SCHRIFTENREIHE DES INSTITUTS FÜR
ANGEWANDTE INFORMATIK | AUTOMATISIERUNGSTECHNIK
KARLSRUHER INSTITUT FÜR TECHNOLOGIE (KIT)

Band 52

LUTZ GRÖLL

Methodik zur Integration von Vorwissen in die Modellbildung

KIT Scientific Publishing

Lutz Gröll

Methodik zur Integration von Vorwissen in die Modellbildung

Schriftenreihe des
Instituts für Angewandte Informatik / Automatisierungstechnik
am Karlsruher Institut für Technologie
Band 52

Eine Übersicht aller bisher in dieser Schriftenreihe erschienenen Bände
finden Sie am Ende des Buches.

Methodik zur Integration von Vorwissen in die Modellbildung

von
Lutz Gröll

Scientific Publishing

Habilitation, Karlsruher Institut für Technologie (KIT)
Fakultät für Maschinenbau, 2014

Tag des Habilitationskolloquiums: 30. April 2014

Impressum

Scientific Publishing

Karlsruher Institut für Technologie (KIT)
KIT Scientific Publishing
Straße am Forum 2
D-76131 Karlsruhe

www.ksp.kit.edu

Print on Demand 2015

ISSN 1614-5267
ISBN 978-3-7315-0303-3
DOI: 10.5445/KSP/1000044645

Methodik zur Integration von Vorwissen in die Modellbildung

Der Fakultät für Maschinenbau
des Karlsruher Instituts für Technologie

vorgelegte
Habilitationsschrift
von

Dr.-Ing. Lutz Gröll
geboren am 18. Mai 1965 in Meißen

Hauptreferent:	Prof. Dr.-Ing. habil. G. Bretthauer
Korreferenten:	Prof. Dr.-Ing. P. Gratzfeld
	Prof. Dr.-Ing. habil. K. Röbenack
	Prof. Dr.-Ing. habil. J. Wernstedt
Tag der Einreichung:	21. Oktober 2013

Vorwort

Die vorliegende Arbeit entstand während meiner Tätigkeit am Institut für Angewandte Informatik des Karlsruher Instituts für Technologie. Sie enthält wesentliche Ergebnisse meiner Forschungsarbeiten zur Verbesserung der Modellbildung durch das Einbeziehen von Vorwissen.

Mein Dank gilt an erster Stelle Herrn Prof. Dr.-Ing. habil. G. Bretthauer. Er gewährte mir umfangreiche Unterstützung und große Freiheiten. Vor allem aber motivierte er mich beständig zum Fertigstellen der Arbeit, wenn bedingt durch intensive Projekt- oder Vorlesungstätigkeit das Schreiben ins Stocken geriet. Letztlich gebührt ihm großer Dank für die gründliche Durchsicht des Manuskriptes und die wertvollen Hinweise. Letztere führten zwar zu einer Ausweitung des Umfangs, dienten aber der Erhöhung der Verständlichkeit.

Weiterhin gilt mein Dank Herrn Prof. Dr.-Ing. P. Gratzfeld, Herrn Prof. Dr.-Ing. habil. K. Röbenack und Herrn Prof. Dr.-Ing. habil. J. Wernstedt für die Übernahme des Korreferats und dessen zeitnaher Erstellung.

Danken möchte ich insbesondere meinen Kollegen Herrn Dr.-Ing. Jörg Matthes und Herrn Dr.-Ing. Markus Reischl. Mit beiden diskutierte ich über die Jahre den ein oder anderen Abschnitt intensiv. Das half, die Sachverhalte mit einfacherer Mathematik darzustellen oder sie ingenieurmäßig besser zu motivieren.

Karlsruhe, im Oktober 2013 Lutz Gröll

Inhaltsverzeichnis

Kapitel 1

Einleitung

Bei der Modellierung technischer und nichttechnischer Systeme können bessere Modelle hinsichtlich der Problemadäquatheit, der Statistik oder der Robustheit erhalten werden, wenn quantitatives und qualitatives Vorwissen über das betrachtete System in den Modellbildungsprozess einbezogen wird. Oft ist dieses Vorwissen aber nicht in geeigneter Form oder sogar nur verbal verfügbar wie etwa „Das System ist passiv.“ oder „Die Kennlinie ist monoton.“ Da mitunter nicht klar ist, wie das Vorwissen mathematisch zu fassen und in die Optimierung einzubeziehen ist, und da diese Thematik in Standardwerken zur Identifikation [232], [572], [406], [508], [314], [313], [628] bestenfalls am Rande betrachtet wird, bleibt das Vorwissen häufig ungenutzt. Den Standardwerken zur Optimierung [71], [320], [417], [194] und im Speziellen zur konvexen Optimierung [63], [475], [93], [90] hingegen fehlt weitgehend der Bezug zu den dynamischen Modellen und deren Eigenschaften, die ihrerseits aber gerade bei der Modellierung im Rahmen der Regelungstechnik von Interesse sind. Um also das im Vorwissen vorhandene Potenzial zu heben, ist das Vorwissen in eine solche mathematische Form zu bringen, die eine Berücksichtigung in einem Gütekriterium erlaubt. Genau hier greift diese Arbeit an, indem sie die Thematik über zwei Teilthemen angeht:

1. Formulierung von Vorwissen in parametrische und strukturelle Restriktionen (Teil 1)
2. Methoden zur Behandlung restringierter Optimierungsprobleme (Teil 2).

Obwohl beide Teile zunächst losgelöst voneinander erscheinen, beeinflussen sie sich doch wechselseitig. Denn was nutzt eine Restriktion, die algorithmisch nur schwer zu handhaben ist. Hieraus ergibt sich die eigentliche Herausforderung für den Ingenieur, da Standardkenntnisse zur Problemlösung gemeinhin nicht ausreichen, um verbales Vorwissen mathematisch geschickt zu formulieren. Vielmehr wird nämlich ein breites Wissen und eine reichhaltige Erfahrung sowohl in der Systemtheorie als auch der Optimierungstheorie benötigt. Ziel der

Arbeit ist es deshalb, dem Ingenieur Wege aufzuzeigen, wie er das Vorwissen einer konkreten Modellierungsaufgabe in ein mathematisches Optimierungsproblem integrieren und dieses durch Umformungen bzw. Vereinfachungen gut handhabbar gestalten kann.

Eine Problematik beim Zusammenführen der Formulierung und Behandlung von Restriktionen liegt in der Vielfalt möglicher Eigenschaften, Modelltypen und Modellierungsklassen. Letztere erstrecken sich von der Approximation und Regression über die Prädiktion bis hin zur Klassifikation. Zudem gibt es innerhalb einer einzelnen Klasse zahlreiche Methoden. So zählen beispielsweise zur Approximation die Datenapproximation mit den Unterklassen „Interpolation“ und „Ausgleich“ sowie die Abbildungsapproximation, die Funktionen, Matrizen und Systeme umfasst. Allein die Systemapproximation kann wiederum in Zugänge „linear nach linear“, „nichtlinear nach linear“, „nichtlinear nach nichtlinear“ oder „kontinuierlich nach zeitdiskret“ unterteilt werden. Um dieser enormen Vielfalt an Eigenschaften und Problemklassen gerecht zu werden, wird als Gemeinsamkeit das restringierte Optimierungsproblem gewählt, in das die Modellierungsaufgaben letztlich münden. Ein restringiertes Optimierungsproblem wird dabei abstrakt durch folgende Gleichungen beschrieben:

$$\begin{aligned}
\textbf{Suchraum} + \textbf{Restriktionen} &= \textbf{zulässige Menge} \\
\textbf{Zielfunktion} + \textbf{Ordnungsrelation} &= \textbf{Gütekriterium} \\
\textbf{Gütekriterium} + \textbf{zulässige Menge} &= \textbf{restringiertes Optimierungsproblem.}
\end{aligned}$$

Beim Stellen des restringierten Optimierungsproblems ist der Schwierigkeitsgrad für das Lösen zu beachten (Problemgröße, Mehrdeutigkeiten, Differenzierbarkeit usw.). Bekannt ist, dass heutzutage lineare Probleme mit 1000 Variablen und bis zu 10000 linearen Restriktionen und lineare Least-Squares-Probleme mit 1000 Variablen und 10000 Gleichungen keine nennenswerten Schwierigkeiten bereiten. Diese Tatsache erweckt den Anschein, die Linearität sei der Gradmesser für die Schwierigkeiten. Doch R. T. Rockafellar [539], einer der weltweit führenden Wissenschaftler auf dem Gebiet der Optimierungstheorie, bemerkte zu dieser Thematik

> „In fact the great watershed in optimization
> isn't between linearity and nonlinearity,
> but convexity and nonconvexity.“

Die Gründe hierfür sind vielschichtig. So sind lokale Minimierer konvexer Probleme stets auch globale und, wenn ein Minimierer existiert, ist dieser bei streng konvexen Funktionen

eindeutig. Ferner sind die Optimalitätsbedingungen unter schwachen Einschränkungen notwendig und hinreichend, liefert die Dualitätstheorie starke Aussagen und es existieren Subgradientenkalküle, die die Behandlung nichtdifferenzierbarer Funktionen gestatten [90], [93], [475], [63]. Für den Praktiker aber ist die Existenz und sehr gute Verfügbarkeit ausgefeilter Algorithmen für spezielle konvexe Problemklassen vordergründig, wobei die SDP-Probleme (SDP steht für „semidefinit programming“) besonders herausragen. Dank ihrer Eigenschaften und der Algorithmen sind alle konvexen Probleme, die sich passend umschreiben lassen, fast so effizient lösbar wie LP- oder LS-Probleme (LP steht für „linear programming“, d. h. Problemlösung mit linearer Zielfunktion und linearen Restriktionen; LS steht für „least squares“, d. h. Kleinste-Quadratsummen-Probleme). Als Konsequenz daraus wird in dieser Arbeit gezielt darauf geachtet, wann immer möglich, eine konvexe Formulierung einer Restriktion zu erhalten.

Aus den Vorbemerkungen ergeben sich für diese Arbeit die Herausforderungen insbesondere aus der Problemvielfalt und der Nichtkonvexität. Beide Herausforderungen werden über die folgenden Teilziele bewältigt:

1. Behandlung eines breiten Spektrums von Vorwissen
2. Vereinheitlichte Darstellung
3. Übersichten zu Begriffen, Formeln und Zusammenhängen
4. Bewertung von Zugängen aus theoretischer Sicht
5. Empfehlungen zu Methoden aus der eigenen praktischen Erfahrung
6. Ableitung modifizierter Verfahren für spezielle Aufgaben
7. Darstellung prägnanter Beispiele zur Anwendbarkeit/Nichtanwendbarkeit
8. Erschließung für den Ingenieur bisher unbekannter Literatur.

Der Aufbau der einzelnen Abschnitte im Teil 1 ist vom Grundansatz fast immer gleich. Ausgangspunkt bildet eine Begriffsklärung, aus der eine Restriktion oder Struktur gefolgert wird, sofern sie nicht in der Erklärung bereits enthalten ist. Die Grundidee und die Schlussfolgerungen werden kurz vorgestellt und durch Anmerkungen zu begleitenden Gedanken ergänzt. Einfache, meist per Hand nachrechenbare Beispiele dienen der Erläuterung, während Gegenbeispiele vor Denkfallen warnen. Eine Restriktion selbst gilt in der Arbeit als formuliert, sobald sie in algebraischer Form in Gleichungen, Definitionen, Kriterien oder Bedingungen vorliegt oder wenn sie auf eine wohlbestimmte Systemstruktur führt. Ebenso wird ein Problem als gelöst angesehen, wenn es in ein Standardproblem der Optimierung überführt wurde, für das entsprechende Software (z. B. Matlab) existiert.

Dem Zweck einer kompakten Darstellung dienen auch zahlreiche Tabellen und Übersichten im Anhang, etwa zu den Restriktionstypen, zur Nomenklatur restringierter Probleme und zur Stabilitätstheorie. Die Ausführungen zur Stabilitätstheorie richten sich dabei an jene Leser, die Stabilitätsrestriktionen für nichtlineare Systeme ableiten wollen oder die die Theorie zur Konvergenzbeweisführung für Modellbildungsalgorithmen (selbstentwickelten Online-Algorithmen, in adaptiven Regler- und Beobachtersystemen) heranziehen möchten. Leser, die sich mit statischen, nichtparametrischen oder linearen Modellen beschäftigen, benötigen die Ausführungen zur Stabilitätstheorie im Anhang nicht, da insbesondere in Kapitel 2 Methoden zur Umsetzung der Stabilitätsrestriktionen für lineare Modelle vorgestellt werden.

Neben den erwähnten Übersichten enthalten ausgewählte Abschnitte zusätzliche Formelzusammenstellungen, die für die Bearbeitung konkreter Probleme ein kompaktes Handwerkszeug bieten, das so in Standardwerken nur verstreut oder gar nicht enthalten ist.

Sofern möglich, werden Aussagen zu Verfahren aus theoretischer Sicht getroffen. Bei nichtkonvexen Problemstellungen ist das allerdings nur bedingt möglich. Gelingt es allerdings, eine Systemeigenschaft als generisch (fast immer zutreffend) einzustufen, entschärft sich ein Problem. Einfach gesprochen heißt das für die Identifikation nämlich: Mit Wahrscheinlichkeit Eins brennt nichts an, wenn eine Restriktion für eine generische Eigenschaft einfach weggelassen wird. Diese Aussage freut den Praktiker, braucht er sich doch um eine Restriktion weniger zu kümmern. Für den Theoretiker sei bemerkt, dass eine Unterscheidung in topologisch generisch und metrisch generisch im Rahmen dieser Arbeit nicht erforderlich ist. Beide Begriffe sind in den betrachteten Fällen nämlich äquivalent, da als Ausschlussrestriktionen nur algebraische und semialgebraische Restriktionen auftreten. Unabhängig davon wird die topologische Betrachtung bevorzugt, weil dichte und magere Mengen anschaulicher als sogenannte Nullmengen sind.

Eine weitere zentrale Technik in dieser Arbeit, die eingesetzt werden kann, wenn es nicht gelingt, eine konvexe Restriktion abzuleiten oder die Generizität einer Restriktion zu zeigen, ist die der Matrixapproximationsprobleme. Die Idee dahinter ist einfach: Lass' die Restriktion zunächst weg und sichere sie a posteriori durch eine Projektion in die zulässige Menge; tausche damit Optimalität gegen Suboptimalität. Anwendung findet die Idee bei bestimmten Prokrustes-Problemen (LS-Formulierung mit Matrixrestriktion), die schwierig zu lösen sind. Gerade die in der Systemtheorie wichtigen Restriktionen an die Lage von Eigenwerten sind nicht konvex und führen leider auf derartig schwierige Probleme. Der Vorteil der Matrixapproximationsprobleme erwächst dabei aus der Verfügbarkeit eines breiten Spektrums gelöster Probleme. Neben Aussagen zur Eindeutigkeit und zur Zahl möglicher Nebenminima existieren zugeschnittene Algorithmen, die beispielsweise die Suche vom n^2-dimensionalen Parameterraum auf einen n- oder manchmal sogar ein- bzw. zweidimensionalen Raum reduzieren.

Entsprechend der Zweiteilung der Arbeit widmet sich der *erste Teil* dem Formulieren von Restriktionen zu den Themen

1. **Restriktionen an dynamische Modelle** (Kapitel 2)
   Im Speziellen werden Statikrestriktionen, Restriktionen an diverse Systemeigenschaften (Steuerbarkeit, Passivität, Minimalphasigkeit usw.) behandelt, vor allem aber Restriktionen, die sich aus der Stabilität linearer zeitinvarianter Systeme für die unterschiedlichen Beschreibungsformen ergeben. Während in der Literatur die Einbeziehung von Stabilitätsrestriktionen für lineare Modelle zunehmend mehr Beachtung findet, bleibt die Thematik für nichtlineare Modelle auf wenige Probleme beschränkt. Das liegt an der Vielfalt unterschiedlicher Stabilitätsdefinitionen, dem Schwierigkeitsgrad und der eingeschränkten Generalisierungsfähigkeit einzelner Lösungen. Aus diesem Grund werden im Kapitel Zugänge zur Identifikation nichtlinearer Modelle vorgestellt, die die Problematik der Stabilitätsrestriktionen für nichtlineare Modelle weitgehend umgehen und dennoch gute Modelle liefern. Ergänzend findet sich im Anhang eine kompakte Darstellung zu Definitionen und Sätzen bezüglich der Stabilität nichtlinearer Systeme. Ein Abschnitt zur Einbeziehung von Vorwissen über Bifurkationen, also das Auftreten qualitativer Änderungen im dynamischen Verhalten, beschließt das Kapitel.

2. **Restriktionen an Funktionen** (Kapitel 3)
   Während die Identifikation aus Sicht der Regelungstechnik vordergründig dynamische Modelle im Fokus hat, werden bei der Modellierung nichttechnischer Systeme dagegen oft statische Zusammenhänge betrachtet. Beschrieben werden die Zusammenhänge durch parametrische und nichtparametrische Funktionen. Auch für die Funktionen kann Vorwissen genutzt werden, um die Zielfunktionen des resultierenden Optimierungsproblems durch Restriktionen zu ergänzen. Letztlich lassen sich so wiederum wirksamere Modelle erstellen. Die Restriktionen für Funktionen ergeben sich dabei aus Eigenschaften wie Nichtnegativität, Monotonie und Konvexität oder aber aus Forderungen nach Stetigkeit, Glattheit oder dem Einhalten exakter Funktionswerte. All diese Aspekte werden in den Abschnitten des Kapitels betrachtet, wobei Anwendungen aus der Interpolation, Approximation und Regression das Einsatzspektrum verdeutlichen. Ergänzt werden diese Abschnitte um Restriktionen zu Kovarianzfolgen und Kovarianzmatrizen. Beide Beschreibungsformen treten vielfach als Zwischenmodell bei der Modellierung dynamischer Systeme auf, oder sie werden zum Erkennen kausaler Zusammenhänge und damit zur Generierung von Vorwissen herangezogen. Darüber hinaus wird an den unterschiedlichen Typen von Restriktionen an die Kovarianzmatrizen auch deutlich, warum es mehrere Möglichkeiten zu ihrer mathematischen Behandlung geben muss.

3. **Eindeutigkeitserzwingende Restriktionen** (Kapitel 4)
   Während sich die Kapitel 2 und 3 dem Erstellen von Restriktionen aus dem Modell heraus widmen, geht es in Kapitel 4 um das Vermeiden der Überparametrisierung eines Modells und um das Sicherstellen der Eindeutigkeit des mathematischen Problems, über das die Modellparameter bestimmt werden. Die Frage der Überparametrisierung wird in der Literatur zur experimentellen Modellbildung aus unterschiedlichen Blickwinkeln heraus betrachtet, was auf die auch in dieser Arbeit erläuterten diversen Abstufungen des Identifizierbarkeitsbegriffs führt. Durch Fixieren von Parametern auf feste Werte oder das Verwenden von Normalformen lässt sich die Überparametrisierungsproblematik lösen. Darüber hinaus wird in dieser Arbeit aber auch gezeigt, dass neben dem Fixieren von Parametern auch andere, in bestimmten Fällen zweckmäßigere Parameterrestriktionen verwendet werden können. Die Frage nach der Eindeutigkeit des mathematischen Problems wird in der ingenieurgeprägten Modellbildungsliteratur selten betrachtet, da überwiegend Quadratmittelansätze betrachtet werden. Nach wie vor dominieren diese Ansätze die Anwendungen. Doch dank leistungsfähiger Rechner und verfügbarer, einfach handhabbarer Software können heutzutage auch nichtquadratische Ansätze effizient gelöst werden. Sie werden bevorzugt, wenn die Fehler nicht normalverteilt sind oder wenn eine hohe Modellrobustheit angestrebt wird. Anders als die quadratischen Probleme, die bei geeigneten Experimentaldaten und parameterlinearen Modellen auf streng konvexe Formulierungen und damit auf eindeutige Lösungen führen, muss Eindeutigkeit bei parameternichtlinearen Modellen oder bei nichtquadratischen Ansätzen nicht mehr vorliegen. Wann sie aber gilt und wie zu verfahren ist, wenn sie nicht gilt, wird im Kapitel erläutert. Hierfür wird insbesondere das mathematische Rüstzeug für den Ingenieur aufbereitet, wobei Aussagen zu Lösungsmengen und deren Reduktion durch Restriktionen den Schwerpunkt bilden.

Der *zweite Teil* der Arbeit beschreibt wichtige Methoden zur Behandlung der in den vorangehenden Kapiteln erstellten restringierten Probleme. Dabei ist zwischen äquivalenten Problemen, bei denen aus dem umgeformten Problem auf die Lösung des Originalproblems geschlossen werden kann, und nichtäquivalenten Problemen, bei denen ein modifiziertes Problem eine Näherungslösung erzeugt, zu unterscheiden:

1. **Reduktionsmethoden** (Kapitel 5)
   Reduktionsmethoden erzeugen ein äquivalentes Problem mit einer geringeren Anzahl von Variablen und/oder Restriktionen. Den Schwerpunkt der Betrachtungen bilden die klassischen Zugänge zur Elimination von Ungleichungs- und Gleichungsrestriktionen. Dabei wird auch erläutert, wie redundante Restriktionen erkannt werden können, die sich bei der Umsetzung von Vorwissen für kompliziertere Systeme ergeben können. Algorithmisch bedeutend ist ferner das Aufspüren singulärer Ungleichungsrestriktionen,

also jener, die nur mit Gleichheit gelten. Dadurch werden nämlich nicht nur die Schwierigkeiten wegen des Fehlens eines halboffenen Suchgebiets umgangen, sondern es kann durch die Entartung zur Gleichung auch eine Variablenreduktion erfolgen. Ergänzend zu den klassischen Eliminationsmethoden werden die Elimination mittels scharfer Ungleichungen und mittels Zielfunktionserweiterung vorgestellt. Während das Anwenden der scharfen Ungleichungen (z. B. Cauchy-Schwarz-Ungleichung, Dreiecksungleichung) einerseits spezielle Problemstellungen und andererseits ein gewisses mathematisches Geschick erfordert, lässt sich eine Zielfunktion relativ formal um hochgewichtete Strafterme für die zu eliminierende Restriktion erweitern. Ob das allerdings einen Vorteil bringt, ist fallweise zu prüfen, da das Erweitern die Differenzierbarkeit, die Topologie oder die Kondition des Problems wesentlich ändern kann. Durch einfache Beispiele werden die grundlegenden Ideen und gegebenenfalls die Schwächen der Zugänge und Algorithmen aufgezeigt; vergleichende Studien oder programmtechnische Details bleiben der zitierten Spezialliteratur vorbehalten.

2. **Erweiterungsmethoden** (Kapitel 6)
   Erweiterungsmethoden erzeugen ein äquivalentes Problem mit einer größeren Anzahl von Variablen und/oder Restriktionen. Doch warum sollte eine solche Technik eingesetzt werden? Kurzum, weil sie Vereinfachungen liefern kann. Aus der Mechanik ist das Freischneiden bekannt, durch das sich ein System in Teilsysteme zerlegen lässt, die aber durch Zwangskräfte miteinander verbunden sind. Ähnlich ist die Idee, eine Zielfunktion $f(x) = x^2$ in $g(x, y) = x \cdot y$ mit der Restriktion $x = y$ umzuschreiben (Idee der Splitting-Methode). De facto die gleiche Idee kommt bei der Pseudodimensionserhöhung zur Anwendung, also mehr Variablen, die ihrerseits aber durch zusätzliche, meist nicht so einfache Restriktionen wie bei der Splitting-Methode gekoppelt sind. Bei der Schlupfvariablenmethode werden neue nichtnegative Variablen eingeführt, um aus einer Ungleichungsrestriktion eine Gleichungsrestriktion zu machen, während bei der Methode der Lagrange-Multiplikatoren die neuen Variablen (Lagrange-Faktoren) die Aufgabe haben, die Restriktionen in eine erweiterte Zielfunktion aufzunehmen. Die Epigraph-Methode geht den umgekehrten Weg: sie nutzt eine neue Variable, um die Zielfunktion in eine zusätzliche Restriktion umzuformen und erreicht dadurch, dass die neue Zielfunktion linear ist. Matrixfaktorisierungen hingegen helfen, komplizierte Rangrestriktionen, auf die ableitungsbasierte Zugänge nicht anwendbar sind, zu eliminieren. Sie nehmen dafür eine höhere Variablenanzahl in Kauf. All diese Methoden werden vorgestellt und erörtert, um vor allem ein Gefühl zu vermitteln, wann welche Methode eingesetzt kann. Insbesondere wird dabei Wert auf die vektor- und matrixvariate Formulierung und Lösung von Problemen gelegt. Hierfür wird auch ein neues Matrixableitungskalkül entwickelt, das sich zur quasifreien Optimierung eignet und für bestimmte Problemklassen eine Alternative zum Lagrange-Formalismus darstellt.

3. **Problemtransformationsmethoden** (Kapitel 7)
   Problemtransformationsmethoden erzeugen ein äquivalentes Problem, bei dem die Anzahl von Variablen und/oder Restriktionen gleich bleibt. Sie zielen auf Vereinfachungen, Änderungen in der Topologie des Gütekriteriums oder dienen der Anpassung einer Problemstellung an die verfügbare Software. Im Abschnitt zu den elementaren Umformungen wird insbesondere auf die monotonen Transformationen eingegangen, denn gerade durch Logarithmieren, fallweises Quadrieren oder Wurzelziehen ergeben sich Vereinfachungen, während das Skalieren von Zielfunktionen und/oder Restriktionen genutzt wird, um die numerischen Eigenschaften eines Problems positiv zu verändern. Beim Dekompositionszugang wird das Problem durch Zerlegung in kleinere Probleme vereinfacht, während bei Minimax-Problemen ein Ändern der Optimierungsreihenfolge zu einem Maximin-Problem eine Transformationsoption darstellen kann. Weniger am Problem direkt als vielmehr indirekt über die Parameter greifen die bijektiven (eingeschränkt auch surjektiven) Transformationen an. Durch sie werden die Probleme reparametrisiert, wodurch häufig eine für die Optimierung bessere Topologie erreichbar ist. Die Momentenmethode, bei der beispielsweise aus Mittelwert und Streuung die Parameter einer nichtgaußschen Verteilung bestimmt werden, nutzt diese Technik ebenso wie das Prony-Verfahren, das die Parameter von Schwingungen über Parameter von Differenzengleichungen berechnet. Neben dem Prony-Verfahren gibt es weitere Techniken, um parameternichtlineare Probleme in parameterlineare umzuformen. Solche Techniken fasst die Arbeit in einer Übersicht zusammen, wobei der thematischen Zuordnung wegen auch Zugänge aufgenommen werden, die eigentlich nicht als Problemtransformationen zu charakterisieren sind, sondern beispielsweise in die Problemerweiterung fallen.
   Einen Schwerpunkt des Kapitels bilden die Umformulierungen zur semidefiniten Optimierung, die in den letzten Jahren dank frei verfügbarer Software und einer Vielzahl neuer Anwendungen große Bedeutung in der Modellbildung erlangt hat. Aufbauend auf den theoretischen Grundlagen wird gezeigt, wie sich einige konvexe Probleme in semidefinite Probleme transformieren lassen. In diesem Zusammenhang ist die Umformulierung der Restriktionen in lineare Matrixungleichungen von zentraler Bedeutung. Dass dabei redundante Restriktionen entstehen oder wegfallen können, soll nichts an der Zuordnung ändern, da sie oft als versteckte (nicht sichtbare) Restriktionen auftreten. Zudem wird am Beispiel der semidefiniten Optimierung auch gezeigt, wie sich die in den Kapiteln zuvor aufgeführten Methoden zur weiteren Problemaufbereitung nutzen lassen, wobei dazu die Eliminationsmethode und die Epigraph-Methode besonders hilfreich sind.

4. **Problemmodifikationsmethoden** (Kapitel 8)
   Problemmodifikationsmethoden ändern ein Problem derart, dass ein einfacheres Problem entsteht, dessen Lösung die des Originalproblems nähert. Das Vereinfachen eines Problems geschieht dabei durch Weglassen (Relaxation) oder Hinzunehmen (Kontraktion) von Restriktionen. Beim Weglassen interessieren besonders jene Fälle, bei denen sich die Lösung gar nicht oder generisch nicht ändert oder bei denen iterative Algorithmen entworfen werden können, die gegen die Lösung des Originalproblems konvergieren. Durch Hinzunahme von Restriktionen lassen sich ebenfalls Probleme vereinfachen, wenn beispielsweise alle Variablen bis auf eine auf feste Werte gesetzt werden, da dann nur eine einvariable Optimierung auszuführen ist. Dadurch, dass dieses Vorgehen in einem nächsten Schritt für eine andere Variable getätigt wird, entstehen iterative Algorithmen. Der Wechsel zwischen den Variablen erfolgt dabei meist zyklisch, bei zwei Variablen also alternierend. Die beiden Techniken – Weglassen und Hinzunehmen von Restriktionen – können auch miteinander kombiniert werden. So erzeugt die Splitting-Methode zunächst ein erweitertes Problem, das durch Weglassen der Variablenzwangsbedingung mit alternierenden, einfacheren Teiloptimierungen iterativ gelöst werden kann. All die angesprochenen Ideen werden hinsichtlich ihres mathematischen Hintergrunds (ausgewählte Relaxationen, Lagrange-Dualität, Dualitätstheoreme) vorgestellt, bewertet und durch fünf Algorithmen untersetzt. Darüber hinaus werden im Kapitel Grundlagen zum Einsatz von Matrixapproximationsproblemen zur Problemmodifikation erklärt. Statt der optimalen Lösung eines Prokrustes-Problems (Matrix-LS-Problem mit Restriktionen an die Matrix) wird eine genäherte Lösung gewählt, die durch Projektion der LS-Lösung auf die zulässige, durch die Restriktionen bestimmte Menge entsteht. Hierbei ist die gewöhnliche LS-Lösung einfach zu berechnen und die Projektion ebenfalls, da für zahlreiche Fälle explizite Lösungen existieren oder auf ein- und zweidimensionale Suchen reduzierte Algorithmen verfügbar sind. Matrixapproximationsprobleme liefern zudem Maße (häufig Radien genannt) zur quantitativen Bewertung dafür, wie stark eine Eigenschaft ausgeprägt ist. Im ersten Teil der Arbeit werden sie unter anderem für Kovarianzmatrizen, Systemmatrizen und auch Matrixtupel formuliert und mit Literaturhinweisen zu ihrer Lösung hinterlegt.

Um den Umfang trotz eines breiten Spektrums an potenziellem Vorwissen und Methoden angemessen zu halten, wird ein eher skizzenhafter Stil verwendet. Die einzelnen Abschnitte sind dabei weitgehend separat lesbar, wobei Fußnoten für nicht so gängige Begriffe oder Formelzusammenhänge ein unnötiges Nachschlagen verhindern sollen. Auf die statistische Aspekte der restringierten Parameterschätzung musste verzichtet werden, s. hierfür [207], [530] (Likelihood-Verhältnistest, Wald-Test, Lagrange-Multiplikator-Test) und für den Nachweis einer bessere Statistik restringierter Schätzer gegenüber freien Schätzern [441], [233], [423].

Ergänzend zu den beiden Teilen der Arbeit werden in Kapitel 9 Empfehlungen für die Identifikation gegeben. Diese richten sich an Neueinsteiger auf dem Gebiet, da sie weitgehend ohne Mathematik auskommen. Gleichzeitig sind die Empfehlungen für die Lehre hilfreich, da sie bestimmte Aspekte durch Thesen, Tabellen und Aufzählungen kompakt zusammenfassen. Letztendlich fließen neben den langjährigen Erfahrungen des Autors auf dem Gebiet der Modellbildung die Ergebnisse der vorliegenden Arbeit mit ein. In diesem Sinn kann das Kapitel 9 als eine Erweiterung der Zusammenfassung in Kapitel 10 betrachtet werden.

Kapitel 2

Restriktionen an dynamische Modelle

Dynamische Modelle beschreiben bekanntermaßen die zeitlichen Aspekte von Systemen. Ihr Anwendungsspektrum reicht von Simulationen zur Prozessverhaltensanalyse oder zur Optimierung über Filtern, die bestimmte Frequenzbereiche unterdrücken, bis hin zur Entwurfsgrundlage für den modellbasierten Reglerentwurf. Das Erstellen der Modelle auf der Basis experimenteller Daten wird dabei gemeinhin als Identifikation bezeichnet. Immer höhere Anforderungen vonseiten der Anwendungen erfordern dabei immer bessere Modelle, wobei sich das Besser auf genauere und weniger streuende Parameter, einen größeren Gültigkeitsbereich, aber vor allem auf ein mit dem A-priori-Wissen konformes Modell bezieht. Solches A-priori-Wissen umfasst Kenntnisse über die Lage einzelner Pole, Wurzeln oder Eigenwerte, aber auch Systemeigenschaften wie Stabilität, Minimalphasigkeit oder Passivität. Die Herausforderung für den Ingenieur besteht nun darin, das A-priori-Wissen zusammenzutragen, um es dann mathematisch zu formulieren. Deshalb werden in jedem Abschnitt zunächst die Eigenschaften definiert und mit Sätzen hinterlegt, aus denen Formeln oder strukturelle Maßnahmen zur Einhaltung der Eigenschaften abgeleitet werden können. Einfache Beispiele verdeutlichen die Grundgedanken, die durch Literaturhinweise zu Anwendungen und Algorithmen ergänzt werden, während Gegenbeispiele auf Trugschlüsse hinweisen. Bedingt durch die hohe Anzahl möglicher Eigenschaften, die es je nach Modelltyp (Zustandsraum-, E/A-Formulierung; linear, nichtlinear) und Zeitbereichstyp (kontinuierlich, zeitdiskret) umzusetzen gilt, ist dieses Kapitel das umfangreichste in der vorliegenden Arbeit. Durch Schwerpunktsetzung auf die Klasse der linearen zeitinvarianten Systeme[1], kurz LTI-Systeme, und Verzicht auf Komplexbeispiele wird der Umfang beschränkt, wodurch die Betrachtung möglichst vieler Eigenschaften gelingt. Damit wird die in der Regelungstechnik bedeutendste Systemklasse abgedeckt, wobei Hinweise auf Besonderheiten, weiterführende Literatur zu nichtlinearen und zeitdiskreten Systeme ebenfalls gegeben werden. Als viel wichtiger als das Auffinden

[1] Der Begriff wird hier einschränkend auf endlichdimensionale Systeme (Systeme mit konzentrierten Parametern) verwendet.

einer Lösung für ein konkretes Anwendungsbeispiel wird in den folgenden Abschnitten das Anwenden prinzipieller Herangehensweisen angesehen. Als erstes stellt sich die Frage, ob es sich um eine generische oder nichtgenerische Eigenschaft handelt. Generische Eigenschaften brauchen nämlich nicht durch Maßnahmen gesichert werden, da die Wahrscheinlichkeit Null ist, ein Modell zu bestimmen, das die Eigenschaft nicht hat. Steuerbarkeit und Beobachtbarkeit sind generische Eigenschaften, Stabilität ist dagegen keine. Muss eine Eigenschaft gesichert werden, so kann das durch Restriktionen erfolgen oder durch Wahl einer speziellen Modellstruktur. Im ersten Fall stellt sich die Frage, nach einer gut handhabbaren Restriktion, im zweiten Fall die Frage, ob durch die Struktur alle möglichen Modelle einer Klasse erfasst werden, oder ob sie nur hinreichend ist. Ebenso ist die Frage zu beantworten, ob die Eigenschaft a priori oder a posteriori gesichert werden soll. Der A-priori-Weg ist dabei tendenziell zu bevorzugen, da er die optimale Lösung liefert (vorausgesetzt algorithmisch werden Nebenminima vermieden). Demgegenüber liefert der A-posteriori-Weg im Allgemeinen nur suboptimale Lösungen, da dabei das Problem zunächst ohne die betreffende Eigenschaftsforderung ermittelt wird und erst im Anschluss eine Projektion in die betreffende Modellklasse die Eigenschaft sicherstellt. Dafür ist das Realisieren der Projektion rechentechnisch oft einfacher, wodurch sich dieser Zugang insbesondere für Online-Anwendungen empfiehlt. Einen ganz anderen Aspekt, der in den einzelnen Abschnitten wiederholt betrachtet wird, stellen die sogenannten Eigenschaftsmaße (Eigenschaftsradien) dar. Sie werden herangezogen, um quantitativ zu bewerten, wie ausgeprägt eine Eigenschaft ist, wie gut steuerbar etwa ein System ist. All diese Fragen und Aspekte sind während des Modellbildungsprozesses ständig neu zu beantworten. Die Vielschichtigkeit resultiert dabei aus der Tatsache, dass selbst für die systemtheoretisch gut verstandene Klasse der LTI-Systeme, die Eigenschaften sehr oft auf nichtkonvexe und nichtlineare Optimierungsprobleme führen. Somit gibt es keine prädestinierten Zugänge und Algorithmen, was die Behandlung schwieriger aber auch spannender macht.

Um dem Leser einen zielgerichteten Zugang für die Umsetzung des A-priori-Wissens über dynamische Systeme zu vermitteln, sind die Abschnitte nach den betreffenden Eigenschaften geordnet. Die Abschnitte lassen sich dabei im Wesentlichen getrennt voneinander lesen.

Im Abschnitt 2.1 wird die Kenntnis des statischen Verhaltens (bekannte statische Verstärkung bei LTI-Systemen) genutzt, um Restriktionen zu formulieren. Im Fall der nichtlinearen Systeme ergeben die sogenannten Ruhelagenbedingungen (Ableitungen sind Null im Kontinuierlichen bzw. zeitverschobene Werte sind gleich im Zeitdiskreten) Einschränkungen an die Modellparameter. Die Ruhelagenbedingungen sind dank der Gleichwertmessung sehr präzise und reduzieren die Freiheitsgrade bei der Identifikation nichtlinearer dynamischer Systeme signifikant, insbesondere dann, wenn an mehreren Arbeitspunkten experimentiert wird. Das Vorgehen wird im Rahmen der Mehrfach-Linearisierungsmethode beschrieben.

Gewissermaßen als Einstieg für die Umsetzung von Restriktionen an Systemeigenschaften wird in Abschnitt 2.2 gezeigt, wie Restriktionen an Wurzeln, Eigenwerte und Pole behandelt werden können. Die Restriktion, wonach Wurzeln, Eigenwerte bzw. Pole in der linken offenen komplexen Halbebene liegen sollen, entspricht nämlich gerade der „Stabilitätsrestriktion“ für LTI-Systeme. Diese Restriktion ist symptomatisch für die angesprochenen Schwierigkeiten, da die Zusammenhänge von den Polynomkoeffizienten zu den Wurzeln nichtlinear und nichtkonvex sind. Neben der linken komplexen Halbebene als Restriktion werden in den Unterabschnitten zu Abschnitt 2.2 auch andere Mengen untersucht, die ihrerseits aus anderen Anwendungen oder Überlegungen resultieren.

Es folgen danach zahlreiche Abschnitte, die sich mit der Umsetzung von Stabilitätsrestriktionen befassen, wobei die Einteilung nach den Systemklassen erfolgt, also
- lineare zeitkontinuierliche Systeme in E/A-Darstellung (Abschnitt 2.3)
- lineare zeitdiskrete Systeme in E/A-Darstellung (Abschnitt 2.4)
- lineare zeitkontinuierliche Systeme in Zustandsraumdarstellung (Abschnitt 2.5)
- lineare zeitdiskrete Systeme in Zustandsraumdarstellung (Abschnitt 2.6)
- lineare zeitkontinuierliche Intervallsysteme (Abschnitt 2.7)
- lineare zeitvariante Systeme (Abschnitt 2.8)
- nichtlineare Systeme (Abschnitt 2.9).

Nach der Systemeigenschaft „Stabilität“, die den Schwerpunkt des Kapitels bildet, werden Eigenschaften dahingehend untersucht, wie das Vorwissen zu den Eigenschaften bei der Identifikation umgesetzt werden kann. Die Klasse der LTI-Systeme steht dabei im Vordergrund. Ergänzende Hinweise oder teils eigene Unterabschnitte zu den linearen zeitvarianten und nichtlinearen zeitinvarianten Systemen vervollständigen die Abschnitte. Zu den hier betrachteten Eigenschaften zählen:
- Passivität (Abschnitt 2.10)
- Externe Positivität (Abschnitt 2.11)
- Steuerbarkeit und Beobachtbarkeit (Abschnitt 2.12)
- Minimalphasigkeit (Abschnitt 2.13)
- Kausalität (Abschnitt 2.14).

Den Abschluss des Kapitels bildet Abschnitt 2.15 zum Thema Bifurkationen. Hierbei handelt es sich um parameterbedingte qualitative Änderungen im Systemverhalten. Beobachtet der Anwender eine solche Verhaltensänderungen, kann er diese einem Bifurkationstyp zuordnen. Dadurch schränkt es wiederum seine Modellklasse ein. Die Kenntnis des Bifurkationstyps ist für die Identifikation nämlich eine wichtige Strukturinformation, aus der sich Einschränkungen bezüglich der Eigenwerte der Ruhelagenlinearisierungen ergeben oder aus der auf das Vorhandensein bestimmter nichtlinearer Terme im Vektorfeld bzw. einer Approximation des Vektorfelds geschlossen werden kann.

2.1 Restriktionen bezüglich des statischen Verhaltens

Das Einbeziehen einer a priori bekannten statischen Verstärkung, die z. B. durch ein Sprungexperiment sehr genau bestimmt werden kann, hilft die Parametervarianzen zu senken und gegebenenfalls den lokalen Gültigkeitsbereich nichtlinearer Modelle zu erweitern [144].
Neben dem experimentellen Zugang leitet sich A-priori-Information über die Statik auch aus Bilanzgleichungen ab. So müssen beispielsweise die E/A-Massen- bzw. -Molströme über alle Elemente (Kohlen-, Wasser-, Sauerstoff) der Reaktanden (Methanol, Wasser, Kohlenstoffdioxid) passen. Solche Zwangsbedingungen werden etwa in [16] in einen empirischen multivariaten polynomialen Modellansatz eingearbeitet.

In den nächsten zwei Abschnitten wird davon ausgegangen, dass die Statikinformation als Verstärkung, Punktpaare oder Kennlinie verfügbar ist. Dieses Wissen wird in Restriktionen umformuliert. Im Abschnitt 2.1.3 wird dann eine Technik vorgestellt, die Statik- und Linearisierungsinformation nutzt, um auf die Parameter eines nichtlinearen Modells zu schließen.

2.1.1 Statikrestriktionen bei linearen Modellen

Bei stabilen[2] LTI-Systemen gilt für die statische Verstärkung $K = G(s)|_{s=0}$. Speziell für die Systeme mit einem Eingang und einem Ausgang, also die sog. SISO-Systeme, ergibt sich bei der Standardparametrisierung die Restriktion $K = b_0/a_0$. Ist zudem $a_0 = 1$, dann gilt $b_0 = K$. Diese Restriktion wird unmittelbar zur Elimination des Parameters b_0 aus dem Identifikationsproblem genutzt (Ersetzen von b_0 durch K).
Bei stabilen zeitdiskreten Systemen gilt für die statische Verstärkung $K = G_z(z)|_{z=1}$, woraus sich folgende lineare Restriktion ableitet

$$b_0 + \ldots + b_m = K(a_0 + \ldots + a_n). \tag{2.1}$$

Für ein FIR-Filter (Filter mit endlicher Impulsantwort)

$$y[k] = \sum_{i=0}^{m} b_i u[k-i] \tag{2.2}$$

bedeutet das, dass für die statischen Verstärkung K gelten muss $K = \sum_{i=0}^{m} b_i$. Demzufolge ist für ein Gewichtsfolgemodell $\{g[k]\}_{k=0}^{\infty}$

$$\sum_{k=0}^{l} g[k] = K \tag{2.3}$$

zu fordern, wobei l die Approximationslänge für die Gewichtsfolge angibt.

[2] Instabile Systeme haben keine statischen Kennlinien, wohl aber eine Ruhelagenmannigfaltigkeit.

Selbstverständlich lassen sich all diese Restriktionen auch als Ungleichungen formulieren, wenn sich die Kenntnis der statischen Verstärkung nur auf Intervalle $K_{\min} \leq K \leq K_{\max}$ oder Vorzeichen $K \geq 0$ bzw. $K \leq 0$ beschränkt. So wird (2.1) bei monischem[3] Nenner beispielsweise zu[4]

$$b_0 + \ldots + b_m - K_{\max}(a_0 + \ldots + a_{n-1} + 1) \leq 0 \tag{2.4a}$$

$$-b_0 - \ldots - b_m + K_{\min}(a_0 + \ldots + a_{n-1} + 1) \leq 0. \tag{2.4b}$$

Das Einbeziehen der Vorzeichenkenntnis und damit das Garantieren eines bestimmten Vorzeichens der statischen Verstärkung ergibt sich in vielen adaptiven Regelungssystemen [367] allein schon aus Stabilitätsforderungen. Dank der Konvexität der Statikrestriktionen wird deren Einhalten bei den adaptiven Reglern üblicherweise per Projektion gelöst. Im einfachsten Fall sichert $\text{Proj}_{K\geq 0}\{K\} = \max\{K, 0\}$ die Nichtnegativität von K. Der Praktiker gibt noch etwas Reserve $\text{Proj}_{K\geq \varepsilon}\{K\} = \max\{K, \varepsilon\}$ dazu.

Anmerkung 2.1 Bei approximierenden LTI-Systemen ist die Vorgabe

$$K = \frac{y(u_{\max}) - y(u_{\min})}{u_{\max} - u_{\min}} \qquad \text{Sekantenapproximation von } K \tag{2.5}$$

im Arbeitsbereich zwischen unterem $(u_{\min}, y(u_{\min}))$ und oberen Arbeitspunkt $(u_{\max}, y(u_{\max}))$ zweckmäßig. Liegt genauere Modellinformation vor, kann auch mit der Tangentenlinearisierung operiert werden.

Beispiel 2.1 (Statische Verstärkungsinformation aus der Linearisierung)
Der Füllstand eines Behälters mit Auslauf genügt $A\frac{\mathrm{d}h}{\mathrm{d}t} + \alpha\sqrt{h} = \dot{q}$ mit der Höhe h, der Fläche A, dem Auslaufkoeffizienten α und dem Zufluss $\dot{q}$. Das im Arbeitspunkt $(\dot{q}_s, h_s)$ per Linearisierung erhaltene PT1-Glied hat die Verstärkung $K = 2\sqrt{h_s}/\alpha$ und die Zeitkonstante $2\sqrt{h_s}A/\alpha$. Mit i. Allg. bekanntem A, bekanntem Arbeitspunkt und zumindest intervallweise bekanntem α ergeben sich so arbeitspunktabhängige Verstärkungsintervalle.

Anmerkung 2.2 Werden keine linearen, sondern lediglich linearisierte Modelle identifiziert, ist eine Gleichwertbehandlung erforderlich, da die Verstärkung, aber auch die anderen Parameter sonst gänzlich falsch bestimmt werden. Zur Gleichwertbehandlung bieten sich an:
- Abziehen des Arbeitspunkts (u_s, y_s), d. h. $u[k] := u[k] - u_s$ und $y[k] := y[k] - y_s$
- Abziehen der Mittelwerte, d. h. $u[k] := u[k] - \bar{u}$ und $y[k] := y[k] - \bar{y}$ /[5]

[3] Ein Polynom heißt monisch, wenn der zur höchsten Potenz gehörende Koeffizient $a_n = 1$ ist.

[4] Für stabile Systeme gilt $a_0 + \ldots + a_{n-1} + 1 \geq 0$, weshalb beim Herleiten keine Vorzeichenumkehr auftritt. Betrachte $\prod_{i=1}^{n}(z - z_i)|_{z=1} = \prod_{i=1}^{n}(1 - z_i) \geq 0$ wegen $|z_i| \leq 1$.

[5] Diese Methode ist selbst bei mittelwertfreier Anregung um den Arbeitspunkt nicht identisch mit der vorgenannten, da es bedingt durch die Nichtlinearitäten zu einer Gleichwertverschiebung am Ausgang kommt.

- skalierte und zentrierte Signale, d. h. $u[k] := \frac{u[k]-\bar{u}}{\hat{\sigma}_u}$ und $y[k] := \frac{y[k]-\bar{y}}{\hat{\sigma}_y}$
  a posteriori ist die Verstärkung zu korrigieren: $G(z) := \frac{\hat{\sigma}_y}{\hat{\sigma}_u} G(z;\hat{\theta})$
- Hochpassfilterung (HP) der Signale, z. B. $u[k] := HP\{u[k]\}$ und $y[k] := HP\{y[k]\}$
- Einsatz inkrementaler Signale, d. h. $u[k] := u[k] - u[k-d]$ und $y[k] := y[k] - y[k-d]$
- Schätzung des aggregierten Gleichwerts mit der Eins-Spalten-Methode[6],

wobei die Formeln für Signale denen der Folgen analog sind.

Eine Anwendung zur Nutzung der Statikinformation bei der Identifikation von Systemen mit mehreren Eingängen und einem Ausgang, also sog. MISO-Systemen, zeigt das nachfolgende Beispiel. Es verdeutlicht, wie die Statikrestriktion aus der MISO-Modellschätzung für die Reduktion auf die SISO-Teilmodelle genutzt werden kann und wie sich Vorwissen über die SISO-Teilverstärkungen in der MISO-Schätzung einbeziehen lässt. Das Vorgehen ist dabei formal auf Systeme mit mehr als zwei Eingängen erweiterbar.

Beispiel 2.2 (Statikinformation bei MISO-Gleichungsfehlermodellen)
Das System

$$Y(z) = \frac{b_1(z)}{a_1(z)} U_1(z) + \frac{b_2(z)}{a_2(z)} U_2(z) \tag{2.6}$$

mit den Graden (m_1, n_1) bzw. (m_2, n_2) kann durch Bilden des Hauptnenners als

$$a_1(\mathrm{q})a_2(\mathrm{q})y[k] = b_1(\mathrm{q})a_2(\mathrm{q})u_1[k] + b_2(\mathrm{q})a_1(\mathrm{q})u_2[k] \tag{2.7a}$$

$$\tilde{a}(\mathrm{q})y[k] = \tilde{b}_1(\mathrm{q})u_1[k] + \tilde{b}_2(\mathrm{q})u_2[k] \tag{2.7b}$$

mit q als Rechtsverschiebeoperator geschrieben werden. Hierfür lautet ein zugehöriges Gleichungsfehlermodell

$$\sum_{i=0}^{n_1+n_2} \tilde{a}_i y[k+i] = \sum_{i=0}^{n_2+m_1} \tilde{b}_{1,i} u_1[k+i] + \sum_{i=0}^{n_1+m_2} \tilde{b}_{2,i} u_2[k+i] + \varepsilon[k]; \qquad k = 0, 1, 2, \ldots \tag{2.8}$$

Aus den Schätzungen zu (2.8) leiten sich zwei Statikrestriktionen für eine nachgeschaltete Bestimmung der Koeffizienten für (2.6) ab

$$\frac{\sum_{i=0}^{n_2+m_1} \hat{\tilde{b}}_{1,i}}{\sum_{i=0}^{n_1+n_2} \hat{\tilde{a}}_i} = \underbrace{\hat{K}_1 = \frac{\sum_{i=0}^{m_1} b_{1,i}}{\sum_{i=0}^{n_1} a_{1,i}}}_{\text{erste Restriktion}} \qquad \text{und} \qquad \frac{\sum_{i=0}^{n_1+m_2} \hat{\tilde{b}}_{2,i}}{\sum_{i=0}^{n_1+n_2} \hat{\tilde{a}}_i} = \underbrace{\hat{K}_2 = \frac{\sum_{i=0}^{m_2} b_{2,i}}{\sum_{i=0}^{n_2} a_{2,i}}}_{\text{zweite Restriktion}}, \tag{2.9}$$

wobei die Restriktionen zweckmäßigerweise in lineare Gleichungsrestriktionen gemäß (2.1) umgeformt werden. Die Beziehungen in (2.9) können aber auch verwendet werden, wenn

[6] Bei dieser Methode wird der Regressionsansatz um einen konstanten Parameter (aggregierten Gleichwert aus u_s und y_s) erweitert. Dieser konstante Parameter führt in der Datenmatrix zu einer Eins-Spalte (Namensgebung).

Vorwissen über die SISO-Teilverstärkungen vorliegt. Dann ergeben sich durch Umformen der jeweils linken Gleichungen in (2.9) zwei lineare Gleichungsrestriktionen, die sich in die MISO-Schätzung einarbeiten lassen.

Statikrestriktionen spielen auch bei der Modellapproximation, speziell der Modellreduktion, eine wichtige Rolle:

$$\text{Statik des Modells} \quad = \quad \text{Statik des reduzierten Modells.}$$

In diesem Fall kann die erforderliche Statikinformation direkt aus dem Modell mit den beschriebenen Techniken aus der Zeit- bzw. Frequenzbereichsdarstellung gewonnen werden.

Abschließend sei nochmals darauf verwiesen, dass die Statikrestriktion das Schätzproblem um einen Freiheitsgrad verringert und somit auf bessere statistische Eigenschaften führt (Informationsverdichtung auf weniger Parameter). Da Statikrestriktionen mathematisch zudem durch lineare Gleichungen beschrieben werden, ist der erforderliche Mehraufwand für ihre Einbeziehung gering.

2.1.2 Statikrestriktionen für parameterlineare nichtlineare Modelle

Neben der Verstärkungsrestriktion zählt auch die Ruhelagenrestriktion zu den Statikrestriktionen. Vielfach sind Ruhelagen $(u_s, y_s)_i$ aus Vorexperimenten bekannt, mit denen das dynamische Modell verträglich sein sollte. Durch Nullsetzen der Ableitungen im kontinuierlichen Fall bzw. Gleichsetzen aller Verschiebungen im diskreten Fall werden die Ruhelagenrestriktionen erzeugt. Für parameterlineare SISO-NARX-Modelle[7]

$$y[k] = \psi^T[k]\theta + \varepsilon[k], \qquad \theta \in \mathbb{R}^r, \tag{2.10}$$

wobei $\psi^T[k]$ der aus Funktionen $\psi_i(y[k-1], \ldots, y[k-n_y], u[k-d], \ldots, u[k-d-n_u+1])$; $i = 1, \ldots, r$ gebildete Regressor ist, ergibt sich die statische Beziehung durch Setzen von $y_s = y[k] = y[k-1] = \ldots = y[k-n_y]$ und $u_s = u[k-d] = \ldots = u[k-d-n_u+1]$. Die ψ_i werden hierdurch zu Funktionen $\bar{\psi}_i(u_s, y_s)$, womit für jedes statische Paar (u_s, y_s), also für jeden Punkt auf der statischen Kennlinie oder für jede isolierte stabile Ruhelage, eine lineare Restriktion für $\theta = (\theta_1, \ldots, \theta_r)^T$ folgt

$$y_s = \bar{\psi}_1(u_s, y_s)\theta_1 + \ldots + \bar{\psi}_r(u_s, y_s)\theta_r. \tag{2.11}$$

[7] NARX steht für „nonlinear autoregressive exogenous“, also für nichtlineare autoregressive Modelle mit äußerer Anregung (mit Eingang).

Statt der Gleichheitsrestriktionen sollten bei gemessenen (u_s, y_s) Restriktionen der Art

$$-\delta \leq y_s - \bar{\psi}_1(u_s, y_s)\theta_1 + \ldots + \bar{\psi}_r(u_s, y_s)\theta_r \leq \delta \tag{2.12}$$

verwendet werden, wobei δ in der Größenordnung der Messfehler zu wählen ist. Liegt nicht nur punktweise Kenntnis über den statischen Zusammenhang vor, sondern sogar die gesamte statische Kennlinie $y_s = h(u_s)$, führt dies auf

$$h(u_s) = \bar{\psi}_1(u_s, h(u_s))\theta_1 + \ldots + \bar{\psi}_r(u_s, h(u_s))\theta_r \qquad \forall u_s. \tag{2.13}$$

Bei Polynomen $y_s = h(u_s)$ und polynomialen Regressorkomponenten ψ_i folgen lineare Restriktionen an θ durch Koeffizientenvergleich bezüglich der Potenzen u_s^l. Da (2.13) für alle u_s gelten muss, liefert jedes frei gewählte u_s eine neue lineare Gleichungsrestriktion. Sollten die Gleichungen inkonsistent sein, ist der nichtlineare dynamische Ansatz nicht mit der Statik verträglich. Im Regelfall wird nur ein konsistentes unterbestimmtes System von Restriktionen entstehen.

2.1.3 Mehrfach-Linearisierungsmethode

Die Mehrfach-Linearisierungsmethode eignet sich zur Identifikation nichtlineare Zustandsraummodelle, die parameterlinear sind. Sie beruht darauf, an verschiedenen Ruhelagen (Arbeitspunkten) lineare Modelle zu schätzen und deren Parameter zur Bestimmung der unbekannten Parameter des nichtlinearen Modells heranzuziehen. Somit können die ausgefeilten Verfahren zur Schätzung linearer Modelle eingesetzt werden.

Sei ein Zustandsraummodell mit messbaren Zustandsvektoren gegeben

$$\dot{x}_i = \theta_i^T f_i(x_1, \ldots, x_n, u_1, \ldots, u_m); \qquad i = 1, \ldots, n; \theta_i \in \mathbb{R}^{n_i} \tag{2.14a}$$

$$y_i = \psi_i^T h_i(x_1, \ldots, x_n, u_1, \ldots, u_m); \qquad i = 1, \ldots, p; \psi_i \in \mathbb{R}^{p_i} \tag{2.14b}$$

und sei (u_s, x_s, y_s) ein Tupel einer stabilen Ruhelage, dann resultieren aus der Ruhelagenbedingung $x(t) \equiv x_s$, d. h. $\dot{x} = 0_n$, die Restriktionen

$$0 = \theta_i^T f_i(x_s, u_s); \;\; i = 1, \ldots, n \tag{2.15a}$$

$$y_{si} = \psi_i^T h_i(x_s, u_s); \; i = 1, \ldots, p. \tag{2.15b}$$

An diesen Ruhelagen kann (2.14) linearisiert werden

$$\Delta\dot{x}_i = \theta_i^T \Big(\sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j}(x_s, u_s)\Delta x_j + \sum_{j=1}^{m} \frac{\partial f_i}{\partial u_j}(x_s, u_s)\Delta u_j \Big); \qquad i = 1, \ldots, n \tag{2.16a}$$

$$\Delta y_i = \psi^T \Big(\sum_{j=1}^{n} \frac{\partial h_i}{\partial x_j}(x_s, u_s)\Delta x_j + \sum_{j=1}^{m} \frac{\partial h_i}{\partial u_j}(x_s, u_s)\Delta u_j \Big); \qquad i = 1, \ldots, p. \tag{2.16b}$$

Ferner wird ein für (u_s, x_s, y_s) linearisiertes Modell geschätzt

$$\Delta\dot{x}_i = \hat{A}\Delta x + \hat{B}\Delta u \qquad \hat{A} = ((\hat{a}_{ij})) \in \mathbb{R}^{n\times n}, \hat{B} = ((\hat{b}_{ij})) \in \mathbb{R}^{n\times m} \tag{2.17a}$$

$$\Delta y = \hat{C}\Delta x + \hat{D}\Delta u \qquad \hat{C} = ((\hat{c}_{ij})) \in \mathbb{R}^{p\times n}, \hat{D} = ((\hat{d}_{ij})) \in \mathbb{R}^{p\times m}. \tag{2.17b}$$

Per Koeffizientenvergleich gilt dann

$$\hat{a}_{ij} \approx \theta_i^T \frac{\partial f_i}{\partial x_j}(x_s, u_s),\ \hat{b}_{ij} \approx \theta_i^T \frac{\partial f_i}{\partial u_j}(x_s, u_s),\ \hat{c}_{ij} \approx \psi_i^T \frac{\partial h_i}{\partial x_j}(x_s, u_s),\ \hat{d}_{ij} \approx \psi_i^T \frac{\partial h_i}{\partial u_j}(x_s, u_s). \tag{2.18}$$

Hieraus leitet sich z. B. ein LS-Schätzproblem für die Parameter der i-ten Zustandsgleichung ab, wobei die Koeffizientenbeziehungen einem Kleinsten-Fehlerquadratsummen-Ausgleich (LS für „least squares“) unterzogen werden, bei dem die Ruhelagenbedingung als Restriktion fungiert:

$$\left\| \begin{bmatrix} \frac{\partial f_i^T}{\partial x_1}(x_s, u_s) \\ \vdots \\ \frac{\partial f_i^T}{\partial x_n}(x_s, u_s) \\ \frac{\partial f_i^T}{\partial u_1}(x_s, u_s) \\ \vdots \\ \frac{\partial f_i^T}{\partial u_m}(x_s, u_s) \end{bmatrix} \theta_i - \begin{bmatrix} \hat{a}_{i1} \\ \vdots \\ \hat{a}_{in} \\ \hat{b}_{i1} \\ \vdots \\ \hat{b}_{im} \end{bmatrix} \right\|_2^2 \stackrel{!}{=} \text{Min} \qquad 0 = \theta_i^T f_i(x_s, u_s). \tag{2.19}$$

Analoges gilt für die ψ_i. Sollte die Datenmatrix keinen vollen Rang haben, ist sie durch Datenblöcke aus der Identifikation an anderen Arbeitspunkten zu ergänzen; daher der Vorsatz „Mehrfach“. Für jede neue Ruhelage kommen dabei neue Restriktionen hinzu, was zu Widersprüchen in den Restriktionen führen kann – weniger bedingt durch eine fehlerhafte Messung der Ruhelagen als vielmehr durch das gemeinhin approximative Verhalten des zu identifizierenden nichtlinearen Modells (z. B. Bilinearmodell als Approximation). Deshalb empfiehlt es sich, die Restriktionen als einen Block in die Datenmatrix und damit in den Fehlerausgleich aufzunehmen, wie es beim WLSE-Grenzwertverfahren in Abschnitt 5.4 der Fall ist.

Sollte ein Parameter in mehreren Gleichungen auftreten, besteht also beispielsweise die zusätzliche Restriktion $\theta_{13} = \theta_{21}$, so kann eine solche Restriktion durch Blockmatrixformulierung eliminiert werden, vgl. Abschn. 5.2.3.

In Analogie zum beschriebenen Vorgehen kann die Mehrfach-Linearisierungsmethode auch auf nichtlineare E/A-Differenzialgleichungen übertragen werden. Ein Koeffizientenvergleich zwischen der Ruhelagenlinearisierung und der geschätzten linearen Differenzialgleichung in den Delta-Größen liefert dann die für den LS-Ausgleich erforderlichen Beziehungen mit der Ruhelagenbedingung als Restriktion.

Eine Anwendung dieser Methode auf einen nichtlinearen Schwinger und einen elastischen Roboterarm, die durch zustandsquadratische Modelle approximiert werden, wird in [70] beschrieben. Das folgende Beispiel soll das Vorgehen skizzieren.

Beispiel 2.3 (Linearisierungsmethode für ein bilineares Modell)
Gegeben sei $\dot{x} = \theta_1 x + \theta_2 xu + \theta_3 u$. Das linearisierte Modell für die Ruhelage (u_s, x_s) lautet

$$\begin{aligned}\Delta\dot{x} &= \theta_1\Delta x + \theta_2 u_s\Delta x + \theta_2 x_s\Delta u + \theta_3\Delta u \\ &= (\theta_1 + \theta_2 u_s)\Delta x + (\theta_2 x_s + \theta_3)\Delta u = \hat{a}\Delta x + \hat{b}\Delta u.\end{aligned}$$

Unter Einbeziehung der Gleichwertrestriktion reicht sogar die Experimentation an einem Arbeitspunkt aus, um alle drei Parameter zu bestimmen

$$\left\| \begin{bmatrix} 1 & u_s & 0 \\ 0 & x_s & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} - \begin{bmatrix} \hat{a} \\ \hat{b} \\ 0 \end{bmatrix} \right\|_2^2 \stackrel{!}{=} \text{Min} \qquad 0 = [\theta_1, \theta_2, \theta_3] \begin{bmatrix} x_s \\ x_s u_s \\ u_s \end{bmatrix}. \tag{2.20}$$

Die alleinige Verwendung von drei oder auch mehreren Gleichwertbeziehungen gemäß

$$\begin{bmatrix} x_{s1} & x_{s1}u_{s1} & u_{s1} \\ x_{s2} & x_{s2}u_{s2} & u_{s2} \\ x_{s3} & x_{s3}u_{s3} & u_{s3} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \tag{2.21}$$

reicht nicht aus, um die Parameter zu bestimmen. Im Fall einer regulären Matrix ergäbe sich nämlich nur die Triviallösung $\theta = 0_3$, im Fall einer singulären Matrix ist θ dagegen nicht mehr eindeutig bestimmt.

2.2 Restriktionen an Wurzeln, Eigenwerte und Pole

Bekanntermaßen wird die Dynamik von LTI-Systemen wesentlich durch die Eigenwerte der Systemmatrix, Wurzeln von Polynomen bzw. Pole von Übertragungsfunktionen bestimmt. Während die Wurzeln der Nennerpolynome das asymptotische Verhalten ausdrücken, haben die Wurzeln der Zählerpolynome Einfluss auf das transiente Verhalten und charakterisieren zudem die Nulldynamik. Das Vorwissen entstammt physikalischen Überlegungen (Integratoren, Differenzierern), folgt aus Schwingungsmoden (z. B. Netzfrequenz, bekannte Eigenfrequenz oder Störfrequenz) oder ergibt sich aus der Verknüpfung einer Strecke mit bekannten Elementen (Regler, Stellglied, Messfilter). Darüber hinaus kann A-priori-Wissen aus dem Abtasttheorem gefolgert werden. Losgelöst von den Quellen des A-priori-Wissens liegt der Schwerpunkt der folgenden, separat lesbaren Unterabschnitte auf dem Formulieren und algorithmischen Umsetzen der Einschränkungen zu Eigenwerten, Wurzeln und Polen entweder explizit durch Restriktionen oder implizit durch strukturelle Ansätze.

2.2.1 Integrator, Differenzierer

Auf das Vorhandensein von Integratoren deuten vielfach bereits physikalische Größen, wie Weg, Geschwindigkeit, Energie, Füllstand, ohne dass dabei eine detaillierte theoretische Analyse notwendig ist. Experimentell äußert sich ein Integrator in offenen Strukturen durch einen nicht endenden Anstieg oder Abfall der Ausgangsgröße bei einer sprungförmigen Änderung der Eingangsgröße. Dieser Anstieg braucht keineswegs linear sein, da eventuell Nichtlinearitäten wirken. Außerdem wird er in praktischen Fällen häufig beschränkt sein. Das macht es schwierig, einen Tiefpass mit großer Zeitkonstante und hoher Verstärkung von einem Integrator zu unterscheiden.
Ein Indiz für Integratoren in Regelkreisen ist der Übergang auf Arbeitspunkte ohne bleibende Regelabweichung. Eine rampenförmige Anregung kann dabei weiteren Aufschluss geben, ob mehrere Integratoren im Regelkreis auftreten.

Im zeitkontinuierlichen LTI-Modell entspricht ein Integrator einer Wurzel des Nennerpolynoms $a(s)$ bei $s = 0$ oder äquivalent $a_0 = 0$. Diese Restriktion lässt sich bei der Identifikation über Differenzialgleichungsmodelle durch Weglassen dieses Koeffizienten einfach behandeln. Analog wird im Fall eines Differenzierers mit dem Zählerpolynom verfahren.
Das zeitdiskrete Modell hat einen Integrator, wenn $z = 1$ Wurzel des Nennerpolynoms $a(z)$ ist oder äquivalent $a(1) = 0$ gilt. Hieraus resultiert die lineare Gleichungsrestriktion

$$[1, 1, \ldots, 1] \begin{bmatrix} a_0 \\ \vdots \\ a_n \end{bmatrix} = 0. \tag{2.22}$$

Alternativ kann das Vorwissen über einen Integrator auch durch eine Hochpassfilterung der Daten mit $(z-1)/z$ umgesetzt werden. Das Filter kürzt dann den Pol bei $z = 1$. Allerdings bewirkt das Filter eine Erhöhung der Störamplituden, da es ein numerischer Differenzierer ist. Bei der restringierten Behandlung tritt diese unerwünschte Störerhöhung dagegen nicht auf.

2.2.2 Bekannte Wurzeln

Sind $s_1, \ldots, s_k$ bekannte Wurzeln eines Polynoms $a(s)$, dann gilt die lineare Restriktion

$$\begin{bmatrix} 1 & s_1 & s_1^2 & \ldots & s_1^n \\ 1 & s_2 & s_2^2 & \ldots & s_2^n \\ \vdots & \vdots & & \vdots & \\ 1 & s_k & s_k^2 & \ldots & s_k^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = 0. \tag{2.23}$$

Handelt es sich um eine komplexe Wurzel, so tritt diese bei den interessierenden Polynomen mit reellen Koeffizienten konjugiert komplex auf. Das heißt, dass eigentlich zwei Wurzeln bekannt sind. Um nun nicht komplex rechnen zu müssen, wird die Restriktion umgeformt. Es gilt: Sei $\alpha \pm j\beta = r(\cos\phi \pm j \sin\phi)$ mit $r = \sqrt{\alpha^2 + \beta^2}$ und $\phi = \arctan(\beta/\alpha)$ ein a priori bekanntes Wurzelpaar, dann impliziert dieses Paar die Restriktion

$$\begin{bmatrix} 1 & r\cos\phi & r^2\cos(2\phi) & \dots \\ 0 & r\sin\phi & r^2\sin(2\phi) & \dots \end{bmatrix} \begin{bmatrix} a_0 \\ \vdots \\ a_n \end{bmatrix} = 0. \tag{2.24}$$

Beispiel 2.4 (Frequenz bei 50 Hz)
Sind Messdaten mit der Netzfrequenz von $f = 50\,\text{Hz}$ gestört und wird mit $1/T_A = 400\,\text{Hz}$ abgetastet, dann sind in (2.24) $r = 1$ und $\phi = 2\pi f T_A = \pi/4$ zu setzen. Der Vorteil der restringierten Behandlung gegenüber einer Datenvorfilterung mit einer zeitdiskreten Bandsperre liegt darin, dass im Gegensatz zur Bandsperre keine anderen Signalbestandteile verfälscht werden und dass somit ein filterbedingter systematischer Schätzfehler vermieden wird. Ein Wurzelpaar bei 50 Hz kann anschließend durch Division aus dem geschätzten Polynom beseitigt werden, was dank der Restriktion stets exakt gelingt.

Weitere Beispiele mit bekannten Wurzeln liefern die Spektroskopie, bei der bestimmte Frequenzen aufgrund der Stoffzusammensetzungen a priori bekannt sind, oder die elektromagnetische Signaturanalyse, bei der aus dem Resonanzmuster einer reflektierten Radarwelle der Flugzeugtyp bestimmt wird. Im letzteren Fall resultiert aus der Sendefrequenz der Quelle ein bekanntes Wurzelpaar [127].

Als k-fache Wurzel werden Werte mit der Eigenschaft

$$s_i \text{ ist Wurzel von } a(s), \text{ nicht aber von } a(s)/(s - s_i)^k \tag{2.25}$$

$$\text{bzw.} \qquad \lim_{s\to s_i} \frac{a(s)}{(s - s_i)^j} = 0; j = 0, \dots, k-1 \quad \text{und} \quad \lim_{s\to s_i} \frac{a(s)}{(s - s_i)^k} \neq 0 \tag{2.26}$$

bezeichnet. Bekannte k-fache Wurzeln können in analoger Weise bearbeitet werden und führen ebenfalls auf lineare Restriktionen. So ist aus der Faktorzerlegung $a(s) = (s - s_i)^k \bar{a}(s)$ durch Ableiten direkt zu erkennen, dass für eine k-fache Wurzel s_i die folgenden Restriktionen

$$\left.\frac{\mathrm{d}^j a(s)}{\mathrm{d}s^j}\right|_{s=s_i} = 0; \qquad j = 0, \dots, k-1. \tag{2.27}$$

gelten.

Beispiel 2.5 (Restriktion für eine Dreifachwurzel)
Für eine dreifache Wurzel s_1 ergibt sich so beispielsweise die Restriktion

$$\begin{bmatrix} 1 & s_1 & s_1^2 & \dots & s_1^n \\ 0 & 1 & 2s_1 & \dots & n s_1^{n-1} \\ 0 & 0 & 2 & \dots & (n-1)n s_1^{n-2} \end{bmatrix} \begin{bmatrix} a_0 \\ \vdots \\ a_n \end{bmatrix} = 0. \tag{2.28}$$

Als eine Anwendung für bekannte Mehrfachwurzeln seien die per Tustin-Transformation aus Übertragungsfunktionen mit den Polynomgraden (m, n) hervorgehenden z-Modelle genannt. Sie besitzen $n - m$ bekannte Zählerwurzeln bei $z = -1$, da die $n - m$ Zählerwurzeln von $G(s)$ im Unendlichen auf $z = -1$ abgebildet werden.

Begründung für Nullstellen im Unendlichen
Nullstellen von $G(s)$ im Unendlichen werden sichtbar über die Nullstellen bei $q = 0$, die $G(1/q)$ hat. Somit hat jedes SISO-System mit Differenzgrad $r = n-m = \min\limits_{1\leq i\leq n}\{c^T A^{i-1}b \neq 0\}$ zudem r Nullstellen im Unendlichen oder besser ausgedrückt: eine Nullstelle im Unendlichen von der Ordnung r. Es gilt

$$\lim_{q\to 0}\frac{1}{q^j}G(1/q) = \lim_{s\to\infty} s^j G(s) = 0 \qquad j = 0, \ldots, r-1. \tag{2.29}$$

Übertragungsmatrizen $G(s)$ besitzen eine Nullstelle im Unendlichen der Ordnung k, wenn $G(1/q)$ eine k-fache Übertragungsnullstelle bei $q = 0$ hat (Definition von Pugh und Ratcliffe [520]). Zudem haben nicht degenerierte Systeme $\Sigma = (A, B, C, D)$, also solche mit vollem Normalrang der Rosenbrock-Matrix, eine Nullstelle im Unendlichen der Ordnung $\min\{m, p\}$ und degenerierte Systeme eine von einer Ordnung kleiner $\min\{m, p\}$.

Ist eine Wurzel des Zählerpolynoms von $G(s)$ bekannt, wird aber eine Zustandsraumschätzung angestrebt, so liefert für bekanntes s_i, vgl. (2.158),

$$\det\begin{bmatrix} s_i I_n - A & -b \\ c^T & d \end{bmatrix} = 0 \tag{2.30}$$

eine entsprechende Restriktion[8]. Gleichung (2.30) kann z. B. als Rangdefektheitsrestriktion oder über die lineare Abhängigkeit (dann verschwindet die Determinante) berücksichtigt werden.

Anmerkung 2.3 Approximationsprobleme bezüglich des nächstgelegenen Polynoms mit bekannter (vorgegebener) Wurzel werden in [239] diskutiert.

[8] Diese Restriktion folgt direkt aus der Schur-Zerlegung

$$\begin{aligned}\det\begin{bmatrix} sI_n - A & -b \\ c^T & d \end{bmatrix} &= \det\left(\begin{bmatrix} I_n & 0 \\ c^T(sI_n - A)^{-1} & I_n \end{bmatrix}\begin{bmatrix} sI_n - A & -b \\ 0 & c^T(sI - A)^{-1}b + d \end{bmatrix}\right) \\ &= 1 \cdot \underbrace{\det(sI_n - A)}_{N(s)} \cdot \underbrace{\det(c^T(sI - A)^{-1}b + d)}_{G(s)} = Z(s).\end{aligned}$$

2.2.3 Bekannter Eigenwert

Der Kenntnis von Polen in der E/A-Darstellung entspricht im Zustandsraum die Kenntnis von Eigenwerten. Obwohl beide Restriktionen systemtheoretisch dasselbe beschreiben, weisen sie hinsichtlich ihrer Behandlung einen sehr unterschiedlichen Schwierigkeitsgrad auf. Bekannte Pole führen auf lineare Gleichungsrestriktionen, während $\lambda(A) = \zeta$ eine nichtkonvexe, komplizierte nichtlineare Restriktion an die Matrixelemente von A darstellt. Diese Restriktion kann zwar scheinbar einfacher als

$$(A - \zeta I_n) \text{ singulär, bzw. äquivalent } \mathrm{Rang}(A - \zeta I_n) < n \tag{2.31}$$

geschrieben werden, was die Situation aber nur unwesentlich verbessert, wenn die Restriktion a priori berücksichtigt werden soll. Einfacher ist es, die Restriktion a posteriori durch das Matrixapproximationsproblem

$$||\hat{A} - X||_F \stackrel{!}{=} \mathrm{Min} \qquad \lambda(X) = \zeta, \tag{2.32}$$

zu berücksichtigen. Hier wird zu einer Schätzung der Systemmatrix $\hat{A}$ die nächstgelegene Matrix X_{opt} gesucht, die den bekannten Eigenwert ζ hat. Umgeschrieben lautet (2.32)

$$||\hat{A} - X||_F \stackrel{!}{=} \mathrm{Min} \qquad \mathrm{Rang}(X - \zeta I_n) < n \tag{2.33}$$

bzw. nach der Substitution $\tilde{X} =: X - \zeta I_n$

$$||(\hat{A} - \zeta I_n) - \tilde{X}||_F \stackrel{!}{=} \mathrm{Min} \qquad \mathrm{Rang}\tilde{X} < n. \tag{2.34}$$

Dieses Problem hat dank des Eckart-Young-Mirsky-Theorems [299] eine geschlossene Lösung, nämlich $\tilde{X}_{\mathrm{opt}} = \sum_{i=1}^{n-1} \sigma_i u_i v_i^T$, wobei $\hat{A} - \zeta I_n = ((u_i))\mathrm{diag}(\sigma_i)((v_i))^T$ eine Singulärwertzerlegung von $\hat{A} - \zeta I_n$ ist. Die Lösung ist eindeutig, falls der kleinste Singulärwert einfach ist. Durch Resubstitution ergibt sich $X_{\mathrm{opt}} = \tilde{X}_{\mathrm{opt}} + \zeta I_n$.

Anmerkung 2.4 Artverwandte Matrixapproximationsprobleme sind:
- ein bekannter Eigenwert: [327]
- ein bekannter Mehrfacheigenwert: [237], [427], [181]
- zwei bekannte, aber unterschiedliche Eigenwerte: [236], [402]
- mehrere bekannte, aber unterschiedliche Eigenwerte: [403]
- n bekannte Eigenwerte: [134]
- Eigenwert, Links- und Rechtseigenvektor: [226].

2.2.4 Wurzeln in einer Halbebene oder einem Streifen

Die Forderung nach Wurzeln in der linken Halbebene und damit nach Hurwitz-Stabilität führt auf komplizierte notwendige und zugleich hinreichende Restriktionen. Das Problem wird deshalb oft über eine Reparametrisierung oder andere Zugänge gelöst, s. Abschn. 2.3. Gut handhabbar sind letztlich nur die notwendigen Bedingungen für Hurwitz-Stabilität, wonach alle Koeffizienten das gleiche Vorzeichen haben müssen. Ohne Einschränkung der Allgemeinheit wird nachfolgend nur der Fall mit positiven Vorzeichen betrachtet, der insbesondere für monische Polynome greift. Durch Betrachtung des Polynoms $\tilde{a}(z) = a(z+\gamma)$ werden die notwendigen Hurwitz-Bedingungen zu notwendigen Bedingungen für die Lage von Wurzeln links der Achse $\Re e\, z = \gamma$. Seien $\tilde{a}_i$ die Koeffizienten von $\tilde{a}(z)$, so ist

$$0 < \tilde{a}_{n-i} = \sum_{k=0}^{i} a_{n-i+k} \binom{n-i+k}{k} \gamma^k; \qquad i = 0, \ldots, n \tag{2.35}$$

notwendig dafür, dass die Wurzeln von $a(z)$ alle streng links von γ liegen. In Matrixnotation schreibt sich (2.35) für $a_n > 0$ als

$$\begin{bmatrix} \tilde{a}_0 \\ \vdots \\ \tilde{a}_n \end{bmatrix} = L \cdot \begin{bmatrix} a_0 \\ \vdots \\ a_n \end{bmatrix} > 0_{n+1} \quad \text{mit} \quad L_{ij} = \begin{cases} \binom{j-1}{j-i} \gamma^{j-i} & i \le j \\ 0 & i > j. \end{cases} \tag{2.36}$$

Wird die Gleichheit mit zugelassen, so liegen sie links von γ oder auf der γ-Achse.
In der gleichen Weise kann eine Bedingung für Wurzeln rechts der Achse $\Re e\, z = \beta$ abgeleitet werden. Seien $\tilde{a}_i$ die Koeffizienten von $\tilde{a}(z) = a(-z+\beta)$, so ist

$$0 < \tilde{a}_{n-i} = \sum_{k=0}^{i} (-1)^k a_{n-i+k} \binom{n-i+k}{k} \beta^k; \qquad i = 0, \ldots, n \tag{2.37}$$

notwendig dafür, dass die Wurzeln von $a(z)$ alle streng rechts von β liegen. Durch Kombination von (2.35) und (2.37) ergibt sich eine Streifenrestriktion.

Anmerkung 2.5 Für Systeme zweiter Ordnung sind die Streifenrestriktionen notwendig und hinreichend zugleich. Ist zudem bekannt, dass die Wurzeln reell sind, liefert die Diskriminante eine ergänzende Restriktion

$$0 \le a_1^2 - 4a_0a_2. \tag{2.38}$$

Anmerkung 2.6 Charakterisierungen für reelle univariate Polynome, die ausschließlich reelle Wurzeln haben und die zudem alle in einem Intervall $[a, b]$ liegen, werden in [384] angegeben.

2.2.5 Wurzeln auf dem Einheitskreis

Bei der Frequenzschätzung von Multisinussignalen über autoregressive Modelle [572]

$$x[k] = \sum_{i=1}^{n} A_k \mathrm{e}^{j2\pi f_i k} + \varepsilon[k] \tag{2.39}$$

ist bekannt, dass die Wurzeln des zugehörigen z-Modells gerader Ordnung auf dem Einheitskreis liegen müssen. Diesbezügliche Restriktionen verbessern die Schätzergebnisse gegenüber der nicht-restringierten Schätzung.

Damit alle Wurzeln eines reellen Polynoms $a(z) = \sum_{i=0}^{n} a_i z^i$ auf dem Einheitskreis liegen, ist ein symmetrischer Koeffizientenvektor, d. h. $a_i = a_{n-i}$ für $i = 0, \ldots, n$, notwendig (Spiegelpolynom). Der Beweis dieser Aussage erfolgt durch Induktion ausgehend von Paaren $z_{i_{1,2}} = \cos\phi_i \pm j \sin\phi_i$. Die Symmetrie lässt sich mit $a = (a_0, \ldots, a_n)^T$ durch

$$(I_{n/2}, 0_{n/2}, -J_{n/2})a = 0_{n/2} \qquad \text{für } n \text{ gerade} \tag{2.40a}$$

$$(I_{(n+1)/2}, -J_{(n+1)/2})a = 0_{(n+1)/2} \qquad \text{für } n \text{ ungerade} \tag{2.40b}$$

bzw. durch

$$a = \begin{bmatrix} I_{n/2} & 0_{n/2} \\ 0_{n/2}^T & 1 \\ J_{n/2} & 0_{n/2} \end{bmatrix} b; \quad b \in \mathbb{R}^{n/2+1} \qquad \text{für } n \text{ gerade} \tag{2.41a}$$

$$a = \begin{bmatrix} I_{(n+1)/2} \\ -J_{(n+1)/2} \end{bmatrix} b; \quad b \in \mathbb{R}^{(n+1)/2} \qquad \text{für } n \text{ ungerade} \tag{2.41b}$$

ausdrücken. J_n ist hierbei die Rotationsmatrix, die nur auf der zentralen Nebendiagonalen mit Einsen besetzt ist.
Die Symmetrierestriktion lässt noch Skalierungen bezüglich a bzw. b zu. Bei der Wahl einer zusätzlichen eindeutigkeitserzwingenden Restriktion ist dann aber zu beachten, dass diese nicht im Widerspruch zu (2.40) steht (konsistentes System von Restriktionen). Während $a_n = 1$ oder $||a||_2 = 1$ zulässig sind, trifft dies für beliebige Linearkombinationen oder bilineare Restriktionen im Allgemeinen nicht zu.

Eine notwendige Bedingung dafür, dass alle Wurzeln eines komplexen Polynoms auf dem Einheitskreis liegen, ist die konjugierte Symmetrie der Polynomkoeffizienten, d. h. $a_i = \bar{a}_{n-i}$; $i = 0, \ldots, n$ bzw. $a = J_{n+1}\bar{a} \in \mathbb{C}^{n+1}$. Solche Probleme treten bei der Reparametrisierung von Richtungsschätzproblemen auf, bei denen die Wurzeln $z_i = \exp(j2\pi f_i) = \exp(j2\pi d\theta_i)$, die

gesuchten Winkel θ_i enthalten. Konjugierte Symmetrie kann durch

$$a = \begin{bmatrix} I_{n/2} & 0_{n/2} & jI_{n/2} \\ 0_{n/2}^T & 1 & 0_{n/2}^T \\ J_{n/2} & 0_{n/2} & -jJ_{n/2} \end{bmatrix} b;\ b = \begin{bmatrix} \Re e\, a_{0:(n/2-1)} \\ a_{n/2} \\ \Im m\, a_{0:(n/2-1)} \end{bmatrix} \in \mathbb{R}^{n+1} \quad n \text{ gerade} \tag{2.42a}$$

$$a = \begin{bmatrix} I_{(n+1)/2} & jI_{(n+1)/2} \\ J_{(n+1)/2} & -jJ_{(n+1)/2} \end{bmatrix} b;\ b = \begin{bmatrix} \Re e\, a_{0:(n-1)/2} \\ \Im m\, a_{0:(n-1)/2} \end{bmatrix} \in \mathbb{R}^{n+1} \quad n \text{ ungerade} \tag{2.42b}$$

behandelt werden.Da b reell ist, reduziert sich die Zahl der Freiheitsgrade gegenüber dem komplexen a um $n+1$.

Auch die konjugierte Symmetrierestriktion lässt noch Skalierungen (Multiplikation aller Polynomkoeffizienten mit einem gemeinsamen Faktor) zu. Um die Skalierbarkeit zu unterdrücken und gewissermaßen eine rechte Seite für eine lineare LS-Formulierung zu schaffen, wird standardmäßig meist die eindeutigkeitserzwingende Restriktion $a_n = 1$ eingesetzt. Das ist hier aber unzulässig, da die Restriktion $a_n = 1$ im Komplexen zwei Restriktionen enthält; $\Re e\, a_n = 1$, $\Im m\, a_n = 0$. Vor der dann naheliegenden Restriktion $\Re e\, a_n = 1$ bzw. äquivalent $b_1 = 1$ sei aber gewarnt. Wenn nämlich für die Frequenzen f_i gilt $\sum_{i=1}^{n} f_i = (2l+1)/2$ mit einer beliebigen natürlichen Zahl l, ist $\Re e\, a_n = 0$. In einem solchen Fall ist es unmöglich, das auf $\Re e\, a_n = 1$ restringierte Polynom durch nachträgliche Skalierung von a in eines mit $\Re e\, a_n = 0$ umzuformen. Ähnliches gilt, wenn $\Im m\, a_n = 1$ gesetzt wird und $\sum_{i=1}^{n} f_i = l$ ist. Derartige Phänomene treten nur bei komplexwertigen Schwingungen auf, die wiederum bei der mathematischen Formulierung von Problemen der Akustik und Kommunikationstechnik entstehen. Unter dem Suchbegriff „direction-of-arrival“finden sich schnell detaillierte Erklärungen zu den Anwendungen und Algorithmen. Ein Zahlenbeispiel soll die angesprochene Restriktionsproblematik nochmals vertiefen.

Beispiel 2.6 (Restriktionsproblematik bei komplexen Polynomen, [471])
Die komplexe Schwingung $x[k] = \exp(j2\pi 0.14k) + (1+j)/\sqrt{2}\exp(j2\pi 0.36k)$ genügt einer Differenzengleichung, die zu $a(z) = -jz^2 - (2 \cdot 0.7705)z + j$ korrespondiert. Hier ist $\Re e\, a_2 = 0$. Zwar ließe sich $\Re e\, a_2 = 1$ per Division von $a(z)$ durch $-j$ erzwingen, doch ist dann a_1 nicht mehr reell, was im Widerspruch zur Spiegelsymmetrie mit reellem Mittelkoeffizienten steht. Als Konsequenz empfiehlt sich statt der Elementfixierung eine Normrestriktion $||b||_2 = 1$, die keine zusätzliche Forderungen an die f_i stellt.

Anmerkung 2.7 Die Menge der symmetrischen bzw. konsymmetrischen Polynome enthält auch solche, die am Einheitskreis gespiegelte Wurzeln besitzen. Wenn aber die Polynomordnung der Problemstellung angepasst ist, keine Mehrfachwurzeln auftreten und die Wurzeln hinreichend voneinander getrennt sind (Winkeldifferenz nicht zu klein), dann bleiben die Wurzeln auch bei kleinen Störungen der Koeffizienten des Konspiegelpolynoms auf dem Einheitskreis. Die Restriktion ist unter diesen zusätzlichen Annahmen also äquivalent zu der

vollständigen Restriktion $|z_k| = 1$ für alle k . Gemeinhin wird die Anzahl derjenigen Realisierungen der Polynomkoeffizienten klein sein, bei denen echt gespiegelte Wurzeln auftreten; der Effizienzverlust ist also eher klein. Für konsistente Schätzer $\hat{\mathbf{a}}$ ohne Mehrfachwurzeln ist die Restriktion bedingt durch das Zusammenziehen der Störintervalle der Polynomkoeffizienten demnach asymptotisch hinreichend.

Anmerkung 2.8 Alle Wurzeln von $a(z)$ sind genau dann vom Betrage Eins, wenn $a(z)$ ein Konspiegelpolynom ist und die Wurzeln von $\frac{\mathrm{d}}{\mathrm{d}z}a(z)$ auf oder im Einheitskreis liegen [432]. Letztere Bedingung ist algorithmisch schwieriger umzusetzen, s. Abschn. 2.2.6, weshalb sie nur in den Fällen hinzugenommen wird, in denen die Konspiegelpolynombedingung versagt.

Anmerkung 2.9 Restriktionen, die sich auf einen Kreis mit Radius $r \neq 1$ beziehen, können mit der im nachfolgenden Abschnitt angewandten Substitutionstechnik abgeleitet werden.

2.2.6 Wurzeln innerhalb des Einheitskreises

Bekanntermaßen ist das Innere des Einheitskreises das Schur-Stabilitätsgebiet, vgl. Abschn. 2.4. Das folgende Beispiel plausibilisiert, warum geschätzte Differenzengleichungsmodelle höherer Ordnung ohne Stabilitätsrestriktionen häufig instabil sind. Es dient somit zugleich als Motivation dafür, Stabilität des Modells durch geeignete Maßnahmen sicherzustellen.

Beispiel 2.7 (Empfindlichkeitsdilemma)
Die Wurzeln von z-Polynomen können sehr empfindlich bezüglich der Polynomkoeffizienten sein. So hat[9]

$$a(z) = 1.3459z^4 - 4.8839z^3 + 6.6388z^2 - 4.0063z + 0.90556 \tag{2.43}$$

die stabilen Wurzeln $z_1 = 0.9004$, $z_2 = 0.8240$ und $z_{3,4} = 0.9522 \pm 0.0160j$. Wird lediglich a_2 auf $a_2 = 6.6387$ (vierte Nachkommastelle!) geändert, lauten die Wurzeln $z_1 = 1.0107$ (instabil), $z_2 = 0.8024$ und $z_{3,4} = 0.9078 \pm 0.0743$. Bei der Identifikation verschärft sich das Problem, da Parameteränderungen als Folge stochastisch gestörter Signale in der Regel deutlich größer sind. Als Alternative zu Restriktionen bietet sich die direkte Identifikation von s-Modellen an [247], bei der das Dilemma weniger stark wirksam ist.[10]

[9] aus Skript von Prof. Schlacher; Uni Linz: http://regpro.mechatronik.uni-linz.ac.at/downloads/ate/ate-kap2_0809.pdf

[10] Bei digitalen Regelungen höherer Ordnung wird das beschriebene Problem (und zugleich das Problem der Stellenauslöschung) dadurch umgangen, dass die Reglerübertragungsfunktion faktorisiert und als eine Kaskade von Differenzengleichungen erster bzw. zweiter Ordnung implementiert wird. Die nachfolgend vorgestellten Reparametrisierungen nutzen ähnliche Ideen.

Das Problem bei der Formulierung geeigneter Einheitskreisrestriktionen liegt darin, dass die gängigen Kriterien wie Jury, Schur, Schur-Cohn [382] oder Schur-Cohn-Fujiwa auf multivariate, polynomiale, nichtkonvexe Ungleichungsrestriktionen führen, vgl. explizite Restriktionen bis vierter Ordnung in [3]. Daher gibt es bisher keinen prädestinierten Zugang zur Einbeziehung der Stabilitätsrestriktion. Eine Ausnahme bilden die folgenden notwendigen und hinreichenden Bedingungen

$$|a_1| > |a_0| \qquad \text{System erster Ordnung} \tag{2.44a}$$

$$\left.\begin{array}{rcl} a_2 + a_1 + a_0 > 0 & \text{oder} & a_2 + a_1 + a_0 < 0 \\ a_2 - a_0 > 0 & & a_2 - a_0 < 0 \\ a_2 - a_1 + a_0 > 0 & & a_2 - a_1 + a_0 < 0 \end{array}\right\} \quad \text{System zweiter Ordnung,} \tag{2.44b--d}$$

wobei (2.44a) direkt und (2.44b) bis (2.44d) nach Bilineartransformation $z = \frac{1+w}{1-w}$ aus den Hurwitz-Bedingungen abzuleiten sind. Für monische Polynome zweiter Ordnung ist dabei der linke Block mit $a_2 = 1$ zu wählen.[11] Eine innere konvexe Approximation des Stabilitätsgebiets im Parameterraum, also eine hinreichende Bedingung, lautet [196]

$$|a_0| + |a_1| + \ldots + |a_{n-1}| = ||(a_0, a_1, \ldots, a_{n-1})^T||_1 < |a_n|. \tag{2.45}$$

Neben dem Einheitskreis sind auch Kreise mit anderen Radien interessant. So folgen Radien kleiner Eins aus dem Abtasttheorem und der A-priori-Kenntnis über die größte Zeitkonstante. Radien größer Eins treten etwa bei der Identifikation instabiler Modelle im geschlossenen Regelkreis auf oder wenn die Stabilitätsgebiete von Δ-Modellen[12] betrachtet werden.

Beispiel 2.8 (A-priori-Information zum Radius)
Liegen die Zeitkonstanten in einem Intervall $[T_{\min}, T_{\max}]$, so ist $r_{\max} = \mathrm{e}^{-T_A/T_{\max}}$ und $r_{\min} = \mathrm{e}^{-T_A/T_{\min}}$. Für PT1-Glieder gilt $T \approx \frac{1}{3}T_{95\%}$, womit eine Zeitkonstantenabschätzung und damit eine Radiusschätzung aus der 95%-Einschwingzeit gewonnen werden kann. Für PT2-Glieder ohne Vorhalt folgt über $T_{95\%} \geq \max\{T_{95\%,1}, T_{95\%,2}\}$ die Abschätzung $|z_i| \leq r = \mathrm{e}^{-3T_A/T_{95\%}}$. Diese Abschätzung ist nicht, wie mitunter behauptet wird, für alle Systeme 2. Ordnung[13] gültig!

[11] Statt (2.44c), also $1 > a_0$, findet sich in vielen Darstellungen die Forderung $|a_0| < 1$, die zusätzlich die redundante Restriktion $-1 < a_0$ enthält, welche direkt aus (2.44b)+(2.44d) folgt.

[12] Unter Δ-Modellen sollen hier solche verstanden werden, bei denen an die Stelle der z-Variablen eine gebrochen rationale Kombination Δ von z tritt. Dies zieht eine Transformation der Ein- und Ausgangssignale und Parameter nach sich und hat als Ziel, die problematische Polclusterung bei $z = 1$ im Fall kleiner Abtastzeiten zu entschärfen, die Parametersensitivität zu reduzieren und die Konditionierung der Schätzprobleme zu verbessern, s. z. B. [174], [468], [38], [261].

[13] $G(s) = \frac{1}{s^2+s+1}$ hat $T_{95\%} = 5.3$. Wegen $|z_i| = \mathrm{e}^{\Re e s_i T_A} = \mathrm{e}^{-0.5T_A}$ und $r = \mathrm{e}^{-3T_A/5.3} = \mathrm{e}^{-0.57T_A}$ ist r kleiner $|z_i|$ und somit nicht wie behauptet eine Abschätzung nach oben.

Der Standardweg, um aus einem Kriterium für $|z_i| \leq 1$ eines für $|z_i| \leq r$ zu folgern, besteht darin, das Polynom $\tilde{a}(z) = a(rz)$ zu untersuchen. Da für dessen Wurzeln $r\tilde{z}_i = z_i$ gilt, folgt aus $|\tilde{z}_i| < 1$ direkt $|z_i| < r$. Somit brauchen in den Kriterien nur die Koeffizienten durch $\tilde{a}_i = a_i r^i$ ersetzt werden. Die notwendigen und hinreichenden Restriktionen für Wurzeln in $\mathcal{B}(0,r) = \{z \in \mathbb{C} : |z| < r\}$ lauten für ein monisches Polynom zweiten Grads demnach

$$r^2 + ra_1 + a_0 > 0, \quad r^2 - ra_1 + a_0 > 0, \quad r^2 - a_0 > 0. \tag{2.46}$$

Gut handhabbare notwendige Kreisgebietsbedingungen $\mathcal{B}(0,r)$ für höhergradige Polynome sind die Intervallrestriktionen

$$|a_i| < \binom{n}{n-i} r^{n-i} \qquad i = 0, \ldots, n-1; a_n = 1 \quad /^{14}. \tag{2.47}$$

Strengere Restriktionen liefert der nachfolgende Satz.

Satz 2.1 (Notwendige Bedingungen für Wurzeln innerhalb von $\mathcal{B}(m,r)$, [605]) Wenn $a(z) = a_n z^n + a_{n-1} z^{n-1} + \ldots + a_0$ mit $a_n > 0$ Wurzeln in $\mathcal{B}(m,r)$, $m \in \mathbb{R}$ hat, gilt

$$R \cdot L \cdot \begin{bmatrix} a_0 \\ \vdots \\ a_n \end{bmatrix} > 0_{n+1} \qquad \text{mit} \qquad R_{ij} = \begin{cases} (-1)^{i-1}\binom{n+1-j}{i-1} & j-1 \leq n+1-i \\ 0 & \text{sonst}; i,j = 1, \ldots, n+1 \end{cases}$$
$$L_{ij} = \begin{cases} (m-r)^{j-i}\binom{j-1}{j-i}(2r)^{i-1} & i \leq j \\ 0 & i > j. \end{cases} \tag{2.48}$$

Mit (2.48) stehen also lineare Restriktionen zur Verfügung (allerdings nur notwendige), die sich in Verbindung mit parameterlinearen LS-Formulierungen vor allem für Online-Anwendungen eignen. Für Probleme, die ohnehin per numerischer Optimierung offline gelöst werden sollen, ist eine Reparametrisierung des Parameterraums der bessere Zugang, da dann notwendige und hinreichende Bedingungen für den Suchraum formuliert werden können. Wesentlicher Vorteil gegenüber den Originalparametern ist, dass die Menge der stabilen Polynome nicht verlassen wird, wodurch eine aufwendige und in jedem Optimierungsschritt erforderliche numerische Wurzelberechnung entfällt. Eine Reparametrisierung liefert der folgende Satz, wobei die Kombination aus Addition und Faktorisierung Vorteile hinsichtlich der Topologie des neuen Suchraums bietet, s. hierzu eine allgemeinere Diskussion in Abschn. 5.1.1.

[14] Per Vietaschen Wurzelsatz: $|a_i| = |\sum_{\beta \in C_{n-i,n}} z_{\beta_1} \cdots z_{\beta_i}| \leq \sum_{\beta \in C_{n-i,n}} |z_{\beta_1}| \cdots |z_{\beta_i}| < \sum_{\beta \in C_{n-i,n}} r^i = \binom{n}{n-i} r^{n-i}$

Satz 2.2 (Reparametrisierung nach Ramachandran und Gargour, [523])
$a(z) = \sum_{i=0}^{n} a_i z^i$ mit $a_n > 0$ ist genau dann ein Schur-Polynom, wenn

$$a(z) = c\prod_{i=1}^{m}(z^2 - 2\alpha_i z + 1) + (z^2 - 1)\prod_{i=1}^{m-1}(z^2 - 2\beta_i z + 1), \quad m = \frac{n}{2} \tag{2.49a}$$

$$1 > \alpha_1 > \beta_1 > \alpha_2 > \beta_2 \ldots > \beta_{n-1} > \alpha_n > -1; c > 1 \qquad \text{bzw.} \tag{2.49b}$$

$$a(z) = c(z+1)\prod_{i=1}^{m}(z^2 - 2\alpha_i z + 1) + (z-1)\prod_{i=1}^{m}(z^2 - 2\beta_i z + 1), \quad m = \frac{n-1}{2} \tag{2.49c}$$

$$1 > \alpha_1 > \beta_1 > \alpha_2 > \beta_2 \ldots > \alpha_n > \beta_n > -1; c > 1. \tag{2.49d}$$

Alternativ kann eine Reparametrisierung mittels Schur-Parametern gemäß Satz 2.3 vorgenommen werden, wobei dann die Indizierung anzupassen ist. Eine Anwendung zur Bestimmung des nächstgelegenen von-Neumann-Polynoms (s. S. 42 Def. 2.2) beschreibt [463].

2.2.7 Wurzeln außerhalb eines Kreises

Wenngleich in der Praxis viel häufiger Restriktionen für Wurzeln innerhalb eines Kreises auftreten, soll hier dennoch kurz auf die Behandlung des Falls eingegangen werden, in dem die Wurzeln eines Polynoms außerhalb eines Kreises liegen müssen. Eine Anwendung liefert das Beispiel 2.8 mit einem entsprechenden inneren Radius. Weitere Anwendungen sind Rekursionsbeziehungen, die zu Polynomen in $\zeta := z^{-1}$ korrespondieren.

Ein Weg, die Restriktion „Wurzeln außerhalb eines Kreises“ zu fassen, ist die Rückführung auf die Restriktion „Wurzeln im Einheitskreis“. Das gelingt durch die Umformung $\tilde{a}(z) = z^n a(r/z)$. Zwischen den Wurzeln gilt dann $r/\tilde{z}_i = z_i$, weshalb unmittelbar aus $|\tilde{z}_i| < 1$ die Eigenschaft $|z_i| > r$ folgt. Die Koeffizientenersetzung lautet demnach $\tilde{a}_i = a_{n-i} r^{n-i}$, und das Polynom der $\tilde{a}_i$ muss Schur-stabil sein.

Die Rückführung der Restriktion ist auch zweckmäßig, wenn bei Online-Verfahren nur notwendige Bedingungen eingesetzt werden sollen, da dann lineare Restriktionen entstehen. Wird hingegen eine Offline-Optimierung angestrebt, so empfiehlt sich eine Reparametrisierung basierend auf den Schur-Parametern [463], s. auch Schur-Cohn-Kriterium [588].

Satz 2.3 (Reparametrisierung mit Schur-Parametern, [324])
Es gilt $a(\zeta) = 1 + \sum_{i=1}^{n} a_i \zeta^i \neq 0$; $a_i \in \mathbb{R}$ für $|\zeta| \leq 1$ genau dann, wenn

$$a_{k,i} = a_{k-1,i} + a_{k,k} a_{k-1,k-i} \quad k = 2, \ldots, n; i = 1, \ldots, k-1 \tag{2.50a}$$

$$a_i := a_{n,i} \quad i = 1, \ldots, n \tag{2.50b}$$

$$|a_{k,k}| < 1 \quad k = 1, \ldots, n, \tag{2.50c}$$

wobei $\theta_k := a_{k,k}$ die Schur-Parameter sind, über denen optimiert wird.

Anmerkung 2.10 Der Algorithmus (2.50a) wird als Levinson-Durbin-Rekursion bezeichnet [572]. Er kann auch zur Berechnung der partiellen Ableitungen von $a_i(a_{11}, \ldots, a_{nn})$ für die Optimierung verwendet werden, indem mit $a_{jj} = 1$ bei sonst festen $a_{ii}; i \neq j$ gerechnet wird.

Beispiel 2.9 (Reparametrisierungen mit Schur-Parametern)
Für Polynome 2. und 3. Ordnung ergeben sich aus (2.50a) die Beziehungen

$$\begin{aligned} a_1 &=: a_{11}(1 + a_{22}) \qquad \text{bzw.} \qquad & a_1 &=: a_{11}(1 + a_{22}) + a_{22}a_{33} \\ a_2 &=: a_{22} & a_2 &=: a_{22} + a_{33}a_{11}(1 + a_{22}) \\ & & a_3 &=: a_{33} \end{aligned}$$

Die Bedingungen $|a_{11}| < 1$ und $|a_{22}| < 1$ für ein Polynom 2. Ordnung sind, wie nicht anders zu erwarten, äquivalent zu den Bedingungen (2.44), wie folgende Überlegungen zeigen. Seien $\tilde{a}_2 = 1, \tilde{a}_1 = a_1, \tilde{a}_0 = a_2$ die Koeffizienten von $\tilde{a}(z) = \tilde{a}_2 z^2 + \tilde{a}_1 z + \tilde{a}_0$, dann ist $|\tilde{a}_0| = |a_2| = |a_{22}| < 1$ nach (2.50c) kompatibel mit (2.44c). Für $0 \leq a_{11} < 1$ folgt $\tilde{a}_1 = a_1 < 1 + a_{22} = 1 + \tilde{a}_0$, also (2.44d), und für $-1 < a_{11} < 0$ folgt $-\tilde{a}_1 = -a_1 < 1 + a_{22} = 1 + \tilde{a}_0$, also (2.44b).

2.2.8 Unbekannte Mehrfachwurzeln und -eigenwerte

Vielfachheiten von Wurzeln oder Eigenwerten treten beispielsweise auf, wenn in einer Wirkungskette zwei oder mehrere gleiche Teilsysteme vorhanden sind. Das Problem dabei ist, dass Systeme mit dieser Eigenschaft schwer identifizierbar sind, da die Wurzellagen sehr sensitiv auf Änderungen in den Polynomkoeffizienten reagieren. Mathematisch sind strukturelle Änderungen etwa in den Jordan-Strukturen oder der Verlust der Diagonalisierbarkeit von Matrizen die Folge. Numerisch führt bereits im ungestörten Fall ein Rechnen mit geringerer Genauigkeit dazu, dass Mehrfachwurzeln nicht mehr als solche in Erscheinung treten. Doch das Dramatische aus Sicht der Identifikation ist, dass es keine gut handhabbaren Restriktionen gibt, mit deren Hilfe die A-priori-Kenntnis über die Vielfachheit in das Schätzproblem einbezogen werden kann. Diese Aussage wird durch das Resultantenkriterium[15] und das daraus abgeleitete Beispiel klarer.

Satz 2.4 (Resultantenkriterium für k-fache Wurzeln, [209])
Die Resultante, d. h. die Determinante der Sylvester-Matrix (2.53), der Polynome $a(s)$ und $\frac{\mathrm{d}^{k-1}}{\mathrm{d}s^{k-1}}a(s)$, verschwindet genau dann, wenn $a(s)$ eine k-fache Wurzel hat.

[15] Das Kriterium ist auch als Diskriminantenkriterium bekannt. Die erste Diskriminante von $a(s)$, also $\mathrm{Disk}(a) \stackrel{\mathrm{def}}{=} \frac{(-1)^{n(n-1)/2}}{a_n}\mathrm{Res}(a, a')$, ist im Fall $n = 2$ dabei die gewöhnliche Diskriminante (Radikand der Wurzel der quadratischen Lösungsformel), denn $\mathrm{Res}(a, a') = -a_2(a_1^2 - 4a_0a_2)$.

Beispiel 2.10 (Zur Handhabbarkeit des Resultantenkriteriums)
Für monische quadratische Polynome erzwingt $a_1^2 - 4a_0 = 0$ eine Doppelwurzel; für $n = 3$ und eine Dreifachwurzel entsteht der unschöne Ausdruck $18a_0a_1a_2 + a_2^2a_1^2 - 4a_1^3 - 4a_2^3a_27a_0^2 = 0$.

Kurzum, bei der gleichungsfehlerorientierten Schätzung ist es außer für $n = 2$ nicht möglich, das Wissen über unbekannte Mehrfachwurzeln effizient zu nutzen. Bei der ausgangsfehlerorientierten Schätzung kann die Vielfachheit über einen geeigneten Strukturansatz berücksichtigt werden, indem etwa in einer Linearfaktorisierung eines Polynoms mehrmals der gleiche Faktor angesetzt wird, was die freie Parameteranzahl dann zwangsläufig reduziert.

Anmerkung 2.11 Für Schätzungen von Systemmatrizen bietet sich noch an, a posteriori ein Matrixapproximationsproblem zu formulieren, s. [427], [237], [404], [13].

2.2.9 Polkürzbarkeit

Bei der Schätzung von s- oder z-Übertragungsfunktionen führen zu hohe Polynomgrade dazu, dass im ungestörten Fall eine Polkürzung möglich ist. Bei Gleichungsfehleransätzen tritt dann ein Rangverlust in der Datenmatrix auf, s. Beispiel 2.11. Im gestörten Fall kommt es zwar zu keiner Polkürzung und damit auch nicht zum Rangverlust, aber hohe Parameterempfindlichkeiten oder ein plötzlicher Programmabsturz infolge einer resultierenden asymptotischen Singularität sind einhergehende Probleme.

Beispiel 2.11 (Polkürzbarkeit und lineare Abhängigkeit)
Die Gleichungsfehler-LS von $G(z) = \frac{z-0.2}{(z-0.8)(z-0.2)}$ führt auf den Ansatz

$$y[k-1] - 0.16y[k-2] + u[k-1] - 0.2u[k-2] = y[k] + \varepsilon[k] \qquad k = 2, \ldots, N. \tag{2.51}$$

Wegen $y[k-1] = 0.8y[k-2] + u[k-2]$ (verschobene Differenzengleichung des gekürzten Systems) ist die 1. Spalte aus den $y[k-1]$ von der 2. und 4. Spalte aus den $y[k-2]$ und $u[k-2]$ abhängig (wohlgemerkt im ungestörten Fall, andernfalls gilt diese Aussage im statistischen Mittel).

Das Problem der Polkürzbarkeit entsteht auch bei der Identifikation von MISO-Systemen über Gleichungsfehleransätze, vgl. Beispiel 2.2. Infolge von Störungen werden die Polynome $\tilde{a}(q)$ und $\tilde{b}_1(q)$ bzw. $\tilde{a}(q)$ und $\tilde{b}_2(q)$ nicht mehr den gemeinsamen Teiler $a_2(q)$ bzw. $a_1(q)$ besitzen. In diesem Fall ist es beispielsweise sinnvoll, das nächstgelegene gekürzte System zu suchen.

Rein formal lässt sich die Polkürzbarkeit über Parameterrestriktionen erzwingen. Allerdings ist bereits das Aufstellen der Restriktionen ohne Computeralgebra kaum möglich. Ferner

gestaltet sich ihre Behandlung in gleichungsfehlerorientierten Algorithmen äußerst schwierig, sodass deren Vorteile verloren gehen; bei ausgangsfehlerorientierten Zugängen tritt das Polkürzbarkeitsproblem in dieser Form nicht auf!
Eine Polkürzbarkeit kann mit Hilfe eines Rangtests der Sylvester-Matrix, die im folgenden Satz eingeführt wird, erkannt werden.

Satz 2.5 (Polkürzbarkeit, [572], [532])
Zwei Polynome

$$a(s) = a_n \prod_{i=1}^{n}(s - s_{a,i}) \qquad \text{und} \qquad b(s) = b_m \prod_{j=1}^{m}(s - s_{b,j}) \tag{2.52}$$

haben genau dann p gemeinsame Wurzeln, wenn die Sylvester-Matrix

$$S(a(s), b(s)) = \begin{bmatrix} a_n & a_{n-1} & a_{n-2} & \dots & a_1 & a_0 & 0 & 0 & \dots & 0 \\ 0 & a_n & a_{n-1} & \dots & a_2 & a_1 & a_0 & 0 & \dots & 0 \\ 0 & 0 & a_n & \dots & a_3 & a_2 & a_1 & a_0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & & & & & \ddots & \vdots \\ 0 & 0 & 0 & & a_n & \dots & & & \dots & a_0 \\ b_m & b_{m-1} & b_{m-2} & \dots & b_0 & 0 & \dots & & & 0 \\ 0 & b_m & b_{m-1} & \dots & b_1 & b_0 & 0 & \dots & & 0 \\ 0 & 0 & b_m & \dots & b_2 & b_1 & b_0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & & & & & & \vdots \\ 0 & 0 & 0 & & b_m & \dots & & & \dots & b_0 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}a\\a\\a\\a\\a\end{matrix}}\right\} m \text{ Zeilen} \\ \left.\vphantom{\begin{matrix}a\\a\\a\\a\\a\end{matrix}}\right\} n \text{ Zeilen} \end{matrix} \tag{2.53}$$

den Rang $n + m - p$ hat.

Für Diagnoseaufgaben, bei denen sich beispielsweise ein Fehler über eine Polkürzbarkeit äußert (Wegfall eines dynamischen Übertragungselements), kann die Determinante von S, die auch Resultante genannt wird, herangezogen werden. Das ist einfacher als eine Rangdetektion in den Schätzgleichungen über die Singulärwerte (Ausnahme: rekursive SVD-Varianten zu LS und TLS).

Nun sollen zwei Möglichkeiten zur A-posteriori-Ordnungsreduktion von Polynomen vorgestellt werden. Als ein populärer Zugang bietet sich das Matrixapproximationsproblem

$$||B - X||_F \stackrel{!}{=} \text{Min} \qquad \text{Rang} X < m + n; X \text{ Sylvester-Matrix} \tag{2.54}$$

an, aus dessen Minimierer kürzbare Polynome resultieren. Dieses nichtkonvexe Problem lässt sich in ein affin strukturiertes TLS-Problem umformulieren und als solches lösen. Wichtungen berücksichtigen die Vielfachheiten n bzw. m des Auftretens der Elemente. Alternativ kann eine zyklische Projektion (Projektion in die Menge der $(m+n)$-Sylvester-Matrizen und

Projektion in die Menge der Matrizen vom Rang kleiner $m+n$) verwendet werden. Dieser einfach zu programmierende Algorithmus liefert eine im Gütewert monoton fallende Folge, die gegen ein Element aus der Restriktion strebt, allerdings nicht notwendigerweise gegen den Minimierer.
Wird eine Polkürzbarkeit durch Polynome höherer Ordnung angestrebt, dann ist die Rangbedingung zu ändern. Anders als beim klassischen Rang-k-Approximationsproblem, bei dem der Gütewert mit fallendem Rang des Approximanden wächst, kann hier der Gütewert zu einem Approximanden kleineren Rangs auch kleiner sein, s. Beispiel in [182].
Eine Alternative zum Matrixapproximationsproblem (2.54) stellt das Gütekriterium

$$||\Delta a||_2^2 + ||\Delta b||_2^2 \stackrel{!}{=} \text{Min} \qquad \begin{array}{l} \Delta a = C_1(d)u - a \\ \Delta b = C_2(d)v - b \end{array} \tag{2.55}$$

dar, wobei $a, b, \Delta a, \Delta b$ Vektoren mit den Koeffizienten von $a(s), b(s), \Delta a(s), \Delta b(s)$ sind und d, u, v die Koeffizientenvektoren des Teilerpolynoms bzw. der Quotientenpolynome bezeichnen. $C_1 : \mathbb{R}^{r+1} \to \mathbb{R}^{(n+1)\times(n-r+1)}$ und $C_2 : \mathbb{R}^{r+1} \to \mathbb{R}^{(m+1)\times(m-r+1)}$ sind Band-Toeplitz-Matrizen, die die Multiplikationen $u(s)d(s) = a(s) + \Delta a(s)$ bzw. $v(s)d(s) = b(s) + \Delta b(s)$ ausführen. Die Optimierung ist letztlich über $d \in \mathbb{R}^{r+1}, u \in \mathbb{R}^{n-r+1}, v \in \mathbb{R}^{m-r+1}$ auszuführen. Drei Zugänge hierfür werden in [132] diskutiert.

2.2.10 Einhaltung des Abtasttheorems

Nach dem Abtasttheorem können bandbegrenzte Signale mit der Grenzfrequenz ω_G vollständig rekonstruiert werden, wenn $\omega_A = 2\pi/T_A$ mindestens der doppelten Grenzfrequenz entspricht.[16] Ein Verletzen des Abtasttheorems führt zum sog. Alias-Effekt, bei dem das Originalsignal durch ein Signal mit falschen Amplituden und scheinbar zu niedrigen Frequenzen rekonstruiert wird. Durch Antialiasing-Maßnahmen, also insbesondere Tiefpassfilterungen, lässt sich der Effekt vermindern. Bei der Identifikation kann ein Nichtbeachten des Abtasttheorems und des entstehenden Effekts zu falschen Modellen führen. Auf der anderen Seite kann es aber selbst bei Beachten des Abtasttheorems für das Nutzsignal bedingt durch Approximation (falsche Modellordnung, Näherung durch LTI-System) und durch Störungen (Rauschen) vorkommen, dass die Modelle Pole haben, deren Eigenvorgänge im Widerspruch zum Whittaker-Kotelnikow-Shannon-Abtasttheorem [607] (auch Nyquist-Shannonsches Abtasttheorem oder kurz Abtasttheorem genannt) stehen. Deshalb sollten solche Modelle durch geeignete Restriktionen für „korrektes Abtasten“ ausgeschlossen werden oder aber a posteriori diesbezüglich überprüft werden. Beispielsweise müssen Eigenvorgänge $\exp(\alpha_i t)\cos(\omega_i t)$

[16] Sofern die untere Frequenz des Bandes nicht wie üblich Null ist, gilt $f_{\text{abtast}} > 2(f_{\max} - f_{\min})$. Das wird in der Radiotechnik zur Bandpassunterabtastung genutzt [607].

nach dem Abtasttheorem

$$0 \leq \omega_i \leq \omega_G \leq \frac{\omega_A}{2} = \frac{\pi}{T_A} \tag{2.56}$$

erfüllen, um rekonstruierbar zu sein. Hieraus folgt mit $\omega_i = |\Im m\, s_i|$

$$|\Im m\, s_i| \leq \frac{\pi}{T_A}. \tag{2.57}$$

Obwohl die Shannon-Abtastfrequenz im Grenzfall die Signalrekonstruktion aus zwei Werten pro Periode zuließe, wird in der Praxis üblicherweise eine Abtastfrequenz gewählt, die mindestens vier Werte pro Periode liefert. Hierdurch verschärft sich (2.57) zu

$$|\Im m\, s_i| \leq \frac{\pi}{2T_A}. \tag{2.58}$$

Eine Hurwitz-Stabilitätsforderung begrenzt diesen Streifen von rechts auf $\Re e\, s_i < 0$. Um eine linke Schranke für den Streifen zu erhalten, wird jetzt angenommen, dass sich das System als Parallelschaltung von Tiefpässen erster Ordnung schreiben lässt. Der Tiefpass mit der kleinsten Zeitkonstante T_G bestimmt dann die höchste Frequenz $\omega_G = 1/T_G$ von Sinusschwingungen, die das Gesamtsystem mehr oder weniger ungedämpft passieren. Nun ist diese Frequenz aber durch das Abtasttheorem beschränkt. Über $\omega_A/2 \geq 1/T_G \geq 1/T_i = -s_i = -\Re e\, s_i$ ergibt sich dann eine Begrenzung des Streifens von links

$$-\frac{\pi}{T_A} \leq \Re e\, s_i. \tag{2.59}$$

Der zulässige Polbereich wird durch die angegebene linke Schranke immer noch sehr großzügig eingegrenzt. Wird nämlich gefordert, dass der schnellste exponentielle Abklingvorgang wenigstens dreimal innerhalb seiner Einschwingzeit abgetastet werden soll, dann muss $3T_A \approx T_{Einschwing} \approx 3T_G$ gelten:

$$-\frac{1}{T_A} \approx -\frac{1}{T_G} \leq -\frac{1}{T_i} = s_i = \Re e\, s_i, \qquad \text{kurzum} \quad -\frac{1}{T_A} \leq \Re e\, s_i. \tag{2.60}$$

Die Restriktionen an die Pole des kontinuierlichen Modells können auch in Restriktionen für das zeitdiskrete Modell umgeformt werden. Für Modelle, die per Tustin-Transformation erstellt werden, ist $s_i := \frac{2}{T_A}\frac{z_i-1}{z_i+1}$ zu ersetzen. Bei Halteglied-Diskretisierungen ist $z_i := \exp(s_i T_A)$ zu verwenden. Für stabile, rein reelle Pole[17] folgt dann mit (2.59)

$$0.043 \approx \mathrm{e}^{-\pi} \leq z_i < 1, \tag{2.61}$$

wobei die linke Schranke mit der schärferen Forderung (2.60) zu $\mathrm{e}^{-1} \approx 0.37$ verschärft wird. (2.61) selbst ist indes schärfer als die Trivialforderung an reelle Pole, keinen negativen Realteil zu besitzen! Wäre dem so, würde ein zeitdiskretes Modell erster Ordnung schwingen, etwa $(-0.8)^k$, was im Kontinuierlichen unmöglich ist.

[17] Die Annahme reeller Pole kann nicht abgeschwächt werden, womit (2.61) auch nicht zu $0.043 \leq \Re e\, z_i < 1$ erweitert werden kann. Wird etwa $\cos(\omega_i t)$ mit $\omega_i = \frac{3\pi}{4T_A}$ also $f_i = \frac{3}{8T_A}$ mit $f_A = \frac{1}{T_A} > 2f_i$ abgetastet, hat das zugeordnete Abtastsystem die Pole $z_{1,2} = -\frac{\sqrt{2}}{2} \pm j\frac{\sqrt{2}}{2}$.

Unter Annahme einer praktischen Abtastung nach (2.58), d. h. $\omega_A \geq 4 \max_i \omega_i$, liegen alle Pole des Abtastsystems in der rechten Halbebene, denn

$$0 = \mathrm{e}^{\Re e\, s_i T_A} \cos(\frac{\pi}{2}) \leq \mathrm{e}^{\Re e\, s_i T_A} \cos(T_A \Im m\, s_i) = \Re e\, z_i.$$

Dies kann bei stabilen Systemen zusammen mit $\Re e\, z_i < 1$ durch eine Streifenrestriktion, vgl. Abschn. 2.2.4, berücksichtigt werden. Für Systeme ohne Schwingungsanteil ist indes eine Kreisbedingung $\mathcal{B}(\frac{1}{2}, \frac{1}{2})$ schärfer.

Die Imaginärteilrestriktion bringt keine Einschränkungen, da in

$$z_i = \mathrm{e}^{T_A \Re e\, s_i}(\cos(T_A \Im m\, s_i) + j \sin(T_A \Im m\, s_i)) \tag{2.62}$$

die Argumente mit $-\pi \leq T_A \Im m\, s_i \leq \pi$ durch (2.57) nicht eingeschränkt werden.

Zusammengefasst ergeben sich einige Konsequenzen aus dem Abtasttheorem:
Die Systemdynamik erfordert Testsignale, die eine ständige Anregung im wirksamen Frequenzbereich bewirken. Hieraus folgt eine Mindestabtastfrequenz. Falls ein zeitdiskretes Modell erstellt werden soll, darf die Abtastfrequenz jedoch nicht zu hoch gewählt werden, da es sonst zu einer Polclusterung kommt, s. [314] für Hinweise zur Wahl der Abtastfrequenz. Gegebenenfalls ist die Abtastfrequenz für die Identifikation durch Weglassen von Abtastwerten künstlich zu verringern. Bei der Identifikation zeitkontinuierlicher Modelle ist eine zu hohe Abtastfrequenz weniger kritisch, da beispielsweise beim Zustandsvariablenfilter-Verfahren eine Integration erfolgt.
Selbst bei geeignet gewählter Abtastfrequenz muss das identifizierte System nicht mit der Abtastfrequenz verträglich sein, was an der Approximation oder den Störungen liegen kann. Die entsprechenden Restriktionen liefern einen Weg, um das zu erkennen und zu vermeiden. Die abschließende Anmerkung gibt einen Hinweis darauf, wie der Bandbreitenspreizung (Oberwellen) bei nichtlinearen Systemen begegnet werden kann.

Anmerkung 2.12 Nichtlineare Systeme bewirken in aller Regel eine Spreizung der Bandbreite des Ausgangssignals gegenüber der des Eingangssignals. Bei der Identifikation wird deshalb die Abtastfrequenz auf den Ausgang bezogen, was sie oft deutlich erhöht. Das Problem lässt sich durch das verallgemeinerte Abtasttheorem von Zhu (1992) [650]

$$y(t) = g^{-1}\Big(\sum_{k=-\infty}^{\infty} g(y(kT_A)) \frac{\sin(\pi(t - kT_A)/T_A}{\pi(t - kT_A)/T_A} \Big)$$

signifikant entschärfen. Hierzu wird eine kompensierende Nichtlinearität $g(.)$ eingesetzt, die das Ausgangssignal so transformiert, das $g(y(t))$ die Bandbreite des Eingangs hat. Anwendungen werden in [602] beschrieben.

