichtigt werden müssen.

13.2.3 Polstellen von Mehrgrößensystemen

Für SISO-Systeme ist bekannt, dass die Polstellen der Übertragungsfunkti-on $G(s)$ im Falle einer minimalen Realisierung genau den Eigenwerten der Matrix \mathbf{A} der Zustandsraumdarstellung entsprechen. Da die Zustandsraum-darstellung für Mehrgrößensysteme identisch zu der für Eingrößensysteme ist, gilt dieser Zusammenhang folglich auch hier. Daher ist es bei einem Mehrgrößensystem in Zustandsraumdarstellung sehr einfach, die System-ordnung und die Polstellen des Systems abzulesen.

Liegt das Mehrgrößensystem als Übertragungsmatrix $\mathbf{G}(s)$ vor, so gestaltet sich die Bestimmung der Polstellen schwieriger, was vor allem daran liegt, dass die Systemordnung und die Vielfachheit der Polstellen nicht direkt er-sichtlich ist. Das Vorgehen zur Bestimmung der Polstellen ist daher, eine

Realisierung von $\mathbf{G}(s)$ im Zustandsraum zu suchen und deren Eigenwerte zu bestimmen. Hierbei muss aber darauf geachtet werden, dass es sich um eine minimale Realisierung handelt, da ansonsten zusätzliche Polstellen entstehen.

Eine sehr einfache, in vielen Fällen aber nicht minimale Realisierung, kann man erhalten, indem man jedes Element $G_{ij}(s)$ der Übertragungsmatrix $\mathbf{G}(s)$ in der für Eingrößensysteme bekannten Weise in ein Zustandsraummodell überführt und anschließend aus allen erhaltenen Teilmustern ein Gesamtmodell als Blockmatrix zusammensetzt. So kann man die Übertragungsmatrix

$$\mathbf{G}(s) = \begin{bmatrix} G_1(s) & 0 \\ 0 & G_2(s) \end{bmatrix} \quad (13.8)$$

durch die Zustandsraummodelle $(\mathbf{A}_1, \mathbf{b}_1, \mathbf{c}_1^T, d_1)$ und $(\mathbf{A}_2, \mathbf{b}_2, \mathbf{c}_2^T, d_2)$ für $G_1(s)$ bzw. $G_2(s)$ in die Zustandsraumdarstellung

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{b}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{b}_2 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \mathbf{c}_1^T & \mathbf{0} \\ \mathbf{0} & \mathbf{c}_2^T \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix} \quad (13.9)$$

überführen.

Dieser Weg führt leider im Allgemeinen nicht auf eine minimale Realisierung. Eigenvorgänge des Systems werden nämlich mehrfach abgebildet, wenn verschiedene $G_{ij}(s)$ dieselben Polstellen besitzen.

Eine minimale Realisierung kann jedoch keine zusätzlichen Eigenwerte im Vergleich zu der obigen sehr einfach gewonnenen Zustandsraumdarstellung besitzen. Daher folgt:

Polstellen von MIMO-Systemen

Polstellen des MIMO-Systems entsprechen – abgesehen von der Vielfachheit – genau den Polstellen der Einzelemente $G_{ij}(s)$. Folglich ist das Mehrgrößensystem $\mathbf{G}(s)$ genau dann stabil, wenn alle Einzelemente $G_{ij}(s)$ stabil sind.

Für die allermeisten Zwecke reicht es aus, die Position der Polstellen, nicht aber deren Vielfachheit zu kennen. In einigen Fällen muss aber auch die

Vielfachheit bekannt sein. Neben der Aufstellung einer minimalen Realisierung ist dies insbesondere für Anwendungen des vollständigen Nyquist-Kriteriums unabkömmlich, dessen Ergebnis von der Anzahl der instabilen Polstellen und damit auch von deren Vielfachheit abhängt.

Aus expliziten Konstruktionsvorschriften für eine minimale Realisierung, die sukzessive nicht steuer- und nicht beobachtbare Zustände entfernt, lässt sich das Folgende herleiten [53]:

Charakteristisches Polynom von MIMO-Systemen

Das charakteristische Polynom $p(s)$ einer minimalen Realisierung eines MIMO-Systems ist der kleinste gemeinsame Nenner von allen Determinanten und Unterdeterminanten von $\mathbf{G}(s)$.

Da die Unterdeterminanten erster Ordnung den einzelnen Elementen G_{ij} entsprechen, ist hier auch der vorherigen Ausführung Rechnung getragen.

Als Beispiel für die Bestimmung der Pole eines MIMO-Systems sei folgende Übertragungsmatrix \mathbf{G} gegeben:

$$\mathbf{G} = \begin{bmatrix} \frac{1}{s+1} & 0 & \frac{1-s}{(s+1)(s+2)} \\ \frac{1}{s-1} & \frac{1}{s+2} & \frac{1}{s+2} \end{bmatrix} \quad (13.10)$$

Die Unterdeterminanten zweiter Ordnung von Gl.(13.10) ergeben sich ohne Beachtung des Vorzeichens zu

$$\frac{s-1}{(s+1)(s+2)^2}, \quad \frac{2}{(s+1)(s+2)}, \quad \frac{1}{(s+1)(s+2)} \quad (13.11)$$

Zusammen mit den Unterdeterminanten erster Ordnung ergibt sich der kleinste gemeinsame Nenner zu

$$p(s) = (s+1) \cdot (s+2)^2 \cdot (s-1). \quad (13.12)$$

Die Polstellen des MIMO-Systems \mathbf{G} ergeben sich folglich zu $-1, -2, -2$ und $+1$. Dieses Ergebnis konnte ohne eingehende Analyse nicht abgelesen werden, da beispielsweise der Eigenwert $s = -1$ ebenfalls zweimal in \mathbf{G} enthalten ist, sich jedoch nicht in einer doppelten Polstelle niederschlägt.

13.2.4 Richtungsabhängige Verstärkung

Bei linearen Systemen bietet der Frequenzgang eine Beschreibung des Verhältnisses der Zeiger der harmonischen Ausgangs- und Eingangsgröße. Für ein stabiles SISO-System ist die Amplitudenverstärkung dabei durch $|G(j\omega)|$ beschrieben und ist daher nur von der Frequenz ω der Eingangsgröße u abhängig, nicht aber von deren Amplitude. Dies liegt daran, dass die einfache Rechnung

$$\frac{|Y(j\omega)|}{|U(j\omega)|} = \frac{|G(j\omega) \cdot U(j\omega)|}{|U(j\omega)|} = |G(j\omega)| \cdot \frac{|U(j\omega)|}{|U(j\omega)|} = |G(j\omega)| \quad (13.13)$$

wegen der Multiplikativität des Betrages $|a \cdot b| = |a| \cdot |b|$ im SISO-Fall Gültigkeit hat.

Für MIMO-Systeme sind $\mathbf{Y}(j\omega)$ und $\mathbf{U}(j\omega)$ vektorwertige Größen, weshalb eine Norm $\|\cdot\|$ anstelle des Betrages Verwendung finden muss. Bei einem analogen Vorgehen ergibt sich hier

$$\frac{\|\mathbf{G}(j\omega) \cdot \mathbf{U}(j\omega)\|}{\|\mathbf{U}(j\omega)\|} \leq \|\mathbf{G}(j\omega)\| \cdot \frac{\|\mathbf{U}(j\omega)\|}{\|\mathbf{U}(j\omega)\|} = \|\mathbf{G}(j\omega)\| \quad , \quad (13.14)$$

da hier keine Multiplikativität, sondern nur Submultiplikativität gegeben ist. Welcher der Werte kleiner gleich der Matrixnorm $\|\mathbf{G}(j\omega)\|$ angenommen wird, hängt dabei von der Richtung des Eingangsvektors $\mathbf{U}(j\omega)$ ab.

Richtungsabhängige Verstärkung

Wenn ein stabiles MIMO-System mit einer Eingangsschwingung $\mathbf{U}(j\omega)$ angeregt wird, hängt die Amplitude der Ausgangsschwingung $\mathbf{Y}(j\omega)$ nicht nur von der Frequenz ω , sondern auch von der Richtung von \mathbf{U} ab.

Dies kann man mit den Eigenvektoren und Eigenwerten quadratischer Matrizen plausibilisieren: Vektoren in Richtung von Eigenvektoren, deren Eigenwerte große Beträge haben, werden stärker gestreckt als Vektoren in Richtung von Eigenvektoren, deren Eigenwerte kleine Beträge haben. Dies lässt sich anwenden auf die Eingangsvektoren $\mathbf{U}(j\omega)$. Da $\mathbf{G}(j\omega)$ jedoch nicht quadratisch sein muss, weicht man auf das Konzept der Singulärwerte aus, die auch für nicht quadratische Matrizen definiert sind.

Singulärwerte

Die Singulärwerte $\sigma_i(\mathbf{M})$ einer Matrix \mathbf{M} sind die positiven Quadratwurzeln der Eigenwerte von $\mathbf{M}^T \mathbf{M}$, also

$$\sigma(\mathbf{M}) = \sqrt{\lambda(\mathbf{M}^T \mathbf{M})} . \quad (13.15)$$

Geometrisch kann man die Singulärwerte als richtungsabhängige Streckungsfaktoren interpretieren, wobei $\bar{\sigma}$ den größten und $\underline{\sigma}$ den kleinsten Singulärwert bezeichnet.

Abhängig von der Richtung des Vektors ergibt sich die tatsächliche Verstärkung zu

$$\underline{\sigma} \leq \frac{\|\mathbf{Y}(j\omega)\|}{\|\mathbf{U}(j\omega)\|} \leq \bar{\sigma} , \quad (13.16)$$

wobei der tatsächliche Wert beispielsweise durch Ausführung der Matrix-Vektor-Multiplikation $\mathbf{G}(j\omega) \cdot \mathbf{U}(j\omega)$ ermittelt wird.

Eine anschauliche Interpretation dieser richtungsabhängigen Verstärkung liefert das Beispiel eines Tanks mit dem Füllstand als Regelgröße y , einem Ventil u_1 im Zulauf und einem Ventil u_2 im Ablauf. Jedes Ventil sei für $u = 1$ vollständig geöffnet und für $u = -1$ vollständig geschlossen, während $u = 0$ dem aktuellen Arbeitspunkt entspricht. Hier hat ein gleichzeitiges Öffnen beider Ventile ($\mathbf{u} = [1 \ 1]^T$) einen viel geringeren Einfluss auf y als ein wechselseitiges Schließen und Öffnen ($\mathbf{u} = [1 \ -1]^T$), obwohl $\|\mathbf{u}\|$ für beide Eingangsvektoren gleich ist.

Die geschilderte Richtungsabhängigkeit der Verstärkung kann den Reglerentwurf erschweren, was anschaulich dadurch eingesehen werden kann, dass die Durchtrittsfrequenzen ω_d mit $|G(j\omega_d)| = 1$ im SISO-Fall eine entscheidende Rolle für die Stabilität des Regelkreises spielt. Durch die Richtungsabhängigkeit entsteht für große Unterschiede zwischen $\underline{\sigma}$ und $\bar{\sigma}$ ein großes Durchtrittsfrequenzband, was die Reglerauslegung erschwert.

Richtungsabhängigkeit kann auch durch schlechte Skalierung der Ein- und Ausgangssignale entstehen oder verstärkt werden. Entsprechende Skalierungsmatrizen können diese Effekte minimieren. Anschaulich kann die Bedeutung der Skalierung so verstanden werden, dass die Werte der Ein- und Ausgangssignale der Prozesse von der gewählten physikalischen Einheit ab-

hängen. Je nach Größe der Einheit ist keine Vergleichbarkeit bei unterschiedlichen Ein- und Ausgangsgrößen gegeben. So ist es nicht sinnvoll, bei zwei zu regelnden Drücken den einen in Pascal, den anderen in bar zu messen.

Zur Durchführung der Skalierung werden Skalierungsmatrizen als Diagonalmatrizen verwendet. Um diese aufzustellen, ermittelt man die betragsmäßig maximal zulässigen Stellgrößen separat für alle Stellgrößen und schreibt diese in eine Diagonalmatrix \mathbf{S}_u . Mit einem analogen Vorgehen für den größten betragsmäßig zulässigen Regelfehler (oder alternativ die größte zu erwartende Sollwertänderung) erhält man die Diagonalmatrix \mathbf{S}_y .

Bezeichnet man die unskalierten Größen mit $\hat{\mathbf{u}}$ und $\hat{\mathbf{y}}$ und die skalierten Größen mit \mathbf{u} und \mathbf{y} , so ergibt sich der Zusammenhang

$$\mathbf{S}_u \mathbf{u} = \hat{\mathbf{u}} \quad , \quad \mathbf{S}_y \mathbf{y} = \hat{\mathbf{y}} \quad (13.17)$$

und damit die skalierte Prozessbeschreibung

$$\mathbf{y} = \mathbf{S}_y^{-1} \hat{\mathbf{G}} \mathbf{S}_u \mathbf{u} = \mathbf{G} \cdot \mathbf{u} \quad , \quad (13.18)$$

wobei die Skalierungsmatrizen durch ihre Konstruktion invertierbar sind.

13.3 Verfahren der dezentralen Regelung

13.3.1 Relative Gain Array

Bei einer dezentralen Regelung werden – wie beschrieben – die jeweiligen Regelgrößen von separaten und voneinander unabhängigen Reglern geregelt und das Problem der Mehrgrößenregelung somit wieder auf das einer SISO-Regelung zurückgeführt.

Hierfür ist es in einem ersten Schritt notwendig, passende Paarungen von Ein- und Ausgangsgrößen für die einschleifigen Regelkreise zu ermitteln, wie in der Permutationsmatrix in Gl.(13.7) kodiert.

Es liegt nahe, diese Paarung durch das Verhältnis von direkter und indirekter Wirkung der jeweiligen Ein- und Ausgangsgrößen zu bestimmen.

Für das lineare MIMO-System $\mathbf{Y}(s) = \mathbf{G}(s) \cdot \mathbf{U}(s)$ wird die direkte Wirkung

von u_j auf y_i durch

$$\left. \frac{\partial y_i}{\partial u_j} \right|_{u_k \neq j = \text{const.}} = [G(s)]_{ij} \quad (13.19)$$

beschrieben, d. h. außer u_j werden alle Eingänge identisch Null gewählt.

Bei einer Regelung mit mehreren Einzelregelkreisen entsteht neben dieser direkten Wirkung jedoch auch eine indirekte, da u_j nicht nur auf y_i , sondern auch auf $y_{k \neq i}$ wirkt. Das erzeugt eine Regelabweichung in den anderen Regelkreisen, weswegen sich auch die anderen $u_{l \neq j}$ ändern, die wiederum einen Einfluss auf y_i haben.

Die Gesamtwirkung (das ist die Summe aus direkter und indirekter Wirkung) von u_j auf y_i ergibt sich, wenn alle anderen Ausgänge $y_{k \neq j}$ perfekt eingeregelt werden. Betrachtet man dazu die ideale Stellgröße $\mathbf{U}(s) = \mathbf{G}^{-1}(s) \cdot \mathbf{Y}(s)$, so ergibt sich

$$\left. \frac{\partial u_j}{\partial y_i} \right|_{y_{k \neq i} = \text{const.}} = [G^{-1}(s)]_{ji} \Rightarrow \left. \frac{\partial y_i}{\partial u_j} \right|_{y_{k \neq i} = \text{const.}} = \frac{1}{[G^{-1}(s)]_{ji}}. \quad (13.20)$$

Hiermit lässt sich ein Maß für die Auswahl der Ein- und Ausgangspaare aufstellen:

Relative Gain Array (RGA)

Das Relative Gain Array (RGA) $\Lambda(s)$ ist das Verhältnis der direkten Wirkung und der Gesamtwirkung der Ein- und Ausgangsgrößen. Es berechnet sich zu

$$\Lambda_{ij}(s) = \frac{\left. \frac{\partial y_i}{\partial u_j} \right|_{u_k \neq j = \text{const.}}}{\left. \frac{\partial y_i}{\partial u_j} \right|_{y_{k \neq i} = \text{const.}}} = [G(s)]_{ij} \cdot [G^{-1}(s)]_{ji} \quad . \quad (13.21)$$

Für alle Kombinationen von Ein- und Ausgängen lässt sich die kompakte Formel

$$\Lambda(s) = \mathbf{G}(s) \cdot \times (\mathbf{G}^{-1}(s))^T \quad (13.22)$$

angeben, wobei $\cdot \times$ die elementweise Multiplikation bezeichnet.

Aus Gl.(13.21) ist zudem ersichtlich, dass die Spalten- und Zeilensumme des RGA immer eins beträgt, da alle direkten Wirkungen zusammen die Gesamtwirkung ergeben. Zudem ist das RGA unabhängig von Skalierungen und konsistent gegenüber Permutationen (d. h. Umnummerierung der Eingänge und Ausgänge).

Es liegt auf der Hand, dass bei der Zuordnung von Stell- und Regelgrößen solche Kombinationen vorteilhaft sind, bei denen die direkte Wirkung näherungsweise der Gesamtwirkung entspricht. Da das Relative Gain Array von s und damit der betrachteten Frequenz abhängt, hat sich die Heuristik als nützlich erwiesen, solche Zuordnungen zu bevorzugen, bei denen die jeweiligen RGA-Elemente bis zur gewünschten Bandbreite des Regelkreises etwa eins sind. Hierbei sind Zuordnungen mit negativen Einträgen zu vermeiden, da dies impliziert, dass die direkte Wirkung und die Gesamtwirkung unterschiedliche Vorzeichen haben, was den Gesamtregelkreis leicht destabilisiert.

Zur Veranschaulichung wird das RGA des MIMO-Systems in Gl.(13.23) ermittelt, um eine gute Paarung von Ein- und Ausgangssignalen zu bestimmen:

$$\mathbf{G}(s) = \begin{bmatrix} \frac{2}{s+1} & \frac{2}{s+2} & \frac{1}{s+5} \\ \frac{3}{s+1} & -\frac{2}{s+1} & 0 \\ \frac{7}{s+2} & -\frac{1}{s+5} & \frac{2}{s+5} \end{bmatrix}. \quad (13.23)$$

Da die gewünschte Bandbreite noch nicht feststeht, wird hier nur das statische Relative Gain Array mit $s = 0$ betrachtet. Für die Berechnung von $\mathbf{\Lambda}$ müssen durch Ausnutzung der Zeilen- und Spaltensumme nur vier Elemente explizit berechnet werden. Aufgrund der elementweisen Multiplikation kann dabei der Nulleintrag in G_{23} in $\mathbf{\Lambda}(0)$ übernommen werden, was einer nicht existenten direkten Wirkung entspricht. Es berechnet sich

$$\mathbf{\Lambda}(0) \approx \begin{bmatrix} 1,05 & \mathbf{0,79} & -0,84 \\ \mathbf{0,87} & 0,13 & 0 \\ -0,92 & 0,08 & \mathbf{1,84} \end{bmatrix}. \quad (13.24)$$

Für eine gute Paarung kommt zur Regelung von y_3 nur u_3 in Frage, da alle anderen Einträge der dritten Zeile negativ oder nahezu null sind. Obwohl zur Regelung von y_1 mit $1,05 \approx 1$ u_1 der beste Kandidat wäre, so ist wegen $\Lambda_{22} = 0,13$ nur die hervorgehobene Paarung zielführend.

Wegen des stärker von 1 abweichenden Eintrages sind im Regelkreispaar $y_3 \rightarrow u_3$ stärkere Kopplungen durch die anderen Regelkreise zu erwarten. Konkret bedeutet der Wert von etwa 2, dass die Wirkung von u_3 auf y_3 halbiert wird, wenn auch die anderen Regelkreise geschlossen werden. Dies muss bei der Auslegung der Einschleifenregler entsprechend berücksichtigt werden.

Die Permutationsmatrix der dezentralen Ausgangsrückführung ist

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} . \quad (13.25)$$

Das RGA als heuristisches Maß für die Paarung von Ein- und Ausgangsgrößen erlaubt außer in Sonderfällen keine Aussage über die Stabilität des geschlossenen Regelkreises, selbst wenn alle Einzelregelkreise stabil arbeiten. Auch liefert das RGA kein Maß für die tolerierbare Abweichung von RGA-Einträgen von dem Idealwert 1.

13.3.2 MIMO-Nyquist und Diagonaldominanz

Ein notwendiges und hinreichendes Kriterium für die Stabilität des geschlossenen Regelkreises ist, dass die Nullstellen des charakteristischen Polynoms alle in der linken offenen s -Halbebene liegen. Diese erhält man aus der Rückführdifferenzmatrix zu

$$\det(\mathbf{F}(s)) = \det(\mathbf{I} + \mathbf{G}(s)\mathbf{K}(s)) = 0 . \quad (13.26)$$

Analog zum Nyquist-Kriterium für SISO-Systeme lässt sich hier ein entsprechendes Pendant für MIMO-Systeme formulieren, welches sich mit punktuellen Modifikationen besonders gut für den Einsatz bei dezentralen Reglern eignet. Eine Analyse der Herleitung des Nyquist-Kriteriums für SISO-Systeme zeigt hierbei, dass für einen Übertrag auf den Mehrgrößenfall nur die Nullstellen n von $1 + G_0(j\omega)$ durch die Nullstellen von $\det(\mathbf{F}(j\omega))$ ersetzt werden müssen. Daraus folgt:

Nyquist-Kriterium für Mehrgrößensysteme

Wenn die Übertragungsfunktion $\mathbf{G}_0(s)$ des aufgeschnittenen Regelkreises

p Pole in der rechten offenen s -Halbebene aufweist, dann gilt:

Der geschlossene Regelkreis ist genau dann stabil, wenn die Ortskurve des Frequenzganges $\det(\mathbf{F}(j\omega))$ beim Durchlaufen der Frequenzwerte von $-\infty$ über null bis $+\infty$ den Ursprung genau p Mal im mathematisch positiven Sinn (d. h. gegen den Uhrzeigersinn) umfährt.

Hier ist zu beachten, dass es sich bei der Determinante um eine skalare Größe handelt, für die wie üblich Ortskurven gezeichnet und Umdrehungen gezählt werden können. Da der Punkt -1 bereits in der Determinanten durch die Identität \mathbf{I} in \mathbf{F} berücksichtigt wurde, sind hier die Umdrehungen um den Ursprung zu zählen.

In dieser Form eignet sich die Mehrgrößen-Version des Nyquist-Kriteriums noch nicht für den Reglerentwurf, da der Verlauf der zu betrachtenden Ortskurve wegen der Determinanten auf undurchsichtige Weise von $\mathbf{K}(s)$ abhängt.

Glücklicherweise lässt sich aber leicht eine zweite, nutzbringendere Version gewinnen. Hierzu wird die Eigenschaft genutzt, dass sich die Determinante als Produkt der Eigenwerte ergibt, also

$$\det(\mathbf{I} + \mathbf{G}_0(j\omega)) = \prod_{i=1}^q (1 + \lambda_i(j\omega)), \quad (13.27)$$

wobei λ_i die q Eigenwerte von $\mathbf{G}_0(j\omega)$ bezeichnet. Daher lässt sich die Phasendrehung der Determinanten-Ortskurve durch die Summe der Phasendrehungen der Terme $(1 + \lambda_i(j\omega))$ ersetzen. Als zweite Variante kann also das folgende Kriterium angegeben werden:

Nyquist-Kriterium für Mehrgrößensysteme (Variante)

Wenn die Übertragungsfunktion $\mathbf{G}_0(s)$ des aufgeschnittenen Regelkreises p Pole in der rechten offenen s -Halbebene aufweist, dann gilt:

Der geschlossene Regelkreis ist genau dann stabil, wenn die Ortskurven der Eigenwerte des Frequenzganges $\mathbf{F}(j\omega)$ beim Durchlaufen der Frequenzwerte von $-\infty$ über null bis $+\infty$ den Ursprung zusammen genau p Mal im mathematisch positiven Sinn (d. h. gegen den Uhrzeigersinn) umfahren.

Für vollständig entkoppelte Systeme ist diese Formulierung des Nyquist-Kriteriums für Mehrgrößensysteme identisch zum bekannten Eingrößenfall.

Dort entspricht \mathbf{G}_0 nämlich einer Diagonalmatrix und die Determinante dem Produkt der Diagonalelemente.

Allerdings lässt sich mithilfe des Gershgorin¹-Theorems eine einfache Abschätzung der Eigenwerte ableiten, durch die das so gewonnene Stabilitätskriterium jedoch nur noch hinreichend und nicht mehr notwendig ist. Dennoch scheint durch diese zweite Version nicht viel gewonnen, da auch der Zusammenhang zwischen den Eigenwerten und $\mathbf{K}(s)$ unklar ist. Das Gershgorin-Theorem liefert dabei Aussagen über das Gebiet, in welchem sich die Eigenwerte einer Matrix \mathbf{M} befinden können.

Gershgorin-Theorem

Zu jeder quadratischen Matrix \mathbf{M} und zu jedem Eigenwert λ_i von \mathbf{M} existiert ein Diagonalelement M_{jj} , sodass der Eigenwert nicht weiter entfernt von M_{jj} ist als die Summe der Beträge der anderen Elemente in derselben Zeile oder Spalte wie M_{jj} :

$$|\lambda_i - M_{jj}| \leq \sum_{k \neq j} |M_{kj}| = r_j \text{ und } |\lambda_i - M_{jj}| \leq \sum_{k \neq j} |M_{kj}| = R_j. \quad (13.28)$$

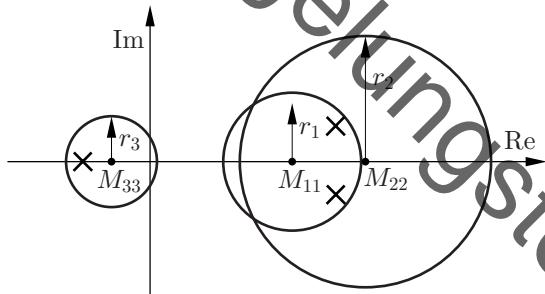


Bild 13-3: Visualisierung des Gershgorin-Theorems

Dieser Zusammenhang der Eigenwerte mit den Diagonalelementen ist in Bild 13-3 visualisiert. Da die Diagonalelemente von $\mathbf{G}_0(j\omega)$ im Falle einer dezentralen Regelung genau die Hauptregelkreise sind, können sich die gesuchten Ortskurven der Eigenwerte nicht allzu weit von den Ortskurven der

¹Семён Аронович Гершгорин (1901-1933), sowjetischer Mathematiker [16]

Hauptregelkreise entfernen, sofern die Ausdrücke r_j oder R_j klein sind. Da für die Stabilitätsbetrachtung nur die Umdrehungen der Ortskurve entscheidend sind, können derartige Abweichungen zugelassen werden, sofern sich die Umdrehungszahl hierdurch nicht ändert. Dies motiviert die Definition der Diagonaldominanz.

Diagonaldominanz

Eine Übertragungsmatrix $\mathbf{G}(j\omega)$ heißt zeilendominant, wenn

$$|G_{ii}(j\omega)| > \sum_{k \neq i} |G_{ik}(j\omega)| \quad (13.29)$$

und spaltendominant, wenn

$$|G_{ii}(j\omega)| > \sum_{k \neq i} |G_{ki}(j\omega)|. \quad (13.30)$$

Eine Übertragungsmatrix ist diagonaldominant, wenn sie entweder zeilen- oder spaltendominant ist.

Mit diesen Vorüberlegungen lässt sich eine leicht zu prüfende hinreichende Stabilitätsbedingung angeben.

Stabilitätskriterium über Diagonaldominanz

Wenn die Übertragungsfunktion $\mathbf{G}_0(s)$ des aufgeschnittenen Regelkreises p Pole in der rechten s -Halbebene aufweist und die Rückführdifferenzmatrix $\mathbf{F}(j\omega)$ diagonaldominant ist, dann gilt:

Wenn die Ortskurven der Diagonalelemente $G_{0,ii}(j\omega)$ beim Durchlaufen der Frequenzwerte von $-\infty$ über null bis $+\infty$ den Punkt -1 zusammen genau p Mal im mathematisch positiven Sinn umfahren, dann ist der geschlossene Regelkreis stabil.

Es gibt entsprechende Verallgemeinerungen, die dieses hinreichende Stabilitätskriterium zu einem notwendigen ergänzen, auf die hier aber nicht näher eingegangen werden soll [44]. Diagonaldominanz kann in diesem Sinne als ein Maß interpretiert werden, dass sicherstellt, dass die Querkopplungen die stabilen Einzelregelkreise nicht destabilisieren.

Der Entwurf einer dezentralen Regelung erfolgt nun üblicherweise iterativ:

Nach der Zuordnung von Stell- und Regelgrößen mithilfe des RGA werden die Einzelregelkreise ausgelegt. Ist die nachfolgende Überprüfung der Diagonaldominanz positiv, so ist der Reglerentwurf abgeschlossen. Ist die Rückführdifferenzmatrix jedoch nicht diagonaldominant, so müssen die Einzelregler entsprechend modifiziert werden.

13.4 Verfahren der zentralen Regelung

13.4.1 Zentrale Regelung im Zustandsraum

Dezentrale Regler haben neben der dezentralen Implementierung den Vorteil, dass die Einzelregelkreise – mit dem Zusatz der Diagonaldominanz – nach bekannten SISO-Verfahren ausgelegt werden können. Nachteilig ist allerdings, dass die Querkopplungen nur über Abschätzungen in Form des Gershgorin-Theorems und nicht explizit berücksichtigt werden. Hierdurch büßen dezentrale Regler oft an Leistungsfähigkeit ein und können bei ungünstigen Konstellationen und starken Querkopplungen faktisch nicht zur Regelung eingesetzt werden.

Bei einer zentralen Regelung werden die Querkopplungen im Reglerentwurf berücksichtigt und eine potentiell voll besetzte Reglerübertragungsmatrix $\mathbf{K}(s)$ entworfen. Hierdurch entspricht der Entwurf nicht mehr dem von mehreren Einzelregelkreisen, weswegen die Entwurfsv erfahren für SISO-Regler teilweise angepasst werden müssen.

Da die Beschreibungsform des Zustandsraumes einen einheitlichen Zugang sowohl für SISO- als auch für MIMO-Systeme bietet, ist es naheliegend, Zustandsregler als zentrale Regler einzusetzen. Bei einer dezentralen Regelung würden einzelne Zustandsregler $u_i = -\mathbf{k}_i \mathbf{x}_i$ eingesetzt, welche die Zustände des i -ten Teilsystems auf die Stellgröße u_i zurückführen.

Fasst man alle Zustände zu einem Gesamtzustand \mathbf{x} zusammen, so kann man stattdessen

$$\mathbf{u} = -\mathbf{K}\mathbf{x} \tag{13.31}$$

als zentralen Regler ansetzen. Hier ist \mathbf{K} eine Rückführmatrix anstelle eines Rückführvektors \mathbf{K} .

Der geschlossene Regelkreis ergibt sich dann zu

$$\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{B}\mathbf{K})\mathbf{x}. \tag{13.32}$$

Die Rückführmatrix \mathbf{K} kann mit denselben Überlegungen wie im SISO-Fall (z. B. durch eine Polvorgabe) ausgelegt werden. Auch die Verfahren der optimalen Regelung in Kapitel 18 sind anwendbar.

Allerdings sind bei diesem Vorgehen ein paar Besonderheiten zu beachten. Es seien hierzu \mathbf{b}_i die vektorwertigen Spalten von \mathbf{B} . Diese beschreiben den Einfluss von u_i auf die gesamte Systemdynamik. Sofern die verschiedenen Regelgrößen nicht voneinander entkoppelt sind, ist davon auszugehen, dass u_i auf alle Regelgrößen einen Einfluss hat. Folglich ist es nicht ungewöhnlich, wenn das Paar $(\mathbf{A}, \mathbf{b}_i)$ steuerbar ist. Dann könnte aber bereits mit dem Zustandsregler $u_i = -\mathbf{k}_i \mathbf{x}$, welcher nur die i -te Stellgröße nutzt, die Polvorgabe vollständig gelöst werden. Die entstehenden Gleichungssysteme sind daher im MIMO-Fall meist überbestimmt.

Die dadurch entstehenden zusätzlichen Freiheitsgrade können verschieden genutzt werden. In der Literatur [30] finden sich neben dem Verfahren der optimalen Regelung Ansätze, die zusätzliche Größen neben den Eigenwerten (wie z. B. Eigenvektoren) durch die Zustandsrückführung \mathbf{K} vorgeben. Andere Verfahren machen Vorgaben für die Struktur von \mathbf{K} und fordern, dass nach Möglichkeit die Ausgänge \mathbf{y} anstelle der Zustände \mathbf{x} zurückgeführt werden sollen, oder möglichst viele Einträge von \mathbf{K} verschwinden. Durch Letzteres fordert man, dass der zentrale Zustandsregler möglichst dezentral arbeiten soll, um die Implementierungsvorteile einer dezentralen Regelung mitzunehmen.

13.4.2 Entkopplungsregler

Neben dem Entwurf zentraler Regler im Zustandsraum besitzt der Ansatz der Entkopplungsregler eine gewisse Prominenz. Zu deren Motivation wird das Eingangsbeispiel der Flüssigkeitsmischstation erneut aufgegriffen. Dort sorgte eine Veränderung des Kaltwasserventils für eine Verringerung der Temperatur und eine Erhöhung des Durchflusses, wodurch starke Querkopplung entstanden.

Die intuitive Lösung, die in heutigen Mischbatterien umgesetzt ist, besteht darin, die Stellgrößen u_1 („Ventilstellung warm“) und u_2 („Ventilstellung kalt“) durch passendere Stellgrößen zu ersetzen. Diese sind $\tilde{u}_1 = u_1 - u_2$ („Differenz der Ventilstellungen“) und $\tilde{u}_2 = u_1 + u_2$ („Summe der Ventilstellungen“). Mit diesen alternativen Stellgrößen wird \tilde{u}_2 die Temperatur

nur unwesentlich beeinflussen, aber großen Einfluss auf den Durchfluss haben. Somit eignet sich \tilde{u}_2 in besonderer Weise zur Durchflussregelung. Die umgekehrte Argumentation gilt für \tilde{u}_1 und die Temperaturregelung.

Da \tilde{u}_1 und \tilde{u}_2 nicht die tatsächlichen Stellgrößen sind, erweitert man den eigentlich dezentralen Regler um ein Regelkreiselement $L(s)$, welches die Zuordnung von \tilde{u}_1 und \tilde{u}_2 zu u_1 und u_2 vornimmt. Das Ziel von $L(s)$ ist es, dafür zu sorgen, dass das Verhalten von \tilde{u}_1 auf y_1 und \tilde{u}_2 auf y_2 näherungsweise dem unabhängigen Eingrößensystem entspricht. Daher wird $L(s)$ auch *Entkopplungsregler* genannt.

Entkopplungsregler

Ein Regelkreiselement $L(s)$, welches aus einem gekoppelten Mehrgrößensystem mehrere voneinander unabhängige Eingrößensysteme erzeugt, heißt *Entkopplungsregler*.

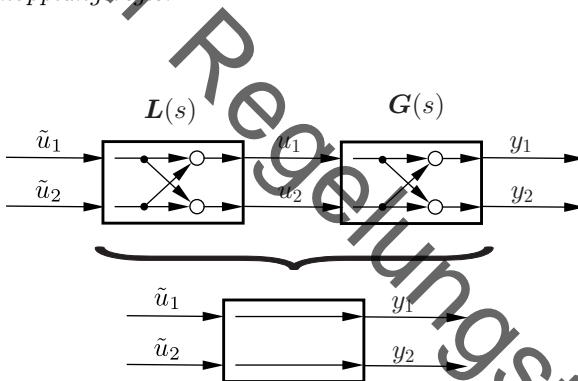


Bild 13-4: Grundidee des Entkopplungsreglers für zwei Stell- und zwei Regelgrößen

Diese Grundidee ist in Bild 13-4 für ein System mit zwei Stell- und zwei Regelgrößen dargestellt. Hieran sieht man auch, dass der Entkopplungsregler strukturell keiner Regelung, sondern einer Steuerung entspricht. Da er aber üblicherweise in Form einer Reihenschaltung in die Rückführung des Regelkreises integriert wird, besitzt die Bezeichnung als Entkopplungsregler eine gewissen Berechtigung.

Beim Beispiel der Flüssigkeitsmischstation würde die beschriebene Zuord-

nung einem statischen Entkopplungsregler

$$\mathbf{L}(0)\tilde{\mathbf{u}} \stackrel{!}{=} \mathbf{u} \quad \Rightarrow \quad \mathbf{L}(0) = \begin{bmatrix} 0,5 & 0,5 \\ -0,5 & 0,5 \end{bmatrix} \quad (13.33)$$

entsprechen.

Die lineare Streckenübertragungsfunktion wird durch geeignete Skalierung im statischen Fall die Form

$$\mathbf{y} = \mathbf{G}(0)\mathbf{u} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \mathbf{u} \quad (13.34)$$

besitzen, da u_1 Temperatur und Durchfluss erhöht (erste Spalte), während u_2 die Temperatur senkt und den Durchfluss erhöht (zweite Spalte).

Somit entspricht der Entkopplungsregler genau

$$\mathbf{L}(0) = \mathbf{G}(0)^{-1} \quad . \quad (13.35)$$

Das ist logisch, da eine ideale Entkopplung bedeutet, dass die Reihenschaltung $\mathbf{L}(s)\mathbf{G}(s)$ eine Diagonalmatrix ist.

Hieraus sieht man auch die Limitierungen des Verfahrens: Der Entwurf eines Entkopplungsreglers entspricht wegen der verfolgten Streckeninvertierung genau der des Vorsteuerungsentwurfs. Zu Schwierigkeiten führen folglich Glieder mit Verzögerung höherer Ordnung und besonders solche mit Totzeit- oder Allpassanteilen. Oft begnügt man sich mit näherungsweiser oder auch mit statischer Entkopplung; bei statischer Entkopplung gelten die Ausdrücke für die Entkopplungsregler entsprechend Gl.(13.35) nur für die Frequenz null.

Der aufgeschnittene Regelkreis ergibt sich unter Verwendung des Entkopplungsreglers und einer dezentralen Regelung zu

$$\mathbf{G}_0(s) = \underbrace{\mathbf{K}(s) \cdot \mathbf{L}(s)}_{\text{zentraler Regler}} \cdot \underbrace{\mathbf{G}(s)}_{\text{Strecke}} = \underbrace{\mathbf{K}(s)}_{\text{dezentral}} \cdot \underbrace{\mathbf{L}(s) \cdot \mathbf{G}(s)}_{\approx \text{diagonal}} \quad . \quad (13.36)$$

Somit ist der Gesamtregler bestehend aus dem dezentralen Regler und dem Entkopplungsregler insgesamt zentral. Der Entwurf des Reglers kann dennoch nach dezentralen Prinzipien erfolgen, wobei der Entkopplungsregler meist sicherstellt, dass die Bedingung an Diagonaldominanz erfüllt sind.

14 Zeitdiskrete Systeme

14.1 Abtastregelungen

14.1.1 Definitionen

Die bisherigen Ausführungen gingen davon aus, dass sich alle Elemente des Regelkreises über Differentialgleichungen beschreiben lassen. Die Signale $f(t)$ als Lösungen dieser Differentialgleichungen sind somit Funktionen $f : \mathbb{R} \rightarrow \mathbb{R}$ der Zeit. Die zugrundeliegende Modellvorstellung ist folglich die einer kontinuierlichen Zeit $t \in \mathbb{R}$, sodass alle Signale zu allen Zeitpunkten existieren und potentiell unterschiedliche Werte annehmen können.

Zeitkontinuierliche Signale und Systeme

Ein Signal $f(t)$ heißt zeitkontinuierlich, wenn für die Zeit t gilt: $t \in \mathbb{R}$.

Ein System heißt zeitkontinuierlich, wenn alle seine Ein- und Ausgangssignale zeitkontinuierlich sind.

Die analoge Wirklichkeit ist zeitkontinuierlich, da physikalische Größen wie Momenten, Spannungen oder Geschwindigkeiten zu allen Zeitpunkten existieren. In der Praxis wird aber nahezu jeder Regelkreis auch Glieder enthalten, welche die Signale nicht kontinuierlich, sondern nur zu diskreten Zeitpunkten übertragen.

Als Beispiel für zeitdiskret arbeitende Systeme muss an erster Stelle der Regler selbst genannt werden. Dieser wird zumeist digital als Algorithmus auf einer Recheneinheit umgesetzt. Dieser Algorithmus wird mit einer Taktrate ausgeführt, da jede Ausführung des Algorithmus Rechenzeit benötigt.

Folglich werden die neu berechneten Stellgrößen des Reglers nicht zu allen Zeitpunkten vorliegen, sondern nur zu diskreten Zeitschritten entsprechend der Taktrate. Ist diese Taktrate konstant, so wären dies die Zeitpunkte

$$t_k = k \cdot T \quad , \quad k \in \mathbb{Z} \tag{14.1}$$

mit der Abtastzeit T , welche zwischen den Aufrufen des Reglers vergeht. Die Zeitpunkte t_k sind nicht kontinuierlich verteilt, sondern bilden eine diskrete Menge. Das gilt auch für den Fall einer nicht-äquidistanten Verteilung der Zeitpunkte.

Zeitdiskrete Signale und Systeme

Ein Signal $f(t)$ heißt zeitdiskret, wenn für die Zeit t gilt: $t \in \mathbb{Z}$.

Ein System heißt zeitdiskret, wenn alle seine Ein- und Ausgangssignale zeitdiskret sind.

Die Eingangs- und Ausgangssignale eines zeitdiskreten Systems sind zwar immer noch Funktionen $f(t)$ der Zeit. Allerdings gilt nun $f(t = t_k) : \mathbb{Z} \rightarrow \mathbb{R}$, sodass sich die Signale als Folgen auffassen lassen. Zeitdiskrete Systeme bilden dann Eingangsfolgen auf Ausgangsfolgen ab.

Um zeitkontinuierliche und zeitdiskrete Signale optisch – ohne explizite Angabe der Zeit – unterscheiden zu können, nutzt man diesen Unterschied zwischen Zeitfunktionen und Folgen in Form folgender Schreibweise:

Folgen und Funktionen

Zeitkontinuierliche Signale sind Funktionen und werden über $f(t)$ dargestellt. Zeitdiskrete Signale sind Folgen und werden über $f_k = f[k]$ dargestellt.

Hierbei entspricht $f[k] = f(t_k)$.

Zeitdiskrete Systeme werden benötigt, um digitale Regler adäquat beschreiben zu können. Aber auch einige Messglieder oder Regelstrecken bedürfen einer zeitdiskreten Beschreibung. Das gilt beispielsweise für digitale Sensoren oder Systeme, denen nur zu bestimmten Zeiten Proben entnommen werden können (z. B. Hochofen beim Abstich).

Auch der Fall kaskadierter Regelkreise ist zu nennen. Dort ist die Regelstrecke des äußeren Regelkreises der innere geschlossene Regelkreis, welcher eine zeitdiskrete Führungsgröße $w[k]$ als Eingang verarbeitet.

Man versucht normalerweise, die digital arbeitenden Teilsysteme mit einer gemeinsamen Abtastzeit (oder zumindest mit einem ganzzahligen Vielfachen) ablaufen zu lassen. Daher wird im Folgenden ausschließlich der Fall der Übertragung zu äquidistanten Zeitpunkten betrachtet. Dies entspricht häufig nicht völlig der Wirklichkeit, vereinfacht aber manche Betrachtungen erheblich.

14.1.2 Abtaster und Halteglied

Zeitkontinuierliche Systeme verarbeiten nur zeitkontinuierliche Eingangssignale, während zeitdiskrete Systeme ausschließlich zeitdiskrete Ausgangssignale haben. Um diese beiden Beschreibungsformen trotzdem in Form von Reihenschaltungen und Rückführungen zu Regelkreisen verbinden zu können, werden daher zusätzliche Elemente benötigt, deren Ein- und Ausgangsgrößen je einmal zeitdiskret und einmal zeitkontinuierlich sind.

Abtaster und Halteglied

Ein Abtaster ist ein System mit zeitkontinuierlichen Eingangsgrößen und zeitdiskreten Ausgangsgrößen. Er wandelt also ein kontinuierliches Signal durch Abtastung in ein diskretes Signal um.

Ein Halteglied ist ein System mit zeitdiskreten Eingangsgrößen und zeitkontinuierlichen Ausgangsgrößen. Es wandelt also ein diskretes Signal durch Halten in ein kontinuierliches Signal um.

Mit diesen Regelkreisgliedern kann ein Regelkreis mit zeitdiskretem Regler und zeitkontinuierlicher Regelstrecke im Wirkungsplan wie in Bild 14-1 dargestellt werden. Dieser Fall soll im Folgenden bevorzugt behandelt werden, da die gewonnenen Hilfsmittel ohne Schwierigkeiten auf andere Kombinationen zeitdiskreter und zeitkontinuierlicher Systeme übertragbar sind.

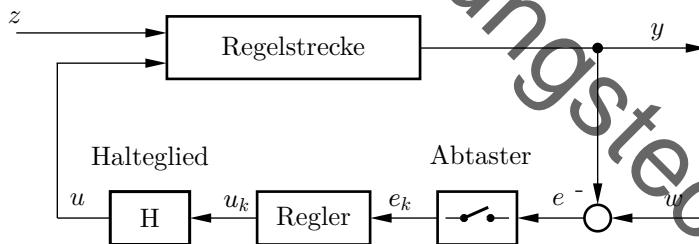


Bild 14-1: Einfache Abtastregelung

Die Umwandlung zeitdiskreter Wertefolgen in kontinuierliche Funktionen, wie sie als Eingangsgrößen kontinuierlich arbeitender Systeme erforderlich sind, ist relativ einfach und es gibt wenige Freiheitsgrade. Das einfachste und in den meisten Fällen benutzte Halteglied 0. Ordnung erzeugt aus einem Eingangswert eine Ausgangsgröße entsprechender Amplitude und hält

diese bis zum Eintreffen des nächsten Eingangswertes konstant – siehe Bild 14-2. Diese Umsetzung sorgt für stückweise konstante Verläufe der zeitkontinuierlichen Ausgangsgrößen.

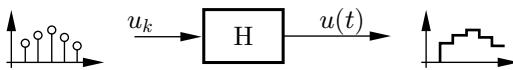


Bild 14-2: Halteglied 0. Ordnung

Für Halteglieder höherer Ordnung, die beispielsweise die Eingangswerte durch Geradenstücke oder Polynome verbinden, müssen zukünftige Eingangswerte vorzeitig bekannt sein. Eine solche aktausale Signalverarbeitung schließt Online-Anwendungen aus. Diese Halteglieder werden daher meist nur für nachträgliche Visualisierungen oder Verarbeitungen genutzt.

Aus dem zeitkontinuierlichen Ausgangssignal des Haltegliedes kann stets die ursprüngliche Eingangsfolge rekonstruiert werden. Dies liegt auch daran, dass in der zeitkontinuierlichen Zeit $t \in \mathbb{R}$ mehr Informationen gespeichert werden können als in der zeitdiskreten Zeit $t_k \in \mathbb{Z}$.

Die zeitliche Diskretisierung eines kontinuierlichen Signals, im folgenden *Abtastung* genannt, kann man so deuten, dass dem kontinuierlichen Signal $y(t)$ eine Folge von Werten $y_k = y[k]$ zugeordnet wird (Bild 14-3).

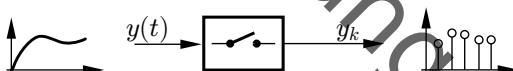


Bild 14-3: Abtaster

14.1.3 Aliasing

Eine entscheidende Frage ist, ob in Analogie zum Fall des Haltegliedes auch für eine Abtastung aus der entstehenden Ausgangsfolge y_k das ursprüngliche Eingangssignal $y(t)$ rekonstruiert werden kann. Da in der zeitkontinuierlichen Zeit $t \in \mathbb{R}$ mehr Informationen gespeichert werden können als in der zeitdiskreten Zeit $t_k \in \mathbb{Z}$, ist dies ohne zusätzliche Annahmen leider nicht der Fall.

Als Beispiel zeigt Bild 14-4 wie ein sinusförmiges Originalsignal mit einer Schwingungsfrequenz von 60 Hz mit einer Frequenz von $f_s = 50 \text{ Hz}$ abgetastet wird.

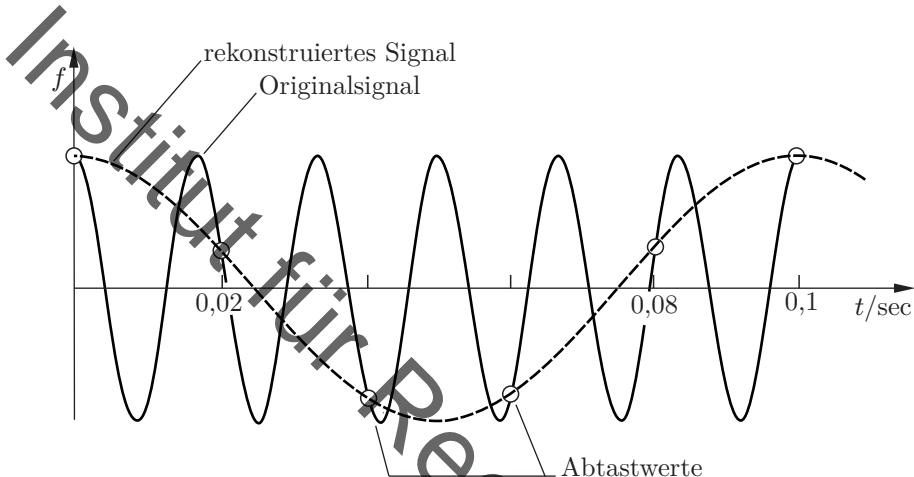


Bild 14-4: Versuch der Rekonstruktion eines Signals aus den Abtastwerten

Wird für die Rekonstruktion des ursprünglichen Signals aus den Abtastwerten vorausgesetzt, dass es sich um ein sinusförmiges Signal handelt, so ergibt sich selbst bei sorgfältigster Durchführung ein sinusförmiges Signal mit einer Frequenz von 10 Hz – der Differenz zwischen Abtastfrequenz und Frequenz des Originalsignals. Dieses Ergebnis verbindet mit dem Original so gut wie nichts. Es ist nicht möglich, das Originalsignal fehlerfrei zu rekonstruieren.

Der Grund hierfür ist offensichtlich, dass die Abtastfrequenz zu gering gewählt wurde. Ein einfaches Gedankenexperiment zeigt, dass für die fehlerfreie Rekonstruktion einer reinen Sinusschwingung durchschnittlich mehr als zwei Funktionswerte pro Periodendauer benötigt werden. Der Grenzfall von genau zwei Funktionswerten pro Periodendauer ist dabei für eine Rekonstruktion gerade nicht ausreichend, da sich die zu bestimmende Sinusschwingung an beiden Abtastzeitpunkten gerade in den Nulldurchgängen befinden könnte. Folglich muss die Abtastfrequenz mehr als doppelt so groß wie die Frequenz des zu rekonstruierenden Signals gewählt werden –

das sogenannte *Shannon¹-Theorem*.

Abtasttheorem nach Shannon

Eine zeitkontinuierliches Signal $f(t)$ kann nur dann aus seinen Abtastwerten $f_k = f(t_k)$ fehlerfrei rekonstruiert werden kann, wenn für die höchste in dieser Funktion enthaltene Frequenz ω_{\max} und die Abtastfrequenz ω_s gilt:

$$2 \cdot \omega_{\max} < \omega_s \quad (14.2)$$

Die Kreisfrequenz $\omega_s/2$ wird auch *Shannon- oder Nyquistfrequenz* genannt. Formuliert man das Shannon-Theorem für die Abtastzeiten anstelle der Frequenzen, so erhält man die Aussage, dass die Abtastzeit kürzer sein muss als die Hälfte der kürzesten Periodendauer im abzutastenden Signal.

Unterabtastung

Verletzt die Abtastzeit das Shannon-Theorem, so spricht man von einer *Unterabtastung*.

Das Beispiel in Bild 14-4 zeigt, dass man durch Unterabtastung nicht nur höherfrequente Signalanteile verliert (den Verlust kann man manchmal durchaus in Kauf nehmen) sondern auch Verfälschungen des Signals im Bereich niedriger Frequenzen erhält.

Aliasing

Liegt eine Unterabtastung vor, so werden durch den Abtastvorgang hohe Frequenzen auf niedrige Frequenzen abgebildet. Dies nennt man *Aliasing*.

Der Name „Aliasing“ röhrt daher, dass sich anschaulich hohe Frequenzen als niedrige Frequenzen „ausgeben“. Dieser Effekt tritt beispielsweise auch in der Computergraphik bei der Abtastung von Bildern auf und führt zu Mustern (wie Treppenstufen), die im Originalbild nicht enthalten sind.

Die durch Aliasing entstehenden Verfälschungen kann man im Allgemeinen nicht durch irgendwelche Nacharbeiten an den Abtastwerten beheben und sie wirken sich kritisch auf die Regelgüte aus. Der Regler wird nämlich üblicherweise so ausgelegt, dass er für niedrige Frequenzen eine hohe Verstärkung, für hohe Frequenzen aber eine niedrige Verstärkung besitzt (siehe

¹Claude Shannon (1916-2001), amerikanischer Elektrotechniker [51]

Abschnitt 10.3). Wird durch Aliasing eine (eigentlich nicht zu verstärkende) hochfrequente Störung auf eine niederfrequente Störung abgebildet, so wird der Regler fälschlicherweise versuchen, diese zu unterdrücken. Dies regt dann die Regelstrecke an, sodass sich eine niederfrequente Regelabweichung einstellt.

Möchte man, um Aliasing zu vermeiden, eine Abtastzeit wählen, die alle im Regelkreis auftretenden Frequenzen gemäß Shannon beinhaltet, so wird man leider enttäuscht. Die regelungstechnisch wichtigsten Signale sind nämlich allesamt nicht bandbegrenzt, d. h. sie enthalten prinzipiell alle möglichen Frequenzen $0 \leq \omega < \infty$.

Dies sieht man am besten im Spektrum eines Signals, das gemäß Definition die Information darüber enthält, welche Frequenz wie stark im Signal vertreten ist. Dort ergibt sich beispielsweise für die Impulsantwort eines PT₁ über die Fouriertransformation

$$f(t) = \begin{cases} e^{-t/T_1} & t > 0 \\ 0 & \text{sonst} \end{cases} \Rightarrow F(j\omega) = \frac{1}{T_1 j\omega + 1} . \quad (14.3)$$

Zwar fällt der Beitrag mit steigender Frequenz – vor allem nach der Eckkreisfrequenz $\omega_E = 1/T_1$ – kontinuierlich ab. Dennoch enthält das Signal prinzipiell alle möglichen Frequenzen, sodass das Shannon-Theorem in jedem Fall verletzt wird.

Dies verschlechtert zwar die Leistungsfähigkeit des Regelkreises; allerdings kann dieser Effekt relativ gut durch zwei Maßnahmen kontrolliert werden: Die passende Wahl der Abtastzeit und eine geeignete Filterung der Messsignale.

Wahl der Abtastzeit

Die Abtastzeit im Regelkreis ist zumindest so hoch zu wählen, dass alle für das Verhalten des Regelkreises relevanten Zeitkonstanten abgedeckt werden. Typischerweise wird dies dann erfüllt, wenn die Abtastfrequenz ω_s mindestens um den Faktor fünf größer als die gewünschte Bandbreite ω_g des geschlossenen Regelkreises ist:

$$\omega_s > ! 5 \cdot \omega_g . \quad (14.4)$$

Auch bei dieser Wahl der Abtastzeit wird es Signalanteile geben, die jenseits

der gewählten Abtastfrequenz liegen und zum Aliasing beitragen. Das gilt insbesondere für Messrauschen, dass durch die Unterabtastung in kritische Frequenzbereiche verschoben wird.

Dem Aliasing wirkt man dadurch entgegen, dass man die unerwünschten höherfrequenten Anteile in den abzutastenden zeitkontinuierlichen Funktionen vor dem Abtasten durch ein Tiefpassfilter unterdrückt.

Anti-Aliasing-Filter

Ein (analoges) Tiefpassfilter, dessen Grenzfrequenzen in Bezug auf das Shannon-Theorem gewählt wurden, heißt auch *Anti-Aliasing-Filter*.

Die Bezeichnung des Filters besagt, dass es Verfälschungen der Abtastwerte durch höherfrequente Signalanteile verhindert oder zumindest enorm reduziert. In automatisierten, rechnergestützten Mess- und Datenerfassungseinrichtungen werden die Anti-Aliasing-Filter automatisch an die vom Benutzer eingestellte Abtastfrequenz angepasst.

14.1.4 Verschaltung zu hybriden Systemen

Mit Abtaster und Halteglied kann eine zeitkontinuierliche Regelstrecke mit einem zeitdiskreten Regler zu einem Gesamtregelkreis wie in Bild 14-1 verbunden werden. Dieser Regelkreis ist dann aber weder ein rein zeitdiskretes noch ein rein zeitkontinuierliches System, sondern besteht aus zeitkontinuierlichen und zeitdiskreten Teilsystemen sowie den Verbindungen dazwischen in Form von Abtaster und Halteglied.

Um ein solches hybrides System mit zeitkontinuierlichen und zeitdiskreten Bestandteilen auf Stabilität und andere Eigenschaften analysieren zu können, muss eine einheitliche Beschreibungsform für beide Teilsysteme gefunden werden.

Dabei gibt es grundsätzlich zwei Möglichkeiten: Die erste Möglichkeit besteht darin, die Elemente des Regelkreises wie in Bild 14-5 zusammenzufassen. Dieser Ansatz fasst das Gesamtsystem als kontinuierlich arbeitendes System auf. Für die aus Abtaster, zeitdiskretem Regler und Halteglied bestehende Regeleinrichtung muss dann ein kontinuierliches Ersatzsystem gefunden werden, welches das Übertragungsverhalten von $e(t)$ nach $u(t)$ beschreibt.

Quasikontinuierliche Beschreibung

Wird ein Regelkreis mit zeitdiskreten und zeitkontinuierlichen Teilsystemen als zeitkontinuierliches Gesamtsystem aufgefasst, welches mit zeitkontinuierlichen Methoden untersucht wird, so spricht man von einer *quasikontinuierlichen* Beschreibung.

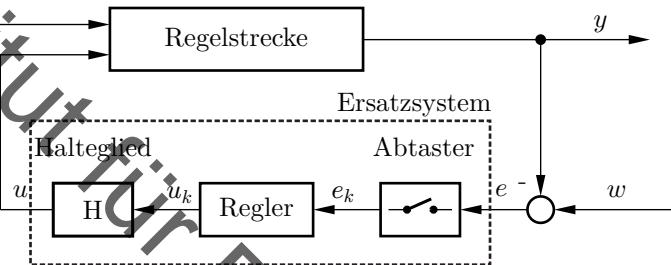


Bild 14-5: Abtastregelung in quasikontinuierlicher Beschreibung

Die zweite Möglichkeit ergibt sich, wenn man abweichend den Regelkreises wie in Bild 14-6 zusammenfasst, wodurch ein zeitdiskretes Gesamtsystem entsteht. In diesem Fall muss die Reihenschaltung aus Halteglied, zeitkontinuierlicher Regelstrecke und Abtaster in ein zeitdiskretes Ersatzsystem überführt werden, welches das Übertragungsverhalten von u_k nach e_k beschreibt.

Zeitdiskrete Beschreibung

Wird ein Regelkreis mit zeitdiskreten und zeitkontinuierlichen Teilsystemen als zeitdiskretes Gesamtsystem aufgefasst, welches mit zeitdiskreten Methoden untersucht wird, so spricht man von einer *zeitdiskreten* Beschreibung.

Beide Ansätze werden in den folgenden Abschnitten vorgestellt. Die quasikontinuierliche Betrachtung benötigt dabei kaum neue mathematische Werkzeuge, während der zeitdiskrete Entwurf stark vom zeitdiskreten Bildbereich profitiert und den H-Inhalten zugeordnet ist. In beiden Fällen ist es aber notwendig, eine passende zeitdiskrete Beschreibungsform für den Regler zu finden sowie zeitdiskrete Beschreibungen in kontinuierliche umrechnen zu können und umgekehrt.

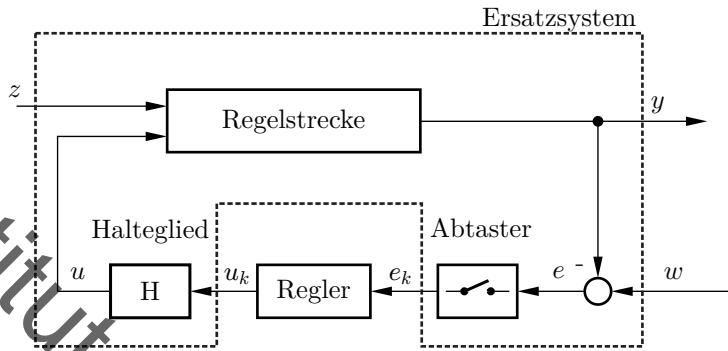


Bild 14-6: Abtastregelung in zeitdiskreter Zeitdiskreter

14.2 Einführung in Differenzengleichungen

Eine digitale Abtastregelung nach Bild 14-1 besteht prototypisch aus einer zeitkontinuierlichen Regelstrecke, einem zeitdiskreten Regler sowie Abtaster und Halteglied als verbindenden Elementen. Während die Systembeschreibung für den zeitkontinuierlichen Teil des Regelkreises aus den bisherigen Ausführungen klar ist, muss für den zeitdiskreten Teil diese noch eingeführt und diskutiert werden.

Die Beschreibungsform für zeitdiskrete dynamische Systeme ist die Differenzengleichung. Diese beinhaltet – im Gegensatz zur Differentialgleichung – keine Ableitungen, sondern verschobene Zeitargumente. In expliziter Darstellung lässt sich eine Differenzengleichung mit einer Eingangsfolge u_k und einer Ausgangsfolge y_k wie folgt schreiben:

$$y_{k+n} = f(y_{k+n-1}, \dots, y_{k+1}, y_k, u_{k+m}, \dots, u_k, k) \quad . \quad (14.5)$$

Die Differenzengleichung besitzt offensichtliche Parallelen und Gemeinsamkeiten zur Differentialgleichung, wie sie in Gl.(2.1) eingeführt worden ist. Daher ergeben sich diverse Definitionen und Ausführungen nahezu wortgleich zum Fall zeitkontinuierlicher Systeme. Diese werden im Folgenden ohne weitere Begründungen angegeben, damit die Unterschiede zwischen den Beschreibungsformen mehr Raum erhalten.

Die Systemordnung der Differenzengleichung in Gl.(14.5) ist n und die Anfangsbedingung umfasst alle Werte der rechten Gleichungsseite für $k = 0$.

Alle Signale (Folgen) werden für $k < 0$ zu null angenommen und können sich in $k = 0$ erstmalig ändern. Eine Unterscheidung zwischen $t = -0$ und $t = +0$ wie im Zeitkontinuierlichen ist nicht notwendig. Allerdings bedient man sich oft einer Indexverschiebung in Form von

$$y_{k+1} = f(y_k, y_{k-1}, \dots, y_{k-n+1}, u_{k+m-n+1}, \dots, u_{k-n+1}, k) , \quad (14.6)$$

da diese Darstellung eher der Arbeitsweise eines Algorithmus entspricht, welcher auf Basis vergangener und aktueller Eingangswerte den neuen Ausgangswert y_{k+1} berechnet.

Die Klassifikation der Differenzengleichungen bezüglich Zeitvarianz und Linearität erfolgt wortgleich wie für die Differentialgleichungen. Folglich besitzt ein zeitdiskretes LTI-System die Form

$$y_{k+n} + a_1 y_{k+n-1} + \dots + a_n y_k = b_0 u_k + b_1 u_{k+1} + \dots + b_m u_{k+m} \quad (14.7)$$

mit u_k als Folge der Eingangswerte und y_k als Folge der Ausgangswerte. In Summenschreibweise lässt sich dies auch als

$$\sum_{i=0}^n a_i y_{k+n-i} = \sum_{i=0}^m b_i u_{k+m-i} \quad (14.8)$$

mit $a_0 = 1$ zusammenfassen.

Der relative Grad $r = n-m$ hat bei zeitdiskreten Systemen eine Bedeutung, die über die im Zeitkontinuierlichen bekannte hinausgeht. Im Gegensatz zu zeitkontinuierlichen Systemen, wo die Totzeit einer separaten Behandlung durch ein verschobenes Eingangsargument bedurfte, kann die Totzeit im Zeitdiskreten nämlich durch

$$y(t) = u(t - T_t) \Rightarrow y_k = u_{k-d} \text{ mit } d = \frac{T_t}{T} \quad (14.9)$$

dargestellt werden, sofern die Totzeit ein Vielfaches der Abtastzeit ist. Auf diesen Fall wird man sich üblicherweise beschränken.

Folglich sorgt eine Totzeit für eine Verschiebung aller Eingangsfolgen um d Zeitschritte in die Vergangenheit. Besitzt eine Differenzengleichung einen relativen Grad von r , so ist der erste von Null verschiedene Koeffizient der Differenzengleichung in Gl.(14.7) genau b_{n-r} – folglich ist die Eingangsfolge um r Zeitschritte in die Vergangenheit verschoben.

Relativer Grad zeitdiskreter Systeme

Bei zeitdiskreten Systemen entspricht der relative Grad der zeitdiskreten Totzeit (in Abtastschritten) des Systems.

Ein zeitdiskretes System ist für $r \geq 0$ kausal, was bedeutet, dass die Totzeit im System nicht negativ sein darf und gemäß Gl.(14.7) keine zukünftigen (und damit unbekannten) Eingangswerte zur Berechnung der Ausgangswerte herausgezogen werden.

Auch Differenzengleichungen lassen sich im Wirkungsplan oder im Zustandsraum darstellen. Im Zustandsraum führt man beispielsweise die Größen y_k bis y_{k+n-1} (oder andere Größen) als Zustandsvektor \mathbf{x}_k ein, der n Einträge besitzt. Die Zustandsraumdarstellung wird dann

$$\mathbf{x}_{k+1} = \mathbf{A}_D \mathbf{x}_k + \mathbf{B}_D \mathbf{u}_k , \quad \mathbf{y}_k = \mathbf{C}_D \mathbf{x}_k + \mathbf{D}_D \mathbf{u}_k \quad (14.10)$$

für den linearen Mehrgrößenfall, wobei der Index D verwendet wird, um Verwechslungen mit der zeitkontinuierlichen Darstellung vorzubeugen.

Bei der Darstellung im Wirkungsplan zeichnet man die zeitdiskreten Übergangsfolgen der linearen Regelmässigkeitsfolgen.

Übergangsfolge und Gewichtsfolge

Die Übergangsfolge und Gewichtsfolge sind als Antworten auf die Einheitssprungfolge und Einheitsimpulsfolge

$$1[k] = \begin{cases} 1 & \text{für } k \geq 0 \\ 0 & \text{für } k < 0 \end{cases} , \quad \delta[k] = \begin{cases} 1 & \text{für } k = 0 \\ 0 & \text{sonst} \end{cases} \quad (14.11)$$

definiert.

Hier ist anzumerken, dass der Einheitsimpuls einen endlichen Wert für $k = 0$ annimmt. Dies ist notwendig, damit – analog dazu wie die Integration des Impulses $\delta(t)$ den Sprung $1(t)$ ergibt – die fortlaufende Summe (als Pendant zur Integration) des Impulses $\delta[k]$ den Sprung $1[k]$ ergibt.

14.3 Autonome zeitdiskrete Systeme

Die Definition einer Ruhelage eines zeitdiskreten Systems erfolgt mit analogen Ideen zum zeitkontinuierlichen Fall. Ein System ist genau dann in

Ruhe, wenn gilt:

$$\mathbf{x}_{k+1} = \mathbf{x}_k \quad . \quad (14.12)$$

Hier beschreibt Gl.(14.12) die Situation, dass sich im Laufe eines Zeitschrittes das System nicht verändert hat. Da die Berechnungsvorschrift Gl.(14.5) im zeitinvarianten Fall eindeutig ist, ergibt sich auch für alle zukünftige Zeitschritte ein gleichbleibender Zustandsvektor.

Alle Ausführungen zur Linearisierung gelten wortgleich auch für die Linearisierung einer nichtlinearen Differenzengleichung. Auch die Definitionen in Gl.(3.11) und Gl.(3.12) für asymptotische Stabilität können wortgleich (mit einer diskreten Zeit $k \in \mathbb{Z}$) übernommen werden. Allerdings unterscheiden sich die mathematischen Kriterien zum Überprüfen der Stabilität eines linearen zeitdiskreten Systems.

Zur Herleitung einer Stabilitätsbedingung wird die einfache autonome Differenzengleichung erster Ordnung

$$x_{k+1} = \lambda x_k \quad , \quad x_0 = {}_0x \quad (14.13)$$

betrachtet. Da Gl.(14.13) eine rekursive Rechenvorschrift ist, kann man ausrechnen, dass

$$x_0 = {}_0x \quad , \quad x_1 = \lambda x_0 = \lambda {}_0x \quad , \quad \dots \quad , \quad x_n = \lambda^n {}_0x \quad (14.14)$$

gilt.