2.3 Übertragungsstabilität für lineare zeitkontinuierliche Systeme

Als SISO-LTI-Modell wird in der Online-Parameterschätzung[18] oft die Differenzialgleichung

$$a_n y^{(n)}(t) + a_{n-1} y^{(n-1)} + \ldots + a_0 y(t) = b_m u^{(m)}(t) + b_{m-1} u^{(m-1)}(t) + \ldots + b_0 u(t) \qquad m \leq n \tag{2.63}$$

mit der „bevorzugten“ Restriktion $a_n = 1$ gewählt. Lediglich für Systeme erster und zweiter Ordnung gelingt es, Stabilität auf einfache Weise durch Positivitätsrestriktionen an die a_i zu erzwingen.

Etwas anders ist die Situation bei ausgangsfehlerorientierten Verfahren, die per Simulation $\hat{y}(t;\hat{\theta})$ berechnen und den Parameter $\hat{\theta}$ entsprechend der Gütefunktion iterativ verbessern. Das Problem dabei ist, dass zwar prinzipiell zahlreiche Stabilitätskriterien existieren, die die Wurzelberechnung umgehen (Bezout, Hermite [394], Hermite-Biehler, Routh, Routh-Hurwitz, Liénard-Chipart [382], Markov [382], Wall-Frank [617], [199]), die sich zur numerischen Prüfung gut eignen, die jedoch nicht direkt parametrische Restriktionen liefern. Werden Restriktionen abgeleitet, sind diese, abgesehen von den Spezialfällen erster und zweiter Ordnung, unschön. So liefern die Kriterien für die Originalparameter eine semialgebraische Menge[19] als Restriktion, also eine nichtkonvexe, unbeschränkte, offene Menge.

2.3.1 Polynomiale Ungleichungen

Die Stabilität einer LTI-Differenzialgleichung wird anhand ihres charakteristischen Polynoms und die einer Übertragungsfunktion anhand des Nennerpolynoms überprüft. Ist das jeweilige Polynom ein Hurwitz-Polynom, dann ist das das betreffende LTI-System exponentiell stabil, ist es ein einfaches Hurwitz-Grenzpolynom, dann sind die zugehörigen LTI-Zustandsraumdarstellungen Lyapunov-stabil.

Definition 2.1 (Hurwitz-Polynom)
$a(s) = \sum_{i=0}^{n} a_i s^i$ heißt Hurwitz-Polynom, wenn alle Wurzeln in der linken offene komplexen Halbebene, d. h. $\Re e\, s_i < 0$ für alle $i = 1, \ldots, n$ gilt. Es heißt Hurwitz-Grenzpolynom, wenn $\Re e\, s_i \leq 0$ für alle $i = 1, \ldots, n$ gilt, und es heißt einfaches Hurwitz-Grenzpolynom, wenn das Grenzpolynom nur einfache Wurzeln auf der imaginären Achse hat.

[18] Die Verfahren „Zustandsvariablenfilter-Methode“, „Poisson-Momentenfunktional-Methode“, „Block-Puls-Methode“ und „Lineare- Integralfilter-Methode nach Sagara“ überführen das Problem durch eine Filterung der Messdaten in ein LS-Problem in den Originalparametern. Umfangreiche Vergleiche der Methoden finden sich in [247], wo das Zustandsvariablenfilter empfohlen wird.

[19] Eine Menge heißt semialgebraisch, wenn sie sich durch endlich viele Polynomungleichungen darstellen lässt.

Notwendige Bedingung für ein Hurwitz-Polynom: Alle Koeffizienten haben das gleiche Vorzeichen, was für $a(s)$ aus der Faktorisierung

$$a(s) = \alpha \prod_{i=1}^{k}(1 + sT_i) \prod_{j=1}^{l}(1 + 2D_jT_js + T_j^2s^2); \qquad \begin{array}{l} T_i, T_j, D_j > 0; \alpha \neq 0; i = 0, \ldots, n \\ 2l + k = n \end{array} \tag{2.64}$$

folgt. Nur für $n = 1, 2$ ist diese notwendige Bedingung auch hinreichend. Als Gegenbeispiel für $n = 3$ sei $a(s) = s^3 + s^2 + s + 1$ genannt.
Problematischer ist es, notwendige und hinreichende Bedingungen für $n > 2$ zu finden, die sich algorithmisch noch einigermaßen handhaben lassen. Aus dem Liénard-Chipart-Kriterium folgen derartige Bedingungen für $a_n > 0$ [209]:

$$\begin{array}{ll} a_0, a_1 > 0 & \text{1. Ordnung} \\ a_0, a_1, a_2 > 0 & \text{2. Ordnung} \\ a_0, a_2, a_3 > 0;\ a_1a_2 - a_0a_3 > 0 & \text{3. Ordnung} \\ a_0, a_1, a_2, a_4 > 0;\ a_1a_2a_3 - a_1^2a_4 - a_0a_3^2 > 0 & \text{4. Ordnung} \\ a_0, a_2, a_4, a_5 > 0;\ a_3a_4 - a_2a_5 > 0 & \text{5. Ordnung} \\ 2a_0a_1a_4a_5 + a_0a_2a_3a_5 + a_1a_2a_3a_4 - a_0a_3^2a_4 - a_0^2a_5^2 - a_1a_2^2a_5 - a_1^2a_4^2 > 0 & \end{array} \tag{2.65}$$

Ein Maß dafür, wie dicht ein monisches Hurwitz-Polynom $a(s)$ an der Stabilitätsgrenze liegt, liefert der reelle Stabilitätsradius für Polynome

$$r_{a(s)} \stackrel{\text{def}}{=} \min\{||a - x||_2 : x \in \mathbb{R}^n, \max_i \Re e\, s_i \geq 0 \text{ mit } x(s_i) = 0\}, \tag{2.66}$$

wobei a und x die Koeffizientenvektoren von $a(s)$ und $x(s)$ darstellen [241], [240]. Das Maß liefert bezüglich eines nominalen (erwarteten) $a(s)$ eine Auskunft, ob eventuell bei der Identifikation auf Restriktionen bezüglich der Stabilität verzichtet werden kann. Zugleich lässt sich ein Anhaltspunkt für die Größe von Intervallrestriktionen ableiten, vgl. Abschn. 2.7. Für Anwendungen in der robusten Regelungstechnik sollte indes der komplexe Stabilitätsradius (Optimierung über $x \in \mathbb{C}^n$) herangezogen werden, vgl. auch die Anmerkungen zum komplexen Stabilitätsradius für Matrizen.

2.3.2 Reparametrisierung

Für Ausgangsfehlerzugänge, die per Optimierung offline gelöst werden, ist die folgende Reparametrisierung mit einfachen Ungleichungsrestriktionen vorteilhaft, vgl. [247]

$$a(s) = \begin{cases} c_0 \prod_{i=1}^{n/2}(s^2 + c_is + c_{n/2+i}) & n = 2k \\ c_0(s + c_n) \prod_{i=1}^{(n-1)/2}(s^2 + c_is + c_{(n-1)/2+i}) & n = 2k + 1 \end{cases} \tag{2.67}$$

$$\begin{array}{ll} \text{mit } c_i > 0;\ i = 1, \ldots, n & \text{Hurwitz-Polynom} \\ \text{mit } c_i \geq 0;\ i = 1, \ldots, n & \text{Hurwitz-Grenzpolynom.} \end{array}$$

Die Faktorisierung in überwiegend quadratische Faktoren lässt konjugiert komplexe Wurzeln zu, ohne ins Komplexe ausweichen zu müssen, und hält die Restriktionen einfach. Die Restriktionen an die c_i sollten direkt behandelt und nicht per Substitution $c_i := d_i^2$ eliminiert werden (bessere Topologie über c_i).

Nachteilig an der angegebenen Parametrisierung ist deren Nichteindeutigkeit, die aus der Vertauschbarkeit der Produkte und damit der Parameter resultiert. Zudem können Einfachwurzeln in unterschiedlicher Weise zu den quadratischen Termen zusammengefasst werden. Folglich besitzt das Gütegebirge zahlreiche isolierte globale Minima und damit viele Sättel.

Eine bisher wenig beachtete Reparametrisierung stellen die vorzeichenmodifizierten Markov-Parameter[20] m_i von

$$\frac{q(s)}{p(s)} = m_0 + \sum_{i=1}^{\infty} (-1)^{i-1} \frac{m_i}{s^i} \tag{2.68}$$

dar, wobei $a(s)$ in die Form $a(s) = p(s^2) + sq(s^2)$ gebracht wird. Zwischen den n Koeffizienten a_i eines monischen Polynoms und den n ersten Markov-Parametern besteht eine Bijektion. Hurwitz-Stabilität von $a(s)$ liegt genau dann vor, wenn die Hankel-Matrizen aus den m_i positiv definit sind [209]. Da die positiv-definiten Matrizen einen konvexen Kegel formen und die Hankel-Matrizen einen linearen Raum, ist die Menge der Hurwitz-Polynome im Raum der Markov-Parameter konvex![21] Das bringt viele Vorteile. So sind LMI-Formulierungen der Restriktionen möglich und auch die Arbeit mit projizierten oder reduzierten Gradienten, die Berechnung zulässiger Richtungen für den m-Parametervektor und der Einsatz von Barriere-Funktionen [93] vereinfacht sich.

Handelt es sich bei der Stabilitätsrestriktion um das Nennerpolynom einer SISO-LTI-Übertragungsfunktion, kann auch das gesamte SISO-System reparametrisiert werden, was über Normalformen gelingt. Dabei sollte die Normalform einfache Restriktionen für die Stabilität liefern und die Menge der Normalformen sollte zusammenhängend sein. Letztgenanntes Argument spricht gegen die balancierte Normalform [133], bei der zwar im generischen Fall disjunkter Hankel-Singulärwerte die Restriktion mit $\sigma_1 > \sigma_2 > \ldots > \sigma_n > 0$ einfach ist, dafür aber 2^n unterschiedliche Fälle zu betrachten sind. Bei der Jordan-Form können zwar durch eine lexikografische Ordnung der Eigenwerte permutationsbedingte globale isolierte Minima unterdrückt werden, doch erfordert die Segre- bzw. Weyl-Charakterisierung der Jordan-Blöcke die Behandlung mehrerer Fälle [161]. Zudem existiert kein numerisch stabiler Algorithmus zur Berechnung der Jordan-Form, deren Parametrisierung überdies unstetig ist[22]. Ähnliche Argumente sprechen gegen andere eigenwertorientierte Normalformen, wie die reelle Jordan-Form, die Vandermonde-Form [649], die rationale Form und die

[20] Als Markov-Parameter werden gemeinhin die Koeffizienten der s^{-i} bezeichnet, also $\tilde{m}_i = (-1)^{i-1} m_i$.

[21] Für Beziehungen zum Kharitonov-Theorem sei auf [297] verwiesen.

[22] $A = \left[\begin{smallmatrix} \varepsilon & 0 \\ 1 & 0 \end{smallmatrix}\right]$ hat die Jordan-Form $\left[\begin{smallmatrix} 0 & 0 \\ 0 & \varepsilon \end{smallmatrix}\right]$, während $A = \left[\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right]$ die Jordan-Form $\left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$ hat.

Elementarteilerform (rationale kanonische Form) [299]. Von den in [89] unter dem Transformationsblickwinkel betrachteten Normalformen vom Spaltentyp (Steuerbarkeitsnormalformen)[23] – Frobenius-Normalform (Luenberger-Normalform), Tridiagonalform, Schwarz-Normalform, Routh-Normalform, Subdiagonalform, Cauer-Normalform – eignet sich besonders die Schwarz-Normalform

$$A = \begin{bmatrix} 0 & -p_{n-1} & & 0 \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & -p_1 \\ 0 & & 1 & -p_0 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad c^T = (c_1, \ldots, c_n), \qquad d = d. \tag{2.69}$$

Dabei ist $-A$ eine sog. Schwarz-Matrix und für die gilt, dass für $p_i > 0; i = 0, \ldots, n-1$ alle Eigenwerte in der rechten offenen komplexen Halbebene liegen [558], womit A folglich stabil ist. Die Parameter p_i sind durch

$$p_0 = \det H_1, \;\; p_1 = \frac{\det H_2}{\det H_1}, \;\; p_2 = \frac{\det H_3}{(\det H_1)(\det H_2)}, \;\; p_{3,\ldots,n-1} = \frac{(\det H_{i-2})(\det H_{i+1})}{(\det H_{i-1})(\det H_i)} \tag{2.70}$$

bestimmt, wobei $\det H_i$ die i-ten Hurwitz-Determinanten von $a(s)$ sind [209]. Für Differenzgrade $r \geq 2$ ist $d = 0$ und $c_1 = \ldots = c_{r-1} = 0$ zu setzen. Per Ähnlichkeitstransformation mit

$$S = (s_1, \ldots, s_n) \qquad \text{mit} \;\; s_1 = b, s_2 = Ab, s_i = As_{i-1} + p_{n-i+2}s_{i-2}; i = 3, \ldots, n \tag{2.71}$$

lässt sich jedes steuerbare Zustandsraummodell in die Schwarz-Normalform überführen [126], [89]. Bei nichtsteuerbaren, aber beobachtbaren Modellen ist die Schwarz-Normalform vom Zeilentyp zu verwenden.

Da die Reparametrisierung nur ein Teilschritt bei der Offline-Identifikation mit Ausgangsfehlerzugängen ist, wird nachfolgend die Grundstruktur eines solchen Algorithms am Beispiel der Schwarz-Normalform kurz skizziert.

Ausgangsfehlerorientierter Algorithmus mit Reparametrisierung:
- Reparametrisierung des Startmodells $G(s) = b(s)/a(s)$ in p_i, c_i mit (2.70) und (2.71)
- Diskretisierung mit Halteglied nullter Ordnung
- Berechnung $\hat{y}(k; p, c, d)$ und des Zielfunktionswerts, z. B. $\sum_{k=0}^{N-1}(\hat{y}[k] - y[k])^2$
- Korrektur der Parameter unter Einhaltung von $p_i > 0$ für $i = 0, \ldots, n-1$
- Fortsetzen der Iteration bis Abbruch
- Rückparametrisierung: $\hat{G}(s) = \hat{b}(s)/\hat{a}(s) = c_{\text{opt}}^T(sI_n - A(p_{\text{opt}}))^{-1}b + d_{\text{opt}}$

[23] Steuerbarkeitsnormalformen vom Zeilentyp, wie auch die Beobachtbarkeitsnormalformen vom Zeilen- bzw. Spaltentyp ergeben gleichwertige Parametrisierungen.

2.4 Übertragungsstabilität für lineare zeitdiskrete Systeme

Die Schätzungen stabiler zeitdiskreter Systeme über Gleichungsfehlerzugänge liefern nicht zwingend stabile Modelle [571]. Mögliche Gründe sind der statistisch bedingte Bias, der Approximationsfehler und der zufällige Fehler für die konkrete Realisierung. Auch eine schlechte Konditionierung der Datenmatrizen infolge zu schneller Abtastungen kann die Ursache sein. Empfehlungen zur Wahl einer zweckmäßigen Abtastzeit finden sich in [314]. Zu den Auswirkungen einer schlechten Konditionierung zählt eine erhöhte Parameterempfindlichkeit bezüglich der Daten, was in Verbindung mit der oft ohnehin hohen Empfindlichkeit der Wurzeln bezüglich der Polynomparameter (s. Bsp. 2.7) das Identifizieren eines ungewollt instabilen Modells begünstigt.

Besonders wichtig ist die Sicherung der Stabilität, wenn das Modell für interne Simulationen benötigt wird, da in diesem Fall selbst eine asymptotische Garantie für ein stabiles Modell nicht ausreicht. Auch für adaptive Regelungen erweist sich die Stabilitätssicherung des identifizierten Modells als zweckmäßig. Sie verhindert nämlich unnötig kleine Reglerverstärkungen und die damit einhergehende schlechte Reglerperformance.

Bekanntermaßen wird die Stabilität von Differenzengleichungsmodellen durch die Lage der Wurzeln des Ausgangspolynoms bezüglich des Einheitskreises bestimmt. Die Polynome erhielten dabei eigene Namen.

Definition 2.2 (Schur- und von-Neumann-Polynom)
Ein Polynom heißt diskret-stabil oder Schur-Polynom, wenn seine Wurzeln innerhalb des Einheitskreises liegen. Es heißt von-Neumann-Polynom, wenn sich die Wurzeln im oder auf dem Einheitskreis befinden, und einfaches von-Neumann-Polynom, wenn die Wurzeln auf dem Einheitskreis einfach sind.

Aus der Definition folgt unmittelbar die Implikation

$$\text{Schur-Polynom} \quad \Rightarrow \quad \text{einfaches von-Neumann-P.} \quad \Rightarrow \quad \text{von-Neumann-Polynom.}$$

Weitere Implikationen, die die Stabilität eines zeitdiskreten Systems in Verbindung zum charakteristische Polynom der Systemmatrix setzen, gibt die nachfolgende Übersicht:

$$\begin{aligned} &\text{Schur-Polynom} \Leftrightarrow \text{exponentiell stabil} \Leftrightarrow \text{asymptotisch stabil} \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{Schur-stabil}) \\ &\text{einfaches von-Neumann-Polynom} \Rightarrow \text{Lyapunov-stabil} \\ &\text{von-Neumann-Polynom} \Leftarrow \text{Lyapunov-stabil} \\ &\text{von-Neumann-Polynom} \Rightarrow \text{generisch-stabil} \end{aligned}$$

Anmerkung 2.13 Der Defekt in der Schlussweise hat folgende Ursache. Satz: Eine Matrix ist genau dann eine nichtderogatorische Matrix (geometrische Vielfachheit jedes Eigenwerts ist Eins), wenn sie zur Begleitmatrix ihres charakteristischen Polynoms ähnlich ist [299]. Folglich korrespondieren zu Polynomen (bzw. Differenzengleichungen) nur nichtderogatorische Matrizen. Im Fall einer Mehrfachwurzel auf dem Einheitskreis ist der betreffende Eigenwert dann defektiv und das System somit instabil. Matrizen mit mehrfachen Eigenwerten auf dem Einheitskreis hingegen müssen nicht zwingend instabil sein, vgl. das Lyapunov-stabile System $x[k+1] = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right] x[k]$ und das instabile $x[k+1] = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right] x[k]$. Beide besitzen das charakteristische Polynom $c(\lambda) = \lambda^2 - 2\lambda + 1$.

Da ebenso wie im zeitkontinuierlichen Fall, die Wurzeln in komplizierter Weise von den Polynomkoeffizienten abhängen, wird sich bei einer Gleichungsfehleridentifikation auf notwendige Bedingungen (zweckmäßigerweise mit Restriktionen an die Einhaltung des Abtasttheorems kombiniert) beschränkt, oder es werden bei Ausgangsfehleridentifikationen Reparametrisierungen genutzt, vgl. Abschnitt 2.2.6.

Eine Alternative bietet der Umweg über die Identifikation kontinuierlicher Modelle mit der Zustandsvariablenfilter-Methode oder vergleichbaren Methoden [247]. Das Stabilitätsgebiet ist dann größer (linke Halbebene statt Einheitskreis), die Parameteranzahl ist meist kleiner, die Schätzprobleme sind bei kleinen Abtastzeiten besser konditioniert, die Restriktionen sind etwas einfacher und die Modelle können einfach auf beliebige Abtastzeiten umgerechnet werden. Die Umrechnung eines z-Modells für eine andere Abtastzeit ist problembehaftet, da die $\mathfrak{Z}$-Transformation nicht bijektiv ist.

Obwohl gerade für die Gleichungsfehleridentifikation die Situation zunächst recht unbefriedigend erscheint, da sich auf notwendige Bedingungen begrenzt werden muss (andernfalls gehen die algorithmischen Vorteile verloren), verbessert sie sich asymptotisch. Erfüllen nämlich die Eingangssignale die im folgenden Satz genannten Bedingungen, so ist Stabilität asymptotisch garantiert.

Satz 2.6 (Stabilität von Gleichungsfehlermodellen, [412])
Wird das stabile System $G(z) = \frac{b_0 + b_1 z^{-1} + \ldots + b_m z^{-m}}{a_0 + a_1 z^{-1} + \ldots + a_n z^{-n}}$ mit einem stationären AR-Prozess $\boldsymbol{u}[k]$ vom Grad $p \leq m+1$ erregt und ist $\boldsymbol{v}[k]$ eine stationäre, hinsichtlich des Spektrums nicht spezifizierte Störung, die unkorreliert mit $\boldsymbol{u}[k]$ ist, dann liefert für $\boldsymbol{y}[k] = G(\mathrm{q})\boldsymbol{u}[k] + \boldsymbol{v}[k]$ das Kriterium[24]

$$\mathrm{E}\left\{ \left((a_0 + a_1 \mathrm{q}^{-1} + \ldots + a_n \mathrm{q}^{-n})\, \boldsymbol{y}[k] - (b_0 + \ldots + b_m \mathrm{q}^{-m})\, \boldsymbol{u}[k] \right)^2 \right\} \stackrel{!}{=} \mathrm{Min} \qquad a_0 = 1 \quad (2.72)$$

ein Schur-stabiles Modell. Wird die Restriktion durch $\sum_{i=0}^{n} a_i^2 = 1$ ersetzt, dann liegen keine Pole des Modells außerhalb des Einheitskreises.

[24] $\mathrm{E}\{\,.\,\}$ bezeichnet hierbei den Erwartungswertoperator.

Aus diesem Satz kann gefolgert werden, dass bei Erregung mittels ergodischer AR-Prozesse (solche sind stationär; die Umkehrung gilt nicht) das Modell bei Schätzung über

$$\sum_{k=n}^{N} \Big((a_0 + a_1 \mathrm{q}^{-1} + \ldots + a_n \mathrm{q}^{-n})\, y[k] - (b_0 + \ldots + b_m \mathrm{q}^{-m})\, u[k] \Big)^2 \stackrel{!}{=} \mathrm{Min} \qquad a_0 = 1 \quad (2.73)$$

asymptotisch die Stabilitätseigenschaft des Systems annimmt. Im Speziellen gilt dies für stationäre Gaußsche AR-Prozesse mit $p \leq m + 1$, da diese ergodisch sind.
Der Satz gibt ferner eine zusätzliche, weitgehend unbeachtete Begründung dafür, warum weißes Rauschen, Pseudorauschbinärsignale und -ternärsignale so nützliche Testsignale sind. Bei ihnen ist nämlich wegen $p = 0$ die Voraussetzung unabhängig von m erfüllt.

Während in diesem und den vorherigen Abschnitten die Sicherung der Übertragungsstabilität betrachtet wird, geht es nachfolgend um die Stabilitätssicherung für Zustandsraummodelle. Die Fragestellung tritt dabei insbesondere bei der Modellreduktion auf. Für die klassische Identifikation ist sie weniger von Bedeutung, da abgesehen von den Unterraumverfahren [490] die Mehrzahl der Zugänge auf E/A-Modellansätzen basiert. Eine wichtige Ausnahme bildet die Identifikation von Zustandsraummodellen mit messbaren Zustandsgrößen (z. B. interne Temperaturen). Der folgende Abschnitt fasst zunächst die Grundlagen kurz zusammen und stellt danach einige Zugänge zur Behandlung der Stabilitätsrestriktionen vor.

2.5 Stabilität für lineare zeitkontinuierliche Zustandsraumsysteme

Die Stabilität von LTI-Systemen im Zustandsraum wird ausschließlich durch die Systemmatrix bestimmt. Deshalb werden die Systemmatrizen namensgleich zur entsprechenden Stabilität der Lösungen definiert. Die Charakterisierung der Matrizen erfolgt aber rein über die Eigenwerte und nicht über das Lösungsverhalten. Somit liefert die Definition per se Restriktionen an die Eigenwerte.

Definition 2.3 (Stabile Matrix)
Eine Matrix $A \in \mathbb{C}^{n\times n}$ heißt stabil oder auch Hurwitz-stabil[25], wenn alle Eigenwerte einen negativen Realteil besitzen. Sie heißt stark stabil, wenn sogar $\lambda_i(A + A^T) < 0, \forall i$ gilt.[26] Sie heißt neutralstabil oder auch Lyapunov-stabil, wenn alle Eigenwerte nichtpositiv sind und jene auf der imaginären Achse nicht defektiv[27] sind. Hat eine Matrix diese Eigenschaft nicht, wird sie Lyapunov-instabil genannt. Matrizen, deren Eigenwerte alle nichtpositiv sind, ohne dass sie bezüglich der Nichtdefektivität der rein imaginären Eigenwerte spezifiziert werden, heißen generisch-stabil. Ferner wird A für $\gamma > 0$ als γ-Lyapunov-stabil (γ-generisch-stabil) bezeichnet, wenn $A + \gamma I_n$ Lyapunov-stabil (γ-generisch-stabil) ist. γ fungiert als Stabilitätsreserve.

Aus der Definition ergeben sich unmittelbar die folgenden Implikationen:

$$\begin{array}{ccccccc} \gamma\text{-Lyapunov-stabil} & \Rightarrow & \text{Hurwitz-stabil} & \Rightarrow & \text{Lyapunov-stabil} & \Rightarrow & \text{generisch-stabil.} \\ \Downarrow & & \Uparrow & & & & \\ \gamma\text{-generisch-stabil} & & \text{stark stabil} & & & & \end{array} \tag{2.74}$$

Eine Formulierung der Stabilitätsrestriktion über die Eigenwerte ist aus Sicht der Optimierung ungünstig. Es besteht nämlich zwischen den Matrixelementen und den Eigenwerten ein komplizierter nichtlinearer, nichtkonvexer, nichtglatter Zusammenhang, der ab fünfter Ordnung i. Allg. auch nicht mehr in geschlossener Form angegeben werden kann. Computeralgebrasysteme scheitern dementsprechend ab fünfter Ordnung und auch die numerische Behandlung mit Gradientenverfahren ist schwierig, da das Ableiten der Matrixelemente nach den Eigenwerten jeweils eine komplette Spektralfaktorisierung erfordert.

[25] Die Bezeichnung „Hurwitz-stabil“ rührt aus der Tatsache, dass das charakteristische Polynom einer solchen Matrix ein Hurwitz-Polynom ist. Die im Hurwitz-Kriterium verwendete Matrix wird dagegen Hurwitz-Matrix genannt.

[26] Der Zusatz „stark“ rührt aus der Eigenschaft, dass nicht nur $||x(t)|| \to 0$, sondern auch $\frac{\mathrm{d}||x(t)||}{\mathrm{d}t} < 0$ gilt. Die Norm fällt also streng monoton, s. Abschnitt 2.5.5 für stabile Systeme, wo das nicht der Fall ist.

[27] Ein k-facher Eigenwert heißt defektiv, wenn es zu ihm keine k linear unabhängigen Eigenvektoren gibt.

Ziel der folgenden Unterabschnitte ist es, Zugänge für die Lösung des LS-Problems

$$||FA - G||_F^2 \stackrel{!}{=} \text{Min} \qquad A \in \mathbb{R}^{n\times n} \text{ stabil in einem Sinne} \tag{2.75}$$

mit den Datenmatrizen F und G vorzustellen. Es ist trivial, diese auf

$$||AF - G||_F^2 \stackrel{!}{=} \text{Min} \qquad A \in \mathbb{R}^{n\times n} \text{ stabil in einem Sinne} \tag{2.76}$$

zu übertragen. In Fällen, in denen die freien Lösungen $A = F^+G$ bzw. $A = GF^+$ die Stabilitätsrestriktion einhalten, gilt das Problem als gelöst. Deshalb wird unterstellt, dass die freien Lösungen die geforderte Stabilität nicht erfüllen. Aufgrund der Nichtkonvexität des Suchraums kann von den nachfolgenden Zugängen nicht erwartet werden, dass sie globale Minimierer (2.75) und (2.76) liefern. Im Speziellen werden betrachtet:

1. Zugang über das Lyapunov-Theorem (Abschn. 2.5.1)
2. Zugang im Fall von Diagonaldominanz (Abschn. 2.5.2)
3. Zugang über den $\mathbb{R}$-Stabilitätsradius (Abschn. 2.5.3)
4. Zugang über stabilitätsinvariante Umformungen (Abschn. 2.5.4)

Den Abschluss der Betrachtungen zur Stabilität im Zustandsraum bildet das Konzept der praktischen Stabilität in Abschn. 2.5.5. Es kann gewissermaßen als ein Gegenpol zur gewöhnlichen Stabilität der LTI-Systeme angesehen werden. Während mit der gewöhnlichen Stabilität Aussagen zum asymptotischen Verhalten einhergehen, bezieht sich die praktische Stabilität vordergründig auf das transiente Verhalten. Im Abschnitt 2.5.5 geht es dabei weniger um Restriktionen für die praktische Stabilität als vielmehr um den Nutzen des Konzepts für die A-posteriori-Modellbewertung.

2.5.1 Zugang über das Lyapunov-Theorem

Die Lyapunov-Theoreme für LTI-Systeme gestatten Stabilitätsaussagen ohne Berechnung der Eigenwerte. Es gibt sie in der Gleichungs- und Ungleichungsform.

Satz 2.7 (Lyapunov-Theorem in Gleichungsform, [614])
Folgende Aussagen sind äquivalent[28]:
(i) A ist Hurwitz-stabil
(ii) $\exists Q \in \mathcal{S}_n^> : \; AP + PA^T = -Q \Rightarrow P \in \mathcal{S}_n^>$, P eindeutig
(iii) $\forall Q \in \mathcal{S}_n^> : \; AP + PA^T = -Q \Rightarrow P \in \mathcal{S}_n^>$, P eindeutig.

[28] Die gleichartige Darstellung $A^T\tilde{P} + \tilde{P}A = -\tilde{Q}$ beruht auf der Lyapunov-Funktion $V(x) = x^T\tilde{P}x$, während für Satz 2.7 $V(x) = x^TP^{-1}x$ zusammen mit der Links-Rechtsmultiplikation mit P genutzt werden kann.

Anmerkung 2.14 Ist für ein beliebig gewähltes $Q \in \mathcal{S}_n^>$, z. B. $Q = I_n$, die Lösung P eindeutig und positiv definit, ist A Hurwitz-stabil. Existiert keine Lösung P, unendlich viele Lösungen oder eine eindeutige, aber nicht positiv definite Lösung, dann ist A nicht Hurwitz-stabil. Beachte aber, dass $P \in \mathcal{S}_n^>$ und A Hurwitz-stabil nicht $Q \in \mathcal{S}_n^>$ impliziert.

Anmerkung 2.15 Das Lyapunov-Theorem lässt sich auf Systeme $E\dot{x} = Ax$ verallgemeinern, wobei dann die Gleichung $APE^T + EPA^T = -Q$ zu betrachten ist.[29]

Satz 2.8 (Lyapunov-Theorem in Ungleichungsform, [92])
$A \in \mathbb{R}^{n\times n}$ ist genau dann Lyapunov-stabil, wenn

$$P \succ 0_{n\times n}, \quad AP + PA^T \preceq 0_{n\times n} \tag{2.77}$$

zulässig ist[30], d. h. $\exists P \in \mathcal{S}_n^> : AP + PA^T \preceq 0_{n\times n}$. $A \in \mathbb{R}^{n\times n}$ ist genau dann Hurwitz-stabil, wenn (2.77) streng zulässig ist, d. h. $\exists P \in \mathcal{S}_n^> : AP + PA^T \prec 0_{n\times n}$.[31]

Der Vorteil, keine Restriktionen an die Eigenwerte mehr zu haben, wird sich durch $\frac{n(n+1)}{2}$ zusätzliche Parameter (freie Parametern von P) und zwei Matrixungleichungen (2.77) erkauft und führt auf das folgende Prokrustes-Problem

$$||FA - G||_F^2 \stackrel{!}{=} \text{Min} \qquad P \succ 0_{n\times n},\ AP + PA^T \preceq 0_{n\times n}. \tag{2.78}$$

Modifikationen zur γ-Lyapunov-Stabilität ergeben sich aus der Äquivalenz

$$A \text{ ist } \gamma\text{-Lyapunov-stabil} \Leftrightarrow \exists P \in \mathcal{S}_n^> : AP + PA^T + 2\gamma P \preceq 0_{n\times n}. \tag{2.79}$$

Die Zielfunktion ist konvex bezüglich A und die Ungleichung $P \succ 0_{n\times n}$ beschreibt eine konvexe Menge (Kegel der nichtnegativ definiten Matrizen). Beide harmonieren wegen ihrer Konvexität. Die Lyapunov-Ungleichung ist aber bilinear und zerstört somit die konvexe Formulierung. Allerdings ist bilinear zwar nichtlinear, aber „weit weniger nichtlinear“ als der Zusammenhang zwischen den Matrixelementen und Eigenwerten. Zudem existiert zur Lösung derartiger Probleme spezielle kommerzielle Software [359]. Um sie zu nutzen, sind die Zielfunktion und die Restriktionen auf die Standardform

$$\begin{aligned} &\frac{1}{2}x^T Hx + g^T x \stackrel{!}{=} \text{Min} \qquad && \sum_{k=1}^{n} b_{k,i}x_{k,i} \le c_i; i = 1,\dots,m_l \\ & && B_{0,i} + \sum_{k=1}^{n} x_k B_{k,i} + \sum_{k=1}^{n}\sum_{l=1}^{n} x_k x_l C_{kl,i} \preceq 0_{p_i\times p_i}; i = 1,\dots,m \\ & && B_{k,i}, C_{kl,i} \in \mathcal{S}_{p_i} \end{aligned}$$

[29] Nutze $A^T PE + E^T PA = -Q$ [66] und Dualitätsprinzip, d. h. $A := A^T$ und $E := E^T$, s. Fußnote 32.

[30] Duale Formulierung: A Lyapunov-stabil $\Leftrightarrow \tilde{P} \succ 0_{n\times n}, \tilde{P}A + A^T\tilde{P} \preceq 0_{n\times n}$ zulässig. Folgt über $P = \tilde{P}^{-1}$. Beweist $V(x) = x^T Px$ Stabilität von $\dot{x} = Ax$, dann beweist $V(x) = x^T P^{-1}x$, die von $\dot{x} = A^T x$.

[31] Die Menge der Matrizen $\mathcal{P} = \{P \in \mathcal{S}_n^> : AP + PA^T \prec 0_{n\times n}\}$ ist konvex und heißt Lyapunov-Kegel.

zu bringen. Per Vektorisierung von A und symmetrischer Vektorisierung von P sowie entsprechender Wahl der Matrizen $B_{k,i}$, $C_{kl,i}$ gelingt das für die zwei Restriktionen in (2.78). Hinweis: Zur Konstruktion der Matrizen $B_{k,i}$, $C_{kl,i}$ empfiehlt es sich, von der Darstellung der Vektorräume $\mathbb{R}^{n\times n}$ bzw. $\mathcal{S}_n$ über kanonische Basismatrizen Gebrauch zu machen, z. B.

$$\mathcal{S}_3 = \left\{ p_1 \begin{bmatrix} 1&0&0\\0&0&0\\0&0&0 \end{bmatrix} + p_2 \begin{bmatrix} 0&0&0\\0&1&0\\0&0&0 \end{bmatrix} + p_3 \begin{bmatrix} 0&0&0\\0&0&0\\0&0&1 \end{bmatrix} + p_4 \begin{bmatrix} 0&1&0\\1&0&0\\0&0&0 \end{bmatrix} + p_5 \begin{bmatrix} 0&0&1\\0&0&0\\1&0&0 \end{bmatrix} + p_6 \begin{bmatrix} 0&0&0\\0&0&1\\0&1&0 \end{bmatrix} : p_1, \ldots, p_6 \in \mathbb{R} \right\}.$$

Als Nachteile dieses Zugangs bleiben der große numerische Aufwand, die Verfügbarkeit einer speziellen Software und die Eingeschränktheit auf LMI-Restriktionen.
Die letzten beiden Nachteile lassen sich durch den Einsatz allgemeiner numerischer Algorithmen für nichtlineare Probleme umgehen. Dafür ist allerdings eine Modifikation des Problems ratsam. So empfiehlt es sich, die Ungleichung $P \succ 0_{n\times n}$ durch $P^{-1} \succ 0_{n\times n}$ zu ersetzen. Dadurch wird verhindert, dass Elemente in P gegen Unendlich streben, wenn der im Realteil größte Eigenwert der stabilen Matrix gegen Null strebt. Weiterhin ist meist die Behandlung durch eine Lyapunov-Gleichung $A^T P + PA + I_n = 0$ zweckmäßiger als durch eine Ungleichung. Letztlich kann die Definitheitsrestriktion als Symmetrierestriktion $P = P^T$ und $k = 1, \ldots, n$ Determinanten-Restriktionen $\det(P^{-1}[k|k]) > 0$ (Sylvester-Kriterium) gefasst werden, wobei $\det(P^{-1}[k|k]) \geq \varepsilon$ mit z. B. $\varepsilon = 10^{-6}$ die strengen Ungleichungen umgeht[32]. Eine Anwendung für diesen Zugang ist die Bestimmung eines optimalen, stabilen Arbeitspunkts bezüglich der Zustände, bei dem A die Jacobi-Matrix ist, vgl. [81][33].

2.5.2 Zugang im Fall von Diagonaldominanz

Eine besonders vorteilhafte Situation liegt vor, wenn die Systemmatrix zeilen- oder spaltendiagonaldominant ist, wenn also jedes Hauptdiagonalelement betragsmäßig größer als die Summe der Beträge aller weiteren Elemente der Zeile oder Spalte ist. In der Praxis treten Systeme mit der Diagonaldominanzeigenschaft bevorzugt auf, wenn es in nahezu gleichartigen gekoppelten Zonen (Zellen, Elementen) zu Ausgleichsvorgängen kommt. Die Diagonaldominanz leitet sich dann aus den Massen- oder Energiebilanzen her, wobei Verlustterme auf die strenge Ungleichheit führen. Beispiele sind Mehrzonenöfen [255], bei denen der Aus-

[32] Die Wahl numerischer Schranken ε, um Tests auf Gleichheit oder Ungleichheit bezüglich Null vorzunehmen (Vermeidung Nulldivision oder negativer Wurzeln) und um aus strengen Ungleichungen nichtstrenge zu machen, erfordert Erfahrungen oder auch empirische Untersuchungen. Sie hängt neben dem Zweck von der Leistungsfähigkeit der eingesetzten Algorithmen, der geforderten Genauigkeit und der Problemgröße ab. Mit dem hier angegebenen Wert arbeitet der Autor vorwiegend bei Problemen mit $n < 10$ und in eigenentwickelten Algorithmen, weil damit zumeist eine praktisch ausreichende Genauigkeit erzielt wird.

[33] In [81] werden die Determinanten-Restriktionen mittels einer logarithmischen Barrierefunktion relaxiert; in [80] wird zur Lösung des Problems der hier dargestellte Weg empfohlen.

gleichsvorgang bezüglich der Energie stattfindet, oder der menschliche Körper, in dem sich Konzentrationen z. B. von Markersubstanzen oder Radioisotopen [325] ausgleichen.

Aus dem Satz von Geršhgorin [431] ergeben sich die folgenden Restriktionen:

$$-a_{ii} > \sum_{i \neq j}^{n} |a_{ij}|;\ a_{ii} < 0;\ i = 1, \ldots, n \qquad \Rightarrow A \text{ stabil} \tag{2.80a}$$

$$-a_{ii} \geq \sum_{i \neq j}^{n} |a_{ij}|;\ a_{ii} \leq 0;\ i = 1, \ldots, n \qquad \Rightarrow A \text{ generisch-stabil} \tag{2.80b}$$

Analog zu dieser Zeilendiagonaldominanzbedingung gibt es eine Spaltendiagonaldominanzbedingung[34]. Der Charme der Diagonaldominanzbedingung liegt in der Einfachheit der Restriktionen, die sich in lineare Restriktionen umformen lassen und somit für die Kriterien (2.75) und (2.76) auf konvexe Probleme führen. Ohne Wissen über das Vorliegen einer Diagonaldominanz sollten die Restriktionen aber nicht verwendet werden, da sie nur hinreichend sind und somit den Suchraum unzulässig einschränken würden.

Neben der Einfachheit der Restriktionen, insbesondere bei Kenntnis der Vorzeichen von a_{ij}, liegt ihr Vorteil bei messbaren Zustandsgrößen in der Möglichkeit einer Systemdekomposition. Für jedes $\dot{x}_i$ (Vektor aus N Messwerten) kann ein separates Schätzproblem unter der jeweiligen Zeilenrestriktion formuliert und gelöst werden, z. B.

$$||\dot{x}_1 - a_{11}x_1 - a_{12}x_2 - \ldots - a_{1n}x_n||_2^2 \stackrel{!}{=} \text{Min} \qquad \begin{array}{l} a_{1j} > 0; j = 2, \ldots, n \\ -a_{11} > a_{12} + \ldots + a_{1n}. \end{array} \tag{2.81}$$

Das Gesamtsystem ist dann trotz Dekomposition automatisch stabil!

Diagonaldominanz lässt sich auch bei Frequenzbereichsschätzungen ausnutzen. Als Beispiel sei die Identifikation von Mehrgrößensystemen genannt. Üblicherweise wird nämlich angestrebt, dass der i-te Eingang auch den i-ten Ausgang am meisten und/oder am schnellsten beeinflusst (Regel: Nebenstrecken sollten langsamer als die Hauptstrecken sein). Bei der Identifikation im geschlossenen Regelkreis rührt die Diagonaldominanz aus der angestrebten partiellen oder vollständigen Entkopplung durch die Regelung her.
Infolge von Installationsfehlern (fehlerhaftes Klemmen von Signalleitungen) oder durch ungünstiges Festlegen der E/A-Größen braucht diese Zuordnung nicht zu stimmen. Durch Permutieren der Eingangsgrößen lässt sie sich aber unter Umständen erzeugen. Hierfür erweisen sich Koppelfaktoren als Zuordnungsmaße ebenfalls als vorteilhaft.

[34] Die Zeilendiagonaldominanzbedingung folgt unabhängig von Geršhgorin aus der Lyapunov-Funktion $V(x) = ||x||_\infty$, während die Spaltendiagonaldominanzbedingung über $V(x) = ||x||_1$ abgeleitet werden kann.

2.5.3 Zugang über den $\mathbb{R}$-Stabilitätsradius

Eine zentrale Idee zur Lösung komplizierter nichtkonvexer Probleme besteht darin, eine Folge konvexer Teilprobleme zu lösen. Hierbei wird ausgehend von der Lösung im k-ten Schritt ein konvexer Suchraum bestimmt, der im Inneren des Originalsuchraums liegt. Dieser sollte natürlich möglichst groß und einfach handhabbar sein, also etwa Polytop, Kugel, Ellipsoid oder Schnitte dieser. Problematisch an diesem Vorgehen ist meist die Berechnung der Ausdehnung des Gebiets; im Fall der Kugel die des Radius. Gegebenenfalls ist in jedem Iterationsschritt eine skalare Optimierung erforderlich, was aber bei ausgefeilten Algorithmen kaum ins Gewicht fällt. Eine Anwendung dieser Idee für Filter mit unendlicher Impulsantwort und damit für Polynome ist in [172] beschrieben. Hier wird das Vorgehen für Matrizen näher erläutert.

Für das LS-Problem

$$||FA - G||_F^2 \stackrel{!}{=} \text{Min} \qquad A \text{ Hurwitz-stabil} \tag{2.82}$$

kann durch sukzessives Lösen konvexer Teilprobleme der Art

$$\begin{aligned} \Delta A_{k+1} = & \operatorname*{argmin}_{||\Delta A||_2 < r_k} ||F(A_k + \Delta A) - G||_F^2 \\ = & \operatorname*{argmin}_{||\Delta A||_2 < r_k} ||F\Delta A - (G - A_k)||_F^2 \end{aligned} \tag{2.83}$$

eine Folge stabiler Matrizen $A_{k+1} = A_k + \Delta A_{k+1}$ mit monoton fallenden Werten $||FA_{k+1} - G||_F^2$ erhalten werden. r_k ist so zu wählen, dass $A_k + \Delta A$ für alle $||\Delta A||_2 < r_k$ stabil ist.

(2.83) gestattet eine SDP-Formulierung und damit den Einsatz von Standardalgorithmen. Vollen Spaltenrang von F vorausgesetzt, hat (2.83) eine eindeutige Lösung.

Der Radius r_k sollte möglichst groß sein, damit der Suchraum für ΔA möglichst groß ist, und r_k sollte einfach zu berechnen sein.

Eine erste, naive Idee bietet die negative Spektralabszisse $r_k = -\alpha(A_k)$ mit

$$\alpha(A) \stackrel{\text{def}}{=} \max_{1 \le i \le n} \Re e\, \lambda_i(A) \qquad \text{Spektralabszisse.} \tag{2.84}$$

Doch $r_k = -\alpha(A_k)$ garantiert mit $||\Delta A||_2 < r_k$ nicht, dass $A + \Delta A$ stabil bleibt, wie folgendes Beispiel zeigt.

Beispiel 2.12 (Spektralabszisse als schlechtes Stabilitätsmaß)
Obwohl $-\alpha(A) = 0.1$ für

$$A = \begin{bmatrix} -0.1 & 1 & 0 & 0 \\ 0 & -0.1 & 1 & 0 \\ 0 & 0 & -0.1 & 1 \\ 0 & 0 & 0 & -0.1 \end{bmatrix}, \qquad \Delta A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 10^{-4} & 0 & 0 & 0 \end{bmatrix}$$

ist, destabilisiert ΔA mit $||\Delta A||_2 = 10^{-4}$ die Matrix A (Nulleigenwert). Zudem ist die Spektralabszisse eine nichtkonvexe, nicht Lipschitz-stetige Funktion der Matrixelemente.

Eine zweite Möglichkeit bietet ein Resultat aus der Lyapunov-Theorie [548]

$$r_k = \frac{1}{2\,||P||_2} \qquad \text{mit} \qquad AP + PA^T = -I_n. \tag{2.85}$$

Da die Lyapunov-Theorie nur hinreichende Bedingungen liefert, kann r_k aber sehr konservativ sein. Bezogen auf Beispiel 2.12 ergibt sich $r_k \approx 3.13 \cdot 10^{-7}$.

Eine dritte Möglichkeit ist der reelle Stabilitätsradius ($\mathbb{R}$-Stabilitätsradius)

$$r_{\mathbb{R}}^{-}(A) \stackrel{\text{def}}{=} \inf\{||\triangle A||_2 : \triangle A \in \mathbb{R}^{n\times n}; A + \triangle A \text{ nicht Hurwitz-stabil}\}. \tag{2.86}$$

Im Beispiel ergibt sich der Wert $r_{\mathbb{R}}^{-}(A) = 10^{-4}$, was übrigens derselbe Wert ist wie bei der vierten Möglichkeit, dem komplexen Stabilitätsradius

$$r_{\mathbb{C}}^{-}(A) \stackrel{\text{def}}{=} \inf\{||\triangle A||_2 : \triangle A \in \mathbb{C}^{n\times n}; A + \triangle A \text{ nicht Hurwitz-stabil}\}. \tag{2.87}$$

Anmerkung 2.16 Für die skizzierte Anwendung kann bestenfalls die einfachere Berechnung des konservativeren $\mathbb{C}$-Stabilitätsradius als Vorteil gewertet werden. Gleichwohl ist der $\mathbb{C}$-Stabilitätsradius als Stabilitätsmaß (Stabilitätsreserve) auch bei reellen Matrizen das bessere Maß! Denn während $r_{\mathbb{R}}^{-}(A)$ nur Unsicherheiten in den Matrixelementen toleriert, liefert $r_{\mathbb{C}}^{-}(A)$ auch Aussagen zur Größe tolerierbarer nichtlinearer oder zeitvarianter Störungen/Unsicherheiten, siehe hierzu Abschnitt A.6.4.

Beide Radien sind Spezialfälle des strukturierten Stabilitätsradius des Tripels $(A, B, C) \in \mathbb{F}^{n\times n} \times \mathbb{F}^{n\times m} \times \mathbb{F}^{p\times n}$ mit $\mathbb{F} := \mathbb{C}$ bzw. $\mathbb{F} := \mathbb{R}$

$$r_{\mathbb{F}}^{\mathcal{S}}(A, B, C) \stackrel{\text{def}}{=} \inf\{||\Delta||_2 : \Delta \in \mathbb{F}^{m\times p}; \lambda_i(A + B\Delta C) \in \mathcal{S}^c\}. \tag{2.88}$$

Hierbei ist $\mathcal{S}$ eine offene Teilmenge in $\mathbb{C}$ und $\mathcal{S}^c$ deren Komplement in $\mathbb{C}$. $r_{\mathbb{F}}^{\mathcal{S}}(A, B, C)$ bietet für die Identifikation einige über das Stabilitätsgebiet hinausgehende Freiheiten, die in Anm. 2.17 aufgezeigt werden, weshalb nachfolgende Betrachtung etwas allgemeiner bleibt.

Eine praktikable Berechnung von $r_{\mathbb{F}}^{\mathcal{S}}(A, B, C)$ wurde durch die Formel von Qui [521] möglich, wodurch sich die n^2-dimensionale Optimierung über die Matrixelemente auf eine zweidimensionale oder gar eindimensionale Optimierung reduzieren lässt:

$$
\begin{aligned}
r_{\mathbb{F}}^{\mathcal{S}}(A,B,C) &= \inf\{||\Delta||_2 : \Delta \in \mathbb{F}^{m\times p}; \exists \lambda_i(A+B\Delta C) \in \mathcal{S}^c\} \\
&= \inf\{\quad \dots \quad ; \exists \lambda_i(A+B\Delta C) \in \mathrm{bd}\mathcal{S}\} \qquad \text{(Stetigkeit der Eigenwerte)} \\
&= \inf\{\dots ; \exists s \in \mathrm{bd}\mathcal{S} : \lambda_i(A+B\Delta C - sI_n) = 0\} \qquad \text{(Spektralverschiebung)} \\
&= \inf_{s\in\mathrm{bd}\mathcal{S}} \inf\{||\Delta||_2 : \Delta \in \mathbb{F}^{m\times p}; \det(sI_n - A - B\Delta C) = 0\} \\
&= \inf_{s\in\mathrm{bd}\mathcal{S}} \inf\{||\Delta||_2 : \Delta \in \mathbb{F}^{m\times p}; \det(I_m - \Delta \underbrace{C(sI_n - A)^{-1}B}_{=M}) = 0\}. \quad /^{35} \\
&= \inf_{s\in\mathrm{bd}\mathcal{S}} \begin{cases} 1/||M||_2 & \text{für } \mathbb{F} = \mathbb{C} \ /^{36} \\ \left[\inf\limits_{\gamma\in(0,1]} \sigma_2\left(\begin{bmatrix} \Re e\, M & -\gamma \Im m\, M \\ \gamma^{-1}\Im m\, M & \Re e\, M \end{bmatrix}\right)\right]^{-1} & \text{für } \mathbb{F} = \mathbb{R}; \\ & \text{Formel von Qui} \end{cases} \\
&= \begin{cases} 1/\sup\limits_{s\in\mathrm{bd}\mathcal{S}} ||M||_2 & \text{für } \mathbb{F} = \mathbb{C} \\ \left[\sup\limits_{s\in\mathrm{bd}\mathcal{S}} \inf\limits_{\gamma\in(0,1]} \sigma_2\left(\begin{bmatrix} \Re e\, M & -\gamma \Im m\, M \\ \gamma^{-1}\Im m\, M & \Re e\, M \end{bmatrix}\right)\right]^{-1} & \text{für } \mathbb{F} = \mathbb{R}. \end{cases}
\end{aligned} \tag{2.89}
$$

Für den reellen Stabilitätsradius $r_{\mathbb{R}}^{-}(A)$ ist nunmehr die linke komplexe Halbebene $\mathcal{S} = \{s \in \mathbb{C} : \Re e\, s < 0\}$ einzusetzen. Ferner ist damit der Rand von $\mathcal{S}$ durch $\mathrm{bd}\mathcal{S} = \{j\omega : \omega \in \mathbb{R}\}$ (imaginäre Achse) gegeben und die Formel (2.89) vereinfacht sich zu

$$r_{\mathbb{C}}^{-}(A) = \min_{\omega\in\mathbb{R}} \sigma_{\min}(j\omega I_n - A) \tag{2.90a}$$

$$r_{\mathbb{R}}^{-}(A) = \min_{\omega\in\mathbb{R}} \max_{\gamma\in(0,1]} \sigma_{2n-1}\left(\begin{bmatrix} A & -\gamma\omega I_n \\ \omega/\gamma I_n & A \end{bmatrix}\right) \tag{2.90b}$$

$$r_{\mathbb{C}}^{-}(A) \le r_{\mathbb{R}}^{-}(A). \tag{2.90c}$$

Das Resultat (2.90a) findet sich bereits in [407], [160]. Algorithmen, die für (2.90a) das globale Minimum und eine destabilisierende Störung finden, sind z. B. in [110], [91] (Bisektionsverfahren) und [277] (inverse Iteration; schwache Besetztheit) beschrieben. Das Bisektionsverfahren, angewandt auf ω, kann auch für (2.90b) herangezogen werden. Der Funktionswert der inneren Maximierung lässt sich dabei per eindimensionaler Suche sicher berechnen, da $\sigma_{2n-1}(.)$ bei festem ω eine auf $(0,1]$ quasikonkave Funktion (impliziert Unimodalität) in γ ist [521].

---

[35] Nutze $\det(A+BC) = \det(A^{-1}(I_n + A^{-1}BC)) = (\det A^{-1}) \cdot \det(I_n + A^{-1}BC)$ und wende Schur-Komplementformel $\det\begin{bmatrix} I_n & -A^{-1}B \\ C & I_m \end{bmatrix} = \det(I_n + A^{-1}BC) = \det(I_m + CA^{-1}B)$ an.

[36] $1 = \sigma_{\min}(I_m) = \min\limits_{\mathrm{Rang}(I_m - X)<m} ||X||_2 = \min\limits_{\det(I_m - \Delta M)=0} ||\Delta M||_2 \le \min\limits_{\det(I_m-\Delta M)=0} ||\Delta||_2||M||_2$ für $\mathrm{Rang} M \ge 1$ (Das zweite Gleich folgt aus dem Schmidt-Mirsky-Theorem [299], die Ungleichheit aus der Submultiplikativität der Norm.) Gleichheit gilt z. B. für $\Delta_{\mathrm{opt}} = v_1 u_1^H/\sigma_{\max}(M)$, also ein Singulärtripel.

Als eine algorithmische Arbeit, die sich auf die Berechnung reeller strukturierter Stabilitätsradien auf der Grundlage von (2.89) bezieht, sei [584] genannt. In Verbindung mit den Ergebnissen aus [521] kann mit Kenntnis des Radius auch die nächstgelegene Matrix bestimmt werden. Mit den gleichen algorithmischen Umsetzungen können auch die in der nachfolgenden Anmerkung aufgeführten Mengen behandelt werden.

Anmerkung 2.17 $r_{\mathbb{F}}^{\mathcal{S}}(A,B,C)$ stützt sich auf einen erweiterten Stabilitätsbegriff, nach dem eine Matrix stabil in $\mathcal{S}$ heißt, wenn ihre Eigenwerte in $\mathcal{S}$ liegen. Beispiele für jeweils offene Mengen $\mathcal{S}$ sind:
- linke komplexe Halbebene; $r_{\mathbb{C}}^{-}(A,B,C) = ||G(s)||_{\mathcal{H}_\infty}^{-1}$ für $G(s) \not\equiv 0$ (Hurwitz-Stabilität)
- Ebene links eines $\gamma; \gamma < 0$ (Mindestabklingrate)
- ein nach links offener Sektor (Mindestdämpfung)
- Innere des Einheitskreises (Schur-Stabilität)
- rechten Hälfte des Einheitskreises (Mindestabtastung).

Die Matrizen B und C können benutzt werden, um den Bildraum und den Nullraum der zulässigen Änderungen zu restringieren. Allgemeine lineare oder affine Restriktionen (z. B. Symmetrie) an die zulässigen Änderungen sind durch $B\Delta C$ nicht realisierbar.

Anmerkung 2.18 Eine Anwendung des Stabilitätsradius beim High-Gain-Beobachterentwurf beschreibt Röbenack in [536], wobei er gleichzeitig die Beziehungen zur H_∞-Norm herstellt und das Bounded-Real-Lemma [647], [366] zur Lösung über eine LMI heranzieht.

2.5.4 Zugang über stabilitätsinvariante Umformungen

Bei der Modellreduktion mittels Padé-Approximation[37] werden Zustandsraumsysteme

$$\dot{x} = Ax + Bu \tag{2.91a}$$

$$y = Cx + Du \tag{2.91b}$$

unter der Maßgabe approximiert, dass die ersten k Markov-Parameter M_i

$$G(s) = C(sI_n - A)^{-1}B + D = D + \sum_{i=1}^{\infty} CA^{i-1}Bs^{-i} = M_0 + \sum_{i=1}^{\infty} M_i s^{-i} \tag{2.92}$$

des approximierenden Systems mit denen des Originalsystems übereinstimmen. Durch dieses Approximationsprinzip werden aber nur k Freiheitsgrade der Parametermatrizen fixiert.

[37] Eine Padé-Approximation einer Funktion $f(x)$ ist eine gebrochen rationale Funktion $g(x) = p(x)/q(x)$ mit $\deg p(x) = m, \deg q(x) = n$ und der Eigenschaft, dass die ersten $m+n$ Glieder der Taylor-Reihe von g und f übereinstimmen. Die Taylor-Reihe kann dabei an einem beliebigen Punkt entwickelt werden, wobei für $x = 0$ die Maclaurin-Reihe mit Potenzen x^k und für $x = \infty$ eine Reihe mit Potenzen x^{-k} folgt.

So kann es geschehen, dass trotz stabilen Originalsystems das approximierende System instabil ist, vgl. die Approximation mit dem populären Lanczos-Algorithmus in [567][38]. Statt nun die kompliziert zu behandelnden Stabilitätsrestriktionen explizit zu formulieren, um dadurch die Freiheitsgrade in A weiter einzuschränken, wird in [567] ein modifizierter Arnoldi-Algorithmus vorgeschlagen. Die Idee dieses Algorithmus ist es, das System sukzessive zu approximieren, wobei die Umformungen so gestaltet sind, dass die Klasse der stabilen Systeme nicht verlassen wird. In [39] wird beschrieben, wie auf ähnliche Weise Passivität des Systems bei der Modellreduktion sichergestellt werden kann.

2.5.5 Praktische Stabilität

Die Stabilität linearer Systeme im Sinne von Lyapunov ist eine asymptotische Eigenschaft, die an den Systemeigenwerten festgemacht wird. In der Praxis können sich Lösungen jedoch temporär sehr weit von der Ruhelage entfernen, bevor sie dann gegen selbige streben. Mit anderen Worten: Das stabile System verhält sich temporär wie ein instabiles. Dieser Effekt tritt besonders dann auf, wenn die Systemmatrix einen großer Abstand zu den normalen Matrizen ($A^TA = AA^T$) hat, vgl. Beispiel 2.13.

Beispiel 2.13 (Stabiles System mit hoher transienter Schranke)
Am nachstehenden System und dessen Lösung (Simulation in Bild 2.1)

$$\dot{x} = \begin{bmatrix} -0.6 & c \\ 0 & -1 \end{bmatrix} x;\ x(0) = x_0 \quad \Rightarrow \quad x(t) = \begin{bmatrix} \mathrm{e}^{-0.6t} & 2.5c(\mathrm{e}^{-0.6t} - \mathrm{e}^{-t}) \\ 0 & \mathrm{e}^{-t} \end{bmatrix} x_0 \tag{2.93}$$

ist unschwer zu erkennen, dass je größer c, desto größer ist das temporäre Anwachsen der Norm. Ein größer werdendes c bedeutet nämlich, dass die Nichtnormalität der Systemmatrix zunimmt, was seinerseits an der wachsende Unsymmetrie der Matrix erkennbar ist.
Der Effekt des temporären Vergrößerns liegt auch beim Regeln oder Beobachten der Zustände des Doppelintegrators vor. Um nämlich aus eine Ruhelage in die andere zu gelangen, muss der Regler die Geschwindigkeit zunächst vergrößern. Es ist also unmöglich, die Fehler in beiden Zustandskoordinaten simultan zu verringern.

Anmerkung 2.19 Tritt der beschriebene Effekt in einem linearisierten Modell ein, resultiert daraus für das nichtlineare System meist ein sehr kleiner Einzugsbereich der stabilen Ruhelagen, vgl. $\dot{x}_1 = x_1(x_1 - 0.6) + cx_2, \dot{x}_2 = -x_2$. Die Linearisierung ergibt gerade (2.93) und zeigt für positive x_2 einen kleinen Einzugsbereich der stabilen Ruhelage $(0, 0)$ [288].

[38] Selbst wenn alle Freiheitsgrade ausgeschöpft werden (gleiche Parameteranzahl in $f(x)$ und $g(x)$), kann die Approximation einer stabilen Übertragungsfunktion instabil sein, vgl. System dritter Ordnung in [458].

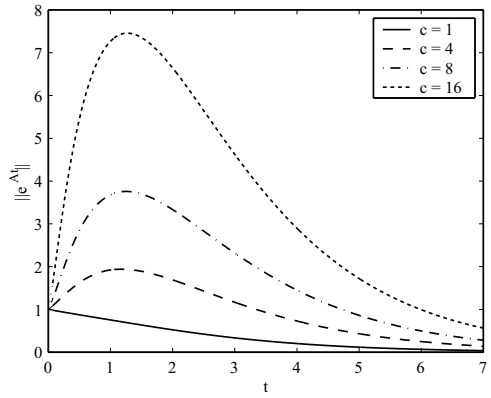

Bild 2.1: Verläufe zu Beispiel 2.13

Das Problem an dem im Beispiel beschriebenen Verhalten und der in der Anmerkung geschilderten Konsequenz ist, dass der Gültigkeitsbereich des linearen Modells oder schlimmer noch der Sicherheitsbereich der Anlage überschritten werden kann, wenn etwa bestimmte Zustandskomponenten kurzzeitig extrem groß werden. Das führte Hinrichsen, Plischke und Pritchard auf die Definition der praktischen Stabilität.

Definition 2.4 (Praktische Stabilität, [288])
Ein autonomes lineares LTI-System heißt praktisch stabil, wenn es die Zeitbereichsbedingung

$$||\mathrm{e}^{At}||_2 \leq \gamma \mathrm{e}^{\beta t}, t \geq 0 \tag{2.94}$$

für vorgegebenes $\gamma \geq 1$ und $\beta \leq 0$ erfüllt.

Anmerkung 2.20 Bei dieser Definition entscheidet die Vorgabe von γ und β darüber, ob das System die Eigenschaft hat oder nicht.

Anmerkung 2.21 Für jedes $\beta > \alpha(A) = \max\limits_{1 \leq i \leq n} \Re e\, \lambda_i$ existieren derartige γ (abhängig von β)[39], wobei das jeweils kleinste γ

$$\gamma_\beta(A) = \inf\{\gamma \in \mathbb{R} : ||\mathrm{e}^{At}||_2 \leq \gamma \mathrm{e}^{\beta t}, t \geq 0\} \tag{2.95}$$

transiente Schranke heißt. Sie entspricht dem Wert der majorisierenden e-Funktion in $t = 0$ und ist eine Kennzahl für das transiente Verhalten.

[39]Die Spektralabszisse $\alpha(A)$ misst das asymptotische Wachstum des Eigenvorgangs: $\alpha(A) = \lim\limits_{t \to \infty} \frac{1}{t} ||\mathrm{e}^{At}||_2$.

Anmerkung 2.22 Eine alternative Kennzahl für das transiente Verhalten ist die Anfangswachstumsrate (logarithmische Norm, Lozinskij-Maß[40] [614], [382])

$$\mu_2(A) \stackrel{\text{def}}{=} \frac{\mathrm{d}^+}{\mathrm{d}t}||\mathrm{e}^{At}||_2\Big|_{t=0} \tag{2.96}$$

$$= \frac{1}{2}\max_{1\le i\le n}\lambda_i(A+A^T), \tag{2.97}$$

die den Anstieg der Spektralnorm von e^{At} in $t=0$ beschreibt.

Da es sowohl sehr schwierig ist, eine Zeitbereichsbedingung aufzustellen, als auch sie algorithmisch zu behandeln, bleiben die Bedingung und die Kennzahlen der A-posteriori-Modellauswertung und dem Reglerentwurf [288] vorbehalten. Werden zur Identifikation Beobachter eingesetzt, so können diese im linearen Fall derart entworfen werden, dass möglichst keine Zustandsgröße ihren Anfangsfehler temporär extrem vergrößert. Das wird durch einen Entwurf auf ein möglichst kleines γ erreicht.
Bei Systemen mit Dreiecksstruktur gibt die Summe der Beträge der Nichtdiagonalelemente (Maß für die Abweichung von den normalen Matrizen) ein Indiz dafür, wie ausgeprägt eine praktische Instabilität ist, s. auch Beispiel 2.13. Das betragsgrößte Nichtdiagonalelement zeigt zudem an, welche Zustandsgröße primär für das schlechte transiente Verhalten verantwortlich ist.

2.6 Stabilität für lineare zeitdiskrete Zustandsraumsysteme

Bisher wurde die Frage behandelt, wie Stabilität bei der Modellbildung für zeitkontinuierliche LTI-Zustandsraumsysteme gesichert werden kann, wobei sich die Anwendungen auf die Modellreduktion, Modellapproximation und die A-posteriori-Modellbewertung beziehen. Die klassische Identifikation wird nur für den Spezialfall der diagonaldominanten Systeme gelöst, während für den allgemeinen Fall nur ein möglicher Restriktionsansatz andiskutiert wird. Ursache hierfür ist der Mangel an leistungsstarken Verfahren zur Identifikation kontinuierlicher Zustandsraummodelle. Anders ist die Situation bei zeitdiskreten Systemen. Dort führt die Identifikation im Zustandsraum bei messbaren, aber auch bei geschätzten Zustandsgrößen auf LS-Probleme vom Typ

$$\frac{1}{2}\left\|\begin{bmatrix} X_{k+1} \\ Y_k \end{bmatrix} - \begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} X_k \\ U_k \end{bmatrix}\right\|_F^2 \stackrel{!}{=} \text{Min} \tag{2.98}$$

[40] Die Kennzahl wurde unabhängig von Lozinskij und Dahlquist eingeführt. Sie lässt sich bezüglich jeder Operatornorm erweitern.

Darüber hinaus sind bei einigen Unterraummethoden folgende Probleme zu lösen [490]

$$\tfrac{1}{2}||O_{0:k-1}A - O_{1:k}||_F^2 \stackrel{!}{=} \text{Min} \qquad O_{0:k} = \begin{bmatrix} C \\ \vdots \\ CA^k \end{bmatrix} \tag{2.99a}$$

$$\tfrac{1}{2}||AC_{0:k-1} - C_{1:k}||_F^2 \stackrel{!}{=} \text{Min} \qquad C_{0:k} = \Big[B, AB, \ldots, A^kB\Big], \tag{2.99b}$$

wobei $O_{0:k}$ eine erweiterte, als Schätzung vorliegende Beobachtbarkeitsmatrix und $C_{0:k}$ eine Steuerbarkeitsmatrix bezeichnet. Fortan wird sich auf (2.98) beschränkt, da (2.99b) als Spezialfall von (2.98) angesehen werden kann und sich (2.99a) ähnlich behandeln lässt.
Das zentrale Problem von (2.98) ist, dass die LS-Lösung A_{LS} keine Schur-stabile Matrix (Eigenwerte im Einheitskreis) sein muss, selbst wenn das zu Grunde liegende System exponentiell stabil ist. Ursache kann eine zu groß gewählte Modellordnung sein, wenn beispielsweise bei ihrer Festlegung über die Hankel-Singulärwerte keine signifikante Lücke erkennbar war. Darüber hinaus bewirken Störungen, dass mitunter instabile Modelle geschätzt werden. Gleichwohl sorgt zumindest die asymptotische Effizienz der Unterraummethoden dafür, dass mit wachsender Tupelzahl unter der Voraussetzung einer ständigen Anregung die Wahrscheinlichkeit für instabile Modelle bei richtig gewählter Modellordnung abnimmt.
Eine zusätzliche Restriktion an den Spektralradius $\varrho(A) \stackrel{\text{def}}{=} \max_i |\lambda_i(A)|$, nämlich $\varrho(A) < 1$, würde zwar Stabilität garantieren, nicht aber die Existenz einer Lösung. Das liegt an der strengen Ungleichung und der Stetigkeit des Problems, welches unter Umständen nur ein Infimum zulässt. Als Ausweg bietet sich $\varrho(A) \leq \gamma$ mit $\gamma < 1$ an. Der Fall $\gamma = 1$ ist dabei interessant, wenn A einen oder mehrere ungedämpften Schwingungsmodi oder einen Integrator modellieren muss.
Das Problem (2.98) mit der Restriktion $\varrho(A) \leq \gamma$ heißt γ-generisch-stabiles Problem. Doch selbst mit diesen Einschränkungen werden die Schwierigkeiten nicht merklich weniger, denn der Spektralradius ist eine nichtpolynomiale, nichtkonvexe und nicht differenzierbare Funktion der Matrixelemente, die für Systeme $n > 3$ zudem praktisch nicht algebraisch zu formulieren ist (mehrseitige computeralgebraische Ausdrücke). Aus diesem Grund werden in den nachfolgenden Abschnitten Zugänge diskutiert, die die angesprochenen Schwierigkeiten umgehen. Hierzu zählen:

1. Zugang über das zeitdiskrete Lyapunov-Theorem (Abschnitt 2.6.1)
2. Ad-hoc-Zugänge zur A-posteriori-Stabilitätssicherung (Abschnitt 2.6.2)
3. Zugang über Modifikation der Zielfunktion (Abschnitt 2.6.3)
4. Zugang über Systemapproximation (Abschnitt 2.6.4)

2.6.1 Zugang über das zeitdiskrete Lyapunov-Theorem

In strenger Analogie zu Abschn. 2.5.1 existieren auch für zeitdiskrete Systeme Eigenwertcharakterisierungen der Systemmatrix für die unterschiedlichen Stabilitätsausprägungen. Der einzige Unterschied ist, dass im diskreten Fall der Einheitskreis den Rand der Menge der Schur-stabilen Matrizen bildet, während dieser im zeitkontinuierlichen Fall die imaginäre Achse ist. Ebenso lassen sich auch aus dem zeitdiskreten Lyapunov-Theorem [543] semidefinite Restriktionen

$$A^T P A - P \preceq 0_{n\times n} \quad \text{und} \quad P \succ 0_{n\times n}, \qquad /^{41} \tag{2.100}$$

für Lyapunov-Stabilität angeben. Die Umkehrung gilt nicht, da für ein Lyapunov-stabiles System (äquivalent: A hat keine Eigenwerte außerhalb des Einheitskreises und Eigenwerte auf dem Einheitskreis sind nicht defektiv [415]) in (2.100) $P \succ 0_{n\times n}$ nicht zwingend folgt. Doch es existiert für jedes solches A ein $P \succ 0_{n\times n}$, sodass sich (2.100) erfüllen lässt, kurzum dass (2.100) zulässig ist. Dies wird anhand einer reellen Block-Spektralfaktorisierung $A = SDS^{-1}$ mit (2×2)-Blöcken $r_i \begin{bmatrix} \cos\phi_i & \sin\phi_i \\ -\sin\phi_i & \cos\phi_i \end{bmatrix}$ mit $r_i \leq 1$ und/oder (1×1)-Diagonalblöcken für $P = I_n$ offensichtlich.
Wie bei der Spektralradius-Restriktion kann auch hier eine Stabilitätsreserve durch ein $\gamma < 1$ eingebaut werden, und zwar durch

$$A^T P A - \gamma^2 P \preceq 0_{n\times n} \quad \text{und} \quad P \succ 0_{n\times n}, \tag{2.101}$$

Mit dieser Restriktion wurde (2.98) mit einem SQP-Algorithmus in [218] gelöst. Der Aufwand ist durch die $\frac{n(n+1)}{2}$ zusätzlichen Unbekannte für P beträchtlich. Das Problem, den globalen Minimierer zu finden, ist wegen der Multimodalitäten noch ungelöst.

2.6.2 Ad-hoc-Zugänge zur A-posteriori-Stabilitätssicherung

Die bisherigen Überlegungen zeigen, dass die Einhaltung der Schur-Stabilität von A nicht direkt, sondern eher durch eine Problemmodifikation erreicht werden sollte. Die Restriktion wird also nicht in das Kriterium eingebaut, sondern stattdessen wird die LS-Lösung A_{LS} im Nachhinein so modifiziert, dass eine Schur-stabile Matrix entsteht. Im Folgenden werden zwei Matrixapproximationsprobleme vorgeschlagen, die auf eine zweidimensionale Suche bzw. eine LMI-Formulierung führen. Zuvor werden die beiden naheliegenden Zugänge Matrixskalierung

[41] $P \succ 0_{n\times n}$ kann nicht zu $P \succeq 0_{n\times n}$ abgeschwächt werden, vgl. $A = \begin{bmatrix} 0.8 & 0 \\ 0 & 1.2 \end{bmatrix}$ und $P = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, da dann $A^T P A - P \preceq 0_{n\times n}$ gilt, obwohl A instabil ist.
$A^T P A - P \preceq 0_{n\times n}$ kann nicht zu $A^T P A - P \prec 0_{n\times n}$ verschärft werden, da A dann zwingend Schur-stabil und damit Element einer offenen Menge ist.

und Eigenwertprojektion vorgestellt. Beide erfordern einen geringen Aufwand, keine spezialisierten Algorithmen, erreichen aber nicht die Qualität der Matrixapproximationsprobleme.

Matrixskalierung und Eigenwertprojektion
Als Ad-hoc-Variante scheint eine Skalierung der LS-Lösung A_{LS}

$$A_{skal} = A_{LS}/\varrho(A_{LS}) \cdot \gamma \qquad \text{mit } \gamma \leq 1 \tag{2.102}$$

naheliegend. Ferner drängt sich eine Projektion der instabilen Eigenwerte auf den γ-Kreis auf

$$A_{eig} = S(\text{Proj}J)S^{-1}, \tag{2.103}$$

die ausgehend von der Jordan-Form $A_{LS} = SJS^{-1}$ bewerkstelligt wird. Die instabilen reellen Eigenwerte in den Jordan-Blöcken werden auf γ oder $-\gamma$ gesetzt und die instabilen konjugiert komplexen auf einen Betrag von γ skaliert.
Bei der A_{skal}-Berechnung wird die Matrix A_{LS} für die Korrektur auf einen einzigen Parameter, nämlich auf $\varrho(A_{LS})$, reduziert. Die Kenntnis über die n^2 Matrixelemente geht dabei fast vollständig verloren. Entsprechend schlecht fällt das Approximationsverhalten für A_{skal} aus. Bei der Eigenwertprojektion (2.103) wird auf n Parameter – die Eigenwerte – reduziert, wodurch bessere Ergebnisse erzielt werden [137]. Noch bessere Ergebnisse (gemessen an den mittleren Parameterabweichungen) lassen sich mit den folgenden Formulierungen erzielen.

Matrixapproximationsprobleme
Die komplette Information von A_{LS} wird durch

$$A_{MAP} = \arg\min_{A} \|A_{LS} - A||_2; \qquad A \in \mathbb{R}^{n\times n} : \varrho(A) \leq \varrho; A_{LS} \text{ instabil.} \tag{2.104}$$

ausgenutzt. Dieses Matrixapproximationsproblem, das die nächstgelegene generisch Schur-stabile Matrix ($\varrho = 1$) oder eine Schur-stabile Matrix mit Stabilitätsreserve ($\varrho < 1$) zu A_{LS} sucht, ist nichtkonvex. Es lässt sich gemäß Formel (2.89) auf ein zweidimensionales Problem reduzieren. Für die numerische Behandlung [110], [291] ist es dabei unbedeutend, ob die nächstgelegene grenzstabile Matrix zu einer instabilen oder grenzinstabile zu einer stabilen gesucht wird. Das Minimum ist ein Stabilitätsmaß[42] und gibt eine Stabilitätsreserve bezüglich der Matrixelemente an, wie sie etwa beim robusten Reglerentwurf benötigt wird.

Im alternativen Matrixapproximationsproblem [434]

$$||(A_{LS} - A)P||_2 \stackrel{!}{=} \text{Min} \qquad P \succ 0_{n\times n}, P - APA^T \succ 0_{n\times n}, \tag{2.105}$$

[42] Der Vorteil dieses Maßes gegenüber dem Eigenwertabstand vom Einheitskreis als Stabilitätsmaß wird am Beispiel $A_{sys} = \text{Jord}_3(0.9)$ deutlich. Hier ist der Eigenwertabstand 0.1, aber bereits eine Störung des Elements a_{31} auf $a_{31} = 0.001$ schiebt einen Eigenwert auf Eins!

wird die Schur-Stabilität über die Lyapunov-Ungleichung ausgedrückt. Die Wichtung in der Zielfunktion durch P ermöglicht eine Umformulierung in ein SDP-Problem.[43] Da sich für $P \to 0_{n\times n}$ ein Infimum ergibt, was seinerseits eine schlechte Konditionierung zur Folge hat, ist es besser, mit der regularisierten Variante

$$||(A_{LS} - A)P||_2 \stackrel{!}{=} \text{Min} \qquad P \succeq \mu I_n, P - APA^T \succeq \mu I_n; \mu > 0 \text{ fest.} \tag{2.106}$$

zu arbeiten. Diese liefert bezüglich A den gleichen Minimierer, da die zulässige Menge homogen in P ist. Mit $X := AP$ folgt mit der Epigraph-Methode aus Abschnitt 6.5 und den Umformungen nach Tabelle 7.4 das LMI-Problem

$$y \stackrel{!}{=} \text{Min} \qquad \begin{bmatrix} yI_n & (A_{LS}P - X)^T \\ A_{LS}P - X & I_n \end{bmatrix} \succeq 0_{2n\times 2n}, \quad \begin{bmatrix} P - \mu I_n & X \\ X^T & P \end{bmatrix} \succeq 0_{2n\times 2n}, \tag{2.107}$$

wobei $A_{\text{opt}} = X_{\text{opt}} P_{\text{opt}}^{-1}$ die gesuchte Lösung liefert.

2.6.3 Zugang über Modifikation der Zielfunktion

In (2.104) wird zwar die komplette Information über A_{LS} genutzt, die datenbezogene Information, mit der A_{LS} geschätzt wird, bleibt aber unberücksichtigt. Durch Einbeziehen dieser Information kommt die Schätzung der auf γ-generische Stabilität restringierten LS-Lösung näher. Die Ergebnisse zur sog. Datenerweiterungsmethode in [137] zeigen das experimentell. Nachfolgend wird nicht das Schätzproblem um „künstliche Daten“ erweitert, um Stabilität zu erzielen, sondern es wird die Zielfunktion erweitert. Bei LS-Problemen kann aber eine Datenerweiterung in eine Zielfunktionserweiterung überführt werden und meist auch umgekehrt (Umkehrung gilt für additiver Modifikation mit quadratischen Termen). Somit sind beide Zugänge sehr ähnlich, wenngleich Anschaulichkeit, Beweisführung, Handhabbarkeit und erzielbare Modellgüte für den nachfolgend beschriebenen Zugang sprechen.

Satz 2.9 (Stabilitätssicherung durch Regularisieren, [218])
Durch Wahl eines geeigneten α im Regularisierungsterm von

$$\left\| X_{k+1} - [A, B] \begin{bmatrix} X_k \\ U_k \end{bmatrix} \right\|_F^2 + \alpha \,\text{spur}(AWA^T) \stackrel{!}{=} \text{Min} \qquad A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m} \tag{2.108}$$

mit $W \in \mathcal{S}_n^>$ lässt sich für den Minimierer A_α die Forderung $\varrho(A_\alpha) < \gamma$ einhalten. Ferner existiert mindestens ein α, für das $\varrho(A_\alpha) = \gamma$ gilt.

[43] $||XP||_2$ ist für beliebiges festes reguläres P eine Norm für $X \in \mathbb{R}^{n\times n}$, kurz $||X||_{2,P}$. Somit kann (2.106) als Minimierung von $||A_{LS} - A||_{2,P}$ interpretiert werden.

Der Satz lässt zunächst offen, wie α zu wählen ist. Weitere Betrachtungen in [218] zeigen, dass ein $\tilde{\alpha}$ über ein verallgemeinertes Eigenwertproblem bestimmt werden kann, für das $\varrho(A_{\tilde{\alpha}}) = \gamma$ gilt. Für jedes größere α liegt dann γ-Stabilität vor. Das Problem an dem in [218] bestimmten $\tilde{\alpha}$ ist aber, dass es nicht zwingend das kleinste α ist, für das A_α γ-generisch-stabil ist. Ursache ist die Nichtmonotonie der Spektralradien $\varrho(A_\alpha)$ der Minimierer A_α bezüglich α. Nun ist aber gerade ein kleines α von Interesse, da dann die Lösung der modifizierten Zielfunktion weniger von der ursprünglichen Zielfunktion abweicht.
Neben α kann der Regularisierungsterm auch über W beeinflusst werden. Für die Wahl $W = I_n$ spricht das Maximum-Entropie-Prinzip, nach dem alle Elemente von A gleichermaßen zu bestrafen sind. Der Regularisierungsterm $\tilde{\alpha}\,\mathrm{spur}(AA^T)$ erwies sich in experimentellen Untersuchungen [218] der Datenerweiterungsmethode nach [137] überlegen, die vom Prinzip her eine Regularisierungsmethode mit $W \neq I_n$ ist. Die Frage nach einem kleinstmöglichen Regularisierungsterm, um A_α γ-generisch-stabil zu machen, blieb bisher unbeantwortet.

2.6.4 Zugang über Systemapproximation

In den Problemen (2.104) und (2.106) werden zwar alle Elemente von A_{LS} in die Approximation einbezogen, und es wird ihr Einfluss auf die Stabilität bewertet, die Information über die Matrizen B_{LS}, C_{LS}, D_{LS}, die ihrerseits auch wesentlich das E/A-Verhalten bestimmen, bleibt ungenutzt. Deshalb kann statt eines Matrix- auch ein Systemapproximationszugang verwendet werden. Eine aus der Modellordnungsreduktion bekannte Zielfunktion hierfür ist

$$||F(\mathrm{e}^{j\omega}) - G(\mathrm{e}^{j\omega})||_{L_2}^2 = \frac{1}{2\pi}\int_0^{2\pi} ||F(\mathrm{e}^{j\omega}) - G(\mathrm{e}^{j\omega})||_F^2 \,\mathrm{d}\omega \stackrel{!}{=} \text{Min.} \tag{2.109}$$

Mit $F(z) = C(zI - A)^{-1}B + D$ wird die zu approximierende, aus einer LS-Schätzung stammende z-Übertragungsmatrix bezeichnet, während G aus der Menge der stabilen rationalen $(p \times m)$-Matrizen vom McMillan-Grad $\leq n$ zu wählen ist.
Die Existenz einer Lösung ist garantiert [48], allerdings besitzt das Problem wegen der Nichtkonvexität der zulässigen Menge i. Allg. zahlreiche lokale Minima. Ein entsprechender Algorithmus in [438] basiert auf der Douglas-Shapiro-Shields-Faktorisierung[44], während ein anderer Algorithmus in [188] eine Padé-ähnliche Reduktionsmethode nutzt.

Anfangs wurde als Vorteil herausgestellt, dass die rationale L_2-Approximation auch die Information über die Matrizen B, C, D nutzt. Dieser Vorteil ist unumstritten, wenn es sich um ein Approximationsproblem wie etwa eine Ordnungsreduktion handelt. Er sollte sicher auch

[44] Jede streng propere $(r \times m)$-Übertragungsmatrix G lässt sich als $G = SP$ schreiben, wobei S eine rationale verlustfreie $(r \times r)$-Matrixfunktion mit dem gleichen McMillan-Grad wie G und P eine instabile $(r \times m)$-Matrix ist. S ist bis auf eine Rechtsmultiplikation mit einer unitären Matrix eindeutig [167].

bei ausgewogenen Approgressionsproblemen zum Tragen kommen. Bei Regressionsproblemen indes kann die A-posteriori-Approximation einen Bias in zuvor biasfreien Schätzungen für B, C und D hervorrufen. Daher scheint die sog. Regularisierungsmethode für Regressionsprobleme besser geeignet, zumal sie zusätzlich datenbezogene Information nutzt.

2.7 Stabilität für lineare Intervallsysteme

Eine Möglichkeit, Stabilität des identifizierten Systems zu sichern, eröffnet sich, wenn die Systemparameter durch Intervalle begrenzt sind. Für die Online-Identifikation ist das insofern interessant, als dass sich Intervallrestriktionen algorithmisch einfach handhaben lassen. Die nachfolgenden Betrachtungen beziehen sich dabei auf:

Intervallpolynome:

$$[a(\zeta)] = [a_0] + [a_1]\zeta + [a_2]\zeta^2 + \ldots + [a_n]\zeta^n; \qquad [a_i] = [\underline{a}_i, \overline{a}_i] \tag{2.110}$$

mit $\zeta =: s$ (zeitkontinuierlich) bzw. $\zeta =: z$ (zeitdiskret)

Affine Polynomkombinationen mit Intervallparametern:

$$p(\zeta; [\theta]) = p_0(\zeta) + [\theta_1]p_1(\zeta) + \ldots + [\theta_k]p_k(\zeta); \qquad [\theta_i] = [\underline{\theta}_i, \overline{\theta}_i] \tag{2.111}$$

Intervallmatrizen:

$$[A] = \begin{bmatrix} [a_{11}] & \ldots & [a_{1n}] \\ \vdots & & \vdots \\ [a_{n1}] & \ldots & [a_{nn}] \end{bmatrix} \in \mathbb{IR}^{n\times n}; \qquad [a_{ij}] = [\underline{a}_{ij}, \overline{a}_{ij}]. \tag{2.112}$$

Parameteraffine Matrizen mit Intervallparametern:

$$A([\theta]) = A_0 + [\theta_1]A_1 + \ldots + [\theta_k]A_k; \qquad [\theta_i] = [\underline{\theta}_i, \overline{\theta}_i]. \tag{2.113}$$

In allen Fällen geht es um die Stabilität ganzer Familien von LTI-Systemen (absolute Stabilität). Es sind also stets unendlich viele Systeme zu untersuchen, was natürlich nur dann gelingt, wenn es Sätze gibt, die diesen Aufwand auf ein endliches Maß reduzieren. Bevor diese Sätze aber Anwendung finden können, gilt es, die Intervalle der physikalischen Prozessparameter in Intervalle der Systemparameter umzurechnen. Einen Weg bietet die Intervallarithmetik [416], [321], die im folgenden einfachen Beispiel zum Einsatz kommt.

Beispiel 2.14 (Relaxation in ein Intervallpolynom)

$$a(s;\theta_1,\theta_2) = \theta_1 s^4 + (\theta_1+1)s^3 + s^2 + \theta_1\theta_2 s + \theta_2 \qquad \theta_1 \in [0,\overline{\theta}_1], \theta_2 \in [\underline{\theta}_2,\overline{\theta}_2] \tag{2.114}$$

wird für nichtnegative Parameter durch das Intervallpolynom

$$[a(s)] = [0,\overline{\theta}_1]s^4 + [1,1+\overline{\theta}_1]s^3 + [1,1]s^2 + [0,\overline{\theta}_1\overline{\theta}_2]s + [\underline{\theta}_2,\overline{\theta}_2] \tag{2.115}$$

relaxiert, d. h. das Intervallpolynom enthält mehr Polynome als die Familie.

Statt über Intervallarithmetik können die Intervalle für die Systemparameter auch durch systematisches Variieren der Prozessparameter ermittelt werden, sofern die Zusammenhänge stetig sind. Auf diesem ingenieurnahen Weg werden oft sogar engere Systemparameterintervalle (schärfere Restriktionen) erhalten. Das folgende Beispiel zeigt einen zu beachtenden Effekt der Intervallarithmetik, der beim ingenieurnahen Weg nicht auftritt.

Beispiel 2.15 (Äquivalente Terme, aber unterschiedliche Intervalle, [321])
Aus $\theta \in [-1,1]$ soll ein Intervall für $\tilde{\theta} = \theta(\theta+2)$ gefolgert werden. Die Arithmetik liefert $[-1,1]\cdot[1,3] = [-3,3]$. Wird $\tilde{\theta}$ als $\tilde{\theta} = \theta^2 + 2\theta$ geschrieben, folgt $[0,1] + 2[-1,1] = [-2,3]$; für $\tilde{\theta} = (\theta+1)^2 - 1$ folgt sogar $[0,4] - 1 = [-1,3]$, also jenes Ergebnis, was sich auch direkt aus den Parametergrenzen von θ ergibt.

Sind die Intervalle a priori bestimmt worden, können die Sätze zur Stabilität von Intervallsystemen herangezogen werden. So lässt sich a priori überprüfen, ob Intervallrestriktionen mit der Stabilitätsforderung konsistent sind. Das Kharitonov-Theorem hat sich hierfür bewährt, da es die Untersuchung der unendlich vielen Systeme $[a(s)]$ auf nur vier Systeme reduziert!

Satz 2.10 (Kharitonov-Theorem (reelle Koeffizienten), [346][45])
Jedes Polynom $a(s) \in \{\sum_{i=0}^{n} a_i s^i : \underline{a}_i \le a_i \le \overline{a}_i; \underline{a}_n \neq 0 \text{ oder } \overline{a}_n \neq 0\}$ mit $n \ge 1$ ist genau dann Hurwitz-stabil, wenn jedes der Kharitonov-Polynome ein Hurwitz-Polynom ist

$$k_1(s) = \underline{a}_0 + \underline{a}_1 s + \overline{a}_2 s^2 + \overline{a}_3 s^3 + \underline{a}_4 s^4 + \underline{a}_5 s^5 + \ldots \tag{2.116a}$$
$$k_2(s) = \underline{a}_0 + \overline{a}_1 s + \overline{a}_2 s^2 + \underline{a}_3 s^3 + \underline{a}_4 s^4 + \overline{a}_5 s^5 + \ldots \tag{2.116b}$$
$$k_3(s) = \overline{a}_0 + \underline{a}_1 s + \underline{a}_2 s^2 + \overline{a}_3 s^3 + \overline{a}_4 s^4 + \underline{a}_5 s^5 + \ldots \tag{2.116c}$$
$$k_4(s) = \overline{a}_0 + \overline{a}_1 s + \underline{a}_2 s^2 + \underline{a}_3 s^3 + \overline{a}_4 s^4 + \overline{a}_5 s^5 + \ldots \tag{2.116d}$$

[45] Der Beweis von Kharitonov ist kompliziert und liefert wenig Einsicht in die Struktur des Problems. Kürzere, elementare Beweise geben Minnichelli, Anagnost und Desoer [457] sowie Neubacher [478]. Ein Überblick findet sich in [51]. Die im Original geforderte Bedingung eines invarianten Polynomgrads, d. h. $0 \notin [\underline{a}_n, \overline{a}_n]$, kann entfallen [330]. Eine Version für Polynome mit komplexen Koeffizienten beschreibt [73].

Anmerkung 2.23 Für Polynome bis 5. Grads werden nicht alle Kharitonov-Polynome benötigt. So reicht für monische Polynome 2. Grads $\underline{a}_{0,1} > 0$. Für monische Polynome 3. Grads genügt das Erfülltsein von (2.116c), für 4. Grads (2.116c) und (2.116d) und für 5. Grads (2.116b) bis (2.116d) [21].

Anmerkung 2.24 Ein Theorem vom Kharitonov-Typ für zeitdiskrete Systeme existiert nicht.[46] Ebenso existiert kein derartiges Theorem für LTI-Systeme mit Totzeit [204]. Es muss auf das Kantentheorem (Satz 2.11) ausgewichen werden.

Anmerkung 2.25 Signalisiert das Kharitonov-Theorem, dass das Intervallpolynom nicht stabil ist, bedeutet dies noch nicht zwingend Instabilität für das System, denn durch die Relaxation (s. Beispiel in [73]) oder zu geringe A-priori-Schranken an die Prozessparameter haben sich möglicherweise zu große Intervalle ergeben.

Während bei Intervallpolynomen die Stabilität spezieller Eckpolynome ausreichend war, sind für affine Polynomkombinationen ganze Kanten (Segmente, Strecken), d. h. einparametrische Familien der Art

$$[p_1(\zeta), p_2(\zeta)] = \{\gamma p_1(\zeta) + (1-\gamma)p_2(\zeta) : \gamma \in [0,1]\} \tag{2.117a}$$

$$= p_2(\zeta) + [\gamma](p_1(\zeta) - p_2(\zeta));\ \gamma \in [0,1] \tag{2.117b}$$

zu untersuchen. Der naheliegende Weg, nur die Eckpolynome $p_1(\zeta)$ und $p_2(\zeta)$ zu untersuchen, scheitert an der Nichtkonvexität der Hurwitz-Polynome

Beispiel 2.16 (Nichtkonvexität der Hurwitz-Polynome, [73])
Die Polynome

$$p_1(s) = 3s^4 + 3s^3 + 5s^2 + 2s + 1 \qquad p_2(s) = s^4 + s^3 + 5s^2 + 2s + 5.$$

sind Hurwitz-stabil. Ihr Mittelpunkt $p(s) = \frac{1}{2}(p_1(s) + p_2(s))$ ist es nicht (Wurzel bei $s = j$).

Dank der Theorie der robusten Regelungen wurden in den letzten Jahrzehnten zahlreiche Kriterien entwickelt, um zu entschieden, ob jedes Element in (2.117a) stabil ist. Die für die Identifikation wichtigsten Theoreme werden hier einschließlich der erforderlichen Vorgehensweisen aufgeführt und erläutert. So eignet sich das Aguirre-Suárez-Theorem [8] für die Optimierung, da es eine Formulierung als Matrixungleichung nutzt. Das Białas-Theorem [74], [75] ist für Tests zu empfehlen, da lediglich die Eigenwerte einer speziellen Matrix, die die Polynomkoeffizienten enthält, zu untersuchen sind. Der Vorteil des Segment-Lemmas [73], das die Einhaltung einer Phasendifferenzbedingung betrachtet, zeigt sich wegen des

[46] $p(z) = z^5 + [-1.5, 0.5]z^4 + 0.95z^3 + -0.12z^2 - 0.25$ ist an den Ecken stabil, nicht aber für $a_4 = -0.5$ [480].

Frequenzbereichsbezugs bei Erweiterungen um Totzeiten. Um die Theoreme anwenden zu können, müssen einparametrische Familien zunächst in die Segmentdarstellung (2.117a) gebracht werden. Wie das geht, zeigt das nachfolgende Beispiel.

Beispiel 2.17 (Umformung in eine Segmentdarstellung)
Für die durch den Parameter θ bestimmte Menge von Polynomen

$$p(\zeta;\theta) = \zeta^3 + 2\theta\zeta^2 + (\theta+1)\zeta + \theta - 1 \qquad \theta \in [2,3]$$

liefert das Einsetzen der Grenzen von θ die Eckpolynome der Segmentdarstellung (2.117a)

$$p_1(\zeta) = \zeta^3 + 4\zeta^2 + 3\zeta + 1 \qquad p_2(\zeta) = \zeta^3 + 6\zeta^2 + 4\zeta + 2.$$

Treten mehrere Parameter auf, so ist entscheidend, ob diese affin oder nichtlinear miteinander verknüpft sind. Während parameteraffine Systeme mit dem Kantentheorem behandelt werden können, führen parameternichtlineare Probleme häufig auf Relaxationen durch Intervallsysteme (die betrachtete Menge ist größerer als die ursprüngliche Menge, dafür aber einfacher strukturiert). Im Fall affiner Parameterabhängigkeiten, wie sie neben der theoretischen Modellierung von Strecken auch bei der Identifikation in einem geschlossenen stabilen Regelkreis entstehen (z. B. realer PD-Regler; Streckenparameter erscheinen mehrfach in der Regelkreisstabilitätsbedingung), empfiehlt sich das Kantentheorem, das die Stabilitätsanalyse auf mehrere Segmente (2.117a) reduziert.

Satz 2.11 (Kantentheorem, [54], [73][47])
Alle $p(\zeta) = p_0(\zeta) + \sum_{i=1}^{k} [\theta_i] p_i(\zeta)$ mit $\underline{\theta}_i \le \theta_i \le \overline{\theta}_i$ sind genau dann auf einer offenen, zusammenhängenden Teilmenge $\mathcal{G} \subset \mathbb{C}$ stabil, wenn alle $k2^{k-1}$ Kantenpolynome[48] auf $\mathcal{G}$ sind. Ist $\mathcal{G}$ die linke Halbebene, dann liegt Hurwitz-Stabilität vor; ist $\mathcal{G}$ der Einheitskreis, dann Schur-Stabilität. Es können aber auch verschobene linke Halbebenen bzw. linke Sektoren (zusätzliche Stabilitäts- bzw. Dämpfungsreserve) betrachtet werden.

Beispiel 2.18 (Anwendung des Kantentheorems)
Die Restriktionen $\underline{\theta}_{1,2} \le \theta_{1,2} \le \overline{\theta}_{1,2}$ sind mit der Stabilitätsrestriktion für das Polynom $p(s) = 5\theta_1 s^4 + (\theta_1 + 5)s^3 + s^2 + 2\theta_2 s + \theta_2$ genau dann konsistent, wenn alle Kantenpolynome

[47] Für die Erweiterung des Kantentheorems auf MIMO-Systeme siehe [621], für komplexe Polynome [76].

[48] Ein Kantenpolynom ist eine einparametrische Untermenge von $p(s)$, die durch Fixieren von $k-1$ Parametern auf einer Intervallgrenze entsteht.

$$
\begin{aligned}
p_{K1}(s;[\theta_2]) &= (5s^3+s^2)+\underline{\theta}_1(5s^4+s^3)+[\theta_2](2s+1)\\
p_{K2}(s;[\theta_2]) &= (5s^3+s^2)+\overline{\theta}_1(5s^4+s^3)+[\theta_2](2s+1)\\
p_{K3}(s;[\theta_1]) &= (5s^3+s^2)+[\theta_1](5s^4+s^3)+\underline{\theta}_2(2s+1)\\
p_{K4}(s;[\theta_1]) &= (5s^3+s^2)+[\theta_1](5s^4+s^3)+\overline{\theta}_2(2s+1)
\end{aligned}
$$

Hurwitz-stabil sind. Für $p_{K1}(s;[\theta_2])$ lässt sich der Test nach Umformung

$$
\begin{aligned}
p_{K1}(s;[\theta_2]) &= (5s^3+s^2)+\underline{\theta}_1(5s^4+s^3)+[\theta_2](2s+1)\\
&=(5s^3+s^2)+\underline{\theta}_1(5s^4+s^3)+(\underline{\theta}_2\gamma+\overline{\theta}_2(1-\gamma))(2s+1) \quad \gamma\in[0,1]\\
&=\gamma\underbrace{\left(5\underline{\theta}_1 s^4+(5+\underline{\theta}_1)s^3+s^2+2\underline{\theta}_2 s+\underline{\theta}_2\right)}_{=p_1(s)}\\
&\quad+(1-\gamma)\underbrace{\left(5\underline{\theta}_1 s^4+(5+\underline{\theta}_1)s^3+s^2+2\overline{\theta}_2 s+\overline{\theta}_2\right)}_{=p_2(s)}
\end{aligned}
$$

beispielsweise mit dem Białas-Theorem bewerkstelligen.

Anmerkung 2.26 Eine Erweiterung des Kantentheorems auf nichtlineare Parameterabhängigkeiten ist nicht zulässig, wie ein Beispiel in [3] zeigt.

Anmerkung 2.27 Die größeren Einsatzmöglichkeiten des Kantentheorems durch Zulassen anderer Stabilitätsgebiete führen dazu, dass viele Tests redundant sind. Speziell für Hurwitz-Stabilität eröffnet das verallgemeinerte Kharitonov-Theorem (Box-Theorem) [73] einen effektiveren Weg. Es betrachtet Polynommengen der Art

$$\Delta(s) = p_1(s)[a_1(s)] + \ldots + p_k(s)[a_k(s)] \tag{2.118}$$

mit festen Polynomen $p_i(s)$ und Intervallpolynomen $[a_i(s)]$.

Die bisherigen Betrachtungen beziehen sich auf Intervallpolynome. Ähnliches gilt für Intervallmatrizen. Deren Behandlung ist allerdings erwartungsgemäß schwieriger. So kann nicht in Analogie zum Kharitonov-Theorem aus der Stabilität von Eckmatrizen auf Stabilität der Intervallmatrix geschlossen werden, vgl. [52], wonach

$$[A] = \begin{bmatrix} [-1.5,-0.5] & -12.06 & -0.06\\ -0.25 & 0 & 1\\ 0.25 & -4 & -1 \end{bmatrix} \quad \text{mit} \quad A_r = \begin{bmatrix} -0.5-r & -12.06 & -0.06\\ -0.25 & 0 & 1\\ 0.25 & -4 & -1 \end{bmatrix} \tag{2.119}$$

für $r \in (0.5-\sqrt{0.06}, 0.5+\sqrt{0.06})$ instabil ist. Die Situation ist sogar noch schlimmer, denn für $n \geq 3$ reichen zum Stabilitätsnachweis noch nicht einmal die $(n-3)$-dimensionalen Randflächen[49] der Intervallmatrizen aus [492].

[49] Eine k-dimensionale Randfläche einer n-quadratischen Intervallmatrix ist eine Untermenge, die erhalten wird, indem n^2-k Elemente auf ihre Ecken gesetzt werden und k Elemente zwischen den Ecken variieren. Im Speziellen ergeben sich für $k=0$ die Eckmatrizen und für $k=1$ die Kantenmatrizen.

Bei intervall-symmetrischen Matrizen[50]

$$[A]_s = \left\{ A \in \mathbb{R}^{n\times n} : \sum_{j\geq i} a_{ij} E_{ij}^s, a_{ij} \in [a_{ij}] \right\} \quad \text{mit } E_{ij}^s = \begin{cases} E_{ij}^{n\times n} + (E_{ij}^{n\times n})^T & i \neq j \\ E_{ij}^{n\times n} & i = j \end{cases} \tag{2.120}$$

genügt es allerdings, 2^{n-1} ausgewählte Eckmatrizen zu betrachten [282], s. auch [456] für algorithmische Aspekte. Rohn [541] erweiterte dieses Resultat dahingehend, dass die Stabilität von 2^{n-1} ausgewählten symmetrischen Eckmatrizen einer symmetrischen Intervallmatrix $[A] = \{A \in \mathbb{R}^{n\times n} : \underline{A} \leq A \leq \overline{A}; \underline{A}, \overline{A} \in \mathcal{S}_n\}$ notwendig und hinreichend für die Stabilität aller symmetrischen und nichtsymmetrischen Matrizen in $[A]$ ist.
Manchmal gelingt der Stabilitätsnachweis [562] über eine einzige Testmatrix $B = ((b_{ij}))$ mit

$$b_{ii} = \overline{a}_{ii} \qquad \text{und} \qquad b_{ij} = \max\{|\underline{a}_{ij}|, |\overline{a}_{ij}|\}. \tag{2.121}$$

Hierfür stehen die Chancen gut, wenn nur wenige Matrixelemente echte Intervalle sind, wenn die Vorzeichenverteilung der $A \in [A]$ weitgehend mit der von B übereinstimmt oder wenn die A diagonaldominant sind. Der Preis für die Einfachheit sind konservative hinreichende Aussagen, wobei ergänzend zu vermerken ist, dass die Testmatrix selbst nicht in der Menge $[A]$ liegen muss.

Zugänge für parameteraffine Matrizen mit Intervallparametern (2.113) nutzen zumeist die Eigenschaft aus, dass die Menge der Systeme $\dot{x} = Ax$, für die $V(x) = x^T Px$ eine Lyapunov-Funktion ist, konvex ist [348].

Anmerkung 2.28 Abschließend sei davor gewarnt, aus der Stabilität von Intervallsystemen auf die von LTV-Systemen zu folgern, wenn sich deren Koeffizientenfunktionen für alle t innerhalb der Intervalle bewegen. Spezielle rheolineare (d. h. parametererregte) Schwingungen sollen als Gegenbeispiel dienen. Die gedämpfte Schaukel ist für alle Pendellängen asymptotisch stabil. Sobald aber die (virtuelle) Pendellänge geeignet periodisch geändert wird, entsteht eine Schwingung und die Ruhelage wird nicht mehr angestrebt. Ein derartiges System $\ddot{x} + 2\xi\dot{x} + (1 + \delta(t))x = 0$ wird in [289] quantitativ untersucht, und es wird gezeigt, dass Signale $\delta(t)$ mit $0 < |\delta(t)| < \varrho < 1$ existieren, unter denen keine asymptotische Stabilität vorliegt. Gleichwohl erfüllt wegen $0 < [1 - \varrho, 1 + \varrho]$ das Intervallsystem die Kharitonov-Bedingungen.

---

[50] $[A] \cap S_n$ ist eine intervall-symmetrische Matrix. Eine intervall-symmetrische Matrix ist aber wegen der Abhängigkeit der Nichtdiagonalelemente gemeinhin keine Intervallmatrix.

$$[A]_s = \begin{bmatrix} [-1,2] & [3,4] & [0,2] \\ * & [-2,3] & [2,7] \\ * & * & [2,4] \end{bmatrix} \text{ ist nicht zu verwechseln mit } [A] = \begin{bmatrix} [-1,2] & [3,4] & [0,2] \\ [3,4] & [-2,3] & [2,7] \\ [0,2] & [2,7] & [2,4] \end{bmatrix}.$$

2.8 Stabilität für lineare zeitvariante Systeme

Modelle für zeitvariante (zeitveränderliche, zeitvariable) Systeme, also jene mit

$$\exists t_0, T: \qquad y(t) = \Phi_{t_0}\{u(t); x_0\} \not\Rightarrow y(t-T) = \Phi_{t_0+T}\{u(t-T); x_0\}, \tag{2.122}$$

sind durch mindestens einen bezüglich t nicht konstanten Parameter gekennzeichnet. Ganz analog werden die in der Mechanik häufig auftretenden ortsvarianten Systeme definiert, die in gleicher Weise behandelt werden können. Neben Systemen, die originär zeitvariabel sind (Pendel mit variabler Länge als Laufkatzenmodell, Schwingermodell mit variabler Frequenz, schaltende Systeme, zyklische Prozesse mit periodischen Koeffizienten, Konten mit variablem Zinssatz), entstehen LTV-Systeme beim Linearisieren nichtlinearer Systeme um Trajektorien. Somit reichen die Anwendungsfelder von Messmodellen (Vortex-Durchflussmessung) über Diagnosemodelle bis hin zu adaptiven Reglern. LTV-Modelle können zudem eine Alternative zur Identifikation nichtlinearer Modelle sein. So weist das System $\dot{x}_1 = -\theta_1 x_1 + u$ und $\dot{x}_2 = \theta_2 x_1 x_2$ bezüglich x_2 bei Erregung durch einen Einheitssprung das gleiche Verhalten wie $\dot{x} = \theta_2(1 - \mathrm{e}^{-\theta_1 t})x$ auf.

Restriktionen für interne Stabilitätseigenschaften sind bereits für einfache LTV-Systeme

$$\dot{x}(t) = A(t)x(t) \qquad A(.) : [t_0, \infty) \to \mathbb{R}^{n\times n}, A(.) \in \mathcal{C}^1 \text{ und beschränkt}, x(t_0) = x_0 \tag{2.123}$$

schwierig zu formulieren, da den unterschiedlichen Ausprägungen des Stabilitätsbegriffs, vgl. Abschn. A.6.2, nur wenige algebraische Kriterien gegenüberstehen. Gemeinhin ist für jedes konkrete System eine zugeschnittene Stabilitätsanalyse zum Ableiten von Parameterrestriktionen erforderlich. Eine Ausnahme bilden folgende spezielle Klassen:

Systeme erster Ordnung, d. h. $\dot{x} = -a(t)x$, sind
– stabil, wenn $a(t) \geq 0$
– asymptotisch stabil, wenn $\int_0^\infty a(\tau)\mathrm{d}\tau = +\infty \qquad (a(t) > 0$ reicht nicht[51]$)$
– exponentiell stabil, wenn $\exists T > 0 : \int_t^{t+T} a(\tau)\mathrm{d}\tau \geq \gamma > 0, \forall t \geq 0$.

Systeme zweiter Ordnung, d. h. $\ddot{x} + a_1(t)\dot{x} + a_0(t) = 0$, s. [170]. Für $\dot{x} = A(t)x$ mit $A(.) : \mathbb{R} \to \mathbb{R}^{2\times 2}$ und $\underline{a}_{ij} \leq a_{ij}(t) \leq \overline{a}_{ij}$ findet sich in [289] ein Satz. Ergebnisse zur gleichmäßigen Stabilität werden in [543] angegeben.

Systeme zweiter Ordnung mit einem Freiheitsgrad, d. h. $\frac{\mathrm{d}}{\mathrm{d}t}[p(t)x] + q(t)x = 0$, lassen sich mit dem Sonin-Polya-Theorem untersuchen [516].

Systeme zweiter Ordnung in Vektorform, d. h. $A_2(t)\ddot{x} + A_1(t)\dot{x} + A_0(t)x = 0_n$, mit aus der Mechanik stammenden Einschränkungen an die $A_i(t)$ werden in [516] analysiert.

[51] stabil: $\dot{x} = -x/(1+t)^2$; asymptotisch stabil: $\dot{x} = -x/(1+t)$; exponentiell stabil: $\dot{x} = -tx$.

Asymptotisch zeitinvariante Systeme, d. h. $\lim_{t\to\infty} A(t) = A$ mit $A(.) \in \mathcal{C}^0$, sind asymptotisch stabil, wenn der Grenzwert A Hurwitz-stabil ist [543]. Allerdings impliziert ein Grenzwert „A ist Lyapunov-stabil“ nicht Lyapunov-Stabilität für $x(t; t_0, x_0)$ in (2.123)![52]

Dominant zeitinvariante Systeme, d. h. für $A(t) = A + C(t)$ mit stetigem $C(t)$ existiert ein $c > 0$ mit $||C(t)|| \leq c$ für alle $t \geq t_0$, sind asymptotisch stabil, wenn A Hurwitz-stabil ist, wobei c von A abhängt. Der komplexe Stabilitätsradius $r_{\mathbb{C}}^-(A)$, gepaart mit der $\mathcal{L}_\infty$-Norm, liefert eine Schranke, d. h. für $c < r_{\mathbb{C}}^-(A)$ folgt dann GAS [289].[53]

T-periodische Systeme, d. h. $\exists T > 0 : A(t + T) = A(t)$, haben die Stabilitätseigenschaften des LTI-Systems $\dot{z} = Rz$ mit $R \in \mathbb{C}^{n\times n}$, wobei R aus der Floquet-Faktorisierung der Transitionsmatrix $\Phi(t, \tau) = P(t)\mathrm{e}^{R(t-\tau)}P^{-1}(\tau)$ stammt [543], [596].

Normale Systeme, d. h. $A(t)A^T(t) = A^T(t)A(t)$, haben die gleichmäßigen Stabilitätseigenschaften, die sich aus den punktweisen Eigenwertfunktionen $\lambda_i(t)$ ergeben [424]. Somit sind schiefsymmetrische Systeme gleichmäßig stabil [543].

Kommutative Systeme, d. h. $\forall t, \tau : A(t)A(\tau) = A(\tau)A(t)$, können untersucht werden über $\Phi(t, \tau) = \exp\left(\int_\tau^t A(\eta)\mathrm{d}\eta\right)$. In diese Klasse fällt $A(t) = \alpha(t)M$ mit $\Phi(t, \tau) = \mathrm{e}^{M\eta}\Big|_{\eta=\int_\tau^t \alpha(\eta)\mathrm{d}\eta}$, wie auch $A(t) = \alpha_1(t)M_1 + \alpha_2(t)M_2$ mit $M_1M_2 = M_2M_1$.

Schaltsysteme, d. h. $A(t) \in \{A_1, \ldots, A_k\}$, sind GAS, wenn eine gemeinsamen Lyapunov-Funktion existiert (hinreichend, aber nicht notwendig) [377].

Systeme mit polytopischem Einschluss, d. h.

$$\dot{x}(t) = A(t)x(t), \quad \forall t : \ A(t) \in \mathrm{conv}\{A_1, \ldots, A_N\}, \tag{2.124}$$

sind UGAS, wenn $\exists P \succ 0_{n\times n} : A_k^T P + PA_k \prec 0_{n\times n}, k = 1, \ldots, N$ gilt [92]. Mit Annahmen an $\dot{A}(t)$ ergeben sich größere Parametergebiete für die Stabilität [205], [131].

Systeme vom Lur'e-Typ, z. B. $y^{(3)}(t) + a_2\ddot{y}(t) + a_1\dot{y}(t) + f(t)y(t) = 0$, können nach Umformung der Art $G(s) = \frac{1}{s^3+a_2s^2+a_1s}$ und zeitvariablem Regler $u = f(t)y(t)$ mit Popov-ähnlichen Kriterien mit und ohne Ableitungsbedingungen an $f(t)$ behandelt werden [363].

Ein zentrales Problem der Analyse von LTV-Systemen stellt die Tatsache dar, dass die Eigenwertfunktionen $\lambda_i(A(t))$ ohne weitere Einschränkungen keine Stabilitätsaussage gestatten, selbst dann nicht, wenn $\exists c < 0, \forall t \geq 0, \forall i : \Re e\lambda_i(A(t)) \leq c < 0$ gilt.[54]

---

[52] Die zu $\ddot{x} - (1/t)\dot{x} + x = 0$ korrespondierende Systemmatrix ist für $t \to \infty$ stabil, während die Lösungen $\sin t - t\cos t$ und $\cos t + t\sin t$ unbeschränkt sind. Um Beschränktheit für im Grenzwert stabile asymptotisch zeitinvariante Systeme zu sichern, sind weitere Voraussetzungen wie absolute Integrierbarkeit, beschränkte Variation, Stabilität von $A(t)$ für alle t und Einfachheit der Eigenwerte notwendig [121].

[53] $\forall t \geq 0 : ||C(t)|| < r_{\mathbb{C}}^-(A)$ reicht nicht, vgl. $\dot{x} = (-1 + (1 - \mathrm{e}^{-t}))x$ mit $r_{\mathbb{C}}^-(A) = 1$ und $x(t) = \mathrm{e}^{\mathrm{e}^{-t}}x(0)$.

[54] Das Konzept der dynamischen Eigenwerte [18], [649], [351], die mit den Elementarmodi $c_i(t)\,\mathrm{e}^{\int_0^t \lambda_i(\tau)\mathrm{d}\tau}$ korrespondieren, bietet eine alternative Behandlungsmöglichkeit.

Beispiel 2.19 (Instabiles LTV-System, obwohl punktweise Hurwitz-stabil[55])

$$x(t) = \begin{bmatrix} \mathrm{e}^t \sin(2t) \\ \mathrm{e}^t \cos(2t) \end{bmatrix} \qquad \text{löst} \qquad \dot{x} = \begin{bmatrix} 1 - 4\cos^2(2t) & 2 + 2\sin(4t) \\ -2 + 2\sin(4t) & 1 - 4\sin^2(2t) \end{bmatrix} x.$$

Somit ist das System bzw. die Lösung $x(t) \equiv 0_2$ trotz der Eigenwerte $\lambda_{1/2}(t) \equiv -1$ instabil (Nemytskii-Vinograd-Beispiel [391]).

Beispiel 2.20 (Stabiles LTV-System, obwohl punktweise instabil, [635])

$$\dot{x} = \begin{bmatrix} -\frac{11}{2} + \frac{15}{2}\sin(12t) & \frac{15}{2}\cos(12t) \\ \frac{15}{2}\cos(12t) & -\frac{11}{2} - \frac{15}{2}\sin(12t) \end{bmatrix} x$$

ist exponentiell stabil, obwohl für die Eigenwerte $\lambda_1(t) \equiv 2$, $\lambda_2(t) \equiv -13$ gilt, woraus fälschlicherweise Instabilität geschlossen werden könnte.

Langsam-zeitvariante Systeme

Obschon die Eigenwertfunktionen für sich allein zur Stabilitätsuntersuchung nicht geeignet sind, besteht die berechtigte Vermutung, dass sie für Systeme mit sich langsam ändernden Parametern (Alterung; temperaturschwankungsbedingt) ausreichen sollten. Das führt auf langsam-zeitvariante Systeme, also Systeme, für die die Vermutung gilt. Doch damit ist Keinem geholfen, denn dem Vorsatz „langsam“ fehlt die mathematische Strenge. Letztlich wird sich mit Kriterien beholfen, die über eine Norm die Änderungsgeschwindigkeit $\dot{A}(t)$ messen und sie in Beziehung zur Parametergröße (gemessen über eine Norm von $A(t)$) und zu den punktweisen Eigenwerten setzen. Kriterien dieser Art gibt es reichlich, s. z. B. [17]. Weitere und andersartige Kriterien finden sich in [543], [374], [363], [415], [307], [400].
Um die Langsamkeit parametrierbar zu machen, wird ein Parameter $\alpha \gg 0$ eingeführt, d. h. $\dot{x} = A(t/\alpha)x$ oder allgemeiner $\dot{x} = f(x, t, t/\alpha)$. Dadurch kann eine Aufspaltung in schnelle und langsame Systemteile gelingen, was einen Zwei-Zeitskalen-Zugang eröffnet, s. Abschn. A.6.4. Auch lassen sich Abschätzungen an α finden oder geeignete Lyapunov-Funktionen konstruieren [447].

Anmerkung 2.29 Das Stabilitätsverhalten von LTV-Systemen kann sich – anders als bei LTI-Systemen – unter äußerer Anregung grundlegend ändern. Vielschichtige Schwierigkeiten entstehen, wenn der führende Polynomkoeffizient in der E/A-Darstellung Nullstellen in t hat. Bereits einfachste Beispiele erster Ordnung zeigen dann Phänomene, die LTI-Systemen völlig fremd sind (Singularitäten unterschiedlichen Typs, keine frei wählbaren Anfangswerte usw.), vgl. die sechs Beispiele in [308]. Insofern sollte die Gültigkeit eines Modells mit zeitvariablem führenden Koeffizienten und möglichen Nullstellen nochmals sorgfältig überprüft werden.

[55] Vergleiche auch $x[k+1] = \begin{bmatrix} -\cos(k\pi/2) & 1+\sin(k\pi/2) \\ -1+\sin(k\pi/2) & \cos(k\pi) \end{bmatrix} x[k]$; $x[0] = [0,1]^T$ mit den Eigenwerten $\lambda_{1,2}(k) = 0$ und unbeschränkter Lösung.

2.9 Stabilität für nichtlineare Systeme

Für nichtlineare Systeme gibt es keine allgemeinen parameterbezogenen algebraischen Stabilitätskriterien, da die Vielfalt der nichtlinearen Systeme das nicht zulässt. Zudem existieren mehrere Stabilitätsdefinitionen, die sich in Anschaulichkeit und Handhabbarkeit unterscheiden. So kann ein System in dem einen Sinn stabil sein, in einem anderen aber nicht. In diesem Abschnitt soll nicht auf die unterschiedlichen Stabilitätsdefinitionen eingegangen werden. Hierzu sei auf die Zusammenstellung im Anhang A.6.2 verwiesen. Stattdessen sollen einige Zugänge aufgezeigt werden, um das schwierige Problem der Stabilitätsrestriktionen bei der Identifikation nichtlinearer Systeme zu umgehen oder abzuschwächen.

1. Zugang: Ableiten von Bedingungen aus der Lösung
Einige Systeme erster und zweiter Ordnung ohne äußere Erregung oder mit einfachen äußeren Erregungen (Sprünge, Sinusfunktionen) lassen sich geschlossen lösen. In Standardwerken wie Kamke [333], Zwillinger [653] oder Polyanin und Zaitsev [512] finden sich für derartige Systeme die entsprechenden Ansätze. Aus der allgemeinen Lösung ergeben sich dann die Parameterrestriktionen.

2. Zugang: Ableiten einer Bedingung aus einem Stabilitätskriterium
Die Schwierigkeit erwächst hierbei aus der Vielzahl der Stabilitätsdefinitionen. Zu diesem Zweck sind in Abschn. A.6.2 Definitionen, Implikationen und Hinweise auf Kriterien zusammengestellt. Anhand eines passenden Kriteriums entscheidet sich, ob Stabilität strukturell über den Modellansatz oder durch parametrische Restriktionen gesichert wird. Speziell für Systeme niedriger Ordnung sind Kriterien bekannt, vgl. das nachfolgende Beispiel, [352], [49], [518], [461], [400], [639] für Systeme zweiter, [264], [496], [5], [400] für Systeme dritter und die Quellenangaben in [264] für einige Systeme vierter Ordnung.

Beispiel 2.21 (Stabilitätsrestriktion für ein System zweiter Ordnung, [614])
Es sei $\dot{x}_1 = x_2, \dot{x}_2 = -f(x_2, \theta_1) - g(x_1, \theta_2)$ mit $f, g \in \mathcal{C}^0$ zu untersuchen. Aus der Lyapunov-Funktion $V(x_1, x_2) = x_2^2 + 2\int_0^{x_1} g(\gamma, \theta_2)\mathrm{d}\gamma$ folgen unter Ausnutzung von $\frac{\mathrm{d}}{\mathrm{d}t}\int_0^{x_1} g(\gamma)\mathrm{d}\gamma = g(x_1)\dot{x}_1$ die hinreichenden Restriktionen für UAS in $\mathcal{I} = [-\gamma_0, \gamma_0]$ zu

$$\begin{aligned} \gamma f(\gamma, \theta_1) \geq 0 \qquad & \forall \gamma \in \mathcal{I} \qquad && \text{aus } \dot{V} \leq 0 \\ \gamma g(\gamma, \theta_2) > 0 \qquad & \forall \gamma \in \mathcal{I} \backslash 0 \qquad && \text{aus } V > 0. \ /^{56} \end{aligned}$$

[56] $\gamma g(\gamma) > 0 \Rightarrow \gamma > 0, g(\gamma) > 0 \Rightarrow \int_0^x g(\gamma)\mathrm{d}\gamma > 0$ für $x > 0$ oder $\gamma < 0, g(\gamma) < 0 \Rightarrow -\int_x^0 g(\gamma)\mathrm{d}\gamma = \int_0^x g(\gamma)\mathrm{d}\gamma > 0$ für $x < 0$ und Stetigkeit der Lyapunov-Funktion V folgt aus der Integration.

3. Zugang: Verwenden spezieller approximativparametrischer Modelle
Für Nlq-Modelle[57], bilineare Modelle, Wiener- oder Hammerstein-Modelle lässt sich Stabilität häufig einfach sicherstellen oder auf die Stabilität der linearen Teilmodelle zurückführen.

4. Zugang: Relaxation eines simulativ ermittelten Stabilitätsgebiets
Für eine gegebene Modellklasse wird durch Parametervariation ein Stabilitätsgebiet simulativ bestimmt. Da dieses Gebiet i. Allg. sehr kompliziert ausfällt, wird es durch ein innenliegendes, üblicherweise konvexes Gebiet (hinreichendes Kriterium) approximiert, z. B. Ellipsoid, Quader, Polyeder. Häufig entstehen dann LSQ- oder LSI-Probleme.

5. Zugang: Vertrauen auf den numerischen Optimierungsalgorithmus
Bei Algorithmen für Ausgangsfehlerkriterien kann vielfach auf stabilitätssichernde Restriktionen verzichtet werden, da diese Algorithmen bei geeigneten Startwerten und guter Schrittweitensteuerung tendenziell Parameterkonstellationen vermeiden, die instabile Lösungen hervorrufen, die ihrerseits schlechte Gütewerte bewirken. Zu beachten ist aber, dass bei nichtlinearen Systemen kleinste Parameteränderungen ausreichen können, um gänzlich andere Lösungen zu generieren (chaotisches Verhalten, Fraktale). Als Beispiel sei die logistische Gleichung $x[k+1] = 4(1-x[k])x[k]$ genannt, die mit den Startwerten $x[0] = 0.4$ bzw. $\tilde{x}[0] = 0.4 + 10^{-8}$ bereits in den Schritten $k = 24, 25, 26$ deutlich voneinander abweichende Lösungen produziert[58], nämlich 0.53, 1.00, 0.02 bzw. 0.36, 0.92, 0.28. Eine Schätzung des richtigen Anfangswertes (benötigter Parameter in Ausgangsfehlerverfahren) ist dann ohne Kenntnis einer geschlossenen Lösung auf rein numerischem Weg kaum erfolgreich.

6. Zugang: Behandlung von Abtastsystemen im Kontinuierlichen
Handelt es sich bei dem zu identifizierenden System um ein Abtastsystem, so ist tendenziell die direkte Identifikation eines kontinuierlichen Modells der indirekten über ein diskretisiertes Modell vorzuziehen, da die zeitdiskreten nichtlinearen Modelle häufig eine kompliziertere nichtlineare Beschreibung erfordern, s. nachfolgendes Beispiel.

---

[57] NLq-Systeme sind zeitdiskrete nichtlineare Zustandsraummodelle, die aus q Schichten sich abwechselnder linearer und nichtlinearer Operatoren bestehen, die einer Sektorbedingung genügen. Sie umfassen eine Vielzahl der bekannten Künstlichen Neuronalen Netztypen. Für diese Klasse sind hinreichende Bedingungen für GAS, E/A-Stabilität und Dissipativität bekannt. Allgemeine Systembeschreibung:

$$\begin{aligned} x[k+1] &= \Gamma_1(V_1\Gamma_2(V_2 \ldots \Gamma_q(V_q x[k] + B_q u[k]) + B_{q+1}u[k]) + \ldots) + B_1 u[k]) \\ y[k] &= \Lambda_1(W_1\Lambda_2(W_2 \ldots \Lambda_q(W_q x[k] + D_q u[k]) + D_{q+1}u[k]) + \ldots) + D_1 u[k]), \end{aligned}$$

mit konstanten kompatiblen Matrizen B_i, D_i, V_i, W_i und Diagonalmatrizen Γ_i, Λ_i, deren Diagonalelemente stetige Funktionen sind, die auf $[0,1]$ abbilden [591].