Die Folge x_n konvergiert genau dann gegen null, wenn die Bedingung $|\lambda| < 1$ erfüllt ist. Dieses Resultat kann auf den allgemeinen Fall eines zeitdiskreten autonomen LTI-Systems n -ter Ordnung übertragen werden. Hierzu setzt man für die Differenzengleichung

$$a_n y_{k+n} + \dots + a_1 y_{k+1} + a_0 y_k = 0 \quad (14.15)$$

den Lösungskandidaten $C\lambda^k$ an. Eingesetzt erhält man

$$a_n y_{k+n} + \dots + a_1 y_{k+1} + a_0 y_k = a_n C \cdot \lambda^{k+n} + \dots + a_1 C \cdot \lambda^{k+1} + a_0 C \cdot \lambda^k = 0 \quad . \quad (14.16)$$

Ausklammern des Lösungskandidaten führt auf

$$C \cdot \lambda^k \cdot (a_n \lambda^n + \dots + a_2 \lambda^2 + a_1 \lambda + a_0) = 0 \quad . \quad (14.17)$$

Da der erste Term des Produkts für $C \neq 0$ nicht verschwindet, muss λ so gewählt werden, dass der Ausdruck innerhalb der Klammer verschwindet und man erhält das identische charakteristische Polynom wie im zeitkontinuierlichen Fall. Allerdings müssen dessen Wurzeln nun einen Betrag kleiner eins aufweisen.

Der Fall mehrfacher Wurzeln des charakteristischen Polynoms ergibt sich wortgleich zu den Ausführungen in Abschnitt 3.4. Die Wurzeln des charakteristischen Polynoms sind zudem – analog zum Zeitkontinuierlichen – identisch mit den Eigenwerten der Systemmatrix \mathbf{A}_D der zeitdiskreten Zustandsraumdarstellung.

Stabilität zeitdiskreter LTI-Systeme

Für ein autonomes zeitdiskretes LTI-System mit einer Ruhelage in $y = 0$ hängt die Stabilität der Ruhelage mit den Wurzeln λ_i des zugehörigen charakteristischen Polynoms wie folgt zusammen:

- Gilt für den Betrag aller Wurzeln $|\lambda_i| < 1$, so ist die Ruhelage stabil.
- Gilt für den Betrag mindestens einer Wurzel $|\lambda_i| > 1$, so ist die Ruhelage instabil.
- Gilt für den Betrag einer mehrfachen Wurzel $|\lambda_i| = 1$, so ist die Ruhelage instabil.
- Ansonsten (d. h. einfache Wurzeln mit $|\lambda_i| = 1$, alle anderen Wurzeln mit $|\lambda_j| < 1$) ist die Ruhelage grenzstabil.

Auch das Linearisierungstheorem gilt für zeitdiskrete Systeme in dem Sinne, dass sich die Stabilität des linearisierten Systems nach den gleichen Regeln auf die Stabilität der Ruhelage des nichtlinearen Systems überträgt.

Für stabile Systeme lässt sich auch der statische Endwert K der Übergangsfolge leicht ermitteln. Hierzu bedenkt man, dass für die Eingangsfolge $u_k = 1[k]$ sich für $k \rightarrow \infty$ die Ausgangsgröße $y_k = K$ einstellen muss. Einsetzen in Gl.(14.7) ergibt direkt

$$\sum_{i=0}^n a_i \cdot K = \sum_{i=0}^m b_i \cdot 1 \quad \Rightarrow \quad K = \frac{\sum_{i=0}^m b_i}{\sum_{i=0}^n a_i} . \quad (14.18)$$

Die statische Verstärkung ist also das Verhältnis der Summen der Koeffizienten von linker und rechter Gleichungsseite.

Alles in allem lässt sich eine Differenzengleichung bezüglich der aufgeführten Eigenschaften und insbesondere der Stabilität mit fast identischen Mitteln wie eine Differentialgleichung untersuchen. Unterschiede finden sich in der mathematischen Stabilitätsbedingung ($\text{Re}(\lambda) < 0$ vs. $|\lambda| < 1$) und in der Darstellung der Totzeit. Aus weiterem folgt insbesondere, dass Regelalgorithmen Differenzengleichungen mit relativem Grad $r = 0$ sein sollten, da ansonsten eine zusätzliche Totzeit in das System eingebracht wird.

14.4 Umrechnen von Differenzen- und Differentialgleichungen

14.4.1 Rückwärtsdifferenzen

Der Regelkreis mit einer Abtastregelung enthält Elemente (wie den Regler), die zeitdiskret arbeiten, und Teilsysteme (wie die Regelstrecke), die zeitkontinuierlich beschrieben werden. Für eine einheitliche Beschreibung des Systems wahlweise als kontinuierliches oder zeitdiskretes System müssen die Differenzengleichung in Differentialgleichungen umgerechnet werden und umgekehrt.

Aus der Numerik – insbesondere den Verfahren der numerischen Simulation – sind zahlreiche Verfahren bekannt, eine Differentialgleichung in eine Differenzengleichung umzuformen und damit für numerische Lösungsalgorithmen zugänglich zu machen. Nur die einfachste Form der Rückwärtsdifferenzen wird hier behandelt, welche mit dem expliziten Euler²-Verfahren identisch ist.

Rückwärtsdifferenzen

Wird eine Differentialgleichung im Zustandsraum angenähert durch

$$\dot{x} = f(x) \quad \Rightarrow \quad \frac{x_k - x_{k-1}}{T} \approx f(x_k) \quad (14.19)$$

so spricht man von *Rückwärtsdifferenzen*. Hierbei bezeichnet T die Abtastzeit.

²Leonhard Euler (1707-1783), Schweizer Mathematiker [26]

Der Vorteil der Rückwärtsdifferenz ist, dass diese sehr einfach ist und keine Werte aus der Zukunft erfordert und somit auf kausale Systeme führt. Anschaulich wird die Differentiation durch einen Differenzenquotienten ersetzt, dessen Stützstelle zeitlich in die Vergangenheit versetzt wurde.

Angewendet auf das Beispiel eines Differenzierers ergibt sich

$$y(t) = K_D \cdot \dot{u}(t) \Rightarrow y_k = \frac{K_D}{T} (u_k - u_{k-1}) \quad (14.20)$$

und für einen Integrierer

$$y(t) = K_I \int_0^t u(\tau) d\tau \Rightarrow y_k = y_{k-1} + K_I \cdot T \cdot u_k . \quad (14.21)$$

Die Differentialgleichung eines Verzögerungsgliedes 1. Ordnung

$$T_1 \dot{y} + y = K \cdot u \quad (14.22)$$

lässt sich umformen in die Differenzengleichung

$$\frac{T_1}{T} (y_k - y_{k-1}) + y_k = K \cdot u_k , \quad (14.23)$$

welche durch Umordnen

$$\left(\frac{T_1}{T} + 1 \right) y_{k+1} + \left(-\frac{T_1}{T} \right) y_k = K \cdot u_{k+1} \quad (14.24)$$

in die Form Gl.(14.7) gebracht werden kann.

Mit diesem Verfahren kann jede Differentialgleichung in eine zugehörige Differenzengleichung umgeformt werden. Allerdings ist hier Vorsicht geboten, da die Eigenschaften von Differential- und Differenzengleichung nicht identisch sein müssen. Um dies herauszuarbeiten, wird das Beispiel des Systems 1. Ordnung aufgegriffen. Das zeitkontinuierliche System ist für $\text{Re}(T_1) > 0$ stabil, da $\text{Re}(\lambda) = \text{Re}(-T_1) < 0$ erfüllt ist. Für das zeitdiskrete System muss für Stabilität $|\lambda| < 1$ erfüllt sein. Mit

$$\lambda = \frac{T_1}{T_1 + T} \quad (14.25)$$

gewinnt man

$$\left| \frac{T_1}{T_1 + T} \right| < 1 \Rightarrow \left| \frac{T_1 + T}{T_1} \right| > 1 \Rightarrow \left| 1 + \frac{T}{T_1} \right| > 1 \quad . \quad (14.26)$$

Diese Gleichung stellt eine Bedingung an das Verhältnis von Abtastzeit T und Zeitkonstante T_1 , die aus dem Zeitkontinuierlichen so nicht bekannt ist. So wird beispielsweise auch für $T_1 < 0$ das eigentlich instabile zeitkontinuierliche System 1. Ordnung für die Abtastzeit $T = (2 + \epsilon) \cdot |T_1|$ auf ein stabiles zeitdiskretes System abgebildet.

Diese Stabilitätsveränderung durch Zeitdiskretisierung ist fatal. Im gerade diskutierten Fall erscheint ein eigentlich instabiles zeitkontinuierliches System nach Rückwärtsdifferenzen als stabil. Auch der umgekehrte Fall von stabilen zeitkontinuierlichen Systemen, die nach Rückwärtsdifferenzen instabil erscheinen, ist möglich und nicht wünschenswert. Am gegebenen Beispiel kann man dabei erahnen, dass diese Effekte vorwiegend bei (zu) großen Abtastzeiten zu erwarten sind.

Rückwärtsdifferenzen als Näherung

Bei Rückwärtsdifferenzen handelt es sich nur um eine Näherung des eigentlichen zeitkontinuierlichen Systemverhaltens, bei der immer Fehler gemacht werden, die mit größerer Abtastzeit wachsen. Die Näherung kann dabei beliebig schlecht werden, sodass auch die Stabilitätseigenschaften des Systems sich durch die Zeitdiskretisierung verändern.

Diese Ausführungen gelten nicht nur für die Rückwärtsdifferenzen, sondern für alle Verfahren, die auf der Annäherung der Ableitung durch einen Differenzenquotienten basieren.

Für die quasikontinuierliche Beschreibung wird der „inverse“ Fall zum bisher diskutierten Verfahren benötigt: Es muss aus einer Differenzengleichung eine Differentialgleichung abgeleitet werden. Die Einschränkungen bezüglich der Gültigkeit der Näherung gelten weiterhin. Bei der Rechnung wird man dabei Terme der Struktur $y_k - y_{k-1}$ durch $T\dot{y}$ ersetzen. So ergibt sich beispielsweise für die Differenzengleichung

$$y_k - y_{k-1} = 2u_k - u_{k-1} \quad (14.27)$$

durch passendes Ergänzen die Differentialgleichung

$$y_k - y_{k-1} = 2u_k - u_{k-1} = u_k + (u_k - u_{k-1}) \Rightarrow \dot{y} = \frac{1}{T}u + \dot{u} \quad (14.28)$$

und damit ein zeitkontinuierlicher PI-Regler.

Leider ist diese Umrechnung nicht eindeutig. Dies liegt daran, dass die Totzeit im Zeitdiskreten einer Index-Verschiebung entspricht. Folglich könnte die Differentialgleichung in dem Beispiel ebenso

$$y_k - y_{k-1} = 2u_k - u_{k-1} \Rightarrow \dot{y}(t) = \frac{1}{T} (2u(t) - u(t-T)) \quad (14.29)$$

lauten. Durch entsprechende Ergänzungen sind hier unendlich viele Kombinationen denkbar.

Praktisch ist diese Uneindeutigkeit aber nicht von Bedeutung. Als Regler werden nämlich in der überwiegenden Mehrzahl der Anwendungsfälle Algorithmen eingesetzt, die das Verhalten der in Kapitel 7.2 eingeführten kontinuierlich wirkenden Regler nachbilden. Außerdem ist die Verwendung von Totzeit-Gliedern als Bestandteil des Reglers wegen der schlechten Stabilitätseigenschaften in nahezu allen Fällen unsinnig, weswegen dieser Fall praktisch ausgeschlossen werden kann.

Allgemein erhält man für einen PID-Regler

$$y = K_R \left(u + \frac{1}{T_n} \int_0^t u \, d\tau + T_v \dot{u} \right) \quad (14.30)$$

durch Ableiten nach der Zeit, Rückwärtsdifferenzen und sortieren

$$y_k = y_{k-1} + K_R \left[\left(1 + \frac{T}{T_n} + \frac{T_v}{T} \right) u_k - \left(1 + 2 \frac{T_v}{T} \right) u_{k-1} + \frac{T_v}{T} u_{k-2} \right] \quad (14.31)$$

als rekursive Rechenvorschrift zur Bestimmung der Werte der Stellgröße eines zeitdiskreten PID-Reglers. Man erkennt, dass neben dem aktuellen Wert der Regelabweichung u_k noch die zu den beiden vorhergehenden Abtastzeitpunkten gehörenden Werte u_{k-1}, u_{k-2} und der zum vorangehenden Zeitpunkt ermittelte Stellgrößenwert y_{k-1} bereithalten werden müssen.

Die Rückführung des Wertes der Stellgröße y_{k-1} kann dabei ggf. vermieden werden, indem nur die Änderung der Stellgröße $\Delta y_k = y_k - y_{k-1}$ einem integrierenden Stellglied zugeführt wird.

14.4.2 Analytische Lösung

Bei der Verwendung der Rückwärtsdifferenzen wird ein Fehler gemacht, der in vielen Fällen nur für kleine Abtastzeiten tolerabel sein wird. Sollen zeitkontinuierliche und zeitdiskrete Systeme auch für große Abtastzeiten ineinander umgerechnet werden, führen Rückwärtsdifferenzen folglich nicht ans Ziel.

Für große Abtastzeiten ist es sinnvoll, unter Annahmen an den Verlauf der Eingangsgrößen $\mathbf{u}(t)$ die Umrechnung zwischen den Zeitbereichen analytisch vorzunehmen. Hierfür bietet sich die Darstellung im Zustandsraum an.

Ausgehend von der Lösung der kontinuierlichen Zustandsdifferentialgleichung mittels Transitionsmatrix (siehe Abschnitt 3.6)

$$\mathbf{x}(t) = e^{\mathbf{A}(t-t_0)} \mathbf{x}(t_0) + \int_{t_0}^t e^{\mathbf{A}(t-\tau)} \mathbf{B} \mathbf{u}(\tau) d\tau \quad t > t_0 \quad (14.32)$$

lässt sich der Zustandsvektor \mathbf{x} zu den Abtastzeitpunkten berechnen. Dazu werden die Integrationsgrenzen zu $t = (k+1)T$ und $t_0 = kT$ angenommen:

$$\mathbf{x}((k+1)T) = e^{\mathbf{A}T} \mathbf{x}(kT) + \int_{kT}^{(k+1)T} e^{\mathbf{A}((k+1)T-\tau)} \mathbf{B} \mathbf{u}(\tau) d\tau . \quad (14.33)$$

Mit der Substitution $\theta = \tau - kT$ vereinfacht sich der Integrand ausdruck, und es ergibt sich

$$\mathbf{x}((k+1)T) = e^{\mathbf{A}T} \mathbf{x}(kT) + e^{\mathbf{A}T} \int_0^T e^{-\mathbf{A}\theta} \mathbf{B} \mathbf{u}(kT + \theta) d\theta . \quad (14.34)$$

Um das Integral berechnen zu können, muss der zeitliche Verlauf von $\mathbf{u}(t)$ bekannt sein. Es liegt nahe, $\mathbf{u}(t)$ während eines Abtastschrittes als konstant anzusetzen. Diese Annahme entspricht einem Halteglied 0. Ordnung im Wirkungsweg von \mathbf{u} und lässt sich formulieren als

$$\mathbf{u}(kT + \theta) = \mathbf{u}(kT) \quad \text{für} \quad 0 \leq \theta < T . \quad (14.35)$$

Damit kann das Integral berechnet werden, und es folgt aus Gl.(14.34)

$$\mathbf{x}((k+1)T) = e^{\mathbf{A}T} \mathbf{x}(kT) + e^{\mathbf{A}T} (\mathbf{I} - e^{-\mathbf{A}T}) \mathbf{A}^{-1} \mathbf{B} \mathbf{u}(kT). \quad (14.36)$$

Dies lässt sich auch in die zeitdiskrete Zustandsraumdarstellung überführen:

Zeitdiskreter Zustandsraum zeitkontinuierlicher Modelle

Gegeben ist ein zeitkontinuierliches Modell mit den Matrizen \mathbf{A} , \mathbf{B} , \mathbf{C} und \mathbf{D} . Dann wird für die Abtastzeit T das äquivalente zeitdiskrete Zustandsraummodell durch die Matrizen

$$\begin{aligned} \mathbf{A}_D &= e^{\mathbf{A}T} \\ \mathbf{C}_D &= \mathbf{C} \\ \mathbf{D}_D &= \mathbf{D} . \end{aligned} \quad (14.37)$$

beschrieben.

Unter der Annahme, das \mathbf{u} zwischen den Abtastwerten konstant ist, ergibt sich zudem die Eingangsmatrix

$$\mathbf{B}_D = (e^{\mathbf{A}T} - \mathbf{I}) \mathbf{A}^{-1} \mathbf{B} . \quad (14.38)$$

Die so hergeleitete zeitdiskrete Darstellung eines eigentlich zeitkontinuierlichen Systems ist exakt und ohne grundsätzlichen numerischen Fehler. Das bedeutet, dass die Lösung des zeitkontinuierlichen Systems und des zeitdiskreten Systems an den Abtastzeitpunkten durch die exakt gleichen Punkte verläuft: $\mathbf{x}(t_k) = \mathbf{x}_k$. Auch bleiben die Stabilitätseigenschaften erhalten, da

$$\operatorname{Re}(\lambda(\mathbf{A})) < 0 \Leftrightarrow |\lambda(e^{\mathbf{A}T})| < 1 \quad (14.39)$$

gilt.

Die Berechnung für \mathbf{B}_D in Gl.(14.38) existiert auch für nicht invertierbare \mathbf{A} , was sich über die Reihenentwicklung der Transitionsmatrix zeigen lässt. Einzige Voraussetzung ist, dass die Eingangsgröße über einen Zeitschritt konstant ist. Ist diese Voraussetzung nicht erfüllt, so ist \mathbf{B}_D falsch und die Zeitverläufe unterscheiden sich.

Unter bestimmten Voraussetzungen lässt sich die gewonnene Berechnungsvorschrift auch umkehren und aus einer zeitdiskreten Darstellung eine zeitkontinuierliche gewinnen. Hierbei zeigt sich, dass diese Voraussetzungen – unabhängig von den Annahmen für u – identisch zum Shannon-Theorem sind.

Transitionsmatrix $e^{\mathbf{A}T}$ und das Shannon-Theorem

Die Zuordnung $\mathbf{A}_D = e^{\mathbf{A}T}$ ist aufgrund der Periodizität der komplexen e -Funktion nicht eindeutig umkehrbar. So würde nämlich wegen $e^{aT+2\pi j} = e^{aT}$ ein im Komplexen um den Faktor $2\pi/T$ verschobenes \mathbf{A} auf das identische \mathbf{A}_D führen. Folglich muss für einen eindeutigen Zusammenhang der Imaginärteil (und das sind genau die Eigenkreisfrequenzen ω_D) auf den Wertebereich $-\pi/T < \omega_D < \pi/T$ beschränkt werden. Mit $\omega = 2\pi/T$ entspricht dies genau dem Abtasttheorem nach Shannon.

Insgesamt gestalten sich die analytischen Verfahren als rechenaufwändiger. In vielen Fällen lohnt sich aber die Investition in eine analytische Lösung, da so bei der Umwandlung kein Fehler gemacht wird. Dies gilt unabhängig von der verwendeten Abtastzeit, weswegen sich das Verfahren insbesondere für solche Fälle eignet, wo die Abtastzeit vergleichsweise groß gewählt werden muss.

Ein Vorteil der Rückwärtsdifferenzen ist allerdings, dass sich diese auch für nichtlineare Differenzengleichungen einsetzen lassen, was bei der analytischen Lösung nur in Sonderfällen möglich ist.

14.5 Quasikontinuierlicher Reglerentwurf

Zur Untersuchung des Regelkreises mit zeitkontinuierlicher Regelstrecke und zeitdiskretem Regler besteht eine Möglichkeit darin, dass Gesamtsystem als kontinuierlich arbeitendes System aufzufassen. Bei dieser quasikontinuierlichen Betrachtung geht man – wie bereits erwähnt – von einer Anordnung nach Bild 14-5 aus und definiert ein kontinuierliches Ersatzsystem für Regler, Abtaster und Halteglied.

Um dieses Ersatzsystem zu gewinnen, wird in einem ersten Schritt die Reihenfolge von Regler und Halteglied wie in Bild 14-7 vertauscht. Dies entspricht zwar nicht der technischen Wirklichkeit, ist in diesem Fall aber zweckmäßig, weil die so entstehende Reihenschaltung von Abtaster und Hal-

teglied als ein einziges lineares Übertragungsglied aufgefasst werden kann, welches unabhängig vom verwendeten Regler ist. Somit kann das Finden eines Ersatzsystems in einen gleichbleibenden Anteil (Abtaster und Halteglied) und einen von Anwendungsfall zu Anwendungsfall variablen Anteil (Regler) unterteilt werden. Diese Änderung der Reihenfolge ist dabei zulässig, weil Regler und Halteglied lineare Übertragungsglieder sind, die miteinander kommutieren.

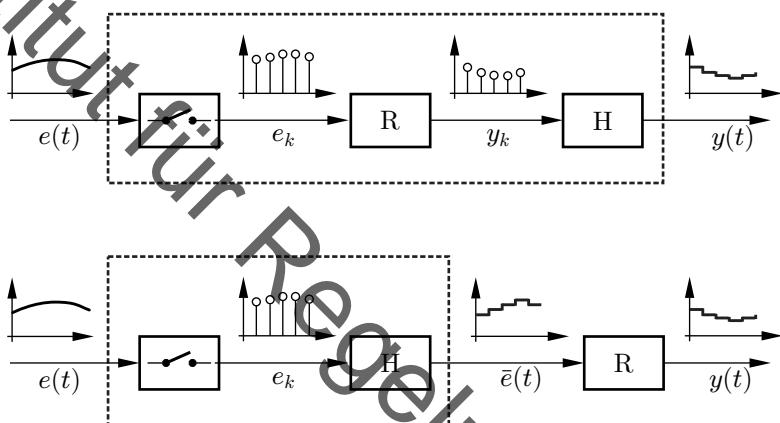


Bild 14-7: Regeleinrichtung zur Abtastregelung

Sucht man ein Ersatzmodell für das Abtaster-Halteglied (d. h. die Reihenschaltung von Abtaster und Halteglied), so geschieht dies am einfachsten dadurch, dass man das Verhalten im Zeitbereich betrachtet. Hierzu zeigt Bild 14-8 das Ausgangssignal $y(t)$, welches bei einem zufällig gewählten Eingangssignal $u(t)$ nach Abtastung und Halten entsteht. Man sieht, dass hierdurch näherungsweise das Ausgangssignal im Mittelwert dem Eingangssignal um die Zeit $T/2$ hinterherläuft. Folglich wirkt das Abtaster-Halteglied ungefähr wie ein Totzeit-Glied mit $T_t = T/2$. Dies hat zur Folge, dass die Stabilitätseigenschaften von Regelkreisen durch die digitale Umsetzung des Reglers schlechter werden.

Diese graphische Erläuterung lässt sich auch über eine Rechnung im Frequenzbereich untermauern. Mit Überlegungen, die hier nicht wiedergegeben werden sollen, gewinnt man zur angenäherten Beschreibung des aus Abtas-

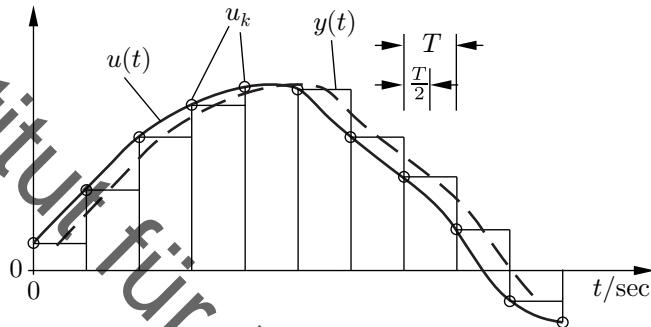
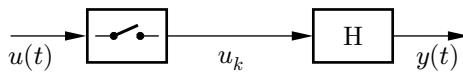


Bild 14-8: Abtaster mit Halteglied und Größenverläufe

ter und Halteglied bestehenden Teilsystems den Frequenzgang

$$G_{AH}(j\omega) = \frac{1 - e^{-j\omega T}}{j\omega T} = \frac{\sin \frac{\omega T}{2}}{\frac{\omega T}{2}} \cdot e^{-j\frac{\omega T}{2}} \approx e^{-j\frac{\omega T}{2}} . \quad (14.40)$$

Für Frequenzen bis etwa $\omega T = 1$ weicht der Betrag des Frequenzganges um weniger als 5% von eins ab, sodass die eingeführte Näherung durch ein Totzeitglied mit einer Totzeit von $T/2$ für niedrige Frequenzen und/oder kleine Abtastintervalle ausreicht.

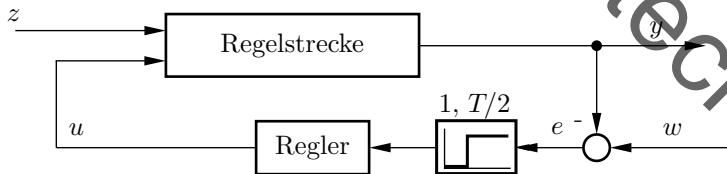


Bild 14-9: Ersatzregelkreis

Mit den im vorangegangenen Abschnitt erläuterten Methoden kann für einen zeitdiskret arbeitenden Regelkreis ohne größere Umstände ein Reg-

ler entworfen werden. Hierzu macht man sich bewusst, dass sich in der quasikontinuierlichen Betrachtung der in Bild 14-9 gezeigte Ersatzregelkreis ergibt. Ausgangspunkt ist nun wahlweise ein zeitkontinuierlicher oder zeitdiskreter Regler. Im Falle eines zeitkontinuierlichen Reglers wird dieser nach den üblichen Verfahren für die zeitkontinuierliche Regelstrecke ausgelegt. Einziger Unterschied ist, dass diese um eine zusätzliche Totzeit von $T/2$ erweitert wird.

Ist die Abtastzeit bereits festgelegt, kann dies direkt in geänderte Anforderungen an eine Phasenreserve übersetzt werden. Daher bietet sich das Frequenzkennlinienverfahren in Abschnitt 11.1 besonders für diesen Reglerentwurf an. Ist die Abtastzeit noch nicht festgelegt, so kann aus der Reglerauslegung (und insbesondere der Phasenreserve, aber auch aus Gl.(14.4)) eine obere Schranke für die Abtastzeit abgeleitet werden.

Der entworfene Regler wird dann über Rückwärtsdifferenzen oder die analytische Lösung mit der festgelegten Abtastzeit in eine Differenzengleichung überführt. Diese Differenzengleichung sollte in einem letzten Schritt noch auf Plausibilität überprüft werden. So sollten die Stabilitätseigenschaften gleich bleiben. Ist das nicht der Fall, sollte die Abtastzeit verringert werden oder Rückwärtsdifferenzen durch die analytische Lösung ersetzt werden.

Bildet stattdessen ein zeitdiskreter Regler den Ausgangspunkt, so wird dieser in einem ersten Schritt mit Rückwärtsdifferenzen oder der analytischen Lösung und einer definierten Abtastzeit in ein zeitkontinuierliches Pendant überführt. Der restliche Entwurf erfolgt wie zuvor.

Das geschilderte Vorgehen lässt vermuten, dass diese Art des Reglerentwurfs nur für kleine Abtastzeiten ausreichend genau sein wird. In solchen Fällen wird die quasikontinuierliche Nachbildung so gut sein, dass zur Dimensionierung des Reglers die für kontinuierliche Regler eingeführten Verfahren angewandt werden können.

Die Aufgabe geht darin über, für den Ersatzregelkreis nach Bild 14-9 einen kontinuierlich wirkenden Regler zu spezifizieren, und diesen dann durch einen geeigneten zeitdiskret wirkenden Algorithmus zu verwirklichen. Hierzu muss das Abtastintervall T so klein in Relation zur Dynamik des Systems sein, dass der Abtastvorgang das Gesamtverhalten nicht wesentlich beeinflusst. Dies ist i. Allg. dann der Fall, wenn das Abtastintervall deutlich kleiner ist als die kleinsten Zeitkonstanten der Regelstrecke.

14.6 Zeitdiskreter Bildbereich

14.6.1 \mathcal{Z} -Transformation

Wenn das Abtastintervall in Relation zur Dynamik des Systems nicht mehr vernachlässigbar klein ist, wird man eine Beschreibung des Gesamtsystems als zeitdiskretes System anstreben. Der geschlossene Regelkreis liegt dann ebenfalls in einer zeitdiskreten Beschreibungsform vor und Regelziele und Reglerentwurf müssen im Zeitdiskreten erfolgen.

Für diesen Entwurf möchte man nicht ausschließlich im zeitdiskreten Zustandsraum arbeiten, sondern ähnliche Vorteile nutzen, wie der s -Bereich der Laplace-Transformation sie bietet. Hierzu gehören insbesondere das einfache Berechnen von Verschaltungen aber auch Darstellungsformen wie das Pol-Nullstellen-Diagramm und Übertragungsfunktionen. Da im Regelkreis eine Reihenschaltung von zeitdiskretem Regler und zeitdiskreter Regelstrecke vorliegt, sind diese Vorteile direkt spürbar.

Der Zugang in den zeitdiskreten Bildbereich erfolgt dabei über die \mathcal{Z} -Transformation.

\mathcal{Z} -Transformation

Gegeben sei eine Folge $f[k]$ mit $f[k] = 0$ für $k < 0$ und $k \in \mathbb{Z}$. Die \mathcal{Z} -Transformation ordnet dieser Folge $f[k]$ im Zeitbereich eine Funktion $F(z)$ mit der komplexen Variable $z \in \mathbb{C}$ im Bildbereich über die folgende Formel zu:

$$F(z) = \sum_{k=0}^{\infty} f[k] \cdot z^{-k} = \mathcal{Z}(f[k]) \quad . \quad (14.41)$$

Der Betrag der komplexen Variablen z muss dabei größer als der Konvergenzradius der Potenzreihe sein, damit die angegebene Summe endlich bleibt.

Als Beispiel soll die \mathcal{Z} -Transformierte des Einheitssprungs $1[k]$ ermittelt werden. Mit $f_k = 1$ für alle $k \geq 0$ ergibt sich

$$F(z) = \sum_{k=0}^{\infty} f_k \cdot z^{-k} = 1 + z^{-1} + z^{-2} + \dots \quad . \quad (14.42)$$

Gl.(14.42) beschreibt eine *geometrische Reihe*.

Geometrische Reihe

Die Reihe

$$S_N = \sum_{k=0}^N q^k \quad (14.43)$$

heißt Geometrische Reihe und besitzt für $|q| < 1$ den Grenzwert

$$\lim_{N \rightarrow \infty} S_N = \sum_{k=0}^{\infty} q^k = \frac{1}{1-q} \quad . \quad (14.44)$$

Setzt man $q = 1/z$, so gewinnt man für $|z| > 1$

$$\mathcal{Z}(1[k]) = F(z) = \frac{1}{1 - z^{-1}} = \frac{z}{z - 1} \quad . \quad (14.45)$$

Als weiteres Beispiel soll zu der Folge

$$f_k = \lambda^k \quad (14.46)$$

die \mathcal{Z} -Transformierte bestimmt werden. Man erhält

$$F(z) = \sum_{k=0}^{\infty} \lambda^k \cdot z^{-k} = \sum_{k=0}^{\infty} (\lambda \cdot z^{-1})^k \quad (14.47)$$

und damit wiederum eine geometrische Reihe mit $q = \lambda z^{-1}$. Die Summe ist

$$F(z) = \frac{1}{1 - \lambda \cdot z^{-1}} = \frac{z}{z - \lambda} \quad (14.48)$$

sofern $|q| < 1$ bzw. $|z| > |\lambda|$ gilt.

Auf eine Folge mit endlich vielen Werten ungleich null wie z. B.

$$f_k = (3, 5, -2, 6, 0, \dots) \quad (14.49)$$

lässt sich Gl.(14.41) ebenfalls anwenden und liefert als \mathcal{Z} -Transformierte

$$F(z) = 3 + 5z^{-1} - 2z^{-2} + 6z^{-3} \quad . \quad (14.50)$$

Bei der praktischen Anwendung der \mathcal{Z} -Transformation kann man ähnlich wie bei der Laplace-Transformation Korrespondenztafeln und Eigenschaften der Transformation nutzen. Die Eigenschaften können dabei direkt aus denen der Laplace-Transformation über den folgenden Zusammenhang abgeleitet werden: Es wird eine zeitkontinuierliche Funktion $f(t)$ betrachtet, die sich als Summe von verschobenen Impulsen darstellen lässt, die Gewichte f_k besitzen:

$$f(t) = \sum_{k=0}^{\infty} f_k \delta(t - kT) . \quad (14.51)$$

Unterzieht man diese Funktion der Laplace-Transformation, so erhält man

$$\begin{aligned} F(s) &= \int_{-0}^{\infty} \sum_{k=0}^{\infty} f_k \delta(t - kT) e^{-st} dt = \sum_{k=0}^{\infty} f_k \int_{-0}^{\infty} \delta(t - kT) e^{-st} dt \\ &= \sum_{k=0}^{\infty} f_k e^{-skT} = \sum_{k=0}^{\infty} f_k \underbrace{(e^{-sT})}_{}^k = \mathcal{Z}(f_k) . \end{aligned} \quad (14.52)$$

Die Laplace-Transformierte dieser Funktion verschobener Dirac-Impulse ist also genau die \mathcal{Z} -Transformierte der Folge f_k , die das Gewicht der Impulse angibt.

Die zeitkontinuierliche Repräsentation einer zeitdiskreten Folge über eine unendliche Reihe verschobener Impulse erscheint gewöhnungsbedürftig und wird für Rechnungen auch nicht herangezogen. Der Zusammenhang zeigt allerdings, dass sich Rechenregeln der Laplace-Transformation (wie bspw. Linearität) über Gl.(14.52) auf die \mathcal{Z} -Transformation übertragen.

Die Umkehrung gilt wegen der Periodizität der e -Funktion mit Einschränkung an den Imaginärteil von s (vgl. auch den Zusammenhang zwischen A und e^{AT}):

Zusammenhang zwischen s - und z -Ebene

Für den zeitkontinuierlichen Bildbereich mit $s \in \mathbb{C}$ und den zeitdiskreten

Bildbereich mit $z \in \mathbb{C}$ gilt der Zusammenhang

$$z = e^{sT} \quad . \quad (14.53)$$

Die Umkehrung ist eindeutig mit einer Beschränkung des Imaginärteils von s auf $-\frac{\pi}{T} < \text{Im}(s) < \frac{\pi}{T}$.

Man bezeichnet den Bereich $-\frac{\pi}{T} < \text{Im}(s) < \frac{\pi}{T}$ auch als *Hauptstreifen*. Dieser wird umso breiter, je kürzer das Abtastintervall T gewählt wird.

Mit diesem Zusammenhang und der Definition der \mathcal{Z} -Transformation können die Korrespondenztafeln ermittelt werden. Einen Auszug aus einer solchen Tafel zeigt Tab. 14-1. Darin sind jeder Zeitfunktion $f(t)$, die für negative Werte der Zeit verschwindet, jeweils zwei Transformierte gegenübergestellt: Einerseits die Laplace-Transformierte $F(s)$ und andererseits die \mathcal{Z} -Transformierten $F(z)$ der Wertefolge f_k , wobei f_k durch Abtasten von $f(t)$ mit dem Abtastintervall T entsteht.

Einige wichtige Korrespondenzen für die Operationen mit den \mathcal{Z} -Transformierten sind in Tab. 14-2 zusammengestellt. Daraus ist u. a. zu ersehen, dass die Faltung von Wertefolgen ähnlich wie im Fall der Laplace-Transformation in das Produkt der Transformierten übergeht. Diese Eigenschaft der \mathcal{Z} -Transformation macht sie zur Beschreibung zeitdiskreter Systeme besonders geeignet.

Auch gelten die gleichen Regeln für das Verschalten von Systemen. Dabei ist allerdings zu beachten, dass dies nur für die Verknüpfung zeitdiskreter, d. h. Wertefolgen verarbeitender, Übertragungssysteme untereinander gilt. Für die Verbindung zeitdiskreter und kontinuierlicher Systeme über Abtaster und Halteglieder gelten andere Regeln.

Die zu einer \mathcal{Z} -Transformierten gehörende Wertefolge kann man i. Allg. dadurch gewinnen, dass man die Transformierte z. B. durch Partialbruchzerlegung in Teilausdrücke überführt, die dann einzeln mit Hilfe von Korrespondenztafeln transformiert werden. Hierdurch kann mit einem analogen Vorgehen zur Laplace-Transformation die Lösung einer linearen Differenzgleichung auf eine allgemeine Eingangsfolge bestimmt werden. Man erhält dabei einen analytischen Ausdruck für die gesuchte Wertefolge.

Interessieren hingegen nur einige Werte am Anfang der Folge, die mit mäßigem Aufwand bestimmt werden sollen, so gelingt dies dadurch, dass man

$F(s)$	$f(t)$ für $t > 0$	$f[k] = f(kT)$ für $k > 0$	$F(z)$
$\frac{1}{s}$	$1(t)$	$1[k]$	$\frac{z}{z - 1}$
$\frac{1}{s^2}$	t	kT	$\frac{Tz}{(z - 1)^2}$
$\frac{n!}{s^{n+1}}$	$t^n, n \in \mathbb{N}$	$(kT)^n$	$T^n z \frac{z^{n-1} + \dots}{(z - 1)^{n+1}}$
$\frac{1}{s - \lambda}$ λ beliebig komplex	$e^{\lambda t}$	$e^{\lambda kT}$	$\frac{z}{z - e^{\lambda T}}$
$\frac{\omega}{s^2 + \omega^2}$	$\sin(\omega t)$	$\sin(\omega kT)$	$z \frac{\sin(\omega T)}{z^2 - 2z \cos(\omega T) + 1}$
$\frac{s}{s^2 + \omega^2}$	$\cos(\omega t)$	$\cos(\omega kT)$	$z \frac{z - \cos(\omega T)}{z^2 - 2z \cos(\omega T) + 1}$
$\frac{\omega}{(s - \alpha)^2 + \omega^2}$	$e^{\alpha t} \sin(\omega t)$	$e^{\alpha kT} \sin(\omega kT)$	$\frac{e^{\alpha T} \sin(\omega T)}{z^2 - 2ze^{\alpha T} \cos(\omega T) + e^{2\alpha T}}$
$\frac{s - \alpha}{(s - \alpha)^2 + \omega^2}$	$e^{\alpha t} \cos(\omega t)$	$e^{\alpha kT} \cos(\omega kT)$	$\frac{e^{\alpha T} \cos(\omega T)}{z^2 - 2ze^{\alpha T} \cos(\omega T) + e^{2\alpha T}}$
$\frac{(s - \alpha) \sin(\varphi) + \omega \cos(\varphi)}{(s - \alpha)^2 + \omega^2}$	$e^{\alpha t} \sin(\omega t + \varphi)$	$e^{\alpha kT} \sin(\omega kT + \varphi)$	$\frac{z \sin(\varphi) + e^{\alpha T} \sin(\omega T - \varphi)}{z^2 - 2ze^{\alpha T} \cos(\omega T) + e^{2\alpha T}}$
$\frac{1}{(s - \lambda)^2}$	$te^{\lambda t}$	$kTe^{\lambda kT}$	$\frac{Tze^{\lambda T}}{(z - e^{\lambda T})^2}$
$\frac{1}{s - \frac{1}{T} \ln a}, a \neq 0$	$a^{t/T}$	a^k	$\frac{z}{z - a}$

Tabelle 14-1: Korrespondenztafel für Laplace- und \mathcal{Z} -Transformierte

Operation	Zeitbereich	Bildbereich
Überlagerung	$a \cdot f_k + b \cdot g_k$	$a \cdot F(z) + b \cdot G(z)$
Ähnlichkeit	$a^k \cdot f_k$	$F\left(\frac{z}{a}\right)$
Verschiebung nach rechts (Verzögern)	f_{k-m}	$z^{-m} \left[F(z) + \sum_{i=1}^m f_{-i} z^i \right]$
Verschiebung nach links (Vorhersagen)	f_{k+m}	$z^m \left[F(z) + \sum_{i=1}^{m-1} f_i z^i \right]$
Differenzenbildung	$f_k - f_{k-1}$	$\frac{z-1}{z} F(z)$
Summation	$\sum_{i=0}^k f_i$	$\frac{z}{z-1} F(z)$
Faltung	$\sum_{i=0}^k f_i g_{k-i}$	$F(z) \cdot G(z)$
Anfangswert	f_0	$\lim_{z \rightarrow \infty} F(z)$
Endwert	f_∞	$\lim_{z \rightarrow 1} (z-1) \cdot F(z)$

Tabelle 14-2: Operationen im Zeitbereich und im Bildbereich der Z-Transformation

die \mathcal{Z} -Transformierte in die Form einer Potenzreihe

$$F(z) = f_0 + f_1 z^{-1} + f_2 z^{-2} + \dots \quad (14.54)$$

bringt. Diese resultiert unmittelbar in der Folge

$$(f_k) = (f_0, f_1, f_2, \dots) . \quad (14.55)$$

Die Form in Gl.(14.54) kann man dabei für gebrochen rationale $F(z)$ durch Polynomdivision erhalten. So gewinnt man z. B. aus

$$F(z) = \frac{z}{z - 0,5} \quad (14.56)$$

durch einfache Division

$$\begin{array}{r} F(z) = z : (z - 0,5) \\ \hline -z + 0,5 \\ \hline 0,5 \\ \hline -0,5 + 0,25z^{-1} \\ \hline 0,25z^{-1} \end{array} \quad (14.57)$$

und daraus die Folge

$$(f_k) = \left(1, \frac{1}{2}, \frac{1}{4}, \dots \right) . \quad (14.58)$$

14.6.2 Zeitdiskrete Übertragungsfunktion

In Analogie zur Laplace-Transformation kann man die \mathcal{Z} -Transformation nicht nur zum Lösen von Differenzengleichungen mit Eingangsfolgen nutzen, sondern auch eine Übertragungsfunktion herleiten. Wendet man auf beide Seiten einer linearen Differenzengleichung die \mathcal{Z} -Transformation an, so erhält man einen Zusammenhang zwischen den \mathcal{Z} -Transformierten der Eingangs- und Ausgangsfolge in der Form

$$\begin{aligned} Y(z) \cdot [a_0 + a_1 z^{-1} + \dots + a_n z^{-n}] \\ = U(z) \cdot [b_0 + b_1 z^{-1} + \dots + b_m z^{-m}] \end{aligned} \quad (14.59)$$

und daraus mit

$$Y(z) = G(z) \cdot U(z) \quad (14.60)$$

die zugehörige \mathcal{Z} -Übertragungsfunktion des zeitdiskreten Übertragungssystems

$$G(z) = \frac{Y(z)}{U(z)} = \frac{b_0 + b_1 z^{-1} + \dots + b_m z^{-m}}{a_0 + a_1 z^{-1} + \dots + a_n z^{-n}} = \frac{\sum_{i=0}^m b_i z^{-i}}{\sum_{i=0}^n a_i z^{-i}} . \quad (14.61)$$

Oft wird für allgemeinere Betrachtungen $m = n$ gesetzt und in Kauf genommen, dass u.U. einige der Koeffizienten a_i oder b_i zu null werden.

Der statische Übertragungsfaktor K lässt sich alternativ zu der Vorgehensweise in Abschnitt 14.2 im Bildbereich bestimmen indem man anstelle der Folge

$$(u_k) = (1, 1, 1, \dots) \quad (14.62)$$

deren \mathcal{Z} -Transformierte

$$U(z) = \frac{z}{z - 1} \quad (14.63)$$

mit der Übertragungsfunktion multipliziert und den Grenzwert der Ausgangsfolge für $k \rightarrow \infty$ mit der entsprechenden Beziehung aus Tab. 14-2 bestimmt.

Man erhält

$$K = Y_\infty = \lim_{z \rightarrow 1} \left[(z - 1) \frac{z}{z - 1} G(z) \right] = G(z = 1) \quad (14.64)$$

und das entspricht genau der Gl.(14.18).

Man kann mit einem nahezu wortgleichen Vorgehen zum zeitkontinuierlichen Fall zeigen, dass die Pole der Übertragungsfunktion $G(z)$ bei einer minimalen Realisierung den Eigenwerten der Systemmatrix \mathbf{A}_D entsprechen.

Der Übertragungsfunktion $G(z)$ kann man zudem in Analogie zu den Verhältnissen bei kontinuierlichen Systemen eine Gewichtsfolge (g_k) zuordnen, sodass gilt

$$G(z) = \mathcal{Z}\{g_k\} , \quad g_k = \mathcal{Z}^{-1}\{G(z)\} . \quad (14.65)$$

Dabei wird die Ausgangsfolge gleich der Gewichtsfolge, wenn die \mathcal{Z} -Transformierte der Eingangsfolge gleich eins wird. Das ist genau für

$$U(z) = 1 , \quad u_k = (1, 0, 0, \dots) , \quad (14.66)$$

d. h. für den Einheitsimpuls gegeben.

Wie im Zeitkontinuierlichen kann man zeigen, dass für die Gewichtsfolge g_k

$$\sum_{k=0}^{\infty} |g_k| < \infty \quad (14.67)$$

gelten muss, damit die Ausgangsfolge des Übertragungssystems für jede beschränkte Eingangsfolge beschränkt bleibt. Das ist aber wegen

$$\left| \sum_{k=0}^{\infty} g_k z^{-k} \right| \leq \sum_{k=0}^{\infty} |g_k| \quad \text{für } |z| \geq 1 \quad (14.68)$$

sichergestellt durch die Forderung

$$|G(z)| = \left| \sum_{k=0}^{\infty} g_k z^{-k} \right| < \infty \quad \text{für } |z| \geq 1 \quad (14.69)$$

und daraus leitet sich als notwendiges und hinreichendes Kriterium ab:

Polstellen und Stabilität zeitdiskreter Systeme

Ein zeitdiskretes Übertragungssystem in minimaler Realisierung mit Übertragungsfunktion $G(z)$ ist genau dann stabil, wenn die Polstellen von $G(z)$ sämtlich innerhalb des Einheitskreises der z -Ebene liegen.

Somit besitzen die Pole einer zeitdiskreten Übertragungsfunktion denselben Informationsgehalt wie im Zeitkontinuierlichen. Dasselbe gilt ebenfalls für

Nullstellen. Die mit den Pol- und Nullstellen verknüpften Eigenschaften lassen sich dabei stets durch den Zusammenhang $z = e^{sT}$ herleiten. So werden Nullstellen in der rechten s -Halbebene auf Punkte außerhalb des z -Einheitskreises abgebildet.

Minimalphasige zeitdiskrete Systeme

Ein zeitdiskretes System ist minimalphasig, wenn sich alle Null- und Polstellen innerhalb des Einheitskreises befinden.

Interessanterweise ist die Totzeit, die im Zeitkontinuierlichen nicht minimalphasig war, nun minimalphasig, da diese nur Pole bei $z = 0$ besitzt.

Im s -Bereich sind alle Systeme mit Polen ausschließlich auf der reellen Achse (d. h. $s = a$, $a \in \mathbb{R}$) nicht schwingungsfähig. Die reelle Achse wird wegen $e^{aT} \in \mathbb{R}^+$ auf die positive reelle Achse abgebildet.

Schwingungsfähige zeitdiskrete Systeme

Ein zeitdiskretes System ist schwingungsfähig, wenn es mindestens einen Pol besitzt, der nicht auf der positiven reellen Achse liegt.

Als Unterschied können Systeme im Zeitdiskreten sogar schwingungsfähig sein, wenn sie nur einen Pol aufweisen, sofern dieser auf der negativen reellen Achse liegt. Im Zeitkontinuierlichen wäre hierfür ein System von mindestens zweiter Ordnung notwendig, da Pole immer konjugiert auftreten. Weitere Eigenschaften ergeben sich allesamt aus $z = e^{sT}$.

Auch zeitdiskreten PID-Reglern kann eine Übertragungsfunktion zugeordnet werden, wobei ein Vergleich mit dem zeitkontinuierlichen Pendant lohnt. Die in Abschnitt 14.4.1 zu Rückwärtsdifferenzen entwickelte Differenzengleichung des PID-Reglers

$$y_k = y_{k-1} + K_R \cdot \left[\left(1 + \frac{T}{T_n} + \frac{T_v}{T} \right) x_k - \left(1 + 2 \frac{T_v}{T} \right) x_{k-1} + \frac{T_v}{T} x_{k-2} \right] \quad (14.70)$$

lässt sich als \mathcal{Z} -Übertragungsfunktion

$$G_R(z) = K_R \cdot \frac{\left(1 + \frac{T}{T_n} + \frac{T_v}{T} \right) z^2 - \left(1 + 2 \frac{T_v}{T} \right) z + \frac{T_v}{T}}{z^2 - z} \quad (14.71)$$

darstellen. Man erkennt, dass diese zwei Polstellen $z_1 = 1$ und $z_2 = 0$ aufweist, obgleich die Übertragungsfunktion des kontinuierlichen PID-Reglers nur einen Pol bei $s_1 = 0$ besitzt, der dem bei $z_1 = 1$ entspricht. Der zweite Pol bei $z_2 = 0$ ist bei der Umsetzung der Differential- in eine (kausale) Differenzengleichung durch Verwenden der Rückwärtsdifferenz bei Annähern der Differentiation entstanden.

14.6.3 Zeitdiskreter Frequenzgang

Aus der z -Übertragungsfunktion lässt sich auch ein zeitdiskreter Frequenzgang gewinnen. Hierzu kann man die Spezialisierung $s = j\omega$ aus dem Zeitkontinuierlichen durch $z = e^{sT}$ ins Zeitdiskrete übertragen. Der zeitdiskrete Frequenzgang beschreibt dann für stabile Systeme auch die Übertragung (abgetasteter) sinusförmiger Funktionen. Man erhält dabei

$$G(e^{j\omega T}) = \frac{b_0 + b_1 e^{-j\omega T} + b_2 e^{-j\omega 2T} + \dots + b_m e^{-j\omega mT}}{a_0 + a_1 e^{-j\omega T} + a_2 e^{-j\omega 2T} + \dots + a_n e^{-j\omega nT}} . \quad (14.72)$$

Der Frequenzgang hat nicht die Kreisfrequenz als Argument sondern eine Exponentialfunktion mit imaginärem Exponenten. Diese ist periodisch in 2π und daher ist der Frequenzgang selbst periodisch in $2\pi/T$ – auch hier taucht die Shannon-Frequenz erneut auf. Als Konsequenz ähneln die Ortskurven denen der Totzeitglieder.

Als Beispiel wird der zeitdiskrete Differenzierer mit der Differenzengleichung

$$y_k = \frac{K_D}{T} (u_k - u_{k-1}) \quad (14.73)$$

und dem zugehörigen Frequenzgang

$$G_D(e^{j\omega T}) = \frac{K_D}{T} (1 - e^{-j\omega T}) \quad (14.74)$$

betrachtet. Die Ortskurve dieses Frequenzgangs hat für $K_D/T = 1$ die Form eines Kreises mit dem Radius 1 und dem Mittelpunkt 1 (Bild 14-10).

Durch Vergleich mit der Ortskurve des Frequenzgangs des kontinuierlichen Differenziergliedes ist sofort zu erkennen, dass beide Ortskurven nur für

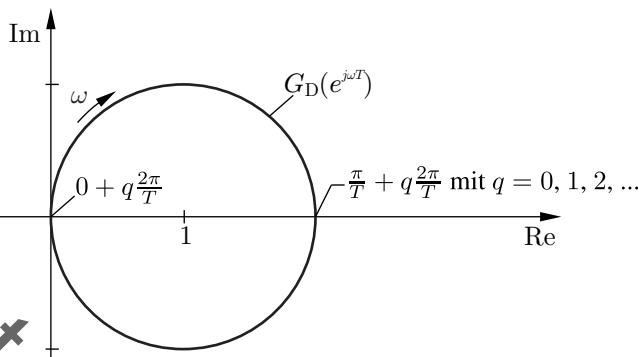


Bild 14-10: Ortskurve des Frequenzgangs des zeitdiskreten Differenzierers

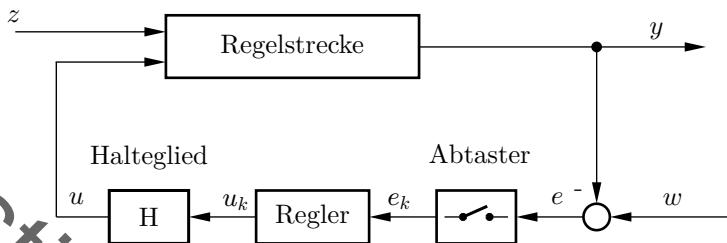
kleine Werte der Frequenz eine ähnlichen Verlauf aufweisen. Insbesondere ist aufgrund der Periodizität aus dem Phasenwinkel der Ortskurve nur schwerlich abzulesen, wo die Pol- und Nullstellen des Systems liegen. Eine einfache Zuordnung wie im Zeitkontinuierlichen von Phasenabfall und Position der Pol-/Nullstellen gibt es in dieser Form nicht.

Auch wenn sich das Nyquist-Kriterium auch für zeitdiskrete Systeme formulieren lässt, so ist dessen Auswertung – insbesondere im Bode-Diagramm – doch insgesamt unhandlich. Daher spielt der zeitdiskrete Frequenzgang – im Gegensatz zum zeitkontinuierlichen – eine stark untergeordnete Rolle in der Systemanalyse und im Reglerentwurf.

14.6.4 Zeitdiskrete Modelle zeitkontinuierlicher Systeme

Mit den Beschreibungsmitteln im Bildbereich kann – ebenso wie über die analytische Lösung im Zustandsraum – ein zeitdiskretes Modell einer zeitkontinuierlichen Regelstrecke inklusive Abtaster und Halteglied ermittelt werden. Hierzu ist der Regelkreis in Bild 14-1 erneut gezeigt.

Es wird angenommen, dass das Halteglied 0. Ordnung ist. In diesem Fall wird das kontinuierliche Übertragungsglied G_s mit Sprungfunktionen angeregt werden, da das Halteglied stückweise konstante Verläufe als Ausgangsgröße besitzt. Damit das abgetastete zeitkontinuierliche Signal $y(t_k)$ dem



zeitdiskreten Signal y_k entspricht, muss das zeitdiskrete System folglich so bestimmt werden, dass die Sprungantworten von zeitkontinuierlichem und zeitdiskretem System an den Abtastzeitpunkten den gleichen Wert aufweisen: $h(t_k) = h_k$.

In Tab. 14-1 stehen in derselben Zeile die jeweils abgetasteten Signale. Daher entspricht das gesuchte $G(z)$ nicht dem Ausdruck, der in derselben Zeile wie $G(s)$ steht – in diesem Fall wären die beiden Impulsantworten an den Abtastzeitpunkten identisch. Stattdessen ist der Weg, zunächst $H(s) = G(s)/s$ zu bestimmen, um die Sprunganregung des Haltegliedes zu berücksichtigen. Anschließend kann man in Tab. 14-1 das zugehörige abgetastete $H(z)$ ablesen. Das gesuchte $G(z)$ ergibt sich nach Tab. 14-2 dann zu

$$G(z) = (1 - z^{-1}) \cdot H(z) = \frac{z - 1}{z} \cdot H(z) \quad . \quad (14.75)$$

Das bedeutet aber auch, dass man sich bei der Überführung von $G(s)$ in ein äquivalentes $G(z)$ für eine Klasse von Anregungssignalen entscheiden muss. Das beschriebene Vorgehen sorgt dafür, dass Abtaster und Halteglied berücksichtigt werden und daher die Sprungantworten korrekt abgebildet werden. Ein Impuls oder eine Rampe würden hingegen falsch dargestellt und das abgetastete zeitkontinuierliche System $y(t_k)$ wäre nicht identisch mit dem zeitdiskreten y_k .

Bestimmen von $G(z)$ aus $G(s)$ mit Abtaster und Halteglied

Das zeitdiskrete System $G(z)$, das an den Abtastzeitpunkten dieselben Werte annimmt wie das zeitkontinuierliche System $G(s)$ inklusive Abtaster und Halteglied, wird wie folgt bestimmt

- $G(s) \rightarrow H(s)$ über $H(s) = G(s)/s$
- $H(s) \rightarrow H(z)$ über Tab. 14-1
- $H(z) \rightarrow G(z)$ über $G(z) = \frac{z-1}{z} \cdot H(z)$

Als Beispiel soll $G(z)$ für eine integrierende Regelstrecke bestimmt werden.
Aus $G_S(s) = \frac{1}{s}$ wird $H(s) = \frac{1}{s^2}$ und mit Tab. 14-1

$$H(z) = \frac{Tz}{(z-1)^2} \quad . \quad (14.76)$$

und daraus mit Gl.(14.75)

$$G(z) = \frac{z-1}{z} \cdot \frac{Tz}{(z-1)^2} = \frac{T}{z-1} \quad . \quad (14.77)$$

Dieses Resultat unterscheidet sich von dem falschen Vorgehen, direkt von $G(s)$ in Tab. 14-1 auf $G(z)$ zu wechseln. Dann erhielte man $\frac{z}{z-1}$ und damit ein System mit u. a. falschem relativen Grad. Die beschriebene Methode ist gleichwertig zu dem Vorgehen über die Transitionsmatrix in 14.4.2.

14.7 Bilineare Transformation

Für die Stabilitätsprüfung zeitkontinuierlicher Systeme gibt es Verfahren wie das Hurwitz- oder Routh-Kriterium, die das explizite Ausrechnen der Polstellen umgehen. Auch für zeitdiskrete Systeme gibt es entsprechende Kriterien, die aber vergleichsweise unübersichtlich ausfallen. Einen oft einfacheren Zugang bietet die sogenannte *bilineare Transformation*.

Bilineare Transformation

Die bilineare Transformation

$$z = \frac{1+w}{1-w} \quad (14.78)$$

ordnet jeder Funktion in z eine Funktion in der neuen Variablen w zu.

Man kann nachrechnen, dass wegen

$$\left| \frac{1+j\omega}{1-j\omega} \right| = 1 \quad (14.79)$$

die bilineare Transformation den Einheitskreis in der z -Ebene auf die imaginäre Achse in der w -Ebene abbildet. Da zudem $w = -1$ auf $z = 0$ abgebildet wird, ist das Abbild des Inneren des z -Einheitskreises die komplexe linke w -Halbebene.

Bilineare Transformation und algebraische Stabilitätskriterien

Die Pole stabiler zeitdiskreter Übertragungsfunktionen müssen in der linken w -Halbebene liegen. Diese Bedingung entspricht der Stabilitätsbedingung für zeitkontinuierliche Systeme und kann daher auch mit den für zeitkontinuierliche Systeme entwickelten Verfahren (wie algebraischen Stabilitätskriterien) geprüft werden.

Ein einfaches Beispiel soll die Behandlung von zeitdiskreten Regelkreisen erläutern. Für den Wirkungsplan in Bild 14-11 gilt

$$G_S(z) = K_S \cdot z^{-2}, \quad G_R(z) = K_R \cdot \frac{z}{z-1} \quad . \quad (14.80)$$

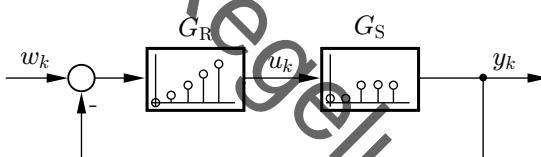


Bild 14-11: Wirkungsplan einer zeitdiskreten Regelung

Man erkennt (ggfs. durch Ausrechnen der Übergangsfunktion oder Ansatz der zugehörigen Differenzengleichung), dass sich die Regelstrecke wie ein Totzeitglied mit $T_t = 2T$ und der Regler wie ein Integrierer verhält. Mit

$$G_0(z) = G_S(z) \cdot G_R(z) = K_S K_R \cdot z^{-2} \cdot \frac{z}{z-1} = \frac{K_0}{z^2 - z} \quad (14.81)$$

wird die Führungsübertragungsfunktion

$$T(z) = \frac{G_0}{1 + G_0} = \frac{1}{\frac{1}{G_0} + 1} = \frac{K_0}{z^2 - z + K_0} \quad (14.82)$$

mit den Polstellen

$$\lambda_{1,2} = \frac{1}{2} \pm \sqrt{\frac{1}{4} - K_0} \quad . \quad (14.83)$$

Aus Gl.(14.83) kann man ablesen, dass für negative K_0 ein Pol außerhalb des Einheitskreises liegt und dass für $K_0 > 0,25$ ein Paar konjugiert-komplexer Polstellen entsteht mit dem Betragsquadrat

$$|\lambda|^2 = \frac{1}{4} + K_0 - \frac{1}{4} = K_0 \quad . \quad (14.84)$$

Stabilität ist nur gesichert, wenn die Polstellen im Innern des Einheitskreises liegen, d.h. wenn $0 < K_0 < 1$ ist.

Zum gleichen Ergebnis gelangt man mit der bilinearen Transformation. Angewandt auf Gl.(14.82) erhält man

$$\begin{aligned} T(w) &= \frac{K_0}{\left(\frac{1+w}{1-w}\right)^2 - \frac{1+w}{1-w} + K_0} \\ &= \frac{K_0(1-w)^2}{(1+w)^2 - (1+w)(1-w) + K_0(1-w)^2} \\ &= \frac{K_0(1-w)^2}{(2+K_0)w^2 + 2(1-K_0)w + K_0} \quad . \end{aligned} \quad (14.85)$$

Wendet man auf den Nenner von Gl.(14.85) das Hurwitz-Kriterium an, so erkennt man, dass für Stabilität $0 < K_0 < 1$ sein muss.

Beide Verfahren liefern also das identische Resultat. Im Allgemeinen wird das direkte Ausrechnen der Polstellen nur für Systeme von 1. oder 2. Ordnung praktikabel sein und schon dort können Fallunterscheidungen von Wurzeln Probleme bereiten. Spätestens ab Ordnung 3 wird die bilineare Transformation in Kombination mit den algebraischen Stabilitätskriterien der schnellere Rechenweg sein.

14.8 Klassischer zeitdiskreter Reglerentwurf

Nachdem nun alle Werkzeuge bereitgestellt sind, mit dem ein Regelkreis als zeitdiskretes System beschrieben werden kann, soll das eigentliche Ziel, der Reglerentwurf, behandelt werden. Der Reglerentwurf erfolgt – im Gegensatz zum Zeitkontinuierlichen – ausschließlich am geschlossenen Regelkreis, da ein Entwurf am aufgeschnittenen Regelkreis wegen des umständlichen Nyquist-Kriteriums nicht zielführend ist.

Wie in Abschnitt 11.3 ist es sinnvoll, das dynamische Verhalten über die Position der Polstellen bzw. Eigenwerte des geschlossenen Regelkreises zu beschreiben. Gesucht ist zunächst das Zielgebiet im Zeitdiskreten, in welchem die Polstellen des geschlossenen Regelkreises vorzugsweise liegen sollten. Der einfachste Weg, diese Gebiet zu bestimmen, ist erneut der Zusammenhang $z = e^{sT}$. Hiermit lässt sich das zeitdiskrete Zielgebiet direkt aus dem Zielgebiet für den zeitkontinuierlichen Fall ableiten.

Das zeitkontinuierliche Zielgebiet ließ sich durch Geraden und Kreise darstellen, wie in Bild 11-7 bzw. 14-12 auf der linken Seite gezeigt ist. Die dieses Gebiet begrenzenden Kurven kann man nun in die z -Ebene abbilden und erhält aus der Forderung

$$\operatorname{Re}(s_i) < -\alpha \quad (14.86)$$

die Vorschrift

$$|z_i| < e^{-\alpha T} , \quad (14.87)$$

die einen Kreis mit dem Radius $e^{-\alpha T}$ beschreibt. Analog ergibt sich die Übertragung des Imaginärteils ω_D .

Die Begrenzungen des durch D definierten Sektors lassen sich beispielhaft für $D = \sqrt{2}/2$ durch

$$s_i = -\beta \pm j\beta \quad (14.88)$$

mit der Laufvariablen β beschreiben. Hieraus wird

$$z_i = e^{(-\beta \pm j\beta) \cdot T} = e^{-\varphi} \cdot e^{\pm j\varphi} \quad (14.89)$$

mit $\varphi = \beta T$ als neuen Laufvariablen. Auf die Abbildung von ω_0 wird meistens verzichtet, da die zugehörigen Eigenschaften durch die anderen Variablen meist gut genug beschrieben werden und das entstehende transformierte Gebiet recht kompliziert ausfällt.

Zielgebiete im Zeitdiskreten

Bild 14-12 zeigt die Zielgebiete für die Pole des geschlossenen Regelkreises in der s - und der z -Ebene. Die Pfeile zeigen dabei die Richtung besser werdender Eigenschaften an.

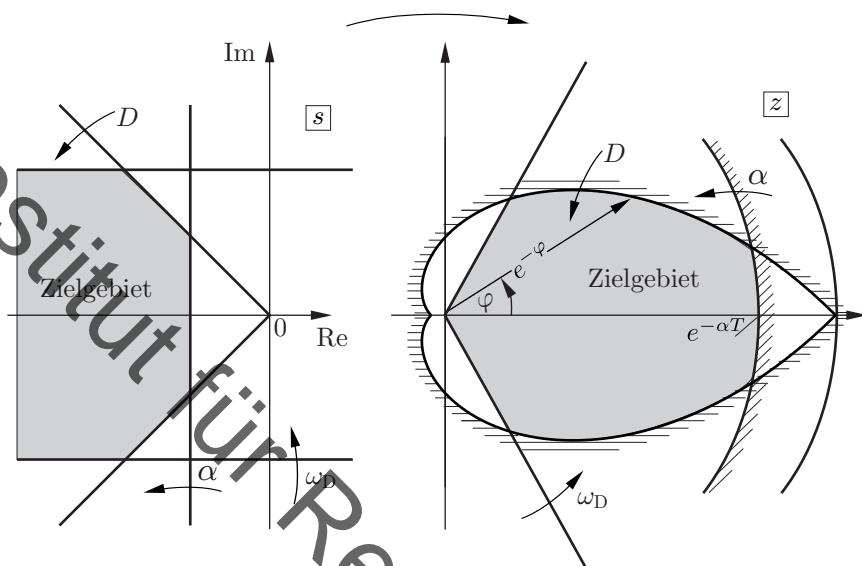


Bild 14-12: Zeitkontinuierliches und zeitdiskretes Zielgebiet

Im Zeitdiskreten scheint es sinnvoll, die Pole möglichst nah an den Ursprung zu setzen. Auch hier muss für eine Gesamtabwägung aber auch die dazu notwendige Reglerverstärkung berücksichtigt werden.

Allgemein kann – im Unterschied zum Zeitkontinuierlichen – die Position $z = 0$ als die „schnellstmögliche“ Position für die Polstellen identifiziert werden. Dies entspricht Polen bei $s = -\infty$. Der Regelgröße klingt dann nicht exponentiell ab, sondern verhält sich wie ein Totzeit-Glied und erreicht nach endlich vielen Zeitschritten ihren Endwert. Diese Besonderheit machen sich die Regler mit endlicher Einstellzeit zunutze, die in Abschnitt 14.9 diskutiert werden.

Beim Versuch, durch einen Regler möglichst viele Polstellen bzw. Eigenwerte in den Zielbereich zu verschieben, sind grundsätzliche Limitierungen der Reglerstruktur zu beachten. So ist allen Regleralgorithmen gemeinsam, dass die \mathcal{Z} -Übertragungsfunktion $G_R(z)$ mindestens so viele Pol- wie Nullstellen aufweist. Andernfalls würde es sich um ein nichtkausales System handeln, das nicht realisierbar ist.

Eine Möglichkeit des Reglerentwurfs ist analog zu Abschnitt 11.3 eine Polvorgabe, wofür sich Zustandsrückführungen in besonderer Weise eignen. Mit dem Stellgesetz

$$\underline{u}_k = -\mathbf{K}x_k \quad (14.90)$$

ergibt sich der geschlossene Regelkreis zu

$$\underline{x}_{k+1} = (\mathbf{A}_D - \mathbf{B}_D \cdot \mathbf{K}) \underline{x}_k \quad (14.91)$$

und damit strukturell gleich wie im Zeitkontinuierlichen. Hieraus folgt direkt, dass das gleiche Vorgehen zur Berechnung des Reglers genutzt werden kann.

Auch systemische Eigenschaften wie Steuerbarkeit und Beobachtbarkeit, auf denen die Lösbarkeit der Polvorgabe basiert, sind daher wortgleich definiert. Die Überprüfungskriterien wie die Rangkriterien nach Kalman in Gl.(11.47) sind ebenfalls identisch. Der einzige Unterschied besteht in den gewünschten Positionen der Eigenwerte von $\mathbf{A}_D - \mathbf{B}_D \cdot \mathbf{K}$.

Auch der Beobachterentwurf kann prinzipiell gleichwertig zum zeitkontinuierlichen Fall erfolgen. Ein besonderes Augenmerk gilt dabei dem Kalmanfilter als optimalem zeitdiskreten Beobachter, welcher aufgrund seiner Prominenz und vielfältigen Anwendung in einem eigenen Kapitel 15 behandelt wird.

Der Entwurf von Ausgangsrückführungen kann in vielerlei Hinsicht wortgleich zum zeitkontinuierlichen Fall vollzogen werden. Das gilt beispielsweise für das Wurzelortskurven-Verfahren, das in der z -Ebene fast genau so angewandt werden kann wie für kontinuierliche Systeme in der s -Ebene: Man geht von den Polen und Nullstellen der Übertragungsfunktion $G_0(z)$ des aufgeschnittenen Regelkreises aus und konstruiert mit den in Abschnitt 11.5 angegebenen Konstruktionsregeln die Wurzelortskurven. Diese ändern sich nicht, da sie auf der Gleichung $G_0 = -1$ basieren, die sowohl im Zeitkontinuierlichen wie im Zeitdiskreten die Stabilität beschreibt. Damit der geschlossene Regelkreis stabil ist, muss der an die Kurven angetragene Parameter K so gewählt werden, dass alle zugehörigen Äste innerhalb des Einheitskreises liegen.