[58] Die logistische Gleichung – von John von Neumann als Zufallsgenerator vorgeschlagen – hat die Lösung $x[k] = \frac{1}{2}[1 - \cos(2^n \arccos(1-2x[0]))]$ (WolframScience, http://www.wolframscience.com/nksonline/page-1098a-text?firstview=1).

Beispiel 2.22 (Diskretisierung verkompliziert das Vektorfeld)
Die Bernoulli-Differenzialgleichung mit polynomialem Vektorfeld

$$T\dot{y}(t) = y(t)u(t) - y^2(t); \qquad u(t), y(t) > 0 \tag{2.125}$$

hat für stückweise konstante Eingangserregung die exakte Diskretisierung[59]

$$y[k+1] = \frac{u[k]y[k]}{y[k] + (u[k] - y[k])\exp(-T_A u[k]/T)} \tag{2.126}$$

mit einem transzendenten Vektorfeld (Bruch und Exponentialfunktion).

Das Ableiten von Stabilitätsrestriktionen ist bei zeitdiskreten meist schwieriger, zumal erheblich weniger nützliche Sätze existieren und die Stabilitätstheorie zeitdiskreter Systeme [375] in der Lehre stiefmütterlich behandelt wird. Hinzu kommt, dass sich viele kontinuierliche nichtlineare Systeme dank einer polynomialen Struktur der Gleichungen bei der Identifikation, aber auch beim Reglerentwurf erheblich einfacher behandeln lassen. Natürlich könnten die komplizierteren Strukturen der diskreten Systeme prinzipiell auch mittels Polynomansätzen approximiert werden, doch sind dabei die erforderlichen Polynomgrade und damit die Parameteranzahl meist hoch, und zudem ist es recht schwierig, die Freiheitsgrade durch Hinzunahme von Restriktionen wieder zu reduzieren.

7. Zugang: Umformulierung als Verbundbeobachterproblem
Viele Modelle, die aus einer theoretischen Prozessanalyse stammen, werden im Zustandsraum beschrieben, wobei nicht alle Zustände messbar sind. Die Zahl der unbekannten Parameter ist dabei häufig gering. Während im Linearen den nicht messbaren Zuständen durch Verwenden der E/A-Darstellung ausgewichen wird, gelingt dies im Nichtlinearen nicht so einfach oder die entstehende E/A-Darstellung ist für eine Identifikation ungeeignet. Dann bieten sich Verbundbeobachter[60] an, die dank ihrer Konstruktion auch für driftende Parameter geeignet sind [517], [317], [397]. Verbundbeobachter können aber auch bei linearen Problemen eine Alternative darstellen, da sie die meist komplizierten nichtlinearen Beziehungen zwischen den Parametern und denen der E/A-Darstellung vermeiden. Ein weiterer Vorteil gegenüber dem E/A-Umweg sind einfachere Restriktionen an die Parameter, die aus ihrer physikalischen Bedeutung resultieren (z. B. $\theta \geq 0$). Per Parameterprojektion lassen sich die Restriktionen

[59] Die Substitution $\xi = 1/x$ führt auf ein für die Abtastperiode $kT_A \leq t < (k+1)T_A$ gültige lineare Differenzialgleichung $T\dot{\xi} + u[k]\xi = 1$. Deren Lösung und die Rücksubstitution liefert das angegebene Ergebnis.

[60] Ein Verbundbeobachter ist ein Beobachter, bei dem die Parameter als zusätzliche Zustandsgrößen aufgefasst werden, das bestehende Zustandsraummodell also durch $\dot{\theta} = 0_p$ erweitert wird. Für Parameter- und Zustandsvektor wird gleichermaßen Konvergenz gefordert. Bei dem sich ebenfalls auf das erweiterte System beziehenden adaptiven Beobachtern wird nur Konvergenz für den Zustandsvektor gefordert (vergleichbar mit adaptiven Reglern mit Identifikationsmodell, wo die Konvergenz der Regelgröße nicht, aber die der Parameter gefordert wird).

analog zur Parameterschätzung in adaptiven Regelungen [367] elegant einhalten. Nachteil der Verbundbeobachter ist der recht komplizierte Entwurf, der sich allerdings für Systeme in Adaptiver Beobachternormalform etwas vereinfacht.

8. Zugang: Identifikation im geschlossenen Regelkreis
Mitunter ist die Regelstrecke instabil und kann somit nur, von Ausnahmen wie dem Integrator abgesehen, mit Hilfe eines stabilisierenden Reglers identifiziert werden. Hierzu bieten sich die robusten und die adaptiven Regler an. Für das, wohlgemerkt überaus einfache, lineare System $\dot{y} = \theta y + u; \theta > 0$ existiert kein linearer stabilisierender Regler, sofern keine obere Schranke für θ bekannt ist. Hingegen garantiert der Regler $u = -k_1 x - k_2 x^3$ Konvergenz in eine der Ruhelagen $x_e = \pm\sqrt{(\theta - k_1)/k_2}$, woraus θ gefolgert werden kann.
Der Einsatz eines MRAC-Reglers[61] oder eines expliziten STC-Reglers ist ebenso möglich, wenngleich die Stabilitätsproblematik dann auf den Entwurf verschoben wird. Insbesondere für STC-Regler können dann Restriktionen, die über Zugang 2 ermittelt werden, notwendig sein, um per Projektionstechniken Stabilität des adaptiven Regelkreises zu sichern.

9. Zugang: Sektorbedingungen bei Lur'e-Regelkreisen
Ein Lur'e-Regelkreis besteht aus einem LTI-System und einer statischen Nichtlinearität. Beide formen ein autonomes System der Art

$$\dot{x} = Ax - b\varphi(c^T x); \qquad \varphi \in \mathcal{C}^0 : \mathbb{R} \to \mathbb{R}. \tag{2.127}$$

Für dieses lassen sich Restriktionen an φ zur absoluten Stabilität[62] über das Popov-Kriterium oder das Kreiskriterium ableiten [473], die gemeinhin im Inneren des Hurwitz-Sektors liegen. Für $\varphi(y) = \gamma y$ bezeichnet das Intervall $(\underline{\gamma}, \overline{\gamma})$ den sogenannten Hurwitz-Sektor, also jene γ, für die $\dot{x} = Ax - b(\gamma c^T x)$ Hurwitz-stabil ist[63]. Obwohl die Kalman-Vermutung, wonach UGAS für

$$\forall y: \quad \underline{\gamma} < \frac{\mathrm{d}\varphi(y)}{\mathrm{d}y} < \overline{\gamma} \qquad \text{mit } \varphi \in \mathcal{C}^1, \varphi(0) = 0 \tag{2.128}$$

vorliegt, falsch ist, ist sie doch für LTI-Systeme bis einschließlich dritter Ordnung wahr (Barabanov-Theorem, [47])[64]. Mithin wird für alle Systemordnungen LAS impliziert, was

[61] MRAC steht für „model reference adaptive control" und STC für „self tuning control".

[62] Ein Regelkreis $\dot{x} = Ax - b\varphi(c^T x)$ heißt absolut stabil im Sektor $[\underline{\gamma}, \overline{\gamma}]$, wenn seine Ruhelage $x = 0_n$ für jede Nichtlinearität φ im Sektor global asymptotisch stabil ist. Anders ausgedrückt: Die Menge aller derartigen Regelkreise mit Nichtlinearitäten φ aus dem Sektor heißt global asymptotisch stabil.

[63] Mitunter wird als Hurwitz-Sektor nur ein Sektor $(0, \overline{\gamma})$ verstanden, was sinnvoll ist, sofern auch Einfachintegratoren zugelassen werden. Eine Substitution $\tilde{\phi}(y) = \phi(y) - \underline{\gamma}$ und $\tilde{G}(s) = G(s)/(1 + \underline{\gamma} G(s))$ stellt hier die Verbindung her. Je nach Voraussetzungen muss die Sektorungleichung nicht streng sein.

[64] Demnach ist das Gegenbeispiel dritter Ordnung in Narendra und Taylor [473] falsch. Leonov [392] entwarf ein System vierter Ordnung, das die Kalman-Bedingung einhält, das aber nicht UGAS ist, da ein stabiler Grenzzyklus auftritt.

direkt aus Lyapunovs Linearisierungsmethode folgt. Für Systeme, die nicht durch die Linearisierungsmethode erfasst werden, sei auf das Yakubovich-Kriterium [640] verwiesen, in dem die Ableitung durch einen Differenzenquotienten ersetzt wird, wodurch φ keine $\mathcal{C}^1$-Funktion mehr sein muss.[65]
Die Kalman-Bedingung ist stärker als die Bedingung der Aizerman-Vermutung

$$\forall y \neq 0: \quad \underline{\gamma} < \frac{\varphi(y)}{y} < \overline{\gamma} \qquad \text{mit } \varphi \in \mathcal{C}^0, \varphi(0) = 0. \tag{2.129}$$

Nach Multiplikation mit dy in (2.128) und anschließender Integration folgt direkt (2.129), während die Umkehrung nicht gilt, vgl. $y\mathrm{e}^{-y}$ im Sektor $[0, \infty)$. Die Aizerman-Vermutung ist bereits für Systeme zweiter Ordnung falsch.[66] Dennoch erweist sich die enthaltene Bedingung in einer schwächeren Form als hilfreiche Restriktion. Die schwache Aizerman-Vermutung bezieht sich auf $[\varepsilon, K], 0 < \varepsilon, K < \overline{\gamma}$ [67] und unterstellt absolute Stabilität in $[0, K]$ für Hurwitz-stabile Systeme mit P-Verhalten und in $[\varepsilon, K]$ für bis auf einen Integrator Hurwitz-stabile Systeme. Sie ist richtig für Systeme zweiter Ordnung, Systeme dritter Ordnung mit höchstens einer Nullstelle und Systeme vierter Ordnung mit reellen stabilen Polen ohne Nullstellen! Die letzte Bedingung ist ein Korollar eines Ergebnis von Gil' [220], wonach für die Gültigkeit der Aizerman-Vermutung die Nichtnegativität der Gewichtsfunktion eines minimalen LTI-Systems hinreichend ist. Ergänzt um eine spezielle Hurwitz-Bedingung reicht die Nichtnegativität auch aus, um absolute $\mathcal{L}_\infty$-Stabilität [198] im Sinne der Aizerman-Bedingung zu gewährleisten [221]. Bedingungen für die Gültigkeit einer Aizerman-Vermutung für Mehrgrößensysteme werden in [289] gegeben, wo unter anderem GAS für $\dot{x} = Ax + B\varphi(Cx)$ garantiert wird, wenn $||\varphi(z)||_2 < r_{\mathbb{C}}^-(A, B, C)||z||_2, \forall z \in \mathbb{C}^p \backslash 0_p$. Überdies finden sich in [289] Gegenbeispiele, die die Nichtgültigkeit für zeitvariante Nichtlinearitäten und beliebige zeitvariante Störungen belegen. Eine Mehrgrößenverallgemeinerung, die auf die simultane Stabilität von Eckmatrizen eines Polytops führt, wird in [94] beschrieben.

Der Vorteil des Zugangs über Sektorrestriktionen (s. auch Abschnitt 3.8) oder allgemeiner von Zugängen, die die Stabilität ganzer Familien von Systemen erfassen, liegt in der Anwendung für approximativparametrische Modelle, da die Funktionen innerhalb der Sektorbedingungen relativ frei wählbar sind. Der Nachteil bei anderen Modellklassen besteht im

[65] Es können sogar stückweise stetige Funktionen mit isolierten Singularitäten betrachtet werden, wenn die Lösung im Sinne von Filippov [192] verstanden wird.

[66] Krasovskii-Beispiel:

$$\begin{array}{ll} \dot{x}_1 & = x_2 + \varphi(x_1) \\ \dot{x}_2 & = -x_1 - x_2 \end{array} \quad \text{mit} \quad \varphi(x_1) = \begin{cases} x_1 - [\mathrm{e}^{-2}/(1+\mathrm{e}^{-1})]x_1 & \text{für } |x_1| < 1 \\ x_1 - [\mathrm{e}^{-2|x_1|}/(1+\mathrm{e}^{-|x_1|})]\mathrm{sign}x_1 & \text{für } |x_1| \geq 1 \end{cases}$$

Wird $\varphi(x_1)$ durch kx_1 mit $0 < k < 1$ ersetzt, ist das System GAS und es gilt $0 < x\varphi(x) < x^2; x \neq 0$. Aber für die Lösung mit $x_{10} = 1, x_{20} = \mathrm{e}^{-1} - 1$, die die Gleichung $x_2 = \mathrm{e}^{-x_1} - x_1$ erfüllt, sind $\lim_{t\to\infty} x_1(t) = \infty$ und $\lim_{t\to\infty} x_2(t) = -\infty$ die Folge.

[67] $[\varepsilon, K], 0 < \varepsilon, K < \overline{\gamma}$ ist einer stärkere Voraussetzung als $(0, \overline{\gamma})$, vgl. $\mathrm{e}^{-y} > 0$, aber $\nexists \varepsilon > 0 : \mathrm{e}^{-y} \geq \varepsilon$.

Aufstellen der erforderlichen Restriktionen, wozu in der Regel Lyapunovs direkte Methode herangezogen wird [614], deren Handhabung aber einige Erfahrung erfordert. Der Nachteil eines hohen systemtheoretischen Wissens gepaart mit der erforderlichen Erfahrung für die praktische Anwendung ereilt aber auch die anderen Zugänge dieses Abschnitts, da es im Nichtlinearen keine prädestinierte und damit auch keine schematische Herangehensweise gibt. Insofern sind die angeführten Zugänge als Ideenskizzen zu sehen, um die eigenen Identifikationsprobleme anders oder besser anzugehen.

2.10 Passivität

Eine wichtige Systemeigenschaft ist die Passivität, die nur kausale Systeme aufweisen können [482]. Passivität bedeutet, dass das System keine Energie generiert, sondern sie nur absorbiert oder speichert, z. B. elektrische Netzwerke aus passiven Bauelementen wie Widerständen, Kondensatoren und Spulen, beachte aber Anmerkung 2.30. Im Folgenden werden drei Konzepte zur Passivität vorgestellt und diskutiert:

1. das physikalische Konzept
2. das operatortheoretische Konzept
3. das abstrakte interne Energiekonzept.

Das physikalische Konzept bezieht sich auf die im System verfügbare Energie $E : \mathcal{X} \to \mathbb{R}_{\geq} \cup \{+\infty\}$ (bei gegebenem Anfangszustand x_0 maximal dem System entziehbare Energie)

$$E(x_0) = \sup_{\substack{t\geq 0 \\ (u,y)}} -\int_0^t u^T(\tau)y(\tau)\mathrm{d}\tau, \tag{2.130}$$

wobei das Supremum über alle Paare zulässiger Funktionen $(u(.), y(.))$ zu erstrecken ist. Aus $t = 0$ folgt die Nichtnegativität der Energiefunktion.

Definition 2.5 (Passivität)
Ein nichtlineares System in Zustandsraumdarstellung heißt passiv, wenn $E(x)$ für alle x endlich ist, andernfalls heißt es aktiv. Existiert darüber hinaus ein Punkt x mit $E(x) = 0$, dann heißt das System stark passiv.

Für äquivalente Zustände zueinander äquivalenter Systeme ist die verfügbare Energie gleich, weshalb die Passivitätseigenschaft unabhängig von der Zustandsraumdarstellung des Systems ist. Der Vorteil des Konzepts liegt in der Anschaulichkeit und der systembeschreibungsfreien Formulierung, was aber den Nachteil einer schlechten mathematischen Handhabbarkeit hat.

Das operatortheoretische Konzept $\Phi(x_0, .) : u_{[0,t]} \mapsto y_{[0,t]}$ variiert nur über u und führt die Passivitätsbetrachtung auf die Charakterisierung der Übertragungsfunktion als positiv reelle Funktion zurück. Es ist nur in der linearen Regelungstheorie anwendbar, da nur dort das Ignorieren der Anfangswerte eine vorteilhafte Behandlung erlaubt. Der Vorteil des Konzepts liegt in der guten mathematischen Handhabbarkeit; nachteilig ist die Einschränkung auf LTI-Systeme, wenngleich Erweiterungen durch das Einbeziehen isolierter statischer Nichtlinearitäten möglich sind. Das Konzept wird in Abschnitt 2.10.1 näher betrachtet.

Das abstrakte interne Energiekonzept ähnelt dem physikalischen Konzept, bezieht sich aber auf eine abstrakte Energiefunktion $V(x)$. Diese Funktion muss bei einem passiven System einen geringeren Zuwachs $V(x,t) - V(x_0,t_0)$ haben als die zugeführte Energie. Ursache ist, dass sich die zugeführte Energie in den internen Energiezuwachs und eine Verlustenergie aufteilt. Da diese Eigenschaft für alle Zeiten gilt, wächst die interne Energie entlang der Trajektorien höchstens so schnell wie die eingeprägte Energie. Das Konzept ist auf nichtlineare und zeitvariante Systeme anwendbar, seine mathematische Handhabbarkeit bewegt sich auf dem Niveau von Lyapunov's direkter Methode. Die Definition und daraus abgeleitete Schlussfolgerungen sind Gegenstand der nachfolgenden Ausführungen.

Definition 2.6 (Passives System, integrale Version, [198])
Ein System

$$\dot{x} = f(x,u,t) \qquad f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_{\geq} \to \mathbb{R}^n \text{ lokal Lipschitz-stetig} \tag{2.131a}$$

$$y = h(x,u,t) \qquad h : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_{\geq} \to \mathbb{R}^m \text{ stetig} \tag{2.131b}$$

mit $f(0_n, 0_m, t) \equiv 0_n$ und $h(0_n, 0_m, t) \equiv 0_m$ heißt passiv, wenn eine stetige positiv semidefinite Funktion $V : \mathbb{R}^n \times \mathbb{R}_{\geq} \to \mathbb{R}_{\geq}$, genannt interne Energiefunktion oder Speicherfunktion[68], existiert, sodass

$$V(x,t) - V(x_0,t_0) \leq \int_{t_0}^{t} u^T(\tau) y(\tau)\, \mathrm{d}\tau \quad \text{mit} \quad x = x(t; x_0, t_0, u) \tag{2.132}$$

für alle $u \in \mathcal{U} \subseteq \mathcal{L}_2^m, x_0 \in \mathbb{R}^n, 0 \leq t_0 \leq t < T_{u,x_0}$ gilt, wobei T_{u,x_0} die obere Zeitgrenze darstellt, für die Lösungen existieren.[69]

Unter sehr milden Anforderungen sind das physikalische Konzept und das interne Energiekonzept identisch [637], wobei sich die weiteren Betrachtungen vornehmlich auf das für die Regelungstechnik wichtige interne Energiekonzept beziehen.

[68] Die Speicherfunktion wird hier wegen der Zusatzforderung $f(0_n, 0_m, t) \equiv 0_n$ und $h(0_n, 0_m, t) \equiv 0_m$ zu einem Lyapunov-Funktionskandidat – deshalb auch der Bezeichnerwechsel auf V. Bei Definitionen ohne die Forderung braucht V nur nichtnegativ zu sein.

[69] Eine Definition für streng passive Systeme ohne Zeitobergrenze wird in [367] gegeben.

Wegen der Energieerhaltungssätze müssten alle technischen Systeme entsprechend der obigen Diskussion passiv sein. Dabei wird aber übersehen, dass durch die Wahl der Systemgrenzen (nicht interessierende Ein-/Ausgänge bleiben unberücksichtigt), der Signale (Produkt aus Eingang und Ausgang muss keine Leistung sein), der Zeitskala (Batterie ist Konstantspannungsquelle bzgl. schnellerer Vorgänge in einem elektrischen Netzwerk) usw. Modelle entstehen, die als nichtpassive Systeme zu charakterisieren sind.

In der Modellbildung ist also zunächst ein für die weitere Anwendung zweckmäßiges Konzept zu wählen. Hierzu muss klar definiert werden, was unter Passivität verstanden wird, denn nicht jede Definition in der Literatur ist geeignet, ein System als passiv zu klassifizieren [637]. Wie so oft bereiten Anfangswerte, Nichtproperheit, Nichtminimalität, Nichtglattheit oder Systeme ohne Nichtnull-Energiezustand (Infimum) Probleme.

Anmerkung 2.30 Ein System mit passiven Elementen ist nicht zwingend passiv, sondern nur dann, wenn das Produkt aus Ein- und Ausgangssignalen physikalisch einer Leistung entspricht (Strom-Spannung, Kraft-Geschwindigkeit). So ist die E/A-Spannungsübertragung eines Doppel-RC-Glieds (Reihenschaltung von R_1C_1-Glied und R_2C_2-Glied) durch

$$G(s) = \frac{1}{R_1C_1R_2C_2s^2 + (R_1C_1 + R_2C_2 + R_1C_2)s + 1} \tag{2.133}$$

gegeben und wegen ihres Differenzgrad von 2 nicht passiv, während die Eingangsstrom-Eingangsspannungsübertragung passiv ist

$$G(s) = \frac{R_1C_1R_2C_2s^2 + (R_1C_1 + R_2C_2 + R_1C_2)s + 1}{(C_1 + C_2)s \cdot (R_2\frac{C_1C_2}{C_1+C_2}s + 1)}. \tag{2.134}$$

Anmerkung 2.31 Passive Systeme sind immer quadratisch, d. h. die Anzahl der Ein- und Ausgänge ist gleich. Sie stellen einen Spezialfall der dissipativen Systeme im Sinne von Willems dar [198], bei denen die Anzahl der Ein- und Ausgänge differieren können und bei denen die Einspeisung keine Leistung, ausgedrückt durch ein Skalarprodukt, sein muss.

Anmerkung 2.32 Ein passives System kann Lösungen haben, die nur auf einem endlichen Zeitintervall definiert sind, vgl. $\dot{x} = -x + x^2u, y = x^3$. Aus $V(x) = \frac{1}{2}x^2$ und $\dot{V} = -x^2 + x^3u = -x^2 + yu$ folgt nach Integration $V(x) - V(x_0) = \int_0^t uy\mathrm{d}\tau - \int_0^t x^2\mathrm{d}\tau$, also $V(x) - V(x_0) < \int_0^t uy\mathrm{d}\tau$. $x(t)$ eine endliche Fluchtzeit für $u(t) = \mathrm{e}^{-t}$; das System ist nicht ISS-stabil.

Anmerkung 2.33 Ein passives System muss nicht minimal sein, vgl. $\dot{x}_1 = -4x_1 + 2x_2, \dot{x}_2 = -x_2 + u, y = x_2$. Es gilt $uy = ux_2 \geq \frac{\mathrm{d}}{\mathrm{d}t}(\frac{x_1^2}{2} + \frac{x_2^2}{2}) = -4x_1^2 + 2x_1x_2 - x_2^2 + x_2u = ux_2 - [x_1, x_2]^T \begin{bmatrix} 4 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. Das positiv reelle $G(s)$ impliziert $E(x_2) \geq 0$ und die vollständige Steuerbarkeit, dass in x_1 endliche Energie eingebracht wird, weshalb $E(x) \geq 0$ gilt.

Anmerkung 2.34 Ein passives System muss nicht proper sein, vgl. $G(s) = K_P(1 + \frac{1}{sT_N} + sT_D)$ mit $K_P > 0$, wobei der PID-Regler nicht in der Systemklasse (2.131) enthalten ist. Da das Skalarprodukt kommutativ ist, kann durch Tauschen der Signale der Differenzierer (z. B. Kondensator) nach (2.130) als passiv eingestuft werden. Die Addition in $G(s)$ entspricht der Parallelschaltung und die erhält die Passivitätseigenschaft. Beachte: Der Differenzierer ist nicht BIBO-stabil.

Anmerkung 2.35 Ein passives System muss nicht glatt sein, vgl. $\theta_1\ddot{y} + \theta_2\dot{y} + \theta_3 \text{sign}(\dot{y}) = u$ mit $\text{sign}(\dot{y}) = \begin{cases} 1 & \dot{y} > 0 \\ [-1, 1] & \dot{y} = 0 \\ -1 & \dot{y} < 0 \end{cases}$ (mechanisches System mit trockener Reibung).

Anmerkung 2.36 Passive Systeme mit positiv definiter Verlustfunktion haben Lyapunov-stabile Ruhelagen $x = 0_n$, denn $u \equiv 0_m$ impliziert, dass V entlang der Eigenvorgänge nicht steigt. Falls das passive System eine Nulldynamik hat, ist diese Lyapunov-stabil, was analog aus $y \equiv 0_m$ folgt. Strenge Passivität oder strenge Ausgangspassivität und Zustandsbeobachtbarkeit implizieren eine asymptotisch stabile Ruhelage $x = 0_n$ [345].

Anmerkung 2.37 Die abstrakte Funktion $V(x)$ ist im Gegensatz zu $E(x)$ in der Regel nicht eindeutig festgelegt [637].

Anmerkung 2.38 $E(x) \neq 0$ für alle x kann gelten (d. h. für kein x ist die Energie Null), während $x = 0 \Rightarrow V(x) = 0$ in der Definitheitsforderung steckt. Das System $\dot{x} = u, y = \mathrm{e}^x$ liefert einen solchen Fall. Da die Energie im Integrator nur von den Zuständen abhängt, gilt nämlich $E(x) = \sup_{x_2} \int_x^{x_2} -\mathrm{e}^\xi \mathrm{d}\xi = \sup_{x_2}(\mathrm{e}^x - \mathrm{e}^{x_2}) = \mathrm{e}^x > 0$ [637].

Neben der integralen Version existiert eine differenzielle Version, die zudem in [345] etwas verfeinert wird. Unter Hinzunahme weiterer Eigenschaften an $V(x, t)$ lassen sich daraus Zusammenhänge zu diversen Stabilitätskonzepten ableiten [345].

Definition 2.7 (Passives System, differenzielle Version, [345])
Für stetig differenzierbares V heißt (2.131) passiv, wenn

$$u^T y \geq \frac{\partial V}{\partial t} f(x, u) + \frac{\partial V}{\partial x^T} f(x, u, t) + \epsilon u^T u + \delta y^T y + \rho\psi(x) \qquad \forall (x, u) \in \mathbb{R}^n \times \mathbb{R}^m \quad (2.135)$$

gilt, wobei $\epsilon, \delta, \rho \in \mathbb{R}_{\geq}$ Konstanten sind und $\psi(x)$ eine positiv semidefinite Funktion ist. $\rho\psi(x)$ heißt Zustandsdissipationsrate. Ferner heißt das System
- verlustfrei (konservativ), wenn Gleichheit für $\epsilon = \delta = \rho = 0$ gilt, (z. B. $\dot{x} = u, y = x$)
- eingangs-streng-passiv, wenn $\epsilon > 0$, (z. B. $\dot{x} = u, y = x + u$)
- ausgangs-streng-passiv, wenn $\delta > 0$, (z. B. $\dot{x} = -x + u, y = x$)
- zustands-streng-passiv, wenn $\rho > 0$, (z. B. $\dot{x} = -x + u, y = x$).

Als Folgerung ergeben sich Definitionen für speicherlose passive Systeme, aus denen die Restriktionen direkt ablesbar sind, s. Abschn. 3.8 für deren Behandlung.

Definition 2.8 (Passivität speicherloser Systeme, [345])
Wenn die folgenden Beziehungen für alle u, t gelten, heißt das System $y = f(u, t)$
- passiv für $u^T y \geq 0$, /[70]
- verlustfrei für $u^T y = 0$,
- passiv mit Eingangsvorsteuerung, wenn $u^T y \geq u^T \phi(u)$ für ein ϕ gilt,
- eingangs-streng-passiv für $\forall u \neq 0 : u^T y \geq u^T \phi(u) > 0$,
- passiv mit Ausgangsrückkopplung, wenn $u^T y \geq y^T \phi(y)$ für ein ϕ gilt,
- ausgangs-streng-passiv für $\forall y \neq 0 : u^T y \geq y^T \phi(y) > 0$.

2.10.1 Passive Systeme und positiv reelle Funktionen

Im vorangehenden Abschnitt wurde die Passivität definiert und an Beispielen erklärt. Zwar liefert die Definition 2.7 eine Möglichkeit, Passivität zu überprüfen, doch ist die Konstruktion einer Speicherfunktion $V(x)$ eine ebenso schwierige Aufgabe wie die Konstruktion einer Lyapunov-Funktion. Für speicherlose Systeme enthält die Definition 2.8 gut handhabbare Restriktionen. Für die Klasse der LTI-Systeme ergeben sich Restriktionen aus der Beschreibung der passiven Systeme durch positiv reelle Funktionen. Wesentliche Zusammenhänge und abgeleitete Restriktionen sind Gegenstand dieses Abschnitts. Einen guten Einstieg bietet auch [357], wo durch Venn-Diagramme die Beziehungen für LTI-Systeme aufzeigt werden. Den Ausgangspunkt bildet hier zunächst die Definition.

Definition 2.9 (Positiv reelle Matrixfunktion, [357])
Die Funktion $G : \mathbb{C} \to (\mathbb{C} \cup \infty)^{m \times m}$ heißt positiv reell (PR)[71], wenn

$$G(s) - \text{ analytisch auf } \{s : \Re e\, s > 0\} \quad (\Rightarrow \text{Stabilität bzw. Grenzstab.}) \tag{2.136a}$$

$$\overline{G(s)} = G(\bar{s}) \text{ für alle } s \in \mathbb{C} \quad (\Rightarrow \text{reelle Gewichtsfunktion}[72]) \tag{2.136b}$$

$$G(s) + G^T(\bar{s}) \succeq 0_{m \times m} \text{ für alle } s : \Re e\, s > 0 \quad (\Rightarrow \text{kein negativer Widerstand}), \tag{2.136c}$$

und heißt streng positiv reell (SPR), wenn $G(s - \varepsilon)$ für ein $\varepsilon > 0$ positiv reell ist.

[70] $u^T f(u) \geq 0$ garantiert einen nichtnegativen Integralwert in (2.135). Ferner, ist $u_0^T f(u_0) < 0$ für ein u_0 auszuschließen, da sonst $u(\tau) \equiv u_0$ das Integral negativ macht.

[71] Andere Namen sind positiv analytische Funktion, passive Funktion, Carathéodory-Nevanlinna-Funktion, Weyl-Funktion, Titchmarsh-Weyl-Funktion oder im reell-rationalen Fall auch Brune-Funktionen.

[72] Vielfach, z. B. [367], wird statt (2.136b) alternativ $G(s) \in \mathbb{R}^{m \times m}$ für alle $s \in \mathbb{R}$ gefordert, was zu (2.136b) nicht äquivalent ist, vgl. $G(s) = \begin{cases} 1 & s \in \mathbb{R} \\ j & s \in \mathbb{C} \backslash \mathbb{R}. \end{cases}$ Zusammen mit (2.136a) wird die Alternativbedingung aber gleichwertig.

Anmerkung 2.39 $G(s)$ braucht nicht proper (beschränkt auf der rechten Halbebene) zu sein. $G(s) = \frac{1}{s}$ ist streng proper, $G(s) = 1$ ist proper und $G(s) = s$ ist nicht proper. Alle sind PR, aber nur $G(s) = 1$ ist SPR. $G(s) = \frac{s+1}{(s+1)(s+2)}$ ist nichtminimal, aber SPR. $G(s) = -\frac{1}{s}$ ist nicht PR, obwohl es keine Pole rechts hat.
In Anwendungen wird zusätzlich zur SPR-Bedingung noch $\text{Rang}(G(s) + G^H(s)) = m$ fast überall (äquivalent $\det \text{Re}\, G(s) \not\equiv 0$) gefordert. $G(s) = \frac{1}{s+a} 1_{2\times 2}$ ist SPR, erfüllt aber die Rangbedingung nicht.
Die Bedingung: $G(s - \varepsilon)$ analytisch auf $\{s : \Re e\, s > 0\}$ ist äquivalent zu $G(s)$ ist analytisch auf $\{s : \Re e\, s \geq 0\}$.

Für reell-rationale Matrixfunktionen (alle Koeffizienten sind reell) lassen sich die Bedingungen vereinfachen. So entfällt (2.136b), da reelle Koeffizienten ein reelles $G(s)$ für reelles s implizieren, s. Fußnote zu (2.136b). Ferner reduziert sich (2.136a) auf eine Stabilitätsbedingung und die Untersuchung von (2.136c) kann auf die Imaginärachse beschränkt werden.

Satz 2.12 (PR/SPR-Kriterium für reell-rationale Funktionen, [345])
Ein reell-rationales $G : \mathbb{C} \to (\mathbb{C} \cup \infty)^{m\times m}$ ist genau dann PR, wenn es Hurwitz-stabil oder grenzstabil[73] ist und

$$\text{Re}\, G(j\omega) \succeq 0_{m\times m} \quad \text{für } 0 \leq \omega \leq \infty \quad /^{74} \tag{2.137}$$

mit Ausnahme der ω gilt, an denen $G(j\omega)$ Teilübertragungspolstellen auf der imaginären Achse und im Unendlichen hat, und wenn derartige Polstellen einfach sind und eine nichtnegativ definite Residuenmatrix $\lim\limits_{s\to j\omega_0} (s - j\omega_0)G(s)$ besitzen[75].
Ein properes $G(s)$ mit $\det \text{Re}\, G(s) \not\equiv 0$ ist genau dann streng positiv reell (SPR), wenn $G(s)$ Hurwitz-stabil ist, (2.137) für alle $\omega \geq 0$ streng ist und entweder $\text{Re}\, G(\infty) \succ 0_{m\times m}$, falls $\text{rg}\, \text{Re}\, G(\infty) = m$, oder $\text{Re}\, G(\infty) \succeq 0_{m\times m}$ und $\lim\limits_{\omega\to\infty} \omega^2 M^T \text{Re}\, G(j\omega) M \succ 0_{(m-q)\times(m-q)}$ für alle $M \in \mathbb{R}^{m\times(m-q)}_{m-q} : M^T \text{Re}\, G(\infty) M = 0_{(m-q)\times(m-q)}$, falls $\text{rg}\, \text{Re}\, G(\infty) = q < m$.

Korollar 2.1 (Erweiterung des SPR-Kriteriums, [269])
Ein reelles $G(s) = k\frac{s^m + b_1 s^{m-1} + \ldots + b_m}{s^n + a_1 s^{n-1} + \ldots + a_n}$ ist genau dann SPR, wenn
- $|n - m| \leq 1$, $k > 0$ und falls $|n - m| = 1$, dann noch $a_1 \neq b_1$
- $G(s)$ analytisch auf $\{s \in \mathbb{C} : \Re e\, s \geq 0\}$
- $\Re e\, G(j\omega) > 0, \forall \omega \geq 0$.

---

[73] Grenzstabil schließt hier einen Einfachpol im Unendlichen mit ein wie etwa beim Differenzierer.

[74] $\text{Re}\, G(s) \stackrel{\text{def}}{=} \frac{1}{2}(G(s) + G^H(s))$ meint den hermiteschen Teil der Toeplitz-Zerlegung und nicht den Realteil $\Re e\, G(s)$ der komplexen Zerlegung! Für SISO-Systeme sind beide äquivalent.

[75] Im Skalarfall bedeutet eine nichtnegativ definite Residuenmatrix ein positives Residuum, denn bei einem Residuum von Null läge kein Pol vor. Im Fall $G(s) = s$ liegt ein Pol im Unendlichen und mit $s = 1/p$ ergibt sich als Residuenmatrix $\lim\limits_{p\to 0}(p - 0)\frac{1}{p} = 1$, also ist $G(s) = s$ PR.

Anmerkung 2.40 Die Bedingung (2.137) kann zur Formulierung eines komplexen Passivitätsradius für $G(s) = C(sI_n - A)^{-1}B + D$ herangezogen werden, der den Abstand zu den nichtpassiven Systemen misst. Die Optimierung über das Quadrupel $(\triangle A, \triangle B, \triangle C, \triangle D)$ lässt sich auf eine zweidimensionale Suche reduzieren [491].

Anmerkung 2.41 Die SPR-Bedingungen in den Büchern Åström und Wittenmark [34], Slotine und Li [570], Ioannou und Sun [311] sowie Tao [593] sind unterschiedlich, aber allesamt nicht allgemeingültig, vgl. Gegenbeispiele in [269]. Die SPR-Bedingungen in Khalil [345] ist korrekt, aber auf propere Übertragungsfunktionen beschränkt. Sie greift nicht für das impropere $G(s) = \frac{s^2+3s+1}{s+1}$ (Differenzgrad -1). Das Beispiel widerlegt auch die Auffassung in Khalil [345], dass SPR-Systeme nur Differenzgrade 0 und 1 haben können.

Anmerkung 2.42 Die PR-Definition ist für Übertragungsfunktionen äquivalent zur Definition positiv reeller Systeme durch [113]

$$\forall u \in \mathcal{U}, t \geq 0: \qquad 0 \leq \int_0^t u^T(\tau) y(\tau)\, \mathrm{d}\tau, \quad \text{wenn } x(0_n) = 0_n. \tag{2.138}$$

Die Beziehung ermöglicht eine Erweiterung der PR-Eigenschaft auf nichtlineare Systeme. Von einer synonymen Bezeichnung als schwache passive Systeme oder dem Einsatz von (2.138) als Definition für Passivität ist abzuraten, da die Forderung in der Tat zu schwach ist [637].

Satz 2.13 (PR-Funktion, Passivität, [629])
Ein lineares wie auch nichtlineares passives System ist positiv reell. Das lineare System hat eine PR-Übertragungsfunktion. Ein vollständig steuerbares (streng) positiv reelles System mit reell-rationalem $G(s)$ ist (streng) passiv mit positiv definiter quadratischer Verlustfunktion.

Anmerkung 2.43 Für die Implikation bei nichtlinearen Funktionen ist die Steuerbarkeit durch Erreichbarkeit zu ersetzen. Die Steuerbarkeit ist essentiell, vgl. $\dot{x} = -x + [0,0]\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$, $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} x + \begin{bmatrix} 1\,0 \\ 0\,0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ mit $G(s) = \begin{bmatrix} 1\,0 \\ 0\,0 \end{bmatrix}$, da $E(x) = \infty$ oder da bei Wahl $u_1 \equiv 0$, $x_0 > 0$ und u_2 so, dass $u_2 y_2 = u_2 x \geq \dot{V}(x) = -\frac{\mathrm{d}V}{\mathrm{d}x} x$ bzw. $u_2 \geq -\frac{\mathrm{d}V}{\mathrm{d}x}$ immer ein Widerspruch erzeugbar ist, s. [637] für eine schaltungstechnische Realisierung eines solchen nichtpassiven Systems.

Aus der Definition ergeben sich Eigenschaften, die für den Strukturansatz und für Restriktionen herangezogen werden können.

Notwendige Bedingungen für PR:
- keine Pole in der rechten offenen Halbebene $\Rightarrow a_{ij,k} \geq 0$
- keine Teilübertragungsfunktionsnullstellen in rechter offener Halbebene[76] $\Rightarrow b_{ij,k} \geq 0$
- Differenzgradbedingung für $G_{ij}(s) = B_{ij}(s)/A_{ij}(s)$: $-1 \leq \operatorname{grd} A_{ij}(s) - \operatorname{grd} B_{ij}(s) \leq 1$
- Phasenbedingung: $-\frac{\pi}{2} \leq \arg G_{ij}(j\omega) \leq \frac{\pi}{2}$ (Ortskurve in $\operatorname{cl} \mathbb{C}_+$)
- $\Sigma = (A, B, C, 0)$ minimal, $\operatorname{Rang} B = m : CB = (CB)^T \succ 0_{m\times m}, CAB + (CAB)^T \preceq 0_{m\times m}$

Notwendige Bedingungen für SPR:
- alle $A_{ij}(s), B_{ij}(s)$ sind Hurwitz-stabil, aber nicht hinreichend, vgl. $G(s) = \frac{s+3}{(s+1)(s+2)}$ (Bed. 1 in Korollar 2.1 verletzt); nur hinreichend für SISO, wenn $m = n$
- Differenzgrade liegen in $\{-1, 0, 1\}$
- Phasenbedingung: $-\frac{\pi}{2} < \arg G_{ij}(j\omega) < \frac{\pi}{2}$
- für $G(s) = \frac{b_{n-1}s^{n-1}+b_{n-2}s^{n-2}+\ldots+b_0}{s^n+a_{n-1}s^{n-1}+\ldots+a_0}$ gilt $a_i, b_j > 0, a_{n-1} > b_{n-2}/b_{n-1}$ [269]
- für $G(s) = \frac{b_{n+1}s^{n+1}+b_n s^n+\ldots+b_0}{s^n+a_{n-1}s^{n-1}+\ldots+a_0}$ gilt $a_i, b_j > 0, a_{n-1} < b_n/b_{n+1}$ [269]
- $\Sigma = (A, B, C, 0)$ minimal, $\operatorname{Rang} B = m : CB = (CB)^T \succ 0_{m\times m}, CAB + (CAB)^T \prec 0_{m\times m}$

Anmerkung 2.44 Aus Def. 2.9 und der Differenzgradbedingung folgt die PR-Normalform

$$G(s) = sD_1 + (D_0 + C(sI_n - A)^{-1}B) \text{ mit } D_1 \succeq 0_{n\times n},\ D_0 + C(sI_n - A)^{-1}B \in \text{PR}. \quad (2.139)$$

Der sD_1-Term entfällt bei klassischen LTI-Systemen, kann bei Deskriptorsystemen aber auftreten. In [396] wird eine Transformation auf die PR-Normalform für $\Sigma = \{E, A, B, C, D\}$ ausgeführt.

Anmerkung 2.45 Die PR-Eigenschaft ist invariant unter Tustin-Diskretisierungen, aber nicht unter Halteglied-Diskretisierungen. Haben ein kontinuierliches LTI-System und seine Diskretisierung den Differenzgrad 1, dann kann das diskrete System nicht positiv reell sein, denn zeitdiskrete positive reelle Systeme müssen immer den Differenzgrad 0 haben.

Anmerkung 2.46 Da der Differenzgrad per Sprungexperiment leicht bestimmt werden kann ($r = 0$, wenn die Sprungantwort bei $t = 0$ unstetig ist; $r = 1$, wenn ihr Anstieg in $t = 0$ unstetig ist; $r \geq 2$ sonst) oder auch aus Daten geschätzt werden kann [643], ergibt sich ein Test, wonach für $\hat{r} \geq 2$ Passivität als Systemeigenschaft verworfen werden kann.

[76] Für SISO-Systeme entspricht dies de facto der schwachen Minimalphasigkeit, abgesehen von Mehrfachnullstellen und denen bei Unendlich. Für MIMO-Systeme besteht kein Zusammenhang zur Minimalphasigkeit, vgl. Abschn. 2.13.2.

2.10.2 Kalman-Yakubovich-Popov- und Positive-Real-Lemma

Das Kalman-Yakubovich-Popov-Lemma (KYP-Lemma) bzw. das Positive-Real-Lemma liefern notwendige und hinreichende Bedingungen an $\Sigma = (A, B, C, D)$, damit die zugehörigen Übertragungsmatrix streng positiv reell bzw. positiv reell ist.[77] Es gibt viele Versionen dieser Lemmata, die sich hinsichtlich der Voraussetzungen, der PR- oder SPR-Variante, der Formulierung in Gleichungs- oder Ungleichungsform unterscheiden. Genannt seien [594], [525], [117], [171], Erweiterungen auf Deskriptorsysteme [117], [64], zeitdiskrete Systeme [525], [636] und nichtlineare Systeme [113] existieren ebenso. Hier wird sich auf das PR-Lemma beschränkt. Das KYP-Lemma inkl. Beweis findet sich z. B. in [345]. Der Nutzen beider Lemmata liegt darin, dass statt der unendlichdimensionalen PR-Bedingungen nur endlichdimensionale PR-Bedingungen an die Matrizen betrachtet werden müssen.

Satz 2.14 (Positive-Real-Lemma, [117])
Es sei $\Sigma = (A, B, C, D)$ ein quadratisches System, $m = p$. Dann gelten für die folgenden Aussagen die Implikationen $1 \Leftrightarrow 2$, $2 \Rightarrow 3$, $3 \wedge 4 \Rightarrow 2$, $3 \wedge 5 \Rightarrow 2$ und $2 \wedge 6 \Rightarrow 7$.
1. Σ ist passiv mit quadratischer Speicherfunktion
2. Die nachfolgende LMI hat eine nichtnegativ definite Lösung P

$$\begin{bmatrix} A^T P + PA & PB - C^T \\ B^T P - C & -(D + D^T) \end{bmatrix} \preceq 0_{(n+m)\times(n+m)} \tag{2.140}$$

   bzw. für $D = 0_{m\times m}$ vereinfacht: $\exists P \succeq 0_{m\times m} : A^T P + PA \preceq 0_{m\times m}$ und $PB = C^T$
3. $G(s) = C(sI_n - A)^{-1}B + D$ ist positiv reell und $\det \operatorname{Re} G(s) \not\equiv 0$
4. Σ ist minimal
5. A Hurwitz-stabil und B vollen Spaltenrang[78]
6. (C, A) ist beobachtbar
7. P ist positiv definit.

Die LMI-Restriktion kann für Projektionen auf die passiven Systeme genutzt werden, wenn A Hurwitz-stabil ist und nicht verändert werden soll. Diese Lösung liefert im Allgemeinen aber nicht das nächstgelegene passive System, denn dazu müsste A frei gelassen werden. Hierin liegt auch der Grund, warum der Zugang einer Projektion auf die stabilen Matrizen und einer anschließenden Projektion auf die passiven Systeme mit festem A meist relativ schlechte Ergebnisse liefert, vgl. [228].

[77] Mitunter wird diese feine Unterscheidung nicht getätigt und beide werden synonym gebraucht.

[78] Diese Bedingung ist stärker als die in [525], (A, B) steuerbar und $\det(j\omega I_n - A) \neq 0$ für $\omega \in \mathbb{R}$. Dafür wird in 2. nicht nur die Existenz einer symmetrischen, sondern einer nichtnegativ definiten Lösung gefolgert.

Anmerkung 2.47 Für die Einhaltung der PR-Bedingung bei sog. parametrischen Kovarianzmodellen (LTI-Zustandssysteme, die eine mit der Messung korrespondierende Ausgangskovarianzfolge „erzeugen“ können) wird anstatt auf die LMI-Technik bevorzugt auf Regularisierung [228] oder Balancierung [133] zurückgegriffen. Spezielle PR-Balancierungen [533], [632] werden auch bei der Modellreduktion passiver Systeme benutzt, da das gewöhnliche Balancierte Abschneiden die Passivität nicht erhält.

2.10.3 Auftrennung in Teilsysteme

Passivität ist invariant unter E/A-Äquivalenz[79], d. h. Zustandstransformationen ändern die Passivität nicht. Reihen-, Parallel- und Rückkopplungsschaltungen [367], [345] (streng) passiver Systeme ergeben wieder ein (streng) passives System. Partielle Parallelschaltungen und partielle Rückkopplungen, bei denen nur Teile des Systems parallel oder rückgekoppelt geschalten werden, sind ebenso wieder passiv [46].

Die Zusammenschalteigenschaft passiver Systeme ermöglicht es, die Passivität an das Gesamtsystem auf Passivitätsforderungen an die Teilsysteme aufzuspalten. Für Wiener- und Hammerstein-Modelle muss also sowohl der lineare dynamische als auch der nichtlineare statische Teil passiv sein. Das Prinzip der Aufspaltung der Passivitätsforderung funktioniert aber nur, wenn die System- und Modellblockstrukturen übereinstimmen und die Art der Zusammenschaltung festliegt. So ergibt zwar die Parallelschaltung passiver Systeme ein passives System, aber umgekehrt erzwingt ein rationales passives System nicht notwendigerweise, dass jedes der Teilsysteme in der Parallelschaltung passiv sein muss.

Beispiel 2.23 (Passives System mit nicht-passivem Teilsystem)
Die Parallelschaltung von $G_1(s) = \frac{-1}{s+1}$ und $G_2(s) = \frac{2}{s+2}$ erzeugt das passive System $G(s) = \frac{s}{(s+1)(s+2)}$, ohne dass $G_1(s)$ passiv ist.

Partialbruchzerlegungen oder Partialbruchfaktorisierungen rationaler Übertragungsfunktionen in Systeme erster und zweiter Ordnung und Restriktionen an die Teilsysteme sind also nur hinreichend für Passivität. Somit kann auf diese Weise zwar Passivität gesichert werden, doch unter Umständen zum Preis einer schlechten Approximation.

[79] Passivität hängt nur von den Eingangs- und Ausgangsgrößen ab, was in den integralen Definitionen deutlicher wird. Die Invarianz folgt aus (2.135) mit $\frac{\partial V(x)}{\partial x^T} f(x,u) = \frac{\partial V(\phi(z))}{\partial z^T} \tilde{f}(z,u)$, wobei $\dot{z} = \tilde{f}(z,u) = (\frac{\partial \phi}{\partial z^T})^{-1} f(\phi(z),u)$ mit bijektiver, stetig differenzierbarer Transformation $x = \phi(z)$.

2.11 Externe Positivität

Bei der Modellierung von elektrischen Netzwerken, chemischen Reaktoren, Wärmeübertragern, Altersstrukturen in Populationsmodellen oder der Ausbreitung bzw. Akkumulation von Substanzen sind die Eingangs-, Zustands- und Ausgangssignale (Verhältnisse, Konzentrationen, Dichten, Alter usw.) nichtnegativ. Solche Systeme werden positiv-linear genannt.

Definition 2.10 (Externe Positivität)
Ein SISO-LTI-System heißt extern positiv, wenn es auf jedes nichtnegative Eingangssignal bei Erregung aus dem Nullzustand mit einem nichtnegativen Ausgangssignal antwortet, wenn es eine nichtnegative Gewichtsfunktion besitzt oder wenn die Systemantwort auf alle monotonen Eingangssignale ebenfalls monoton ist. Analoges gilt im Zeitdiskreten. Bei autonomen Systemen entfällt die Spezifizierung „extern“; sie heißen einfach positiv, wenn jede Bewegung des Zustandsvektors, die im nichtnegativen Orthanten startet, dort verweilt.

Die Gewichtsfunktion bzw. -folge ist somit geeignet, die externe Positivität durch Restriktionen zu fassen. Im Gegensatz zur PR-Eigenschaft bleibt die externe Positivität bei Abtastung erhalten. Beide Eigenschaften sind zudem dadurch gekennzeichnet, dass ihre E/A-Energie für alle $t > 0$ positiv ist.

Anmerkung 2.48 Externe Positivität impliziert keine positiv reelle Übertragungsfunktion, denn externe Positivität ist nicht an einen Differenzgrad $|r| \leq 1$ gebunden und nicht auf stabile Systeme beschränkt, vgl. $G(s) = \frac{1}{s-2}$ mit $g(t) = \mathrm{e}^{2t} \geq 0$. Umgekehrt impliziert PR nicht externe Positivität, wie $G(s) = \frac{ds+b}{s+a} = d + \frac{b-da}{s+a}$ mit $a, b, d > 0$ und $b < da$ zeigt, denn, sobald $u(t) = 0$ für ein t ist, bewirkt die negative Gewichtsfunktion des zweiten Summanden, dass das Ausgangssignal negativ wird.

Eine einfache hinreichende Koeffizientenbedingung für eine nichtnegative Gewichtsfunktion bei Systemen mit durchweg reellen Polen $s_n \leq \ldots \leq s_1 < 0$ liefert das Zählerpolynom $b(s); b_i \geq 0$ [221]

$$\frac{\mathrm{d}^k}{\mathrm{d}s^k} b(s) \geq 0 \qquad s_n \leq s \leq s_1, k = 0, \ldots, m. \tag{2.141}$$

Restriktionen für E/A-Modelle werden in [185] beschrieben.

Eine Konsequenz aus der Systemeigenschaft sind für Zustandsraumsysteme die Restriktionen

$$A \in \mathbb{R}_{\geq}^{n\times n}, \quad B \in \mathbb{R}_{\geq}^{n\times m}, \quad C \in \mathbb{R}_{\geq}^{p\times n}, \quad D \in \mathbb{R}_{\geq}^{p\times m}. \tag{2.142}$$

für den Fall zeitdiskreter Systeme. Im Fall zeitkontinuierlicher Systeme sind B, C, D gleichfalls nichtnegativ, wohingegen A eine Metzler-Matrix ist ($a_{ij} \geq 0$ für $i \neq j$) [418].[80] Die

[80] Äquivalent: $\exists \gamma > 0 : (\gamma I_n + A) \in \mathbb{R}_{\geq}^{n\times n}$. Überdies ist eine Metzler-Matrix genau dann Hurwitz-stabil, wenn $-A^{-1} \in \mathbb{R}_{\geq}^{n\times n}$ [418].

Diagonalelemente können negativ sein und sind wegen der Stabilität gemeinhin negativ. Weitere Restriktionen können hinzukommen, etwa bei stochastischen Prozessen, da Wahrscheinlichkeiten stets ≤ 1 sind. Das schränkt A oft auf die Klasse der stochastischen bzw. doppelstochastischen[81] Matrizen ein. Eine Anwendung zum Generieren von Fuzzy-Regeln, bei der die Bedingung „A ist eine stochastische Matrix“ aus Forderungen an die Zugehörigkeiten resultiert, wird in [316] beschrieben.

Klar ist, dass die nichtnegativen Matrizen polyedrische konvexe Kegel formen und die stochastischen bzw. doppelstochastischen Matrizen konvexe Mengen sind. Somit bereitet das Lösen von LS-Problemen keine nennenswerte Schwierigkeit.

Bisher noch nicht vollständig gelöst ist die Frage nach globaler struktureller Identifizierbarkeit positiv-linearer Systeme. Die eng verwandten Fragen der Realisierungstheorie (Minimalrealisierungen) positiv-linearer Systeme wurden aber bereits beantwortet [295]. Das Problem dabei ist, dass die üblichen Rangbetrachtungen nicht zulässig sind und mit dem sogenannten nichtnegativen Rang[82] gearbeitet werden muss.

2.12 Steuerbarkeit und Beobachtbarkeit

2.12.1 Steuerbarkeit für lineare zeitinvariante Systeme

Definition 2.11 (Steuerbarkeit)
Ein LTI-System $\Sigma = \{A, B, C, D\}$ heißt vollständig steuerbar, genauer zustandssteuerbar, wenn zu jedem Anfangszustand $x(0)$ und jedem Endzustand $x(T)$ eine stückweise stetige Eingangsgröße $u(t)$; $0 \leq t \leq T$ mit der beliebigen endlichen Zeit T existiert, die $x(0)$ in $x(T)$ überführt.

Äquivalent dazu heißt das System vollständig steuerbar, besser nullpunktsteuerbar, wenn es aus jedem beliebigen Anfangszustand in endlicher Zeit T durch eine stückweise stetige Steuerung $u(t)$ in den Nullzustand überführt werden kann. In beiden Fällen wird nicht gefordert, dass das System für $t > T$ im Endzustand verharren muss. Zudem unterliegt $u(t)$ keinerlei Stellgrößenbeschränkungen wie etwa beim Problem der Optimalsteuerung.

[81] Eine quadratische nichtnegative Matrix heißt stochastisch, wenn alle Zeilensummen Eins sind. Eine stochastische Matrix heißt doppelstochastisch, wenn alle Spaltensummen Eins sind.

[82] Ist $A \in \mathbb{R}_{\geq}^{m \times n}$, dann heißt die kleinste natürliche Zahl r_+, für die eine nichtnegative Faktorisierung $A = FG$ mit Rang-r_+-Matrizen $F \in \mathbb{R}_{\geq}^{m \times r_+}$ und $G \in \mathbb{R}_{\geq}^{r_+ \times n}$ existiert, der nichtnegative Rang. Für $A \in \mathbb{R}_{\geq}^{n \times n}$ und $n \leq 3$ stimmen der gewöhnliche Rang und der nichtnegative Rang überein; für $n = 4$ gibt es Beispiele, in denen $r_+ > \operatorname{Rang} A$ gilt [142].

Anmerkung 2.49 Definition 2.11 ist mit Blick auf andere Systemklassen weitreichender. Sie impliziert klar die Nullpunktsteuerbarkeit. Die Umkehrung folgt über die Zerlegung der Steuerung in eine nach $x(T-\tilde{T})=0$; $0<\tilde{T}<T$ und eine von $x(T-\tilde{T})=0$ nach $x(T)$. Den ersten Teil bewerkstelligt die Nullpunktsteuerbarkeit, den zweiten das Signal $u(t)=B^T(\mathrm{e}^{A(\tilde{T}-t)})^T Q_{\tilde{T}}^{-1}x(T)$ mit der Gramschen Steuerbarkeitsmatrix $Q_{\tilde{T}}$.

Die Steuerbarkeit kann durch die Kriterien von Kalman [331], [543] und Hautus [276], [543] (auch als Popov-Belevitch-Hautus-Rangtest bekannt) verifiziert werden. Allerdings eignen sich diese Kriterien eher für niedrigdimensionale Systeme und die computeralgebraische Auswertung. Für hochdimensionale Systeme und jene, die in der Nähe der Steuerbarkeitsgrenze liegen, können numerische Probleme etwa bei der Rangauswertung der Kalman-Matrix (Steuerbarkeitsmatrix zweiter Art oder Erreichbarkeitsmatrix) auftreten [495]. Dann liefert ein Zugang über die Gramsche Steuerbarkeitsmatrix (Steuerbarkeitsmatrix erster Art)

$$Q_T=\int_0^T \mathrm{e}^{A\tau}BB^T(\mathrm{e}^{A\tau})^T\,\mathrm{d}\tau \tag{2.143}$$

eine Alternative. Denn das LTI-System ist genau dann steuerbar, wenn Q_T regulär ist [543]. Für Hurwitz-stabile A kann $Q_\infty=\lim\limits_{T\to\infty}Q_T$ numerisch stabil über die Lyapunov-Gleichung $AQ_\infty+Q_\infty A^T=-BB^T$ bestimmt werden.

Anmerkung 2.50 Die vollständige Steuerbarkeit von LTI-Systemen ist zwar invariant unter regulären Transformationen im Eingangs-, Zustands- und Ausgangsraum sowie Zustandsrückführungen der Art $u:=Kx+u$, nicht aber invariant für Reihen- und Parallelschaltung sowie die Vertauschung von Reihengliedern. Für LTI-Systeme mit $m=\dim u=\dim y$ bleibt die Steuerbarkeit unter statischen Rückkopplungen erhalten, nicht aber unter dynamischen.

Die steuerbaren Paare $(A,B)\in\mathbb{R}^{n\times n}\times\mathbb{R}^{n\times m}$ formen eine Residualmenge (Komplement einer mageren Menge[83]). Das ist eine Konsequenz der Rangbedingung an die Kalman-Matrix[84]

$$\mathrm{Rang}[B,AB,A^2B,\ldots,A^{n-1}B]<n \qquad \Leftrightarrow \qquad \text{System nicht steuerbar,} \tag{2.144}$$

[83] Eine Menge $\mathcal{M}$ in einem topologischen Raum heißt nirgends dicht, wenn ihr Abschluss kein Inneres hat, d. h. wenn $\mathrm{int}\,\mathrm{cl}\mathcal{M}=\emptyset$ gilt (int steht für „interior“; cl für „closure“). Das Komplement einer nirgends dichten Menge in $\mathbb{R}$ ist dicht; die Umkehrung gilt nicht, vgl. $\mathcal{M}=\mathbb{Q}$. Das Gegenteil von „nirgends dicht“ ist nicht „dicht“! So ist (a,b) weder nirgends dicht (enthält innere Punkte) noch dicht in $\mathbb{R}$ (viel Platz außerhalb). Den Hurwitz-stabilen Matrizen geht es in $\mathbb{R}^{n\times n}$ genauso.
Eine Menge $\mathcal{M}$ heißt mager oder Menge erster Baire-Kategorie, wenn sie sich als abzählbare Vereinigung nirgends dichter Mengen darstellen lässt, $\mathcal{M}=\cup_{i=1}^{\infty}\mathcal{M}_i, \mathrm{int}\,\mathrm{cl}\mathcal{M}_i=\emptyset$. Abzählbare Teilmengen $\mathcal{M}\subset\mathbb{R}$ sind somit stets mager, vgl. $\mathcal{M}=\mathbb{R}$. Nicht magere Mengen heißen auch Mengen der zweiten Baire-Kategorie.

[84] Die übliche Rangbedingung $\mathrm{Rang}(B,\ldots,A^{n-1}B)<n$ lässt sich zu $\mathrm{Rang}(B,\ldots,A^{n-k}B)$ verschärfen, wobei $k=\min\{n-\mathrm{Rang}B, q-1\}$ und q als Grad des Minimalpolynoms von A sind.

nach der die nichtsteuerbaren Systeme eine magere Menge sind, genauer eine algebraische Varietät (Menge, die durch Polynomgleichungen beschrieben wird). Steuerbarkeit ist somit eine topologisch generische Eigenschaft. Praktisch bedeutet das, dass mit Wahrscheinlichkeit Eins ein steuerbares Modell geschätzt wird. Somit braucht die Steuerbarkeit nicht über Restriktionen bei der Identifikation gesichert werden. Dennoch ist die Kenntnis darüber, wie gut ein System steuerbar ist, für die Bewertung des Identifikationsergebnisses von Bedeutung. Wenn sich nämlich ein System nahe der Nichtsteuerbarkeit befindet, dann sind eine schlechte Konditionierung des Schätzproblems und hohe Parametervarianzen die Folge. Die Ursachen können in der mathematischen Modellstruktur, im physikalischen System selbst oder in den Testsignalen liegen, durch die gerade die schwach steuerbaren Modi ungenügend angeregt wurden. Gegebenenfalls ist die Identifikation eines Modells mit geringerer Ordnung eine Alternative.

2.12.2 Steuerbarkeitsradius

Ein System ist vollständig steuerbar oder eben nicht. Darüber, wie gut es steuerbar ist, geben die Kriterien keine Auskunft. Um Steuerbarkeit zu quantifizieren und damit das System, das identifizierte Modell oder eine Approximation von System bzw. Modell beurteilen zu können, sind Maße erforderlich. Abgeleitet aus Eigenwert-Eigenvektor-orientierten Kriterien können Aussagen über bevorzugte Eingänge und schlecht ansteuerbare Zustandskombinationen gewonnen werden. Die Gramsche-Steuerbarkeitsmatrix gibt Hinweise zur benötigten Energie. Beispielhaft seien hier die Dominanzmaße von Litz [405], die Maße von Lückel und Müller [414] sowie die beiden bezüglich des Steuerbarkeitsbegriffs konsistenten Maße[85] von Benninger und Rivoir [65] sowie der komplexe Steuerbarkeitsradius nach Paige [495] genannt.[86] Welche Konsequenzen eine schlechte Steuerbarkeit hat und warum eine quantitative Bewertung des identifizierten Modells zweckmäßig ist, verdeutlicht das folgende Beispiel.

Beispiel 2.24 (Probleme bei schlechter Steuerbarkeit, [108])
Das System

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} \delta \\ \varepsilon \end{bmatrix} u$$

ist genau dann steuerbar, wenn $\delta \neq 0$ und $\varepsilon \neq 0$ gilt. Demnach wird es nicht steuerbar für $\varepsilon = 0$ und sogar nicht stabilisierbar (und damit natürlich auch nicht steuerbar) für $\delta = 0$. Eine Eigenwertplatzierung auf -2 und -4 durch Zustandsrückführung führt auf

[85] Das Maß hat den Wert 0, wenn das System nicht steuerbar ist.

[86] Das naheliegende Maß $\sigma_{\min}([B, AB, \ldots, A^{n-1}B])$ ist ungünstig, da es bedingt durch die Produkte unter Umständen eine viel zu kleine zulässige Störung in $\Delta A, \Delta B$ impliziert, um ein steuerbares in ein nichtsteuerbares System zu überführen, s. [140] für ein Beispiel.

$k_p = (\frac{-15}{2\delta}, \frac{3}{2\varepsilon})$, d. h. die Verstärkung strebt nach Unendlich, wenn das System schlechter steuerbar wird. In [108] wird zudem gezeigt, dass bei einem LQ-optimalen Reglerentwurf die P-Matrix der Riccati-Gleichung für $\delta \to 0$ in der Norm gegen Unendlich strebt und somit das Lösungsverfahren extrem schlecht konditioniert wird.

Analog zum Stabilitätsradius wird der komplexe Steuerbarkeitsradius nach Paige

$$\delta_2^{\mathbb{C}}(A,B) \stackrel{\text{def}}{=} \min ||[\triangle A, \triangle B]||_2 \quad \begin{array}{l}\text{mit } \dot{x} = (A+\triangle A)x + (B+\triangle B)u \text{ nicht steuerbar} \\ \text{und } [\triangle A, \triangle B] \in \mathbb{C}^{n\times(n+m)}\end{array} \tag{2.145}$$

definiert. Seine Berechnung lässt sich auf ein skalares nichtkonvexes Problem in $\mathbb{C} \cong \mathbb{R}^2$ zurückführen

$$\delta_2^{\mathbb{C}}(A,B) = \min_{\zeta\in\mathbb{C}} \sigma_{\min}([\zeta I_n - A, B]) \qquad \text{(Eising [180])}, \tag{2.146}$$

für dessen Lösung Algorithmen in [259] und [364] angegeben sind. Das Ergebnis ist eine direkte Folgerung aus dem Mirsky-Theorem [299], das über $\sigma_{\min}$ den Abstand einer Vollrangmatrix zu den rangdefizienten Matrizen misst. Somit kann (2.146) formal als eine Verstetigung des Hautus-Kriteriums mit Nichtsteuerbarkeit bei $\sigma_{\min} = 0$ interpretiert werden.

Der Steuerbarkeitsradius offenbart einen weiteren interessanten Aspekt: Mit wachsender Systemdimension kann die Steuerbarkeit deutlich nachlassen. So ergibt sich aus

$$A = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & & & \ddots & 1 \\ 0 & \dots & \dots & 0 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad K_{(A,b)} = \begin{bmatrix} 0 & \dots & 0 & 1 \\ \vdots & & & 0 \\ & & & \\ 0 & 1 & & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix} \tag{2.147}$$

wegen des Vollrangs der Kalman-Matrix $K_{(A,b)}$ Steuerbarkeit. Die perfekte Konditionierung der Kalman-Matrix lässt keinerlei Steuerbarkeitsprobleme erahnen. Demgegenüber signalisiert[87] $\delta_2^{\mathbb{C}}(A,B) = \sin(\frac{\pi}{n+1})$ ein Absinken des Steuerbarkeitsradius mit der Systemdimension. Das entspricht auch der intuitiven Vorstellung, wonach ein höherdimensionaler Zustandsvektor durch einen Eingang bei gleicher Struktur schwieriger ansteuerbar ist als ein niedrigdimensionaler Vektor.

Nachteilig am komplexen Steuerbarkeitsradius ist, dass das nächstgelegene nichtsteuerbare Paar (A,B) zu einem steuerbaren reellen Paar (A,B) durchaus komplex sein kann. Hier bietet der reelle Stabilitätsradius einen Ausweg.

[87] Die Singulärwerte $\sigma_i([\zeta I_n - A, b])$ sind die Wurzeln der Eigenwerte $\lambda_i([\zeta I_n - A, b]^T[\zeta I_n - A])$. Für die Eigenwerte von $[\zeta I_n - A, b]^T[\zeta I_n - A] = (\zeta^2+1)I_n + \text{Jord}_n(-\zeta) + \text{Jord}_n^T(-\zeta)$ gilt $\lambda_i = \zeta^2 + 1 - 2\zeta\cos(\frac{\pi i}{n+1})$ (Spektralverschiebungstheorem und Eigenwerte der Differenzenmatrix $\text{Jord}_n(-\zeta) + \text{Jord}_n^T(-\zeta)$ [209]. Minimieren über ζ liefert $\sigma_i = \sin(\frac{\pi i}{n+1})$, wovon $i = 1$ der kleinste Singulärwert ist.

Satz 2.15 (Reeller Steuerbarkeitsradius, [301], [521])
Für $\Sigma = \{A, B, C, D\}$ mit steuerbarem $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ ist der reelle Steuerbarkeitsradius

$$\delta_2^{\mathbb{R}}(A,B) \stackrel{\text{def}}{=} \min ||[\triangle A, \triangle B]||_2 \quad \begin{array}{l}\text{mit } \dot{x} = (A+\triangle A)x + (B+\triangle B)u \text{ nicht steuerbar} \\ \text{und } [\triangle A, \triangle B] \in \mathbb{R}^{n\times(n+m)}.\end{array} \tag{2.148}$$

durch

$$\delta_2^{\mathbb{R}}(A,B) = \min_{\zeta\in\mathbb{C}} \max_{\gamma\in(0,1]} \sigma_{2n-1}\left(\begin{bmatrix} \Re e\, W & -\gamma \Im m\, W \\ \gamma^{-1}\Im m\, W & \Re e\, W \end{bmatrix}\right) \tag{2.149}$$

mit $W = [\zeta I_n - A, B]$ gegeben, wobei die maximierende Funktion über γ quasikonkav (und damit unimodal) ist. Die normkleinsten Änderungen $[\triangle A, \triangle B]$ errechnen sich anhand der Blockmatrix in (2.149) mit γ_{opt} und ζ_{opt} und sind höchstens vom Rang 2.

Das nachfolgende Beispiel zeigt, dass der Ingenieur den reellen Steuerbarkeitsradius bevorzugen sollte. Der komplexe Steuerbarkeitsradius fällt nämlich unter Umständen sehr klein aus (konservatives Maß), was schlechte Steuerbarkeit anzeigt, obwohl die erforderlichen reellen (damit ingenieurrelevanten) Parameteränderung bis zur Nichtsteuerbarkeit keineswegs so klein sind. Hieraus darf allerdings nicht gefolgert werden, dass dann auch der komplexe Stabilitätsradius für den Ingenieur weniger brauchbar ist, s. Anmerkung 2.16.

Beispiel 2.25 (Reeller und komplexer Steuerbarkeitsradius, [338])
Für

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & -\gamma^2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u \qquad \gamma \geq 1.$$

erfüllt der komplexe Steuerbarkeitsradius $\delta_2^{\mathbb{C}}(A_\gamma, B) < 1/\gamma$ wegen $\triangle A = 0_{2\times 2}$ und $\triangle b = (0, 1/(j\gamma))^T$, während für den reellen Steuerbarkeitsradius $\delta_2^{\mathbb{R}}(A_\varepsilon, B) = 1$ mit $\triangle A_{\text{opt}} = \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix}$ und $\triangle b_{\text{opt}} = 0_2$ gilt. Für große γ wird also $\delta_2^{\mathbb{C}}$ immer kleiner, während $\delta_2^{\mathbb{R}}$ konstant bleibt. Obere und untere Schranken für $\delta_2^{\mathbb{R}}(A, B)$ werden in [140] angegeben, wobei $\delta_2^{\mathbb{C}}(A, B) \leq \delta_2^{\mathbb{R}}(A, B)$ trivial ist.

Eine Erweiterung der angesprochenen Probleme auf Deskriptorsysteme und auf Polynommatrixsysteme wie $M\ddot{x} + D\dot{x} + Kx = Bu$ ist möglich [598], [452]. Des Weiteren können simultane Maße für Beobachtbarkeit und Steuerbarkeit durch Blockmatrixdarstellungen konstruiert werden [140]. Erweiterungen hinsichtlich strukturierter Störungen, bei denen beispielsweise die Symmetrie erhalten bleiben soll, oder hinsichtlich separabler Störungen, bei denen nur A bzw. nur B gestört werden, finden sich in [338].

Anmerkung 2.51 Bei der quantitativen Untersuchung der Steuerbarkeit ist zu beachten, dass das Ergebnis von den gewählten Zustandskoordinaten abhängt. Diese sollten systemangepasst gewählt werden, was aber bei der sog. Blackbox-Identifikation zumeist nicht der

Fall ist. Als Alternative bieten sich Maße basierend auf der Polkürzbarkeit an, vgl. Abschn. 2.2.9 und [435]. Zu jeder polkürzbaren Übertragungsfunktion lässt sich nämlich sowohl ein nichtsteuerbares als auch ein nichtbeobachtbares System angeben, was einerseits über die gemeinsame Übertragungsfunktion dualer Systeme, aber andererseits auch über die Reihenfolge des Aufeinandertreffens von Pol- und Nullstelle (Polstelle vor gleicher Nullstelle nicht beobachtbar wegen der Blockierungseigenschaft; Nullstelle vor gleicher Polstelle nicht steuerbar) erklärt werden kann.

2.12.3 Steuerbarkeit für andere Systemklassen

Für kontinuierliche LTI-Systeme ist es gleichbedeutend, die Steuerbarkeit zwischen zwei beliebigen Punkten, zum Nullpunkt hin oder vom Nullpunkt weg zu definieren. Anders ist die Situation bei zeitdiskreten LTI-Systemen, bei LTV-Systemen und bei nichtlinearen Systemen. Im Folgenden wird auf einige Aspekte bezüglich der Steuerbarkeit für diese Systemklassen hingewiesen. Für den Ingenieur ist es dabei im Wesentlichen wichtig zu wissen, worauf er achten muss. Für formelmäßige Details und Kriterien ist ein vertiefendes Studium der Zusammenhänge unerlässlich.

Steuerbarkeit zeitdiskreter LTI-Systeme
Ein n-dimensionales zeitdiskretes LTI-System heißt vollständig steuerbar, besser nullpunktsteuerbar, wenn es zu jedem Anfangszustand n Eingangswerte $u[0], \ldots, u[n-1]$ gibt, durch die dieser Zustand in den Nullzustand überführt werden kann. Es heißt vollständig erreichbar (steuerbar aus dem Ursprung), wenn jeder beliebige Endzustand aus dem Nullzustand durch eine Folge von n Eingangswerten erreichbar ist.

Zudem ist bekannt, dass für zeitdiskrete LTI-Systeme die vollständige Erreichbarkeit die vollständige Steuerbarkeit impliziert und dass die Umkehrung nur im Fall regulärer Systemmatrizen gilt [543].

Steuerbarkeit von LTV-Systemen
Die Steuerbarkeit von LTV-Systemen interessiert bei der Identifikation primär aus struktureller Sicht, denn wie bei LTI-Systemen wird fast sicher ein steuerbares Modell geschätzt, da die Menge der nichtsteuerbaren LTV-Modelle mager ist. Die Situation ändert sich, wenn das LTV-System durch eine Linearisierung eines nichtlinearen Systems entlang einer nichtkonstanten Lösung hervorgeht. Das die Lösung generierende Signal (Testsignal, Reglersollsignal, Reglerstellsignal) sollte dann so gewählt werden, dass es nichtsteuerbare Gebiete meidet oder diese nur in einzelnen isolierten Punkten kreuzt. Zudem empfiehlt sich bei aktiver Experimentation ein genügend großer Abstand zu nichtsteuerbaren Gebieten, um eine hinreichende Systemanregung zu gewährleisten. Bisher sind diese Fragen aus Identifikationssicht kaum behandelt wurden.

Einen Einstieg in die zugehörige Systemtheorie gibt [543]. Ein einfaches und leistungsstarkes Kriterium für Steuerbarkeit ist die Verallgemeinerung des Kalman-Kriteriums [577]. Wie im Fall der Stabilität darf nämlich nicht aus der punktweisen Steuerbarkeit für jedes $t \geq t_0$ auf Steuerbarkeit geschlossen werden, vgl. [543]. Im Übrigen hängt die Steuerbarkeit bei LTV-Systemen vom betrachteten Intervall ab, d. h. sie kann erscheinen und verschwinden, vgl. $\dot{x} = \max\{0, \sin t\}u$.

Steuerbarkeit nichtlinearer Systeme
Für nichtlineare Systeme ist bereits der Nachweis struktureller Steuer- bzw. Beobachtbarkeit schwierig. Erschwerend kommen zahlreiche begriffliche Ausprägungen (asymptotisch, lokal, ...) hinzu [388], [548]. Auch ist zwischen der Steuerbarkeit zu einem Punkt und von einem Punkt weg zu unterscheiden. So ist der Nullpunkt des Systems $\dot{x} = xu$ von jedem Punkt aus anzusteuern, aber kein anderer Punkt vom Nullpunkt zu erreichen.

Auf keinen Fall darf aus der Nichtsteuerbarkeit des linearisierten Modells auf die Nichtsteuerbarkeit des nichtlinearen Modells geschlossen werden. Als Klassiker gilt das Beispiel „Fahrzeugvorderachse“ [479], deren vollständige Steuerbarkeit kein Autofahrer in Frage stellt. Zudem darf für pseudolinearen Systeme $\dot{x} = A(x)x + b(x)u$ nicht aus der Steuerbarkeit für alle Paare $(A(x), b(x))$ auf globale Steuerbarkeit geschlossen werden. So liegt für

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} x_2 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u, \tag{2.150}$$

trotz regulärer Steuerbarkeitsmatrix $K_{(A,b)} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ für alle Punkte x auf der Menge $\overline{\mathcal{S}} = \{x_1 = -1, x_2 \in \mathbb{R}\}$ keine Steuerbarkeit vor. Analog zu LTV-Systemen ist also eine punktweise Betrachtung unzulässig. Vielmehr ist die nichtlineare Steuerbarkeitsrangbedingung zu verwenden [548].

2.12.4 Beobachtbarkeit

Die Ausführungen zur Steuerbarkeit lassen sich unter Ausnutzung der Dualität direkt auf die Beobachtbarkeit übertragen.

Definition 2.12 (Duale Systeme)
Das linke System in (2.151) heißt primales System, und das rechte wird (algebraisch) duales System genannt. Analoges gilt für zeitdiskrete LTI-Systeme.

$$\dot{x}_p = Ax_p + Bu \qquad \text{und} \qquad \dot{x}_d = A^T x_d + C^T u \tag{2.151a}$$

$$y_p = Cx_p + Du \qquad\qquad y_d = B^T x_d + D^T u. \tag{2.151b}$$

Der nachfolgende Satz stellt den Zusammenhang zwischen Systemeigenschaften des primalen und dualen Systems her. Dabei folgen die Äquivalenzen direkt aus den zugehörigen Kriterien unter Ausnutzung der Rangbeziehung $\text{Rang}X = \text{Rang}X^T$ bzw. der Eigenwertbeziehung $\lambda_i(A) = \lambda_i(A^T)$. Der Nutzen des Satzes liegt darin, dass numerische Algorithmen, die etwa für den Steuerbarkeitsradius erstellt wurden, durch Transponieren der entsprechenden Matrizen auch zur Berechnung des Beobachtbarkeitsradius angewandt werden können. Desgleichen gilt für andere Aspekte, wie die nachfolgenden Anmerkungen zeigen.

Satz 2.16 (Dualität qualitativer Eigenschaften)
Das primale LTI-System hat genau dann die linke Eigenschaft, wenn das duale LTI-System die rechte Eigenschaft aufweist [328], [543].

Erreichbarkeit	$\Leftrightarrow$	Beobachtbarkeit	(2.152a)
Steuerbarkeit	$\Leftrightarrow$	Rekonstruierbarkeit	(2.152b)
Stabilisierbarkeit	$\Leftrightarrow$	Detektierbarkeit	(2.152c)
Beobachtbarkeit	$\Leftrightarrow$	Erreichbarkeit	(2.152d)
Rekonstruierbarkeit	$\Leftrightarrow$	Steuerbarkeit	(2.152e)
Detektierbarkeit	$\Leftrightarrow$	Stabilisierbarkeit	(2.152f)
Stabilität	$\Leftrightarrow$	Stabilität	(2.152g)

Für kontinuierliche LTI-Systeme sind Erreichbarkeit und Steuerbarkeit äquivalent, ebenso wie Beobachtbarkeit und Rekonstruierbarkeit. Für zeitdiskrete LTI-System impliziert die Erreichbarkeit („Steuerung von der Null weg“) die Steuerbarkeit („Steuerung zur Null hin“). Ebenso impliziert die Beobachtbarkeit die Rekonstruierbarkeit.

Anmerkung 2.52 Wie bei der Steuerbarkeit gibt es im Nichtlinearen viele begriffliche Ausprägungen für die Beobachtbarkeit [573], [388]. Letztlich geht es darum, ob zwei Anfangszustände durch Wahl eines geeigneten Eingangssignals anhand des Ausgangssignals unterschieden werden können. In der Anwendung wird diese Frage oft nur für benachbarte Zustände, also lokal, gestellt. Aus Sicht der Identifikation und der Beobachtertheorie wird häufig mehr gefordert: Unterscheidbarkeit muss für jedes oder fast jedes Eingangssignal vorliegen.

Anmerkung 2.53 Wie die Steuerbarkeit ist die auch Beobachtbarkeit eine generische Systemeigenschaft. Für glatte nichtlineare Systeme ist Beobachtbarkeit ebenfalls generisch [6].

Ergänzend zu den im Abschnitt 2.12.2 erwähnten Anwendungen der Maße sei hier die Distanz zum nächstgelegenen System mit einem nicht-beobachtbaren reinen Schwingungsmode genannt, das beim Beobachterentwurf für Lipschitz-nichtlineare Systeme genutzt wird [2]

$$d(A, C) = \min_{\omega \in \mathbb{R}} \sigma_{\min}\Big(\begin{bmatrix} j\omega I_n - A \\ C \end{bmatrix}\Big). \tag{2.153}$$

2.13 Minimalphasigkeit

2.13.1 Minimalphasigkeit für lineare Eingrößensysteme

In der Nachrichtentechnik, Akustik und Elektrotechnik ist es die minimale Phasenverzögerungseigenschaft, die einer Systemklasse ihren Namen gibt. Mit der Minimalphaseneigenschaft geht auch eine minimale Energieverzögerung unter allen LTI-Systemen gleichen Frequenzgangs einher.

Definition 2.13 (Minimalphasigkeit i. S. der Nachrichtentechnik)
Ein kausales, minimales Hurwitz-stabiles System $G(s)$ mit positiver statischer Verstärkung heißt minimalphasig, wenn es von allen Systemen mit dem gleichen Betragsfrequenzgang für alle ω die kleinste Phasenlaufzeit[88] (Phasenverzögerung)

$$t_{ph} = -\frac{\arg G(j\omega)}{\omega} = -\frac{1}{\omega}\arctan\frac{\Im m\, G(j\omega)}{\Re e\, G(j\omega)} \tag{2.154}$$

oder äquivalent die kleinste Phase hat. Systeme mit negativer statischer Verstärkung heißen minimalphasig, wenn $-G(s)$ minimalphasig ist. Ein kausales Hurwitz-stabiles System, das die Minimalphaseneigenschaft nicht hat, heißt nichtminimalphasig.

Anmerkung 2.54 Die Einschränkung auf Hurwitz-stabile Systeme ist der erforderlichen Existenz der Phasenlaufzeit (praktische Ermittelbarkeit des Frequenzgangs) geschuldet.[89] Zudem folgt, dass die Nullstellen einen nichtpositiven Realteil haben müssen, um eine minimale Phasenlaufzeit aufzuweisen. Der zusätzliche Ausschluss von Nullstellen auf der imaginären Achse (keine Blockierung bestimmter Sinusschwingungen), führt auf den Begriff der minimalphasigen Filter. Sie sind demnach dadurch gekennzeichnet, dass $G(s)$ und $G^{-1}(s)$ Hurwitz-stabil und totzeitfrei sind.

Die Minimalphaseneigenschaft garantiert, dass aus der Phasenkenntnis der Betragsfrequenzgang (Amplituden(verstärkungs)gang) und damit die Übertragungsfunktion bis auf einen Faktor folgt und dass dem Betragsfrequenzgang genau ein Phasenfrequenzgang zugeordnet werden kann [607], wobei die Hilbert-Transformation die Verbindung herstellt

$$\arg G(j\omega)|_{\min} = -\mathfrak{H}\{\ln|G(j\omega)|\} \tag{2.155a}$$

$$\ln|G(j\omega)| = \ln|G(j\infty)| + \mathfrak{H}\{\arg G(j\omega)\}. \tag{2.155b}$$

[88] Die Phasenlaufzeit ist die Zeit, die eine einzelne kosinusförmige Schwingung beim Durchgang durch ein Übertragungssystem benötigt, d. h. die Zeit, um die ein Nulldurchgang verschoben wird. Sie ist anschaulicher als die Phase selbst.

[89] Einfache konjugierte Polpaare auf der imaginären Achse liefern stationär zwei Schwingungen; Mehrfachpole auf der imaginären Achse haben instationäre Lösungen.

Die Beziehung ist in äquivalenter Darstellung auch als Bode's Gain-Phase-Relation [33] bekannt. Da der umkehrbare Zusammenhang gültig bleibt, wenn Pole- und/oder Nullstellen auf der imaginären Achse liegen – ja sogar bei einem Nullstellenüberschuss – definierte Bode die Minimalphasigkeit für grenzstabile und totzeitfreie $G(s)$ und $G^{-1}(s)$, s. [33], [197].

In der Regelungstechnik ist es nicht die minimale Phasenverzögerung, die die Systeme hervorhebt, sondern die Eigenschaft, wonach alle Wurzeln des Zählerpolynoms in der linken offenen komplexen Halbebene liegen. Durch diese vom Frequenzgang losgelöste Betrachtung entfallen Forderungen an die statische Verstärkung und an die Stabilität. Ferner gelingt die Erweiterung auf MIMO-Systeme und nichtlineare Systeme. Allerdings entsteht damit zusätzlich zur Inkonsistenz in den Definitionen infolge der Hinzunahme oder des Weglassens der imaginären Achse eine weitere Inkonsistenz, da nunmehr die Pole unbetrachtet bleiben. Zeitz [644][90] plädiert deshalb dafür, in der regelungstechnischen Lehre und Forschung den Minimalphasenbegriff zu meiden und stattdessen den wohldefinierten Begriff der Hurwitz-Stabilität des Zählerpolynoms oder der asymptotischen Stabilität der Nulldynamik zu verwenden. Alternativ dazu wird in dieser Arbeit das Inkonsistenzdilemma durch Nachsätze wie „i.S. der Nachrichtentechnik„, „nach (Name)“ oder „für (Systembeschreibung)“ aufgelöst.

Die bisherigen Betrachtungen beziehen sich auf Minimalphasensysteme. Gute Modelle für ihr Gegenstück – die Nichtminimalphasensysteme – sind aus regelungstechnischer Sicht aber viel bedeutsamer, da der Reglerentwurf besonders bei Nullstellen mit einem kleinen positiven Realteil erheblich schwieriger ist. Das Vorwissen über ein Nichtminimalphasenverhalten kann aus der theoretischen oder experimentellen Analyse stammen. Es wird durch Restriktionen umgesetzt, die diejenigen für Minimalphasensysteme negieren. Die theoretische Analyse nennt als Quellen für Nichtminimalphasigkeit

- parallele Prozesse mit entgegengesetzter Wirkung und unterschiedlicher Dynamik
  Beispiel: Zusätzliche nasse Kohle senkt zunächst die Temperatur eines Ofens, bevor sie zur Temperaturerhöhung beiträgt.
  Erklärung: Die Parallelität bedingt in der Modellierung eine Addition zweier Übertragungsfunktionen und das Ausführen einer Hauptnennerbildung erzeugt dann (vorzeichenbedingt) wenigstens eine Nichtminimalphasennullstelle.

- unteraktuierte Mehrkörpersysteme [469]
  Beispiel: inverses Pendel auf Wagen, rückwärtsfahrendes Auto

- nichtkollokierte (nicht am gleichen Ort) Aktor-Sensoranordnungen flexibler Strukturen
  Erklärung: Komplexes Nullstellenpaar teilt sich im Unendlichen und erscheint als reelle Nullstellen, die sich von $\pm\infty$ in Richtung Null bewegen.

[90] Die Arbeit erschien, während die vorliegende Arbeit begutachtet wurde. Sie wurde zusätzlich zu den Gutachterhinweisen aufgenommen und führte zu marginalen Änderungen der ursprünglichen Darstellung.

- sich nicht neutralisierende Nichtlinearitäten
  Beispiel Pumpspeicherwerk: Vergrößern der Durchflussfläche verringert zunächst die Leistung, bevor sie steigt.[91]

- Allpässe und Systeme mit Allpassfaktor
  Ein Allpass ist ein System mit der gleichen Anzahl von symmetrisch zur imaginären Achse angeordneten Pol- und Nullstellen, sodass $|G_A(j\omega)| = K$ für alle Frequenzen ω ist. Ein System mit Allpassfaktor hat die Struktur $G(s) = G_{\min}(s)G_A(s)$ und entsteht durch Faktorisierung in ein Minimalphasensystem und einen Allpass. In der Filtertheorie ermöglicht die Allpassfaktordarstellung einen unabhängigen Entwurf von Betrags- und Phasenfrequenzgang.

- Medikation
  Erklärung: Einer Verbesserung des Gesundheitszustands folgt eine Verschlechterung durch Nebenwirkungen bei zu langer Medikamenteneinnahme. Das ist ein Beispiel, in dem die Systemantwort in die „falsche“ Richtung etwas Gutes ist.

- Regelkreise mit instabiler Strecke und/oder instabilem Regler
  Erklärung: Pole erscheinen als Nullstellen in Eingangs- und Ausgangsstörübertragungsfunktion; instabile Regler sind bei Verletztheit der Parity-Interlacing-Property unvermeidbar [615].

- Padé-Approximation (Totzeit wird durch Allpass genähert)
  Beispiel: Neben der bewusst ausgeführten Approximation können bei der Identifikation ignorierte Totzeiten bzw. Totzeitanteile instabile Nullstellen hervorrufen.

- parameter- oder zeitinduzierte Systeme
  In Analogie zur Instabilität von LTV-Systemen mit punktweiser Hurwitz-Stabilität können zeitliche Parameteränderungen oder durch in der Modellierung unberücksichtigte andere Parameter ein Nichtminimalphasenverhalten erzeugen. Ein Beispiel ist in [563] angegeben. Zur Definition und Theorie der Nichtminimalphasigkeit für LTV-Systeme sei auf [66] verwiesen.

Im Rahmen der experimentellen Analyse lassen sich Nichtminimalphasensysteme mit einer ungeraden Anzahl von Nullstellen auf der rechten Halbebene an der Sprungantwort erkennen. Sie starten nämlich in die dem stationären Endwert entgegengesetzte Richtung (engl.: wrong way behavior), was unmittelbar aus dem Anfangs- und Endwerttheorem der

[91] Wegen Av = konst reduziert sich die Geschwindigkeit, bevor sie auf den alten Wert $v = \sqrt{2gh}$ steigt. Da die Leistung proportional zu Av^3 ist, führt das Absinken von v zu einer größeren Leistungsreduktion als der Flächenzuwachs zu einer Erhöhung. Mit dem Ansteigen von v stellt sich im Verlauf natürlich eine höhere Leistung ein.

Laplace-Transformation folgt (experimenteller Hinweis auf Nichtminimalphasigkeit). Bei einer geraden Anzahl von Nullstellen kommt es zwar zu einem Anfangsverlauf in Richtung des stationären Endwerts, um alsbald meistens eine Richtungsänderung vorzunehmen, die der bei ungerader Anzahl entspricht. Die Stärke dieses Effekts hängt entscheidend von der Pol-Nullstellenverteilung ab. Weitere Details zum Unter- und Überschießen sowie zu lokalen Extrema und Nullstellen der Sprungantwort werden in [293] gegeben, während in [551] darauf verwiesen wird, dass die Ergebnisse nicht formal auf MIMO-Systeme übertragbar sind.

2.13.2 Minimalphasigkeit für lineare Mehrgrößensysteme

Die Erweiterung der Minimalphasigkeit auf MIMO-Systeme geschieht nicht wie naheliegend über die Minimalphasigkeit der Teilübertragungsfunktionen, sondern über die Übertragungsnullstellen. Da die zugehörigen Begriffe nicht einheitlich gebraucht werden und da viele Definitionen zueinander nicht äquivalent sind, wird nachfolgend auf ausgewählte Zusammenhänge näher eingegangen. So kann die für die Anwendung passende Definition und abgeleitete Restriktion gewählt werden. Einen Überblick zu Nullstellendefinitionen und deren numerische Berechnungsmöglichkeiten geben Schrader und Sain [552].

Definition 2.14 (Smith-McMillan-Form)
Jede propere[92] gebrochen rationale Übertragungsmatrix $G(s) \in \mathbb{R}(s)^{p\times m}$ ist zu ihrer Smith-McMillan-Form[93]

$$U^{-1}(s)G(s)V^{-1}(s) = \mathrm{diag}_{p\times m}\left(\frac{\varepsilon_1(s)}{\psi_1(s)}, \ldots, \frac{\varepsilon_r(s)}{\psi_r(s)}, 0, \ldots, 0\right) =: G^{\mathrm{SM}}(s) \tag{2.156}$$

mit unimodularen $U(s) \in \mathbb{R}(s)^{p\times p}, V(s) \in \mathbb{R}(s)^{m\times m}$ äquivalent, wobei r den Normalrang[94] bezeichnet und $\psi_i(.), \varepsilon_i(.) \in \mathbb{R}(s)$ monische und koprime Polynome sind, die die Teilerbedingungen $\varepsilon_i(.)|\varepsilon_{i+1}(.)$ und $\psi_{i+1}(.)|\psi_i(.)$ für $i = 1, \ldots, r$ erfüllen.

Definition 2.15 (Übertragungsnullstellen/-pole)
Die Wurzeln der Polynome $\varepsilon(s) = \prod_{i=1}^{r} \varepsilon_i(s)$ bzw. $\psi(s) = \prod_{i=1}^{r} \psi_i(s)$ heißen endliche Übertragungsnullstellen (ÜN) bzw. Übertragungspole (ÜP) von $G(s)$. Die Übertragungsnullstellen bzw. -pole im Unendlichen ergeben sich entsprechend der Anzahl der ÜN bzw. ÜP von

[92] Eine Übertragungsmatrix heißt proper, wenn alle Teilübertragungsfunktionen proper sind, also $\forall i, j : \lim_{s\to\infty} |G_{ij}(s)| = M_{ij} < \infty$ gilt.

[93] Die angegebene Standardvariante der Smith-McMillan-Form beschreibt die endliche Pol-Nullstellen-Struktur korrekt, kann aber die Pol-Nullstellen-Struktur im Unendlichen zerstören [610].

[94] Der Normalrang einer gebrochen rationalen Matrix $G(s)$ ist der maximale Rang, der sich für feste Werte der komplexen Variablen s einstellt.

$G(1/q)$ bei $q = 0$. Für propere $G(s)$ wird der Grad von $\psi(s)$ McMillan-Grad δ_M genannt, und er gibt die Dimension für eine Minimalrealisierung $\Sigma = (A, B, C, D)$ von $G(s)$ an.[95]

Definition 2.16 (Minimalphasigkeit für Übertragungsmatrizen)
Ein gebrochen rationales $G(s)$ heißt minimalphasig (i. S. der Regelungstechnik), wenn alle endlichen ÜN in der offenen linken Halbebene $\mathbb{C}_-$ liegen. Für linksinvertierbare Systeme[96] gilt äquivalent: $G(s)$ ist minimalphasig, wenn seine Linksinverse Hurwitz-stabil ist.

Eigenschaften der Übertragungsnullstellen:

1. Blockierungseigenschaft: Übertragungsnullstellen s_i besitzen die Eigenschaft, die Übertragung von $u = u_0 \mathrm{e}^{s_i t}; t \geq 0$ zu blockieren, d. h. bei geeignetem Anfangswert bleibt der Ausgang trotz Eingangssignal Null. Da bei der Übertragungsfunktion der Anfangswert per se Null ist, muss er durch Aufnahme eines Deltafunktionsanteils zum Eingangssignal u geeignet einstellt werden [15]. Bei der Identifikation sollten Signalanregungen mit $\mathrm{e}^{s_i t}$ deshalb unterbleiben.
   Definitionen der ÜN über die Blockierungseigenschaft schließen Nullstellen im Unendlichen ein, da unendliche Frequenzen von echt properen Systemen nicht übertragen werden. Problematisch an Definitionen über die Blockierung sind die Vielfachheiten der Nullstellen und die Einschränkung auf linksinvertierbare Systeme, denn nicht linksinvertiere Systeme können alle Frequenzen bei geeignetem Eingangssignal unterdrücken.

2. Die Übertragungsnullstellen sind die s_i, für die $\operatorname{Rang} G(s_i) < \operatorname{Normalrang} G(s)$ gilt, wobei s_i nicht gleichzeitig ein Pol sein darf, da sonst der Rang von $G(s_i)$ nicht definiert ist, s. Beispiel in Punkt 3 mit $s_1 = -1$.

3. Zwischen den Zählernullstellen der $G_{ij}(s)$ und den ÜN von $G(s)$, die direkt aus der zugehörigen Smith-McMillan-Form ablesbar sind, besteht kein Zusammenhang

$$G(s) = \begin{bmatrix} \frac{1}{s+1} & \frac{1}{(s+1)(s+2)} \\ \frac{s}{(s+1)(s+2)} & \frac{2s+1}{(s+1)(s+2)} \end{bmatrix} \qquad G^{\mathrm{SM}}(s) = \begin{bmatrix} \frac{1}{(s+1)(s+2)} & 0 \\ 0 & \frac{s+1}{s+2} \end{bmatrix};$$

   links $s_1 = 0, s_2 = -1/2$, rechts $s_1 = -1$.
   Somit kann die Kenntnis der Minimalphaseneigenschaft der Teilübertragungsfunktionen nicht über Bedingungen basierend auf (2.157) bei der Identifikation umgesetzt werden. Stattdessen ist das P-kanonische Modell zu wählen. Bei aktiver Experimentation lässt sich dann sogar die Identifikation auf $m \cdot p$ SISO-Identifikationen reduzieren.

[95] Für nicht-propere Systeme sei auf [328] verwiesen, wo auch die wichtigsten Eigenschaften zum McMillan-Grad zusammengestellt sind.

[96] Ist $\Sigma : \mathcal{U} \to \mathcal{Y}$ gegeben, so heißt $\Sigma^l : \mathcal{Y} \to \mathcal{U}$ Linksinverse, wenn $\Sigma^l(\Sigma(u(t)) = u(t)$ fast überall gilt.

4. Im Gegensatz zu SISO-Systemen können Übertragungsnullstellen dieselben Werte wie die Übertragungspole annehmen, ohne dass eine Kürzung oder ein Steuer- bzw. Beobachtbarkeitsverlust auftritt, vgl. $G(s) = \mathrm{diag}(\frac{s+2}{(s+1)^2}, \frac{s+1}{(s+2)^2})$ und vgl. $G(s) = \left(\begin{smallmatrix} 1 & s \\ 0 & 1 \end{smallmatrix}\right)$ mit je einem ÜP und einer ÜN im Unendlichen $G^{SM}(s) = \mathrm{diag}(1/s, s)$.
   Ferner gilt nicht wie bei SISO-Systemen, dass die totale Anzahl der Pole und der Nullstellen (endliche und unendliche) gleich ist. So hat $G(s) = (I_m, G_2^T(s))^T$ die ÜP von $G_2(s)$, aber keine ÜN, da I_m sowohl bei $G(s)$ und auch $G(1/q)$ für vollen Rang sorgt.

5. Übertragungsnullstellen und damit auch Minimalphasigkeit sind invariant unter regulären Eingangs-, Ausgangs- und Zustandstransformationen sowie unter Ausgangsrückführungen. Die Anzahl der Übertragungsnullstellen kann durch Zustandsrückführung geändert werden, indem etwa ein beobachtbarer Eigenwert auf eine Übertragungsnullstelle gelegt wird und diese damit zu einer Ausgangsentkopplungsnullstelle wird.
   Die Invarianz der Nullstellen gegenüber Ausgangsrückführungen hat zur Folge, dass der Regelkreis ab einer bestimmten Kreisverstärkung instabil werden muss (Pole wandern auf der Wurzelortskurve zu den Nullstellen).
   Im Gegensatz zu Übertragungsnullstellen lassen sich meist einige der Teilübertragungsnullstellen durch Ausgangsrückführung beeinflussen, was aus der Wirkung der anderen Ausgänge resultiert. Auf diese Weise lässt sich unter Umständen eine Nichtminimalphasigkeit einer Teilübertragungsfunktion beseitigen.

6. Für steuer- und beobachtbare Systeme stimmen die Übertragungsnullstellen mit den Eigenwerten der Nulldynamik überein.

Außer bei Systemen mit wenigen, einfachen Teilübertragungsfunktionen eignet sich der Zugang über die Smith-McMillan-Form nicht zur Formulierung geeigneter Restriktionen. Wird indes eine Minimalrealisierung für $G(s)$ angesetzt, so kann wegen der Minimalität und dem vollen Normalrang bei quadratischen Systemen die Polynombedingung (2.158) herangezogen werden. Einen anderer Weg, Restriktionen für stabile MIMO-Minimalphasensysteme de facto zu umgehen, stellt eine geeignete Parametrisierung in der balancierten Normalform [450], [484] dar, der in [133] vorgeschlagen wird.

Um Minimalphasigkeit auch für Zustandsraumsysteme $\Sigma = (A, B, C, D)$ fassen zu können, werden weitere Definitionen benötigt.

Definition 2.17 (Entkopplungsnullstellen)
Die $s_i \in \mathbb{C}$ mit $\mathrm{Rang}[s_i I - A, B] < n$, heißen Eingangsentkopplungsnullstellen (EEN) oder nicht steuerbare Eigenwerte. Die $s_i \in \mathbb{C}$ mit $\mathrm{Rang}[s_i I - A^T, C^T] < n$ heißen Ausgangsentkopplungsnullstellen (AEN) oder nicht beobachtbare Eigenwerte. s_i, die beide Bedingungen erfüllen, heißen E/A-Entkopplungsnullstellen (EAEN). Die Elemente der Multimenge $\{\mathrm{EN}\} = \{\mathrm{EEN}, \mathrm{AEN}\} \backslash \{\mathrm{EAEN}\}$ heißen Entkopplungsnullstellen.

Definition 2.18 (Invariante Nullstellen)
Die Smith-Normalform der Rosenbrock-Matrix $R(s) = \begin{bmatrix} sI_n - A & -B \\ C & D \end{bmatrix}$ definiert die invarianten Polynome[97] $p_1(s), \ldots, p_r(s)$, wobei r der Normalrang von $R(s)$ ist. Die Wurzeln dieser Polynome heißen Invariante Nullstellen[98].
Äquivalent: Die Invarianten Nullstellen sind jene s_i mit

$$\text{Rang} \begin{bmatrix} s_i I_n - A & -B \\ C & D \end{bmatrix} < \text{Normalrang} \begin{bmatrix} sI_n - A & -B \\ C & D \end{bmatrix}. \tag{2.157}$$

Definition 2.19 (Systemnullstellen, [15])
Die Systemnullstellen (SN) sind die Wurzeln des größten gemeinsamen Teilers aller Nichtnullminoren der Ordnung $n+k$ von $R(s)$ mit $\text{Rang}R(s) = n+k$, wobei die Minoren jeweils die Untermatrix $sI_n - A$ enthalten müssen.
Äquivalent: Die Systemnullstellen sind die Vereinigung der Übertragungsnullstellen und der Entkopplungsnullstellen, kurz $\{\text{SN}\} = \{\text{ÜN}\} \cup \{\text{EN}\}$.

Beispiel 2.26 (Systemnullstelle, die keine Invariante Nullstelle ist)
Für $A = \text{diag}(1, -1, -3)$, $b = [0, -1, -1]^T$, $C = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 2 & 1 \end{bmatrix}$, $d = [0, 0]^T$ ist $s = 1$ eine EEN. Weiterhin ist der Normalrang von $R(s)$ gleich 4. Da $s = 1$ den Rang von $R(s)$ nicht verkleinert, kann $s = 1$ keine Wurzel eines invarianten Polynoms, also keine Invariante Nullstelle sein. Es gibt somit Entkopplungsnullstellen, die keine Invarianten Nullstellen sind, wohl aber Systemnullstellen, d. h. $\{\text{IN}\} \subseteq \{\text{SN}\}$.

Definition 2.20 (Minimalphasigkeit für LTI-Zustandsraumsysteme)
$\Sigma = (A, B, C, D)$ heißt minimalphasig (i. S. der Regelungstechnik), wenn keine Systemnullstelle in der abgeschlossenen rechten komplexen Halbebene $\text{cl}\mathbb{C}_+$ liegt, bzw. äquivalent, wenn
- Σ stabilisierbar ist, $\forall s \in \text{cl}\mathbb{C}_+ : \text{Rang}(sI_n - A, B) = n$, d. h. keine EEN in $\text{cl}\mathbb{C}_+$,
- Σ detektierbar ist, $\forall s \in \text{cl}\mathbb{C}_+ : \text{Rang}(sI_n - A^T, C^T) = n$, d. h. keine AEN in $\text{cl}\mathbb{C}_+$,
- keine ÜN in $\text{cl}\mathbb{C}_+$ liegt.[99]

[97] Sei $\Delta_i(s)$ der größte gemeinsame Teiler aller $(i \times i)$-Minoren einer Matrix $P(s) \in \mathbb{R}(s)^{m \times p}$, dann bezeichnet die Menge $\{\Delta_i(s)\}$ die Determinantenteiler von $P(s)$ und die Menge $\{p_i(s)\}$ mit $p_i(s) = \Delta_i(s)/\Delta_{i-1}(s)$, $p_0(s) \equiv 1$ die invarianten Polynome.

[98] Der Name Invariante Nullstellen bezieht sich auf deren Zusammenhang mit den invarianten Polynomen und nicht auf deren Invarianzen unter regulären Transformationen des Eingangs, Ausgangs und Zustands sowie unter Ausgangs- und Zustandsrückführungen. Er soll deshalb als Eigenname verstanden werden.

[99] Verschärfungen stellen die strenge Minimalphasigkeit (zusätzlich $\det(CB) \neq 0$, d.ḣ. relativer Grad 1) und die Hyperminimalphasigkeit (zusätzlich $CB \succeq 0_{m \times m}$, d. h. Positivität der Hochfrequenzverstärkung CB) dar [24].

Obwohl alle Bedingungen algebraisch sind, ist eine Formulierung als gut handhabbare Restriktion schwierig. Eine Ausnahme bilden die quadratischen Systeme mit vollem Normalrang, für die sich (2.157) auf

$$\det \begin{bmatrix} sI_n - A & -B \\ C & D \end{bmatrix} = 0 \qquad \Leftrightarrow \qquad \det(C\mathrm{adj}(sI_n - A)B + D) = 0 \tag{2.158}$$

zurückführen lässt, also auf eine bekannte Wurzelrestriktion für ein Polynom[100]. Für reguläres D folgt aus (2.158) mit der Schurschen Determinantenregel, dass die Nullstellen mit den Eigenwerten von $A - BD^{-1}C$ zusammenfallen. Dies lässt sich mit geringen Modifikationen sogar auf nichtquadratische Systeme auf die Eigenwerte von $A - BD^{+}C$ erweitern, wenn D vollen Rang hat [600]. Für Minimalrealisierungen folgt das aus der Links- bzw. Rechtsinverse $G^{+}(s) = D^{+} - D^{+}C(sI_n - A + BD^{+}C)^{-1}BD^{+}$.

Die Nichtkonvexität der Restriktion beschränkt ein direktes Einbeziehen der Restriktion (2.158) auf Parameterschätzprobleme mit niedriger Systemordnung und nur wenige freien Parameter in A, B, C, D. Anders ist die Situation, wenn (2.158) für Überwachungsaufgaben oder A-posteriori-Auswertungen verwendet wird, da die Auswertung der Beziehung keine Schwierigkeiten bereitet. So kann dann analog zum Stabilitäts- und Steuerbarkeitsradius auch der sog. Minimalphasenradius eingeführt werden [381]. Er misst den Abstand zu den Nichtminimalphasensystemen. Anhand der minimierenden Matrizenquadrupels $(\triangle A, \triangle B, \triangle C, \triangle D)$ können Rückschlüsse gezogen werden, ob eingangs- oder ausgangsseitige Parameteränderungen einen stärkeren Einfluss auf das Minimalphasenverhalten haben. Auf diesem Weg liefert die Identifikation Anhaltspunkte für ein Überdenken der Aktor- und Sensoranordnung. Werden Aktor- und Sensor dynamisch gesehen weit voneinander entfernt (Idee des flachen Ausgangs bzw. Eingangs), treten wenige (in der Regel stabile), bestenfalls keine Nullstellen auf. Bei flexiblen Strukturen werden dagegen Aktor und Sensor oft am gleichen Ort angeordnet, was ein Alternieren von Pol- und Nullstellen entlang der imaginären Achse bewirkt. Das bietet regelungstechnische Vorteile, da jede Nullstelle ihren benachbarten Pol anzieht. Falls eine solche gerätetechnische Anordnung nicht möglich ist, kann der Minimalphasenradius in Verbindung mit einer Modalanalyse helfen, einen anderen geeigneten Sensorort zu finden (Nulldurchgänge der Welle).

Aus (2.158) lassen sich auch Abschätzungen für die Anzahl q der Invarianten Nullstellen gewinnen [158]:

$$q \leq \begin{cases} n & D \neq 0_{p\times m} \\ n - \max\{m, p\} & D = 0_{p\times m}, \mathrm{Rang}\, B = m, \mathrm{rg}C = p \\ n - m - (m - \mathrm{Rang}(CB)) & D = 0_{m\times m}, \mathrm{Rang}\, B = m, \mathrm{rg}C = p, m = p \end{cases} \tag{2.159}$$

[100] Im SISO-Fall für $\Sigma = (A, b, c^T)$ ist $\det \begin{bmatrix} sI_n - A & b \\ c^T & 0 \end{bmatrix}$ gerade das Zählerpolynom.

sowie generisch für $\operatorname{Rang} B = m$ und $\operatorname{Rang} C = p$

$$q = \begin{cases} n & D \neq 0_{m\times m}, m = p \\ n-m & D = 0_{m\times m}, m = p \\ 0 & m \neq p. \end{cases} \tag{2.160}$$

Aus (2.159) folgt, dass quadratische Systeme ohne Durchgriffsanteil mit genauso vielen Zuständen wie Eingängen keine Nullstellen haben. Ebenso besitzen Systeme mit voller Zustandsinformation ($C = I_n$) keine Nullstellen, weshalb bei Reglern mit vollständiger Zustandsrückführung keine Nichtminimalphasenprobleme auftreten können. Aus (2.160) folgt, dass nichtquadratische Systeme generisch keine Nullstellen haben. Restriktionen an die Minimalphasigkeit sind in den genannten Fällen also nicht zweckdienlich.
Aus (2.160) folgt weiterhin, dass ein nichtsprungfähiges SISO-LTI-System aus einer Zustandsraumidentifikation generisch den Differenzgrad $r = 1$ hat. Das A-priori-Wissen über den Differenzgrad wird deshalb zweckmäßigerweise über einen zeitkontinuierlichen E/A-Modellansatz mit dem Zählergrad $n - r$ berücksichtigt.

Eine wichtige Anwendung erfahren die Minimalphasensysteme durch nachfolgenden Satz, der über den Weg der Passifizierung die Möglichkeit eröffnet, die Systeme in Verbindung mit nichtlinearen Stellgliedern zu stabilisieren (Hyperstabilitätszugang).

Satz 2.17 (Feedback-Äquivalenz zu passiven Systemen, [198])
Für quadratische $\Sigma = (A, B, C)$ und $\operatorname{Rang} B = m < n$ sind folgende Aussagen äquivalent:[101]
- Σ ist minimalphasig und $\operatorname{Rang}(CB) = m$, kurzum Σ ist streng minimalphasig.
- Σ kann per Ausgangsrückführung $u = Ky + Lv$ streng zustandspassiv gemacht werden.
- Σ kann per Zustandsrückführung $u = Mx + Lv$ streng zustandspassiv gemacht werden.

2.13.3 Minimalphasigkeit für lineare zeitdiskrete Systeme

Analoge Aussagen zur Minimalphasigkeit gelten für zeitdiskrete Systeme. Auch hier ist zwischen Definitionen, die sich auf die Phasenverzögerung beziehen, und jenen, die sich auf Nullstellen gründen, zu unterscheiden. Bei Letzteren ist für Minimalphasigkeit zu fordern, dass alle endlichen Nullstellen im Inneren des Einheitskreises liegen.

Auf keinen Fall darf bedenkenlos aus der Minimalphasigkeit eines kontinuierlichen Systems auf die des zugehörigen Abtastsystems geschlossen werden und umgekehrt, was eine Folgerung aus dem nächsten Satz ist.

[101] Der Satz ist eine direkte Konsequenz des Feedback-Kalman-Yakubovich-Lemmas [198] und ein Korollar eines Satzes für nichtlineare Systeme [113]. Unter der Annahme schwacher Minimalphasigkeit kann Passivität mit quadratischer Verlustfunktion $V(x) = x^T Q x$ durch Rückführung erreicht werden.

Satz 2.18 (Nichtminimalphasigkeit von Abtastsystemen, [32])
Für ein LTI-SISO-Systems mit Differenzgrad $r > 2$ ist das korrespondierende Abtastsystem mit einem Halteglied nullter Ordnung für hinreichend kleine T_A immer nichtminimalphasig.

Typische Abtastzeiten in der Identifikation und Regelung sind in diesem Zusammenhang hinreichend klein. Für die Identifikation zeitdiskreter Abtastsysteme mit einem Halteglied nullter Ordnung mit $r > 2$ stellt die Minimalphasigkeit also keine durch Restriktionen zu garantierende Eigenschaft dar.

Anmerkung 2.55 Für ein nichtlineares SISO-System mit relativem Grad $r > 2$ ist dessen Haltegliedd iskretisierung für hinreichend kleine T_A immer nichtminimalphasig, was aus Satz 2.18 sowie der Vertauschbarkeit von Linearisierung und Abtastung folgt.

Anders ist die Situation, wenn mit der Tustin-Transformation $s = \frac{2}{T_A}\frac{z-1}{z+1}$ diskretisiert wird. Da die Abbildung konform ist, wird jede Pol- und Nullstelle der linken komplexen Halbebene in den Einheitskreis abgebildet. Die Transformation und deren Umkehrung vererbt also die Minimalphaseneigenschaft. Es ist lediglich zu beachten, dass bei nichtsprungfähigen Systemen zusätzliche Zählerwurzeln bei $z_i = -1$ entstehen, die von den Systemnullstellen bei $s_i = \infty$ herrühren. Eine Konsequenz der Tustin-Transformation ist, dass dann z-Modelle unabhängig vom s-Modell immer sprungfähig zu wählen sind.
Bedingungen für Minimalphasigkeit lassen sich für die Zählerkoeffizienten analog zur Schur-Stabilität formulieren, wobei das um die bekannten Wurzeln bei $z = -1$ reduzierte Polynom zu betrachten ist. Das Zählerpolynom ist per Restriktion auf die bekannten Wurzeln bei $z = -1$ zu zwingen, vgl. Abschn. 2.2.2, denn gerade bei Mehrfachwurzeln wirken sich kleine Abweichungen in den Polynomkoeffizienten stark auf die Wurzeln aus.
Während Tustin-z-Modelle geeignet sind, die Übertragung von sinoidalen oder polygonzugähnlichen Signalen zu modellieren, erweisen sie sich für stückweise konstante Signale als ungeeignet, da sie gegenüber der Nullter-Ordnung-Halteglied-Diskretisierung eine Totzeit von etwa $-T_A/2$ (bedeutet nichtkausales Verhalten) bewirken. Vergleiche hierzu die Näherung

$$\frac{z+1}{2}\underbrace{\mathfrak{Z}\left\{\frac{1-\mathrm{e}^{-sT_A}}{s}\cdot G(s)\right\}}_{\text{Transf. mit Halteglied}} \approx \text{Tustin}\{G(s)\} = G(s)\big|_{s=\frac{2}{T_A}\frac{z-1}{z+1}}, \tag{2.161}$$

die für $G(s) = 1/s$ exakt ist, und den Faktor

$$\frac{z+1}{2} = \frac{\mathrm{e}^{sT_A}+1}{2} \approx \frac{1+sT_A+1}{2} = 1 + s\frac{T_A}{2} \approx \mathrm{e}^{sT_A/2}. \tag{2.162}$$

Um zu erreichen, dass sich das Tustin-z-Modell ähnlich wie das Halteglied-z-Modell verhält, ist dieses mit einer zusätzlichen Totzeit $\mathrm{e}^{-sT_A/2} = z^{-1/2}$ zu versehen. Dazu wird nicht

durch den Faktor $\frac{z+1}{2}$ dividiert, da dann in der Realisierung eine Eingangssignalfilterung mit unendlicher und nur grenzstabiler Gewichtsfolge entsteht, sondern es wird mit der Näherung

$$\mathfrak{Z}\left\{\frac{1-\mathrm{e}^{-sT_A}}{s}\cdot G(s)\right\} \approx \text{Tustin}\{G(s)\}z^{-1/2} \approx \text{Tustin}\{G(s)\}\frac{1}{2}(1+z^{-1}) \tag{2.163}$$

gearbeitet. Bei der Identifikation oder Simulation ist also lediglich die Eingangsfolge $\{u_k\}$ durch die Ersatzfolge $\{\tilde{u}_k\} = \{(u_k + u_{k-1})/2\}$ zu ersetzen.

Anmerkung 2.56 Einige Modelle wie z. B. die Autokorrelationsfunktion enthalten nicht die volle Phaseninformation des Signals. Sie sind somit für Minimalphasenrestriktionen nicht verwendbar. Anders ist es beim Spektrum $X^*(\mathrm{e}^{j\omega}) = \mathfrak{F}\{\{x[k]\}\}$, das die Fourier-Transformierte einer Folge $\{x[k]\}$ beschreibt, und beim komplexen Cepstrum[102] $c[d] = \mathfrak{F}^{-1}\{\ln X^*(\mathrm{e}^{j\omega})\}$.

2.13.4 Minimalphasigkeit für nichtlineare Systeme

Die klassischen Definitionen der Minimalphasigkeit sind nicht sinnvoll auf nichtlineare Systeme übertragbar, da der Frequenzgang nicht erklärt ist und die Laplace-Transformation wie auch Eigenwertkonzepte nicht anwendbar sind. Deshalb beziehen sich Erweiterungen auf

1. den durch Zustandsrückführung maximal unbeobachtbar zu machenden Systemteil,
2. die verbleibende Dynamik, wenn der Ausgang konstant auf Null gezwungen wird oder
3. die Dynamik einer Minimalrealisierung der Linksinverse.

Für LTI-Systeme führt die Stabilitätsforderung an die Dynamiken zu äquivalenten Definitionen der Minimalphasigkeit im Sinne der Regelungstechnik; für nichtlineare Systeme haben sich die Zugänge 2 und 3 etabliert.

Eine erste Definition für Minimalphasigkeit stützt sich auf den Begriff der Nulldynamik (Punkt 2), die im Minimalphasenfall, grob gesagt, stabil sein muss.

Definition 2.21 (Nulldynamik, [315], [198])
Die Nulldynamik beschreibt das Verhalten jener Lösungen $x(t)$, für die ein $u(t)$ und x_0 existiert, sodass $y(t) \equiv 0$ gilt. Sie ist also durch die Menge der Signale $\{(x(t), u(t), y(t)) : y(t) \equiv 0\}$ charakterisiert, wobei für $u(t)$ gemeinhin stückweise stetige Signale angenommen werden. [103]

[102] Das Cepstrum ist eine in der Nachrichtentechnik und Akustik häufig verwendete Darstellungsform, bei der Quelle und Filter nicht mehr über eine Faltung (Zeitbereich) oder ein Produkt (Frequenzbereich), sondern dank des Logarithmus über eine Summe miteinander verknüpft sind. Eine additive Verknüpfung liegt auch für Minimalphasen- und Allpassteil vor.

[103] Ist $\mathcal{U}$ eine Umgebung von $x = 0_n$, so heißt $\dot{x} = f^*(x); x \in \mathcal{M}^* = \{x \in \mathcal{U} : h(x) = 0_p\}$ lokale Nulldynamik mit dem Nulldynamikvektorfeld f^* und der Nulldynamikmannigfaltigkeit $\mathcal{M}^*$.

Im Nichtlinearen ist eine Aussage wie „muss stabil sein“ wegen der unterschiedlichen Stabilitätsbegriffe unzulänglich, zumal auch der für die Stabilität erforderliche Bezug auf Ruhelagen oder Mengen fehlt. Auch ist aus der Nulldynamikdefinition nicht klar, wie die Nulldynamik einer Stabilitätsbehandlung zugänglich gemacht werden kann. Für eingangsaffine Systeme

$$\dot{x} = f(x) + g(x)u;\ x(0) = x_0 \qquad x \in \mathbb{R}^n, u \in \mathbb{R}, y \in \mathbb{R} \tag{2.164a}$$

$$y = h(x) \qquad h \in \mathcal{C}^r, \tag{2.164b}$$

gelingt Letzteres über die Byrnes-Isidori-Form [315]

$$\begin{aligned}
\dot{\xi}_1 &= \xi_2 \\
&\cdots \\
\dot{\xi}_{r-1} &= \xi_r \\
\dot{\xi}_r &= b(\xi,\eta) + a(\xi,\eta)u \qquad \text{mit } \xi = [\xi_1,\ldots,\xi_r]^T \\
\dot{\eta} &= q(\xi,\eta) \qquad\qquad \eta = [\eta_1,\ldots,\eta_{n-r}]^T \\
y &= \xi_1 \qquad\qquad \text{Im Fall } r = n \text{ verschwindet } \eta.
\end{aligned} \tag{2.165}$$

Bezeichne r den relativen Grad (Ausgangsableitung $y^{(r)}$, in der der Eingang erstmalig auftaucht), dann existiert eine lokale Koordinatentransformation gemäß $\xi_k = y^{(k-1)} = \phi_k(x) = L_f^{k-1}h(x); k = 1,\ldots,r$ und mit $n-r$ weiteren Funktionen $\eta_k = \phi_k(x); k = r+1,\ldots,n$ mit $L_g\phi_k(x) = 0$ auf Byrnes-Isidori-Form[104]. Die lokale Koordinatentransformation kann dabei immer so ausgeführt werden, dass eine Ruhelage x_e des Systems (2.164) in eine Nullruhelage von (2.165) abgebildet wird.

Aus (2.165) folgt, dass $y \equiv 0$ zwingend $\xi(0) = 0_r$ und $u = -b(0_r,\eta)/a(0_r,\eta)$ erfordert (einen wohl-definierten relativen Grad, also $a(0_r,\eta_e) \neq 0$ für die Ruhelage η_e vorausgesetzt). $\dot{\eta} = q(\xi,\eta); \eta(0) = \eta_0$, also der durch u nicht direkt beeinflussbare Teil, heißt interne Dynamik von (2.165) und $\dot{\eta} = q(0_r,\eta); \eta(0) = \eta_0$ heißt Nulldynamik-Untersystem von (2.165).

Anmerkung 2.57 Während die Nulldynamik eindeutig bestimmt ist, existieren je nach Transformation für (2.165) unterschiedliche Nulldynamik-Untersysteme. Sei $\Sigma = (A,B,C,0)$ mit relativem Grad r gegeben, dann ist die Nulldynamik durch $y(t) \equiv 0_p$ und $u(t) = -(CA^{r-1}B)^{-1}CA^r x(t)$ sowie alle Lösungen von

$$\dot{x} = \underbrace{(I_n - B(CA^{r-1}B)^{-1}CA^{r-1})Ax}_{=f^*(x)}; \quad x_0 \in \mathcal{M}^* = \mathcal{N}\left(\begin{bmatrix} C \\ \vdots \\ CA^{r-1} \end{bmatrix}\right) \tag{2.166}$$

charakterisiert. Die Nulldynamik-Untersysteme nach (2.165) lauten

$$\dot{\eta} = V^+(I_n - B(CA^{r-1}B)^{-1}CA^{r-1})AV\eta; \qquad \eta_0 \in \mathbb{R}^{n-mr} \tag{2.167}$$

[104] In der Literatur häufig auch als Byrnes-Isidori-Normalform bezeichnet, obwohl sie keine Normalform im algebraischen Sinn ist, da sie keinen eindeutigen Repräsentanten für jede Äquivalenzklasse darstellt.

und hängen von der Wahl der Matrix $V \in \mathbb{R}^{n\times(n-mr)}$ mit $\mathcal{R}(V) = \mathcal{M}^*$ ab. Die Eigenwerte aller Systeme (2.167) sind unabhängig von V, entscheiden über die Stabilität der Nulldynamik und entsprechen – ergänzt um $m \cdot r$ Nulleigenwerte – denen von (2.166). Für die Transformation $x \mapsto \eta$, die Beweise und ergänzende Ausführungen zur allgemeineren Klasse der LTV-Systeme siehe [66].

Definition 2.22 (Minimalphasigkeit nach Byrnes/Isidori, [112], [315])
Das System (2.165) und damit auch (2.164) heißt lokal asymptotisch (exponentiell) minimalphasig in der Ruhelage η_e der Nulldynamik $\dot{\eta} = q(0_r, \eta)$, wenn diese lokal asymptotisch (exponentiell) stabil ist. Existiert eine globale Koordinatentransformation auf (2.165) und ist die Nulldynamik global asymptotisch stabil, dann heißt (2.164) global minimalphasig.[105]

Anmerkung 2.58 Die Definition 2.22 erweitert in ihrer MIMO-Variante [315] die regelungstechnisch motivierten Definitionen 2.16 und 2.20, da die Eigenwerte der Nulldynamik in der Byrnes-Isidori-Form mit den Nullstellen übereinstimmen. Sie ist aber wie Def. 2.16 und 2.20 nicht mit Def. 2.13 äquivalent. So wird der reine Differenzierer nicht abgedeckt und gegenüber Def. 2.13 werden instabile Systeme zugelassen.