14.9 Regler mit endlicher Einstellzeit

14.9.1 Entwurf

Die bisher vorgestellten Entwurfsverfahren stellten Übertragungen von Entwurfsverfahren für zeitkontinuierliche System dar. Hiervon unterscheidet sich der Regler mit endlicher Einstellzeit, der in seiner Form so nur bei zeitdiskreten Systemen möglich ist.

Regler mit endlicher Einstellzeit „(Dead-beat)“

Der Grundgedanke des Verfahrens ist, einen Regler so zu dimensionieren, dass die Regelgröße nach einem Führungsgrößensprung in m (und damit endlich vielen) Abtastintervallen auf die neue Führungsgröße gebracht und auf diesem Wert gehalten wird.

Ein solches Entwurfsziel – ein Einschwingvorgang endlicher Dauer – ist mit zeitkontinuierlichen Reglern nicht zu erreichen, weil bei linearen kontinuierlichen Systemen solche Einschwingvorgänge (außer bei der Totzeit) prinzipiell unendlich lange dauern.

Dieses Ziel erreicht man, indem man ein System n -ter Ordnung mit einer geeignet gewählten Folge von m Eingangssprungfunktionen ($m \geq n$) auf einen Endwert bringt. Dieser Endwert wird dann – sofern das System kein differenzierendes Verhalten aufweist – nach Abschluss dieser Folge von Sprüngen beibehalten.

Wenn man einen zeitdiskreten Algorithmus angeben kann, der die vorher als geeignet ermittelte Folge von Stellgrößenwerten aus dem Verlauf der Regelgröße erzeugt, ist das beschriebene Problem gelöst.

Die Ableitung des Reglers für endliche Einstellzeit wird in zwei Schritten durchgeführt: Zunächst wird eine Stellgrößenfolge u_k für $k = 0, \dots, m$ bestimmt, die die Regelgröße auf die geänderte Führungsgröße bringt. Aus dem Verlauf der Regelgröße bzw. der Übertragungsfunktion der Regelstrecke und der Stellgrößenfolge wird dann die Übertragungsfunktion des Reglers gewonnen. Falls die Regelstrecke dabei ein Totzeitglied enthält, soll dieses zunächst abgetrennt und später berücksichtigt werden.

Die Übertragungsfunktion der Regelstrecke (ohne Totzeit) wird in ein zeit-

diskretes Modell $G(z)$ umgerechnet und in die Form

$$G(z) = \frac{Y(z)}{U(z)} = \frac{\sum_{i=0}^n b_i z^{-i}}{\sum_{i=0}^n a_i z^{-i}} = \frac{B(z^{-1})}{A(z^{-1})} \quad (14.92)$$

gebracht.

Wenn die Regelgröße in m Abtastschritten ihren Wert von $y_0 = 0$ auf die Führungsgröße $w_0 = 1$ ändern soll, so muss die Folge der Regelgrößenwerte von der Form

$$(y_k) = (0, y_1, y_2, \dots, y_{m-1}, 1, 1, \dots) \quad (14.93)$$

sein und die \mathcal{Z} -Transformierte

$$Y(z) = 0 + y_1 z^{-1} + y_2 z^{-2} + \dots + y_{m-1} z^{-(m-1)} + \frac{z^{-m}}{1 - z^{-1}} \quad (14.94)$$

besitzen.

Wenn die Regelgröße Ausgangsgröße eines zeitkontinuierlichen Übertragungssystems ist und nach m Abtastschritten auch zwischen den Abtastzeitpunkten gleich der Führungsgröße sein soll, so muss die Stellgröße u nach m Abtastschritten ebenfalls ihren Endwert erreicht haben und dann konstant bleiben. D. h. ihre \mathcal{Z} -Transformierte muss die Form

$$U(z) = u_0 + u_1 z^{-1} + u_2 z^{-2} + \dots + u_{m-1} z^{-(m-1)} + u_m \frac{z^{-m}}{1 - z^{-1}} \quad (14.95)$$

besitzen. In die Beziehung

$$Y(z) = G(z) \cdot U(z) = \frac{B(z^{-1})}{A(z^{-1})} \cdot U(z) \quad (14.96)$$

wird nun $Y(z)$ nach Gl.(14.94) eingesetzt und beide Seiten der Gleichung mit $1 - z^{-1}$ multipliziert. Man erhält

$$\begin{aligned} & \left(y_1 z^{-1} + y_2 z^{-2} + \dots + y_{m-1} z^{-(m-1)} \right) (1 - z^{-1}) + z^{-m} = \\ & = \frac{B(z^{-1})}{A(z^{-1})} \cdot U(z) \cdot (1 - z^{-1}) \\ & = B(z^{-1}) \cdot Q(z^{-1}) \end{aligned} \quad (14.97)$$

mit einem noch unbekannten Polynom $Q(z^{-1})$.

Die linke Seite in Gl.(14.97) ist ein endliches Polynom vom Grade m in z^{-1} . Wegen der Gleichheit beider Seiten einer Gleichung muss auch die rechte Seite in Gl.(14.97) ein endliches Polynom vom Grade m in z^{-1} sein. Diese rechte Seite lässt sich als Produkt des schon bekannten Zählerpolynoms $B(z^{-1})$ von $G(z)$ mit dem noch unbekannten $Q(z^{-1})$ auffassen.

Da $B(z^{-1})$ vom Grade n ist und die rechte Seite von Gl.(14.97) ein endliches Polynom vom Grade m sein soll, so muss folglich

$$Q(z^{-1}) = \frac{U(z) \cdot (1 - z^{-1})}{A(z^{-1})} \quad (14.98)$$

ein endliches Polynom in z^{-1} vom Grade $m - n$ sein.

Aus dieser Feststellung folgt u. a., dass die Zahl m der Abtastschritte, innerhalb derer die Regelgröße den Sollwert erreicht, nicht kleiner sein kann als die Ordnung n der Übertragungsfunktion der Regelstrecke.

Wenn die Einstellzeit minimal sein soll, muss $m = n$ gesetzt werden, und damit ergibt sich

$$Q(z^{-1}) = q_0 \quad (14.99)$$

als Polynom vom Grade null.

Eine Aussage über den unbekannten Koeffizienten q_0 erhält man, wenn man Gl.(14.99) in Gl.(14.97) einsetzt,

$$\left(y_1 z^{-1} + y_2 z^{-2} + \dots + y_{m-1} z^{-(m-1)} \right) (1 - z^{-1}) + z^{-m} = B(z^{-1}) \cdot q_0 \quad (14.100)$$

und fordert, dass die Gleichung für $z = 1$ erfüllt ist.

Die linke Seite von Gl.(14.100) wird dann zu eins und das Polynom $B(z^{-1})$ ist gleich der Summe aller seiner Koeffizienten b_i , sodass

$$1 = \sum_{i=0}^n b_i \cdot q_0 \quad (14.101)$$

beziehungsweise

$$Q(z^{-1}) = q_0 = \frac{1}{\sum b_i} \quad (14.102)$$

gelten muss. Dieser Ausdruck ist wohldefiniert, sofern $\sum b_i \neq 0$ gilt. Das ist genau dann der Fall, wenn die Regelstrecke keinen differenzierenden Anteil (also keinen Nullstelle bei $z = 1$) aufweist.

Mit diesem Ergebnis kann bereits die Stellgrößenfolge u_k für $k = 0, \dots, n$ bzw. deren \mathcal{Z} -Transformierte

$$U(z) = \frac{A(z^{-1})}{1 - z^{-1}} Q(z^{-1}) = \frac{A(z^{-1})}{(1 - z^{-1}) \cdot \sum b_i} \quad (14.103)$$

angegeben werden, mit der die Regelgröße in minimaler Zeit von null auf eins verändert werden kann. Man sieht aus Gl.(14.103), dass der Regler seine Nullstellen $A(z^{-1}) = 0$ auf die Positionen der Polstellen der Regelstrecke legt. Hieraus kann man bereits sehen, dass dieser Reglerentwurf mit minimaler Einstellzeit wegen der Pol-Nullstellen-Kürzungen nur für stabile Regelstrecken möglich sein wird.

Für die Regelgröße gilt entsprechend

$$Y(z) = \frac{B(z^{-1})}{1 - z^{-1}} Q(z^{-1}) = \frac{B(z^{-1})}{(1 - z^{-1}) \cdot \sum b_i} \quad . \quad (14.104)$$

Der Regler als Ausgangsrückführung muss die Beziehung

$$U(z) = G_R(z)(W(z) - Y(z)) = G_R(z)(W(z) - G(z)U(z)) \quad (14.105)$$

erfüllen. Wegen des vorausgesetzten sprungförmigen Führungsgrößenverlaufs mit der Sprunghöhe eins ist

$$W(z) = \frac{1}{1 - z^{-1}} \quad . \quad (14.106)$$

Die Reglerübertragungsfunktion folgt durch Einsetzen der hergeleiteten Gleichungen ineinander. Man erhält

$$\begin{aligned} G_R(z) &= -\frac{U(z)}{U(z) \cdot G(z) - W(z)} = -\frac{\frac{A(z^{-1})}{1 - z^{-1}} \cdot Q(z^{-1})}{\frac{B(z^{-1})}{1 - z^{-1}} \cdot Q(z^{-1}) - \frac{1}{1 - z^{-1}}} \quad (14.107) \\ &= \frac{A(z^{-1}) \cdot Q(z^{-1})}{1 - B(z^{-1}) \cdot Q(z^{-1})} \quad . \end{aligned}$$

und im Fall minimaler Einstellzeit mit Gl.(14.102) den Zusammenhang

$$G_R(z) = \frac{A(z^{-1})}{\frac{1}{q_0} - B(z^{-1})} = \frac{A(z^{-1})}{\sum b_i - B(z^{-1})} . \quad (14.108)$$

Für Regelungen mit endlicher aber nicht minimaler Einstellzeit wird $Q(z^{-1})$ ein Polynom höheren als nullten Grades mit einer entsprechenden Zahl von Koeffizienten. Diese Koeffizienten können so gewählt werden, dass geeignet definierte Zusatzbedingungen, z. B. Begrenzungen der Stellgröße oder Minimum der Summe der Quadrate der Stellgrößenfolge, eingehalten werden.

Wegen der Linearität aller Beziehungen arbeiten Regelkreise, die mit dem hier beschriebenen Verfahren ausgelegt worden sind, für jede Art von Führungsgrößensprung mit endlicher bzw. minimaler Einstellzeit, obgleich das Entwurfsverfahren nur für einen Sprung von null auf eins abgeleitet wurde.

14.9.2 Stabilität

Eine wichtige Aussage zur Stabilität aller auf endliche Einstellzeit entworfenen Regelkreise erhält man aus der Führungsübertragungsfunktion

$$T(z) = \frac{Y(z)}{W(z)} = \frac{B(z^{-1})}{1 - z^{-1}} \cdot Q(z^{-1}) \cdot (1 - z^{-1}) = B(z^{-1}) \cdot Q(z^{-1}) . \quad (14.109)$$

Folglich ist $T(z)$ ein endliches Polynom in z^{-1} vom Grade m . Durch Erweitern mit z^m erhält man

$$T(z) = \frac{B(z^{-1}) \cdot Q(z^{-1}) \cdot z^m}{z^m} = \frac{D(z)}{z^m} . \quad (14.110)$$

Dabei ist $D(z)$ ein Polynom in z vom Grade m , dessen Nullstellen die Nullstellen der Führungsübertragungsfunktion sind. Die für die Stabilität wichtigen Polstellen ergeben sich aus dem Nenner, und man erkennt, dass die Führungsübertragungsfunktion $T(z)$ eine m -fache Polstelle bei $z = 0$ besitzt.

Stabilität des Reglers mit endlicher Einstellzeit

Da alle Polstellen des geschlossenen Regelkreises unter Einsatz eines Reglers mit endlicher Einstellzeit bei $z = 0$ liegen, ist der entworfene Regelkreis stets stabil. Eine wichtige Einschränkung sind mögliche instabile Pol-Nullstellen-Kürzungen. Diese treten hier genau dann auf, wenn die Regelstrecke instabil ist. Deswegen wird der Regelkreis – auch wenn Gl. (14.110) dies nicht erwarten lässt – nur für stabile Regelstrecken stabil sein.

Im Falle instabiler Regelstrecken wird durch Gl. (14.107) ein Regler spezifiziert, dessen Übertragungsfunktion eine oder mehrere Nullstellen außerhalb des Einheitskreises aufweist, die entsprechende Polstellen der Regelstreckenübertragungsfunktion kompensieren. Da diese Kompensation praktisch nie vollkommen gelingt, ist ein solcher Regelkreis instabil.

Für Regelstrecken mit einer Totzeit $T_t = d \cdot T$, d ganzzahlig, muss das beschriebene Entwurfsverfahren nur geringfügig modifiziert werden. Die Regelstrecke soll durch

$$G_d(z) = G(z) \cdot z^{-d} = \frac{B(z^{-1})}{A(z^{-1})} \cdot z^{-d} \quad (14.111)$$

beschrieben werden, die Stellfolge bleibt unverändert und die Regelgröße wird zu

$$Y(z) = G_d(z) \cdot U(z) = G(z) \cdot U(z) \cdot z^{-d} \quad . \quad (14.112)$$

Für die Bestimmung des Hilfspolynoms $Q(z^{-1})$ gelten alle oben angegebenen Überlegungen weiter. Die Reglerübertragungsfunktion erhält man in analoger Weise zu

$$G_R(z) = -\frac{U(z)}{U(z)G(z) \cdot z^{-d} - W(z)} = \frac{A(z^{-1})Q(z^{-1})}{1 - B(z^{-1})Q(z^{-1}) \cdot z^{-d}} \quad (14.113)$$

Hieraus ergibt sich die Führungsübertragungsfunktion des Regelkreises zu

$$T(z) = B(z^{-1}) \cdot Q(z^{-1}) \cdot z^{-d} = \frac{D(z)}{z^m} \cdot z^{-d} \quad , \quad (14.114)$$

d. h. die Aussagen über die Stabilität von auf endliche Einstellzeit ausgelegten Regelkreisen gelten unverändert.

14.9.3 Beispiel

Als einfaches Beispiel soll ein Regler mit minimaler Einstellzeit für eine Regelstrecke mit Verzögerung erster Ordnung und Totzeit mit der Übertragungsfunktion

$$G_S(s) = \frac{K_S}{1 + sT_S} e^{-sT_t}, \quad T_S = a \cdot T, \quad T_t = d \cdot T \quad (14.115)$$

ermittelt werden.

Das in Abschnitt 14.6.4 beschriebene Vorgehen liefert die zugehörige Übertragungsfunktion des zeitdiskreten Modells zu

$$G_d(z) = K_S \cdot \frac{1 - z_1}{z - z_1} \cdot z^{-d} = \frac{K_S(1 - z_1)z^{-1}}{1 - z_1z^{-1}} \cdot z^{-d} \quad (14.116)$$

mit $z_1 = e^{-\frac{1}{a}}$. Wegen

$$A(z^{-1}) = 1 - z_1z^{-1}, \quad B(z^{-1}) = K_S(1 - z_1)z^{-1} \quad (14.117)$$

gilt mit

$$Q(z^{-1}) = \frac{1}{\sum b_i} = \frac{1}{K_S(1 - z_1)} \quad (14.118)$$

für die Stellfolge nach Gl.(14.103)

$$U(z) = \frac{1 - z_1z^{-1}}{(1 - z^{-1})K_S(1 - z_1)} \quad . \quad (14.119)$$

Daraus erhält man z. B. durch Ausdividieren

$$\begin{aligned} U(z) &= \frac{1}{K_S(1 - z_1)} (1 + (1 - z_1)z^{-1} + (1 - z_1)z^{-2} + \dots) \\ &= \frac{1}{K_S} \left(\frac{1}{1 - z_1} + z^{-1} + z^{-2} + \dots \right) \end{aligned} \quad (14.120)$$

und daraus ohne weitere Rechnung die Werte der Folge selbst.

Die Reglerübertragungsfunktion ergibt sich mit Gl.(14.113) zu

$$\begin{aligned} G_R(z) &= \frac{A(z^{-1})}{\frac{1}{Q(z^{-1})} - B(z^{-1})z^{-d}} = \frac{1 - z_1 z^{-1}}{K_S(1 - z_1) - K_S(1 - z_1)z^{-1}z^{-d}} \\ &= \frac{1}{K_S(1 - z_1)} \cdot \frac{1 - z_1 z^{-1}}{1 - z^{-(1+d)}} . \end{aligned} \quad (14.121)$$

Für $d = 5$ wird $z_1 = e^{-0,2} \approx 0,819$, und mit $d = 2$ erhält man als Reglerübertragungsfunktion

$$G_R(z) \approx \frac{5,517}{K_S} \cdot \frac{1 - 0,819 z^{-1}}{1 - z^{-3}} \quad (14.122)$$

und für den Zusammenhang zwischen Stellgröße und Regelabweichung aus $U(z) = G_R(z)(W(z) - Y(z))$ die Differenzengleichung

$$u_k - u_{k-3} = -\frac{5,517}{K_S} (w_k - y_k - 0,819(w_{k-1} - y_{k-1})) . \quad (14.123)$$

Der Verlauf der Regelgröße des geschlossenen Regelkreises ist für $K_S = 1$, $T = 1 \text{ sec}$ und einen Einheitssprung der Führungsgroße in Bild 14-13 gezeigt. Man sieht, dass der gewünschte Endwert durch zwei passende Stellausschläge nach drei Zeitschritten erreicht wird. Dies entspricht genau der Systemordnung der zeitdiskreten Regelstrecke (2 von der Totzeit, 1 von dem PT₁).

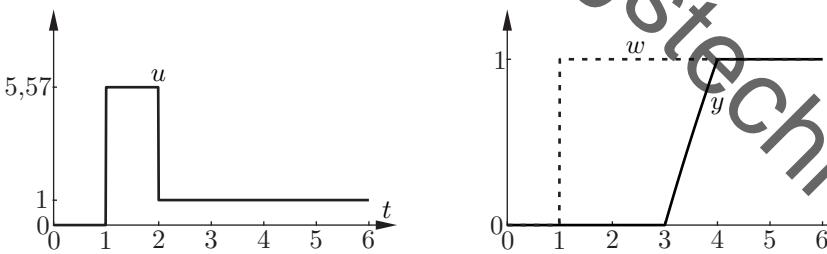


Bild 14-13: Verlauf der Regelgröße beim Regler mit endlicher Einstellzeit

15 Kalmanfilter

15.1 Allgemeines

Mit den in den vorangegangenen Kapiteln vorgestellten Verfahren ist bereits ein breites Methodenspektrum für den Reglerentwurf vorhanden. Für eine erfolgreiche Regelung müssen dabei bestimmte Voraussetzungen erfüllt sein. Hierzu gehören ein Modell der Regelstrecke, Aktoren, mit denen auf das System eingewirkt werden kann, und Sensoren, die die Regelgrößen erfassen. Gerade der letzte Punkt ist in der Praxis oftmals nur mit einer unzureichenden Qualität gegeben. Das liegt daran, dass präzise Sensoren teuer oder nicht verfügbar sind oder die physikalischen Messprinzipien fehleranfällig sind. Auch die notwendige Kalibrierung der Sensoren ist aufwändig und kann häufig nicht mit der erforderlichen Genauigkeit durchgeführt werden. Die Qualität der Regelung hängt aber entscheidend von den zur Verfügung stehenden Informationen über die Regelgrößen ab, sodass schlechte Sensoren große Auswirkungen auf die Regelgüte haben.

Im Abschnitt 11.4 wurde bereits diskutiert, dass Beobachter auch als Filter fungieren können, weil sie durch ein Parallelmodell der Regelstrecke die Messungen plausibilisieren und somit das Rauschen reduzieren. An dieser Stelle knüpft das Kalmanfilter an, welches nach Rudolf Kálmán als einem seiner Erfinder benannt ist.

Kalmanfilter

Das Kalmanfilter ist ein Filter, welches die Modellinformationen und die Messungen vereint und hieraus eine optimale Filterung und Zustands schätzung ableitet, welche auch Informationen über die Vertrauenswürdigkeit der Schätzung enthält.

Der entstehende hocheffiziente Algorithmus eignet sich sowohl zur Zustands schätzung als auch zur Filterung von Messwerten oder aber zur Fusionierung von Informationen verschiedener Sensoren. Daher wird das Kalmanfilter in vielzähligen Anwendungen wie bei Navigationslösungen oder Trackingverfahren eingesetzt. Als eine der prominentesten Anwendungen gilt dabei der Einsatz auf der ersten Mondlandefähre, wo dem auf dem Apollo Guidance Computer implementierten Kalmanfilter eine Schlüssel rolle bei der erfolgreichen Mondlandung zugeschrieben wird [37].

In seiner Optimalität als bestmögliches Filter gemäß einer bestimmten Aufgabenstellung ähnelt das Kalmanfilter in einigen Punkten den Verfahren, die in Kapitel 18 unter dem Stichwort der optimalen Regelung diskutiert werden. In der Formulierung der Gleichungen des Kalmanfilters kann man wahlweise im Zeitkontinuierlichen oder im Zeitdiskreten vorgehen. Da das zeitdiskrete Kalmanfilter sich signifikant größerer Beliebtheit erfreut, wird zweiterer Weg gewählt.

Die klassische Herleitung inklusive Beweis der Optimalität erfolgt mit umfangreichen Methoden der Stochastik. Hier wird ein anderer Zugang gewählt, der an die Ausführungen zur Identifikation in Kapitel 8 anknüpft und mit rudimentären Kenntnissen der Statistik auskommt. Auf die statistischen Implikationen des Kalmanfilters wird mit Verweis auf [52] an passender Stelle verwiesen.

15.2 Herleitung

Das Kalmanfilter ist von seiner Grundstruktur ein zeitdiskreter Beobachter und einem Luenberger-Beobachter sehr ähnlich. Formuliert man den Luenberger-Beobachter für den zeitdiskreten MIMO-Fall, so erhält man durch Gegenüberstellung mit dem zeitkontinuierlichen Fall

$$\begin{aligned}\dot{\boldsymbol{x}} &= \mathbf{A}\boldsymbol{x} + \mathbf{B}\boldsymbol{u} \quad , \quad \boldsymbol{y} = \mathbf{C}\boldsymbol{x} \\ \Rightarrow \hat{\dot{\boldsymbol{x}}} &= \mathbf{A}\hat{\boldsymbol{x}} + \mathbf{B}\boldsymbol{u} + \mathbf{L}\mathbf{C}(\boldsymbol{x} - \hat{\boldsymbol{x}})\end{aligned}\tag{15.1}$$

zunächst die Beschreibung

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \mathbf{A}\boldsymbol{x}_k + \mathbf{B}\boldsymbol{u}_k \quad , \quad \boldsymbol{y}_k = \mathbf{C}\boldsymbol{x}_k \\ \Rightarrow \hat{\boldsymbol{x}}_{k+1} &= \mathbf{A}\hat{\boldsymbol{x}}_k + \mathbf{B}\boldsymbol{u}_k + \mathbf{L}\mathbf{C}(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k) \quad .\end{aligned}\tag{15.2}$$

Hier wurde der zeitdiskrete Beobachter genau parallel zum zeitkontinuierlichen gebildet. Die zeitdiskreten Zustandsraummatrizen werden im Gegensatz zu Kapitel 14 nicht mehr speziell durch das rechte Subskript D gekennzeichnet, da im Folgenden nur zeitdiskrete Matrizen Verwendung finden. Die Matrizen der zeitdiskreten Darstellung erhält man üblicherweise aus denen der kontinuierlichen Darstellung über die analytische Lösung (siehe Abschnitt 14.4.2).

Der zeitdiskrete Beobachter weist hier eine Schwäche auf, die im Zeitkontinuierlichen keine Rolle spielte. Um das einzusehen, teilt man Gl.(15.2) in

zwei Teile auf:

$$\hat{x}_{k+1} = \underbrace{\mathbf{A}\hat{x}_k + \mathbf{B}u_k}_{\text{Prädiktion}} + \underbrace{\mathbf{L}(y_k - \hat{y}_k)}_{\text{Korrektur}} . \quad (15.3)$$

Der erste Term beschreibt die Streckenkopie und damit die reine Vorwärts-simulation auf Basis des bisher geschätzten Zustands. Er kann daher als bestmögliche Vorhersage (Prädiktion) \hat{x}_{k+1} des Zustandes zum nächsten Zeitschritt $k+1$ auf Basis der Schätzung des Zustandes zum aktuelle Zeitschritts \hat{x}_k verstanden werden.

Der zweite Term enthält die Abweichung zwischen dem tatsächlich gemes-senen Ausgangssignal y_k und dem Ausgang \hat{y}_k , welcher von der Prädiktion vorhergesagt wurde. Auf Basis dieser Abweichung korrigiert das Filter die Prädiktion mit der noch zu bestimmenden Gewichtung \mathbf{L} .

Diese Aufteilung zeigt, dass ein Algorithmus wie in Gl.(15.3) nicht sinnvoll ist. Dort wird nämlich die Korrektur von \hat{x}_{k+1} – also zum Zeitschritt $k+1$ – auf Basis der Abweichung im Ausgangssignal zum Zeitschritt k durchgeführt. In einer algorithmischen Denkweise ist es aber viel logischer, eine Vorhersage für den Zeitschritt $k+1$ zu machen und diese Vorhersage dann mit dem zeitlich passenden Messwert zu korrigieren. Dies entspräche

$$\hat{x}_{k+1} = \underbrace{\mathbf{A}\hat{x}_k + \mathbf{B}u_k}_{\text{Prädiktion}} + \underbrace{\mathbf{L}(y_{k+1} - \hat{y}_{k+1})}_{\text{Korrektur}} . \quad (15.4)$$

Im Zeitkontinuierlichen tritt dieses Problem nicht auf, da \dot{x} und x am selben Zeitpunkt definiert sind.

Die Formulierung Gl.(15.4) ist nicht explizit nach \hat{x}_{k+1} aufgelöst, da \hat{y}_{k+1} ebenfalls von \hat{x}_{k+1} abhängt. Als zeitdiskreter Algorithmus aufgeschrieben mit zwei separaten Schritten der Prädiktion und Korrektur stellt dies aber kein Problem dar.

Zeitdiskreter Beobachter mit Prädiktion und Korrektur

Ein zeitdiskreter Beobachter mit Prädiktion und Korrektur kann wie folgt algorithmisiert werden:

- 0) Wähle eine Startschätzung \hat{x}_0^+ und setze den Index $k = 0$
- 1) Führe eine Prädiktion durch: $\hat{x}_{k+1}^- = \mathbf{A}\hat{x}_k^+ + \mathbf{B}u_k$

- 2) Bestimme den vorhergesagten Ausgang: $\hat{y}_{k+1} = \mathbf{C}\hat{x}_{k+1}^-$
- 3) Erhalte die Messung zum Zeitschritt $k + 1$: y_{k+1}
- 4) Korrigiere die Schätzung: $\hat{x}_{k+1}^+ = \hat{x}_{k+1}^- + \mathbf{L}(y_{k+1} - \hat{y}_{k+1})$
- 5) Setze $k \rightarrow k + 1$ und wiederhole bei 1).

In diesem Algorithmus beschreibt die Notation f^- Signale vor der Korrektur und f^+ nach der Korrektur. Das Ergebnis ist eine Schätzung des Zustands \hat{x}^+ oder aber ein gefiltertes Ausgangssignal $\hat{y}^+ = \mathbf{C}\hat{x}^+$.

Die offene Frage ist, wie die Verstärkung \mathbf{L} bestimmt werden soll. Hier sind prinzipiell auch Verfahren der Polvorgabe denkbar. Das Kalmanfilter hingegen verfolgt den Ansatz, die beiden algorithmischen Anteile der Prädiktion und Korrektur analog zur Modellidentifikation als lineares Ausgleichsproblem aufzufassen. Der korrigierte Zustand \hat{x}^+ soll nämlich sowohl zur Modellvorhersage als auch zur Messung passen.

Damit ergibt sich das least-square-Problem

$$\min_{\hat{x}^+} \left(\underbrace{\|\hat{x}^+ - \hat{x}^-\|_{\mathbf{W}_x}^2}_{\text{Prädiktion}} + \underbrace{\|\mathbf{C}\hat{x}^+ - y\|_{\mathbf{W}_y}^2}_{\text{Korrektur}} \right). \quad (15.5)$$

Die beiden Gewichtungsmatrizen \mathbf{W}_x und \mathbf{W}_y bestimmen, ob der Schätzer eher dem Modell oder der Messung vertraut.

Im Vergleich zur Modellidentifikation kann \hat{x}^- als Startschätzung für die zu bestimmenden Parameter \hat{x}^+ gedeutet werden. Das Gleichungssystem zu Gl.(15.5) ergibt sich vektorwertig zu

$$\underbrace{\left\| \begin{bmatrix} \mathbf{I} \\ \mathbf{C} \end{bmatrix} \cdot \hat{x}^+ - \begin{bmatrix} \hat{x}^- \\ y \end{bmatrix} \right\|_{\mathbf{W}}^2}_{\simeq \|\mathbf{M}\boldsymbol{\vartheta} - \mathbf{b}\|^2}, \quad \mathbf{W} = \begin{bmatrix} \mathbf{W}_x & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_y \end{bmatrix} \quad (15.6)$$

und ist damit überbestimmt. Strukturell gleicht es aber genau einer Parameteridentifikation $\|\mathbf{M}\boldsymbol{\vartheta} - \mathbf{b}\|_{\mathbf{W}}^2$ und kann ebenso über die gewichtete Pseudoinverse Gl.(8.15) gelöst werden. Multipliziert man diese mit den vorliegenden Matrizen aus, so gewinnt man nach einigen Umformungen

$$\hat{x}^+ = \mathbf{M}_{\mathbf{W}}^\dagger \mathbf{b} = \hat{x}^- + \underbrace{(\mathbf{W}_x^{-1} + \mathbf{C}^T \mathbf{W}_y^{-1} \mathbf{C})^{-1} \mathbf{C}^T \mathbf{W}_y^{-1} (\mathbf{y} - \hat{y})}_{\mathbf{L}} \quad (15.7)$$

und damit tatsächlich eine Korrektur in der Struktur eines Beobachters.

Zeitdiskreter Beobachter mit least-square

Ein zeitdiskreter Beobachter mit least-square kann wie folgt algorithmiert werden:

- 0) Wähle eine Startschätzung \hat{x}_0^+ , Gewichte \mathbf{W}_x und \mathbf{W}_y und setze den Index $k = 0$
- 1) Führe eine Prädiktion durch: $\hat{x}_{k+1}^- = \mathbf{A}\hat{x}_k^+ + \mathbf{B}u_k$
- 2) Bestimme den vorhergesagten Ausgang: $\hat{y}_{k+1} = \mathbf{C}\hat{x}_{k+1}^-$
- 3) Erhalte die Messung zum Zeitschritt $k + 1$: y_{k+1}
- 4) Bestimme \mathbf{L} über eine Pseudoinverse gemäß Gl.(15.7)
- 5) Korrigiere die Schätzung: $\hat{x}_{k+1}^+ = \hat{x}_{k+1}^- + \mathbf{L}(y_{k+1} - \hat{y}_{k+1})$
- 6) Setze $k \rightarrow k + 1$ und wiederhole bei 1).

Die Frage nach passenden Gewichten \mathbf{W}_x und \mathbf{W}_y für das Optimierungsproblem verbleibt. Betrachtet man Gl.(15.5), so erscheint es sinnvoll, durch die Gewichte zu beschreiben, wie sehr man den Messungen (in Form von \mathbf{W}_y) beziehungsweise der Modellvorhersage (\mathbf{W}_x) vertrauen kann. Hierbei wird ein großes Vertrauen sich in großen Einträgen widerspiegeln.

Um die Vertrauenswürdigkeit mathematisch zu fassen, bemüht das Kalmanfilter Bewertungen aus der Statistik. Hierzu versteht man Signale als zufällig, die aufgrund von Zufallsprozessen unterschiedliche Werte annehmen.

Stochastische Signale

Gegeben ist ein vektorwertiges Signal $f(t)$. Das Signal heißt stochastisch, wenn für jedes t der Wert $x = f(t)$ nicht fest definiert ist, sondern zufällige Werte x annimmt.

Das bedeutet nicht, dass der Wert $x = f(t)$ beliebig ist. Stattdessen unterliegen für jedes feste t die Werte x einer Wahrscheinlichkeitsverteilung, die beschreibt, wie wahrscheinlich es ist, einen konkreten Wert x zu beobachten.

Wahrscheinlichkeitsdichte und Erwartungswert

Gegeben ist ein stochastisches Signal $\mathbf{x} = \mathbf{f}(t)$. Die Wahrscheinlichkeitsdichte $\rho(\mathbf{x})$ beschreibt, mit welcher relativen Häufigkeit ein Wert von \mathbf{x} angenommen wird.

Der Erwartungswert $E[\mathbf{f}]$ mit

$$E[\mathbf{f}] = \int \mathbf{x} \rho(\mathbf{x}) d\mathbf{x} \quad (15.8)$$

beschreibt, welcher Signalwert für \mathbf{f} im Mittel erwartet werden kann.

Würde man die zufällige Messung von $\mathbf{f}(t)$ für ein festes t mehrfach wiederholen und die gefundenen Ergebnisse mitteln, so erhielte man für unendlich viele Wiederholungen im Grenzwert den Erwartungswert.

Für nicht stochastische Signale (das sind deterministische Signale) entspricht der Erwartungswert dem Signalwert selbst. Insofern kann der Erwartungswert als der „echte“ Signalwert verstanden werden.

Die zufälligen Auswertungen von \mathbf{f} können – auch wenn sie im Mittelwert den Erwartungswert ergeben – deutlich vom Erwartungswert abweichen. Dies wird durch die *Kovarianz* des stochastischen Signals beschrieben.

Kovarianz

Die Kovarianz berechnet sich aus dem Erwartungswert zu

$$\sigma^2[\mathbf{f}] = E[(\mathbf{f} - E[\mathbf{f}])(\mathbf{f} - E[\mathbf{f}])^T] \quad (15.9)$$

und beschreibt, welche quadratische Abweichung vom Erwartungswert im Mittel erwartet werden kann.

Die Kovarianz ist für vektorwertige Signale $\mathbf{x} \in \mathbb{R}^n$ eine $n \times n$ -Matrix, die auch Kovarianzmatrix genannt wird und bei stochastischen Signalen positiv definit ist.

Wenn die einzelnen Komponenten des Vektors statistisch voneinander unabhängig sind (d. h. der zufällige Wert von f_i ist unabhängig vom zufälligen

Wert f_j für jedes feste t), so wird die Kovarianz zu einer Diagonalmatrix

$$\sigma^2 [f] = \Sigma = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{bmatrix} \quad (15.10)$$

mit den positiven Varianzen der einzelnen Komponenten des Vektors auf der Diagonalen. Die Wurzeln der Diagonaleinträge σ_i werden auch als *Standardabweichung* (auf Englisch „standard deviation“) bezeichnet. Als Beispiel sind Erwartungswert und Varianz für eine sogenannte Normalverteilung, die auch Gauß¹-Verteilung genannt wird, zusammen mit der Wahrscheinlichkeitsdichte in Bild 15-1 dargestellt.

Sehr viele technisch interessante Größen gehorchen mit guter Näherung einer Normalverteilung. Eine solche Verteilung ist immer dann zu erwarten, wenn sich viele voneinander statistisch unabhängige Effekte in einer Größe überlagern. So strebt die Verteilung einer Summe voneinander unabhängiger Größen für zunehmende Zahl von Summanden gegen die Normalverteilung, auch wenn die einzelnen Summanden einer gänzlich anderen Verteilung gehorchen. Die Normalverteilung ist damit eine gute Näherung für Messrauschen.

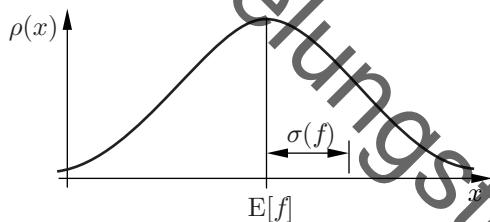


Bild 15-1: Normalverteilung mit Erwartungswert und Varianz

Unabhängig von der tatsächlichen Verteilung soll untersucht werden, wie sich Erwartungswert und Varianz im Rahmen des least-square des Kalman-filters Gl.(15.5) verhalten. Hierzu wird angenommen, dass bei einem linea- ren Ausgleichsproblem der Ausgangsvektor b mit einem zufälligen Signal v additiv überlagert und somit verrauscht wird:

$$M\vartheta = b + v \quad . \quad (15.11)$$

¹Carl Friedrich Gauß (1777-1855), deutscher Mathematiker und Physiker [15]

Aufgrund der Linearität des Erwartungswertes ergibt sich $E[\boldsymbol{\vartheta}]$ mithilfe der gewichteten Pseudoinversen zu

$$E[\boldsymbol{\vartheta}] = E\left[\mathbf{M}_W^\dagger(\mathbf{b} + \mathbf{v})\right] = E\left[\mathbf{M}_W^\dagger \mathbf{b}\right] + \mathbf{M}_W^\dagger E[\mathbf{v}] = \boldsymbol{\vartheta} + \mathbf{M}_W^\dagger E[\mathbf{v}] . \quad (15.12)$$

Das bedeutet, dass bei mittelwertfreiem Rauschen $E[\mathbf{v}] = 0$ der Erwartungswert der Parameteridentifikation der tatsächlichen unverrauschten Lösung entspricht. Ein solcher Schätzer heißt *erwartungstreu*.

Das Weiteste kann man zeigen [52], dass die Kovarianz $\sigma^2[\boldsymbol{\vartheta}]$ genau dann minimal wird, wenn die Gewichtung des least squares mit der inversen Kovarianz von \mathbf{v} vorgenommen wird.

Schätzung mit minimaler Kovarianz

Gegeben ist das Parameteridentifikationsproblem

$$\mathbf{M}\boldsymbol{\vartheta} = \mathbf{b} + \mathbf{v} , \quad E[\mathbf{v}] = 0 , \quad \sigma^2[\mathbf{v}] = \mathbf{R} . \quad (15.13)$$

Dann ist mit $\mathbf{W} = \mathbf{R}^{-1}$ und der gewichteten Pseudoinverse

$$\mathbf{M}_{\mathbf{R}^{-1}}^\dagger = (\mathbf{M}^T \mathbf{R}^{-1} \mathbf{M})^{-1} \mathbf{M}^T \mathbf{R}^{-1} \quad (15.14)$$

die Lösung $\boldsymbol{\vartheta} = \mathbf{M}_{\mathbf{R}^{-1}}^\dagger(\mathbf{b} + \mathbf{v})$ erwartungstreu und mit minimaler Kovarianz.

Die Gewichtung $\mathbf{W} = \mathbf{R}^{-1}$ ist in diesem Sinne die bestmögliche Gewichtung, da sie die Abweichungen der Schätzung vom unverrauschten Wert im quadratischen Mittel minimiert. Da \mathbf{R} positiv definit ist, existiert auch \mathbf{R}^{-1} und ist ebenfalls positiv definit. Somit erfüllt diese Gewichtung die Voraussetzungen des least-square.

Auch anschaulich ist die Gewichtung $\mathbf{W} = \mathbf{R}^{-1}$ sinnvoll: Große Einträge in \mathbf{R} entsprechen einer großen Kovarianz und damit einer breiten Streuung der Zufallswerte um den eigentlichen korrekten Wert herum. Diesen Signalkomponenten kann folglich wenig vertraut werden und sie sollten daher im least-square gering gewichtet werden. Genau das geschieht durch die Inversenbildung.

Die resultierende minimale Kovarianz von $\boldsymbol{\vartheta}$ lässt sich über Gl.(15.9) zu

$$\sigma^2[\boldsymbol{\vartheta}] = (\mathbf{M}^T \mathbf{R}^{-1} \mathbf{M})^{-1} \quad (15.15)$$

berechnen. Somit ist eine Möglichkeit gegeben, auch die Verlässlichkeit der Parameteridentifikation in Form einer Kovarianz direkt anzugeben.

Man sieht an dieser Stelle, dass schlecht konditionierte \mathbf{M} , für die \mathbf{M}^{-1} große Einträge annimmt, auf eine große Kovarianz in ϑ führen. Dies stützt die Argumentation in Kapitel 8, dass die Matrix \mathbf{M} unterschiedliche dynamische Verläufe enthalten sollte.

Da die beschriebene Gewichtung mit der inversen Kovarianz die bestmögliche ist, bietet es sich an, diese auch in dem least-square Beobachter zu verwenden. Tut man dies, so erhält man das sogenannte *Kalmanfilter*.

Optimalität des Kalmanfilters

Das Kalmanfilter ist das lineare Filter, das erwartungstreu ist und die Kovarianz des Schätzfehlers minimiert.

Die letzte Hürde, um die Bestimmungsgleichung des Kalmanfilters zu gewinnen, ist es, die Kovarianz der additiven Störung wie in Gl.(15.11) für den zeitdiskreten Beobachter in Gl.(15.6) zu bestimmen. Hierzu reicht es wegen

$$\mathbf{W}^{-1} = \begin{bmatrix} \mathbf{W}_x & \\ & \mathbf{W}_y \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{W}_x^{-1} & \\ & \mathbf{W}_y^{-1} \end{bmatrix} \quad (15.16)$$

aus, separat die Kovarianz von $\hat{\mathbf{x}}^-$ und die Kovarianz von \mathbf{y} zu bestimmen.

Der Ansatz von \mathbf{W} in Form von zwei Blöcken \mathbf{W}_x und \mathbf{W}_y ist dabei nur dann zielführend, wenn $\hat{\mathbf{x}}^-$ und \mathbf{y} voneinander statistisch unabhängig sind. Dies ist in vielen praktischen Problemen aber zumindest näherungsweise erfüllt.

Die Kovarianz der Messung \mathbf{y} ist dabei relativ klar zu bestimmen. Hierzu modelliert man ein additiv überlagertes, mittelwertfreies Messrauschen in der Messgleichung in Form von

$$\mathbf{y}_k = \mathbf{C}\hat{\mathbf{x}}_k + \mathbf{v}_k \quad , \quad E[\mathbf{v}_k] = 0 \quad , \quad \sigma^2[\mathbf{v}_k] = \mathbf{R} \quad . \quad (15.17)$$

Damit ist $\sigma^2[\mathbf{y}] = \mathbf{R}$ und es sollte $\mathbf{W}_y = \mathbf{R}^{-1}$ gewählt werden.

Für den zweiten Term wird die Kovarianz von $\hat{\mathbf{x}}^-$ gesucht – also die Kovarianz der Zustandsprädiktion. Man kann sich vorstellen, dass diese nicht konstant sein wird. Zu Beginn (d. h. kurz nach Einschalten des Beobachters), liegt nur eine möglicherweise schlechte Startschätzung $\hat{\mathbf{x}}(0)$ vor und

die Kovarianz ist entsprechend groß. Nach einigen Iterationen, wenn Prädiktion und Messung näherungsweise übereinstimmen, sinkt diese Kovarianz. Sie ist also zeitvariant als \mathbf{P}_k^- anzusetzen.

Außerdem ist die Kovarianz von $\hat{\mathbf{x}}^+$ nach der Korrektur bekannt. Diese lässt sich nämlich über Gl.(15.15) direkt berechnen, da dieser Korrekturschritt analog zu einer Bildung der Pseudoinversen abläuft.

Dieses Wissen um die Veränderung der Kovarianz durch die Korrektur kann mit einer Startschätzung für die Kovarianz und dem Modellwissen zu einer Bestimmung der Kovarianz der Prädiktion genutzt werden. Hierzu ergänzt man das lineare Systemmodell um ein Modellrauschen \mathbf{w}_k :

$$\hat{\mathbf{x}}_{k+1}^- = \mathbf{A}\hat{\mathbf{x}}_k^+ + \mathbf{B}\mathbf{u}_k + \mathbf{w}_k \quad , \quad E[\mathbf{w}_k] = 0 \quad , \quad \sigma^2[\mathbf{w}_k] = \mathbf{Q} \quad . \quad (15.18)$$

In Gl.(15.18) ist der Term $\mathbf{B}\mathbf{u}_k$ exakt bekannt und hat daher eine Kovarianz von null. Die Kovarianz $\sigma^2[\hat{\mathbf{x}}_k^+] = \mathbf{P}_k^+$ ist aus dem vorherigen Zeitschritt als Kovarianz nach der Korrektur bekannt. Ebenso ist die Kovarianz $\sigma^2[\mathbf{w}_k] = \mathbf{Q}$ über Gl.(15.18) gegeben.

Mit der Rechenregel Gl.(15.9) und der Linearität des Erwartungswertes ergibt sich nun für die Kovarianz des Zustands die Formel

$$\mathbf{P}_{k+1}^- = \mathbf{A}\mathbf{P}_k^+\mathbf{A}^T + \mathbf{Q} \quad . \quad (15.19)$$

Hiermit kann der Algorithmus des Kalmanfilters abschließend angegeben werden:

Algorithmus des Kalmanfilters

Ein Kalmanfilter kann wie folgt algorithmisiert werden:

- 0) Wähle eine Startschätzung $\hat{\mathbf{x}}_0^+$, \mathbf{P}_0^+ , Gewichte \mathbf{R} und \mathbf{Q} und setze den Index $k = 0$
- 1) Führe eine Zustandsprädiktion durch: $\hat{\mathbf{x}}_{k+1}^- = \mathbf{A}\hat{\mathbf{x}}_k^+ + \mathbf{B}\mathbf{u}_k$
- 2) Prädiziere die Kovarianz des Zustands: $\mathbf{P}_{k+1}^- = \mathbf{A}\mathbf{P}_k^+\mathbf{A}^T + \mathbf{Q}$
- 3) Bestimme den vorhergesagten Ausgang: $\hat{\mathbf{y}}_{k+1} = \mathbf{C}\hat{\mathbf{x}}_{k+1}^-$
- 4) Setze die least-square Gewichtung auf die Blockmatrix mit \mathbf{R} und \mathbf{P}_{k+1}^-
- 5) Erhalte die Messung zum Zeitschritt $k + 1$: \mathbf{y}_{k+1}

- 6) Bestimme \mathbf{L} über die Pseudoinverse nach Gl.(15.7) und Schritt 4)
- 7) Korrigiere die Schätzung: $\hat{\mathbf{x}}_{k+1}^+ = \hat{\mathbf{x}}_{k+1}^- + \mathbf{L}(\mathbf{y}_{k+1} - \hat{\mathbf{y}}_{k+1})$
- 8) Korrigiere die Kovarianz gemäß Gl.(15.15).
- 9) Setze $k \rightarrow k + 1$ und wiederhole bei 1).

Es gibt hier noch effizientere Formeln, die die konkrete Struktur der Pseudoinversen ausnutzen, um Matrixinversionen zu sparen. Der grundlegenden Algorithmus ist jedoch stets derselbe.

Das Kalmanfilter ist immer dann sinnvoll, wenn eine Minimierung der Varianz ein sinnvolles Ziel der Zustandsschätzung ist und eine Gewichtung der stochastischen Signale gemäß ihrer Varianz sinnvoll ist. Das ist insbesondere bei normalverteiltem Modell- und Messrauschen der Fall. Solches Rauschen wird durch Erwartungswert und Kovarianz vollständig charakterisiert. Man kann beweisen, dass für solches Rauschen das Kalmanfilter das bestmögliche Filter ist und hier auch durch nichtlineare Filter keine Verbesserung erzielt werden kann [52].

Es kann aber auch Rauschprozesse geben, bei denen die Varianz nicht genügend Informationen über die Wahrscheinlichkeitsverteilung beinhaltet. In solchen Fällen muss auf alternative Algorithmen, wie sie in Abschnitt 15.4 angesprochen werden, ausgewichen werden.

15.3 Auslegung und Beispiel

Ein großer Vorteil des Kalmanfilters ist seine zugängliche Parametrierung. Die wählbaren Einstellparameter sind dabei:

- $\hat{\mathbf{x}}_0^+$: Die Startschätzung. Hier nimmt man üblicherweise die bestmögliche Schätzung, die für den Systemzustand bekannt sind. Auch wenn dieser Wert nicht immer leicht zu bestimmen ist, so ist doch klar, was zu tun ist und dieser Parameter ist sehr intuitiv.
- \mathbf{P}_0^+ : Die Kovarianz der Startschätzung. Hier sollte man angeben, wie sicher man sich mit der Startschätzung ist. Ist man sich mit bestimmten Einträgen x_i des Zustandes unsicher, so sollte P_{ii} mit einem entsprechend großen Eintrag versehen werden. Nulleinträge auf der Diagonalen von \mathbf{P} sollten in jedem Fall vermieden werden. Eine Standard-Parametrierung verwendet ein Vielfaches der Einheitsmatrix. Es sollte jedoch auch auf die Einheiten innerhalb des Zustandes

σ geachtet werden.

- **R:** Die Kovarianzmatrix des Messrauschens. Diese kann meist mithilfe eines genaueren Sensors, wie diese typischerweise auch zur Kalibrierung eingesetzt werden, ermittelt werden. Hierzu vergleicht man die hochgenauen Messdaten mit denen des vorliegenden Sensors und bestimmt die Varianz des Messfehlers. **R** wird für viele Sensoren in den Datenblättern angegeben.
- **Q:** Die Kovarianz des Modellrauschens. Dieser Parameter ist typischerweise der am schwierigsten einzustellende. **Q** beschreibt, wie unsicher das Modell in den einzelnen Zuständen ist. Werden beispielsweise bei der Modellierung von x_2 größere Vereinfachungen getroffen und linearisiert, dann ist Q_{22} mit einem größeren Gewicht zu versehen. Dies zu quantifizieren, ist aber oftmals sehr schwierig und bedarf verschiedener Iterationen. Ein typischer erster Versuch ist, $\mathbf{Q} = \lambda \mathbf{I}$ als ein Vielfaches der Einheitsmatrix zu setzen, wobei λ in Relation zu **R** gesetzt wird. Hierdurch zerfällt die Parametrierung auf einen einzigen Wert λ , der angibt, ob das Kalmanfilter tendenziell eher den Messungen oder den Modellvorhersagen vertrauen soll.

Die Einstellparameter sind bis auf **Q** relativ intuitiv zu wählen und zu bestimmen. Der Wahl von **Q** nähert man sich üblicherweise in wiederholten Simulationen und Versuchen, bis eine gewünschte Leistungsfähigkeit erreicht wurde.

Als Beispiel für den Einsatz eines Kalmanfilters wird ein Fahrzeug betrachtet, dessen 2D-Position verfolgt werden soll – siehe Bild 15-2. Der verwendete Sensor kann vom Fahrzeug dabei die Position in x_1 -Richtung nur sehr ungenau, in x_2 -Richtung allerdings sehr genau bestimmen. Dies zeigt sich in der (dimensionslosen) Kovarianzmatrix

$$\mathbf{R} = \begin{bmatrix} 2 & 0 \\ 0 & 0,01 \end{bmatrix} \quad . \tag{15.20}$$

Das Fahrzeug selbst bewegt sich mit einem nicht exakt bekannten Lenkwinkel ϑ und betragsmäßig konstanter Geschwindigkeit v .

In Bild 15-3 ist die echte abgetastete Trajektorie des Fahrzeugs und die gemessene Trajektorie dargestellt. Hierbei wird die Kovarianz als Sicherheit der Schätzung in Form einer gestrichelten Ellipse eingezeichnet. Die Länge der Ellipse gibt dabei die Standardabweichung an, so dass sich innerhalb

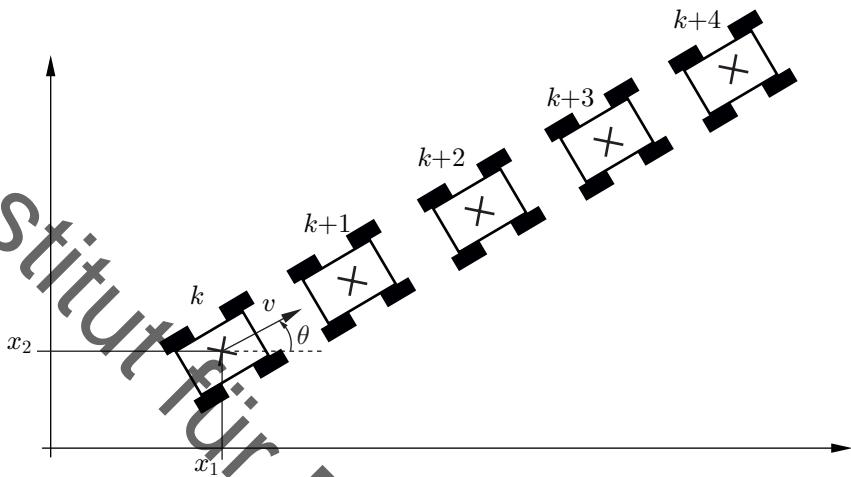


Bild 15-2: Bewegung eines einfachen Fahrzeugs über mehrere Zeitschritte

der Ellipse im statistischen Mittel etwa 68 % der Werte befinden werden.

Man sieht, dass die gemessene Trajektorie wenig mit der tatsächlichen zu tun hat und vor allem zwischen $k+3$ und $k+4$ unphysikalische Fahrmanöver zeigt. Für eine Regelung – wie beispielsweise autonomes Fahren – ist diese Form der Messung unbrauchbar.

Um dieses Problem zu beheben, wird ein Kalmanfilter eingesetzt. Zunächst wird ein zeitdiskretes Modell des Prozesses hergeleitet. Dabei wird das Fahrzeug zeitkontinuierlich als je ein Integrator in x_1 - und x_2 -Richtung modelliert. Die Geschwindigkeit ist der Eingang und die zwei Zustände x_1 und x_2 sind die gleichnamigen Koordinaten. Überführt man das Modell mit der Transitionsmatrix (Abschnitt 14.4.2) in ein zeitdiskretes Modell, so erhält man

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} T \cos(\vartheta) \\ T \sin(\vartheta) \end{bmatrix}, \mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \mathbf{d} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (15.21)$$

Im Vektor \mathbf{b} wurde angenommen, dass der Lenkwinkel ϑ exakt bekannt ist. Da dieser einer Unsicherheit unterliegt, wird \mathbf{b} in Wirklichkeit leicht anders sein. Diese Abweichung – obgleich sie systematisch und kein Zufallsprozess ist – kann im Kalmanfilter näherungsweise über das Modellrauschen

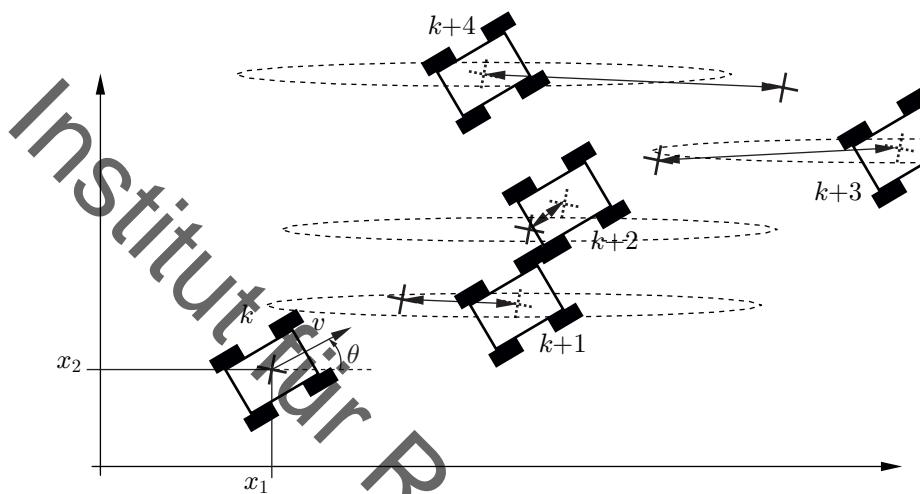


Bild 15-3: Echte und gemessene Trajektorie des Fahrzeugs

\mathbf{Q} beschrieben werden. Da wegen $\cos^2 + \sin^2 = 1$ durch das unbekannte ϑ hervorgerufene Abweichungen in x_1 - und x_2 -Richtung in der Kovarianz etwa gleich groß ausfallen werden, wird $\mathbf{Q} = \lambda \mathbf{I}$ gewählt, wobei λ ein Einstellparameter ist.

Da für die Startschätzung ${}_0\hat{\mathbf{x}} = [0 \ 0]^T$ keine Informationen vorliegen, wird eine Startvarianz von ${}_0\mathbf{P} = \mathbf{I}$ gewählt. Der tatsächliche Startzustand ${}_0\hat{\mathbf{x}} = [1 \ -1]^T$ liegt somit außerhalb der Standardabweichung σ .

Der Ablauf des ersten Schritts des Kalmanfilters ist für $\lambda = 0,2$ in Bild 15-4 gezeigt. Ausgehend von der Startschätzung ${}_0\hat{\mathbf{x}}$ wird mit dem Modell eine Prognose für die zu erwartende Position $\hat{\mathbf{x}}^-$ abgegeben. Hierbei werden auch die Kovarianzen prädiziert, die auf die übliche Weise eingezeichnet sind. Dann erhält das Filter die Messung \mathbf{y} , dessen Standardabweichung ebenfalls eingezeichnet ist und deren Wert erheblich von der Prädiktion abweicht. Das Kalmanfilter korrigiert daraufhin die Prognose zu $\hat{\mathbf{x}}^+$. Hierbei übernimmt es fast vollständig die Messung in x_2 -Richtung, da diese Messung sehr genau, die Prädiktion aber ungenau ist. Anders verhält es sich mit der Position in

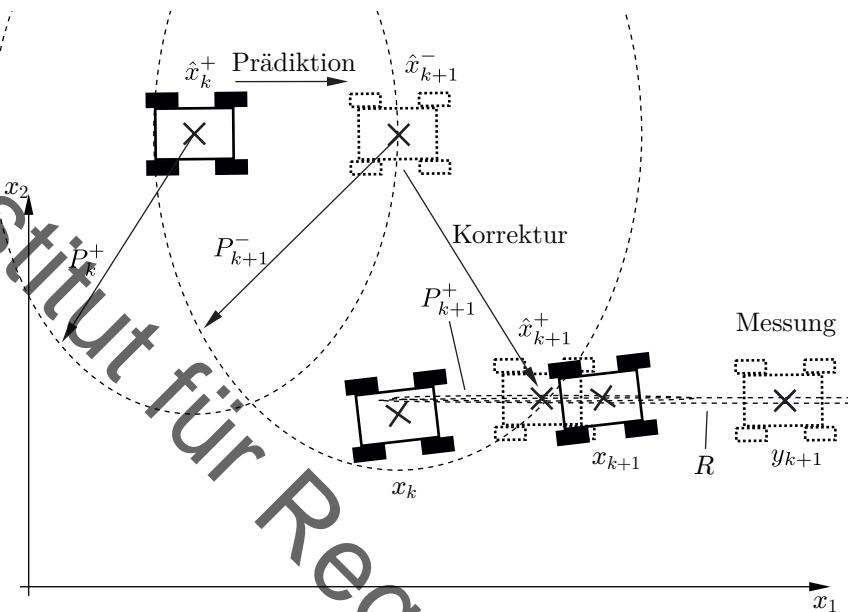


Bild 15-4: Visualisierung eines Kalmanschritts für die Positionsschätzung eines Fahrzeugs

x_1 -Richtung, wo beide Komponenten (Messung und Prädiktion) unsicher sind und daher ein optimaler mittlerer Wert zwischen beiden gewählt wird. Das starke Rauschen in der x_1 -Komponente wird so reduziert.

Beim resultierenden korrigierten Zustand ist sich (aufgrund der guten Messung) das Filter mit der x_2 -Richtung sehr sicher, was sich auch in der korrigierten Kovarianzmatrix (ebenfalls als Ellipse gezeichnet) zeigt. In x_1 -Richtung hingegen ist sich das Filter noch unsicher, da Messung und Prädiktion bisher nicht zusammenpassten.

Die rekursive Ausführung des in Bild 15-4 gezeigten Kalmanschritts führt zu einer geglätteten Trajektorie des Fahrzeuges. Diese ist in Bild 15-5 gezeigt. Man sieht deutlich, dass die resultierende Trajektorie durch den Modellabgleich ein sehr viel plausibleres Fahrverhalten darstellt. Trotz der ungenauen Startschätzung (die außerhalb von Bild 15-5 liegt) ist bereits die erste

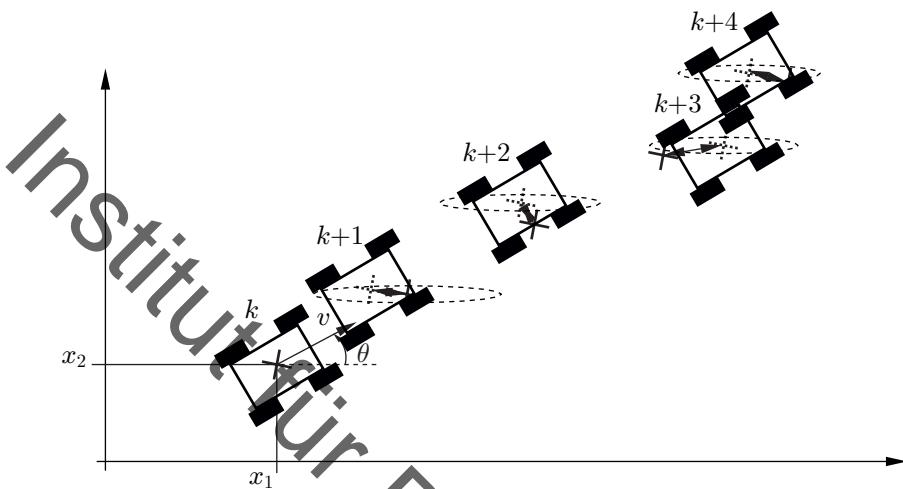


Bild 15-5: Echte und geschätzte Trajektorie des Fahrzeugs

Schätzung nahe an der tatsächlichen Trajektorie. Somit bietet das Kalmanfilter sowohl eine Lösung zur Glättung von Messsignalen als auch zur schnellen Rekonstruktion unbekannter Systemzustände.

15.4 Limitierungen und Erweiterungen

Das vorgestellte Kalmanfilter ist immer dann eine gute Beobachterwahl, wenn – gemäß Herleitung – das least-square Problem in Gl.(15.5) die Aufgabenstellung des Filters gut beschreibt. Außerdem muss es sinnvoll sein, die Gewichtung des least-square anhand der Varianzen vorzunehmen. Diese beiden Annahmen sind in vielen Anwendungsfällen zumindest näherungsweise erfüllt – insbesondere für lineare Systeme mit normalverteiltem Rauschen, wo es sich beim Kalmanfilter um den Schätzer mit der minimalen Varianz handelt. Es gibt jedoch auch Fälle, wo die beschriebenen Voraussetzungen nicht erfüllt sind und der Standard-Algorithmus des Kalmanfilters erweitert werden muss.

Hierzu wird das least-square in Gl.(15.5) mit der Gewichtung des Kalman-

filters betrachtet:

$$\min_{\hat{x}^+} \left(\underbrace{\|\hat{x}^+ - \hat{x}^- \|_{P^{-1}}^2}_{\text{Prädiktion}} + \underbrace{\|\mathbf{C}\hat{x}^+ - \mathbf{y}\|_{R^{-1}}^2}_{\text{Korrektur}} \right) . \quad (15.22)$$

Zunächst sollen nicht die Gewichte, sondern nur die beiden Terme des Optimierungsproblems analysiert werden.

Der erste Term beschreibt, dass der korrigierte Zustand von der Prädiktion wenig abweichen soll. Dieser Term ist stets sinnvoll und kann ausgewertet werden, sobald irgendein Modell vorliegt, über welches \hat{x}^- bestimmt werden kann. Nichtlinearitäten im Modell sind hier unproblematisch, da es sich um eine reine Vorwärtssimulation handelt.

Der zweite Term beschreibt, dass der zum korrigierten Zustand gehörige Ausgang zur Messgröße passen soll. Auch dies ist als Forderung immer sinnvoll. Allerdings setzt die Schreibweise in Gl.(15.22) voraus, dass die Ausgangsgleichung $\mathbf{y} = \mathbf{Cx}$ linear ist. Wäre die Ausgangsgleichung nichtlinear mit $\mathbf{y} = \mathbf{h}(\mathbf{x})$, so ergäbe sich

$$\min_{\hat{x}^+} \left(\|\hat{x}^+ - \hat{x}^- \|_{P^{-1}}^2 + \|\mathbf{h}(\hat{x}^+) - \mathbf{y}\|_{R^{-1}}^2 \right) , \quad (15.23)$$

was kein lineares Ausgleichsproblem mehr ist, da \hat{x}^+ über \mathbf{h} nichtlinear in die Optimierung mit eingeht.

Die Lösung besteht nun darin, die Ausgangsgleichung zu linearisieren. Hierzu muss ein Linearisierungspunkt gewählt werden, der möglichst nah an \hat{x}^+ liegt, um die Linearisierungsfehler gering zu halten. Die passende Wahl ist \hat{x}^- als Prädiktion des Zustandes. Es ergibt sich für den Korrekturterm

$$\begin{aligned} \|\mathbf{h}(\hat{x}^+) - \mathbf{y}\|_{R^{-1}}^2 &\approx \|\mathbf{h}(\hat{x}^-) + \mathbf{C}(\hat{x}^+ - \hat{x}^-) - \mathbf{y}\|_{R^{-1}}^2 \\ \text{mit } \mathbf{C} &= \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \Big|_{\mathbf{x}=\hat{x}^-} , \end{aligned} \quad (15.24)$$

was auf ein lineares Ausgleichsproblem in \hat{x}^+ führt.

Als Unterschied muss in jedem Zeitschritt linearisiert werden. Das führt aber nur zu einem marginal größeren Rechenaufwand, da das Gewicht \mathbf{W}

ohnehin sich bereits mit jedem Zeitschritt ändert und daher die Pseudoinversen nicht offline im Voraus berechnet werden können.

Die beschriebene Modifikation des Kalmanfilters wirkt sich üblicherweise nur moderat nachteilig auf die Leistungsfähigkeit aus, sodass in vielen Fällen mit einem EKF für ein nichtlineares System vergleichbare Resultate wie mit einem Kalmanfilter für ein lineares System erzielt werden können. Abweichungen gibt es vor allem in der Konvergenzphase des Filters, wenn die Zustände möglicherweise stärker korrigiert werden müssen. Sobald der Zustand konvergiert ist, sind die Linearisierungsfehler aber so klein, dass die Nichtlinearität der Ausgangsgleichung kein wesentliches Problem darstellt. Ausnahmen bilden enorm nichtlineare Ausgangsgleichungen mit starken Steigungswechseln oder fehlender Differenzierbarkeit.

Die beiden Prädiktions- und Korrekturterme des Kalmanfilters besitzen also eine relativ breite Gültigkeit. Etwas anders verhält es sich mit der Gewichtung. Betrachtet man die Gewichtung der Korrektur \mathbf{R}^{-1} , so wird die inverse Kovarianzmatrix genutzt, um die quadratische Abweichung passend zu gewichten. Das ist nur dann sinnvoll, wenn die Kovarianz ein sinnvolles Maß zur Beschreibung der Vertrauenswürdigkeit des Messwertes darstellt.

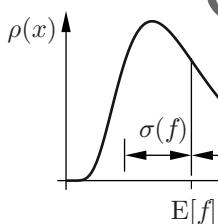


Bild 15-6: Nicht normalverteiltes Rauschen mit Varianz und Erwartungswert

Als Gegenbeispiel wird die Wahrscheinlichkeitsverteilung $\rho(x)$ des Sensors in Bild 15-6 betrachtet. Das Rauschen, mit dem der Sensor belegt ist, ist sehr unsymmetrisch und der Erwartungswert entspricht nicht dem Maximum. Die Kovarianz und die quadratische Norm sind allerdings symmetrisch in dem Sinne, dass sie Abweichungen, die zu klein oder zu groß sind, gleich gewichten. Eine solche Gewichtung über die Kovarianz verliert aber die Informationen über eine mögliche „Vorzugsseite“ des Rauschprozesses.

Das Beispiel in Bild 15-6 mag etwas konstruiert wirken und tatsächlich

weisen die meisten Sensoren ein Rauschen ähnlich einer Normalverteilung auf. Die beschriebene Problematik einer Nicht-Normalverteilung ergibt sich jedoch regelmäßig an einer anderen Stelle im Kalmanfilter, und zwar bei der Wahrscheinlichkeitsverteilung des geschätzten Zustandes \hat{x} .

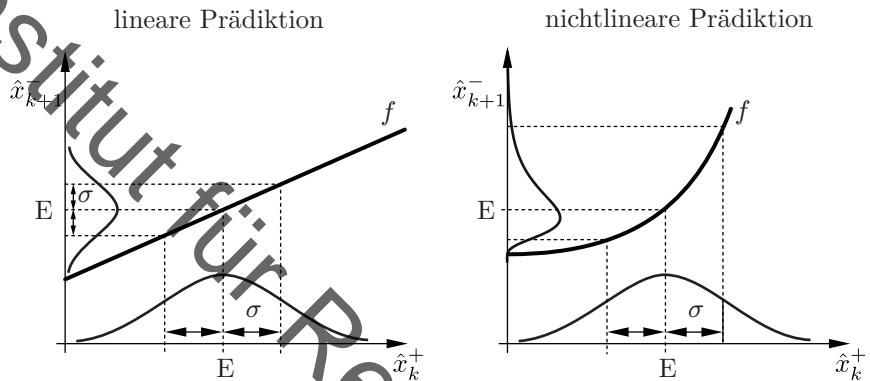


Bild 15-7: Verlust der Normalverteilung bei nichtlinearer Prädiktion

Um dies zu erläutern wird angenommen, dass der Zustand \hat{x}_k^+ normalverteilt sei mit der Kovarianz P_k^+ . Die Prädiktion von \hat{x}_{k+1}^- wird nun einmal mit einem linearen und einmal mit einem nichtlinearen Modell f durchgeführt. Beide Formen der Prädiktion sind für den eindimensionalen Fall in Bild 15-7 gegenübergestellt.