Die Einschränkung der Definition 2.22 auf eingangsaffine Systeme und der Umweg über die Transformation lassen sich durch Beibehaltung der Nulldynamikidee umgehen.

Definition 2.23 (Minimalphasigkeit nach Ebenbauer/Allgöwer, [177])
Das System

$$\dot{x} = f(x,u);\ x(0) = x_0 \qquad x \in \mathbb{R}^n, u \in \mathbb{R}^m, y \in \mathbb{R}^p \tag{2.168a}$$

$$y = h(x) \tag{2.168b}$$

hat die Minimalphaseneigenschaft bezüglich der Ruhelage x_e, wenn x_e asymptotisch stabil unter der Restriktion $y(t) = 0$ für $t \geq 0$ ist.

Die Überprüfung der Minimalphaseneigenschaft erfolgt ähnlich wie bei Lyapunov's direkter Methode mittels einer differenzierbaren positiv definiten Funktion $V : \mathbb{R}^n \to \mathbb{R}$ über eine Dissipationsungleichung [177].

Das Problem an den bisher dargestellten Erweiterungen des Minimalphasenbegriffs stellt die asymptotische Stabilitätsforderung an die Nulldynamik dar, die ohne Zusatzforderungen für

[105] Das Konzept der globalen Nulldynamik i. S. von Byrnes/Isidori und damit der globalen Minimalphasigkeit ist insofern in Frage zu stellen, da nur selten für alle x stets der gleiche relative Grad existiert. Die Menge der eingangsaffinen Systeme mit dieser Eigenschaft ist nämlich mager. Ein weiteres Problem erwächst aus der Tatsache, dass der relative Grad für Systeme mit einem Ausgang generisch $r = 1$ ist [335]. Somit ist die Ordnung der Nulldynamik meist nur um Eins niedriger als die Systemdynamik.

regelungstechnische Anwendungen oft nicht ausreichend ist. Mit dem Konzept der Eingangs-zu-Zustands-Stabilität (input-to-state stability, ISS), s. hierzu Anhang A.6.2, lässt sich das beheben. So sichert die ISS-Eigenschaft unter anderem, anders als asymptotische Stabilität der Nulldynamik, dass, wenn die Zustände ξ nicht Null sind, aber nach Null streben, auch die Zustände η nach Null streben. Weiterhin impliziert sie Ausgangs-zu-Eingangs-und-Zustands-Stabilität nach Tab. A.6.2, was anschaulich bedeutet, dass mit kleiner werdendem Ausgang und dessen Ableitungen auch die Zustände und Eingänge kleiner werden (Idee der stabilen Linksinverse bei LTI-Systemen). Gerade diese Eigenschaft begründet die Bedeutung der LTI-Minimalphasigkeit in adaptiven Regelungen nach dem „Certainty-Equivalence-Prinzip“ [462], bei dem der Regler basierend auf der aktuellen Parameter-/Zustandsschätzung unter der Annahme „das Modell ist korrekt“ bestimmt wird. Ein ausgangsstabilisierender adaptiver Regler sorgt nämlich dann dafür, dass bei einem kleinen Schätzausgangsfehler, die Zustände und Schätzwerte klein bleiben.[106] Letztlich münden die angeführten Forderungen in der folgenden Definition.

Definition 2.24 (Starke Minimalphasigkeit nach Liberzon/Morse/Sontag, [401])
Das System (2.168) heißt stark minimalphasig, wenn es den relativen Grad[107] r hat und schwach gleichmäßig detektierbar[108] von der Ordnung $r-1$ ist.

Der Nachweis der starken Minimalphasigkeit erfolgt wiederum über eine Dissipationsungleichung (hinreichende Bedingung).

Das Problem für den Anwender dieser Definitionen besteht in ihrer Nichtäquivalenz zueinander und der Entscheidung für die der Anwendung angepasste Definition. Insofern empfiehlt sich ein Studium der Arbeiten [401] und [177], in denen sich zahlreiche Beispiele für zulässige und unzulässige Implikationen zwischen den drei Definitionen der Minimalphasigkeit, den beiden hinreichenden, aber unterschiedlichen Dissipationsungleichungen und der Ausgangs-zu-Eingangs-und-Zustands-Stabilität nach Tabelle A.6.2 finden. Aus Sicht der Identifikation bleibt festzuhalten, dass eine Detektion der Nichtminimalphasigkeit über die Sprungantwort in Analogie zum Linearen erfolgen kann [224], vgl. Schlussabsatz in Abschnitt 2.13.1. Ferner

[106] Dabei ist es nicht zwingend, dass die Parameter gegen die wahren Parameter streben, um ein befriedigendes Regelungsverhalten zu erzielen. Allerdings ist es bei der Adaption, etwa beim STC-Entwurf (self-tuning controller), vonnöten und/oder von Vorteil, die Minimalphasigkeit und häufig auch die Stabilität des Modells durch Projektion in die zulässigen Parametermengen oder durch spezielle Modellansätze sicherzustellen.

[107] Eine Erweiterung des relativen Grads auf nicht eingangsaffine Systeme wird in [401] gegeben.

[108] Das Schwach bezieht sich darauf, dass zur Detektierbarkeit nicht nur y, sondern auch dessen Ableitungen zugelassen sind. Das Gleichmäßig meint, dass die Detektierbarkeit gleichmäßig bezüglich u (unabhängig von u) ist. Die Ordnung gibt an, bis zu welcher Ableitung Ausgangssignale für die Detektierbarkeit verwendet werden.

liegt Minimalphasigkeit bei Systemen, deren Ausgang flach ist, ohnehin vor, da solche Systeme keine Nulldynamik haben. Passive Systeme mit positiv definiter Speicherfunktion sind per se schwach minimalphasig, da sich die Verlustfunktion entlang aller Trajektorien, die mit der Restriktion $y \equiv 0$ konsistent sind, verringert und somit die Nulldynamik Lyapunov-stabil ist. Durch Restriktionen hinsichtlich der Passivität werden Restriktionen bezüglich der Minimalphasigkeit also hinfällig. Für andere Systemklassen bleibt der Nachteil, die Nulldynamik durch Formulierung von Restriktionen für asymptotische Stabilität parametrisch fassen zu müssen. Letzteres ist nicht ganz einfach, gestaltet sich für ein- und zweidimensionale Systeme aber überschaubar [548].
Obwohl es zwar erfreulich ist, in der Identifikation ein nichtlineares Modell strukturell oder über Restriktionen sicher als (nicht)minimalphasig bestimmen zu können, bleibt der Nutzen im Nichtminimalphasenfall in Anbetracht der Regelungsprobleme fraglich. In vielen Fällen ist es deshalb ratsam, bereits beim Entwurf durch Positionierung von Aktoren und Sensoren und durch Festlegung der Regelgröße ein minimalphasiges Verhalten anzustreben. Bei Stabilität der Nulldynamik kann dann nämlich in aller Regel auf zusätzliche Restriktionen verzichtet werden, da diese dann implizit durch die experimentelle Modellanpassung eingehalten werden.

2.14 Kausalität

Ein kausales System ist dadurch gekennzeichnet, dass jedes Ausgangssignal nicht von zukünftigen Werten der Eingangssignale abhängt.

Definition 2.25 (Kausales System/Signal)
Ein System heißt kausal, wenn für alle Eingangssignale mit der Eigenschaft $u_1(t) \equiv u_2(t)$ für $t \leq t_1$ bei beliebigem t_1 die entsprechenden Ausgangssignale für $t_0 \leq t \leq t_1$ die Eigenschaft[109] haben $y_1\{x_0, u_1(t); t_0\} \equiv y_2\{x_0, u_2(t); t_0\}$. Systeme, die nicht kausal sind, werden auch akausal genannt. Ein Signal bzw. eine Folge heißt kausal, wenn es für alle $t < 0$ bzw. sie für alle $k < 0$ Null ist.

Eine zwingende Konsequenz aus der Definition ist, dass ein System genau dann kausal ist, wenn die Impulsantwort kausal ist [607]. Ferner müssen alle physikalischen Systeme kausal sein. Zeitdiskrete Filter, etwa zur Signalverarbeitung, können durchaus akausal sein, z. B.

$$y_k = \frac{u_{k+1} - u_{k-1}}{2T_A} \qquad \text{numerischer Differenzierer erster Ordnung.} \tag{2.169}$$

[109] Der Signalverlauf für $t > t_1$ kann sich unterscheiden. Bei kausalen Systemen dürfen also die Signalwerte in $t > t_1$ keinen Einfluss auf den Verlauf bis t_1 haben.

Aber auch physikalische Modelle kontinuierlicher Systeme können nicht kausal sein, wie das folgende Beispiel zeigt.

Beispiel 2.27 (Nichtkausales physikalisches Modell)
Das Skin-Effekt-Widerstandsmodell[110] $R(j\omega) = \sqrt{\frac{\mu_0 \varrho}{2}} \frac{l}{\pi d} \sqrt{|\omega|}$ ist nicht kausal, da ein reelles nichtkonstantes Spektrum niemals zu einem kausalen System gehören kann, vgl. Tabelle 2.1. Nichtsdestotrotz ist das Skin-Effekt-Widerstandsmodell für eine Reihe elektrotechnischer Anwendungen geeignet, für einige andere aber nicht (Modellierung passiver Netzwerke). Ein akausales Modell ist nämlich nie passiv [482].

Doch wodurch treten Akausalitäten auf, wenn sie doch physikalisch nicht existieren? Ursachen sind unter anderem das Vertauschen der Ursache-Wirkungsbeziehungen, Aliasing-Phänomene oder vereinfachte Modelle. Auch sind bei sehr komplexen Prozessen die Ursache-Wirkungsbeziehungen nicht immer offensichtlich. Ein Vertauschen der Signalwirkungsrichtung begründet dann Akausalitäten. Zu beachten ist auch, dass unterschiedliche Dynamiken in den Messpfaden zu einer Verzerrung der Ursache-Wirkung-Beziehungen führen können. Mit Hilfe der Kreuzkovarianzfunktion lassen sich Akausalitäten erkennen (Extremum für negative Verschiebungen).
Bei Schätzungen können durch Akausalitäten instabile Modelle entstehen, da die Verfahren über die Parameter eine fehlerhafte Ursache-Wirkungsbeziehung auszugleichen versuchen, siehe hierzu Beispiel 2.28. Eine so erzeugte künstliche Instabilität hat fatale Auswirkungen in simulativen oder adaptiven Anwendungen.

Beispiel 2.28 (Instabilität bei vertauschter Wirkungskette)
Das stabile System $G(z) = \frac{0.2}{z-0.8}$ mit dem Anfangswert $y[0] = 1$ wird durch $u[0] = 2$, $u[1] = 2$ und $u[2] = 1$ erregt, womit $y[1] = 1.2$ und $y[2] = 1.36$ folgen. Für die Gleichungsfehlermodelle $\theta_1 y[k] + \theta_2 u[k] = y[k+1]$ und $\tilde{\theta}_1 u[k] + \tilde{\theta}_2 y[k] = u[k+1]$ (vertauschte Signale) ergeben sich die Gleichungssysteme

$$\begin{bmatrix} 1 & 2 \\ 1.2 & 2 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1.2 \\ 1.36 \end{bmatrix} \quad \text{und} \quad \begin{bmatrix} 2 & 1 \\ 2 & 1.2 \end{bmatrix} \begin{bmatrix} \tilde{\theta}_1 \\ \tilde{\theta}_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

mit den Lösungen $\theta = [0.8, 0.2]^T$ und $\tilde{\theta} = [3.5, -5]^T$. Das mit fehlerhafter Ursache-Wirkungsbeziehung ermittelte Modell ist wegen $\tilde{\theta}_1 > 1$ instabil.

Zur Einbeziehung der Kausalitätseigenschaft bei der Identifikation empfehlen sich spezielle Ansätze (Nullfunktion für $t < 0$, keine prädiktiven Elemente in z-Übertragungsfunktionen,

[110] Fließt Wechselstrom durch einen Leiter, induziert der sich periodisch ändernde Strom eine gegen das Leiterinnere zunehmende Spannung in Flussrichtung, welche die im Zentrum fließenden Elektronen bremst und dadurch bewirkt, dass der Strom an die Außenwand des Leiters gedrängt wird. Dieser Effekt zeigt sich mit steigender Frequenz immer ausgeprägter.

$x(t)$ reell	$\Leftrightarrow X(-j\omega) = \bar{X}(j\omega)$
$x(t)$ reell und gerade	$\Leftrightarrow X(j\omega)$ reell und gerade
$x(t)$ reell und ungerade	$\Leftrightarrow X(j\omega)$ imaginär und ungerade
$x(t) \geq 0$ für alle t	$\Rightarrow \lvert X(0)\rvert \geq X(j\omega)$
$x(t)$ Autokovarianzfunktion	$\Leftrightarrow X(j\omega) \geq 0$ für alle ω
$x(t)$ reell und kausal	$\Rightarrow X(j\omega)$ entweder komplex ($X(j\omega) \not\subseteq \mathbb{R}, X(j\omega) \not\subseteq j\mathbb{R}$) oder reell und konstant
$x(t)$ komplex und kausal[111]	$x_g(t) = \frac{1}{2}x(t)$ für $t \geq 0$ und $x_g(t) = x_g(-t)$ für $t < 0$ $x_g(t) = x_u(t)\operatorname{sgn} t$ mit $x_g(t) = \frac{1}{2}(x(t) + x(-t))$ und $x_u(t) = \frac{1}{2}(x(t) - x(-t))$
$x(t)$ stabil und kausal	$\Leftrightarrow X(j\omega) = \mathfrak{F}\{x_g(t)\} + j\mathfrak{H}^{-1}\{\mathfrak{F}\{x_g(t)\}\}$ $\Leftrightarrow \Re e\, X(j\omega) = \Re e\, X(j\omega)\Big\rvert_{\omega=\infty} + \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\Im m\, X(j\eta)}{\omega - \eta}\,\mathrm{d}\eta$ /[112] $\Leftrightarrow \Im m\, X(j\omega) = -\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\Re e\, X(j\eta)}{\omega - \eta}\,\mathrm{d}\eta$ /[113] $\Leftrightarrow \int_0^\infty \Re e\, X(j\omega)\cos(\omega t)\mathrm{d}\omega = -\int_0^\infty \Im m\, X(j\omega)\sin(\omega t)\mathrm{d}\omega;\ t > 0$

Tabelle 2.1: Zeit-Frequenzbereichsbeziehungen [607]

nichtnegative Totzeit), da eine explizite Behandlung der unendlichdimensionalen Frequenzbereichsrestriktionen sehr kompliziert wird.

Nach Tabelle 2.1 ist das zentrale Merkmal eines kausalen Signals, dass Real- und Imaginärteil des Spektrums nicht voneinander unabhängig, sondern über die Hilbert-Transformation verknüpft sind. Da in kausalen LTI-Systemen die Gewichtsfunktion kausal ist und da Gewichts- und Frequenzgangfunktion über die Fourier-Transformation gekoppelt sind, ist ein LTI-System genau dann kausal, wenn $\Im m\, G(j\omega) = -\mathfrak{H}\{\Re e\, G(j\omega)\}$ gilt, wenn also $\Re e\, G(j\omega)$ komplett ausreicht, um $G(j\omega)$ zu konstruieren. Hierin liegt der Schlüssel für Tests, ob ein Spektrum zu einem kausalen System gehört, und der Schlüssel für Restriktionen an Spektralschätzungen.

---

[111] Komplexwertige Signale werden z. B. für die Beschreibung von Übertragungssystemen mit Modulation verwendet. In diesem Zusammenhang spielen analytische Signale eine wichtige Rolle, die gewissermaßen das Pendant zu den kausalen Signalen sind. Bei ihnen verschwindet das Spektrum für negative Frequenzen komplett. Es gelten ähnliche Beziehungen für die Hilbert-Transformierten.

[112] Der Integralterm steht für die Hilbert-Transformation $\mathfrak{H}\{f(y)\} = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{f(x)}{y-x}\,\mathrm{d}x = f(y) * \frac{1}{\pi y}$.

$\mathfrak{H}\{f(y)\} = \frac{1}{\pi}\lim\limits_{\substack{\varepsilon\to+0\\ z\to\infty}}\big(\int_{-z}^{y-\varepsilon} \frac{f(x)}{y-x}\,\mathrm{d}x + \int_{y+\varepsilon}^{z} \frac{f(x)}{y-x}\,\mathrm{d}x\big) = -\frac{1}{\pi}\lim\limits_{\varepsilon\to+0}\int_{\varepsilon}^{\infty} \frac{f(y+x)-f(y-x)}{x}\,\mathrm{d}x$ ist als Cauchyscher Hauptwert zu verstehen. Falls auf die Distributionentheorie ausgewichen werden muss, ist im Faltungsterm die singuläre „principal value Distribution“ $\mathrm{p.v.}\frac{1}{\pi y}(\phi) = \lim\limits_{\varepsilon\to+0}\int_{|y|>\varepsilon} \frac{\phi(y)}{\pi y}\,\mathrm{d}y$ zu verwenden.

[113] Der Integralterm steht für die inverse Hilbert-Transformation $\mathfrak{H}^{-1} := -\mathfrak{H}$, denn $\mathfrak{H}\{\mathfrak{H}\{x(t)\}\} = -x(t)$.

2.15 Bifurkationen

Bei praktischen Systemen wurde beobachtet, dass kleine Änderungen von Parametern qualitative (abrupte) Änderungen des Verhaltens verursachen können. Als Beispiel seien chemischen Kolonnen genannt, die auf konstante Arbeitspunkte eingestellt sind und auf einmal anfangen zu schwingen, ohne dass sich die Stellgrößen änderten. Ursachen für derartige Phänomene können nicht geregelte Temperaturen oder Konzentrationen sein, die im Grunde als (zeitvariante) Parameter wirken. Solche parameterbedingten qualitativen Verhaltensänderungen werden als Bifurkationen bezeichnet.

Definition 2.26 (Bifurkation, [372])
Die Änderung der topologischen Eigenschaften des Flusses eines Systems (Anzahl der stationären Punkte, periodische Orbits) in Abhängigkeit von einem oder mehreren Parametern heißt Bifurkation. Die grafische Darstellung der Tupel (x, θ), die der Gleichung $f(x; \theta) = 0$ genügen, wird Bifurkationsdiagramm für $\dot{x} = f(x; \theta)$ genannt. Jene Punkte im Bifurkationsdiagramm, in denen Bifurkationen auftreten, heißen Bifurkationspunkte.
Bifurkationen, die in der Nähe einzelner Trajektorien ablaufen, heißen lokal. Betreffen sie sofort einen großen Teil des Phasenraums, werden sie als globale Bifurkationen bezeichnet. Eine Bifurkation heißt katastrophisch, wenn ein Attraktor[114] samt Einzugsgebiet verschwindet, andernfalls nichtkatastrophisch (engl. subtle), wobei dabei die zahlenmäßige Differenz von Attraktoren (inkl. periodischer und seltsamer) und Repelloren[115] gleichbleibt.
Die kleinste Parameterzahl, die variiert werden muss, um eine bestimmte Bifurkation zu erzeugen, heißt Kodimension.

Beispiel 2.29 (Sattel-Knoten-Bifurkation)
Das System $\dot{x} = \theta - x^2$ hat für den Bifurkationsparameter $\theta < 0$ keine Ruhelagen, im Bifurkationspunkt $(x, \theta) = (0, 0)$ eine Ruhelage und für $\theta > 0$ je einen stabilen Knoten und einen Sattel[116]. Sattel und Knoten erscheinen aus dem Nichts oder treffen im Bifurkationspunkt zusammen, je nachdem, ob θ auf- oder absteigend betrachtet wird.

Anmerkung 2.59 Die Tatsache, dass das Bifurkationsverhalten von der Richtung abhängt, in der der Parameter geändert wird, führt auch bei anderen Bifurkationstypen dazu, dass in der Literatur die verbale Formulierung des Verhaltens unterschiedlich ausfällt.

[114] Ein Attraktor beschreibt eine Untermenge im Phasenraum, auf die sich die Lösungen zubewegen und die sie unter der Systemdynamik nicht mehr verlassen.

[115] Ein Repellor ist das Gegenteil eines Attraktors, also eine Untermenge, die auf Lösungen abstoßend wirkt (instabiles Verhalten).

[116] Eine instabile Ruhelage kann im eindimensionalen Fall als Sattel interpretiert werden. Der Name erschließt sich im zweidimensionalen Fall $\dot{x}_1 = \theta - x_1^2, \dot{x}_2 = -x_2$ klarer, da dort ein Sattel und ein stabiler Fixpunkt auftreten.

Anmerkung 2.60 Bei lokalen Bifurkationen kollidieren Attraktoren, Repelloren und Sättel. Anders ausgedrückt: Es kommt in Abhängigkeit von einem oder mehreren Parametern zu einer Verzweigung der Lösungskurven (Ruhelagenkurven, Menge der Ruhelagen über dem Parameter), z. B. in einen stabilen und einen instabilen Zweig von Ruhelagen.
Bei globalen Bifurkationen, wie der Sattelverbindungsbifurkation[117], kollidieren nicht die Attraktoren, Repelloren und Sättel selbst, sondern deren In- und Outsets. Da bereits in $\mathbb{R}^3$ als Attraktoren neben Fixpunkten, Grenzzyklen[118] und Tori (jeweils hyperbolisch und nicht-hyperbolisch) auch seltsame und chaotische Attraktoren (Charakterisierung über Lyapunov-Exponenten) auftreten, sind weitere, noch unentdeckte Bifurkationstypen in höheren Dimensionen zu erwarten. Ein Beispiel stellt die von Turaev und Shilnikov (1995) entdeckte Blue-Sky-Bifurkation dar, bei der ein Grenzzyklus mit einem nicht-hyperbolischen Zyklus kollidiert (globale Bifurkation).
Bei einer Isola-Bifurkation[119] verschmilzt eine offene und eine geschlossene Lösungskurve zu einer offenen Lösungskurve, s. Beispiel in [509].

Anmerkung 2.61 Hat ein System eine Bifurkation, dann hat es meist noch weitere [14]. Parameterinduzierte Hysteresen (Hysterese im Parameter-Ruhelagen-Diagramm) werden durch zwei katastrophische Bifurkationen hervorgerufen, also etwa zwei Sattel-Knoten- oder eine Sattel-Knoten- und eine transkritische Bifurkation. In aller Regel liegt dann zwischen zwei Attraktoren ein Repellor. Von den Bifurkationen sind die meisten katastrophisch; auch sind lokale Bifurkationen häufiger als globale.

Beispiel 2.30 (Bifurkation in linearen Systemen)
Obwohl Bifurkationen vorrangig im Rahmen der nichtlinearen Systeme behandelt werden, treten sie auch bei linearen Systemen auf. So hat $\dot{x}_1 = \theta x_2, \dot{x}_2 = -x_1 - 2x_2$ für $\theta < 0$ einen Sattel, in $\theta = 0$ die gesamte Gerade $x_2 = -x_1/2$ als Ruhelage und für $0 < \theta$ eine stabile Ruhelage. Bifurkationen für zweidimensionale LTI-Systeme lassen sich statt an den vier Systemparametern besser über Spur und Determinante charakterisieren, s. Spur-Determinanten-Bifurkationsdiagramm (-restriktionen) in [302].
Für nichtautonome quadratische Systeme ist ein Tripel (x_e, u_e, y_e) ein statischer Bifurkationspunkt, nur wenn gilt

$$\text{Rang} \begin{bmatrix} -A & B \\ -C & 0_{m\times m} \end{bmatrix} \neq m + n \qquad \text{(notwendige Bedingung).} \tag{2.170}$$

[117] Eine Sattelverbindung kann heteroklinisch sein, d. h. eine von einem Sattel weggehende Trajektorie geht direkt zu einem zweiten Sattel hin (Grenzzyklus kollidiert mit zwei Sätteln). Sie kann aber auch homoklin sein, d. h. eine weggehende Trajektorie geht zum gleichen Sattel hin, bildet also eine Schleife (Grenzzyklus kollidiert mit Sattel). Von einer simultanen homoklinischen Verbindung wird gesprochen, wenn sich um einen Sattel zwei derartige Schleifen bilden.

[118] Ein Grenzzyklus ist ein isolierter periodischer Orbit in der Menge aller periodischen Orbits.

[119] Eine isolierte Lösung einer Differenzialgleichung heißt Isola.

(2.170) impliziert Invariante Nullstellen im Ursprung (nicht-degenerierter Fall) [373].[120]

Lokale Bifurkationen treten auf, wenn die Eigenwerte der Jacobi-Matrix in Abhängigkeit von θ die imaginäre Achse passieren. Generisch ist dies ein einzelner reeller Eigenwert oder ein einzelnes konjugiertes Paar, weshalb für lokale Bifurkationen die Betrachtung sich weitgehend auf zweidimensionale Systeme beschränken kann. Globale Bifurkationen lassen sich hingegen nicht über die Eigenwerte erfassen, sondern sind über die Lösungen abzuleiten, was sehr kompliziert werden kann. Zudem sind Bifurkationen in höherdimensionalen Systemen sehr vielschichtig. Insgesamt wird sich deshalb im Folgenden auf den Fall $n = 2$ beschränkt. Damit dann eine Bifurkation auftritt bzw. damit ein Parameter ein Bifurkationspunkt ist, muss wenigstens eine der folgenden vier Bedingungen gelten [302]:

1. Eine Ruhelage x_e hat eine Linearisierung mit einem Nulleigenwert, d. h.

$$\det \left.\frac{\partial f}{\partial x^T}\right|_{x_e} = 0 \qquad \text{(generische Bedingung)}. \tag{2.171}$$

   Die Bedingung charakterisiert die sog. Ruhelagenbifurkationen, zu denen die Sattel-Knoten-Bifurkation (generischer Bifurkationstyp für hinreichend glatte Vektorfelder [101]), die transkritische Bifurkation und die Heugabel-Bifurkation (Beispiel: Wattscher Zentrifugalrotator) zählen. Kollidieren im Bifurkationspunkt mehr als zwei Ruhelagen, also etwa drei wie bei der Heugabel-Bifurkation, so wird eine solche Bifurkation degeneriert und nicht generisch genannt. Eine Bedingung für nicht-degenerierte Bifurkationen ist, dass der Nulleigenwert der Linearisierung einfach ist.

2. Die Linearisierung hat ein rein imaginäres Eigenpaar, d. h.

$$\text{spur}\left.\frac{\partial f}{\partial x^T}\right|_{x_e} = 0, \quad \det \left.\frac{\partial f}{\partial x^T}\right|_{x_e} > 0 \qquad \text{(generische Bedingung)}. \tag{2.172}$$

   Hierzu zählt die Andronov-Hopf-Bifurkation[121]

3. Es gibt eine Sattelverbindung.

4. Es gibt eine Zyklenauslöschung, d. h. im Kollisionspunkt entsteht ein semistabiler Grenzzyklus.

[120] Der degenerierte Fall mit unzureichenden unabhängigen Eingängen $\text{Rang}\, B < m$ bzw. redundanten Ausgängen $\text{Rang}\, C < m$ ist insbesondere bei der linearisierten Variante mit von Interesse.

[121] Vielfach wird nur von einer Hopf-Bifurkation, benannt nach dem holländischen Mathematiker Eberhardt Hopf, gesprochen. Gleichwohl entdeckte der russische Mathematiker A.A. Andronov in den 1930iger Jahren das Phänomen.

Anmerkung 2.62 De facto als Gegenstück werden die Morse-Smale-Vektorfelder auf zweidimensionalen Mannigfaltigkeiten definiert: endliche Anzahl von Ruhelage, die alle hyperbolisch sind; endliche Anzahl von hyperbolischen Grenzzyklen (linear anziehend oder abstoßend); keine Sattelverbindungen; alle Trajektorien, die keine kritischen Elemente (Ruhelage oder Grenzzyklus) sind, streben für $t \to \infty$ oder für $t \to -\infty$ gegen genau ein kritisches Element. Morse-Smale-Vektorfelder sind strukturell stabil und dicht (Peixoto-Theorem [302]).

Die lokalen Bifurkationen und der Satz von Hartman und Grobman [548], [596], [275] über die strukturelle Stabilität hyperbolischer Ruhelagen legen nahe, als Maß für die Bifurkationsfreiheit die normkleinste Störung zu wählen, die vonnöten ist, um wenigstens einen Eigenwert der Jacobi-Matrix auf die imaginäre Achse zu schieben. Für autonome stabile LTI-Systeme ist das gerade der Stabilitätsradius. Unter Einschränkungen der Struktur ist der strukturierte Singulärwert (μ-Funktion) ein Maß, das aus der Theorie der robusten Regelungen bekannt ist. Bei niedrigdimensionalen Problemen führt die Realteil-Null-Bedingung der Linearisierung in aller Regel auf eine Hyperfläche im Parameterraum Θ, also eine Menge $\mathcal{W}$ mit einer um Eins kleineren Dimension (für $\theta \in \mathbb{R}^1$ auf Punkte, für $\theta \in \mathbb{R}^2$ auf eindimensionale Kurven usw.). Mittels numerischer Optimierung kann der kürzeste Abstand zu $\mathcal{W}$ bestimmt werden [165]. Daraus und aus der Struktur von $\mathcal{M}$ ergeben sich Hinweise für die Wahl des Arbeitsbereichs von Regelungen, die möglichst weit von Bifurkationen entfernt operieren sollten.

Neben diesem A-posteriori-Maß bietet sich bei bekanntem Bifurkationstyp an, vgl. Tabelle 2.2, das Bifurkationsverhalten strukturell über sog. Bifurkationsnormalformen zu sichern. Je nach Bifurkationstyp werden dabei bestimmte Approximationen des Vektorfelds verwendet [260], [30]. Auf diese Weise lässt sich A-priori-Information nutzen, um einerseits die richtigen nichtlinearen Terme zu verwenden und andererseits die Parameteranzahl für die Identifikation klein zu halten. Bei den Bifurkationsnormalformen handelt es sich um Differenzialgleichungen, deren Vektorfelder Polynome sind. Im Bifurkationspunkt verhalten sich die Approximationen dann topologisch wie das Originalsystem. Vielfach werden die Normalformen entsprechend dem Satz über die Zentrumsmannigfaltigkeit [345], [548] auf niedrigere Dimensionen reduziert. So lautet die auf die eindimensionale Zentrumsmannigfaltigkeit reduzierte Sattel-Knoten-Bifurkationsnormalform $\dot{x} = \mu - x^2 + \ldots$. Die Punkte stehen für Monome höherer Ordnung. Die reduzierte Normalform mit der kleinsten Polynomordnung, also $\dot{x} = \mu - x^2$, heißt verkürzte Normalform. Normalformen sind in der Regel nicht eindeutig, vgl. Beispiel 2.31.

Die nachfolgende Darstellung beschränkt sich auf autonome Systeme $\dot{x} = f(x;\theta)$, die in $(x,\theta)_{Bif}$ eine Bifurkation haben mögen. In einem ersten Schritt wird das System so transformiert, dass die Bifurkation in $y = 0$ stattfindet, also $y = x - x_{Bif}$. Vielfach wird auch noch der Parametervektor in den Ursprung transformiert $\mu = T(\theta - \theta_{Bif})$. Die Transfor-

Name	Merkmal	Beispiel	glob.	lok.	n.-kat.	kat.
Punktauslöschung (engl.: static fold)	Attraktor und Repellor kollidieren und löschen sich aus	$\dot{x} = \theta + x^2$		x		x
Zyklenauslöschung (engl.: cyclic fold)	stabiler und instabiler Grenzzyklus kollidieren und löschen sich aus	$\dot{r} = (\frac{1}{4}r^4 - \frac{1}{2}r^2 + \theta)r$ $\dot{\phi} = 1; \quad \theta_{Bif} = \frac{1}{4}$		x		x
Sattel- Verbindungsbifurkation	Auflösen einer Sattel-Sattel-Separatrix (Sattel-Sattel-Verbindung)	$\dot{x}_1 = \theta + 2x_1x_2$ $\dot{x}_2 = 1 + x_1^2 - x_2^2$	x			x
transkritische Bifurkation	Attraktor und Repellor kollidieren und tauschen ihr Anzieh-Abstoßverhalten	$\dot{x} = (\theta - x)x$		x	x	
Sattelknoten-Bifurkation	ein Sattel und ein Knoten kollidieren und löschen sich aus	$\dot{x}_1 = \theta + x_1^2$ $\dot{x}_2 = -x_2$		x		x
superkritische Pitchfork-Bifurcation	aus einem werden zwei Attraktoren und ein Repellor, der beide trennt	$\dot{x} = (\theta - x^2)x$		x	x	
subkritische Pitchfork-Bifurcation	aus einem Attraktor und zwei Repelloren wird ein Repellor	$\dot{x} = (\theta + x^2)x$		x		x
superkritische Hopf-Bifurcation	Attraktor wird zu Repellor, umgeben von einem stabilen Grenzzyklus	$\dot{r} = \theta r + r^3$ $\dot{\phi} = 1 + \theta + r^2$		x	x	
subkritische Hopf-Bifurcation	Repellor wird zum Attraktor, umgeben von instabilem Grenzzyklus	$\dot{r} = \theta r - r^3$ $\dot{\phi} = 1 + \theta + r^2$		x		x
Spitzenbifurkation (engl.: cusp bifurcation)	drei stationäre Punkte werden zu einem (gefaltete Fläche $f(x) = 0$)	$\dot{x} = \theta_1 + \theta_2 x - x^3$		x		x

Tabelle 2.2: Zusammenstellung einiger Bifurkationen, [302], [509], [565], [548]

mation T dient der Aufspaltung von $\mu^T = (\mu_1^T, \mu_2^T)$ in einen Teilparametervektor μ_1 mit der Kodimension der Bifurkation und einen Restparametervektor. Ausgangspunkt für die Bifurkationsnormalformen ist demnach ein System

$$\dot{y} = f(y; \mu_1) \qquad \text{mit } f(0_n) = 0_n; \mu_{1,Bif} = 0_{codim}. \tag{2.173}$$

Mit Hilfe einer klassischen Ähnlichkeitstransformation auf reelle Jordan-Form $z = Sy$ und einer anschließenden Folge von (identitätsnahen) Koordinatentransformationen vom Typ $z := z + h_k(z)$ mit h_k aus der Menge der multivariaten Polynome vom Grad k wird (2.173) auf

$$\dot{z} = Jz + \phi_2(z) + \phi_3(z) + \dots \tag{2.174}$$

transformiert, s. [509], wobei $\phi_2(z)$ für alle multivariaten Polynome in den $z_1, \dots, z_n$ vom Grad 2 und $\phi_3(z)$ die vom Grade 3 enthält. J ist die auf reelle Jordan-Form transformierte Jacobi-Matrix von f. Das Ziel der Transformationen ist es, dass die ϕ_k möglichst eine geringe Anzahl von Monomen aufweisen. Die Transformation selbst ist dabei für die Identifikation gar nicht so bedeutend. Vielmehr ist es wichtig zu wissen, welche nichtlinearen Terme auftreten müssen, um ein bestimmtes Verhalten zu realisieren. Speziell für $n = 2$ ergeben sich dann die Normalformen für einen Nulleigenwert, ein rein imaginäres Paar und einen Doppelnulleigenwert

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \sum_{k=2}^{N} \begin{bmatrix} a_k z_1 z_2^{k-1} \\ b_k z_2^k \end{bmatrix} + O(||z||_2^{N+1}) \tag{2.175a}$$

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} 0 & -\beta \\ \beta & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \sum_{k=1}^{[\frac{N-1}{2}]} (z_1^2 + z_2^2)^k \begin{bmatrix} (a_k - b_k) z_1 \\ (a_k + b_k) z_2 \end{bmatrix} + O(||z||_2^{N+1}) \tag{2.175b}$$

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \sum_{k=2}^{N} \begin{bmatrix} a_k z_1^k \\ b_k z_1^k \end{bmatrix} + O(||z||_2^{N+1}) \tag{2.175c}$$

Eine Hopf-Bifurkation braucht für den Grenzzyklus ein konjugiertes Paar, womit die Normalform (2.175b) angezeigt ist. Aus der verkürzten Form mit $a_1 = 1, b_1 = 0$ ergibt sich beispielsweise eine subkritische Hopf-Bifurkation

$$\dot{z}_1 = \mu z_1 - \beta z_2 + z_1(z_1^2 + z_2^2) \tag{2.176}$$

$$\dot{z}_2 = \beta z_1 + \mu z_2 + z_1(z_1^2 + z_2^2) \tag{2.177}$$

mit μ als Bifurkationsparameter und $\sqrt{\beta}$ als Radius des Grenzzyklus (In Tabelle 2.2 wurde eine Darstellung in Polarkoordinaten gewählt.).

Die bei der Transformation auftretenden Freiheiten zur Reduktion von Termen in $\phi_k(z)$ können auf unterschiedliche Normalformen führen.

Beispiel 2.31 (Takens- und Bogdanov-Normalform)
Die Bogdanov-Takens-Bifurkation ist eine Ruhelagenbifurkation einer zweiparametrischen Familie von Differenzialgleichungen, in der die kritische Ruhelage einen zweifachen Nulleigenwert hat. Zwei verkürzte Normalformen lauten:

Takens	Bogdanov
$\dot{x}_1 = x_2 + ax_1^2$	$\dot{x}_1 = x_2$
$\dot{x}_2 = bx_1^2$	$\dot{x}_2 = cx_1^2 + dx_1x_2$

Eine weitergehende Diskussion zu Normalformen der Bogdanov-Takens-Bifurkation findet sich in [619].

Für die Modellbildung ergibt sich aus der kurzen Ideenskizze zu Bifurkationsnormalformen folgendes Vorgehen: Zunächst ist zu prüfen, welche Bifurkation modelliert werden soll. Über eine Literaturrecherche ist im Anschluss eine Normalform für die entsprechende Bifurkation zu ermitteln. Alternativ lassen sich Darstellungen nutzen [302], [509], [565], die den Normalformen sehr ähnlich sind. Um die Normalformen herum werden dann Transformationen gebaut, die die Normalformkoordinaten in die Originalkoordinaten überführen $y = S^{-1}z$ ($S^{-1} = W$ ist eine Matrix mit n^2 freien Parametern) und dabei ggf. eine Anpassung an die Dimension des Originalproblems vornehmen. Im Anschluss wird eine Translation in den bekannten Bifurkationspunkt ausgeführt, wobei die Verschiebungsparameter, falls sie unbekannt sind, dem Schätzproblem hinzuzufügen sind. Ein ähnliches Vorgehen wird exemplarisch am Beispiel eines Rührkesselreaktormodells in der dreiteiligen Darstellung [531] gezeigt, in der auch einige Normalformen zusammengestellt sind.

Anmerkung 2.63 Die angesprochenen Normalformen beziehen sich auf lokale Bifurkationen. Für globale Bifurkationen gibt es ebenfalls Normalformen, wenngleich die zugehörigen Theorien ungleich schwieriger werden. Speziell für homokline Bifurkationen sei auf Arbeiten von Turaev[122] verwiesen.

[122] Normalformen für homokline Bifurkationen: www.wias-berlin.de/annual_report/2000/node30.html

Kapitel 3

Restriktionen an Funktionen

Wenn von Identifikation aus Sicht der Regelungstechnik gesprochen wird, fällt der erste Gedanke zumeist auf die dynamischen Modelle, die für den Entwurf leistungsstarker modellbasierter Regler benötigt werden. Das vorherige Kapitel widmete sich den dynamischen Modellen und belegt, dass nicht nur die Anzahl potenzieller Restriktionen groß ist, sondern wegen der zum Teil komplizierten Zusammenhänge zwischen den Parametern eine Vielzahl von Zugängen existiert. Sind die Regelungen allerdings für einen größeren Arbeitsbereich auszulegen, dann ist eine gute Beschreibung der statischen Kennlinie eine Voraussetzung für eine leistungsfähige Regelung (z. B. Regelung mit Gain-Scheduling oder mit statischer Vorsteuerung). In anderen Anwendungen spielen die dynamischen Zusammenhänge hingegen oft keine Rolle, wenn beispielsweise die statischen Kennlinien zur Interpolation, Prädiktion oder zum Ableiten von Kennwerten und Parametern verwendet werden. Mathematisch gesehen handelt es sich bei der Modellierung von Kennlinien oder allgemeiner Kennfeldern um die Approximation bzw. Regression von Funktionen. Wie dabei durch das Aufstellen und Einbeziehen von Restriktionen das Modellierungsergebnis verbessert werden kann, ist Gegenstand dieses Kapitels. Ziel ist es also, eine Funktion an die Prozessdaten (x_i, y_i) anzupassen und zwar so, dass sie vorab bekannte Eigenschaften einhält, z. B. Monotonie, Konvexität, Unimodalität, Positivität. Die Funktion kann dabei parametrisch beschrieben werden durch $f(x;\theta)$ bzw. nichtparametrisch durch $f_1, \ldots, f_M$ (f_j sind Stützwerte der Funktion). Da Funktionen und Folgen (Funktionen mit $\mathbb{N}$ oder $\mathbb{Z}$ als Definitionsbereich) neben der Beschreibung von statischen Zusammenhängen auch zur Beschreibung dynamischer Zusammenhänge (Sprungantwort, Korrelationsfolge, Spektren) eingesetzt werden können, ergänzt dieses Kapitel das auf Restriktionen an Differenzial- und Differenzengleichungen geprägte Kapitel 2 zur Einbeziehung von Vorwissen an dynamische Modelle. Nicht behandelt werden mehrdeutige Funktionen wie Kennlinien mit Gedächtnis (Lose, auch Spiel genannt; Hysterese). Hinsichtlich Lose sei auf [378] und zur Hysterese auf [369] verwiesen.

Die Herausforderung besteht nun darin, das Vorwissen mathematisch so umzusetzen, dass die resultierende Optimierungsaufgabe gut lösbar ist. Prinzipiell bieten sich hierzu folgende Zugänge an:

- expliziter Zugang über Restriktionen
- impliziter Zugang über Strukturansätze
- Kompromisszugang über Strafterme.

Diese Zugänge werden in Abschnitt 3.1 diskutiert. Es folgen Abschnitte, in denen das Behandeln einzelnen Funktionseigenschaften erörtert wird, wobei die Abschnitte relativ separat voneinander gelesen werden können. Die Eigenschaften sind:

- punktweise Restriktionen (Abschnitt 3.2)
- Stetigkeitsrestriktionen (Abschnitt 3.3)
- intervallweise Restriktionen (Abschnitt 3.4)
  mit den speziellen Restriktionen Nichtnegativität, Monotonie und Konvexität
- Unimodalitätsrestriktion (Abschnitt 3.5)
- Glattheitsrestriktionen (Abschnitt 3.6)
- Symmetrierestriktionen (Abschnitt 3.7)
- Sektorrestriktionen (Abschnitt 3.8).

Auszüge aus den Abschnitten 3.1 bis 3.8 flossen in die Übersichtsarbeit [249] ein.

Die angesprochenen Eigenschaften beziehen sich direkt auf die betreffenden Funktionen. Sie sind dem Ingenieur vertraut und aus der Kurvendiskussion bekannt. Etwas komplizierter ist die Situation, wenn es sich um Eigenschaften handelt, die sich dadurch ergeben, dass eine Funktion transformiert wird. Als klassisches Beispiel sei die Autokovarianzfunktion genannt, die sich als Fourier-Rücktransformierte der Spektraldichtefunktion (Leistungsdichtefunktion) darstellen lässt (Satz von Bochner[1] [83], Wiener-Khintchine-Beziehungen). Die Nichtnegativität der Dichte/Leistung äußert sich bedingt durch die Transformation in der nichtnegativen Definitheit der Kovarianzfunktion. Da die Beziehung zwischen Autokovarianzfunktion und Leistungsdichtespektrum eineindeutig ist, muss entweder die Nichtnegativität im Spektralbereich oder die Definitheit im indirekten Zeitbereich sichergestellt werden. Wie Letzteres geschehen kann, wird in Abschnitt 3.9 behandelt. Dabei wird sich aber auf das nichtparametrische Modell „Autokovarianzfolge“ konzentriert.

[1] Der Satz von Bochner besagt, dass Kovarianzfunktionen schwach stationärer Zufallsprozesse durch Fourier-Transformation eines positiven endlichen Maßes dargestellt werden können. Hat das Maß eine Dichte, die Spektraldichte oder auch Leistungsdichtespektrum genannt wird, dann ergeben sich Wiener-Khintchine-Beziehungen.

3.1 Formulierungszugänge

Die Eigenschaften an Funktionen können explizit über Restriktionen beschrieben oder implizit mittels Strukturansätzen sichergestellt werden. So lässt sich die Nichtnegativität einer Funktion $f : \mathcal{D} \to \mathbb{R}$ durch die unendlich vielen Restriktionen $\forall x \in \mathcal{D} : f(x;\theta) \geq 0$ (meist relaxiert durch endlich viele Restriktionen $f(x_k) \geq 0;\ k = 1,\ldots,N$) erzwingen. Alternativ sichert ein Ansatz $f(x;\theta) = [g(x;\theta)]^2$ mit beliebig wählbarem g ebenfalls Nichtnegativität. Der Vorteil der Restriktionen ist, dass sie die Zielfunktion nicht ändern, sondern nur den zulässigen Suchraum einschränken. Allerdings erfordern sie das Lösen einer restringierten Optimierungsaufgabe. Der Vorteil der Strukturansätze ist, dass sie ohne Restriktionen auskommen, da die geforderte Eigenschaft per se eingehalten wird. Als nachteilig erweist sich bei einigen Ansätzen der sog. Konvexitätsverlust. So ist zwar $f(x;\theta) = (\theta_0 + \theta_1 x)^2$ stets nichtnegativ, doch zum Preis eines nichtkonvexen LS-Problems $Q(\theta) = \sum_{i=1}^{N}(y_i - f(x_i;\theta))^2$. Überdies ist $Q(\theta)$ ein Polynom vierten Grads in den Parametern, weshalb Algorithmen langsamer konvergieren. Letzteres lässt sich durch Modifikation der Zielfunktion (Beträge statt Quadrate) oder der Algorithmen entschärfen. Schwerer wiegt dagegen der Konvexitätsverlust, denn damit sind lokale Minimierer nicht mehr automatisch auch globale. Zudem ist die Optimalitätsbedingung erster Ordnung dann nur noch notwendig.
Welchem der beiden Zugänge im konkreten Fall der Vorrang zu geben ist, hängt darüber hinaus auch vom Zusammenspiel mit anderen zu integrierenden Eigenschaften, den verfügbaren Softwaretools oder dem Modellierungsziel ab. Der Autor empfiehlt tendenziell den Restriktionszugang, da heutzutage leistungsfähige Algorithmen zur nichtlinearen restringierten Optimierung existieren [71]. Auch sind beispielsweise LS-Probleme mit linearen Gleichungs- und Ungleichungsrestriktionen nicht nennenswert schwieriger lösbar als das nicht restringierte LS-Problem (s. LSE- bzw. LSI-Problem in Tab. A.3). Gleiches gilt für parameterlineare Ausgleichs- und Regressionsprobleme in der Manhattan- oder Chebyshev-Norm[2] unter linearen Restriktionen, die sich in LP-Probleme (s. Tab. A.8) umformen lassen.

Expliziter und impliziter Zugang sichern ein „definitives Einhalten“ einer Eigenschaft. Der sog. Kompromisszugang begnügt sich mit einem „weitgehenden Einhalten“. Das wird durch ein Erweitern der Zielfunktion $Q(\theta)$ um additive Strafterme für die r Eigenschaften zu

$$\tilde{Q}(\theta) = Q(\theta) + \sum_{i=1}^{r} \gamma_i q_i(\theta) \tag{3.1}$$

erreicht. Dabei sind die γ_i Wichtungsfaktoren, mit denen der Grad des Einhaltens der jeweiligen Eigenschaft gesteuert werden kann und die im Fall von Minimierungsaufgaben positiv

[2] Manhattan-Norm (l_1-Norm): $||x||_1 \stackrel{\text{def}}{=} \sum_{i=1}^{n} |x_i|$ und Chebyshev-Norm (l_∞-Norm): $||x||_\infty \stackrel{\text{def}}{=} \max_{1\leq i\leq n} |x_i|$.

gewählt werden. Beispiele für Strafterme sind

$$q_1(\theta := f) = \sum_{k=1}^{N} (\min\{0, f_k\})^2;\ f = ((f_k)) \in \mathbb{R}^N \tag{3.2a}$$

$$q_2(\theta := f_j) = (f_j - a)^2 \tag{3.2b}$$

$$q_3(\theta := f) = ||Tf||_2 \tag{3.2c}$$

$$q_4(\theta) = (\max\{0, g(\theta)\})^\alpha;\ \alpha > 0 \tag{3.2d}$$

$$q_5(\theta) = |h(\theta)|^\alpha;\ \alpha > 0. \tag{3.2e}$$

Die ersten drei Terme gelten für nichtparametrische Modelle, in denen die Stützwerte f_k als Semiparameter aufgefasst werden. Hierbei soll der Präfix „semi“ verdeutlichen, dass es sich nur i. S. der Optimierung, nicht aber i. S. der Identifikation um Parameter handelt. Der Term q_1 bestraft ein Verletzen der Nichtnegativität an allen Stellen x_k. Typisch für die Behandlung einer punktweisen Restriktion ist der Term q_2. Konkret soll durch ihn die Forderung, wonach $f(x_j)$ den Wert a haben soll, einbezogen werden. q_3 ist ein Term, der bei inversen Problemen Anwendung findet. Im Fall $T = I_p$ bewertet er die l_2-Norm der Stützwertefolge (Maß für die Energie). Zahlreiche Variationen hinsichtlich der Norm und der Matrix T sind möglich. Auch wird gern die quadrierte Norm zwecks einfacherer Lösbarkeit der Optimierungsaufgabe verwendet. Mit der Bandmatrix $T = \frac{1}{(\Delta t)^2}\text{band}([1, -2, 1])$ lässt sich mittels numerischer zweiter Ableitung beispielsweise die Glattheit ausdrücken. Die Terme q_4 bzw. q_5 bestrafen das Verletzen einer algebraischen Ungleichungsrestriktion $g(\theta) \leq 0$ bzw. einer Gleichungsrestriktion $h(\theta) = 0$ und beziehen sich auf parametrische Modelle.

Im Grenzfall $\gamma_i \to \infty$ wird das definitive Einhalten der i-ten Eigenschaft erreicht. Die mathematisch strenge Formulierung hierfür liefert unter einigen schwachen Voraussetzungen der Satz von Pietrzykowski [320]. Nicht sofort offensichtlich – aber sehr nützlich – ist die Tatsache, dass in bestimmten Fällen bereits mit einem endlichen γ ein definitives Einhalten einer Restriktion garantiert werden kann. Das Konzept der exakten Penalty-Funktionen ist der Schlüssel hierfür [71].

Für den Zugang über Strafterme spricht zunächst der algorithmische Vorteil, kann doch dadurch eine freie statt einer restringierten Optimierungsaufgabe gelöst werden. Dank der additiven Kombination der Terme weicht $\tilde{Q}(\theta)$ weniger stark von $Q(\theta)$ ab als bei einigen impliziten Zugängen (s. vorgenannte Grenzbetrachtung). Ein weiterer Vorteil ist, dass wünschenswerte und/oder nicht streng quantifizierbare Eigenschaften einbezogen werden können. So ist die Glattheit einer Lösungsfunktion oft wünschenswert, doch kann a priori nicht gesagt werden, wie glatt die Lösung eigentlich sein muss. Schlussendlich eignet sich der Zugang bedingt für sich widersprechende Eigenschaften (sollte eigentlich vermieden werden), da er wenigstens eine Lösung liefert, während der explizite und implizite Zugang wegen des Widerspruchs ohne Lösung bleiben.

3.2 Punktweise Restriktionen

Eine punktweise Restriktion an eine Funktion ist eine Forderung, wonach für ein festgelegtes Argument der zugehörige Funktionswert oder eine Ableitung einen bestimmten Wert oder ein Ordinatenintervall exakt einzuhalten hat. Solche Forderungen können notwendig sein, da ohne sie etwa bei der Funktionsapproximation in der L_2- oder L_∞-Norm zwar eine geringe Abweichung über einen Bereich erzielt wird, aber die Funktion in prädistinierten Punkten (Nulldurchgänge, Minimierer) nicht mit der Originalfunktion übereinstimmt. Beispielsweise sollte die Approximierende eines Sinus bei $x = 0$ den Wert 0 exakt und nicht nur in Näherung liefern oder eben bei $x = \pi/6$ den Wert $1/2$.
Bei der Identifikation von Übergangsfolgen betrifft derartiges A-priori-Wissen meist den Anstieg bei $h'(0) \leq h'_{\max}$ (bekannter Anstieg bei Maximalaussteuerung) oder das asymptotische Verhalten $h(\infty) = K$. Weiterhin treten häufig Randrestriktionen auf. So beginnt eine Verteilungsfunktion einer stetigen skalaren Zufallsgröße am linken Intervallrand des Trägers bei 0 und endet am rechten bei 1. Typisch sind Randrestriktionen auch für Diagramme, in denen Verhältnisse dargestellt werden. Von der trivialen Restriktion $0 \leq f(x_{\text{rand}}) \leq 1$ abgesehen, kann der untere Rand oft durch das Anfangsmischungsverhältnis genauer spezifiziert werden, während am oberen Rand theoretische Überlegungen hinsichtlich des stöchiometrischen Gleichgewichts oder des vollständigen Umsatzes eine Eingrenzung rechtfertigen. Beispiele, in denen Randrestriktionen erster Ordnung zum Tragen kommen, sind Ortskurven (z. B. senkrechter Anstieg bei $\omega = 0$) oder Bézier-Kurven (Anstieg folgt aus Bézier-Punkten).

Im Fall parameterlinearer Modelle führen punktweise Restriktionen auf lineare Gleichungs- und Ungleichungsrestriktionen, z. B. für $f(x) = \theta_0 + \theta_1 x + \theta_2 x^2$ mit $f(2) = 3$ auf $\theta_0 + 2\theta_1 + 4\theta_2 = 3$. Ist darüber hinaus die Zielfunktion $Q(\theta)$ konvex über den Parametern, bleibt das entstehende restringierte Minimierungsproblem konvex und damit einfach. Werden mehrere gleichartige Restriktionen $f(x_k) = y_k$ gestellt, dann liegt eine sog. diskrete Approximation vor, die mit Interpolationsmethoden gelöst wird.
Das kontinuierliche Approximationspendant

$$||F(x) - f(x)||_{L^2} = \sqrt{\int_{-1}^{1} (F(x) - f(x))^2 \mathrm{d}x} \to \min_{f \in \mathcal{P}_n}, \tag{3.3}$$

bei dem im Intervall $[-1, 1]$ zu einer gegebenen Funktion $F(x)$ das bestapproximierende Polynom f ($\mathcal{P}_n$ Menge der Polynome bis zum Grad n) in der L^2-Norm gesucht ist, liefert eine gewichtete Summe [506], [163]

$$f^{\text{opt}}(x) = \sum_{i=0}^{n} \theta_i^{\text{opt}} \phi_i(x) \tag{3.4}$$

mit den Legendre-Koeffizienten

$$\theta_i^{\text{opt}} = \frac{2i+1}{2} \int_{-1}^{1} F(x) \phi_i(x) \mathrm{d}x \tag{3.5}$$

von Legendre-Polynomen

$$\phi_i(x) = \sum_{j=0}^{\lfloor i/2 \rfloor} (-1)^j \frac{(2i-2j)!\; x^{i-2j}}{2^i\; j!\; (i-j)!\; (i-2j)!} \tag{3.6}$$

als Lösung. Im Fall der $L^\infty([-1,1])$- bzw. $L^1([-1,1])$-Approximation ergeben sich gewichtete Summen von Chebyshev-Polynomen erster bzw. zweiter Art als Näherung (Near-Minimax-Approximationstheorem [430]). Kommen Randrestriktionen hinzu, ändern sich zwar die Polynomtypen, doch letztlich müssen auch wieder nur die Gewichte θ_i^{opt} berechnet werden (diverse Integrale über $F(x) \cdot \phi_i(x)$ mit den jeweiligen Basispolynomen ϕ_i).[3] Die Einschränkung auf $[-1,1]$ wird durch die affine Transformation

$$z = \frac{2}{b-a}\left(x - \frac{a+b}{2}\right) \tag{3.7}$$

aufgehoben (Bijektion von $[a,b]$ auf $[-1,1]$). Eine Anpassung einer Randrestriktion der Art $f(1) = F(1)$ an die Standard-Randrestriktion $f(1) = 0$ erfolgt über $f(x) := f(x) - F(1)$. Zur Behandlung allgemeinerer Randrestriktionen siehe auch Beispiel 3.2.

Obwohl Polynome bestimmte Approximationsaufgaben gut bewältigen, sind sie für periodische Funktionen ungeeignet. Dafür werden sinnvollerweise periodische Ansatzfunktionen (Fourier-Partialsummen) verwendet. Aber auch für Funktionen mit Sprungstellen, Funktionen mit Polen und Funktionen mit Asymptoten sind Polynome ungeeignet. Für diese Funktionstypen mit punktweisen Restriktionen empfehlen sich Funktionsansätze der Art

$$f(x;\theta) = u(x) + v(x)g(x;\theta) \tag{3.8}$$

Hierbei ist $u(x)$ eine beliebige Funktion, die die punktweisen Restriktionen erfüllt, und $v(x)$ eine Funktion, die an den Punkten Null ist und damit dort den Einfluss der frei wählbaren Funktion $g(x;\theta)$ vollständig unterdrückt. Die nachfolgenden zwei Beispiele verdeutlichen das Vorgehen.

Beispiel 3.1 (Struktureller Ansatz für punktweise Restriktionen)
Sei $f(0) = 1$, $f(1) = f'(1) = 0$ gefordert, dann erfüllt $u(x) = (1-x)^2$ die punktweisen Restriktionen $u(0) = 1, u(1) = 0, u'(1) = 0$ und $v(x) = x(1-x)^2$ nullt wegen $v(0) = v(1) = 0$, $v'(1) = 0$ jede Funktion $g(x;\theta)$ an den Stellen. Falls z. B. $f(2) = 0$, $f'(2) = \infty$ gefordert wird, dann leistet $u(x) = c|2-x|^{1/n}; n \geq 2$ und $v(x) = (2-x)^2$ das Gewünschte. Analog dazu erzeugt $u(x) = c(2-x)^{-2n}$ einen Pol und $u(x) = (2-x)^{-2n+1}$ einen Polwechsel. Eine Asymptote $a(x)$ für $x \to \infty$, z. B. $a(x) = 3x + 2$, kann über $u(x) = a(x)$ und $v(x) = \mathrm{e}^{-x}$ gesichert werden. Alternativ wird oft auch $v(x)g(x;\theta)$ durch eine echt rationale Funktion $r(x;\theta)$ ersetzt, wobei zusätzliche Bedingungen an den Nenner reelle Wurzeln ausschließen und so unerwünschte Polstellen unterdrücken.

[3] In ähnlicher Weise kann als Restriktion die Mittelwertfreiheit auf einem Intervall einbezogen werden [222].

Beispiel 3.2 (Struktureller Ansatz für Randrestriktionen)
Mit $f(x) = u(x)+v(x)g(x;\theta)$ und Polynomen $u(x)$, $v(x)$, $g(x;\theta)$ kann (3.3) unter zusätzlichen Randrestriktionen an den Stellen -1 und $+1$ auf bekannte Ergebnisse zurückgeführt werden.

$$v(x) = (1+x)^{\alpha}(1-x)^{\beta}; \qquad \alpha, \beta \in \mathbb{N}$$

stellt sicher, dass Bedingungen bis zur $(\alpha-1)$-ten bzw. $(\beta-1)$-ten Ableitung berücksichtigt werden können (Im Fall $\alpha = 0$ werden an den linken Rand keine Bedingungen gestellt.). Über eine allgemeine Hermite-Interpolation (Verfahren, das ein Polynom liefert, welches nicht nur durch die Punkte geht, sondern auch noch gestellte Ableitungsbedingungen einhält) kann $u(x)$ aus den Restriktionen $u^{(k)}(-1) = F^{(k)}(-1); k = 0, \ldots, \alpha-1$ und $u^{(k)}(1) = F^{(k)}(1); k = 0, \ldots, \beta-1$ berechnet werden. $F(x) := F(x) - u(x)$ liefert dann gemäß (3.3) das von Restriktionen befreite Problem

$$||F(x) - v(x)g_{\tilde{n}}(x;\theta)||_2 \stackrel{!}{=} \mathrm{Min}_{g_{\tilde{n}}(x)}$$

mit $\tilde{n} = n - \alpha - \beta$. $g_{\tilde{n}}(x;\theta)$ ist ein Polynom, dass sich stets durch eine Partialsumme $g_{\tilde{n}}(x;\theta) = \sum_{i=0}^{\tilde{n}} \theta_i \phi_i(x)$ mit Funktionen $\phi_i(x)$ aus einer orthogonalen Basis der Polynome bis zum Grade $\tilde{n}$ darstellen lässt. Einzig störend gegenüber (3.3) ist der Faktor $v(x)$. Dieser lässt sich aufheben, wenn statt der gewöhnlichen Orthogonalität in $\mathcal{L}^2([-1,1])$ eine auf dem Skalarprodukt $\langle r, s\rangle = \int_{-1}^{1} v^2(x)r(x)s(x)\mathrm{d}x$ begründete Orthogonalität gefordert wird. Das bedeutet aber, dass die $\phi_i(x)$ Jacobi-Polynome sein müssen.
Anwendungen aus dem computergestützten grafischen Entwurf mit Bézier-Kurven und -flächen werden von Y. J. Ahn vorgestellt, wobei [12] als Einstieg dienen kann.

Ein zu (3.8) alternativer Ansatz für das Einhalten einer oder mehrerer Polrestriktionen und von Asymptotenrestriktionen nutzt rationalen Funktionen. Zu diesem Zweck werden Ansatzfunktionen vom Typ

$$f(x) = a(x) + \frac{p(x;\theta)}{(x-x_1)\cdots(x-x_l)q(x;\theta)} \tag{3.9}$$

mit bekannter polynomialer Asymptote und bekannten Polstellen x_i und den unbekannten Parametern θ herangezogen. Der Ansatz

$$f(x) = c + \frac{\theta_1 x + \theta_0}{x^2 + \theta_4 x + \theta_3} \qquad \theta_4^2 - 4\theta_3 < 0 \tag{3.10}$$

garantiert etwa $f(\infty) = c$, wobei die Zusatzrestriktion reelle Pole ausschließt.

Losgelöst vom Restriktionsgedanken lässt sich eine rationale Approximation (auch Padé-Approximation genannt) in einigen Fällen durch Vorwissen rechtfertigen (indirekte Proportionalitäten in idealer Gasgleichung, reziprokale Substitution in der Bernoulli-Differenzialgleichung, Lösungen von Diffusionsgleichungen usw.). Ohne Vorwissen muss fallweise geprüft

werden, ob bei gleicher Zahl der Parameter ein Polynom, eine rationale Funktion oder ein anderer Funktionstyp die besseren Ergebnisse liefert. So fiel beispielsweise für eine industrielle μ-Controller-Anwendung, in der eine Approximation von Mollier-Diagrammen benötigt wurde, die Entscheidung nach eigenen Untersuchungen zugunsten einer rationalen Funktion. Hinsichtlich des Modellerstellungsaufwands unterscheiden sich Polynom und rationale Funktion kaum. Zwar ist das Polynom parameterlinear und damit bei einem Ausgangsfehler-LS-Kriterium direkt lösbar, doch bereitet auch die parameternichtlineare Ausgangsfehler-LS für die rationale Funktion keine größeren Probleme, da gängige Programmsysteme über Funktionen zur Lösung nichtlinearen LS-Probleme und zum Kurvenfit (Funktionsanpassung) verfügen. In zeitkritischen oder adaptiven Anwendungen kann die rationale Funktion auch als parameterlineare implizite Funktion behandelt werden (Durchmultiplizieren mit dem Nenner). So wird etwa statt

$$\sum_{i=1}^{N} \left(\frac{\theta_1}{x_i + \theta_2} - y_i \right)^2 \stackrel{!}{=} \text{Min} \tag{3.11}$$

nun

$$\sum_{i=1}^{N} (\theta_1 - \theta_2 y_i - x_i y_i)^2 \stackrel{!}{=} \text{Min} \tag{3.12}$$

betrachtet. Für Approximationsaufgaben (Modell genähert, keine Störungen) ist eine solche Umformung legitim, für Regressionsaufgaben (Modell exakt, Störungen) ergeben sich Abweichungen hinsichtlich Erwartungswert und Kovarianzen, da Störungen auf den Messdaten y_i nunmehr auf beiden Seiten der Gleichung auftreten.

3.3 Stetigkeitsrestriktionen

Eine Stetigkeitsrestriktion k-ter Ordnung in einem festgelegten Argument $x = w$ ist eine Forderung, wonach die k-te links- und rechtsseitige Ableitung in w gleich sind. Anders als bei den punktweisen Restriktionen wird keine Forderung an den konkreten Wert $f^{(k)}(w)$ gestellt. Stetigkeitsrestriktionen, in denen w nicht festgelegt wird, sondern in denen w ein zusätzlicher Parameter $\theta_0 := w$ ist, werden Stetigkeitsrestriktionen mit freiem Argument genannt. Analog werden Unstetigkeitsrestriktionen eingeführt.

Die Stetigkeitsforderung ist typisch für stückweise Approximationsprobleme. Sie tritt bei der Spline-Interpolation auf oder auch bei Kennlinien, bei denen sich in bestimmten Punkten die Struktur bzw. die Parameter ändern, weshalb dann eine punktübergreifende Approximation wenig erfolgversprechend ist. Stetigkeitsrestriktionen erster Ordnung werden gefordert, damit der Graph von f keine sichtbaren „Knicke“ aufweist. Unstetigkeitsforderungen beziehen sich in der Regel auf Vorwissen über die Sprunghöhe.

Im Folgenden werden zunächst die Stetigkeitsrestriktionen mit festem Argument behandelt. Für parameterlineare Modellansätze ist das in eleganter Weise möglich. Hierzu werden die zu verheftenden Funktionen an der bekannten Stelle w gleichgesetzt, wodurch eine lineare Restriktion an die Parameter der zu verheftenden Funktionen entsteht. Ebenso wird gegebenenfalls mit den Ableitungen verfahren. Das folgende Beispiel verdeutlicht das Vorgehen.

Beispiel 3.3 (Restriktionsbasierte Verheftung linearer Funktionen)
Gegeben seien Paare (x_k, y_k); $k = 1, \ldots, N$, die durch eine Funktion

$$f(x) = \begin{cases} \theta_1 x + \theta_0 & x < w \\ \theta_3 x + \theta_2 & x \geq w \end{cases} \qquad \underbrace{\theta_1 w + \theta_0 = \theta_3 w + \theta_2}_{\text{Stetigkeitsforderung}}$$

zu approximieren sind. Ohne Einschränkung kann angenommen werden, dass die Daten hinsichtlich w sortiert wurden, dass also für die ersten N_1 Paare $x_k < w$ gilt. Ein Ansatz lautet

$$\begin{bmatrix} 1 & x_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N_1} & 0 & 0 \\ 0 & 0 & 1 & x_{N_1+1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_N \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} \cong \begin{bmatrix} y_1 \\ \vdots \\ y_{N_1} \\ y_{N_1+1} \\ \vdots \\ y_N \end{bmatrix}.$$

Das in der Regel inkonsistente Gleichungssystem kann i. S. minimaler Residuen über $||A\theta - y|| \stackrel{!}{=} \text{Min}_\theta$ unter der linearen Restriktion $[1, w, -1, -w]\,\theta = 0$ mit zweckmäßig gewählter Norm gelöst werden. Für den Fall einer LS-Formulierung ($||\,.\,||_2^2$) entsteht ein LSE-Problem (s. Tab. A.3).

Neben der restringierten Behandlung gelingt für Stetigkeitsrestriktionen nullter Ordnung und parameterlineare Modelle auch eine restriktionsfreie Behandlung, wenn die Teilfunktionen mit dem Argument $(x - w)$ angesetzt werden und deren Absolutglieder in $x = w$ gleich sind. Eine Alternative zu Beispiel 3.3 zeigt Beispiel 3.4, in dem die Restriktion durch eine andere Darstellung der linearen Funktion vermieden wird und ein unbekannter Parameter weniger auftritt. Im Fall der linearen Funktionen ist die Variante aus Beispiel 3.4 zu bevorzugen. Demgegenüber ist die Variante aus Beispiel 3.3 dahingehend allgemeingültiger, dass sie auch zur Verheftung von Funktionen unterschiedlichen Typs verwendet werden kann.

Beispiel 3.4 (Ansatzbasierte Verheftung linearer Funktionen)
Mit dem Ansatz

$$f(x) = \begin{cases} \theta_1 (x - w) + \tilde{\theta}_0 & x < w \\ \theta_3 (x - w) + \tilde{\theta}_0 & x \geq w \end{cases}$$

ergibt sich für Paare (x_i, y_i) entsprechend Beispiel 3.3

$$\begin{bmatrix} 1 & x_1 - w & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_{N_1} - w & 0 \\ 1 & 0 & x_{N_1+1} - w \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_N - w \end{bmatrix} \begin{bmatrix} \tilde{\theta}_0 \\ \theta_1 \\ \theta_3 \end{bmatrix} \cong \begin{bmatrix} y_1 \\ \vdots \\ y_{N_1} \\ y_{N_1+1} \\ \vdots \\ y_N \end{bmatrix}.$$

Statt einer LS, empfiehlt sich für den Fall, dass auch die x_i gestört sind, eine Totale LS, eine Totale LS mit fester Spalte (hier Einsspalte) oder besser noch eine strukturierte Totale LS, bei der sämtliche Variationen der Null- und Einselemente ausgeschlossen werden [303].

Nach der Behandlung der Stetigkeitsrestriktionen mit festem Argument wird nun gezeigt, wie Stetigkeitsrestriktionen mit variablem Argument umgesetzt werden können. Hierzu bieten sich zwei Wege an. Zum einen kann das Problem bezüglich $\theta_0 := w_i$ mit $w_i \in [\underline{\theta}_0, \overline{\theta}_0]$ diskretisiert werden, was dann die Lösung von $i = 1, \ldots, M$ Problemen mit festgelegtem w_i erfordert. Das empfiehlt sich, wenn vorab klar ist, wo der Verheftungspunkt θ_0 grob liegt. $M = 10$ ist dabei oft ausreichend, wenn $[\underline{\theta}_0, \overline{\theta}_0]$ etwa ein Zehntel des gesamten Approximationsbereichs ausmacht. Zum anderen kann die in der Regelungstheorie gebräuchliche Technik der Signalkonstruktion mit Heaviside-Funktionen

$$1(x) = \frac{1}{2}(\operatorname{sgn} x + 1) \tag{3.13}$$

herangezogen werden. Auf diesem Weg entsteht etwa aus einer Sprungrestriktion $f(\theta_0 + 0) - f(\theta_0) = 2$ der Ansatz

$$f(x) = g_1(x; \theta_1) + 1(x - \theta_0) g_2(x; \theta_2);\ \theta_1 \in \mathbb{R}^{p_1}, \theta_2 \in \mathbb{R}^{p_2}$$

mit der Restriktion $g_2(\theta_0; \theta_1) - g_1(\theta_0; \theta_2) = 2$, die explizit, implizit oder per Strafterm behandelt werden kann. Dem Vorteil eines kompakten Ansatzes steht als Nachteil die Nichtkonvexität bzgl. θ_0 selbst bei parameterlinearen Ansätzen für $g_i(x; \theta_i)$ gegenüber. Dadurch werden zur Lösung der Probleme numerische Algorithmen benötigt, die ihrerseits in aller Regel die Ableitungen bzgl. θ_0 erfordern, weshalb auf Glättungen der Funktionen $1(x)$ ausgewichen wird. Hierfür empfehlen sich:

$$1_{A.1}(x) = \frac{1}{1 + \mathrm{e}^{-\kappa x}} \qquad \text{Sigmoidfunktion}^4 \tag{3.14a}$$

$$1_{A.2}(x) = \frac{1}{2}(1 + \operatorname{sgn}_{A.1}(x)) \qquad \text{mit } \operatorname{sgn}_{A.1}(x) = \tanh(\kappa x);\ \kappa \approx 1000.\ /^5 \tag{3.14b}$$

Das nachfolgende Beispiel zeigt eine Anwendung der beschriebenen Technik, wobei über den Glättungsparameter κ im Beispiel das Verschleifen der Knickstellen (in anderen Beispielen das von Sprungstellen) eingestellt werden kann.

Beispiel 3.5 (Verheftung stückweise definierter Funktionen)
Ein nichtsymmetrisches Totzone-Modell

$$f(x) = \begin{cases} g_1(x;\theta_1) & x < \theta_{01} \\ 0 & \theta_{01} \le x \le \theta_{02} \\ g_2(x;\theta_2) & x > \theta_{02} \end{cases}$$

lässt sich umformen in

$$f_A(x) = g_1(x;\theta_1)1_A(\theta_{01} - x) + g_2(x;\theta_2)1_A(x - \theta_{02}).$$

Das beschriebene Vorgehen gestattet Erweiterungen auf den $\mathbb{R}^n$. In $\mathbb{R}^2$ treten dann Verheftungen entlang von Geraden und Kurven auf. Anwendungen ergeben sich bei der Identifikation hybrider und stückweiser affiner Systeme [497].

3.4 Intervallweise Restriktionen

Eine intervallweise Restriktion an eine Funktion oder deren Ableitung ist eine Forderung, die für alle (unendliche viele) Argumente x aus Intervallen $(-\infty, \infty), (\infty, \overline{x}], [\underline{x}, \infty)$ oder $[\underline{x}, \overline{x}]$ gelten muss. Es ergeben sich somit unendlich viele Restriktionen. Die entstehenden Optimierungsprobleme werden daher als semi-unendliche Probleme bezeichnet („semi-“, da immerhin die Anzahl der Parameter θ_i endlich ist). Die intervallweisen Restriktionen

$$\forall x \in [\underline{x}, \overline{x}]: \quad a_k \le f^{(k)}(x;\theta) \le b_k; \quad k \in \mathbb{N} \tag{3.15}$$

umfassen unter anderem:
- die 1-Restriktionen ($k = 0, a_0 = 0,\ b_0 = 1$), s. hierzu Beispiel 3.6
- die Nichtnegativitätsrestriktionen ($k = 0, a_0 = 0,\ b_0 = \infty$),
- die Monotonierestriktionen (z.B. $k = 1, a_1 = 0, b_1 = \infty$ für monoton steigende Funktionen),
- die Konvexitäts- bzw. Konkavitätsrestriktionen ($k = 2$).

[4] Etwas allgemeiner heißt eine beschränkte, differenzierbare reelle Funktion Sigmoidfunktion, wenn ihre erste Ableitung durchweg positiv oder durchweg negativ ist und wenn sie genau einen Wendepunkt hat. Weitere Vertreter sind $\tanh(x), \arctan(x), \mathrm{erf}(x), \frac{x}{\sqrt{1+x^2}}$.

[5] Verglichen mit $\mathrm{sgn}_{A.2}(x) = \frac{2}{\pi}\arctan(\kappa x)$ hat $\mathrm{sgn}_{A.2}(x)$ den Vorteil, dass wegen $\frac{\mathrm{d}}{\mathrm{d}x}\tanh(x) = 1 - \tanh^2(x)$ für die Ableitung nur das Quadrat des ohnehin zu berechnenden Wertes $\tanh(x)$ benötigt wird.

Im Fall multivariater Funktionen wird von bereichsweisen Restriktionen gesprochen. In Fällen, in denen die Intervallgrenzen selbst noch frei sind, wird von adaptiven intervallweisen Restriktionen gesprochen.[6]

Beispiel 3.6 (100%-Restriktionen)
Anteile und Konzentrationen liegen immer zwischen 0 und 1 bzw. zwischen 0% und 100%. Dichtefunktionen sind immer nichtnegativ. Für Anteile, Konzentrationen und Dichten gilt außerdem, dass deren Summe bzw. Integral 1 bzw. 100% ist. Bei linearer Regression der Stoffausbeuten y bezüglich eines Prozessparameters $x \in [\underline{x}, \overline{x}]$ führt das bei drei Stoffen auf $y_{i,j} = \theta_{1,j} x_i + \theta_{2,j}$ mit $i = 1, \ldots, N$; $j = 1, 2, 3$ und die Restriktionen

$$\theta_{1,1} + \theta_{1,2} + \theta_{1,3} = 0 \qquad \text{und} \qquad \theta_{2,1} + \theta_{2,2} + \theta_{2,3} = 100\%.$$

Im Weiteren genügt es, immer nur ein Intervall zu betrachten, und somit die mehrfachen Intervallrestriktionen bei der stückweisen Approximation zu reduzieren. Zum einen kann die im vorherigen Abschnitt beschriebene Verheftungstechnik zum simultanen Behandeln der Restriktionen verwendet werden, zum anderen ist ein separates Behandeln möglich, wenn die Funktion stückweise definiert wird. Statt mit einer Restriktion der Art

$$f(x) \leq \begin{cases} 3 & \text{für } (0,1] \\ 5 & \text{für } (1,4] \\ 9 & \text{für } (4,7], \end{cases} \tag{3.16}$$

wird dann mit der Restriktion

$$f(x) \leq 3 + \underbrace{(5-3)}_{=2} \cdot 1(x-1) + \underbrace{(9-5)}_{=4} \cdot 1(x-4) \quad x \in (0,7] \tag{3.17}$$

gearbeitet. Alternativ wird $f(x)$ stückweise definiert, und es ergibt sich

$$f_{\text{opt}}(x) = \begin{cases} f_1(x, \theta_{1,\text{opt}}) & \text{für } (0,1] \\ f_2(x, \theta_{2,\text{opt}}) & \text{für } (1,4] \\ f_3(x, \theta_{3,\text{opt}}) & \text{für } (4,7], \end{cases} \tag{3.18}$$

womit aber dreimal ein Problem für ein Intervall zu lösen ist.

Eine Standardtechnik zur Behandlung der unendlich vielen Restriktionen (3.15) ist die Rückführung auf eine endliche Anzahl punktweiser Restriktionen

$$a_k + \varepsilon_{a_k} \leq f^{(k)}(z_i; \theta) \leq b_k - \varepsilon_{b_k}; i = 1, \ldots, M; \varepsilon_{a_k}, \varepsilon_{b_k} > 0 \text{ klein bez. } |a_k| \text{ bzw. } |b_k|, \tag{3.19}$$

wobei die $\varepsilon_{a_k}, \varepsilon_{b_k}$ gewissermaßen als Sicherheitspolster dienen.

[6] Solche Probleme entstehen, wenn möglichst große Teilsegmente gleicher Monotonie oder Krümmung angestrebt werden, s. [485] für adaptive monotone Spline-Approximationen.

Die z_i können hierbei mit den Stützstellen x_i der Messwertpaare (x_i, y_i) übereinstimmen, sie können aber auch nach einem beliebigen Schema (meist äquidistant) ausgewählt werden. Letzteres Vorgehen hat den Charme, dass die Restriktionen auch für Bereiche gelten, in denen wenige Messwerte vorliegen. Ferner kann die Dichte der z_i höher gewählt werden als die der x_i. Bei hinreichend hoher Dichte der z_i werden die unendlich vielen Restriktionen praktisch immer eingehalten. Allerdings sind für n-variates x und jeweils L Stützstellen in jeder Koordinate immerhin $M = L^n$ Restriktionen zu behandeln. Die Wahl parameterlinearer Modellansätze (globale Polynome, B-Splines usw.) garantiert dann wenigstens durchweg lineare Restriktionen für alle Ableitungen, z. B.

$$f(z_i) = \theta_0 + \theta_1 z_i + \theta_2 z_i^2 \geq 0;\ i = 1, \ldots, L \qquad \text{nichtnegative Funktion} \tag{3.20a}$$

$$f'(z_i) = \theta_1 + \theta_2 2 z_i \geq 0;\ i = 1, \ldots, L \qquad \text{monoton wachsende Funktion.} \tag{3.20b}$$

Die Ableitungen $f^{(k)}(x)$ werden dabei bei parametrischen und semiparametrischen Modellen durch die Ableitungen an den Stützstellen $f^{(k)}(z_i)$ ersetzt, während bei nichtparametrischen Modellen numerische Ableitungsformeln zum Einsatz kommen, die ihrerseits lineare Funktionen der zu bestimmenden Stützwerte sind. Für äquidistante Stützstellen sind numerische Ableitungsformeln in [101] zu finden. Eine Anwendung der beschriebenen Technik für parametrische Modelle zeigt das folgende Beispiel, während für nichtparametrische Modelle die Beispiele 3.12 und 3.13 das Vorgehen verdeutlichen.

Beispiel 3.7 (Polynomansatz für Antriebskennfeld)
Zur Längsregelung von Fahrzeugen mit der Beschleunigung als Stellgröße ist ein Kompensator der Nichtlinearitäten des Antriebskennfelds erforderlich. Dazu muss das Kennfeld invertiert werden, was bei Monotonie gelingt. Die Monotonie ist auch durch die Physik gerechtfertigt, denn bei konstanter Drehzahl bedingt eine größere Gaspedalstellung stets eine größere Beschleunigung. Es bezeichne x_1 die Motordrehzahl, x_2 die Gaspedalstellung und y die Beschleunigung. Eigene empirische Untersuchungen zur Wahl des Modellansatzes zeigten, dass ein vollständiges bivariates Polynom dritten Grads ergänzt um zwei Terme in x_2

$$\sum_{k=0}^{3} \sum_{l=0}^{3-k} \theta_{k,l} x_1^k x_2^l + \theta_{4,0} x_2^4 + \theta_{5,0} x_2^5 = y \tag{3.21}$$

den besten Kompromiss aus Parameteranzahl und Approximationsfehler liefert. Monotonie bezüglich x_2 kann durch nichtnegative partielle Ableitungen in x_2 an allen Messwertpaaren $(x_{1,i}, x_{2,i})$ erreicht werden, sodass sich $i = 1, \ldots, N$ lineare Ungleichungen ergeben

$$\frac{\partial y}{\partial x_2}(x_{1,i}, x_{2,i}) = \sum_{k=0}^{3} \sum_{l=0}^{3-k} \theta_{k,l} (l x_{1,i}^k x_{2,i}^{l-1}) + \theta_{4,0} (4 x_{2,i}^3) + \theta_{5,0} (5 x_{2,i}^4) \geq 0. \tag{3.22}$$

Statt einer gewöhnlichen LS ist also das folgende LSI-Problem zu lösen

$$\sum_{i=1}^{N} \left(y_i - \sum_{k=0}^{3} \sum_{l=0}^{3-k} \theta_{k,l} x_{1_i}^k x_{2,i}^l + \theta_{4,0} x_{2,i}^4 + \theta_{5,0} x_{2,i}^5 \right)^2 \stackrel{!}{=} \text{Min} \qquad \text{unter (3.22).} \tag{3.23}$$

Unter realen Messdaten aus Fahrversuchen mit einem VW Passat zeigten der Modellansatz ohne Restriktionen und auch andere empirische Ansätze ohne Restriktionen nicht die geforderte Monotonie. Mit Restriktionen hingegen gab es keine Probleme bei der Kennlinieninvertierung, die dann auch im autonomen Fahrzeug „AnnieWAY“ eingesetzt wurde [626], s. Bild 3.1 für die erreichte Approximation.[7]

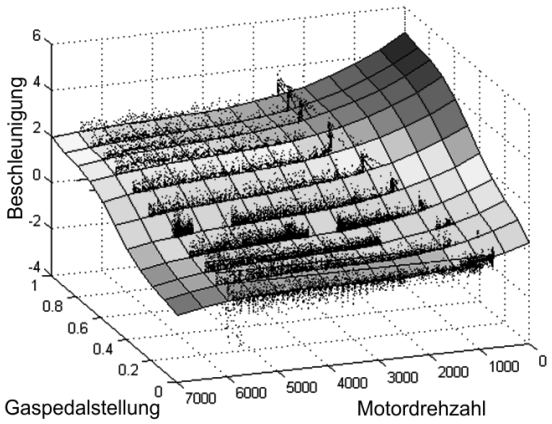

Bild 3.1: Approgression eines Antriebskennfelds

Der beschriebene Zugang über Rückführung auf endlich viele punktweise Restriktionen ist relativ einfach, weist aber einige Nachteile auf. Bei multivariaten Funktionen erhöht sich rasch die Anzahl der Restriktionen. Überdies muss a posteriori geprüft werden, ob die Restriktion eingehalten wird, was in $\mathbb{R}$ und $\mathbb{R}^2$ grafisch erfolgen kann, in $\mathbb{R}^n$ für Nichtnegativität aber schwierig ist. Auch der Zugang über eine Interpolation eines nichtparametrischen Zwischenmodells führt häufig auf eine ungewollt hohe Anzahl stückweise definierter Abschnitte. Deshalb sind Zugänge eine Alternative, die die Einhaltung der intervallweisen Restriktionen über eine Kombination aus speziellem Ansatz und/oder Parameterrestriktionen a priori sicherstellen. Die Zugänge überführen die unendlich vielen Restriktionen, die durch die $x \in [\underline{x}, \overline{x}]$ generiert werden, in endlich viele Restriktionen an die Parameter oder gar in ein vollständig restriktionsfreies Problem. Da die Vielzahl der Anwendungen, Restriktionen und Randbedingungen (Erfahrung, verfügbare Software, Rechenleistung PC oder μ-Controller, etc.) aber keine allgemeinen Regeln zulässt, wann welcher Ansatz optimal ist, werden in den folgenden Abschnitten die Grundideen beispielhaft skizziert. Ihre Zuordnung richtet sich nach dem Restriktionstyp, also Nichtnegativität, Monotonie oder Konvexität.

[7] Entnommen aus Diplomarbeit von H. Hynkova [306], die in der Arbeitsgruppe des Autors entstand.

3.4.1 Nichtnegativitätsrestriktionen

In diesem Abschnitt wird ausschließlich die Umsetzung der intervallweise gültigen Nichtnegativitätsforderung für parametrische Modelle untersucht, da nichtparametrische Modelle trivialerweise auf die Stützwertrestriktionen $f_i \geq 0$ führen. Im Speziellen wird die Behandlung durch elementare Transformationen kurz vorgestellt, während im Anschluss ausführlicher gezeigt wird, wie für ein Polynom, dass global oder auf einem Intervall definiert wird, die Nichtnegativität sichergestellt werden kann.

1. Behandlung durch elementare Transformationen
Transformationen zur Elimination von Ungleichungsrestriktionen, s. Tabelle 5.1, lassen sich auch zum Beseitigen der unendlichdimensionalen intervallweisen Restriktionen nutzen. So garantiert

$$f(x) = \mathrm{e}^{g(x;\theta)} > 0 \qquad \forall x \in \mathbb{R}^n \text{ und } g \text{ beliebig wählbar} \qquad \text{Positivität} \tag{3.24a}$$

$$f(x) = [g(x;\theta)]^2 \geq 0 \qquad \forall x \in \mathbb{R}^n \text{ und } g \text{ beliebig wählbar} \qquad \text{Nichtnegativität.} \tag{3.24b}$$

Um die entstehenden Optimierungsprobleme handhabbar zu machen, wird der Funktionenraum von g auf eine d-parametrische Familie von Funktionen, oft einen d-dimensionalen Vektorraum, eingeschränkt (orthogonales Funktionensystem, Splines, KNN-basierte Funktionen). Das semiunendliche Optimierungsproblem reduziert sich so auf ein d-parametrisches Problem.

Anmerkung 3.1 Dichtefunktionen f müssen nicht nur nichtnegativ sein, sondern auch die Normierungsbedingung (Integral über f ergibt 1) einhalten. Dies bewerkstelligt

$$f(x) = \frac{\mathrm{e}^{g(x)}}{\int_{\mathcal{X}} \mathrm{e}^{g(x)}\,\mathrm{d}x} \qquad \mathcal{X} \text{ ist Definitionsbereich von } f, \tag{3.25}$$

wobei g ebenfalls $\mathcal{X}$ als Definitionsbereich haben muss. Der Wert Null wird dabei über $\mathcal{X}$ nicht angenommen, weshalb diese Transformation nicht für alle Dichteprobleme optimal ist. Ferner ist die Transformation nicht bijektiv, denn $g_1(x) = g(x) + c$ impliziert $f(x) = \mathrm{e}^{g_1} / \int \mathrm{e}^{g_1} = \mathrm{e}^{g} / \int \mathrm{e}^{g}$. Die Hinzunahme einer Restriktion wie $g(x_0) = 0$; $x_0 \in \mathcal{X}$ bzw. $\int_{\mathcal{X}} g = 0$ erzwingt aber Bijektivität. Indem wiederum g aus einer d-parametrischen Familie von Funktionen gewählt wird, kann das Problem auf ein endlichdimensionales zurückgeführt werden [164].

Da an keiner Stelle $g(x;\theta)$ eingeschränkt wird, kann jede positive (nichtnegative) Funktion $f(x)$ so erzeugt werden und das immerhin für multivariate Funktionen. Es muss allerdings eine für die Aufgabenstellung angepasste Funktionenklasse für g gewählt werden. Bei ungünstiger Wahl führt das zu hohen Approximationsordnungen und vielen Parametern. Wegen

der nichtlinearen Transformationen in (3.24) entstehen parameternichtlineare Modelle, die gemeinhin auf nichtkonvexe Optimierungsprobleme führen. Das ist der Preis der hohen Flexibilität dieses Zugangs.

Weniger Flexibilität wie der beschriebene Zugang bieten Polynomansätze. Sie haben aber den Vorteil, dass bei derselben Zahl von Freiheitsgraden ihre Graphen gleichmäßiger krümmen. Hierfür wird auch eine nicht so gute Anpassung in Kauf genommen. Außerdem erübrigen sich in der Anwendung Fallunterscheidungen über die lokal wirksamen Approximationsfunktionen. Damit vereinfacht sich das Ableiten und ganz allgemein das Weiterrechnen etwa für eine Regelkreisanalyse bzw. -synthese. Um diese Vorteile nutzen zu können, wird nachfolgend die Vorgehensweise zur Sicherung der Nichtnegativität von Polynomen beschrieben, wobei einige begriffliche und theoretische Aspekte zuvor bearbeitet werden müssen.

2. Behandlung für globalgültige Polynome
Die globalen n-variaten Polynome bis zum Grad d werden beschrieben durch

$$\mathcal{P}_{n,d} = \Big\{ p(x) = \sum_{i=1}^{\binom{n+d}{d}} p_i x_1^{\alpha_{i1}} x_2^{\alpha_{i2}} \cdots x_n^{\alpha_{in}}; \alpha_{ij} \in \mathbb{N}_0, \sum_{j=1}^{n} \alpha_{ij} \leq d \Big\}. \tag{3.26}$$

In der Menge $\mathcal{P}_{n,d}$ liegen die positiv semidefiniten Polynomen, die sich genau durch die gewünschte Nichtnegativitäteigenschaft auszeichnen. Allerdings sind sie schwierig zu handhaben, wie nachfolgende Überlegungen zeigen.

Definition 3.1 (psd-Polynom)
Ein Polynom $p(x)$ heißt psd-Polynom (psd steht für „positiv semidefinit“), wenn $p(x) \geq 0$ für alle $x \in \mathbb{R}^n$ gilt. Die Menge der psd-Polynome vom Grad d wird mit $\mathcal{P}_{n,d}^{\geq}$ bezeichnet.

Eine Restriktion an das Minimum $\min_x p(x) \geq 0$ sichert zwar die psd-Eigenschaft, doch scheitert ein solcher Zugang an der Existenz geschlossener Formeln zur Nullstellenberechnung für höhergradige Polynome im univariaten Fall und multivariat erst recht. Klar ist nur, dass für $\mathcal{P}_{n,d}^{\geq}$ der Grad d stets gerade sein muss, da sich bei ungerader Potenz für hinreichend betragsgroße Argumente in einer Variablen negative Funktionswerte ergeben.
Auch der Zugang über Hilberts 17. Problem[8], wonach ein Polynom genau dann ein psd-Polynom ist, wenn es sich als Summe von Quadraten rationaler Funktionen darstellen lässt, ist schwierig, da die Klasse der Polynome verlassen wird. Einen Ausweg zur Parametercharakterisierung von $\mathcal{P}_{n,d}^{\geq}$ bieten die sos-Polynome, die ihrer Struktur wegen stets psd-Polynome sein müssen.

[8] Eine Übersicht zu allen Problemen findet sich in http://en.wikipedia.org/wiki/Hilbert%27s_problems.

Definition 3.2 (sos-Polynom)
Ein psd-Polynom $p(x)$ vom Grad $d \in \{2m : m \in \mathbb{N}\}$ heißt sos-Polynom (sum of squares of polynomials), wenn es sich als Summe von Quadraten aus Polynomen q_i schreiben lässt

$$p(x) = \sum_{i=1}^{r} q_i^2(x_1, \ldots, x_n) \qquad \deg q_i \leq d/2. \tag{3.27}$$

Die Menge der sos-Polynome wird durch $\Sigma_{n,d}^2$ bezeichnet.[9]

Beispiel 3.8 (psd- und sos-Polynom)
$p(x_1, x_2) = x_1^2x_2^4 + x_1^4x_2^2 - 3x_1^2x_2^2 + 1$ (Motzkin-Polynom [464]) ist ein psd-Polynom, aber kein sos-Polynom. Um das zu zeigen, wird p mit einem positiven Faktor multipliziert, sodass ein sos-Polynom aus $\Sigma_{2,8}^2$ folgt

$$\begin{aligned}(x_1^2 + x_2^2 + 1) \cdot p(x_1, x_2) =& (x_1^2x_2 - x_2)^2 + (x_1x_2^2 - x_1)^2 + (x_1^2x_2^2 - 1)^2 \\ &+ \frac{1}{4}(x_1x_2^3 - x_1^3x_2)^2 + \frac{3}{4}(x_1x_2^3 + x_1^3x_2 - 2x_1x_2)^2.\end{aligned}$$

Wegen der Quadrate auf der rechten Seite folgt sofort $p \in \mathcal{P}_{2,6}^{\geq}$, während $p \notin \Sigma_{2,6}^2$ aus der Indefinitheit der korrespondierende Matrix S in Satz 3.1 folgt, die ihrerseits durch das negative Diagonalelement -3 aus dem Term $-3(x_1x_2)^2$ resultiert.

Der Vorteil der sos-Polynomen begründet sich durch eine notwendige und hinreichende Bedingung für die sos-Eigenschaft. Der folgende Satz stellt nämlich eine Beschreibung der sos-Polynome als quadratische Form vor, die ihrerseits mit der Existenz einer nichtnegativ definiten Matrix einen LMI-Zulässigkeitstest gestattet.

Satz 3.1 (sos-Polynomdarstellung als „quadratische Form“, [515])
$p \in \Sigma_{n,2m}^2$ gilt genau dann, wenn eine Matrix $S \succeq 0_{\binom{n+m}{m} \times \binom{n+m}{m}}$ existiert, sodass

$$p(x_1, \ldots, x_n) = [1, x_1, x_2, \ldots, x_1^2, x_1x_2, \ldots, x_n^m] S [1, x_1, x_2, \ldots, x_1^2, x_1x_2, \ldots, x_n^m]^T. \tag{3.28}$$

Der Rang von S gibt die Anzahl r der benötigten Quadrate in (3.27) an.

Zu klären ist nunmehr, wann die Mengen der sos-Polynome und der psd-Polynome gleich sind. Die Antwort gibt der folgende Satz.

Satz 3.2 (Hilbert-Theorem für psd-Polynome, [286])
$\Sigma_{n,d}^2 = \mathcal{P}_{n,d}^{\geq}$ gilt nur für $n = 1$ (univariate psd-Polynome), für $d = 2$ (quadratische psd-Polynome) oder für $n = 2$ und $d = 4$ (quartische bivariate psd-Polynome).

[9] Im Fall homogener Polynome vom Grad d lässt sich zeigen, dass alle q_i homogen vom Grad $d/2$ sein müssen. Zudem formt $\Sigma_{n,d}^2$ einen abgeschlossenen konvexen Kegel.

Die ernüchternde Aussage des Satzes ist, dass nur in wenigen Fällen die sos-Polynome die psd-Polynome äquivalent beschreiben. Im univariaten Fall ermöglicht der Satz aber, zum unendlichdimensionalen Nichtnegativitätsproblem ein äquivalentes konvexes endlichdimensionales Problem zu formulieren. Genauer: Da zwischen den Polynomkoeffizienten p_i und den Matrixelementen s_{kl} affine Zusammenhänge über (3.28) bestehen, reduziert sich die Suche über den sos-Polynomen auf eine Suche im Schnitt der nichtnegativ definiten Matrizen mit einer affinen Menge von Matrizen. Für $p \in \Sigma^2_{1,2m}$ heißt das

$$S \in \mathcal{S}^{\succeq}_{m+1} \tag{3.29a}$$

$$p_i = \sum_{k+l=i+2} s_{kl} = \text{spur}(SH_{i+1}); \quad i = 0, 1, \ldots, 2m, \tag{3.29b}$$

wobei $H_k \in \mathbb{R}^{(m+1)\times(m+1)}$ die k-te Hankel-Basismatrix[10] ist.

Beispiel 3.9 (LS-Formulierung für Polynome mit Nichtnegativitätsrestriktion)
Mit den obigen Ausführungen lautet das LS-Ausgleichsproblem für N Tupel (x_i, y_i):

$$\left\| \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} 1 & x_1 & \ldots & x_1^{2m} \\ \vdots & \vdots & & \vdots \\ 1 & x_N & \ldots & x_N^{2m} \end{bmatrix} \begin{bmatrix} p_0 \\ \vdots \\ p_{2m} \end{bmatrix} \right\|_2^2 \stackrel{!}{=} \text{Min} \quad \begin{array}{l} p_i = \text{spur}(SH_{i+1}); i = 0, \ldots, 2m \\ S \succeq 0_{(m+1)\times(m+1)}. \end{array} \tag{3.30}$$

Nach der Behandlung globalgültiger Polynome soll diese nun für intervallgültige Polynome beschrieben werden. Sie baut auf den bisherigen Ergebnissen auf, ist aber schwieriger, da ungerade Polynomgrade nicht mehr von vornherein auszuschließen sind. Lediglich für den univariaten Fall lassen sich praktikable Aussagen formulieren.

3. Behandlung für intervallgültige Polynome
Aus der Vielzahl von Sätzen hierfür [514], sollen drei Sätze vorgestellt werden, die die Grundlage für eine algorithmische Umsetzung liefern. Das Lukács-Theorem bietet sich beispielsweise für abgeschlossene Intervalle an.

Satz 3.3 (Lukács-Theorem, [511])
Sei p ein psd-Polynom auf $[a, b]$ vom Grad d, dann gilt

$$p(x) = \begin{cases} f_1^2(x) + (x-a)(b-x)f_2^2(x) & d \text{ gerade} \\ (x-a)f_3^2(x) + (b-x)f_4^2(x) & d \text{ ungerade} \end{cases} \tag{3.31}$$

für reelle Polynome $f_1, \ldots, f_4$, wobei der Grad der Summanden d nicht übersteigt, aber d wenigstens einmal annimmt.

[10] Hankel-Basismatrizen formen die kanonische Basis für den Vektorraum der Hankel-Matrizen. Bei der k-ten Basismatrix ist die k-te Nebendiagonale mit Einsen besetzt, während der Rest Nullen enthält. Beispiele: $H_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, $H_2 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, $H_3 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$, $H_4 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$, $H_5 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

Anmerkung 3.2 Im Original bezieht sich der Satz auf $[-1, 1]$, woraus diese Version per Koordinatentransformation folgt. Für gerade Ordnung ist der Satz auch unter Fekete-Theorem bekannt. Verfeinerungen liefert das Karlin-Shapley-Theorem [514].

Das Lukács-Theorem ist interessant, da es eine Reparametrisierung des Suchraums

$$(p_0, \ldots, p_d) \in \mathbb{R}^{d+1} : \; p(x) = p_d x^d + \ldots + p_1 x + p_0 \geq 0 \; \forall x \in [a, b] \tag{3.32}$$

hin zu einem freien Suchraum aufzeigt

$$(f_{1,0}, \ldots, f_{1,d/2}, f_{2,0}, \ldots, f_{2,d/2-1})^T \in \mathbb{R}^{d+1} \qquad d \text{ gerade} \tag{3.33a}$$

$$(f_{3,0}, \ldots, f_{3,(d-1)/2}, f_{4,0}, \ldots, f_{4,(d-1)/2})^T \in \mathbb{R}^{d+1} \qquad d \text{ ungerade.} \tag{3.33b}$$

Nachteilig ist, dass wegen der Quadrierung der f_i-Polynome nichtkonvexe Probleme entstehen. Eine Mehrdeutigkeit in den f-Parametern bedeutet dabei keineswegs eine Mehrdeutigkeit in den p-Parametern. Da die Menge der auf $[a, b]$ nichtnegativen Funktionen nämlich konvex ist, wird bei konvexen Zielfunktionen und problemangepasste Experimentationsdaten die Lösung bezüglich der p-Parameter eindeutig sein.

Eine Alternative zum auf Suchverfahren zugeschnittenen Ansatz, der auf einer Reparametrisierung in ein freies Problem beruht, stellt eine SDP-Formulierung mit LMI-Restriktionen dar. Das folgende Beispiel zeigt, welche LMI-Restriktionen dies bewerkstelligen. Es kombiniert die Koeffizientenbedingung und die Aussage des Lukács-Theorems. Zugleich zeigt es, warum der unterschiedliche Funktionsansatz für gerade und ungerades d in (3.31) auf Ebene der LMI-Restriktionen nicht mehr in Erscheinung tritt.

Beispiel 3.10 (LMI-Restriktionen für psd-Polynome auf Intervallen)
Für ungerades d impliziert $p = (x-a)f_3^2(x) + (b-x)f_4^2(x)$ im Lukács-Theorem, dass $p \in (x-a)\Sigma_{1,m}^2 + (b-x)\Sigma_{1,m}^2$ gilt. Für gerades d folgt aus $(b-a)p = (x-a)p + (b-x)p = (x-a)(f_1^2 + (b-x)^2 f_2^2) + (b-x)(f_1^2 + (x-a)f_2^2)$ ebenfalls $p \in (x-a)\Sigma_{1,m}^2 + (b-x)\Sigma_{1,m}^2$, denn schließlich sind $(f_1^2 + (b-x)^2 f_2^2)/(b-a)$ und $(f_1^2 + (x-a)f_2^2)$ Elemente aus $\Sigma_{1,m}^2$. Mit den entsprechenden Restriktionen für $\Sigma_{1,m}^2$ und den Koeffizientenbeziehungen entsprechend (3.29) ergibt sich

$$S_1, S_2 \in \mathcal{S}_{m+1}^{\geq}; p_i = \text{spur}\Big(S_1(H_i - aH_{i+1})\Big) + \text{spur}\Big(S_2(bH_{i+1} - H_i)\Big). \tag{3.34}$$

p ist unter diesen Restriktionen ein psd-Polynom auf $[a, b]$ vom Grade $\leq d$, wenn bei $m = d/2$ (d gerade) bzw. $m = (d-1)/2$ (d ungerade) verwendet wird.

Für halboffene Intervalle greift das Pólya-Szegö-Theorem.

Satz 3.4 (Pólya-Szegö-Theorem, [511], [514])
Sei p ein psd-Polynom vom Grad d auf $[a, \infty)$ bzw. $(-\infty, b]$, dann existieren globale sos-Polynome f_1, f_2, sodass

$$p(x) = f_1(x) + (x-a)f_2(x) \qquad \text{bzw.} \qquad p(x) = f_1(x) + (b-x)f_2(x), \tag{3.35}$$

wobei der Grad der Summanden d nicht übersteigt, aber wenigstens einmal erreicht.

Die Forderung, wonach der Grad wenigstens einmal erreicht werden muss, lässt zu, dass $f_1(x)$ nicht den Grad d oder $f_2(x)$ den Grad $d-1$ haben muss, wenn das jeweils andere Polynom den maximal zulässigen Grad hat. Da aber die Mengen von Polynomen kleinerer Grade mager sind, ist es praktisch legitim, $p \in \Sigma^2_{1,d} + (x-a)\Sigma^2_{1,d-1}$ zu fordern, was auf die folgenden Restriktionen führt

$$S_1 \in \mathcal{S}^{\geq}_{d/2}, S_2 \in \mathcal{S}^{\geq}_{d/2-1}, \quad p_i = \text{spur}(S_1 H_k) + \text{spur}\Big(S_2(H_k - aH_{k+1})\Big) \quad d \text{ gerade} \tag{3.36a}$$

$$S_1, S_2 \in \mathcal{S}^{\geq}_{(d-1)/2}, \qquad p_i = \text{spur}\Big((S_1+S_2)H_k - aS_2H_{k+1}\Big) \qquad d \text{ ungerade.} \tag{3.36b}$$

Anmerkung 3.3 Neben Intervallen kann die Nichtnegativität auch auf dem Komplement eines endlichen Intervalls [157] oder über speziellen konvexen Mengen betrachtet werden, was zum Beispiel im Rahmen der Stabilitätstheorie interessant ist. Zweckmäßig zu behandelnde Mengen sind dabei Simplexe, Ellipsen, konvexe Polytope [130].

Für den Ingenieur gilt es insgesamt festzuhalten, dass die Grundidee in der Formulierung von Ansätzen liegt, die sos-Polynome nutzen. Über diesen Weg können dann freie, allerdings nichtkonvexe Optimierungsprobleme abgeleitet werden oder aber konvexe Optimierungsprobleme mit LMI-Restriktionen. Letztere sind zu bevorzugen, setzen aber Kenntnisse der semidefiniten Optimierung voraus, s. Abschnitt 7.8.

3.4.2 Monotonierestriktionen

Vielen Modellen ist die Monotonieeigenschaft per Definition und der Physik heraus eigen, z. B. Verteilungsfunktion, Kurven von Verstärkern, oder sie ergibt sich aus logischen Überlegungen, z. B. beim Längenwachstum von Kindern. Ähnliches gilt für monoton fallende Funktionen, z. B. die Fehleranzahl bei Softwaretests. In der Regelungstechnik ist strenge Monotonie einer stetigen statischen Abbildung eine oft gestellte hinreichende Voraussetzung für Regelkreisstabilität. Ob ein Signal bzw. eine Funktion dabei monoton fällt oder wächst, ändert nichts an der prinzipiellen mathematischen Behandlung.

Monoton wachsende Funktionen[11] $f : \mathcal{X} \subseteq \mathbb{R} \to \mathbb{R}$ lassen sich charakterisieren durch

$$\forall x_1, x_2 \in \mathcal{X} : x_1 \leq x_2 \Rightarrow f(x_1) \leq f(x_2) \tag{3.37a}$$

$$\text{äquivalent: } (f(x_1) - f(x_2))(x_1 - x_2) \geq 0 \tag{3.37b}$$

und äquivalent für stückweise stetig differenzierbare Funktionen durch

$$f'(x) \geq 0 \qquad \text{für alle } x \in \mathcal{X} \text{ bis auf isolierte Punkte.} \tag{3.38}$$

Mitunter braucht die Monotonie nicht gesondert behandelt werden, wie etwa bei nichtparametrischen Maximum-Likelihood-Schätzern, kurz ML-Schätzern, von Verteilungsfunktionen, da diese per Konstruktion (Summe der relativen Häufigkeiten) eine monotone Stützwertfolge liefern. Anders verhält es sich bei ML-Schätzern für die Verteilungsdichte über ein Histogramm. Dann ist beispielsweise für die Exponentialverteilung keineswegs eine monotone Folge von Schätzwerten garantiert. Für eine ausführliche Bearbeitung der Monotoniesicherung bei Dichteschätzern einschließlich der Normierungsbedingung sei auf [50], [245], [537] verwiesen. Im Folgenden werden sechs Möglichkeiten zur Behandlung der Monotonierestriktion vorgestellt.

1. Behandlung mittels Integralansatz
Jede stetig differenzierbare streng monotone Funktion auf $[0, 1]$ kann durch[12]

$$f(x) = c_0 + c_1 \int_0^x \mathrm{e}^{\int_0^\xi g(\tau;\theta)\mathrm{d}\tau}\mathrm{d}\xi \quad \text{mit} \quad \int_0^x g^2(\xi;\theta)\mathrm{d}\xi < \infty \tag{3.39}$$

dargestellt werden, wobei für g ein gängiger Funktionsansatz ohne Restriktionen verwendet werden kann (erfüllt zwangsläufig die quadratische Integrabilitätsbedingung). Die Idee hinter (3.39) ist sehr anschaulich: Da die e-Funktion unabhängig vom Exponenten stets positiv ist, kann das äußere Integral nur wachsen. Details zu diesem Ansatz finden sich in [524], wo zudem durch Hinzunahme des Terms $\int_0^x g^2(\xi)\mathrm{d}\xi$ die relative Krümmung von y bestraft wird.

[11] Die monotonen Funktionen formen einen Kegel. Sie müssen nicht stetig sein; haben aber höchstens abzählbar viele Unstetigkeiten erster Art (Sprünge). Zwischen den monotonen und injektiven Funktionen besteht der Zusammenhang: Eine stetige injektive Funktion $f : \mathbb{R} \to \mathbb{R}$ ist streng monoton. Für unstetige injektive Funktionen gilt das nicht, vgl. $f(x) = \begin{cases} x & \text{für } x \in \mathbb{R}\backslash\mathbb{Z} \\ 2x & \text{für } x \in \mathbb{Z} \end{cases}$ oder $f(x) = \begin{cases} x & \text{für } x \in \mathbb{R}\backslash(0,1) \\ 1-x & \text{für } x \in (0,1). \end{cases}$
Umgekehrt ist jede streng monotone Funktion $f : \mathbb{R} \to \mathbb{R}$ injektiv.
Eine Funktion kann nirgends monoton sein [214].

[12] (3.39) ist die Lösung der Differenzialgleichung $f'' = gf'$, die diese Funktionenfamilie umfasst.

2. Behandlung mit speziellen Ansatzfunktionen
Eine Vielzahl von Funktionen (Logarithmus, Exponentialfunktion, lineare Funktion, Potenzfunktion) sind global oder auf großen Intervallen monoton. Ob sie dabei wachsend oder fallend sind, hängt dabei nur von einem oder wenigen Parametern ab. Da die Funktionen die Daten gut anpassen sollen und somit geeignet gewählt werden, kann auf die Restriktionen an die Parameter meist verzichtet werden. Die Schwierigkeit besteht dann also darin, eine geeignete Funktionsklasse zu wählen, wenn diese nicht durch A-priori-Wissen aus dem technischen Hintergrund bekannt ist. Da solche Ansatzfunktionen gemeinhin mit wenigen Parametern auskommen, spielt ein parameternichtlinearer Zusammenhang nur eine untergeordnete Rolle. Insbesondere schnelle Löser für nichtlineare LS-Formulierungen sind in den gängigen Programmsystemen vorhanden.

Neben der Monotonie weisen viele Kennlinien aus der Mechanik, Pneumatik, Hydraulik und Thermodynamik zudem einen gesättigten Anstieg aus, vgl. Bild 3.2. Die Kennlinien sind monoton wachsend mit abflachendem Anstieg, d. h. die zweite Ableitung geht gegen Null. Folglich ergibt sich mit wachsender x-Koordinate eine Asymptote (gesättigter Anstieg). Gegebenenfalls gelten diese Eigenschaften für die Umkehrfunktion, was in der weiteren Betrachtung nur die Rolle von (x, y) tauscht.

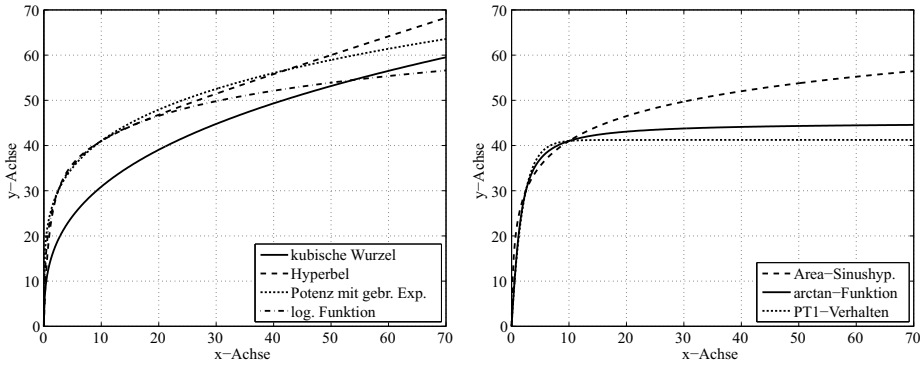

Bild 3.2: Monotone Kennlinien mit gesättigtem Anstieg, [85]

Prinzipiell könnten die Kennlinien durch Polynome oder Splines mit entsprechender Formulierung der Restriktionen genähert werden. Nachteilig daran ist aber, dass das Invertieren von Polynomen gemeinhin nicht geschlossen möglich ist (Nullstellenproblem). Zudem erweist sich die Behandlung polynomialer Kennlinien im Absolutglied von Differenzialgleichungen als recht knifflig, oder sie ist gar unmöglich. Deshalb seien hier einige Ersatzfunktionen genannt, die praktische Anwendung bei nichtlinearen Widerstandskennlinien sowie magnetischen und dielektrischen Kennlinien finden, vgl. Tabelle 3.1.

Ersatzfunktion	Formel
spezielle kubische Funktion	$\frac{x}{x_b} = \frac{y}{y_b} + \left(\frac{y}{y_b}\right)^3$
Hyperbeln	$y = \frac{ax}{b+x} + cx$
Potenz mit gebrochenem Exponent	$y = kx^\beta; \beta > 0$
logarithmische Funktion	$\frac{y}{y_b} = \ln\left(1 + \frac{x}{x_b}\right)$
Area-Sinus-Hyperbolicus	$\frac{y}{y_b} = \operatorname{arsinh}\frac{x}{x_b}$
Arkus-Tangens-Funktion	$\frac{y}{y_b} = \arctan\frac{x}{x_b}$

Tabelle 3.1: Monotone Kennlinien mit gesättigtem Anstieg, [85]

Zunächst erscheinen die Ansätze deutlich komplizierter; ihre Behandlung vereinfacht sich aber durch geeignete Substitutionen entsprechend. Das wird an der magnetischen Kennlinie einer Spule mit Eisenkern unter Vernachlässigung der Hysterese deutlich, die sich durch $\frac{\psi}{\psi_b} = \operatorname{arsinh}\frac{i}{i_b}$ ausdrücken lässt. Die Differenzialgleichung des Einschaltvorgangs an Wechselspannung kann nämlich durch Substitution in eine lineare Differenzialgleichung überführt und gelöst werden [86].
Welche der Ersatzfunktionen sich für den konkreten Fall am besten eignet, hängt von Einsatzfall und der erzielbaren Approximationsgüte ab. Ein Durchprobieren der Ersatzfunktionen gepaart mit einer Parameterschätzung stellt auf modernen Rechnern kein Problem dar. Dabei empfehlen sich Ausgangsfehler-LS-Ansätze, wobei eine höhere Fehlerwichtung im „Sättigungsbereich“ zweckmäßig ist, da er für die Anwendung oft von größerem Interesse ist.

3. Behandlung für intervall- und globalgültige Polynome
Für Monotonierestriktionen bei (globalen) Polynomapproximationen sei auf [334] verwiesen. Das Polynom hat bei globalen Approximationen dabei stets einen ungeraden Grad. Prinzipiell können die Zugänge aus Abschnitt 3.4.1 herangezogen werden, was über die Integration der Ansatzfunktionen für Nichtnegativitätsrestriktionen gelingt.

4. Behandlung von Interpolationsproblemen
Ausgleichsprobleme mit einer intervallgültigen Funktion (Polynom) haben zwar den Vorteil, dass Fallunterscheidungen gegenüber stückweise teilintervallbezogenen Funktionen entfallen, aber den Nachteil, dass die Stützstellen nicht exakt getroffen werden wie bei Interpolationsproblemen. Zunächst ist klar, dass die Monotonie der Stützwerte eine notwendige Voraussetzung für eine monotone Interpolation ist. Sie ist aber nicht hinreichend. Für globale Polynominterpolationen leuchtet das sofort ein, da das bekannte Problem der Überanpassung auf oszillierenden Graphen führen kann. Für die stückweise lineare Interpolation ist die

Monotonie der Stützwerte hinreichend, für kubische kubischen Spline-Interpolation hingegen wieder nicht. Im bivariaten Fall versagt die lineare Dreiecksinterpolation im Beispiel der im ersten Quadranten monotonen Funktion $f(x_1, x_2) = (x_1 x_2)^{1/4}$, deren Dreiecksinterpolation über $(0, 1; 0), (1, 0; 0), (1, 1; 1), (0.1, 0.1; 0.316)$ nicht monoton ist [58]. Aus diesen Erkenntnissen leitet sich die Notwendigkeit zusätzlicher Restriktionen für die Mehrzahl monotoner Interpolation ab. Im folgenden Beispiel wird gezeigt, wie das dann umgesetzt werden kann.

Beispiel 3.11 (Behandlung durch B-Spline-Approximation vom Grad k)
Gegeben seien Wertepaare (x_i, y_i); $i = 1, \ldots, N$ mit $x_1 \leq x_2 \leq \ldots \leq x_N$. Ferner seien $m + k + 2$ Knoten p_j mit $x_1 = p_0 \leq p_1 \leq \ldots \leq p_{m+k+1} = x_N$ für die B-Splines sowie die Randwerte $y(x_1) = a$ und $y(x_N) = b$ vorgegeben. Dem Ziel einer monotonen B-Spline-Approximation wird dann das linearrestringierte LS-Problem für die θ_j

$$\sum_{i=1}^{N} \left(y_i - \sum_{j=0}^{m+k+1} \theta_j B_{j,k}(x_i) \right)^2 \stackrel{!}{=} \text{Min} \qquad \begin{array}{l} \sum_{j=0}^{m+k+1} \theta_j B_{j,k}(x_1) = a \\ \sum_{j=0}^{m+k+1} \theta_j B_{j,k}(x_N) = b;\ \theta_j \leq \theta_{j+1} \end{array} \tag{3.40}$$

gerecht, wobei

$$B_{j,0}(x) = \begin{cases} 1 & p_j \leq x < p_{j+1} \\ 0 & \text{sonst} \end{cases} \tag{3.41}$$

$$B_{j,l}(x) = \frac{x - p_j}{p_{j+l} - p_j} B_{j,l-1}(x) + \frac{p_{j+l+1} - x}{p_{j+l+1} - p_{j+1}} B_{j+1,l-1}(x); \qquad 0/0 := 0 \tag{3.42}$$

eine rekursive Definition der normalisierten B-Splines nach Cox [148] und de Boor [88] ist. Die letzte Restriktion im Gütekriterium ist hinreichend; für notwendige Bedingungen s. [202]. Für eine monoton fallende Funktion ist $\theta_j \geq \theta_{j+1}$ zu fordern.

5. Behandlung im mehrvariablen Fall
Während Monotonie im einvariablen Fall eindeutig festgelegt ist, muss im mehrvariablen Fall zunächst immer genau definiert werden, in welcher Weise der Monotoniebegriff erweitert wird. Das ist erforderlich, da für Vektoren anders als Skalare keine Totalordnung naturgegeben ist, sondern gemeinhin nur Halb- oder Quasiordnungen betrachtet werden. Letztere müssen genau spezifiziert werden. Begriffe wie Quasi- oder Pseudomonotonie [266] oder Monotonie i. S. von Volkmann [616] sind die Folge, wobei der Zusatz „quasi“ nicht immer in gleicher Weise verwendet wird. Drei für den Ingenieur interessanten Varianten seien hier genannt.

Variante 1 nutzt die natürliche Halbordnung von Vektoren (s. Anhang A.1), indem sie fordert, dass für einen Punkt, der in all seinen Koordinaten größer ist als ein anderer auch der zugehörige Funktionswert größer ist, kurzum

$$\mathcal{G}_1 = \{ f : \mathbb{R}^n \to \mathbb{R} : x_i \leq x_j \Rightarrow f(x_i) \leq f(x_j);\ \forall i, j \}. \tag{3.43}$$

Anwendung findet dieser Zugang bei der monotonen Regression[13]. Ob eine multivariate Funktion dabei monoton steigt oder fällt oder nicht monoton ist, hängt von der Wahl des Koordinatensystems ab. So ist $f(x_1, x_2) = x_1 - x_2$ in den Basiskoordinaten weder monoton fallend noch wachsend, in den Koordinaten $x_1' = x_1$ und $x_2' = -x_2$ aber streng monoton wachsend.

Variante 2 verallgemeinert die Eigenschaft (3.37b) zu einer kumulierten Zuwachsbedingung mittels Skalarprodukt [487]

$$\mathcal{G}_2 = \{f : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}^n : (f(x_1) - f(x_2))^T (x_1 - x_2) \geq 0 \text{ für alle } x_1, x_2 \in \mathcal{X}\}. \tag{3.44}$$

Variante 3 ist für differenzierbare Funktionen geeignet und beruht auf der Verallgemeinerung der nichtnegativen Anstiege monoton steigender Funktionen zu

$$\mathcal{G}_3 = \left\{ f : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}^n : \frac{\partial f}{\partial x^T} + \frac{\partial f^T}{\partial x} \succeq 0_{n \times n} \text{ für alle } x \in \mathcal{X} \right\}. \tag{3.45}$$

Hieraus ergeben sich bei parametrischen Funktionsansätzen die entsprechenden Parameterrestriktionen. Derartige monotone Funktionen wie etwa $f(x_1, x_2) = (\frac{1}{3}x_1^3 + x_1 x_2^2; \frac{1}{3}x_2^3 + x_2 x_1^2)$ stellen diverse mehrvariable Sektorbedingungen für den Entwurf von Reglern und Beobachtern sicher [648].

6. Behandlung von nichtparametrischen Modellen
Für nichtparametrische Modelle können im mehrvariablen Fall die Varianten 1 und 2 direkt genutzt werden, während bei Variante 3 die Ableitungen zuvor durch numerische Approximationen zu ersetzen sind. Die Varianten führen bei LS-Ansätzen auf einfach zu behandelnde LSI-Probleme. An einem Beispiel für eine nichtparametrische monotone Regression für das Modell $\boldsymbol{y}_i = f(x_i) + \boldsymbol{\varepsilon}_i$; $i = 1, \ldots, N$ mit $x_i \in \mathbb{R}^n$ und identisch mittelwertfrei normalverteilten $\boldsymbol{\varepsilon}_i$ soll das gezeigt werden. Für statistische Betrachtungen sei auf [50], [151] verwiesen.

Beispiel 3.12 (LSI-Formulierung für monotone Regression)
Für $x_1 = (2, 1)$, $x_2 = (3, 3)$, $x_3 = (-1, -2)$, $x_4 = (-2, 1)$, $x_5 = (2, -1)$, $x_6 = (1, 3)$ seien die Messwerte $y_1, \ldots, y_6$ gegeben. Damit existieren die 11 Relationen (i.S. des elementweisen Vergleichs) $x_1 \leq x_2$; $x_3 \leq x_1, x_2, x_5, x_6$; $x_4 \leq x_1, x_2, x_6$; $x_5 \leq x_1, x_2$; $x_6 \leq x_2$, die gleichermaßen von den zu bestimmenden Stützwerte f_i einzuhalten sind. Wird ferner berücksichtigt, dass sich $f_i \leq f_j$ als $f_i - f_j \leq 0$ schreiben lässt, führt das monotone Regressionsproblem auf

[13] Eine Verallgemeinerung stellt die isotone Regression dar, bei der die natürliche Halbordnung zwischen den Vektoren x_i und x_j durch eine beliebige Quasiordnung ersetzt wird [50]. Es finden sich aber auch Definitionen für isotone Funktionen, die auf einer beliebigen oder einer speziellen Halbordnung [72] beruhen. Darüber hinaus lässt sich Monotonie auch für eine Teilmenge der Variablen definieren [58].

das folgende LSI-Problem

$$\left\| \begin{bmatrix} y_1 \\ \vdots \\ y_6 \end{bmatrix} - \begin{bmatrix} f_1 \\ \vdots \\ f_6 \end{bmatrix} \right\|_2^2 \stackrel{!}{=} \text{Min} \qquad \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \\ -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_1 \\ \vdots \\ f_6 \end{bmatrix} \leq 0_{11}. \tag{3.46}$$

Für die unterschiedlichen Möglichkeiten eine Funktion (Kennlinie) zu definieren, nämlich globalgültig, intervallgültig, abschnittsweise oder nichtparametrisch werden Techniken vorgestellt und diskutiert, mit denen die wichtige Eigenschaft der Monotonie sichergestellt werden kann. Erweiterungen für mehrvariable Funktionen werden ebenso diskutiert. Letztlich sei nochmals daran erinnert, dass die im übergeordneten Abschnitt 3.4 beschriebene Technik des Rückführens auf punktweise Restriktionen für parameterlineare Ansätzen sehr einfach, aber effektiv ist. Da die Behandlung von Konvexitätsrestriktionen mit ähnlichen Techniken erfolgt wie die der Monotonierestriktionen, fällt der nachfolgende Abschnitt kurz aus.

3.4.3 Konvexitätsrestriktionen

Konvexitätsrestriktionen und ihr Pendant die Konkavitätsrestriktionen lassen sich bei univariaten Funktionen leicht erkennen. Eine konvexe Funktion ist wie $y = x^2$ nach oben geöffnet, eine konkave wie $y = \sqrt{x}$ nach unten.
Handelt es sich bei der zu identifizierenden Funktion um eine Kennlinie, die in der Rückkopplung eines Regelkreises liegt, ist zu beachten, dass bei großen Reglerverstärkungen sich das Regelkreisverhalten (Führungsgröße-Regelgröße) statisch invers wie die Kennlinie verhält. Statt Konvexität ist dann Konkavität zu fordern. Ähnlich ist zu verfahren, wenn statt der konkaven Strecke der zugehörige vor- oder nachgeschaltete konvexe Kompensator geschätzt werden soll.
Konvexität/Konkavität können durch die zweiten Ableitungen oder durch monoton wachsende bzw. fallende erste Ableitungen spezifiziert werden. Dabei muss die Hesse-Matrix, die selbst eine Funktion der x-Vektoren ist, für alle zulässigen x nichtnegativ bzw. nichtpositiv definit sein. Vielfach wird diese unendlichdimensionale Restriktion auf punktweise Restriktionen heruntergebrochen (notwendige Bedingungen), oder es werden hinreichende parametrische Restriktionen abgeleitet. Einzig quadratische multivariate Funktionen sind einfacher zu behandeln, da bei diesen die Hesse-Matrix unabhängig von x ist.

Die Polygonzuginterpolation zählt zu den populärsten Methoden der Datenapproximation, da lediglich die Stützwerte durch Geradenstücke zu verbinden sind. Das erleichtert das Abspeichern und das Berechnen von Zwischenwerten. Wenn nun von einer Funktion bekannt ist, dass sich in einem Intervall deren Ableitung mit wachsenden x vergrößert (verkleinert), dann ist die Funktion konvex (konkav). Als Beispiel seien konvexe, monoton fallende Dichten genannt [246], für die geschätzte Häufigkeiten vorliegen. Andere Beispiele liefern einige Gewichts- oder Übergangsfolgen, die durch Messung oder Entfaltung entstehen. Aufgrund der Schätzungen oder Störungen wird die gewöhnliche Polygonzuginterpolation dann nicht die Konvexitätseigenschaft aufweisen. Durch Verschieben einzelner oder mehrere Stützwerte kann die Eigenschaft aber erzwungen werden, vgl. nachfolgendes Beispiel.

Beispiel 3.13 (Konvexe Regression mit stückweise linearen Funktionen)
Seien $(x_i, y_i); i = 1, \ldots, N$ disjunkte, geordnete Messpunkte mit $x_i < x_{i+1}$, dann sind die $\hat{f}_i := f_{i,\text{opt}}$ aus

$$\frac{1}{2}\sum_{i=1}^{N}(y_i - f_i)^2 \stackrel{!}{=} \text{Min} \qquad \frac{f_i - f_{i-1}}{x_i - x_{i-1}} \leq \frac{f_{i+1} - f_i}{x_{i+1} - x_i};\ i = 2, \ldots, N-1 \tag{3.47}$$

LS-Schätzungen für die Stützwerte der stückweisen linearen konvexen Funktion (das nichtparametrische Modell). Die Konvexität wird über eine Restriktion an die Anstiege von f gesichert. Das Problem (3.47) ist eindeutig lösbar, da die Zielfunktion streng konvex und koerzitiv ist und der zulässige Bereich eine konvexe Menge darstellt, s. Satz 4.1 und Satz 4.7. Die Schätzungen sind unter recht allgemeinen Voraussetzungen konsistent [273]. Algorithmisch wird (3.47) in ein LSI-Problem (s. Tab. A.3) umgeformt und als solches gelöst.

Zusammengefasst lautet die Idee der nichtparametrischen konvexen Regression also: Verschiebe zunächst die Stützwerte so wenig wie möglich, aber so, dass die Geradenstückenanstiege immer größer werden und verbinde anschließend die neuen Stützwerte durch Geraden.

3.5 Unimodalitätsrestriktionen

Viele Dichtefunktionen sind eingipflig (unimodal, genau ein lokaler und gleichzeitig globaler Maximierer). Auch ist oft a priori bekannt, dass bestimmte Zusammenhänge zunächst ein monotones Anwachsen einer Größe und anschließend ein fortwährendes Absenken dieser Größe beschreiben. Beispiele sind die Drehzahl-Drehmomentenkennlinie, Chromatografieprofile mit nur einem Peak oder physiologische Kenngrößen des Menschen über dem Alter (Körpergröße: Wachstum bis etwa 18 Jahre und danach durch Knorpelreduktion bedingtes Schrumpfen).

Definition 3.3 (Unimodale Funktion)
Eine Funktion $f : \mathcal{X} \cap [a,b] \to \mathbb{R}$ heißt (streng) unimodal nach oben auf $[a,b]$, wenn f in x_{opt} ein (eindeutiges) globales Maximum hat und für $x_1, x_2 \in \mathcal{X} \cap [a,b]$ gilt[14]:
1. $x_1 < x_2 \leq x_{\text{opt}} \Rightarrow (f(x_1) < f(x_2))$ bzw. $f(x_1) \leq f(x_2)$
2. $x_{\text{opt}} \leq x_1 < x_2 \Rightarrow (f(x_1) > f(x_2))$ bzw. $f(x_1) \geq f(x_2)$.
Analog lassen sich streng unimodale Funktionen nach unten definieren, wenn f genau ein globales Minimum hat und in obigen Konklusionen die Relationszeichen umgekehrt werden.

Anmerkung 3.4 In der Statistik werden Verteilungen stetiger Zufallsgrößen, deren Dichte nach oben bzw. unten streng unimodal ist, kurz als streng unimodal bzw. streng antimodal bezeichnet. Letztere heißen auch U-förmige Verteilungen, wenn der Modalwert nicht am Rand liegt. Die Definition streng unimodaler Verteilungen erfolgt auch über die Verteilungsfunktion, wonach diese in $(-\infty, m]$ konvex und in $[m, \infty)$ konkav für einen eindeutigen Modus m ist. Für stetige Zufallsgrößen ist das äquivalent zur Definition über die Dichten. Beachte: Der Begriff der strengen Unimodalität von Verteilungen wird in der Statistik auch verwendet, um unimodale Verteilungen zu charakterisieren, deren Verteilungsfunktionen/Dichten bei Faltung mit einer anderen beliebigen unimodalen Verteilung wieder eine unimodale Verteilung ergeben. Die Cauchy-Verteilung ist dabei streng im ersten, aber nicht streng in diesem Sinne.

Anmerkung 3.5 Stetige streng unimodale Funktionen, deren Extremum auf dem Rand liegt, sind streng monoton. Strenge Unimodalität nach unten (oben) und strenge Quasikonvexität (Quasikonkavität) für Funktionen $f : [a,b] \subset \mathbb{R} \to \mathbb{R}$ sind äquivalent [487]. Ist f zudem differenzierbar, liegt sogar Äquivalenz zu den auf $[a,b]$ streng pseudokonvexen (pseudokonkaven) Funktionen i. S. von Mangasarian [428] vor.

[14] $[a,b]$ muss keine Teilmenge des Definitionsbereichs $\mathcal{X}$ sein. Dieser kann auf $[a,b]$ unterbrochen, ja sogar eine endliche diskrete Teilmenge sein, z. B. bei äquidistanten Abtastfolgen $\{f[k]\} = \{2,1,0,1,2,3,\ldots\}$. Stetigkeit von f wird nicht gefordert.

Es ist offensichtlich, dass die Menge der unimodalen Funktionen auf $[a, b]$ nicht konvex ist. Die Teilmenge der nach unten streng unimodalen Funktionen mit demselben Minimierer ist hingegen konvex. Hieraus leitet sich als Strategie ab, zunächst für mögliche, fest vorgegebenen Positionen der Modi konvexe Teilprobleme zu lösen, um anschließend diejenige Lösung mit dem kleinsten Gütewert zu wählen. Die Lösung der Teilprobleme selbst geschieht, indem zwei monotoniebezogene Restriktionen formuliert werden, denn links des Minimums fällt die Funktion und rechts davon steigt sie monoton. Im Fall der B-Spline-Approximation gemäß Beispiel 3.11 bedeutet das, dass sich in der Folge der Parameter $\{\theta_j\}_{j=0}^{m+2k+1}$ das Relationszeichen höchstens einmal umkehrt. Es sind also maximal $m + 2k + 1$ Fälle zu untersuchen. Praktisch sind dies indes weniger, da für die Lage des Modus sicher nur wenige Knoten in Frage kommen. Gemeinhin werden die Knoten in der Nähe des potenziellen Modus dichter gewählt, da dort die Krümmung am größten ist. Natürlich kann auch die Lage der Knoten in die Optimierung einbezogen werden.

Die Definition der Unimodalität für mehrvariable Funktionen führt wegen der fehlenden Totalordnung in x auf die gleichen Schwierigkeiten wie die Erweiterung des Monotoniebegriffs. Dennoch scheint es legitim und anschaulich, eine stetige Funktion streng unimodal nach oben zu nennen, wenn sie genau ein streng globales Maximum besitzt. Nur leider ist diese Definition für eine mathematische Auswertung ungeeignet. Klar ist auch, dass Unimodalität in einem Punkt bezüglich jeder Koordinatenachse – auch als Orthounimodalität (abgeleitet von Orthant) bezeichnet – keineswegs Unimodalität impliziert, vgl. den um 45° gedrehten Sattel $f(x_1, x_2) = x_1^3 - x_1^2 x_2 + 2x_2^2$ im Punkt $(6, 9)$, in dem die Achsenschnitte $f_1(x_1) = x_1^3 - 9x_1^2 + 162$ und $f_2(x_2) = 216 - 36x_2 + 2x_2^2$ jeweils bei $x_1 = 6$ bzw. $x_2 = 9$ ihr Minimum annehmen.

Eine etwas strengere Definition der Unimodalität gelingt über α-Schnitte, die konvexe Superniveaumengen erzeugen müssen, die mit größer werdendem α immer kleiner werden.

Definition 3.4 (A-Unimodalität, [601])
$f : \mathbb{R}^n \to [0, \infty)$ heißt A-unimodal oder auch quasikonkav[15] [195], wenn die Superniveaumenge $\mathcal{L}_\alpha^+ = \{x : f(x) \geq \alpha\}$ für alle $\alpha > 0$ konvex ist.

Nur ist auch diese Definition für die Auswertung ungeeignet, allerdings steht eine für die Anwendung geeignete Charakterisierung zur Verfügung [195]

$$\forall \gamma \in [0, 1] : \qquad f(\gamma x_1 + (1 - \gamma) x_2) \geq \min\{f(x_1), f(x_2)\}. \tag{3.48}$$

Anmerkung 3.6 A-Unimodalität (Quasikonkavität) stimmt nicht mit dem anschaulichen Verständnis der Unimodalität überein. So ist eine obere Halbkugel, die mit einer stetigen Einkerbung entlang eines Längengrads versehen wird, zwar nach wie vor unimodal, aber nicht A-unimodal, denn die Einkerbung führt auf nichtkonvexe Superniveaumengen.

[15] Wichtige Eigenschaft: Jedes strenge lokale Maximum ist ein globales Maximum.

Für stetig differenzierbare Funktionen kann das Konzept der Unimodalität zudem über die Pseudokonkavität i. S. Mangasarians erweitert werden

$$\forall x_1, x_2 \in \mathcal{C}: \qquad f(x_1) > f(x_2) \Rightarrow \langle \nabla f(x_1), x_1 - x_2 \rangle < 0. \tag{3.49}$$

Anmerkung 3.7 Pseudokonkave Funktionen sind auch quasikonkav. Modifikationen der Pseudokonkavität basierend auf einem Subdifferenzialkalkül werden in [267] beschrieben.

Im Ergebnis der Diskussion zur Unimodalität sollte das A-priori-Wissen also dahingehend beurteilt werden, ob statt Unimodalität nicht die stärkere Eigenschaft der A-Unimodalität vorliegt. Wenn Ja, dann bietet (3.48) den Schlüssel zur Formulierung restringierter Probleme.

3.6 Glattheitsrestriktionen

Ein kritisches Problem bei der Modellbildung für Funktionen ist die Überanpassung. So neigen Polynomansätze zu hohen Grads zu einem starken Oszillieren, wobei die Fehler an den Stützstellen klein oder sogar Null sind, aber der Verlauf zwischen den Stützstellen deutlich vom intuitiv erwarteten Verlauf abweicht. Eine ähnliche Erscheinung ist mitunter auch bei LS-Lösungen zu beobachten, wenn der Parametervektor die Stützwerte einer Funktion sind, die zwar das Residuum minimieren, aber übermäßig schwanken. Der Ingenieur wünscht sich dann einen geglätteten Verlauf. Eine Möglichkeit, das zu erreichen, stellen die in Abschnitt 3.1 besprochenen Kompromisszugänge dar. Doch unabhängig davon, wie die Glattheitsrestriktion bearbeitet wird, zunächst muss sie mathematisch formuliert werden. Eine schwache Form der Glattheit ist die Stetigkeit, die in Abschnitt 3.3 behandelt wurde. Zudem wurde in Abschnitt 3.3 auch eine Technik vorgestellt, mit der sich Knicke und Sprünge glätten lassen. Für parametrische Modelle kann die Glattheit über Restriktionen an die zweite oder eine höhere Ableitung formuliert werden. Dabei gilt, je betragskleiner die zweite Ableitung desto sanfter die Funktionsverläufe. In diesem Abschnitt werden deshalb drei Zugänge beschrieben, mit denen die Forderung nach hinreichender Glattheit von nichtparametrischen Modellen sichergestellt werden kann.

1. Behandlung über Lipschitz-Bedingung
Die Forderung, dass die N Stützwerte $(x_i, y_i) \in \mathbb{R}^n \times \mathbb{R}$ nicht zu sehr schwanken sollen, lässt sich über die bekannte Lipschitz-Bedingung mit einer vorzugebenden Lipschitz-Konstanten L umsetzen. Das approximierende nichtparametrische Modell $(x_i, \hat{f}_i)$ ergibt sich dann aus

$$\sum_{i=1}^{N} (y_i - f_i)^2 \stackrel{!}{=} \text{Min} \qquad f_1, \ldots, f_N : |f_i - f_j| \leq L||x_i - x_j||_2; \forall i, j. \tag{3.50}$$

Eine Kombination mit Monotonierestriktionen wird in [58] beschrieben.

2. Behandlung über numerische Ableitungen
Eine andere Möglichkeit, die Glattheit auszudrücken, basiert auf der numerischen Ableitung zweiter Ordnung $f_i'' = f_{i+1} - 2f_i + f_{i-1}$. Dies führt auf

$$||Af|| \leq \gamma \qquad \text{mit } A = \begin{bmatrix} -2 & 1 & 0 & \dots & 0 \\ 1 & -2 & 1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & -2 & 1 \\ 0 & \dots & 0 & 1 & -2 \end{bmatrix}. \tag{3.51}$$

Im Fall der euklidischen Norm wird dadurch die Schwankung der zweiten Ableitungen im Mittel beschränkt, im Fall der Chebyshev-Norm wird der Betrag der zweiten Ableitung für jeden Punkt beschränkt. Das Prinzip kann ebenso auf Ableitungen anderer Ordnung angewandt werden. Auch lassen sich Ableitungsrestriktionen verschiedener Ordnung miteinander kombinieren. Neben einer direkten Berücksichtigung als Restriktion, werden die Terme der Struktur $||Af||$ oder $||Af||^2$ gern gewichtet in das Gütekriterium aufgenommen, s. hierzu auch Abschnitt 3.1. Eine typische Anwendung hierfür sind Varianten der Tikhonov-Regularisierung von LS-Problemen [597].

3. Behandlung über modifizierte Splines
Eine Alternative zu den expliziten Restriktionen für Glattheit sind bei der Interpolation modifizierte kubische Splines [368]. Obwohl Splines von Hause aus weniger zum Überschießen (zur Welligkeit) neigen wie globale Polynominterpolationen, tritt dieser unerwünschte Effekt dennoch auf. Ein zweckmäßiger Weg besteht dann darin, auf die Stetigkeit der zweiten Ableitung zu verzichten und stattdessen die erste Ableitung stärker zu restringieren. Während bei den kubischen Splines nur die Stetigkeit $f_i'(x_i) = f_{i+1}'(x_i)$ gefordert wird, wird diese bei den modifizierten kubischen Splines um eine zusätzliche punktweise Restriktion ergänzt

$$f_i'(x_i) = f_{i+1}'(x_i) = \begin{cases} \dfrac{2}{\frac{x_{i+1}-x_i}{y_{i+1}-y_i} + \frac{x_i-x_{i-1}}{y_i-y_{i-1}}} & \text{keine Vorzeichenänderung des Anstiegs} \\ 0 & \text{bei Vorzeichenänderung des Anstiegs.} \end{cases} \tag{3.52}$$

Tritt keine Vorzeichenänderung auf, so wird für die ersten Ableitungen der Splines an einer Stützstelle gefordert, dass sie dem harmonischen Mittel der benachbarten Sekantenanstiege entsprechen müssen. Andere Heuristiken sind gleichfalls möglich. Dank dieser Modifikation wird bei monotonen bzw. unimodalen Stützwertfolgen auch eine monotone bzw. unimodale Interpolation erreicht. Im Ergebnis entsteht ein Verlauf, der dem erwarteten Verlauf bei visueller Betrachtung der Stützwerte oft sehr nahe kommt.

Da die Glattheitsforderung meist nicht dem A-priori-Wissen, sondern eher der ingenieurmäßigen Intuition entstammt, wird sie bevorzugt nicht hart, sondern per Kompromisszugang über einen Strafterm einbezogen.

3.7 Symmetrierestriktionen

Statische Kennlinien weisen oft Symmetrien auf:

$$f(x) = f(2x_0 - x) \qquad \text{Achsensymmetrie bzgl. } x = x_0 \tag{3.53a}$$

$$f(x) = 2y_0 - f(2x_0 - x) \qquad \text{Punktsymmetrie bzgl. } (x_0, y_0). \tag{3.53b}$$

Das Wissen um solche Symmetrien kann gerade bei approximativparametrischen Modellen die Güte erheblich verbessern, nämlich immer dann, wenn die Daten den Gültigkeitsbereich nicht gleichmäßig abdecken. Liegen im Fall einer symmetrischen nichtnegativen Funktion experimentationsbedingt nur wenige Daten im zweiten Quadranten aber viele im ersten Quadranten, so ergibt sich ohne Beachtung der Symmetrierestriktion oft ein stark asymmetrisches Regressionsergebnis. Mit einer Symmetrierestriktion hingegen passiert das nicht. Zudem verbessert sich die Schätzgüte, da in jedem Quadranten symmetriebedingt nun alle Daten wirksam werden.

Neben der Anwendung bei statischen Kennlinien können Symmetrierestriktionen auch erfolgreich bei Phasendiagrammen (Approximation des Lorenz-Systems im (x_k, x_{k+3})-Diagramm) oder bei Bifurkationsdiagrammen (Duffing-Ueda-Oszillators) eingesetzt werden [9]. Eine Anwendung der Symmetrierestriktionen bei der Regression mit Support-Vector-Machines und nichtlinearen AR-Modellen beschreibt [183].

Der Achsensymmetrie im Zweidimensionalen entspricht die Flächensymmetrie im Dreidimensionalen, der Punktsymmetrie die Drehsymmetrie um 180°. Hinzu kommt die Rotationssymmetrie um eine Achse, wobei identische Abbilder für jeden Drehwinkel vorliegen müssen, und die Radialsymmetrie, bei der identische Abbilder für mindestens einen Winkel existieren. Die Rotationssymmetrie ist dabei typisch für Ausbreitungsverläufe (Temperatur, Schall, Konzentration) von Punktquellen.

Sind Symmetrien für ein Problem erkannt worden, können sie über Gleichungsrestriktionen oder Penalty-Terme eingebaut werden. Vielfach ist es aber geschickter, ein System von Ansatzfunktionen zu wählen, dass die Symmetrie per se sicherstellt. Für achssymmetrische Probleme um $x = 0$ sind das die geraden Funktionen, für die rotationssymmetrischen Probleme die radialen Basisfunktionen. Ganz allgemein können Ansatzfunktionen auch dadurch erzeugt werden, dass beliebige Funktionen $g(x)$ (diese können mit zusätzlichen Eigenschaften versehen werden) auf die Menge der Funktionen $f(x)$ projiziert werden. Als Beispiel sei das Symmetrisieren einer Ansatzfunktion g durch $f(x) = \frac{1}{2}(g(x) + g(-x))$ genannt. Anschaulich bedeutet dies, dass das Platzieren einer Ansatzfunktion g etwa in einem Bereich, wo viele Daten vorliegen, stets ein Platzieren der gespiegelten Ansatzfunktion nach sich zieht. Der Faktor $\frac{1}{2}$ kann oft weggelassen werden, da f ohnehin meist noch gewichtet wird. Auf keinen Fall dürfen $g(x)$ und $g(-x)$ unabhängig voneinander gewichtet werden, sondern nur als Summe, weil sonst die Symmetrie verloren geht.

Für die mathematische Behandlung ist es entscheidend, ob die Symmetrieparameter x_0 bzw. (x_0, y_0) bekannt oder unbekannt sind. Bei bekannten Parametern kann das betreffende Problem durch Koordinatentransformation vereinfacht werden; bei unbekannten Parametern sind diese zusätzlich zu dem f beschreibenden Parametervektor θ in das Schätzproblem aufzunehmen.

3.8 Sektorrestriktionen

Eng verwandt mit den Intervall- und den Symmetrierestriktionen sind die Sektorrestriktionen

$$\forall x, t: \qquad k_1 x^2 \leq x f(x,t) \leq k_2 x^2 \tag{3.54}$$

oder de facto äquivalent[16]

$$\forall x \neq 0, t: \qquad k_1 x \leq f(x,t) \leq k_2 x, \tag{3.55}$$

bei denen $f(x,t)$ im durch die Geraden mit den Anstiegen k_1, k_2 aufgespannten Sektor liegen muss. Verallgemeinerungen auf mehrvariate Probleme, das Zulassen strenger Ungleichungen, das Einschränken der x, wie bei $\sin x$ auf $|x| \leq \pi$, oder die Wahl von $k_2 = \infty$ sind gängige Erweiterungen. So entspricht die Sektorbedingung $[0, \infty)$ der 1-3-Quadrantenbeziehung

$$x f(x,t) \geq 0, \qquad f(0,t) \equiv 0, \tag{3.56}$$

die unter anderem speicherlose passive Systeme kennzeichnet. Sie erfährt ihre Bedeutung durch einen Satz zur Stabilisierbarkeit passiver Systeme [113] und spielt eine wichtige Rolle bei der Frage nach der absoluten Stabilität von Lur'e-Regelkreissystemen [4], s. auch Abschnitt 2.9. Die Zeitabhängigkeit liegt bei Anwendungen, in denen f identifiziert wird, oft nicht vor. Wenn doch, wird sie vernachlässigt oder indirekt durch Online-Identifikationsverfahren über zeitveränderliche Parameter berücksichtigt.

Die Punktrestriktion $f(0) = 0$ lässt sich durch eine geeignete Funktionswahl sichern oder über eine parametrische Restriktion erzwingen. Die Ungleichung kann für Polynome in Nichtnegativitätsrestriktionen überführt und als solche behandelt werden.

Anmerkung 3.8 Gewissermaßen das Pendant zur 1-3-Quadrantenbeziehung ist die 2-4-Quadrantenbeziehung

$$x f(x) < 0, \qquad f(0) = 0, \tag{3.57}$$

die für skalare $\dot{x} = f(x)$ notwendig und hinreichend für GAS ist.

[16] Für eine Äquivalenz muss zudem gelten: $f(0,t) \equiv 0$ oder f stetig oder k_1, k_2 beschränkt.

Die Verallgemeinerung der Sektorbedingungen auf den Mehrgrößenfall [211] erfolgt dadurch, dass die positiven Verstärkungen durch positiv definite Verstärkungsmatrizen und Produkte durch Skalarprodukte ersetzt werden:
- $[0_{m\times m}, \infty_{m\times m}]$, wenn $x^T f(x,t) \geq 0$
- $[K_1, \infty_{m\times m}]$, wenn $x^T[f(x,t) - K_1 u] \geq 0$
- $[0_{m\times m}, K_2]; K_2 \succ 0_{m\times m}$, wenn $f^T(x,t)[f(x,t) - K_2 u] \leq 0$
- $[K_1, K_2]; K_2 \succ K_1$, wenn $[f(x,t) - K_1 u]^T[f(x,t) - K_2 u] \leq 0$.

Für den Regelungstechniker sind Sektorrestriktionen also insbesondere unter Stabilitätsgesichtspunkten von Regelkreissystemen (LTI-System und statischen Nichtlinearität) interessant. Dabei kann es sein, dass zunächst offline oder per A-priori-Wissen der Sektor für das LTI-System bestimmt wird und die Adaption bzw. Identifikation der Nichtlinearität dann in den Grenzen erfolgen muss. Gegebenenfalls sind sogar Kombinationen mit Passivitätsrestriktionen an das LTI-System erforderlich. Es ist aber auch möglich, die Sektorbedingungen für die A-posteriori-Modellvalidierung eines Regelkreissystems zu verwenden. Für nichtregelungstechnische Anwendungen ist oft die Punktrestriktion $f(0) = 0$ bedeutender, da die Daten gemeinhin die Quadranten- oder Sektorforderung einhalten. Fernerhin kann die Sektorrestriktion oft zu einer Punktrestriktion $f'(0) \leq k_2$ zuzüglich einer Konkavitätsrestriktion (Kurve mit abflachendem Anstieg) verschärft werden.

3.9 Restriktionen an Kovarianzfolgen und -matrizen

Das Autokovarianzfolgenmodell bevorzugt der Ingenieur als Zwischenmodell für die Erstellung parametrischer Filtermodelle oder auch, um Aussagen über einen Zufallsprozess ableiten zu können. Während Abschnitt 3.9.1 die theoretischen Grundlagen betrachtet, bezieht sich Abschnitt 3.9.2 auf die unterschiedlichen Schätzungen (Optionen in der Standardsoftware) für Autokovarianzfolgen. Wichtige Erkenntnis dieses Abschnitts für den Anwender ist es, dass die Abtastperiode geeignet zu wählen ist und nicht die komplette Autokovarianzfolge für die Weiterverarbeitung verwendet werden sollte, sondern nur jenen Teil, für den die Schätzwerte hinreichend valide sind.

Eng verwandt mit der Autokovarianzfolge ist die Autokovarianzmatrix. Sie verkörpert die stochastischen Beziehungen zweiter Ordnung zwischen den Komponenten eines Zufallsvektors und entspricht im Fall ergodischer vektorieller Zufallsprozesse der Autokovarianzfolge an der Stelle $\lambda = 0$. Der Einsatz von Restriktionen zur Verbesserung der Schätzungen wird in Abschnitt 3.9.3 beschrieben.

3.9.1 Grundlagen

Für einen schwach stationären zeitdiskreten Zufallsprozess $\{\boldsymbol{x}[k]\}_{k=1}^{N}$ mit $\mu_{\boldsymbol{x}} \stackrel{\text{def}}{=} \mathrm{E}\{\boldsymbol{x}[k]\}$ ist die (gewöhnliche, d. h. ensemblegemittelte) Autokovarianzfolge (AKF) $\{c_{\boldsymbol{x}}[\lambda]\}_{\lambda=-\infty}^{\infty} \in \mathbb{C}^{\mathbb{Z}}$ definiert durch

$$c_{\boldsymbol{x}}[\lambda] \stackrel{\text{def}}{=} \mathrm{cov}(\boldsymbol{x}[k], \boldsymbol{x}[k+\lambda]) = \mathrm{E}\{(\boldsymbol{x}[k]-\mu_{\boldsymbol{x}})(\overline{\boldsymbol{x}[k+\lambda]-\mu_{\boldsymbol{x}}})\}; \quad \lambda \in \mathbb{Z}.^{17} \tag{3.58}$$

Sie ist konsymmetrisch, d. h. $c_{\boldsymbol{x}}[-\lambda] = \bar{c}_{\boldsymbol{x}}[\lambda]$, und es gilt $|c_{\boldsymbol{x}}[\lambda]| \leq c_{\boldsymbol{x}}[0]$, da (Cauchy-Schwarz-Ungleichung)

$$|c_{\boldsymbol{x}}[\lambda]| = |\mathrm{cov}(\boldsymbol{x}[k], \boldsymbol{x}[k+\lambda])| \leq \sqrt{\mathrm{var}(\boldsymbol{x}[k])\mathrm{var}(\boldsymbol{x}[k+\lambda])} = \sqrt{c_{\boldsymbol{x}}[0]\cdot c_{\boldsymbol{x}}[0]} = c_{\boldsymbol{x}}[0]. \tag{3.59}$$

Aus (3.58) wird überdies gefolgert, dass für alle $n \geq 1$ und beliebige $a = ((a_i)) \in \mathbb{C}^n$ gilt

$$\mathrm{var}\Big(\sum_{k=1}^{n} a_k \boldsymbol{x}[k]\Big) = \sum_{k=1}^{n}\sum_{l=1}^{n} a_k \bar{a}_l \mathrm{cov}(\boldsymbol{x}[k], \boldsymbol{x}[l]) = \sum_{k=1}^{n}\sum_{l=1}^{n} a_k \bar{a}_l c_{\boldsymbol{x}}[l-k] = a^H C_n a \geq 0 \tag{3.60}$$

bzw. äquivalent

$$C_n = \begin{bmatrix} c_{\boldsymbol{x}}[0] & c_{\boldsymbol{x}}[1] & \dots & c_{\boldsymbol{x}}[n-1] \\ \bar{c}_{\boldsymbol{x}}[1] & c_{\boldsymbol{x}}[0] & \ddots & \vdots \\ \vdots & \ddots & \ddots & c_{\boldsymbol{x}}[1] \\ \bar{c}_{\boldsymbol{x}}[n-1] & \dots & \bar{c}_{\boldsymbol{x}}[1] & c_{\boldsymbol{x}}[0] \end{bmatrix} \succeq 0_{n\times n}. \tag{3.61}$$

Definition 3.5 (Toeplitz-Folge)
Eine zweiseitige unendliche Folge $\{c_{\boldsymbol{x}}[\lambda]\}_{\lambda=-\infty}^{\infty}$ mit der Eigenschaft (3.61), dass die korrespondierende Folge von Toeplitz-Matrizen $\{C_n\}_{n=1}^{\infty}$ nichtnegativ (positiv) definit ist, heißt nichtnegativ (positiv) definite Folge oder auch Toeplitz-Folge.

Die zeitdiskrete Fourier-Transformation einer Autokovarianzfolge heißt Leistungsspektraldichte $S_{\boldsymbol{x}}(\omega)$. Zwischen beiden gelten die Wiener-Khintchine-Beziehungen[18]

$$S_{\boldsymbol{x}}(\omega) := \tilde{S}_{\boldsymbol{x}}(\mathrm{e}^{j\omega T_A}) \stackrel{\mathcal{D}'}{=} \sum_{\lambda=-\infty}^{\infty} c_{\boldsymbol{x}}[\lambda]\mathrm{e}^{-j\omega T_A \lambda} \qquad /^{19} \tag{3.62a}$$

$$c_{\boldsymbol{x}}[\lambda] \stackrel{\mathcal{D}'}{=} \frac{1}{2\pi}\int_{[-\pi,\pi)} S_{\boldsymbol{x}}(\omega)\mathrm{e}^{j\omega T_A \lambda}\,\mathrm{d}\omega, \qquad /^{20} \tag{3.62b}$$

[17] Allgemeiner wird $c_{\boldsymbol{xy}}[k,l] \stackrel{\text{def}}{=} \mathrm{cov}(\boldsymbol{x}[k], \boldsymbol{y}[l]) = \mathrm{E}\{(\boldsymbol{x}[k]-\mathrm{E}\{\boldsymbol{x}[k]\})(\overline{\boldsymbol{y}[l]-\mathrm{E}\{\boldsymbol{y}[l]\}})\}$ die Kreuzkovarianzfunktion für instationäre Prozesse definiert. Diese gilt für alle k, l unabhängig davon, ob $k > l, k < l, k = l$ gilt. Die Wahl $\lambda := l - k$ ist üblich, wenngleich gelegentlich $\lambda := k - l$ verwendet wird. Der Überstrich bedeutet, dass das konjugiert Komplexe der betreffenden Variable zu nehmen ist.

[18] In $\tilde{S}_{\boldsymbol{x}}(\mathrm{e}^{j\omega T_A})$ (Substituiere $z := \mathrm{e}^{j\omega T_A}$) zeigt sich der Zusammenhang zur zweiseitigen z-Transformation.

$S_{\boldsymbol{x}}(\omega)$ ist eine reelle und gerade Funktion, denn

$$S_{\boldsymbol{x}}(\omega) = 2\,\Re e\left(\frac{c_{\boldsymbol{x}}[0]}{2} + \sum_{\lambda=1}^{\infty} c_{\boldsymbol{x}}[\lambda]\mathrm{e}^{-j\omega T_A\lambda}\right) = c_{\boldsymbol{x}}[0] + 2\sum_{\lambda=1}^{\infty} c_{\boldsymbol{x}}[\lambda]\cos(\lambda\omega T_A). \tag{3.63}$$

Zudem gilt $S_{\boldsymbol{x}}(\omega) \geq 0$ für alle ω. Wäre nämlich $S_{\boldsymbol{x}}(\omega)$ auf einem beliebigen Intervall $[\omega_u, \omega_o]$ negativ, hätte der über einen idealen Bandpass $G(j\omega) = \begin{cases} 1 & \omega \in [\omega_u, \omega_o] \cup [-\omega_o, -\omega_u] \\ 0 & \text{sonst} \end{cases}$ gefilterte Zufallsprozess $\{\boldsymbol{y}_k\}$ die Spektraldichte $S_{\boldsymbol{y}}(\omega) = |G(j\omega)|^2 S_{\boldsymbol{x}}(\omega)$, die eine negative Varianz $\mathrm{var}(\boldsymbol{y}_k) = \int_{[-\pi,\pi)} S_{\boldsymbol{y}}(\omega)\mathrm{d}\omega$ implizieren würde (Widerspruch, da Varianz immer positiv). Umgekehrt ist $S(\omega) \geq 0$ hinreichend, um eine AKF zu konstruieren, denn

$$\sum_{k,l=1}^{n} a_k\bar{a}_l c[k-l] = \sum_{k,l=1}^{n} a_k\bar{a}_l \int_{[-\pi,\pi)} \mathrm{e}^{j(k-l)\omega T_A} S_{\boldsymbol{x}}(\omega)\,\mathrm{d}\omega = \int_{[-\pi,\pi)} \left|\sum_{k=1}^{n} a_k \mathrm{e}^{jk\omega T_A}\right|^2 S_{\boldsymbol{x}}(\omega)\,\mathrm{d}\omega \geq 0. \tag{3.64}$$

All diese Eigenschaften fasst der folgende Satz zusammen.

Satz 3.5 (Charakterisierung AKF)
Folgende Aussagen sind äquivalent:
1. Es existiert ein stationärer Zufallsprozess mit der Autokovarianzfolge $\{c_{\boldsymbol{x}}[\lambda]\}_{\lambda=-\infty}^{\infty}$.
2. $\{c_{\boldsymbol{x}}[\lambda]\}_{\lambda=-\infty}^{\infty}$ ist eine nichtnegativ definite Folge: $\forall n \geq 1:\ \mathrm{Toe}(c_{\boldsymbol{x}}[0], \ldots, c_{\boldsymbol{x}}[n-1]) \succeq 0_{n\times n}$.
3. $S_{\boldsymbol{x}}(\omega) \geq 0$ gilt für alle $\omega \in [-\pi, \pi)$.

Eine Konsequenz aus Satz 3.5 ist, dass „$\{c_{\boldsymbol{x}}[\lambda]\}_{\lambda=-\infty}^{\infty}$ sei eine Autokovarianzfolge“ eine unendlichdimensionale Restriktion darstellt, nämlich entweder $C_k \succeq 0_{k\times k};\ k = 1, 2, \ldots$ oder $\tilde{S}_{\boldsymbol{x}}(\mathrm{e}^{j\omega T_A}) \geq 0$ für alle $\omega \in [-\pi, \pi)$. Es bleibt die Frage, ob sich die Situation verbessert, wenn zusätzlich gefordert wird, dass die AKF endlich sein soll.

Definition 3.6 (Endliche AKF)
Eine endliche AKF $\{c_{\boldsymbol{x}}[\lambda]\}_{\lambda=-n}^{n}$ ist eine AKF mit $c_{\boldsymbol{x}}[\lambda] = 0$ für $|\lambda| > n$ und $c_{\boldsymbol{x}}[n] \neq 0$.

Nachfolgende Überlegungen zeigen, dass Satz 3.5 keine direkte Möglichkeit bietet, um für $\{c_{\boldsymbol{x}}[\lambda]\}_{\lambda=-n}^{n}$ auf endlich viele Restriktionen schließen zu können. Wie jede AKF impliziert auch die endliche AKF $S_{\boldsymbol{x}}(\omega) \geq 0$. Aber $S_{\boldsymbol{x}}(\omega) \geq 0$ impliziert nicht, dass die AKF endlich

[19] $\sum_{\lambda=0}^{\infty} |c_{\boldsymbol{x}}[\lambda]| < \infty$ ist hinreichend dafür, dass (3.62a) für gewöhnliche Funktionen gilt. Für den stationären Prozess $\{\boldsymbol{x}_k\} = \{\cos(\omega k + \boldsymbol{u})\}$ mit $\boldsymbol{u} \sim Gl(-\pi, \pi);\ 0 \leq \omega \leq \pi$ ist die Bedingung nicht erfüllt. Er hat keine Spektraldichte in den gewöhnlichen Funktionen, aber in den verallgemeinerten. Das $\mathcal{D}'$ über dem Gleichheitszeichen verdeutlicht, dass das Gleich i. S. der Distributionentheorie [566], [215] zu verstehen ist.

[20] Im Gegensatz zur Spektraldichte existiert die Spektralverteilungsfunktion, kurz Spektralfunktion, $SV_{\boldsymbol{x}}(\omega)$ stets und ist eindeutig. (3.62b) kann damit auch als Stieltjes-Integral geschrieben werden, $c_{\boldsymbol{x}}[\lambda] = \int_{[-\pi,\pi)} \mathrm{e}^{j\omega\lambda}\mathrm{d}SV_{\boldsymbol{x}}(\omega)$. Für absolutstetige Spektralfunktionen gehen die Integrale ineinander über.

ist. Eine endliche AKF impliziert $C_{n+1} \succeq 0_{(n+1)\times(n+1)}$, was wiederum $C_k \succeq 0_{k\times k}$ für $k < n+1$ impliziert [299]. Aber $C_{n+1} \succeq 0_{(n+1)\times(n+1)}$ impliziert nicht, dass eine endliche AKF existiert. So ist $\mathrm{Toe}(1, 0.7)$ sogar positiv definit, aber $\mathrm{Toe}(1, 0.7, 0, 0)$ indefinit, d. h. $\{1, 0.7, 0, 0, \ldots\}$ kann nach Satz 3.5 keine AKF sein. Sowohl $C_{n+1} \succeq 0_{(n+1)\times(n+1)}$ als auch $C_{n+1} \succ 0_{(n+1)\times(n+1)}$ sind nur notwendig für die Existenz einer endlichen AKF, während $C_{n+1} \succ 0_{(n+1)\times(n+1)}$ hinreichend für die Existenz einer nicht notwendig endlichen AKF (Carathéodory-Toeplitz-Erweiterungsproblem [114]) ist. Deshalb sind die folgenden Sätze von Bedeutung. So stellt Satz 3.6 die Verbindung zwischen einer endlichen AKF und ihrer Generierung über ein „moving average"-Filter, kurz MA-Filter, her, weshalb endliche AKF auch MA(n)-AKF heißen.

Satz 3.6 (Charakterisierung endlicher AKF, [572])
Es existiert genau dann ein stationärer Zufallsprozess $\{\boldsymbol{x}[k]\}_{k=1}^{\infty}$ mit $\{c_{\boldsymbol{x}}[\lambda]\}_{\lambda=-n}^{n}$, wenn $c_{\boldsymbol{x}}[\lambda] = \sum_{k=0}^{n-\lambda} b_k \bar{b}_{k+\lambda}$ für Zahlen $b_0, \ldots, b_n \in \mathbb{C}$ und $\lambda = 0, \ldots, n$ gilt.[21]

Satz 3.6 gibt $n+1$ algebraische Restriktionen für $c_{\boldsymbol{x}}[\lambda]$ an, die eine endliche AKF gewährleisten. Leichter anwendbar als diese nichtlinearen Restriktionen sind die aus ihnen abgeleiteten Bedingungen [572]

$$\sum_{\lambda=1}^{n} |c_{\boldsymbol{x}}[\lambda]| \leq \frac{1}{2} c_{\boldsymbol{x}}[0] \quad \text{(hinreichend)} \qquad \sum_{\lambda=1}^{n} |c_{\boldsymbol{x}}[\lambda]| \leq \frac{n}{2} c_{\boldsymbol{x}}[0] \quad \text{(notwendig)}. \tag{3.65}$$

Beide Bedingungen sind zwar in endlich viele Ungleichungen umformbar, sind aber abgesehen von $n = 1$ nicht gleichzeitig notwendig und hinreichend und somit nur eingeschränkt nutzbar.

Zum Abschluss wird die algebraische Struktur endlicher AKF angegeben, wobei die Abgeschlossenheit und Konvexität hervorzuheben sind, die nach den Sätzen in den Abschnitten 4.1 und 4.2.1 für die Existenz einer Lösung und deren Eindeutigkeit wesentlich sind. Die äquivalente Formulierung der Endliche-AKF-Restriktion durch eine einzige Matrixrestriktion gibt Satz 3.7.

Satz 3.7 (Algebraische Struktur endlicher AKF, [173])
Die Menge der reellen endlichen AKF $\mathcal{C}_{n+1}$ formt einen abgeschlossenen konvexen Kegel. Der Dualkegel $\mathcal{C}_{n+1}^{D}$ ist isomorph zu den nichtnegativ definiten Toeplitz-Matrizen der Ordnung $n+1$.

Für das prinzipielle Aufbereiten derartiger Matrixrestriktionen zu linearen Matrixungleichungen im Rahmen der semidefiniten Optimierung sei auf Abschnitt 7.8 verwiesen, für das Aufbereiten dieses Spezialproblems seien die Arbeiten [173], [636], [100] genannt.

[21] Die b_i sind nicht eindeutig bestimmt, vgl. die $\{c_{\boldsymbol{x}}[\lambda]\}_0^3 = \{15, -2, 5, 2\}$ und $b_{0:3} = [1, 3, -1, 2]$ bzw. $b_{0:3} = [3.46, -0.579, 1.54, 0.579]$.

3.9.2 Schätzproblematik von Autokovarianzfolgen

Als Schätzungen für die AKF aus einer Realisierung $\{x[1], \ldots, x[N]\}$ kommen spezielle empirische AKF in Frage, die sich hinsichtlich der verwendeten Mittelwertschätzung und des Skalierungsfaktors unterscheiden, was seinerseits unterschiedliche statische Eigenschaften nach sich zieht, vgl. die nachstehenden Diskussion:

$$\hat{c}_{\boldsymbol{x}}^{nnd}[\lambda] = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{k=1}^{N-\lambda}(x[k]-\hat{\mu}_{\boldsymbol{x}})(\overline{x[k+\lambda]-\hat{\mu}_{\boldsymbol{x}}}) & \lambda = 0,\ldots,N-1 \\ \overline{\hat{c}_{\boldsymbol{x}}^{nnd}[-\lambda]} & \lambda = 1-N,\ldots,-1 \end{cases} \tag{3.66a}$$

$$\hat{c}_{\boldsymbol{x}}[\lambda] = \begin{cases} \dfrac{1}{N-\lambda}\displaystyle\sum_{k=1}^{N-\lambda}(x[k]-\hat{\mu}_{\boldsymbol{x}})(\overline{x[k+\lambda]-\hat{\mu}_{\boldsymbol{x}}}) & \lambda = 0,\ldots,N-1 \\ \overline{\hat{c}_{\boldsymbol{x}}[-\lambda]} & \lambda = 1-N,\ldots,-1 \end{cases} \tag{3.66b}$$

$$\hat{c}_{\boldsymbol{x}}^{sep}[\lambda] = \begin{cases} \dfrac{1}{N-\lambda}\displaystyle\sum_{k=1}^{N-\lambda}(x[k]-\hat{\mu}_{1:N-\lambda})(\overline{x[k+\lambda]-\hat{\mu}_{\lambda+1:N}}) & \lambda = 0,\ldots,N-1 \\ \overline{\hat{c}_{\boldsymbol{x}}^{sep}[-\lambda]} & \lambda = 1-N,\ldots,-1 \end{cases} \tag{3.66c}$$

mit

$$\hat{\mu}_{\boldsymbol{x}} = \frac{1}{N}\sum_{k=1}^{N} x[k], \qquad \hat{\mu}_{1:N-\lambda} = \frac{1}{N-\lambda}\sum_{k=1}^{N-\lambda} x[k], \qquad \hat{\mu}_{\lambda+1:N} = \frac{1}{N-\lambda}\sum_{k=\lambda+1}^{N} x[k]. \tag{3.67}$$

Die empirischen AKF haben folgende Eigenschaften:

- Die empirischen AKF liefern Schätzungen für die zeitgemittelte AKF

$$c_{\boldsymbol{x}}^{time}[\lambda] = \lim_{N\to\infty}\frac{1}{2N+1}\sum_{k=-N}^{N}(x[k]-\hat{\mu}_{\boldsymbol{x}})(\overline{x[k+\lambda]-\hat{\mu}_{\boldsymbol{x}}}) \tag{3.68}$$

  Für im weiteren Sinne ergodische Prozesse sind das auch Schätzungen für die gewöhnliche AKF (3.58).

- Eine Normierung liefert die empirischen Autokorrelationsfolgen $\{\hat{\varrho}_{\boldsymbol{x}}[\lambda]\} = \{\hat{c}_{\boldsymbol{x}}[\lambda]/\hat{c}_{\boldsymbol{x}}[0]\}$, die für die angegebenen drei Fälle bei ergodischen Zufallsprozessen asymptotisch erwartungstreue Schätzungen sind.

- $\{\hat{c}_{\boldsymbol{x}}^{nnd}[\lambda]\}_{\lambda=-n}^{n}$ ist stets eine endliche AKF, aber biasbehaftet. Ist λ bzgl. N groß, so gehen in die Berechnung nur wenige Werte ein und da für große λ auch der Absolutwert einer exponentiell abklingenden AKF klein ist, ist der relative Bias sehr groß. Allerdings ist der Erwartungswert des mittleren quadratischen Fehlers geringer als bei Verwendung von (3.66b). Ein weiterer Vorteil ist, dass für große λ die empirische AKF gegen Null strebt, wodurch Artefakte für große λ vermieden werden und eine wichtige Eigenschaft absolut summierbarer AKF, d. h. jener mit einer gewöhnlichen Spektraldichte, eingehalten wird.

- $\{\hat{c}_{\boldsymbol{x}}[\lambda]\}_{\lambda=-n}^{n}$ ist für absolut summierbare AKF asymptotisch erwartungstreu und konsistent, aber gemeinhin keine endliche AKF. Die Schätzung ist nicht erwartungstreu, denn für einen stationären weißen Rauschprozess ist bekannt, dass die Varianz $c_{\boldsymbol{x}}[0]$ nur erwartungstreu geschätzt wird, wenn statt durch N durch $N-1$ geteilt wird, da durch die Mittelwertbildung ein Freiheitsgrad verloren geht. Doch selbst mit einer Korrektur des Faktors auf $\frac{1}{N-\lambda-1}$, ist die Schätzung abgesehen vom unkorrelierten Fall i. Allg. nicht erwartungstreu, denn $\mathrm{E}\{\hat{\boldsymbol{c}}_{\boldsymbol{x}}[0]\} = c_{\boldsymbol{x}}[0]\Big(1 - \frac{2}{N-1}\sum_{\lambda=1}^{N-1}(1-\frac{\lambda}{N})\varrho_{\boldsymbol{x}}[\lambda]\Big)$, s. auch [550]. Das liegt daran, dass die Realisierungen für die AKF-Schätzung nicht unabhängig, sondern eben gerade korreliert sind.

- $\hat{c}_x^{sep}[\lambda]$ basiert auf der Aufteilung der Daten in $x_1, \ldots, x_{N-\lambda}$ und $x_{\lambda+1}, \ldots, x_N$. Für diese werden empirische Kovarianz- bzw. Korrelationskoeffizienten berechnet, wobei die zweite Datengruppe die in den Formeln auftretenden y-Werte ersetzt. Das Problem dabei ist, dass zur Mittelwertberechnung und zur Berechnung der Standardabweichung für den Korrelationskoeffizienten nicht die gesamten Daten, sondern nur die Teilmengen herangezogen werden (schlechtere Schätzungen bei weniger Daten). Der Vorteil von $\hat{c}_x^{sep}[\lambda]$ liegt in einer Abschwächung von Instationaritätseinflüssen, etwa eines Trends. Dennoch sind Trendkorrekturen (Entfernen eines linearen oder exponentiellen Trends) oder die Betrachtung der numerisch differenzierten Zeitreihe oft die bessere Wahl. Von den statistischen Eigenschaften her verhält sich $\hat{c}_{\boldsymbol{x}}^{sep}[\lambda]$ ähnlich wie $\hat{c}_{\boldsymbol{x}}[\lambda]$.

Ein Manko aller AKF-Schätzungen ist, dass die Varianz der AKF bei Gaußprozessen praktisch unabhängig vom Argument λ ist [550], wodurch betragskleine Schätzwerte der AKF eine große relative Varianz aufweisen. Deshalb empfiehlt es sich, für die Weiterverarbeitung (beispielsweise MA-Parameterschätzung) nur Werte mit einem vertretbaren Stör-Nutzsignal-Verhältnis zu benutzen. Das verkürzt die nutzbare Länge der AKF. Die verkürzte AKF wird dann entweder als endliche AKF betrachtet oder sie wird um meist endlich viele Glieder erweitert, sodass die erweiterte AKF bestimmte gewünschte Eigenschaften besitzt. Die erweiterte AKF interpoliert dabei die Werte der verkürzten AKF (Carathéodory-Toeplitz-Problem). In beiden Fällen ist zu sichern, dass die verkürzte AKF eine endliche AKF ist, damit ein die AKF generierendes diskretes MA-Filter existiert. Eine mögliche Problemformulierung mit der endliche-AKF-Restriktion wird im Beispiel 3.14 gegeben.

Beispiel 3.14 (Projektion auf endliche AKF)
Sei $\hat{c} = (\hat{c}_{\boldsymbol{x}}[0], \ldots, \hat{c}_{\boldsymbol{x}}[n])^T$ ein Vektor der geschätzten reellen Folgenwerte, dann kann die nächstgelegene AKF c_{opt} aus

$$\left\| \begin{bmatrix} c - \hat{c} \\ \hat{c}_{n+1:n+m} \end{bmatrix} \right\|_W \stackrel{!}{=} \text{Min} \qquad \{c_{\boldsymbol{x}}[\lambda]\}_{\lambda=-n}^{n} \text{ ist endliche AKF} \tag{3.69}$$

erhalten werden, s. [173] zur Wahl von W, zur Problemeinordnung und zu Varianten, die die verbale Restriktion in eine mathematische übersetzen.

Anmerkung 3.9 Im Beispiel 3.14 werden mehr als die benötigten $n+1$ Schätzwerte von $\{\hat{c}_{\boldsymbol{x}}[\lambda]\}_{\lambda=0}^{n}$ in die Zielfunktion einbezogen. Das ist auch typisch für andere Verfahren, die die AKF als Zwischenmodell nutzen. Auf diesem Weg wird dem neuen Schätzproblem mehr Information zur Verfügung gestellt, wodurch der Einfluss der schlechten Statistik der AKF etwas abgemildert werden kann. Zu beachten ist aber, dass „viel, hilft viel“ bei AKFs gefährlich ist, denn mit zu großem λ sinkt der relative Informationsgehalt (Betrag der AKF im Verhältnis zur Störung).
Weiterhin ist auf die Abtastrate zu achten. Ist diese zu klein, gibt es Konditionierungsprobleme und Pole nahe Eins; ist sie zu groß, sinkt die AKF zu schnell (relatives Informationsproblem, Begrenzung der Modellordnung).

3.9.3 Restriktionen an die Kovarianzmatrix

Autokovarianzmatrizen, kurz Kovarianzmatrizen, werden unter anderem genutzt, um statische lineare Zusammenhänge zwischen Prozessgrößen zu erkennen. Indem für einzelne Vektorkomponenten nichtlineare Abbildungen anderer Komponenten gewählt werden, also z. B. $x_2 = x_1^2$, können auch nichtlineare Zusammenhänge detektiert werden. Insofern sind Kovarianzmatrizen ein Mittel, um A-priori-Wissen zu generieren. Die entdeckten Zusammenhänge lassen sich zur Prädiktion, zur Reduktion von Messgrößen oder zur Konstruktion aggregierter Größen heranziehen. Aufgrund dieser vielfältigen Einsatzmöglichkeiten und dem Potenzial zur Verbesserung von Kovarianzmatrixschätzungen durch den Einsatz von Restriktionen wird diese Thematik in diesem Abschnitt behandelt.

Die wichtigste Eigenschaft einer Kovarianzmatrix ist ihre nichtnegative Definitheit. Sie ist also symmetrisch oder im komplexen Fall hermitesch und hat nur Eigenwerte, die größer oder höchstens gleich Null sind. Während die Berechnungs- oder Bildungsvorschriften die Symmetrie sicherstellen, können durch schätzfehlerbehaftete AKF-Werte, durch gestörte Daten oder durch numerische Fehler negative Eigenwerte auftreten, die die Kovarianzmatrix indefinit machen. Das ist meist der Fall, wenn nur sehr wenige Werte im Verhältnis zur Größe der Kovarianzmatrix für die Berechnung verfügbar sind oder die zu schätzende Kovarianzmatrix

a priori bereits Nulleigenwerte hat. Neben der Definitheitsrestriktion können aber je nach Anwendung auch weitere Restriktionen an die Kovarianzmatrix gestellt werden, die letztlich auf ein besseres Modell führen. Einige dieser Restriktionen sind nachfolgend aufgeführt:

- Nichtnegative Definitheit
  Sei $\{x[k]\}_{k=1}^{N}$ ein Ausschnitt einer Realisierung eines mittelwertfreien AR-Prozesses, so kann die Kovarianzmatrix für einen Vektor von $\lambda \ll N$ aufeinander folgenden Werten – diese ist eine symmetrische Toeplitz-Matrix – aus den Kovarianzfunktionsschätzwerten

  $$\hat{c}_{\boldsymbol{x}}[\lambda] = \frac{1}{N} \sum_{k=1}^{N-\lambda} x[k]x[k+\lambda]; \qquad \lambda = 0, \ldots, p-1 \tag{3.70}$$

  mit $\hat{\Sigma} = ((\hat{c}_{\boldsymbol{x}}[i-j]))$ bestimmt werden. Abgesehen von $\lambda = 0$ sind die Schätzungen $\hat{c}_{\boldsymbol{x}}[\lambda]$ biasbehaftet, aber zumindest asymptotisch erwartungstreu. Die so geschätzte Kovarianzmatrix ist nichtnegativ definit. Werden die Kovarianzfunktionsschätzer durch die biasfreien Schätzer

  $$\hat{c}_{\boldsymbol{x}}[\lambda] = \frac{1}{N-\lambda} \sum_{k=1}^{N-\lambda} x[k]x[k+\lambda] \tag{3.71}$$

  ersetzt, kann $\hat{\Sigma}$ indefinit werden! Beide Zugänge versagen für $\lambda \to N$.

- Strukturfixiert, aber skalierungsfrei

  $$\Sigma = \sigma^2 A; \qquad A \text{ hat feste Struktur} \tag{3.72}$$

  Weißes Rauschen mit unbekannter Streuung wird nach einem bekannten Filter gemessen. Das Filter bestimmt die Struktur von A. Die ML-Schätzung lautet

  $$\hat{\Sigma} = \frac{\text{spur}(A^{-1}S)}{n} A \tag{3.73}$$

  mit S als gewöhnlicher ML-Kovarianzmatrixschätzer. Ersetze in (3.81) X durch $\sigma^2 A$ und minimiere das freie Problem bezüglich σ^2.

- Affine Struktur von Σ

  $$\Sigma = \sum_{i=0}^{m} \theta_i A_i; \qquad m \leq n^2;\ A_i \text{ sind häufig Binärmatrizen} \tag{3.74}$$

  Beispiele sind symmetrische Toeplitz-Matrizen (stationäre Zeitreihen), Block-Toeplitz-Matrizen (multidimensionale Zufallsprozesse), zirkulante Matrizen[22] (schrittmotorengetriebene Objektträger).

[22] Eine Matrix heißt zirkulant, wenn die Elemente benachbarter Zeilen um eine Position verschoben sind und das dabei herausgeschobene Element als Anfangs- (bei Rechtsverschiebung) oder Endelement (bei Linksverschiebung) erscheint.

Das Modell $\boldsymbol{x} = \mu + K\boldsymbol{f} + \boldsymbol{\varepsilon}$ tritt in der Faktoranalyse auf, wobei $\mathrm{E}\{\boldsymbol{f}\} = 0$, $\mathrm{E}\{\boldsymbol{\varepsilon}\} = 0$, $\operatorname{cov} \boldsymbol{f} = D = \operatorname{diag}_n(d_i)$, $\mathrm{E}\{\boldsymbol{f}\boldsymbol{\varepsilon}^T\} = 0$, $\operatorname{cov} \boldsymbol{\varepsilon} = \sigma^2 I_n$ und $K \in \mathbb{R}^{n\times p}$ als bekannt angenommen wird. Damit ist $\operatorname{cov} \boldsymbol{x} = KDK^T + \sigma^2 I_n$, und die Matrizen A_i haben die Struktur $A_0 = I_n$ und $A_i = KE_{ii}^{(p\times p)}K^T$, sodass $\theta = (\sigma^2, d_1, \ldots, d_p)^T$.
Für Kovarianzmatrizen in Toeplitz-Struktur findet sich in [572] eine Parametrisierung.

- Bekannte Elemente in Σ
  Typisch sind Einsen auf der Hauptdiagonale (Korrelationsmatrix), wenn mit normierten Zufallsgrößen gearbeitet wird, oder Nullen, wenn a priori bekannt ist, das Zufallsgrößen stochastisch unabhängig sind (z. B. aus der Physik heraus oder bei abgeleiteten Größen in der multivariaten Statistik wie x und x^2 oder wie Mittelwert und Streuung)

- Strukturvorgaben an Σ^{-1}
  In zahlreichen Problemen, wie etwa dem verallgemeinerten LS-Problem, wird nicht Σ, sondern Σ^{-1} benötigt. Es ist deshalb mitunter von Vorteil, Σ^{-1} direkt zu schätzen, statt $\hat{\Sigma}$ zu invertieren. Die Mehrzahl der angegebenen Restriktionen können je nach Sachverhalt dann auch an Σ^{-1} gestellt werden. Besonders vorteilhaft erweist sich eine restringierte Schätzung von Σ^{-1}, wenn x ein N-dimensionaler Vektor eines mittelwertfreien AR-Prozesses ist. Wie oben aufgeführt, bereitet die direkte Schätzung aus nur einer Realisierung Schwierigkeiten. Wird hingegen die Restriktion [156]

$$\Sigma^{-1} = \frac{1}{\sigma^2}(U^TU - V^TV) \tag{3.75}$$

  mit den Dreiecks-Toeplitz-Matrizen

$$U = \begin{bmatrix} 1 & & & & & & \\ a_1 & 1 & & & & & \\ \vdots & a_1 & 1 & & & & \\ a_n & & a_1 & \ddots & & & \\ 0 & a_n & & \ddots & \ddots & & \\ \vdots & \ddots & \ddots & & \ddots & 1 & \\ 0 & \ldots & 0 & a_n & \ldots & a_1 & 1 \end{bmatrix}, V = \begin{bmatrix} 0 & & & & & \\ \vdots & \ddots & & & & \\ 0 & & \ddots & & & \\ a_n & \ddots & & \ddots & & \\ \vdots & \ddots & \ddots & & \ddots & \\ a_1 & \ldots & a_n & 0 & \ldots & 0 \end{bmatrix}, \tag{3.76}$$

  genutzt, vereinfacht sich das Problem und verbessert sich die Schätzstatistik. Die benötigten Variablen a_i können über die Yule-Walker-Gleichungen (Differenzengleichung, der die Autokovarianzfunktionswerte genügen) [550] per LS bestimmt werden, während für σ^2 eine klassische Minimum-Varianzschätzung [572] herangezogen werden kann.
  Ergänzend sei auf den Spezialfall des AR-Prozesses erster Ordnung mit $c_{\boldsymbol{x}}[\lambda] = \sigma_{\varepsilon}^2 \varrho^{|\lambda|}$

verwiesen, für den $\Sigma_{\boldsymbol{x}} = \sigma_{\varepsilon}^2((\varrho^{|i-j|}))$ und

$$\Sigma_{\boldsymbol{x}}^{-1} = \frac{1}{(1-\varrho^2)\sigma_{\varepsilon}^2} \begin{bmatrix} 1 & -\varrho & 0 & \dots & 0 & 0 \\ -\varrho & 1+\varrho^2 & -\varrho & \dots & 0 & 0 \\ 0 & -\varrho & 1+\varrho^2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1+\varrho^2 & -\varrho \\ 0 & 0 & 0 & \dots & -\varrho & 1 \end{bmatrix} \tag{3.77}$$

gilt.

- Bekannte Eigenvektoren
  Der Fall bekannter Eigenvektoren u_j – eine zirkulante Kovarianzmatrix hat diskrete Fourier-Vektoren als Eigenvektoren – lässt sich über die dyadische Zerlegung $\Sigma = \sum_{j=1}^{n} d_j u_j u_j^T$ auf den Fall mit affiner Struktur mit $A_i = u_j u_j^T$ zurückführen. Dank der speziellen Struktur existiert eine direkte Lösung

$$\hat{\Sigma} = UDU^T, \qquad d_i = (U^T S U)_{ii}, \qquad U = ((u_j)). \tag{3.78}$$

- Schranken an die Elemente von Σ bzw. Σ^{-1}
  Interessant sind Restriktionen an die Hauptdiagonalelemente, wenn Σ a priori singulär ist. Schranken an Elemente von Σ^{-1} lassen sich geeignet umformulieren

$$e_i^T \Sigma^{-1} e_i \le \gamma \Leftrightarrow \begin{bmatrix} \Sigma & e_i \\ e_i^T & \gamma \end{bmatrix} \succeq 0. \tag{3.79}$$

- Schranken für den maximalen Korrelationskoeffizienten, d. h.

$$\frac{|\sigma_{ij}|}{\sqrt{\sigma_{ii}\sigma_{jj}}} \le \varrho_{\max} \Leftrightarrow \begin{bmatrix} \sqrt{\varrho_{\max}}\sigma_{ii} & \sigma_{ij} \\ \sigma_{ij} & \sqrt{\varrho_{\max}}\sigma_{jj} \end{bmatrix} \succeq 0. \tag{3.80}$$

- Anzahl der Nulleigenwerte
  Die Anzahl der Nulleigenwerte ergibt sich aus Vorwissen über mögliche lineare Abhängigkeiten der Zufallsgrößen oder auch aus der begrenzten Anzahl benutzter Tupel für die Kovarianzschätzung.

- Anzahl gleicher kleinster Eigenwerte
  $\Sigma = U[\mathrm{diag}(\sigma_1^2, \dots, \sigma_p^2, 0, \dots, 0) + \sigma^2 I_n]U^T; U \in \mathcal{O}_n$ ist die Struktur einer Kovarianzmatrix von n Zufallsgrößen (Sensorsignale), die aus p unkorrelierten Quellen resultieren und denen ein stochastisch unabhängiges Rauschen gleicher Varianz (Messstörung, Quantisierungsrauschen) überlagert ist. Gemeinhin ist die Störvarianz kleiner der Nutzvarianz, sodass $n-p$ gleiche Eigenwerte die Kovarianzmatrix prägen. Für die Behandlung dieses Problems sei auf [549] verwiesen.

Nach der Auflistung möglicher Restriktionen, die eine Kovarianzmatrix einhalten muss, werden nachfolgend zwei für viele Restriktionen anwendbare Zugänge vorgestellt:

1. Restringiertes Maximum-Likelihood-Schätzproblem
2. Matrixapproximationsproblem (Projektionsproblem).

Beim ersten Zugang steht die Restriktion direkt im ML-Schätzproblem, beim zweiten wird zunächst ein freies oder ein mit einfachen Restriktionen versehenes Problem gelöst, dessen Lösung a posteriori auf eine Menge restringierter Matrizen projiziert wird. Alternative Formulierungen von Matrixapproximationsproblemen werden in Abschnitt 8.3 vorgestellt.

Restringiertes Maximum-Likelihood-Schätzproblem
Die Lösung des restringierten ML-Schätzproblems vereinfacht sich zu

$$\mathrm{sp}(SX^{-1}) + \ln\det X \stackrel{!}{=} \mathrm{Min} \qquad X \text{ genüge den gestellten Restriktionen,} \tag{3.81}$$

wobei S den ML-Kovarianzmatrixschätzer bezeichnet und X_{opt} die restringierte Schätzung für Σ liefert. Positiv definite Lösungen existieren, wenn S positiv definit ist und die zulässigen X eine abgeschlossene Teilmenge der nichtnegativen Matrizen formen. Sie brauchen nicht eindeutig sein [106]. Die Lösungen sind meist iterativ zu ermitteln. Lässt sich X über einen Vektor θ parametrisieren, dies ist zum Beispiel bei den affinen Restriktionen, der Skalierungsrestriktion aber auch bei der Restriktion "Σ ist eine positiv definite Toeplitz-Matrix" durch

$$X = \begin{bmatrix} 1 & \dots & 1 \\ e^{j\omega_1} & \dots & e^{j\omega_n} \\ \vdots & & \vdots \\ e^{j(n-1)\omega_1} & \dots & e^{j(n-1)\omega_n} \end{bmatrix} \begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \vdots & 0 & d_n \end{bmatrix} \begin{bmatrix} 1 & \dots & 1 \\ e^{-j\omega_1} & \dots & e^{-j\omega_n} \\ \vdots & & \vdots \\ e^{-j(n-1)\omega_1} & \dots & e^{-j(n-1)\omega_n} \end{bmatrix}^T + \sigma^2 I_n \tag{3.82}$$

mit $\theta = (\omega_1, \dots, \omega_n, d_1, \dots, d_n, \sigma^2)^T$ der Fall, dann ergibt sich aus

$$\mathrm{sp}(S[X(\theta)]^{-1}) + \ln\det X(\theta) \stackrel{!}{=} \mathrm{Min} \qquad \theta \in \mathbb{R}^p;\ p \text{ Anzahl der freien Parameter.} \tag{3.83}$$

die notwendige Bedingung für ein Minimum

$$\mathrm{spur}\Big(X^{-1}(\theta)\frac{\partial X(\theta)}{\partial \theta_i}\Big) - \mathrm{spur}\Big([X(\theta)]^{-1}S[X(\theta)]^{-1}\frac{\partial X(\theta)}{\partial \theta_i}\Big) = 0; \qquad i = 1, \dots, p. \tag{3.84}$$

Hiervon ausgehend werden für den Fall affiner Restriktionen Algorithmen abgeleitet, die sich für simultan diagonalisierbare[23] A_i erheblich vereinfachen [22]. Algorithmen für positiv

[23] Matrizen heißen simultan diagonalisierbar, wenn sie mit der gleichen Transformation auf Diagonalform transformiert werden können. Bei einer Ähnlichkeitstransformation erfordert das, dass alle paarweisen Produkte kommutativ sind.

definite Toeplitz-Matrizen werden in [395], [318] diskutiert, wo sich auch Vergleiche mit anderen Zugängen finden. Analog kann mit Restriktionen an Σ^{-1} verfahren werden. Wenn die Restriktionen wie in (3.79) und (3.80) als lineare Matrixungleichungen formulierbar sind, lässt sich (3.81) als ein sog. Max-Det-Problem, s. Tabelle A.5, schreiben, für das effektive Algorithmen existieren [608].

Zugang über Matrixapproximation

Die Formulierung als Matrixapproximationsproblem bezüglich S, wobei S wiederum eine freie ML-Kovarianzmatrixschätzung ist, lautet

$$||X - S||_F \stackrel{!}{=} \text{Min} \qquad X \text{ genüge den gestellten Restriktionen.} \tag{3.85}$$

Für die Restriktion „X sei nichtnegativ definit“ folgt über die Spektralfaktorisierung von $S = U\text{diag}(\lambda_i)U^T$ und mit der Orthogonalitätsinvarianz der Frobenius-Norm unmittelbar

$$\hat{\Sigma} = X_{\text{opt}} = U\text{diag}(\max\{\lambda_i(S), 0\})U^T, \tag{3.86}$$

s. auch [283] für große Probleme. Sind neben der nichtnegativen Definitheit weitere affine Restriktionen gestellt, empfehlen sich zyklische Projektionen, wobei vor der Projektion auf die nichtnegativ definiten Matrizen eine Dykstra-Korrektur vorzunehmen ist [285]. Die Lösungen des Matrixapproximationsproblems sind im Allgemeinen keine restringierten ML-Schätzungen. Zudem müssen die zyklischen Projektionen, Konvergenz vorausgesetzt, insbesondere bei nichtkonvexen Teilrestriktionen (z. B. Rangrestriktionen) nicht gegen ein Minimum konvergieren, vgl. Abschn. 8.2.1.

Für den Ingenieur bleibt festzuhalten, dass der Vorteil, viele Matrixapproximationen geschlossen lösen zu können und damit numerisch schnelle und einfach zu programmierende Algorithmen zu haben, durch den approximativen Charakter der Lösung erkauft wird. Es empfiehlt sich deshalb den restringierten ML-Zugang zu verwenden, gegebenenfalls in Kombination mit einer Startlösung aus dem Matrixapproximationsproblem.

Kapitel 4

Eindeutigkeitserzwingende Restriktionen

Eindeutigkeitserzwingende Restriktionen reduzieren die Menge aller Lösungen auf eine prädestinierte Lösung, wobei bei komplexeren Problemen die Eindeutigkeit meist nur lokal gefordert wird. Searl [560] spricht hierbei von „Restriktionen an die Lösung“. Am wohl bekanntesten ist die Parameterfixierung, bei der ein Parameter auf einen bestimmten Wert, meist die Eins, gesetzt wird. Neben der Parameterfixierung kommen auch andere lineare Restriktionen zum Einsatz. Dabei bietet die Summe-Null-Bedingung, nach der die Summe über alle oder eine bestimmte Anzahl von Parametern Null ist, rechentechnische Vorteile.
Ist die Lösungsmenge eines Problems konvex, wird vielfach als eindeutigkeitserzwingende Lösung diejenige mit der geringsten Norm gewählt, d. h. die Minimumnormlösung

$$||x|| \stackrel{!}{=} \text{Min} \qquad x \in \mathcal{X}_{\text{opt}}. \tag{4.1}$$

Alle streng konvexen Normen garantieren dann Eindeutigkeit, wobei für LS-Lösungsmengen $\mathcal{X}_{\text{opt}}$ die euklidische Norm bevorzugt wird. Die sich dabei ergebende eindeutige (normkleinste) Lösung wird oft Pseudonormallösung genannt.

Bevor aber über die Eindeutigkeit der Lösung eines Problems nachgedacht wird, muss zunächst deren Existenz gegeben sein. So hat $1/x \stackrel{!}{=} \text{Min}$ kein Minimum und folglich keinen Minimierer. An diesem Beispiel ist das sofort klar. Doch für komplizierte Probleme sind Sätze wünschenswert, die die Existenz garantieren. Solche Probleme sind nicht nur akademischer Natur, sondern treten durchaus auf, vgl. das Totale LS-Problem

$$|||[\Delta A, \Delta b]|||_F \stackrel{!}{=} \text{Min}; \;\; (A + \Delta A)x = b + \Delta b. \tag{4.2}$$

mit den Zahlenwerten $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ und $b = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$. $[\Delta A, \Delta b] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \varepsilon & 0 \end{bmatrix}$ und $x_{\text{opt}} = \begin{bmatrix} 3 \\ 1/\varepsilon \end{bmatrix}$ lösen das Gleichungssystem für beliebig kleines ε, ohne das der Grenzwert eine Lösung ist.

Bei der Modellierung entsteht eine solche Situation zumeist infolge einer Überparametrisierung oder durch zu wenig Prozessanregung. Dass diese Fälle nicht generisch sind, also fast nie auftreten, ist erfreulich. Allerdings bewirken auch Probleme, die in der Nähe nichtgenerischer Probleme liegen, numerischer Schwierigkeiten oder hohe Parameterempfindlichkeiten. Insofern kann die strukturelle Kenntnis über Optimierungsprobleme helfen, durch Wahl der Zielfunktion, durch zusätzliche Restriktionen oder durch Verändern der Versuchsstrategie die genannten Schwierigkeiten zu vermeiden und zu besseren Modellen zu gelangen.

In Abschnitt 4.1 wird deshalb zunächst der Satz von Weierstraß diskutiert, der sich der Existenz von Lösungen widmet. Da für die wichtige Klasse der konvexen Probleme weitergehende Aussagen möglich sind, werden diese in Abschnitt 4.1.2 zusammengefasst.

Selbst wenn die Existenz einer Lösung gesichert ist, kann sie immer noch mehrdeutig sein. Statt nun ein solches Problem durch eindeutigkeitserzwingende Restriktionen eindeutig zu machen, ist es geschickter, von vornherein ein anderes, eindeutiges Problem zu formulieren. Konvexe Probleme mit streng konvexen Normen oder noch besser Formulierungen in Hilbert-Räumen sind hier von Vorteil. Die theoretischen Begründungen gibt Abschnitt 4.2.

Oft lässt sich der Einsatz eindeutigkeitserzwingender Restriktionen nicht vermeiden. Für Online-Anwendungen empfiehlt es sich dann, die Restriktion von vornherein zu berücksichtigen. So lassen sich mögliche numerische Probleme vermeiden und die Analyse des Algorithmus wird gemeinhin leichter. Der Charme einer A-posteriori-Einbeziehung der Restriktion liegt hingegen darin, dass ausgehend von der gesamten bekannten Lösungsmenge, die der Anwendung angepasste prädestinierte Lösung ermittelt werden kann. Überdies animieren größere Lösungsmengen zu einem intensiveren Nachdenken über das eigentliche Problem.

Im Abschnitt 4.3 werden die A-priori- und A-posteriori-Identifizierbarkeit untersucht. Die erste sollte eher als strukturelle Identifizierbarkeit bezeichnet werden, da sie fragt, ob die Parameter eines Modells unter idealen Annahmen an die Signale (keine Störung, beliebige Anregung) losgelöst vom Gütekriterium prinzipiell eindeutig bestimmbar sind. Grob gesagt geht es um Minimalparametrisierungen und die Nichtexistenz eines zweiten Modells mit gleichem Verhalten. Bei der A-posteriori-Identifizierbarkeit hingegen wird gefragt, ob bei vorliegender struktureller Identifizierbarkeit auch unter realen Bedingungen, also bezüglich der Daten und des Gütekriteriums, die Eindeutigkeit erhalten bleibt. Da dies aber in enger Beziehung zu den Erkenntnissen aus Abschnitt 4.2 steht, ist die Thematik recht kurz gefasst.

Den Abschluss bildet ein Abschnitt, der eine ganz andere Seite der eindeutigkeitserzwingenden Restriktionen aufzeigt. Stets gibt es bei mehrdeutigen Problemen die Möglichkeit, Eindeutigkeit durch unterschiedliche Restriktionen zu erzwingen. Meist wird dann diejenige Restriktion gewählt, die sich besonders einfach behandeln lässt. Doch diese muss nicht mit der für das Problem zweckmäßigsten Restriktion übereinstimmen. Da es diesbezüglich keine Theorie gibt, sollen ausgewählte Beispiele helfen, den Blick in diese Richtung zu schärfen.

4.1 Existenz von Minima

Einer der bedeutendsten Sätze zur Existenz von Minima ist der Satz von Weierstraß, aus dem wichtige Schlussfolgerungen abgeleitet werden. Da konvexe Probleme schärfere Aussagen gestatten, ist ihnen der darauffolgende Abschnitt gewidmet. Letztlich geht es darum, den Suchraum entweder kompakt zu wählen oder mit einem passenden Gütekriterium über abgeschlossenen Mengen zu arbeiten oder konvexe Problemformulierungen anzustreben.

4.1.1 Satz von Weierstraß

Satz 4.1 (Weierstraß, [71])
Sei $\mathcal{M}$ eine nichtleere Menge in einem metrischen Raum und $f : \mathcal{M} \to \mathbb{R}$ eine auf $\mathcal{M}$ unterhalbstetige Funktion[1], dann garantiert jede der nachfolgenden Bedingungen die Existenz eines Minimierers x_{opt} mit $f(x_{\text{opt}}) = \inf_{x \in \mathcal{M}} f(x)$:

1. $\mathcal{M}$ ist kompakt.
2. $\mathcal{M}$ ist abgeschlossen und f koerzitiv[2]
3. Es existiert ein α, sodass $\mathcal{L}_\alpha \stackrel{\text{def}}{=} \{x \in \mathcal{M} : f(x) \leq \alpha\}$ nichtleer und kompakt ist.[3]

Anmerkung 4.1 Wird in diesem Satz die Halbstetigkeit von unten durch die von oben ersetzt und wird in der dritten Bedingung das Relationszeichen geändert, ergibt sich ein analog lautender Satz für den Maximierer.

Anmerkung 4.2 Auf die Halbstetigkeit von unten kann nicht verzichtet werden.
So hat $f(x) = \begin{cases} 1 & \text{für } x = 0 \\ x^2 & \text{für } x \in (0, 1] \end{cases}$ kein Minimum über der kompakten Menge $\mathcal{M} = [0, 1]$.

Alle drei Bedingungen von Satz 4.1 werden nachfolgend kurz erläutert.

Zur Kompaktheitsbedingung
Eine direkte Schlussfolgerung aus dem Satz lautet: Jede stetige Funktion nimmt auf einer kompakten Menge ein globales Minimum und Maximum an, ist also dort beschränkt.

In Beispiel 4.1 wird nun gezeigt, wie die Kompaktheit des Suchraums (ersten Bedingung in Satz 4.1) genutzt werden kann, um beim orthogonalen Prokrustes-Problem[4] auf die Existenz von mindestens zwei getrennten lokalen Minima zu schließen.

---

[1] f heißt unterhalbstetig in x, wenn $f(x) \leq \liminf_{y \to x} f(y)$.

[2] $f : \mathcal{M} \to \mathbb{R}$ heißt koerzitiv, wenn f nach unten beschränkt ist und $\lim_{||x|| \to \infty} f(x) = \infty$ für alle Folgen gilt.

[3] L_α heißt Subniveaumenge zum Wert α.

[4] Prokrustes war ein Riese und Bösewicht in der griechischen Mythologie, der Reisenden ein Bett anbot. Die Reisenden wurden aber zuvor per Streckung oder Gliedmaßenabhacken dem Bett angepasst. Der Anpassungsgedanke verbunden mit der Restriktion (Bett) gab der Problemklasse den Namen.

Beispiel 4.1 (Orthogonales Prokrustes-Problem)
$||AX - B||_F \stackrel{!}{=} \text{Min}; X \in \mathbb{R}^{n\times n} : X^T X = I_n$ hat mindestens zwei getrennte lokale Minima. Die Menge der orthogonalen Matrizen ist abgeschlossen (algebraische Varietät $X^T X = I_n$) und beschränkt (Teilmenge der Einheitssphäre $X \in \mathbb{C}^{n\times n} : ||X||_2 = 1$). Weiterhin folgt aus $\det X = \pm 1$ und der Stetigkeit der Determinante bezüglich der Matrixelemente, dass die Menge nicht zusammenhängend ist. Beide Teilmengen sind abgeschlossen, da $\det X = 1$ bzw. $\det X = -1$ algebraische Restriktionen darstellen. Die Menge der orthogonalen Matrizen zerfällt also in zwei nicht zusammenhängende kompakte Mengen, was nach dem Satz 4.1 mindestens zwei getrennte lokale Minima impliziert.

Eine konsequente Anwendung der Kompaktheitsbedingung garantiert die Existenz von Lösungen bei metrischen Projektionen auf abgeschlossene Mengen, s. Satz 4.2. Die Bedeutung von Satz 4.2 für die Modellbildung liegt darin, dass sich viele Probleme als Bestapproximationen in diversen Metriken darstellen lassen, wobei es gleichgültig ist, ob über Vektoren, Matrizen oder Polynome minimiert wird [298], [568].

Satz 4.2 (Metrische Projektionen auf abgeschlossene Mengen, [62])
Ist $\mathcal{M}$ eine nichtleere abgeschlossene Menge in einem normierten endlichdimensionalen Vektorraum $\mathcal{V}$, dann ist die Menge aller Projektionen $\text{Proj}_{\mathcal{M},\rho}(p)$ (Punkte mit kürzestem Abstand zu p gemessen in der Norm $||\,.\,||_\rho$) für alle $p \in \mathcal{V}$ nicht leer.

Anmerkung 4.3 Eine metrische Projektion auf eine nicht abgeschlossene Menge[5] kann eine Lösung haben, muss es aber nicht. Trivialerweise sei die Projektion auf $[0, 1)$ genannt, die für $x < 1$ Lösungen hat, für $x \geq 1$ aber nicht. Diese Problematik zeigt sich nicht immer so offen. Sie ist aber typisch für Rang-k-Approximationen (eine Matrizenmenge soll genau den Rang k haben), kann aber durch Formulieren einer Max-Rang-k-Approximation (eine Matrizenmenge soll höchstens den Rang k haben) umgangen werden, s. Beispiel 4.2. Die Menge der Max-Rang-k-Matrizen ist nämlich abgeschlossen.

Beispiel 4.2 (Rang-2-Approximation versus Max-Rang-2-Approximation)
Gesucht ist die nächstgelegene symmetrische Toeplitz-Matrix vom Range 2:

$$\left\| \begin{bmatrix} 2 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 2 \end{bmatrix} - \begin{bmatrix} x_1 & x_2 & x_3 \\ x_2 & x_1 & x_2 \\ x_3 & x_2 & x_1 \end{bmatrix} \right\|_F \stackrel{!}{=} \text{Min} \qquad \text{rg}\, X = 2. \tag{4.3}$$

Die von der geforderten Toeplitz-Struktur abweichenden Elemente sind $a_{12}, a_{21}, a_{23}, a_{32}$. Ihre Mittelung gibt 2, womit die Optimallösung $X_{\text{opt}} = \left[\begin{smallmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{smallmatrix}\right]$ lauten würde, doch dies nicht tut,

[5] Die Begriffe „abgeschlossen“ und „offen“ sind bei Mengen keine logischen Gegensätze wie in der Alltagssprache! Es gibt nämlich Mengen, die weder offen noch abgeschlossen sind (z. B. halboffene Intervalle), und solche, die sowohl offen als auch abgeschlossen sind (z. B. leere Menge).

da sie vom Range 1 ist. Gleichwohl sind in ihrer Umgebung symmetrische Rang-2-Toeplitz-Matrizen. Bei der Restriktion $\operatorname{rg} X \leq 2$ (d. h. $\max \operatorname{rg} X = 2$) tritt dieses Problem nicht auf, denn eine Rang-1-Matrix genügt der Restriktion. Stünden in A im Übrigen statt der Dreien Vieren, gäbe es sowohl für $\operatorname{rg} X = 2$ als auch für $\max \operatorname{rg} X = 2$ eine eindeutige Lösung.

Zur Koerzivitätsbedingung

In vielen Identifikationsproblemen ist die zulässige Menge nicht kompakt. So greift für $f(x) = x^2 \stackrel{!}{=} \text{Min}; x \in \mathbb{R}$ die erste Bedingung nicht, wohl aber die zweite, denn $\mathcal{M} = \mathbb{R}$ ist abgeschlossen und f koerzitiv. Im Speziellen sichert ein konvexes, stetiges, koerzitives f über einer abgeschlossenen konvexen Menge die Existenz einer Lösung.

Die Koerzivitätsforderung sichert durch die Beschränktheit nach unten, dass die Zielfunktion nicht beliebig klein werden kann, und verhindert durch das unbeschränkte Wachsen von f in allen Richtungen, dass f im Unendlichen ein Infimum hat. Was passieren kann, wenn Koerzitivität nicht gegeben ist, zeigen die nachfolgenden Beispiele.

Beispiel 4.3 (Konvexes, nicht-koerzitives Problem)

Eine nicht-koerzitive (da nach unten unbeschränkt) Zielfunktion hat

$$\begin{bmatrix} x_1, x_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \stackrel{!}{=} \text{Min},$$

denn für $x_2 \to -\infty$ strebt sie nach minus Unendlich. Der Fall tritt bei konvexen quadratischen Problemen $x^T A x + 2b^T x + c \stackrel{!}{=} \text{Min}$ mit $A \in \mathcal{S}_n^{\geq}$ immer auf, wenn $b \notin \mathcal{R}(A)$.

Beispiel 4.4 (Konvexes, nicht-koerzitives Problem, [23])

Gegeben sei das konvexe Problem

$$f(X) = \left\| \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ x_{12} & x_{22} \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \right\|_F^2 \stackrel{!}{=} \text{Min} \quad \begin{bmatrix} x_{11} & x_{12} \\ x_{12} & x_{22} \end{bmatrix} \in \mathcal{S}_2^{\geq}$$

$$\text{äquivalent} \qquad x_{11}^2 + (x_{12} - 1)^2 \stackrel{!}{=} \text{Min} \quad x_{11} \geq 0, x_{11} x_{22} - x_{12}^2 \geq 0$$

f ist nicht koerzitiv, da für die nichtnegativen Matrizen $X_k = \begin{bmatrix} 1/k & 1 \\ 1 & k \end{bmatrix}$; $k = 1, 2, \ldots$ aus $\lim\limits_{k\to\infty} ||X_k||_F = \infty$ nicht $\lim\limits_{k\to\infty} f(X_k) = \infty$, sondern $\lim\limits_{k\to\infty} f(X_k) = \lim\limits_{k\to\infty} 1/k^2 = 0$ folgt.

f hat kein Minimum, denn $\inf f(X) = 0$ wird nur für Matrizen mit $x_{11} = 0, x_{12} = 1$ und $x_{22} = \textit{beliebig}$ erfüllt, die aber alle wegen $\det X = -1$ indefinit sind.

Bemerkenswert ist, dass f auf einer konvexen, abgeschlossenen Menge $\mathcal{M}$ konvex und auch nach unten beschränkt ist und trotzdem kein Minimum besitzt. Wie im Beispiel davor ist der Rangdefekt von A die Ursache.

Ergänzend bleibt festzuhalten, dass die Koerzivitätsforderung eine hinreichende oft zu einschränkende Bedingung ist. Als Beispiel hierzu sei das nichtnegativ definite Prokrustes-Problem genannt, für das in [234] schwächere hinreichende und in [633] sogar notwendige und hinreichende Bedingungen für die Lösungsexistenz angegeben werden.

Zur Subniveaumengen-Bedingung
Für das Beispiel $f(x) = x^2 \stackrel{!}{=} \text{Min}; x \in (-2, 2)$ ist weder die erste noch zweite Voraussetzung erfüllt, aber die dritte, denn die Subniveaumenge $\mathcal{L}_1 = [-1, 1]$ ist nichtleer und kompakt.
Eine Abschwächung der Kompaktheitsforderung an die Subniveaumenge auf Beschränktheit ist nicht möglich, vgl. $f(x) = x \stackrel{!}{=} \text{Min}; x \in (0, 1)$. Beschränktheit ist im Allgemeinen eine notwendige Bedingung, vgl. $f(x) = \exp(-x) \stackrel{!}{=} \text{Min}; x \in \mathbb{R}$, aber keine zwingend notwendige Bedingung, vgl. $f(x) = \begin{cases} \exp(-|x|) & x \neq 0 \\ 0 & x = 0 \end{cases}$.
Beschränktheit von L_α auf $\mathcal{M}$ oder auf einer Übermenge von $\mathcal{M}$ impliziert zusammen mit der Stetigkeit von f bei Abgeschlossenheit von $\mathcal{M}$, dass L_α kompakt ist.

4.1.2 Existenzsätze für konvexe Probleme

Da die Formulierung konvexer Probleme, d. h. konvexe Funktionen über konvexen Mengen, viele Vorteile bei der Lösung bietet (konvexe Lösungsmengen, vielfach Eindeutigkeit, effiziente Algorithmen), ist die Kenntnis einiger wichtiger Sätze für konvexe Probleme hilfreich. Sie sagen nämlich nicht nur, wann eine Lösung existiert, sondern auch wo, nämlich für bestimmte Probleme an Extrempunkten und Ecken.

Definition 4.1 (Extrempunkt und Ecke)
$x \in \mathcal{C}$ heißt Extrempunkt einer konvexen Menge $\mathcal{C}$, wenn keine $y, z \in \mathcal{C} \backslash \{x\}$ existieren, sodass $x = \alpha y + (1 - \alpha)z$; $\alpha \in (0, 1)$. Extrempunkte konvexer polyedrischer Mengen werden Ecken genannt.

Damit lässt sich ein erster Satz formulieren, der die anschauliche Tatsache verallgemeinert, wonach eine konvexe Funktion über einem abgeschlossenen Intervall ihr Maximum am Intervallrand annimmt.

Satz 4.3 (Optimalität von Extrempunkten, [71])
Auf kompakten konvexen Mengen wird das Maximum (Minimum) einer konvexen (konkaven) Funktion an einem Extrempunkt angenommen.

Aus diesem Satz ergeben sich unmittelbar die folgenden Schlussfolgerungen.

1. Maximierungen konvexer Funktionen über Einheitskugeln $||x|| \leq 1$ sind denen über Einheitssphären $||x|| = 1$ äquivalent.

2. Ist $f : \mathcal{C} \to \mathbb{R}$ konvex über $\mathcal{C} = \text{conv}\{x_0, \ldots, x_p\}$ (konvexes Polytop in $\mathbb{R}^n$ oder $\mathbb{R}^{m\times n}$), dann gilt $f(x) \leq \max\limits_{i=1,\ldots,p} f(x_i)$ für alle $x \in \mathcal{C}$. Ist f konkav unter sonst gleichen Voraussetzungen, gilt $f(x) \geq \min\limits_{i=1,\ldots,p} f(x_i)$ für alle $x \in \mathcal{C}$ [431].

3. Eine lineare Funktion f nimmt ihr Maximum (Minimum) über einem konvexen Polytop an einer Ecke an. Mit anderen Worten: Unter allen Ecken gibt es wenigstens eine, die ein Element der Menge aller globalen Maximierer (Minimierer) ist.
   Beachte: Die Ecken sind nicht zwingend die einzigen Extremallösungen; betrachte Rechteck mit zu einer Seite parallelen Höhenlinien der lineare Funktion.

Eine zu Schlussfolgerung 3 ähnliche Aussage verzichtet auf die Kompaktheit nach Satz 4.3 und ist als Fundamentalsatz bekannt.

Satz 4.4 (Fundamentalsatz der linearen Optimierung, [71])
Hat ein abgeschlossener konvexer polyedrischer Bereich $\mathcal{C}$ wenigstens eine Ecke und nimmt eine lineare Funktion f ein Minimum (Maximum) über $\mathcal{C}$ an, dann nimmt sie es an einer Ecke an.

Anmerkung 4.4 Der Satz behält für abgeschlossene Quader seine Gültigkeit, wenn f eine multilineare Funktion (z. B. Determinante) ist [51]. Er spielt im Rahmen der Robustheitsanalyse eine wichtige Rolle.

4.2 Eindeutigkeit von Minima

In der Einführung zum Kapitel wurde bereits auf die Bedeutung der Eindeutigkeit von Extremwerten hingewiesen. Das mathematische Rüstzeug, um Modellierungsprobleme so zu formulieren, dass sie eindeutig lösbar sind, wird in den beiden folgenden Abschnitten bereitgestellt. Hierfür werden Sätze formuliert, die die erforderlichen Bedingungen nennen. Die Beispiele vertiefen die Aussagen, sind aber kurz und überschaubar gehalten, um den mathematischen Kern nicht zu verlieren. Der nächste Abschnitt stellt zunächst einige Grundlagen zusammen, die bei den sog. Prokrustes-Problemen in Abschnitt 4.2.2 angewendet werden.

4.2.1 Sätze aus der konvexen Optimierung

Dieser Abschnitt fasst wichtige Sätze zur konvexen Optimierung zusammen, wobei auf einige Verallgemeinerungen des Konvexitätsbegriffs [513] eingegangen wird und nur Gâteaux-Differenzierbarkeit [195] gefordert wird (weniger starke Verallgemeinerung des Ableitungsbegriffs für mehrvariable Funktionen), s. Abschnitt 6.6.1. Statt wie im einvariablen Fall, wo

ein stationärer Punkt über die Forderung $f'(x_{\text{stat}}) = 0$ definiert ist, wird nun für einen stationären Punkt gefordert, dass für das Differenzial $Gf(x_{\text{stat}}; h) = 0$ für alle h gilt.
Eine wichtige Eigenschaft konvexer Funktionen $f : \mathcal{C} \to \mathbb{R}$ über konvexen Mengen $\mathcal{C}$ ist, dass deren Subniveaumengen $\mathcal{L}_\alpha$ konvex sind [71]. Wird also das Niveau immer weiter abgesenkt, dann schrumpft zwar die Subniveaumenge, aber sie bleibt konvex. Ist der kleinste Funktionswert erreicht (wenn dieser für ein x angenommen wird), dann ist die Subniveaumenge immer noch konvex. Diese Aussage wird im nachfolgenden Satz formuliert und erweitert, denn die Menge der Funktionen mit konvexen Subniveaumengen ist eine Obermenge der konvexen Funktionen, nämlich die Menge der sog. quasikonvexen Funktionen[6].

Satz 4.5 (Lösungsmengen konvexer Probleme)
Die Menge der Minimierer einer konvexen Funktion $f : \mathcal{C} \to \mathbb{R}$ ist konvex oder leer. Ist f pseudokonvex[7] und Gâteaux-differenzierbar, dann ist die Menge der stationären Punkte konvex. Die Menge der globalen Minimierer einer quasikonvexen Funktion ist konvex.

Für die Rechtfertigung der Erweiterung wird von der Annahme ausgegangen, dass x_1 und x_2 zwei beliebige globale Minimierer seien, d. h. $f(x_1) = f(x_2) = \min f(x)$. Die Sätze 4.6 bzw. 4.7 garantieren diese Globalitätsannahme für die ersten beiden Teilaussagen des Satzes, für die dritte steckt sie in der Voraussetzung. Globalität impliziert nunmehr $f(x_1) \leq f(\gamma x_1 + (1-\gamma)x_2)$, während Quasikonvexität (schwächste Konvexitätseigenschaft) $f(\gamma x_1 + (1-\gamma)x_2) \leq \max\{f(x_1), f(x_2)\}$ und die Globalitätsannahme $\max\{f(x_1), f(x_2)\} = f(x_1)$ implizieren. Damit gilt $f(x_1) \leq f(\gamma x_1 + (1-\gamma)x_2) \leq f(x_1)$ oder gleichbedeutend $f(\gamma x_1 + (1-\gamma)x_2) = f(x_1) = \min f(x)$, was schlussendlich die Satzaussage bestätigt.

Die Tatsache, dass die Lösungsmenge sogar für streng konvexe Funktionen leer sein kann (Infimumproblem), zeigt das Beispiel $f(x) = \begin{cases} 1, & x = 0 \\ x^2, & 0 < x \leq 1 \end{cases}$. Es widerlegt zudem die Aussage „Eine streng konvexe Funktion hat auf einer abgeschlossenen konvexen Menge genau ein globales Minimum.“. Die präzise Formulierung für den „gemeinten Aussageinhalt“ liefern die nachfolgenden Sätze.

Satz 4.6 (Globale Optimalität lokaler Minimierer, [513])
Jeder lokale Minimierer eines konvexen f, eines stetigen pseudokonvexen f oder eines streng quasikonvexen f über einer konvexen Menge ist ein globaler.

[6] Eine Funktion heißt quasikonvex, wenn alle Subniveaumengen $\mathcal{L}_\alpha = \{x \in \mathcal{V} : f(x) \leq \alpha\}$ konvex sind, bzw. äquivalent, wenn $\forall \gamma \in [0,1] : f(\gamma x + (1-\gamma)y) \leq \max\{f(x), f(y)\}$ [195]. Sie heißt streng quasikonvex, wenn die vorgenannte Ungleichung für alle $\gamma \in (0,1)$ und $x \neq y$ streng gilt.

[7] Eine Funktion $f : \mathcal{C} \to \mathbb{R}$ heißt pseudokonvex, wenn $\forall x, y \in \mathcal{C}, \exists c \in (0,1], \exists \alpha > 0$, sodass

$$f(x) > f(y) \Rightarrow f((1-\gamma)x + \gamma y) \leq f(x) - \gamma c \quad \forall \gamma \in [0, \alpha].$$

Anmerkung 4.5 Die Einschränkung „über einer konvexen Menge“ ist wichtig, da $f(x) = x^2$ über der nichtkonvexen Menge $\mathcal{M} = [-1, 1] \cup [2, 3]$ in $x_{\text{loc}} = 2$ einen lokalen Minimierer hat, der nicht der globale ist.

Satz 4.7 (Globale Optimalität stationärer Punkte, [487])
Jeder stationäre Punkt x_{stat} für pseudokonvexes und damit auch für ein konvexes $f : \mathcal{C} \to \mathbb{R}$, das in x_{stat} Gâteaux-differenzierbar ist, ist ein globaler Minimierer. Für streng pseudokonvexes und damit auch für streng konvexes f ist der globale Minimierer eindeutig.

Die Sätze 4.6 und 4.7 gelten nicht für das quasikonvexe $f(x) = \begin{cases} |x| & |x| \leq 1 \\ 1 & |x| > 1, \end{cases}$ betrachte z. B. $x_{\text{stat}} = 2$. Für diese schwächere Form von Konvexität gilt aber der folgende Satz.

Satz 4.8 (Eindeutigkeit streng quasikonvexer Probleme, [487])
Ein quasikonvexes f hat auf einer konvexen Menge höchstens einen isolierten lokalen Minimierer; ein streng quasikonvexes f höchstens einen lokalen Minimierer. Sofern für streng quasikonvexes f ein lokaler Minimierer existiert, ist dieser der eindeutig bestimmte globale Minimierer. Für stetiges, streng quasikonvexes und koerzitives f ist die Existenz eines eindeutigen globalen Minimierers gesichert[8].

Als eine klassische Anwendung für diese Sätze erweisen sich Projektionsprobleme. Dabei geht es darum, zu einem in einem normierten Raum $\mathcal{V}$ gegebenen Punkt x die Menge $\text{Proj}_{\mathcal{M},\rho}(p)$ aller zu p nächstgelegenen Punkte x_{opt} in einer Menge $\mathcal{M} \subset \mathcal{V}$ zu bestimmen

$$\text{Proj}_{\mathcal{M},\rho}(p) = \underset{x \in \mathcal{M}}{\text{argmin}} ||p - x||_\rho. \tag{4.4}$$

Für die Existenz einer Lösung ist Abgeschlossenheit von $\mathcal{M}$ hinreichend, da Koerzivität durch die Struktur der Zielfunktion sichergestellt wird. Dabei versteckt sich die Abgeschlossenheit mitunter in der Mengenbezeichnung. So ist nämlich ein linearer Unterraum $\mathcal{M}$ in einem endlichdimensionalen Raum $\mathcal{V}$ per se abgeschlossen. Zudem ist die Zielfunktion wegen der Konvexität der Normen über jeder konvexen Teilmenge von $\mathcal{M}$ konvex. Ist dann zusätzlich die Menge noch konvex, greifen die angeführten Sätze. Problematisch erscheint die Eindeutigkeit, da keine Norm eine streng konvexe Funktion ist, da die strenge Ungleichheit für strenge Konvexität nicht nur bei $x = y$, sondern auch bei $y = \beta x; \beta \geq 0$ verletzt wird. Dessen ungeachtet wird die folgende Definition verwendet.

Definition 4.2 (Streng konvexe Norm, [299])
Eine Norm heißt streng konvex, wenn gilt

$$||\gamma\, x + (1 - \gamma)\, y|| < \gamma\, ||x|| + (1 - \gamma)\, ||y|| \qquad \gamma \in (0, 1); \forall x \neq 0_{\mathcal{L}}, y \notin \{\beta\, x : \beta \geq 0\}. \tag{4.5}$$

[8] $f(x) = |x|$ hat z. B. diese Eigenschaften.

In der Definition werden die problematischen Punkte einfach ausgeschlossen[9]. Aus der Definition ergibt sich aber, dass jede streng konvexe Norm eine streng quasikonvexe Funktion ist. Für die unproblematischen Punkte folgt $||\gamma\, x + (1-\gamma)\, y|| < \gamma\, ||x|| + (1-\gamma)\, ||y|| \leq \max\{||x||, ||y||\}$ für $x \neq y$ und für die problematischen Punkte gilt

$$||\gamma x + (1-\gamma)\beta x|| = (\gamma + (1-\gamma)\beta)||x|| < \begin{cases} ||x|| & 0 \leq \beta < 1 \\ \beta||x|| & \beta > 1 \end{cases} = \max\{||x||, \beta||x||\},$$

wobei $\beta = 1$ wegen $x \neq y$ nicht betrachtet werden muss.

Mit der strengen Quasikonvexität streng konvexer Normen folgt die Eindeutigkeit metrischer Projektionen auf konvexe Mengen und im Speziellen auf lineare Unterräume in endlichdimensionalen Räumen. Streng konvexe Normen sind deshalb für die eindeutigkeitserzwingende Minimum-Norm-Restriktion bei konvexen Lösungsmengen prädestiniert.

Korollar 4.1 (Eindeutigkeit metrischer Projektionen)
Ist $\mathcal{C}$ eine nichtleere abgeschlossene konvexe Menge und $||.||_\rho$ eine streng konvexe Norm, dann ist die Projektion eines Punkts p auf $\mathcal{C}$ eindeutig, d. h. $x_{\text{opt}} = \text{Proj}_{\mathcal{C},\rho}(p)$.

Wird für p der Nullpunkt des Raumes gewählt, dann ist x_{opt} das eindeutig bestimmte normkleinste Element der Menge $\mathcal{C}$.

Korollar 4.2 (Normkleinstes Element konvexer Mengen)
Jede nichtleere abgeschlossene konvexe Menge $\mathcal{C}$ in einem streng normierten Raum $(\mathcal{R}, ||.||_\rho)$ hat ein eindeutig bestimmtes normkleinstes Element $x_{\text{opt}} = \text{Proj}_{\mathcal{C},\rho}(0_\mathcal{R})$.

Nachteilig an den metrischen Projektionen ist, dass wenige allgemeingültige, gut handhabbare Charakterisierungen für die Optimallösung existieren. Existenz und Eindeutigkeit sind zwar wichtig, für eine effiziente Berechnung aber oft zu wenig. Deshalb wird in der Anwendung oft die klassische Projektion in Hilbert-Räumen bevorzugt, wo sich die Lösung durch die sogenannten Projektionssätze näher spezifizieren lässt.

Satz 4.9 (Projektionssatz)
Ist $\mathcal{C} \subset \mathcal{H}(\mathbb{C})$ eine nichtleere abgeschlossene konvexe Menge in einem Prä-Hilbert-Raum und $x \in \mathcal{H}(\mathbb{C})$ ein beliebiger Punkt, dann existiert eine eindeutige Projektion $x_{\text{opt}} \in \mathcal{C}$ mit $||x - x_{\text{opt}}|| < ||x - y||$ für alle $y \in \mathcal{C} \backslash x_{\text{opt}}$ [50]. Der Punkt $x_{\text{opt}} = \text{Proj}_\mathcal{C}(x)$ ist durch die

[9] Aus diesem Grund wird in [62] auch von im Wesentlichen streng konvexen Normen gesprochen, was an sich präziser ist. Diese längliche Bezeichnung hat sich aber nicht durchgesetzt.
Äquivalente Definition: Eine Norm heißt streng konvex, wenn ihre Normkugeln Rotunde sind (konvexe Mengen ohne Kanten und Fassetten, wo jeder Randpunkt Extrempunkt ist).

Bedingungen $x_{\text{opt}} \in \mathcal{C}$ und

$$\begin{aligned}
\text{für } \mathcal{C} \text{ konvex}^{10} \quad & \forall y \in \mathcal{C} : \Re e\, \langle x - x_{\text{opt}}, y - x_{\text{opt}} \rangle \leq 0 && [62]\\
\text{konvexer Kegel} \quad & \forall y \in \mathcal{C} : \langle x - x_{\text{opt}}, x_{\text{opt}} \rangle = 0 \text{ und } \Re e\, \langle x - x_{\text{opt}}, y \rangle \leq 0 && [50]\\
\text{affiner Unterraum} \quad & \forall y \in \mathcal{C} = \mathcal{L}_a : \langle x - x_{\text{opt}}, a - y \rangle = 0 \\
& \Leftrightarrow x_{\text{opt}} = a + \text{Proj}_{\mathcal{L}}(x - a) && [454]\\
\text{linearer Unterraum}^{11} \quad & \forall y \in \mathcal{C} = \mathcal{L} : \langle x - x_{\text{opt}}, y \rangle = 0,\ \text{d. h. } x - x_{\text{opt}} \perp \mathcal{L} && [417]\\
& \Leftrightarrow \langle x - x_{\text{opt}}, x_{\text{opt}} \rangle = 0,\ \text{d. h. } x - x_{\text{opt}} \perp x_{\text{opt}} && [62]\\
& \Leftrightarrow x_{\text{opt}} = x_1,\ \text{wobei } x = x_1 + x_2 \text{ mit } x_1 \in \mathcal{L}, x_2 \in \mathcal{L}^{\perp} && [62]\\
& \Leftrightarrow x_{\text{opt}} = \sum_{i=1}^{n} \frac{\langle x, l_i \rangle}{\langle l_i, l_i \rangle} l_i; \mathcal{L} = \text{lin}\{l_1, \ldots, l_n\}; l_i \perp l_j; i \neq j\ /^{12}
\end{aligned}$$

eindeutig charakterisiert.

Beispiel 4.5 (Nächstgelegene symmetrische Matrix)
Da die symmetrischen Matrizen $\mathcal{S}_n$ einen linearen Unterraum in $\mathbb{R}^{n \times n}$ formen, folgt für $||A - X||_F \stackrel{!}{=} \text{Min}; X \in \mathcal{S}_n$ mit der erstgenannten Projektionsformel für lineare Unterräume

$$\begin{aligned}
0 &= \text{spur}((A - X_{\text{opt}})Y) \qquad \forall Y \in \mathcal{S}_n \qquad \text{mit } \langle X, Y \rangle \stackrel{\text{def}}{=} \text{spur}(XY^T)\\
&= \text{spur}((\tfrac{1}{2}(A + A^T) + \tfrac{1}{2}(A - A^T) - X_{\text{opt}})Y)\\
&= \text{spur}((\tfrac{1}{2}(A + A^T) - X_{\text{opt}})Y) + \text{spur}(\tfrac{1}{2}(A - A^T)Y)\\
&= \text{spur}((\tfrac{1}{2}(A + A^T) - X_{\text{opt}})Y) \qquad \Rightarrow X_{\text{opt}} = \tfrac{1}{2}(A + A^T).
\end{aligned}$$

Die Implikation folgt, da $\text{spur}(BY) = 0; \forall Y \in \mathcal{S}_n$ nur für $B = 0_{n \times n}$ gelten kann. Für $B \neq 0_{n \times n}$ verhindert nämlich $Y = B^T$ wegen $\text{spur}(BB^T) > 0$ (vollständige Quadrate), dass $\text{spur}(BY) = 0$ gilt.
Mit der dritten Projektionsformel für lineare Unterräume und der Kenntnis, dass $\mathcal{S}_n^{\perp}$ gerade die schiefsymmetrischen Matrizen sind, folgt aus der Zerlegung $A = \frac{1}{2}(A + A^T) + \frac{1}{2}(A - A^T)$ mit $\frac{1}{2}(A + A^T) \in \mathcal{S}_n$ und $\frac{1}{2}(A - A^T) \in \mathcal{S}_n^{\perp}$ die angegebene Lösung X_{opt} sofort. Alternativ kann sie aus dem Pythagoras $||A - X||_F^2 = ||\frac{1}{2}(A - A^T)||_F^2 + ||\frac{1}{2}(A + A^T) - X||_F^2 \geq ||\frac{1}{2}(A - A^T)||_F^2$ mit Gleichheit für das gesuchte X_{opt} abgelesen werden.

---

[10] Für allgemeine konvexe Mengen nennt sich dieser Satz auch Kolmogorov-Kriterium. Geometrisch bedeutet $\Re e\, \langle x - x_{\text{opt}}, y - x_{\text{opt}} \rangle \leq 0$ für $x \notin \mathcal{C}$ einen rechten bzw. stumpfen Winkel $\angle(x, x_{\text{opt}}, y)$, d. h. alle $y \in \mathcal{C}$ liegen auf oder auf der entgegengesetzten Seite der Tangentialebene durch x_{opt} wie x.

[11] Endlichdimensionale Unterräume sind per se abgeschlossen. Sind die linearen Unterräume nicht abgeschlossen, dann gilt: Wenn ein minimierendes $x_{\text{opt}} \in \mathcal{L}$ existiert, ist es eindeutig und durch $x - x_{\text{opt}} \perp \mathcal{L}$ charakterisiert [417].

[12] Die Faktoren $\langle x, l_i \rangle / \langle l_i, l_i \rangle$ heißen Fourier-Koeffizienten.

4.2.2 Sätze für Prokrustes-Probleme

Dieser Abschnitt widmet sich den Prokrustes-Problemen, womit Matrix-LS-Probleme mit Restriktionen an die Matrix bezeichnet werden. Prokrustes-Probleme finden praktische Anwendungen bei der Modellreduktion, -approximation und -identifikation. Der Schwierigkeitsgrad und die Lösungsmethode hängen von der betreffenden Restriktion ab, weshalb sich in den Publikationen [284], [283], [68], [62], [299], [178], [235], [527], [201] zumeist nur auf eine bestimmte Restriktion beschränkt wird. In den Publikationen werden auch konkrete Anwendungen genannt. Ziel dieses Abschnitts ist es, Prokrustes-Probleme zu lösen, bei denen die Restriktion sehr allgemein gefasst ist, sodass durch Einsetzen einer speziellen, verträglichen Restriktion die Lösung unmittelbar geschlussfolgert werden kann.
Den Anfang bilden zwei Sätze für das freie LS-Probleme, die die aus der Identifikation bekannte Bedingung nach vollem Spaltenrang der Datenmatrix A herausstreichen.

Satz 4.10 (Kompaktheit der Subniveaumengen der LS)
Die $\mathcal{L}_\alpha$ von $f(X)=\frac{1}{2}||AX-B||_F^2$; $X\in\mathbb{R}^{m\times n}$, $A\in\mathbb{R}^{p\times m}$, $B\in\mathbb{R}^{p\times n}$ sind genau dann kompakt, wenn A vollen Spaltenrang hat.

Der Satz ergibt sich als Schlussfolgerung aus einem Satz in [71], wonach alle Subniveaumengen einer konvexen Funktion kompakt sind, sobald eine Subniveaumenge kompakt ist. Eine solche kompakte Subniveaumenge ist der eindeutige globale Minimierer (einelementige Mengen sind kompakt), dessen Existenz und Eindeutigkeit durch den vollen Spaltenrang garantiert wird.

Satz 4.11 (Strenge Konvexität der LS)
Für $A\in\mathbb{R}^{p\times m}$, $B\in\mathbb{R}^{p\times n}$ ist $f(X)=\frac{1}{2}||AX-B||_F^2$ eine konvexe Funktion bezüglich $X\in\mathbb{R}^{m\times n}$, die genau dann streng konvex ist, wenn A vollen Spaltenrang hat.[13]

Durch Kombination der Aussagen aus Satz 4.10 und 4.11 mit den Eigenschaften abgeschlossener bzw. konvexer Mengen und Aussagen zur Optimalität aus Abschnitt 4.2.1 entstehen vier Sätze für große Klassen von Prokrustes-Problemen, und zwar für die Klassen: abgeschlossene Menge, konvexe Menge, konvexer Kegel und affiner Raum.

[13] Konvexität folgt direkt aus

$$\gamma||AX-B||_F^2+(1-\gamma)||AY-B||_F^2-||A[\gamma X+(1-\gamma)Y]-B||_F^2=\gamma(1-\gamma)||A(X-Y)||_F^2\geq 0.$$

Zudem gilt $A(X-Y)\neq 0_{p\times n}$ für $X\neq Y$ und Vollrang A, womit die Ungleichung für $\gamma\in(0,1)$ streng ist. Umgekehrt lassen sich bei rangdefektivem A Matrizen X,Y mit $X\neq Y$ mit $A(X-Y)=0_{p\times n}$ konstruieren. Alternativ kann $f(X)$ als Summe quadratischer Funktionen $\tilde{f}(x)=\frac{1}{2}x^TQx+r^Tx+s$ geschrieben werden, die für $Q\succeq 0_{m\times m}$ konvex und für $Q\succ 0_{m\times m}$ streng konvex sind [71].

Satz 4.12 (Prokrustes-Probleme über abgeschlossenen Mengen)
Das Problem

$$\frac{1}{2}||AX - B||_F^2 \stackrel{!}{=} \text{Min} \qquad X \in \mathcal{M} \subseteq \mathbb{R}^{m\times n}; A \in \mathbb{R}^{p\times m}, B \in \mathbb{R}^{p\times n} \tag{4.6}$$

hat für nichtleeres, abgeschlossenes $\mathcal{M}$ mindestens eine Lösung, wenn A spaltenregulär ist.[14]

Satz 4.13 (Prokrustes-Probleme über konvexen Mengen, [23])
Sofern

$$\frac{1}{2}||AX - B||_F^2 \stackrel{!}{=} \text{Min} \qquad X \in \mathcal{C} \subseteq \mathbb{R}^{m\times n}; A \in \mathbb{R}^{p\times m}, B \in \mathbb{R}^{p\times n} \tag{4.7}$$

für nichtleeres, abgeschlossenes konvexes $\mathcal{C}$ eine Lösung hat, ist die Lösungsmenge konvex. Eine notwendige und hinreichende Bedingung für eine Lösung $X_{\text{opt}} \in \mathcal{C}$ lautet

$$\forall X \in \mathcal{C}: \quad X_{\text{opt}} + X \in \mathcal{C} \Rightarrow \text{spur}(A^T(AX_{\text{opt}} - B)X^T) \geq 0. \tag{4.8}$$

Für spaltenreguläres A ist die Lösung eindeutig[15]

Dass auf die Existenz einer Lösung in der Voraussetzung dieses Satzes nicht verzichtet werden kann, zeigte Beispiel 4.4. Für nichtleere, konvexe Kegel (nichtnegative Matrizen, nichtnegativ definite Matrizen, schwach besetzte Matrizen, usw.) kann die Lösung näher spezifiziert werden, wobei das Konzept der positiven Polarkegel (Dualkegel)

$$\mathcal{C}^{\oplus} \stackrel{\text{def}}{=} \{y \in \mathcal{H}(\mathbb{R}) : \langle x, y\rangle \geq 0, \forall x \in \mathcal{C}\} \tag{4.9}$$

auf den Hilbert-Raum $(\mathbb{R}^{m\times n}, \text{spur}(XY^T))$ bezogen wird.

Satz 4.14 (Prokrustes-Probleme über konvexen Kegeln, [538], [23])
Das Problem

$$\frac{1}{2}||AX - B||_F^2 \stackrel{!}{=} \text{Min} \qquad X \in \mathcal{C} \subseteq \mathbb{R}^{m\times n}; A \in \mathbb{R}^{p\times m}, B \in \mathbb{R}^{p\times n} \tag{4.10}$$

hat für nichtleere abgeschlossene konvexe Kegel $\mathcal{C}$ genau dann eine Lösung X_{opt}, wenn

$$X_{\text{opt}} \in \mathcal{C}, \qquad A^T(AX_{\text{opt}} - B) \in \mathcal{C}^{\oplus}, \qquad \text{spur}(A^T(AX_{\text{opt}} - B)X_{\text{opt}}^T) = 0. \tag{4.11}$$

Für spaltenreguläres A ist die Lösung eindeutig.[16]

[14] Die Satzaussage ergibt sich direkt aus der Kombination der Kompaktheitsaussage zu den Subniveaumengen nach Satz 4.10 und dem Weierstraß-Satz 4.1.

[15] $\text{spur}(A^T(AX_{\text{opt}} - B)X^T)$ bezeichnet das Differenzial $Df(X_{\text{opt}}; H)$, das laut (4.8) in alle zulässigen Richtungen H zunehmen muss.

[16] In [538] wird ein allgemeinerer Satz über die Fenchel-Dualität bewiesen, während in [23] elementare Konvexitätseigenschaften zum Beweis genutzt werden. Strenge Konvexität von $\frac{1}{2}||AX-B||_F^2$ bei spaltenregulärem A impliziert Eindeutigkeit, vgl. Satz 4.11.

Anmerkung 4.6 Der Satz 4.14 bleibt in ähnlicher Weise gültig, wenn der Ursprung des konvexen Kegels in X_0 verschoben wird, d. h., wenn $\mathcal{C}-X_0$ ein abgeschlossener konvexer Kegel ist. Dies wird insbesondere im typischen Fall der affinen Räume $\mathcal{L}_{X_0} = X_0+\mathcal{L}$ deutlich, wobei $\mathcal{L}$ für einen linearen Raum steht. Diese Menge umfasst alle Matrizen, bei denen bestimmte Elemente fest, aber nicht notwendig Null sind und schließt die linearen Räume mit ein (symmetrische, schiefsymmetrische Matrizen, Toeplitz- und Hankel-Matrizen).

Der nachfolgende Satz ist ein Korollar des Satzes 4.14, da lineare Räume $\mathcal{L}$ konvexe Kegel mit $\mathcal{C}^{\oplus} = \mathcal{L}^{\perp}$ sind, und sich affine Räume durch Subtraktion des Stützpunkts X_0 in linear überführen lassen, s. Anmerkung 4.6.

Satz 4.15 (Prokrustes-Probleme über affinen Räumen)
Das Problem

$$\frac{1}{2}||AX - B||_F^2 \stackrel{!}{=} \text{Min} \qquad X \in (X_0 + \mathcal{L}) \subseteq \mathbb{R}^{m\times n}; A \in \mathbb{R}^{p\times m}, B \in \mathbb{R}^{p\times n} \tag{4.12}$$

hat für lineare Räume $\mathcal{L}$ eine affine Lösungsmenge $\mathcal{X}_{\text{opt}}$, die durch

$$\mathcal{X}_{\text{opt}} = \{X \in \mathcal{L}_{X_0} : \; X - X_0 \in \mathcal{L},\, A^T(AX - B) \in \mathcal{L}^{\perp}\} \tag{4.13}$$

gekennzeichnet ist. Für spaltenreguläres A ist die Lösung eindeutig.

Der Beweis des Satzes 4.15 kann auch ohne Zuhilfenahme von Satz 4.14 direkt mit dem $\mathcal{L}$-Gradienten (6.29) erfolgen, das auf

$$\frac{1}{2}||AY - (B - AX_0)||_F^2 \stackrel{!}{=} \text{Min} \qquad Y \in \mathcal{L} \subseteq \mathbb{R}^{m\times n}; A \in \mathbb{R}^{p\times m}, B \in \mathbb{R}^{p\times n} \tag{4.14}$$

mit $Y = X - X_0$ angewendet wird. Die Stationaritätsbedingung lautet damit

$$\text{Proj}_{\mathcal{L}}\Big(A^T(AY - (B - AX_0))\Big) = 0_{m\times n}, \tag{4.15}$$

wobei $\text{Proj}_{\mathcal{L}}(.)$ den Gradienten der Zielfunktion orthogonal auf $\mathcal{L}$ projiziert[17]. Die Projektion ist genau dann Null, wenn die zu projizierende Größe im orthogonalen Komplement $\mathcal{L}^{\perp}$ liegt, also wenn $A^T(AY - (B - AX_0)) \in \mathcal{L}^{\perp}$ bzw. nach Resubstitution $A^T(AX - B) \in \mathcal{L}^{\perp}$ gilt. Ferner ist $X - X_0 \in \mathcal{L}$ die geforderte lineare Restriktion.

Die Bedeutung der hier angeführten Sätze ist für den Ingenieur vielschichtig. Aus Satz 4.10 kann nämlich gefolgert werden, dass ein voller Spaltenrang von A für restringierte LS-Probleme hinreichend für beschränkte Subniveaumengen ist. Bei abgeschlossenen zulässigen Mengen ergeben sich sogar kompakte Subniveaumengen. Nach Satz 4.11 ist der volle Spaltenrang von A notwendig und hinreichend für strenge Konvexität und damit für Eindeutigkeit

[17] $\text{Proj}_{\mathcal{L}}(X) \stackrel{\text{def}}{=} \text{argmin}_{Y\in\mathcal{L}}||X - Y||_F$

der Lösung. Zudem lässt sich Satz 4.11 dahingehend verallgemeinern, dass $f(X) = ||AX-B||$ für jede streng konvexe Norm bei spaltenregulärem A streng quasikonvex und koerzitiv ist[18], was wiederum Eindeutigkeit sichert. Folglich haben alle l_p-Probleme mit $1 < p < \infty$ bei spaltenregulärem A eine eindeutige Lösung. Für die l_1- und l_∞-Approximationen gilt diese schöne Eigenschaft leider nicht. Die Sätze 4.13 bis 4.15 liefern Charakterisierungen der Lösung, d. h., sie sagen konkret, welche Bedingungen die Lösung erfüllen muss. Eine solche Aussagekraft gelingt dank des Bezugs auf einen Hilbert-Raum. Nur Satz 4.12 fällt in seiner Aussagekraft schwächer aus, da die abgeschlossenen Mengen wenig innere Struktur aufweisen. Aber auch er unterstreicht die Bedeutung des vollen Spaltenrangs von A, sichert er doch für Prokrustes-Probleme auf abgeschlossenen Mengen immerhin die Existenz einer Lösung.

4.3 Identifizierbarkeit

4.3.1 Strukturelle Identifizierbarkeit

Bei der Frage nach der strukturellen Identifizierbarkeit (auch theoretische Identifizierbarkeit oder A-priori-Identifizierbarkeit genannt) wird ein deterministisches parametrisches oder approximativparametrisches Modell dahingehend untersucht, ob sich ein Parametervektor eindeutig bestimmen lässt. Dabei kann das Eingangssignal bestmöglich (rauschfrei, vollständige Systemanregung, Messen ohne Quantisierung, passendes Verfahren) gewählt werden. Diese Annahmen sind natürlich streng, dafür kann die strukturelle Identifizierbarkeit vor dem eigentlichen Experiment überprüft werden. Sie ist eine notwendige Voraussetzung für Identifizierbarkeit unter praktischen Bedingungen.

Definition 4.3 (Punktweise Identifizierbarkeit, [155], [296])
Das Modell

$$\dot{x}(t;\theta) = f(x,u;\theta) \qquad x_0 = x(t_0;\theta); \qquad t \geq t_0 \tag{4.16a}$$

$$y(t;\theta) = h(x,u;\theta) \qquad \theta \in \mathcal{F} \tag{4.16b}$$

mit eindeutigen Lösungen[19] für alle $\theta \in \mathcal{F}$ heißt im Punkt θ eindeutig identifizierbar, wenn ein Signal $u(t); t \geq t_0$ existiert, sodass gilt

$$\forall x_0 \in \mathbb{R}^n, \forall \tilde{\theta} \in \mathcal{F}: \quad y(t,u;\tilde{\theta}) \equiv y(t,u,\theta) \quad \Rightarrow \quad \tilde{\theta} = \theta \tag{4.17}$$

[18] $\forall X \neq Y : ||A(\gamma X + (1-\gamma)Y) - B|| = ||\gamma(AX-B) + (1-\gamma)(AY-B)|| < \max\{||AX-B||, ||AY-B||\}$ für $\gamma \in (0,1)$ zeigt strenge Quasikonvexität.

[19] Ohne die Eindeutigkeit ist statt (4.17) zu fordern, dass zunächst in θ eine Lösung existiert und dass $\tilde{\theta} = \theta$ impliziert wird, wenn der Schnitt der Lösungsmengen für θ und der für $\tilde{\theta}$ nicht leer ist. Nichteindeutigkeit kann bei geometrischen Problemen auftreten, vgl. $\dot{x}(\dot{x}-x) = 0$; $\theta = x_0$ mit $x(t) \equiv \theta$ und $x(t) = \theta e^t$.

oder äquivalent

$$\forall x_0 \in \mathbb{R}^n, \forall \tilde{\theta} \in \mathcal{F}: \quad \tilde{\theta} \neq \theta \quad \Rightarrow \quad y(t,u;\tilde{\theta}) \not\equiv y(t,u,\theta). \tag{4.18}$$

Für autonome System entfällt konsequenterweise die Existenzforderung an $u(t)$, für statische Systeme die Forderung $\forall x_0 \in \mathbb{R}^n$. Gilt die Implikation mit der Abschwächung $\forall \tilde{\theta} \in \mathcal{U}_\varepsilon(\theta)$, dann heißt das Modell in θ lokal eindeutig identifizierbar. Ist das Modell in θ nicht lokal eindeutig identifizierbar, so heißt es in θ nicht identifizierbar.