Es ist gut zu erkennen, dass die Normalverteilung im linearen Fall erhalten bleibt. Die Standardabweichung ändert sich zwar mit der Steigung der Funktion f – dies wird aber im Kalmanfilter in der Prädiktion P_{k+1}^- berücksichtigt.

Im nichtlinearen Fall weist aber (trotz der Normalverteilung des Eingangs \hat{x}_k^+) die Prädiktion \hat{x}_{k+1}^- im Allgemeinen keine Normalverteilung mehr auf. Je nach Nichtlinearität der Funktion f können dabei leicht Situationen wie in Bild 15-6 entstehen.

Bei einem nichtlinearen Prädiktionsmodell kann folglich der Prädiktionsterm $\|\hat{x}^+ - \hat{x}\|^2$ trotz Nichtlinearität aufgestellt und linear ausgewertet werden. Allerdings kann die Gewichtung mit P_{k+1}^- unpassend sein.

Zusätzlich ergibt sich das Problem \mathbf{P}_{k+1}^- zu berechnen, da die Berechnungsformel Gl.(15.19)

$$\mathbf{P}_{k+1}^- = \mathbf{A}\mathbf{P}_k^+\mathbf{A}^T + \mathbf{Q}$$

nur für den linearen Fall Gültigkeit besitzt.

Die einfache Lösung ist es erneut, zu linearisieren und Gl.(15.19) mit der Linearisierung im Arbeitspunkt auszuwerten. Der Linearisierungspunkt ist allerdings nicht ganz klar: die Linearisierung \mathbf{A} soll beschreiben, wie sich die Kovarianz \mathbf{P} im Laufe der Prädiktion von \mathbf{x}_k^+ hin zu \mathbf{x}_{k+1}^- verändert. Somit kommen diese beiden Punkte als Arbeitspunkte für die Linearisierung in Frage. Allerdings wird die unkorrigierte Schätzung \mathbf{x}_{k+1}^- in den allermeisten Fällen einer größeren Unsicherheit unterliegen und der entstehende Linearisierungsfehler lässt sich daher schlechter kontrollieren. Daher wird im Standardansatz um den Punkt \mathbf{x}_k^+ linearisiert, um \mathbf{A} zu erhalten.

Algorithmus des Erweiterten Kalmanfilters

Ein Kalmanfilter, welches nichtlineare Systeme durch eine Linearisierung an den notwendigen Stellen behandelt, wird *Erweitertes Kalmanfilter* (englisch: Extended Kalman filter, EKF) genannt. Es kann wie folgt algorithmisiert werden:

- 0) Wähle eine Startschätzung $\hat{\mathbf{x}}_0^+$, \mathbf{P}_0^+ , Gewichte \mathbf{R} und \mathbf{Q} und setze den Index $k = 0$
- 1) Führe eine nichtlineare Zustandsprädiktion durch: $\hat{\mathbf{x}}_{k+1}^- = \mathbf{f}(\hat{\mathbf{x}}_k^+, \mathbf{u}_k)$
- 2) Linearisiere um den Punkt $\hat{\mathbf{x}}_k^+$ und erhalte \mathbf{A}_k .
- 3) Prädiziere die Kovarianz des Zustands mit dem linearisierten Modell: $\mathbf{P}_{k+1}^- = \mathbf{A}_k\mathbf{P}_k^+\mathbf{A}_k^T + \mathbf{Q}$
- 4) Bestimme den vorhergesagten Ausgang: $\hat{\mathbf{y}}_{k+1} = \mathbf{h}(\hat{\mathbf{x}}_{k+1}^-)$
- 5) Setze die least-square Gewichtung auf die Blockmatrix mit \mathbf{R} und \mathbf{P}_{k+1}^-
- 6) Erhalte die Messung zum Zeitschritt $k + 1$: \mathbf{y}_{k+1}
- 7) Linearisiere um den Punkt $\hat{\mathbf{x}}_{k+1}^-$ und erhalte \mathbf{C}_{k+1} .
- 8) Bestimme \mathbf{L} über eine Pseudoinverse gemäß Gl.(15.7) und Schritt 5) und 7)

- 9) Korrigiere die Schätzung: $\hat{x}_{k+1}^+ = \hat{x}_{k+1}^- + \mathbf{L}(\mathbf{y}_{k+1} - \hat{\mathbf{y}}_{k+1})$
- 10) Korrigiere die Kovarianz gemäß Gl.(15.15).
- 11) Setze $k \rightarrow k + 1$ und wiederhole bei 1).

Wichtig ist hier zu beachten, dass ein EKF nicht einfach das ganze System linearisiert. Sowohl Schritt 1) als auch Schritt 4) erfolgen anhand der nichtlinearen Gleichungen, die gleichzeitig den Arbeitspunkt definieren, um welchem linearisiert wird. Die Linearisierung erfolgt dann entlang der prädictierten Trajektorie zur Bestimmung der Kovarianz und der lokalen Abweichungen durch den Korrekturschritt.

Das EKF behandelt nichtlineare Systeme konsistent innerhalb der Struktur des Kalmanfilters. Der Algorithmus muss dabei nur marginal verändert werden. Bei eher schwach ausgeprägten Nichtlinearitäten liefert das EKF – trotz der systematischen Fehler wie sie in Bild 15-6 aufgezeigt wurden – erstaunlich gute Ergebnisse.

Daher gehört das EKF zu den beliebtesten Filtern und Beobachtern und Weiterentwicklungen müssen sich meist einem EKF als Standardimplementierung in einem Benchmarktest stellen. Derartige Weiterentwicklungen setzen üblicherweise an den Kovarianzmatrizen an und versuchen, die Wahrscheinlichkeitsverteilung nicht über nur zwei Kennwerte (Erwartungswert und Standardabweichung), sondern genauer zu beschreiben [52].

Die *Partikelfilter* nutzen hierfür eine hohe Anzahl an Partikeln als „Zustandshypothesen“. Die Veränderung dieser Partikel durch die Abbildung \mathbf{f} wird dann einzeln berechnet und so eine Verteilung der Prädiktionshypothesen gewonnen. Diese kann dann für eine Optimierung genutzt werden, die prinzipiell beliebige Rauschprozesse berücksichtigt.

Nachteilig ist die stark erhöhte Rechenzeit, die mit der Anzahl der verwendeten Partikel steigt. Hier setzt das *Uncented Kalmanfilter* (UKF) an, in welchem eine möglichst geringe Anzahl an Partikeln bezüglich einer Normalverteilung optimal platziert wird. Diese Algorithmen sollten dann eingesetzt werden, wenn genügend Rechenleistung zu Verfügung steht und das EKF aufgrund zu starker Nichtlinearitäten oder merkwürdiger Rauschprozesse an seine Grenzen kommt.

16 Nichtlineare Systeme

16.1 Phasenportraits

Bisher wurden fast ausschließlich lineare Regelungssysteme betrachtet. So weit Nichtlinearitäten eine Rolle spielten, wurde davon ausgegangen, dass ein lineares Ersatzsystem die zu untersuchenden Zusammenhänge genügend genau beschreibt. Ein wesentlicher Grund für diese Vorgehensweise liegt darin, dass es zur Analyse und Synthese linearer Systeme viele gut handhabbare Methoden gibt. Außerdem zeigt das Linearisierungstheorem, dass in einer lokalen Umgebung des stationären Arbeitspunktes die lineare Näherung die wesentlichen dynamischen Eigenschaften des nichtlinearen Systems abbilden kann. Diese Näherung gilt aber nur in einer Umgebung um die Ruhelage.

Global vs. lokal

Wird ein nichtlineares System in der Umgebung seiner Ruhelage über die Linearisierung analysiert, so können ausschließlich *lokale* Eigenschaften, die nur in einer gewissen Umgebung gültig sind, nachgewiesen werden.

Das Verhalten des Systems im gesamten Definitionsbereich der Variablen wird *globales* Verhalten genannt.

Oft muss ein nichtlineares System in seinem globalen Verhalten untersucht werden. Diese Situation tritt beispielsweise bei Folgeregelungen auf, bei denen durch geänderte Führungsgrößen der bisherige Arbeitspunkt verlassen wird. Eine globale Analyse nichtlinearer Systeme bedarf dabei komplexerer Verfahren, welche sich in geringerem Maße systematisieren lassen als die für lineare Systeme.

Keines der bisher genutzten Analysewerkzeuge lässt sich für nichtlineare Systeme nutzen. Der Hauptgrund hierfür ist, dass das in Kapitel 2 eingeführte Superpositionsprinzip im Kontext nichtlinearer Systeme seine Gültigkeit verliert. Hierdurch enthalten Sprungantworten und andere Zeitverläufe nicht die vollständigen Systeminformationen, da die Sprunghöhe ebenfalls das dynamische Verhalten beeinflusst. Auch der über die Laplace-Transformation abgeleitete Bildbereich ist nicht weiter anwendbar, da die Laplace-Transformation nur für lineare Systeme definiert ist. Es gibt zudem

keine Systemmatrix \mathbf{A} , deren Eigenwerte Aufschluss geben.

Folglich müssen neue Analysewerkzeuge gesucht werden, wenn globale Eigenschaften nichtlinearer Systeme analysiert werden sollen. Zweckmäßigerweise beschränkt man sich hierbei auf autonome zeitkontinuierliche Systeme in Zustandsraumdarstellung. An erster Stelle der Verfahren ist das *Phasenportrait* zu nennen.

Phasenportrait

Gegeben ist ein autonomes nichtlineares System $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ mit $\mathbf{x}(0) = {}_0\mathbf{x}$ und $\mathbf{x} \in \mathbb{R}^n$. Dann ergibt sich für jedes ${}_0\mathbf{x}$ eine andere Lösungstrajektorie im \mathbb{R}^n . Stellt man diese als Kurve im Zustandsraums mit dem Laufparameter t dar, so erhält man die zugehörige *Phasenlinie*. Das *Phasenportrait* ist die graphische Darstellung dieser Phasenlinien, wobei die Anfangszustände der gezeigten Phasenlinien so gewählt werden, dass sie die relevante Dynamik gut abdecken.

Im Falle von Ruhelagen verkommen die Phasenlinien zu Punkten im Zustandsraum. Zu beachten ist, dass die Zeit t nicht explizit im Phasenportrait auftaucht, sondern einen Laufparameter für jede einzelne Phasenlinie bildet, die man oft mit einem Pfeil in Richtung wachsender t markiert.

Zwei beliebige Phasenlinien schneiden sich nicht, da die Lösung der Differentialgleichung eindeutig ist. Zusätzlich können die Phasenlinien um das vollständige Gradientenfeld erweitert werden. Das Phasenportrait kann sowohl numerisch als auch stückweise analytisch konstruiert werden.

Phasenportraits eignen sich insbesondere für die Analyse von Systemen zweiter Ordnung, da dort der Zustandsraum eine Ebene ist und die graphische Darstellung somit leichter zu deuten ist. Für ein autonomes System zweiter Ordnung lässt sich die Zustandsgleichung in zwei Differentialgleichungen aufteilen

$$\dot{x}_1 = f_1(x_1, x_2)$$

$$\dot{x}_2 = f_2(x_1, x_2)$$

(16.1)

Die Steigung m einer Phasenlinie im Punkt (x_1^*, x_2^*) ist dann durch

$$m = \left. \frac{dx_2}{dx_1} \right|_{\mathbf{x}=(x_1^*, x_2^*)^\text{T}} = \frac{f_2(x_1^*, x_2^*)}{f_1(x_1^*, x_2^*)} \quad (16.2)$$

gegeben.

Phasenportraits sind ebenfalls für $n \neq 2$ möglich, allerdings aufgrund der graphisch schwierigeren Darstellung oft weniger zweckmäßig. Bild 16-1 zeigt das Phasenportrait des bereits mehrfach behandelten reibungsbehafteten Pendels (siehe Bild 3-1 und Gl.(3.10)) mit dem Auslenkungswinkel x_1 und der Winkelgeschwindigkeit x_2 . Im Phasenportrait ist das Gradientenfeld sowie vier Phasenlinien dargestellt, welche zu Systemantworten bei Anfangszuständen von $\mathbf{x}(t_0) = (\pm \frac{\pi}{2}, 0)^T$ und $\mathbf{x}(t_0) = (0, \pm \frac{\pi}{2})^T$ gehören.

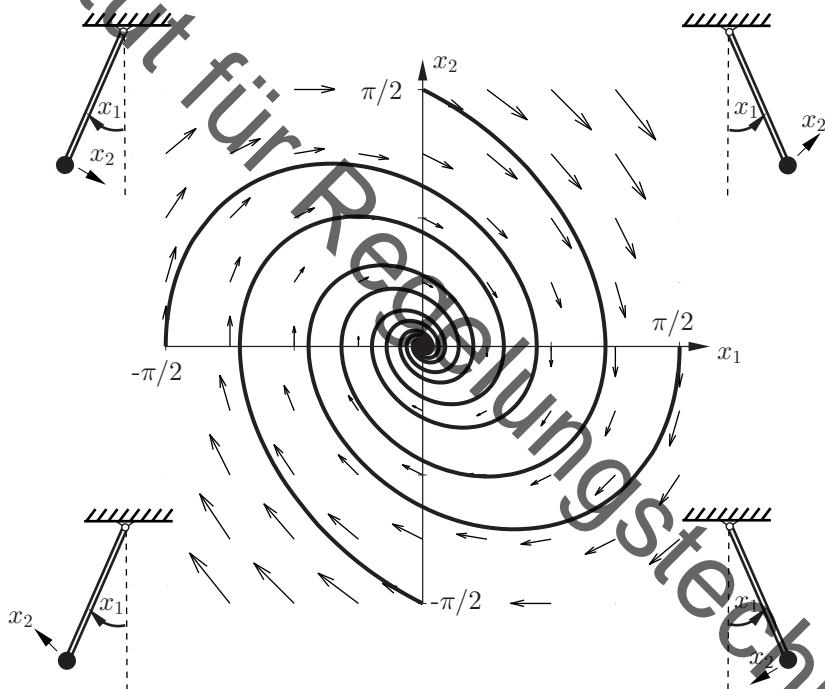


Bild 16-1: Phasenportrait des mathematischen Pendels

Wie an den eingezeichneten Phasenlinien zu erkennen ist, konvergiert das System für die exemplarischen Anfangszustände erwartungsgemäß gegen die untere Ruhelage. Das Gradientenfeld führt auf immer engeren Bahnen in den Ursprung.

Aus dem Linearisierungstheorem ist bekannt, dass ein nichtlineares System um einen Arbeitspunkt herum durch ein lineares System angenähert werden kann. Daher entsprechen die Phasenportraits nichtlinearer Systeme lokal um eine Ruhelage den Phasenportraits linearer Systeme. Folglich ist es zielführend, häufig auftretende Phasenbilder linearer Systeme in unterschiedliche Klassen zu unterteilen, die in Tab. 16-1 gegeben sind.

Die Linearisierung des Pendelsystems um die untere Ruhelage resultiert in einer Systemmatrix mit zwei komplex-konjugierten Eigenwerten mit negativem Realteil, womit die lokale Übereinstimmung des Phasenportraits mit einem *stabilen Strudelpunkt* folgt. Dieser lässt sich ebenfalls in Bild 16-1 eindeutig identifizieren.

Offensichtlich hängen die Eigenwerte des (lokal äquivalenten) linearen Systems mit der Laufrichtung der Phasenlinien zusammen: Während Phasenlinien im Falle stabiler Ruhelagen zum Gleichgewichtspunkt führen, ist eine abstoßende Wirkung für instabile Ruhelagen zu erkennen. Komplexe Eigenwerte führen zu Schwingungen und damit zu Strudelpunkten. Eine genauere Betrachtung der Phasenlinien zeigt zudem den Einfluss der zugehörigen Eigenvektoren, die für Knoten- und Sattelpunkte als Kontraktionsrichtungen auftreten.

16.2 Einzugsbereich

Das Phasenportrait ermöglicht es, die Systemtrajektorien in einem großen Bereich des Zustandsraum zu visualisieren. Dies erlaubt es, sich nicht nur auf die Umgebung einer Ruhelage zu beschränken, sondern auch globale Eigenschaften des System zu detektieren.

Als Beispiel zeigt Bild 16-2 ein Phasenportrait mit drei Ruhelagen. Während die Ruhelage im Ursprung einem Sattel gleicht und instabil ist, gleichen die Phasenbilder um die anderen Ruhelagen stabilen Strudelpunkten. Offenbar übertragen sich lokalen Eigenschaften nicht auf den gesamten Zustandsraum.

Offenbar ist es zudem vom Anfangszustand abhängig, ob das gezeigte System in den Endwert in $x = 1$ oder $x = -1$ läuft (die Ruhelage $x = 0$ wird nur für ${}_0\mathbf{x} = \mathbf{0}$ erreicht). Die Menge aller Startzustände, deren Phasenlinien letztlich in eine definierte Ruhelage laufen, ist der sogenannte *Einzugsbereich*.

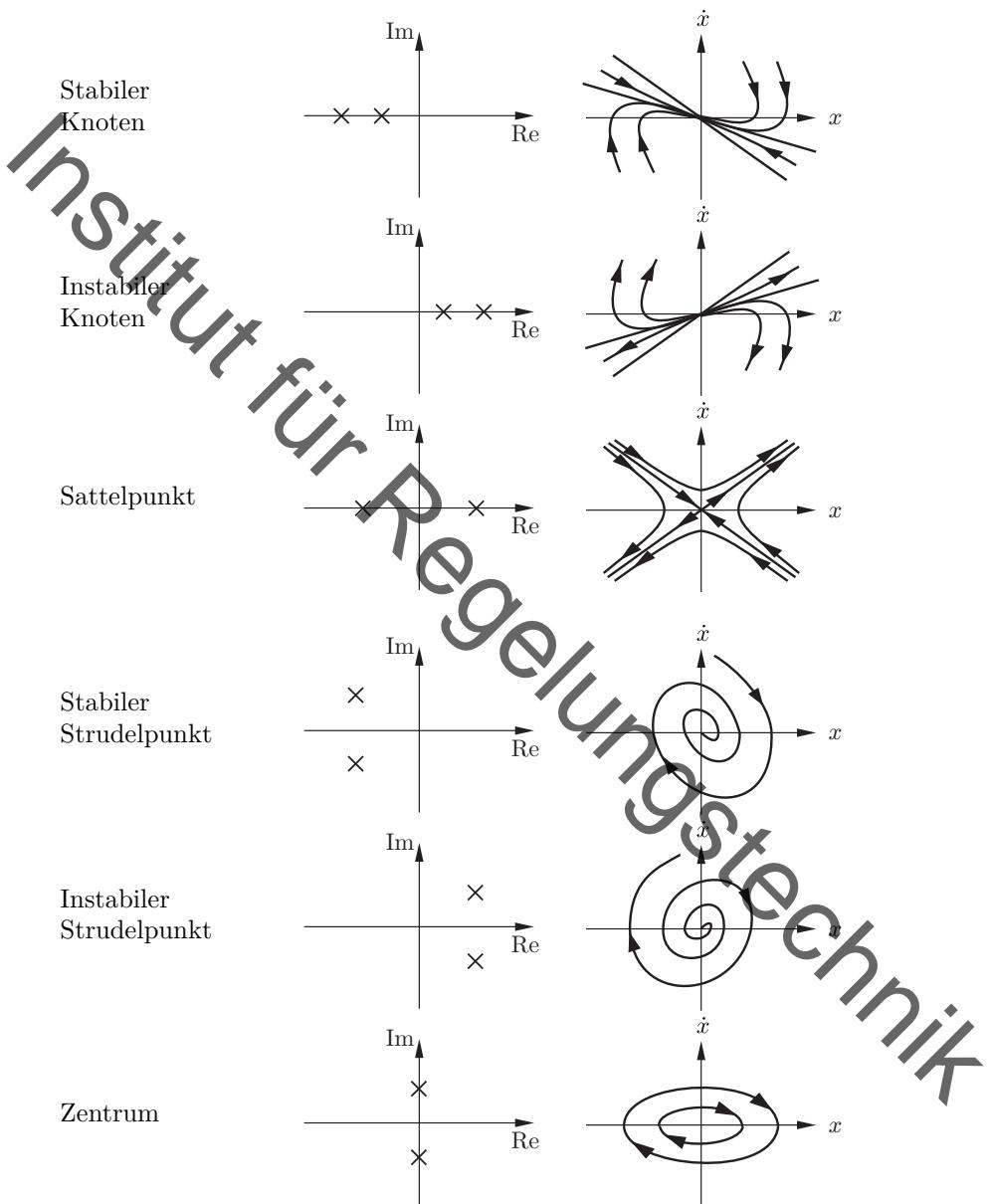


Tabelle 16-1: Phasenbilder linearer Systeme

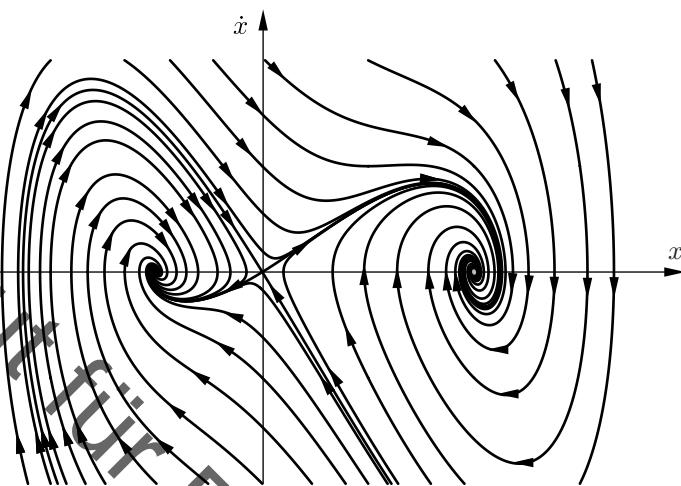


Bild 16-2: Phasenportrait mit drei Ruhelagen

Einzugsbereich

Im Falle einer stabilen Ruhelage \boldsymbol{x}_0 bilden die Anfangszustände $\boldsymbol{x}^*(t_0)$ aller Systemtrajektorien, welche Gl.(3.12) erfüllen, den sogenannten Einzugsbereich \mathcal{E} einer solchen Ruhelage.

$$\mathcal{E}(\boldsymbol{x}_0) = \left\{ \boldsymbol{x}^*(t_0) \in \mathbb{R}^n \mid \lim_{t \rightarrow \infty} \boldsymbol{x}^*(t) = \boldsymbol{x}_0 \right\} \quad (16.3)$$

Im gezeigten Fall zerlegen die beiden Einzugsbereiche der linken und rechten Ruhelage den Zustandsraum \mathbb{R}^n . Das ist nicht immer der Fall, da es Punkte im Zustandsraum geben kann, die in keine Ruhelage konvergieren.

Globale Stabilität

Ist der Einzugsbereich eine nicht weiter spezifizierte Umgebung, so heißt die Ruhelage *lokal stabil*. Entspricht der Einzugsbereich dem ganzen euklidischen Raum $\mathcal{E} = \mathbb{R}^n$, ist die Ruhelage *global stabil*.

Da die Systemtrajektorien mitunter hoch sensitiv gegenüber Änderungen in den Anfangsbedingungen sind (chaotisches Verhalten), fällt die Bestimmung des Einzugsbereichs im Allgemeinen sehr aufwändig aus.

16.3 Lyapunov-Funktionen

16.3.1 Direkte Methode nach Lyapunov

Allgemeine Aussagen über die Stabilität nichtlinearer Systeme, die ggf. auch globale Stabilität umfassen, lassen sich mit dem Stabilitätskriterium nach Lyapunov¹ treffen. Motiviert durch die Beobachtung, dass stabile Gleichgewichtslagen physikalischer Systeme mit (lokalen) Minima der im System enthaltenen Energie zusammenfallen, gilt es eine sogenannte Lyapunov-Kandidatenfunktion $V(\mathbf{x})$ zu finden, welche in der betrachteten Ruhelage minimal ist.

Zudem dissipiert ein physikalischer Prozess auf dem Weg zu seinem energetisch günstigsten Zustand Energie. Kann in Analogie nachgewiesen werden, dass der aufgestellte Kandidat für eine Lyapunov-Funktion mit der Zeit abnimmt, impliziert dies die Stabilität der Ruhelage.

Übertragen auf das Beispielsystem des Pendels wird diesem Energie in Form von Reibung entzogen, bis es in seiner unteren Ruhelage zum Stillstand kommt. Da für jede weitere Bewegung Energie in das System eingeprägt werden müsste, handelt es sich zugleich um den Zustand lokal geringster Energie.

Eine formale Herleitung der obigen Zusammenhänge kann beispielsweise [2] entnommen werden. Eine allgemeingültige Form für nichtlineare Systeme lässt sich dann wie folgt formulieren:

Kandidatenfunktion

Eine Funktion $V(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$ heißt (Lyapunov-)Kandidatenfunktion oder *Kandidat* für eine Lyapunov-Funktion, wenn sie differenzierbar ist und in einer Umgebung $\mathcal{U}(\mathbf{x}_0)$ der Ruhelage \mathbf{x}_0 positiv definit ist, d. h. es gilt:

$$V(\mathbf{x}) > V(\mathbf{x}_0) \quad \forall \mathbf{x} \in \mathcal{U}(\mathbf{x}_0) \setminus \{\mathbf{x}_0\} \quad . \quad (16.4)$$

Kandidatenfunktionen sind mögliche verallgemeinerte Energiefunktionen, die in der Ruhelage ein Minimum aufweisen. Liegt dieses Minimum nicht bei $V(\mathbf{x}_0) = 0$, so wird durch eine einfache Verschiebung $\tilde{V}(\mathbf{x}) = V(\mathbf{x}) - V(\mathbf{x}_0)$ die Bedingung erfüllt.

¹Александр Михайлович Ляпунов(1857-1918), russischer Mathematiker [31]

Das Stabilitätstheorem nach Lyapunov besagt nun, dass eine Ruhelage stabil ist sofern eine Kandidatenfunktion gefunden wird, deren Energie kontinuierlich abnimmt.

Direkte Methode nach Lyapunov

Wenn für das autonome System $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}_0)$ eine Kandidatenfunktion $V(\mathbf{x})$ existiert, für die gilt

$$\text{i)} \quad \dot{V}(\mathbf{x}) \leq 0 \quad \forall \mathbf{x} \in \mathcal{U}(\mathbf{x}_0) \setminus \{\mathbf{x}_0\} \quad (16.5)$$

dann ist diese Ruhelage lokal (asymptotisch) stabil oder grenzstabil. Ist die strikte Ungleichung

$$\text{ii)} \quad \dot{V}(\mathbf{x}) < 0 \quad \forall \mathbf{x} \in \mathcal{U}(\mathbf{x}_0) \setminus \{\mathbf{x}_0\} \quad (16.6)$$

erfüllt, so ist die Ruhelage lokal (asymptotisch) stabil. Gilt zudem

$$\text{iii)} \quad \mathcal{U}(\mathbf{x}_0) = \mathbb{R}^n \setminus \{\mathbf{x}_0\} \quad (16.7)$$

$$\text{iv)} \quad V(\mathbf{x}) \rightarrow \infty \quad \text{für } \|\mathbf{x}\| \rightarrow \infty \quad (16.8)$$

dann ist die Ruhelage global (asymptotisch) stabil.

Für die nach Gl.(16.6) lokale stabile Ruhelage \mathbf{x}_0 ist zu beachten, dass der zuvor eingeführte Einzugsbereich $\mathcal{E}(\mathbf{x}_0)$ ungleich der Umgebung $\mathcal{U}(\mathbf{x}_0)$ sein kann. Tatsächlich gilt im Allgemeinen nicht einmal $\mathcal{E}(\mathbf{x}_0) \subset \mathcal{U}(\mathbf{x}_0)$.

16.3.2 Beispiel und Anwendungshinweise

Die direkte Methode nach Lyapunov sei an einem mathematischen Beispiel verdeutlicht. Betrachtet werde das nichtlineare System

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} x_1(x_2 - 1) \\ x_2(x_1 - 1) \end{bmatrix} \quad . \quad (16.9)$$

Das zugehörige Phasenportrait mit zwei exemplarischen Systemtrajektorien ist in Bild 16-3 dargestellt. Zudem sind die Bereiche, in denen $\frac{d}{dt}V(\mathbf{x}) = \frac{d}{dt}(0,5(x_1^2 + x_2^2)) < 0$ gilt, grau hinterlegt, wobei unterschiedliche Grautöne die Niveaulinien angeben.

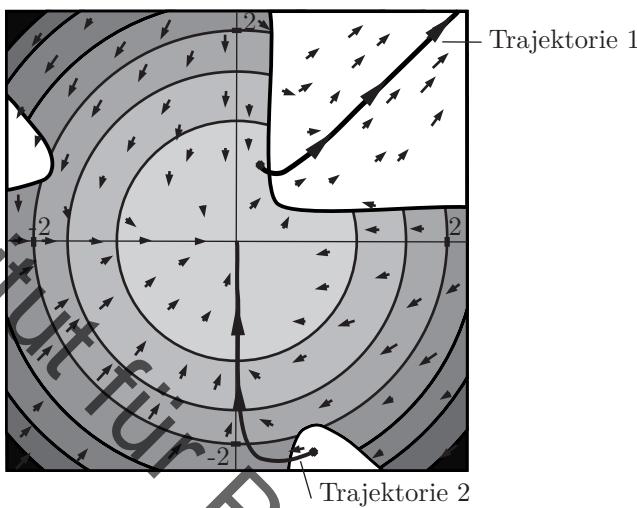


Bild 16-3: Einzugsbereich und Lyapunov-Funktionen

Die Ruhelage $\mathbf{x} = [0 \ 0]^T$ ist lokal asymptotisch stabil. Offensichtlich existieren jedoch Punkte mit $\dot{V} < 0$, für welche die Zustandstrajektorien von der Ruhelage weg führen (Trajektorie 1), da der Bereich, in welchem $\dot{V} < 0$ gilt, verlassen wird. Genauso existieren Punkte außerhalb der Umgebung $\mathcal{U}(\mathbf{x}_0)$, für die sich konvergente Systemlösungen ergeben (Trajektorie 2). Das abgeleitete Kriterium dient daher nur einer Stabilitätsbeurteilung.

Die direkte Methode nach Lyapunov gibt hinreichende Bedingungen für die Stabilität von Ruhelagen und dient daher in dieser Form nicht zum Nachweis von Instabilität. Dies liegt an der freien Wahl der Kandidatenfunktion $V(\mathbf{x})$. Lässt sich für eine spezifische Wahl von $V(\mathbf{x})$ keine Stabilität nachweisen, muss ein anderer Kandidat gesucht werden.

Neben physikalisch motivierten Funktionen, wie einer skalierten Beschreibung der Systemenergie, werden hierbei oft quadratische Formen gemäß

$$V(\mathbf{x}) = \mathbf{x}^T \mathbf{P} \mathbf{x} \quad (16.10)$$

verwendet, welche über die positiv definite Matrix $\mathbf{P} \succ 0$ skaliert werden.

Zwei Beispiele sollen die Methode illustrieren. Für das System $\dot{X} = -X^3$ ergibt die Stabilitätsanalyse über die Linearisierung

$$\dot{x} = -3X_0^2 \cdot x = 0 \cdot x \quad (16.11)$$

mit einem Eigenwert bei Null.

Über die Linearisierung kann also keine Aussage über die Stabilität getroffen werden. Mit der Kandidatenfunktion $V(X) = X^2$, die offensichtlich differenzierbar und positiv definit ist, ergibt sich

$$\dot{V}(X) = \frac{d}{dt}(X^2) = 2X \cdot \dot{X} = 2X \cdot (-X^3) = -2X^4 \quad (16.12)$$

und das ist für alle $X \neq 0$ negativ. Somit ist die Ruhelage $X = 0$ global asymptotisch stabil – ein Resultat, dass weit über die Betrachtung der Linearisierung hinaus geht.

Es ist aber auch der umgekehrte Fall möglich, dass (meist im Fall einer nicht ausreichenden Kandidatenfunktion) das gewünschte Resultat nicht gefunden werden kann.

Als Lyapunov-Kandidat für das Pendelbeispiel werde abschließend eine energiebasierte Form

$$V(\mathbf{x}) = \frac{g}{l} (1 - \cos(x_1)) + \frac{x_2^2}{2} \quad (16.13)$$

gewählt. Diese nimmt für die untere Ruhelage $x_0 = [0 \quad 0]^T$ ihr Minimum an. Während der quadratische Summand stets positiv ist, oszilliert der erste Summand zwischen $\frac{2g}{l}$ und 0. Somit ist die Bedingung Gl.(16.4) lokal erfüllt. Ableiten der Kandidatenfunktion und Einsetzen der Systemdynamik Gl.(3-1) liefert schließlich

$$\begin{aligned} \dot{V}(\mathbf{x}) &= \frac{g}{l} \sin(x_1) \dot{x}_1 + x_2 \dot{x}_2 \\ &= \frac{g}{l} \sin(x_1) x_2 + x_2 \left(-\frac{g}{l} \sin(x_1) - \frac{\mu}{Ml} x_2 \right) \\ &= -\frac{\mu}{Ml} x_2^2 \leq 0 \end{aligned} \quad (16.14)$$

Hieraus kann nach Gl.(16.5) geschlossen werden, dass die untere Ruhelage mindestens grenzstabil ist.

Asymptotische Stabilität kann jedoch so nicht direkt nachgewiesen werden, da Gl.(16.6) für $x_1 \neq 0$ und $x_2 = 0$ nicht erfüllt ist. Das liegt daran, dass das System für $\dot{\theta} = 0$ kurzzeitig keine Energie verliert.

Hieran zeigt sich auch, dass die Methode nach Lyapunov nur hinreichende Kriterien liefert, da die untere Ruhelage des reibungsbehafteten Pendels nach der Stabilitätsanalyse über die Linearisierung asymptotisch stabil ist. Für das reibungsfreie System liefert die Methode noch weniger nützliche Aussagen, da die Ableitung von $V(x)$ nunmehr identisch null ist.

16.4 Grenzzyklen

16.4.1 Definition

Bei nichtlinearen Systemen kann neben Ruhelagen noch ein weiteres Phänomen auftreten, das Grenzzyklus genannt wird. Hierzu wird als Beispiel das Phasenportrait des so genannten Van-der-Pol²-Oszillators in Bild 16-4 betrachtet, der für gewisse Parameter über die Gleichung $\ddot{x} + 0,5(x^2 - 1)\dot{x} + x = 0$ beschrieben wird.

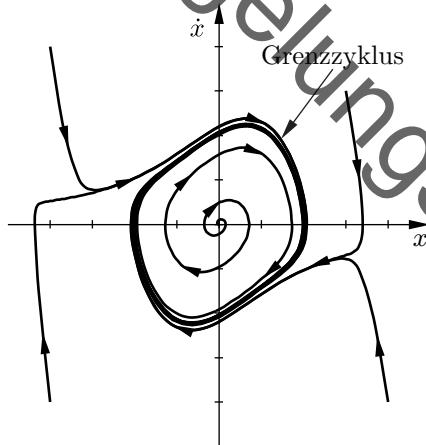


Bild 16-4: Phasenportrait des Vanderpol-Oszillators

²Balthasar van der Pol (1889-1959), niederländischer Physiker [59]

Dieser Oszillator besitzt eine instabile Ruhelage in $x = 0$. Die die Ruhelage verlassenden Trajektorien laufen jedoch weder gegen eine andere Ruhelage noch ins Unendliche, sondern konvergieren gegen eine periodische Lösung, die sich als geschlossene Kurve im Phasenraum zeigt – den *Grenzzyklus*.

Grenzzyklen

Grenzzyklen sind isolierte periodische Lösungen der Zustandsgleichung, d. h. es gilt: $\dot{\mathbf{x}}(t) = \mathbf{x}(t + T)$ mit der endlichen Periodendauer $T > 0$.

Zuweilen wird auch von einer *Arbeitsbewegung* anstelle eines Grenzzyklus gesprochen. Das Adjektiv „isoliert“ beschreibt, dass sich in der unmittelbaren Nähe eines Grenzzyklus keine andere periodische Lösung befindet. Somit stellt das Zentrum keinen Grenzzyklus dar.

In Analogie zu den Ruhelagen des Systems können Grenzzyklen weiter differenziert werden. Abhängig vom lokalen Systemverhalten um den Grenzzyklus wird zwischen stabilen, semi-stabilen und instabilen Zyklen unterschieden. Die entsprechenden Phasenbilder sind in Tab. 16-2 dargestellt.

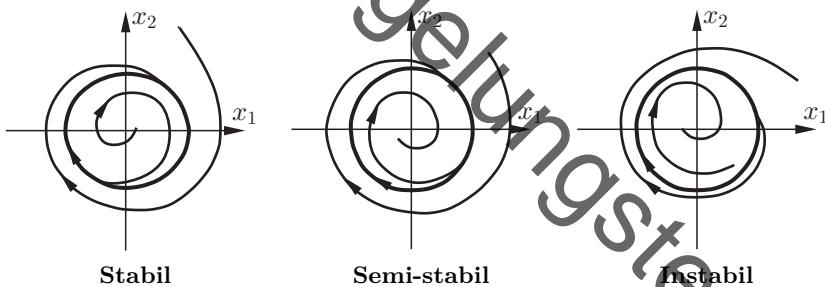


Tabelle 16-2: Phasenbilder eines Grenzzyklus

16.4.2 Grenzzyklen durch schaltende Komponenten

Grenzzyklen sind kein rein mathematisches Phänomen, sondern tauchen bereits bei sehr einfachen, praxisrelevanten Regelkreisen regelmäßig auf. So werden aus Kostengründen in einigen Regelkreisen keine stetig wirkenden Stellglieder verwendet, sondern Aktoren, die nur wenige Schaltzustände

de herzustellen und auszuführen haben. Beispielsweise ist ein elektrischer Schalter, der nur auf *EIN* oder *AUS* geschaltet wird, einfacher, billiger und meist zuverlässiger als ein stetig wirkendes Strom- oder Spannungsstellgerät. Für Ventile und viele andere Stellorgane gilt Ähnliches. Auch Pulsweitenmodulatoren fallen unter diese Kategorie.

Schaltende Systeme

Ein System wird *schaltend* genannt, wenn es durch eine statische, nichtlineare Kennlinie $y = f(u)$ beschrieben wird, wobei die Kennlinie f unstetig ist.

Weit verbreitet sind schaltende Regler bei Temperaturregelungen in Gebäuden, Wärmegeräten und Öfen, aber auch zur Spannungsregelung an kleinen Gleichstromgeneratoren oder zur Füllstandsregelung an Behältern.

Wird ein schaltender Regler in einem Regelkreis eingesetzt, so leuchtet ein, dass dieser nicht mit den Mitteln der Linearisierung behandelt werden kann. Ihr Einsatz führt in den meisten Fällen auf Grenzzyklen im geschlossenen Kreis.

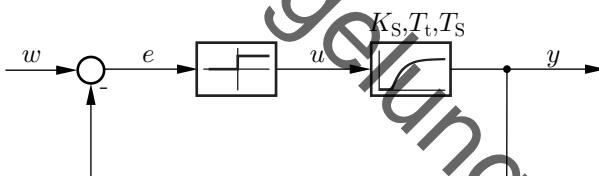


Bild 16-5: Regelkreis mit Zweipunktregler

Als Beispiel soll ein Regelkreis bestehend aus einer PT_1T_t -Regelstrecke und einem sogenannten Zweipunktregler nach Bild 16-5 untersucht werden. Aus der Kennlinie des Zweipunktreglers ist zu erkennen, dass die Stellgröße für positive Werte der Regeldifferenz e ihren positiven Maximalwert annimmt, und dass sie für negative Werte der Regeldifferenz null ist.

Nimmt man an, dass alle Signale für lange Zeit null gewesen sind und die Regelung zum Zeitpunkt $t = 0$ in Betrieb gesetzt wird, so ergeben sich die in Bild 16-6 dargestellten Verläufe von Stell- und Regelgröße. Zum Vergleich ist die Sprungantwort der Regelstrecke auf einen Sprung von $u = u_m$ zum Zeitpunkt $t = 0$ mit eingetragen.

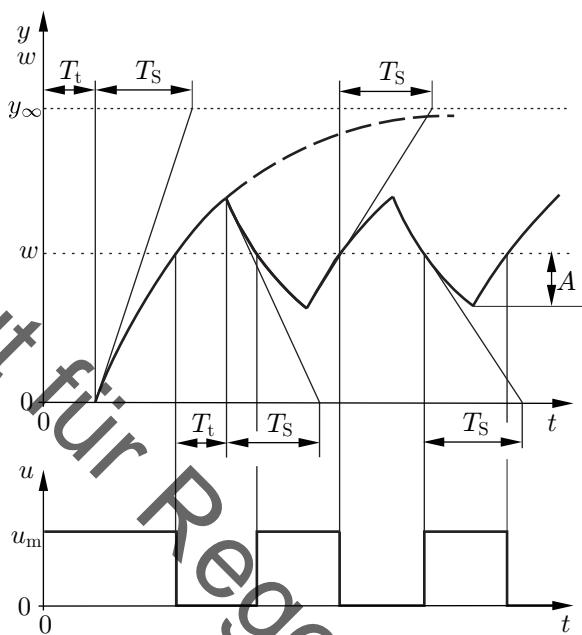


Bild 16-6: Regel- und Stellgröße für den Regelkreis mit Zweipunktregler

Bild 16-6 zeigt, dass zum Zeitpunkt $t = 0$ die Stellgröße auf ihren Maximalwert gebracht wird, weil die Regeldifferenz positiv ist. Der Verlauf der Regelgröße entspricht dem einer Sprungantwort. Sobald $y = w$, d. h. y die Führungsgröße w erreicht, schaltet der Regler die Stellgröße auf ihren Minimalwert (hier $u = 0$). Die zugehörige Wirkung im Verlauf der Regelgröße wird wegen der Totzeit der Regelstrecke erst um T_t nach dem Zeitpunkt des Umschaltens sichtbar. Man erkennt, dass sich im Regelkreis eine periodische Lösung und damit ein Grenzzyklus einstellt.

Der entstehende Grenzzyklus ist ein wesentlicher Nachteil schaltender Regler gegenüber den stetig wirkenden, da dieser Grenzzyklus mit einer Regelabweichung und kontinuierlichen Energieaufwendungen des Aktors verbunden ist.

16.4.3 Grenzzyklen durch Integrator-Windup

Auch bei nicht-schaltenden Reglern können Grenzzyklen entstehen, wofür das sogenannte Integrator-Windup das vermutlich prominenteste Beispiel ist. Hierfür wird eine verzögernde Regelstrecke mit Stellgrößenbeschränkung betrachtet, die durch einen Regler mit integrierendem Anteil geregelt wird – aus Gründen der Einfachheit wird dabei von einem I-Regler ausgegangen (siehe Bild 16-7).

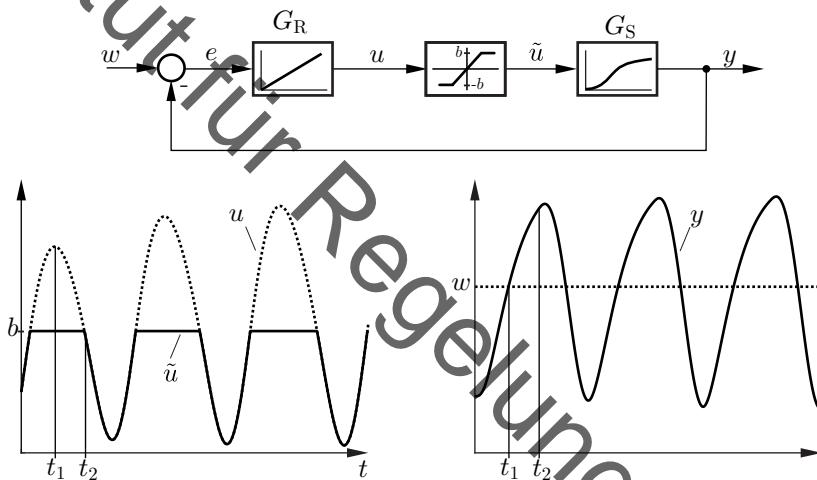


Bild 16-7: Regelkreis und Zeitverläufe beim Integrator-Windup

Bei einem Sollwertsprung ergibt sich eine deutliche Regelabweichung e , die den I-Regler bei hinreichend großer Verstärkung in die Begrenzung b laufen lässt. Hierdurch beginnt sich die angeforderte Stellgröße u des Reglers von der tatsächlich aufgeprägten Stellgröße \tilde{u} zu unterscheiden. Da der I-Regler nicht weiß, dass er bereits den maximalen Wert anfordert, erhöht er u immer weiter, um die bleibende Regelabweichung auszugleichen.

Erreicht die Regelgröße die Führungsgröße (t_1 in Bild 16-7), so versucht der Regler die Stellgröße zu verringern. Dies tut der I-Regler auch – allerdings reduziert er nur u und nicht \tilde{u} . Somit gibt die Begrenzung für eine zusätzliche Zeit bis t_2 noch eine unverminderte maximale Stellgröße aus.

Der Integrator ist also sozusagen „vollgelaufen“ und muss zunächst u abbauen, bevor \tilde{u} sich ändert. Das ist sehr nachteilig, da Arbeitsbewegungen, trüges Verhalten oder starkes Überschwingen die Folge sind.

Integrator-Windup

Wird eine Regelstrecke mit einer Stellgrößenbeschränkung durch einen Regler mit integrierendem Anteil geregelt, so werden Abweichungen zwischen angeforderter und aufgeprägter Stellgröße zu nachteiligem Verhalten im Regelkreis führen. Dieses Phänomen wird (Integrator-) *Windup* genannt.

Zur Bewertung, ob derartige Grenzyklen insgesamt tolerierbar sind, ist es notwendig, Amplitude und Periodendauer des Grenzyklus näherungsweise zu ermitteln und den Grenzyklus auf Stabilität zu untersuchen.

16.5 Stabilitätsanalyse von Grenzyklen

Eine rigorose Analyse von Grenzyklen ist sehr aufwändig und erfordert mathematische Werkzeuge, die über das im Ingenieursstudium vermittelte Wissen hinausgehen. Dies liegt daran, dass Lösungstrajektorien, die sich nach einer leichten Störung der Trajektorie des Grenzykluses einstellen, selbst keine periodischen Lösungen sind. Deshalb kann ein Abstand zwischen dieser gestörten Trajektorie und dem Grenzyklus nur schwer ermittelt werden. Aus diesem Grunde beschränkt man sich bei der Analyse von Grenzyklen auf eine Klasse spezieller nichtlinearer Regelungssysteme, welche für die regelungstechnische Praxis von großer Bedeutung ist und für die ein sehr handliches Näherungsverfahren existiert.

16.5.1 Beschreibungsfunktionen

Im Folgenden werden nur Regelungssysteme mit der in Bild 16-8 wiedergegebenen Struktur betrachtet, die einem Wiener³-Hammerstein-Modell entspricht [39]. Sie bestehen nach passendem Zusammenfassen der Einzelglieder aus der Reihenschaltung zweier Teilsysteme. Das erste ist ein statisches (nichtlineares) Übertragungsglied, das durch seine Kennlinie beschrieben

³Norbert Wiener (1894-1964), amerikanischer Mathematiker und Philosoph [63]

werden kann. Die dynamischen Anteile des Regelkreises sind hingegen linear im zweiten Teilsystem zusammengefasst.

Diese Art der Dekomposition tritt beispielsweise auf, wenn eine lineare Regelstrecke mit einem schaltenden Regler geregelt wird. Das so isolierte nichtlineare Glied muss aber keineswegs immer die Funktion eines Reglers und der abgetrennte lineare Teil die einer Regelstrecke haben. Durch Abtrennen nichtlinearer Eigenschaften von Regelstrecken erhält man die gleiche Struktur. Lediglich der Einfachheit halber wird im Folgenden vorwiegend von der Vorstellung einer linearen Regelstrecke und eines statischen nichtlinearen Reglers ausgegangen.

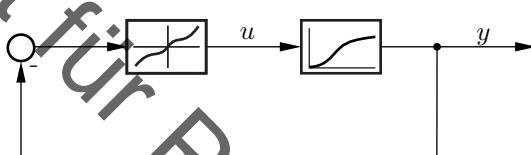


Bild 16-8: Regelkreis als Reihenschaltung einer statischen Nichtlinearität mit einem dynamischen linearen System

Es wird nun angenommen, dass ein Grenzzyklus vorliegt und der Regelkreis daher eine Dauerschwingung $u = f(t)$ ausführt. Als periodische Funktion kann man diese Dauerschwingung in seine Fourierreihe wie in Gl.(16.15)

$$f(t) = \sum_{k=-\infty}^{\infty} c_k \cdot e^{j\omega k t} \quad (16.15)$$

zerlegen. Dies erlaubt es, die resultierende Ausgangsgröße des nichtlinearen Teils ($u(t)$) in Bild 16-8) durch sinusförmige Teilschwingungen unterschiedlicher Amplitude und unterschiedlicher Frequenz darzustellen.

Falls der lineare Teil des Regelkreises Tiefpassverhalten aufweist, dämpft er Teilschwingungen in u mit höherer Frequenz stärker als solche mit geringer Frequenz. Am wenigsten dämpft er die Teilschwingung mit der niedrigsten Frequenz.

Ist die Kennlinie symmetrisch zum Nullpunkt, so verschwindet der konstante Faktor $c_0 = 0$ und die niedrigste Frequenz ist die Grundschwingung c_1 mit Frequenz ω . Durch geeignete Definition von Abweichungsgrößen kann

man dabei in vielen Fällen erreichen, dass die Kennlinien zumindest näherungsweise punktsymmetrisch sind.

Ist das Tiefpassverhalten des linearen Teilsystems stark ausgeprägt, so verläuft auch die periodische Ausgangsgröße des linearen Teils ($y(t)$ in Bild 16-8) mit guter Näherung sinusförmig. Dabei ist die resultierende Frequenz mit der Eingangsfrequenz ω identisch. Die daraufhin resultierende Ausgangsgröße u des nichtlinearen Teils ist wiederum nicht sinusförmig, aber besitzt aufgrund des statischen Verhaltens erneut ω als Grundschwingung.

Somit ist die eingangs getroffene Annahme einer periodischen Lösung mit Frequenz ω erfüllt. Zudem kann offenbar die Betrachtung des nichtlinearen Gliedes auf seine Übertragungseigenschaften für die Grundschwingung beschränkt werden. Man untersucht also den Zusammenhang zwischen einer sinusförmigen Eingangsgröße und der Grundschwingung der daraus resultierenden Ausgangsgröße und gewinnt so im Wege der „harmonischen Linearisierung“ eine lineare Näherung des nichtlinearen Gliedes.

Definition von Beschreibungsfunktion

Die Beschreibungsfunktion verknüpft harmonische Größen gleicher Frequenz miteinander und zwar ganz ähnlich einem Frequenzgang. Sie ist definiert durch

$$B = \frac{y_g}{u} = \frac{\text{Zeiger der Grundschwingung der Ausgangsgröße}}{\text{Zeiger der sinusförmigen Eingangsgröße}} . \quad (16.16)$$

Da die Beschreibungsfunktion hier auf statische Glieder angewandt wird, ist sie nicht von der Frequenz der erregenden Schwingung, sondern nur von der Amplitude abhängig. Hierin unterscheidet sie sich vom Frequenzgang, der bekanntlich nur von der Frequenz und nicht von der Amplitude der Eingangsgröße abhängt.

Zu einem nichtlinearen Glied mit vorgegebener Kennlinie bestimmt man die Beschreibungsfunktion, indem man für sinusförmige Eingangsgrößen unterschiedlicher Amplitude die infolge der Kennlinie entstehenden Ausgangsgrößen ermittelt. Zu den Ausgangsgrößen wird das erste Glied einer Fourierreihe bestimmt und dieses durch seinen Zeiger beschrieben. Der Quotient aus diesem Zeiger und dem Zeiger der Eingangsgröße ergibt nach Gl.(16.16) die Beschreibungsfunktion als i. Allg. komplexwertige Funktion der Amplitude der Eingangsgröße.

Wenn wie vorausgesetzt die Kennlinie des nichtlinearen Gliedes eine ungerade Funktion ist, die punktsymmetrisch zum Ursprung des Koordinatensystems verläuft, und ferner die Eingangsgröße mittelwertfrei ist, dann ist auch die resultierende Ausgangsgröße mittelwertfrei. Diese Voraussetzungen sind nicht notwendig, vereinfachen aber die durchzuführenden Rechnungen erheblich.

Viele technisch interessante Aufgabenstellungen fallen in die Kategorie statischer punktsymmetrischer Kennlinien. So werden aus Kostengründen schaltende Regler vorgesehen. Begrenzungen gibt es in jedem technischen System aufgrund der Stellgrößenbeschränkungen. Systeme mit Totzone (Ansprechschwelle) entstehen z. B. bei hydraulischen Stellantrieben, wenn deren Steuerorgane eine positive Überdeckung aufweisen.

Die Beschreibungsfunktionen für zahlreiche einfache Kennlinienformen sind tabelliert; die Tabellen 16-3 und 16-4 geben einige Beschreibungsfunktionen $B = f(U)$ mit U als Amplitude der sinusförmigen Eingangsgröße $u(t)$ wieder. Man erkennt, dass die Beschreibungsfunktionen aller eindeutigen Kennlinien (Tab. 16-3) rein reell sind, während zu mehrdeutigen (Hysterese-) Kennlinien komplexwertige Beschreibungsfunktionen (Tab. 16-4) gehören. Das liegt daran, dass bei eindeutigen Kennlinien keine Phasenverschiebung zwischen Ein- und Ausgangssignal entstehen kann.

Der Verlauf der Berechnung einer Beschreibungsfunktion soll am Beispiel des Zweipunkt-Gliedes erläutert werden. Bild 16-9 zeigt die Kennlinie, zwei sinusförmige Eingangsgrößen unterschiedlicher Amplitude und die resultierende Ausgangsgröße des Zweipunkt-Gliedes. Die Ausgangsgröße ist unabhängig von der Amplitude der Eingangsgröße, weil ein Zweipunkt-Glied von beliebig kleinen Werten der Eingangsgröße umgeschaltet wird. Die resultierende Rechteckschwingung kann durch die Fourierreihe

$$y(t) = \frac{4}{\pi} d \cdot \left[\sin(\omega t) + \frac{1}{3} \sin(3\omega t) + \frac{1}{5} \sin(5\omega t) + \dots \right] \quad (16.17)$$

dargestellt werden. Die Grundschwingung darin ist

$$y_g(t) = \frac{4}{\pi} d \sin(\omega t) \quad (16.18)$$

und der zugehörige Zeiger ist $\underline{y}_g = -j \frac{4}{\pi} d$. Der Zeiger der Eingangsgröße

Bez.	Kenmlinie	Ausgangsgröße	Beschreibungsfunktion	Ortskurve
Zweipunkt-schalter			$B = \frac{4}{\pi} \frac{d}{U}$ $\text{Re}(B) = B$	
Begrenzung Saturierung			$B = \frac{2k}{\pi} \left[\arcsin\left(\frac{b}{U}\right) + \frac{b}{U} \sqrt{1 - \left(\frac{b}{U}\right)^2} \right]$ $\text{Re}(B) = B$	
Totzone Ansprechschwelle			$B = \frac{2k}{\pi} \left[\frac{\pi}{2} - \arcsin\left(\frac{b}{U}\right) - \frac{b}{U} \sqrt{1 - \left(\frac{b}{U}\right)^2} \right]$ $\text{Re}(B) = B$	
Dreipunkt-schalter			$B = \frac{4}{\pi} \frac{d}{U} \sqrt{1 - \left(\frac{b}{U}\right)^2}$ $\text{Re}(B) = B$	

Tabelle 16-3: Beschreibungsfunktionen 1

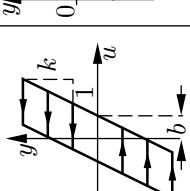
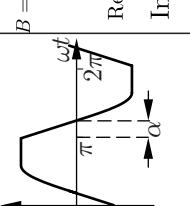
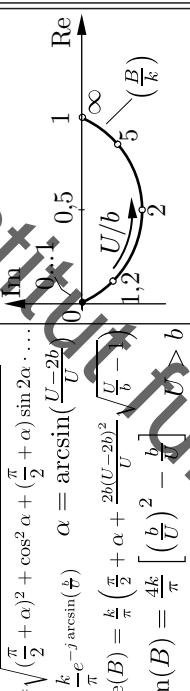
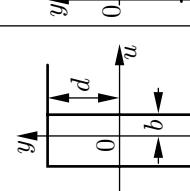
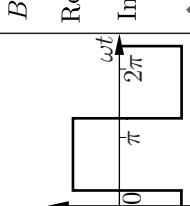
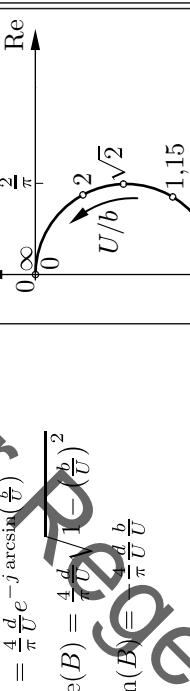
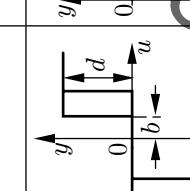
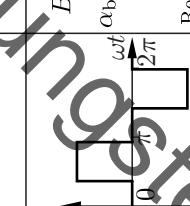
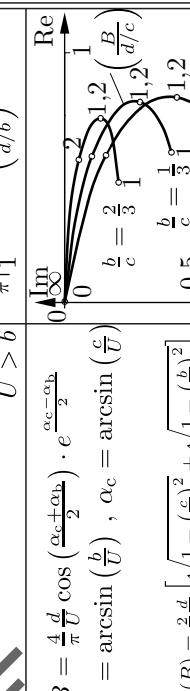
Bez.	Kennlinie	Ausgangsgröße	Beschreibungsfunktion	Ortskurve
Hysterese			$B = \sqrt{\left(\frac{\pi}{2} + \alpha\right)^2 + \cos^2 \alpha} + \left(\frac{\pi}{2} + \alpha\right) \sin 2\alpha \dots$ $\frac{k}{\pi} e^{-j \arcsin(\frac{\pi}{B})}$ $\text{Re}(B) = \frac{k}{\pi} \left(\frac{\pi}{2} + \alpha + \frac{2b(U-2b)}{U} \right) \sqrt{\left(\frac{U}{b}\right)^2 - \left(\frac{b}{U}\right)^2}$ $\text{Im}(B) = \frac{4k}{\pi} \left[\left(\frac{b}{U}\right)^2 - \frac{b}{U} \right]$	
Zweipunktschalter mit Hysterese			$B = \frac{4}{\pi} \frac{d}{U} e^{-j \arcsin(\frac{b}{U})}$ $\text{Re}(B) = \frac{4}{\pi} \frac{d}{U} \sqrt{1 - \left(\frac{b}{U}\right)^2}$ $\text{Im}(B) = \frac{4}{\pi} \frac{d}{U} \frac{b}{U}$	
Dreipunktschalter mit Hysterese			$B = \frac{4}{\pi} \frac{d}{U} \cos\left(\frac{\alpha_c + \alpha_b}{2}\right) \cdot e^{\frac{\alpha_c - \alpha_b}{2}}$ $\alpha_b = \arcsin\left(\frac{b}{U}\right), \quad \alpha_c = \arcsin\left(\frac{c}{U}\right)$ $\text{Re}(B) = \frac{2}{\pi} \frac{d}{U} \left[\sqrt{1 - \left(\frac{b}{U}\right)^2} + \sqrt{1 - \left(\frac{c}{U}\right)^2} \right]$ $\text{Im}(B) = -\frac{2}{\pi} \frac{d}{U} \frac{c-b}{U}$	

Tabelle 16-4: Beschreibungsfunktionen 2

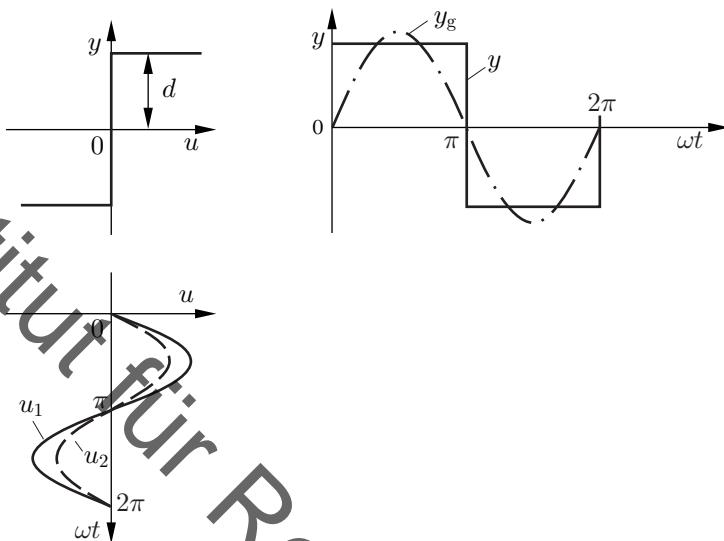


Bild 16-9: Zweipunktschalter, Eingangs- und Ausgangssignal

$u(t) = U \sin(\omega t)$ ist $\underline{u} = -jU$ und damit folgt die Beschreibungsfunktion

$$B = \frac{y_g}{\underline{u}} = \frac{4}{\pi} \frac{d}{U} . \quad (16.19)$$

Sie ist in Tab. 16-3 enthalten und dort auch als Ortskurve dargestellt.

In vielen Fällen kann eine grobe Abschätzung der Beschreibungsfunktion hilfreich sein. So kann man im oben behandelten Fall des Zweipunkt-Gliedes aus der eindeutigen Kennlinie auf eine reelle Beschreibungsfunktion schließen. Da die Amplitude der Ausgangsgröße unabhängig ist von der der Eingangsgröße, strebt der Quotient nach Gl.(16.16) für große Amplituden der Eingangsgröße gegen null und für kleine Amplituden gegen unendlich.

In ähnlicher Weise kann man abschätzen, dass die Beschreibungsfunktion eines Gliedes mit Begrenzung für kleine Amplituden der Eingangsgröße (kleiner als b in Tab. 16-3) gleich der Steigung der Kennlinie ist und dass sie für sehr große Amplituden der Eingangsgröße gegen null geht.

Bei der Kennlinie mit einer Totzone gilt in gewissem Sinne das Umgekehr-

te: Für kleine Amplituden der Eingangsgröße ist die Ausgangsgröße und damit auch die Beschreibungsfunktion null, für genügend große geht sie gegen die Steigung der Kennlinie, d. h. das Übertragungsglied verhält sich näherungsweise linear.

Auch für Kennlinien mit Hysterese können derartige Abschätzungen hilfreich sein. So ist leicht zu erkennen, dass ein Übertragungsglied mit einer Zweipunkt-Kennlinie mit Hysterese für Eingangsgrößen mit einer Amplitude unterhalb der halben Hysteresebreite (kleiner als b in Tab. 16-4) eine Ausgangsgröße und damit auch eine Beschreibungsfunktion von null zur Folge hat. Weiter nimmt die Ausgangsgröße für Amplituden der Eingangsgröße, die wenig größer sind als die halbe Hysteresebreite, einen endlichen Wert an und ihre Grundschwingung gegenüber der Eingangsgröße ist um -90° in der Phase verschoben.

16.5.2 Zwei-Ortskurven-Kriterium

Mit der Beschreibungsfunktion ist eine lineare Näherung für das im Regelkreis enthaltene nichtlineare Glied gefunden worden, die das Verhalten dieses Gliedes für den Fall beschreibt, dass der Regelkreis eine Dauerschwingung ausführt. Bezuglich dieses Grenzzyklus sind die interessanten Fragestellungen, ob dieser überhaupt auftritt, ob er stabil ist und wie groß Amplitude und Frequenz sind. Alle diese Fragen können im Rahmen der Gültigkeit der eingeführten Näherung mit einer Modifikation des vereinfachten Nyquist-Kriterium – dem sogenannten Zwei-Ortskurven-Kriteriums – beantwortet werden.

In Abschnitt 9.3 ist das „Vereinfachte Nyquist-Kriterium“ hergeleitet worden. Dieses besagt, dass (Anwendbarkeit vorausgesetzt) $|G_0(\omega_\pi)| < 1$ gelten muss, damit der geschlossene Regelkreis stabil ist. Wenn nun der Frequenzgang des aufgeschnittenen Regelkreises aus zwei miteinander multiplizierten Teilstrecken besteht, z. B. dem des Reglers und dem der Regelstrecke, allgemeiner also

$$G_0 = G_1 \cdot G_2 = |G_1| \cdot e^{j\varphi_1} \cdot |G_2| \cdot e^{j\varphi_2} = |G_1| \cdot |G_2| \cdot e^{j(\varphi_1 + \varphi_2)} \quad (16.20)$$

so ergibt sich

$$\varphi_0(\omega_\pi) = \varphi_1(\omega_\pi) + \varphi_2(\omega_\pi) = \pm(1 + 2n)\pi \quad , \quad n \in \mathbb{N}_0 \quad (16.21)$$

bzw. für o.B.d.A. $n = 0$:

$$\varphi_1(\omega_\pi) = -\pi - \varphi_2(\omega_\pi) = \varphi'_2(\omega_\pi) \quad (16.22)$$

mit φ'_2 als dem Winkel, der zu dem Frequenzgang

$$G'_2 = |G'_2| \cdot e^{j\varphi'_2} = -\frac{1}{G_2} = \left| \frac{1}{G_2} \right| \cdot e^{j(-\pi - \varphi_2)} \quad (16.23)$$

gehört. Weiter ergibt sich

$$|G_0(j\omega_\pi)| = |G_1(\omega_\pi)| \cdot |G_2(\omega_\pi)| < 1 \Leftrightarrow |G_1(\omega_\pi)| < |G'_2(\omega_\pi)|. \quad (16.24)$$

Diese beiden Überlegungen münden im *Zwei-Ortskurven-Kriterium*.

Zwei-Ortskurven-Kriterium

Gegeben sei der aufgeschrittene Regelkreis $G_0 = G_1 \cdot G_2$, welcher die Voraussetzungen für die Anwendbarkeit des vereinfachten Nyquist-Kriteriums erfüllt. Trägt man die Ortskurven der beiden Frequenzgänge G_1 und $G'_2 = -1/G_2$ einschließlich ihrer Frequenzparametrierung in ein gemeinsames Diagramm ein, so ist die Frequenz ω_π genau die Frequenz, für die die Phasenwinkel beider Frequenzgänge bei gleichem ω gleich groß sind. Der geschlossene Regelkreis ist genau dann stabil, wenn der Betrag des durch Inversion eines Teilstreckengangs gewonnenen Frequenzgangs größer ist als der des nichtinvertierten Teilstreckengangs.

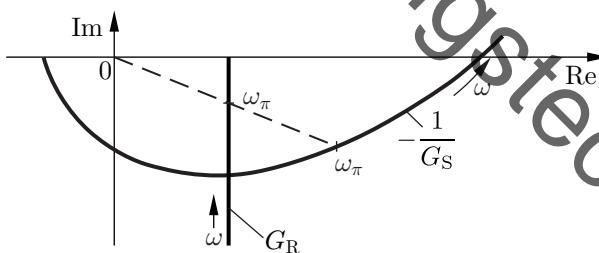


Bild 16-10: Zwei-Ortskurven-Verfahren, Beispiel

Bild 16-10 zeigt ein Anwendungsbeispiel für einen PI-Regler und ein Verzögerungsglied höherer Ordnung. Der Frequenzgang der Regelstrecke wurde invertiert und ω_π bereits eingezeichnet. Offenbar ist der geschlossene

Regelkreis stabil, da der invertierte Frequenzgang an dieser Frequenz den größeren Betrag besitzt.

Als Vorteil dieser Darstellungsform gegenüber der Darstellung der Ortskurve des Frequenzganges des aufgeschnittenen Regelkreises wird deutlich, dass bei Änderung des Reglers nur die zum Regler gehörende Ortskurve erneut zu zeichnen ist.

Ein durch seine Ortskurve gegebener Frequenzgang $G(j\omega)$ wird dabei analog zur Inversion einer komplexen Zahl negativ invertiert: Man invertiert den Betrag, d. i. die Länge des Zeigers, und spiegelt den Zeiger von G an der imaginären Achse, um die Richtung des Zeigers von $-1/G$ zu erhalten – siehe Bild 16-11.

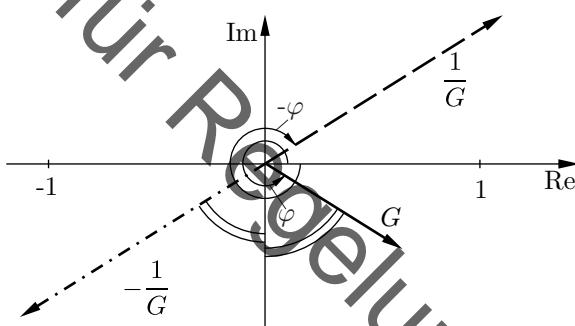


Bild 16-11: Negative Inversion von Ortskurven

Eine Dauerschwingung am geschlossenen Regelkreis ist entsprechend dem Zwei-Ortskurven-Kriterium dann zu erwarten, wenn die beiden Beträge in $|G_1|$ und $|-1/G_2|$ gleich groß sind. Dann schneidet nämlich die Ortskurve des Frequenzganges des aufgeschnittenen Regelkreises den kritischen Punkt -1 . In diesem Fall sind die Frequenzgänge G_1 und G'_2 sowohl hinsichtlich ihres Betrages als auch ihres Phasenwinkels gleich. Folglich schneiden ihre Ortskurven sich in Punkten mit der gleichen Frequenz ω_π .