Die konsequente Erweiterung der punktweisen Identifizierbarkeit stellt die strukturelle Identifizierbarkeit dar, die sich nunmehr auf alle Punkte θ bezieht, oder um präziser zu sein, auf fast alle. Diese Einschränkung ist sinnvoll, da in den meisten Fällen eine Menge atypischer Punkte existiert, in denen die Identifizierbarkeit verloren geht. Ein anderes Problem älterer Definitionen zur strukturellen Identifizierbarkeit betrifft den Bezug auf ein gegebenes u und ein gegebenes x_0 [60], [243]. Dabei widerspricht die Einschränkung des u der Vorstellung, dass strukturelle Identifizierbarkeit die Frage beantworten soll, ob es prinzipiell möglich ist, die Parameter zu identifizieren. Die Frage, ob es mit einem bestimmten Signal möglich ist, ist eine andere, verwandte und gleichfalls interessante Frage. Für lineare Systeme sind die angesprochenen Feinheiten ohnehin weniger von Bedeutung.

Definition 4.4 (Strukturelle Identifizierbarkeit, [155], [296])
Das Modell (4.16) heißt global (lokal) strukturell identifizierbar, wenn es in allen Punkten $\theta \in \mathcal{F}$ bis auf eine magere Teilmenge punktweise (lokal) eindeutig identifizierbar ist. Es heißt explizit strukturell identifizierbar, wenn für jede Komponente des Parametervektors ein geschlossener Formelausdruck aus den Messsignalen existiert (keine Reihendarstellung, keine implizites Gleichungssystem, keine algorithmische Formulierung)[20].

Beispiel 4.6 (Strukturelle Identifizierbarkeit)
Eine SISO-Übertragungsfunktion ist in allen Punkten $(1, a_1, \ldots, a_n, b_0, \ldots, b_m)$, in denen es zu einer Pol-Nullstellenkürzung kommt, nicht identifizierbar. Da aber die Menge der Pol-Nullstellen-kürzbaren Punkte mager in $\mathbb{R}^n \times \mathbb{R}^{m+1}$ ist, liegt globale strukturelle Identifizierbarkeit vor.
$\dot{x} = \theta_1(x^2 - 1) - \theta_2$ ist für alle $\theta \in \mathcal{F}_1 = \{\theta \in \mathbb{R}^2 : \theta_1(\theta_1 + \theta_2) > 0\}$ nicht eindeutig identifizierbar, da (4.17) nicht für alle x_0 gilt, nämlich nicht für die Ruhelagen, $x(t;\theta) \equiv x(t;\tilde{\theta}) \equiv x_0 = \sqrt{1 + \theta_2/\theta_1}$ für $\theta, \tilde{\theta} \in \{\theta : \theta_2/\theta_1 = c\}$. Da $\mathcal{F}_1$ in $\mathbb{R}^2$ keine magere Menge ist,

[20] Diese Definition ist ähnlich der algebraischen Beobachtbarkeit, wonach sich jeder Zustand als $x_i = p_i(y, \dot{y}, \ldots, y^{n-1}, u, \ldots, u^{n-1})$ darstellen lässt, wobei die p_i Polynome sind. Hier werden statt Polynomen auch andere Funktionen zugelassen.
So ist $\dot{x} = -x, y = x^3$ klassisch beobachtbar, da alle Zustände unterscheidbar sind, $x = \sqrt[3]{y}$, ist aber nicht algebraisch beobachtbar.

liegt keine globale strukturelle Identifizierbarkeit vor. $\dot{x} = \theta_1(x^2 - 1) - \theta_2$ ist hingegen für $\mathcal{F} := \mathcal{F}_2 = \{\theta \in \mathbb{R}^2 : \theta_1(\theta_1 + \theta_2) \leq 0\}$ global strukturell identifizierbar, da für $\mathcal{F}_2$ keine Ruhelagen möglich sind [155].

$\dot{x} = -\theta^2 x$ ist für $\mathcal{F} = \mathbb{R}$ lokal strukturell identifizierbar und für $\mathcal{F} = \mathbb{R}^{\geq}$ global strukturell identifizierbar.
Ist ein LTI-System in einem θ eindeutig identifizierbar, so ist es strukturell identifizierbar.

Anmerkung 4.7 Die Definition 4.3 ist sehr stringent, da sie für alle x_0 gelten muss und demzufolge auch für alle x_0 zu überprüfen ist. Eine Abschwächung von $\forall x_0$ in (4.17) auf „für fast alle x_0“ empfiehlt sich nicht, da dann gerade die ausgeschlossenen Punkte besondere Schwierigkeiten bereiten können. So ist $\dot{x} = \theta x u$ bis auf $x_0 = 0$ global identifizierbar, doch für $x_0 = 0$ geht wegen $x(t) \equiv 0$ gar nichts mehr, denn dann ermöglicht kein $u(t)$ eine Identifikation von θ. Eine andere Abschwächung betrifft die Identifizierbarkeit von θ bei konkretem x_0. Das ist zweckmäßig, wenn es die Experimentation gestattet, über $u(t)$ einen bestimmten Anfangszustand (üblicherweise eine Ruhelage) einzustellen. Probleme mit fixem x_0 sind einfacher zu handhaben.

Anmerkung 4.8 In der Definition 4.4 wird mitunter als Ausschlussmenge eine Nullmenge (Lebesgue-Maß Null; metrisch geprägter Begriff) [26], [162] statt wie hier eine magere Menge (topologisch geprägter Begriff) [155], [296] gefordert. Beide Forderungen sind nicht gleichwertig! Es gibt magere Mengen, die keine Nullmengen sind, und Nullmengen, die nicht mager sind, s. [28]. Aber: Bei der Identifikation treten als Ausschlussmengen überwiegend algebraische Mengen oder semialgebraische Mengen mit leerem Inneren auf, die sowohl mager als auch Nullmengen sind, womit dann Äquivalenz zwischen den Definitionen besteht.

Anmerkung 4.9 Definition 4.3 kann in Analogie auf Zufallsvariablen gemäß $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \Rightarrow F_{\boldsymbol{\theta}_1}(\xi) \neq F_{\boldsymbol{\theta}_2}(\xi)$ erweitert werden, wobei $F_{\boldsymbol{\theta}_i}(\xi)$ für die jeweiligen Verteilungsfunktionen steht.

Anmerkung 4.10 Wird das Modell (4.16) um die Gleichung $\dot{\theta} = 0_p$ ergänzt und wird der Zustandsvektor x um θ zu $x := \begin{pmatrix} x \\ \theta \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R}^p$ erweitert, dann zeigen sich die Parallelen von struktureller Identifizierbarkeit und globaler Beobachtbarkeit sowie Nichtidentifizierbarkeit und Nichtunterscheidbarkeit.[21]
Bezogen auf den erweiterten Zustandsvektor liegt der Unterschied zwischen beiden Konzepten darin, dass die Identifizierbarkeit gewissermaßen eine auf θ reduzierte Beobachtbarkeit darstellt.

[21] Zwei Zustände $x_1(t_0), x_2(t_0) \in \mathcal{M}$ heißen nichtunterscheidbar, wenn $y_1(t, t_0, x_1, u) \equiv y_2(t, t_0, x_2, u)$ für jeden zulässigen Eingang u gilt. Ein nichtlineares System heißt global beobachtbar, wenn aus der Identität der Ausgangssignale für alle u nur $x_1(t_0) = x_2(t_0)$ folgt.

Anmerkung 4.11 Die Identifizierbarkeit kann statt für θ auch für eine (interessierende, spezielle) Abbildung $\xi = \Psi(\theta)$ definiert werden. Das Konzept der sog. Abbildungsidentifizierbarkeit ist eng verwandt mit dem Konzept der schätzbaren Funktionen [459]. Es enthält die Identifizierbarkeit eines einzelnen Parameters $\xi = e_i^T\theta = \theta_i$ als Spezialfall.

Beispiel 4.7 (Nicht identifizierbar, aber abbildungsidentifizierbar)
$\dot{x} = -\theta_1 x + \theta_2 u$; $x_0 = 0$; $y = \theta_3 x$ ist strukturell nicht identifizierbar, da aus $y(t) = \theta_2\theta_3 \int_0^t \exp(-\theta_1(t-\tau)u(\tau)\mathrm{d}\tau$ unabhängig von u die Komponenten θ_2 und θ_3 nicht separiert werden können. Identifizierbarkeit für die Komponente θ_1 und die Abbildung $\xi = \theta_2\theta_3$ liegt indes vor.

Anmerkung 4.12 Strukturelle Nichtidentifizierbarkeit kann bedeuten, ein Modell und die zugehörige Identifikationsstrategie zu verwerfen, s. hierzu Beispiel 4.8. Mitunter lässt sich das Verwerfen aber durch Einbeziehen von weiterem Vorwissen umgehen. In [253] wird beispielsweise gezeigt, wie durch Annahmen zur potenziellen Fahrzeuggröße und/oder Geschwindigkeit Identifizierbarkeit bei der Lagebestimmung für sich bewegende Fahrzeuge erzielbar ist, wenn mit Lidar-Sensoren gemessen wird.

Beispiel 4.8 (Identifikationsstrategie für Doppel-RC-Glied)
Um die vier Werte R_1, R_2, C_1, C_2 aus dem Übertragungsverhalten zu identifizieren, muss $G(s)$ mindestens vier Parameter aufweisen. Die naheliegende Spannungsübertragung muss verworfen werden, da sie nur auf ein PT2-Glied (drei Parameter)[22] führt, vgl. Abschn. 2.10. Die Strom-Spannungsübertragung gibt ein PT2T'-Glied (vier Parameter) und löst das Problem. Für weitere Ausführungen sei auf [312] verwiesen.

4.3.2 Methoden zur Überprüfung struktureller Identifizierbarkeit

Ein SISO-Modell $\Sigma = (A(\theta), b(\theta), c^T(\theta), d(\theta))$ sei gegeben, und es soll geprüft werden, ob der Parametervektor θ strukturell über E/A-Daten identifizierbar ist. Ohne Restriktionen an θ ist dessen Komponentenanzahl notwendigerweise auf $n + m + 1$, also die Parameteranzahl der Übertragungsfunktion beschränkt. Mit Hilfe des Leverrier-Faddeev-Algorithmus [209] kann die Übertragungsfunktion berechnet werden (zweckmäßig mit einem Computeralgebra-System)

$$G(s) = \frac{b_m(\theta)s^m + b_{m-1}(\theta)s^{m-1} + \ldots + b_0(\theta)}{s^n + a_{n-1}(\theta)s^{n-1} + \ldots + a_0(\theta)}. \tag{4.19}$$

Diese ist strukturell identifizierbar bezüglich der a_i, b_j. Für eine bijektive Abbildung $\Phi(\theta) = (a_0(\theta), \ldots, a_{n-1}(\theta), b_0(\theta), \ldots, b_m(\theta))^T$ ist auch θ strukturell identifizierbar, denn Φ^{-1} ist dann

[22] Strenggenommen liefert das PT2-Glied in diesem Fall sogar nur zwei Gleichungen für vier Werte, da der Parameter $b_0 = 1$ nicht von R_1, R_2, C_1, C_2 abhängt.

ebenfalls bijektiv und bildet somit Residualmengen (Komplement einer mageren Menge) in Residualmengen ab. Lokale Identifizierbarkeit folgt für $\theta \in \mathbb{R}^{n+m+1}$ und stetig differenzierbares Φ aus dem Satz über inverse Funktionen [487],[23], also aus der Regularität der Funktionalmatrix

$$\det \frac{\partial \Phi}{\partial \theta^T} \neq 0. \tag{4.20}$$

Globale Identifizierbarkeit folgt aus (4.20) aber nicht, wie nachfolgendes Beispiel zeigt.

Beispiel 4.9 (Reguläre Funktionalmatrix sichert nur lokale Identifizierbarkeit)
Das folgende System liefert über die Übertragungsfunktion

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} \theta_1 & 1 \\ 1 & \theta_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u \qquad \Rightarrow \quad G(s) = \frac{1}{s^2 - (\theta_1 + \theta_2)s + (\theta_1\theta_2 - 1)}$$
$$y = x_1$$

den Zusammenhang $a_1 = -\theta_1 - \theta_2$ und $a_0 = \theta_1\theta_2 - 1$, woraus die Regularität der Funktionalmatrix $\frac{\partial (a_1,a_0)^T}{\partial (\theta_1,\theta_2)}$ aus $\det \frac{\partial (a_1,a_0)^T}{\partial (\theta_1,\theta_2)} = -\theta_1 + \theta_2 \neq 0$ für $\theta_1 \neq \theta_2$ folgt. In diesem Fall liegt in der mageren Ausschlussmenge (Gerade $\theta_2 = \theta_1$) nicht Nichtidentifizierbarkeit vor, sondern punktweise Identifizierbarkeit. Für alle anderen Paare (θ_1, θ_2) sind wegen

$$\theta_1 = -\frac{a_1}{2} \pm \sqrt{\frac{a_1^2}{4} - (1 + a_0)}, \qquad \theta_2 = -\frac{a_1}{2} \mp \sqrt{\frac{a_1^2}{4} - (1 + a_0)}$$

nur lokal eindeutig identifizierbar, womit θ schlussendlich lokal strukturell identifizierbar ist.

Anmerkung 4.13 Weitere Zugänge für LTI-Systeme über Taylor-Reihenentwicklung, Ähnlichkeitstransformationen und Markov-Parameter gibt [296]. Ein auf einem Rangtest basierendes Kriterium zur lokalen Identifizierbarkeit, das unter Zusatzannahmen auf globale Identifizierbarkeit erweitern wird, findet sich in [36]. Es ist besonders geeignet, wenn die Parameter θ affin in A, B, C, D eingehen und linearen Gleichungsrestriktionen unterliegen.

Da sich E/A-Darstellung für Zustandsraummodelle mit Polynomvektorfeldern ebenfalls relativ einfach ableiten lassen, funktioniert der beschriebene Weg auch für Polynomsysteme, was am folgenden Beispiel gezeigt wird.

Beispiel 4.10 (Identifizierbarkeit eines Polynomsystems, [547])
Gegeben sei

$$\begin{aligned} \dot{x}_1 &= -\theta_2 x_1 - \theta_3 x_2 - \theta_0 u \\ \dot{x}_2 &= \theta_3 x_1 x_2 - \theta_1 x_1 \\ y &= x_1 \end{aligned}$$

[23] Die Bedingung sichert einen lokalen C^1-Diffeomorphismus. Sie ist hinreichend, vgl. $a = \theta^3$ bzw. $a = \theta^{1/3}$, wo die Ableitung in $\theta = 0$ verschwindet bzw. nicht existiert. Dennoch ist θ in beiden Fällen global identifizierbar.

Mit der Reihung $\dot{x}_2 \succ \dot{x}_1 \succ x_2 \succ x_1 \succ \ddot{y} \succ \dot{y} \succ y \succ \ddot{u} \succ \dot{u} \succ u$ ergibt sich

$$\ddot{y} = -\theta_1\dot{u} - \theta_2\dot{y} + \theta_3 y\dot{y} + \theta_0\theta_3 uy + \theta_2\theta_3 y^2 + \theta_1\theta_3 y \tag{4.21a}$$

$$= a_1\dot{u} + a_2\dot{y} + a_3 y\dot{y} + a_4 uy + a_5 y^2 + a_6 y. \tag{4.21b}$$

Damit liegt für Erregung aus dem Nullzustand globale Identifizierbarkeit vor. Von den 6 Parametern in (4.21b) sind nur die ersten vier frei, was bei einer Identifikation über ein Gleichungsfehlermodell (4.21b) die Restriktionen $a_5 = -a_2a_3$ und $a_6 = -a_1a_3$ erfordert.

Für differenzialalgebraische Modelle wird in [223] ein Zugang beschrieben, der das Identifizierbarkeits- in ein Lösbarkeitsproblem überführt. Dabei wird analysiert, ob das System implizierter Gleichungen für θ, erweitert um eines für $\tilde{\theta}$, nur die Lösung $\theta = \tilde{\theta}$ hat. Wenn ja, liegt globale strukturelle Identifizierbarkeit vor. Die Lösbarkeitsuntersuchung erfolgt computeralgebraisch. Als Vorteil erweist sich, dass dabei Differenzialgleichungsbedingungen an $u(t)$ erhalten werden, die besagen, wie $u(t)$ nicht gewählt werden darf.

Kriterien für die Identifizierbarkeit von Zustandsraummodellen (mit und ohne Eingangsgröße $u(t)$) beruhen auf der Äquivalenz von Systemen. Prinzipiell prüfen sie, ob eine Variablentransformation Φ existiert, sodass beim gleichen Anfangswert, aber anderen Parametern das gleiche Ausgangssignal erzeugt wird. Von den ähnlich gearteten Kriterien sei hier beispielhaft das nachfolgende genannt.

Satz 4.16 (Notwendiges Nichtidentifizierbarkeitskriterium, [162])
Gegeben sei

$$\dot{x}(t;\theta) = f(x;\theta), \qquad x(0;\theta) = x_0(\theta) \tag{4.22a}$$

$$y(t;\theta) = h(x;\theta), \qquad f, h \in C^1. \tag{4.22b}$$

Wenn unter den Voraussetzungen:
- $\theta, \tilde{\theta} \in \mathcal{U}$; $\mathcal{U}$ beschränkte offene Teilmenge in $\mathbb{R}^p$
- $\mathcal{Z}$ offene Umgebung von $x_0(\tilde{\theta})$ in $\mathcal{M}$
- $\mathcal{M}$ offene zusammenhängende Menge in $\mathbb{R}^n$ mit $x(t,\theta) \in \mathcal{M}$ für alle $\theta \in \mathcal{U}$
- x_0 keine Ruhelage
- $\Phi : \mathcal{Z} \to \mathbb{R}^n$ ist ein C^1-Diffeomorphismus, sodass $\forall z \in \mathcal{Z}, \exists \theta \neq \tilde{\theta}$

$$\operatorname{rang}\frac{\partial \Phi}{\partial z^T}(z) = n \tag{4.23a}$$

$$x_0(\theta) = \Phi(x_0(\tilde{\theta})) \tag{4.23b}$$

$$f(\Phi(z),\theta) = \frac{\partial \Phi}{\partial z^T}(z) \cdot f(z,\tilde{\theta}) \tag{4.23c}$$

$$h(\Phi(z),\theta) = h(z,\tilde{\theta}) \tag{4.23d}$$

gilt, ist (4.22) durch ein Experiment $(x_0(\theta), [0, t_1])$ nicht identifizierbar (t_1 hängt von der Größe der Umgebung ab).

Da die Anwendung dieses Kriteriums nicht ganz einfach ist, soll das Vorgehen für das Beispielsystem aus [162] ausführlich gezeigt werden, wo lediglich das Endergebnis angegeben ist. Zunächst muss also die Menge der parametrisierten Variablentransformationen bestimmt werden, die die geforderten Bedingungen erfüllt, um daraus die algebraischen Parameterbedingungen ableiten zu können.

Beispiel 4.11 (Anwendung des Nichtidentifizierbarkeitskriteriums)
Gegeben sei das auf Identifizierbarkeit zu untersuchende Modell

$$\dot{x}_1 = \theta_1 x_2, \qquad x_1(0) = 1 \tag{4.24a}$$
$$\dot{x}_2 = \theta_2 x_1 x_2 + \theta_3 x_2, \qquad x_2(0) = 1 \tag{4.24b}$$
$$y = x_2. \tag{4.24c}$$

Es folgt mit $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \Phi_1(z_1, z_2) \\ \Phi_2(z_1, z_2) \end{pmatrix}$ und $z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$

$$\Phi_1(1,1) = 1, \quad \Phi_2(1,1) = 1 \qquad \text{aus (4.23b)} \tag{4.25a}$$
$$\Phi_2(z) = z_2 \qquad \text{aus (4.23d)} \tag{4.25b}$$
$$\theta_1 \Phi_2(z) = \frac{\partial \Phi_1}{\partial z_1} \cdot \tilde{\theta}_1 z_2 + \frac{\partial \Phi_1}{\partial z_2} \cdot (\tilde{\theta}_2 z_1 z_2 + \tilde{\theta}_3 z_2) \qquad \text{aus (4.23c)} \tag{4.25c}$$
$$\theta_2 \Phi_1(z) \Phi_2(z) + \theta_3 \Phi_2(z) = \frac{\partial \Phi_2}{\partial z_1} \cdot \tilde{\theta}_1 z_2 + \frac{\partial \Phi_2}{\partial z_2} \cdot (\tilde{\theta}_2 z_1 z_2 + \tilde{\theta}_3 z_2) \qquad \text{aus (4.23c).} \tag{4.25d}$$

Daraus wiederum ergeben sich der Diffeomorphismus

$$\Phi_1(z) = \tilde{\theta}_2/\theta_2 z_1 + (\tilde{\theta}_3 - \theta_3)/\theta_2 \qquad \text{mit } \Phi_2(z) = z_2 \text{ aus (4.25d)} \tag{4.26a}$$
$$\Phi_2(z) = z_2 \tag{4.26b}$$

und die Bedingungen

$$\theta_2 + \theta_3 = \tilde{\theta}_2 + \tilde{\theta}_3 \qquad \text{mit } \Phi_1(1,1) = 1,\ \Phi_2(1,1) = 1 \text{ aus (4.25d)} \tag{4.27a}$$
$$\theta_1 \theta_2 = \tilde{\theta}_1 \tilde{\theta}_2, \qquad \text{mit } \Phi_1(z), \Phi_2(z) \text{ aus (4.26) in (4.25c)} \tag{4.27b}$$

die außer $\theta = \tilde{\theta}$ weitere Lösungen zulassen, sodass (4.24) nicht identifizierbar ist. Für festgehaltene $\theta_1, \theta_2, \theta_3$ sind nur 2 Gleichungen für $\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3$ verfügbar, was eine unendliche Anzahl von Vektoren $\tilde{\theta}$ impliziert. Somit liegt auch keine lokale Identifizierbarkeit vor, bei der endlich viele $\tilde{\theta}$ zulässig sind.

4.3.3 Sicherung der Identifizierbarkeit über kanonische Formen

Die Menge aller möglichen E/A-Verhalten eines LTI-Systems $\Sigma = (A, B, C, D)$ ist eine Mannigfaltigkeit der Dimension $(m + p)n + mp$. Da der Parameterraum aber die Dimension $(n + m + p)n + mp$ hat, sind n^2 Freiheitsgrade zu restringieren, um ein eindeutiges Modell zu erhalten. Hierzu dienen kanonische Formen, bei denen es sich um zweckmäßige Systemdarstellungen handelt.

Definition 4.5 (Kanonische Form)
Eine kanonische Form für eine Äquivalenzrelation "$\sim$" auf einer Menge $\mathcal{M}$ ist eine Abbildung $T : \mathcal{M} \to \mathcal{M}$, die $T(x) \sim x$ und $x \sim y \Rightarrow T(x) = T(y)$ für alle $x, y \in \mathcal{M}$ erfüllt. Äquivalent kann eine kanonische Form durch die Untermenge $\mathcal{B} = T(\mathcal{M}) \subset \mathcal{M}$ beschrieben werden, die jedem $x \in \mathcal{M}$ genau ein Element $b \in \mathcal{B}$ mit $b \sim x$ zuordnet. Kanonische Formen werden synonym auch als Normalformen bezeichnet.

Im Abschnitt 2.3.2 wurden bereits einige Normalformen für LTI-Systeme aufgeführt und hinsichtlich ihrer Anwendbarkeit bei Stabilitätsrestriktionen bewertet. Ergänzend sei hier noch auf einige Normalformen für Mehrgrößensysteme verwiesen, die in jene vom Hermite- und vom Kronecker-Typ, auch Typ I und Typ II genannt, unterteilt werden [328]. Letztlich entscheiden Anwendung, Algorithmus und interessierende Systemeigenschaft darüber, welche Normalform gerade die zweckmäßigste ist.

Kanonische Formen sind nicht auf LTI-Zustandsraummodelle beschränkt. Es gibt sie auch für Matrizen (Echolon-Normalform), Polynommatrizen (Smith-Normalform), Übertragungsfunktionsmatrizen (Smith-McMillan-Normalform), in der Boolschen Logik (disjunktive oder konjunktive Normalform), für Klassen nichtlinearer Systeme (Bilineare Systeme, Bifurkationsnormalform, Künstliche-Neuronale-Netze-Normalformen) und viele mehr. Je nach Problem empfiehlt sich deshalb, zunächst eine Recherche unter den Suchbegriffen „normal form" oder „canonical form" oder „standard form" zzgl. des Systemklassenbegriffs vorzunehmen.

Anmerkung 4.14 Die strenge Forderung nach Eindeutigkeit wird oft aufgebrochen, sodass jedes Element aus $\mathcal{M}$ nicht zu genau einem Element in $\mathcal{B}$, sondern nur zu einer kleinen, leicht beschreibbaren Menge äquivalenter Elemente in $\mathcal{B}$ äquivalent sein muss. So werden bei der Jordan-Form Permutationen der Eigenwerte zugelassen. Dabei fordern einige Autoren, dass die Jordan-Blöcke zu gleichen Eigenwerten nacheinander stehen und der Größe nach geordnet sind [161], andere hingegen fordern weder das Nacheinander-Stehen, noch die Ordnung in den Jordan-Blöcken [299].

4.3.4 Alternative Identifizierbarkeitskonzepte

Bei der strukturellen Identifizierbarkeit ging es darum, ob die Parameter eines Modells unter Idealannahmen eindeutig bestimmbar sind. Dabei wurde ausschließlich das Modell betrachtet. Stimmen System- und Modellstruktur überein, dann gilt die Identifizierbarkeitseigenschaft des Modells gleichfalls für das System. Das ist der Standardfall/die Standardannahme bei der Identifikation. Bei einer Unterparametrisierung des Modells oder wenn das System nicht in der Menge der Modelle enthalten ist, kann das Modell zwar strukturell identifizierbar sein, was dann aber nicht bedeutet, dass mit der gegebenen Modellstruktur auch das System identifiziert werden kann. So liegt ein PT1-System $G(s) = \frac{K}{1+sT}$ nicht in der Modellklasse $G_M(s) = \frac{b_0}{s^2+a_1 s+a_0}$ und ist somit nicht über $G_M(s)$ identifizierbar. Bei einer Überparametrisierung, bei der die Menge der Systeme eine Teilmenge der Modelle ist und Modellidentifizierbarkeit vorliegt, ist die Situation besser. So ist das PT1-System durch $G_M(s) = \frac{b_0}{a_2 s^2+a_1 s+1}$ über $G_M(s)$ identifizierbar und in der Modellklasse $G_M(s) = \frac{b_1 s+b_0}{a_2 s^2+a_1 s+1}$ immerhin noch identifizierbar bezüglich einer Äquivalenzrelation. Die Äquivalenzklassen sind durch alle PT2T'-Modelle charakterisiert, die gekürzt das gleiche PT1-Modell ergeben

$$[G(s)] \stackrel{\text{def}}{=} \left\{ \lim_{s \to s_i} \frac{b_1(s)}{a_1(s)} = \lim_{s \to s_i} \frac{b_2(s)}{a_2(s)}, \forall s_i \in \mathbb{R} \right\}. \tag{4.28}$$

Doch selbst wenn das Modell nicht identifizierbar ist, kann das System identifizierbar sein, nämlich dann, wenn die System- und Modellstruktur unterschiedlich sind. So ist ein Zustandsraummodell n-ter Ordnung nicht strukturell identifizierbar, aber es identifizierbar bezüglich des E/A-Verhaltens als Äquivalenzrelation, weshalb das System als Übertragungsfunktionsstruktur identifizierbar ist. Das wird bei den Unterraumalgorithmen (N4SID, MOEPS) [490] ausgenutzt, die je nach Verfahren ein spezielles Zustandsraummodell bestimmen, aus dem dann die Übertragungsfunktion ermittelt werden kann.

Fortan mögen System- und Modellstruktur gleich sein, und es möge strukturelle Identifizierbarkeit vorliegen. Dann stellen sich folgende weiteren Fragen:

Sind die Parameter mit der Identifikationsmethode identifizierbar?
Während Gleichungsfehlerkriterien konvexe Probleme mit streng konvexen Normen formulieren und damit unter Koerzivitätsannahmen eindeutige Lösungen liefern, führen Ausgangsfehlerkriterien in der Regel auf nichtkonvexe Probleme. Hiermit können die Konvergenz zu lokalen Minima oder die Existenz globaler Mehrfachlösungen verbunden sein. Ein weiterer Aspekt betrifft die Fehlerformulierung. So geht im Frequenzbereich beispielsweise die Information über das Einschwingverhalten verloren, während Autokorrelationen keine Mittelwertinformation enthalten. All dies kann zu Problemen bei der Identifizierbarkeit führen.

Beispiel 4.12 (Zeit-, aber nicht frequenzbereichsidentifizierbar)
Ein PT2-Glied wird durch einen Sinus angeregt. Im Frequenzbereich liegt keine A-posteriori-Identifizierbarkeit vor, da aus Amplitudenverstärkung und Phasenverschiebung nur zwei Parameter identifizierbar sind, das PT2-Glied aber drei hat. Im Zeitbereich mit Ausgangsfehler-LS liegt Identifizierbarkeit vor, da aus dem Einschwingverhalten zusätzliche Information gezogen werden kann.

Sind die Parameter mit den gemessenen Daten identifizierbar?
Dies führt bei LS-orientierten Zugängen auf Rangbetrachtungen an die Datenmatrix, vgl. Satz 4.11. Um den Vollrang zu erhalten, müssen die Signale entsprechend reich an Information sein. So kann ein Hammerstein-Modell mit einem Pseudo-Rausch-Binär-Signal (PRBS) nicht identifiziert werden, da nur zwei Amplituden auftreten (nämlich 1 und -1), für eine Parabel aber allein drei gebraucht werden. Hierfür werden deshalb Pseudo-Rausch-Multilevel-Signale genommen. Ferner beschränkt die Länge eines PRBS die Ordnung der Differenzengleichung, bis zu der die Parameter eindeutig identifiziert werden können. Die Abtastung an sich gefährdet bei bestimmten Systemklassen deren Identifizierbarkeit. Auch hier können Restriktionen weiterhelfen, z. B. das Festlegen der Totzeit auf ganzzahlige Vielfache der Tastzeit, $T_t = kT_A$.

Beispiel 4.13 (Mehrdeutigkeit durch Abtastung, [247])
Bei stückweise konstanter, beliebiger Erregung und äquidistanter Abtastung mit T_A ist das System $G(s) = V\dfrac{1+sT'}{1+sT}\mathrm{e}^{-sT_t}$ nicht identifizierbar. Das z-Modell

$$G_z(z) = \begin{cases} V\dfrac{\frac{T'}{T}z + (1-a) - \frac{T'}{T}}{z-a}\, z^{-d} & \text{für } \varepsilon = 0 \\ V\dfrac{(1-a^{1-\varepsilon}(1-\frac{T'}{T}))z + a^{1-\varepsilon}(1-\frac{T'}{T}) - a}{z-a}\, z^{-d-1} & \text{für } 0 < \varepsilon < 1 \end{cases} \tag{4.29}$$

mit $\varepsilon = T_t/T_A - [T_t/T_A]$, $d = [T_t/T_A]$, $a = \mathrm{e}^{-T_A/T_t}$ beschreibt den Zusammenhang zwischen den Abtastfolgen und ist linksstetig bezüglich T_t. In ihm treten T' und der Totzeitanteil εT_A gemeinsam in nur einem Parameter $a^{1-\varepsilon}(1-\frac{T'}{T})$ auf. Somit haben z. B. das s-Modell mit $V = 1, T = 2, T_A = 0.25$ gleich und $T' = 1, T_t = 0.25$ bzw. $T' = 0.89482908, T_t = 0.05$ das gleiche z-Modell $G_z(z) = \frac{0.5z-0.3825}{z-0.8825}z^{-1}$.

Sind die Parameter „wohlverhaltend“[24] identifizierbar?
Wohlverhaltend ist hier einerseits aus Sicht der Empfindlichkeit zu verstehen und andererseits in dem Sinn, dass die Identifizierbarkeit asymptotisch erhalten bleibt. Angenommen die Parameter sind eindeutig identifizierbar, so können sie doch sehr empfindlich gegen Änderungen

[24] Andere Begriffe sind robust oder stabil, wenngleich dann die Adjektive nicht im strengen systemtheoretischen Sinn zu verstehen sind.

in den Daten reagieren. Dies führt auf Konditionsbetrachtungen, Regularisierungen (meist solche vom Tikhonov-Typ) oder Sondermaßnahmen wie bei Identifikation im geschlossenen Regelkreis. Dort rührt das Problem daher, dass der Regler gerade bei Festwertregelungen ja immer versucht, so schnell wie möglich wieder einen konstanten Wert einzuregeln. Damit werden die Spalten der Matrizen genauso schnell linear abhängig.
Eine Art von Empfindlichkeitsproblemen tritt auf, wenn im ungestörten Fall das lokale Minimum fast so klein ist wie das globale, vgl. Beispiel 4.14. Werden infolge von Störungen beide Minima gleich groß, so verursachen kleinste Änderungen ein Springen der Minimierer.

Beispiel 4.14 (De facto gleiche Sprungantworten zweier Modelle, [247])
Betrachtet werden die beiden Modelle

$$G_1(s) = \frac{1 + 0.2s}{1 + 0.4s + 0.4s^2}\,\mathrm{e}^{-0.6s} \qquad \text{bzw.} \qquad G_2(s) = \frac{1.001 - 0.143s}{1 + 0.388s + 0.395s^2}\,\mathrm{e}^{-0.278s}, \quad (4.30)$$

die als Gütewert der Identifikation $Q_1 = 0$ und $Q_2 = 0.002$ liefern. Der nichtminimalphasige Vorhalt kompensiert einen Teil der Totzeit: $0.2 - 0.6 \approx -0.143 - 0.278$. Zweckmäßigere Systemanregung oder Restriktionen verhindern diesen Effekt.

Die Identifizierbarkeitsbetrachtungen zum Gütekriterium, zu den Experimentalbedingungen und damit letztlich auch zur Qualität der Daten werden zur A-posteriori-Identifizierbarkeit zusammengefasst. In Anlehnung an [60] wird sie wie folgt definiert.

Definition 4.6 (A-posteriori-Identifizierbarkeit)
Ein strukturell identifizierbares System heißt a posteriori identifizierbar, wenn das zugeordnete Optimierungsproblem unter den gegebenen Daten eine eindeutige Lösung hat.

Anmerkung 4.15 Manchmal wird A-priori-Identifizierbarkeit als jene mit perfekten Daten und A-posteriori-Identifizierbarkeit als die mit realen Daten charakterisiert. Zum groben Merken ist das gut, berücksichtigt aber nicht das Identifikationsverfahren (Algorithmus zum Lösen des Optimierungsproblems).

Die bisherigen Betrachtungen sind rein deterministisch. Sind die Daten als Realisierungen eines Zufallsprozesses aufzufassen, so führt das auf die stochastische Identifizierbarkeit. Im Gegensatz zu [99], wo das stochastische Konvergenzkonzept noch frei wählbar ist, wird es in der Definition 4.7 genau spezifiziert. Die Wahl fiel dabei auf ein sehr starkes Konvergenzkonzept, das den praktischen Anforderungen an eine Parameterschätzung gerecht wird.

Definition 4.7 (Stochastische Identifizierbarkeit)
Strukturelle Identifizierbarkeit vorausgesetzt, heißt ein System stochastisch identifizierbar, wenn ein asymptotisch erwartungstreuer und konsistenter[25] Parameterschätzer existiert.

[25] Konsistenz bedeutet fast sichere Konvergenz oder etwas anschaulicher ausgedrückt, ein asymptotisches Verringern der Streuungen und damit das Streben gegen eine Dirac-Verteilung (Eindeutigkeit).

Mitunter wird die stochastische Identifizierbarkeit nicht auf alle Schätzer, sondern nur bezogen auf einen konkreten Schätzer (Verfahren) betrachtet und geprüft, weil das wesentlich einfacher ist. Das nachfolgende Beispiel zeigt, dass zwischen der datenbezogenen A-posteriori-Identifizierbarkeit und der stochastischen Identifizierbarkeit zu unterscheiden ist. Die Ursache für die Diskrepanz liegt im Beispiel in der fehlenden Dynamikanregung.

Beispiel 4.15 (A posteriori identifizierbar, aber nicht stochastisch identifizierbar) Ein PT1-Glied ist per Ausgangsfehler-LS aus dem Verlauf der Gewichtsfunktion a posteriori identifizierbar, aber nicht stochastisch identifizierbar. Wird die Gewichtsfunktion mit weißem Rauschen $S(\omega) = S_0$ gestört, dann ist die asymptotisch Kovarianzmatrix ihrer Linearisierung bezüglich der Parameter V, T gegeben durch

$$\operatorname{cov}\begin{bmatrix} \hat{\mathbf{V}} \\ \hat{\mathbf{T}} \end{bmatrix} = S_0 \left(\int_0^\infty \begin{bmatrix} \frac{\partial V \mathrm{e}^{-t/T}}{\partial V} \\ \frac{\partial V \mathrm{e}^{-t/T}}{\partial T} \end{bmatrix} \begin{bmatrix} \frac{\partial V \mathrm{e}^{-t/T}}{\partial V} \\ \frac{\partial V \mathrm{e}^{-t/T}}{\partial T} \end{bmatrix}^T \mathrm{d}t \right)^{-1} = \begin{bmatrix} \frac{4}{T} & -\frac{4}{V} \\ -\frac{4}{V} & \frac{8T}{V^2} \end{bmatrix}.$$

Diese ist ganz offensichtlich nicht Null, womit keine Konsistenz vorliegen kann. Fälschlicherweise wird manchmal angenommen, dass für Abtastzeiten $T_A \to 0$ und damit Tupelzahlen $N \to \infty$ Konsistenz erreichbar wäre. Doch dem ist auch nicht so. Die Eigenschaft der Konsistenz hängt nämlich nicht allein von der Beobachtungsdauer $T \to \infty$ oder von $T_A \to 0$ ab, sondern davon, inwieweit nutzbare Information gewonnen wird. Aus diesem Grund kann aus einem gestörten Sprungexperiment die Zeitkonstante nicht stochastisch identifiziert werden, die statische Verstärkung aber gleichwohl.

Anmerkung 4.16 Probleme mit der stochastischen Identifizierbarkeit treten insbesondere bei rückgekoppelten Systemen auf. Das liegt daran, dass das Rauschen auf dem Ausgangssignal eines Teilsystems bedingt durch die Rückkopplung auch, wenngleich gefiltert, im Eingangssignal auftritt. Hierdurch wird die bei der Identifikation oft geforderte Unabhängigkeit von Ein- und Ausgangsrauschen verletzt. Ähnliche Probleme treten bei der Frequenzbereichsidentifikation in rückgekoppelten Systemen auf [98]. Selbst die Zeitbereichsidentifikation, die im Gegensatz zur Frequenzbereichsidentifikation Information aus dem Einschwingverhalten ziehen kann, gestaltet sich bei rückgekoppelten Systemen schwierig. Ursache ist hier der Regler, der, wenn seine Ordnung zu niedrig ist, lineare Abhängigkeiten zwischen Ein- und Ausgang schafft, die die zu identifizierende Systemordnung bei linearen LS-Verfahren begrenzen. Zudem wirkt der Regler der Systemanregung entgegen, die bei Festwertregelungen asymptotisch sogar gänzlich verschwindet. Es empfiehlt sich deshalb, in Phasen zu geringer Anregung die Identifikation zu stoppen oder aber einen Kompromiss zwischen Regelverhalten und Identifikation zu schließen (adaptive duale Regelung [191]).

4.4 Ergänzende Aspekte

4.4.1 Lösungsreduzierende Restriktionen

Manchmal genügt es, statt Eindeutigkeit nur zu fordern, dass die Lösungsmenge endlich ist oder dass alle Lösungen zueinander äquivalent sind. Bei einer endlichen Lösungsmenge kann dann a posteriori relativ pragmatisch eine prädestinierte Lösung ausgewählt werden. Oft entpuppen sich auch einige Lösungen als Scheinlösungen (Lösungen des mathematischen Problems, aber nicht des technischen Problems) und entfallen dadurch ohnehin.
Eine Möglichkeit zur Reduktion bietet die Normierung $||x||_2 = 1$, die eine Lösungsmenge $\mathcal{X}_{\text{opt}} = \{\gamma x_{\text{opt}} : \gamma \in \mathbb{R}\}$ mit $x_{\text{opt}} \in \mathbb{R}^n$ auf die beiden Lösungen $\pm x_{\text{opt}}/||x_{\text{opt}}||_2$ reduziert.

Das Optimierungsproblem

$$\frac{x^T A x}{||x||_2^2} \stackrel{!}{=} \text{Min} \qquad x \in \mathbb{R}^n; A \in \mathcal{S}_n \tag{4.31}$$

ist bezüglich des Minimums äquivalent zu

$$x^T A x \stackrel{!}{=} \text{Min} \qquad x \in \mathbb{R}^n : ||x||_2^2 = 1; A \in \mathcal{S}_n. \tag{4.32}$$

Außerdem besteht zwischen den Lösungsmengen $\mathcal{X}_{\text{opt},1}$ von (4.31) und $\mathcal{X}_{\text{opt},2}$ von (4.32) der Zusammenhang $\mathcal{X}_{\text{opt},1} = \{\gamma \mathcal{X}_{\text{opt},2} : \gamma \in \mathbb{R}\}$. Die Lösungen von (4.31) liegen also auf Strahlen, die durch die Lösungen von (4.32) gehen. Die lösungsreduzierende Restriktion $||x||_2^2 = 1$ bietet noch einen weiteren Vorteil. So ist nämlich (4.32) wesentlich einfacher zu behandeln, denn mit der Lagrange-Multiplikatoren-Methode folgt direkt, dass x_{opt} jeder auf Eins normierte Eigenvektor zum kleinsten Eigenwert von A ist.
Die Umformulierung einer skalierungsinvarianten Bruch-Optimierungsaufgabe $\frac{f(x)}{g(x)} \stackrel{!}{=} \text{Min}$ mit $\frac{f(\gamma x)}{g(\gamma x)} \stackrel{!}{=} \text{Min}$ in $f(x) \stackrel{!}{=} \text{Min}; g(x) = 1$ lässt sich auch auf matrixvariate Probleme übertragen. Für $\text{sp}((X^T B X)^{-1} X^T A X) \stackrel{!}{=} \text{Min}$ bewerkstelligt die Restriktion $X^T B X = I_n$ die Umformung in $\text{sp}(X^T A X) \stackrel{!}{=} \text{Min}; X^T B X = I_n$.

4.4.2 Einfluss von Restriktionen auf die Lösung

Restriktionen beeinflussen das Approximationsverhalten unter Umständen stark. Dies soll an einem Beispiel gezeigt werden.

Beispiel 4.16 (Ellipsen-Fit)
Die Kegelschnittgleichung lautet $F(x, y, \theta) = ax^2 + bxy + cy^2 + dx + ey + f = 0$ mit $\theta = [a, b, c, d, e, f]^T$ bzw. $F(x, y, \theta) = [x, y] Q [x, y]^T + [d, e][x, y]^T + f = 0$ mit $Q = \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix}$.

Mit dem algebraischen Fehler $\varepsilon_i = ax_i^2 + bx_iy_i + cy_i^2 + dx_i + ey_i + f$ ergibt sich das LS-Problem $\sum_{i=1}^{N} \varepsilon_i^2 \stackrel{!}{=} \text{Min}$. Dieses Problem hat die triviale Lösung $\theta = 0$, die durch Restriktionen zu vermeiden ist. Dafür bieten sich an:

Eins-Fixierung	$f = 1$
Eins-Normierung	$\|\|\theta\|\|_2 = 1$
Gander [208]	$a + c = 1$
Bookstein [87]	$a^2 + \frac{1}{2}b^2 + c^2 = 1$
Fitzgibbon [193]	$4ac - b^2 = 1$

Die Restriktionen $f = 1$ und $a + c = 1$ sind lineare Restriktionen, allgemein beschrieben durch $p^T\theta = 1$. Die anderen Restriktionen sind quadratisch ohne linearen Term, allgemein $\theta^T B\theta = 1$, speziell z. B.

$$4ac - b^2 = 1 \Leftrightarrow \theta^T \underbrace{\begin{bmatrix} 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}}_{B} \theta = 1.$$

Das Gütekriterium liest sich dann kompakt als

$$||A\theta||_2 \stackrel{!}{=} \text{Min} \qquad p^T\theta = 1 \text{ oder } \theta^T B\theta = 1$$

und führt für die linearen Restriktionen auf ein LSE-Problem und für die quadratische auf ein spezielles LSQ-Problem. Für die quadratischen Restriktionen ist jeder auf $\theta^T B\theta = 1$ normierte Eigenvektor zum größten positiven Eigenwert von $(A^TA)^{-1}B$ eine Lösung. Im Reellen gibt es derer genau zwei, die sich nur im Vorzeichen unterscheiden, wenn der größte Eigenwert einfach ist. Beide Vektoren beschreiben aber ein und dieselbe Lösungskurve. Die Herleitung dieser Aussage erfolgt gradlinig über das Karush-Kuhn-Tucker-Theorem [71] oder die Rayleigh-Ritz-Ungleichung [299].

Der Nachteil der Restriktionen $f = 1$ und der Eins-Normierung, die einer TLS-Formulierung entspricht, liegt darin, dass beide nicht invariant unter euklidischen Transformationen sind. Diese Invarianz wird nur erreicht, wenn die algebraischen Restriktionen gleichzeitig Restriktionen an die Eigenwerte von Q sind, oder genauer ausgedrückt, wenn sie bezüglich der Eigenwerte symmetrisch sind. Es gilt $\text{spur}Q = a + c = \lambda_1 + \lambda_2$ und $||Q||_F^2 = a^2 + \frac{1}{2}b^2 + c^2 = \lambda_1^2 + \lambda_2^2$ und $\det Q = ac - \frac{1}{4}b^2 = \lambda_1\lambda_2$. Von diesen Restriktionen zwingen nur die auf der Determinantenrestriktion basierenden (z. B. $4ac - b^2 = 1$) zusammen mit $a > 0$ die Matrix Q auf positive Definitheit. Die anderen Restriktionen können prinzipiell auch einen anderen Kegelschnitt als eine Ellipse liefern. In solchen Fällen ist allerdings der Modellansatz Ellipse oder

die Qualität der Daten anzuzweifeln. Abschließend soll das Bild 4.1 in Anlehnung an [79] das Gesagte unterstreichen und als Warnung vor einem zu unbedarften Umgang mit Restriktionen dienen. Losgelöst von der weniger guten Anpassung der Ellipse an die das Dreieck bildenden Punkte durch die Eins-Fixierung und Eins-Normierung auf der linken Bildseite, fällt auf, dass beide Restriktionen keine invarianten Approximationen liefern. So ist im linken oberen Teilbild deutlich zu erkennen, dass sich die Ellipsen bei Verschieben der Punkte nicht gleichsam mitverschieben, sondern gänzlich anders liegen. Das ist auf der rechten Seite nicht der Fall. Der Vergleich von linkem oberem und unterem Teilbild zeigt zudem, dass sich auch bei Drehung der Punkte die approximierende Ellipse nicht einfach mitdreht. Die Erfüllung der Invarianzeigenschaft ist bei geometrischen Modellbildungsproblemen also ebenso bedeutend wie bei Pfad- und Trajektorienregelung von Fahrzeugen [627].

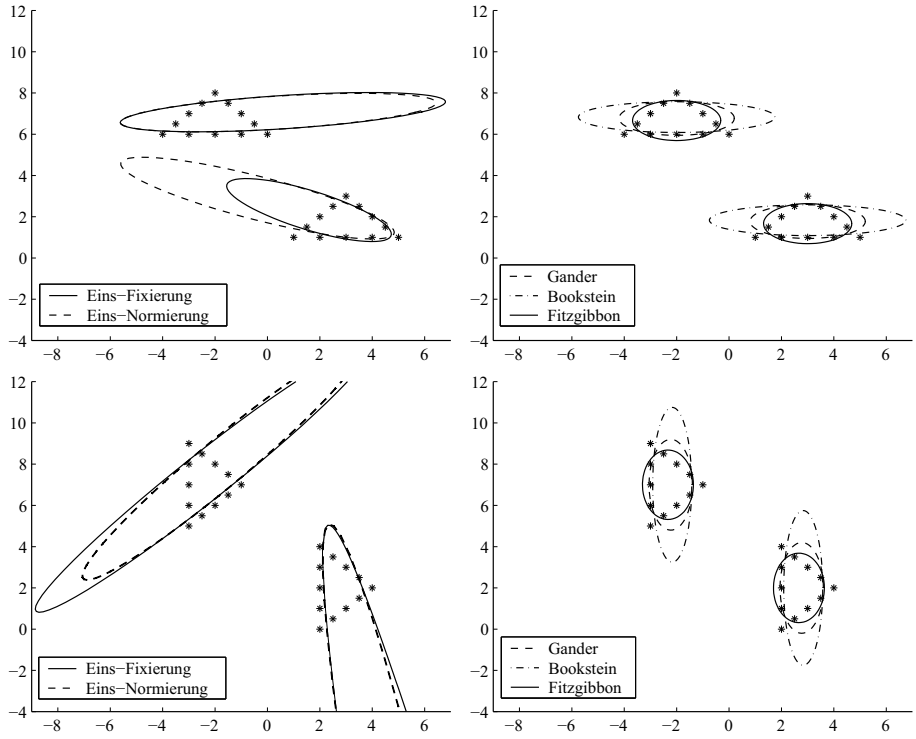

Bild 4.1: Translations- und Rotationsinvarianz

4.4.3 Ordnungsrelationen für eindeutigkeitserzwingende Lösungen

Die Pseudonormallösung[26] des Standard-LS-Problems basiert auf einer Ordnungsrelation, die über die euklidische Norm die normkleinste Lösung aus der Menge aller LS-Lösungen prädestiniert. Für konvexe Lösungsmengen liefern streng konvexe Normen, wie z. B. die euklidische Norm, eine eindeutige Lösung. Andere Ordnungsrelationen können natürlich ebenso herangezogen werden, um eine Lösung hervorzuheben. Vollständige Ordnungen liefern dabei auf abgeschlossenen Lösungsmengen stets ein eindeutiges minimales und maximales Element. Für symmetrische Matrizen empfiehlt sich die Löwner-Halbordnung [299], nach der $A \succeq B$ gilt, wenn $A - B$ nichtnegativ definit ist. Diese und weitere für die Modellbildung wichtige Ordnungsrelationen werden in [651] genannt, wobei besonders auf die strikte Spektralordnung eingegangen wird, vgl. auch Beispiel 4.17. Die strikte Spektralordnung ordnet Matrizen, indem sie bezüglich ihrer Singulärwerte lexikografisch ordnet. Sie ist eine Quasiordnung, da mehrere Matrizen die gleichen Singulärwerte haben können. Basierend auf dieser Ordnung wird für Matrixapproximationsprobleme $||A - X|| \stackrel{!}{=} \text{Min}; X \in \mathcal{M}$, der strikte Spektralapproximand durch $\sigma(A - X^{\text{st}}_{\text{opt}}) \underset{\text{lex}}{\leq} \sigma(A - X_{\text{opt}})$ für alle $X_{\text{opt}} \in \mathcal{X}_{\text{opt}}$ eingeführt. $X^{\text{st}}_{\text{opt}}$ ist also gewissermaßen der größte Approximand.

Beispiel 4.17 (Strikter Spektralapproximand, [651])

Es sei die in der Spektralnorm nächstgelegene symmetrische Matrix zu $\begin{bmatrix} 0 & -2 & 0 \\ 2 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ gesucht. Es gilt dann $\mathcal{X}_{\text{opt}} = \left\{ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \alpha \end{bmatrix} : -1 \leq \alpha \leq 3 \right\}$. Im Gegensatz zur normkleinsten Lösung $X^{\text{min}}_{\text{opt}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ folgt $X^{\text{st}}_{\text{opt}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Der strikte Spektralapproximand ist für dieses Problem sicherlich die bessere Wahl, denn zum einen stimmt er mit dem in der Frobenius-Norm nächstgelegenen symmetrischen Approximanden überein, zum anderen ist er intuitiver.

[26] Das eindeutig bestimmte $||.||_2$-kleinstes Element der Lösungsmenge von $||Ax - b||_2^2 \stackrel{!}{=} \text{Min}$ heißt Pseudonormallösung.

4.4.4 Eindeutigkeitserzwingende Restriktionen für Übertragungsfunktionen

Das Übertragungsfunktionsmodell nimmt eine prädestinierte Stellung bei der Identifikation von Modellen für den Reglerentwurf ein. Damit stellt sich die Frage nach der Wahl der eindeutigkeitserzwingenden Restriktionen.

In der Literatur zur adaptiven Regelung wird bevorzugt $a_n = 1$ gesetzt. Zusammen mit einem Zustandsvariablenfilter der Ordnung n braucht keine Ableitung berechnet werden. In die LS-Matrizen gehen nur die gefilterten Ableitungen ein. Selbst die n-te Ableitung kann aus $y(t)$ und gefilterten Ableitungen berechnet werden. Die Übertragungsfunktion der gefilterten n-ten Ableitung lautet nämlich

$$\frac{s^n Y(s)}{s^n + f_{n-1}s^{n-1} \ldots + f_0} = Y(s) - \sum_{i=0}^{n-1} \underbrace{\frac{f_i}{s^n + f_{n-1}s^{n-1} \ldots + f_0}(s^i Y(s))}_{\text{gefilterte Ableitungen}} \tag{4.33a}$$

$$= Y(s) - \sum_{i=0}^{n-1} \underbrace{\frac{f_i s^i}{s^n + f_{n-1}s^{n-1} \ldots + f_0} Y(s)}_{\text{Realisierung als Signalfilter}} \tag{4.33b}$$

mit den Filterparametern f_i. Da beim klassischen adaptiven Reglerentwurf vom ungestörten oder schwach gestörten Modell ausgegangen wird und da der adaptive Regler sich auch an „falsche Parameter“ (mit statistischer Abweichung und Bias versehen) anpasst, steht hierbei eine dem Anwendungszweck bezogene einfache Realisierung im Vordergrund.

Eine andere Situation liegt vor, wenn die Parameter zum indirekten Messen dienen oder wenn aus ihnen Prozesskoeffizienten bestimmt werden. Dann sollten sie um den wahren Wert möglichst wenig streuen. In [312] wird daher vorgeschlagen, insbesondere $b_m = 1$ oder $a_n = 1$ zu restringieren, da diese Parameter die größten Schätzvarianzen aufweisen würden. Doch so einfach ist das nicht, denn es ist sowohl die Varianz relativ zum Parameterwert als auch der Bias zu betrachten. Genaugenommen müsste auch noch die Transformation in die eigentlichen Prozessparameter einfließen, da nichtlineare Transformationen Varianz und Bias erheblich verändern können. Da schlussendlich Varianz und Bias entscheidend vom verwendeten Verfahren abhängen, misslingen allgemeingültige Empfehlungen für die betreffende Restriktion. Ungeachtet dessen werden nachfolgend einige Aspekte diskutiert.
Eine Argumentation über das Stör-Nutzsignal-Verhältnis legt eine Restriktion mit $a_n = 1$ nahe, da bei der Modellannahme mit ungestörtem Eingang und additiver Ausgangsstörung die gefilterte n-te Ableitung des Ausgangs das ungünstigste Stör-Nutzsignal-Verhältnis haben wird. Wenn nun $a_n = 1$ gewählt wird, dann steht das ungünstige Signal auf der rechten Seite des LS-Problems (virtueller Ausgangsvektor), was seinerseits eine größere Streuung,

aber einen kleineren Bias der Schätzung nach sich zieht. Bei $a_0 = 1$ steht es auf der linken Seite (Datenmatrix), und es bewirkt eine kleinere Streuung, aber einen größeren Bias. Wird also mit einem biasreduzierenden Identifikationsverfahren gearbeitet, dann ist $a_0 = 1$ die bessere Wahl unter der Voraussetzung, dass a_0 und a_n etwa gleich groß sind. Empirische Untersuchungen in [247] an der Zustandsvariablenfilter-Methode und in [360] an der Linearen-Integralfilter-Methode nach Sagara belegen die Argumentation.

Ein Vorteil der Restriktion $a_0 = 1$ ist, dass bei ungenauer Kenntnis der Modellordnung strukturelle Identifizierbarkeit eines Systems kleinerer Ordnung erhalten bleibt. Die führenden Koeffizienten werden dann eben zu Null, vgl. Diskussion am Anfang von Abschn. 4.3.4.

Wenn aus der A-priori-Information eine Abschätzung für die Größenordnung der Parameter getroffen werden kann, so ist eine Restriktion, die den kleinsten Nennerparameter auf Eins setzt, eine Alternative. Da der kleinste Parameter fest ist, kann er bei der Schätzung nicht negativ werden, was ein instabiles Modell zur Folge hätte. Eine solche Restriktion kann helfen, ohne zusätzliche stabilitätssichernde Restriktionen auszukommen.

Handelt es sich bei der gesuchten Übertragungsfunktion um eine Approximation (Modellordnungsreduktion) aus Ein- und Ausgangsdaten, dann führen die Restriktionen $a_i = 1$ oder $||a||_2 = 1$ oder auch $||a||_2^2 + ||b||_2^2 = 1$ im Fall $||a||_2 > ||b||_2$ zu einer Unterapproximation der Frequenzgangamplituden (gewichteter Betragsgang des Modells liegt im Mittel unter dem des realen Systems), während $b_i = 1$ oder $||b||_2 = 1$ oder auch $||a||_2^2 + ||b||_2^2 = 1$ im Fall $||a||_2 < ||b||_2$ eine Überapproximation liefern [460]. Als Empfehlung für passfähige Approximationen werden deshalb in [460] die Restriktionen $\sum_{i=1}^{r} a_{i-1} b_{i-1} = 1$ mit $r \leq m + 1$ vorgeschlagen, die allerdings die Lösung von Eigenwertproblemen erfordern.

Kapitel 5

Reduktionsmethoden

Während sich Kapitel 2 bis Kapitel 4 mit dem Aufstellen von Restriktionen beschäftigen, werden in Kapitel 5 bis Kapitel 8 Methoden zur Vereinfachung des entstehenden Optimierungsproblems behandelt. Sofern keine analytische Lösung des Problems möglich ist, muss auf eine numerische Lösung ausgewichen werden. Dabei sind unterschiedliche Aspekte zu betrachten:

1. Soll ein Standardlöser genutzt werden?
2. Wird eine rekursive Lösung (Onlinelösung) angestrebt?
3. Handelt es sich um ein großes schwachbesetztes Problem?

Den Schwerpunkt bildet in dieser Arbeit der erste Aspekt. Der zweite Aspekt wird durch Hinweise begleitet und in einzelnen Abschnitten vertieft, während sich auf den dritten Aspekt nur über Literaturverweise bezogen wird. Indirekt wird der zweite Aspekt auch durch den ersten abgedeckt. Waren früher rekursive Algorithmen das alleinige Mittel der Wahl für Onlinelösungen, hat sich das heutzutage dank leistungsstarker Computer und effizienter Algorithmen geändert. Bei nicht übermäßig schnellen Prozessen ist es nunmehr möglich, nichtrekursive, ja sogar iterative Algorithmen zur Online-Identifikation einzusetzen. So werden – gute Startwerte vorausgesetzt – bei lokaler quadratischer Konvergenzordnung in der Regel nur ein bis zwei Iterationen pro Parameteraktualisierung benötigt und selbst bei lokaler linearer Konvergenzordnung reicht wegen der praktisch erforderlichen Parametergenauigkeit (wenige Nachkommastellen) gemeinhin eine einstellige Zahl von Iterationen aus. Die guten Startwerte liefert dabei die vorangegangene Berechnung. Eindeutigkeit der Lösung oder zumindest gute Kenntnisse zum Lösungsverhalten sollten aber vorliegen, um etwa ein Springen oder ein Wegdriften der Lösung ausschließen zu können.

Entscheidend für die Wahl eines effizienten Algorithmus ist die Zuordnung einer Identifikationsaufgabe zu einer möglichst gut lösbaren Klasse von Optimierungsproblemen. Zur Bewältigung dieser Aufgabe hat der Autor die Tabellen A.8 bis A.7 auf den Seiten 363ff. erstellt, die Standardprobleme und Spezialprobleme zusammenfassen, für die zugeschnittene Algorithmen existieren. Es ist empfehlenswert, sich vor Beginn einer Problemaufbereitung an den „gelösten“ Problemen zu orientieren, weil dadurch das zu erreichende Ziel für die Umformungen vorgegeben wird. Sollte ein Spezialproblem durch die verfügbare Software nicht abgedeckt sein, kann eine Literaturrecherche unter den in den Tabellen genannten Namen und Abkürzungen weiterhelfen. Oftmals stellen Autoren auch ihre Algorithmen online oder gegeben sie kostenfrei auf Anfrage für wissenschaftliche Zwecke weiter.

Unter Reduktionsmethoden werden diejenigen zusammengefasst, die die Anzahl der Variablen und/oder Restriktionen verringern, aber dennoch ein äquivalentes Problem im weiteren Sinne lösen, d. h., dass aus der Lösung des reduzierten Problems auf die des Originalproblems geschlossen werden kann. Damit unterscheiden sich die Reduktionsmethoden von Problemmodifikationen per Relaxation, bei denen etwa unhandliche Restriktionen weggelassen werden, wohlwissend, dass die relaxierte Lösung nicht mit der des Originalproblems übereinstimmen muss.

In Abschnitt 5.1 werden zuerst Eliminationsmethoden für Ungleichungen aufgeführt. An die Methoden zur Elimination von Ungleichungen schließen sich in Abschnitt 5.2 jene für Gleichungen an. Während sich bei der Elimination von Ungleichungsrestriktionen oft nur deren Anzahl verringert, verringert sich bei der klassischen Elimination von Gleichungsrestriktionen zusätzlich die Variablenanzahl. Da das Erkennen und Reduzieren redundanter linearer Gleichungsrestriktionen zum Repertoire der linearen Algebra (lineare Unabhängigkeit, Rangtest etc.) gehört, werden ohne Einschränkungen nichtredundante Gleichungsrestriktionen angenommen.

Während in Abschnitt 5.1 die Elimination von Ungleichungen betrachtet wird, geht es in Abschnitt 5.3 um die Elimination mit Ungleichungen. Das klingt zunächst paradox, ist aber für bestimmte Problemklassen möglich, wobei sogar eine vollständige Problemlösung gelingen kann, ohne dass Ableitungen verwendet werden müssen. Da diese elegante Technik dem Ingenieur wenig vertraut ist, soll sie an einfachen Beispielen erläutert werden. Nachteilig ist, dass für die aktive Anwendung der Technik gute Kenntnisse in der Theorie der Ungleichungen verlangt werden. Für die passive Anwendung, also dem Verstehen von Herleitungen, genügt es indes, die dahinter liegende Idee verinnerlicht zu haben.

Im Abschnitt 5.4 wird eine andere Idee zur Reduktion von Restriktionen genutzt. Bei den sog. Penalty-Verfahren werden die Gleichungsrestriktionen über Strafterme in das Gütekriterium aufgenommen, wodurch sie als explizite Restriktionen verschwinden. Die Anzahl der Variablen bleibt dabei gleich. Eigentlich müssten die Penalty-Verfahren der Problemmodifikation

zugeordnet werden, da sie die geforderte Bedingung nach Lösungsäquivalenz nicht erfüllen. Werden die Strafterme aber unendlich stark gewichtet, muss die Lösung des modifizierten Gütekriteriums die Restriktionen exakt einzuhalten. Das rechtfertigt, die Penalty-Verfahren den Reduktionsmethoden zugeordnet. In der Praxis wird die unendlich starke Wichtung durch ein sukzessives Erhöhen der Wichtung oder durch eine sehr hohe Anfangswichtung umgesetzt. Zur Lösung sind ausgefeilten Algorithmen erforderlich, da die hohen Wichtungen die Kondition des Problems beträchtlich verschlechtern. Als Kompromiss werden deshalb die Wichtungen nur so hoch gewählt, dass die Abweichungen zur exakten Lösung vernachlässigbar wird.

Eng verwandt mit den Penalty-Verfahren sind die Barriere-Verfahren [320], [71], [213]. Auch bei diesen wird eine Folge von Hilfsproblemen betrachtet, wobei die Ungleichungsrestriktionen über gewichtete Strafterme eliminiert werden. Die Verfahren erzeugen eine Folge innerer Punkte, die die Ungleichungen streng erfüllen. Je näher ein Punkt an der Restriktionsgrenze liegt, desto größer wird der Strafterm. Um aber sicherzustellen, dass die Verfahren auch gegen zulässige Punkte am Rand konvergieren und dass zudem das Minimum durch den Strafterm nicht verfälscht wird, muss das Gewicht des Strafterms asymptotisch gegen Null gehen. Da die numerische Implementierung der Barriere-Verfahren nicht ganz einfach ist und ihre Effizienz entscheidend von dem Optimierungsproblem abhängt, empfehlen sie sich für Vereinfachungen von Problemstellungen im Sinne dieser Arbeit nicht. Nichtsdestotrotz sei auf die populären Innere-Punkt-Methoden [476] für SDP-Probleme verwiesen, die sich auf die Idee der Barriere-Verfahren stützen.

5.1 Elimination von Ungleichungsrestriktionen

Ungleichungen sind von ihrer algorithmischen Behandlung in aller Regel aufwendiger als Gleichungen, da sie mit Fallunterscheidungen sowohl bei der analytischen als auch numerischen Lösung einhergehen. Zudem können Ungleichungen anders als Gleichungen, die einfach so gelten, streng, scharf, singulär, aktiv, konsistent, redundant sein, s. Tabelle A.9 auf Seite 364 für Erklärungen. Es ist deshalb wünschenswert, ihre Anzahl zu reduzieren. Dem widmen sich die folgenden Abschnitte.

Im Abschnitt 5.1.1 wird die Technik der Reparametrisierung vorgestellt. Dabei wird eine Variable, die einer Restriktion unterliegt durch eine Funktion einer anderen Variablen ersetzt, die dann keiner Restriktion mehr unterliegt. So ist für $x \geq 0$ die Substitution $x := y^2$ eine Reparametrisierung, da egal welchen Wert y annimmt, $x \geq 0$ wegen des Quadrates immer gilt. Die Erweiterung der eben beschriebenen Reparametrisierung auf Matrizen wird verwendet, wenn eine Matrix semidefinit sein soll. Im Abschnitt 5.1.1 wird auch auf einige Nachteile dieser Technik verwiesen.

Die mit der Gauß-Elimination für Gleichungen verwandte Technik heißt bei Ungleichungen Fourier-Motzkin-Elimination. Sie wird in Abschnitt 5.1.2 behandelt. Von zentraler Bedeutung sind dabei die Inferenzregel

$$f_1(x) \geq 0,\ f_2(x) \geq 0;\ \gamma_1, \gamma_2 \geq 0 \qquad \Rightarrow \qquad \gamma_1 f_1(x) + \gamma_2 f_2(x) \geq 0 \tag{5.1}$$

und die Eliminationsregel

$$f_1(x) \leq y,\ f_2(x) \geq y \qquad \Rightarrow \qquad f_1(x) \leq f_2(x). \tag{5.2}$$

Gegenstand von Abschnitt 5.1.3 sind redundante Ungleichungen, die wie der Name sagt, überflüssig sind. Für Probleme, die mit Standardsoftware gelöst werden sollen, bereiten redundante Ungleichungen anders als redundante Gleichungen selten Schwierigkeiten (redundante Gleichungen bewirken Rangverlust, Konditionsprobleme usw.), da sie gemeinhin immer streng erfüllt sind. Für rekursive Algorithmen erhöhen sie nur den Aufwand. Im Abschnitt wird deshalb vor allem erklärt, wie redundante Ungleichungen erkannt werden können.

Im Abschnitt 5.1.4 werden singuläre Ungleichungen (z. B. $x^2 \leq 0$ mit der einzigen Lösung $x = 0$) behandelt, die prinzipiell vermieden werden sollten. Das Problem an ihnen ist, dass sie keine inneren Punkte definieren und somit eigentlich Gleichungen sind. Als Folge sind dann Voraussetzungen etwa der Methode der Lagrange-Multiplikatoren verletzt und numerische Algorithmen verzweifeln, Suchrichtungen in das Gültigkeitsgebiet der Ungleichung zu erzeugen. Werden singuläre Ungleichungen detektiert, so ermöglichen sie als Gleichungen die Reduktion der Variablenanzahl.

Der Abschnitt 5.1.5 zeigt, wie vorzeichenunbeschränkte Variablen in linearen Ungleichungssystemen verwendet werden können, um die Zahl funktionaler Ungleichungen zu verringern, also solcher, die von mehreren Variablen abhängen. Es entstehen dann Ungleichungen in nur einer Variablen, was rekursive Algorithmen signifikant vereinfachen kann.

Die im Abschnitt 5.1.6 vorgestellte Technik, aus einer Ungleichung einfach eine Gleichung zu machen, klingt zunächst unzulässig. Normalerweise ist dem auch so. Wenn allerdings Betrachtungen zur Eindeutigkeit eine Reduktion von Freiheitsgraden erfordern, kann es zweckmäßig sein, nicht einen Parameter auf Eins zu setzen, sondern einen Freiheitsgrad zu verwenden, um aus einer Ungleichung eine Gleichung zu machen.

5.1.1 Elimination von Ungleichungen durch Reparametrisierung

Durch Reparametrisierung können einige einfache lineare Ungleichungsrestriktionen, die sich auf x beziehen, beseitigt werden, indem x durch Funktionen ersetzt wird, in denen eine neue, von Restriktionen freie Variable y auftritt. Hierbei wird die Nichtnegativität bestimmter Funktionen, z. B. $y^2 \geq 0$, $|y| \geq 0$ und $\mathrm{e}^y > 0$ für alle $y \in \mathbb{R}$, oder ein beschränkter Wertebereich ausgenutzt, z. B. $\arctan y \in (-\frac{\pi}{2}, \frac{\pi}{2})$, $\tanh y = \frac{\mathrm{e}^y - \mathrm{e}^{-y}}{\mathrm{e}^y + \mathrm{e}^{-y}} \in (-1, 1)$ oder $\sin y \in [-1, 1]$. Da die y dann keinen Restriktionen mehr unterliegen, können Algorithmen für freie Probleme eingesetzt werden, sofern es gelingt, alle Restriktionen zu beseitigen. In Tabelle 5.1 hat der Autor einige gebräuchliche Substitutionen zur Elimination von Ungleichungen zusammengestellt, die die angeführten Reparametrisierungsfunktionen nutzen. Funktionen mit gleicher Wirkung können dabei gegeneinander ausgetauscht werden. Welche dann die richtige Wahl ist, lässt sich selten vorab sagen. Die Schwierigkeiten mit dem Ableiten der Betragsfunktion bleiben beispielsweise unbedeutend, wenn die Optimallösung y_{opt} nicht nahe Null liegt. Dafür ändert die Betragssubstitution die Topologie der Gütefunktion weniger als die Quadratsubstitution. Monotone Grundfunktionen wie $\tanh y$ oder $\arctan y$ sind tendenziell einem $\sin y$ vorzuziehen, da die Lösungsmenge $\mathcal{Y}_{\mathrm{opt}}$ wegen der Nichtperiodizität weniger Minimierer enthält. Gute Erfahrungen hat der Autor mit der Substitution $x =: y^2$ bei der Umsetzung von Stabilitätsrestriktionen für SISO-Systeme in Ausgangsfehlerverfahren gemacht [247], zu deren Lösung nur ein Algorithmus für freie Probleme verfügbar war. Bewährt hat sich für Parameterintervallrestriktionen (physikalische oder technische Parameterbegrenzungen) ebenso die Substitution $x =: \tanh x$. Im Folgenden zeigen vier Beispiele Anwendungen für das Reparametrisieren, weisen aber auch auf damit einhergehende Schwierigkeiten hin.

Beispiel 5.1 (Ähnlichkeitsskalierung)
Das Problem

$$f(X; A) = ||XAX^{-1}||_2 \stackrel{!}{=} \mathrm{Inf} \qquad X \text{ diagonal mit } x_{ii} > 0 \text{ und } \det X = 1$$

tritt bei der Robustheitsanalyse auf und kann durch Anwenden der e^y-Substitution entsprechend der dritten Zeile in Tabelle 5.1 vereinfacht werden. Das Problem ist nicht konvex (Determinantenrestriktion)[1]. Durch Reparametrisierung entfallen die Positivitätsrestriktionen, und es ergibt sich ein konvexes Problem [603]

$$||\mathrm{e}^Y A \mathrm{e}^{-Y}||_2 \stackrel{!}{=} \mathrm{Inf} \qquad Y \in \mathbb{R}^{n \times n},\ Y \text{ diagonal},\ \mathrm{spur} Y = 0.$$

[1] In der Literatur wird mitunter auf die Restriktionen $x_{ii} > 0$ und $\det X = 1$ verzichtet. Das hat dann aber zusätzliche Mehrdeutigkeiten zur Folge, denn wenn $x_{ii,\mathrm{opt}}$ minimiert, liefert $-x_{ii,\mathrm{opt}}$ wegen der Quadrate in $||XAX^{-1}||_2 = \sqrt{\lambda_{\max}(XAX^{-1}(X^{-1}A^TX))} = \sqrt{\lambda_{\max}(AX^{-2}A^TX^2)}$ den gleichen Gütewert und wenn X_{opt} minimiert, hat jede Skalierung γX_{opt} den gleichen Gütewert. Die Determinantenrestriktion unterdrückt die Skalierungsmöglichkeit.

Restriktion	Substitution
$x \geq a$	$x =: a + y^2$ oder $x =: a + \lvert y \rvert$
$x \geq a, x \geq b$	$x =: \max\{a, b\} + y^2$
$x > a$	$x =: a + \exp y$
$x \leq b$	$x =: b - y^2$
$x < b$	$x =: b - \exp y$
$x_1 \leq x_2 \leq x_3$	$x_1 =: y_1$ $x_2 =: y_1 + y_2^2$ $x_3 =: y_1 + y_2^2 + y_3^2$
$x_1 < x_2 < x_3$	$x_1 =: y_1$ $x_2 =: y_1 + \exp(y_2)$ $x_3 =: y_1 + \exp(y_2) + \exp(y_3)$
$\lvert x \rvert < 1$	$x =: \frac{1 - \exp y}{1 + \exp y}$ oder $x =: \tanh x$ oder $x =: \frac{2}{\pi} \arctan y$
$a \leq x \leq b$	$x =: \frac{a+b}{2} + \frac{a-b}{2} \sin y$
$a < x < b$	$x =: a + (b-a)\frac{\exp y}{\exp y + \exp(-y)}$

Tabelle 5.1: Elimination linearer Ungleichungsrestriktionen

Die Spurrestriktion ist konvex bezüglich Y und korrespondiert wegen $\det X = \det \mathrm{e}^Y = \prod_{i=1}^n \mathrm{e}^{y_{ii}} = \mathrm{e}^{\sum_{i=1}^n y_{ii}} = \mathrm{e}^{\mathrm{spur} Y} = \mathrm{e}^0 = 1$ mit der Determinantenrestriktion.
Für detaillierte Betrachtungen zur Existenz eines Minimums statt eines Infimums sei auf [42] verwiesen; algorithmische Aspekte werden in [488] behandelt.

Im Beispiel 5.1 verbesserte die Reparametrisierung die Topologie hin zu einem konvexen Problem. Ein konvexes Problem wird auch bei der Geometrischen Optimierung mit der gleichen Reparametrisierung erreicht, s. Fußnote auf Seite 276. Leider gelingen solch positive Änderungen nur selten. Vielmehr sei vor einer zu unbedarften Anwendung der Reparametrisierung gewarnt, da sich die Topologie des Gütegebirges unter Umständen erheblich ändert, vgl. Beispiel 5.2.

Beispiel 5.2 (Nachteile durch Reparametrisierung)
$x =: y^2$ beseitigt zwar die Restriktion $x \geq 0$, doch wird dadurch beispielsweise eine quadratische Zielfunktion in x zu einer vierten Grads in y, vgl. $f(x) = (x-4)^2$ und $\tilde{f}(y) = (y^2-4)^2$. Mit dramatischen Folgen: langsamere Konvergenz nahe $y = \pm 2$, Nichtausreichen der Opti-

malitätsbedingung zweiter Ordnung, um einen stationären Punkt als lokalen Minimierer zu klassifizieren, und Entstehen eines stationären Punkts $y = 0$ (lokales Maximum von $\tilde{f}$), der nicht mit dem Minimierer $x_{\text{opt}} = 4$ korrespondiert!

Anmerkung 5.1 Als Alternative zur Elimination der Doppelungleichung $l \leq x \leq u$ nach Tabelle 5.1 bietet sich eine Reduktion auf eine einzige quadratische Ungleichung an, nämlich $(x - \frac{u+l}{2})^2 \leq (\frac{u-l}{2})^2$. Für Ausführungen hierzu und die Umformulierung eines LSI-Problems in eine Folge Tikhonov-regularisierter quadratischer Probleme sei auf [451] verwiesen.

Anmerkung 5.2 Mithin ist es nicht so, dass Intervallrestriktionen (ein- wie auch zweiseitig) algorithmisch schwierig zu handhaben sind. So arbeiten aktive Mengenstrategien bei diesen Restriktionen sehr effektiv, sodass bei deren Einsatz eine Elimination mehr schadet, als nutzt.

Die Matrixerweiterung der Reparametrisierung von $x \geq 0$ durch $x = y^2$ lautet $X = Y^T Y$ und eliminiert die Restriktion $X \succeq 0_{n\times n}$. Der Preis für die Elimination der Restriktion ist eine in der Regel tiefere Verkettung der neuen Unbekannten in der Zielfunktion, was das Ableiten und numerische Optimieren erschwert. Die folgenden Beispiele zeigen zwei Anwendungen der beschriebenen Eliminationstechnik.

Beispiel 5.3 (Herleitung des ML-Kovarianzmatrixschätzers, Variante 1)
Die Log-Likelihood-Funktion zur Schätzung der Kenngrößen einer multivariaten Normalverteilung lautet (s. auch Beispiel 8.2)

$$l(\mu, \Sigma) = -\frac{Nn\ln(2\pi)}{2} - \frac{N}{2}\ln\det\Sigma - \frac{1}{2}\text{spur}(\Sigma^{-1}Z(\mu)); \quad Z(\mu) = \sum_{i=1}^{N}(y_i - \mu)(y_i - \mu)^T.$$

Beim Herleiten der ML-Schätzer für μ und Σ ist positive Definitheit von $\hat{\Sigma}_{ML}$ zu fordern, die durch Variablensubstitution $\Sigma = Y^T Y$ erfolgen kann. Mit

$$\tilde{l}(\mu, X) = -\frac{Nn\ln(2\pi)}{2} - \frac{N}{2}\ln\det(Y^T Y) - \frac{1}{2}\text{spur}((Y^T Y)^{-1}Z(\mu))$$

ergeben sich dann die Bedingungen erster Ordnung mit den Ableitungsregeln aus [421] zu

$$\frac{\partial\tilde{l}(\mu, Y)}{\partial\mu} = (Y^T Y)^{-1}(\sum_{i=1}^{N} y_i - N\mu) = 0_n$$
$$\frac{\partial\tilde{l}(\mu, Y)}{\partial Y} = -N \cdot Y(Y^T Y)^{-1} + Y(Y^T Y)^{-1}(Y^T Z(\mu)Y)^{-1} = 0_{n\times n}.$$

Es folgt $\hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} y_i$. Die untere Bedingung gibt $(Y_{\text{opt}}^T Z(\hat{\mu})Y_{\text{opt}})^{-1} = N \cdot I_n$ nach Linksmultiplikation mit Y_{opt}^T und liefert damit die bekannte Lösung $\hat{\Sigma}_{ML} = Y_{\text{opt}}^T Y_{\text{opt}} = \frac{1}{N}Z(\hat{\mu})$. Sind die notwendigen Bedingungen formuliert, so erweist es sich bei der Umformung als Vorteil, sich nicht an Y, sondern an $Y^T Y$ zu orientieren.

Beispiel 5.4 (Nichtnegativ definites Prokrustes-Problem)
In [634] wird gezeigt, dass alle lokalen Minimierer von $||AY^TY - B||_F$ gleichfalls globale Minimierer $X_{\text{opt}} = Y_{\text{opt}}^T Y_{\text{opt}}$ von

$$||AX - B||_F \stackrel{!}{=} \text{Min} \qquad X \succeq 0_{n\times n}$$

sind. Das ist bemerkenswert, ist doch bei einer derartigen surjektiven Parametrisierung gemeinhin nur gesichert, dass unter den lokalen Minimierern einer ist, der mit dem globalen Minimierer korrespondiert.

5.1.2 Fourier-Motzkin-Elimination

Die Fourier-Motzkin-Elimination[2] ist eine Lösungstechnik für lineare Ungleichungssysteme $Ax \leq b$. Sie ist die konsequente Erweiterung der Gauß-Elimination auf Ungleichungen und dient zur Zulässigkeitsprüfung und zum Ermitteln der Lösungsmenge, die durch ein Dreiecksystem von Ungleichungen beschrieben wird. Das Vorgehen wird in der folgenden Handlungsanweisung kurz beschrieben und im nachfolgenden Beispiel erläutert. Für eine mathematisch strengere Darstellung sei auf [385] und auf die hinsichtlich des Aufwands verbesserten Varianten von Černikov [120], Kohler [353], Imbert [310], [309] verwiesen.

Fourier-Motzkin-Eliminationsalgorithmus:

- Teile das Ungleichungssystem in drei Teile:
  - Ungleichungen, die die zu eliminierende Variable nicht enthalten
  - Ungleichungen, die obere Schranken für die zu eliminierende Variable sind
  - Ungleichungen, die untere Schranken für die zu eliminierende Variable sind
- Erstelle ein neues System von Ungleichungen aus allen Ungleichungen, die die zu eliminierende Variable nicht enthalten, und allen Paaren von unteren und oberen Schranken. Die maximale Zahl neuer Ungleichungen, die aus p Ungleichungen entstehen kann, beträgt $\frac{1}{4}p^2$; sie entsteht bei einer hälftigen Aufteilung in untere und obere Schranken.
- Beginne von vorn und eliminiere die nächste Variable, und zwar solange bis für die letzte Variable ein Intervall verbleibt. Sollte kein Intervall existieren (untere Schranke größer als obere), dann hat das System keine Lösung.

Dieser Algorithmus wird im folgenden Beispiel erläutert.

[2] Entwickelt von Fourier (1826), wiederentdeckt von Motzkin (1936).

Beispiel 5.5 (Fourier-Motzkin-Elimination)
Zur Veranschaulichung der Fourier-Motzkin-Elimination dient das folgende System von Ungleichungen

$$\begin{aligned} -4x - 3y &\leq -15 \\ 8x + 4y &\leq 56 \\ -8x + 4y &\leq 20 \\ 3x - 2y &\leq 9. \end{aligned}$$

Auf dieses System werden die Schritte

1. Schritt: Separieren von y
2. Schritt: Eliminieren von y durch Kombinieren der Ungleichungen
3. Schritt: Vereinfachen

angewendet und es ergibt sich

$$\underbrace{\begin{aligned} y &\geq 5 - \tfrac{4}{3}x \\ y &\leq 14 - 2x \\ y &\leq 5 + 2x \\ y &\geq -\tfrac{9}{2} + \tfrac{3}{2}x \end{aligned}}_{\text{nach dem 1. Schritt}} \qquad \underbrace{\begin{aligned} 5 - \tfrac{4}{3}x &\leq 14 - 2x \\ -\tfrac{9}{2} + \tfrac{3}{2}x &\leq 14 - 2x \\ 5 - \tfrac{4}{3}x &\leq 5 + 2x \\ -\tfrac{9}{2} + \tfrac{3}{2}x &\leq 5 + 2x \end{aligned}}_{\text{nach dem 2. Schritt}} \qquad \underbrace{\begin{aligned} x &\leq \tfrac{27}{2} \\ x &\leq \tfrac{37}{7} \\ x &\geq 0 \\ x &\geq -19. \end{aligned}}_{\text{nach dem 3. Schritt}}$$

Hierbei entsteht die oberste Ungleichung nach dem 2. Schritt beispielsweise durch Kombination aus $y \geq 5 - \frac{4}{3}x$ und $y \leq 14 - 2x$. Im 3. Schritt wird nach x aufgelöst, und alle x-Ungleichungen liefern ein Intervall für x, während durch Zusammenfassen der rechten Seiten des Blocks „nach dem 1. Schritt“ eine Ungleichung für y entsteht:

$$\begin{aligned} L &= \{(x,y) : 0 \leq x \leq \tfrac{37}{7}, \max\{5 - \tfrac{4}{3}x, -\tfrac{9}{2} + \tfrac{3}{2}x\} \leq y \leq \min\{14 - 2x, 5 + 2x\}\} \\ &\subset \{(x,y) : 0 \leq x \leq \tfrac{37}{7}, \tfrac{9}{17} \leq y \leq \tfrac{19}{2}\} \end{aligned}$$

Die relaxierte Ungleichung $\frac{9}{17} \leq y \leq \frac{19}{2}$ entsteht, indem die Schnittpunkte der in der Maximum- bzw. Minimumfunktion stehenden Terme bestimmt werden.
Alternativ kann L auch über ein y-Intervall mit y-abhängigen x-Restriktionen beschrieben werden, wenn aus den Ungleichungen x eliminiert wird. $y \leq \frac{19}{2}$ folgt dann aus der Addition von $8x + 4y \leq 56$ und $-8x + 4y \leq 20$, während sich $\frac{9}{17} \leq y$ aus der Kombination von $-4x - 3y \leq -15$ und $3x - 2y \leq 9$ ergibt.

Die Fourier-Motzkin-Elimination erweist sich auch für Probleme mit nichtlinearen Gleichungen und linearen Ungleichungen als vorteilhaft. So bleibt die Linearität der Ungleichungen bei der Fourier-Elimination von z. B. x_1 erhalten, während sie beim formalen Einsetzen von $x_1 = x_2^2$ in die Ungleichungen verloren geht. Bei affinen Gleichungsrestriktionen $x_i = c^T x + d$ mit $c_i = 0$ lässt sich x_i dagegen durch direktes Einsetzen einfacher als über den Algorithmus eliminieren.

Beispiel 5.6 (Relaxation durch achsparallelen Hyperkubus)
In Beispiel 5.5 wurde das Polytop durch ein umschließendes achsparalleles Rechteck relaxiert, das eine schwächere, notwendige, aber einfachere Restriktion darstellt. Für komplexere Probleme können die unteren bzw. oberen Schranken der x_i statt per Elimination durch eine Folge $i = 1, \ldots, n$ von LP-Problemen

$$x_i \stackrel{!}{=} \text{Min bzw. Max} \qquad Ax \leq b \tag{5.3}$$

ermittelt werden. Hierfür müssen die A, b aber mit Zahlen belegt sein, während bei der Elimination auch mit variablen Koeffizienten (Vorzeichen und Größenordnungen seien bekannt) gearbeitet werden kann.

5.1.3 Elimination redundanter Ungleichungsrestriktionen

Oft werden in der Praxis zunächst alle denkbaren und sinnvollen Restriktionen einfach hingeschrieben. Dabei kommt es vor, dass Gleichungen und Ungleichungen überflüssig, also redundant, sind. So ist $x \geq 0$ wegen $\mathcal{F} = \{x \in \mathbb{R} : x \geq 0, x \geq 1\} = \{x \in \mathbb{R} : x \geq 1\}$ redundant. Neben dem erhöhten Aufwand durch zu viele Restriktionen sind redundante Restriktionen für einige Verfahren auch ein Problem. So führen überflüssige Restriktionen zu linearen bzw. funktionalen Abhängigkeiten, die wiederum Rangverluste entstehen lassen können und die damit verbundenen Schwierigkeiten hervorrufen. Es ist deshalb teils notwendig, zumindest aber zweckmäßig, redundante Gleichungen und Ungleichungen zu erkennen und zu eliminieren. Für lineare Gleichungsrestriktionen bieten sich die bekannten Rangtests an, für Ungleichungen beschreibt dieser Abschnitt das Vorgehen.

Definition 5.1 (Redundante Ungleichung)
Eine Ungleichung $g_i(x) \leq 0$ heißt redundant in $\mathcal{F} = \{x \in \mathbb{R}^n : g(x) \leq 0_p, h(x) = 0_m\}$, wenn sich $\mathcal{F}$ bei Weglassen von $g_i(x) \leq 0$ nicht ändert.

Beispiel 5.7 (Redundanz bei LP-Problemen)
Ist $\mathcal{F} = \{x \in \mathbb{R}^n : Ax \geq b, x \geq 0_n\}$, hat die i-te Zeile von A nur nichtnegative Koeffizienten und ist $b_i < 0$, dann ist die i-te Zeile redundant und kann somit weggelassen werden.

Normalerweise ist Redundanz nicht einfach zu erkennen. Allerdings ist klar, dass, wenn es kein x die i-ten Ungleichung verletzt, muss diese redundant sein, da sie immer gilt. Dabei sind alle x jener zulässigen Menge zu betrachten, die die i-te Restriktion nicht enthält. Andernfalls würde die i-te Restriktion ja immer dafür sorgen, dass sie selbst nicht verletzt wird. Auf dieser Überlegung basiert die Vorgehensweise zum Erkennen redundanter Restriktionen.

Methode zum Erkennen redundanter Ungleichungsrestriktionen
Ist das Maximum von

$$g_i(x) \stackrel{!}{=} \text{Max} \qquad x \in \mathbb{R}^n : g_j(x) \leq 0; j = 1, \ldots, p; j \neq i;\ h(x) = 0_m \tag{5.4}$$

kleiner oder gleich Null, dann ist $g_i(x) \leq 0$ redundant, andernfalls nicht. Im Fall linearer Gleichungs- und Ungleichungsrestriktionen sind also einige LP-Probleme zu lösen.

Eine besonders günstige Situation liegt vor, wenn $x_i \geq 0$ in einem System von Restriktionen $Ax = b, x \geq 0_n$ redundant ist. Die Variable x_i wird dann nichtextremale Variable genannt. Sie kann wie eine freie Variable behandelt werden und somit selbst eliminiert werden. Nichtextremale Variable treten dabei gar nicht so selten auf. So gelingt beispielsweise bei drei nichtnegativen Variablen und zwei unabhängigen Gleichungen stets eine Reduktion auf zwei nichtnegative Variable und eine Gleichung. Über das LP-Problem

$$x_i \stackrel{!}{=} \text{Min} \qquad Ax = b,\ x \geq 0_n \tag{5.5}$$

kann x_i als nichtextremal erkannt werden, siehe das sog. Nichtextremale-Variablen-Theorem [419]. Wurde x_i als nichtextremal erkannt, kann es eliminiert werden, indem x_i als affine Funktion durch eine der Gleichungen aus $Ax = b$ ausgedrückt und in die verbleibenden Gleichungen und die Zielfunktion eingesetzt wird, s. hierzu das nachfolgende Beispiel.

Beispiel 5.8 (Nichtextremale Variable)
Es gelte $x_1 \geq 0, x_2 \geq 0, x_3 \geq 0$ und

$$x_1 + 3x_2 + 4x_3 = 4 \tag{5.6a}$$
$$2x_1 + x_2 + 2x_3 = 5. \tag{5.6b}$$

Subtraktion der zweiten von der ersten Gleichung gibt $x_1 = 1 + 2x_2 + 2x_3$, weshalb $x_1 \geq 1$ immer gilt und somit $x_1 \geq 0$ redundant ist. Für $x_1 \stackrel{!}{=} \text{Min}$ unter (5.6) und $x_i \geq 0$ folgt über $x_{\text{opt}} = [2, 0, \frac{1}{2}]^T$ ebenfalls Redundanz von x_1, da $x_{1,\text{opt}} = 2 \geq 0$. Die Elimination von x_1 mit $x_1 = 1 + 2x_2 + 2x_3$ reduziert die Restriktionen auf

$$5x_2 + 6x_3 = 3, x_2 \geq 0, x_3 \geq 0$$

und eine Zielfunktion, z. B. $f(x) = x_1 + x_2 + x_3$, um eine Variable, im Beispiel auf $\tilde{f}(\tilde{x}) = 1+3x_2+3x_3$ mit $\tilde{x} = [x_2, x_3]^T$. Die Konstante in dieser Zielfunktion spielt für die Bestimmung des Minimierers keine Rolle und kann weggelassen werden.
Gleichwohl hätte auch $x_1 = 4 - 3x_2 - 4x_3$ eingesetzt in (5.6b) oder $x_1 = \frac{1}{2}(5 - x_2 - 2x_3)$ eingesetzt in (5.6a) zur Elimination genutzt werden können. Auch in diesen Fällen ergibt sich die Gleichungsrestriktion $5x_2 + 6x_3 = 3$; die Zielfunktionen nach Elimination unterscheiden sich aber! Fazit: eine Gleichung, eine Ungleichung und eine Variable weniger.

5.1.4 Elimination singulärer Ungleichungsrestriktionen

Die Ungleichung $x^2 \leq 0$ hat im Gegensatz zu $x^2 \leq -1$ eine zulässige Menge, gilt aber für kein x streng, da $x^2 < 0$ nicht möglich ist. Sie wird deshalb als singuläre Ungleichung bezeichnet. Singuläre Ungleichungen charakterisieren also den Rand der zulässigen Menge, wie etwa $\begin{bmatrix} 0 & x \\ x & y \end{bmatrix} \succeq 0_{2\times 2}$ durch $(x = 0, y \in \mathbb{R}_{\geq})$.

Definition 5.2 (Singuläre Ungleichung)
Eine Ungleichung heißt singulär oder per se aktiv[3], wenn es im nichtleeren zulässigen Bereich $\mathcal{F}$ kein x gibt, für das die Ungleichung streng ist. Sie formuliert deshalb implizit eine Gleichungsrestriktion.

Singuläre Ungleichungen sollten aus dem Problem entfernt und als Gleichungen geführt werden, da Gleichungen einfacher zu behandeln sind und sich zur Variablenelimination eignen. Vielfach müssen sie sogar aus algorithmischen Aspekten ausgeschlossen werden, um vergebliche Versuche der Suchrichtungsfreigabe in das strenge Ungleichheitsgebiet zu verhindern, das ja leer ist. Numerisch nicht ausgereifte Algorithmen stellt diese Art von Singularität sonst vor Probleme.

Singuläre Ungleichungen können über $g_i(x) \stackrel{!}{=} \text{Min}; x \in \mathcal{F}$ anhand von $g_i(x_{\text{opt}}) = 0$ ermittelt werden, vgl. auch die Anmerkung in Abschn. 6.5 zu inkonsistenten Ungleichungen. Im Fall linearer Gleichungs- und Ungleichungsrestriktionen reduziert sich die Analyse also auf die Lösung einiger LP-Probleme. Für solche Probleme in Standardform charakterisieren per se aktive Restriktionen, die sog. Nullvariablen, also jene Variable x_i, die für $Ax = b; x \geq 0_n$ Null als einzigen zulässigen Wert haben. Die Elimination einer Nullvariablen x_i geschieht einfach durch Streichen der i-ten Spalte von A und den Verzicht auf die Restriktion $x_i \geq 0$. Neben der beschriebenen allgemeinen Methode zum Erkennen impliziter Gleichungen kann für den Spezialfall das Nullvariablentheorem [419] herangezogen werden, dass eine algorithmische Bestimmung aller Nullvariablen über die Simplexmethode gestattet.

Beispiel 5.9 (Nullvariablen)
Es gelte $x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0$ und

$$\begin{aligned} 2x_1 - 5x_2 + 4x_3 - 8x_4 &= 10 \\ x_1 - 2x_2 + 3x_3 - 4x_4 &= 5. \end{aligned}$$

Kombinieren der Gleichungen liefert $x_2 + 2x_3 = 0$ und erzwingt $x_2 = x_3 = 0$ wegen $x_{2,3} \geq 0$.

[3] Die Bezeichnung „per se aktiv“ erfolgt in Anlehnung an Optimierungsalgorithmen, in denen eine Ungleichung aktiv genannt wird, wenn sie im aktuellen Schritt als Gleichung behandelt wird.

5.1.5 Elimination linearer funktionaler Ungleichungen

In rekursiven Algorithmen sind lineare funktionale Ungleichungen $c_i^T x \geq d_i$ schwieriger zu handhaben als einfache Vorzeichenrestriktionen $x_i \geq 0$; bei der Nutzung von Standardsoftware spielt das eine eher untergeordnete Rolle. Die Möglichkeit, die Anzahl der funktionalen Restriktionen zu reduzieren, bietet sich, wenn die funktionalen Restriktionen vorzeichenfreie Variable enthalten, d. h. Variable ohne explizite Vorzeichenrestriktion $x_i \geq 0$. Statt nun vorzeichenfreie Variablen auf klassische Weise $x_i =: y_i - z_i;\ y_i, z_i \geq 0$ in nichtnegative umzuwandeln und damit die Anzahl der Variablen und Restriktionen zu erhöhen, ist ihre Elimination zweckmäßiger. Dabei verringert sich die Zahl der Variablen und Restriktionen nicht zwingend. Bei Problemen ohne Gleichungsrestriktion reduziert sich aber die Anzahl der Restriktionen $c_i^T x \geq d_i$ um die Anzahl der vorzeichenfreien Variablen, vgl. Beispiel 5.10.

Beispiel 5.10 (Elimination linearer funktionaler Ungleichungen)
Im linken, zu lösenden Problem (5.7) sind x_1 und x_2 vorzeichenfrei. Nach Einführen der Schlupfvariablen x_3, x_4 (s. auch Abschnitt 6.3) ergibt sich die rechte Darstellung

$$\begin{array}{lll} x_1 + 2x_2 \stackrel{!}{=} \text{Min} & x_2 \geq 3 - x_1 & \Rightarrow \quad x_1 + 2x_2 \stackrel{!}{=} \text{Min} \qquad x_1 + x_2 \geq 3 \\ & x_2 \leq 3 + x_1 & \qquad\qquad\qquad\quad -x_1 + x_2 + x_3 = 3 \\ & x_1 \leq 3 & \qquad\qquad\qquad\quad x_1 + x_4 = 3 \\ & & \qquad\qquad\qquad\quad x_3, x_4 \geq 0. \end{array} \tag{5.7}$$

Die Elimination von $x_1 = 3 - x_4$ und anschließend von $x_2 = 3 + x_1 - x_3 = 6 - x_3 - x_4$ liefert

$$-2x_3 - 3x_4 + 15 \stackrel{!}{=} \text{Min} \qquad -x_3 - 2x_4 \geq -6;\quad x_3, x_4 \geq 0 \tag{5.8}$$

(15 kann bei der Minimiererberechnung weggelassen werden). Wie (5.7, links) treten drei Restriktionen auf, allerdings haben sich die vom Typ $c_i^T x \geq d_i$ um zwei (Anzahl vorzeichenfreier Variablen) verringert. Die Lösungen lauten: $x_3 = 6, x_4 = 0 \Rightarrow x_1 = 3, x_2 = 0$.

Das im Beispiel praktizierte Vorgehen lässt sich auf allgemeinere Probleme mit linearen Restriktionen übertragen. Dabei empfiehlt es sich das folgende Vorgehen.

Methode zur Elimination linearer funktionaler Ungleichungsrestriktionen

- Elimination der Gleichungsrestriktionen gemäß Beispiel 5.12
- Einführen von soviel Schlupfvariablen, wie vorzeichenfreie Variablen vorhanden sind
- Elimination der vorzeichenfreien Variablen

Anmerkung 5.3 Die Schlupfvariablen sind natürlich in Ungleichungen einzuführen, die die vorzeichenfreien Variablen enthalten. Außerdem ist darauf zu achten, dass die entstehenden Gleichungen bezüglich der vorzeichenfreien Variablen auflösbar sind (Invertierbarkeit der zugeordneten Koeffizientenmatrix). Im Beispiel hätte also auch die erste und zweite Ungleichungsrestriktion ausgewählt werden können. Die Lösung für die Schlupfvariablen wäre dann eine andere, die für x_1 und x_2 natürlich nicht.

5.1.6 Ersetzen durch Gleichungsrestriktion

Bei der Modellierung werden die zu fordernden Eigenschaften zunächst meist verbal notiert, um sie danach strukturell über den Modellansatz oder durch Parameterrestriktionen mathematisch umzusetzen. In Verbindung mit dem Gütekriterium ist dabei darauf zu achten, dass eine möglichst kleine Lösungsmenge oder noch besser eine eindeutige Lösung entsteht. Hierfür eignen sich Gleichungsrestriktionen besser als Ungleichungsrestriktionen, da sie einschränkender sind. Es ist deshalb ratsam, die Menge der Ungleichungsrestriktionen dahingehend zu untersuchen, ob sie nicht Ungleichungen enthält, die durch Gleichungen ersetzt werden können, um die Lösungsmenge zu verkleinern. Gerade für geometrische Probleme (Geraden-, Kreis-, Ellipsen-Fit) in der Bildverarbeitung und Quellenlokalisation oder bei Hurwitz-Polynomen ist das möglich. So lässt sich unter den notwendigen Bedingung $a_i \geq 0$ für Hurwitz-Polynome eine Ungleichung durch $a_j = 1$ ersetzen. Das nachfolgende Beispiel vertieft die Idee am Ellipsen-Fit, vgl. auch Beispiel 4.16.

Beispiel 5.11 (Ellipsen-Fit)
Gegeben seien Messpunkte in der (x, y)-Ebene, gesucht ist eine Ellipse, die diese Messpunkte approximiert. Wird das Quadrat des Fehlers der impliziten Kegelschnittgleichung unter der Nebenbedingung „Ellipse“ minimiert, ergibt sich

$$\sum_{i=1}^{N}(ax_i^2 + bx_iy_i + cy_i^2 + dx_i + ey_i + f)^2 \stackrel{!}{=} \text{Min} \qquad 4ac - b^2 \geq 0.$$

Das Problem ist nicht richtig gestellt, da eine Ellipse nur durch fünf (Mittelpunkt 2, Halbachsenlängen 2, Ausrichtung 1) und nicht wie hier durch sechs Parameter beschrieben wird. Statt nun einen Parameter zu fixieren, ist es geschickter, die Ungleichungsnebenbedingung in die Gleichungsnebenbedingung $4ac - b^2 = 1$ umzuformen. Mit $\phi_i = [x_i^2, x_iy_i, y_i^2, x_i, y_i, 1]^T$ und $\Phi = [\phi_1, \ldots, \phi_N]^T$ sowie $\theta = [a, b, c, d, e, f]^T$ ergibt sich

$$||\Phi\theta||_2^2 \stackrel{!}{=} \text{Min} \qquad \theta^T \underbrace{\begin{bmatrix} 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}}_{=:B} \theta = 1,$$

wobei θ_{opt} der auf $\theta^T B\theta = 1$ normierte Eigenvektor von $\Phi^T\Phi\theta = \lambda B\theta$ ist, der zu dem einzigen positiven verallgemeinerten Eigenwert gehört. Die Lösung ist eindeutig [193].

5.2 Elimination von Gleichungsrestriktionen

Gleichungsrestriktionen, insbesondere lineare Gleichungsrestriktionen sind in der Modellbildung und in Optimierungsproblemen sehr dankbar. Sie bedeuten eine Einschränkung der Freiheitsgrade, sodass grob gesagt, jede Gleichungsrestriktion zur Reduktion des Problems um eine Variable genutzt werden kann. Aus statistischer Sicht ist das ebenfalls von Vorteil, da die Messdaten pro Schätzvariable relativ gesehen mehr Information enthalten, wodurch diese weniger streuen (gemessen in einem skalaren Maß ihrer Kovarianzmatrix).

Die ersten fünf Abschnitte 5.2.1 bis 5.2.5 beschäftigen sich mit Möglichkeiten der Elimination linearer Restriktionen:

1. Eliminationsmethode (anwendbar bei Standardrestriktion $Ax = b$)
2. Nullraummethode (anwendbar bei Standardrestriktion und Matrixgleichungen)
3. Blockmatrixformulierung (anwendbar bei Summen und elementweisen Restriktionen)
4. Vektorisierung (anwendbar bei Matrixgleichungen und elementweisen Restriktionen)
5. kanonische Zerlegung (anwendbar bei Matrizen mit elementweisen Restriktionen)

Der Unterschied zwischen der Eliminationsmethode und der Nullraummethode besteht darin, dass bei ersterer die Variablen des durch Elimination vereinfachten Problems eine Teilmenge der Variablen des Originalproblems sind, während sie im anderen Fall affine Kombinationen der Originalvariablen darstellen.

Im Abschnitt 5.2.6 wird die Elimination polynomialer Ungleichungen beschrieben. Hierfür können computeralgebraisch umgesetzte Algorithmen verwendet werden, wobei dabei die Anzahl der Gleichungsrestriktionen und die Anzahl der Variablen gering sein muss. Darüber hinaus wird in dem Abschnitt gezeigt, wie insbesondere eine quadratische Restriktion (die Einschränkung der geringen Variablenanzahl entfällt dann) durch Reparametrisierung eliminiert werden kann.

Die in den Abschnitten 5.2.7 und 5.2.8 zur Elimination genutzte Idee basiert darauf, zunächst zusätzliche Gleichungsrestriktionen über die Optimalitätsbedingung zu erzeugen, um diese dann zur Reduktion der Variablenanzahl zu verwenden. Das gelingt für sogenannte separable Probleme, die in den Abschnitten vorgestellt werden. Da die separablen Quadratmittelprobleme in der Modellbildung von besonderem Interesse sind, werden sie in Abschnitt 5.2.8 auch hinsichtlich möglicher Anwendungen und spezieller Algorithmen separat behandelt.

Einen geometrisch motivierten Modellierungszugang stellt die orthogonale Distanzregression bzw. -approximation dar, bei der zu Daten eine optimale Kurve oder Fläche bestimmt wird.

Gesucht werden die Kurven- oder Flächenparameter und die Minimalabstände für jeden Datenpunkt. Als Restriktionen dienen die impliziten funktionalen Kurven- oder Flächenbeschreibungen, die alle korrigierten Datenpunkte (beispielsweise die Menge aller Kreisgleichungen für die Lotpunkte) erfüllen müssen. Für bestimmte Kurven oder Flächen lassen sich diese Gleichungsrestriktionen aber komplett eliminieren, wenn auf die Parameterdarstellung von Kurven oder Flächen ausgewichen wird. Abschnitt 5.2.9 zeigt das Vorgehen und nennt Beispiele. Vom Wesen her basiert der Zugang also auf einer Reparametrisierung.

5.2.1 Eliminationsmethode

Die Eliminationsmethode ist eine Reduktionsmethode, die lineare oder auch nichtlineare Gleichungsrestriktionen nutzt, um die Anzahl der Variablen im Optimierungsproblem zu verkleinern. Dabei sind die Variablen des reduzierten Problems im Unterschied zur Nullraummethode eine Teilmenge der Variablen des ursprünglichen Problems.

Lässt sich $h(x) = 0_m$ auf die Form $x_1 = \tilde{h}(x_2)$ mit $x_1 \in \mathbb{R}^m$, $x_2 \in \mathbb{R}^{n-m}$ bringen, wobei hier ohne Einschränkung der Allgemeinheit $x^T = (x_1^T, x_2^T)$ gelten möge, dann reduziert sich das Standardproblem auf eines der Dimension $n - m$

$$f(\tilde{h}(x_2), x_2) \stackrel{!}{=} \text{Min} \qquad g(\tilde{h}(x_2), x_2) \leq 0. \tag{5.9}$$

Speziell für

$$h(x) = Cx - d = 0 \qquad C \in \mathbb{R}_m^{m \times n} \tag{5.10}$$

kann die Elimination aufbauend auf der Zerlegung $CP = [F, G]$ mit $F \in \mathbb{R}_m^{m \times m}$ vorgenommen werden. P bezeichnet dabei eine Permutationsmatrix, die den Vektor x mit $P^T x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$; $x_1 \in \mathbb{R}^m, x_2 \in \mathbb{R}^{n-m}$ und damit die Matrix C so umordnet, dass F regulär ist. Für die Gleichungsrestriktion gilt dann $Fx_1 + Gx_2 = d$ bzw. $x_1 = F^{-1}(d - Gx_2)$, wodurch die Variable x_1 eliminiert werden kann. Es folgt

$$f\Big(P\begin{bmatrix} F^{-1}(d - Gx_2) \\ x_2 \end{bmatrix}\Big) \stackrel{!}{=} \text{Min} \qquad x_2 \in \mathbb{R}^{n-m} : g\Big(\begin{bmatrix} F^{-1}(d - Gx_2) \\ x_2 \end{bmatrix}\Big) \leq 0. \tag{5.11}$$

Falls C keinen vollen Zeilenrang hat, kann ein zulässiges Problem durch Streichen der redundanten Zeilen auf eines mit vollem Zeilenrang gebracht werden.

Anwendung findet die Eliminationsmethode beispielsweise bei LSE-Problemen, also linearen LS-Problemen mit linearen Gleichungsrestriktionen. Hierfür stehen effiziente Algorithmen zur Verfügung [147], die vergleichbare Ergebnisse wie die Nullraummethode liefern [507]. Exemplarisch für weitere Anwendungen soll die Elimination für LP-Probleme gezeigt werden.

Beispiel 5.12 (Elimination für LP-Probleme)
Ist $c^T x \stackrel{!}{=} \text{Min}; Ax = b, x \geq 0_n; A \in \mathbb{R}_r^{m \times n}$ zulässig, dann kann die Zahl der $m + n$ Restriktionen um m und die der Variablen um r gesenkt werden.
Dieser Reduktionsgewinn lässt sich dabei wie folgt erzielen. Im ersten Schritt werden für $r < m$ entsprechend $m - r$ redundante Gleichungen gestrichen, womit $\tilde{A} \in \mathbb{R}_r^{r \times n}$ entsteht. Im zweiten Schritt wird eine QR-Zerlegung mit Spaltentausch ausgeführt, d. h. $\tilde{A}P = Q[R_1, R_2]$ bzw. $R_1\tilde{x}_1 = Q^T b - R_2\tilde{x}_2$, $R_1 \in \mathbb{R}_r^{r \times r}$ und $P^T x = \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix}$. Mit $c^T P = [\tilde{c}_1^T, \tilde{c}_2^T]$ folgt unter Vernachlässigung des Zielfunktionsterms $\tilde{c}_1^T R_1^{-1} Q^T b$ ein kanonisches LP-Problem $(\tilde{c}_2^T - \tilde{c}_1^T R_1^{-1} R_2)\tilde{x}_2 \stackrel{!}{=} \text{Min}; -R_1^{-1}R_2\tilde{x}_2 \geq -R_1^{-1}Q^T b, \tilde{x}_2 \geq 0_{n-r}$ mit nur noch n Restriktionen in $n - r$ Variablen.

Abschließend sei vor einem allzu formalen Umgang bei der Elimination gewarnt! So können Eliminationen die Kondition eines Problems verschlechtern, wodurch sich Fehler in den Eingangsdaten oder durch Rundungen stark verstärken, was unter Umständen unbrauchbare Lösungen hervorruft, siehe hierzu die Diskussion in Beispiel 5.13.

Ein anders gelagertes Problem erwächst, wenn bei der formalen Elimination, versteckte Restriktionen gleich mit eliminiert werden. Beispiel 5.14 soll den Leser hierfür sensibilisieren.

Beispiel 5.13 (Verschlechterung der Kondition durch Elimination)
Die verallgemeinerten Regressionsprobleme

$$||v||_2 \stackrel{!}{=} \text{Min} \qquad b = Ax + Lv \tag{5.12}$$

und

$$||Ax - b||_W \stackrel{!}{=} \text{Min} \qquad \text{mit} \quad W = L^T L = (\text{cov}\boldsymbol{\varepsilon})^{-1} \tag{5.13}$$

sowie

$$||L^{-1}(Ax - b)||_2 \stackrel{!}{=} \text{Min} \tag{5.14}$$

sind für reguläres L äquivalent. Im Gegensatz zu (5.13) und (5.14) ist die auf Paige [494] zurückgehende Darstellung (5.12) auch definiert, wenn A oder L nicht vollen Rang haben. Abgesehen vom rangdefizienten Fall ist die Auswertung der ersten Formulierung mit QR-Faktorisierungen numerisch besser, da bei schlecht konditionierter Matrix L die Lösung x wegen $L^{-1}A$ schlecht bestimmt wird. Für algorithmische Details sei auf [230] verwiesen.

Beispiel 5.14 (Fehler bei formaler Elimination)
Obwohl $x_1^2 + x_2^2 \stackrel{!}{=} \text{Min}; (x_1 - 1)^3 = x_2^2$ den Minimierer $(1, 0)$ hat, führt die formale Elimination von x_2 auf ein Problem, das kein Minimum besitzt, nämlich auf $x_1^2 + (x_1 - 1)^3 \stackrel{!}{=} \text{Min}$. Der Fehler steckt im Nichtbeachten der impliziten Restriktion $x_1 \geq 1$ aus $(x_1 - 1)^3 = x_2^2$, die gleichfalls mit eliminiert wurde. Bei korrekter Elimination ergibt sich demnach $x_1^2 + (x_1 - 1)^3 \stackrel{!}{=} \text{Min}; x_1 \geq 1$ mit dem Minimierer $x_1 = 1$.

5.2.2 Nullraummethode

Die Nullraummethode ist eine Reduktionsmethode, die Gleichungsrestriktionen $Cx = d$ eliminiert.[4] Die neuen Variablen sind Linearfaktoren für Vektoren, die den Nullraum von C aufspannen. Im Gegensatz zur Eliminationsmethode sind die Variablen des reduzierten Problems gemeinhin keine Teilmenge der Variablen des ursprünglichen Problems.

Die allgemeine Lösung für $Cx = d$ lautet

$$x = C^{+}b + Qy \qquad Q \in \mathbb{R}^{n\times(n-r)}_{(n-r)}, \mathcal{R}(Q) = \mathcal{N}(C); y \in \mathbb{R}^{n-r}; C \in \mathbb{R}^{m\times n}_{r}. \tag{5.15}$$

Hier bezeichnet Q eine Matrix mit vollem Spaltenrang, deren Spaltenvektoren den Nullraum von C aufspannen. Q lässt sich dabei über eine QR-Faktorisierung oder eine Singulärwertzerlegung von C berechnen. Per Substitution (5.15) kann damit ein Problem in den neuen Variablen y formuliert werden.

An zwei Beispielen soll das Vorgehen demonstriert werden, wobei das erste eine Handrechnung ist und das zweite direkt einen Algorithmus formuliert.

Beispiel 5.15 (Elimination einer Spurrestriktion)
Gegeben sei $X \succeq 0_{2\times2}, \text{spur}X = 1$.
Aus der Spurrestriktion $x_{11} + x_{22} = 1$, d. h. $[1, 1, 0][x_{11}, x_{22}, x_{12}]^T = 1$, kurz $Cx = d$, folgt

$$\begin{bmatrix} x_{11} \\ x_{22} \\ x_{12} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{bmatrix} + \begin{bmatrix} -1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

und nach Substitution

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{12} & x_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} + y_1 \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} + y_2 \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \succeq \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Beispiel 5.16 (Nullraummethode für das LSE-Problem)
Basierend auf der Zerlegung von $x_{LSE} = x_1 + x_2$ mit $x_1 \in \mathcal{N}(C)^{\perp}$[5] und $x_2 \in \mathcal{N}(C)$ leitet sich die sogenannte Nullraummethode für das LSE-Problem ab, von der es zahlreiche Varianten gibt. In Tabelle 5.2 ist ein Algorithmus dargestellt, der sich auf eine QR-Faktorisierung der Nebenbedingungen stützt. Zu vermerken ist auch die Formulierung in Schritt 4, bei der x_2 nicht direkt über $||Ax_2 - (b - Ax_1)||_2 \stackrel{!}{=} \text{Min}; x_2 \in \mathbb{R}^n$, sondern aufwandsreduziert über die Zwischenvariable y_2 (kleinerer Dimension als x_2) bestimmt wird.

[4] Matrixrestriktionen $CX = D$ oder allgemeinere lineare Matrixrestriktionen können über den Weg der Vektorisierung oder basierend auf ihren allgemeinen Lösungsdarstellungen völlig analog behandelt werden.

[5] $\mathcal{N}(C)^{\perp}$ bezeichnet das orthogonale Komplement von $\mathcal{N}(C)$, d. h. die Menge der Vektoren, die senkrecht auf $\mathcal{N}(C)$ stehen.

Die Nullraummethode erweist sich in Tests [507] für Standardprobleme als numerisch stark (Aufwand, Genauigkeit). Für große, schwach besetzte Probleme ist sie hingegen ebenso ungeeignet wie für Probleme, die bei festem A für verschiedene C zu lösen sind, da bei denen dann AQ_2 ständig neu zu berechnen ist.

1.	Berechne QR-Faktorisierung von C^T $C^T = [Q_1, Q_2]\begin{bmatrix} R_1 \\ 0_{(n-r)\times r} \end{bmatrix}$; $Q_1 \in \mathcal{O}_{n,r}, Q_2 \in \mathcal{O}_{n,n-r}$
2.	Löse das Dreieckssystem $R_1^T y_1 = d$; $y_1 \in \mathbb{R}^r$
3.	Setze $x_1 = Q_1 y_1$
4.	Löse $\|\|(AQ_2)y_2 - (b - Ax_1)\|\|_2 \stackrel{!}{=} \text{Min} \quad y_2 \in \mathbb{R}^{n-r}$
5.	$x_{LSE} = x_1 + Q_2 y_2$ Anmerkung: $\mathcal{R}(Q_2) = \mathcal{N}(C)$

Tabelle 5.2: LSE-Algorithmus nach der Nullraummethode [78]

5.2.3 Elimination durch Blockmatrixformulierung

Parameterlineare Schätzprobleme, bei denen Teilprobleme über gemeinsame Parameter miteinander gekoppelt sind, führen oft auf Gütekriterien der Art

$$\sum_{i=1}^{m} w_i^2 ||A_i x_i - b_i||_2^2 \stackrel{!}{=} \text{Min} \qquad \begin{array}{l} x_i \in \mathbb{R}^{n_i}; A_i \in \mathbb{R}^{m_i \times n_i}, b_i \in \mathbb{R}^{m_i} \\ \text{mit } (x_i)_j = (x_k)_l \text{ für Tupel } (i,j,k,l). \end{array} \tag{5.16}$$

Hierbei stimmen einzelne Komponenten der Vektoren x_i mit denen der Vektoren x_k überein. Die Zielfunktion von (5.16) kann äquivalent geschrieben werden als

$$\left\| \begin{bmatrix} w_1 I_{n_1} & & 0 \\ & \ddots & \\ 0 & & w_m I_{n_m} \end{bmatrix} \left(\begin{bmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_m \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} - \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} \right) \right\|_2^2 \stackrel{!}{=} \text{Min}.$$

Bei der Blockmatrixformulierung werden nun die Restriktionen genutzt, um ein kleineres Schätzproblem mit einem kleinsten gemeinsamen Parametervektor zu formulieren. Dabei sind entsprechende Spalten der Blockmatrix zu streichen, was einem Aufsplitten der Matrizen A_i entspricht. Da eine allgemeine Umformulierung dieses Problems aufgrund der komplexen Indizierung die Darstellung nur erschwert, soll das Prinzip exemplarisch erläutert werden.

Die Vektoren x_1, x_2, x_3 seien zerlegbar in $x_1^T = [x_{1A}^T, x_{1B}^T]$, $x_2^T = [x_{2A}^T, x_{2B}^T]$, $x_3^T = [x_{3A}^T, x_{3B}^T]$. Entsprechend partitioniert seien auch die Matrizen A_i. Ferner mögen die Restriktionen $x_{2A} = x_{1B}$ und $x_{3B} = x_{1A}$ gelten. Der kleinste gemeinsame Vektor lautet dann $x^T = (x_{1A}^T, x_{1B}^T, x_{2B}^T, x_{3A}^T)$ und führt auf die Blockmatrixformulierung

$$\left\| \begin{bmatrix} w_1 I & 0 & 0 \\ 0 & w_2 I & 0 \\ 0 & 0 & w_3 I \end{bmatrix} \left(\begin{bmatrix} A_{1A} & A_{1B} & 0 & 0 \\ 0 & A_{2A} & A_{2B} & 0 \\ A_{3B} & 0 & 0 & A_{3A} \end{bmatrix} \begin{bmatrix} x_{1A} \\ x_{1B} \\ x_{2B} \\ x_{3A} \end{bmatrix} - \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \right) \right\|_2^2 \stackrel{!}{=} \text{Min.}$$

Der Vorteil der Blockmatrixformulierung soll am folgenden Beispiel gezeigt werden.

Beispiel 5.17 (Blockmatrixformulierung für 3D-Diffusionsmodelle)
In einem 3D-Diffusionsmodell (keine Konvektion) gilt [446]

$$4\ddot{C}_i(t) \cdot (t - t_0)^2 + 6\dot{C}_i(t) \cdot (t - t_0) - \dot{C}_i(t)\frac{r_i^2}{K} = 0; \qquad i = 1, \ldots, n$$

für den Konzentrationsverlauf $C_i(t)$ am Messort x_i, wenn zum Zeitpunkt t_0 ein sprungförmiger Stoffeintrag erfolgt und die Quelle vom Messort den Abstand r_i hat (K ist der Diffusionskoeffizient). Wird das Diffusionsmodell so umgeschrieben, dass t_0, t_0^2, r_i in einem Parametervektor auftreten, dann entsteht nach Einsetzen der Zeitpunkte t_k ein überstimmtes Gleichungssystem für eine LS-Formulierung pro Messort

$$\underbrace{\left[-8\ddot{C}_i(t_k)t_k - 6\dot{C}_i(t_k), \quad -4\ddot{C}_i(t_k), \right.}_{A_i(:,2)} \left. -\dot{C}_i(t_k) \right] \begin{bmatrix} t_{0,i} \\ t_{0,i}^2 \\ r_i^2/K \end{bmatrix} \cong \underbrace{-4\ddot{C}_i(t_k)t_k^2 - 6\dot{C}_i(t_k)t_k.}_{b_i(:)}$$

Bei dieser Formulierung wird jedoch für jeden Messort ein eigenes $\hat{t}_{0,i}$ geschätzt. Statt über alle Schätzungen $\hat{t}_{0,i}$ zu mitteln, um sich so auf den gemeinsamen Startzeitpunkt t_0 zu einigen, ist es besser, die Restriktion $t_{0,1} = \ldots = t_{0,n}$ zu stellen, um von vornherein nur einen Startzeitpunkt zuzulassen. Durch eine Blockmatrixformulierung können die hinzugenommenen Restriktionen aber elegant eliminiert werden

$$\left\| \begin{bmatrix} A_1(:,2) & -\dot{C}_1(:) & 0 & \ldots & 0 \\ A_2(:,2) & 0 & -\dot{C}_2(:) & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ A_n(:,2) & 0 & \ldots & 0 & -\dot{C}_n(:) \end{bmatrix} \begin{bmatrix} t_0 \\ t_0^2 \\ r_1^2/K \\ \vdots \\ r_n^2/K \end{bmatrix} - \begin{bmatrix} b_1(:) \\ b_2(:) \\ \vdots \\ b_n(:) \end{bmatrix} \right\|_2^2 \stackrel{!}{=} \text{Min.}$$

Bezeichne θ den Parametervektor, so sichert $\theta_1^2 = \theta_2$ die Widerspruchsfreiheit bezüglich t_0 und $\theta_3, \ldots, \theta_{n+2} \geq 0$ die Nichtnegativität von r_i^2/K. Aus den geschätzten r_i^2/K kann in einem zweiten Schritt die Lage der Quelle als gemeinsamer Schnitt aller über K skalierbaren Kreise ermittelt werden [446], [254].

5.2.4 Elimination durch Vektorisierung

Für affine Restriktionen an Matrizen (Toeplitz-Matrix, symmetrische Matrix, Dreiecksmatrix) bietet sich die Vektorisierung an, bei der die Matrix in einen Vektor angeordnet wird

$$\text{vec} : \mathbb{R}^{m\times n} \to \mathbb{R}^{mn}; \quad \text{vec}\, X \stackrel{\text{def}}{=} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \text{mit } x_j \in \mathbb{R}^m. \tag{5.17}$$

Die an die Matrix gestellten Forderungen lassen sich dann äquivalent als Restriktionen $Cx = d$ mit $x := \text{vec}\, X$ ausdrücken, die ihrerseits in einem weiteren Schritt zur Variablenreduktion herangezogen werden können. Damit allerdings eine zweckmäßige Umformulierung des Gesamtproblems gelingt, darf die Vektorisierung nicht die Zielfunktion komplizierter gestalten. Aus diesem Grund kommt die Vektorisierung überwiegend zum Einsatz, wenn in der Zielfunktion die Frobenius-Norm auftritt. Gemeinsam mit der wichtigen Regel [382]

$$\text{vec}\,(AXB) = (B^T \otimes A)\,\text{vec}\, X, \tag{5.18}$$

in der $\otimes$ das Kronecker-Produkt $A \otimes B \stackrel{\text{def}}{=} ((a_{ij}B))$ bezeichnet, gelingt eine zweckmäßige Umformulierung, da dann die euklidischen Vektornorm in der Zielfunktion die Rolle der Frobenius-Norm einnimmt.
Das nachfolgende Beispiel zeigt das Aufstellen der Restriktionen und und das umformulierte Problem, das sich mit den Methoden aus Abschnitt 5.2.1 und 5.2.2 lösen lässt.