Zur Analyse von Regelkreisen mit nichtlinearen Gliedern wird das Zwei-Ortskurven-Kriterium dadurch modifiziert, dass nicht mehr zwei Frequenzgänge, sondern ein Frequenzgang und eine Beschreibungsfunktion miteinander in Beziehung gebracht werden. Dies war mit dem vereinfachten Nyquist-Kriterium nicht möglich, da Beschreibungsfunktionen (mit Parameter U)

und Ortskurven (mit Parameter ω) nicht miteinander multipliziert werden können. Die folgenden Voraussetzungen müssen für die Anwendbarkeit dieses Verfahren gelten:

Voraussetzungen für die Anwendbarkeit des Zwei-Ortskurven-Kriteriums zur Analyse von Grenzzyklen

Besteht der aufgeschnittene Regelkreis aus der Reihenschaltung

eines linearen (dynamischen) Systems G mit ausgeprägtem Tiefpassverhalten und

- eines (nichtlinearen) statischen Systems B mit nullpunktssymmetrischer Kennlinie,

so kann das Zwei-Ortskurven-Kriterium zur Analyse von Grenzzyklen am geschlossenen Regelkreis angewandt werden. Man spricht auch vom Verfahren der *harmonischen Balance* [2].

Die Voraussetzungen des vereinfachten Nyquist-Kriteriums sind dabei indirekt mit erfüllt, da Tiefpassfilter nie instabile Polstellen haben (da der eingeschwungene Zustand per definitionem existiert) und daher mit der fallenden Amplitude auch eine monoton fallende Phase aufweisen.

Grenzzyklen gemäß der harmonischen Balance

Stellt man entweder

- $-1/G(j\omega)$ und $B(U)$ oder
- $-1/B(U)$ und $G(j\omega)$

als Ortskurven dar, so entspricht jeder Schnittpunkt einem Grenzzyklus, dessen Frequenz ω_A an der Ortskurve des Frequenzgangs und deren Amplitude U_A an der Ortskurve der Beschreibungsfunktion abzulesen ist.

Das Ablesen von ω_A und U_A kann Bild 16-12 entnommen werden. Dies liegt daran, dass ein Grenzzyklus als Dauerschwingung sich in unmittelbarer Analogie zum Stabilitätsrand bei linearen Regelkreisen aus

$$G(j\omega_A) \cdot B(U_A) = -1 \tag{16.25}$$

ergibt. Für die Fälle $\omega_A = 0$ oder $U_A = 0$ liegt kein Grenzzyklus vor.

Anhand von Bild 16-12 ist zu erkennen, dass für Regelkreise mit Verzögerungsgliedern zweiter oder erster Ordnung zusammen mit nichtlinearen

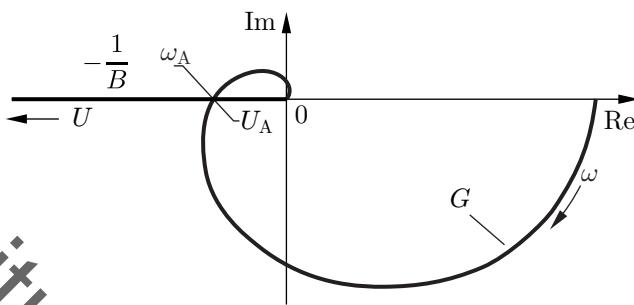


Bild 16-12: Ermittlung der Arbeitsbewegung für einen Regelkreis mit Verzögerung 4. Ordnung und Zweipunkt-Glied

Gliedern mit eindeutiger Kennlinie keine Arbeitsbewegung ermittelt wird, weil die zugehörige Frequenzgangortskurve die negativ-reelle Achse nicht schneidet. Demgegenüber ist jedoch bekannt, dass Zweipunkt-Glieder in derartigen Regelkreisen immer zu Arbeitsbewegungen führen; im Falle der Verzögerung erster Ordnung ist allerdings die Frequenz ω_A sehr groß. Daraus folgt, dass für Verzögerungsglieder niedriger Ordnung die eingangs getroffenen Voraussetzung des ausgeprägten Tiefpassverhaltens nicht erfüllt ist.

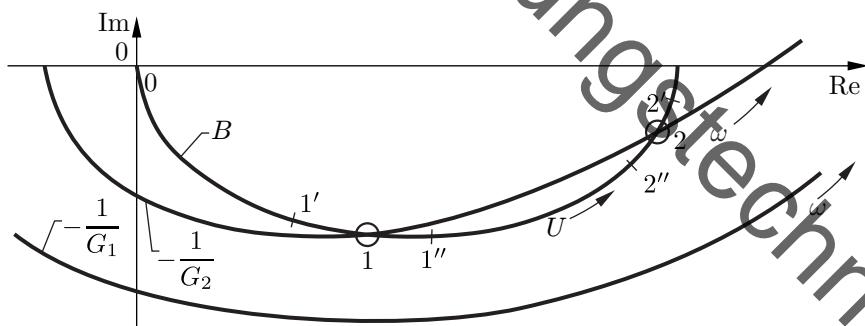


Bild 16-13: Beschreibungsfunktion für Glied mit Hysterese und Ortskurven für Verzögerungsglieder

Bild 16-13 zeigt anhand der Ortskurven für einen Regelkreis mit Verzöge-

rung höherer Ordnung und ein nichtlineares Glied mit Hysterese, dass die Ortskurven von Beschreibungsfunktion und negativ-inversem Frequenzgang sich sowohl mehrmals als auch gar nicht schneiden können. Die Ortskurven der Beschreibungsfunktion B und des negativ-inversen Frequenzgangs $-1/G_1$ haben keinen gemeinsamen Schnittpunkt; ein Grenzzyklus ist nicht zu erwarten. Da außerdem die Länge jedes an die Ortskurve der Beschreibungsfunktion vom Koordinatenursprung aus gezogenen Zeigers kleiner ist als die des unter gleichem Winkel an die Ortskurve $-1/G_1$ gezogenen Zeigers, ist in Übertragung der für die Stabilitätsanalyse bei linearen Regelkreisen gewonnenen Erkenntnisse nicht damit zu rechnen, dass das untersuchte System aufklingende Schwingungen ausführt.

Die Verhältnisse sind anders für die Ortskurven B und $-1/G_2$, die zwei gemeinsame Schnittpunkte 1 und 2 aufweisen. Daraus ist zu schließen, dass dieser Regelkreis zwei unterschiedliche Grenzzyklen besitzt. Frequenz und Amplitude sind durch die bei den Schnittpunkten in 1 und 2 vorgefundenen Parameterwerte an den Ortskurven gegeben.

Ein Schnittpunkt der beiden Ortskurven besagt zwar, dass der Regelkreis einen bestimmten Grenzzyklus besitzt; er besagt aber nicht, dass dieser auch stabil ist. Stabil ist eine Grenzzyklus, der sich nach einer Störung wieder auf die vorher vorhandenen Werte von Amplitude und Frequenz einstellt. Im Gegensatz dazu stehen instabile Grenzzyklen, die infolge von Störungen auf größere Amplitudenwerte aufklingen oder auf kleinere abklingen bzw. ganz verlöschen, wobei mit den Veränderungen der Amplitude meist auch solche der Frequenz verbunden sind. Zur Prüfung der Stabilität von Arbeitsbewegungen empfiehlt es sich, die Auswirkungen kleiner Störungen wie folgt zu untersuchen.

Stabilitätsanalyse von Grenzzyklen

Grenzzyklen können mit der harmonischen Balance anhand zweier Bedingungen auf Stabilität geprüft werden:

- i) Für wachsende U besitzt der Zeiger gleicher Richtung des invertierten Frequenzgangs einen größeren Betrag als der nicht-invertierte.
- ii) Für sinkende U besitzt der Zeiger gleicher Richtung des invertierten Frequenzgangs einen kleineren Betrag als der nicht-invertierte.

Sind beide Bedingungen erfüllt, ist der Grenzzyklus stabil. Ist eine erfüllt, so ist er semi-stabil. Ist keine erfüllt, so ist er instabil.

Dieser Zusammenhang soll in Bild 16-13 anhand von Schnittpunkt 1 erläutert werden. Anschaulich wird Stabilität dadurch geprüft, indem man beobachtet, wie der Regelkreis reagiert, wenn der Grenzzyklus durch eine Störung die zum Punkt $1'$ gehörende Amplitude annimmt. Der zugehörige Zeiger gleicher Richtung an $-1/G_2$ ist größer als der Zeiger an B ; das zugehörige lineare System wäre stabil, d. h. es würde abklingende Schwingungen ausführen. Zu abklingenden Schwingung gehören aber Punkte auf B , die noch weiter vom Schnittpunkt 1 entfernt liegen als der geprüfte Punkt $1'$, d. h. eine so gestörter Grenzzyklus wird verlöschen. Eine entsprechende Prüfung des Punktes $1''$ ergibt das entgegengesetzte Bild. Der Zeiger an $-1/G_2$ ist kürzer als der an B , die Schwingung wird aufklingen und sich damit vom Punkt 1 in Richtung wachsender Amplitudenwerte entfernen. Zusammenfassend ist festzustellen, dass Grenzzyklus 1 nicht stabil ist.

Im Gegensatz zu der durch den Punkt 1 in Bild 16-13 beschriebene Grenzzyklus ist der durch den Punkt 2 gekennzeichnete stabil. Eine Störung in Richtung wachsender Amplitudenwerte, etwa nach $2'$, ergibt ein System, dessen Schwingungen abklingen; abklingende Schwingungen führen den Prozess aber nach 2 in den Schnittpunkt zurück. Für eine entgegengesetzt wirkende Störung gilt das Entsprechende; von $2''$ wird der Prozess auch nach 2 zurückgeführt. Der Schnittpunkt 2 bezeichnet daher einen stabilen Grenzzyklus.

16.5.3 Maßnahmen gegen Grenzzyklen

Aus der Methode der Beschreibungsfunktionen können auch Maßnahmen abgeleitet werden, um das oft unerwünschte Auftreten von Grenzzyklen zu verhindern oder wenigstens deren Frequenz und in erster Linie Amplitude abzusenken. Hierzu muss ein Schnittpunkt zwischen den beiden Frequenzgängen soweit möglich verhindert werden. Ist dies nicht möglich, so sollte der Schnittpunkt zumindest bei geringen Werten für U (für geringe Amplituden des Grenzzyklus) und ω (für eine geringe Schaltfrequenz) liegen.

Um dieses Ziel zu erreichen, ist es oft am zweckmäßigsten, den Regelkreis mit zusätzlichen Rückführungen zu versehen, durch welche die Ortskurve des linearen Teils passend verändert wird. So ist es ein bewährtes Vorgehen bei Zweipunktreglern, diese mit Rückführungen wie in Bild 16-14 zu versehen. Im selben Bild sind die Ortskurven der Frequenzgänge G_S und

G_r von Regelstrecke und Rückführung sowie die Ortskurve der negativ-inversen Beschreibungsfunktion $-1/B$ des Zweipunktreglers mit Hysterese eingetragen.

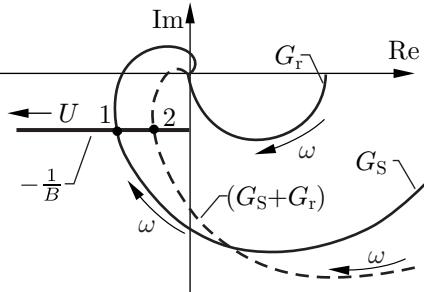
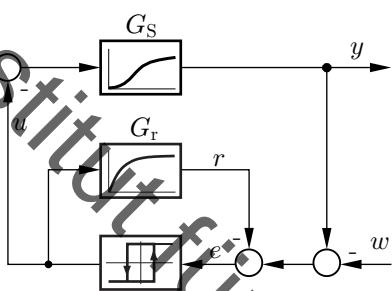


Bild 16-14: Regelkreis mit Rückgeführtem Zweipunktregler inklusive Ortskurven

Falls der Regler ohne Rückführung betrieben wird, ergibt sich ein Grenzzyklus, der durch den Schnittpunkt 1 der Ortskurven $-1/B$ und G_S charakterisiert wird. Durch Zuschalten der Rückführung wird das dynamische Verhalten des linearen Teils der Regelstrecke verändert. Man erkennt in Bild 16-14, dass Regelstrecke und Rückführung parallel geschaltet sind, sodass ihre Frequenzgänge zu addieren sind, um den Frequenzgang des linearen Teils des Regelkreises zu gewinnen. Die Ortskurve der Summe $G_S + G_r$ schneidet die der negativ-inversen Beschreibungsfunktion im Punkt 2. Dieser Schnittpunkt bezeichnet einen Grenzzyklus mit leicht höherer Frequenz und wesentlich kleinerer Amplitude als der zum Punkt 1 gehörende. Dies ist in den Zeitverläufen in Bild 16-15 zu erkennen.

Bei diesem Vorgehen kann die Parallelschaltung eines schwach verzögerten Gliedes möglicherweise die Voraussetzungen des ausgeprägten Tiefpassverhaltens zunichte machen. Außerdem ist bei der Berechnung der Amplitude der Regelgröße zu beachten, dass aus der Parametrierung U der Beschreibungsfunktion nur die Amplitude des Signals e am Eingang des Zweipunktreglers abgelesen wird. Die Amplitude der Regelgröße oder Stellgröße ist hieraus mit Hilfe von $B(U_A)$ und $G_S(j\omega_A)$ zu bestimmen.

Ein weiteres Beispiel für eine geeignete Rückführung zur Verbesserung des dynamischen Verhaltens bei nichtlinearen Kennlinien liefert das sogenannte

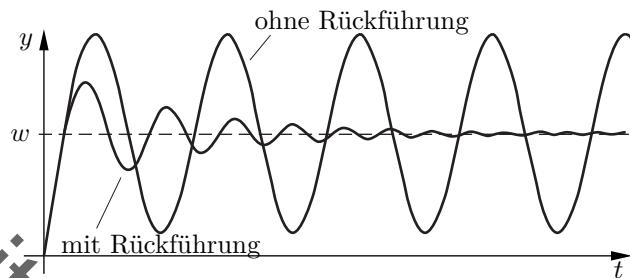


Bild 16-15: Zweipunktregler mit Hysterese mit und ohne Rückführung

Anti-Windup. Wie der Name bereits vermuten lässt, werden hiermit Ge- genmaßnahmen bezeichnet, die das Integrator-Windup unterbinden sollen. Ein bewährtes Vorgehen ist es auch hier, eine zusätzliche Rückführung zu verwenden, indem gemäß Bild 16-16 die Differenz $u - \tilde{u}$ aus angeforderter und aufgeprägter Stellgröße zurückgeführt wird.

Wählt man die Verstärkung $m >> 0$ groß, so wird bei einem Überschreiten der Stellgrößenbeschränkung der Integrator durch die Rückführung in ein schnell abklingendes PT₁ verwandelt. Somit können Abweichungen zwischen u und \tilde{u} unterbunden und dem Windup entgegengewirkt werden.

Die positiven Auswirkungen dieser Maßnahme sieht man in den Zeitverläufen in Bild 16-16. Dort folgt auf einen Sollwertsprung in $t = 0$ ein Sprung zurück in $t = t_1$ auf den Ausgangswert. Der I-Regler läuft beim ersten Sollwertsprung in die Beschränkung und meint daher, mit einem um etwa 50 % erhöhtem Stellwert u das Regelziel erreicht zu haben. Dies führt zu einer um Δt verzögerten Reaktion beim zweiten Sollwertsprung. Durch die Rückführung mit $m >> 0$ wird u an \tilde{u} angeglichen und dieser Irrtum des I-Reglers korrigiert. Mit Anti-Windup erfolgt daher eine unverzögerte Reaktion des Reglers auf die zweite Sollwertänderung.

Die harmonische Balance liefert auch in diesem Fall zusätzliche Argumente, warum die anschaulich festgelegte Maßnahme zur Vermeidung des Windups aus systemtheoretischer Perspektive sinnvoll ist: Die Rückführung m stellt eine Parallelschaltung eines P-Elements zur Regelstrecke G_S dar. Hierdurch wird die Ortskurve des dynamischen Anteils in der komplexen Ebene zu positiven Realteilen hin verschoben, wodurch sich diese von möglichen

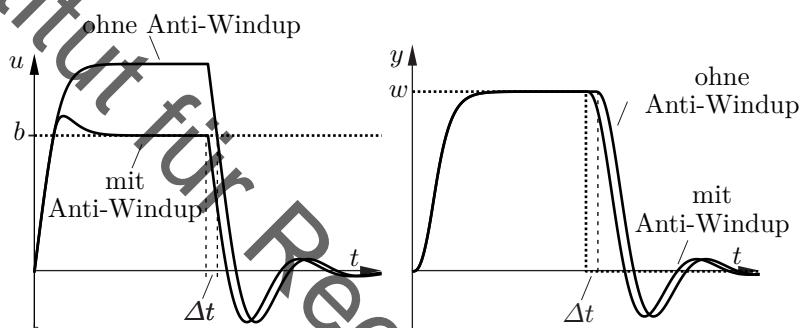
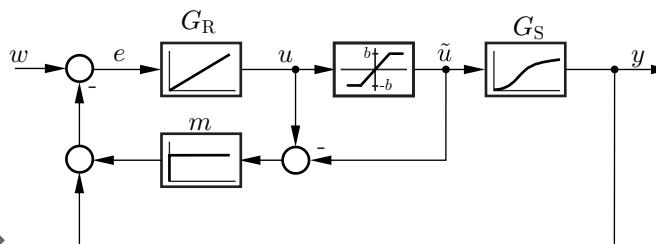


Bild 16-16: Regelkreis und Zeitverläufe mit und ohne Anti-Windup

Schnittpunkten mit $-1/B$, welches negative Realteile aufweist, entfernt. Auch wenn durch diese Parallelschaltung die Anforderung an das ausgeprägte Tiefpassverhalten verletzt werden, so lässt sich aus dem Verfahren dennoch die begründete Vermutung ableiten, dass die vorgeschlagene Maßnahme einen positiven Einfluss auf das Systemverhalten haben wird. Somit kann die harmonische Balance auch dazu dienen, geeignete Maßnahmen zur Beeinflussung von Grenzzyklen abzuleiten.

17 Nichtlineare Regelung

Aufbauend auf den zuvor abgeleiteten Systemeigenschaften nichtlinearer Systeme wurde eine Vielzahl an Reglerverfahren entwickelt, welche sich für verschiedene Problemstellungen eignen. Im Folgenden seien die grundlegenden Konzepte dargestellt, die beispielsweise zur Stabilisierung einzelner Ruhelagen aber auch ganzer Systemtrajektorien genutzt werden.

Neben der Abbildung der nichtlinearen Systemdynamik stellt sich in der Praxis zudem die Frage, ob eine Stabilisierung des Prozess auch bei unvollständigem Systemwissen erzielt werden kann. Dies führt auf die Klasse robuster Reglerentwürfe, die zu Teilen anhand der gleichen Konzepte studiert werden können.

Der Reglerentwurf erfolgt dabei nicht im Bildbereich, da dieser nur für lineare Systeme definiert ist, sondern ausschließlich im Zustandsraum. Dabei werden im Wesentlichen das aus Kapitel 16 bekannte Konzept der Lyapunov-Funktionen mit dem in Abschnitt 17.1 vorzustellenden Konzept der *Exakten Linearisierung* verschieden kombiniert.

17.1 Exakte Linearisierung

Es werden nichtlineare SISO-Systeme der Form

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, u) \quad , \quad y = g(\mathbf{x}, u) \quad (17.1)$$

betrachtet. Für die äquivalente Behandlung von MIMO-Systemen sei auf [2, 49] verwiesen.

Die Aufgabe einer Steuerung oder Regelung ist es, der Regelgröße y ein Sollverhalten aufzuprägen. Die passende Stellgröße u kann offenbar einfacher ermittelt werden, wenn ein unmittelbarer, umkehrbarer Zusammenhang zwischen u und y besteht. Ließe sich nämlich die Ausgangsgleichung in Gl.(17.1) $y = g(\mathbf{x}, u)$ umkehren, d. h. eindeutig nach u auflösen, so gewinne man

$$u = h(\mathbf{x}, v) \quad \Rightarrow \quad y = g(\mathbf{x}, h(\mathbf{x}, v)) = v \quad (17.2)$$

mit der Umkehrfunktion h von g und der Variablen v . Hiermit kann man direkt y ein beliebiges Sollverhalten aufprägen. Wählt man nämlich das

Stellgesetz $u = h(\mathbf{x}, v)$ mit v identisch der Führungsgröße, so ergibt sich direkt $y = v$ mit perfektem Führungsverhalten.

Leider wird dieser Zusammenhang in aller Regel nicht existieren bzw. sich nicht umkehren lassen. Dies sieht man beispielsweise aus dem Vergleich mit dem linearen Zustandsraummodell in Gl.(2.18). Dort taucht in der Ausgangsgleichung $y = \mathbf{c}^T \mathbf{x} + du$ die Eingangsgröße u nur auf, sofern $d \neq 0$ gilt, d. h. das System sprungfähig ist. Diese Eigenschaft werden die meisten Regelstrecken wie Verzögerungsglieder nicht aufweisen.

Die Tatsache, dass u oft keinen unmittelbaren Einfluss auf y hat, sieht man auch an den zugehörigen Sprungantworten der linearen Systeme. In Abschnitt 4.6 wurde nachgewiesen, dass bei einem System mit relativem Grad r die r -te Ableitung der Sprungantwort unstetig ist. Das bedeutet, dass ein unmittelbarer Zusammenhang zwischen u und der r -ten Ableitung $y^{(r)}$ der Ausgangsgröße existiert – zu niedrigeren Ableitungen allerdings nicht. Diesen Zusammenhang nutzt man zu Verallgemeinerung des relativen Grades auf nichtlineare Systeme.

Relativer Grad

Der relative Grad r ist diejenige Ordnung der zeitlichen Ableitung des Systemausgangs y , für die eine unmittelbare Abhängigkeit von der Eingangsgröße u besteht, d. h. es gibt eine Funktion g_r mit

$$y^{(r)} = g_r(\mathbf{x}, u) \quad (17.3)$$

Die Definition fällt im linearen Fall mit der bisherigen zusammen, gilt aber auch für nichtlineare Systeme. Um im Regelungsentwurf den Zusammenhang zwischen y und u ausnutzen zu können, leitet man also y so lange ab, bis der Eingang u erscheint.

Dies soll an einem Beispiel demonstriert werden. Es gelten die Systemgleichungen

$$\dot{x}_1 = x_3 - x_2 \quad , \quad \dot{x}_2 = x_2^2 + x_3 + u \quad , \quad \dot{x}_3 = u \quad , \quad y = x_1 \quad . \quad (17.4)$$

Leitet man die Ausgangsgleichung so oft ab, bis der Eingang u erscheint,

so erhält man hier

$$\begin{aligned} y &= x_1 \\ \dot{y} &= \dot{x}_1 = x_3 - x_2 \\ \ddot{y} &= \dot{x}_3 - \dot{x}_2 = -x_2^2 - x_3 - u + u = -x_2^2 - x_3 \\ \dddot{y} &= -2x_2\dot{x}_2 - \dot{x}_3 = u(-1 - 2x_2) - 2x_2^3 - 2x_2x_3 \quad . \end{aligned} \tag{17.5}$$

Sofern $x_2 \neq -0,5$ gilt, gibt es den Zusammenhang zwischen u und \ddot{y}

$$\ddot{y} = u(-1 - 2x_2) - 2x_2^3 - 2x_2x_3 = g_r(\mathbf{x}, u) \quad . \tag{17.6}$$

Der relative Grad ist also für $x_2 \neq -0,5$ genau 3. Für $x_2 = -0,5$ hingegen wird auch bei weiterem Ableiten der Eingang nie in den Gleichungen erscheinen. Der relative Grad ist nicht wohldefiniert (bzw. $r = \infty$).

Nimmt man an, dass der relative Grad wohldefiniert ist, kann man versuchen $g_r(\mathbf{x}, u)$ umzukehren und nach u aufzulösen. Das entstehende Regelgesetz ist dann die bereits erwähnte Exakte Linearisierung.

Exakte Linearisierung

Gegeben ist ein System mit relativem Grad r , d. h. $y^{(r)} = g_r(\mathbf{x}, u)$. Dann heißt das Stellgesetz

$$u = h_r(\mathbf{x}, v) \tag{17.7}$$

mit der Umkehrfunktion h_r von g_r Exakte Linearisierung.

Die Bezeichnung des Stellgesetzes als „Linearisierung“ wird klar, wenn man den geschlossenen Regelkreis unter diesem Regelgesetz aufstellt. Hier erhält man

$$u = h_r(\mathbf{x}, v) \Rightarrow y^{(r)} = g_r(\mathbf{x}, h_r(\mathbf{x}, v)) = v \tag{17.8}$$

und damit $y^{(r)} = v$ ein lineares System. Dieses entspricht einem r -fachen Integrator mit Eingangsgröße v . Da v ein Parameter der Exakten Linearisierung ist und nicht der physikalischen Stellgröße u entspricht, bezeichnet man v auch als *fiktive Stellgröße*.

Die Exakte Linearisierung darf allerdings nicht mit der Linearisierung einer Differentialgleichung verwechselt werden. Es handelt sich bei der Exakten

Linearisierung nämlich um keine Approximation wie bei einer Linearisierung mittels Taylorreihe. Stattdessen erzwingt das Regelgesetz in Gl.(17.7) eine lineare Systemdynamik. Daher wird diese Art der Linearisierung mit dem Adjektiv „exakt“ versehen. Eine andere übliche Bezeichnung ist auch „feedback linearization“ oder „Linearisierung durch Zustandsrückführung“, die den Charakter der Regelung hervorheben.

Zudem muss festgehalten werden, dass in der bisherigen Form die Exakte Linearisierung zwar ein lineares System $y^{(r)} = v$ erzeugt, das aber nicht bedeutet, dass hiermit das nichtlineare System vollständig linearisiert wurde.

Zur Illustration wird das erste Beispiel in Gl.(17.4) minimal abgeändert, indem als Ausgang $y = x_3$ statt $y = x_1$ genutzt wird. Dann ergibt sich sofort $\dot{y} = u$, ein relativer Grad von 1 und damit ein lineares System erster Ordnung. Offenbar beschreibt aber $\dot{y} = u$ nicht die vollständige Systemdynamik aus Gl.(17.4), welche aus drei Gleichungen besteht. Zwei dieser Gleichungen bleiben folglich unlinearisiert zurück.

17.2 Interne Dynamik

Die Exakte Linearisierung überführt also ein nichtlineares System von Ordnung n in ein lineares Teilsystem von Ordnung r , während $n-r$ Gleichungen potentiell unlinearisiert verbleiben. In diesem Sinne kann man die Exakte Linearisierung als Zustandstransformation auffassen: Aus den Koordinaten \mathbf{x} und u werden neue Koordinaten \mathbf{z} und v , in welchen das System eine möglichst lineare Struktur besitzt.

Exakte Linearisierung als Zustandstransformation

Die Exakte Linearisierung entspricht einer Zustandstransformation auf die neuen Zustandskoordinaten $\mathbf{z} = \varphi(\mathbf{x})$ mit

$$\begin{aligned}\mathbf{z} &= [y \quad \dot{y} \quad \dots \quad y^{(r-1)}]^T \\ &= \left[g(\mathbf{x}) \quad \frac{d}{dt}g(\mathbf{x}) \quad \dots \quad \frac{d^{r-1}}{dt^{r-1}}g(\mathbf{x}) \right]^T = \varphi(\mathbf{x}),\end{aligned}\tag{17.9}$$

Für $r = n$ besitzt der Vektor \mathbf{z} die gleiche Beschreibungskraft wie der Vektor \mathbf{x} und es entsteht eine eindeutige Zustandstransformation. Folglich kann das nichtlineare System in den neuen Koordinaten \mathbf{z} vollständig durch das lineare System $y^{(n)} = v$ beschrieben werden. Alle Eigenschaften, die bei

einer Zustandstransformation erhalten bleiben, übertragen sich von der n -fachen Integratorkette auf das nichtlineare System.

Exakte Eingangs-Zustands-Linearisierung (EZL)

Entspricht der relative Grad der Systemordnung ($r = n$), so entstehen bei der exakten Linearisierung ebenso viele Gleichungen wie Zustandsgleichungen des ursprünglichen Systems. Die Systemdynamik wird in Folge der eindeutigen Zustandstransformation $\mathbf{z} = \varphi(\mathbf{x})$ vollständig durch das lineare Ersatzsystem beschrieben. Es wird von einer *Exakten Eingangs-Zustands-Linearisierung* (EZL) gesprochen.

Der Wirkungsplan einer EZL ist in Bild 17-1 dargestellt. Das transformierte System entspricht einer Integratorkette und liegt zusätzlich auch in Regelungsnormalform vor.

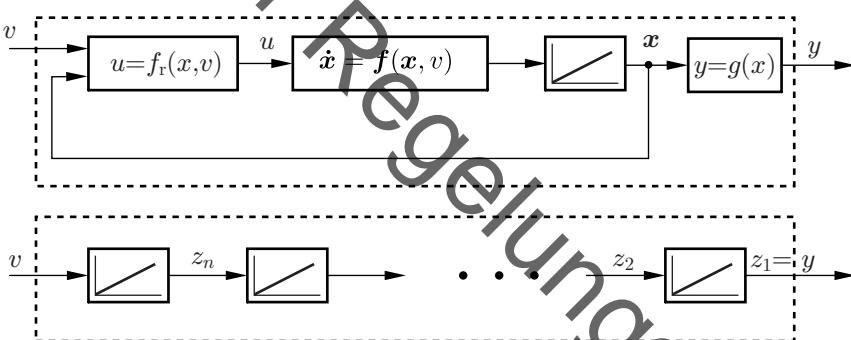


Bild 17-1: Wirkungsplan einer exakten Eingangs-Zustands-Linearisierung

Ist der relative Grad jedoch kleiner als die Systemordnung ($r < n$), so hat der Vektor \mathbf{z} weniger Einträge als \mathbf{x} . Die Transformation $\mathbf{z} = \varphi(\mathbf{x})$ ist nicht eindeutig, da verschiedene \mathbf{x} auf das gleiche \mathbf{z} abgebildet werden. Nur die in der Transformation enthaltenden Zustände werden linearisiert.

Exakte Eingangs-Ausgangs-Linearisierung (EAL)

Ist der relative Grad kleiner als die Systemordnung ($r < n$), so entstehen bei der exakten Linearisierung weniger Gleichungen als Zustandsgleichungen des ursprünglichen Systems. Die Systemdynamik wird in Folge

der nicht eindeutigen Zustandstransformation $z = \varphi(x)$ nicht vollständig durch das lineare Ersatzsystem beschrieben. Es wird von einer *Exakten Eingangs-Ausgangs-Linearisierung* (EAL) gesprochen.

Der Zustandsraum ist mit z also nicht vollständig, sondern muss für eine vollständige Beschreibung des Systems um zusätzliche Zustände erweitert werden. Da die Zustandstransformation bei der EAL nur r Zustände eindeutig transformiert, sind noch $n - r$ Zustände η frei wählbar. Diese können bei einer EAL nicht so gewählt werden, dass das Gesamtsystem in eine reine Integratorkette überführt wird. Sie müssen unabhängig von den r festgelegten Zuständen z sein, damit $[z \ \eta]^T$ dieselbe Beschreibungskraft wie x besitzt und die Zustandstransformation umkehrbar ist.

Interne und externe Dynamik

Die Dynamik der zusätzlichen Zustände $\eta \in \mathbb{R}^{n-r}$, die benötigt werden, um den Zustandsraum $x \mapsto [z \ \eta]^T$ zu vervollständigen, wird *interne Dynamik* genannt. In Abgrenzung davon spricht man bei der Dynamik in z von der *externen Dynamik*.

Bei der Wahl von η genießt man gewisse Freiheiten. Die Zustände der internen Dynamik müssen die Zustandstransformation eindeutig machen und daher unabhängig voneinander und von z sein. Da die Dynamik von y durch z bereits vollständig beschrieben wird, hat folglich η keinen sichtbaren Einfluss auf das Ein-Ausgangsverhalten. Daher führt auch die Bezeichnung als „interne Dynamik“.

Dennoch ist die Wahl von η nicht rein willkürlich, sondern es gibt geschicktere und ungeschicktere Wahlmöglichkeiten, für die auf [2] verwiesen wird.

Das exakt linearisierte System zerfällt bei der EAL in eine Blockstruktur mit r Zuständen z für die lineare externe Dynamik und $n - r$ Zuständen η für die im Allgemeinen nichtlineare interne Dynamik. Zur Darstellung dieses Struktur bedient man sich der *Byrnes-Isidori¹-Normalform*.

¹ Alberto Isidori (*1942), italienischer Regelungstechniker [21]

Byrnes-Isidori-Normalform (BINF)

Die Darstellung eines exakt linearisierten System in der Form

$$\begin{aligned} \dot{\tilde{z}} &= \begin{bmatrix} 0 & 1 & 0 & \dots \\ \vdots & \ddots & \ddots & \\ & & 0 & 1 \\ 0 & \dots & 0 & 0 \end{bmatrix} z + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} v \\ \dot{\eta} &= h(z, \eta, v) \\ y &= [1 \ 0 \ \dots \ 0]^T z \end{aligned} \quad (17.10)$$

heißt *Byrnes-Isidori-Normalform (BINF)*.

Der Wirkungsplan einer EAL ist in Bild 17-2 dargestellt. Das System zerfällt in eine Integratorkette und einen Parallelzweig der internen Dynamik, welcher von z und v gespeist wird. Bei geschickter Wahl von η ist es dabei möglich, den Einfluss von v auf $\dot{\eta}$ zu eliminieren [49].

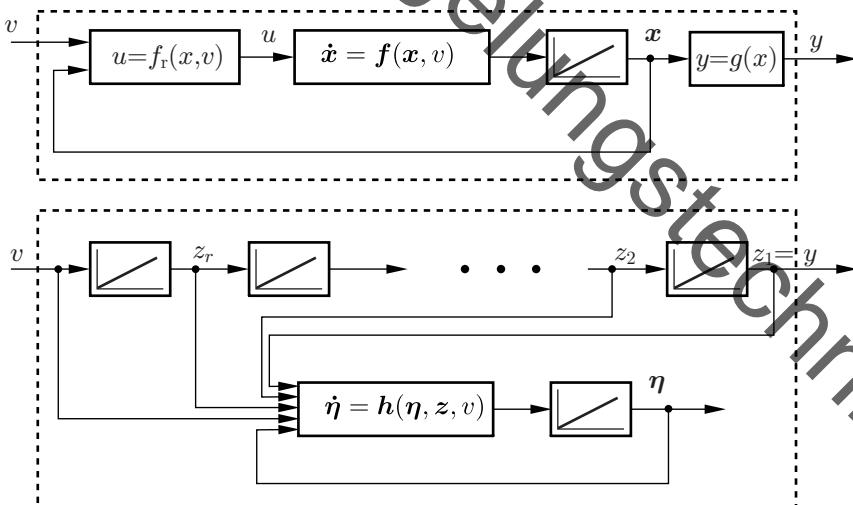


Bild 17-2: Wirkungsplan einer exakten Eingangs-Ausgangs-Linearisierung

Die Exakte Linearisierung soll für das erste Beispielsystem in Gl.(17.4) berechnet werden. Aus Gl.(17.6) ergibt sich aufgelöst nach u

$$u(-1 - 2x_2) - 2x_2^3 - 2x_2x_3 \stackrel{!}{=} v \Rightarrow u = -\frac{2x_2^3 + 2x_2x_3 + v}{1 + 2x_2}. \quad (17.11)$$

Die Wahl dieses Eingangs u prägt dem System das Verhalten $\ddot{y} = v$ auf. Solange $x_2 \neq -0,5$ gilt, ist $n = r$ und das entstandene System eine EZL. Mit den neuen Koordinaten $\mathbf{z} = [y \ \dot{y} \ \ddot{y}]^T$ und dem v aus Gl.(17.11) ergibt sich dann die neue Zustandsraumdarstellung

$$\dot{\mathbf{z}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{z} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \mathbf{v}, \quad y = z_1. \quad (17.12)$$

Das neue System ist offenbar wesentlich einfacher als das nichtlineare System zu regeln, da die Methoden der linearen Regelungstheorie direkt angewendet werden können. Außerdem ist zu erkennen, dass es nicht möglich ist, y völlig frei zu wählen. Stattdessen können nur r -fach differenzierbare Verläufe von y realisiert werden. Dies stellt für praktische Regelungszwecke aber keine Einschränkung dar.

Allerdings ist die Exakte Linearisierung (wie zuvor der relative Grad) nur für $x_2 \neq -0,5$ wohldefiniert. Für $x_2 = -0,5$ würde das Stellgesetz eine unendliche große Stellgröße anfordern, was für den praktischen Einsatz unbrauchbar ist. In der Interpretation als Zustandstransformation verliert die Exakte Linearisierung hier ihre Eindeutigkeit.

Es stellt sich anhand des Beispiels die Frage, ob allgemeine Bedingungen angegeben werden können, die die Existenz einer Exakten Linearisierung – vor allem einer EZL – garantieren.

17.3 Flachheit

Die Berechnung der Exakten Linearisierung erfolgt durch das Ableiten des Systemausgangs y . Da y als gemessener Systemausgang durch die Platzierung von Sensoren verändert werden kann, ist y im gewissen Rahmen frei wählbar. Die Wahl von y spielt aber eine entscheidende Rolle dabei, ob eine gefundene Exakte Linearisierung eine EZL oder eine EAL ist.

Es stellt sich daher die Frage, welche systemtheoretische Eigenschaft sicherstellt, dass es einen möglichen Systemausgang gibt, für den eine EZL gefunden werden kann. Diese Systemeigenschaft ist die *Flachheit* eines Systems, welche eine vollständige Invertierung der (nichtlinearen) Zustandsgleichungen erlaubt. Sie ist daher für den Entwurf nichtlinearer Regler und Steuerungen von herausragender Bedeutung. Die Bezeichnung „Flachheit“ geht dabei auf differentialgeometrische Bezüge zurück.

Flachheit

Ein allgemeines nichtlineares, dynamisches SISO-System

$$\dot{x} = f(x, u) \quad (17.13)$$

mit Zustandsgrößen $x \in \mathbb{R}^n$ und Stellgröße u heißt *flach*, wenn ein (fiktiver) Ausgang γ definiert werden kann, für den die folgenden Bedingungen erfüllt sind:

- i) Der flache Ausgang γ hängt nur von den Zuständen x , dem Eingang u und einer endlichen Anzahl ε an zeitlichen Ableitungen des Eingangs ab:

$$\gamma = \psi_0(x, u, \dot{u}, \dots, u^{(\varepsilon)}). \quad (17.14)$$

- ii) Alle Zustände lassen sich in Abhängigkeit des flachen Ausgangs γ und seiner ersten $n - 1$ Ableitungen darstellen:

$$x = \psi_1(\gamma, \dot{\gamma}, \dots, \gamma^{(n-1)}). \quad (17.15)$$

- iii) Der Eingangsvektor u taucht in den zeitlichen Ableitungen von γ in auflösbarer Form auf:

$$u = \psi_2(\gamma, \dot{\gamma}, \dots, \gamma^{(n)}). \quad (17.16)$$

Die Ausgangsgröße γ wird als linearisierender oder *flacher Ausgang* bezeichnet, das System als *flach*.

Sind die obigen Bedingungen nur lokal erfüllt, so wird auch von einer lokalen Flachheit des Systems gesprochen.

Je nach System stellt das Finden eines flachen Ausgangs eine hohe Herausforderung dar, sodass oft nur heuristische Vorgehensweisen genutzt werden können. Im Falle von Mehrgrößensystemen kommt noch eine vierte Bedingung für Flachheit hinzu, die im Eingrößenfall automatisch erfüllt ist [2].

Eine genaue Betrachtung der Anforderungen an Flachheit zeigt, dass gemäß Gl.(17.15) und Gl.(17.16) sowohl der Zustand \mathbf{x} als auch der Eingang u vollständig über γ und seine Ableitungen parametrisiert werden können – das ist genau die Interpretation der EZL als Zustandstransformation. Hier stellt Gl.(17.14) die Umkehrbarkeit dieser Transformation sicher. Tatsächlich lässt sich der folgende Satz für beliebige SISO-Systeme beweisen:

Flachheit und EZL

Ein System ist genau dann flach mit flachem Ausgang γ , wenn die exakte Linearisierung für γ wohldefiniert und eine EZL ist.

Für jedes flache System lässt sich folglich eine EZL finden, indem man den flachen Ausgang γ als Systemausgang nutzt. Es gilt dann $v = \gamma^{(n)}$ und γ mit seinen Ableitungen beschreibt als neuer Zustand \mathbf{z} das System eindeutig. Durch eine passende Vorgabe von v kann das System dann aus jedem Startzustand \mathbf{z}_1 mit einer genug stetig differenzierbare Trajektorie in einen beliebigen Zielzustand \mathbf{z}_2 überführt werden.

Da die Zustandstransformation $\mathbf{z} = \varphi(\mathbf{x})$ eindeutig ist, gilt dasselbe auch für den Startzustand \mathbf{x}_1 und den Zielzustand \mathbf{x}_2 . Die Zustandstrajektorien müssen lediglich die Bedingungen

$$\begin{aligned}\mathbf{x}_1 &= \psi_1(\gamma_d(t_0), \dots, \gamma_d^{(n-1)}(t_0)), \\ \mathbf{x}_2 &= \psi_1(\gamma_d(T_d), \dots, \gamma_d^{(n-1)}(T_d))\end{aligned}\tag{17.17}$$

gemäß Gl.(17.15) erfüllen, mit $t_0 \leq t \leq T_d$. Damit erfüllt jedes flache System aber die Definition von Steuerbarkeit.

Steuerbarkeit und Flachheit

Jedes flache System ist steuerbar bzw. die Flachheit eines Systems ist hinreichend für dessen Steuerbarkeit. Für lineare Systeme gilt auch die Umkehrung, dass jedes lineare steuerbare System flach ist.

Flachheit kann insofern als eine Art Verallgemeinerung des Steuerbarkeitsbegriffs für nichtlineare Systeme verstanden werden. Daher ist es interessant

zu untersuchen, welche Auswirkungen nicht steuerbare Systeme auf die exakte Linearisierung haben können. Hierzu wird das lineare Beispielsystem

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & -1 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} u \quad , \quad y = x_1 \quad (17.18)$$

betrachtet, das dem Beispielsystem in Gl.(17.4) strukturell ähnelt. Das System ist nicht steuerbar, da

$$[\mathbf{b} \quad \mathbf{Ab} \quad \mathbf{A}^2\mathbf{b}] = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad (17.19)$$

linear abhängige Zeilen besitzt. Da das System linear und nicht steuerbar ist, ist es auch nicht flach. Folglich gibt es keine Exakte Linearisierung – egal welchen Ausgang γ man betrachtet. Versucht man dennoch, diese mit dem gegebenen y zu berechnen, so erhält man

$$\begin{aligned} y &= x_1 \\ \dot{y} &= \dot{x}_1 = x_3 - x_2 \\ \ddot{y} &= \dot{x}_3 - \dot{x}_2 = u + x_2 - x_3 - u = -\dot{y} \end{aligned} \quad (17.20)$$

und u taucht nie in y oder seinen Ableitungen auf. Die Bedingung aus Gl.(17.16) für Flachheit ist nicht erfüllt und der relative Grad ist nicht wohldefiniert. Für andere Ausgänge wie $y = x_2$ gibt es einen wohldefinierten Grad. Dieser wird allerdings nie $r = n$ erreichen können, sodass nur eine EAL möglich ist.

Aus der Theorie linearer Systeme ist bekannt, dass jede minimale Realisierung steuerbar ist. Somit trifft diese Eigenschaft auf die allermeisten linearen Regelstrecken zu. Tatsächlich gilt Ähnliches auch für die Flachheit – zumindest lokal. So lässt sich für eine Vielzahl realer Anwendungen eine flache Systembeschreibung finden, die zumindest in einem gewissen Arbeitsbereich wohldefiniert mit umkehrbaren Transformationen ist. Beispiele für flache Systeme sind serielle Manipulatoren in der Robotik, Kransysteme, Drohnen oder auch bestimmte Fahrzeugmodelle. Die Einschränkungen des Gültigkeitsbereichs sind dabei meist verkraftbar. Bei Drohnen besitzt

die flache Systembeschreibung beispielsweise nur für einen aufrechten Flug (und nicht kopfüber) volle Gültigkeit.

Obgleich viele Systeme flach sind, so ist der flache Ausgang oft nicht die reale Regelgröße des Prozesses ($y = \gamma$). Dies ist aus einer linearen Betrachtung recht einleuchtend. Flachheit ist dort gleichbedeutend mit Steuerbarkeit – also sind lineare Regelstrecken normalerweise flach. Bei einer EZL gilt aber $r = n$, d. h. der relative Grad entspricht dem Systemgrad. Folglich darf ein lineares System mit flachem Ausgang $y = \gamma$ keine Nullstellen aufweisen und ist daher ein reines Verzögerungsglied. Diese Eigenschaft ist für wesentlich weniger Regelstrecken gegeben.

Dieser scheinbare Widerspruch ist darin begründet, dass für Flachheit irgendein Ausgang γ existieren muss, der zur EZL führt, d. h. dass keine Nullstellen vorliegen. Das bedeutet nicht, dass der tatsächliche Systemausgang diese Bedingung erfüllt. In der Praxis muss der flache Ausgang aus den messbaren Größen rekonstruiert werden oder in messbare Größen umgerechnet werden.

17.4 Normalformen

Einen einfachen Zugang zur Überprüfung der Flachheit bietet neben der Exakten Linearisierung auch die Verwendung von Normalformen. So konnte an der bereits eingeführten Byrnes-Isidori Normalform Gl.(17.10) direkt der relative Grad abgelesen werden, wodurch direkt erkennbar ist, ob y ein flacher Ausgang ist. Es gibt auch entsprechende Normalformen, aus der die Flachheit selbst direkt ersichtlich ist.

Aus der Regelung von linearen Systemen ist bekannt, dass bestimmte Zustandsraumdarstellungen für den Reglerentwurf besonders vorteilhaft sind. Hier ist an erster Stelle die Regelungsnormalform Gl.(2.25) zu nennen, die beispielsweise die Polplatzierung vereinfacht und stets steuerbar ist. Für nichtlineare Systeme kann ein entsprechendes nichtlineares Pendant identifiziert werden.

Nichtlineare Regelungsnormalform

Für ein SISO-System ist die *nichtlineare Regelungsnormalform* als

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} x_2 \\ x_3 \\ \vdots \\ x_n \\ \alpha(\mathbf{x}) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \beta(\mathbf{x}) \end{bmatrix} u \quad (17.21)$$

definiert.

Offenbar besitzt Gl.(17.21) eine zur Regelungsnormalform äquivalente Struktur. Vorteilhaft ist, dass sich die Exakte Linearisierung für $\beta(\mathbf{x}) \neq 0$ mit

$$u = \frac{-\alpha(\mathbf{x}) + v}{\beta(\mathbf{x})} \quad (17.22)$$

direkt hinschreiben lässt. Folglich ist jedes System in nichtlinearer Regelungsnormalform flach.

Viele technische Systeme werden allerdings nicht in Form von Gl.(17.21) vorliegen. Für andere Darstellungen reicht es für den Nachweis von Flachheit aber aus zu zeigen, dass sie sich in die *nichtlineare Regelungsnormalform* bringen lassen. Das ist insbesondere für die sogenannte *strenge Rückkopplungsform* der Fall.

Strenge Rückkopplungsform

Für ein SISO-System ist die *strenge Rückkopplungsform* als

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} \alpha_1(x_1) & + \beta_1(x_1)x_2 \\ \alpha_2(x_1, x_2) & + \beta_2(x_1, x_2)x_3 \\ \vdots & \vdots \\ \alpha_{n-1}(x_1, x_2, \dots, x_{n-1}) & + \beta_{n-1}(x_1, x_2, \dots, x_{n-1})x_n \\ \alpha_n(x_1, x_2, \dots, x_{n-1}, x_n) & + \beta_n(x_1, x_2, \dots, x_{n-1}, x_n)u \end{bmatrix} \quad (17.23)$$

definiert.

Die Gestalt der Zustandsgleichungen weist in dieser Normalform eine Art

untere Dreiecksform auf, da α_i und β_i nur von x_j mit $j \leq i$ abhängen. Die nichtlineare Regelungsnormalform ist zudem offenbar ein Spezialfall der strengen Rückkopplungsform für $\alpha_i = 0$ und $\beta_i = 1$ für $i = 1 \dots n - 1$.

Man kann zeigen, dass sich die strenge Rückkopplungsform für $\beta_i \neq 0$ in die nichtlineare Regelungsnormalform transformieren lässt. Beide Darstellungen beschreiben Systeme, die zu der in der Praxis relevanten Unterklasse der sogenannten *eingangslinearen* oder auch *eingangsaaffinen* Systeme gehören.

Eingangsaaffine Systeme

Ein SISO-System heißt *eingangsaaffin*, wenn es linear im Eingang u ist, d. h. es gilt

$$\dot{x} = f(x, u) = \alpha(x) + \beta(x)u \quad . \quad (17.24)$$

Die allermeisten technischen Systeme gehören zur Klasse der eingangsaaffinen Systeme. Für diese große Systemklasse reicht es tatsächlich aus, für den Nachweis von Flachheit (und damit der Auffindbarkeit einer EZL) zu untersuchen, ob sich das System in nichtlinearer Regelungsnormalform schreiben lässt.

Flachheit eingangsaaffiner Systeme

Ein eingangsaaffines System ist genau dann flach, wenn es sich in die nichtlineare Regelungsnormalform mit $\beta(x) \neq 0$ transformieren lässt.

Dies ist eine stärkere Aussage als zuvor: Zwar ist jedes in nichtlinearer Regelungsnormalform transformierbare System flach, aber nicht jedes flaches System lässt sich in nichtlineare Regelungsnormalform transformieren. Ein Gegenbeispiel hierfür ist $\dot{y} = u^3$. Schränkt man die Klasse der nichtlinearen Systeme aber auf eingangsaaffine ein, so gilt die Äquivalenz zwischen Flachheit und möglicher Darstellung in nichtlinearer Regelungsnormalform.

17.5 Flachheitsbasierte Steuerung und Regelung

17.5.1 Entwurf für flache Systeme

Flache Systeme lassen sich im Allgemeinen leichter als beliebige nichtlineare Prozesse in einen gewünschten Zielzustand überführen. Dies liegt in der Invertierbarkeit der Systemgleichungen Gl.(17.14) bis Gl.(17.16) begründet.

Dieser Zusammenhang kann genutzt werden, um besonders performante Steuergesetze abzuleiten.

Liegt eine gewünschte Solltrajektorie $\gamma_d(t)$ des flachen Ausgangs derart vor, dass eine hinreichende Anzahl von n stetigen Ableitungen gebildet werden kann, so lassen sich die zum Einprägen der Trajektorie notwendigen Stellgrößen u direkt – ohne das Lösen von Differentialgleichungen – berechnen. Entsprechen die gewählten flachen Ausgänge dabei den Regelgrößen des Systems ($y = y$), kann ein exaktes Folgeverhalten erzielt werden. Dies ergibt sich direkt aus der Definition der Flachheit nach Gl.(17.16): Die Stellgröße muss entsprechend zu

$$u = \psi_2 \left(y_d, \dots, y_d^{(n)} \right) \quad (17.25)$$

mit der Solltrajektorie y_d gewählt werden.

Dieser Entwurf einer idealen Steuerung wird auch als flacher Vorsteuерungsentwurf bezeichnet. Praktisch funktioniert dieser allerdings nur für sehr genau modellierbare Strecken, auf die keine äußeren Störungen einwirken. Zudem ist das Vorgehen als reine Steuerung nur für stabile Strecken anwendbar, da ansonsten bereits kleine Störungen eine Divergenz der Systemtrajektorie zur Folge hätten.

Daher wird die flachheitsbasierte Steuerung in der Praxis meist als Vorsteuerung genutzt und um einen stabilisierenden Regler erweitert. Für flache Systeme kann in diesen Fällen ein Reglerentwurf auf Basis der exakten Linearisierung erfolgen.

Die entsprechenden Reglerentwürfe passen sowohl für SISO- als auch MIMO-Systeme. Aufgrund der kompakteren Darstellung seien im Folgenden jedoch nur SISO-Systeme betrachtet. Eine Erweiterung auf Mehrgrößensysteme kann beispielsweise [2] entnommen werden.

Nach Abschnitt 17.1 können die Systemgleichungen eines Eingangs-Zustands linearisierbaren Systems unter der Transformation $z = (y, \dot{y}, \dots, y^{(n-1)}) = \varphi(x)$ aus Gl.(17.9) in die lineare Darstellung

$$\dot{z} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & 1 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix} z + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} v \quad (17.26)$$

überführt werden. Bezuglich des neuen Eingangs v liegt das System in der linearen Regelungsnormalform vor. Da bei einer EZL der relative Grad der Systemordnung entspricht, wird das nichtlineare System zudem vollständig durch das lineare Ersatzsystem beschrieben. Eine Stabilisierung des Systems kann daher in bekannter Weise durch eine (in v lineare) Zustandsrückführung

$$v = -\mathbf{k}^T \mathbf{z} = -[k_1 \ k_2 \ \dots \ k_n] \mathbf{z} \quad (17.27)$$

erfolgen. Die Koeffizienten der Rückführvektors \mathbf{k} können dabei wortgleich nach den Methoden der linearen Regelungstheorie gewählt werden. Das nichtlineare Stellgesetz folgt unter Rücktransformation der virtuellen Zustände Gl.(17.9) zu

$$u = f_r \left(\mathbf{x}, v = -\mathbf{k}^T \varphi(\mathbf{x}) \right) = f_r \left(\mathbf{x}, v = -\sum_{i=1}^n k_i \frac{d^{i-1}}{dt^{i-1}} g(\mathbf{x}) \right). \quad (17.28)$$

Wird ein flacher Ausgang betrachtet, kann das Stellgesetz um die flache Vorsteuerung Gl.(17.25) erweitert werden.

Da sowohl die Regelung als auch die Steuerung ein (transformiertes) lineares System zum Entwurf verwenden, sind alle Vorteile und Aussagen aus Abschnitt 12.2 auch hier gültig. Folglich können so hochgenaue Folgeregelungen mit gutem Ansprechverhalten realisiert werden. Dies resultiert in der in Bild 17-3 dargestellten Regelkreisstruktur.

Hierbei wird der virtuelle Eingang v zu

$$v = \underbrace{y_d^{(n)}}_{\text{Vorsteuerung}} - \underbrace{\sum_{i=1}^n k_i (z_i - y_d^{(i-1)})}_{\text{Regelung}} \quad (17.29)$$

gewählt. Während der erste Summand eine reine Steuerung $y^{(n)} = y_d^{(n)}$ darstellt, führt der zweite Summand Abweichungen von der Solltrajektorie zurück. Das zugehörige Stellgesetz ergibt sich analog zum Vorgehen ohne Vorsteuerung Gl.(17.28).

Für den speziellen Fall eines Systems in nichtlinearer Regelungsnormalform

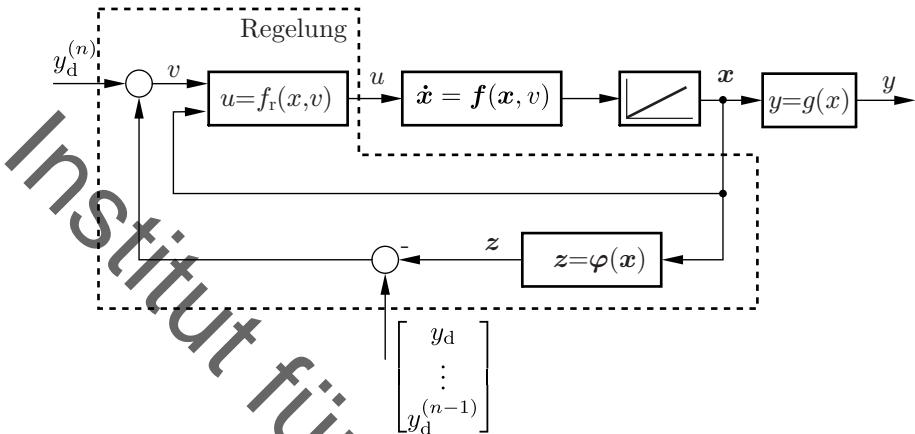


Bild 17-3: Struktur mit EZL, flacher Vorsteuerung und Rückführung

folgt die Stellgröße zu

$$u = -\frac{1}{\beta(\mathbf{x})} \left(\alpha(\mathbf{x}) - y_d^{(n)} + \sum_{i=1}^n k_i \left(\frac{d^{i-1}}{dt^{i-1}} g(\mathbf{x}) - y_d^{(i-1)} \right) \right) \quad (17.30)$$

17.5.2 Beispiel

Das geschilderte Vorgehen soll an dem folgenden Beispielsystem verdeutlicht werden, das bereits in nichtlinearer Regelungnormalform vorliegt:

$$\dot{x}_1 = x_2 \quad , \quad \dot{x}_2 = -\sin(x_1) + u \quad , \quad y = x_1 \quad (17.31)$$

Das System soll vom Arbeitspunkt innerhalb einer Sekunde von $y = 0$ nach $y = \pi$ überführt werden.

Eine Stabilitätsuntersuchung über die Linearisierung zeigt, dass das linearisierte System in $y = \pi$ nicht stabil arbeitet. Da offenbar $u = \sin(x_1) + v$ eine EZL des Systems und $\gamma = y$ ein flacher Ausgang ist, gilt die Regelsatzstruktur in Bild 17-3.

Eine mögliche Solltrajektorie ergibt sich aus den Bedingungen

$$y(0) = 0 \quad , \quad \dot{y}(0) = 0 \quad , \quad y(1) = \pi \quad , \quad \dot{y}(1) = 0 \quad (17.32)$$

in Kombination mit einem polynomialen Ansatz, welcher die Bedingungen an Differenzierbarkeit erfüllt:

$$y_d(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 = 3\pi t^2 - 2\pi t^3 \quad . \quad (17.33)$$

Eine Polvorgabe nach -1 für das exakt linearisierte System ergibt das Stellgesetz

$$\ddot{y} = u = -2\dot{y} - y = -2x_2 - x_1 \quad . \quad (17.34)$$

Da der Systemausgang gleichzeitig ein flacher Ausgang ist, ergibt sich mit Vorsteuerung entlang der Trajektorie

$$\begin{aligned} \ddot{y} &= v = \ddot{y}_d - 2(\dot{y} - \dot{y}_d) - (y - y_d) \\ &= 6\pi - 12\pi t - 2(x_2 - 6\pi(t - t^2)) - (x_1 - 3\pi t^2 + 2\pi t^3) \end{aligned} \quad (17.35)$$

und damit das Stellgesetz

$$u = \sin(x_1) + v = \underbrace{\sin(x_1) - 2x_2 - x_1}_{\text{Regelung}} + \underbrace{6\pi - 9\pi t^2 - 2\pi t^3}_{\text{Steuerung}} \quad (17.36)$$

17.5.3 Umgang mit interner Dynamik

Ist der relative Grad kleiner als die Systemordnung ($r < n$), kann das System nicht vollständig linearisiert werden – es verbleibt eine interne Dynamik. Die aus der EAL resultierende BINF Gl.(17.10) kann nach Abschnitt 17.1 in eine steuerbare, lineare Komponente

$$\dot{z} = \mathbf{A}z + \mathbf{b}v \quad (17.37)$$

sowie einen unlinearisierten Anteil

$$\dot{\eta} = h(z, \eta, v) \quad (17.38)$$

unterteilt werden. Die externe Dynamik ist abermals linear, wodurch ein analoges Vorgehen zum Reglerentwurf mittels EZL möglich ist. Die virtuelle Stellgröße ist folglich gemäß

$$v = y_d^{(r)} - \sum_{i=1}^r k_i (z_i - y_d^{(i-1)}) \quad (17.39)$$

gegeben, wobei lediglich Ausgangableitungen entsprechend des relativen Grads r berücksichtigt werden. Anders als im Falle der EZL ist das nicht-lineare System nicht weiter über das lineare Ersatzsystem Gl.(17.37) bestimmt.

Für die Stabilität des Regelkreises müssen auch die Systemtrajektorien der internen Dynamik berücksichtigt werden. Eine allgemeingültige Aussage ist meist schwierig zu treffen, so dass auf eine lokale Analyse der internen Dynamik zurückgegriffen wird.

Wird die Zustandsrückführung Gl.(17.39) derart ausgelegt, dass die externe Dynamik stabilisiert wird, verschwindet der Folgefehler

$$\mathbf{z}_d - \mathbf{z} = \begin{bmatrix} y_d - y & \dots & y_d^{(r-1)} - y^{(r-1)} \end{bmatrix}^T \xrightarrow{t \rightarrow \infty} \mathbf{0} \quad (17.40)$$

über der Zeit. Ab diesem Zeitpunkt genügt eine Stabilitätsbetrachtung des Systems

$$\dot{\boldsymbol{\eta}} = \mathbf{h}(\mathbf{z}_d, \boldsymbol{\eta}, v(\mathbf{z}_d)) \quad , \quad (17.41)$$

welche als *Nulldynamik* bezeichnet wird.

Nulldynamik

Die interne Dynamik $\boldsymbol{\eta}$ ist auch von der externen Dynamik \mathbf{z} und v abhängig. Setzt man für die externe Dynamik den Sollzustand ein, so vereinfacht sich die interne Dynamik zu Gl.(17.41). Sie ist nicht mehr von v und \mathbf{z} abhängig und wird *Nulldynamik* genannt.

Die Nulldynamik vereinfacht sich für das Stellgesetz in Gl.(17.39) zu

$$\dot{\boldsymbol{\eta}} = \mathbf{h}\left(\mathbf{z}_d, \boldsymbol{\eta}, y_d^{(r)}\right) \quad . \quad (17.42)$$

Die Nulldynamik hat entschiedenen Einfluss auf die Stabilität des entworfenen Regelkreises. Es lässt sich folgender Satz beweisen:

Stabilität von Regelungen über EZL und EAL

Für einen Reglerentwurf über EZL gilt: Der Regelkreis ist genau dann global stabil, wenn das exakt linearisierte System stabil ist.

Für einen Reglerentwurf über EAL gilt: Der Regelkreis ist genau dann

lokal stabil, wenn das exakt linearisierte System stabil ist und die Nulldynamik lokal stabil ist.

Da die Nulldynamik nicht von der Ausgestaltung der Vorsteuerung oder Regelung abhängig ist, solange diese den Sollzustand z_d einstellt, beeinflusst die Wahl von \mathbf{k}^T nicht die Stabilität der Nulldynamik. Somit kann die lokale Stabilität der Nulldynamik als Voraussetzung für die Anwendbarkeit eines flachheitsbasierten Regelentwurfs für Systeme ohne EZL aber mit EAL verstanden werden.

Für die Erläuterung des Namens „Nulldynamik“ – und gleichzeitig eine Erweiterung des Regelungstechnischen Horizontes – ist ein Ausflug in die lineare Systemtheorie notwendig. Dies soll an einem Beispiel erfolgen, welches gleichzeitig das Aufstellen der EAL veranschaulicht.

Es wird das lineare System

$$\dot{x}_1 = x_2 - u \quad , \quad \dot{x}_2 = -x_1 - 2x_2 + 3u \quad , \quad y = x_1 \quad (17.43)$$

betrachtet. Die Exakte Linearisierung berechnet sich aus $\dot{y} = x_2 - u$ direkt zu $u = x_2 - v$. Da der relative Grad eins beträgt, handelt es sich um eine EAL. Wegen $y = x_1$ ist der Zustand x_2 bei der Zustandstransformation nach $z = y$ nicht erfasst und wird über $\eta = x_2$ der internen Dynamik zugeschlagen. Für den Sollwert $y = 0$ folgt $u = x_2 = \eta$ und damit die Nulldynamik

$$\dot{\eta} = \dot{x}_2 = -x_1 - 2\eta + 3u = 0 - 2\eta + 3\eta = \eta \quad (17.44)$$

und damit ein Eigenwert bei eins. Die Nulldynamik ist folglich instabil und der Regelentwurf über die EAL nicht anwendbar.

Berechnet man die Übertragungsfunktion von Gl.(17.43), so gewinnt man

$$G(s) = \frac{Y(s)}{U(s)} = \frac{-s + 1}{s^2 + 2s + 1} \quad . \quad (17.45)$$

Offenbar findet sich die Nulldynamik mit dem Eigenwert bei eins in der Position der Nullstelle von $G(s)$ wieder. Das ist kein Zufall, sondern lässt sich allgemein beweisen [2].

Nulldynamik und Nullstellen

Bei linearen Systemen entsprechen die Eigenwerte der Nulldynamik den Nullstellen der Übertragungsfunktion.

Dieser Zusammenhang passt zu den bisherigen Beobachtungen des relativen Grades. Dieser entspricht im Linearen der Differenz von Nenner- und Zählergrad. Folglich ist die Dimension $n - r$ der internen Dynamik gleich der Anzahl der Nullstellen des Systems.

Die Nulldynamik kann folglich als eine Verallgemeinerung der Nullstellen für nichtlineare Systeme verstanden werden. Daher wird in der Literatur gelegentlich die Forderung nach einer lokal stabilen Nulldynamik mit der Vokabel der Minimalphasigkeit assoziiert. Davon wird hier Abstand genommen, da die Motivation über den Phasengang, der für nichtlineare Systeme nicht existiert, in die Irre führt.

Aus der Interpretation der Nulldynamik als nichtlineare Nullstellen lassen sich jedoch weitere Schlüsse auf die Wirkungsweise der exakten Linearisierung ziehen. Die Forderung, dass eine Regelstrecke nur Nullstellen mit negativem Realteil haben darf, ist aus dem perfekten Vorsteuerungsentwurf in Abschnitt 12.2 bekannt. Das lag an den Pol-Nullstellen-Kürzungen, die bei diesem Steuerungsentwurf vorgenommen werden.

Die Vermutung, dass auch die Exakte Linearisierung eine solche Kürzung in Nichtlinearen vornimmt, lässt sich durch eine Analyse der Beobachtbarkeit bestätigen. Betrachtet man den Wirkungsplan der Byrnes-Isidori-Normalform in Bild 17-2, so erkennt man, dass die interne Dynamik η einen parallelen Zweig zur Integratorkette $y^{(r)}$ bildet. Folglich wird man η aus der reinen Betrachtung von y und seinen Ableitungen nicht rekonstruieren können. Daher ist der mittels EAL geschlossene Regelkreis nicht beobachtbar.

Fehlende Beobachtbarkeit bedeutet im Linearen, dass eine Pol-Nullstellen-Kürzung in der Übertragungsfunktion vorliegt. Da die Regelstrecke selbst im Allgemeinen beobachtbar sein wird, entstammt diese der EAL. Die Pol-Nullstellen-Kürzung zeigt sich dabei auch in der von n auf r reduzierten Ordnung des Ein-Ausgangs-Verhaltens nach Anwendung der EAL.

Hieraus kann abgeleitet werden, dass die Reglerentwurfsverfahren über Exakte Linearisierung eine hohe Modellgüte voraussetzen. Bei Abweichungen zwischen tatsächlichem und modelliertem Systemverhalten wird die Exakte Linearisierung das geplante lineare Systemverhalten nicht einstellen. Die

fehlerhaft kompensierten Nullstellen können – insbesondere bei dominanten Nullstellen – zu langsamem oder schwach gedämpften Lösungsanteilen führen.

17.6 Integrator Backstepping

17.6.1 Reglerentwurf über Lyapunov

Der Reglerentwurf auf Basis einer exakten Linearisierung setzt eine hohe Modellgüte voraus. Nichtmodellierte Systemdynamiken sowie Parameterabweichungen reduzieren die Regelgüte und können den Regelkreises destabilisieren.

Es gibt zudem noch eine zweite Problematik, die am Beispiel des Systems

$$\dot{y} = -ay^3 + u \quad a > 0 \quad (17.46)$$

erläutert werden soll. Der Entwurf über exakte Linearisierung führt hier auf ein Stellgesetz $u = ay^3 - ky$ und damit auf ein lineares System $\dot{y} = -ky$. Dessen Stabilität kann mit linearen Methoden untersucht werden und man berechnet einen Eigenwert bei $-k$. Aufgrund des linearen Verhaltens im geschlossenen Kreis ist die Reglereinstellung vergleichsweise intuitiv.

Um diese vorteilhafte lineare Struktur zu erreichen, kompensiert der Regler durch den Term ay^3 die Systemdynamik $-ay^3$. Das ist aber problematisch, da $-ay^3$ ein stabilisierender dynamischer Anteil ist. Das lässt sich beispielsweise über die direkte Methode nach Lyapunov $u = 0$ mit

$$V(y) = y^2 \quad \Rightarrow \quad \dot{V}(y) = -2ay^4 < 0 \quad (17.47)$$

schnell nachweisen.

Folglich entfernt die Exakte Linearisierung auch stabile Anteile, sofern diese nichtlinear sind. Dies wird bei unbekannten Parametern besonders kritisch. Ist der stabile Anteil a weniger groß als erwartet, überkompensiert der Regler diesen und erzeugt ein möglicherweise sogar instabiles System.

Möchte man stabile nichtlineare Anteile nicht kompensieren sondern im Regelkreis behalten, so ist eine Regelung mittels Exakter Linearisierung nicht zielführend. Verzichtet man aber auf diese, kann Stabilitätsbeurteilung und Reglerauslegung nicht auf Basis linearer Kriterien erfolgen, da der geschlossene Regelkreis selbst nichtlinear ist.

In diesem Fall ist es zweckmäßig, auf die direkte Methode nach Lyapunov zurückzugreifen.

Lyapunov-Regler

Wird ein (nichtlinearer) Zustandsregler mittels der direkten Methode nach Lyapunov ausgelegt, so bezeichnet man den entstehenden Regler als *Lyapunov-Regler*.

Man bezeichnet hiermit folglich keine spezielle Reglerstruktur, sondern einer Art der Reglerauslegung. Formal gilt es für die Regelstrecke $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, u)$ ein Stellgesetz $u = h(\mathbf{x})$ zu finden, so dass für das autonome System $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, h(\mathbf{x}))$ Stabilität für die gewünschte Ruhelage \mathbf{x}_d nachgewiesen werden kann.

Das für eine gute Reglerauslegung zusätzlich zur Stabilität geforderte gute dynamische Verhalten muss dabei ebenfalls durch die gewählte Lyapunovfunktion festgelegt werden. Um das zu erreichen, verwendet man üblicherweise quadratische Kandidatenfunktionen $V(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_d)^T(\mathbf{x} - \mathbf{x}_d)$, die sowohl radial unbeschränkt, als auch positiv außerhalb der Ruhelage sind. Für die zeitliche Ableitung wird dann der Zusammenhang

$$\dot{V}(\mathbf{x}) \stackrel{!}{=} -2\tau V(\mathbf{x}) < 0 \quad \forall \mathbf{x} \neq \mathbf{0} \quad (17.48)$$

gefördert. Dies impliziert ein exponentielles Abklingen der Lyapunovfunktion, wobei über die Zeitkonstante $\tau \in \mathbb{R}^+$ die Aggressivität des Reglers eingestellt werden. Tatsächlich kann man schnell nachrechnen, dass dieser Ansatz für ein lineares System einer Vorgabe aller Pole nach $-\tau$ entspricht.

Dieses Entwurfsverfahren besitzt leider einige elementare Schwächen, von denen vor allem zwei zu nennen sind. Zum einen hängt das gefundene Stellgesetz stets von der gewählten Lyapunovfunktion ab. Auch wenn man diese auf einen quadratischen Kandidaten einschränkt, so spielt die Wahl des Zustands \mathbf{x} eine entscheidende Rolle. Aufgrund von Zustandstransformationen ergeben sich dann abweichende Kandidatenfunktionen mit teils erheblichem Unterschied in den Regelgesetzen – sowohl im Bezug auf deren Kompaktheit als auch deren Leistungsfähigkeit.

Zum anderen ist die einschränkendere Forderung $\dot{V} = -2\tau V$ oftmals gar nicht zu erfüllen, obgleich es strukturell genügen würde, $\dot{V} < 0$ zu garantieren.

Die Nachteile sollen an einem Beispiel illustriert werden. Es wird die (flache) Regelstrecke

$$\dot{x}_1 = -x_1^3 + x_1 + x_2 \quad , \quad \dot{x}_2 = u \quad , \quad y = x_1 \quad (17.49)$$

betrachtet, die im Nullpunkt stabilisiert werden soll.

Mit der quadratischen Kandidatenfunktion $V(x_1, x_2) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2$ berechnet sich

$$\dot{V} = x_1 \dot{x}_1 + x_2 \dot{x}_2 = -x_1^4 + x_1^2 + x_1 x_2 + x_2 u \stackrel{!}{=} -\tau(x_1^2 + x_2^2) \quad . \quad (17.50)$$

Die Gleichung ist im Allgemeinen nicht lösbar, da alle Terme in u mit x_2 multipliziert werden. Ein Teilen durch x_2 ist aber nicht erlaubt, da ja der Nullpunkt stabilisiert werden soll. Lässt man die Forderung $\dot{V} < -\tau V$ fallen und fordert nur $\dot{V} \leq 0$, so ist die Lösung für u dennoch sehr schwierig zu sehen.

Insgesamt ist dieser Ansatz für den Entwurf von Lyapunov-Reglern zu unsystematisch. Es bleibt unklar, ob V eine geeignete Kandidatenfunktion ist und mit welchen Methoden ein passendes u überhaupt gefunden werden kann.

17.6.2 Schrittweises Erstellen des Stellgesetzes

Abhilfe für die Schwächen der oben beschriebenen Methode bietet das sogenannte *Backstepping*-Verfahren, das über ein strukturelles und schrittweises Vorgehen die Lyapunov-Funktion und das Stellgesetz erstellt. Das Verfahren entwickelt ausgehend von einem flachen Ausgang $y = z_1$ sukzessive einen passenden Zustand z , welcher besonders günstig zur Konstruktion einer Lyapunov-Funktion ist. Im Gegensatz zur Exakten Linearisierung werden Stabilisierungsaufgabe und Zustandstransformation nicht als zwei separate Schritte aufgefasst, sondern gemeinsam vollzogen.

Dieses Vorgehen soll an einem Beispiel erläutert werden, wofür erneut das System in Gl.(17.49) herangezogen wird. Der flache Ausgang ist $y = z_1$, welcher als erster Zustand z_1 der neuen Backstepping-Koordinaten genutzt wird.

Wie soll nun der zweite Zustand z_2 gewählt werden? Würde man $z_2 = x_2$ wählen, so wäre die Dynamik für $x_2 = 0$ genau $\dot{x}_1 = -x_1^3 + x_1$ und damit

instabil im Ursprung. Das ist für die Lyapunov-Funktion schlecht, da diese nicht in zwei separate Terme für x_1 und x_2 aufgeteilt werden kann. Ziel ist es daher, den Zustand z_2 so zu wählen, dass die Dynamik von $x_1 = z_1$ für $z_2 = 0$ genau stabil ist.

Dies kann man erreichen, indem man den Beitrag der bisher nicht in z erfassten Zustände als (fiktive) Stellgröße \tilde{u} auffasst. Dieses \tilde{u} wählt man dann entsprechend für eine stabile Dynamik in z . Der neue Zustand folgt als Abweichung dieser gewünschten fiktiven Stellgröße von dem tatsächlichen Eintrag.

Im konkreten Beispiel ist x_2 nicht in $z_1 = x_1$ enthalten und wird daher als fiktive Stellgröße \tilde{u} aufgefasst:

$$\dot{z}_1 = -z_1^3 + z_1 + \underbrace{x_2}_{\equiv \tilde{u}}. \quad (17.51)$$

Für diese eindimensionale Regelstrecke muss ein Regler entworfen werden. Dies kann mit beliebigen Methoden, vorzugsweise aber über den Ansatz eines Lyapunov-Reglers geschehen. Man sieht, dass die Wahl von $\tilde{u} \stackrel{!}{=} -z_1$ wegen $\dot{V} = -z_1^4 < 0$ auf ein stabiles System führt. Nun entspricht aber $\tilde{u} = x_2$ nicht dieser Wunschgröße. Diese Abweichung wird der neue Zustand z_2 , der sich hier dann zu

$$x_2 = -z_1 + z_2 \quad (17.52)$$

ergibt. In den neuen Koordinaten $z_1 = x_1$ und $z_2 = x_2 + x_1$ ergibt sich dann das System

$$\begin{aligned} \dot{z}_1 &= -z_1^3 + z_2 \quad (\text{stabil, wenn } z_2 \rightarrow 0) \\ \dot{z}_2 &= -z_1^3 + z_2 + u \end{aligned} \quad (17.53)$$

Diese neuen Koordinaten haben den Vorteil, dass die Dynamik in z_1 für $z_2 \rightarrow 0$ stabil ist, was bei den x -Koordinaten nicht gegeben war. Der Lyapunov-Kandidat in z -Koordinaten führt dann auf

$$V(z_1, z_2) = \frac{1}{2}z_1^2 + \frac{1}{2}z_2^2 \quad \Rightarrow \quad \dot{V} = -z_1^4 + z_1 z_2 - z_2 z_1^3 + z_2^2 + z_2 u. \quad (17.54)$$

Wie in Gl.(17.50) tritt u nur im Produkt mit dem zweiten Zustand (hier z_2 , zuvor x_2) auf. Da aber die erste Zustandsgleichung für z_1 nun stabil ist,

stellt dies in Gl.(17.50) kein Problem dar, da alle Instabilität verursachenden Terme selbst einen Vorfaktor z_2 tragen. Ein stabilisierendes u lässt sich daher aus Gl.(17.54) leicht zu bspw.

$$u = -z_1 + z_1^3 - 2z_2 \quad (17.55)$$

bestimmen.

Dieses skizzierte Vorgehen lässt sich allgemein im folgenden Algorithmus fassen:

Ablaufschema des Integrator-Backsteppings

- 1) Wählen von $z_1 = y$ mit einem (nach Möglichkeit) flachen y und $i = 1$
- 2) Aufstellen der Differentialgleichung für z_i .
- 3) Auffassen der Größen, die nicht z sind, als fiktiven Eingang \tilde{u} .
- 4) Entwerfen eines Reglers für \tilde{u} (z. B. mittels Lyapunovfunktion).
- 5) Aufstellen des Fehlerterms z_{i+1} im fiktiven Eingang und weiter bei 2) mit $i = i + 1$.

Aus der Beschreibung des Ablaufschemas folgt, dass sich dieses Verfahren besonders einfach auf Systeme in der strengen Rückkopplungsform Gl.(17.23) anwenden lässt. Dies liegt daran, dass jede Zeile \dot{z}_i nur von einem weiteren Zustand x_{i+1} abhängt, welcher als fiktive Stellgröße dienen kann.

Die Stellgröße u taucht in der letzten Gleichung auf und das Stellgesetz kann dann nach u aufgelöst werden. Dies ist günstig für die Anwendbarkeit des Backstepping-Verfahrens: Taucht u zu früh auf, bleiben Zustandsgleichungen ungeregelt und es entsteht eine interne Dynamik. Hier gelten dann analoge Stabilitätsbedingungen wie für den Reglerentwurf mittels BAL.

Da sich alle flachen eingangsaffinen Systeme in nichtlineare Regelungsform und diese in die strenge Rückkopplungsform überführen lassen, ist das Verfahren dabei allgemein auf alle eingangsaffinen und flachen Systeme anwendbar.

Die Bezeichnung als „Integrator-Backstepping“ wird klarer, wenn man den Wirkungsplan betrachtet, der bei der Anwendung des Verfahrens auf ein System in strenger Rückkopplungsform entsteht. Beginnend bei der Ausgangsgröße $z_1 = y$ wird gemäß Schritt 2 des Algorithmus nach der Ableitung

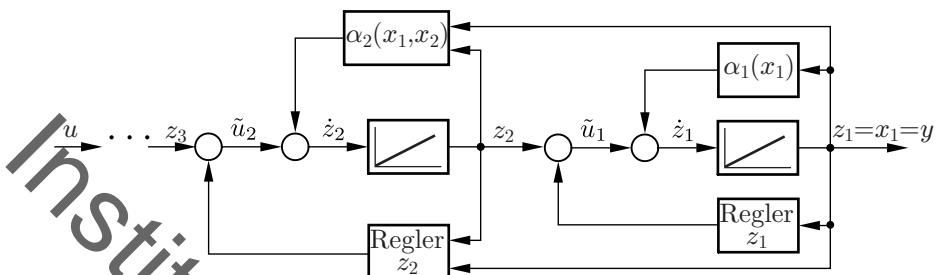


Bild 17-4: Integrator-Backstepping für die strenge Rückkopplungsform

\dot{z}_1 und damit dem Eingang in einen Integrator gesucht, dessen Ausgang z_1 ist. Es wird also dieser Integrator „rückwärts“ entgegen der eigentlichen Ursache-Wirkungs-Richtung durchlaufen, woher das Verfahren seinen Namen hat. Dieses rückwärtige Durchlaufen der Integratoren endet erst beim Auftauchen der Stellgröße u .

Da die virtuelle Stellgröße \tilde{u} in weiten Bereichen frei wählbar ist, kann das Backstepping-Verfahren für ein gegebenes System auf unterschiedliche Regelgesetze führen. Dabei sind allgemeinere Reglerentwürfe als bei der Exakten Linearisierung möglich. Jeder über Exakte Linearisierung und Polvorgabe entworfene Regler lässt sich dabei prinzipiell auch über das Backstepping-Verfahren umsetzen, indem man in den Einzelschritten gemäß Gl.(17.48) eine definierte lineare Systemdynamik erzwingt.

17.7 Sliding Mode Regelung

17.7.1 Der eindimensionale Fall

Die Ideen zum flachheitsbasierten Reglerentwurf aus Abschnitt 17.1 lassen sich mit dem Konzept der Lyapunov-Regler aus Abschnitt 17.6 dergestalt kombinieren, dass Unsicherheiten in der Modellbeschreibung besonders effektiv berücksichtigt werden können. Der prominenteste Vertreter dieser Idee ist die sogenannte *Sliding Mode Regelung*. Im Folgenden wird das Verfahren lediglich für Systeme mit einer Stellgröße u betrachtet, es sei allerdings angemerkt, dass sich die Ergebnisse auf Systeme mit mehreren Stellgrößen u_i übertragen lassen (siehe [54]).

Aus Gründen der Übersichtlichkeit wird von einer nichtlinearen Streckenbeschreibung in nichtlinearer Regelungsnormalform

$$\begin{aligned}\dot{x}_i &= x_{i+1} \quad \text{mit } i = 1, \dots, n-1 \\ \dot{x}_n &= \alpha(\mathbf{x}) + \beta(\mathbf{x})u\end{aligned}\tag{17.56}$$

ausgegangen. Es wird angenommen, dass die wahren Funktionen $\alpha(\mathbf{x})$ und $\beta(\mathbf{x})$ nicht exakt bekannt sind. Stattdessen ist nur ein Modell mit Funktionen $\bar{\alpha}(\mathbf{x})$ und $\bar{\beta}(\mathbf{x})$ vorhanden, welches sich durch Modellfehler $\Delta\alpha(\mathbf{x})$ und $\Delta\beta(\mathbf{x})$ von der wahren Regelstrecke unterscheidet.

Unsicherheiten und nominelles Modell

Ist die Beschreibung f einer Regelstrecke nicht exakt bekannt, so kann diese zerlegt werden in ein bekanntes Modell \bar{f} , das *nominelles Modell* genannt wird, sowie den Modellfehler Δf , der *Unsicherheit* genannt wird.

Für einen sinnvollen Reglerentwurf müssen die Unsicherheiten abgeschätzt werden können. Im folgenden wird von einer Abschätzung in der Form

$$\begin{aligned}\alpha(\mathbf{x}) &= \bar{\alpha}(\mathbf{x}) + \Delta\alpha(\mathbf{x}) \quad \text{mit } |\Delta\alpha(\mathbf{x})| \leq A(\mathbf{x}) \\ \beta(\mathbf{x}) &= \bar{\beta}(\mathbf{x}) + \Delta\beta(\mathbf{x}) \quad \text{mit } B_{\min} \leq \Delta\beta(\mathbf{x}) \leq B_{\max}\end{aligned}\tag{17.57}$$

ausgegangen.

Die nun betrachteten Regelung zielt darauf ab, die Regelabweichung mit einer vorgegebenen Fehlerdynamik trotz Modellunsicherheit zu eliminieren. Da die Unsicherheit im genauen Zahlenwert unbekannt ist, kann dabei nur das nominelle Modell $\bar{\alpha}, \bar{\beta}$ und die Abschätzung der Unsicherheit A, B für den Entwurf verwendet werden.

Die Beschreibung des Vorgehens teilt sich nun in zwei Bereiche: Zunächst wird die Regelung für ein System mit eindimensionaler Dynamik betrachtet. Im zweiten Teil wird gezeigt, wie man eine mehrdimensionale Dynamik auf eine Dimension und damit den ersten Fall reduzieren kann.

Für den ersten Teil geht man von einer eindimensionalen Dynamik in der Variablen $s(t)$ aus:

$$\dot{s} = \alpha(s) + \beta(s)u \quad \text{mit } \alpha = \bar{\alpha} + \Delta\alpha.\tag{17.58}$$

Im Weiteren wird lediglich eine Unsicherheit in α und nicht in β betrachtet; das folgende Verfahren lässt sich allerdings auf diesen Fall erweitern.

Um $s = 0$ zu stabilisieren, könnte man den intuitiven Ansatz einer Polvorgabe mit Exakter Linearisierung verfolgen und für die Stellgröße $u = \frac{-1}{\beta}(\bar{\alpha} + \lambda s)$ wählen. Damit würde man allerdings nur im nominellen Fall $\Delta\alpha = 0$ das Ziel erreichen. Ansonsten erhält man ein PT₁-Verhalten mit nicht verschwindendem Eingang

$$\dot{s} + \lambda s = \Delta\alpha \quad (17.59)$$

und damit eine bleibende Regelabweichung von $e = \Delta\alpha/\lambda$.

Ein nichtlinearer Regler mit schaltender Charakteristik kann dahingegen alle Modelle stabilisieren, die innerhalb des Unsicherheitsbereiches liegen. Dazu wird die Stellgröße

$$u = \frac{-1}{\beta}(\bar{\alpha} + K \operatorname{sign}(s)) \quad \text{mit} \quad K = A + \eta \quad (17.60)$$

gewählt. Dies führt zur Beschreibung des geschlossenen Kreises in der Form

$$\dot{s} = \Delta\alpha - K \operatorname{sign}(s). \quad (17.61)$$

Die Stabilität des Punktes $s = 0$ kann mit einer Lyapunov-Funktion geprüft werden. Wählt man $V = \frac{1}{2}s^2$, so ist mit Gl.(17.61) die zeitliche Ableitung

$$\dot{V} = s\dot{s} = s\Delta\alpha - K|s| \leq -\eta|s| \leq 0. \quad (17.62)$$

für alle s stets negativ, d. h. der Punkt $s = 0$ ist stabil. Dies bedeutet ferner, dass die Modellunsicherheit sogar eine beliebige (begrenzte) Dynamik enthalten darf und damit nicht nur Parameterunsicherheiten erfasst werden.

Die Reglerverstärkung K setzt sich zum einen aus der Übersteuerung der Modellunsicherheiten und zum anderen aus der Größe η zusammen, mit der die Geschwindigkeit zum Erreichen von $s = 0$ eingestellt werden kann.

Mit dem gewählten Reglertyp erhält man nämlich bei ideal (d. h. ohne Verzögerung) schaltendem Stellglied einen Regler mit endlicher Einstellzeit. Dazu werden die beiden Fälle $s > 0$ und $s < 0$ betrachtet:

für $s > 0$ gilt:

$$s\dot{s} \leq -\eta|s| \Rightarrow \dot{s} \leq -\eta \Rightarrow s(t_e) - s(t_0) \leq -\eta(t_e - t_0) \quad (17.63)$$

und damit $(t_e - t_0) \leq \frac{s(t_0)}{\eta}$. Für $s < 0$ hingegen gilt:

$$s\dot{s} \leq -\eta|s| \Rightarrow \dot{s} \geq \eta \Rightarrow s(t_e) - s(t_0) \geq \eta(t_e - t_0) \quad (17.64)$$

und damit $(t_e - t_0) \leq \frac{-s(t_0)}{\eta}$.

Mit η und der Anfangsbedingung $s(t_0)$ kann also eine obere Schranke für die Zeit zum Erreichen der Ruhelage $s = 0$ angegeben werden, und das bei einer beliebigen Unsicherheit innerhalb der angenommenen Schranke für $\Delta\alpha$.

17.7.2 Chatter

Im Folgenden wird ein Beispiel für Gl.(17.58) untersucht. Für die Funktion wird $\bar{\alpha} = 100$ und $\Delta\alpha = \sin(s)$ gewählt, wobei für den zweiten Term eine Abschätzung $|\Delta\alpha| \leq 1$ gemacht werden kann. In Bild 17-5 ist links die Dynamik bei Verwendung eines ideal schaltenden Reglers angegeben.

Aus den Ausführungen zu Grenzzyklen in Abschnitt 16.5 ist bekannt, dass schaltende Regler in den meisten Fällen zur Grenzzyklen führen. Betrachtet man die negativ invertierte Beschreibungsfunktion des idealen Zweipunkt-schalters, so bedeckt diese die gesamte negative reelle Achse. Der lineare dynamische Anteil des Sliding-Mode-Regelkreises ist aber ein PT_1 , welches diese nie schneidet. Daher ist es plausibel, dass in diesem Fall kein Grenzzyklus auftritt.

Allerdings wird bereits die kleinste Verzögerung beim Schaltvorgang (sprich eine zusätzliche Totzeit im System) zu einem PT_1T_t führen, welches entsprechende Schnittpunkte mit der negativen reellen Achse besitzt. Reale Regler, die auf Digitalrechnern umgesetzt sind oder andere Formen der Latzen aufweisen, weisen dabei immer eine endliche Schaltzeit auf. Hier erhält man einen Grenzzyklus in der Nähe der Ruhelage in Form eines Hin- und Herschalten des Reglers.

Chatter

Arbeitsbewegungen des Sliding-Mode-Reglers, die aufgrund verzögter Schaltvorgänge entstehen, werden auch als *Chatter* bezeichnet.

Der Fall mit Chatter ist für das gleiche Regelgesetz in Bild 17-5 rechts dargestellt.

$$\alpha(s) = 100 + \sin(s)$$

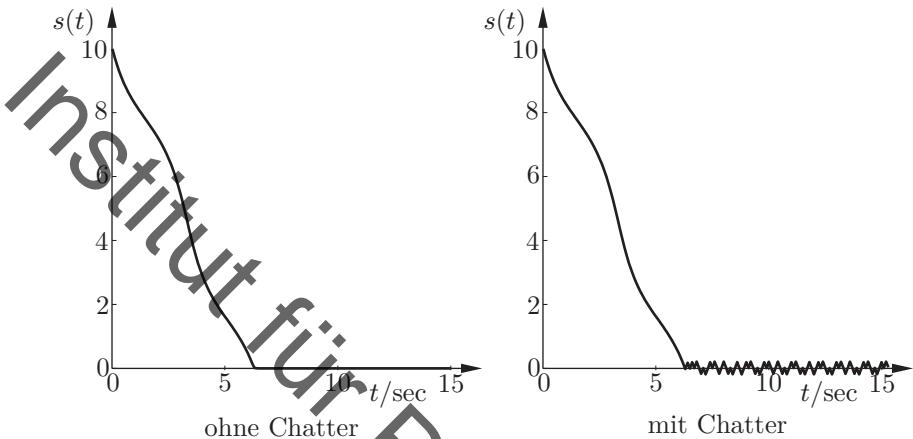


Bild 17-5: Verlauf von $s(t)$ ohne und mit Chatter

Das Chattern sollte vermieden werden, da es zu einer hohen Belastung der Aktoren führt. Aus der harmonischen Balance lässt sich ein Ansatz ableiten, um den Effekt des Chatters zu vermeiden. Ersetzt man das unstetige signum-Signal durch eine Begrenzung wie in Bild 17-6 dargestellt, so ergibt sich eine negativ invertierte Beschreibungsfunktion, die auf der reellen Achse nicht bei 0, sondern erst bei $-\varphi$ beginnt. Ist φ hinreichend groß, so ist aus der harmonischen Balance eine Verhinderung des Chatters zu erwarten. Allerdings möchte man φ nicht zu groß wählen, da die Abweichungen zwischen Schaltglied und Begrenzung sonst immer größer werden. Mit der Begrenzung definiert man nämlich einen Schlauch ϕ um den Zustand $s = 0$, in dem der Regeleingriff proportional zur Abweichung ist. Außerhalb des Schlauches ϕ gilt die bisher betrachtete Stabilitätsuntersuchung, allerdings nimmt man innerhalb des Schlauches eine bleibende Regelabweichung in Kauf, da der Regler dort einem P-Regler entspricht.

Insofern muss man eine Abwägung zwischen hochfrequenten Regeleingriffen und bleibender Regelabweichung treffen. In jedem Fall lässt sich aber eine Stabilisierung trotz Modellunsicherheiten erreichen.

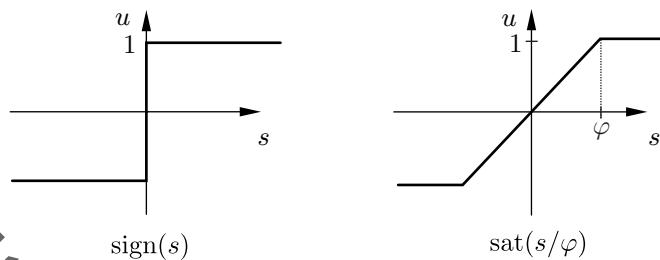


Bild 17-6: Beschreibung der Stellgröße durch die Signum-Funktion (links), eine Begrenzung (rechts)

17.7.3 Systeme höherer Ordnung

Im zweiten Teil wird nun auch das System höherer Ordnung aus Gl.(17.56) auf den zuvor beschriebenen eindimensionalen Fall zurückgeführt. Geht man von Gl.(17.56) aus mit den Zuständen x_1, \dots, x_n , so führt man einen neuen Zustand s ein mit

$$s = x_n + k_{n-1}x_{n-1} + \dots + k_1x_1 \quad (17.65)$$

Dieser Zustand s lässt sich als skalarer Abstand von einer Hyperebene durch den Ursprung deuten. Die Gl.(17.65) entspricht für $s = 0$ nämlich

$$0 = x^{(n-1)} + k_{n-1}x^{(n-2)} + \dots + k_1x, \quad (17.66)$$

also einer Zwangsbedingung, die als Ebenengleichung die Dimension des \mathbb{R}^n um eins reduziert. Die Zustand s auf der linken Gleichungsseite entspricht dann dem Abstand eines Punktes von dieser Ebene.

Der Regler verfolgt das Ziel, genau wie im eindimensionalen Fall $s = 0$ zu erreichen und das System auf diese Hyperebene zu führen. Ist diese passend (d. h. stabil) gewählt, so laufen Lösungen auf dieser Ebene in den Ursprung. Daher röhrt auch der Name als *Sliding Surface* oder *Gleitebene*, die dem Verfahren seinen Namen gibt.

Gleitebene

Sei $\mathbf{x} \in \mathbb{R}^n$, dann beschreibt die Gleichung

$$s = x_n + k_{n-1}x_{n-1} + \dots + k_1x_1 \quad (17.67)$$

für $s = 0$ die *Gleitebene* des Sliding-Mode-Reglers.

Dabei sind die k_i die Koeffizienten des charakteristischen Polynoms mit denen die Eigenwerte des Systems auf der Gleitebene festgelegt werden können.

In Bild 17-7 sind die Hyperebenen für jeweils ein Beispiel aus dem \mathbb{R}^2 und \mathbb{R}^3 angegeben. Ebenfalls abgebildet sind ausgewählte Phasenlinien bzw. Trajektorien, die sich zunächst dem Unterraum nähern und dann auf diesem verbleiben und im Ursprung münden.

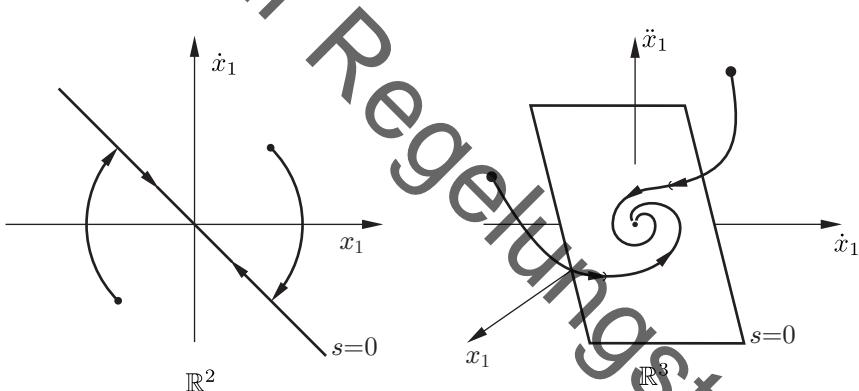


Bild 17-7: Definition eines Unterraums im \mathbb{R}^2 und \mathbb{R}^3

Die Lage der Ebenen ergibt sich aus der Wahl einer stabilen Dynamik. Im \mathbb{R}^2 wählt man

$$s = \dot{x}_1 + k_1x_1 = 0 \quad \text{mit} \quad k_1 > 0 \quad (17.68)$$

und erhält somit eine stets fallende Gerade im \mathbb{R}^2 . Diese Betrachtung lässt sich genauso auf Unterräume des \mathbb{R}^n übertragen.

Die Regelung muss also dafür sorgen, dass der Unterraum $s = 0$ erreicht und beibehalten wird. Hierzu geht man ähnlich dem in Abschnitt 17.7.1

beschriebenen Verfahren vor: Man wählt die Lyapunov-Funktion $V = \frac{1}{2}s^2$ und setzt zur Bestimmung der Stabilität die Definition von s in

$$\begin{aligned}\dot{V} &= s\dot{s} = s(\dot{x}_n + k_{n-1}x_n + \dots + k_1x_2) \\ &= s(\alpha(\mathbf{x}) + \beta(\mathbf{x})u + k_{n-1}x_n + \dots + k_1x_2)\end{aligned}\quad (17.69)$$

Wählt man nun

$$u = \frac{1}{\beta}(\bar{\alpha} + K \operatorname{sign}(s) + k_{n-1}x_n + \dots + k_1x_2) \quad \text{mit} \quad K = A + \eta, \quad (17.70)$$

so erreicht man auch hier, dass \dot{V} stets kleiner als Null für alle s ist und somit für alle Unsicherheiten $\Delta\alpha$ die Gleichgewichtslage $s = 0$ erreicht wird. Durch Einsetzen der Stellgröße erhält man wieder

$$\dot{V} = s\dot{s} = s\Delta\alpha - K|s| \leq -\eta|s| \leq 0. \quad (17.71)$$

Genauso wie im eindimensionalen Fall kann auch hier anstelle des schaltenden Reglergesetzes ein Glied mit Begrenzung benutzt werden, um hochfrequente Eingriffe und damit Anregungen höherer Dynamik zu vermeiden.

18 Lineare Optimale Regelung

18.1 Allgemeines

Im Kapitel 10 wurden verschiedene Kennwerte wie Einschwingzeit oder Überschwingweite vorgestellt, die eine gute Regelung charakterisieren. Dabei wurde festgestellt, dass bereits bei einfachen Systemen die Verringerung eines der definierten Kennwerte oftmals die Erhöhung eines anderen zur Folge hat. Folglich stellen diese Kennwerte in sich widersprüchliche Anforderungen.

Dieser Zielkonflikt in der Regelungstechnik wurde auch im Kapitel 11 beobachtet: So ist prinzipiell eine hohe Reglerverstärkung wünschenswert um Störungen zu unterdrücken und Folgefehler zu minimieren. Andererseits ist dies mit diversen Nachteilen wie Stellgrößenbeschränkungen, energetischen Aufwänden oder Instabilität verbunden. Hierdurch wird die Wahl geeigneter Reglerparameter eine hohe Kunst, da diese widersprüchlichen Anforderungen zueinander gewichtet werden müssen.

Einen alternativen Lösungsansatz zur Wahl der Reglerparameter bieten die Verfahren der *optimalen Regelung*.

Optimale Regelung

Die Idee der optimalen Regelung besteht darin, die Güte einer Regelung in einem einzigen skalaren Zahlenwert zu erfassen und diesen dann zu minimieren.

Die zahlreichen Verfahren der optimale Regelung unterscheiden sich in verschiedenen Unterpunkten. So kann versucht werden, das entstehende Minimierungsproblem analytisch oder numerisch zu lösen. Analytische Lösungen sind dabei meist nur für lineare Regelstrecken und lineare Regler zu finden. Dieses Kapitel stellt die wichtigsten Vertreter dieses Ansatzes vor. Lässt man die Einschränkungen, die für das Finden einer analytischen Lösung notwendig sind, fallen, so sind allgemeinere Entwürfe möglich, von denen die Modellprädiktive Regelung in nächsten Kapitel 19 der wichtigste Vertreter ist.

Bei all diesen Ansätzen geschieht die Abbildung der Regelgüte auf einen einzigen Zahlenwert durch sogenannte Gütemaße. Das wichtigste dieser Gü-

temaße ist die quadratische Regelfläche, welche strukturell die Form

$$J = \int_0^{\infty} e(t)^2 dt \quad , \quad (18.1)$$

besitzt. Diese führt – analog zur Methode der kleinsten Fehlerquadrate bei der Systemidentifikation – auf strukturell günstige Optimierungsprobleme, wie sie in Abschnitt 18.2 verwendet werden.

Dieses Gütemaß bevorzugt allerdings durch die quadratische Gewichtung ein Verhalten, das geringe Einschwingzeiten aufweist aber größeres Überschwingen zulässt. In einigen Anwendungsfällen kann dies unerwünscht sein. Als Beispiel dazu zeigt Bild 18-1 einen mit einer Drehmaschine hergestellten Absatz an einer Welle. Ein Überschwingen führt dazu, dass stellenweise die Sollkontur unterschritten wird, was in jedem Fall zu vermeiden ist.

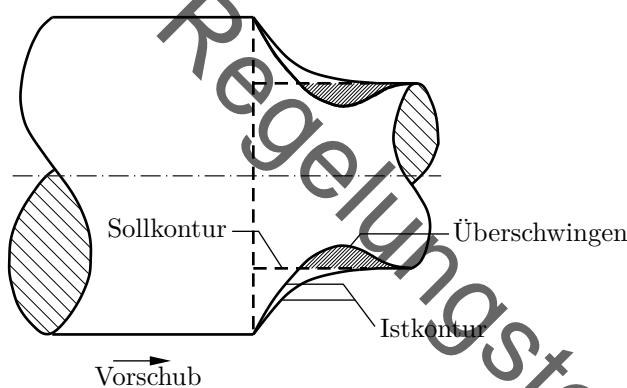


Bild 18-1: Überschwingen an einer Drehmaschine

Es haben sich daher auch Gütemaße, die die aufgezählten Kennwerte zueinander anders gewichten, einen Platz bei bestimmten Anwendungen gesichert. Hier ist vor anderen die zeitbeschwere betragsslineare Regelfläche (ITAE-Kriterium, integral of time-multiplied absolute value of error) zu nennen:

$$J = \int_0^{\infty} |e| \cdot t dt \quad . \quad (18.2)$$

Alternativ können die Gütemaße auch im Frequenzbereich formuliert werden. Dieser Ansatz findet sich u. a. bei der \mathcal{H}_∞ -Regelung, die in Abschnitt 18.3 thematisiert wird. Unabhängig vom konkret verwendeten Gütemaß gilt eine Regelung im Sinne eines dieser Gütemaße dann als optimal, wenn sie so ausgelegt ist, dass das betreffende Gütemaß einen minimalen Wert annimmt.

18.2 Linear-quadratische Regler

18.2.1 Herleitung

Das grundlegendste Verfahren der optimalen Regelung ist der linear-quadratische Regler, welcher auch als LQ-Regler oder LQR abgekürzt wird. Ausgangspunkt hierfür ist die Struktur einer Zustandsrückführung $u = -Kx$ wie in Abschnitt 11.3 und die Zustandsraumbeschreibung

$$\begin{aligned}\dot{x} &= Ax + Bu \\ y &= Cx\end{aligned}\tag{18.3}$$

eines MIMO-LTI-Systems ohne Durchgriff.

Ziel der Regelung ist es, die Systemzustände so auf den Eingang zurückzuführen, dass ein selbstgewähltes positives Gütemaß minimiert wird. Für die Reglerauslegung geht man davon aus, dass der Regler in der Lage sein soll, das System aus jedem beliebigen Anfangszustand $x \in \mathbb{R}^n$ in den Nullzustand $x = \mathbf{0}$ zu überführen. Die Wahl des Nullvektors als Führungsgröße ist dabei keine Einschränkung des Verfahrens, da andere Sollzustände x_{soll} durch Vorsteuerungen oder Verschiebung des Zustandsvektors $\tilde{x} = x - x_{\text{soll}}$ leicht umgesetzt werden können.

Je schneller und mit je weniger Stellenergie der Nullzustand erreicht werden kann, desto besser ist die Regelung. Deshalb wird das Gütemaß häufig auch Kostenfunktion genannt. Man wählt dabei das quadratische Gütemaß

$$J = \int_0^\infty x^T \mathbf{Q} x + u^T \mathbf{R} u \, dt\tag{18.4}$$

Die allgemeine Form der skalaren Kostenfunktion J enthält die Gewichtungsmatrizen \mathbf{Q} und \mathbf{R} . Diese müssen so gewählt werden, dass beide Terme immer positiv werden, damit auch die Kosten positiv bleiben. Für den

Moment wird deshalb angenommen, dass sowohl \mathbf{Q} als auch \mathbf{R} positiv definit (und damit bspw. auch invertierbar) sind. Zudem können \mathbf{Q} und \mathbf{R} stets als symmetrisch angenommen werden, da es zu jedem unsymmetrischen \mathbf{Q} und \mathbf{R} symmetrische Matrizen (den symmetrischen Anteil) gibt, die letztlich auf die gleichen Kosten führen.

Häufig sind \mathbf{Q} und \mathbf{R} nur auf der Hauptdiagonalen mit (positiven) Werten besetzt, sodass lediglich die gewichteten Quadrate der x_i und der u_i übrig bleiben. Dann bewertet der erste Term von J die gewichteten Flächen unter den quadrierten Zustandsgrößen und bietet damit ein Maß für die Abweichung der Zustandsgrößen vom Arbeitspunkt. Der zweite Term enthält ein Maß für die Stellenergien der einzelnen Eingänge.

Gesucht ist nun diejenige Funktion $\mathbf{u}_{\text{opt}}(t)$ aus der Menge aller möglichen Funktionen $\mathbf{u}(t)$, welche die Kostenfunktion J minimiert unter gleichzeitiger Berücksichtigung der Zustandsdifferentialgleichung. Es handelt sich also um ein Optimierungsproblem mit Gleichungsnebenbedingungen, die aus den Modellgleichungen herrühren.

Zur Lösung bietet sich die Methode der Lagrangefunktion an. Mit den Lagrangemultiplikatoren $\lambda(t)$ als Zeitfunktionen aus dem \mathbb{R}^n führt dies auf

$$\mathcal{L}(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) = \int_0^{\infty} \mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{u}^T \mathbf{R} \mathbf{u} + 2\boldsymbol{\lambda}^T (\dot{\mathbf{x}} - \mathbf{A}\mathbf{x} - \mathbf{B}\mathbf{u}) dt . \quad (18.5)$$

Aus der klassischen Optimierung ist bekannt, das im Minimum alle Ableitungen verschwinden müssen. Die Berechnung der partiellen Ableitungen zunächst nur für $\boldsymbol{\lambda}$ und \mathbf{u} ergibt unter Berücksichtigung der Symmetrie von \mathbf{Q} und \mathbf{R}

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}} = 2 \int_0^{\infty} \dot{\mathbf{x}} - \mathbf{A}\mathbf{x} - \mathbf{B}\mathbf{u} dt \stackrel{!}{=} \mathbf{0} , \quad \frac{\partial \mathcal{L}}{\partial \mathbf{u}} = \int_0^{\infty} 2\mathbf{u}^T \mathbf{R} - 2\boldsymbol{\lambda}^T \mathbf{B} dt \stackrel{!}{=} \mathbf{0} . \quad (18.6)$$

Da die Gleichungen unabhängig vom betrachteten Anfangszeitpunkt sind, müssen die Integranden verschwinden und es ergibt sich die bereits bekannte Gleichungsnebenbedingung für \mathbf{x} ,

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} , \quad (18.7)$$

und das optimale Stellgesetz für \mathbf{u} in Abhängigkeit von $\boldsymbol{\lambda}$

$$\mathbf{u} = \mathbf{R}^{-1} \mathbf{B}^T \boldsymbol{\lambda} \quad . \quad (18.8)$$

Für das Ausrechnen dieser optimalen Stellgröße wird also $\boldsymbol{\lambda}$ benötigt. Da \mathbf{u} nach Gl.(18.8) linear in $\boldsymbol{\lambda}$ ist und ein linearer Zustandsregler gesucht ist, muss aber auch ein linearer Zusammenhang zwischen $\boldsymbol{\lambda}$ und \mathbf{x} bestehen. Daher führt man die sogenannte Riccati¹-Matrix \mathbf{P} ein mit

$$\boldsymbol{\lambda}(t) = -\mathbf{P}\mathbf{x}(t) \quad . \quad (18.9)$$

die diesen linearen Zusammenhang herstellt. Um \mathcal{L} auch nach \mathbf{x} ableiten zu können, berechnet man zunächst mit der partiellen Integration

$$\mathcal{L} = \int_0^\infty \boldsymbol{\lambda}^T \dot{\mathbf{x}} dt = - \int_0^\infty \dot{\boldsymbol{\lambda}}^T \mathbf{x} dt + [\boldsymbol{\lambda}^T \mathbf{x}]_0^\infty \quad . \quad (18.10)$$

Ist der geschlossene Regelkreis stabil, so wird $\mathbf{x} \rightarrow 0$ für $t \rightarrow \infty$ gelten und es ergibt sich unter Verwendung von Gl.(18.9)

$$[\boldsymbol{\lambda}^T \mathbf{x}]_0^\infty = \boldsymbol{\lambda}^T(\infty) \mathbf{x}(\infty) - {}_0\boldsymbol{\lambda}^T {}_0\mathbf{x} = - {}_0\mathbf{x}^T \mathbf{P}_0 \mathbf{x} \quad . \quad (18.11)$$

Dieser Term ist unabhängig von \mathbf{x} , da der Anfangszustand ${}_0\mathbf{x}$ fest ist. Damit ergibt sich aus der Forderung $\frac{\partial \mathcal{L}}{\partial \mathbf{x}} \stackrel{!}{=} 0$ die Gleichung

$$0 = -\dot{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T \mathbf{A} + \mathbf{x}^T \mathbf{Q} \Rightarrow \dot{\boldsymbol{\lambda}} = \mathbf{Q}\mathbf{x} - \mathbf{A}^T \boldsymbol{\lambda}. \quad (18.12)$$

Setzt man nun alle Gleichungen ineinander ein, gewinnt man für die Matrix \mathbf{P} die folgende Gleichung, die *stationäre Riccati-Gleichung* genannt wird:

Stationäre Riccati-Gleichung

Die Gleichung

$$\mathbf{P} \mathbf{A} + \mathbf{A}^T \mathbf{P} + \mathbf{Q} - \mathbf{P} \mathbf{B} \mathbf{R}^{-1} \mathbf{B}^T \mathbf{P} = 0. \quad (18.13)$$

heißt *stationäre Riccati-Gleichung* in \mathbf{P} .

¹Jacopo Riccati (1676-1754), italienischer Mathematiker [46]

Die Riccati-Matrix \mathbf{P} muss im Allgemeinen numerisch ermittelt werden, da keine geschlossene Lösung existiert. Dies ist jedoch nur einmal bei jeder Reglerauslegung nötig. Die Lösung einer quadratischen Gleichung für Matrizen lässt sich dabei (insbesondere mit Rechnerunterstützung) mit vertretbarem Aufwand durchführen. Der Rechenaufwand lässt sich zudem reduzieren, da man zeigen kann, dass \mathbf{P} symmetrisch ist, sofern \mathbf{Q} und \mathbf{R} symmetrisch sind.

Die stationäre Riccati-Gleichung ist unabhängig von der gewählten Anfangsbedingung \mathbf{x}_0 . Die tatsächlich auftretenden Kosten berechnen sich dagegen zu

$$J_{\text{opt}} = \int_0^{\infty} \mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{u}^T \mathbf{R} \mathbf{u} dt = {}_0 \mathbf{x}^T \mathbf{P}_0 \mathbf{x} \quad (18.14)$$

und sind somit abhängig von den Anfangsbedingungen. Sie haben aber eine untergeordnete Bedeutung für die Regelung und müssen meist nicht explizit berechnet werden.

Die Lösung der Riccati-Gleichung \mathbf{P} liefert die optimale Rückführmatrix:

$$\mathbf{K} = \mathbf{R}^{-1} \mathbf{B}^T \mathbf{P} \quad . \quad (18.15)$$

Die Lösbarkeit des Optimierungsproblems über die Riccati-Gleichung unterliegt gewissen Voraussetzungen, von denen einige bereits angeklungen sind. Zum einen muss das System stabilisierbar sein, da ansonsten $\mathbf{x} \rightarrow 0$ für $t \rightarrow \infty$ nicht gilt. Zudem besitzt die Riccati-Gleichung als quadratische Gleichung naturgemäß zwei Lösungen. Von diesen beiden Lösungen ist dasjenige \mathbf{P} zu wählen, welches positiv definit ist. Dies liegt daran, dass man zeigen kann, dass $V(\mathbf{x}) = \mathbf{x}^T \mathbf{P} \mathbf{x}$ eine Lyapunov-Funktion des geschlossenen Regelkreises ist. Zuletzt muss die durch Ableiten gefundene Lösung des Optimierungsproblems auch ein Minimum und kein Maximum darstellen, wofür die zweite Ableitung der Kostenfunktion positiv definit sein muss.

Man kann zeigen [30], dass all diese Voraussetzungen erfüllt sind, wenn das Paar (\mathbf{A}, \mathbf{B}) steuerbar ist.

Lösbarkeit des LQR-Problems

Ist das Paar (\mathbf{A}, \mathbf{B}) steuerbar, so besitzt Gl.(18.13) eine eindeutige positiv definite Lösung. Der nach Gl.(18.15) berechnete Zustandsregler führt dann in jedem Fall auf einen stabilen geschlossenen Regelkreis.

18.2.2 Wahl der Gewichtungsmatrizen

Die Stabilität eines LQ-Reglers ist also theoretisch garantiert. Hiermit ist aber noch keine Aussage über die praktische Brauchbarkeit der Regelung verbunden.

Der durch \mathbf{K} eindeutig beschriebene Zustandsregler wird das lineare System optimal in Bezug auf die gewählte Kostenfunktion J regeln. Es bleibt die Wahl der Gewichtungsmatrizen \mathbf{Q} und \mathbf{R} . Von besonderem Interesse ist dabei das Verhältnis der beiden Terme im Integral zueinander: Wird die Abweichung der Zustandsgrößen vom Arbeitspunkt stark gewichtet, so erhält man eine schnelle Regelung mit großen Stellausschlägen. Wird dagegen der zweite Term, der als Maß für die Stellenergie gilt, höher bewertet, so dauert es länger bis Störungen ausgeregelt sind, dafür wird aber auch weniger Stellaufwand benötigt.

In einigen Fällen schränkt die Forderung, dass \mathbf{Q} und \mathbf{R} positiv definit sein müssen, zu sehr ein. Das ist unter anderem dann der Fall, wenn der Verlauf der Ausgangsgrößen gemäß

$$J = \int_0^{\infty} \mathbf{y}^T \mathbf{Q}_y \mathbf{y} + \mathbf{u}^T \mathbf{R} \mathbf{u} dt = \int_0^{\infty} \mathbf{x}^T \underbrace{\mathbf{C}^T \mathbf{Q}_y \mathbf{C}}_{=\mathbf{Q}} \mathbf{x} + \mathbf{u}^T \mathbf{R} \mathbf{u} dt \quad (18.16)$$

gewichtet werden soll. Für viele \mathbf{C} wird die dabei auftretende Matrix \mathbf{Q} nicht positiv definit sein. Dies liegt daran, dass typischerweise \mathbf{C} bei unvollständiger Zustandsmessung viele Nulleinträge besitzt. Hierdurch wird im Allgemeinen \mathbf{Q} nur positiv semidefinit sein.

Positive Semidefinitheit

Eine quadratische Matrix \mathbf{W} ist genau dann positiv semidefinit, wenn

$$\mathbf{x}^T \mathbf{W} \mathbf{x} \geq 0 \quad \forall \mathbf{x} \neq \mathbf{0} \quad (18.17)$$

gilt. Für eine symmetrische Matrix bedeutet dies, dass alle (reellen) Eigenwerte $\lambda \geq 0$ positiv sind.

Der Unterschied zur positiven Definitheit aus Gl.(8.14) besteht in der Zulässigkeit der Gleichheit bei den entsprechenden Ungleichungen. Lässt man allgemein positiv semidefinite Matrizen \mathbf{Q} zu, so ändert dies zunächst nichts an der Lösbarkeit des Optimierungsproblems, welche an die Steuerbarkeit gekoppelt ist.

Allerdings kann die Stabilität für allgemein positive semidefinite \mathbf{Q} nicht mehr garantiert werden. Dies zeigt das einfache Beispiel von $\mathbf{Q} = \mathbf{0}$, welches für beliebige Regelstrecken und beliebige \mathbf{R} sofort auf die Lösung $\mathbf{u} = \mathbf{0}$ führt. Die Regelung wird, da Abweichungen keine Kosten verursachen, nicht aktiv und instabile Systeme bleiben folglich instabil.

Dieses Phänomen tritt genau dann auf, wenn es (divergente) Lösungen im Zustandsraum gibt, die keine Kosten verursachen, da sie in $\mathbf{x}^T \mathbf{Q} \mathbf{x}$ nicht „beobachtet“ werden können.

Stabilität bei semidefiniter Wahl von \mathbf{Q}

Gibt es eine Zerlegung von $\mathbf{Q} = \mathbf{C}^T \tilde{\mathbf{C}}$, sodass das Paar $(\mathbf{A}, \tilde{\mathbf{C}})$ beobachtbar ist, so ist der geschlossene Regelkreis auch dann stabil, wenn \mathbf{Q} nur positiv semidefinit ist.

Die Gewichtung der Stellgrößen \mathbf{R} darf nicht positiv semidefinit erfolgen, da dann die Inverse \mathbf{R}^{-1} nicht existent wäre.

18.2.3 Linear-quadratisch-Gaußsche Regler

Betrachtet man die im LQR verwendete Kostenfunktion genauer, so erkennt man einige Parallelen zur Struktur des Kalmanfilters. Nicht umsonst besitzen die Gewichtungsmatrizen für beide Verfahren identische Bezeichner \mathbf{Q} und \mathbf{R} . Tatsächlich ist der Zusammenhang zwischen beiden Verfahren sehr eng. So könnte man beispielsweise auf die Idee kommen, das Verfahren des LQR zur Auslegung eines Beobachters zu nutzen. Wie aus der Darstellung zum Beobachterentwurf in Abschnitt 11.4.3 ersichtlich ist, kann potentiell jedes Verfahren, das zum Entwurf einer vollständigen Zustandsrückführung geeignet ist, auch zum Entwurf eines Beobachters dienen. Hierzu bedient man sich der Dualität zwischen (\mathbf{A}, \mathbf{B}) und $(\mathbf{A}^T, \mathbf{C})$.

LQR für den dualen Beobachterentwurf

Entwirft man einen Beobachter für das duale Problem über das Verfahren des LQ-Reglers, so gewinnt man die zeitkontinuierliche Formulierung des Kalmanfilters.

Strukturell sind beide Verfahren sich also sehr ähnlich. Daher erscheint es denkbar, dass die Kombination beider Verfahren ebenfalls bestimmten Optimalitätsbedingungen genügt. Hierzu lassen sich einige Vorüberlegungen anstellen.

Der LQR liefert den optimalen Regler unter der Annahme, dass der tatsächliche Zustand x bekannt ist. Im praktischen Einsatz wird dies aber nicht der Fall sein, sondern nur der Systemausgang y verfügbar sein.

Sucht man anstelle eines optimalen Zustandsreglers nach einer optimalen Ausgangsrückführung, so kann man diese in einen Zustandsregler und einen Zustandsbeobachter aufteilen. Dies kann man so interpretieren, dass der LQ-Regler nun \hat{x} als Eingang nutzt und nicht mehr x . Die auf dieser Basis generierte Stellgröße weicht vom optimalen Wert einer Zustandsrückführung ab. Es stellt sich die Frage, welche optimale Kombination von Zustandsbeobachter und Zustandsregler die Kostenfunktion unter Berücksichtigung des Unterschiedes von x und \hat{x} minimiert.

LQG-Problem

Gegeben ist ein lineares System mit mittelwertfreiem Gaußschem Modellrauschen $w(t)$ und Messrauschen $v(t)$

$$\begin{aligned}\dot{x} &= Ax + Bu + w \\ y &= Cx + v\end{aligned}\quad . \quad (18.18)$$

Das linear-quadratisch-Gaußsche (LQG) Regelungsproblem sucht nach der linearen Ausgangsrückführung $\mathbf{K}(s)$, welche den Erwartungswert

$$J = E \left[\int_0^{\infty} x^T Q x + u^T R u dt \right] \quad (18.19)$$

minimiert.

Das Separationstheorem aus Abschnitt 11.4.3 besagte, dass Beobachter und

Regler unabhängig voneinander entworfen werden können. Daher liegt die Vermutung nahe, dass sich diese Frage auf zwei separierbare Optimierungsprobleme aufteilen wird – eins für den Regler und eins für den Beobachter. Tatsächlich lässt sich zeigen, dass die optimale Ausgangsrückführung genau einem LQ-Regler mit Gewichtung \mathbf{Q} und \mathbf{R} aus Gl.(18.19) in Kombination mit einem Kalmanfilter mit Gewichtung $\mathbf{Q} = \sigma^2 [\mathbf{v}]$ und $\mathbf{R} = \sigma^2 [\mathbf{w}]$ entspricht.

LQG-Regler

Der eindeutige, optimale LQG-Regler besteht aus einem LQ-Regler, dessen Zustand von einem Kalmanfilter geschätzt wird.

Dieses Ergebnis unterstreicht die Leistungsfähigkeit von LQR und Kalmanfilter. Allerdings gibt es auch Nachteile einer Kombination beider Verfahren. So gilt die Optimalität nur im Rahmen der perfekten Einstellparameter für das Kalmanfilter, die sich im Allgemeinen nicht bestimmen lassen.

Außerdem geht beim LQG-Regler die wichtige Robustheit der Regelung verloren. So kann man zeigen, dass – obgleich der LQR eine garantierte Amplitudenreserve von mindestens 2 besitzt – für einen Regelkreis mit LQG keine minimale Amplitudenreserve garantiert werden kann [10].

18.3 \mathcal{H}_∞ -Regelung

18.3.1 Closed Loop Shaping

Das vorgestellte Optimierungsproblem des LQR wird im Zeitbereich vorgenommen. Der LQR arbeitet dabei zeitkontinuierlich mit der quadratischen Regelfläche als einer Funktionennorm.

Alternativ ist ein Ansatz im Frequenzbereich möglich, der das Optimierungsproblem im Bildbereich betrachtet. Diese Idee erscheint noch naheliegender, wenn man das *Parseval²-Theorem* kennt, welches die quadratische Regelfläche mit dem Frequenzbereich verbinden kann.

²Marc-Antoine Parseval (1755-1836), französischer Mathematiker [42]

Theorem von Parseval

Für fouriertransformierbare Funktionen $x(t)$ und $y(t)$ gilt:

$$J = \int_{-\infty}^{\infty} x(t) \cdot y(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} X^*(\omega) \cdot Y(\omega) d\omega , \quad (18.20)$$

wobei X^* die komplexe Konjugation bezeichnet.

Der Beweis ergibt sich, indem man die inverse Laplace-Transformation Gl.(4.40) für $y(t)$ einsetzt und die Integrationsreihenfolge vertauscht:

$$J = \int_{-\infty}^{\infty} x(t) \frac{1}{2\pi} \int_{-\infty}^{\infty} Y(\omega) e^{j\omega t} d\omega dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} Y(\omega) \int_{-\infty}^{\infty} x(t) e^{j\omega t} dt d\omega \quad (18.21)$$

Mit $\int_{-\infty}^{\infty} x(t) e^{j\omega t} dt = X(-\omega) = X^*(\omega)$ erhält man

$$J = \frac{1}{2\pi} \int_{-\infty}^{\infty} Y(\omega) \cdot X^*(\omega) d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} X^*(\omega) \cdot Y(\omega) d\omega \quad (18.22)$$

wegen der Kommutativität des ursprünglichen Produkts $x(t) \cdot y(t)$.

Ein interessanter Sonderfall ist $y = x$ mit

$$\int_{-\infty}^{\infty} x^2(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) \cdot X^*(\omega) d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} |X(\omega)|^2 d\omega \quad (18.23)$$

Folglich entspricht der quadratischen Regelfläche ein ganz analoges Integral im Frequenzbereich. Daher liegt die Vermutung nahe, dass sich ähnliche Resultate und eine ähnliche Leistungsfähigkeit des synthetisierten Reglers auch bei Optimierungsverfahren im Frequenzbereich erzielen lassen sollten.

Vor der Formulierung eines passenden Optimierungsproblems und eines damit verbundenen Gütekriteriums im Frequenzbereich soll an dieser Stelle der Reglerentwurf im Frequenzbereich kurz wiederholt werden.

Beim Entwurf im Frequenzkennlinienverfahren 11.1 wurde der aufgeschnittene Regelkreis $G_0(j\omega)$ so festgelegt, dass der durch $S(j\omega)$ und $T(j\omega)$ repräsentierte geschlossene Regelkreis die gewünschten Eigenschaften besaß. Die resultierenden Anforderungen an $G_0(j\omega)$ wurden über Gl.(11.2) und Gl.(11.7) aus Abschätzungen der Anforderungen an die Frequenzgänge des geschlossenen Regelkreises ermittelt. Für eine gegebene Regelstrecke $G_S(j\omega)$ wurde der Regler $G_R(j\omega)$ dann so angepasst, dass sich das gewünschte $G_0(j\omega)$ ergab.

Dieser Reglerentwurf adressiert folglich die Anforderungen nicht direkt, sondern nur über den Umweg des offenen Regelkreises. Dies hat den gewichtigen Vorteil, dass die Berechnung des Reglers $G_R(j\omega)$ mit sehr einfachen Mitteln wie einer graphischen Subtraktion von $G_0(j\omega)$ und $G_S(j\omega)$ im Bode-Diagramm durchgeführt werden kann. Ein entscheidender Nachteil ist jedoch, dass durch den Umweg über $G_0(j\omega)$, der auch Abschätzungen enthält, Reserven und Leistungsfähigkeit des Reglers verloren gehen.

Mit rechnergestützten Optimierungsalgorithmen lassen sich statt des aufgeschnittenen Regelkreises direkt die Frequenzgänge des geschlossenen Regelkreises adressieren. Daher wird dieses Verfahren abgrenzend zum Frequenzkennlinienverfahren („Open Loop Shaping“) auch als „*Closed Loop Shaping*“ bezeichnet. Für eine gegebene Strecke $G_S(j\omega)$ wird direkt ein Regler $G_R(j\omega)$ synthetisiert, der die Anforderungen an die Frequenzgänge des geschlossenen Regelkreises erfüllt.

Aufgrund des stark nichtlinearen Zusammenhangs von $G_R(j\omega)$ und dem geschlossenen Regelkreis kann dabei die Berechnung des Reglers nur noch mit Computerunterstützung erfolgen. Ein weiterer Vorteil des Verfahrens ist allerdings, dass sich auf diese Weise auch Regler für Regelstrecken, die die Anforderungen des vereinfachten Nyquist-Kriteriums nicht erfüllen, synthetisieren lassen.

Aus Abschnitt 10.3 ist bekannt, dass möglichst kleine Amplituden von $S(j\omega)$ über alle Frequenzen erstrebenswert sind. Um diese Anforderung über ein vergleichbares Gütemaß zu fassen, bedient man sich der sogenannten \mathcal{H}_∞ -Norm [44]:

\mathcal{H}_∞ -Norm

Die \mathcal{H}_∞ -Norm einer Übertragungsfunktion ohne Polstellen auf der ima-

ginären Achse ist definiert als

$$\|G(s)\|_\infty = \sup_{\omega} |G(j\omega)| . \quad (18.24)$$

Sie entspricht den größtem Wert des Amplitudengangs und damit bei einer stabilen Übertragungsfunktion der betragsmäßig größten Verstärkung über alle Frequenzen. Ist das Übertragungssystem in Form einer Ortskurve gegeben, so entspricht die \mathcal{H}_∞ -Norm dem maximalen Abstand der Ortskurve zum Ursprung.

In der Definition muss dabei das Supremum \sup anstelle des Maximums \max genutzt werden. Der mathematische Unterschied besteht darin, dass das Maximum nicht zwingend existieren muss, da Grenzwerte wie $\omega \rightarrow \infty$ von der Definition des Maximums nicht abgedeckt werden. Das Supremum hingegen existiert auch in solchen Fällen.

Der Regelkreis mit perfekter Störgrößenunterdrückung würde $|S| = 0$ für alle ω und damit $\|S\|_\infty = 0$ erzielen. Eine solche Regelung würde $|G_0(j\omega)| = \infty$ für alle Frequenzen implizieren und kann daher aus den in Abschnitt 10.3 ausgeführten Gründen nicht erwartet werden.

Tatsächlich gibt es noch weitere fundamentale Einschränkungen, die die Form des Amplitudengangs der Sensitivitätsfunktion limitieren.

Bode Sensitivitäts-Integral

Sei $G_0(j\omega)$ der aufgeschnittene Regelkreis mit p_+ Polstellen λ_i in der rechten s -Halbebene. Dann lässt sich mathematisch nachweisen, dass für die Sensitivitätsfunktion S des geschlossenen Regelkreises gilt:

$$\int_0^\infty \lg |S(j\omega)| d\omega = \pi \cdot \sum_{i=1}^{p_+} \operatorname{Re}(\lambda_i) - \frac{\pi}{2} \lim_{s \rightarrow \infty} s G_0(s) . \quad (18.25)$$

Gl.(18.25) besagt, dass die integrale Fläche unter dem Amplitudengang der Sensitivitätsfunktion $|S|$ nicht frei gewählt werden kann, sondern Beschränkungen unterliegt. Da diese Fläche möglichst klein sein soll, zeigt Gl.(18.25), dass Störungen für instabile Systeme offensichtlich schwerer zu unterdrücken sind, da auf der rechten Gleichungsseite positive Ausdrücke erscheinen.

Genau gegenteilig verhält es sich mit Systemen mit einem relativen Grad von $r = 1$, da für diese der Grenzwert $\lim_{s \rightarrow \infty} sG_0(s)$ nicht verschwindet und somit negative Ausdrücke auf die rechte Gleichungsseite gelangen können.

Das bedeutet, dass ein PT₁ leichter als ein PT₂ zu regeln ist und ein stabiles PT₁ leichter als ein instabiles System erster Ordnung.

Wasserbeteffekt

Besitzt G_0 ausschließlich Pole mit nicht positivem Realteil und einen relativen Grad von $r \geq 2$, dann folgt aus dem Bode-Sensitivitätsintegral

$$\int_0^\infty \lg |S(j\omega)| d\omega = 0 \quad , \quad (18.26)$$

was auch als *Wasserbeteffekt* bezeichnet wird.

Die anschauliche Interpretation von Gl.(18.26) ist, dass die Amplitude der Sensitivitätsfunktion nicht über den ganzen Frequenzbereich klein gezwungen werden kann. Das Erreichen einer Verbesserung, d. h. die Unterdrückung von $|S(j\omega)|$ in einem Frequenzbereich, geht mit einer gleichwertigen Verschlechterung in einem anderen Frequenzbereich einher – siehe Bild 18-2. Daher hat sich die Bezeichnung als Wasserbeteffekt für dieses Phänomen eingebürgert. Anschaulich halten sich die Abschwächung und Verstärkung von Störungen die Waage und können nur in andere Frequenzbereiche verlagert werden. Hier ist zu beachten, dass die Flächen in Bild 18-2 aufgrund der logarithmischen Skalierung der Frequenzachse nicht gleich groß erscheinen, bei einer linearen Skalierung aber identisch sind.

Da folglich die Übertragungsfunktionen des geschlossenen Regelkreises nicht über den gesamten Frequenzbereich betragsmäßig klein sein können, bietet sich die Formulierung von frequenzabhängigen Anforderungen an. Typische Anforderungen an $S(j\omega)$ schließen folgende Aspekte ein:

1. Größter Betrag von $S(j\omega)$ über alle Frequenzen: $\|S(j\omega)\|_\infty \leq M$
2. Begrenzung des stationären Regelfehlers *Stationäre Genauigkeit, stationary accuracy*
3. Einstellen einer gewünschten minimalen Bandbreite ω_g

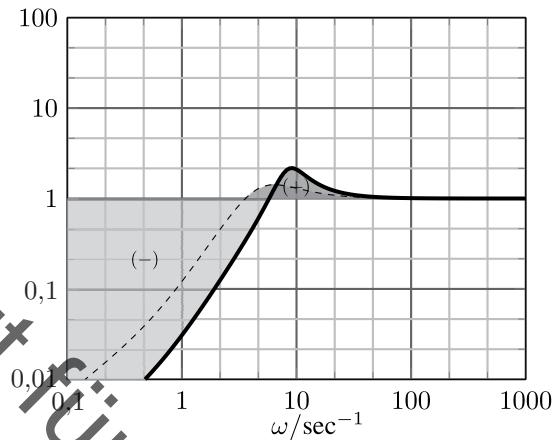


Bild 18-2: Amplitudengang mit Kennzeichnung positiver (+) und negativer (-) Beiträge zum Sensitivitätsintegral

4. Verlauf von $S(j\omega)$ in ausgewählten Frequenzbereichen

Ein möglicher Verlauf der Sensitivitätsfunktion ist in Bild 18-3 abgebildet und die entscheidenden Größen für die ersten drei Aspekte markiert. Die Abbildung zeigt die beschriebenen Anforderungen in Form eines frequenzabhängigen Betragsverlaufs, den der Betrag von $S(j\omega)$ über alle Frequenzen nicht überschreiten darf. Dies wird mathematisch mithilfe einer Gewichtsfunktion $W(j\omega)$ und der bereits definierten \mathcal{H}_∞ -Norm ausgedrückt als:

$$|S(j\omega)| \leq \frac{1}{|W(j\omega)|} \quad \forall \omega \quad \Rightarrow \quad \|W(j\omega) \cdot S(j\omega)\|_\infty \leq 1 \quad (18.27)$$

Ist die Gewichtsfunktion $W(j\omega)$ eine Konstante, so wird angestrebt, dass die Amplitude von $S(j\omega)$ gleichmäßig über alle Frequenzen kleiner als der Kehrwert dieser Konstanten ist. Ist $W(j\omega)$ eine Funktion wie in Bild 18-3, so wird angefordert, dass $|S(j\omega)|$ diese frequenzvariable Form nicht überschreitet.

Alternativ zur Überschreitung von $1/|W(j\omega)|$ kann $|W(j\omega)|$ als frequenzabhängige Gewichtung interpretiert werden, wobei hohe Werte von $|W(j\omega)|$ anzeigen, dass es wichtiger ist, die Sensitivität für diese Frequenzen zu verringern als in Frequenzbereichen, wo $|W(j\omega)|$ kleine Werte annimmt.

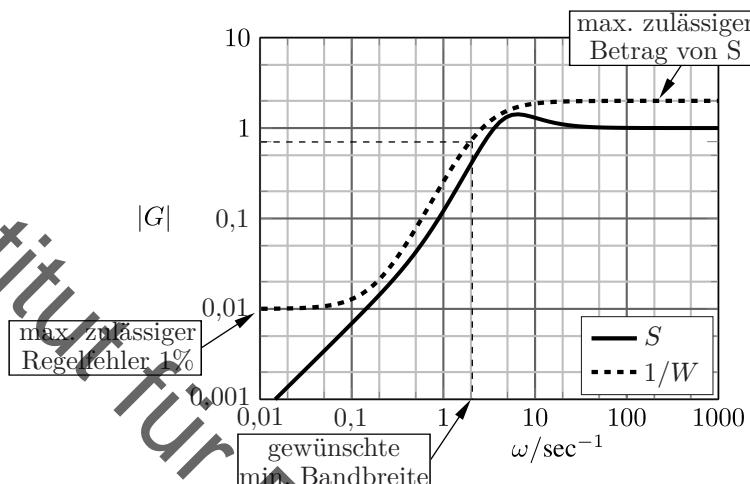


Bild 18-3: Anforderungen der Gewichtsfunktion W an den Amplitudengang der Sensitivität

Da bei diesem Vorgehen die Sensitivität gewichtet wird, spricht man bei diesem Entwurf eines \mathcal{H}_∞ -Reglers auch vom Verfahren der *Weighted Sensitivity*.

18.3.2 Generalized Plant

Die Bedingung in Gl.(18.27) ist ein Optimierungsproblem, welches sich aber recht speziell auf einen (wichtigen) Anwendungsfall der Gewichtung der Sensitivität bezieht. Aus dem LQR ist aber bekannt, dass auch die Gewichtung der verwendeten Stellenergie eine sinnvolle Komponente eines regelungstechnischen Optimierungsproblems sein kann.

Daher ist es erstrebenswert, einen möglichst generellen Prototyp von Optimierungsproblemen zu finden und die (numerische) Lösung an diesem zu diskutieren. Hierbei wird auf die mächtige Darstellungsform der *Generalized Plant* zurückgegriffen.

Generalized Plant

Die Generalized Plant verallgemeinert den klassischen Regelkreis. Ausgehend von der verallgemeinerten Regelstrecke \mathcal{P} und dem zu entwerfenden Regelkreiselement \mathcal{K} werden die auftretenden Signale in vier verschiedene Kategorien unterteilt:

- u Eingänge von \mathcal{P} , welche gleichzeitig Ausgänge von \mathcal{K} sind.
- w Eingänge von \mathcal{P} , welche keine Ausgänge von \mathcal{K} sind und deren Einfluss auf z gering gehalten werden soll.
- v Ausgänge von \mathcal{P} , welche gleichzeitig Eingänge von \mathcal{K} sind.
- z Die Zielgrößen, die das Regelkreiselement minimieren soll.