Beispiel 5.18 (Vektorisierung einer Toeplitz-Matrix mit Einsdiagonale)
Es bezeichne $\mathcal{T}_3$ die spezielle Teilmenge der Toeplitz-Matrizen mit Einsdiagonale. Dann können die Bedingungen $x_{11} = 1$, $x_{22} = 1$, $x_{33} = 1$ sowie $x_{12} = x_{23}$ (gleichbedeutend $x_{12} - x_{23} = 0$) und $x_{21} = x_{32}$ kompakt als $Cx = d$ mit $x = \text{vec}\, X$ geschrieben werden:

$$\begin{bmatrix} 1 & x_1 & x_2 \\ x_3 & 1 & x_1 \\ x_4 & x_3 & 1 \end{bmatrix} \quad \Rightarrow \quad \underbrace{\begin{bmatrix} 1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1 \\ 0&1&0&0&0&-1&0&0&0 \\ 0&0&0&1&0&0&0&-1&0 \end{bmatrix}}_{C} \underbrace{\begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ x_{12} \\ x_{22} \\ x_{32} \\ x_{13} \\ x_{23} \\ x_{33} \end{bmatrix}}_{x=\text{vec}X} = \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}}_{d} \tag{5.19}$$

Anwenden der Vektorisierungsregel (5.18) liefert eine äquivalente Umformulierung eines linearrestringierten Matrix-LS-Problems in ein LSE-Problem

$$\frac{1}{2}||AX - B||_F^2 \stackrel{!}{=} \text{Min}; X \in \mathcal{T}_3 \quad \Leftrightarrow \quad \frac{1}{2}||(I_n \otimes A)x - \text{vec}\, B||_2^2 \stackrel{!}{=} \text{Min}; Cx = d. \tag{5.20}$$

5.2.5 Elimination über kanonische Zerlegungen

Eine Einschränkung für die im vorherigen Abschnitt beschriebene Vektorisierungstechnik ist, dass die Zielfunktion nach der Umformung eine zweckmäßige Darstellung hat. Bei Frobenius-Norm-Formulierungen entstehen klassische Euklidische-Norm-Formulierungen. Etwas komplizierter, aber noch handhabbar wird die Umformung bei Spalten- oder Zeilensummennorm-Formulierungen. Für Spektralnormformulierungen eignet sich die Vektorisierung hingegen nicht, da sich die Spektralnorm nicht zweckmäßig über Vektornormen ausdrücken lässt[6] und da die inverse Vektorisierung (Vektor-zu-Matrix-Abbildung; reshaping) die Zielfunktion komplizierter gestaltet. Zudem können weitere im Problem auftretende Restriktionen (nichtlineare, LMI-Ungleichungen) gegen den Einsatz der Vektorisierungstechnik sprechen. Eine Alternative sind dann bei affinen Restriktionen an Matrizen kanonische Zerlegungen. Dabei wird ausgenutzt, dass sich affine Räume $\mathcal{L}_{A_0} \subseteq \mathbb{R}^{m\times n}$ als Linearkombinationen von Basismatrizen plus eine konstante Matrix darstellen lassen:

$$X = A_0 + \sum_{i=1}^{\dim \mathcal{L}} x_i A_i \tag{5.21}$$

mit $\dim \mathcal{L}$ freien Variablen x_i. Überall dort, wo X im Problem auftritt, wird es gemäß (5.21) ersetzt, z. B. wird aus BXC dann $BA_0C + \sum_{i=1}^{\dim \mathcal{L}} x_i BA_iC$ mit den konstanten Matrizen BA_iC. Die Optimierung erfolgt über die freien Variablen x_i.
Die kanonische Zerlegung selbst ist einfach zu erstellen. Hierzu werden alle fixen Elemente der Matrix X in A_0 eingetragen. Anschließend werden die Basismatrizen A_i angelegt, wobei Einsen an Stellen gleicher Elemente oder Wichtungsfaktoren an die Stellen abhängiger Elemente eingetragen werden. Im nachfolgenden Beispiel sind einige typische Fälle dargestellt.

Beispiel 5.19 (Kanonische Zerlegung affiner Matrizen)
Für die Toeplitz-Matrix aus Beispiel 5.18 ergibt sich die Darstellung

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + x_1 \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} + x_4 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

[6] Die Operatornormdefinition $||X||_2 = \max_{y\neq 0_n} \frac{||Xy||_2}{||y||_2}$ stellt zwar den Zusammenhang zur Vektornorm her, die weitere Umformung von X zu $x = \mathrm{vec}X$ führt aber auf unzweckmäßige Ausdrücke, zumal mit y eine weitere Variable ins Spiel kommt.

Auch für etwas komplizierter strukturierte Matrizen gelingt die Zerlegung

$$\begin{aligned} X &= \begin{bmatrix} 3+x_1 & 4+x_2 & x_3 \\ x_2 & x_1 & x_2+2x_1 \\ -x_3 & x_2 & 5 \end{bmatrix} \\ &= \begin{bmatrix} 3 & 4 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 5 \end{bmatrix} + x_1 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Durch zusätzliche Restriktionen $Cx = d$, wie sie etwa aus

$$\text{spur}X = 1 \qquad \Leftrightarrow \qquad x_{11} + \ldots + x_{nn} = 1 \tag{5.22}$$

hervorgehen, kann die Variablenzahl weiter gesenkt werden. Das geschieht beispielsweise über die Nullraummethode mit $x = C^+d + Qy$, wobei $Q = ((q_{ij}))$ eine spaltenreguläre Matrix ist, deren Spalten den Nullraum von C aufspannen und $y \in \mathbb{R}^p$ die neuen Variablen bezeichnet. Aus (5.21) wird so

$$X = \Big(A_0 + \sum_{i=1}^{\dim \mathcal{L}} (C^+d)_i A_i\Big) + \sum_{j=1}^{p} y_j \Big(\sum_{i=1}^{\dim \mathcal{L}} q_{ij} A_i\Big), \tag{5.23}$$

wobei in den großen Klammern wieder vorab berechenbare Matrizen stehen, sodass sich die gleiche Form wie in (5.21) nur mit reduzierter Summation ergibt.

5.2.6 Elimination polynomialer Gleichungen

Einen Zugang zur Elimination polynomialer Gleichungen bieten das Konzept der Gröbner-Basen[7] [103]. Primär löst es das Idealzugehörigkeitsproblem, von dessen Lösung aber viele andere Probleme abhängen. So lässt sich etwa eine endliche Menge multivariater Polynome so umformen, dass ein äquivalentes, gut handhabbares System von Polynomen entsteht. Bevorzugt wird der Buchberger-Algorithmus [103], der auf ein Dreieckssystem führt

$$p_1(x_1, \ldots, x_n) = 0 \tag{5.24a}$$

$$\vdots \tag{5.24b}$$

$$p_{n-1}(x_{n-1}, x_n) = 0 \tag{5.24c}$$

$$p_n(x_n) = 0. \tag{5.24d}$$

[7] Eine Gröbner-Basis ist ein endliches Erzeugendensystem zu einem Ideal (Teilmenge eines Ringes, die abgeschlossen bzgl. $\mathbb{R}$-Linearkombinationen ist) in einem Polynomring.

Dieser Algorithmus bzw. seine Varianten stellen Verallgemeinerungen des Gaußschen Eliminationsverfahrens dar und gehören zum Standard moderner Computeralgebrasysteme [212]. Beginnend mit dem univariaten Polynom p_n lassen sich die Wurzeln x_n ermitteln. Werden diese in p_{n-1} eingesetzt, ist wiederum nur ein univariates Polynom, diesmal in der Variablen x_{n-1} zu lösen. Der Prozess wird fortgesetzt, bis alle Wurzeln bekannt sind.

Beispiel 5.20 (Umformung von Polynomgleichungen)
Gegeben seine die Restriktionen

$$x^3y^3 - 2 = 0$$
$$x^3y^2 + y - 1 = 0.$$

Mit der Ordnung $y \succ x$ ergibt sich die Gröbner-Basis $\mathcal{G}_1 = \{y - 3 - 4x^3, 2 + 5x^3 + 4x^6\}$, während $x \succ y$ auf $\mathcal{G}_2 = \{-y+3+4x^3, y^2-y+2\}$ führt. Im letzten Fall kann y durch Lösen der quadratischen Gleichung $y^2 - y + 2 = 0$ bestimmt werden. Das Einsetzen der Lösungen führt auf $-y_{1,2} + 3 + 4x^3 = 0$, woraus je drei Lösungen x_i für y_1 und y_2 folgen. Desgleichen kann mit der Basis $\mathcal{G}_1$ verfahren werden, indem zunächst die Lösungen einer quadratischen Gleichung in x^3 bestimmt werden, aus denen sich danach die Werte y ergeben.

Anmerkung 5.4 Im Beispiel führten die beiden Gleichungen direkt auf Lösungen, wie es für die Ruhelagenbestimmung erforderlich ist. Ist die Zahl der Restriktionen aber kleiner als die der Variablen, so wird die Gröbner-Basis im Regelfall kein univariates Polynom enthalten. So liefert $y - x^2 = 0$ und $z - x^3 = 0$ die Basis $\mathcal{G} = \{-y + x^2, -z + xy, -y^2 + xz, -z^2 + y^3\}$ bei Wahl der Ordnung $x \succ y \succ z$. Im vierten Polynom der Basis wurde dabei x eliminiert. Beachte, dass reelle Nullstellen eines univariaten Polynoms aus der Gröbner-Basis keine Ruhelagen eines Systems sein müssen (komplexe Scheinlösungen in den anderen Variablen).

Neben der Methode mit Gröbner-Basen existieren für einige quadratische Polynomgleichungen allgemeine Lösungsformeln, die zur Elimination herangezogen werden können. Zwei Beispiele sind

$$x^TAx = b^2;\ A \in \mathcal{S}_n^> \qquad \Rightarrow \qquad x = by/\sqrt{y^TAy} \quad \text{für } y \in \mathbb{R}^n \text{ beliebig} \tag{5.26}$$

und

$$X^TAX = I_n;\ A \in \mathcal{S}_m^> \qquad \Rightarrow \qquad X = Y(Y^TAY)^{-1/2} \quad \text{für } Y \in \mathbb{R}_n^{m\times n} \text{ beliebig} \tag{5.27}$$

und nicht zuletzt die zahlreichen Ergebnisse zur Matrix-Ricatti-Gleichung. In ähnlicher Weise können auch einige Normgleichungen beseitigt werden

$$||x||_\rho = 1 \qquad \Rightarrow \qquad x = y/||y||_\rho \quad \text{für } y \in \mathbb{R}^n \text{ beliebig.} \tag{5.28}$$

5.2.7 Elimination mittels der Optimalitätsbedingungen

Bisher wurden zur Elimination nur Gleichungsrestriktionen aus dem Gütekriterium herangezogen. Neben diesen lassen sich weitere Gleichungsrestriktionen über die Optimalitätsbedingung erster Ordnung generieren. Zwei Beispiele sollen die Leistungsfähigkeit dieser Reduktionsmethode verdeutlichen. Darüber hinaus wird auch bei den separablen NLS-Problemen die gleiche Idee genutzt. Sie wird in Abschnitt 5.2.8 beschrieben.

Beispiel 5.21 (Separables LSQ-Problem)
Für das separable LSQ-Problem

$$||[A_1, A_2]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - b||_2^2 \stackrel{!}{=} \text{Min} \qquad x_1 \in \mathbb{R}^r, x_2 \in \mathbb{R}^{n-r} : \frac{1}{2}x_2^T H x_2 + g^T x_2 + f = 0$$

folgt aus der Optimalitätsbedingung erster Ordnung $x_{1,\text{opt}} = A_1^+(b - A_2 x_2)$, was nach Elimination

$$||(I_m - A_1 A_1^+)(A_2 x_2 - b)||_2^2 \stackrel{!}{=} \text{Min} \qquad x_2 \in \mathbb{R}^{n-r} : \frac{1}{2}x_2^T H x_2 + g^T x_2 + f = 0.$$

liefert. Als Anwendung sei der Ellipsen-Fit mit den quadratischen Restriktionen nach Fitzgibbon und Bookstein aus Beispiel 4.16 genannt.

Beispiel 5.22 (Bakengestützte Positionsbestimmung)
Gesucht sind die Position x und die Rotationsmatrix A für die Verdrehung vom lokalen System (Fahrzeug) gegenüber einem globalen System, wobei gestörte Messungen der Vektoren p_i (lokale Bakenkoordinaten) vorliegen und die globalen Bakenkoordinaten b_i bekannt sind. Als Kriterium empfiehlt sich $\sum_{i=1}^N ||x + Ap_i - b_i||_2^2 \stackrel{!}{=} \text{Min}; A^T A = I_3, \det A = 1$.
Da x separabel und frei ist, liefert die Optimalitätsbedingung erster Ordnung $x_{\text{opt}} = \bar{b} - A_{\text{opt}}\bar{p}$, wobei $\bar{b}$ bzw. $\bar{p}$ die Mittelwerte der b_i bzw. p_i sind. Mit der hierzu kompatiblen Forderung $x = \bar{b} - A\bar{p}$ wird x in der Zielfunktion eliminiert, und es folgt

$$\sum_{i=1}^N ||A(p_i - \bar{p}) - (b_i - \bar{b})||_2^2 \stackrel{!}{=} \text{Min}; \qquad A^T A = I_3, \det A = 1.$$

Für Details zur Lösung für A, zur Berechnung der Winkel aus A und zur Einbeziehung zusätzlicher Ebenenrestriktion sei auf [252] verwiesen.

Beide Beispiele zeigen, dass durch die analytische Berechnung der notwendigen Bedingungen einer Optimierungsaufgabe, s. hierzu auch Abschn. 6.6, zusätzliche Restriktionen generiert werden können. Die Restriktionen lassen sich dann wiederum einsetzen, um die Variablenanzahl und damit den Schwierigkeitsgrad der Originalaufgabe zu reduzieren. Besonders häufig wird die ausgeführte Technik bei den sog. separablen LS-Problemen eingesetzt, die Gegenstand des nachfolgenden Abschnitts sind.

5.2.8 Separable nichtlineare Quadratmittelprobleme

Wie im vorangegangen Abschnitt gezeigt wird, lässt sich bei separablen Problemen der Schwierigkeitsgrad senken, wenn die notwendigen Bedingungen erster Ordnung zur Elimination genutzt werden kann. Besonders einfach und zugleich nützlich ist das bei Quadratmittelproblemen, in denen ein oder mehrere Variablen linear in den zu quadrierenden Termen auftreten.

Definition 5.3 (Separables NLS-Problem)
Ein Problem $f(x,z) \stackrel{!}{=} \text{Min}$ mit $(x,z) \in \mathcal{X} \times \mathcal{Z} \subseteq \mathbb{R}^n \times \mathbb{R}^p$ als zulässigen Parameterraum, in dem x für feste z und/oder z für feste x explizit bestimmbar ist, heißt separables Problem. Ist f eine Summe von Quadraten, so wird von einem separablen NLS-Problem gesprochen. Lässt sich die explizite Lösbarkeit dazu nutzen, einen der Variablenvektoren zu eliminieren, heißt das Problem reduzierbar.

Als eine Auswahl solcher separabler Quadratmittelprobleme seien [410], [409], [502]

$$||A - X \otimes Z||_F^2 \stackrel{!}{=} \text{Min} \qquad A \in \mathbb{R}^{(m\cdot p)\times(n\cdot q)}, X \in \mathbb{R}^{m\times n}, Z \in \mathbb{R}^{p\times q} \tag{5.29}$$

$$||A(z)x - b||_2^2 \stackrel{!}{=} \text{Min} \qquad x \in \mathbb{R}^n, z \in \mathbb{R}^p; A : \mathbb{R}^p \to \mathbb{R}^{m\times n}. \tag{5.30}$$

und Beispiel 5.21 genannt. Problem (5.30) tritt dabei in der Modellbildung sehr oft auf, was zahlreiche Anwendungen belegen:

1. Identifikation von Sprung- oder Impulsfunktionen, z. B. [489]

   $$y_i = \theta_1 + \theta_2 e^{-\theta_4 t_i} + \theta_3 e^{-\theta_5 t}$$

   oder Frequenzschätzungen

   $$y_i = \theta_1 + \theta_2 \sin(\theta_3 t_i + \theta_4) + \theta_5 \sin(\theta_6 t_i + \theta_7)$$

2. Approximationsprobleme für Kennlinien der Art

   $$y_i = \sum_{j=1}^{n} \theta_j a_j(\theta_{n+1}, \ldots, \theta_{n+p}, u_i),$$

   mit gebrochenrationalen Funktionen, trigonometrischen Funktionen, Exponential- oder Gauß-Funktionen

3. Identifikation bilinearer zeitdiskreter Modelle [613]

4. Spline-Approximationen mit freien Knoten [559]

5. Parametrisierung implizit gegebener Mannigfaltigkeiten; Konturprobleme
   Falls die Mannigfaltigkeiten Lösungen partieller Differenzialgleichungen sind, die mittels finiter Differenzenapproximationen berechnet wurden, ergeben sich schwach besetzte, hochdimensionale Fitting-Probleme mit gerasterten Punkten [242].

6. Umformung eines LS-Problems mit variabler Metrik $||Ax - b||^2_{W(z)} \stackrel{!}{=} \text{Min}$ in ein äquivalentes Problem $||W^{1/2}(z)Ax - W^{1/2}(z)b||^2_2 \stackrel{!}{=} \text{Min}; x \in \mathbb{R}^n, z \in \mathbb{R}^p$, [71]

Zur Lösung separabler LS-Probleme bieten sich folgende Zugänge an:

1. Standard-SQP-Löser (empfehlenswert, wenn durch Elimination die Variablenzahl nicht signifikant reduziert wird)

2. ALS-Algorithmus (empfehlenswert, wenn das Problem für x und z bei fester anderer Variablen explizit lösbar ist; s. Abschnitt 8.2.3)

3. Variablen-Projektionsmethode (populäre Methode mit signifikanter Beschleunigung der Optimierung in bestimmten Fällen).

Im Folgenden wird die Variablen-Projektionsmethode näher vorgestellt. Sie nutzt aus, dass, wenn x in (5.30) linear in das Residuum eingeht, x durch eine Projektion eliminiert werden kann. Das neue, dann nur von z abhängige Problem wird für z optimiert. Aus der Optimallösung z_{opt} wird anschließend die Optimallösung x_{opt} berechnet. Die Rechtfertigung für dieses Vorgehen liefert der nachfolgende Satz.

Satz 5.1 (Variablen-Projektionsmethode, [231])
Unter der Annahme, dass $\mathcal{Z} \subset \mathbb{R}^p$ eine offene Menge ist, auf der $A(z)$ den konstanten Rang $r \leq \min\{m, n\}$ hat, gilt:

- Ist z_{opt} ein stationärer Punkt bzw. ein globaler Minimierer von

$$||(I_m - A(z)A^+(z))b||^2_2 \stackrel{!}{=} \text{Min} \qquad z \in \mathcal{Z} \subset \mathbb{R}^p, \tag{5.31}$$

  dann ist das Paar $(z_{\text{opt}}, x_{\text{opt}})$ mit

$$x_{\text{opt}} = A^+(z_{\text{opt}})b \tag{5.32}$$

  stationärer Punkt bzw. ein globaler Minimierer von (5.30) in $\mathcal{Z} \times \mathbb{R}^n$.

- Ist $(x_{\text{opt}}, z_{\text{opt}})$ globaler Minimierer von (5.30) in $\mathcal{Z} \times \mathbb{R}^n$, dann ist z_{opt} globaler Minimierer von $\mathcal{Z}$ (5.31). Ist x_{opt} eindeutig, dann hat x_{opt} die Form (5.32).

Der Nutzen des Satzes für die numerische Umsetzung und andere Aspekte werden in der folgenden Anmerkung diskutiert.

Anmerkung 5.5

- Statt der Moore-Penrose-Inverse kann jede $(1,3)$-g-Inverse $A^{(1,3)}(z)$ im orthogonalen Projektor $P_{\mathcal{N}(A(z))} = (I_m - A(z)A^+(z))$ verwendet werden, wenngleich $A^+(z)$ die bevorzugte Wahl ist.

- Aus gegebenen Startwerten z lassen sich optimale Startwerte x ermitteln, die zusammen als Startwerte für Standard-SQP-Löser dienen.

- Die Reduktion des Suchraums wird i. Allg. durch eine Erhöhung des Nichtlinearitätsgrads des Problems erkauft, was problematisch sein kann.

- Der Gewinn bei Gradientenverfahren fällt signifikant aus, wenn die Anzahl der linear eingehenden Parameter in Näherung der Anzahl der nichtlinear eingehenden entspricht oder gar deutlich größer ist. Unter günstigen Konstellationen ist eine Aufwandsreduzierung um den Faktor 1000 möglich [559].

- Viele Optimierungsprogramme sind so gestaltet, dass im Programmcode nur die Funktionswertberechnung und/oder die Gradientenberechnung zu ersetzen sind. Der Gradient für (5.31) lautet dabei [231]

$$\frac{\partial}{\partial z}||(I_m - A(z)A^+(z))b||_2^2 = -2(I_n \otimes (A^+(z)b)^T)\frac{\partial A^T(z)}{\partial z}(I_m - A(z)A^+(z))b, \quad (5.33)$$

der für (5.30)

$$\frac{\partial}{\partial x}||A(z)x - b||_2^2 = 2A^T(z)(A(z)x - b) \quad (5.34)$$

$$\frac{\partial}{\partial z}||A(z)x - b||_2^2 = 2(I_n \otimes x^T)\frac{\partial A^T(z)}{\partial z}(A(z)x - b). \quad /^8 \quad (5.35)$$

Die Berechnung der Moore-Penrose-Inverse in (5.33) bestimmt den numerischen Aufwand entscheidend. Sie wird bei der analytischen Gradientenberechnung nur einmal benötigt. Bei der numerischen Gradientenberechnung mit finiten Differenzen erster Ordnung fällt sie $(p+1)$-mal an, nämlich einmal für den Gütewert und p-mal für die Gütewerte der verstimmten Parameter. Eine genauere Betrachtung zeigt, dass in (5.33) nur $A^+(z)b$ auftritt, was gerade die Lösung y des LS-Problems $||A(z)y - b||_2 \stackrel{!}{=} \text{Min}$ ist, die sich numerisch effektiv berechnen lässt. Das Problem vereinfacht sich weiter, wenn

[8] Äquivalente Darstellung: $(I_n \otimes x^T)\frac{\partial A^T(z)}{\partial z} = \left[\begin{array}{ccc} \frac{\partial A(z)}{\partial z_1}x & \dots & \frac{\partial A(z)}{\partial z_k}x \end{array}\right]^T$.

z nur linear in $A(z)$ auftritt, da dann die Matrix $\dfrac{\partial A^T(z)}{\partial z}$ konstant ist. Für die Verwendung numerischer Gradienten spricht der geringe Programmieraufwand; bei einigen Programmen ist die numerische Gradientenberechnung lediglich über eine Variable zu aktivieren.

- Da eine explizite Angabe von $A^+(z)$, von Trivialfällen abgesehen, scheitert, eignet sich diese Methode wenig für geschlossene Lösungen. Bei der numerischen Berechnung braucht $A^+(z)$ oder besser $A^+(z)b$ hingegen nur für in jedem Schritt feste Zahlen z_k berechnet werden, was keine Schwierigkeiten macht.

- Das Vorgehen ist auch für separable Restriktionen anwendbar

$$C(z)x = d(z) \Rightarrow x = C^+(z)d(z) + (I_n - C^+(z)C(z))q \quad q \in \mathbb{R}^n.$$

  Nach der Elimination von x ergibt sich ein separables NLS-Problem, in dem q separabel zu z und linear zum Residuum ist.

- Wird (5.30) zusätzlich mit Ungleichheitsrestriktionen $Cz \geq d$ gestellt, so ist (5.31) mit denselben Restriktionen äquivalent bezüglich z und $Q_{\min}$, was Kaufman in [341] zeigt. In [342] wird das Problem mit mehreren rechten Seiten behandelt.

5.2.9 Kurven- und Flächenparametrisierung

Die parametrische Darstellung von Kurven und Flächen findet ihre Anwendung bei der orthogonalen Distanzregression, für die gilt:

$$\sum_{i=1}^{N} ||\Delta x_i||_2^2 \stackrel{!}{=} \text{Min} \qquad f(x_i + \Delta x_i; \theta) = 0, \quad x_i, \Delta x_i \in \mathbb{R}^n; i = 1, \ldots, N; \theta \in \mathbb{R}^p. \tag{5.36}$$

Algebraisch bedeutet das, dass N Punkte x_i so durch Δx_i zu korrigieren sind, dass die korrigierten Punkte alle ein und derselben durch θ parametrisierten impliziten Gleichung genügen. Geometrisch sind die $||\Delta x_i||_2$ die orthogonalen Abstände zur Kurve $f(x; \theta) = 0$. Ist die Zahl N der Punkte, für die der Ausgleich erfolgen soll, groß, so ist im nichtlinearen Optimierungsproblem auch die Anzahl zu behandelnder N Restriktionen groß. Durch eine parametrische Darstellung $x(t; \theta)$ der Kurven bzw. Flächen können alle Restriktionen beseitigt werden. Es ergibt sich dann das äquivalente $N + p$ bzw. $(2N + p)$-parametrische Problem [623]

$$\sum_{i=1}^{N} ||x_i - x(t_i; \theta)||_2^2 \stackrel{!}{=} \text{Min}_{\theta, t_i} \qquad \begin{array}{l} t_i \in \mathbb{R}; i = 1, \ldots, N \text{ für Kurven} \\ t_i \in \mathbb{R}^2; i = 1, \ldots, N \text{ für Flächen} \end{array} \tag{5.37}$$

wobei die Funktion $x(.;.)$ sicherstellt, dass alle Punkte $x(t_i;\theta)$ auf der Kurve liegen. Viele gebräuchliche Kurven, die z. B. als stationäre Lösungen partieller Differenzialgleichungen auftreten oder aus Anwendungen heraus folgen, besitzen eine parametrische Darstellung. Stellvertretend sind in Tabelle 5.3 einige aufgeführt, wobei auf Einschränkungen wie $r \geq 0$ für den Radius usw. verzichtet wurde. Weitere Kurven wie Spiralen, Schraubenlinien, Evolventen, Rollkurven (Zykloide, Strophoide, Kardioide, Astroide), Pascalsche Schnecke, Lissajous-Figuren finden sich in [216], [101] oder anderen Nachschlagewerken. Die parametrischen Darstellungen sind dabei nicht eindeutig. So kann z. B. t durch $2t$ ersetzt werden. Auch können statt der hier vorwiegend aufgeführten polarparametrischen Darstellungen solche über gebrochenrationale Funktionen verwendet werden. Die Darstellungsart hat dabei Einfluss auf die Komplexität des Gütegebirges. Der Vorteil der Formen in Tabelle 5.3 ist, dass $x(t;\theta) = A(t)\theta$ linear in θ ist, was ggf. algorithmisch ausgenutzt werden kann. Um Segmente von Kurven bzw. Flächen zu beschreiben, muss der freie Parametervektor t nur eingeschränkt werden.

Die orthogonale Distanzregression kann auch zur Identifikation linearer dynamischer Systeme verwendet werden, wenn Messungen oder Schätzungen von Frequenzgangpunkten G_i an verallgemeinerten Frequenzpunkten Ω_i vorliegen [612]. Ω_i steht dabei für $j\omega_i$ bei zeitkontinuierlichen Systemen, $\exp(-j\omega_i T_A)$ bei zeitdiskreten, $\tanh(j\omega_i\tau)$ (Richardson-Bereich) für Mikrowellensysteme [535] oder $\sqrt{j\omega_i}$ für Diffusionsphänomene. Das Gütekriterium lautet dann

$$\sum_{i=1}^{N} \left\| G_i - \frac{\sum_{k=0}^{m} b_k \Omega_i^k)}{\sum_{k=0}^{n} a_k \Omega_i^k} \right\|^2 \stackrel{!}{=} \text{Min} \qquad a = ((a_k)) \in \mathbb{R}^{n+1}, b = ((b_k)) \in \mathbb{R}^{m+1}; \Omega_i \text{ frei.} \tag{5.38}$$

Eine zusätzliche Restriktion an die a_i, b_i, z. B. $a_i = 1$ oder $||(a_0, \ldots, a_n)||_2 = 1$, wird benötigt, um Identifizierbarkeit zu garantieren, vgl. Abschn. 4.3. Der Unterschied zum Ausgangsfehlerzugang mit komplexwertigem Ausgang besteht darin, dass hier die N Frequenzen Ω_i so variiert werden können, dass stets der kürzeste Abstand des Messpunkts zur Ortskurve in die Zielfunktion eingeht; beim Ausgangsfehlerzugang sind die Ω_i fest. Letzterer wird bei fehlerfreier Frequenzeinstellung bevorzugt.

Kurve/Fläche	parametrische Beschreibung
Linie in $\mathbb{R}^n$	$x(t) = a + tb; b \neq 0_n, t \in \mathbb{R}$ a Stützvektor, b Richtungsvektor
k-dim. Ebene	$x(t) = a + Qt; Q \in \mathbb{R}_k^{n \times k}, t \in \mathbb{R}^k$
Kreis	$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cos t \\ 0 & 1 & \sin t \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ r \end{bmatrix}$
Ellipse	$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cos t & 0 \\ 0 & 1 & 0 & \sin t \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ r \\ s \end{bmatrix}$
Hyperbel	$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cosh t & 0 \\ 0 & 1 & 0 & \sinh t \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ r \\ s \end{bmatrix}$
Parabel	$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 & t & 0 \\ 0 & 1 & 0 & t^2 \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ 1 \\ s \end{bmatrix}$
Kugel	$\begin{bmatrix} x(\phi,\psi) \\ y(\phi,\psi) \\ z(\phi,\psi) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \sin\psi\cos\phi \\ 0 & 1 & 0 & \sin\psi\sin\phi \\ 0 & 0 & 1 & \cos\psi \end{bmatrix} \begin{bmatrix} [x_0, y_0, z_0]^T \\ r \end{bmatrix}$
achsparalleler Ellipsoid	$\begin{bmatrix} x(\phi,\psi) \\ y(\phi,\psi) \\ z(\phi,\psi) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \sin\psi\cos\phi & 0 & 0 \\ 0 & 1 & 0 & 0 & \sin\psi\sin\phi & 0 \\ 0 & 0 & 1 & 0 & 0 & \cos\psi \end{bmatrix} \begin{bmatrix} [x_0, y_0, z_0]^T \\ [p, q, r]^T \end{bmatrix}$
achsparalleler Kreiszylinder	$\begin{bmatrix} x(\phi,t) \\ y(\phi,t) \\ z(\phi,t) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \sin\phi & 0 \\ 0 & 1 & 0 & \cos\phi & 0 \\ 0 & 0 & 1 & 0 & t \end{bmatrix} \begin{bmatrix} [x_0, y_0, z_0]^T \\ [r, 1]^T \end{bmatrix}$
achsparalleler Kreiskegel	$\begin{bmatrix} x(\phi,\psi) \\ y(\phi,\psi) \\ z(\phi,\psi) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \sin\phi \\ 0 & 1 & 0 & \cos\phi \\ 0 & 0 & 1 & \tan\psi \end{bmatrix} \begin{bmatrix} (x_0, y_0, z_0)^T \\ r \end{bmatrix}$
achsparalleler Torus	$\begin{bmatrix} x(\phi,\psi) \\ y(\phi,\psi) \\ z(\phi,\psi) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 & \cos\psi\cos\phi \\ 0 & 1 & 0 & 1 & \cos\psi\sin\phi \\ 0 & 0 & 1 & 0 & \sin\psi \end{bmatrix} \begin{bmatrix} [x_0, y_0, z_0]^T \\ [r, \varrho]^T \end{bmatrix}$

Tabelle 5.3: Parametrische Darstellungen spezieller Kurven und Flächen [623], [11]

5.3 Elimination mittels scharfer Ungleichungen

In diesem Abschnitt wird eine Technik vorgestellt, die für MinMax-, MaxMin-, MinMin- und MaxMax-Probleme mit separablen zulässigen Mengen die Eigenschaften scharfer Ungleichungen[9] zur Elimination nutzt. Dabei werden nicht nur Restriktionen eliminiert, sondern es wird auch die innere Optimierung selbst gelöst. Den Ausgangspunkt bildet

$$f(x,y) \leq g(x) \text{ sei } \forall x \in \mathcal{X} \text{ eine scharfe Ungleichung} \quad \Rightarrow \quad \max_{y \in \mathcal{Y}} f(x,y) = g(x). \tag{5.39}$$

Eine Anwendung der Technik soll am Beispiel der robusten LS (RLS) gezeigt werden, s. Formulierung (5.40). Die RLS liefert dabei die LS-Lösung[10] unter den ungünstigsten zulässigen Störungen. Der Zusatz „robust“ erklärt sich außer an der Formulierung auch durch eine Senkung der Empfindlichkeit der LS-Lösung. Plausibel wird das, durch den in (5.41) auftretenden Term $r_A||x||_2$, der als Regularisierungsterm aufgefasst werden kann und der einen Kompromiss zwischen Fehlerausgleich und Länge des Parametervektors erzeugt. An dieser Stelle soll aber nicht die RLS näher betrachtet werden, sondern vielmehr deren Umformulierung in eine vereinfachte Problemstellung beispielhaft gezeigt werden.

Satz 5.2 (Alternative Formulierung der Robusten LS, [123])
Das robuste LS-Problem

$$\max_{\substack{||\Delta A||_F \leq r_A \\ ||\Delta b||_2 \leq r_b}} ||(A+\Delta A)x - (b+\Delta b)||_2 \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n; A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, m \geq n \tag{5.40}$$

mit den Fehlerschranken r_A und r_b ist äquivalent zu

$$||Ax-b||_2 + r_A||x||_2 + r_b \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n; A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, m \geq n. \tag{5.41}$$

Beispiel 5.23 (Umformulierung der Robusten LS)
Mit der Dreiecksungleichung (scharfe Ungleichung) und der Normverträglichkeit folgt

$$\begin{aligned}\max_{\substack{||\Delta A||_F \leq r_A \\ ||\Delta b||_2 \leq r_b}} ||(A+\Delta A)x - (b+\Delta b)||_2 &\leq \max_{\substack{||\Delta A||_F \leq r_A \\ ||\Delta b||_2 \leq r_b}} \{||Ax-b||_2 + ||\Delta A||_F \cdot ||x||_2 + ||\Delta b||_2\} \\ &= ||Ax-b||_2 + r_A||x||_2 + r_b.\end{aligned}$$

Es bleibt zu zeigen, für welche Werte die obere Schranke angenommen wird. Dies sind

$$\begin{aligned}\Delta A_{\text{opt}} &= \frac{(Ax-b)}{||Ax-b||_2} \cdot \frac{x^T}{||x||_2} r_A \\ \Delta b_{\text{opt}} &= -\frac{Ax-b}{||Ax-b||_2} r_b,\end{aligned}$$

[9] Eine scharfe Ungleichung ist eine Ungleichung, die durch keine andere Ungleichung verbessert werden kann. So ist $0 \leq x^2$ scharf, während $-1 \leq x^2$ nicht scharf ist. Auch $0 < \mathrm{e}^x$ ist scharf, denn weder $\exists \varepsilon > 0 : \varepsilon < \mathrm{e}^x$ noch $0 \leq \mathrm{e}^x$ sind mögliche Verbesserungen.

[10] Die Einordnung zu den LS-Problemen statt zu den TLS-Problemen, wie die Struktur vermuten ließe, rührt auch daher, dass für $r_A = 0$ und $r_b = 0$ ein Standard-LS-Problem entsteht.

da wegen der Kollinearität von $Ax - b$, $\Delta A_{\text{opt}} x$ und $-\Delta b_{\text{opt}}$ gilt:

$$\begin{aligned} ||(A + \Delta A_{\text{opt}})x - (b + \Delta b_{\text{opt}})||_2 &= ||(Ax - b) + \Delta A_{\text{opt}} x - \Delta b_{\text{opt}}||_2 \\ &= ||Ax - b||_2 + ||\Delta A_{\text{opt}} x||_2 + ||\Delta b_{\text{opt}}||_2 \\ &= ||Ax - b||_2 + r_A||x||_2 + r_b. \end{aligned}$$

Die in Umformungen ebenfalls gern verwendete Beziehung

$$\max_{||y||_2 \leq r} y^T f(x) = r||f(x)||_2 \qquad x, y \in \mathbb{R}^n \tag{5.42}$$

beruht auf der Cauchy-Schwarz-Ungleichung – einer scharfen Ungleichung –

$$|\langle u, v\rangle| \leq ||u||_{\langle .,.\rangle}||v||_{\langle .,.\rangle}; \qquad u, v \in \mathcal{H}(\mathbb{R}); \qquad (=) \text{ gdw. } u = \gamma v \vee u = 0_{\mathcal{H}} \vee v = 0_{\mathcal{H}}, \tag{5.43}$$

mit dem Spezialfall $\langle u, v\rangle =: y^T f(x)$, $||u||_{\langle .,.\rangle} =: ||f(x)||_2$ und $||v||_{\langle .,.\rangle} =: ||y||_2$. Die Nützlichkeit von (5.42) belegt eine alternative Herleitung der RLS-Reformulierung [624]. In [624] wird die RLS zudem dahingehend erweitert, dass statt des Paares „euklidische Norm und Frobenius-Norm“ entsprechende Paare „separable Matrixnorm, passfähige Vektornorm“[11] („City-Block-Norm, Spaltensummennorm“, „Chebyshev-Norm, Zeilensummennorm“, „euklidische Norm, Spektralnorm“, „Hölder-Matrixnorm, Hölder-Norm“) verwendet werden. Derartige Probleme werden Robuste-kleinste-Norm-Probleme (RLN) genannt und u. a. zur Datenanpassung eingesetzt. Deren Herleitung und Umformung erfolgt analog zum RLS-Problem ebenso wie die eines artverwandten MinMin-Problems [124].

5.4 Penalty-Verfahren

Eine gern genutzte Technik, um Restriktionen aus dem Optimierungsproblem zu entfernen, stellt das Penalty-Verfahren dar. Für die Restriktionen werden dabei Strafterme formuliert, die zur eigentlichen Zielfunktion hinzuaddiert werden. Durch ein permanentes Vergrößern der Bestrafung wird ein immer besseres Einhalten der Restriktionen erreicht. Das entstehende freie Problem ermöglicht somit die Anwendung klassischer rekursiver und adaptiver Verfahren der Identifikation bzw. Regelung für restringierte Probleme.

[11] Eine Matrixnorm $||.||_\rho$ in $\mathbb{R}^{m\times n}$ heißt separabel, wenn es für Vektoren $x \in \mathbb{R}^m$ und $y \in \mathbb{R}^n$ zwei Vektornormen $||.||_\alpha$ und $||.||_\beta$ gibt, sodass

$$\forall x \in \mathbb{R}^m, y \in \mathbb{R}^n : \quad ||xy^T||_\rho = ||y||_\alpha ||x||_\beta^D, \quad ||xy^T||_\rho^D = ||x||_\alpha^D ||y||_\beta.$$

Die Norm $||X||_\rho^D \stackrel{\text{def}}{=} \max_{||Y||_\rho \leq 1} \text{spur}(Y^T X)$ in $\mathbb{R}^{m\times n}$ heißt duale Norm zu $||.||_\rho$.

Definition 5.4 (Penalty-Verfahren, [71])
Das Penalty-Verfahren (Strafverfahren) transformiert ein restringiertes, lösbares Problem

$$f(x) \stackrel{!}{=} \text{Min}; x \in \mathcal{F} = \{x \in \mathbb{R}^n : g_i(x) \leq 0; i = 1, \ldots, p; h_j(x) = 0; j = 1, \ldots, m\} \tag{5.44}$$

mit stetigem f durch Hinzunahme eines stetigen Strafterms $l(x) > 0$ für $x \notin \mathcal{F}$ und $l(x) = 0$ für $x \in \mathcal{F}$ in eine Folge freier Probleme

$$x^{(k)} = \min_{x \in \mathbb{R}^n} \underbrace{f(x) + \gamma^{(k)} l(x)}_{p(x;\gamma^{(k)})} \qquad \gamma^{(0)} > 0 \text{ und } \gamma^{(k+1)} > \gamma^{(k)}, \tag{5.45}$$

deren Lösung für $\gamma^{(k)} \to \infty$ gegen $x_{\text{opt}} = \text{argmin}_{x \in \mathcal{F}} f(x)$ strebt. $\gamma^{(k)} > 0$ heißt dabei Penalty-Parameter, $p(x; \gamma^{(k)})$ Penalty-Funktion. Es ist gleichfalls üblich, nicht $p(x; \gamma^{(k)})$, sondern $l(x)$ als Penalty-Funktion zu bezeichnen.

Anmerkung 5.6 Üblicherweise wird mit einem großen $\gamma^{(0)}$ gestartet. Beim Vergrößern der $\gamma^{(k)}$ wird mindestens mit einem Verdoppeln gearbeitet. Der Iterationsabbruch erfolgt, wenn die Restriktion eingehalten wird und sich die Lösung nur marginal ändert.

Anmerkung 5.7 Eine sehr gebräuchliche Straffunktion ist

$$l(x) = \sum_{i=1}^{p} \underbrace{(\max\{0, g_i(x)\})^{\alpha}}_{\text{für } g_i(x) \leq 0} + \sum_{j=1}^{m} \underbrace{|h_j(x)|^{\alpha}}_{\text{für } h_j(x) = 0} \qquad \text{mit } \alpha > 0. \tag{5.46}$$

Sie ist für $\alpha = 1$ nicht differenzierbar, wohl aber für $\alpha = 2$.

Das Penalty-Verfahren wird durch den nachfolgenden Satz von Pietrzykowski gerechtfertigt.

Satz 5.3 (Konvergenz des Penalty-Verfahrens, [320])
Es sei $f : \mathbb{R}^n \to \mathbb{R}$ eine stetige Funktion und x_{loc} ein strenger lokaler Minimierer von (5.44) und $l : \mathbb{R}^n \to \mathbb{R}_{\geq}$ eine stetige Straffunktion. Dann gibt es ein $\gamma^{(0)} > 0$, sodass für jedes $\gamma \geq \gamma^{(0)}$ die Funktion $f(x) + \gamma l(x)$ einen lokalen Minimierer $x(\gamma)$ hat, der für $\gamma \to \infty$ gegen x_{loc} konvergiert.

Ein spezielles Penalty-Verfahren ist das sog. WLSE-Grenzwertverfahren (weighted least squares with equality constraints), das sich für große, schwachbesetzte LSE-Probleme und Online-Anwendungen eignet. Es nutzt für lineare Gleichungsrestriktionen einen quadratischen Penalty-Term. Umgeschrieben stellt sich das Problem als eine um die Gleichungsrestriktionen erweiterte LS dar, in der die erweiterten Gleichungen eine sehr hohe Zeilenwichtung erfahren:

$$x_{LSE} = \text{argmin}\{||Ax - b||_2^2 : Cx = d; C \in \mathbb{R}^{p \times n}, A \in \mathbb{R}^{m \times n}\} \tag{5.47a}$$

$$= \lim_{\gamma \to \infty} \text{argmin} \left\| \begin{bmatrix} \gamma C \\ A \end{bmatrix} x - \begin{bmatrix} \gamma d \\ b \end{bmatrix} \right\|_2^2 \tag{5.47b}$$

$$= \lim_{\gamma \to \infty} \text{argmin} \left\| \begin{bmatrix} C \\ A \end{bmatrix} x - \begin{bmatrix} d \\ b \end{bmatrix} \right\|_{\gamma^2 I_p \oplus I_m}^2. \tag{5.47c}$$

Die Approximationsgenauigkeit der Lösung, sprich der Abstand zum Originalproblem, lässt sich durch die Wichtungsfaktoren steuern. Erfahrungsgemäß reichen in vielen praktischen Problemen wenige Nachkommastellen, da die Ungenauigkeit der Lösung ohnehin meist mehr durch die Störungen und Modellvernachlässigungen bestimmt wird.

Anmerkung 5.8 Die Idee der quasi unendlich großen Wichtung kann auch auf andere Problemstellungen übertragen werden. So verhindert eine sehr große Spaltenwichtung in TLS-Problemen, dass eine Korrektur in festen Spalten, d. h. in für den Fehlerausgleich nicht zur Verfügung stehenden Spalten, vorgenommen wird. Beim Geradenmodell $\theta_1 u(:) + \theta_2 1_N \cong y(:)$ mit gestörtem $u(:), y(:)$-Vektoren betrifft das die Einsspalte.

Dem Vorteil des Penalty-Verfahrens, ein freies Problem zu formulieren, steht als Nachteil $\gamma \to \infty$ gegenüber, was numerische Probleme infolge der Konditionsverschlechterung bewirkt, siehe hierzu das nachfolgende Beispiel.

Beispiel 5.24 (Konditionierungsproblem beim Penalty-Verfahren)
Für $x_1^2 + x_2^2 \stackrel{!}{=} \text{Min}; x_2 = 1$ wird die Penalty-Funktion $p(x, \gamma) = x_1^2 + x_2^2 + \gamma(x_2 - 1)^2$ minimiert. Deren Lösung $x_{\text{opt}}(\gamma) = (0, \frac{\gamma}{\gamma+1})^T$ strebt gegen die Optimallösung $x = [0, 1]^T$. Da die Eigenwerte $\lambda_1 = 2, \lambda_2 = 2(1+\gamma)$ der Hesse-Matrix von $p(x, \gamma)$ dabei auseinander driften, verschlechtert sich die Konditionierung des Problems zunehmend.

Obwohl die Konditionsverschlechterung den Penalty-Verfahren eigen ist, können die damit verbundenen Auswirkungen durch eine geschickte Formulierung abgemildert werden, wie die Diskussionen im Beispiel 5.25 zeigt.

Beispiel 5.25 (Wichtungsmethode für das LSE-Problem)
Zur numerischen Lösung von (5.47b) für festes, aber großes γ werden gewöhnlich QR-Algorithmen mit Spaltenpivotisierungen benutzt. Allerdings reicht dies im Allgemeinen nicht, sodass formal auch Zeilenpivotisierungen vorzunehmen sind, da auch die Zeilennormen stark differieren. In [407] wird gezeigt, dass die äquivalenten Probleme

$$\left\| \begin{bmatrix} \gamma C \\ A \end{bmatrix} x - \begin{bmatrix} \gamma d \\ b \end{bmatrix} \right\|_2 \stackrel{!}{=} \text{Min} \qquad \gamma \gg 1 \tag{5.48a}$$

$$\left\| \begin{bmatrix} A \\ \gamma C \end{bmatrix} x - \begin{bmatrix} b \\ \gamma d \end{bmatrix} \right\|_2 \stackrel{!}{=} \text{Min} \qquad \gamma \gg 1 \tag{5.48b}$$

in völlig unterschiedlicher numerischer Genauigkeit gelöst werden. Die C-über-A-Variante (5.48a) ist hierbei numerisch zu bevorzugen [79], [145]. Zudem wird $\gamma \approx \epsilon_{mach}^{-1/2} \ldots \epsilon_{mach}^{-3/2}$ [408], [146], [19] empfohlen, wobei die ϵ_{mach} die Zahlenauflösung im Computerprogramm ist.

Eine Konsequenz aus den beiden Beispielen zur Konditionsproblematik ist, der Wunsch nach einem endlichen γ, für das lokale Minimierer des Penalty-Problems und des restringierten übereinstimmen. Das führt auf die folgende Definition.

Definition 5.5 (Exakte Penalty-Funktion)
Eine Penalty-Funktion heißt exakt, wenn für alle $\gamma \geq \gamma_{\min}$ die lokalen Minimierer der Penalty-Funktion mit denen von (5.44) gleich sind.

Ist $\gamma_{\min}$ a priori bekannt oder kann es einfach berechnet werden, braucht nur eine einzige Optimierung mit einem $\gamma \geq \gamma_{\min}$ ausgeführt werden. Im Fall eines unbekannten $\gamma_{\min}$ wird die Iteration abgebrochen, wenn sich $x^{(k)}$ nicht mehr ändert, was ab $\gamma^{(k)} \geq \gamma_{\min}$ der Fall ist.

Das Problem der exakten Penalty-Funktionen ist, dass sie in der Regel nicht differenzierbar sind, da Differenzierbarkeit und Exaktheit nicht vereinbar sind [320]. Somit bleibt als Ausweg der Einsatz zugeschnittener Algorithmen, die mit der Nichtdifferenzierbarkeit klarkommen (Subgradienten- oder Schnittebenenverfahren) [213], die eine ε-Glättung der Nichtdifferenzierbarkeitsstellen vornehmen oder die die Nichtdifferenzierbarkeit mit dem Argument einer Lebesgue-Maß-Null-Menge ignorieren (ggf. Abfangen per Fallunterscheidung). Für bestimmte Problemklassen empfiehlt sich darüber hinaus der Einsatz modifizierter, differenzierbarer exakter Penalty-Verfahren auf der Basis erweiterter Lagrange-Funktionen [320], [71], allerdings geschieht dies zum Preis zusätzlicher Lagrange-Parameter. Abschließend soll eine Anwendung für ein exaktes Penalty-Verfahren vorgestellt werden.

Beispiel 5.26 (Optimalsteuerung für zeitdiskretes Problem)
Es sei $\xi[k+1] = A_z\xi[k] + b_z u[k]$ vom Zustand $\xi[0] = \xi_0$ in $\xi[N] = \xi_N$ zu überführen. Durch $\xi[N] = A_z^N\xi[0] + A_z^{N-1}b_z u[0] + \ldots + A_z b_z u[N-2] + b_z u[N-1]$ wird die Nebenbedingung beschrieben, kurz $A = [b_z, A_z b_z, \ldots, A_z^{N-1}b_z]$, $b = \xi_N - A_z^N\xi_0$, $x = [u[N-1], \ldots, u[0]]^T$ mit $Ax = b$. Das Optimierungsproblem – ein sog. Minimum-Norm-Problem – lautet dann

$$||x||_p \stackrel{!}{=} \text{Min}; Ax = b,$$

wobei x und damit $\{u[k]\}_{k=0}^{N-1}$ bzgl. Durchsatz[12] deren Summe im Falle eines Massestromes (1-Norm), Energie (2-Norm) oder Amplitude (∞-Norm) minimiert wird. Das exakte Penalty-Problem hierzu ist $||x||_p + \gamma||Ax - b||_q \stackrel{!}{=} \text{Min}$, wobei $\gamma_{\min}$ von p, q und A abhängt. Den Beweis und eine einfache Abschätzung für $r_{\min}$ mittels $\lambda_{\min}(A^T A)$ gibt [304], wo die Nichtdifferenzierbarkeit mit dem Maß-Null-Argument algorithmisch ignoriert wird.

[12] Die Bezeichnungen „Durchsatz“ und „Energie“ leiten sich aus speziellen physikalischen Stellgrößen ab und werden als Verallgemeinerungen angesehen. So kennzeichnet die Summe positiver Stellwerte $u[k]$ eines Volumen- bzw. Massestroms den Durchsatz, während die Spannungs- oder Stromquadrate über einem Widerstand zur Leistung proportional sind und deren Integral/Summe folglich eine Energie verkörpert.

Kapitel 6

Erweiterungsmethoden

Das Gegenstück zu den Reduktionsmethoden aus Kapitel 5 sind die Erweiterungsmethoden, bei denen durch Erhöhen der Anzahl der Parameter und/oder der Restriktionen Vereinfachungen des Problems erreicht werden. Auch hierbei wird wieder eine Lösungsäquivalenz im weiteren Sinn gefordert, d. h. aus den Lösungen des erweiterten Problems müssen sich die des Originalproblems näherungsfrei ableiten lassen. Obwohl eine Erhöhung der Parameteranzahl zunächst paradox erscheint, kann sie entscheidende Vorteile bringen, wenn sich dadurch die Topologie des Gütegebirges vereinfacht oder gar eine geschlossene Lösung des Problems möglich wird. Die Vorgehensweisen, mit denen die Vereinfachungen erreicht werden, sind dabei vielfältig.

Bei der Pseudodimensionalitätserhöhung in Abschnitt 6.1 werden zusätzliche Restriktionen in Kauf genommen, um etwa topologische Vorteile oder eine einfachere Funktionenklasse zu erhalten. Der Zugang ähnelt sehr parametertransformationsbasierten Zugängen, nur das hier keine Bijektion genutzt wird, sondern der neue Parameterraum größer ist.

Ebenfalls durch zusätzliche Restriktionen sollen bei der Splitting-Methode in Abschnitt 6.2 die Vereinfachungen erreicht werden. Dabei wird eine Variable, die in der Zielfunktion in einer zu hohen Potenz auftritt, in zwei Variablen aufgeteilt, die alsdann zusätzlich durch eine Gleichheitsrestriktion wieder aneinander gekoppelt werden. Das Problem wird also hinsichtlich des Potenzgrads vereinfacht, der sich durch die Umformung halbiert.

Die Schlupfvariablenmethode in Abschnitt 6.3 gehört zum Standardrepertoire der linearen Optimierung. Sie überführt Ungleichungs- und Ungleichheitsrestriktionen in Gleichungen unter Hinzunahme einer neuen Variablen, die einer sehr einfachen Restriktion genügt. Die Gleichungen können danach eventuell zur Reduktion herangezogen werden, sodass schlussendlich ein Problem mit einfacheren Restriktionen an die neuen Variablen verbleibt. Erweiterungen auf Optimierungsprobleme mit nichtlinearen Ungleichungsrestriktionen sind in direkter Weise möglich [71].

In Abschnitt 6.4 werden Variablen auf eine spezielle Weise dargestellt, um Betragsfunktionen, paarweise Minimum- und Maximumsfunktionen in der Zielfunktion und/oder den Restriktionen zu beseitigen.

Die Epigraph-Methode in Abschnitt 6.5 stellt eine in der konvexen Optimierung beliebte Vorgehensweise dar. Das Ziel dabei ist die Umformung in eine lineare Zielfunktion zum Preis einer zusätzlichen Variablen und einer zusätzlichen Ungleichungsrestriktion.

Die Methode der Lagrange-Multiplikatoren zählt zu den am häufigsten verwendeten Methoden zur Lösung restringierter Probleme. Sie ist weitgehend bekannt, sodass in Abschnitt 6.6 primär matrixvariate Funktionen betrachtet werden.

Der letzte Abschnitt dieses Kapitels widmet sich den Rangrestriktionen. Diese sind schwierig zu behandeln, da der Rang nur natürliche Zahlen annimmt, weshalb sich ableitungsbasierte Zugänge verbieten. Durch Faktorisierungen können die Rangrestriktionen aber strukturell eingehalten und damit eliminiert werden, allerdings zum Preis einer erhöhten Parameteranzahl.

6.1 Pseudodimensionalitätserhöhung

Hierbei handelt es sich um eine Erweiterungsmethode, die den Parameterraum $\Theta \subseteq \mathbb{R}^p$ in einen Parameterraum $\tilde{\Theta} \subset \mathbb{R}^{p+l}$, $l \geq 1$ transformiert, der in einen höherdimensionalen euklidischen Raum eingebettet ist. Der Raum $\tilde{\Theta}$ füllt, bedingt durch zusätzliche Restriktionen, den höherdimensionalen Raum nicht aus. Es handelt sich also nur um eine scheinbare Dimensionserhöhung. Für den Fall, dass $\Theta = \mathbb{R}^p$ und $\tilde{\Theta}$ eine Mannigfaltigkeit ist, gilt $\dim \tilde{\Theta} = p$. Als ein einfaches Beispiel sei die Riemannsche Zahlenkugel genannt, deren Sphäre als zweidimensionale Mannigfaltigkeit im $\mathbb{R}^3$ jedem Punkt der (x, y)-Ebene eindeutig einen Punkt auf der Sphäre zuordnet. Weitere Beispiele sind die Hesse-Normalform für Ebenen in $\mathbb{R}^3$ oder die implizite Geradendarstellung in $\mathbb{R}^2$, z. B. $ax + by - c = 0$, $a^2 + b^2 = 1$ genannt. Letztere ermöglicht im Gegensatz zur expliziten Geradengleichung $y = mx + n$ senkrechte Geraden wie $x = 5$ zu erfassen. Ebenso löst die Erweiterung zu Eigenpaaren (α, β) mit $\alpha^2 + \beta^2 = 1$ die Nulldivisions- bzw. Unendlichproblematik in verallgemeinerten Eigenwertproblemen. Ein Klassiker in der Photogrammetrie [556], der Bildverarbeitung, Robotik [305] und anderen Wissenschaften ist die Quaternionendarstellung für Drehungen im $\mathbb{R}^3$.

Beispiel 6.1 (Quaternionen zur Rotationsmatrixdarstellung)
Bei 3D-Lokalisierungsproblemen sind ein Positionsvektor x und die Verdrehung um den Kurswinkel χ, Neigungswinkel ψ und Rollwinkel ϕ zu bestimmen. Die Verdrehung lässt sich

dann als Matrix formulieren

$$A = \begin{bmatrix} \cos\chi\cos\psi & -\sin\chi\cos\phi - \cos\chi\sin\psi\sin\phi & \sin\chi\sin\phi - \cos\chi\sin\psi\cos\phi \\ \sin\chi\cos\psi & \cos\chi\cos\phi - \sin\chi\sin\psi\sin\phi & -\cos\chi\sin\phi - \sin\chi\sin\psi\cos\phi \\ \sin\psi & \cos\psi\sin\phi & \cos\psi\cos\phi \end{bmatrix}.$$

Eine direkte Optimierung über die Winkel ist aber wegen der komplizierten Struktur von A nicht angebracht. Gleichermaßen kann die Matrix über Einheitsquaternionen $q = q_0 + q_1\mathrm{i} + q_2\mathrm{j} + q_3\mathrm{k}$ parametrisiert werden zu [556]

$$A = \begin{bmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(a_1q_2 - q_0q_3) & 2(q_1q_3 + q_0q_2) \\ 2(q_1q_2 + q_0q_3) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2(q_2q_3 - q_0q_1) \\ 2(q_1q_3 - q_0q_2) & 2(q_2q_3 + q_0q_1) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{bmatrix} \qquad q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1.$$

In den neuen Parametern sind die Matrixelemente algebraische Funktionen, ja sogar nur quadratische Polynome ebenso wie die zusätzliche Restriktion. Beim partiellen Ableiten von linearen Zielfunktionen in A nach den q_i entstehen folglich lineare Terme in den q_i mit Ausnahme jener Terme λq_i, die von einem Lagrange-Ansatz für die Restriktion herrühren. Kurzum das Lokalisierungsproblem lässt sich nach einigen Umformungen auf ein Eigenwertproblem zurückführen, wobei für Details auf [252] verwiesen sei.

6.2 Splitting-Methode

Die Spitting-Methode nutzt die Idee, eine in einem Problem mehrfach auftretende unabhängige Variable in mehrere unabhängige Variable aufzuteilen, diese aber gleichzeitig wieder über Gleichheitsrestriktionen zu koppeln. Sie ist somit eine Erweiterungsmethode, da die Zahl der Variablen und Restriktionen zunimmt. Mit der Splitting-Methode lassen sich beispielsweise polynomiale LS-Probleme in multilineare LS-Probleme umformen. Anwendung findet sie besonders in der Datenanalyse [349], in der unter anderem die simultane Hauptkomponentenanalyse zum Erkennen von Zusammenhängen zwischen Messdaten dient.

Beispiel 6.2 (Splitting bei der simultanen Hauptkomponentenanalyse)
Aus Optimierungssicht erfordert die simultane Hauptkomponentenanalyse die Lösung von

$$\sum_{k=1}^{p} ||A_k - X^T B_k X||_F^2 \stackrel{!}{=} \text{Min} \qquad X \in \mathbb{R}_m^{m\times n}; A_k \in \mathcal{S}_n, B_k \in \mathcal{S}_m, m \le n, \tag{6.1}$$

was äquivalent ist zu

$$\sum_{k=1}^{p} ||A_k - X^T B_k Y||_F^2 \stackrel{!}{=} \text{Min} \qquad X, Y \in \mathbb{R}_m^{m\times n} : X = Y; A_k \in \mathcal{S}_n, B_k \in \mathcal{S}_m, m \le n. \tag{6.2}$$

Die Zahl der Restriktionen hat sich um $m \cdot n$ erhöht, dafür ist aber das Problem quadratisch in X und Y, während (6.1) ein Polynom vierten Grads in X darstellt.

Eine Möglichkeit die Restriktion algorithmisch zu sichern, besteht in der Verwendung eines Penalty-Terms $\frac{\alpha}{2}||Y-X||_F^2$; $\alpha \gg 0$, der Abweichungen von X zu Y bestraft. Für den ALS-Algorithmus (s. Abschn. 8.2.3: Im Wechsel wird eine LS für X bei fixem Y und eine für Y bei fixem X gelöst.), ergibt sich dann[1]

$$X_{j+1} = \Big(\sum_{k=1}^{p} B_k^T Y_j Y_j^T B_k + \alpha I_m\Big)^{-1}\Big(\sum_{k=1}^{p} B_k^T Y_j A_k + \alpha Y_j\Big) \qquad X_0, Y_0 \neq 0_{m\times n} \tag{6.3a}$$

$$Y_{j+1} = \Big(\sum_{k=1}^{p} B_k^T X_{j+1} X_{j+1}^T B_k + \alpha I_m\Big)^{-1}\Big(\sum_{k=1}^{p} B_k^T X_{j+1} A_k + \alpha X_{j+1}\Big). \tag{6.3b}$$

Die Konvergenz kann zwar unter Umständen langsam sein; dafür geht die Herleitung und programmtechnische Implementierung schnell.

Eine Erweiterung erfährt die Splitting-Methode, wenn nicht X, sondern eine Funktion von X durch Y ersetzt wird, beispielsweise $X^{-1} =: Y$. Aus einem freien Problem wird über

$$\text{spur}(AX-B)^T(A-BX^{-1}) \stackrel{!}{=} \text{Min} \qquad X \in \mathbb{R}_n^{n\times n} \tag{6.4a}$$

$$\Leftrightarrow \qquad \text{spur}(AX-B)^T(A-BY) \stackrel{!}{=} \text{Min} \qquad X, Y \in \mathbb{R}_n^{n\times n} : XY = I_n \tag{6.4b}$$

ein restringiertes, das über einen Penalty-Term per ALS-Algorithmus gelöst werden kann.

Als eine Anwendung für (6.4a) sei das symmetrische, positiv definite Schätzproblem aus der Robotik (Massenmatrix, Steifheitsmatrix) [128] genannt, bei dem zusätzlich $X \in \mathcal{S}_n^>$ gefordert wird. Gegenüber einer Formulierung als symmetrisches, positiv definites Prokrustes-Problem, in das nur die Fehlermatrix $AX - B$ für die Prädiktion von B aus den Daten A eingeht, berücksichtigt (6.4a) auch den Fehler $A - BX^{-1}$ bei Umkehrung (Prädiktion von A aus Daten B). Das Kriterium ist somit bei Fehlern in beiden Datenmatrizen besser geeignet. In [128] zeigt sich ein weiterer Vorteil der zusätzlichen Restriktion. Bei der Bearbeitung mittels der Lagrange-Methode für $XY = I_n$ und $X \in \mathcal{S}_n^>$ entstehen nämlich einfachere Ausdrücke, als wenn (6.4a) mit der Definitheitsforderung benutzt wird, da die Anwendung der Kettenregel auf X^{-1} entfällt.

Eine weitere Erweiterung erfährt die Splitting-Methode, wenn sie mehrfach angewandt wird. So lässt sich jede univariate Polynomungleichung in ein System quadratischer Ungleichungen überführen, wie Beispiel 6.3 zeigt. Das System quadratischer Ungleichungen kann wiederum herangezogen werden, um die konvexe Hülle der Lösungen oder eine Relaxation oder LMI-Umformulierungen zu generieren.

Beispiel 6.3 (Splitting eines Polynoms in quadratische Terme)
Aus $x_1^2 x_2^4 x_3^3 \leq 7$ wird $x_5 = x_1 x_2$, $x_6 = x_2 x_3$, $x_7 = x_5 x_6$, $x_8 = x_7^2$, $x_8 x_3 \leq 7$.

[1] Statt über die angegebenen expliziten Formeln vom Typ $X = M^{-1}C$ kann X über das Gleichungssystem $MX = C$ numerisch effizienter berechnet werden (aufwendige Matrixinversion entfällt).

Anmerkung 6.1 Die Splitting-Idee muss nicht direkt in einen Algorithmus münden, sondern kann auch bei Problemumformungen helfen. So dient die Äquivalenz

$$f_1(x) - f_2(x) \stackrel{!}{=} \text{Min}; x \in \mathcal{F}_1 \cap \mathcal{F}_2 \quad \Leftrightarrow \quad f_1(x) - f_2(y) \stackrel{!}{=} \text{Min}; x = y; x \in \mathcal{F}_1, y \in \mathcal{F}_2 \quad (6.5)$$

zur der Herleitung der Fenchel-Dualität[2] [71].

6.3 Schlupfvariablenmethode

Die Schlupfvariablenmethode ist eine Erweiterungsmethode, die Ungleichungsrestriktionen durch Einführen zusätzlicher Variablen y_i, genannt Schlupfvariablen, in Gleichungsrestriktionen wandelt. Umformungen sind

$$g_i(x) \leq 0; i \in \{1, \ldots, p\} \quad \Rightarrow \quad g_i(x) + y_i = 0, y_i \geq 0 \quad (6.6a)$$

$$g_i(x) \geq 0 \quad \Rightarrow \quad -g_i(x) + y_i = 0, y_i \geq 0 \quad (6.6b)$$

$$g_i(x) \neq 0 \quad \Rightarrow \quad g_i(x) \cdot y_i = 1, y_i \in \mathbb{R} \quad (6.6c)$$

$$g_i(x) < 0 \quad \Rightarrow \quad g_i(x) + y_{i,1} = 0, y_{i,1} \geq 0, g_i(x) \cdot y_{i,2} = 1, y_{i,2} \in \mathbb{R}. \quad (6.6d)$$

Alternativ kann die strenge Ungleichung (6.6d) durch $g_i(x) \leq \varepsilon_i$ mit hinreichend kleinen, festen, frei wählbaren ε_i zu einer nichtstrengen Ungleichung gemacht werden. Somit ist dann nur eine Schlupfvariable zur Umformung der Ungleichung erforderlich. Darüber hinaus können bei strengen linearen Ungleichungen Alternativsätze (von Farkas, Carver, Gale, Stiemke, Gordan, Motzkin [159], [213]) verwendet werden, um die Probleme in nichtstrenge zu überführen. Sollte $y \geq 0$ als störende Restriktion empfunden werden, dann schafft die Substitution $y =: z^2$ Abhilfe. Zwei Beispiele zeigen die Anwendung der Methode.

Beispiel 6.4 (Umformung des monotonen nichtnegativen Kegels)
Durch $\{x \in \mathbb{R}^n : 0 \leq x_1 \leq x_2 \leq \ldots \leq x_n\}$ wird der sog. monotone nichtnegative Kegel definiert. Diese Restriktion tritt unter anderem bei der optimalen Versuchsplanung von Zeit- oder Ortsmesspunkten oder der Festlegung variabler Stützstellen (Knoten) bei der Interpolation auf. Das Anwenden der Schlupfvariablenmethode liefert hier

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 1 & 1 & \ddots & \vdots \\ \vdots & \ddots & 1 & 0 \\ 1 & \ldots & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}; \quad y_i \geq 0, \text{ kurz } x = Ay; y \in \mathbb{R}^n_{\geq}. \quad (6.7)$$

[2] Die Fenchel-Dualität besagt im eindimensionalen Fall, dass der minimale senkrechte Abstand zwischen einer konvexen und konkaven Funktion gleich dem maximalen Abstand zweier paralleler Tangenten an diese Funktionen ist. Mittels Fenchel-Dualität können also Minimierungsaufgaben in unter Umständen einfachere Maximierungsaufgaben umformuliert werden.

Wird anschließend in der Zielfunktion und den Gleichungsrestriktionen x durch Ay ersetzt, vereinfacht sich das Problem auf die einfacher zu behandelnde Orthant-Restriktion[3] $y \in \mathbb{R}^n_{\geq}$.

Beispiel 6.5 (Schlupfvariable bei nichtlinearer Restriktion)
Die Zielfunktion $f(x) = (x_1^3 + x_2^3)^{1/2}$ enthält die implizite Restriktion $x_1^3 + x_2^3 \geq 0$, welche sich mit der Schlupfvariablen x_3 in $x_3 = x_1^3 + x_2^3$ und $x_3 \geq 0$ überführen lässt.

6.4 Variablenzerlegung

In diesem Abschnitt werden Variablenzerlegungen vorgestellt, mit denen sich Beträge von Variablen aus der Zielfunktion oder den Nebenbedingungen beseitigen lassen. Auf diese Weise können auch die Schwierigkeiten beim Ableiten von Betragsfunktionen vermieden werden.

Jede reelle Zahl kann als Differenz zweier nichtnegativer Zahlen geschrieben werden, deren Summe ihr Betrag ist. Diese Darstellung lässt sich zum Beseitigen von Betragsfunktionen nutzen. Sie beruht auf folgenden Beziehungen

$$|x| = \max\{x, -x\} = \max\{0, x\} + \max\{0, -x\} \tag{6.8a}$$
$$x = \max\{0, x\} - \max\{0, -x\}. \tag{6.8b}$$

Die Idee ist es nun, x durch zwei komplementäre nichtnegative Variablen $y := \max\{0, x\}$ und $z := \max\{0, -x\}$ darzustellen, bei der die eine die positiven bzw. negativen x Werte beschreibt, während die andere Null ist. Dies führt auf

$$|x| = y + z \tag{6.9a}$$
$$x = y - z \tag{6.9b}$$
$$y, z \geq 0 \tag{6.9c}$$
$$(y \cdot z = 0 \quad \text{Komplementaritätsbedingung}). \tag{6.9d}$$

Kompliziertere Terme wie $|x_1 + 2x_2|$ lassen sich über den Zwischenschritt $w = x_1 + 2x_2$ und anschließende Substitution von $|w|$ entfernen. Die nichtlineare Komplementaritätsbedingung stört und wird deshalb weggelassen. Meist bleibt dies ohne Konsequenzen, denn wenn sowohl $|x|$ als auch x substituiert werden, genügen y_{opt} und z_{opt} ohnehin der dann redundanten Bedingung. Falls nicht, erzeugt

$$\tilde{y}_{\text{opt}} = y_{\text{opt}} - \min\{y_{\text{opt}}, z_{\text{opt}}\} \quad \text{und} \quad \tilde{z}_{\text{opt}} = z_{\text{opt}} - \min\{y_{\text{opt}}, z_{\text{opt}}\} \tag{6.10}$$

[3] Der nichtnegative Orthant in $\mathbb{R}^n$ stellt die Verallgemeinerung des ersten Quadranten (nichtnegativer Quadrant) der Ebene dar.

eine zulässige Lösung mit der Komplementaritätseigenschaft. Ein solcher Fall entsteht beispielsweise, wenn nur $|x|$ durch $y + z$ ersetzt wird, und x im Problem nicht auftritt. Mit der Ersetzung wird das Problem dann bezüglich y, z mehrdeutig, wodurch die Komplementaritätsbedingung verletzt werden kann. In diesen Fällen empfiehlt sich die Substitution $|x| =: y, y \geq 0$, bei der der Parameterraum nicht vergrößert wird.
Das nachfolgende Beispiel zeigt, wie mit der beschriebenen Technik aus einem nichtlinearen, nicht-differenzierbaren Problem ein LP-Problem wird.

Beispiel 6.6 (Beseitigen von Beträgen)

$$
\begin{aligned}
|x_1 + 2| + |x_2| &\stackrel{!}{=} \text{Min} && x_1 + x_2 \geq 3,\ -x_1 + |x_2| \geq 7 \\
|w| + |x_2| &\stackrel{!}{=} \text{Min} && x_1 + x_2 \geq 3,\ -x_1 + |x_2| \geq 7, x_1 + 2 = w \\
y_1 + z_1 + y_2 + z_2 &\stackrel{!}{=} \text{Min} && x_1 + y_2 - z_2 \geq 3,\ -x_1 + y_2 + z_2 \geq 7,\ x_1 + 2 = y_1 - z_1 \\
& && y_1, y_2, z_1, z_2 \geq 0 \\
y_1 + z_1 + y_2 + z_2 &\stackrel{!}{=} \text{Min} && y_1 - z_1 + y_2 - z_2 \geq 5, -y_1 + z_1 + y_2 + z_2 \geq 5; y_1, y_2, z_1, z_2 \geq 0
\end{aligned}
$$

Für dieses Problem ergibt sich $(y_1, z_1, y_2, z_2, w, x_1, x_2)_{\text{opt}} = (0, 0, 5, 0, 0, -2, 5)$.

In Beispiel 6.7 wird gezeigt, wie durch eine einfache Umformung auch paarweise Minimum- und Maximumfunktionen eliminiert werden können. Zum Preis zusätzlicher Variablen werden so wiederum die Ableitungsschwierigkeiten bzw. die erforderlichen Fallunterscheidungen vermieden.

Beispiel 6.7 (Beseitigen paarweiser Minimum- und Maximumfunktionen)
Dies kann durch Beseitigen der Beträge in nachstehender Darstellung erfolgen

$$f(x, y) = \min\{x, y\} = x + y - \frac{|x - y|}{2} \tag{6.11a}$$

$$f(x, y) = \max\{x, y\} = x + y + \frac{|x - y|}{2}. \tag{6.11b}$$

6.5 Epigraph-Methode

Für jedes Problem gilt bezüglich des Gütewerts und des Vektors $x \in \mathbb{R}^n$ die Äquivalenz

$$
\begin{aligned}
f(x) &\stackrel{!}{=} \text{Min} && g(x) \leq 0_p \\
& && h(x) = 0_m
\end{aligned}
\qquad \Leftrightarrow \qquad
\begin{aligned}
x_{n+1} &\stackrel{!}{=} \text{Min} && f(x) = x_{n+1} \\
& && g(x) \leq 0_p \\
& && h(x) = 0_m.
\end{aligned}
\tag{6.12}
$$

Der Gütewert und die Optimierer ändern sich nicht, wenn $f(x) = x_{n+1}$ durch $f(x) \leq x_{n+1}$ ersetzt wird. Die Ungleichung muss im Minimum nämlich immer mit Gleichheit gelten, denn

sonst könnte x_{n+1} weiter verringert werden, wäre also nicht das Minimum. Die Ungleichungsrestriktion hat gegenüber der Gleichungsrestriktion den Vorteil, dass bei konvexem f durch $f(x) \leq x_{n+1}$ eine konvexe Restriktion formuliert wird. Auf diese Weise wird die sogenannte Epigraph-Form[4] für konvexe Probleme erzeugt

$$\begin{array}{ll} x_{n+1} \stackrel{!}{=} \text{Min} & f(x) - x_{n+1} \leq 0 \\ & g(x) \leq 0_p \\ & Ax = b, \end{array} \tag{6.13}$$

die sich durch eine spezielle lineare Zielfunktion und konvexe Restriktionen auszeichnet. Über die Epigraph-Form lassen sich beispielsweise sehr elegant nicht-differenzierbare Funktionen wie die Maximum-Funktion

$$\max_{1 \leq i \leq m} f_i(x) \stackrel{!}{=} \text{Min} \quad x \in \mathcal{F} \subseteq \mathbb{R}^n \quad \Leftrightarrow \quad x_{n+1} \stackrel{!}{=} \text{Min} \quad \begin{array}{l} x \in \mathcal{F} \times \mathbb{R} \subseteq \mathbb{R}^{n+1} : \\ f_i(x_{1:n}) \leq x_{n+1}; i = 1, \ldots, m \end{array} \tag{6.14}$$

oder die Betragsfunktion beseitigen

$$|f(x)| \stackrel{!}{=} \text{Min} \qquad x \in \mathcal{F} \subseteq \mathbb{R}^n \quad \Leftrightarrow \quad x_{n+1} \stackrel{!}{=} \text{Min} \qquad \begin{array}{l} x \in \mathcal{F} \times \mathbb{R} \subseteq \mathbb{R}^{n+1} : \\ -x_{n+1} \leq f(x_{1:n}) \leq x_{n+1}. \end{array} \tag{6.15}$$

Zwei typische Anwendungen für die Epigraph-Methode sind Umformungen, die die l_1- und die l_∞-Approximation betreffen.

Beispiel 6.8 (l_1-Approximation)
Bei der l_1-Approximation wird nicht über die Summe der Quadrate, sondern über die Summe der Beträge der Residuen minimiert. Die Summe der Beträge ist geometrisch anschaulicher als die Summe der Quadrate. Unter statistischem Blickwinkel weist der resultierende Schätzer eine größere Robustheit auf, da eine große Abweichung nicht im Quadrat, sondern nur im Betrag eingeht. Die Umformung geschieht dabei wie folgt.

$$\begin{array}{lll} & ||Ax - b||_1 \stackrel{!}{=} \text{Min} & x \in \mathbb{R}^n; A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m \\ \text{äquivalent:} & \sum_{i=1}^{m} |a_i^T x - b_i| \stackrel{!}{=} \text{Min} & A = [a_1, \ldots, a_m]^T \\ \text{(6.15) anwenden:} & \sum_{i=1}^{m} y_i \stackrel{!}{=} \text{Min} & -y_i \leq a_i^T x - b_i \leq y_i; i = 1, \ldots, m \\ \text{in Ungleichungsform:} & \begin{bmatrix} 0_n \\ 1_m \end{bmatrix}^T \begin{bmatrix} x \\ y \end{bmatrix} \stackrel{!}{=} \text{Min} & \begin{bmatrix} A & -I_m \\ -A & -I_m \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \leq \begin{bmatrix} b \\ -b \end{bmatrix} \end{array}$$

[4] Der Epigraph einer Funktion ist definiert als $\text{epi} f \stackrel{\text{def}}{=} \{(x,t) : x \in \text{dom} f, f(x) \leq t\}$, kurzum als die Menge aller Punkte, die oberhalb und auf dem Graphen von f liegen.

Beispiel 6.9 (l_∞-Approximation, Chebyshev-Approximation)
Die Formulierung einer Minimax-Aufgabe über ein l_∞-Approximation wird unter anderem zur Kennlinienapproximation eingesetzt. Dadurch wird gesichert, dass die größte Abweichung in einem Bereich oder für Daten minimal ist. Das Minimum gibt dann die maximale absolute Abweichung an. Der Autor nutzte die l_∞-Approximation im Rahmen einer hochgenauen Kompensationslinearisierung für Thermoelemente. Die Umformung geschieht wie folgt.

$$\begin{aligned}
&& ||Ax-b||_\infty &\stackrel{!}{=} \text{Min} && x\in\mathbb{R}^n; A\in\mathbb{R}^{m\times n}, b\in\mathbb{R}^m \\
\text{äquivalent:} && \max_{1\le i\le m} |a_i^T x - b_i| &\stackrel{!}{=} \text{Min} && A=[a_1,\ldots,a_m]^T \\
\text{(6.14) anwenden:} && x_{n+1} &\stackrel{!}{=} \text{Min} && |a_i^T x_{1:n} - b_i| \le x_{n+1}; i=1,\ldots,m \\
\text{in Ungleichungsform:} && \begin{bmatrix} 0_n \\ 1 \end{bmatrix}^T \begin{bmatrix} x_{1:n} \\ x_{n+1} \end{bmatrix} &\stackrel{!}{=} \text{Min} && \begin{bmatrix} A & -1_m \\ -A & -1_m \end{bmatrix} \begin{bmatrix} x_{1:n} \\ x_{n+1} \end{bmatrix} \le \begin{bmatrix} b \\ -b \end{bmatrix}.
\end{aligned}$$

Ähnlich wie die l_1-Approximationen lassen sich auch die sog. Summe-von-Normen-Probleme

$$\sum_{i=1}^k c_i ||A_i x - b_i|| \stackrel{!}{=} \text{Min} \qquad x\in\mathbb{R}^n; A_i\in\mathbb{R}^{m_i\times n}, b\in\mathbb{R}^{m_i} \tag{6.16}$$

umformulieren. Sie sind äquivalent zu

$$\begin{bmatrix} c_1 & \ldots & c_k \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} \stackrel{!}{=} \text{Min} \qquad ||A_i x - b_i|| \le y_i; i=1,\ldots,k \tag{6.17}$$

Für die l_2-Norm ergibt sich also ein SOCP-Problem, für die l_1- bzw. l_∞-Norm folgen LP-Probleme. Anwendung findet die Formulierung (6.16) im sog. Weber-Problem (Fermat-Weber-Problem, Steiner-Weber-Problem, Standortproblem) [71].

Beispiel 6.10 (Weber-Problem, metrischer Mittelwert)
Eine Anwendung von (6.16) ist das Weber-Problem mit $A_i = I_n$, bei dem die c_i Kräfte an den Hebelarmen $||x-b_i||_2$ sind und x_{opt} der Drehmomentenschwerpunkt ist. Sind alle c_i gleich, dann ist x_{opt} der Punkt, für den die Summe der Distanzen zu den Punkten b_i minimal wird, kurzum der metrische Mittelwert in der l_2-Metrik. Im Gegensatz zum algebraischen Mittelwert $\bar{x} = \frac{1}{k}\sum_{i=1}^k b_i = \text{argmin} \sum_{i=1}^k ||x-b_i||_2^2$, also der LS-Lösung, existiert für den metrischen Mittelwert keine explizite Lösung. Das Gleiche gilt für das Weber-Problem. Der Weber-Punkt x_{opt} ist aber für $c_i > 0$ eindeutig bestimmt (Konvexität, Koerzivität).

Anmerkung 6.2 Sind $g_i(x) \le 0$ konvexe Restriktionen über einer konvexen Menge, dann prüft das konvexe Optimierungsproblem

$$y \stackrel{!}{=} \text{Min} \qquad x\in\mathcal{C}, g_i(x) \le y; i=1,\ldots,p$$

die Ungleichungen auf Konsistenz. Die liegt vor, wenn $y_{\text{opt}} \le 0$ ist. Im Fall $y_{\text{opt}} = 0$ gibt es singuläre Ungleichungen.

6.6 Methode der Lagrange-Multiplikatoren

Die Idee der Methode der Lagrange-Multiplikatoren besteht darin, die Restriktionen in die Zielfunktion über zusätzliche Variablen aufzunehmen und danach für die so entstehende Lagrange-Funktion die stationären Punkten (Punkte, in denen die Ableitung Null ist) zu bestimmen. Unter ihnen befinden sich unter bestimmten Voraussetzungen die lokalen Optimierer des restringierten Originalproblems. Die Methode ist in ihrer Standardform dem Ingenieur gut bekannt, weshalb hier zusätzlich auf spezielle Aspekte (Nichtdifferenzierbarkeit, Differenzierbarkeit mit Einschränkungen an die Mengen, matrixvariate Funktionen) eingegangen werden soll. Zunächst werden im Abschnitt 6.6.1 wichtige Aussagen zum freien Problem wiederholt und hinsichtlich der Abschwächung der Differenzierbarkeitsforderung kurz diskutiert. Es schließt sich in Abschnitt 6.6.2 eine Vorstellung und Bewertung ausgewählter Matrixableitungskalküle an. Diese werden bei der Anwendung der Methode der Lagrange-Multiplikatoren auf matrixvariate Funktionen benötigt, um im Ergebnis des Ableitens der Lagrange-Funktion gut handhabbare Matrixgleichungen zu erhalten.

Im Abschnitt 6.6.3 wird ein neues Matrixableitungskalkül vorgestellt, der sich zur Lösung sogenannter quasifreie Probleme (Probleme mit speziellen Restriktionen, die wie freie Probleme gelöst werden können) eignet. Dieses Kalkül erweitert ein in Abschnitt 6.6.2 vorgestelltes Kalkül auf Matrizen mit nicht algebraisch unabhängigen Elementen und umgeht Schwierigkeiten, die andere diesbezügliche Erweiterungen aufweisen. Die im Kalkül verwendete $\mathcal{L}$-Ableitung kann, wenn neben affin-strukturierte Matrizen (symmetrische Matrix, Einsdreiecksmatrix usw.) weitere Restriktionen auftreten, formal mit der Methode der Lagrange-Multiplikatoren kombiniert werden.

Nach den speziellen Aspekten für freie Probleme werden im Abschnitt 6.6.4 die Begriffe Lagrange-Multiplikator und Lagrange-Funktion eingeführt. Dann wird der Unterschied zwischen Sätzen vom Kuhn-Tucker-Typ und jenen von Fritz-John-Typ genannt [71], die beide die Verbindung zwischen den stationären Punkten der Lagrange-Funktion und den lokalen Minimierern herstellen. Aufgeführt werden indes nur drei Sätze vom Fritz-John-Typ, wobei die klassischen Differenzierbarkeitsvoraussetzungen getroffen werden. Schwächere Voraussetzungen erfordern die im Abschnitt 6.6.1 erwähnten verallgemeinerten Ableitungen und führen auf artverwandte Sätze, die sich in der Spezialliteratur finden.

Das Ausnutzen der Vorteile der Matrixableitungskalküle bei der Methode der Lagrange-Multiplikatoren geschieht über spezielle Lagrange-Terme in der Lagrange-Funktion. Da dem Autor bisher keine größere Zusammenstellung solcher Terme in der Literatur bekannt ist, wird im Abschnitt 6.6.5 eine solche Tabelle angegeben. An zwei Beispielen wird zudem deren Handhabung bei der Lösung von Modellbildungsaufgaben gezeigt.

6.6.1 Zum freien Problem

Aus der Optimierung einvariabler Funktionen ist bekannt, dass, wenn eine reelle, auf (a, b) definierte Funktion f ein lokales Minimum in $x_{\text{loc}} \in (a, b)$ hat und f eine Ableitung in x_{loc} besitzt, diese Null sein muss. Besitzt f in x_{loc} eine zweite Ableitung, so muss diese nichtnegativ sein. Diese Aussagen decken aber bei Weitem nicht alle Fälle ab, s. Anmerkung.

Anmerkung 6.3 So verschwindet für $f(x) = x^3$ bei $x = 0$ die Ableitung, allerdings hat f dort kein Minimum, da die Satzprämisse verletzt ist. $f(x) = x^4$ hat ein strenges Minimum bei $x_{\text{opt}} = 0$, obwohl die zweite Ableitung nicht positiv definit ist, da die Bedingung nur hinreichend ist. $f(x) = |x|$ hat zwar bei $x_{\text{opt}} = 0$ ein Minimum, aber der Satz greift nicht, da die Ableitung nicht existiert. Er greift auch nicht für $f(x) = 1/x$ in $(0, 10]$, denn der Minimierer bei $x_{\text{opt}} = 10$ ist kein innerer Punkt. Dennoch führen die Beispiele zu dem bekannten Fazit, dass ein lokaler Minimierer an einem stationären Punkt, einem Randpunkt oder einem Punkt auftritt, in dem f nicht differenzierbar ist.

Die Erweiterung der beiden notwendigen Bedingungen von den einvariablen Funktionen auf mehrvariable Funktionen liefert der folgende Satz.

Satz 6.1 (Optimalitätsbedingungen für das freie Problem, [487], [195])
Hat $f : \mathcal{D} \subseteq \mathcal{V} \to \mathbb{R}$ an einem inneren Punkt $x_{\text{loc}} \in \mathcal{D}$ in einem normierten Raum $\mathcal{V}$ ein lokales Minimum und ist f in x_{loc} Gâteaux-differenzierbar[5], dann gilt $Gf(x_{\text{loc}}; v) = 0$ für alle $v \in \mathcal{V}$ und speziell $\nabla f(x_{\text{loc}}) = 0_{\mathcal{V}}$ für Prä-Hilbert-Räume. Wenn f in x_{loc} zudem zweimal F-differenzierbar ist, so ist $D^2 f(x_{\text{loc}})(v, v) \geq 0$ für alle $v \in \mathcal{V}$ (notwendige Bedingungen erster und zweiter Ordnung). Umgekehrt: Ist x_{stat} ein stationärer Punkt[6], in dem f zweimal F-differenzierbar ist, und existiert ein $c > 0$ mit $D^2 f(x_{\text{stat}})(v, v) \geq c||v||^2$ bzw. speziell für $\dim \mathcal{V} < \infty$ mit $D^2 f(x_{\text{stat}})(v, v) > 0$ für alle $h \in \mathcal{V} \backslash 0_{\mathcal{V}}$, dann ist x_{stat} ein strenger lokaler Minimierer (hinreichende Bedingung).

Anmerkung 6.4 Ein gemeinsames lokales Minimum entlang aller Koordinatenachsen impliziert kein lokales Minimum von f, vgl. $f(x_1, x_2) = x_1^3 - x_1^2 x_2 + 2x_2^2$ im Punkt $(6, 9)$ (gedrehter Sattel). Weniger bekannt ist, dass aus der Existenz eines lokalen Minimums entlang aller Linienschnitte kein lokales Minimum der Funktion folgen muss, vgl. Beispiel 6.11.

---

[5] Eine G-Ableitung erfordert $x \in \text{int}\mathcal{D}$, Richtungsdifferenzierbarkeit von f entlang aller $v \in \mathcal{V}$ und Linearität der Abbildung $Gf(x_{\text{stat}}; .)$ [487], [195]. Oft wird die strengere Forderung nach Fréchet-Differenzierbarkeit gestellt, bei der nicht nur alle Richtungen, sondern alle beliebigen Kurven durch den Punkt zu betrachten sind. Mit dem einseitigen G-Differenzial lautet die notwendige Bedingung $\forall v \in \mathcal{V} : G_+ f(x_{\text{loc}}; v) \geq 0$.

[6] Ein Punkt x_{stat} heißt stationärer Punkt, wenn $\forall v \in \mathcal{V} : Gf(x_{\text{stat}}; v) = 0$. Neben den stationären Punkten, die über positiv (negativ) definite Formen $D^2 f(x_{\text{stat}})(v, v)$ als lokale Minimierer (Maximierer) charakterisiert werden, gibt es die Sattelpunkte, bei denen $D^2 f(x_{\text{stat}})(v, v)$ indefinit ist und die singulären Punkte mit einer semidefiniten Form (Hesse-Matrix singulär; mindestens ein Nulleigenwert).

Beispiel 6.11 (Vermutung zum lokalen Maximum, [195])
Die Vermutung, wonach durchweg lokale Minima/Maxima entlang aller geradlinigen Schnitte in ein und demselben Punkt ein lokales Minimum/Maximum der Funktion in diesem Punkt nach sich zieht, widerlegt die Funktion $f(x,y) = (x^2 - y)(y - 2x^2)$. Entlang aller Schnitte $y = ax$ nimmt $f(x, ax) = (x^2 - ax)(ax - 2x^2)$ in $x = 0$ ein lokales Maximum an. Dennoch ist $(0, 0)$ kein Maximum, da entlang des Schnittes $y = 1.5x^2$ aus $f(x, 1.5x^2) = \frac{1}{4}x^4$ ein Minimum vorliegt.

Anmerkung 6.5 Im Rahmen der nichtglatten Optimierung treten an die Stelle der stationären Punkte die verallgemeinerten stationären Punkte. Ist $f : \mathcal{C} \subseteq \mathcal{H}(\mathbb{R}) \to (-\infty, \infty]$ eine konvexe Funktion, dann heißt x_{stat} verallgemeinerter stationärer Punkt, wenn $0_{\mathcal{H}} \in \partial f(x_{\text{stat}})$ gilt. Hierbei steht $\partial f(x)$ für das Subdifferenzial i. S. der konvexen Analysis, auch bekannt als Fenchel-Moreau-Subdifferenzial [266][7]. Das Subdifferenzial umfasst alle Subgradienten $g \in \mathcal{H}(\mathbb{R})$ mit der Eigenschaft

$$f(z) \geq f(x) + \langle g, z - x\rangle \quad \forall z \in \mathcal{H}(\mathbb{R}). \tag{6.18}$$

Geometrisch sind das in $\mathbb{R}$ die Anstiege aller Stützgeraden eines Punkts und speziell im differenzierbaren Fall ist es der gewohnte Tangentenanstieg.
Da für den Betrag $0 \in \partial|x|_{x=0} = [-1, 1]$ gilt, ist $x_{\text{stat}} = 0$ wegen der Konvexität zudem globaler Minimierer.
Beachte: Die symmetrische Ableitung $f_s'(x) \stackrel{\text{def}}{=} \lim_{h\to 0} \frac{f(x+h)-f(x-h)}{2h}$ eignet sich nicht zur Verallgemeinerung stationärer Punkte, wie $\frac{\mathrm{d}_s}{\mathrm{d}x}|x|_{x=0} = 0$ nahelegen würde. So hat $f(x) = x - |2x|$ in $x = 0$ eine lokales Maximum, aber es gilt $f_s'(0) = 1$.

Anmerkung 6.6 Den Ableitungskalkülen für nichtglatte Funktionen kann durch Glättung ausgewichen werden, vgl. hierzu (3.14) für Glättungen der Sprungfunktion $1(x)$ und der Signumfunktion $\operatorname{sgn}(x)$. Sprung- und Signumfunktion sind wiederum Basiselemente zur Glättung der Betragsfunktion $|x| = x \cdot \operatorname{sgn}(x) = 2x \cdot 1(x) - x$. Alternativ glättet $f_{abs}(x) = \sqrt[p]{|x|^p + \varepsilon}$ mit $p \geq 2$ den Betrag weitgehender. Gemeinhin reichen wenige Basiselemente, um beliebige nichtglatte Funktionen durch glatte zu approximieren.

[7] Für konvexe Funktionen bietet sich noch das ε-Subdifferenzial an, für Lipschitz-stetige f das Michel-Penot-Subdifferenzial und das Clarke-Subdifferenzial, für semiglatte f das Bouligand-Subdifferenzial, für stetige f das Fréchet-, das Mordukhovich- oder das Jeyakumar-Luc-Subdifferenzial u. v. a. [376], [40], [322]. In diesen Arbeiten finden sich auch Verallgemeinerungen der Optimalitätskriterien für freie und restringierte Probleme. Letztlich unterscheiden sich die Subdifferenziale primär hinsichtlich der Art der verwendeten Richtungsableitungen.

6.6.2 Matrixableitungen

Die Formulierung freier und restringierter Identifikationsprobleme führt bevorzugt auf Zielfunktionen mit Vektoren oder Matrizen als Variablen. Aus diesem Grund wurden schon beginnend in den 1930iger Jahren ein Vektorableitungs- und in den 1940igern ein Matrixableitungskalkül entwickelt [229], [175]. Zusammenfassende Darstellungen liefern die Bücher von Rogers [540] sowie Magnus und Neudecker [421]. Regeln finden sich auch in [625]. Die Mehrzahl der Arbeiten zu Matrixableitungen zielt auf das Lösen der Optimierungsaufgabe und betrachtet die Berechnung stationärer Punkte sowie den Extremalnachweis über die zweite Ableitung. Hierbei werden die Matrixableitungen als pure Rechenschemata betrachtet, und der Bezug zur Fréchet-Ableitung wird nicht hergestellt. Das hat zur Folge, dass Magnus und Neudecker die Transponierte der Jacobi-Matrix der vektorisierten Matrixfunktion als Gradient bezeichnen, während Flett [195] den Gradientenbegriff an ein hinterlegtes Skalarprodukt knüpft.

Die gängigen Ableitungsschemata lassen sich mit der Fréchet-Ableitung[8] für $f : \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q}$ in Verbindung bringen [503]

$$Df(X;H) = \frac{\mathrm{d}}{\mathrm{d}t} f(X+tH)\Big|_{t=0} \qquad \text{Richtungsableitungsform} \tag{6.20a}$$

$$= H \star \begin{bmatrix} \frac{\partial f(X)}{\partial x_{11}} & \cdots & \frac{\partial f(X)}{\partial x_{1n}} \\ \vdots & & \vdots \\ \frac{\partial f(X)}{\partial x_{m1}} & \cdots & \frac{\partial f(X)}{\partial x_{mn}} \end{bmatrix} \stackrel{\text{def}}{=} \sum_{i=1}^{m}\sum_{j=1}^{n} h_{ij} \frac{\partial f(X)}{\partial x_{ij}} \tag{6.20b}$$

$$= \begin{bmatrix} \frac{f_{11}(X)}{\partial X} & \cdots & \frac{f_{1q}(X)}{\partial X} \\ \vdots & & \vdots \\ \frac{f_{p1}(X)}{\partial X} & \cdots & \frac{f_{pq}(X)}{\partial X} \end{bmatrix} \copyright H \stackrel{\text{def}}{=} \begin{bmatrix} \langle \operatorname{vec} H, \operatorname{vec} \frac{\partial f_{11}(X)}{\partial X} \rangle \ldots \langle \operatorname{vec} H, \operatorname{vec} \frac{\partial f_{1q}(X)}{\partial X} \rangle \\ \vdots \qquad\qquad\qquad \vdots \\ \langle \operatorname{vec} H, \operatorname{vec} \frac{\partial f_{p1}(X)}{\partial X} \rangle \ldots \langle \operatorname{vec} H, \operatorname{vec} \frac{\partial f_{pq}(X)}{\partial X} \rangle \end{bmatrix} \tag{6.20c}$$

$$= \operatorname{vec}^{-1}_{p\times q}\Big(\frac{\partial \operatorname{vec} f(X)}{\partial \operatorname{vec}^T X} \cdot \operatorname{vec} H \Big) \qquad \frac{\partial \operatorname{vec} f(X)}{\partial \operatorname{vec}^T X} \text{ - Jacobi-Matrix.} \tag{6.20d}$$

Die Fréchet-Ableitung selbst ist eindeutig[9], nicht aber ihre Darstellung! So differieren die

[8] Seien $\mathcal{V}, \mathcal{W}$ normierte Vektorräume, dann heißt eine Abbildung $f : \mathcal{D} \subseteq \mathcal{V} \to \mathcal{W}$ in $x_0 \in \operatorname{int}\mathcal{D}$ Fréchet-differenzierbar, wenn es einen stetigen linearen Operator $Df(x_0;\cdot) \in \mathcal{L}(\mathcal{V},\mathcal{W})$ gibt, sodass

$$\lim_{h\to 0} \frac{||f(x_0+h) - f(x_0) - Df(x_0;h)||}{||h||} = 0. \tag{6.19}$$

$Df(x_0;\cdot) =: Df(x_0) : \mathcal{X} \to \mathcal{L}(\mathcal{V},\mathcal{W})$ heißt Fréchet-Ableitung von f in x_0 und $Df(x_0;h) =: Df(x_0)(h) \in \mathcal{W}$ bezeichnet das Fréchet-Differenzial von f in x_0.

[9] Diese Aussage gilt allgemeiner für alle endlichdimensionalen Vektorräume, da dort alle Normen zueinander äquivalent sind. In unendlichdimensionalen Vektorräumen ist die Ableitung hinsichtlich der Norm zu

Ableitungen in der Anordnung der partiellen Ableitungen und im verwendeten Produkt[10]. Während das $\star$-Produkt und das ©-Produkt den Operator direkt für Matrizen beschreiben, macht die dritte Darstellung vom isometrischen Isomorphismus zwischen dem linearen Raum der Matrizen und der Vektoren Gebrauch. Mit speziellen Permutationsmatrizen und dem Kronecker-Produkt lassen sich die drei Schemata ineinander überführen.
Der Vorteil von (6.20d) liegt darin, dass die Kettenregel in $\mathbb{R}^k$ mit der gewöhnlichen assoziativen Matrizenmultiplikation korrespondiert. Im Gegensatz dazu sind die anderen beiden Produkte nicht assoziativ. Ein weiterer Vorteil von (6.20d) ist, dass die Jacobi-Determinante der Identität $f(X) = X$ wegen $\frac{\partial \mathrm{vec} f(X)}{\partial \mathrm{vec}^T X} = I_{mn}$ Eins ist. Das privilegiert (6.20d) immer dann, wenn Funktionaldeterminanten gebraucht werden, also etwa bei Dichtetransformationen matrixvariater Funktionen oder bei der Integration über matrixvariate Funktionen. Demgegenüber empfehlen sich (6.20b) und (6.20c), da sie auf besser handhabbare Gleichungen in der unbekannten Matrix führen. Beide Ableitungsschemata sind für skalare $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ gleich, reduzieren sich zur Skalarproduktdarstellung

$$Df(X;H) = \left\langle \frac{\partial f(X)}{\partial X}, H \right\rangle = \mathrm{spur}\left(H^T \frac{\partial f(X)}{\partial X} \right) \tag{6.21}$$

und führen von der Stationären-Punkt-Bedingung $\forall H \in \mathbb{R}^{m\times n} : Df(X;H) = 0$ auf

$$\frac{\partial f(X)}{\partial X} = 0_{m\times n} \qquad \text{Stationäre-Punkt-Bedingung in Gradientenform.} \tag{6.22}$$

Für die zweiten Ableitungen ist die Sache differenzierter und muss fallweise betrachtet werden. Tendenziell gelingt der Nachweis der Definitheit aber für zweite Ableitungen i. S. von (6.20d) besser.

Anmerkung 6.7 Kleinste Quadratmittelprobleme mit komplexen Variablen (Ortskurvenapproximationen, Frequenzschätzungen s. Bsp. 2.6), führen auf $\frac{1}{2}||Ax-b||_2^2 \stackrel{!}{=}$ Min. Mittels formaler Ableitung

$$\frac{\partial}{\partial x}\frac{1}{2}||Ax-b||_2^2 = A^H(Ax-b) = 0_n \tag{6.23}$$

wird in der Literatur gelegentlich die Normalgleichung hergeleitet. Diese Herleitung ist falsch, wenngleich die entstehende Normalgleichung richtig ist! Die Ableitung nach einer komplexen Variablen erfordert, dass die Cauchy-Riemannschen Differenzialgleichungen [101] erfüllt sind, was hier nicht der Fall ist.[11] Die meisten komplexwertigen Funktionen in der Physik und

spezifizieren, und es ist Vorsicht geboten, da dann lineare Operatoren nicht mehr notwendigerweise stetig sind.

[10] Es gibt auch Darstellungen [585], [443], [534], die ohne partielle Ableitungen und Produkte auskommen wie $Df(X;H) = \int_0^1 \mathrm{e}^{(1-t)X} H \mathrm{e}^{tX} \mathrm{d}t$ für $f(X) = \exp X$.

[11] Setze $A := 1$, $b := 0$, dann ist $f(x) = \frac{1}{2}((\Re e\, x)^2 + (\Im m\, x)^2)$, weshalb $\frac{\partial \Re e\, f(x)}{\partial \Re e\, x} = \Re e\, x \neq \frac{\partial \Im m\, f(x)}{\partial \Im m\, x} = 0$ das Nichterfüllen anzeigt.

Technik erfüllen diese Differenzialgleichungen nicht. Deshalb wird auf das Wirtinger-Kalkül [557], [96]

$$\frac{\partial f(z)}{\partial z} \overset{\text{def}}{=} \frac{1}{2}\Big(\frac{\partial f(z)}{\partial \Re e\, z} - j\frac{\partial f(z)}{\partial \Im m\, z}\Big) \qquad \text{verallgemeinerte komplexe Ableitung} \tag{6.24}$$

und

$$\frac{\partial f(z)}{\partial \overline{z}} \overset{\text{def}}{=} \frac{1}{2}\Big(\frac{\partial f(z)}{\partial \Re e\, z} + j\frac{\partial f(z)}{\partial \Im m\, z}\Big) \qquad \text{verallgemeinerte konjugierte Ableitung}^{12} \tag{6.25}$$

ausgewichen, das sich problemlos auf den vektoriellen Fall erweitern lässt. Diese Ableitungen machen vom Isomorphismus zwischen $\mathbb{C}$ und $\mathbb{R}^2$ Gebrauch. Für analytische Funktionen[13] f gilt $\frac{\partial f(z)}{\partial z} = \frac{\mathrm{d}f(z)}{\mathrm{d}z}$ und für nichtanalytische $\frac{\partial f(z)}{\partial z} = 0$. Aus der verallgemeinerten konjugierten Ableitung in vektorieller Form folgt die Normalgleichung (6.23). Die Herleitung wird somit exakt, dennoch ist eine rein algebraische Herleitung von (6.23) über orthogonale Projektion und Pythagoras eleganter[14].

Anmerkung 6.8 Ist die Stationäre-Punkt-Bedingung inkonsistent, so hat das Problem keine Lösung. Beispiel: Für das reelle konvexe QP-Problem $f(x) = \frac{1}{2}x^T Ax + b^T x + c \overset{!}{=} \text{Min}$ mit $A \succeq 0_{n\times n}$ gilt $\frac{\partial f(x)}{\partial x} = Ax + b = 0_n$. Ist $b \notin \mathcal{R}(A)$, dann ist das Gleichungssystem unlösbar und f nach unten unbeschränkt. Ist $b \in \mathcal{R}(A)$, dann ist für $A \succ 0_{n\times n}$ die Lösung $x_{\text{opt}} = -A^{-1}b$ eindeutig und für $\text{rg}A < n$ mehrdeutig $\mathcal{X}_{\text{opt}} = -A^{+}b + \mathcal{N}(A)$.

6.6.3 Quasifreie Optimierungsprobleme

Unter freien Problemen werden jene verstanden, bei denen die $x \in \mathbb{R}^n$ bzw. $X \in \mathbb{R}^{m\times n}$ keinen Restriktionen unterworfen sind. Handelt es sich bei der Restriktion aber um eine Unterraum- oder affine Raum-Restriktion, so kann ein solches Problem wie ein freies behandelt werden, allerdings ist ein auf die Restriktion zugeschnittenes Kalkül zu verwenden. Der Vorteil ist, dass auf den Einsatz von Lagrange-Multiplikatoren verzichtet werden kann.
Im vorhergehenden Abschnitt wurde angenommen, dass alle Matrixelemente von X mathematisch unabhängig voneinander sind und zudem variabel sind. Bei einer symmetrischen

[12] Statt $\frac{\partial f(z)}{\partial \overline{z}}$ (Der Überstrich meint die konjugiert-komplexe Variable.) wird häufig die Notation $\overline{\partial} f(z)$ verwendet, wobei $\overline{\partial}$ der sog. Cauchy-Riemann-Operator ist.

[13] Eine Funktion, die in jedem Punkt des Definitionsbereichs eine lokal konvergente Potenzreihe hat, heißt analytisch. So ist $f(z) = z$ analytisch, $f(z) = \overline{z}$ aber nicht.

[14] $||(Ax - P_{\mathcal{R}(A)}b) - P^{\perp}_{\mathcal{R}(A)}b||_2^2 = ||Ax - P_{\mathcal{R}(A)}b||_2^2 + ||P^{\perp}_{\mathcal{R}(A)}b||_2^2 \Rightarrow Ax - P_{\mathcal{R}(A)}b = A(x - A^{(1,3)}b) = 0_n \Rightarrow x \in A^{(1,3)}b + \mathcal{N}(A)$
Die Äquivalenz zu $A^H(Ax - b) = 0_n$ folgt über $x \in (A^HA)^-A^Hb + \mathcal{N}(A^HA) = (A^HA)^-A^Hb + \mathcal{N}(A)$ und der Tatsache, dass $(A^HA)^-A^H$ eine (1,3)-Inverse von A ist, da nämlich $A(A^HA)^-A^HA = A$ [528] und $(A(A^HA)^-A^H)^H = A(A^HA)^-A^H$ gilt.

Matrix sind sie es nicht, und bei einer Matrix mit einer Einsdiagonale sind nicht alle Elemente variabel. Bevor nachfolgend ein recht allgemeines Kalkül vorgestellt wird, soll am Beispiel der symmetrischen Matrizen zunächst das in der Literatur standardmäßig empfohlene Kalkül mit dem Ableitungsschema

$$\frac{\partial_S f(X)}{\partial X} = \frac{\partial f(x)}{\partial X} + \left(\frac{\partial f(x)}{\partial X}\right)^T - \mathrm{Diag}\left(\frac{\partial f(x)}{\partial X}\right) \tag{6.26}$$

analysiert werden. Hierbei sind alle Ableitungen der rechten Seite als unabhängig und variabel zu betrachten. Die Konstruktion liefert eine symmetrische Matrix und auch das Abziehen der Hauptdiagonalen erscheint logisch, da deren Elemente sonst zweimal gezählt würden. Gleichwohl hat (6.26) zwei Schwächen (Nichtisometrie und Unstetigkeit):

1. Formel (6.26) ist eine Matrixdarstellung des auf symmetrischer Vektorisierung basierenden Kalküls analog zu (6.20d)[15]. Somit sind die beiden Hilbert-Räume $(\mathcal{S}_n, \mathrm{spur}(XY))$ und $(\mathbb{R}^{n(n+1)/2}, y^T x)$ isomorph. Sie sind aber nicht isometrisch zueinander. Das bereitet solange keine Schwierigkeiten, solange die Ableitungen nur Null sein müssen. Bei quantitativen Betrachtungen treten jedoch Probleme auf, wie Varfolomeev [611] an einem Optimalsteuerproblem zeigt. Das liegt daran, dass sich Parameteränderungen wegen der in den Räumen unterschiedlichen Metriken unterschiedlich auswirken.

2. Wenn A symmetrisch ist, dann ist die freie Ableitung von $\frac{\partial}{\partial X}\mathrm{spur}(AX) = A = A^T$ für alle freien Punkte X symmetrisch. Es wäre zu erwarten, dass das symmetrische Ableitungskalkül (6.26) in allen Punkten $X = X^T$ ebenfalls das symmetrische A liefert und damit Stetigkeit bezüglich der freien Ableitung sicherstellt. Er erzeugt aber $\frac{\partial_S}{\partial X}\mathrm{spur}(AX) = 2A - \mathrm{Diag}(A)$ im Gegensatz zum neuen $\mathcal{L}$-Ableitungskalkül mit $\nabla_{\mathcal{S}_n}\mathrm{spur}(AX) = A = A^T$.

Beide Schwächen werden durch das hier erstmals eingeführte Kalkül der $\mathcal{L}$-Ableitung behoben, mit der Matrixableitungen bei linearabhängigen Matrixelementen berechnet werden können. Strenggenommen ist die nachfolgend als $\mathcal{L}$-Ableitung bezeichnete Ableitung nichts anderes als eine Fréchet-Ableitung, bei der als Vektorraum des Arguments eine Menge $\mathcal{L} \subseteq \mathbb{R}^{m\times n}$ linear strukturierter Matrizen (symmetrische, schiefsymmetrische, Diagonal- oder Dreiecksmatrizen, Toeplitz-Matrizen usw.) fungiert und bei der die Beschreibung der linearen Abbildung für die Ableitung durch das Standardskalarprodukt vorgegeben wird.

Definition 6.1 ($\mathcal{L}$-Ableitung) Es sei $f : \mathcal{L} \subset \mathbb{R}^{m\times n} \to \mathbb{R}$ eine Abbildung zwischen den normierten Räumen $(\mathcal{L}, ||.||_F)$ und $(\mathbb{R}, |.|)$ und $\mathcal{U} \subset \mathcal{L}$ eine offene Teilmenge. Die Abbildung f

[15] Die freien Elemente der symmetrischen Matrix, also ein Dreieck, werden dabei in einen Vektor geordnet und dann wird wie gewöhnlich nach Vektoren abgeleitet.

heißt dann $\mathcal{L}$-differenzierbar in $X \in \mathcal{U}$, wenn eine Matrix $\nabla_{\mathcal{L}} f(X) \in \mathcal{L}$ – genannt $\mathcal{L}$-Gradient – existiert, sodass[16]

$$\lim_{\substack{H \to 0 \\ H \in \mathcal{L}}} \frac{|f(X+H) - f(X) - \langle \nabla_{\mathcal{L}} f(X), H \rangle|}{||H||_F} = 0 \tag{6.27}$$

gilt. Die Abbildung $D_{\mathcal{L}} f(X;.) : \mathcal{L} \to \mathbb{R}; H \mapsto \langle \nabla_{\mathcal{L}} f(X), H \rangle$ heißt $\mathcal{L}$-Ableitung.

Definitionsbedingt besitzt die $\mathcal{L}$-Ableitung die Eigenschaften einer Fréchet-Ableitung!

Für die Berechnung der $\mathcal{L}$-Ableitung wird die stets eindeutige, orthogonale Projektion eines Elements Y auf $\mathcal{L}$ benötigt

$$\text{Proj}_{\mathcal{L}}(Y) \stackrel{\text{def}}{=} \underset{X \in \mathcal{L}}{\text{argmin}} \; ||Y - X||_F. \tag{6.28}$$

Die Berechnung selbst ist einfach, denn hierzu braucht nur der Gradient[17] $\nabla f(X)$ der erweiterten Abbildung $f : \mathbb{R}^{m \times n} \to \mathbb{R}$ (alle Matrixelemente sind frei und hängen nicht voneinander ab) auf $\mathcal{L}$ projiziert werden. Diese wichtige neue Formel soll als Satz hervorgehoben werden.

Satz 6.2 (Formel zur Berechnung des $\mathcal{L}$-Gradienten)

$$\nabla_{\mathcal{L}} f(X) = \text{Proj}_{\mathcal{L}}(\nabla f(X)) \qquad \text{mit} \;\; \nabla f(X) = \frac{\partial f(X)}{\partial X}. \tag{6.29}$$

Beweis 6.1 Ein jedes $\nabla f(X)$ kann eindeutig entsprechend $\nabla f(X) = \text{Proj}_{\mathcal{L}}(\nabla f(X)) + \text{Proj}_{\mathcal{L}^\perp}(\nabla f(X))$ in Komponenten in $\mathcal{L}$ und im orthogonalem Komplement $\mathcal{L}^\perp$ von $\mathcal{L}$ zerlegt werden. Dann gilt $\langle \text{Proj}_{\mathcal{L}^\perp}(\nabla f(X)), H \rangle = 0$ für alle $H \in \mathcal{L}$. Aus der F-Differenzierbarkeit folgt über

$$\begin{aligned}
0 &= \lim_{H \to 0} \frac{|f(X+H) - f(X) - \langle \nabla f(X), H \rangle|}{||H||_F} \qquad \forall X, H \in \mathbb{R}^{m \times n} \\
&= \lim_{H \to 0} \frac{|f(X+H) - f(X) - \langle \text{Proj}_{\mathcal{L}}(\nabla f(X)) + \text{Proj}_{\mathcal{L}^\perp}(\nabla f(X)), H \rangle|}{||H||_F} \\
&= \lim_{\substack{H \to 0 \\ H \in \mathcal{L}}} \frac{|f(X+H) - f(X) - \langle \text{Proj}_{\mathcal{L}}(\nabla f(X)), H \rangle|}{||H||_F} \qquad \forall X \in \mathbb{R}^{m \times n} \text{ und somit } \forall X \in \mathcal{L}
\end{aligned}$$

unmittelbar, dass $\text{Proj}_{\mathcal{L}}(\nabla f(X))$ der $\mathcal{L}$-Gradient ist. Da das Skalarprodukt nicht geändert wurde, korrespondiert $\nabla_{\mathcal{L}} f(X)$ ebenfalls mit dem Standardskalarprodukt. ■

[16] $\langle X, Y \rangle \stackrel{\text{def}}{=} \text{spur}(Y^T X)$ bezeichnet das Standardskalarprodukt in $\mathcal{L}$.

[17] Nach [195] heißt $\nabla f(X)$ Gradient, wenn $Df(X;H) = \langle \nabla f(X), H \rangle$ gilt. Mit $H = tE_{ij}$ folgt aus der Definition der Fréchet-Ableitung

$$\begin{aligned}
0 = \lim_{tE_{ij} \to 0} \frac{|f(X+tE_{ij}) - f(X) - \langle \nabla f(X), tE_{ij} \rangle|}{||tE_{ij}||_F} &= \lim_{t \to 0} \frac{|f(X+tE_{ij}) - f(X)|}{t} - (\nabla f(X))_{ij} \\
&= \frac{\partial f(X)}{\partial x_{ij}} - (\nabla f(X))_{ij} \qquad \Rightarrow \nabla f(X) = \frac{\partial f(X)}{\partial X}.
\end{aligned}$$

Als Konsequenz von Formel (6.29) ergeben sich für die symmetrischen Matrizen $\mathcal{L} = \mathcal{S}_n$, die schiefsymmetrischen Matrizen $\mathcal{L} = \mathcal{S}_n^\perp$, die diagonalen Matrizen $\mathcal{L} = \mathcal{D}_n$ und die Dreiecksmatrizen $\mathcal{L} = \Delta_n$ durch explizite Darstellung der Projektion die Formeln:

$$\nabla_{\mathcal{S}_n} f(X) = \frac{1}{2}\left[\frac{\partial f(X)}{\partial X} + \left(\frac{\partial f(X)}{\partial X}\right)^T\right] \tag{6.30}$$

$$\nabla_{\mathcal{S}_n^\perp} f(X) = \frac{1}{2}\left[\frac{\partial f(X)}{\partial X} - \left(\frac{\partial f(X)}{\partial X}\right)^T\right] \tag{6.31}$$

$$\nabla_{\mathcal{D}_n} f(X) = \mathrm{Diag}\left(\frac{\partial f(X)}{\partial X}\right) \tag{6.32}$$

$$\nabla_{\Delta_n} f(X) = \Delta\left(\frac{\partial f(X)}{\partial X}\right). \tag{6.33}$$

Für allgemeinere Mengen $\mathcal{L} = \{\sum_{i=1}^k \alpha_i B_i : \alpha_i \in \mathbb{R}\}$, wobei $\{B_1, \ldots, B_k\}$ eine Basis des k-dimensionalen Unterraums in $\mathbb{R}^{m\times n}$ sei, gilt

$$\nabla_{\mathcal{L}} f(X) = \sum_{i=1}^k \alpha_{i,\mathrm{opt}} B_i, \tag{6.34}$$

wobei die $\alpha_{i,\mathrm{opt}}$ Lösungen von

$$\begin{bmatrix} \mathrm{Spur}(B_1^T B_1) & \ldots & \mathrm{Spur}(B_1^T B_k) \\ \vdots & & \vdots \\ \mathrm{Spur}(B_k^T B_1) & \ldots & \mathrm{Spur}(B_k^T B_k) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{bmatrix}_{\mathrm{opt}} = \begin{bmatrix} \mathrm{Spur}(B_1^T \nabla f(X)) \\ \vdots \\ \mathrm{Spur}(B_k^T \nabla f(X)) \end{bmatrix} \tag{6.35}$$

sind, was seinerseits Lösung von $||\sum_{i=1}^k \alpha_i B_i - \nabla f(X)||_F \stackrel{!}{=} \mathrm{Min}; \alpha_1, \ldots, \alpha_k \in \mathbb{R}$ ist.

Anwendung erfährt Formel (6.29) bei der quasifreien Optimierung.

Definition 6.2 (Quasifreies Optimierungsproblem)
Ein Problem heißt quasifrei, wenn es eine Restriktion der Art $X \in \mathcal{L}$ ($\mathcal{L}$ ist ein linearer Raum) hat. Dann lässt es sich wie ein freies Problem behandeln, wobei in der Stationären-Punkt-Bedingung

$$\nabla_{\mathcal{L}} f(X) = 0_{m\times n} \qquad \text{und} \qquad X \in \mathcal{L} \tag{6.36}$$

lediglich die gewöhnliche Ableitung durch die $\mathcal{L}$-Ableitung zu ersetzen ist.

Beispiel 6.12 (Symmetrisches Prokrustes-Problem)
Das LS-Problem unter der Nebenbedingung, dass X eine symmetrische Matrix sein soll,

$$\frac{1}{2}||AX - B||_F^2 \stackrel{!}{=} \mathrm{Min} \quad X \in \mathcal{S}_n,$$

lässt sich elegant mit der $\mathcal{L}$-Ableitung lösen. Aus $\nabla_{\mathcal{S}_n} f(X) = \frac{1}{2}(A^T AX - A^T B + (A^T AX - A^T B)^T) = 0_{n\times n}$ und der Konvexität folgt unmittelbar, dass alle symmetrischen Lösungen X der Gleichung Minimierer sind.

Anmerkung 6.9 Affine Restriktionen $\mathcal{L}_A = L_0 + \mathcal{L}$ werden in der quasifreien Optimierung durch Rückführung auf lineare Restriktionen gelöst. Die Ersetzung $X = L_0 + Y$ in der Zielfunktion schafft mit $Y \in \mathcal{L}$ eine solche Restriktion, womit dann $X_{\text{opt}} = L_0 + Y_{\text{opt}}$ folgt.

6.6.4 Optimalitätsbedingungen

In diesem Abschnitt werden Optimalitätsbedingungen vorgestellt, die sich mit der Methode der Lagrange-Multiplikatoren ergeben und die ihrerseits wieder genutzt werden können, um die Optimallösung zu berechnen. Als Lagrange-Multiplikatoren λ_i, μ_i (λ_i für die Gleichungsrestriktionen, μ_i für die Ungleichungsrestriktionen) werden dabei zusätzliche Variablen bezeichnet, die die Zielfunktion zu einer Lagrange-Funktion

$$L(x; \lambda_0, \lambda_1, \ldots, \lambda_m; \mu_1, \ldots, \mu_p) = \lambda_0 f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{i=1}^{p} \mu_i g_i(x); \qquad \mu_i \geq 0 \qquad (6.37)$$

erweitern. Für diese Zielfunktion werden in herkömmlicher Weise stationäre Punkte gesucht.

Prinzipiell sind zwei Arten von Multiplikatorzugängen zu unterscheiden, jene vom Karush-Kuhn-Tucker-Typ (KKT) und jene vom Fritz-John-Typ (FJ) [71]. Beim KKT-Typ entfällt der Faktor λ_0 vor $f(x)$ (d. h. $\lambda_0 = 1$), beim FJ-Typ dagegen nicht. Der Preis für einen Parameter weniger beim KKT-Typ sind die Constraint-Qualifications (zusätzliche Bedingungen, um Existenz und Eindeutigkeit der Lagrange-Multiplikatoren zu sichern, s. [71]), die beim FJ-Typ entfallen. Normalerweise sind die Constraint-Qualifications erfüllt, weshalb sie schnell in Vergessenheit geraten. Das folgende Beispiel zeigt Fälle, in denen sie nicht erfüllt sind.

Beispiel 6.13 (Nichteindeutigkeit, Nichtexistenz von Lagrange-Multiplikatoren)
$(x_1 - 1)^2 + (x_2 - 1)^2 \stackrel{!}{=} \text{Min}$; $(x_1 - 1)^2 - 2(x_2 - 1)^2 = 0$ hat $(1, 1)$ als globalen Minimierer, der Lagrange-Multiplikator ist aber beliebig, wie $\frac{\partial L}{\partial x_1} = 2(1 + \lambda)(x_1 - 1) = 0$ und $\frac{\partial L}{\partial x_2} = 2(1 - 2\lambda)(x_2 - 1) = 0$ für $(1, 1)$ sofort zeigen.
$x_2 \stackrel{!}{=} \text{Min}$; $(x_2 - x_1^2)(2x_2 - x_1^2) = 0$ hat den globalen Minimierer $(0, 0)$, in dem kein Lagrange-Multiplikator existiert. Begründung: Die Restriktion zerfällt in die beiden Fälle $x_2 = x_1^2$ und $x_2 = \frac{1}{2}x_1^2$, woraus $x_1 = 0$ und damit der globale Minimierer resultiert. $\frac{\partial L}{\partial x_2} = 1 + \lambda[(2x_2 - x_1^2) + 2(x_2 - x_1^2)] = 0$ kann indes für $(0, 0)$ nie erfüllt werden.
Singuläre Ungleichungsrestriktionen wie $(g(x))^2 \leq 0$ bereiten ebenso Schwierigkeiten wie Gleichungsrestriktionen, die den Suchraum auf einzelne isolierte Punkte einschränken [71].

Aus der Vielzahl von Sätzen, die die lokalen Minimierer des Gütekriteriums mit den sollen hier nur zwei leistungsstarke Sätze vom Fritz-John-Typ aufgeführt werden.

Satz 6.3 (Notwendige Optimalitätsbedingung 1. Ordnung, [270])
Wenn x_{loc} lokaler Minimierer von

$$f(x) \stackrel{!}{=} \text{Min} \qquad x \in \mathcal{V} : g_i(x) \leq 0,\ i = 1, \ldots, p;\quad h_i(x) = 0,\ i = 1, \ldots, m \tag{6.38}$$

ist und $h_1, \ldots, h_m$ in der Umgebung von x_{loc} stetig sind und alle Funktionen f, g_i, h_i in x_{loc} F-differenzierbar sind, dann existiert ein Vektor $(\mu_1, \ldots, \mu_p, \lambda_0, \ldots, \lambda_m)^T \in \mathbb{R}^{m+p+1} \backslash \{0\}$, sodass

$$\sum_{i=1}^{p} \mu_i \nabla g(x_{\text{loc}}) + \lambda_0 \nabla f(x_{\text{loc}}) + \sum_{i=1}^{m} \lambda_i \nabla h(x_{\text{loc}}) = 0_{\mathcal{H}} \tag{6.39a}$$

$$\lambda_0, \mu_1, \ldots, \mu_p \geq 0 \tag{6.39b}$$

$$\mu_i g_i(x_{\text{loc}}) = 0 \quad i = 1, \ldots, p. \tag{6.39c}$$

Anmerkung 6.10 In Differenzialen (braucht keine Gradientendarstellung und damit keinen Hilbert-Raum) kann der Gleichungsfall auch ausgedrückt werden durch

$$Df(x_{\text{loc}}; v) = 0 \qquad \text{für alle } v \in \mathcal{V}, \text{ sodass} \quad Dh_i(x_{\text{loc}}; v) = 0; i = 1, \ldots, m. \tag{6.40}$$

Anmerkung 6.11 Geometrisch bedeutet (6.39), dass $\nabla f(x_{\text{loc}})$ eine Linearkombination der Gradienten der Restriktionen ist oder anders ausgedrückt: Der Zielfunktionsgradient steht in x_{loc} senkrecht auf dem Tangentialraum der Restriktionen.

Anmerkung 6.12 Auf die Stetigkeitsforderung in Satz 6.3 kann nicht verzichtet werden und F-Differenzierbarkeit lässt sich nicht auf G-Differenzierbarkeit abschwächen [190]. Treten im Problem (6.38) jedoch keine Gleichungsrestriktionen auf, dann gelten abgeschwächte Voraussetzungen, wie folgender Satz zeigt.