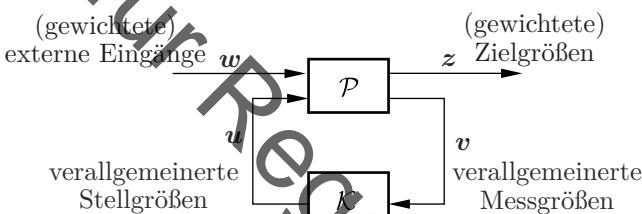


Bild 18-4: Grundsätzlicher Aufbau der Generalized Plant

Diese Zuordnung ist in Bild 18-4 dargestellt. Hierbei wird bewusst von einem „Regelkreiselement“ gesprochen, da die Darstellung so allgemein gehalten wird, dass sie für Ausgangsrückführung, Zustandsrückführungen aber auch Vorsteuerungen oder Beobachter genutzt werden kann. Bei einer klassischen Ausgangsrückführung bezeichnet u die Stellgrößen, da diese vom Regler berechnet werden und somit Ausgänge von \mathcal{K} sind. Unter w würden dann alle weiteren Eingänge des Regelkreises gefasst. Das könnten beispielsweise Sollwerte, Störgrößen oder Messrauschen sein. Beim Entwurf einer Ausgangsrückführung wäre v die gemessene Regelabweichung (inklusive Messrauschen). Davon zu unterscheiden ist dann die Zielgröße z als die tatsächliche Regelabweichung (ohne Messrauschen).

Eine analoge Zuordnung ist aber auch für z. B. einen LQR möglich. Hier bleibt u als Stellgröße gleich. Der externe Eingang w ist der Anfangswert x_0 , der nicht vom Regler festgelegt wird, aber das Systemverhalten entscheidend beeinflusst und die optimalen Kosten festlegt. Zielgröße ist die

Kostenfunktion J in Gl.(18.4), Eingang in \mathcal{K} ist der zurückzuführende Zustand $v = x$.

Offenbar lassen sich also auch bereits bekannte Optimierungsprobleme unter die Struktur der Generalized Plant fassen. Auch das Verfahren der Weighted Sensitivity lässt sich in die Struktur der Generalized Plant überführen, indem unter der Annahme einer Sollwertfolge $w(t) = w(t)$ und einer Ausgangsrückführung $v(t) = w(t) - y(t)$ der mit $W(j\omega)$ gewichtete Ausgangsfehler (d. h. $Z(s) = W_P(s)(W(s) - Y(s))$) minimiert werden soll.

Im linearen Fall lässt sich die Generalized Plant \mathcal{P} durch die Verknüpfung zwischen den genannten Ein- und Ausgangsgrößen als eine mehrdimensionale Übertragungsmatrix \mathbf{P} mit

$$\begin{bmatrix} z \\ v \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{P}_{11}(s) & \mathbf{P}_{12}(s) \\ \mathbf{P}_{21}(s) & \mathbf{P}_{22}(s) \end{bmatrix}}_{\mathbf{P}(s)} \cdot \begin{bmatrix} w \\ u \end{bmatrix} \quad (18.28)$$

darstellen. Die Aufgabe des Entwurfs von \mathcal{K} kann so beschrieben werden, dass eine Übertragungsmatrix $\mathbf{K}(s)$ gesucht wird, welche die Auswirkung von $\mathbf{W}(s)$ auf $\mathbf{Z}(s)$ minimiert.

Die Übertragungsmatrix \mathbf{F} dieses geschlossenen Regelkreises der Generalized Plant lässt sich allgemein zu

$$\mathbf{F}(s) = \frac{\mathbf{Z}(s)}{\mathbf{W}(s)} = \mathbf{P}_{11} + \mathbf{P}_{12}\mathbf{K}(\mathbf{I} - \mathbf{P}_{22}\mathbf{K})^{-1}\mathbf{P}_{21} \quad (18.29)$$

berechnen. Das allgemeine Optimierungsproblem lautet dann

$$\min_{\gamma} \|\mathbf{F}(s)\|_{\infty} \leq \gamma. \quad (18.30)$$

Da es sich hierbei um ein Mehrgrößensystem handelt, muss die \mathcal{H}_{∞} -Norm noch für MIMO-Systeme erweitert werden. Dies geschieht unter Berücksichtigung der Richtungsabhängigkeit durch die größten Singulärwert mit

$$\|\mathbf{G}(s)\|_{\infty} := \sup_{\omega} \bar{\sigma}(\mathbf{G}(j\omega)) \quad , \quad (18.31)$$

wobei Gl.(18.31) für SISO-Systeme auf Gl.(18.24) zurückfällt.

Das aus Gl.(18.29) und Gl.(18.30) bestehende Optimierungsproblem lässt sich mit entsprechenden Algorithmen effizient lösen, auf die an dieser Stelle nicht eingegangen werden soll. Details finden sich beispielsweise in [53].

Festzuhalten ist lediglich, dass die Algorithmen ausgehend vom einem Startwert für γ und einer Zustandsraumdarstellung der Generalized Plant einen Regler synthetisieren, welcher die Bedingung Gl.(18.30) mit einer Gleichheit löst. Entsprechende Nebenbedingungen stellen dabei sicher, dass der so gefundene Regler in jedem Fall stabil ist. Man fährt dann fort, ein identisches Problem mit einem reduzierten Wert für γ zu lösen, was als γ -Iteration bezeichnet wird.

Hierbei gibt man sich gegebenenfalls auch mit einem suboptimalen Regler zufrieden, der die Bedingung nicht für das minimale γ aber ein hinreichend kleines zufriedenstellend löst. Nachteilig an dem Verfahren ist, dass die gewonnenen Lösungen oftmals Regler hoher Ordnung liefern.

Ordnung der \mathcal{H}_∞ -Regler

Die Reglerordnung entspricht der Ordnung der Generalized Plant. Im Fall der Weighted Sensitivity ist das die Summe der Ordnung der Regelstrecke und der Ordnung der Gewichtsfunktion.

Schon für eine PT₂-Regelstrecke mit einer PDT₁-Gewichtungsfunktion W wie in Bild 18-3 ergäbe sich also bereits ein Regler mit drei Polstellen. Diese mangelnde Kontrolle über die Struktur des resultierenden Reglers kann bei der Implementierung des Reglers ein Nachteil sein. Daher werden oft Verfahren zur Modellordnungsreduktion der Optimierung nachgelagert, um einen sehr ähnlichen Regler niedriger Ordnung zu erhalten.

Ebenfalls ist zu beachten, dass die Algorithmen auf einer Zustandsraumdarstellung der Regelstrecke basieren, obgleich der Reglerentwurf im Bildbereich durchgeführt wird. Das rechnerische Vorgehen im Zustandsraum setzt dabei voraus, dass das Streckenmodell nicht rein graphisch, sondern analytisch als gebrochen rationale Funktion vorliegt. Regelkreise mit Totzeitgliedern sind daher ausgeschlossen.

18.3.3 Mixed Sensitivity

Eine beliebte und recht allgemeine Form, um mit Gl.(18.30) einen Regler zu entwerfen, ist eine Erweiterung der Weighted Sensitivity. Die Grundidee ist dabei, dass entsprechende Optimalitätskriterien für den geschlossenen Regelkreis nicht nur für die Sensitivität formuliert werden können. Ein entsprechender Verlauf ist auch für $T(j\omega)$ erwünscht, um gutes Führungsverhalten sowie eine geringe Auswirkung von Messrauschen auf den Regelfehler zu garantieren.

Eine weitere interessante Eigenschaft im klassischen Regelkreis ist das Übertragungsverhalten von sämtlichen Eingangsgrößen auf die Stellgröße u , welches durch die Funktion

$$KS(s) = \frac{U(s)}{W(s)} = G_R(s) \cdot S(s) \quad (18.32)$$

beschrieben wird und sich als das Produkt des Reglers mit der Sensitivitätsfunktion ergibt. Es liegt nahe, dass der Einfluss der Eingangsgrößen auf die Stellgröße ebenfalls gering sein soll, da die verfügbare Stellenergie begrenzt ist. In praktischen Anwendungen müssen folglich oft nicht nur eine, sondern mehrere Übertragungsfunktionen betrachtet werden.

Die *Mixed Sensitivity Methode* kann diese Anforderung mithilfe der Generalized Plant berücksichtigen. Hierzu wird die folgende Generalized Plant aufgestellt:

Unter der Annahme einer Folgeregelung ergeben sich die Ausgänge \mathbf{Z} zu

$$\mathbf{Z} = [Z_1(s) \ Z_2(s)] = [W_u(s)U(s) \ W_p(s)(W(s) - Y(s))] \quad (18.33)$$

Die Auslegung der gewichteten Filter $W_p(s)$ und $W_u(s)$ folgt dabei dem Entwurfsverfahren der Weighted Sensitivity, die bereits für SISO-Systeme besprochen wurde und analog angewendet werden kann. Die generalisierte Strecke $\mathbf{P}(s)$ ergibt sich damit zu

$$\begin{aligned} \mathbf{P}_{11} &= [0 \ W_p(s)]^T, & \mathbf{P}_{12} &= [W_u(s) \ -W_p(s)G(s)]^T \\ P_{21} &= 1, & P_{22} &= -G(s) \end{aligned} \quad (18.34)$$

Durch die Betrachtung der Singulärwerte ergibt sich die Gütfunktion im

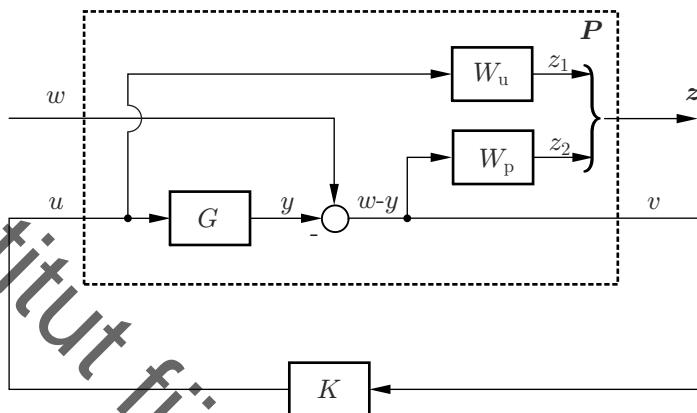


Bild 18-5: Struktur der Generalized Plant für die Mixed Sensitivity

vorliegenden Fall zu

$$\|\mathbf{F}(s)\|_\infty = \sqrt{(W_p S)^2 + (W_u K S)^2} \quad (18.35)$$

Das Größenverhältnis zwischen den beiden Filtern W_p und W_u kann also in gewisser Analogie zur optimalen Zustandsregelung als Gewichtung zwischen Stellauflauf und Regelabweichung gedeutet werden.

Neben der Mixed-Sensitivity, welche die Übertragungsfunktionen S und $K S$ miteinander gewichtet, sind auch zahlreiche andere Kombinationen denkbar, die aufgrund der Formulierung als Generalized Plant alle im identischen Optimierungsrahmen behandelt werden können. So berücksichtigt der populäre $S-KS-T$ -Entwurf zusätzlich noch eine Gewichtung der komplementären Sensitivität. Außerdem können gegebenenfalls auch Unsicherheiten der Regelstrecke mit in das Optimierungsproblem eingebracht werden.

19 Modellprädiktive Regelung (MPR)

19.1 Allgemeines

Die einfache Form der Optimierungsprobleme von LQR und LQG aus Kapitel 18 sorgen dafür, dass diese sich direkt analytisch lösen lassen. Das ist überraschend, da nach einer allgemeinen Funktion $u(t)$ gesucht wird, die als Funktion potentiell unendlich viele Freiheitsgrade besitzt und eine Lösung somit aus numerischer Sicht nicht unbedingt erwartet werden kann.

Der Preis für diese analytische Lösbarkeit ist jedoch, dass viel Potential verschenkt wird, welches sich bei der Betrachtung eines allgemeineren Optimierungsproblems ergäbe.

An dieser Stelle setzt das Verfahren der Modellprädiktiven Regelung (MPR, im Englischen „Model Predictive Control“, MPC) an, welches Ende der 1970er Jahre entwickelt wurde. Dabei bezeichnet MPR kein spezifisches Regelungsverfahren sondern vielmehr eine ganze Klasse an Regelungsmethoden, die auf der expliziten Nutzung eines Modells des zu regelnden Prozesses basieren, um mit diesem das Verhalten relevanter Prozessgrößen vorhersagen und in einer zu minimierenden Kostenfunktion bewerten zu können. Die Methodik ist dabei auf alle Prozessklassen (linear und nichtlinear) anwendbar. Für eine detaillierte Abhandlung sei auf [32] verwiesen.

Grundidee der MPR ist es, ein Optimierungsproblem ähnlich zu dem eines LQR zu formulieren, dies jedoch mit expliziten Beschränkungen, z. B. an die Stellgrößen, zu versehen. Das Optimierungsproblem für $u(t)$ ist dann im Allgemeinen nicht mehr lösbar.

Die MPR ermöglicht dennoch eine Lösung, indem von den potentiell unendlich vielen Freiheitsgraden von $u(t)$ auf eine endliche Menge von Entscheidungsvariablen gewechselt wird. Dies erreicht die MPR folgendermaßen:

Grundidee der MPR

Im allgemeinen Fall sind Optimierungsprobleme nach der Funktion $u(t)$ nicht lösbar. Daher betrachtet die MPR nur solche $u(t)$, welche stückweise konstant verlaufen und ihren Funktionswert nur endlich oft ändern.

Die in der MPR genutzten Stellgrößen entsprechen also einem zeitdiskreten Signal, welches mit einem Halteglied 0.ter Ordnung versehen wurde.

Durch die endliche Anzahl an zulässigen Änderungen der Stellgröße besitzt das Optimierungsproblem hierdurch endliche viele Entscheidungsvariablen und kann daher mit entsprechender Rechnerunterstützung numerisch gelöst werden.

Hierzu benötigt man ein (zeitdiskretes) Modell, welches den Zusammenhang zwischen Stellgrößen und Regelzielen angibt, sowie eine Kostenfunktion, welche die Erfüllung der Regelziele bewertet.

Da die Stellgröße sich nur endlich oft ändern darf, ist ein einmaliges Lösen des Optimierungsproblems keine Option. In diesem Fall würde die MPR nämlich nach Erfolgen der letzten Stellgrößenänderung nur noch einen konstanten Verlauf ausgeben und somit als Steuerung agieren. Hieraus ergeben sich die folgenden drei zentralen Merkmale einer MPR.

Merkmale einer MPR

Allen unter dem Namen MPR zusammengefassten Verfahren sind die folgenden Merkmale gemeinsam:

- Explizite Nutzung eines Prozessmodells zur Prädiktion zukünftiger Prozesszustände.
- Die Berechnung des Stellsignals basiert auf der Minimierung einer Kostenfunktion.
- In jedem Abtastschritt wird eine optimale Folge an Stellsignalen berechnet. Es wird jedoch nur der erste Wert dieser Stellgrößenfolge an den Prozess ausgegeben und die Berechnungen im folgenden Zeitschritt mit einem um einen Abtastschritt verschobenen Horizont wiederholt. Dieses Vorgehen wird als *Prinzip des zurückweichenden Horizonts* (engl. „receding horizon“) bezeichnet.

Die MPR ist das in der Industrie am weitesten verbreitete höhere Regelungsverfahren, da es gegenüber anderen Verfahren erhebliche Vorteile aufweist, welche man an dieser Stelle bereits erahnen kann. So kann durch die explizite Nutzung eines Modells die MPR vielfache Phänomene berücksichtigen, sofern diese im Modell abgebildet sind. Dies umfasst eine einfache Behandlung von Mehrgrößensystemen, da diese durch die Berücksichtigung im Modell entkoppelt werden können. Sind zukünftige Sollwerte bekannt, so können sogar Totzeiten kompensiert werden.

Dies alles leistet die MPR ohne zusätzliche Maßnahmen sondern einzig

durch Inkludieren des Modells. Hierdurch erhält die MPR einen intuitiven Charakter, welcher eine Bedienung auch bei begrenzter regelungstechnischer Ausbildung ermöglicht. Der vermutlich wichtigste Vorteil ist aber die explizite Berücksichtigung von Begrenzungen der Stell-, Zustands- und Ausgangsgrößen, die in das Optimierungsproblem nun eingebracht werden können.

Begrenzungen der Stellgrößen und der Stellgrößenänderungen durch einen limitierten Arbeitsbereich des Aktors oder eine eingeschränkte Aktordynamik sind in fast jeder realen Anwendung vorhanden. Sie werden in konventionellen Regelungsverfahren einerseits dadurch behandelt, dass man versucht, ein Erreichen der Grenzen durch eine entsprechend zurückhaltende Reglereinstellung zu vermeiden. Da dies nicht für jeden Anwendungsfall garantiert werden kann, benötigen Regler mit integrierendem Anteil zusätzlich Anti-Windup-Strategien (siehe Abschnitt 16.5). Da das Erreichen dieser Grenzen grundsätzlich zur Instabilität des Regelkreises führen kann, bietet ihre explizite Berücksichtigung einen großen Vorteil.

Begrenzungen von Prozesszustands- und -ausgangsgrößen sind oftmals direkt mit der Produktqualität, dem Profit oder dem sicheren Anlagenbetrieb verknüpft. In der Prozessindustrie werden die optimalen Arbeitspunkte der Anlagen häufig durch ein Optimierungsproblem mit dem Ziel der Gewinnmaximierung bestimmt.

Aufgrund der Natur solcher beschränkter Optimierungsprobleme liegen die optimalen Betriebspunkte zumeist nahe an den Betriesgrenzen der Anlagen, deren Verletzung katastrophale Folgen haben kann. Daher ist ein gewisser Sicherheitsabstand der Arbeitspunkte zu den Betriesgrenzen notwendig, den die MPR, die sich dieser Grenzen explizit bewusst ist, verringern kann.

Alle Modellgestützten Prädiktiven Regelungsverfahren sagen mit einem geeigneten Modell das Verhalten des Prozesses in Abhängigkeit von zukünftigen Stellgrößenverläufen vorher.

Notation prädizierter Größen

Der Ausdruck $\mathbf{a}(k+j|k)$ bezeichnet den Vektor \mathbf{a} zum Zeitpunkt $k+j$, welcher vom Zeitpunkt k aus prädiziert wurde. Die Darstellung $\mathbf{a}(\cdot|k)$ bezeichnet den gesamten prädizierten Verlauf.

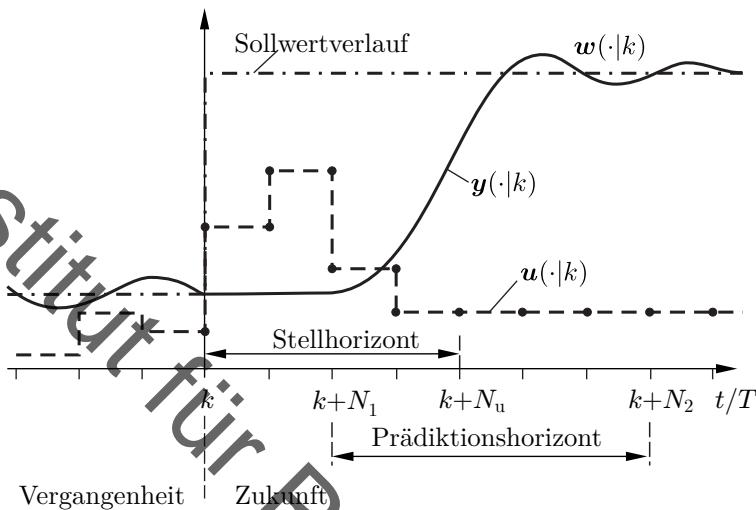


Bild 19-1: Prinzip der MPR

Ziel der Regelung ist es, dem zukünftigen Sollwertverlauf $w(\cdot|k)$ bestmöglich zu folgen. Folglich sollte im Sinne einer optimalen Regelung die Abweichung des präzisierten Verlaufs der Regelgröße von der Solltrajektorie minimieren werden. Analog zum LQR wird man dabei neben der Regelabweichung auch die aufzuwendende Stellenergie bestrafen wollen. Beide Größen können im Gegensatz zum LQR nicht auf dem kompletten, unendlichen Zeitintervall in die Kostenfunktion eingebracht werden, da ein numerischer Optimierer Größen endlicher Dimension verlangt – siehe auch Bild 19-1.

Horizonte in der MPR

Der Zeithorizont, über welchem die Kostenfunktion der MPR ausgewertet wird, ist begrenzt auf den *Prädiktionshorizont* zwischen $k + N_1$ bis $k + N_2$ für die Regelabweichung und den *Stellhorizont* von k bis $k + N_u - 1$.

Die Grenzen N_1 , N_2 und N_u der Horizonte sind dabei Parameter der Regelung, welche festzulegen sind. Für $u(\cdot|k)$ wird die strukturelle Annahme getroffen, dass sich $u(\cdot|k)$ ab dem Zeitpunkt $k + N_u - 1$ nicht mehr ändert,

d. h. $\mathbf{u}(k + N_u + j | k) = \mathbf{u}(k + N_u - 1 | k)$ für alle $j \geq 0$ gilt. Dann gelten folgende Empfehlungen:

Wahl der Horizonte in der MPR

Es ist sinnvoll, den unteren Prädiktionshorizont N_1 so groß wie die Systemtotzeit zu wählen, da die zu berechnenden aktuellen Eingangsgrößen erst nach dieser Totzeit eine Auswirkung auf die Regelgröße haben.

Der obere Prädiktionshorizont N_2 ist so zu wählen, dass er die wesentliche Dynamik des Prozessmodells erfasst.

Der maximal sinnvolle Stellhorizont N_u beträgt N_2 verringert um die Streckentotzeit, da weiter in der Zukunft liegende Stellgrößen für kausale Systeme offensichtlich keinen Einfluss auf $\mathbf{y}(\cdot | k)$ im Prädiktionshorizont besitzen.

Die Quantifizierung der Abweichung des Regelgrößenverlaufs und der Solltrajektorie erfolgt durch die skalare Kostenfunktion J . Ähnlich zu LQR führt die Verwendung einer quadratischen Abweichung auf ein strukturell günstiges Optimierungsproblem. Zur einfacheren Notation nutzt man dabei die Schreibweise

$$\|\mathbf{s}\|_{\mathbf{A}}^2 = \mathbf{s}^T \mathbf{A} \mathbf{s} \quad (19.1)$$

für die gewichtete quadratische 2-Norm. Für positiv definite \mathbf{A} ist $\|\cdot\|_{\mathbf{A}}$ eine Norm; für positiv semidefinite \mathbf{A} ist $\|\cdot\|_{\mathbf{A}}$ nur eine Halbnorm.

In Analogie zum LQR nutzt man nun die Kostenfunktion

$$\begin{aligned} J(\mathbf{u}(\cdot | k)) &= \left\| \begin{bmatrix} \mathbf{y}(k + N_1 | k) \\ \vdots \\ \mathbf{y}(k + N_2 | k) \end{bmatrix} - \begin{bmatrix} \mathbf{w}(k + N_1 | k) \\ \vdots \\ \mathbf{w}(k + N_2 | k) \end{bmatrix} \right\|_{\mathbf{Q}}^2 \\ &\quad + \left\| \begin{bmatrix} \Delta \mathbf{u}(k | k) \\ \vdots \\ \Delta \mathbf{u}(k + N_u - 1 | k) \end{bmatrix} \right\|_{\mathbf{R}}^2 \end{aligned} \quad (19.2)$$

mit den Gewichtungsmatrizen \mathbf{Q} und \mathbf{R} . Diese wird man in den meisten Fällen zu Diagonalmatrizen wählen, deren Einträge wiederum Parameter

der MPR sind. Damit ergibt sich

$$J = \sum_{i=N_1}^{N_2} \|y(k+i|k) - w(k+i|k)\|_{Q(i)}^2 + \sum_{i=0}^{N_u-1} \|\Delta u(k+i|k)\|_{R(i)}^2. \quad (19.3)$$

Alle Einträge von \mathbf{Q} und \mathbf{R} sind (strikt) positiv.

Ableichend zum LQR hat es sich eingebürgert, nicht die Stellgröße \mathbf{u} , sondern deren Änderung $\Delta \mathbf{u}(\cdot|k)$ zu bestrafen. Dies liegt daran, dass im Gegensatz zum LQR der Zustand nicht nach $\mathbf{0}$ überführt werden soll, sondern eine Folgeregelung umgesetzt werden soll, weswegen die Stellgröße im Allgemeinen für $k \rightarrow \infty$ nicht null wird. Der mit \mathbf{R} gewichtete Term soll große Stellgrößenschwankungen vermeiden und somit die aufzuwendende Stellenergie (die wiederum direkte Produktionskosten verursacht) begrenzen.

Die Berechnung der optimalen Stellgrößenfolge ergibt sich aus der Minimierung von J , d. h.

$$\mathbf{u}(\cdot|k)_{\text{opt}} = \arg \min_{\mathbf{u}(\cdot|k)} J(\mathbf{u}(\cdot|k)) \quad (19.4)$$

unter den Nebenbedingungen der (zeitdiskreten, nichtlineare) Systemdynamik und weiteren Beschränkungen

$$\begin{aligned} \mathbf{x}(k+j+1|k) &= \mathbf{f}(\mathbf{x}(k+j|k), \mathbf{u}(k+j|k)), \\ \mathbf{y}(k+j|k) &= \mathbf{h}(\mathbf{x}(k+j|k), \mathbf{u}(k+j|k)), \\ \mathbf{u}(k+j|k) &= \mathbf{u}(k+j-1|k) + \Delta \mathbf{u}(k+j|k) \\ g_{\text{in}}(\mathbf{x}(k+j|k), \mathbf{y}(k+j|k), \mathbf{u}(k+j|k), \Delta \mathbf{u}(k+j|k)) &\leq \mathbf{0} \\ g_{\text{eq}}(\mathbf{x}(k+j|k), \mathbf{y}(k+j|k), \mathbf{u}(k+j|k), \Delta \mathbf{u}(k+j|k)) &= \mathbf{0}, \end{aligned} \quad (19.5)$$

wobei die Funktionen \mathbf{f} und \mathbf{h} das für die Prädiktion verwendete Prozessmodell angeben, und die Funktionen \mathbf{g}_{in} und \mathbf{g}_{eq} die zu berücksichtigenden Nebenbedingungen bezeichnen, welche sich durch Beschränkungen der Stell-, Zustands- und Ausgangsgrößen ergeben. Hierbei steht die Tiefstellung „in“ für das englische Wort „inequality“ und „eq“ für „equality“.

Der prädiktive Regler gibt zu jedem Abtastschritt nur den ersten Wert $\mathbf{u}(k|k)_{\text{opt}}$ der optimalen Stellfolge an den Prozess aus. Im folgenden Abschnitt wiederholt er die Berechnung von $\mathbf{u}(\cdot|k)_{\text{opt}}$, wobei er das Zeitfenster, über das die Kostenfunktion gebildet wird, um einen Abtastschritt verschiebt (Prinzip des zurückweichenden Horizonts). Damit ergibt sich die in Bild 19-2 dargestellte Regelungsstruktur.

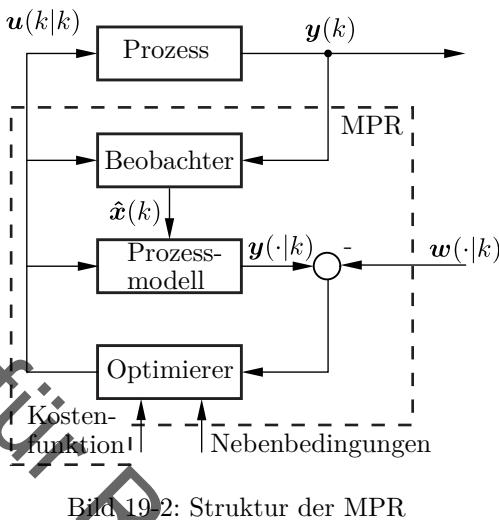


Bild 19-2: Struktur der MPR

Ein wichtiger Sonderfall tritt auf, wenn alle Beschränkungen – und damit insbesondere auch das zur Prädiktion verwendete Prozessmodell – linear oder affin sind.

Lineare MPR

Die MPR wird als *Lineare MPR* bezeichnet, wenn alle auftretenden Beschränkungen linear oder affin sind.

Damit darf die Lineare MPR nicht mit einem linearen Verhalten des Reglers verwechselt werden. Trotz eines linearen internen Modells ist dessen Verhalten bei einer Berücksichtigung von Ungleichheitsnebenbedingungen in der Optimierung nichtlinear.

Die Unterscheidung ist daher nicht vom Regelverhalten motiviert, sondern entstammt numerischen Überlegungen: Optimierungsprobleme mit quadratischen Zielfunktionen und linearen Nebenbedingungen führen auf sogenannte Quadratische Programme, die sich effizient lösen lassen. Somit sind Lineare MPRs mit vergleichsweise überschaubarem Aufwand und weitaus weniger Investitionen in die numerischen Gegebenheiten umsetzbar.

19.2 Quadratische Programme

Zur Formulierung des linearen Prozessmodells gibt es verschiedene Möglichkeiten. Abseits anderer Ansätzen wie zeitdiskreten Übergangsfolgenmodellen oder Übertragungsfunktionen erweist sich dabei die Verwendung eines diskreten Zustandsraummodells der Form

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) \\ \mathbf{y}(k) &= \mathbf{C}\mathbf{x}(k) \end{aligned} \quad (19.6)$$

für die Formulierung der MPR-Problemstellung als besonders vorteilhaft und soll aus diesem Grund im Folgenden betrachtet werden.

Offenbar ist die bestmögliche Prädiktion für $\mathbf{x}(k+1|k)$ und $\mathbf{y}(k+1|k)$ durch

$$\begin{aligned} \mathbf{x}(k+1|k) &= \mathbf{A}\mathbf{x}(k) + \mathbf{A}\mathbf{u}(k-1) + \mathbf{A}\Delta\mathbf{u}(k|k) \\ \mathbf{y}(k+1|k) &= \mathbf{C}[\mathbf{a}\mathbf{x}(k) + \mathbf{A}\mathbf{u}(k-1) + \mathbf{A}\Delta\mathbf{u}(k|k)] \end{aligned} \quad (19.7)$$

gegeben. Daraus folgt direkt

$$\begin{aligned} \mathbf{y}(k+j|k) &= \mathbf{C} \left[\mathbf{A}^j \mathbf{x}(k) + \left(\sum_{i=1}^j \mathbf{A}^{i-1} \right) \mathbf{A}\mathbf{u}(k-1) \right] \\ &\quad + \mathbf{C} \left[\sum_{m=0}^{j-1} \left(\left(\sum_{i=1}^{j-m} \mathbf{A}^{i-1} \right) \mathbf{A} \Delta\mathbf{u}(k+m|k) \right) \right] . \end{aligned} \quad (19.8)$$

Damit ergibt sich für die Vorhersage im gesamten Prädiktionshorizont unter

der Beachtung, dass $\Delta u(k+j|k) = 0$ für $j \geq N_u$ gilt

$$\begin{aligned} \mathbf{y}(\cdot|k) &= \begin{bmatrix} \mathbf{y}(k+N_1|k) \\ \vdots \\ \mathbf{y}(k+N_2|k) \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{CA}^{N_1} \\ \vdots \\ \mathbf{CA}^{N_2} \end{bmatrix} \mathbf{x}(k) + \begin{bmatrix} \mathbf{C} \left(\sum_{i=0}^{N_1-1} \mathbf{A}^i \right) \mathbf{A} \\ \vdots \\ \mathbf{C} \left(\sum_{i=0}^{N_2-1} \mathbf{A}^i \right) \mathbf{A} \end{bmatrix} \mathbf{u}(k-1) \quad (19.9) \\ &\quad + \begin{bmatrix} \Phi(N_1) & \cdots & \Phi(N_1 - N_u + 1) \\ \vdots & \ddots & \vdots \\ \Phi(N_2) & \cdots & \Phi(N_2 - N_u + 1) \end{bmatrix} \cdot \begin{bmatrix} \Delta \mathbf{u}(k|k) \\ \vdots \\ \Delta \mathbf{u}(k+N_u-1|k) \end{bmatrix} \\ \text{mit } \Phi(i) &= \begin{cases} \mathbf{C} \left(\sum_{j=0}^{i-1} \mathbf{A}^j \right) \mathbf{A} & \text{für } i \geq 1 \\ \mathbf{0} & \text{für } i < 1 \end{cases}. \end{aligned}$$

Da $\mathbf{x}(k)$ und $\mathbf{u}(k-1)$ bekannt sind und zum Zeitpunkt k nicht mehr verändert werden können, teilt sich die Kostenfunktion in zwei Teile:

Erzwungene und freie Regelgröße

Der Anteil

$$\mathbf{f}(\cdot|k) = \begin{bmatrix} \mathbf{CA}^{N_1} \\ \vdots \\ \mathbf{CA}^{N_2} \end{bmatrix} \mathbf{x}(k) + \begin{bmatrix} \mathbf{C} \left(\sum_{i=0}^{N_1-1} \mathbf{A}^i \right) \mathbf{A} \\ \vdots \\ \mathbf{C} \left(\sum_{i=0}^{N_2-1} \mathbf{A}^i \right) \mathbf{A} \end{bmatrix} \mathbf{u}(k-1) \quad (19.10)$$

wird als *freie Regelgröße* bezeichnet, da er den erwarteten Verlauf $\mathbf{y}(\cdot|k)$ darstellt, wenn keine Änderungen der Eingangsgrößen vorgenommen werden.

Der Anteil $\Phi \Delta \mathbf{u}(\cdot|k)$ wird dagegen *erzwungene Regelgröße* genannt.

Für die Kostenfunktion ergibt sich damit

$$J = \|\mathbf{f}(\cdot|k) + \Phi \Delta \mathbf{u}(\cdot|k) - \mathbf{w}(\cdot|k)\|_{\mathbf{Q}}^2 + \|\Delta \mathbf{u}(\cdot|k)\|_{\mathbf{R}}^2 . \quad (19.11)$$

Mit der freien Regelabweichung $e(\cdot|k) = \mathbf{f}(\cdot|k) - \mathbf{w}(\cdot|k)$ erhält man

$$\begin{aligned} J &= e(\cdot|k)^T \mathbf{Q} e(\cdot|k) + 2e(\cdot|k)^T \mathbf{Q} \Phi \Delta \mathbf{u}(\cdot|k) \\ &\quad + \Delta \mathbf{u}(\cdot|k)^T (\Phi^T \mathbf{Q} \Phi + \mathbf{R}) \Delta \mathbf{u}(\cdot|k) \end{aligned} \quad (19.12)$$

und damit ein sogenanntes *Quadratisches Programm*.

Quadratisches Programm

Das Optimierungsproblem mit Entscheidungsvariable \mathbf{x} mit einer quadratischen Kostenfunktion

$$\min_{\mathbf{x}} \mathbf{x}^T \mathbf{H} \mathbf{x} + \mathbf{q} \mathbf{x} + c \quad (19.13)$$

und linearen Gleichungs- und Ungleichungsnebenbedingungen

$$\begin{aligned} \mathbf{A} \mathbf{eq} \mathbf{x} - \mathbf{b} \mathbf{eq} &= \mathbf{0} \\ \mathbf{A} \mathbf{in} \mathbf{x} - \mathbf{b} \mathbf{in} &\leq \mathbf{0} \end{aligned} . \quad (19.14)$$

heißt *Quadratisches Programm*

Die in Gl.(19.12) vorliegende Entscheidungsvariable ist $\Delta \mathbf{u}(\cdot|k)$. Sofern die Beschränkungen an \mathbf{u} , \mathbf{x} und \mathbf{y} in Form von einzuhaltenden Minimal- bzw. Maximalwerten gegeben sind, lassen sich diese linearen Beschränkungen aufgrund des linearen Prozessmodells in äquivalente lineare Beschränkungen der Stellgrößen umrechnen lassen. Daher entspricht das zu lösende Optimierungsproblem einem quadratischen Programm.

19.3 Unbeschränkte lineare MPR

Zunächst wird angenommen, dass abseits des bereits in die Kostenfunktion eingesetzten Prozessmodells keine weiteren Nebenbedingungen existieren. Bei dieser Abwesenheit von Beschränkungen lässt sich die Lösung $\Delta \mathbf{u}(\cdot|k)_{\text{opt}}$ analytisch berechnen.

Analog zur Herleitung des LQR leitet man hierzu den Kostenfunktion ab

und erhält die Bedingung

$$\frac{\partial J}{\partial \Delta u(\cdot|k)} = 2\Phi^T Q^T e(\cdot|k) + 2(\Phi^T Q \Phi + R) \Delta u(\cdot|k)_{\text{opt}} \stackrel{!}{=} \mathbf{0} \quad . \quad (19.15)$$

Hieraus gewinnt man

$$\Delta u(\cdot|k)_{\text{opt}} = -(\Phi^T Q \Phi + R)^{-1} \Phi^T Q^T e(\cdot|k) \quad . \quad (19.16)$$

Damit diese Lösung auch ein Minimum ist, muss die zweiten Ableitung (also die Hessematrix) positiv definit sein. Diese berechnet sich zu

$$H = \Phi^T Q \Phi + R \quad (19.17)$$

und ist genau dann positiv definit, wenn Q mindestens positiv semidefinit und R positiv definit ist. Dies entspricht mathematisch genau den Bedingungen an ein konkaves Optimierungsproblem.

Nebenbei ist auch die Invertierbarkeit der Hessematrix sichergestellt und $\Delta u(\cdot|k)_{\text{opt}}$ wohldefiniert.

Da jeweils nur der aktuelle Stelleingriff $\Delta u(k|k)_{\text{opt}}$ ausgegeben wird, ergibt sich das Reglergesetz

$$\Delta u(k|k)_{\text{opt}} = -K_{\text{MPC}} e(\cdot|k) \quad (19.18)$$

mit

$$K_{\text{MPC}} = [I \quad \mathbf{0} \quad \cdots \quad \mathbf{0}] (\Phi^T Q \Phi + R)^{-1} \Phi^T Q^T \quad . \quad (19.19)$$

Da K_{MPC} lediglich von den System- und Gewichtungsmatrizen abhängt und $e(\cdot|k)$ sich linear aus $x(k)$ berechnet, erhält man somit eine lineare Zustandsregelung, deren Eigenschaften sich mit der linearen Regelungstheorie untersuchen lassen.

¹Ludwig Otto Hesse (1811-1874), deutscher Mathematiker [18]

Linear unbeschränkte MPR

Besitzt eine lineare MPR neben der Systemdynamik keine weiteren Beschränkungen, so ergibt sich ein linearer Zustandsregler, der nach Gl.(19.19) einmalig berechnet werden kann.

Die resultierende Regelungsstruktur ist (vereinfacht für volle Zustandsgrößenmessung) in Bild 19-3 dargestellt.

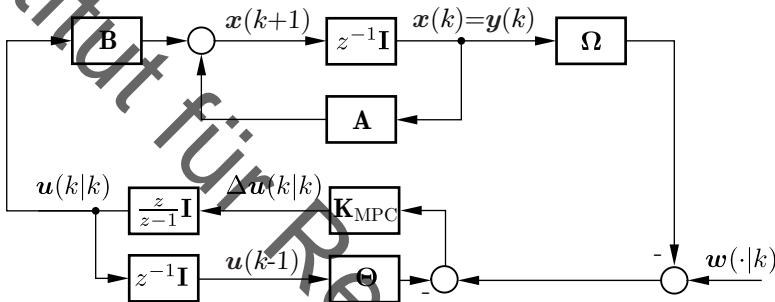


Bild 19-3: Struktur der linearen unbeschränkten MPR

Die verwendeten Matrizen Θ und Ω ergeben sich dabei zu

$$\Theta = \begin{bmatrix} C(A^{N_1-1} + \dots + I)A \\ \vdots \\ C(A^{N_2-1} + \dots + I)A \end{bmatrix} \quad \text{und} \quad \Omega = \begin{bmatrix} CA^{N_1} \\ \vdots \\ CA^{N_2} \end{bmatrix}. \quad (19.20)$$

19.4 Beschränkte lineare MPR

Wie bereits in den vorangehenden Abschnitten erwähnt, liegt der Hauptvorteil der MPR in der expliziten Berücksichtigung von Beschränkungen der Stell-, Zustands- und Regelgrößen begründet. Daher sind Prozesse mit Beschränkungen das Hauptanwendungsgebiet der MPR und man sieht folglich selten den Sonderfall einer unbeschränkten linearen MPR.

Für den allgemeinen, beschränkten Fall stehen für Quadratische Programme effiziente numerische Verfahren zur Verfügung. Deren Ideen sollen hier in Grundzügen angesprochen werden, da sich hieraus neben Implementierungshinweisen auch interessante Schlüsse auf die Struktur des entstehen-

den nichtlinearen Reglers ergeben. Bild 19-4 verdeutlicht schematisch die Lösung des Optimierungsproblems.

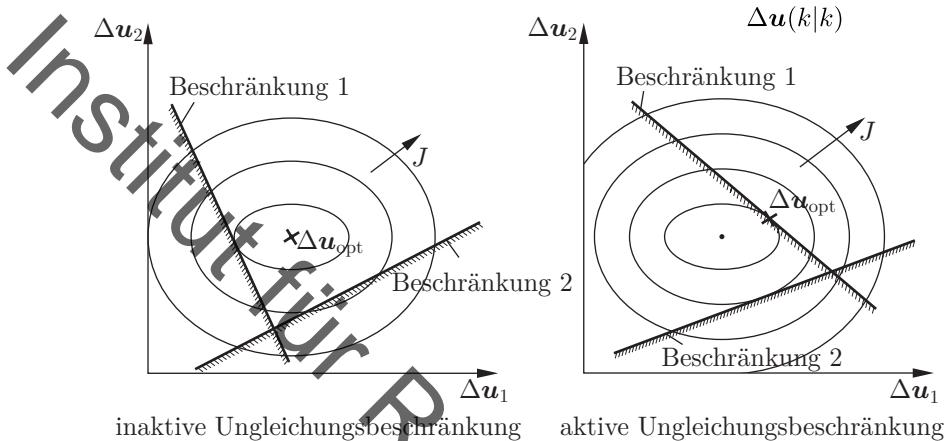


Bild 19-4: Beschränkte quadratische Optimierung

Da Gleichungsnebenbedingungen immer erfüllt sein müssen, schränken sie den Freiheitsgrad des Optimierungsproblems, d. h. die Anzahl an frei wählbaren Optimierungsvariablen Δu_i , ein.

Die Lösung des Optimierungsproblems bei ausschließlichem Vorhandensein von Gleichungsnebenbedingungen lässt sich somit immer als ein unbeschränktes Optimierungsproblem auf einem Subraum des von den Variablen Δu_i aufgespannten Raumes formulieren.

Strukturell ist genau dies beim Herleiten der unbeschränkten linearen MPR geschehen, indem die Modellgleichungen in die Kostenfunktion eingesetzt worden sind. Treten nun zusätzlich Ungleichungsnebenbedingungen auf, so können diese entweder aktiv oder inaktiv sein.

Aktive und inaktive Ungleichungsnebenbedingungen

Bei *inaktiven* Ungleichungsnebenbedingungen liegt das analytische Optimum Δu_{opt} im zulässigen Gebiet. Ihr Vorhandensein ist somit irrelevant und sie können ignoriert werden.

Verletzt das analytische Optimum jedoch eine der Ungleichungsnebenbedingungen, so wird diese *aktiv*. Dann ist es offensichtlich, dass die optimale Lösung auf dieser Beschränkung liegen muss, und diese aktive Beschränkung somit zur Gleichungsnebenbedingung wird.

Je nach dem, ob eine Ungleichungsnebenbedingung aktiv oder inaktiv ist, kann sie also weggelassen werden oder wird zu einer Gleichungsnebenbedingung, für die wie bekannt verfahren werden kann. Der Kern der so genannten „Active-set“-Strategien liegt im Auffinden der active sets – also der Menge der aktiven Ungleichungsnebenbedingungen – durch eine gezielte Iteration. Der Vorteil dieser Methoden ist, dass ausgehend von einem zulässigen Startpunkt der Optimierung in jedem Iterationsschritt eine bessere, ebenfalls zulässige Lösung gefunden wird. Dies ist vor allem im Hinblick auf ein verfrühtes Abbrechen der Optimierung (z. B. aus Echtzeitgründen) im realen Betrieb von großer Bedeutung.

Der Nachteil dieser Methoden ist, dass ihr Rechenbedarf exponentiell mit der Größe des Optimierungsproblems anwachsen kann. Außerdem erfolgen die maßgeblichen Reduktionen der Kostenfunktion zumeist erst am Ende der Iterationsschleife, so dass bei einem vorzeitigen Abbruch unter Umständen nur sehr suboptimale Lösungen zur Verfügung stehen.

Von Vorteil ist allerdings, dass bei einer wiederholten Lösung eines ähnlichen Optimierungsproblems (wie dies bei der MPR wegen des zurückweichenden Horizonts der Fall ist) die active sets der vorherigen Iteration als Startlösung verwendet werden können. Hierdurch können Geschwindigkeitsvorteile bei späteren Iterationen erzielt werden.

Stellen sich active-set-Strategien als nicht ausreichend heraus, so gibt es noch andere Klassen von Verfahren zur Lösung Quadratischer Programme wie beispielsweise die „Interior Point“-Methoden. Ihr Vorteil ist der nur polynomial mit der Problemgröße anwachsende Rechenbedarf und die gleichmäßige Konvergenz zum Optimum.

Sie besitzen jedoch den Nachteil, dass die Zwischenlösungen in den Iterationsschritten unzulässig sein können. In jedem Fall muss – abweichend zum Fall der unbeschränkten linearen MPR – das Optimierungsproblem in jedem Zeitschritt online gelöst werden. Hieraus ergeben sich abhängig von der verfügbaren Rechenleistung und der Abtastrate der Regelung möglicherweise Einschränkungen an die Wahl der Einstellparameter.

Da ein Optimierungsproblem mit ausschließlich Gleichungsnebenbedingungen zu einem äquivalenten unbeschränkten Optimierungsproblem auf einem Unterraum der Optimierungsvariablen führt, muss zu jedem active set ein bestimmtes lineares Regelgesetz gehören.

Welche Ungleichungsnebenbedingungen bei der aktuellen Minimierung von $J(\Delta u(\cdot|k))$ am Optimum aktiv sind, hängt dabei von der zukünftigen Sollwertfolge $w(\cdot|k)$ und dem aktuellen Zustand $x(k)$ ab.

Struktur der beschränkten linearen MPR

Da zu jedem active set genau ein lineares Regelgesetz gehört, ergibt sich das nichtlineare Stellgesetz der beschränkten linearen MPR als stückweise linear.

Diese Tatsache macht man sich die *explizite MPR* zu Nutze, die $x(k)$ und $w(\cdot|k)$ als Parameter des Optimierungsproblems auffasst. Hiermit berechnet man vorab alle möglichen linearen Regelgesetze mit ihren zugehörigen Gültigkeitsbereichen in $x(\cdot)$ und $w(\cdot|k)$.

Explizite MPR

Die explizite MPR extrahiert im laufenden Betrieb anhand der aktuellen Werte $x(k)$ und $w(k)$ lediglich das zugehörige lineare Reglergesetz aus einer vorab berechneten Datenbank.

Dies kann zu erheblichen Rechenzeitersparnissen führen. Es ist jedoch offensichtlich, dass die Zahl an möglichen active sets und damit die Anzahl an linearen Reglergesetzen exponentiell mit der Problemgröße wächst und sich dieses Verfahren damit nur auf Systeme mit wenigen Zustandsgrößen und Nebenbedingungen anwenden lässt.

19.5 Einstellparameter

Als Einstellparameter ergeben sich die Gewichtungsmatrizen \mathbf{Q} und \mathbf{R} , die Horizonte N_1 , N_2 und N_u und die Dynamik der Solltrajektorie $w(\cdot|k)$. Die Einflüsse dieser Parameter auf das Reglerverhalten sollen im Folgenden qualitativ untersucht werden.

Bei der Wahl von \mathbf{Q} und \mathbf{R} ist vor allem das Verhältnis der entsprechenden Einträge von \mathbf{Q} und \mathbf{R} von Bedeutung, da dieses Verhältnis die Forderung nach einem guten Folgen der Solltrajektorie gegenüber der dafür aufzuwen-

denden Stelleingriffe abwägt. Eine typische Wahl ist daher analog zum LQR

$$\mathbf{Q} = \mathbf{I}, \quad \mathbf{R} = \lambda \cdot \mathbf{I} \quad . \quad (19.21)$$

Mit steigendem λ werden die Stelleingriffe stärker bestraft. Dies führt zu einem ruhigen Verlauf der Stellgrößen ohne starke Schwankungen, was die Regelung üblicherweise verlangsamt. Im Grenzfall $\lambda \rightarrow \infty$ werden Stelleingriffe vollständig unterdrückt, was äquivalent zu einer Öffnung des Regelkreises ist. Die Pole des Regelkreises entsprechen dann den Polen des offenen Regelkreises, was bei instabilen Regelstrecken zur Instabilität führt.

Wie bereits in Abschnitt 19.1 erwähnt, ist eine Wahl des unteren Prädiktionshorizontes N_1 kleiner als die Totzeit der Regelstrecke nicht gut, da die aktuell zu berechnenden Stellgrößen erst nach der Systemtotzeit eine Auswirkung auf die Regelgröße besitzen. Somit erhöht dies nur die Länge der Vektoren \mathbf{y} und \mathbf{x} und damit die Rechenzeit ohne Gegenleistung.

Allerdings kann es auch gefährlich sein, N_1 wesentlich größer als die Streckentotzeit zu wählen, da man so den wesentlichen dynamischen Übergang abschneidet und aus den Kosten entfernt. In vielen Anwendungen ist aber gerade dieser Übergang von besonderem Interesse.

Daher wählt man in den meisten Fällen $N_1 = d$ mit der Streckentotzeit d in Abtastschritten und im Zweifel N_1 lieber zu klein als zu groß.

Der obere Prädiktionshorizont N_2 sollte die wesentliche Dynamik der Regelstrecke erfassen. Eine zu kleine Wahl von N_2 kann zur Instabilität des Regelkreises führen, da der Regler nicht in der Lage ist, die Auswirkungen aktueller Stelleingriffe vollständig zu erfassen. Für $N_2 \rightarrow \infty$ konvergiert die Problemstellung der MPR gegen die der optimalen Zustandsregelung, die stets einen stabilisierenden Regler liefert, falls das Optimierungsproblem zulässig ist. Es ist daher offensichtlich, dass ein steigender oberer Prädiktionshorizont N_2 den Regelkreis stabilisiert. Allerdings steigt mit N_2 auch der Rechenaufwand, weswegen N_2 nicht ohne Not übermäßig groß gewählt werden sollte.

Allgemein hat die MPR mit der Stabilisierung instabiler Systeme in vielen Fällen Probleme. Dies liegt u. a. an numerischen Gründen, da für instabile Systeme mit $|\lambda| > 1$ die verwendeten Potenzierungen der Matrizen \mathbf{A}^k in Gl.(19.8) numerisch schlecht konditioniert sind.

Sind zukünftige Sollwertverläufe bekannt, kommt dem oberen Prädiktionshorizont eine weitere wichtige Rolle zu: Da in der Kostenfunktion $\mathbf{y} - \mathbf{w}$ über den Prädiktionshorizont bestraft wird, entstehen bei bekanntem \mathbf{w} zum Zeitpunkt k bereits Kosten, selbst wenn sich erst $\mathbf{w}(k+N_2|k)$ erstmalig ändert. Folglich wird ein Sollwertsprung in $\mathbf{w}(k+N_2|k)$ bereits in $\Delta\mathbf{u}(k|k)$ mit einfließen. Hierdurch kann die MPR prädiktiv bereits vor Eintreten einer Sollwertänderung reagieren und somit ein aktausales Führungsverhalten erreichen.

Der Stellhorizont N_u bestimmt die Freiheitsgrade der Optimierung. Eine Erhöhung von N_u bewirkt somit üblicherweise eine schnellere Dynamik des Regelkreises. Entsprechend der Einstellung von N_2 ist der sinnvolle Wertebereich von N_u nach oben beschränkt. Da N_u die Anzahl der Entscheidungsvariablen direkt erhöht, ist der Einfluss von N_u auf die benötigte Rechenzeit besonders entscheidend. Daher versucht man oft, mit geringen N_u auszukommen. Für viele gutmütige Regelstrecken ist dabei eine Wahl von $N_u \leq 3$ bereits ausreichend.

Ein Sonderfall ergibt sich für den unbeschränkten Fall ohne Stellgrößengewichtung ($\mathbf{R} = \mathbf{0}$) für die Wahl $N_1 = N_u \geq n$ bei hinreichend großem N_2 , wobei n die Ordnung des Regelkreises bezeichnet. Da aus der Theorie der Regler mit endlicher Einstellzeit bekannt ist, dass das System in maximal n Schritten auf den gewünschten Sollwert gebracht werden kann, beträgt der optimale Wert der Kostenfunktion J_{opt} für diese Parameterwahl $J_{\text{opt}} = 0$, da $\mathbf{y}(k+j|k) \equiv \mathbf{w}(k+j|k)$ für $j \geq n$ möglich ist. Die resultierende MPR muss daher ein Regler mit endlicher Einstellzeit sein.

19.6 Stationäre Genauigkeit

Eine wichtige Eigenschaft eines Regelkreises ist dessen stationäre Genauigkeit. Dies wird in der klassischen linearen Reglerauslegung durch ein integrierendes Verhalten des aufgeschnittenen Regelkreises sichergestellt. Aufgrund des Prinzips der MPR, die die an den Prozess ausgegebenen Stellgrößen so berechnet, dass die prädizierte Regelabweichung minimiert wird, wird diese Regelabweichung – sofern der vorgegebene Sollwert realisierbar ist – im stationären Fall theoretisch zu null.

Dies gilt unter der Voraussetzung, dass das reglerinterne Modell den Prozess exakt abbildet. Dies ist jedoch im Allgemeinen nicht der Fall. Während der

reale Prozess durch

$$\begin{aligned}\mathbf{x}(k+1) &= \mathbf{A}\mathbf{x}(k) + \mathbf{A}\mathbf{u}(k) + \mathbf{H}_1\mathbf{d}(k) \\ \mathbf{y}(k) &= \mathbf{C}\mathbf{x}(k) + \mathbf{H}_2\mathbf{d}(k)\end{aligned}\quad (19.22)$$

beschrieben ist, bildet das reglerinterne Modell

$$\begin{aligned}\mathbf{x}(k+1|k) &= \mathbf{A}\mathbf{x}(k) + \mathbf{A}\mathbf{u}(k) \\ \mathbf{y}(k|k) &= \mathbf{C}\mathbf{x}(k|k)\end{aligned}\quad (19.23)$$

den realen Prozess nicht vollständig ab. Die Abweichungen $\mathbf{H}_1\mathbf{d}(k)$ und $\mathbf{H}_2\mathbf{d}(k)$ umfassen dabei sowohl die Wirkung externer Störgrößen als auch allgemeine Abweichungen bei der Modellierung der Systemdynamik.

Betrachtet man den stationären Zustand \mathbf{x}_∞ des Regelkreises, so ist dieser durch

$$\begin{aligned}\mathbf{x}_\infty &= \mathbf{A}\mathbf{x}_\infty + \mathbf{A}\mathbf{u}_\infty + \mathbf{H}_1\mathbf{d}_\infty \\ \mathbf{y}_\infty &= \mathbf{C}\mathbf{x}_\infty + \mathbf{H}_2\mathbf{d}_\infty\end{aligned}\quad (19.24)$$

gegeben. Die Prädiktion der freien Regelgröße mit Hilfe des reglerinternen Modells ergibt sich damit zu

$$\mathbf{f}(\cdot|k)_\infty = \begin{bmatrix} \mathbf{y}_\infty \\ \vdots \\ \mathbf{y}_\infty \end{bmatrix} - \begin{bmatrix} \mathbf{C} \left(\sum_{i=0}^{N_1-1} \mathbf{A}^i \right) \mathbf{H}_1 + \mathbf{H}_2 \\ \vdots \\ \mathbf{C} \left(\sum_{i=0}^{N_2-1} \mathbf{A}^i \right) \mathbf{H}_1 + \mathbf{H}_2 \end{bmatrix} \mathbf{d}_\infty \quad (19.25)$$

Mit $\mathbf{w}(\cdot|k)_\infty = [\mathbf{w}_\infty \ \cdots \ \mathbf{w}_\infty]^T$ ergibt sich für die freie Regelabweichung ohne Beschränkung der Allgemeinheit:

$$\mathbf{e}(\cdot|k)_\infty = \begin{bmatrix} \mathbf{I} \\ \vdots \\ \mathbf{I} \end{bmatrix} (\mathbf{y}_\infty - \mathbf{w}_\infty) - \begin{bmatrix} \mathbf{C} \left(\sum_{i=0}^{N_1-1} \mathbf{A}^i \right) \mathbf{H}_1 + \mathbf{H}_2 \\ \vdots \\ \mathbf{C} \left(\sum_{i=0}^{N_2-1} \mathbf{A}^i \right) \mathbf{H}_1 + \mathbf{H}_2 \end{bmatrix} \mathbf{d}_\infty \quad . \quad (19.26)$$

Aufgrund der stückweisen Linearität des linearen Modellprädiktiven Reglers ist dem stationären Zustand \mathbf{x}_∞ die lineare Rückführmatrix $\mathbf{K}_{\text{MPC},\infty}$ zugeordnet. Aus der Stationarität des Regelkreises folgt die Bedingung

$$\Delta \mathbf{u}_{\text{opt},\infty} = -\mathbf{K}_{\text{MPC},\infty} \mathbf{e}(\cdot | k)_\infty = \mathbf{0} \quad (19.27)$$

und damit

$$\mathbf{K}_{\text{MPC},\infty} \begin{bmatrix} \mathbf{I} \\ \vdots \\ \mathbf{I} \end{bmatrix} (\mathbf{y}_\infty - \mathbf{w}_\infty) = \mathbf{K}_{\text{MPC},\infty} \begin{bmatrix} \mathbf{C} \left(\sum_{i=0}^{N_1-1} \mathbf{A}^i \right) \mathbf{H}_1 + \mathbf{H}_2 \\ \vdots \\ \mathbf{C} \left(\sum_{i=0}^{N_2-1} \mathbf{A}^i \right) \mathbf{H}_1 + \mathbf{H}_2 \end{bmatrix} \mathbf{d}_\infty \quad .$$

Diese Betrachtung zeigt, dass sich bei einer Abweichung $\mathbf{d}_\infty \neq \mathbf{0}$ des reglerinternen Modells von dem realen Prozessverhalten zwangsläufig eine bleibende Regelabweichung $\mathbf{y}_\infty - \mathbf{w}_\infty \neq \mathbf{0}$ einstellt.

Störgrößenmodell und stationäre Genauigkeit

Aufgrund von Modellabweichungen ist ohne zusätzliche Maßnahmen die MPR nicht stationär genau. Der Schlüssel zum Erreichen stationärer Genauigkeit ist dementsprechend eine korrekte Prädiktion. Um dies zu erreichen ist das Prozessmodell um ein Störgrößenmodell zu erweitern.

Eine gängige Form eines Störgrößenmodells ist

$$\begin{aligned} \mathbf{x}(k+1|k) &= \mathbf{A}\mathbf{x}(k) + \mathbf{A}\mathbf{u}(k) + \mathbf{G}_1\mathbf{d}(k) \\ \mathbf{d}(k+1|k) &= \mathbf{A}_d\mathbf{d}(k) \\ \mathbf{y}(k|k) &= \mathbf{C}\mathbf{x}(k) + \mathbf{G}_2\mathbf{d}(k) + \mathbf{D}\mathbf{u}(k) \end{aligned} \quad . \quad (19.28)$$

Dies ist äquivalent zum erweiterten Zustandsraummodell

$$\begin{aligned} \tilde{\mathbf{x}}(k+1|k) &= \begin{bmatrix} \mathbf{x}(k+1|k) \\ \mathbf{d}(k+1|k) \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{G}_1 \\ \mathbf{0} & \mathbf{A}_d \end{bmatrix} \tilde{\mathbf{x}}(k) + \mathbf{A}\mathbf{u}(k) \\ \mathbf{y}(k|k) &= [\mathbf{C} \quad \mathbf{G}_2] \tilde{\mathbf{x}}(k) + \mathbf{D}\mathbf{u}(k) \end{aligned} \quad . \quad (19.29)$$

Man kann mit diesem erweiterten Zustandsraummodell und einem geeigneten Beobachter die nicht messbaren Störungen schätzen und in die Prädiktion einfließen lassen.

Durch diesen Störgrößenansatz lassen sich auch auftretende stationäre Regelabweichungen hervorgerufen durch inexakte Modellmatrizen \mathbf{A} , \mathbf{B} , \mathbf{C} oder \mathbf{D} beseitigen, da $\boldsymbol{\Omega} \mathbf{x}_\infty$ und $\boldsymbol{\Theta} \mathbf{u}_\infty$ konstante Vektoren sind, die sich durch geeignete Wahl von \mathbf{A}_d , \mathbf{G}_1 und \mathbf{G}_2 durch eine konstante Störung \mathbf{d}_∞ abbilden lassen.

Ein einfacher und effektiver Ansatz zur Wahl des Störgrößenmodells ist der einer konstanten Ausgangsstörung $\mathbf{G}_1 = \mathbf{0}$ und $\mathbf{A}_d = \mathbf{G}_2 = \mathbf{I}$. Die Schätzung der Störung kann dann aus der Abweichung der aktuellen Messung $\mathbf{y}_m(k)$ von der Prädiktion aus dem vorangegangenen Zeitschritt zu

$$\mathbf{d}(k) = \mathbf{y}_m(k) - \mathbf{y}(k|k-1) \quad (19.30)$$

erfolgen. Hierdurch bringt $\mathbf{A}_d = \mathbf{I}$ zusätzlich Pole bei $z = 1$ in den Regelkreis ein und stellt somit indirekt ein integrierendes Verhalten des aufgeschnittenen Regelkreises sicher.

In einigen Fällen sind jedoch kompliziertere Ansätze als dieser einer konstanten Ausgangsstörung notwendig. Aus den gemachten Ausführungen ist offensichtlich, dass die Wahl des Störgrößenmodells ein wichtiger Parameter der MPR ist, welcher zusammen mit den in Einstellparametern geeignet zu wählen ist.

19.7 Stabilität

Die wichtigste Eigenschaft eines Regelkreises ist dessen Stabilität. Nur im Falle der unbeschränkten linearen MPR ist der resultierende Regler linear und kann damit direkt durch Berechnung der Eigenwerte der Dynamikmatrix des geschlossenen Kreises auf Stabilität untersucht werden. In jedem anderen Fall ist der Regler entweder aufgrund des Prozessmodells oder der Beschränkungen nichtlinear und aus diesem Grund nur mit der nichtlinearen Stabilitätstheorie zu untersuchen. Da dies über den Rahmen dieser Einführung in die MPR hinaus geht, sollen an dieser Stelle nur einige Grundzüge zur Garantie der Stabilität skizziert werden.

Grundidee des Stabilitätsnachweises ist die Direkte Methode nach Lyapunov. Man versucht dabei zu zeigen, dass die Kostenfunktion der MPR eine Lyapunov-Funktion ist. Leider ist diese keine gültige Kandidatfunktion, da der betrachtete Zeithorizont endlich ist. Daher versucht man – abseits der Möglichkeit von $N_2 \rightarrow \infty$ – die Restkosten von $N_2 + 1$ bis ∞ geeignet zu kontrollieren. Dabei muss beachtet werden, dass – ausgehend von einer angenommenen existenten Lösung des Optimierungsproblem zum Zeitpunkt $k = 0$ – nicht automatisch sichergestellt ist, dass eine solche Lösung auch für nachfolgende Zeitschritte existiert.

Eine Möglichkeit zur Sicherung der Stabilität des Regelkreises ist die Einführung einer Beschränkung des Endzustands auf den zu stabilisierenden Zustand, welcher im Folgenden zu $\mathbf{0}$ angenommen wird:

$$\mathbf{x}(k + N_2 | k) = \mathbf{0} \quad . \quad (19.31)$$

Diese Forderung führt zu Stabilität, ist jedoch sehr restriktiv und kann unter Umständen zu nicht zulässigen Optimierungsproblemstellungen führen, d. h. es kann keine Trajektorie $\mathbf{u}(\cdot | k)$ gefunden werden, die alle Nebenbedingungen erfüllt.

Ein zweiter Ansatz ist die Gewichtung der Abweichung des Endzustandes vom zu stabilisierenden Zustand in der Kostenfunktion durch

$$J = \sum_{i=N_1}^{N_2} \|\mathbf{y}(k + i | k) - r(k + i | k)\|_{\mathbf{Q}(i)}^2 + \sum_{i=0}^{N_u-1} \|\Delta \mathbf{u}(k + i | k)\|_{\mathbf{R}(i)}^2 + \mathbf{x}^T(k + N_2 | k) \mathbf{T} \mathbf{x}(k + N_2 | k). \quad (19.32)$$

Bei hinreichend großer Wahl der Matrix \mathbf{T} garantiert dies Stabilität. Für $\mathbf{T} \rightarrow \infty$ geht dieser Ansatz in die Forderung $\mathbf{x}(k + N_2 | k) = \mathbf{0}$ über.

Komplexere Ansätze benutzen sowohl die Gewichtung des Endzustandes als auch eine zusätzliche Endbeschränkung

$$\mathbf{x}(k + N_2 | k) \in \mathcal{X}_E , \quad (19.33)$$

wobei \mathcal{X}_E ein geschlossenes Gebiet um den zu stabilisierenden Punkt darstellt. Grundüberlegung ist die garantierte Stabilität optimaler Zustandsregler. Gibt es also eine optimale Zustandstrajektorie aus X_E zum zu stabilisierenden Punkt, die keine Beschränkung verletzt und sind die Endkosten $\|\mathbf{x}(k + N_2 | k)\|_{\mathbf{T}}^2$ eine obere Schranke der optimalen Restkosten des Optimierungsproblems für $N_2 \rightarrow \infty$, so besitzt die resultierende MPR garantierte Stabilität.

19.8 Nichtlineare MPR

Auch bei Verwendung eines linearen Prozessmodells bei Berücksichtigung von Prozessbegrenzungen ist der resultierende Regler nichtlinear. Die Unterscheidung zwischen Linearer MPR und Nichtlinearer MPR erfolgt aber nicht entlang dieser Trennlinie sondern anhand der numerischen Erfordernisse zum Lösen des Optimierungsproblems.

Nichtlineare MPR

Eine MPR heißt dann *Nichtlineare MPR*, wenn das zugehörige Optimierungsproblem aufgrund nichtlinearer Beschränkungen (meistens wegen eines nichtlinearen Modells der Regelstrecke) keinem Quadratischen Programm, sondern einem Nichtlinearen Programm entspricht.

Eine mögliche Behandlung nichtlinearer Prozessmodelle besteht darin, das Prozessmodell in jedem Zeitschritt erneut zu linearisieren und eine Lineare MPR mit Quadratischer Programmierung zu verwenden.

Alternativ nutzt eine nichtlineare Optimierung explizit ein nichtlineares Modell. Nichtlineare Optimierungsverfahren sind im Prinzip iterative Suchverfahren, die ausgehend vom Punkt \mathbf{x}^k den nächsten Punkt \mathbf{x}^{k+1} nach

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha^* \cdot \mathbf{s} \quad (19.34)$$

bestimmen, d. h. der nächste Punkt wird auf der Geraden mit der Richtung s nach einem Schritt der Länge α^* gefunden.

Das Minimum wird auf der Grundlage der notwendigen Bedingung

$$\nabla f(\mathbf{x}) = \mathbf{0} \quad (19.35)$$

bestimmt. Hierbei sollen zunächst nur Aufgaben ohne Nebenbedingungen behandelt werden.

Alle Optimierungsverfahren benötigen einen Startpunkt \mathbf{x}^0 , an dem die Suche beginnen soll. Die Suche selbst erfolgt schrittweise, indem zu Beginn eines Schrittes eine Suchrichtung ermittelt wird und dann in dieser Richtung ein eindimensionales Optimierungsproblem gelöst wird.

Die scheinbar einfache Aufgabe, eine günstige Suchrichtung zu bestimmen, erweist sich bei näherer Betrachtung als nicht trivial, sodass sich viele eingesetzte und angebotene Verfahren vor allem in diesem Punkte unterscheiden. Ebenfalls von allgemeiner Bedeutung ist, ob Ableitungen der Funktion $f(\mathbf{x})$ analytisch bestimmbar sind oder durch Differenzenquotienten angenähert werden müssen. Schließlich benötigen einige Verfahren auch Information über die zweiten Ableitungen der Funktion $f(\mathbf{x})$ in der Form der Hessematrix. Solche Aussagen lassen sich durch entsprechende höhere Differenzenquotienten i. A. nur mit empfindlichem Rechenaufwand beschaffen.

Das einfache Gradientenverfahren, auch Verfahren des steilsten Abstiegs genannt, verwendet als Suchrichtung allein den Gradienten nach

$$\mathbf{s} = -\nabla f(\mathbf{x}^k) \quad . \quad (19.36)$$

Obwohl dieses Vorgehen sinnvoll erscheint, so lehrt doch die Anwendungspraxis, dass diese einfache Richtungswahl häufig zu deutlich schlechteren Ergebnissen führt als alternative Verfahren. Daher ist es auch nur geeignet, die grundsätzliche Verfahrensweise bei der mehrdimensionalen Optimierung zu beleuchten.

Bild 19-5 zeigt an einem einfachen Beispiel, dass der Gradient nicht unbedingt die vorteilhafteste Suchrichtung anzeigt: Wenn als Startpunkt \mathbf{x}_A^0 gewählt wird, so erhält man eine Suchrichtung, die genau auf das Minimum der Funktion zeigt. Da die hier als Beispiel gewählte Funktion selbst eine quadratische ist, führt die Suche entlang der angegebenen Richtung mit Hilfe der Minimierung einer quadratischen Parabel in einem Schritt zum Minimum.

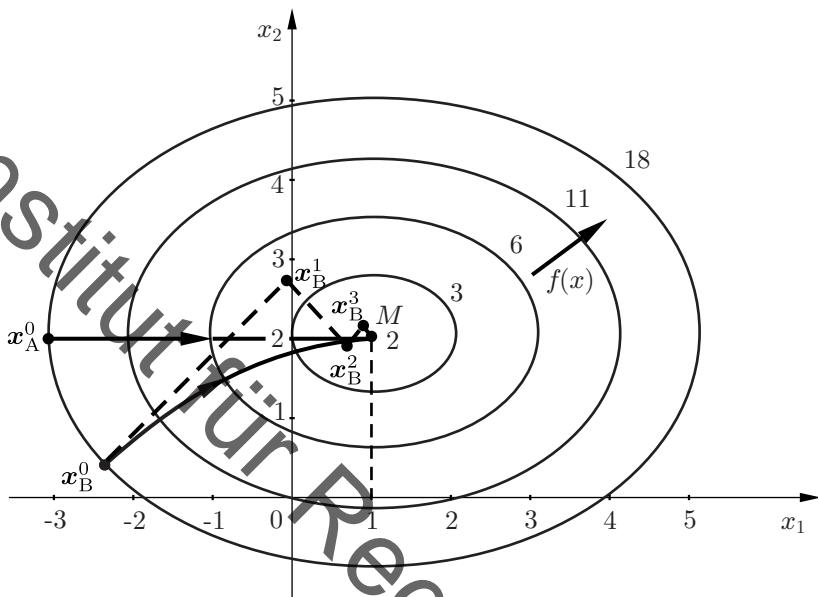


Bild 19-5: Beispiel zum Gradientenverfahren

Ganz anders verläuft die Suche vom Startpunkt x_B^0 aus. Statt auf der leicht gekrümmten, zum Minimum führenden Linie verläuft die Suche in einer Art Zickzackkurs über mehrere Zwischenpunkte x_B^i bis in die Nähe des Minimums.

Diese Schwäche des Gradientenverfahrens hat die Entwicklung von Verfahren gefördert, die mithilfe der zweiten Ableitungen bessere Ergebnisse liefern. Hierzu zählen insbesondere die sogenannten Newton²-Verfahren.

Das Newton-Verfahren zur Optimierung von Funktionen beruht auf dem Gedanken, die zu minimierende Funktion $f(\mathbf{x})$ in der Umgebung des Punktes \mathbf{x}^k durch ein Paraboloid anzunähern mit

$$f(\Delta \mathbf{x}) = \frac{1}{2} \Delta \mathbf{x}^T \mathbf{H}(\mathbf{x}^k) \Delta \mathbf{x} + \mathbf{c}^T(\mathbf{x}^k) \cdot \Delta \mathbf{x} + b \quad (19.37)$$

²Isaac Newton (1643-1727), englischer Physiker [40]

und den Gradienten dieser Näherung zu null zu setzen

$$\nabla f(\Delta \mathbf{x}^*) = \mathbf{H}(\mathbf{x}^k) \Delta \mathbf{x}^* + \mathbf{c}(\mathbf{x}^k) = \mathbf{0}, \quad (19.38)$$

um eine Schätzung \mathbf{x}^* aus

$$\Delta \mathbf{x}^* = -\mathbf{H}(\mathbf{x}^k)^{-1} \cdot \mathbf{c}(\mathbf{x}^k) \quad (19.39)$$

für das gesuchte Minimum zu erhalten.

Wäre die Hessematrix im Spezialfall gleich der Einheitsmatrix, ginge in obiger Gleichung wegen $\mathbf{c}(\mathbf{x}^k) = \nabla f(\mathbf{x}^k)$ das Newton-Verfahren in das einfache Gradientenverfahren über.

Da die Hessesche Matrix \mathbf{H} meist nicht oder nur als Näherung verfügbar ist und das Bilden der Inversen auch problematisch sein kann, wird das Newton-Verfahren in seiner reinen Form selten eingesetzt. Von großer praktischer Bedeutung sind allerdings sogenannte Quasi-Newton-Verfahren, die durch Auswerten von Ergebnissen, die bei der Minimumsuche zwangsläufig entstehen, Informationen über die Hessematrix bzw. ihre Inverse beschaffen.

Eine Sonderstellung nehmen die Optimierungsaufgaben zur Minimierung einer Summe von quadratischen Termen ein, das sind solche Aufgaben, bei denen die Gütfunktion eine spezielle Struktur besitzt, nämlich

$$f(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^m f_i(\mathbf{x})^2. \quad (19.40)$$

Fasst man die $f_i(\mathbf{x})$ zu einem Vektor $\mathbf{f}(\mathbf{x})$ zusammen und bestimmt die zugehörige Jacobi-Matrix $\mathbf{J}(\mathbf{x})$, gegebenenfalls durch Differenzenquotienten, gewinnt man sowohl den Gradienten $\nabla \mathbf{f}(\mathbf{x})$ als auch die Hessematrix $\mathbf{H}(\mathbf{x})$ durch relativ einfache Vektor-Matrix-Operationen.

Damit sind in diesem Fall die Voraussetzungen zum Einsatz des Newton-Verfahren gegeben. Die Kombination von Minimierung einer Summe von Quadraten und des Newton-Verfahrens wird auch als Gauss-Newton-Optimierung bezeichnet. Wenn anwendbar, ist dieses Verfahren üblicherweise effektiver als die sonst einzusetzenden Quasi-Newton-Verfahren.

Optimierungsaufgaben mit Beschränkungen in der Form von Gleichungs- und Ungleichungsnebenbedingungen können durch ähnliche Ideen wie die

der Active-Set-Strategien auf eine Folge von Optimierungsaufgaben ohne Beschränkungen überführt werden.

Wie auch bei der Quadratischen Programmierung kann die verallgemeinerte Lagrange-Funktion

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^T \mathbf{g}(\boldsymbol{x}) + \boldsymbol{\mu}^T \mathbf{h}(\boldsymbol{x}) \quad (19.41)$$

zur nichtlinearen Optimierung mit Nebenbedingungen verwendet werden. Die Nebenbedingungen $\mathbf{g}(\boldsymbol{x})$ und $\mathbf{h}(\boldsymbol{x})$ können nun auch nichtlinear in \boldsymbol{x} sein.

Notwendige Bedingungen für ein Minimum von $f(\boldsymbol{x})$ unter Beachtung der Nebenbedingungen sind die sogenannte Karush³-Kuhn-Tucker-Bedingungen:

$$\begin{aligned} \mathcal{L}_x(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= f_x(\boldsymbol{x}) + \mathbf{g}_x(\boldsymbol{x})^T \boldsymbol{\lambda} + \mathbf{h}_x(\boldsymbol{x})^T \boldsymbol{\mu} = \mathbf{0} \\ \mathcal{L}_{\lambda}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \mathbf{g}(\boldsymbol{x}) = \mathbf{0} \\ \mathbf{h}(\boldsymbol{x}) &\leq \mathbf{0} \\ \mathbf{h}(\boldsymbol{x})^T \boldsymbol{\mu} &= 0 \\ \boldsymbol{\mu} &\geq \mathbf{0}. \end{aligned} \quad (19.42)$$

Die drei letzten Zeilen der Gleichung behandeln die Ungleichungsnebenbedingungen. Sie sind so zu verstehen, dass diese alle erfüllt sein müssen (dritte Zeile), wobei auch hier zwischen inaktiven ($h_i < 0$) und aktiven ($h_i = 0$) unterschieden wird. In der vierten Zeile sollen durch entsprechende Auslegung des Vektors $\boldsymbol{\mu}$ nur die aktiven Ungleichungsnebenbedingungen mit Faktoren $\mu_i > 0$ multipliziert werden, die anderen μ_i sind zu null zu setzen, weil sie zu inaktiven Bedingungen gehören.

Die Nebenbedingungen schränken sowohl den Suchraum für die Lösung des Optimierungsproblems ein als auch die Richtung, in der – ausgehend von einem Start- oder Zwischenpunkt – nach einer Lösung zu suchen ist. Bei der Umsetzung in Optimierungsalgorithmen geht man üblicherweise so vor, dass wiederum von einem Start- oder Zwischenpunkt ausgehend eine Verbesserung der Lösung angestrebt wird. Dabei kann man in der Umgebung dieses Punktes und damit für die Ermittlung der „zulässigen Richtung“ die inaktiven Ungleichungsnebenbedingungen zunächst außer Acht lassen

³William Karush (1917-1997), amerikanischer Mathematiker [24]

und die aktiven als Gleichungsnebenbedingungen formulieren und mit den anderen Gleichungsnebenbedingungen zusammenfassen.

Die dann zu minimierende Lagrange-Funktion

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x}) \quad (19.43)$$

ist deutlich einfacher als die zunächst eingeführte. Ihr Minimum ist erreicht, wenn

$$\begin{aligned}\mathcal{L}_x(\mathbf{x}, \boldsymbol{\lambda}) &= f_x(\mathbf{x}) + \mathbf{g}_x(\mathbf{x})^T \boldsymbol{\lambda} = \mathbf{0} \\ \mathcal{L}_{\lambda}(\mathbf{x}, \boldsymbol{\lambda}) &= \mathbf{g}(\mathbf{x}) = \mathbf{0}.\end{aligned} \quad (19.44)$$

Zur schrittweisen Lösung des Optimierungsproblems wird die Lagrange-Funktion durch eine quadratische Funktion angenähert und deren Minimum gesucht. Dieses Minimum wird, sofern es keine der Nebenbedingungen verletzt, als Ausgangspunkt des nächsten Optimierungsschrittes benutzt.

Das aus diesem Prinzip resultierende Verfahren wird als Sequentielle Quadratische Programmierung (SQP) bezeichnet und gilt als effektives Verfahren zur Lösung nichtlinearer Optimierungsaufgaben mit Nebenbedingungen.

Neben den skizzierten Herausforderungen der nichtlinearen Optimierung ergeben sich bei der Nichtlinearen MPR zusätzliche Probleme bei rege lungstechnischen Fragestellungen. So erschweren einige Ansätze die Stabilitätsbetrachtung erheblich. Zudem müssen üblicherweise viele Tests und Simulationen durchgeführt werden, um das teilweise schwer vorhersagbare Verhalten der Regelung in einem großen Arbeitsbereich untersuchen zu können.

20 Iterativ Lernende Regelung

20.1 Allgemeines

Viele industrielle Anlagen wiederholen eine identische Aufgabe stets aufs Neue. Prominente Beispiele sind Produktionsprozesse wie Fräsen oder Gießen aber auch robotische Montagearbeiten an Fließbändern.

Eine konventionelle Regelung solcher zyklisch ablaufender Prozesse besteht aus einem Regler für den sicheren Betrieb des Prozesses, kombiniert mit einer Vorsteuerung zur Verbesserung des Führungsverhaltens. Diese Regelungskonzepte verwenden jedoch nur die aktuellen Informationen über die Führunggröße und die Regelabweichung, nicht jedoch die Informationen über die Leistungsfähigkeit der Regelung in den vorangegangenen Zyklen.

Das Konzept der Iterativ Lernenden Regelung (ILR) wurde in den 1980er Jahren entwickelt und stellt einen systemtheoretischen Rahmen bereit, dem zyklischen Prozesscharakter explizit Rechnung zu tragen und ein Regelung zu entwickeln, welche aus den Regelfehlern der vergangenen Zyklen Rückschlüsse auf notwendige Eingriffe im aktuellen Zyklus zieht.

Iterativ Lernende Regelung (ILR)

Eine ILR nutzt die Zeit zwischen zwei ablaufenden Zyklen dazu, aus den abgespeicherten Verläufen der Stellgröße und der Regelabweichung eine Modifikation der Vorsteuerung im nächsten Zyklus zu berechnen.

Obwohl damit die ILR innerhalb eines Zyklus als Steuerung agiert, da die vorgesteuerte Stellgröße für den gesamten Zyklus im Vorhinein berechnet wird, so entsteht durch das Einbeziehen vergangener Zyklen dennoch ein geschlossener Wirkungsablauf. Somit ist die Bezeichnung als Iterativ Lernende *Regelung* gerechtfertigt. Der Wirkungsplan der ILR ist dabei in Bild 20-1 gezeigt.

Die Solltrajektorie $w(t)$ des i -ten Zyklus wird im Speicher H abgespeichert, welcher dann die Trajektorie des i -ten Zyklus w_i bereitstellt. Der Index i zeigt dabei an, dass es sich nicht mehr um Zeitverläufe, sondern einen zyklusbezogenen gespeicherten Verlauf handelt. Analog wird der Verlauf der Ausgangsgröße aufgezeichnet und somit die Regeldifferenz e_i des i -ten Zyklus gebildet. Diesen verwendet die ILR zusammen mit der ebenfalls

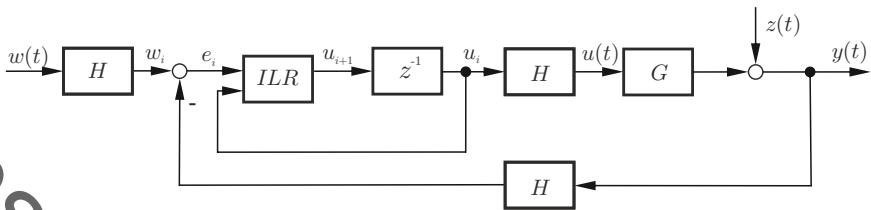


Bild 20-1: Wirkungsplan einer Iterativ Lernenden Regelung

gespeicherten Stellgröße u_i dazu, den neuen Stellgrößenverlauf u_{i+1} zu ermitteln und in einen Speicher zu schreiben. Diese Berechnung wird in der Zeit zwischen zwei Zyklen vorgenommen. Der Indexwechsel $i \rightarrow i + 1$ (als Multiplikation mit z^{-1}) zeigt den Beginn des folgenden Zyklus an. Hier wird die vorberechnete Trajektorie u_{i+1} abgerufen und der Prozess G damit beaufschlagt. Bei G kann es sich um die Regelstrecke oder aber beim Einsatz einer unterlagerten klassischen innerzyklischen Regelung um einen geschlossenen Regelkreis handeln. In diesem Fall ist die Stellgröße u der ILR im Sinne einer Kaskadenregelung zu verstehen. Eine ILR kann prinzipiell auch über mehrere vergangene Zyklen lernen [47].

20.2 Systemtheoretische Betrachtung

Um die skizzierte Idee einer ILR mathematisch zu fassen, benötigt man einen formalen Rahmen für eine kompakte Beschreibung der Aufgabenstellung. Hierfür müssen folgende Annahmen bezüglich des zu regelnden Prozesses G getroffen werden.

- (1) Der Prozess ist zyklisch mit einer festen Zykluszeit T_c .
- (2) Die Solltrajektorie w ist im Vorhinein für $[0, T_c]$ definiert.
- (3) Der Anfangszustand $x_i(0)$ ist für jeden Zyklus i identisch.
- (4) Die Systemdynamik ändert sich über die Zyklen nicht.
- (5) Für das gegebene w existiert ein u , welches das System zur gewünschten Trajektorie führt.
- (6) G ist stabil.

Für fast alle dieser Annahmen gibt es Konzepte, um sie abzuschwächen oder zu verallgemeinern. Einige dieser Konzepte sollen kurz skizziert werden. So fordert (6) nicht die Stabilität der Regelstrecke, sondern nur des Prozesses G , der entsprechend unterlagert stabilisiert werden kann. Für nicht gemäß (5) realisierbare w wird der Regelfehler e_i zwar nie verschwinden, er kann aber dennoch sinnvoll minimiert werden. Eine in der Literatur oft synonym zum Begriff der ILR verwendete Bezeichnung spricht von repetitiven Regelungen. Hierunter ist in Abgrenzung zur ILR zu verstehen, dass dort der Anfangszustand des Folgezustandes der Endzustand des vorherigen Zyklus ist: $\mathbf{x}_i(0) = \mathbf{x}_{i-1}(T_c)$.

Unter den Annahmen (1) bis (6) kann die ILR wie folgt formuliert werden:

Lerngesetz der ILR

Gegeben ist die (potentiell nichtlineare und zeitvariante) Strecke in Zustandsraumdarstellung

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}, u, t) \quad , \quad y = g(\mathbf{x}, u, t) \quad , \quad f(\cdot, \cdot, t + T_c) = f(\cdot, \cdot, t) \quad (20.1)$$

und eine vorgegebene Solltrajektorie

$$w_i = \begin{cases} w_c(t) & \text{für } t \in [t_i, t_i + T_c] \\ w_{int}(t) & \text{für } t \in [t_i + T_c, t_{i+1}] \\ 0 & \text{für } t < 0 \end{cases} \quad (20.2)$$

Dann ist die Iterativ Lernende Regelung ein Lerngesetz Γ mit

$$u_{i+1}(t) = \Gamma(u_i(t), y_i(t), w_i(t)) . \quad (20.3)$$

Bild 20-2 veranschaulicht die Solltrajektorie. Die Unterteilung in w_c und w_{int} stellt dabei sicher, dass Regelfehler nur für die Zeit während des Zyklus und nicht zwischen den Zyklen betrachtet werden.

Mathematisch gesehen stellt das Lerngesetz Γ eine Fixpunktiteration zur Berechnung der optimalen Stelltrajektorie u^* dar. Das liegt daran, dass aufgrund des für jeden Zyklus gleichen Anfangszustandes und der identischen Systemdynamik in jedem Zyklus der Verlauf von \mathbf{x}_i nur abhängig von u_i

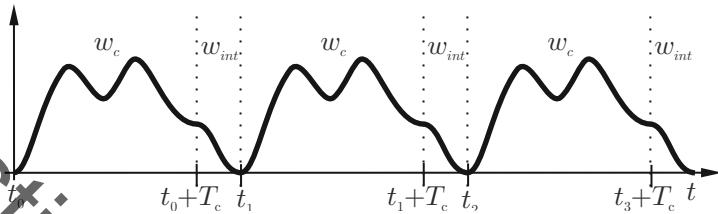


Bild 20-2: Sollverlauf einer iterativ lernenden Regelung

ist. Daher kann man

$$y_i = g(\mathbf{x}_i, u_i, t) = h(u_i) \quad (20.4)$$

für eine Funktion h schreiben. Die Forderung $y_i = w_i$ soll nun durch ein $u_i = u^*$ erfüllt werden, d.h. $h(u^*) = w$. Die Bestimmung von u^* erfolgt dann im laufenden Betrieb als Grenzwert des Lerngesetzes $u_i \rightarrow u^*$. Auf diesem mathematischen Hintergrund wird im Zusammenhang mit der ILR in der Literatur oft nicht von Stabilität, sondern von Konvergenz gesprochen, da wegen der zeitkontinuierlichen Sichtweise $u(t)$ Folgen von Funktionen betrachtet werden. Man bezeichnet das Lerngesetz Γ als konvergent, sofern

$$\|u_k(t) - u_\infty(t)\| \rightarrow 0 \quad (20.5)$$

für alle $u_0(t)$ mit einer noch zu definierenden Norm $\|\cdot\|$ gilt. Hierbei muss nicht zwingend $u_\infty(t) = u^*(t)$ gelten – gegebenenfalls stellt sich eine bleibende Regelabweichung ein. Die Art der verwendeten Norm ist für den Nachweis der Konvergenz entscheidend, da $u(t)$ Element eines Funktionenraums ist und Normen dort nicht äquivalent sind. Aus regelungstechnischer Sicht wird monotone Konvergenz

$$\|u_{k+1}(t) - u_\infty(t)\| \leq \|u_k(t) - u_\infty(t)\| \quad (20.6)$$

gefordert, welche eine Verbesserung mit jedem weiteren Lernzyklus impliziert. Die Norm wird dann entsprechend so gewählt, dass sich für das betrachtete System monotone Konvergenz nachweisen lässt. Die Konvergenz

des Lerngesetzes impliziert, sofern sie auch trotz auftretender Störungen erhalten bleibt, die Stabilität der ILR. Hiermit ist die Stabilität *über die Zyklen* gemeint, d. h. dass die vorgesteuerte Trajektorie $u_k(t)$ für $k \rightarrow \infty$ auch bei Störungen beschränkt bleibt. Innerhalb eines Zyklus beeinflusst das Lerngesetz die Stabilität nicht, da die ILR dort als Steuerung agiert.

Da das Hantieren mit Normen und der Konvergenznachweis für den praktischen Entwurf und die Systemanalyse oft zu langwierig sind, ist ein anderer Zugang zur systemtheoretischen Analyse von Stabilität notwendig. Diesen bietet die sogenannte Lifted-System Darstellung, die sich durch eine zeitdiskrete Betrachtung von der Problematik von Folgen von Funktionen löst. Hierbei wird angenommen, dass im Speicher H ohnehin nur zeitdiskret abgetastete Werte hinterlegt werden können.

Lifted-System Darstellung

Die gespeicherten Werte $y(k)$ von einer skalaren Größe $\tilde{y}(t)$ eines Zyklus können als Vektor zusammengefasst werden, dessen Dimension bei einer äquidistanten Abtastung mit $\Delta t = \frac{T_c}{N}$ genau $N + 1$ entspricht:

$$\begin{aligned} \mathbf{y} &= [\tilde{y}(t=0) \quad \tilde{y}(t=\Delta t) \quad \dots \quad \tilde{y}(t=N\Delta t)]^T \\ &= [y(0) \quad \dots \quad y(N)]^T \in \mathbb{R}^{N+1} \end{aligned} \quad (20.7)$$

Die Abbildung dieser Werte vom vergangenen Zyklus auf den nächsten Zyklus stellt eine zeitdiskrete Abbildung vom \mathbb{R}^{N+1} in den \mathbb{R}^{N+1} dar, die sich mit den Methoden zeitdiskreter Systeme untersuchen lässt.

Dieses Vorgehen lässt dabei nur dann Schlüsse auf die Dynamik des zyklischen Systems zu, wenn bei der Abtastung die einschlägigen Kriterien wie das Shannon-Theorem eingehalten werden. Die abgetasteten Folgen repräsentieren dann stückweise konstante Zeitfunktionen, welche durch ein Halteglied 0. Ordnung erzeugt werden. Formuliert man den Prozess G in der Lifted-System Darstellung, so kann man die zeitdiskrete Faltung in der Form

$$y(k) = g(k) * u(k) = \sum_{j=0}^{\infty} g(k-j) u(j) \quad (20.8)$$

nutzen, um für lineare Prozesse G die algebraische Beschreibung

$$\underbrace{\begin{bmatrix} y(0) \\ y(1) \\ y(2) \\ y(3) \\ \vdots \\ y(N) \end{bmatrix}}_y = \underbrace{\begin{bmatrix} g(0) & 0 & \dots & \dots & \dots & 0 \\ g(1) & g(0) & 0 & & & \vdots \\ g(2) & g(1) & g(0) & 0 & & \vdots \\ g(3) & g(2) & g(1) & g(0) & \ddots & \vdots \\ \vdots & & & & \ddots & 0 \\ g(N) & g(N-1) & g(N-2) & \dots & & g(0) \end{bmatrix}}_{=G} \underbrace{\begin{bmatrix} u(0) \\ u(1) \\ u(2) \\ u(3) \\ \vdots \\ u(N) \end{bmatrix}}_u \quad (20.9)$$

nach Markov¹ zu gewinnen.

Markov-Matrix

Die Matrix \mathbf{G} in Gl.(20.9) heißt *Markov-Matrix* und die Einträge $g_k = g(k)$ auch *Markov-Parameter*.

Der Markov-Parameter g_k ist der k -te Wert der Gewichtsfolge (zeitdiskrete Impulsantwort). Aufgrund des Kausalitätsprinzips ist \mathbf{G} stets eine untere Dreiecksmatrix, da der Ausgang $y(k)$ zum Zeitpunkt k nur von den Eingängen $u(j)$ zu Zeitpunkten $j \leq k$ abhängen kann. Die Markov-Matrix enthält die vollständigen dynamischen Informationen des Systems.

In analoger Weise bestimmt man eine Beschreibungsform der ILR. Setzt man hier anstelle von Gl.(20.3) mit $e_i(t) = w_i(t) - y_i(t)$ das allgemeine lineare, zeitvariante Lerngesetz

$$u_{i+1}(t) = S(t)u_i(t) + \Gamma(t) \left(e_i(t) + \dot{e}_i(t) + \int e_i(t) dt + \dots \right) \quad (20.10)$$

mit Integrationen und Differentiationen beliebiger endlicher Ordnung an, so erhält man die allgemeine Form

$$u_{i+1} = \mathbf{S}u_i + \boldsymbol{\Gamma}e_i \quad (20.11)$$

¹Андрей Андреевич Марков, russischer Mathematiker [33]

mit der Diagonalmatrix \mathbf{S} mit $S_{kk} = S(t = k\Delta t)$ und einer (potentiell) vollbesetzten Matrix $\mathbf{\Gamma}$.

Im Gegensatz zu \mathbf{S} ist $\mathbf{\Gamma}$ nicht diagonal, da auch zeitliche Ableitungen des Fehlers berücksichtigt werden, die in der Approximation als Differenzenquotient zu Einträgen außerhalb der Diagonalen führen.

Im Gegensatz zu \mathbf{G} hat $\mathbf{\Gamma}$ nicht zwingend eine untere Dreiecksstruktur, da durch Abspeicherung der Signale des letzten Zyklus eine akausale Signalverarbeitung zwischen den Zyklen ermöglicht wird.

Der geschlossene Regelkreis kann mit der Lifted-System Darstellung in einen Wirkungsplan überführt werden, der in Bild 20-3 gezeigt ist. Hierbei können bei Bedarf noch zusätzliche Elemente \mathbf{F} zur Filterung von möglicherweise auftretendem Messrauschen oder \mathbf{V} zur Glättung der Referenztrajektorie eingeführt werden.

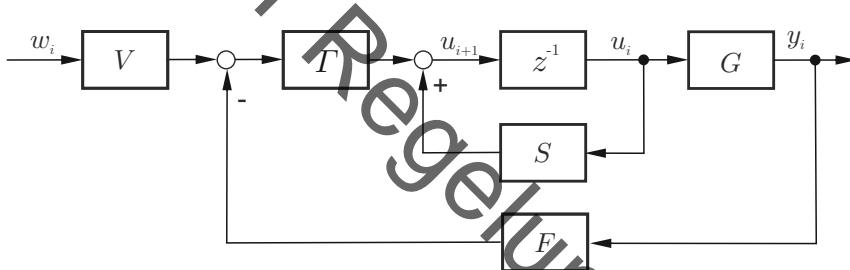


Bild 20-3: Wirkungsplan der ILR in der Lifted-System Darstellung

Ein Vergleich mit der klassischen Zustandsraumdarstellung mit Zustandsrückführung zeigt, dass die Darstellung vollständig analog ist, wenn die Zuordnung

$$\mathbf{A} = \mathbf{S} \quad , \quad \mathbf{B} = \mathbf{\Gamma} \quad , \quad \mathbf{C} = \mathbf{G} \quad , \quad \mathbf{D} = \mathbf{0} \quad , \quad \mathbf{K} = \mathbf{F} \quad (20.12)$$

und \mathbf{V} als Führungsgrößenfilter vorgenommen wird. Damit lässt sich auch sofort das charakteristische Polynom des geschlossenen Regelkreises als

$$\det(\lambda\mathbf{I} - \mathbf{S} + \mathbf{\Gamma}\mathbf{F}\mathbf{G}) \quad (20.13)$$

angeben.

Sofern keine Pol-Nullstellen-Kürzungen auftreten, lässt sich hieraus direkt die Stabilität der ILR bestimmen, da für alle λ aufgrund der zeitdiskreten Dynamik von Zyklus zu Zyklus für Stabilität $|\lambda| < 1$ gelten muss. Dieses Kriterium ist mit den bekannten Werkzeugen leicht zu überprüfen und stellt den üblichen Zugang zur Stabilitätsuntersuchung und Auslegung für Iterativ Lernende Regelungen dar. Zu beachten ist auch hier, das die Stabilität der Abbildung $\mathbf{u}_i \rightarrow \mathbf{u}_{i+1}$ untersucht wird, was nur im Falle eines stabilen Prozesses G die Stabilität des Gesamtregelkreises zur Folge hat.

Zur Überprüfung der Voraussetzung dieser Bedingung muss nachgerechnet werden, ob möglicherweise Nullstellen die Polstellen kompensieren. Die Nullstellen des Führungsverhaltens können dabei gemäß [47] über die Formel

$$\det(\mathbf{G}) \cdot \det(\mathbf{V}) = 0 \quad (20.14)$$

bestimmt werden. Da \mathbf{V} frei wählbar ist, kann $\det(\mathbf{V}) \neq 0$ beim Entwurf sichergestellt werden. Insbesondere ist die Bedingung beim Verzicht auf ein Führungsgrößenfilter $\mathbf{V} = \mathbf{I}$ erfüllt. Aufgrund der unteren Dreiecksstruktur kann die Determinante von \mathbf{G} sofort zu

$$\det(\mathbf{G}) = (g(0))^{N+1} \quad (20.15)$$

berechnet werden. Folglich gilt, dass das Führungsverhalten für zeitdiskrete Systeme G mit Durchgriff (d. h. $g(0) \neq 0$) keine, ansonsten $N+1$ -Nullstellen bei $z = 0$ aufweist. Diese Fälle sind für die Stabilität unproblematisch.

Neben der Stabilität lassen sich in diesem Rahmen Steuer- und Beobachtbarkeit untersuchen. Die entsprechenden Kriterien nach Kálmán erbringen dabei die Bedingungen, dass die Matrizen

$$\mathbf{Q}_S = [\mathbf{\Gamma} \quad \mathbf{S}\mathbf{\Gamma} \quad \dots], \quad \mathbf{Q}_B = \begin{bmatrix} \mathbf{G} \\ \mathbf{GS} \\ \vdots \end{bmatrix} \quad (20.16)$$

vollen Rang besitzen. Die Steuerbarkeit lässt sich durch den frei wählbaren Lernoperator $\mathbf{\Gamma}$ sicherstellen, während Beobachtbarkeit genau nur für Systeme mit Durchgriff gegeben ist. Da die Abbildung $\mathbf{u}_i \rightarrow \mathbf{u}_{i+1}$ analysiert wird, bedeutet das, dass bei geeigneten Lernoperatoren $\mathbf{\Gamma}$ der Verlauf

von \mathbf{u}_i vorgegeben werden kann, was aus der Anschauung leicht verifiziert werden kann. Die Beobachtbarkeit wiederum bedeutet, dass sich aus der Messung des Ausgangs \mathbf{y}_i der Eingang \mathbf{u}_i bestimmen lässt.

Die Beobachtung von \mathbf{u}_i ist technisch nicht notwendig, da \mathbf{u}_i ohnehin bekannt ist. Dennoch hat fehlende Beobachtbarkeit eine Relevanz, da sie impliziert, dass Anteile des Stellvektors \mathbf{u}_i sich nicht am Ausgang \mathbf{y}_i bemerkbar machen. Dies kann für Systeme ohne Durchgriff leicht verifiziert werden, da der letzte Stelleingriff $u(N)$ keine Auswirkungen auf \mathbf{y}_i haben kann.

20.3 Entwurf des Lernoperators

Die Lifted-System Darstellung bietet nun einen passenden Rahmen, um die beiden Lernoperatoren \mathbf{S} und $\boldsymbol{\Gamma}$ zu entwerfen. Der Operator \mathbf{S} definiert dabei gemäß Gl.(20.11), welcher Anteil der Stelltrajektorie des letzten Zyklus in den nächsten Zyklus a priori (d. h. ohne Betrachtung der entstandenen Fehler \mathbf{e}_i) übernommen werden soll. Folglich stellt ein Wert von $S_{kk} = 1$ ein vollständiges Beibehalten dar, während Werte kleiner eins als Vergessensfaktor interpretierbar sind. Werte außerhalb des Intervalls $[0, 1]$ sind aus der Anschauung nicht für ein effizientes Lernen geeignet.

Eine geschickte Wahl von \mathbf{S} kann stationäre Genauigkeit im Regelkreis sicherstellen. Unter der Annahme eines stabilen Betriebs der ILR ist bekannt, dass der geschlossene Regelkreis dann stationär genau arbeitet, wenn der aufgeschnittene Regelkreis einen integrierenden Anteil besitzt. Die Analogie der Zustandsraumdarstellung zeigt, dass die Rolle von \mathbf{A} als aufgeschnittenem Regelkreis durch \mathbf{S} ersetzt wird. Daher ist es sinnvoll, möglichst viele Eigenwerte von \mathbf{S} zu eins zu setzen. Da \mathbf{S} eine Diagonalmatrix ist, ist eine naheliegende Wahl $\mathbf{S} = \mathbf{I}$ mit $N + 1$ Eigenwerten bei eins.

Die Betrachtung des Lernoperators $\boldsymbol{\Gamma}$ ist komplexer, da dieser potentiell voll besetzt ist, was bis zu $(N + 1)^2$ Einstellparameter erlaubt. Zentrale Bedingung hierbei ist die Stabilität der ILR gemäß Gl.(20.13). Unter der Vereinfachung von $\mathbf{F} = \mathbf{I}$ müssen also die Eigenwerte der Matrix $\boldsymbol{\Gamma}\mathbf{G} - \mathbf{S}$ passend gewählt werden. Um eine aufwändige Berechnung der Eigenwerte zu umgehen, bietet es sich an dieser Stelle an, die Diagonalstruktur von \mathbf{S} sowie die untere Dreiecksstruktur von \mathbf{G} so zu nutzen, dass diese Eigenwerte direkt abgelesen werden können. Hierzu sei $r = n - m$ der relative Grad von G . Dieser entspricht für zeitdiskrete Systeme dem Zeitindex r , für den der

Eingang $u_i(0)$ erstmalig auf den Ausgang $y_i(k)$ wirkt und damit der Totzeit. Für Systeme mit Durchgriff gilt $r = 0$. Die Markov-Matrix ist dann eine untere Dreiecksmatrix, die auf der r -ten Nebendiagonalen beginnt. Man definiert nun

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma(0) & \gamma(-1) & \dots & \gamma(-r) & 0 & \dots & 0 \\ \gamma(1) & \gamma(0) & \gamma(-1) & & \gamma(-r) & 0 & \vdots \\ \gamma(2) & \gamma(1) & \gamma(0) & \gamma(-1) & & & \vdots \\ \gamma(3) & \gamma(2) & \gamma(1) & \gamma(0) & \ddots & & \vdots \\ \vdots & & & & \ddots & & \\ \gamma(N) & \gamma(N-1) & \gamma(N-2) & \dots & & & \gamma(0) \end{bmatrix} \quad (20.17)$$

als eine untere Dreiecksmatrix, die bei der $-r$ -ten Nebendiagonalen beginnt. Einfaches Ausmultiplizieren zeigt, dass $\boldsymbol{\Gamma}\mathbf{G}$ ebenfalls eine untere Dreiecksmatrix mit der Diagonalen

$$(\boldsymbol{\Gamma}\mathbf{G})_{ii} = \gamma(-r)g(r) \quad \text{für } i < N + 1 - r \quad \text{und} \quad (\boldsymbol{\Gamma}\mathbf{G})_{ii} = 0 \quad \text{sonst} \quad (20.18)$$

ist. Folglich ergeben sich die Eigenwerte des geschlossenen Regelkreises bei einer Wahl von \mathbf{S} mit

$$S_{ii} = \tilde{s} \quad \text{für } i < N + 1 - r \quad \text{und} \quad S_{ii} = 0 \quad \text{sonst} \quad (20.19)$$

als einen $N + 1 - r$ -fachen Pol bei $z = \tilde{s} - \gamma(-r)g(r)$ und einem r -fachen Pol bei $z = 0$. Der r -fache Pol bei Null entspricht genau den Einträgen im Stellvektor \mathbf{u} die gemäß der Beobachtbarkeitsanalyse keinen Einfluss auf den Ausgangsvektor \mathbf{y} haben. Hier ist es notwendig, von der Vorgabe $S = 1$ abzuweichen, da ohne Fehlerrückführung durch $\boldsymbol{\Gamma}$ gelernt wird und ein Wert von $S = 1$ die letzten r Pole an den Stabilitätsrand verschiebt.

Wahl von $\gamma(-r)$

Die Stabilität der ILR hängt nur von dem Lernparameter $\gamma(-r)$ und dem ersten von Null verschiedenen Markov-Parameter $g(r)$ ab. Üblicherweise wird $\gamma(-r)$ über eine Polvorgabe festgelegt und Stabilitätsgrenzen für $\gamma(-r)$ können direkt angegeben werden.

Die Tatsache, dass bei der Struktur der Lernmatrix in Gl.(20.17) allein $\gamma(-r)$ die Stabilitätseigenschaften festlegt, ermöglicht es, die Lernoperatoren stark zu vereinfachen. Für den Fall $r = 0$ ist es möglich, $\Gamma = \gamma_P \mathbf{I}$ als Vielfaches der Einheitsmatrix zu wählen. Diese Variante der ILR wird auch als P-ILR bezeichnet, da sie zurückübersetzt in die zeitkontinuierlichen Be trachtung genau dem Lerngesetz

$$u_{i+1}(t) = Su_i(t) + \gamma_P(e_i(t)) \quad (20.20)$$

und damit einem P-Regler entspricht. Analog entwickelt man Bezeichnungen wie D-ILR und PD-ILR, wobei letztere in der allgemeinen Struktur

$$\Gamma = \begin{bmatrix} \gamma_P & \gamma_D & 0 & \dots & 0 \\ 0 & \gamma_P & \gamma_D & & \vdots \\ \vdots & & & & 0 \\ \vdots & & & & \gamma_D \\ 0 & \dots & \dots & 0 & \gamma_P \end{bmatrix} \quad (20.21)$$

geschrieben wird und für einen relativen Grad von $r = 1$ geeignet ist. Diese vereinfachten Lernoperatoren erfüllen auch die formulierten Steuerbarkeitsanforderungen – höchstens die letzten r Stellgrößen fallen hier erneut weg.

Eine offene Frage bleibt, ob neben der Vereinfachung des Lernoperators auf bis zu eine einzige Nebendiagonale es auch möglich ist, die Freiheitsgrade der weiteren $\gamma(i)$ gewinnbringend zu nutzen. Zur Analyse ist es hier zielführend, sich erneut der zeitdiskreten Faltung zu bedienen. Die Gleichung $u_{i+1} = u_i + \Gamma(w_i - y_i)$ lässt sich dann als

$$u_{i+1}(k) = u_i(k) + \sum_{j=0}^{k+r} \gamma(k-j) (w_i(j) - y_i(j)) \quad (20.22)$$

schreiben. Setzt man Gl.(20.22) in Gl.(20.8) ein, so erhält man

$$y_{i+1}(k) = \sum_{j=0}^k g(k-j) \left(u_i(j) + \sum_{l=0}^{j+r} \gamma(j-l) (w_i(l) - y_i(l)) \right) . \quad (20.23)$$

Folglich sind die verschiedenen $y(k)$ beim iterativen Lernen miteinander gekoppelt, d. h. ein schlechtes Lernverhalten des k -ten Zeitschrittes wirkt

sich möglicherweise auch negativ auf die anderen Zeitschritte aus. Dies kann man verhindern, indem man Gl.(20.23) entfaltet und fordert, dass $y_{i+1}(k) = f(y_i(l))$ nur für $k = l$ gilt. Hierdurch wird das Lernverhalten der einzelnen Zeitschritte voneinander entkoppelt. Dies führt auf ein lineares Gleichungssystem in den Koeffizienten $\gamma(j-l)$, dessen Lösung sich komponentenweise und rekursiv als

$$\gamma^{(i)} = -\frac{1}{g(r)} \sum_{j=-r}^{i-1} \gamma(j) g(i+r-j) \quad (20.24)$$

angeben lässt. Hier ist r der relative Grad und $\gamma(-r)$ wird beispielsweise über eine Polvorgabe festgelegt.

Entfaltungs-ILR

Das Entwurfsverfahren über die Entfaltung nach Gl.(20.23) wird auch als *Entfaltungs-ILR* bezeichnet.

20.4 Normoptimale ILR

Eine alternative Berechnung und Auslegung der ILR bietet die sogenannte Normoptimale ILR. Diese bestimmt die Stelltrajektorie über die Optimierung einer Kostenfunktion, analog zur Modellbasierten Prädiktiven Regelung in Abschnitt 18. Ein klassischer Kandidat für die Kostenfunktion ist

$$\|\mathbf{w}_{i+1} - \mathbf{y}_{i+1}\|_{\mathbf{Q}}^2 = \|\mathbf{w}_{i+1} - \mathbf{G}\mathbf{u}_{i+1}\|_{\mathbf{Q}}^2 = \|\mathbf{w}_{i+1} - \mathbf{y}_i - \mathbf{G}\Delta\mathbf{u}_{i+1}\|_{\mathbf{Q}}^2 \quad (20.25)$$

mit $\mathbf{u}_{i+1} = \mathbf{u}_i + \Delta\mathbf{u}_{i+1}$. Das Systemwissen wird durch die Markov-Matrix \mathbf{G} eingebracht.

Die Praxis zeigt hierbei, dass die volle Applizierung von $\Delta\mathbf{u}_{i+1}$ aufgrund von Abweichungen von Modell und Regelstrecke zu schwingendem oder instabilem Verhalten führen kann. Daher verwendet man das Stellgesetz

$$\mathbf{u}_{i+1} = \mathbf{u}_i + \eta \Delta\mathbf{u}_{i+1}. \quad (20.26)$$

Durch den Verstärkungsfaktor $\eta \in [0, 1]$ kann das Konvergenzverhalten eingestellt werden, wobei $\eta = 1$ vollständigem Lernen entspricht, während eine Verringerung von η die Robustheitseigenschaften der ILR verbessert. Entscheidender Vorteil der Normoptimalen ILR ist analog zur MPR die Möglichkeit, Beschränkungen beim Lernen explizit zu berücksichtigen.

Literaturverzeichnis

- [1] J. Ackermann. *Robuste Regelung*. Springer Berlin, Heidelberg, 1 edition, 2014.
- [2] J. Adamy. *Nichtlineare Systeme und Regelungen*. Springer Vieweg, Berlin, 3 edition, 2018.
- [3] H. W. Bode. *Network analysis and feedback amplifier design*. Van Nostrand, 1945.
- [4] T. Burton. *Wind Energy Handbook*. Wiley, 2 edition, 2011.
- [5] A. Cayley. The Collected Mathematical Papers. *Cambridge Library Collection - Mathematics*, 14, 1889.
- [6] K. L. Chien, J. A. Hrones, and J. B. Reswick. On the automatic control of generalized passive systems. *Transactions ASME*, 74:175–185, 1952.
- [7] R. Curtain and H. Zwart. *Introduction to Infinite-Dimensional Systems Theory*. Springer Science, 1 edition, 2020.
- [8] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Springer, 2 edition, 2008.
- [9] DIN. IEC 60050-351 Leittechnik. www.electropedia.org, Area 351, Berlin 2014.
- [10] J. C. Doyle. Guaranteed Margins for LQG Regulators. *IEEE Transactions on Automatic Control*, 23(4):756–757, 1978.
- [11] L. Dörschel. *Regelung verteilparametrischer Systeme*. Habilitationsschrift, RWTH Aachen, 2022.
- [12] W. R. Evans. *Control-System Dynamics*. McGraw-Hill, 1954.
- [13] J. B. J. Fourier. *Théorie analytique de la chaleur*. Editions Jacques Gabay, 1822.
- [14] O. Föllinger. *Regelungstechnik*. VDE Verlag, 12 edition, Berlin, Offenbach 2016.

- [15] C. F. Gauss. *Theoria combinationis observationum erroribus minimis obnoxiae*. Commentationes Societatis Regiae Scientiarum Gottingensis recentiores 5 (classis mathematicae), 1823.
- [16] S. Geršgorin. Über die Abgrenzung der Eigenwerte einer Matrix. *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na*, 6:749–754, 1931.
- [17] P. Hartman. A lemma in the theory of structural stability of differential equations. *Proceedings of the American Mathematical Society*, 11(4): 610–620, 1960.
- [18] O. Hesse. *Vorlesungen über analytische Geometrie des Raumes*. Leipzig, 1876.
- [19] A. Hurwitz. Bedingungen, unter welchen eine Gleichung nur Wurzeln mit negativen reellen Teilen besitzt. *Mathematische Annalen*, 46:273–285, 1895.
- [20] R. Isermann. *Mechatronische Systeme*. Springer-Verlag, 2 edition, Berlin 2008.
- [21] A. Isidori. *Nonlinear Control Systems*. Springer, 1995.
- [22] C. G. J. Jacobi. *Gesammelte Werke*. Herausgegeben von K. W. Borchardt, A. Clebsch und K. Weierstraß, 1881–1891.
- [23] C. Jordan. *Traité des substitutions et des équations algébriques*. Gauthier-Villars, 1870.
- [24] W. Karush. Minima of Functions of Several Variables with Inequalities as Side Constraints. *Master Thesis, University of Chicago*, 1939.
- [25] R. E. Kálmán. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.
- [26] L. Euler. *Introductio in analysin infinitorum*. E101 und E102, 1748.
- [27] J. Li, F. L. Lewis, and J. Fan. *Reinforcement Learning*. Springer Cham, 1 edition, 2023.
- [28] D. G. Luenberger. Observers for multivariable systems. *IEEE Transactions on Automatic Control*, 11:190–197, 1966.

- [29] J. Lunze. *Regelungstechnik 1*. Springer-Verlag, 11 edition, Berlin 2016.
- [30] J. Lunze. *Regelungstechnik 2*. Springer-Verlag, 9 edition, Berlin 2016.
- [31] A. Lyapunov. Problème général de la stabilité du mouvement. *Annales de la faculté des sciences de Toulouse*, 2:203–474, 1892.
- [32] J. M. Maciejowski. *Predictive Control with Constraints*. Prentice-Hall, Harlow, 2001.
- [33] A. A. Markov. *Warscheinlichkeitsrechnung*. B. G. Teubner, 1912.
- [34] P. S. Marquis de Laplace. *Théorie analytique des Probabilités*. Ve. Courcier, 1812.
- [35] Mars Climate Orbiter Mishap Investigation Board. NASA Phase I Report, 1999.
- [36] J. C. Maxwell. On governors. *Proceedings of the Royal Society of London*, 16:270–283, 1868.
- [37] L. A. McGee and S. F. Schmidt. Discovery of the Kalman Filter as a Practical Tool for Aerospace and Industry. *NASA Technical Memorandum*, 86847, 1985.
- [38] J. P. Meijaard, J. M. Papadopoulos, A. Ruina, and A. L. Schwab. Linearized dynamics equations for the balance and steer of a bicycle: a benchmark and review. *Proceedings of the Royal Society A*, 463: 1955–1982, 2007.
- [39] O. Nelles. *Nonlinear System Identification*. Springer, 2 edition, 2020.
- [40] I. Newton. *De Methodis Serierum et Fluxionum*. Henry Woodfall, 1736.
- [41] H. Nyquist. Regeneration theory. *Bell System Technical Journal*, 11: 126–147, 1932.
- [42] M.-A. Parseval. Mémoire sur les séries et sur l'intégration complète d'une équation aux différences partielles linéaires du second ordre, à coefficients constants. *Mémoires présentés à l'Institut des Sciences*, 1: 636–648, 1806.

- [43] J. Pöschel. *Etwas mehr Analysis*. Springer Spektrum, 1 edition, 2014.
- [44] J. Raisch. *Mehrgrößenregelung im Frequenzbereich*. Oldenbourg, München, 1994.
- [45] K. Reinisch. *Analyse und Synthese kontinuierlicher Steuerungssysteme*. VEB Verlag Technik Berlin, 1 edition, 1979.
- [46] J. Riccati. Animadversiones in aequationes differentiales secundi gradus. *Actorum Eruditorum, quae Lipsiae publicantur, Supplementa*, 8: 66–73, 1724.
- [47] H. Rockel. *Analyse und Synthese parametrischer iterativ lernender Regelungen*. Dissertation, TU Clausthal, 2005.
- [48] E. Routh. *A Treatise on the Stability of a Given State of Motion: Particularly Steady Motion*. Macmillan, 1877.
- [49] K. Röbenack. *Nichtlineare Regelungssysteme - Theorie und Anwendung der exakten Linearisierung*. Springer Vieweg, 1 edition, 2018.
- [50] K. Schmitz and H. Murrenhoff. *Grundlagen der Fluidtechnik, Teil 1: Hydraulik*. Shaker, 9 edition, 2018.
- [51] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.
- [52] D. Simon. *Optimal State Estimation: Kalman, H_∞ , and Nonlinear Approaches*. John Wiley & Sons, 2 edition, 2006.
- [53] S. Skogestad and I. Postlethwaite. *Multivariable feedback control*. Wiley, 2 edition, 2006.
- [54] J.-J. Slotine and W. Li. *Applied Nonlinear Control*. Prentice Hall, 1994.
- [55] S. Stemmler. *Intelligente Regelungsstrategien als Schlüsseltechnologie selbstoptimierender Fertigungssysteme*. Dissertation, RWTH Aachen University, 2020.
- [56] B. Taylor. *Methodus Incrementorum Directa & Inversa*. William Innys, 1717.

- [57] H. Unbehauen. *Regelungstechnik I-III*. Friedr. Vieweg & Sohn, 1982–1985.
- [58] H. Vallery, J. Veneman, E. van Asseldonk, R. Ekkelenkamp, M. Buss, and H. van Der Kooij. Compliant actuation of rehabilitation robots. *IEEE Robotics & Automation Magazine*, 15(3):60–69, 2008.
- [59] B. van der Pol and J. van der Mark. The Heartbeat considered as a Relaxation oscillation, and an Electrical Model of the Heart. *Phil. Mag. Suppl.*, 6:763–775, 1928.
- [60] W. von Ockham. *Super IV libros Sententiarum*. Jean Trechsel, 1495 (Abschrift).
- [61] K. Voß, M. P. Werner, J. Gesenhues, V. Kučikas, M. van Zandvoort, S. Jockenhoevel, T. Schmitz-Rode, and D. Abel. Towards technically controlled bioreactor maturation of tissue-engineered heart valves. *Biomedical Engineering / Biomedizinische Technik*, 67(6):461–470, 2022.
- [62] C. Waldmann and O. Funke. The triple/nanoauv initiative a technology development initiative to support astrobiological exploration of ocean worlds. *CEAS Space Journal*, 12:115–122, 2020.
- [63] N. Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT Press, 1948.
- [64] J. G. Ziegler and N. B. Nichols. Optimum Settings for Automatic Controllers. *Transactions ASME*, 64:759–767, 1942.

Bezeichnungen

Die folgende Auflistung erhebt keinen Anspruch auf Vollständigkeit aller verwendeten Variablen, sondern hat das Ziel, die wichtigen Regelungstechnischen Vokabeln zu bündeln. Da es eine große Menge englischsprachiger Literatur zum Thema Regelungstechnik gibt, umfasst diese Auflistung ebenso wie das Stichwortverzeichnis zusätzlich wo möglich die englische Übersetzung. Hiermit soll der Zugang zu dieser Literatur erleichtert werden.

Schreibweisen

a, b, \dots	Skalare	<i>scalar</i>
$\mathbf{x}, \mathbf{y}, \dots$	Spaltenvektoren	<i>column vector</i>
$\mathbf{x}^T, \mathbf{y}^T, \dots$	Zeilenvektoren	<i>row vector</i>
$\mathbf{A}, \mathbf{B}, \dots$	Matrizen	<i>matrices</i>
$\underline{x}, \underline{y}, \dots$	Zeiger	<i>pointer</i>
X, Y, \dots	Absolutgrößen	<i>absolute values</i>
x, y, \dots	Abweichungsgrößen	<i>deviation values</i>
x_k, y_k, \dots	Folgen	<i>sequence</i>
X_0, Y_0, \dots	Arbeitspunktwerte	<i>operating point</i>
${}_0x, {}_0y, \dots$	Anfangswerte	<i>initial condition</i>

Regelungstechnische Vokabeln und ihre Formelzeichen

A	Systemmatrix	<i>system matrix</i>
A_R	Amplitudenreserve	<i>gain margin</i>
$B(U)$	Beschreibungsfunction	<i>describing function</i>
B	Eingangsmatrix	<i>input matrix</i>
C	Ausgangsmatrix	<i>output matrix</i>
D	Dämpfung	<i>damping</i>
D	Durchgangsmatrix	<i>feedthrough matrix</i>
e	Regelabweichung	<i>control deviation</i>
$\mathcal{F}\{f\} = F(j\omega)$	Fourier-Transformation	<i>Fourier transform</i>
$g(t)$	Gewichtsfunktion	<i>weighting function</i>
$G(s), G(z)$	Übertragungsfunktion	<i>transfer function</i>
$G(j\omega)$	Frequenzgang	<i>frequency response</i>

G_0	aufgeschnittener Regelkreis	<i>open-loop</i>
$h(t)$	Übergangsfunktion	<i>unit step response</i>
$\mathcal{L}\{f\} = F(s)$	Laplace-Transformation	<i>Laplace transform</i>
Q	Gewichtungsmatrix	<i>weighting matrix</i>
R	Regelfaktor	<i>control factor</i>
R	Gewichtungsmatrix	<i>weighting matrix</i>
T	Zeitkonstante	<i>time constant</i>
u	Stellgröße	<i>control variable</i>
$V(x)$	Lyapunov-Funktion	<i>Lyapunov function</i>
w	Führungsgröße	<i>reference variable</i>
x	Zustandsgröße	<i>state variable</i>
y	Regelgröße	<i>controlled variable</i>
z	Störgröße	<i>disturbance variable</i>
$\mathcal{Z}(f) = F(z)$	Z-Transformation	<i>Z transform</i>
α_R	Phasenreserve	<i>phase margin</i>
$\delta(t)$	Einheitsimpuls	<i>Dirac impulse</i>
η	Nullstelle	<i>zero</i>
λ	Eigenwert, Polstelle	<i>eigenvalue, pole</i>
ω_0	Kennkreisfrequenz	<i>characteristic angular frequency</i>
ω_d, ω_π	Durchtrittsfrequenzen	<i>crossover frequency</i>
ω_D	Eigenkreisfrequenz	<i>natural angular frequency</i>
ω_E	Eckkreisfrequenz	<i>corner angular frequency</i>
ω_g	Grenzfrequenz	<i>cut-off frequency</i>
$1(t)$	Einheitssprung	<i>Heaviside function</i>
*	Faltung	<i>convolution</i>

Index

- Absolutgrößen, *absolute values*, 50
Abtaster, *sampler*, 377, 396, 411
Abtasttheorem, *sampling theorem*, 380, 395, 565
Abtastzeit, *sampling time*, 375, 381, 395, 565
Abweichungsgrößen, *deviation values*, 50
Active set, *active set*, 547
Aliasing, *aliasing*, 380
Allpass, *all-pass*, 154, 155, 193, 337
Amplitudengang, *gain response*, 119, 524
Amplitudenreserve, *gain margin*, 248, 278, 284, 522
Anfangsbedingung, *initial condition*, 23, 102, 107, 384, 452
Anschwingzeit, *step response time*, 260
Aperiodischer Grenzfall, *critically damped*, 180, 299
Arbeitspunkt, *operating point*, 52, 536
Ausgangsrückführung, *output feedback*, 263

Backstepping, *backstepping*, 502, 504
Bandbreite, *bandwidth*, 177, 381, 526
Bandpass, *band-pass*, 134
Beobachtbarkeit, *observability*, 313, 315, 417, 499, 520, 568
Beobachter, *observer*, 310, 417, 426

Beschreibungsfunktion, *describing function*, 464, 508
Betragsbedingung, *magnitude condition*, 322
Bilineare Transformation, *bilinear transform*, 412
Bleibende Regelabweichung, *steady state error*, siehe Stationäre Genauigkeit
Bode-Diagramm, *Bode plot*, 126

Charakteristisches Polynom, *characteristic polynomial*, 68, 361, 567
Chatter, *chatter*, 508

D-Element, *D element*, 168, 390, 409
Definit
 positiv semi-, *positive semi-definit*, 519, 544
 positiv, *positive definit*, 212, 433, 453, 518
Diagonaldominant, *diagonal dominant*, 370, 374
Differentialgleichung, *differential equation*, 22
Differenzengleichung, *difference equation*, 384
Differenzierendes Verhalten, *differentiating behaviour*, 169
Dominant
 Pol- und Nullstellen, *dominant poles and zeros*, 152, 499
 Signal, *dominant signal*, 150, 273

- DT₁-Element, *real derivative element*, 188
Durchgriff, *siehe* Sprungfähig
Durchtrittsfrequenz, *crossover frequency*, 240, 248, 270, 273, 286, 363
Dämpfung, *damping*, 44, 178, 297, 415
Eckkreisfrequenz, *corner angular frequency*, 49, 129, 182
Eigenkreisfrequenz, *natural angular frequency*, 69, 297, 395
Eingeschwungener Zustand, *steady-state*, 116
Einschwingzeit, *settling time*, 260
Einstellregeln
 Chien, Hrones und Reswick, 276
 Ziegler und Nichols, 276
Einzugsbereich, *region of attraction*, 450
Endliche Einstellzeit, *dead-beat*, 418
Entkopplungsregler, *decoupling control*, 373
Erwartungstreu, *unbiasedness*, 433
Erwartungswert, *expected value*, 430
Exakte Linearisierung
 Exakte Linearisierung, *exact linearization*, 481, 491
Eingangs-Ausgangs, *input-output*, 483
Eingangs-Zustands, *input-state*, 483, 493
Faltung, *convolution*, 91, 402
Festwertregelung, *fixed set-point control*, 188, 335
Filter, *filter*, 133, 318, 426
Flachheit, *flatness*, 487, 488, 492
Folgeregelung, *follow-up control*, 188, 335, 532, 539
Fourier-Transformation, *Fourier transform*, 131
Fourierreihe, *Fourier series*, 463, 465
Frequenzgang, *frequency response*, 118, 131, 409
Frequenzspektrum, *Signal frequency spectrum*, 132
Führungsgröße, *reference variable*, 7
Führungsgrößenfilter, *prefilter*, 340, 567
Führungsübertragungsfunktion, *siehe* Komplementäre Sensitivität
Generalized Plant, *generalized plant*, 528, 532
Geometrische Reihe, *geometric series*, 400
Gershgorin-Theorem, *Gershgorin theorem*, 369
Gewichtsfolge, *unit impulse response*, 386, 406, 566
Gewichtsfunktion, *weighting function*, 89
Gewichtung, *weighting*, 218, 429, 436, 515, 519, 527, 531, 538
Gleitebene, *Sliding Surface*, 511
Grenzfrequenz, *cut-off frequency*, 177
Grenzzyklus, *limit cycle*, 458, 508
Halteglied, *holding element*, 377, 396, 411, 565

- Hilfsstellgröße, *auxiliary control variable*, 348
- Hochpass, *high-pass*, 134
- Homogene Lösung, *homogeneous solution*, 85
- Horizont, *horizon*, 537
- Hurwitz-Kriterium, *Hurwitz criterion*, 217, 413
- I-Element, *I element*, 33, 164, 390
- Identifikation, *identification*, 200
- ILR
- ILR, *ILC*, 561
 - Entfaltungs-, *deconvoluting ILC*, 572
 - Normoptimal, *norm optimal ILC*, 572
- Impulsantwort, *impulse response*, 89, 566
- Integrator-Windup, *windup*, 402, 476, 536
- Integrierendes Verhalten, *integrating behaviour*, 166, 167, 247, 348, 462
- Interior point, *interior point*, 547
- Interne Dynamik, *internal dynamics*, 484, 496
- IT₁-Element, 187
- Kalmanfilter
- Kalmanfilter, *Kalman filter*, 426, 435, 520
 - Erweitert, *Extended Kalmanfilter*, 445
- Kaskadenregelung, *cascade control*, 15, 345, 561
- Kausalität, *causality*, 26, 171, 188, 306, 315, 337, 386
- Kennkreisfrequenz, *characteristic angular frequency*, 44, 178
- Kennlinie, *characteristic curve*, 33, 64, 459, 462, 472
- Knoten, *node*, 450
- Komplementäre Sensitivität, *complementary sensitivity*, 141, 275, 531
- Korrespondenztafel, 99, 402
- Kostenfunktion, *cost function*, 515, 535
- Kovarianz, *covariance*, 431, 435
- Laplace-Transformation, *Laplace transform*, 97, 401
- Lifted-System Darstellung, *lifted system representation*, 565
- Linearisierung, *linearization*, 60, 62, 442, 447
- Linearisierungstheorem, *linearization theorem*, 72, 388, 449
- LQG, *LQG*, 521
- LQR, *LQC*, 515, 536
- Lyapunov-Funktion, *Lyapunov function*, 453, 454, 501, 502, 507
- Matrix
- Ausgangs- *output*, 29
 - Durchgangs-, *feedthrough*, 29
 - Eingangs- *input*, 29
 - Jacobi, *Jacobian*, 60
 - Markov, *Markov matrix*, 566
 - System- *system*, 29
 - Transitions-, *transition matrix*, 78, 395, 427
- Messglied, *sensor*, 7
- Messgröße, *measured variable*, 7, 529

Methode der kleinsten Fehlerquadrat, *least-square*, 210, 429, 514
Minimale Realisierung, *minimal realization*, 47, 95, 254, 314, 359
Minimalphasig, *minimumphase*, 153, 155, 194, 271, 337, 408, 499
Modell, *model*, 20
nominal, *nominal model*, 506
MPR
MPR, *MPC*, 534
explizit, *explicit MPC*, 548
linear, *linear MPC*, 540, 544, 548
nichtlinear, *nonlinear MPC*, 555

Normalform
Beobachtungs-, *observer canonical form*, 313
Byrnes-Isidori, *Byrnes-Isidori canonical form*, 485
Diagonale, *diagonal canonical form*, 80
Jordansche, *Jordan canonical form*, 82
nichtlineare Regelungs-, *nonlinear control canonical form*, 491, 505
Regelungs-, *control canonical form*, 32, 493
strenge Rückkopplungsform, *strict feedback form*, 491, 504
Normalverteilung, *normal distribution*, 432, 436, 444
Nulldynamik, *zero dynamics*, 497, 498
Nyquist-Kriterium
vereinfacht, *simplified Nyquist cri-*

terion, 241, 247, 469
vollständig, *full Nyquist criterion*, 225, 231, 367

Optimale Regelung, *optimal control*, 513, 522, 534, 545
Ortskurve, *Nyquist plot*, 123, 231, 232, 409

P-Element, *P element*, 33, 157
PA₁-Element, *first-order all-pass*, 193
Parallelschaltung, *parallel connection*, 138, 355
Partialbruchzerlegung, *partial fraction decomposition*, 105, 402
Partikuläre Lösung, *particular solution*, 85
PD-Element, *PD element*, 170, 306
PDT₁-Element, *lead-lag element*, 191
Phasengang, *phase response*, 119
Phasenporträt, *phase portrait*, 448
Phasenreserve, *phase margin*, 248, 284, 398
PI-Element, *PI element*, 171
PID-Element, *PID element*, 173, 392, 408
Pol-Nullstellen-Diagramm, *pole-zero plot*, 110
Pol-Nullstellen-Kürzung, *pole-zero cancellation*, 254, 422, 499
Polvorgabe, *pole placement*, 301, 302, 371, 414, 496, 501, 570
Polynomdivision, *polynomial long division*, 405
PPT₁-Element, *lead-lag element*, 191

- Prädiktion, *prediction*, 428, 536
 Pseudoinverse, *generalized inverse*, 210, 429
 PT_1 -Element, *first-order lag element*, 175, 390
 PT_1T_t -Element, 196
 PT_2 -Element, *second-order lag element*, 177
 PT_t -Element, *dead-time element*, 196
 Quadratisches Programm, *quadratic program*, 540, 543
 quasikontinuierlich, *quasi continuous*, 382, 395
 Querkopplung, *cross coupling*, 359
 Rampenantwort, *ramp response*, 187
 Rauschen, *noise*, 7, 529
 Regelfaktor, *control factor*, 266
 Regelgröße, *controlled variable*, 7, 529
 Regelkreis
 aufgeschnitten, *open loop*, 140, 230, 262, 357
 geschlossen, *closed loop*, 8, 262
 offen, *open loop*, 8
 Regelstrecke, *controlled system*, 7, 257
 Regler, *feedback control*, 7, 257
 Reihenschaltung, *serial connection*, 138, 355
 Relativ Gain Array, *relative gain array*, 365
 Relativer Grad, *relative degree*, 26, 114, 385, 480
 Residuum, *residue*, 105, 151
 Resonanzfrequenz, *resonance frequency*, 183
 Resonanzüberhöhung, *resonant peak*, 183
 Riccati-Gleichung, *Riccati equation*, 517
 Robustheit, *robustness*, 251, 271, 500, 505, 572
 Routh-Kriterium, *Routh criterion*, 218, 413
 Rückführdifferenzmatrix, *return difference matrix*, 357, 367
 Rückführung, *feedback*, 138, 356
 Rückwärtsdifferenzen, *backward difference*, 389, 391
 Ruhelage, *equilibrium state*, 53, 386, 450
 Sattel, *saddle*, 450
 Schleppfehler, 188
 schwingungsfähig, *oscillatory behaviour*, 69, 408
 Sensitivität, *sensitivity*, 141, 275, 525, 526
 Separationstheorem, *separation principle*, 316, 521
 Shannonfrequenz, *Nyquist frequency*, 380
 Signal
 Signal, *signal*, 5
 stochastisch, *stochastic signal*, 430
 zeitdiskret, *discrete time signal*, 375
 zeitkontinuierlich, *continuous time signal*, 375
 Singulärwert, *singular value*, 362, 530
 Skalierung, *scaling*, 363
 Sliding Mode Regler, *Sliding Mo-*

de Controller, 505, 512
Sollwert, *reference value*, 7
Sprungantwort, *step response*, 89
Sprungfähig, *system with feedthrough*,
113, 255, 341
Stabilität
asymptotisch, *asymptotically stable*, 56, 454
BIBO-stabil, *BIBO stable*, 94
global, *globally stable*, 452, 454,
497
grenzstabil, *marginally stable*, 57,
454
instabil, *unstable*, 57
Kriterien, *stability criteria*, 71,
214, 388, 407
lokal, *locally stable*, 452
semi-stabil, *semistable*, 458
von Grenzzyklen, 458, 472, 474
Stabilitätsrand, *margin of stability*, 71, 221, 234
Stationäre Genauigkeit, *stationary accuracy*, 164, 167, 260, 348,
552, 569
Stellgröße, *control variable*, 7, 529
Stellgrößenbeschränkung, *input constraint*, 269, 462, 465, 535, 545
Steuerbarkeit, *controllability*, 302,
304, 417, 488, 518, 568
Steuerung, *open-loop control*, 9, 336,
343, 561
Strudel, *spiral*, 450
Störgröße, *disturbance*, 7, 167
Störgrößenaufschaltung, *disturbance feedforward control*, 343
Störübertragungsfunktion, *siehe Sensitivity*

System
System, *system*, 5
autonom, *autonomous*, 53, 386
eingangsaffin, *input affine*, 492
linear, *linear*, 25
LTI, *LTI*, 25, 385
MIMO, *MIMO*, 30, 351, 355, 367,
535
nichtlinear, *nonlinear*, 25
Ordnung, *order*, 22, 531
schaltend, *switching system*, 459,
507
SISO, *SISO*, 30
zeitdiskret, *discrete time system*,
375
zeitinvariant, *time-invariant*, 24
zeitkontinuierlich, *continuous time system*, 375
Taylorreihe, *Taylor series*, 60
Tiefpass, *low-pass*, 134, 337, 382,
464, 472
Totzeit, *dead-time*, 196, 251, 385,
396, 508, 535
 T_u - T_g -Modell, 206
Überanpassung, *overshitting*, 205
Übergangsfolge, *unit step response*, 386
Übergangsfunktion, *unit-step response*, 89
Überschwingen, *overshoot*, 181, 260,
341, 514
Übertragungsfunktion, *transfer function*, 107, 405
Übertragungsmatrix, *transfer matrix*, 108, 355

- Unsicherheit, *uncertainty*, 344, 506, 533, 552
Unterabtastung, *undersampling*, 380
Unterschwingen, *undershoot*, 181
Verstärkung, *gain*, 44, 96, 112, 151, 247, 266, 267, 271, 362, 388
Verzögerungsglied, *lag element*, 175
Vorsteuerung, *feedforward control*, 334, 374, 493
Wasserbetteffekt, *waterbed effect*, 526
Winkelbedingung, *angle condition*, 322
Wirkungsplan, *functional diagram*, 3, 32, 44
Wohldefiniert, *well-defined*, 481, 488
Wurzelortskurve, *root locus*, 322, 417
 \mathcal{Z} -Transformation, \mathcal{Z} transform, 399, 401
Zeiger, *pointer*, 116, 464
Zeitdiskret, *discrete time*, 383
Zeitkonstante, *time constant*, 49
Zentrum, *center*, 450
Zurückweichender Horizont, *receding horizon*, 535
Zustandsraumdarstellung, *state space representation*, 29, 63, 394
Zustandsrückführung, *state feedback*, 263, 301, 515
Zustandsschätzung, *state estimation*, 306
Zustandstransformation, *state transformation*, 76, 79, 304, 482, 493, 502
Zwei-Ortskurve-Kriterium, 472