Satz 6.4 (Notwendige Optimalitätsbed. ohne Gleichungsrestriktionen, [190])
Wenn x_{loc} Minimierer von (6.38) bei $m = 0$ ist und alle f, g_i konvexe Richtungsableitungen in x_{loc} besitzen, dann existiert ein Vektor $(\mu_1, \ldots, \mu_p, \lambda_0)^T \in \mathbb{R}^{p+1} \backslash \{0_{p+1}\}$, sodass

$$\sum_{i=1}^{p} \mu_i g'(x_{\text{loc}}; v) + \lambda_0 f'(x_{\text{loc}}; v) \geq 0,\ \forall v \in \mathcal{V} \tag{6.41a}$$

$$\lambda_0, \mu_1, \ldots, \mu_p \geq 0 \tag{6.41b}$$

$$\mu_i g_i(x_{\text{loc}}) = 0 \quad i = 1, \ldots, p. \tag{6.41c}$$

Anmerkung 6.13 Ungleichungsrestriktionen bedingen in aller Regel Fallunterscheidungen für die Lagrange-Multiplikatoren μ_i dahingehend, ob $\mu_i = 0$ bei inaktiver Ungleichung oder $\mu_i \geq 0$ bei aktiver Ungleichung ist, vgl. [71].

Anmerkung 6.14 Für konvexe Probleme sind die KKT- und FJ-Bedingungen hinreichend für globale Optimalität. Zugleich sind sie notwendig für globale Optimalität (im KKT-Fall vorbehaltlich einer Constraint-Qualification) [71].

Analog zum freien Problem gibt es auch im restringierten Fall eine notwendige Bedingung zweiter Ordnung, die für nichtkonvexe Probleme benötigt wird.

Satz 6.5 (Notwendige Optimalitätsbedingung 2. Ordnung, [71])
Zweifache stetige Differenzierbarkeit in $\mathcal{U}_\varepsilon(x_{\text{loc}})$ für f, h_i und alle in x_{loc} aktiven Ungleichungen $g_{i,\text{akt}}$ vorausgesetzt, muss

$$y^T\Big(\lambda_{0,\text{loc}} D^2 f(x_{\text{loc}}) + \sum_{i=1}^{m} \lambda_{i,\text{loc}} D^2 h_i(x_{\text{loc}}) + \sum_{i=1}^{m} \mu_{i,\text{loc}} D^2 g_i(x_{\text{loc}})\Big) y \geq 0 \tag{6.42}$$

für alle $\{y \in \mathbb{R}^n : y^T \nabla h_i(x_{\text{loc}}) = 0; i = 1, \ldots, m, y^T \nabla g_{i,\text{akt}}(x_{\text{loc}}) = 0\}$ gelten.

Die Methode der Lagrange-Multiplikatoren wird überwiegend nur für Funktionen über dem $\mathbb{R}^n$ eingeführt. Sie lässt sich aber auf komplexe Argumente, auf Matrizen und unendlichdimensionale Probleme (Variationsrechnung) problemlos erweitern, sofern die Zielfunktion reell ist. Üblicherweise werden bei der Einführung dieser Methode relativ hohe Forderungen an die Differenzierbarkeit gestellt. Diese lassen sich in unterschiedlicher Weise abschwächen, was für praktische Anwendungen in der Modellbildung auch notwendig ist. So geht die Differenzierbarkeit bei bestimmten Normen (z. B. Chebyshev-Norm, Betragssummennorm), Gütekriterien mit einer Totzone für den Fehler (etwa für adaptive Regelungen), der Parameterschätzung in Modellen mit nicht-differenzierbaren Nichtlinearitäten (Sättigung) oder etwa bei der Konstruktion exakter Penalty-Funktionen verloren. Auswege sind dann die separate Behandlung der Nicht-Differenzierbarkeitsstellen oder die Verwendung alternativer Ableitungskalküle [195], [71], s. auch Anmerkung 6.5.

6.6.5 Lagrange-Terme für matrixvariate Funktionen

Um die Vorteile der Vektor- oder Matrixableitungskalküle effektiv nutzen zu können, werden statt der Summendarstellung auch für die Lagrange-Terme vektor- oder matrixorientierte Darstellungen verwendet. In Tabelle 6.1 sind häufig verwendete Terme zusammengestellt, die zum Teil aus [421] stammen, aber auch während eigener Arbeiten entwickelt wurden. Nachfolgend wird exemplarisch eine Anwendung für die erste Zeile aus Tabelle 6.1 gezeigt. Ein QPE-Problem (quadratic programming with (linear) equality) entsteht, wenn die Norm in einer LSE-Formulierung aufgelöst wird oder wenn eine Taylor-Approximation zweiter Ordnung einer nichtlinearen Zielfunktion und eine Taylor-Approximation erster Ordnung der Gleichungsrestriktion vorgenommen wird, wie es der Einsatz von SQP-Algorithmen (Sequentielle Quadratische Programmierung [71]) erfordert.

Beispiel 6.14 (QPE-Problem)
Zum QPE-Problem

$$\frac{1}{2} x^T A x + b^T x + \gamma \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : Cx = d; A \in \mathcal{S}_n^>$$

Bedingung	Lagrange-Term $\psi(X;\Lambda) =$
$Cx = d;\ C \in \mathbb{R}^{m\times n}, x \in \mathbb{R}^n$	$= \lambda^T(Cx - d)$
$CX = D;\ C \in \mathbb{R}^{p\times m}, X \in \mathbb{R}^{m\times n}$	$= \text{spur}(\Lambda^T(CX - D));\ \Lambda \in \mathbb{R}^{p\times n}$
$X \in \mathbb{R}^{m\times n} : x_{ij} = 0$ für $(i,j) \in \mathcal{M}$	$= \text{spur}(\Lambda^T X);\ \Lambda \in \mathbb{R}^{m\times n} : \lambda_{ij} = 0$ für $(i,j) \notin \mathcal{M}$
$X = X^T;\ X \in \mathbb{R}^{n\times n}$	$= \text{spur}(\Lambda(X - X^T));\ \Lambda = \Lambda^T \in \mathbb{R}^{n\times n}$ $= \text{spur}(\Lambda X);\ \Lambda = -\Lambda^T \in \mathbb{R}^{n\times n}$
$X = -X^T;\ X \in \mathbb{R}^{n\times n}$	$= \text{spur}(\Lambda(X + X^T));\ \Lambda = \Lambda^T \in \mathbb{R}^{n\times n}$
$X \succeq 0_{n\times n}$	$= -\text{spur}(\Lambda X);\ \Lambda \succeq 0_{n\times n}$
$X^T X = I_n;\ X \in \mathbb{R}^{m\times n}$	$= \text{spur}(\Lambda(X^T X - I_n));\ \Lambda = \Lambda^T \in \mathbb{R}^{n\times n}$
$F(X) = F^T(X) \in \mathbb{R}^{p\times p}$	$= \text{spur}(\Lambda(F(X) - F^T(X));\ \Lambda = \Lambda^T \in \mathbb{R}^{p\times p}$
$F(X) = C; C = C^T \in \mathbb{R}^{p\times p};\ X \in \mathbb{R}^{m\times n}$	$= \text{spur}(\Lambda(F(X) - C));\ \Lambda = \Lambda^T \in \mathbb{R}^{p\times p}$

Tabelle 6.1: Lagrange-Terme für matrixvariate Funktionen

lässt sich die Lagrange-Funktion mit dem Lagrange-Multiplikatorvektor λ formulieren

$$L(x;\lambda) = \frac{1}{2}x^T Ax + b^T x + \gamma + \lambda^T(Cx - d)$$

und nach Ableiten folgt

$$\begin{bmatrix} A & C^T \\ C & 0_{m\times n} \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix}_{\text{loc}} = \begin{bmatrix} -b \\ d \end{bmatrix}.$$

Das zweite Beispiel zur Anwendung der Tabelle 6.1 greift das Problem aus Beispiel 5.3 wieder auf. Diesmal wird die Definitheitsrestriktion an die Kovarianzmatrix nicht durch eine Variablensubstitution umgangen (Variante 1), sondern sie wird direkt über die Methode der Lagrange-Multiplikatoren behandelt. In Beispiel 8.2 wird dann noch eine weitere Lösungsvariante aufgezeigt, die auf einer Relaxation der Restriktion beruht.

Beispiel 6.15 (Herleitung des ML-Kovarianzmatrixschätzers, Variante 2)
Die auf der Log-Likelihood-Funktion (8.10c) basierende Lagrange-Funktion lautet mit $Z(\mu) = \sum_{i=1}^N (y_i - \mu)(y_i - \mu)^T$

$$L(\mu, \Sigma; \Lambda) = -\frac{Nn\ln(2\pi)}{2} - \frac{N}{2}\ln\det\Sigma - \frac{1}{2}\text{spur}(\Sigma^{-1}Z(\mu)) - \text{spur}(\Lambda\Sigma).$$

Die KKT-Bedingungen lauten mit der $\mathcal{S}_n$-Ableitung für symmetrisches Σ

$$\nabla_{\mathcal{S}_n} L = -\frac{N}{4}((\Sigma^{-1})^T + \Sigma^{-1}) + \frac{1}{4}((\Sigma^{-1}Z\Sigma^{-1})^T + (\Sigma^{-1}Z\Sigma^{-1})) - \Lambda = 0_{n\times n} \tag{6.43a}$$

$$\text{spur}(\Sigma\Lambda) = 0 \qquad \Leftrightarrow \Sigma\Lambda = 0_{n\times n} \ \text{ unter (6.43c) [365]} \tag{6.43b}$$

$$\Sigma, \Lambda \succeq 0_{n\times n}. \tag{6.43c}$$

Regularität von Σ in (6.43b) impliziert $\Lambda = 0_{n\times n}$, und $\hat{\Sigma}_{ML} = \frac{1}{N} Z(\hat{\mu})$ kann abgelesen werden. Soll die gewöhnliche Ableitung eingesetzt werden, so ist die Symmetrierestriktion an Σ über $L_2 = L(\mu, \Sigma; \Lambda) + \text{spur}(\Lambda_2 \Sigma)$ mit schiefsymmetrischen Λ_2 einzubeziehen. (6.43a) wird damit zu

$$\nabla L_2 = -\frac{N}{2}(\Sigma^{-1})^T + \frac{1}{2}(\Sigma^{-1} Z \Sigma^{-1})^T + \Lambda_2^T - \Lambda = 0_{n\times n}.$$

Mit $\Lambda = 0_{n\times n}$ folgt auch $\Lambda_2 = 0_{n\times n}$, da die ersten beiden Summanden symmetrisch sind und somit nie durch Addition mit der schiefsymmetrischen Matrix Λ_2^T Null werden können, es sei denn, sie sind in Summe Null und Λ_2 ist die Nullmatrix. Links- und Rechtsmultiplikation mit Σ liefert wiederum $\hat{\Sigma}_{ML} = \frac{1}{N} Z(\hat{\mu})$.

Am Beispiel ist zu erkennen, dass die Schwierigkeit bei Anwenden der Methode der Lagrange-Multiplikatoren nicht im Aufstellen der Lagrange-Funktion und dem daraus erforderlichen Ableiten der Optimalitätsbedingung liegt, sondern im Lösen der entstehenden Gleichungen, die je nach Aufgabe ein erhöhtes Maß an Kenntnissen der Matrixalgebra erfordern.

Im folgenden Abschnitt wird nun eine Problemklasse behandelt, für die die Methode der Lagrange-Multiplikatoren nicht anwendbar ist. Das liegt daran, dass Ränge von Matrizen nur ganzzahlige Werte annehmen können, womit sich ableitungsbasierte Methoden verbieten.

6.7 Faktorisierungen von Max-Rang-k-Restriktionen

Max-Rang-k-Restriktionen, d. h. Restriktionen der Art $\text{rg}X \leq k$, sind von zentraler Bedeutung für die Lösung von Reduktionsproblemen, multivariater Regression (Hauptkomponentenregression, Partielle LS), Approgressionsproblemen, Multidimensionaler Skalierung, Fehler-in-den-Variablen-Problemen usw. Häufig werden sie als Matrixapproximationsprobleme (low rank matrix approximation)

$$||A - X|| \stackrel{!}{=} \text{Min} \qquad X \in \mathcal{M} \subseteq \mathbb{R}^{m\times n} : \text{rg}X \leq k < r; A \in \mathbb{R}_r^{m\times n}. \tag{6.44}$$

formuliert. Obwohl die Restriktion nicht konvex ist, hat das Problem für $\mathcal{M} = \mathbb{R}^{m\times n}$ und einige Normen eine explizite Lösung. Diese leitet sich für die Frobenius-Norm aus dem Eckart-Young-Theorem und für die Spektralnorm aus dem Schmidt-Mirsky-Theorem [299] ab.

Satz 6.6 (Eckart-Young-Theorem, Rang-k-Approximation, [62])
Das Problem

$$||A - X||_F \stackrel{!}{=} \text{Min} \qquad X \in \mathbb{R}^{m\times n} : \text{rg}X = k < \text{rg}A; A \in \mathbb{R}^{m\times n} \tag{6.45}$$

hat das Minimum $Q_{\min} = \sqrt{\sigma_{k+1}^2(A) + \ldots + \sigma_{\text{rg}(A)}^2(A)}$ bei $X_{\text{opt}} = \sum_{i=1}^{k} \sigma_i(A)\, u_i v_i^T$, wobei (σ_i, u_i, v_i) das Singulärwert(vektor)tripel bezeichnet. Der Minimierer ist für $\sigma_k(A) > \sigma_{k+1}(A)$ eindeutig und für $\sigma_k(A) = \sigma_{k+1}(A)$ existiert eine Lösungsmenge, die X_{opt} enthält.

Aus Satz 6.6, der nur den Fall $\mathrm{rg}X = k$ behandelt, folgt, dass das Minimum für alle Ränge $k' = k - l; l = 1, \ldots, k$ größer oder höchstens gleich groß $Q_{\min}$ für $\mathrm{rg}X = k$ ist, da l Singulärwerte dazukommen. Damit ist jeder Rang-k-Approximand stets auch ein Max-Rang-k-Approximand. Aus der Lösung ist auch ersichtlich, dass bei symmetrischem A und/oder nichtnegativ definitem A (z. B. Kovarianzmatrix) der Max-Rang-k-Approximand $X_{\mathrm{opt}} = \sum_{i=1}^{k} \sigma_i(A)\, u_i u_i^T$ auch diese Eigenschaft aufweist. Anders ist die Situation, wenn X eine Toeplitz- oder Hankel-Matrix (ARMA-Kovarianzmatrizen, Datenmatrix eines Differenzengleichungsmodells) oder eine affin strukturierte Matrix mit $\mathrm{rg}X \leq k$ sein soll. Erstens erhält dann die reduzierte Singulärwertzerlegung nicht mehr die Struktur von A, was explizite Restriktionen für X bezüglich $\mathcal{M}$ erfordert. Zweitens ist selbst bei linearen Räumen $\mathcal{M}$ und $\sigma_k(A) > \sigma_{k+1}(A)$ Eindeutigkeit nicht mehr zwingend, vgl. Max-Rang-1-Approximation von $A = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 3 \end{smallmatrix}\right]$ durch eine symmetrische Toeplitz-Matrix mit $X_{\mathrm{opt},1} = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$ und $X_{\mathrm{opt},2} = \left[\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right]$. Dadurch wird das Problem erheblich schwieriger, wenngleich die Max-Rang-k-Restriktion im Gegensatz zur Rang-k-Restriktionen wenigstens den Vorteil hat, dass (6.44) für abgeschlossene $\mathcal{M}$ stets eine Lösung besitzt ($\{X : \mathrm{rg}X \leq k\}$ ist nämlich abgeschlossen).

Beispiel 6.16 (Rang-k-Approximation ohne Lösung)
Das folgende Problem für symmetrische Toeplitz-Matrizen

$$\left\| \begin{bmatrix} 2 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 2 \end{bmatrix} - \begin{bmatrix} x_1 & x_2 & x_3 \\ x_2 & x_1 & x_2 \\ x_3 & x_2 & x_1 \end{bmatrix} \right\|_F \stackrel{!}{=} \mathrm{Min} \qquad \mathrm{rg}X = 2 \tag{6.46}$$

hat keine Lösung, denn die nächstgelegene symmetrische Toeplitz-Matrix $X = \left[\begin{smallmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{smallmatrix}\right]$ ist eine Rang-1-Matrix, die durch beliebig kleine Störungen ε auf der ersten Ober- und Unterdiagonalen zu einer Rang-2-Matrix wird; $Q_{\mathrm{inf}} = \sqrt{4}$. Mithin ist das angegebene X aber Lösung des Max-Rang-2-Problems.

Die Schwierigkeit beim Lösen von Max-Rang-k- und Rang-k-Restriktionen mit zusätzlichen Restriktionen ($\mathcal{M} \neq \mathbb{R}^{m \times n}$) ist, dass der Rang keine stetige Funktion der Matrixelemente ist und somit klassische Lagrange-Zugänge ausscheiden. Eine kleine Ausnahme bilden quadratische Matrizen und der rangdefiziente Fall $\mathrm{rg}X \leq n-1$, der äquivalent durch $\det X = 0$ ($\det(.)$ ist stetige Funktion der Matrixelemente) beschreibbar ist. Für allgemeine Probleme bedarf es deshalb anderer Zugänge.

Das Beseitigen von Max-Rang-k-Restriktionen kann über die Rangfaktorisierung [62]

$$\mathrm{rg}\, X \leq k;\ k < \min\{m, n\} \qquad \Leftrightarrow \qquad X =: YZ; Y \in \mathbb{R}^{m \times k}, Z \in \mathbb{R}^{k \times n} \tag{6.47}$$

bzw. äquivalent über die Rangzerlegung[18] [62]

$$X = \sum_{i=1}^{k} y_i z_i^T \qquad y_i \in \mathbb{R}^m, z_i \in \mathbb{R}^n \tag{6.48}$$

geschehen. Statt über mn ist nunmehr über $(m+n)k$ Variablen zu optimieren. Alternativen zur Rangzerlegung sind Singulärwert- und Spektralzerlegung, wenn zusätzlich Restriktionen an die Singulär- oder Eigenwerte gestellt werden oder wenn für die betreffende Matrizenklasse $\mathcal{M}$ spezielle Spektraleigenschaften (z. B. symmetrische Eigenvektoren etc.) zur Reduktion der Parameteranzahl genutzt werden können. Beispielsweise formuliert

$$\begin{aligned} ||A - \sum_{i=1}^{k} \lambda_i x_i x_i^T||_F \stackrel{!}{=} \text{Min} \quad & x_i \in \mathbb{R}^n, \lambda_1 \geq \ldots \geq \lambda_k \geq 0 \\ & \sum_{i=1}^{k} \lambda_i e_j^T x_i x_i^T e_{j+s-1} = \sum_{i=1}^{k} \lambda_i e_1^T x_i x_i^T e_s; \quad \begin{array}{l} s = 1, \ldots, n-1 \\ j = 2, \ldots, n-s+1 \end{array} \end{aligned} \tag{6.49}$$

mit A als empirischer Kovarianzmatrix die Suche nach der nächstgelegenen positiv semidefiniten Max-Rang-k-Kovarianzmatrix mit Toeplitz-Struktur.

Neben den Faktorisierungen und Zerlegungen lässt sich die Max-Rang-k-Restriktion auch über eine Nullraumrestriktion erzwingen. Hat nämlich der Nullraum von X mindestens die Dimension $n-k$, so sichert das Rangtheorem für X einen Maximalrang von k /[19]

$$\text{rg}X \leq k;\ k < \min\{m,n\} \quad \Leftrightarrow \quad XU = 0_{m\times(n-k)}, U \in \mathbb{R}^{n\times(n-k)} : U^T U = I_{n-k}. \tag{6.50}$$

Nachteilig ist neben der Erhöhung der Variablenzahl die Restriktion $U^T U = I_{n-k}$, die die sog. Stiefel-Mannigfaltigkeit beschreibt. Da X und U aber separabel sind, gilt

$$\min_{\substack{XU=0 \\ U^T U = I \\ X \in \mathcal{M}}} ||A - X|| = \min_{U^T U = I} \min_{\substack{XU=0 \\ X \in \mathcal{M}}} ||A - X||. \tag{6.51}$$

Die innere Minimierung über X kann für die häufig benutzte Frobenius-Norm, die elliptischen Normen und elementweise gewichtete quadratische Normen geschlossen berechnet werden, vgl. [553] für Details. Für Optimierungsalgorithmen über die um X bereinigte, neue Zielfunktion in U mit der Stiefel-Mannigfaltigkeit-Restriktionen $U^T U = I_{n-k}$ sei weiterhin auf [178] und [135] verwiesen. Eine Anwendung für die elementweise strukturierte Max-Rang-k-Matrixapproximation wird in [429] beschrieben.

Ähnlich der vorhergehenden Variante ist

$$\text{rg}X \leq k;\ k < \min\{m,n\} \quad \Leftarrow \quad X \begin{bmatrix} B \\ -I_{n-k} \end{bmatrix} = 0_{m\times(n-k)}. \tag{6.52}$$

[18] Der Rang einer Summe von k Rang-1-Matrizen übersteigt k nicht.

[19] Statt des Nullraums von X kann auch der von X^T restringiert werden. Die Behandlung ist dann ähnlich.

Der Verzicht auf die quadratischen Restriktionen hat hier allerdings zur Folge, dass keine Äquivalenz mehr gilt. Das ist aber weniger kritisch als es scheint, denn bei den durch die Restriktion nicht erfassten Max-Rang-k-Matrizen handelt es sich um sogenannte nichtgenerische Fälle (s. z. B. nichtgenerische Fälle der TLS bei $k = n - 1$, die keine Skalierung der letzten Komponente des kleinsten Singulärwertvektors zulassen). Im praktischen Gebrauch treten solche Fälle gemeinhin nicht auf. Von Vorteil sind die nur $k(n - k)$ zusätzlichen Variablen, die ihrerseits durch immerhin $m(n - k)$ Gleichungen restringiert sind.

Die Bilinearität bei der Rangfaktorisierung und in (6.52) lässt sich algorithmisch ausnutzen. In LS-Problemen bieten sich dann ALS-Algorithmen an. Die mit der Bilinearität einhergehende Separabilität kann aber auch zur Elimination einer der Variablen genutzt werden. Zudem können nunmehr, da der Rang nicht mehr explizit als Restriktion auftaucht, auch SQP-Algorithmen benutzt werden. Ein Vergleich der beschriebenen Zugänge und ihrer algorithmischen Umsetzung findet sich in [554]. Die Algorithmenentwicklung für diese wichtige Problemklasse ist aber bei Weitem noch nicht abgeschlossen.

Als grobe Alternative für Probleme, die sich in Matrixapproximationsprobleme umformen lassen, kann natürlich immer die einfach zu programmierende zyklische Projektionsmethode eingesetzt werden, vgl. Abschn. 8.2.1. Die Projektion auf die Matrizen mit $\mathrm{rg}X \leq k$ erfordert eine Singulärwertzerlegung, die Projektion auf eine Matrix aus einem affinen Raum $\mathcal{L}$ meist einfache Mittelungen der Matrixelemente. Unter schwachen Voraussetzungen ist dann lineare Konvergenz zu einer Matrix $\tilde{X}$ aus $\mathcal{L}$ mit $\mathrm{rg}X \leq k$ gesichert, doch muss $\tilde{X}$ nicht notwendigerweise der Minimierer des Problems sein. Nichtsdestotrotz werden in der Praxis oft brauchbare Lösungen mit diesem Verfahren erzielt, nämlich bei Strukturadäquatheit des Problems und günstigen Stör-Nutzsignal-Verhältnissen.

Kapitel 7

Problemtransformationsmethoden

Bei Problemtransformationen werden äquivalente Umformungen der Restriktionen und/oder der Zielfunktion genutzt, um ein Problem in eine für ein verfügbares Softwaretool erforderliche Standardform zu bringen, es einer Systemklasse zuzuordnen, für die geschlossene Lösungen oder Lösungsaussagen existieren, oder um sich andere Vorteile zu verschaffen (bessere Topologie, bessere Konditionierung usw.).

Im Abschnitt 7.1 werden einige elementaren Umformungen nur kurz vorgestellt, da es sich hierbei um recht gut bekannte Techniken handelt. Lediglich die monotonen Transformationen werden etwas ausführlicher behandelt, liefern sie doch eine Möglichkeit, der Nichtdifferenzierbarkeitsproblematik teilweise auszuweichen.

Abschnitt 7.2 liefert einen Satz, der es erlaubt, aus den Minimierern eines bijektiv transformierten Problems auf die des Originalproblems zu schließen. Ein Beispiel dazu und eine Anmerkung zu surjektiven Transformationen vervollständigen den Abschnitt.

Das Dekompositionsprinzip ist Inhalt von Abschnitt 7.3. Dem mit der Identifikation vertrauten Ingenieur ist es von der Reduktion von MIMO-Problemen auf mehrere MISO-Probleme bekannt. Es ist als solches aber viel allgemeiner einsetzbar. Die hierzu erforderlichen Formeln, werden im Abschnitt zusammengestellt.

Der Abschnitt 7.4 stellt Minimax-Theoreme vor, die ihren Ursprung in der Spiel- und Entscheidungstheorie haben. Losgelöst von ihren ursprünglichen Anwendungen können die Theoreme angewendet werden, wenn über einen Teil der Variablen zu maximieren und über einen anderen Teil zu minimieren ist. Dann sagen die Theoreme, ob die Reihenfolge beider Optimierungen vertauschbar ist, wodurch sich lösungstechnische Vorteile ergeben können.

Im Abschnitt 7.5 werden einige Vorgehensweisen zusammengestellt, mit deren Hilfe sich parameternichtlineare Probleme in parameterlineare umformen lassen. Das ist nützlich, da sich parameterlineare Probleme nicht nur einfacher, sondern auch rekursiv gut lösen lassen.

Besonderes Augenmerk wird hier zum einen auf die Notwendigkeit gelegt, bei der Pseudoparametermethode Restriktionen zum Zwecke der Eindeutigkeit und die Rücksubstituierbarkeit einzuführen, und zum anderen auf die Notwendigkeit, bei der Ordinatentransformationsmethode eine Wichtung zu nutzen.

Eine wiederentdeckte Methode, die die adaptive Regelungstechnik aber seit ihren Anfängen prägt, stellt die Transformation des Optimierungsproblems in ein Ruhelagenproblem von Differenzialgleichungen dar, s. Abschnitt 7.6. Darüber hinaus findet sich diese Idee auch in der sog. Speed-Gradienten-Methode (Entwurfsverfahren für nichtlineare Regelungen) wieder, die seit den 1970er Jahren von russischen Wissenschaftlern ständig weiterentwickelt wird (Stabilitätsnachweise, andere Systemklassen) [198].

Abschnitt 7.7 zeigt die Umformulierung der restringierten (linearen) Least Squares in ein orthogonales Projektionsproblem. Diese Formulierung ist für das rekursive Lösen von Vorteil.

Den Schwerpunkt dieses Kapitels bildet Abschnitt 7.8. Er widmet sich der semidefiniten Optimierung (engl. semidefinite programming; SDP), worunter das Lösen eines Problems mit linearer Zielfunktion unter semidefiniten Restriktionen an Matrizen oder affine Matrixfunktionen verstanden wird. Hierbei handelt es sich um eine große Klasse konvexer Probleme, für die in den letzten Jahren sehr leistungsstarke Algorithmen entwickelt wurden. Ausgehend von den Darstellungsformen und den Vorteilen sind Hinweise zur Transformation konvexer Probleme in SDP-Formulierungen und Angaben zu deren Vereinfachung das zentrale Thema dieses Abschnitts. Ergänzt wird die Thematik durch Betrachtungen zur Unterklasse der konischen Probleme zweiter Ordnung (engl. second order cone programming; SOCP), die sich noch besser lösen lassen.

7.1 Elementare Problemtransformationen

Zu den elementaren Problemtransformationen zählen

- Umformungen wie in Tabelle 7.1
- das Reellmachen komplexer Probleme durch neue Variablen $y = \Re e\, x$, $z = \Im m\, x$, vgl.

$$||Ax - b||_2 \stackrel{!}{=} \text{Min}; x \in \mathbb{C}^n \Leftrightarrow \left\| \begin{bmatrix} \Re e\, A & -\Im m\, A \\ \Im m\, A & \Re e\, A \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} - \begin{bmatrix} \Re e\, b \\ \Im m\, b \end{bmatrix} \right\|_2 \stackrel{!}{=} \text{Min}, \quad (7.1)$$

- monotone Transformationen.

	Bedingung	Umformung
1	$g(x) \geq a$	$-g(x) \leq -a$
2	$a \leq x \leq b$	$a - x \leq 0$ $x - b \leq 0$
3	$x \leq \min\{a, b\}$	$x \leq a$ $x \leq b$
4	$\|g(x)\| \leq a$	$g(x) \leq a$ $-a \leq g(x)$
5	$g(x) = a$	$g(x) \leq a$ $-g(x) \leq -a$
6	$g(x) \geq 0; g : \mathbb{R}^n \to \mathbb{R}$	$\min\{g(x), 0\} = 0$
7	$x_1 : x_2 : x_3 = ab : b : 1$	$\begin{bmatrix} 1 & -a & 0 \\ 0 & 1 & -b \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$

Tabelle 7.1: Äquivalente Umformungen von Restriktionen

Da elementare Umformungen vielfach offensichtlich sind, fehlen entsprechende Zusammenstellungen in der Literatur. Tabelle 7.1 schließt diese Lücke. Das Umkehren einer Ungleichungsrelation (Umformung 1), das Überführen einer Doppelungleichung in zwei Einzelungleichungen (Umformung 2), das Auflösen eines Minimums (Umformung 3) oder das Auflösen eines Betrages (Umformung 4) sind leicht überprüfbar. Weitere artverwandte Umformungen ergeben sich aus den Rechenregeln für Ungleichungen.Eine Anwendung der zu Umformung 3 analogen Maximumsauflösung zeigt Beispiel 7.10. Die Umformung 5 wird für Algorithmen in der Regel nicht verwendet, da Gleichungsrestriktionen besser zu handhaben sind. Sie findet aber in Beweisen Anwendung, in denen aus der gleichzeitigen entgegengesetzten Gültigkeit von Ungleichungen auf Gleichheit geschlossen werden kann. Umformungen vom Typ 6 werden für die Konstruktion von Straftermen verwendet, vgl. (3.2). Verhältnisgleichungen sollten durch Überkreuzmultiplizieren gemäß Umformung 7 vereinfacht werden, um das komplizierte Rechnen mit rationalen Funktionen zu vermeiden. Nachfolgend wird auf die Umformungen durch monotone Transformationen näher eingegangen.

Satz 7.1 (Monotone Transformation der Zielfunktion, [422])
Ist $\tilde{f}(x) = \eta(f(x))$ eine stetige reellwertige Funktion, η monoton wachsend auf dem Bild von f und hat f ein globales Minimum (Maximum) in x_{opt}, dann hat auch $\tilde{f}$ an dieser Stelle ein solches. Wächst η sogar streng auf dem Bild von f, dann hat $\tilde{f}$ genau dann ein globales Minimum (Maximum) an der Stelle x_{opt}, wenn f dort ein solches hat. Diese Aussagen gelten gleichermaßen für freie und restringierte Probleme, da η die zulässige Menge nicht ändert.

Anmerkung 7.1 Transformationen für Restriktionen müssen so gestaltet sein, dass sich die zulässige Menge nicht ändert. Für jede einzelne Gleichungs- und Ungleichungsrestriktion kann eine spezifische Transformation gewählt werden.

Ist $g_i(x) \leq 0$ und $\eta : \mathbb{R} \to \mathbb{R}$ eine Transformation mit $\eta(y) \leq 0 \Leftrightarrow y \leq 0$ für alle $y \in \mathcal{WB}(g_i)$, dann gilt $g_i(x) \leq 0 \Leftrightarrow \tilde{g}_i(x) = \eta(g_i(x)) \leq 0$.

Ist $h_j(x) = 0$ und $\eta : \mathbb{R} \to \mathbb{R}$ eine Transformation mit $\eta(y) = 0 \Leftrightarrow y = 0$ für alle $y \in \mathcal{WB}(h_j)$, dann gilt $h_j(x) = 0 \Leftrightarrow \tilde{h}_j(x) = \eta(h_j(x)) = 0$.

Aus dem Satz 7.1 lassen sich folgende Beziehungen herleiten:

1. Addition einer Konstanten oder Multiplikation mit einem positiven Faktor

$$\inf_{x\in\mathcal{X}}(f(x)+c) = \inf_{x\in\mathcal{X}} f(x) + c \qquad \sup_{x\in\mathcal{X}}(f(x)+c) = \sup_{x\in\mathcal{X}} f(x) + c \tag{7.2a}$$

$$\inf_{x\in\mathcal{X}}(\gamma f(x)) = \gamma \inf_{x\in\mathcal{X}} f(x) \qquad \sup_{x\in\mathcal{X}}(\gamma f(x)) = \gamma \sup_{x\in\mathcal{X}} f(x) \tag{7.2b}$$

2. Quadrieren vor dem Minimieren der euklidischen Norm
   $||Ax-b||_2^2$ ist für reelle Variablen überall F-differenzierbar, während $||Ax-b||_2$ an der Stelle $Ax = b$ nicht F-differenzierbar ist. Zudem hat ein Algorithmus bei $||Ax - b||_2$ im Fall kleiner Residuen $\text{abs}(Ax_{\text{opt}} - b) \approx 0$ mit einer „Ecke“ vergleichbar der Betragsfunktion zu kämpfen, während diese bei $||Ax-b||_2^2$ verschwindet.
   Allgemeiner: Transformiere Zielfunktionen so, dass sie LS-ähnlich werden.

3. Wurzelziehen bei Produktfunktionen
   $f(x_1, x_2) = -x_1^\alpha x_2^\beta$ in $\mathbb{R}^2_>$ ist mit $\alpha > 0$, $\beta > 0$ für $\alpha + \beta \leq 1$ konvex und für $\alpha + \beta > 1$ quasikonvex. Ein Wurzelziehen mit hinreichend hohem Wurzelexponenten reduziert die Summe der Exponenten auf bzw. unter Eins und lässt eine konvexe Funktion entstehen.
   Beispiel: $\text{argmin}_{\mathcal{X}}(-x_1x_2) = -\text{argmax}_{\mathcal{X}}(x_1x_2) = -\text{argmax}_{\mathcal{X}}\sqrt{x_1x_2} = \text{argmin}_{\mathcal{X}} -\sqrt{x_1x_2}$

4. Logarithmieren von Likelihood-Funktionen
   Die Likelihood-Funktion für Regressionsmodelle mit normalverteilter Störung führt auf Exponentialterme, die durch Logarithmieren beseitigt werden können. Statt über der Likelihood-Funktion wird dann über der Log-Likelihood-Funktion maximiert.

5. Umformung mit monoton fallender Transformation
   Bei monoton fallenden η werden die Minimierer zu Maximierern und umgekehrt. Die zwei Standardumformungen hierfür sind

$$\inf_{x\in\mathcal{X}}(-f(x)) = -\sup_{x\in\mathcal{X}} f(x) \qquad \text{und} \qquad \left(\inf_{x\in\mathcal{X}} f(x)\right)^{-1} = \sup_{x\in\mathcal{X}} \frac{1}{f(x)};\ f(x) > 0. \tag{7.3}$$

Das Beispiel 7.1 zeigt, wie monotone Transformationen und Umformungen geschickt eingesetzt werden können, um eine Maximierung ohne die Methode der Lagrange-Multiplikatoren zu lösen, und es umgeht die Problematik der eingeschränkten Differenzierbarkeit.

Beispiel 7.1 (Anwendung monotoner Transformationen)
Maximiere das Verhältnis der Summe der Kathetenlängen zur Hypothenusenlänge in einem rechtwinkligen Dreieck im $\mathbb{C}^n$.

$$
\begin{aligned}
Q = \frac{||x||_2 + ||y||_2}{||x+y||_2} &\stackrel{!}{=} \text{Max} && x \perp y \\
\left(\frac{||x||_2 + ||y||_2}{||x+y||_2}\right)^2 &\stackrel{!}{=} \text{Max} && \text{Quadrieren ist monoton, da } Q > 0 \\
\frac{||x||_2^2 + 2||x||_2||y||_2 + ||y||_2^2}{||x||_2^2 + ||y||_2^2} &\stackrel{!}{=} \text{Max} && \text{Pythagoras } ||x+y||_2^2 = ||x||_2^2 + ||y||_2^2 \\
1 + 2\frac{||x||_2||y||_2}{||x||_2^2 + ||y||_2^2} &\stackrel{!}{=} \text{Max} && \\
\frac{||x||_2||y||_2}{||x||_2^2 + ||y||_2^2} &\stackrel{!}{=} \text{Max} && \text{Weglassen von Konstante und Faktor} \\
\frac{||x||_2^2 + ||y||_2^2}{||x||_2||y||_2} &\stackrel{!}{=} \text{Min} && \text{Inversion dreht Optimierungsziel} \\
\frac{||x||_2}{||y||_2} + \frac{||y||_2}{||x||_2} &\stackrel{!}{=} \text{Min} && \\
\gamma + 1/\gamma &\stackrel{!}{=} \text{Min} && \gamma = ||x||_2/||y||_2 \\
\Rightarrow \gamma = 1 &\Rightarrow ||x||_2 = ||y||_2 &&
\end{aligned}
$$

Damit ist $Q_{\max} = \dfrac{2||x||_2}{\sqrt{||x+y||_2^2}} = \dfrac{2||x||_2}{\sqrt{||x||^2 + ||y||_2^2}} = \dfrac{2||x||_2}{\sqrt{2||x||_2^2}} = \sqrt{2}$. Bezeichne nun $P^{\perp}_{\mathcal{R}(x)} = \left(I_n - \dfrac{xx^H}{x^Hx}\right)$ den orthogonalen Projektor in das orthogonale Komplement von x, dann ist die Menge aller Optimallösungen ablesbar

$$
\left\{(x,y) \in \mathbb{C}^n \times \mathbb{C}^n : y = \frac{||x||_2}{||P^{\perp}_{\mathcal{R}(x)} z||_2} P^{\perp}_{\mathcal{R}(x)} z; z \in \mathbb{C}^n \backslash \{\zeta x; \zeta \in \mathbb{C}\}\right\}. \tag{7.4}
$$

Das Beispiel 7.2 nutzt den Einsatz monotoner Transformationen zur numerischen Verbesserung der Problemformulierung.

Beispiel 7.2 (Problemskalierung)
Skalieren der Zielfunktion und der Restriktionen lässt die Menge der Minimierer unverändert:

$$
\begin{array}{rcl}
f(x) \stackrel{!}{=} \text{Min};\ g(x) \le 0_p & & \alpha f(x) \stackrel{!}{=} \text{Min};\ \beta_i g_i(x) \le 0; \beta_i \in \mathbb{R}^{>}, \alpha \in \mathbb{R}^{>} \\
h(x) = 0_m & \Leftrightarrow & \gamma_j h_j(x) = 0_m; \gamma_j \in \mathbb{R}\backslash 0
\end{array} \tag{7.5}
$$

Die Faktoren sollten aus numerischer Sicht (Abbruchkriterien, Genauigkeit, Penalty-Terme) so gewählt werden, dass für typische Werte x die Funktionswerte $g_i(x), h_j(x)$ in etwa gleich groß sind. Ein anderer Aspekt ist der Rechenaufwand. So ist $\text{Proj}_{\mathcal{C}}(a) = a - \max\{b^T x - c, 0\}\frac{b}{||b||_2^2}$ die Projektion auf $\mathcal{C} = \{x \in \mathbb{R}^n : b^T x \leq c\}$, die sich im Fall eines Einheitsvektors b auf $\text{Proj}_{\mathcal{C}}(a) = a - \max\{b^T x - c, 0\}b$ vereinfacht. Eine einmalige äquivalente Änderung der Restriktion zu $\mathcal{C} = \{x \in \mathbb{R}^n : b^T x/||b||_2 \leq c/||b||_2\}$ erzeugt einen solchen Einheitsvektor, womit gerade in zyklischen Projektionsalgorithmen Rechenaufwand gespart werden kann.

7.2 Bijektive Variablentransformation

Variablentransformationen sind ein probates Mittel für Problemumformungen und -vereinfachungen. Bekannt sind sie aus dem Lösen biquadratischer Gleichungen $a_2x^4+a_1x^2+a_0 = 0$, wo $y := x^2$ eine quadratische Gleichung entstehen lässt. Auch die berühmte Weierstraß-Substitution $y = \tan(\frac{x}{2})$, mit deren Hilfe sich einige trigonometrische Terme dank

$$\sin x = \frac{2y}{1+y^2} \qquad \text{und} \qquad \cos x = \frac{1-y^2}{1+y^2} \tag{7.6}$$

in rationale umformen lassen, zählt hierzu. Den Vorteil bijektiver Variablentransformationen für die Lösung freier und restringierter Probleme kennzeichnet der folgende Satz.

Satz 7.2 (Bijektive Variablentransformation, [422])
Ist $f : \mathcal{X} \to \mathbb{R}$ und g eine bijektive Funktion mit $g : \mathcal{Y} \to \mathcal{X}$ und $\tilde{f}(y) = f(g(y))$, dann gilt: Hat f sein Minimum bei x_{opt}, so nimmt $h(y)$ sein Minimum bei $y_{\text{opt}} = g^{-1}(x_{\text{opt}})$ an. Ist y_{opt} Minimierer von $\tilde{f}(y)$, dann ist $x_{\text{opt}} = g(y_{\text{opt}})$ Minimierer von $f(x)$. x_{opt} ist genau dann eindeutig, wenn dies auch y_{opt} ist.

Anmerkung 7.2 Selbstverständlich gilt Satz 7.2 analog für Maximierungsprobleme, und zwar unabhängig davon, ob sie frei oder restringiert sind.

Beispiele für bijektive Variablentransformationen sind orthogonale Transformationen in Verbindung mit QR- oder SVD-Faktorisierungen. Aber auch im Allgemeinen nicht-bijektive Abbildungen können auf $\mathcal{X}$ bijektiv sein. So beschreibt $C^{\perp} \in \mathbb{R}_{n-p}^{n\times(n-p)}$ eine nicht-surjektive Abbildung von $\mathbb{R}^{n-p}$ nach $\mathbb{R}^n$, die aber für $f : \mathbb{R}^n \to \mathbb{R}$ mit der linearen Restriktion $Cx = 0$, d. h. mit $x = g(y) = C^{\perp}y$, bijektiv ist, denn in diesem Fall ist $\mathcal{X} = \mathcal{N}(C) = \mathcal{R}(C^{\perp})$.
Eine weitere Anwendung liefert die Momentenmethode[1], bei der die Momente $\hat{\mu}_k \stackrel{\text{def}}{=} \sum_{i=1}^{N} x_i^k$ zur Bestimmung der Verteilungsparameter herangezogen werden.

[1] Eng verwandt mit der Momentenmethode ist die Quantilmethode, die die in aller Regel deutlich robusteren Quantilschätzer zur Bestimmung der Verteilungsparameter nutzt und die auch angewendet wird, wenn wie bei der Cauchy-Verteilung einzelne Momente nicht existieren.

Beispiel 7.3 (Schätzung der Parameter einer Gamma-Verteilung)
Für N gammaverteilte Stichprobenwerte seien $b > 0, p > 0$ für $f(x) = b^p \mathrm{e}^{-bx} x^{p-1}/\Gamma(p), x > 0$ zu bestimmen. Aus $\mu_1 = \frac{p}{q}$ und $\mu_2 = \frac{p(p+1)}{b^2}$ folgt $\hat{b} = \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2}$ und $\hat{p} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}$.

Den schönen Eigenschaften der bijektiven Variablentransformation steht ein beschränktes Einsatzgebiet gegenüber, weshalb mit Abschwächungen gearbeitet wird. So ist die Matrixexponentialfunktion für $n \geq 2$ nicht global injektiv (damit auch nicht global bijektiv), dafür aber lokal injektiv. Eine weitere Abschwächung besteht darin, die Bijektivität auf eine Teilmenge zu reduzieren, bestimmte Punkte oder Mengen wegzulassen (z. B. bei der Cayley-Transformation [618]) und diese separat zu behandeln. Wenn die weggelassenen Mengen mager sind, ist dies häufig durchaus legitim, allerdings können numerische Probleme in der Nähe derartiger Mengen auftreten.

Anmerkung 7.3 Häufig führt kein Weg an surjektiven Variablentransformationen vorbei. Dann reduziert sich Satz 7.2 auf die Tatsache, dass eine Lösung x_{opt} die Existenz mindestens einer Lösung y_{opt} mit $x_{\mathrm{opt}} = g(y_{\mathrm{opt}})$ impliziert und dass, wenn y_{opt} ein globaler Minimierer ist, auch $x_{\mathrm{opt}} = g(y_{\mathrm{opt}})$ ein eben solcher ist. Die Mehrdeutigkeit impliziert hierbei aber mehrere globale Minimierer y_{opt} und damit potenziell die Existenz stationärer Punkte, die nicht mit dem globalen Minimierer x_{opt} korrespondieren, vgl. Beispiel 5.2.

7.3 Dekomposition

Dekomposition bezeichnet eine Strategie, die ein größeres Problem in mehrere kleinere (Herabsetzen der Problemdimension) zerlegt und/oder die kompliziert beschriebene Mengen (Strukturen) in einfachere aufspaltet. Eine typische Anwendung stellt die Dekomposition von LS-Problemen dar. Hierfür gilt:

$$\min_{x_1,\ldots,x_m} \sum_{i=1}^{m} w_i^2 ||A_i x_i - b_i||_2^2 = \sum_{i=1}^{m} w_i^2 \min_{x_i} ||A_i x_i - b_i||_2^2. \tag{7.7}$$

Sie wird bei der Identifikation von MIMO-LTI-Modellen über MISO-Teilmodelle genutzt. Obwohl dabei meist P-kanonische Strukturen herangezogen werden [356], gelingt eine Dekomposition auch bei V-kanonischen MIMO-Modellen, wenn die einwirkenden Ausgänge als Eingänge angesehen werden. In [255] wird hierzu für eine Mehrzonenofen-Anwendung beschrieben.

Die meisten Dekompositionszugänge beruhen auf den folgenden Regeln:

$$\inf_{x\in\mathcal{X}_1\cup\mathcal{X}_2} f(x) = \min\{\inf_{x\in\mathcal{X}_1} f(x), \inf_{x\in\mathcal{X}_2} f(x)\} \tag{7.8a}$$

$$\sup_{x\in\mathcal{X}_1\cup\mathcal{X}_2} f(x) = \max\{\sup_{x\in\mathcal{X}_1} f(x), \sup_{x\in\mathcal{X}_2} f(x)\} \tag{7.8b}$$

$$\inf_{x_1\in\mathcal{X}_1,x_2\in\mathcal{X}_2}(f(x_1)+g(x_2)) = \inf_{x_1\in\mathcal{X}_1} f(x_1) + \inf_{x_2\in\mathcal{X}_2} g(x_2) \tag{7.8c}$$

$$\sup_{x_1\in\mathcal{X}_1,x_2\in\mathcal{X}_2}(f(x_1)+g(x_2)) = \sup_{x_1\in\mathcal{X}_1} f(x_1) + \sup_{x_2\in\mathcal{X}_2} g(x_2) \tag{7.8d}$$

$$\inf_{x_1\in\mathcal{X}_1,x_2\in\mathcal{X}_2}(f(x_1)-g(x_2)) = \inf_{x_1\in\mathcal{X}_1} f(x_1) - \sup_{x_2\in\mathcal{X}_2} g(x_2) \tag{7.8e}$$

$$\sup_{x_1\in\mathcal{X}_1,x_2\in\mathcal{X}_2}(f(x_1)-g(x_2)) = \sup_{x_1\in\mathcal{X}_1} f(x_1) - \inf_{x_2\in\mathcal{X}_2} g(x_2) \tag{7.8f}$$

$$\inf_{x_1\in\mathcal{X}_1,x_2\in\mathcal{X}_2}(f(x_1)\cdot g(x_2)) = \inf_{x_1\in\mathcal{X}_1} f(x_1)\cdot \inf_{x_2\in\mathcal{X}_2} g(x_2); \qquad f(x_1)\geq 0, g(x_2)\geq 0 \tag{7.8g}$$

$$\sup_{x_1\in\mathcal{X}_1,x_2\in\mathcal{X}_2}(f(x_1)\cdot g(x_2)) = \sup_{x_1\in\mathcal{X}_1} f(x_1)\cdot \sup_{x_2\in\mathcal{X}_2} g(x_2); \qquad f(x_1)\geq 0, g(x_2)\geq 0 \tag{7.8h}$$

$$\inf_{x_1\in\mathcal{X}_1,x_2\in\mathcal{X}_2} g(x_1,x_2) = \inf_{x_1\in\mathcal{X}_1}\{\inf_{x_2\in\mathcal{X}_2} g(x_1,x_2)\} = \inf_{x_2\in\mathcal{X}_2}\{\inf_{x_1\in\mathcal{X}_1} g(x_1,x_2)\} \tag{7.8i}$$

$$\sup_{x_1\in\mathcal{X}_1,x_2\in\mathcal{X}_2} g(x_1,x_2) = \sup_{x_1\in\mathcal{X}_1}\{\sup_{x_2\in\mathcal{X}_2} g(x_1,x_2)\} = \sup_{x_2\in\mathcal{X}_2}\{\sup_{x_1\in\mathcal{X}_1} g(x_1,x_2)\} \tag{7.8j}$$

Regel (7.8a) sagt grob gesprochen, dass das Minimum einer Funktion über der Vereinigung zweier Teilmengen, ermittelt werden kann, indem die Minimierung über jede Teilmenge ausgeführt wird und im Ergebnis das kleinste der beiden Minima zu nehmen ist (Regel (7.8b) analog). Da die Vereinigung zweier konvexer Mengen nicht konvex ist, ist selbst bei konvexem f ein nichtkonvexes Problem zu lösen. Vielfach einfacher ist dann die Lösung zweier leichterer konvexer Probleme. Die Regeln (7.8c) bis (7.8h) beziehen sich auf separable Variablen und aufspaltbare Zielfunktionen. Sie zeigen, wie die Optimierung über die Einzelvariablen die Komplexität senkt. Die Regeln (7.8i) und (7.8j) beziehen sich ebenfalls auf separable Variablen und zeigen, dass durch das Nacheinander-Ausführen zweier Optimierungen Vereinfachungen erreichbar sind. Dieses Prinzip wird bei separablen Quadratmittelproblemen in Abschnitt 5.2.8, genutzt oder allgemeiner bei der Eliminationsmethode mittels Optimalitätsbedingung in Abschnitt 5.2.7.

7.4 Minimax-Theoreme

In einigen Problemen (Maximum-Entropie-Identifikation, Min-Max-optimale-Prädiktion, robuste Filterung [564]) ist über einen Teil der Variablen zu maximieren und über einen anderen zu minimieren, z. B. Maximieren der mittleren Transinformation bezüglich des Nutzsignals und Minimieren bezüglich der Störung [642]. Damit stellt sich die Frage, wann die Reihenfolge von Maximieren und Minimieren getauscht werden darf, wann also

$$\sup_{x\in\mathcal{X}}\inf_{y\in\mathcal{Y}} f(x,y) = \inf_{y\in\mathcal{Y}}\sup_{x\in\mathcal{X}} f(x,y) \tag{7.9}$$

gilt. Solche Aussagen liefern die Minimax-Theoreme.

Satz 7.3 (Kakutani-Minimax-Theorem, [329])
Sind $\mathcal{X} \subset \mathbb{R}^n, \mathcal{Y} \subset \mathbb{R}^m$ konvexe Körper (nichtleere, kompakte, konvexe Mengen) in normierten Vektorräumen und ist f stetig und für jedes feste y in x konkav und für jedes feste x in y konvex, dann gilt

$$\max_{x\in\mathcal{X}}\min_{y\in\mathcal{Y}} f(x,y) = \min_{y\in\mathcal{Y}}\max_{x\in\mathcal{X}} f(x,y). \tag{7.10}$$

Weiterentwicklungen zielen auf das Aufweichen der Kompaktheitsforderung, der Konvexitätsforderung an beide Mengen, der Stetigkeit und der Konkav-konvex-Forderung. Einen Überblick geben Frenk, Kassay und Kolumbán [200] sowie Tuy [606], der selbst eine neue Version liefert. Für die Anwendung ist insbesondere eine aus dem Tuy-Theorem abgeleitete Verfeinerung des Minimax-Theorems von Sion [569] nützlich.

Satz 7.4 (Sion-Tuy-Minimax-Theorem, [606])
Es seien $\mathcal{X} \subset \mathbb{R}^n, \mathcal{Y} \subset \mathbb{R}^m$ abgeschlossene konvexe Teilmengen und f quasikonkav in x und quasikonvex in y.
Ist $f(x,y)$ oberhalbstetig in x und halbstetig in y in jedem Liniensegment, dann gilt

$$\max_{x\in\mathcal{X}}\inf_{y\in\mathcal{Y}} f(x,y) = \inf_{y\in\mathcal{Y}}\sup_{x\in\mathcal{X}} f(x,y) \qquad \text{für } \mathcal{X} \text{ kompakt.} \tag{7.11}$$

Ist $f(x,y)$ halbstetig in x in jedem Liniensegment und unterhalbstetig in y, dann gilt

$$\sup_{x\in\mathcal{X}}\inf_{y\in\mathcal{Y}} f(x,y) = \min_{y\in\mathcal{Y}}\sup_{x\in\mathcal{X}} f(x,y) \qquad \text{für } \mathcal{Y} \text{ kompakt.} \tag{7.12}$$

Anmerkung 7.4 Die Minimax-Theoreme stehen in Beziehung zu den bekannten Fixpunktsätzen [329], [200] und zur starken Dualität (Abschn. 8.1.5). Letzteres ist anschaulich, sagen sie doch, wann in der von-Neumann-Ungleichung (8.17) keine Dualitätslücke auftritt.

7.5 Umformung in parameterlineare Modelle

Ausgleichsprobleme lassen sich für parameterlineare Modelle $f(w,t_k) = g^T(w,t_k)\theta(t_k)$ elegant mit LS-Lösungsverfahren bearbeiten. Tritt eine additive Störung zum rechten Term auf, bleibt das entstehende Regressionsproblem ebenfalls einfach, da es linear bezüglich der Störung ist. Für normalverteilte Störungen empfiehlt sich dann der Gauß-Markov-Schätzer. Bei Regressionsproblemen, die zwar linear in den Parametern, aber nichtlinear bezüglich der Störungen sind und solchen, bei denen die Störungen zwar linear, aber über Gleichungen

miteinander verknüpft sind (z. B. lineare Differenzengleichungsmodelle), hilft die Parameterlinearität zu einer schnellen approximativen Lösung. Allerdings bleibt offen, wie weit in diesen Fällen die LS-Lösung von einer ML-Schätzung entfernt liegt. Ungeachtet dessen sprechen einfache und numerisch schnelle Algorithmen gerade auch bei μ-Controller-Anwendungen für die Verwendung parameterlinearer Modelle. Mit einigen Umformungen kann zudem manches parameternichtlineare Problem durch Transformation in ein parameterlineares überführt werden. In Tabelle 7.2 stellt der Autor von ihm erfolgreich eingesetzte Methoden bei diversen Modellbildungsaufgaben, Standardmethoden und Zugänge aus der Spezialliteratur zusammen und zeigt deren Basisideen beispielhaft. Nachfolgend wird jedoch nur auf die Pseudoparametermethode und die Ordinatentransformation näher eingegangen werden.

Definition 7.1 (Pseudoparameter)
Werden Modellparameter zum Zwecke einer einfacheren Berechnung durch neue Parameter ersetzt, wobei deren Anzahl größer als die des Modells ist, dann wird von Pseudoparametern gesprochen. Der Begriff wird mitunter auch bei Reparametrisierungen verwendet, bei denen die Parameteranzahl gleich bleibt, wenn der Bedeutungszusammenhang zu den physikalischen Parametern sichtlich verloren geht. Synonym werden die neuen (rein mathematischen) Parameter auch Ersatzmodellparameter genannt.

Anmerkung 7.5 Pseudoparameter erfordern gegebenenfalls Restriktionen, die Ausführbarkeit und Widerspruchsfreiheit der Rücksubstitution sichern.
Im Punkt 2 von Tabelle 7.2 garantiert die Restriktion die Widerspruchsfreiheit für die Rücksubstituierbarkeit. Beim Kugelfit über das Modell $(2x_i^T)c + r^2 - ||c||_2^2 \cong ||x_i||_2^2$ mit c als Mittelpunkt und r als Radius und den Pseudoparametern $\theta_{1:3} = c$ und $\theta_4 = r^2 - ||c||_2^2$ sichert die Restriktion $\theta_{1:3}^T\theta_{1:3} + \theta_4 \geq 0$ die Existenz eines reellen r.

Beispiel 7.4 (Pseudoparameter)
Wird beim Ellipsenfit zusätzlich das Vorwissen eingearbeitet, dass das Zentrum der Ellipse x_c auf einer bekannten Geraden $y = p + \gamma q$ liegt, dann entsteht nach Elimination von x_c und der Ersetzung $x := x - p$ die Gleichung $f(\gamma, A, g) = (x - \gamma q)^T A(x - \gamma q) + g = 0$. Sie ist quadratisch in γ, weshalb γ im Gleichungsfehlerquadrat $f^2(\gamma, A, g)$ in vierter Potenz auftritt. Wird indes in $f(\gamma, A, g) = x^T Ax - 2\gamma q^T Ax + \gamma^2 q^T Aq + g = 0$ der Pseudoparameter $h = \gamma^2 q^T Aa + g$ eingeführt, dann erscheint γ in $\tilde{f}(\gamma, A, h) = x^T Ax - 2\gamma q^T Ax + h = 0$ nur noch linear und in $\tilde{f}^2(\gamma, A, h)$ quadratisch. Da g hier keinen Einschränkungen unterliegt, sind auch für h keine erforderlich. Mithin vereinfacht sich die Situation dank des Pseudoparameters h.

---

[2] Eine Funktion heißt q-homogen, wenn $f(ax) = a^q f(x)$ gilt.

[3] Regressionen mit Modellen $y = au^b$ als auch mit den parameterlinearisierten Versionen über $\ln y = \ln a + b \ln u$ heißen potenzielle Regressionen. Analog werden $y = ae^{bu}$ bzw. $y = am^u$ als auch $\ln y = \ln a + bu$ bzw. $\ln y = \ln a + \ln m \cdot u$ als exponentielle Regressionen bezeichnet.

Prinzip	parameternichtlinear	parameterlinear
Reparametrisierung	$y = abu_1 + b^{1/3}u_2$	$y \cong \theta_1 u_1 + \theta_2 u_2$ $\theta_1 = ab; \theta_2 = b^{1/3}$
Einführen von Pseudoparametern	$y = au_1 + \frac{1}{b}u_2 + \frac{a}{b}u_3$	$y \cong au_1 + \theta_1 u_2 + \theta_2 u_3$ Restriktion: $a\theta_1 = \theta_2$
Ordinatentransf. $\phi(y) = f^{-1}(y)$ $\phi(y) = \ln y$ /3	$y = b\, f\left(g_n(u) + \sum_{i=0}^{n-1} a_i g_i(u)\right)$ f ist q-homogen[2]	$f^{-1}(y) \cong cg_n(u) + \sum_{i=0}^{n-1} b_i g_i(u)$; $b_i = a_i b^{1/q}, c = b^{1/q}$
	$y = a[f(u)]^b \mathrm{e}^{cg(u)}$	$\ln y \cong d + b \ln f(u) + cg(u); d > 0$ $d = \ln a$
$\phi(y) = 1/y$	$y = \frac{f(u)}{\sum_{i=0}^{n} a_i g_i(u)}$	$\frac{1}{y} \cong \sum_{i=0}^{n} a_i \frac{g_i(u)}{f(u)}$
$\phi(y) = \psi(u) \cdot y$	$y = \frac{\sum_{i=0}^{m} b_i f_i(u)}{g_n(u) + \sum_{i=0}^{n-1} a_i g_i(u)}$	$g_n(u)y \cong \sum_{i=0}^{m} b_i f_i(u) - \sum_{i=0}^{n-1} a_i g_i(u) y$
Approximieren	$a\dot{y}(t) + y(t) = ku(t - T_t)$	$y(t) \cong -a\dot{y}(t) + Ku(t) - c\dot{u}(t)$ $T_t < T_A, c = KT_t$
Darstellen als Lösung einer Differenzengl. Prony-Prinzip	$y(t) = \sum_{i=1}^{n} \alpha_i \exp(\beta_i t) \sin(\omega_i t)$ $t = kT_A$	$y[k+n] \cong -a_{n+1} y[k+n-1] \ldots$ $-a_1 y[k+1] - a_0 y[k]$
	$y(t) = \frac{\gamma}{1 + \beta \exp(-\alpha t)}$ logistische Funktion	$\frac{1}{y[k]} \cong a_0 + a_1 \frac{1}{y[k-1]}$ $a_0 = (1 - a_1)/\gamma, a_1 = \exp(-\alpha)$ Berechnung von β über zweite Schätzung mit $\hat{\alpha}, \hat{\gamma}$
Darstellen als Lösung einer Differenzialgl.	$y(t) = a \int_0^t \exp(bu^2) \mathrm{d}u$	$\ln \dot{y} \cong c + bu^2$ $c = \ln a$
Lie-symmetrische Transformation [141]	$\dot{y}(t) = Ku(t)[y(t) - y_{eq}]^p$ oder allgemeiner $\dot{y} = f(t)g(y)$	$\frac{\mathrm{d}}{\mathrm{d}t}(yz) \cong c_1 \dot{y} + c_2 uy + c_3 u$ $c_1 = -\frac{z_0}{K}, c_2 = \frac{2-p}{1-p}, c_3 = \frac{y_{eq}}{1-p}$ $z(t) = \int_0^t u(\tau) \mathrm{d}\tau$ z_0 Integrationskonstante
Nutzung prädiktiver Signale	$y = a \exp(bu) + cu^2$	$y \cong a\hat{z}_k + cu^2;\ \hat{z}_k = \exp(\hat{b}_{k-1} u_k)$ $bu \cong \ln\left(\frac{y - \hat{c}_k u^2}{\hat{a}_k}\right)$
Zweistufiges Vorgehen	$y[k] = G(q, \theta)u[k]$	über Gewichtsfolge [104] über Zustandsfolge [490]

Tabelle 7.2: Methoden zur Umformung in parameterlineare Modelle

Für einige einfache Modelle, vgl. Punkt 3 in Tabelle 7.2 ist eine Ordinatentransformation ein probates Mittel zur Umformung in ein parameterlineares Modell, s. auch die weiterführende Literatur Myers [470], Glenberg [225], Neter [477]. Das nachfolgende Beispiel wendet ebenfalls diese Technik an, ist aber insofern zusätzlich interessant, da es über den Zwischenschritt der expliziten Lösung von nichtlinearen Differenzialgleichungen einen Weg zu deren Parameteridentifikation zeigt.

Beispiel 7.5 (Ordinatentransformation für Wachstumsprozesse)
Tabelle 7.3 stellt für das lineare (1), exponentielle (2), monomolekulare (3), logistische (4) und Gompertz-Modell (5) die Zusammenhänge dar. Die auf der rechten Gleichungsseite stehenden ln-Terme über y_0 werden schlussendlich als Pseudoparameter aufgefasst. Für eine detaillierte Analyse zu den Modellen (4) und (5), insbesondere zu deren qualitativem dynamischen Verhalten bei bestimmten Parameterwerten und damit zu einer guten Beurteilungsmöglichkeit der Parameterschätzungen, sei auf [481] verwiesen.

	Differenzialgl.	Lösung	Parameterlinearisierung
1	$\dot{y} = r$	$y(t) = y_0 + rt$	$y(t) = y_0 + rt$
2	$\dot{y} = ry$	$y(t) = y_0 \mathrm{e}^{rt}$	$\ln y(t) = \ln y_0 + rt$
3	$\dot{y} = r(1-y)$	$y(t) = 1 - (1-y_0)\mathrm{e}^{-rt}$	$\ln\left(\frac{1}{1-y(t)}\right) = \ln\left(\frac{1}{1-y_0}\right) + rt$
4	$\dot{y} = r(1-y)y$	$y(t) = \frac{1}{1 + \frac{1-y_0}{y_0}\mathrm{e}^{-rt}}$	$\ln\left(\frac{y(t)}{1-y(t)}\right) = \ln\left(\frac{y_0}{1-y_0}\right) + rt$
5	$\dot{y} = r(-\ln y)y$	$y(t) = \exp(\ln y_0\, \mathrm{e}^{-rt})$	$-\ln(-\ln y(t)) = -\ln(-\ln y_0) + rt$

Tabelle 7.3: Ordinatentransformation bei der Identifikation von Wachstumsraten [483]

Die Ordinatentransformation formuliert äquivalente Probleme, falls keine Störungen und Fehler vorliegen. Andernfalls geht die Äquivalenz verloren, was manchmal übersehen wird. Dann folgen aus dem parameterlinearen Modell nicht die gleichen Parameter wie aus dem parameternichtlinearen (Bias in Folge der Nichtlinearitäten). In Abhängigkeit von der Größe der Störungen und den Genauigkeitsanforderungen erfordert das eine Nachoptimierung. Durch eine geeignete Wichtung kann aber in vielen Fällen darauf verzichtet werden. So approximiert nämlich

$$\sum_{i=1}^{N} w_i^2 \Big(\phi(y_i) - \phi(f(u_i;\theta))\Big)^2 \stackrel{!}{=} \text{Min} \qquad \text{mit} \qquad w_i = \frac{1}{\frac{\mathrm{d}\phi}{\mathrm{d}y}(y_i)}, \tag{7.13}$$

die Zielfunktion

$$\sum_{i=1}^{N} (y_i - f(u_i;\theta))^2 \stackrel{!}{=} \text{Min}. \tag{7.14}$$

Das folgt aus der Taylor-Approximation $\phi(x) \approx \phi(x_0) + \phi'(x_0)(x - x_0)$ mit $x_0 := y_i$ und $x := f(u_i;\theta)$ über $\phi(f(u_i;\theta)) - \phi(y_i) \approx \phi'(y_i)(f(u_i;\theta) - y_i)$; $i = 1, \ldots, N$.

Die Formeln lassen sich wie folgt plausibilisieren: Stauchungen oder Streckungen durch die Transformation machen die Fehler kleiner oder größer als die zugehörigen Fehler im Originalmodell. Die Wichtung bewirkt nun, dass kleiner gewordene Fehler im transformierten Modell höher gewichtet werden; für größere gilt das Gegenteil. Letztlich wird durch die Wichtung erreicht, dass beide Gütefunktionen über alle θ näherungsweise gleich sind. Das bedingt, dass auch die Optimallösungen von (7.14) und (7.13) in Näherung gleich sind.

7.6 Umformulierung in eine Differenzialgleichung

Die Idee der Umformung eines Optimierungsproblems in eine Differenzialgleichung ergibt sich aus der Eigenschaft von Systemruhelagen, die bei zeitkontinuierlichen Systemen durch Nullsetzen der Ableitungen charakterisiert werden. Gleichfalls sind Extremalpunkte durch Nullsetzen des Gradienten gekennzeichnet. Werden diese beiden Aspekte kombiniert, ergibt sich die folgende Aussage: Angenommen $f : \mathbb{R}^n \to \mathbb{R}$ sei eine einmal stetig-differenzierbare freie Zielfunktion und $\nabla_x f = \frac{\partial f}{\partial x}$ der Gradient in x bezüglich des Standardskalarprodukts in $\mathbb{R}^n$, dann stimmen die Ruhelagen der Differenzialgleichung

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\nabla_x f; \qquad x(0) = x_0 \tag{7.15}$$

mit den stationären Punkten von f überein. Letztlich ist die Umformulierung eines freien Minimierungsproblems in eine Differenzialgleichung nichts anderes als die kontinuierliche Version des Verfahrens des steilsten Abstiegs (Gradientenverfahren). Für einen hinreichend guten Startwert x_0 – bei konvexen Problemen für jeden Startwert – strebt die Lösung der Differenzialgleichung (7.15) gegen einen Minimierer der Zielfunktion (positive Definitheitsforderung beim Minimum korrespondiert mit asymptotischer Lyapunov-Stabilität der Differenzialgleichung). Statt eines Optimierungsalgorithmus wird nunmehr ein numerischer Integrationsalgorithmus (Runge-Kutta-Verfahren und ähnliche) benötigt. Eine praktische Anwendung beschreibt das folgende Beispiel.

Beispiel 7.6 (Projektion auf eine ebene Kurve, [250])
Die Pfadregelung autonomer Fahrzeuge erfordert die Berechnung eines Projektionspunkts auf der Sollkurve, z. B. Projektion des Hinterachsmittelpunkts. Statt einer numerischen Optimierung wird in [250] die Formulierung 7.15 zusammen mit einer Vorsteuerung verwendet. Durch die Vorsteuerung mit der bekannten Geschwindigkeit des Autos wird gewissermaßen der Startwert für die Optimierung immer aktualisiert. Da der Standardzugang aber zu einem Schleppfehler führt, wenn die Krümmung ungleich Null ist, wird das Vorwissen über den Krümmungsverlauf ins Modell integriert. Letztlich entsteht dadurch ein Beobachter.

Ein Vorteil dieser Problemtransformation kommt besonders in adaptiven Reglern zum Tragen, bei denen online eine Parameterbestimmung erfolgt. Da solche Regler für nichtlineare Systeme überwiegend anhand der zeitkontinuierlichen Modellbeschreibung entworfen werden, braucht der zeitkontinuierliche Bereich dank der Differenzialgleichungsformulierung bei der Modellbildung nicht verlassen zu werden. Das wiederum vereinfacht den Stabilitätsnachweis gegenüber einer Kombination aus zeitdiskreter Modellbildung über Optimierung und zeitkontinuierlichem Reglerentwurf erheblich. Gleichwohl bleibt der Nachweis in aller Regel anspruchsvoll, da weder das Gewissheitsprinzip für Modell- und Reglerentwurf noch das Separationsprinzip für Beobachter- und Reglerentwurf im Nichtlinearen gilt.

Ein weiterer Vorteil dieser Problemtransformation ergibt sich bei der Optimierung über Restriktionen, die glatte Mannigfaltigkeiten oder konvexe Mengen beschreiben. Als ein Beispiel sei die Optimierung über der Stiefel-Mannigfaltigkeit genannt [135], d. h. über der Menge der orthonormalen Matrizen $X \in \mathbb{R}_n^{m\times n}$ mit $X^TX = I_n$. Mit Hilfe eines Projektionsoperators wird (7.15) so modifiziert, dass $x(t)$ niemals die zulässige Menge verlässt (verletzt), aber dennoch den Zielfunktionswert fortwährend verbessert

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\mathrm{Proj}(\nabla_x f). \tag{7.16}$$

Die Projektion erfolgt bei Gleichungsrestriktionen $h(x) = 0_m$ in den Tangentialraum von h im aktuellen x. Für Ungleichungsrestriktionen $\mathcal{F} = \{x \in \mathbb{R}^n : g(x) \leq 0_p\}$ ist

$$\mathrm{Proj}(\nabla_x f) = \begin{cases} \nabla_x f & x \in \mathrm{int}\mathcal{F} \text{ oder } \langle \nabla_x g, \nabla_x f\rangle \leq 0 \\ \left(I_p - \frac{\nabla_x g \nabla_x^T g}{||\nabla_x g||_2^2}\right)\nabla_x f & x \in \mathrm{bd}\mathcal{F} \text{ und } \langle \nabla_x g, \nabla_x f\rangle > 0 \end{cases} \tag{7.17}$$

ein möglicher Projektionsoperator [367]. Im unteren Fall erfolgt die Projektion in den Tangentialraum von g, womit sich die Trajektorie $x(t)$ auf dem Rand (boundary) $\mathrm{bd}\mathcal{F}$ der Restriktion bewegen wird, sofern sie sich nicht durch das Erfülltsein der Freigabebedingung $\langle \nabla_x g, \nabla_x f\rangle \leq 0$ von diesem lösen kann.

Anmerkung 7.6 Statt orthogonaler Projektionen können auch schiefe Projektionen verwendet werden. Das kann bei adaptiven Regelungssystemen von Vorteil sein, wenn sich dadurch im Stabilitätsnachweis mit einer Lyapunov-Funktion Terme in der abgeleiteten Lyapunov-Funktion aufheben. Durch eine ε-Glättung des Operators im Ungleichungsfall lässt sich die für einige Anwendungen unangenehme Unstetigkeitsproblematik am Rand von $\mathcal{F}$ umgehen [367][4]. In adaptiven Regelungssystemen ist oft zusätzlich eine Dämpfung für das Parameteraktualisierungsgesetz nach (7.16) vorzusehen, um Stabilität für das Gesamtsystem zu sichern. Selbstverständlich sind auch Modifikationen in Anlehnung an das Gauß-Newton- oder das Newton-Verfahren möglich.

[4] Die Unstetigkeitsproblematik wird an $(x-1)^2 \stackrel{!}{=} \mathrm{Min}; x \leq 0$ mit $x_{\mathrm{opt}} = 0$ sichtbar. So ist die Auftreffgeschwindigkeit $\dot{x}(t_{\mathrm{end}} - 0) = -2(x-1)|_{x=0} = 2$, um sodann $\dot{x}(t_{\mathrm{end}} + 0) = 0$ zu sein.

7.7 Formulierung als Projektionsproblem

Das rekursive Lösen restringierter LS-Probleme gilt als schwierig, ist es doch bereits nichtrekursiv schwierig, wenn das Problem nichtkonvex ist (lokale Minima mit unbekannter Anzahl). Die Annahme strenger Konvexität wird dabei meist gestellt, da dann Eindeutigkeit der Lösung folgt. Das erfordert eine konvexe Problemformulierung kombiniert mit einer Vollrangforderung an die Datenmatrix, die für Modelle ohne Überparametrisierung und je nach Modelltyp zusätzlich bei ständiger Systemanregung gegeben ist. Ohne Konvexität oder Vollrangforderung ist die Formulierung eines erfolgreichen Algorithmus nur möglich, wenn das restringierte LS-Problem sehr gut verstanden wird, denn nur dann können Adhoc-Maßnahmen ein Konvergieren zu lokalen Nebenlösungen verhindern.

Die Idee des hier vorgestellten Zugangs ist es, über einen herkömmlichen rekursiven LS-Lösungsalgorithmus die gewöhnliche LS-Lösung schritthaltend zu berechnen und sie dann in jedem Schritt auf die durch transformierte Restriktionen eingeschränkte zulässige Menge zu projizieren. Sofern für die Projektion geschlossene Lösungen existieren oder die Projektion numerisch schnell geht, steht damit ein Online-Algorithmus zum rekursiven Lösen der restringierter LS-Probleme bereit.

Das restringierte LS-Problem

$$||Ax - b||_2^2 \stackrel{!}{=} \text{Min} \qquad x \in \mathbb{R}^n : h(x) = 0, g(x) \leq 0; A \in \mathbb{R}_n^{m \times n} \tag{7.18}$$

ist bezüglich des Minimierers äquivalent zu

$$||Ax - b||_2^2 - b^T b + b^T A(A^T A)^{-1} A^T b \stackrel{!}{=} \text{Min} \qquad x \in \mathbb{R}^n : h(x) = 0, g(x) \leq 0; A \in \mathbb{R}_n^{m \times n}, \tag{7.19}$$

was seinerseits äquivalent ist zu

$$||(A^T A)^{-1} A^T b - x||_{A^T A}^2 \stackrel{!}{=} \text{Min} \qquad x \in \mathbb{R}^n : h(x) = 0, g(x) \leq 0; A \in \mathbb{R}_n^{m \times n}. \tag{7.20}$$

Das bedeutet nichts Anderes, als dass der Minimierer von (7.18) die Projektion der LS-Lösung $x_{LS} = (A^T A)^{-1} A^T b$ im Sinne der durch $A^T A$ bestimmten elliptischen Norm auf die Restriktion ist. Um allerdings die bekannten Formeln, insbesondere die für die orthogonale Projektion auf konvexe Mengen, algorithmisch nutzen zu können, ist die elliptische Norm in die euklidische zu überführen. Mithin ist (7.20) mit $y = (A^T A)^{1/2} x$ äquivalent zu

$$||(A^T A)^{1/2} x_{LS} - y||_2^2 \stackrel{!}{=} \text{Min} \qquad y \in \mathbb{R}^n : h((A^T A)^{-1/2} y) = 0, g((A^T A)^{-1/2} y) \leq 0, \tag{7.21}$$

was sich auch für eine direkte Berechnung der restringierten LS-Lösung eignet. Die Darstellung kann aber auch vorteilhaft verwendet werden, um bestehende rekursive Algorithmen

für Restriktionen tauglich zu machen. Dies soll am Standard-Rekursionsalgorithmus für die LS und das Modell $y[k] = \phi[k-1]\theta + \varepsilon[k]$

$$\hat{\theta}[k] = \hat{\theta}[k-1] + \frac{P[k-2]\phi[k-1]}{1+\phi^T[k-1]P[k-2]\phi[k-1]}\Big(y[k] - \phi^T[k-1]\hat{\theta}[k-1]\Big) \tag{7.22a}$$

$$P[k-1] = P[k-2] - \frac{P[k-2]\phi[k-1]\phi^T[k-1]P[k-2]}{1+\phi^T[k-1]P[k-2]\phi[k-1]}; \qquad k \geq 1 \tag{7.22b}$$

mit gegebenem $\hat{\theta}[0]$ und einer beliebigen positiv definiten Startmatrix $P[-1]$, z. B. $\epsilon_{mach}^{-1/2} I_n$, erklärt werden. Der Algorithmus berechnet im N-ten Schritt den Minimierer $\hat{\theta}[N]$ von

$$\frac{1}{2}\sum_{k=1}^{N}(y[k] - \phi^T[k]\theta)^2 + \frac{1}{2}\Big(\theta - \hat{\theta}[0]\Big)P^{-1}[-1]\Big(\theta - \hat{\theta}[0]\Big) \stackrel{!}{=} \text{Min}, \tag{7.23}$$

vgl. [232]. (7.22b) wird dabei über das Matrix-Inversionslemma, angewandt auf

$$P^{-1}[k-1] = P^{-1}[k-2] + \phi[k-1]\phi^T[k-1] = P^{-1}[-1] + \sum_{i=1}^{k-1}\phi[i-1]\phi^T[i-1], \tag{7.24}$$

hergeleitet. Damit entspricht $P^{-1}[k-1]$ bis auf den zu vernachlässigenden Term $P^{-1}[-1]$ der Matrix A^TA in (7.21). Der Algorithmus braucht also nur um den Schritt

$$||P^{-1/2}[k-1]\hat{\theta}[k] - y||_2^2 \stackrel{!}{=} \text{Min} \qquad y \in \mathbb{R}^n : h(P^{1/2}[k-1]y) = 0, g(P^{1/2}[k-1]y) \leq 0, \tag{7.25}$$

sowie die erforderliche Resubstitution

$$\hat{\theta}_C[k] = P^{1/2}[k-1]y_{\text{opt}} \tag{7.26}$$

ergänzt werden. Nachteil des Verfahrens ist die zusätzliche Berechnung der Quadratwurzeln von $P[k-1]$ und $P^{-1}[k-1]$. Die numerisch robuste UD-Faktorisierung als Alternative zum rekursiven Standard-LS-Algorithmus liefert dagegen die Quadratwurzeln über elementweises Wurzelziehen des sog. D-Faktors praktisch fast gratis mit. Prinzipiell kann der A-priori-Fehler im nächsten Schritt auch mit dem „besseren“ restringierten $\hat{\theta}_C[k]$ berechnet werden. Die entstehende Folge der $\hat{\theta}[k]$ weicht dann von der Lösungsfolge der LS-Probleme unter Restriktionen ab.

Bei anderen Rekursionsalgorithmen, wie etwa Givens-Aufdatierungen [347], kann analog verfahren werden, d. h., dass die entsprechend transformierte aktuelle LS-Lösung auf die transformierte Restriktion projiziert wird. Aus der optimalen Projektion folgt dann per Rücktransformation die aktuelle restringierte LS-Lösung.

7.8 Semidefinite Optimierung

Die semidefinite Optimierung bezeichnet das Lösen von Problemen mit eine linearen Zielfunktion unter Semidefinitheitsrestriktionen an Matrizen oder affine Matrixfunktionen. Aus der Matrixalgebra ist bekannt, dass eine symmetrische Matrix $A \in \mathbb{R}^{n \times n}$ semidefinit (auch nichtnegativ definit genannt) ist, wenn $x^T Ax \geq 0$ für alle $x \in \mathbb{R}^n$ gilt. Die Menge der semidefiniten Matrizen wird mit $\mathcal{S}_n^{\succeq}$ bezeichnet, was auch durch $A \succeq 0_{n \times n}$ ausgedrückt werden kann. Sie formt einen konvexen Kegel, d. h. nichtnegative Vielfache eine Matrix liegen wieder in der Menge. Zudem ist auf den semidefiniten Matrizen wie auch auf den symmetrischen Matrizen die sog. Löwner-Halbordnung erklärt, wonach $A \succeq B$ gilt, wenn $A - B \in \mathcal{S}_n^{\succeq}$ ist. Da lineare Zielfunktionen konvex sind und da eine Semidefinitheitsrestriktion konvex ist und somit auch die Schnitte von mehreren solchen Restriktionen konvex sind, gehört die semidefinite Optimierung zur konvexen Optimierung.

Die Vorteile der semidefiniten Optimierung sind zahlreich:

1. Obwohl die Klasse der SDP-Probleme viel allgemeiner ist als die Klasse der LP-Probleme, sind SDP-Probleme dank der Inneren-Punkt-Methoden nicht viel schwieriger als LP-Probleme zu lösen. Nesterov und Nemirovskii [476] zeigen, dass sich die Inneren-Punkt-Methoden für LP-Probleme, die erfolgreich von Karmarkar [337] eingeführt wurden, für alle konvexen Optimierungen nutzen lassen. Speziell für SDP-Probleme sind geeignete Barriere-Funktionen verfügbar und leicht anzuwenden. Hierin liegt die exponierte Stellung der LMI-Restriktionen[5] gegenüber anderen konvexen Restriktionen begründet.

2. SDP-Probleme lassen sich in polynomialer Zeit lösen. Oft reichen 5 bis 50 Iterationen [476], um die Optimallösung zu finden. Da pro Iteration nur ein lineares LS-Problem zu lösen ist und da hierfür schnelle, robuste, ggf. schwache Besetztheit berücksichtigende Algorithmen verfügbar sind, bereiten auch Probleme mit mehreren hundert Parametern kaum Probleme.

3. Das SDP-Problem wird vom theoretischen Standpunkt aus gut verstanden. Es liegt unter einer Slater-Constraint-Qualification starke Dualität vor [61].

4. LMI-Zugänge erweisen sich für Ingenieuranwendungen (Modelle unter quadratischen Restriktionen, Versuchsplanung, Robuste Regelung, Stabilitätseinzugsbereiche) als leistungsstark [92], [93].

[5] LMI steht als Abkürzung für „linear matrix inequality“.

5. SDP- und LMI-Relaxationen sind für diskrete Optimierungsprobleme, die oft als NP-schwer einzustufen sind (z. B. Travelling-Salesman-Problem, Max-Cut-Problem), interessant, da sie schnell gute Schranken für das Minimum liefern. Sie sind aber auch hilfreich für nichtkonvexe, multimodale Probleme, die sie durch polynomiale Zielfunktionen und Restriktionen ergeben. So wird in [279] wird ein Algorithmus mit sukzessiver LMI-Relaxation vorgestellt, der eine Folge generiert, die monoton gegen das globale Minimum einer multivariaten polynomialen Zielfunktion mit polynomialen Restriktionen konvergiert.

6. Einige Probleme mit unendlichdimensionalen konvexen Restriktionen können durch LMI-Restriktionen in endlichdimensionale Restriktionen überführt werden. Dadurch verringert sich der Schwierigkeitsgrad erheblich.

7. Programmsysteme (Abk. für Problemklassen s. Tab. A.8):[6]
   - SeDuMi (öffentlich; Matlab; LP, SOCP, SDP)
   - CSDP, SDPA (öffentlich; C; SDP)
   - MOSEK (kommerziell; C mit Matlab-Interface; LP, SOCP, GP, ...)[7]
   - solver.com (kommerziell; Excel-Interface; LP, SOCP)
   - GPCVX (öffentlich; Matlab; GP)
   - CVXOPT (öffentlich; Python/C; LP, SOCP, SDP, GP, ...)

Als Nachteile der semidefiniten Optimierung sind für den Ingenieur die Einarbeitung in den mathematischen Hintergrund und die Anpassung seiner Probleme auf die programmspezifische Form zu nennen. Die nachfolgenden Abschnitte geben deshalb eine Einstieg in die Thematik, vermitteln die wichtigsten Grundideen und zeigen an einfachen Beispielen das Vorgehen. Als erste weiterführende Literatur kann [93] empfohlen werden.

---

[6] Eine Zusammenstellung findet sich auf http://www-user.tu-chemnitz.de/~helmberg/semidef.html.

[7] GP steht für Geometric Programming, d. h. Probleme der Form

$$f(x) \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n_> : h_i(x) = 1, g_j(x) \leq 1; i = 1, \ldots, m; j = 1, \ldots, p$$

wobei f, g_j Posynome und h_i Monome sind. Posynome sind Funktionen $f(x) = \sum_{k=1}^{K} c_k x_1^{a_{1k}} \cdots x_n^{a_{nk}}$ mit $f : \mathbb{R}^n_> \to \mathbb{R}$, $c_k > 0$ und $[a_{1k}, \ldots, a_{nk}]^T \in \mathbb{R}^n$; die Summanden heißen Monome. Neben der Positivität liegen die Unterschiede zu den mehrvariablen Polynomen in den nicht natürlichzahligen Exponenten. Durch die Reparametrisierung $x = \ln y$ entstehen konvexe Probleme $\tilde{f}(y) = \sum_{k=1}^{K} \mathrm{e}^{a_k^T y + b_k}$ und $b_k = \ln c_k$.

7.8.1 Beschreibungsformen von semidefiniten Problemen

Das algorithmische Lösen semidefiniter Probleme erfordert, dass Zielfunktion und Restriktion einer bestimmten Darstellung genügen. Meist wird sich dabei auf die kanonische Beschreibungsform bezogen. Nachteil dieser Form ist die große Anzahl von Matrizen, die häufig viele Nullen enthalten. Während das für die theoretische Behandlung unbedeutend ist, erschwert es die programmtechnische Umsetzung und verringert die Übersichtlichkeit. Durch die Matrixform kann der Nachteil vermieden werden. Allerdings erfordert die dadurch entstehende Variationsbreite von Restriktionsformulierungen eine modifizierte theoretische Behandlung, und außerdem sind nur eingeschränkte algorithmische Umsetzungen verfügbar. Die Lücke wird in einigen Softwaretools durch Zusatzmodule geschlossen, die Matrixform-Darstellungen automatisiert in kanonische Darstellungen überführen. Im Folgenden werden die beiden Formen vorgestellt.

Kanonische Form

In der semidefiniten Optimierung wird eine lineare Zielfunktion unter der Restriktion minimiert, dass eine affine Kombination symmetrischer Matrizen nichtnegativ definit ist

$$c^T x \stackrel{!}{=} \text{Min} \qquad F(x) = F_0 + \sum_{i=1}^{n} x_i F_i \succeq 0_{m\times m}; \quad x = ((x_i)) \in \mathbb{R}^n, F_0, \ldots, F_n \in \mathcal{S}_m. \ /^8 \quad (7.27)$$

(7.27) wird als die kanonische Form eines SDP-Problems bezeichnet. Die Restriktion beschreibt eine lineare Matrixungleichung, d. h. eine Relation im Sinne der Löwner-Halbordnungen $\succeq, \succ, \preceq, \prec$ zwischen zwei symmetrischen Matrixfunktionen [72], in denen die Unbekannten x_i affin auftreten. Solch eine Restriktion ist nichtlinear, aber konvex, denn

$$F(x), F(y) \in \mathcal{S}_m^{\succeq}; 0 \le \gamma \le 1: \quad F(\gamma x + (1-\gamma)y) = \gamma F(x) + (1-\gamma)F(y) \in \mathcal{S}_m^{\succeq}. \quad (7.28)$$

Matrixform

Während die kanonische Form die Grundlage für die algorithmische Behandlung liefert, wird aus Gründen der Interpretierbarkeit und Anschaulichkeit in der Darstellung die Matrixform bevorzugt, ggf. ist eine matrixvariate Zielfunktion $\text{spur}(CX)$ zu formulieren. In der Matrixform werden die Unbekannten x_i zu Vektoren oder Matrizen zusammengefasst, die ihrerseits wiederum eine symmetrische Matrix formen. So wird statt der kanonischen Restriktion

$$x_1\begin{bmatrix}1&0&0\\0&0&0\\0&0&0\end{bmatrix} + x_2\begin{bmatrix}0&0&0\\0&1&0\\0&0&0\end{bmatrix} + x_3\begin{bmatrix}0&0&0\\0&0&0\\0&0&1\end{bmatrix} + x_4\begin{bmatrix}0&1&0\\1&0&0\\0&0&0\end{bmatrix} + x_5\begin{bmatrix}0&0&1\\0&0&0\\1&0&0\end{bmatrix} + x_6\begin{bmatrix}0&0&0\\0&0&1\\0&1&0\end{bmatrix} \succeq 0_{3\times 3} \quad (7.29)$$

besser

$$X \succeq 0_{3\times 3} \quad (7.30)$$

[8] Die Restriktion ist zu $\lambda_{\min}(F(x)) \ge 0$ äquivalent. Ihre zulässige Menge hat i. Allg. einen nichtglatten Rand, vgl. $Ax \ge b \Leftrightarrow \text{diag}(Ax - b) \succeq 0_{m\times m}$.

geschrieben. Beide Restriktionen fordern, dass die symmetrische (3×3)-Matrix nichtnegativ definit sein soll. Allgemein kann jede Matrixform stets in die kanonische Form überführt werden, der umgekehrte Weg indes ist nur sinnvoll, wenn sich die x_i zu Vektoren oder Matrizen zusammenfassen lassen. Beispiele für LMI-Restriktionen in Matrixnotation sind die aus der Regelungstheorie bekannte Lyapunov-Ungleichung

$$A^T P + PA \succeq 0_{n \times n}, \tag{7.31}$$

die diskrete Lyapunov-Ungleichung oder die Kalman-Yakubovich-Ungleichung. In (7.31) ist zu beachten, dass entweder A oder P konstant sein muss, da ansonsten die linke Seite der Gleichung nicht mehr linear in den Unbekannten ist. Sind sowohl A als auch P Unbekannte, entsteht eine ungleich schwieriger zu behandelnde nichtkonvexe Restriktion.
Da die Umformung einer Doppelungleichung in zwei einfache Ungleichungen keinerlei Schwierigkeiten bereitet, ist es aus Gründen der stärkeren Aussagekraft der Doppelungleichung zweckmäßig, diese in der Problemformulierung zu belassen und nicht aufzulösen. So wird für die Sphärikrestriktion an einen Ellipsoiden (wenig Abweichung von einer Kugel) besser

$$\gamma_1 I_n \preceq X \preceq \gamma_2 I_n \qquad \text{anstatt} \quad \gamma_2 I_n - X \succeq 0_{n \times n}, X - \gamma_1 I_n \succeq 0_{n \times n} \tag{7.32}$$

geschrieben. Ebenso ist es unüblich, zwei einzelne LMI-Restriktionen durch ihre direkte Summe darzustellen, d. h.

$$A_1(x) \succeq 0_{m \times m}, A_2(x) \succeq 0_{p \times p} \qquad \text{anstatt} \quad \begin{bmatrix} A_1(x) & 0 \\ 0 & A_2(x) \end{bmatrix} \succeq 0_{(m+p) \times (m+p)}. \tag{7.33}$$

Aufmerksamkeit ist der Tatsache zu schenken, dass eine strenge Ungleichung nicht zu einer nicht strengen gemacht werden kann. Dies entspräche einem Abschwächen der Restriktion (7.27), was nur zulässig ist, wenn $F(x_{\text{opt}}) \succ 0_{m \times m}$ gilt. So muss unter der Restriktion $F(x) \succ 0_{m \times m}$ keine Lösung existieren, da der zulässige Bereich eine offene Menge ist. Die Zielfunktion hat dann unter Umständen nur ein Infimum.

Eine alternative Darstellung zu (7.27) lautet

$$c^T x \stackrel{!}{=} \text{Min} \qquad Ax = b, x \in \mathcal{C}. \tag{7.34}$$

Diese Darstellung hat große Ähnlichkeit mit der Standardform eines LP-Problems. Der einzige Unterschied ist, dass statt des Kegels der nichtnegativen Matrizen im LP-Problem als Kegel der nichtnegative Orthant benutzt wird. Umgekehrt lässt sich jedes LP-Problem in ein SDP-Problem umwandeln

$$c^T x \stackrel{!}{=} \text{Min}; Ax \leq b, A \in \mathbb{R}^{m \times n} \quad \Leftrightarrow \quad c^T x \stackrel{!}{=} \text{Min}; F(x) = \text{diag}(-Ax + b) \succeq 0_{m \times m}. \tag{7.35}$$

7.8.2 Umformung konvexer Probleme in semidefinite Probleme

Das Überführen konvexer Probleme in SDP-Probleme geschieht in drei Schritten:

1. Erzeugen einer linearen Zielfunktion mit der Epigraph-Methode (Abschnitt 6.5)
2. Umformen der Restriktion in eine LMI
3. Vereinfachen, wenn möglich (Abschnitt 7.8.3)

Das soll am folgenden Beispiel veranschaulicht werden.

Beispiel 7.7 (Minimieren der Spektralnorm)
Das konvexe Problem

$$||A(x)||_2 \stackrel{!}{=} \text{Min} \qquad x \in \mathbb{R}^n; A(.) : \mathbb{R}^n \to \mathbb{R}^{p \times q} \text{ affin in } x \tag{7.36}$$

lässt sich mit der Epigraph-Form und der Umformung 9 in Tabelle 7.4 über

$$x_{n+1} \stackrel{!}{=} \text{Min} \qquad x \in \mathbb{R}^{n+1}; ||A(x)||_2 \leq x_{n+1} \qquad \text{(erster Schritt)} \tag{7.37a}$$

$$\Leftrightarrow \qquad x_{n+1} \stackrel{!}{=} \text{Min} \qquad \begin{bmatrix} x_{n+1} I_p & A(x) \\ A^T(x) & x_{n+1} I_q \end{bmatrix} \succeq 0_{(p+q)\times(p+q)} \qquad \text{(zweiter Schritt).} \tag{7.37b}$$

in ein SDP-Problem überführen. Ein Zusatzvorteil: In (7.36) existiert die Ableitung nicht überall, wodurch bei den erforderlichen Fallunterscheidungen schwierig handhabbare Ausdrücke entstehen; (7.37b) kennt solche Probleme nicht.

Die LMI-Restriktionsumformung eignet sich nur für konvexe Restriktionen, da eine LMI eine konvexe Restriktion verkörpert und somit bei geforderter Äquivalenz zur Ausgangsrestriktion diese auch konvex sein muss. Für das Umformen gibt es keinen einheitlichen Weg; Möglichkeiten sind in den Tabellen 7.4 und 7.5 angegeben, die Ergebnisse aus [93] zusammenfassen und erweitern. In Tabelle 7.4 ist zu sehen, dass sich parabolische Flächen, Ellipsen und Kegel sehr elegant durch LMI-Restriktionen ausdrücken lassen. Insbesondere die Ellipsen spielen in geometrischen Anwendungen eine wichtige Rolle, z. B. umhüllende Ellipsen, während die Ellipsoide als ihre n-dimensionalen Erweiterungen bevorzugt bei der Erstellung statistischer Modelle (Streuellipsoide) zum Einsatz kommen. Tabelle 7.5 enthält die wichtige Schur-Komplementformel. Hier allerdings nicht in der gewöhnlichen Richtung, dass aus einer Blockmatrix zwei Bedingungen für deren Semidefinitheit kleinerer Dimension abgeleitet werden, sondern die umgekehrte Richtung, also die Aggregation von zwei Bedingungen in einer. Die gewöhnliche Anwendung der Schur-Komplementformel wird jedoch im nächsten Abschnitt benötigt.

Anmerkung 7.7 Das Umformen in LMI-Restriktionen lässt sich den Erweiterungsmethoden zuordnen, da die Zahl der Restriktionen um triviale und redundante Restriktionen zunimmt. Das zeigt das nächste Beispiel.

	Bedingung	Umformung
1	$Ax \underset{(=)}{\geq} b; \quad A \in \mathbb{R}^{m \times n}$ lineare Restriktion	$\text{diag}(Ax-b) \underset{(=)}{\succ} 0_{m \times m}$
2	$y \geq ax^2$ Parabelfläche	$\begin{bmatrix} 1/a & x \\ x & y \end{bmatrix} \succeq 0_{2 \times 2}$
3	$\frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1$ Ellipsenfläche	$\begin{bmatrix} 1 & x & y \\ x & a^2 & 0 \\ y & 0 & b^2 \end{bmatrix} \succeq 0_{3 \times 3}$
4	$\frac{x^2}{a^2} + \frac{y^2}{b^2} \leq z^2, 0 < z < c$ Kegel	$\begin{bmatrix} \zeta & x & y & 0 & 0 \\ x & a^2 & 0 & 0 & 0 \\ y & 0 & b^2 & 0 & 0 \\ 0 & 0 & 0 & \zeta & 0 \\ 0 & 0 & 0 & 0 & \sqrt{c}-\zeta \end{bmatrix} \succeq 0_{5 \times 5}; \; \zeta = z^2$
5	$\|\|Ax-b\|\|_2 \leq c^T x + d$ SOC-Restriktion	$\begin{bmatrix} (c^T x + d) I_m & Ax-b \\ (Ax-b)^T & c^T x + d \end{bmatrix} \succeq 0_{(m+1) \times (m+1)}$
6	$(Ax-b)^T (Ax-b) - c^T x - d \leq 0$ konvexe quadratische Restriktion	$\begin{bmatrix} I_m & Ax-b \\ (Ax-b)^T & c^T x + d \end{bmatrix} \succeq 0_{(m+1) \times (m+1)}$
7	$(x-b)^T A^{-1} (x-b) < \gamma; A \succ 0_{n \times n}$ Ellipsoid	$\begin{bmatrix} A & x-b \\ (x-b)^T & \gamma \end{bmatrix} \succ 0_{(n+1) \times (n+1)}$
8	$(a^T x)^2 / (b^T x) \underset{(=)}{\geq} y \underset{(=)}{\geq} 0$	$\begin{bmatrix} y & a^T x \\ a^T x & b^T x \end{bmatrix} \underset{(=)}{\succ} 0_{2 \times 2}$
9	$\|\|A(x)\|\|_2 \underset{(=)}{\leq} \gamma$, [382] $A(.)$ affine Funktion in x	$\begin{bmatrix} \gamma I_m & A(x) \\ A^T(x) & \gamma I_n \end{bmatrix} \underset{(=)}{\succ} 0_{(m+n) \times (m+n)}$
10	$\|\|X\|\|_2 \underset{(=)}{\leq} \gamma; \quad X \in \mathbb{R}^{m \times n}, \gamma \in \mathbb{R}^{>}$ Korollar aus (7)	$\begin{bmatrix} \gamma I_m & X \\ X^T & \gamma I_n \end{bmatrix} \underset{(=)}{\succ} 0_{(m+n) \times (m+n)}$
11	$\|\|X\|\|_F \underset{(=)}{\leq} \gamma; \quad X \in \mathbb{R}^{m \times n}, \gamma \in \mathbb{R}^{>}$ Korollar aus (12a) im Fall $<$ und (12c) im Fall $\leq$ mit Y als Schlupfvariable	$\begin{bmatrix} Y & X \\ X^T & I_n \end{bmatrix} \underset{(=)}{\succ} 0_{(m+n) \times (m+n)}, \; \text{spur} Y \underset{(=)}{\leq} \gamma^2$

Tabelle 7.4: Äquivalente Umformungen in LMI-Restriktionen (Teil 1)

	Bedingung	Umformung
12a	Restriktionen vom Schur-Komplement-Typ, [299], [92] $C(x) \succ 0_{m\times m}$, $A(x) - B(x)C^{-1}(x)B^T(x) \succ 0_{n\times n}$	$\begin{bmatrix} A(x) & B(x) \\ B^T(x) & C(x) \end{bmatrix} \succ 0_{(m+n)\times(m+n)}$
12b	Für $C(x) \succ 0_{m\times m}$ gilt: $A(x) - B(x)C^{-1}(x)B^T(x) \succeq 0_{n\times n}$	$\begin{bmatrix} A(x) & B(x) \\ B^T(x) & C(x) \end{bmatrix} \succeq 0_{(m+n)\times(m+n)}$
12c	$C(x) \succeq 0_{m\times m}$, $A(x) - B(x)C^{+}(x)B^T(x) \succeq 0_{n\times n}$, $B(x)(I_m - C(x)C(x)^{+}) = 0_{n\times m}$	$\begin{bmatrix} A(x) & B(x) \\ B^T(x) & C(x) \end{bmatrix} \succeq 0_{(m+n)\times(m+n)}$
13	$C(x) \succ 0_{m\times m}$, $\text{spur}(B(x)C^{-1}(x)B^T(x)) < 1$ Korollar aus (12a) mit Y als Schlupfvariable	$\begin{bmatrix} Y & B(x) \\ B^T(x) & C(x) \end{bmatrix} \succ 0_{(m+n)\times(m+n)}$, $\text{spur}Y < 1$
14	$A_i(x) \underset{(=)}{\succ} 0_{n_i\times n_i}; i = 1,\ldots,m$ $A_i(.)$ affine Funktionen in x	$\text{diag}(A_i(x)) \underset{(=)}{\succ} 0_{\sum n_i\times\sum n_i}$
15	$\|\|X(I_n + X)^{-1}\|\|_2 < 1$	$X + X^T + I_n \succ 0_{n\times n}$ /[9]

Tabelle 7.5: Äquivalente Umformungen in LMI-Restriktionen (Teil 2)

Beispiel 7.8 (Triviale und redundante Restriktionen bei der LMI-Umformung) Die Einheitskreisfläche lässt sich als LMI wie folgt ausdrücken

$$x^2 + y^2 \leq 1 \Leftrightarrow \begin{bmatrix} 1 & x & y \\ x & 1 & 0 \\ y & 0 & 1 \end{bmatrix} \succeq 0_{3\times 3}. \tag{7.38}$$

Bekanntermaßen ist eine Matrix genau dann nichtnegativ definit, wenn alle Hauptminoren nichtnegativ sind [645]. Das bedeutet für die Matrix in (7.38)

$$1 \geq 0 \qquad \text{für die drei Hauptminoren 1. Ordnung} \tag{7.39a}$$
$$1 \geq x^2, 1 \geq y^2, 1 \geq 0 \qquad \text{für die drei Hauptminoren 2. Ordnung} \tag{7.39b}$$
$$1 \geq x^2 + y^2 \qquad \text{für den Hauptminor 3. Ordnung} \tag{7.39c}$$

(7.39a) ist trivial und (7.39b) schwächer als (7.39c), also redundant.

[9] $X^T + X + I \succ 0 \Leftrightarrow (X^T + I)(X + I) \succ X^T X \Leftrightarrow I \succ (X^T + I)^{-1}X^T X(X + I)^{-1}$
$\Rightarrow 1 > ||X(X + I)^{-1}||_F^2 \geq ||X(X + I)^{-1}||_2^2$
umgekehrt: $1 > ||X(X + I)^{-1}||_2 \geq \sigma_i(X(X + I)^{-1}) = \sqrt{\lambda_i((X^T + I)^{-1}X^T X(X + I)^{-1})}$
$1 > \lambda_i((X^T + I)^{-1}X^T X(X + I)^{-1}) \Leftrightarrow I \succ (X^T + I)^{-1}X^T X(X + I)^{-1} \Leftrightarrow X^T + X + I \succ 0$

7.8.3 Vereinfachungen von linearen Matrixungleichungen

Während sich im vorangegangen Abschnitt mit der Frage befasst wurde, wie eine konvexe Restriktion in eine LMI überführt werden kann, widmet sich dieser Abschnitt der Frage, wie LMIs durch Umformungen vereinfacht oder auf eine verfügbare Software zugeschnitten werden können. Basisumformungen finden sich in den Standardwerken zur Matrixalgebra [72], [299], [300], [382]. Als Beispiel sei die folgende Äquivalenz auf $\mathcal{S}_n^>$ genannt

$$A \succeq B \Leftrightarrow A^{-1} \preceq B^{-1}. \tag{7.40}$$

Der Trägheitssatz von Sylvester[10] kann verwendet werden, um Koeffizientenmatrizen

$$S \in \mathbb{R}_n^{n\times n}, X \in \mathcal{S}_n : S^T X S \succ 0_{n\times n} \Leftrightarrow X \succ 0_{n\times n} \tag{7.41}$$

zu eliminieren. Noch leistungsstärker werden er und seine Erweiterungen in Kombination mit den Schur-Komplementformeln in Tabelle 7.5. Für Probleme, in denen sich die Variable X nicht über die gesamte Matrix erstreckt, also beispielsweise bei

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22}+X & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \quad \text{oder} \quad \begin{bmatrix} A_{11} & A_{12} & A_{13}+X^T \\ A_{21} & A_{22} & A_{23} \\ A_{31}+X & A_{32} & A_{33} \end{bmatrix}, \tag{7.42}$$

kann die Zahl der Restriktionen reduziert werden. Hierzu werden zunächst alle Variablen durch ein geeignetes S in eine möglichst kleine Blockmatrix in der linken oberen Ecke transformiert, z. B.

$$\begin{bmatrix} 0 & I & 0 \\ I & 0 & 0 \\ 0 & 0 & I \end{bmatrix}^T \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22}+X & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} 0 & I & 0 \\ I & 0 & 0 \\ 0 & 0 & I \end{bmatrix} = \begin{bmatrix} A_{22}+X & A_{21} & A_{23} \\ A_{12} & A_{11} & A_{13} \\ A_{32} & A_{31} & A_{33} \end{bmatrix} \tag{7.43}$$

oder

$$\begin{bmatrix} I & 0 & 0 \\ 0 & 0 & I \\ 0 & I & 0 \end{bmatrix}^T \begin{bmatrix} A_{11} & A_{12} & A_{13}+X^T \\ A_{21} & A_{22} & A_{23} \\ A_{31}+X & A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & I \\ 0 & I & 0 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{13}+X^T & A_{12} \\ A_{31}+X & A_{33} & A_{32} \\ A_{21} & A_{23} & A_{22} \end{bmatrix}. \tag{7.44}$$

Anschließend werden die Schur-Komplementformeln eingesetzt. Für die Zulässigkeit des Problems muss die rechte untere Blockmatrix (diese ist nunmehr eine reine Koeffizientenmatrix)

[10] Die (reelle) Matrixversion des Satzes lautet: Ist A symmetrisch, so bleibt die Trägheit (Eigenwertverteilung bezüglich der imaginären Achse) invariant unter Kongruenztransformationen. Kurzum, A und $S^T AS$ haben bei regulärem S links, rechts und auf der reellen Achse die gleiche Anzahl von Eigenwerten [382].

der gleichen Relation wie die ursprüngliche LMI genügen. Das kann vorab leicht überprüft werden. Ist also Zulässigkeit gegeben, dann gilt für das erste Problem die Äquivalenz

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22}+X & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \succ 0 \Leftrightarrow A_{22}+X-(A_{21},A_{23}) \begin{bmatrix} A_{11} & A_{13} \\ A_{31} & A_{33} \end{bmatrix}^{-1} \begin{bmatrix} A_{12} \\ A_{32} \end{bmatrix} \succ 0 \quad (7.45)$$

und für das zweite

$$\begin{bmatrix} A_{11} & A_{12} & A_{13}+X^T \\ A_{21} & A_{22} & A_{23} \\ A_{31}+X & A_{32} & A_{33} \end{bmatrix} \succ 0 \Leftrightarrow \begin{bmatrix} A_{11} & A_{13}+X^T \\ A_{31}+X & A_{33} \end{bmatrix} - A_{12}A_{22}^{-1}A_{21} \succ 0. \quad (7.46)$$

In beiden Fällen ist die rechtsstehende LMI von geringerer Dimension und besteht ihr zweiter Term nur aus Koeffizientenmatrizen, die – einmal berechnet – lediglich eine neue Koeffizientenmatrix darstellt.

Neben den diskutierten Umformungen können auch Variablentransformationen helfen, Probleme zu vereinfachen, wie das folgende Beispiel zeigt.

Beispiel 7.9 (Stabilisierungsproblem)
Die Stabilisierbarkeit eines LTI-Systems durch Zustandsrückführung kann als – Finde ein $P \succ 0_{n\times n}$ und $K \in \mathbb{R}^{m\times n}$ mit $(A+BK)P+P(A+BK)^T \prec 0_{n\times n}$ – formuliert werden. Dieses Problem ist wegen der Produktverknüpfung von P und K bilinear und damit nichtkonvex. Mit der Variablentransformation $X = P$ und $Z = KP$ entsteht das konvexe Problem: Finde ein $X \succ 0_{n\times n}$ und ein $Z \in \mathbb{R}^{m\times n}$ mit $AX + BZ + XA^T + ZB^T \prec 0_{n\times n}$.

Abschließend sei noch das Eliminationslemma (auch als Projektionslemma bezeichnet) genannt, durch das sich das Zulässigkeitsproblem für eine LMI auf eine Auswertung der Koeffizientenmatrizen reduziert.

Satz 7.5 (Eliminationslemma, [92])
Es gilt:

$$\begin{aligned} &\exists X \in \mathbb{R}^{m\times n} : A + B^T X C + C^T X^T B \succ 0_{n\times n}; \quad A \in \mathcal{S}_k, B \in \mathbb{R}^{m\times k}, C \in \mathbb{R}^{n\times k} \\ &\qquad\qquad \Leftrightarrow \\ &B_\perp^T A B_\perp \succ 0_{(k-\mathrm{rg}B)\times(k-\mathrm{rg}B)} \text{ mit } BB_\perp = 0_{m\times(k-\mathrm{rg}B)}; B_\perp \in \mathbb{R}^{k\times(k-\mathrm{rg}B)}_{k-\mathrm{rg}B} \\ &C_\perp^T A C_\perp \succ 0_{(k-\mathrm{rg}C)\times(k-\mathrm{rg}C)} \text{ mit } CC_\perp = 0_{n\times(k-\mathrm{rg}C)}; C_\perp \in \mathbb{R}^{k\times(k-\mathrm{rg}C)}_{k-\mathrm{rg}C} \end{aligned} \quad (7.47)$$

Anmerkung 7.8 Eine Erweiterung des Lemmas zur Elimination der Unbekannten aus einer Menge von Ungleichungen $A_i + B_i^T X C_i + C_i^T X^T B_i \succ 0_{n\times n}$ ist unzulässig. Allerdings existiert für den Spezialfall mit mehreren A_i, aber $B_i = B$ und $C_i = C$ eine Erweiterung [486].

7.8.4 Konische Probleme zweiter Ordnung

Für eine spezielle Klasse semidefiniter Probleme, bei der die Restriktion einen konvexen m-dimensionalen Standardkegel zweiter Ordnung (Eiskremkegel, Lorentz-Kegel)

$$\begin{aligned}\mathcal{C}_m^2 &\stackrel{\text{def}}{=} \left\{ \begin{bmatrix} u \\ t \end{bmatrix} : u \in \mathbb{R}^{m-1}, t \in \mathbb{R}, ||u||_2 \leq t \right\} \\ &= \{z \in \mathbb{R}^m : z_m \geq \sqrt{z_1^2 + \ldots + z_{m-1}^2}\}.\end{aligned}$$

beschreibt, lassen sich effiziente Innere-Punkt-Algorithmen [476] formulieren, die denen auf LMI-Formulierungen basierenden SDP-Algorithmen überlegen sind. Die zugehörige Problemklasse wird dann wie folgt definiert.

Definition 7.2 (SOCP-Problem, [93])
Ein nichtlineares konvexes Probleme der Art

$$\begin{aligned} f^T x \stackrel{!}{=} \text{Min} \qquad & x \in \mathbb{R}^n : ||A_i x - b_i||_2 \leq c_i^T x + d_i; i = 1, \ldots, p \\ & A_i \in \mathbb{R}^{(m_i-1)\times n}, b_i \in \mathbb{R}^{(m_i-1)}, c_i, f \in \mathbb{R}^n, d_i \in \mathbb{R} \end{aligned} \tag{7.48}$$

heißt SOCP-Problem (SOCP = second-order cone program) und

$$||A_i x - b_i||_2 \leq c_i^T x + d_i \tag{7.49}$$

SOC-Restriktion der Ordnung m_i.

SOCP-Probleme schließen LP-Probleme und konvexe quadratisch restringierte quadratische Probleme ein (setze diverse Matrizen, Vektoren Null; nutze Epigraph-Form), sind aber weniger allgemein als SDP-Probleme. Tabelle 7.6 aggregiert aus mehreren Arbeiten einige Restriktionen, die sich in SOCs umformen lassen. In Verbindung mit der Epigraph-Form konvexer Probleme ergibt sich über diesen Weg für viele Anwendungen eine SOCP-Formulierung [411]. Hierzu seien nur die robuste LS aus Satz 5.2 und die Log-Chebyshev-Approximation genannt, die im folgenden Beispiel dargestellt ist.

Beispiel 7.10 (Log-Chebyshev-Approximation, [93])
Minimiere die maximale Abweichung in log-Skalierung

$$\max_{1\leq i\leq N} |\log(a_i^T x) - \log b_i| \stackrel{!}{=} \text{Min} \qquad x \in \mathbb{R}^n. \tag{7.50}$$

Klar ist, dass $|\log(a_i^T x) - \log b_i| = \log \max\{a_i^T x/b_i, b_i/a_i^T x\}$ gilt, womit die Zielfunktion (Ausnutzung der Monotonie von log) als $\max_i \max\{a_i^T x/b_i, b_i/a_i^T x\}$ geschrieben werden kann. Somit ergibt sich

$$\begin{aligned} x_{n+1} \stackrel{!}{=} \text{Min} \qquad & a_i^T x/b_i \leq x_{n+1} \\ & b_i/a_i^T x \leq x_{n+1}; 1 \leq i \leq N. \end{aligned} \tag{7.51}$$

	Bedingung	Umformung
1	$a^T x + b \leq 0$ lineare Restriktion	$0 \leq -a^T x - b$ Hier ist in (7.49) $A_i = 0$, $b_i = 0$ zu setzen.
2	$\lvert a^T x - b\rvert \leq y$ $a, x \in \mathbb{C}^n, b \in \mathbb{C}$	$\left\| \begin{bmatrix} \Re e\, a^T & -\Im m\, a^T \\ \Im m\, a^T & \Re e\, a^T \end{bmatrix} \begin{bmatrix} \Re e\, x \\ \Im m\, x \end{bmatrix} + \begin{bmatrix} \Re e\, b \\ \Im m\, b \end{bmatrix} \right\|_2 \leq y$
3	$\lvert\lvert x\rvert\rvert_2^2 \leq y$	$\left\| \begin{bmatrix} 2x \\ 1 - y \end{bmatrix} \right\|_2 \leq 1 + y$
4	$1/y \leq c^T x + d,\ y \geq 0$	$\left\| \begin{bmatrix} 2 \\ c^T x + d - y \end{bmatrix} \right\|_2 \leq c^T x + d + y,\ 0 \leq y$
5	$x^T A x \leq yz, y \geq 0, z \geq 0$ $A \succeq 0_{n\times n}, \mathrm{rg} A = r$ hyperbolische Restriktion	$\left\| \begin{bmatrix} 2Lx \\ y - z \end{bmatrix} \right\|_2 \leq y + z$ $L \in \mathbb{R}_r^{r\times n} : L^T L = A$; L ist Cholesky-Faktor
6	$x^T A x + 2b^T x + c \leq 0$ $A \succeq 0_{n\times n}, \mathrm{rg} A = r$ konvexe quadratische Restriktion	$\left\| \begin{bmatrix} 2Lx \\ 1 + c + 2b^T x \end{bmatrix} \right\|_2 \leq 1 - c - 2b^T x$ $L \in \mathbb{R}_r^{r\times n} : L^T L = A$
7	$x^T A x + 2b^T x + c \leq 0$ $A \succ 0_{n\times n}$ streng konvexe quadratische R.	$\lVert A^{1/2} x + A^{-1/2} b\rVert_2 \leq \sqrt{b^T A^{-1} b - c}$
8	$\begin{bmatrix} a(x) I_n & b(x) \\ b^T(x) & a(x) \end{bmatrix} \succeq 0_{(n+1)\times(n+1)}$ spezielle LMI-Restriktion	$\lvert\lvert b(x)\rvert\rvert_2 \leq a(x)$ $a(x), b(x)$ affine Funktionen in x

Tabelle 7.6: Äquivalente Umformungen in SOC-Restriktionen, [93], [475], [71]

Mit der Umformregel 4 aus Tabelle 7.6 ($y := a_i^T x / b_i$) folgt schließlich das SOCP-Problem

$$x_{n+1} \stackrel{!}{=} \text{Min} \qquad \begin{array}{l} a_i^T x / b_i \leq x_{n+1} \\ \left\| \begin{bmatrix} 2 \\ x_{n+1} - a_i^T x / b_i \end{bmatrix} \right\|_2 \leq x_{n+1} + a_i^T x / b_i; ; 1 \leq i \leq N. \end{array} \tag{7.52}$$

In den Beispielen des gesamten Kapitels und seinen Tabellen zeigt sich einmal mehr, dass sich kompliziert erscheinende Zielfunktionen und Restriktionen durch Transformationen und Umformungen beseitigen oder zumindest vereinfachen lassen. Wichtig hierfür ist die Kenntnis darüber, für welche Problemklassen zugeschnittenen Algorithmen existieren (s. Übersichten in Abschnitt A.4) und ob sie für die Anwendung in Programmsystemen verfügbar sind (einfache Rechercheaufgabe). Mit dieser Kenntnis kann das Umformen zielführender erfolgen, es erfordert aber dennoch einige Erfahrung. Liegt diese Erfahrung nicht vor, können nacheinander die im Kapitel angeführten Techniken auf ihre Eignung geprüft werden. Sollten Restriktionen mit linearen und/oder quadratischen Restriktionen auftreten, sind SDP-Formulierungen

zu erwägen. Vorsicht ist allerdings geboten, wenn die im quadratischen Term auftretende Matrix A indefinit ist. Dann ist die zulässige Menge nicht mehr konvex und die Aussagen zur Eindeutigkeit laut Abschnitt 4.1.2 greifen nicht. Letztlich kann empfohlen werden, bei Recherchen zu Lösungen des technischen Problems die Suchbegriffe um mathematische zu ergänzen, z. B. durch „semidefinite programming“ oder „LMI“, um auf artververwandte Publikationen aufmerksam zu werden.

Kapitel 8

Problemmodifikationsmethoden

In Kapitel 7 wurden Transformationen in vereinfachte Probleme diskutiert, aus denen der Minimierer des Originalproblems folgt. Bei der Problemmodifikation hingegen wird sich mit einer Näherung für den Minimierer begnügt, wenn dafür ein einfacheres Problem entsteht (weniger Nebenminima, weniger Restriktionen, keine diskreten Variablen, schneller, numerisch stabiler usw.). Der Nachteil einer genähertem Lösung fällt in der Praxis vielfach nicht so sehr ins Gewicht, da die Lösungen der modifizierten Probleme oft nahe der optimalen Lösung liegen. Außerdem bleibt bei Offline-Anwendungen stets die Möglichkeit einer Nachoptimierung mit der genäherten Lösung als Startwert. Bei Online-Anwendungen in der Regelungstechnik reduziert eine genäherte Lösung zwar die Performance etwas (z. B. weniger genauer Vorsteueranteil), allerdings ist sie dank der Modifikation realisierbaren Modelladaption im Allgemeinen immer noch höher als bei konventioneller robuster Reglerauslegung.

Anwendungen finden Problemmodifikationsmethoden bei bestimmten nichtlinearen Restriktionen, nichtdifferenzierbaren Funktionen, der Identifikation mit diskreten Variablen (Ordnungen, Grade, Indizes, Dimensionen, Ränge, diskrete Totzeiten) und der Parameterschätzung in hybriden Systemen (Kombination aus Steuerungskomponenten und klassische Differenzialgleichungen). Wie typische Modifikationen aussehen, was für Konsequenzen sich daraus ergeben und welche Methoden der Ingenieur kennen sollte, darauf gibt das Kapitel Antworten. Der Schwerpunkt liegt allerdings – wie in den anderen Kapiteln auch – auf der Bereitstellung wichtiger mathematischer Grundlagen und Zusammenhänge sowie der Aufbereitung neuer Ergebnisse, die in der Ingenieurliteratur bisher weitgehend unbekannt sind.

Abschnitt 8.1 widmet sich Relaxationen, bei denen ein Minimierungsproblem

- durch Vergrößern der zulässigen Menge (Weglassen von Restriktionen oder Umhüllen der zulässigen Menge durch eine größere Menge) oder
- durch Ändern der Zielfunktion hin zu einer Minorante

vereinfacht wird. Dabei wird das Minimum der Modifikation das des Originalproblems nicht übersteigen. Von besonderem Interesse sind natürlich solche Relaxationen, bei denen die Minima von Originalproblem und Modifikation übereinstimmen, s. Abschnitt 8.1.2.

Von den Relaxationszugängen bedarf das Weglassen von Restriktionen keiner Erklärung. Das Umhüllen ist ebenfalls anschaulich, wobei eine gute Relaxation natürlich eine enge Hülle erfordert. Zweckmäßig ist die konvexe Hülle. Leider lässt sie sich aber nur für ausgewählte Mengen direkt angeben. Aus diesem Grund wurden Algorithmen entwickelt [354], die eine Folge kompakter konvexer Mengen $\mathcal{C}_k$ generieren, die asymptotisch gegen die konvexe Hülle von $\mathcal{F}$ konvergieren, d. h.[1]

$$\text{conv}\mathcal{F} \subseteq \mathcal{C}_{k+1} \subseteq \mathcal{C}_k \qquad \text{und} \qquad \bigcap_{k=1}^{\infty} \mathcal{C}_k = \text{conv}\mathcal{F}. \tag{8.1}$$

Die Konstruktion einer Minorante ist meist schwierig. Mit der in der Lagrange-Relaxation genutzten Technik gelingt es allerdings, relativ formal eine Minorante zu konstruieren. Ob dann aber die gewünschte Vereinfachung möglich wird, hängt entscheidend vom Originalproblem und der entstehenden Minorante ab.

In Abschnitt 8.1.3 wird mit der LP-Relaxation ein Beispiel für eine Relaxation basierend auf einer Umhüllung und mit der Rang-1-Relaxation eine für das Weglassen einer Restriktion gezeigt. Die Lagrange-Relaxation, die ein Lagrange-duales Problem nutzt, wird in Abschnitt 8.1.4 vorgestellt. Für einige Probleme kann gezeigt werden, dass die Extrema der Relaxation und des Originals übereinstimmen. Dann wird von starker Dualität gesprochen, die Gegenstand von Abschnitt 8.1.5 ist.

Der Gegenspieler zur Relaxation ist die Kontraktion, bei der die zulässige Menge eingeschränkt wird und/oder im Fall der Minimierung eine Majorante bestimmt wird. Das Ziel ist es wiederum, ein einfacher zu lösendes Problem zu erhalten. Durch wiederholte Anwendung der Kontraktion lassen sich ebenso wie bei der Relaxation konvergente Algorithmen ableiten. Leider zeigt sich, dass einige der in den Ingenieurwissenschaften und der Statistik eingesetzte Algorithmen zwar konvergieren, aber nicht notwendigerweise gegen den richtigen Wert. Sie werden deshalb und wegen ihrer meist nur linearen Konvergenz kaum betrachtet. Nichtsdestotrotz erweisen sie sich in der Praxis als nützlich und tauglich. Deshalb sollen hier einige populäre Zugänge in Abschnitt 8.2 vorgestellt werden, wenngleich nicht auf kritische Anmerkungen verzichtet werden kann.

Der Vorteil, durch Relaxation ein einfacheres Problem zu lösen, wird mit dem Nachteil einer eventuell außerhalb der zulässigen Menge liegenden Lösung erkauft. Es ist dann Standard,

[1] Die konvexe Hülle einer Menge ist die kleinste konvexe Menge, die Menge $\mathcal{F}$ umschließt, also alle Randpunkte mit enthält. Sie wird mit conv$\mathcal{F}$ bezeichnet. Von drei Punkten ist die konvexe Hülle die durch die Punkte gegebene Dreiecksfläche; bei einer Kugel mit Loch ist sie die Kugel ohne Loch.

a posteriori durch eine metrische Projektion auf die zulässige Menge eine approximative Lösung für das Originalproblem zu bestimmen. Aber auch einige kontraktionsbasierte Algorithmen nutzen die Projektionen, weil für diese Existenz- und Eindeutigkeitsaussagen gut möglich sind (Satz 4.2 und Korollar 4.1) und weil für viele Probleme geschlossene und damit schnell und einfach berechenbare Lösungen existieren, vgl. den Projektionssatz 4.9 und die Umformungen in Abschnitt 6.5 für Projektionen in der 1- oder ∞-Norm. Neben dem klassischen Matrixapproximationsproblem existieren zur A-posteriori-Behandlung alternative Formulierungen, auf die in Abschnitt 8.3 kurz eingegangen wird.

8.1 Relaxation

8.1.1 Prinzip und Regeln zu Relaxation und Kontraktion

Bei einer Relaxation wird ein schwieriges Problem durch eine leichteres (entspanntes) ersetzt. Zwei offensichtliche Möglichkeiten bieten sich an: Vergrößere die zulässige Menge $\mathcal{F}$ durch eine Obermenge Φ und/oder ersetze die Zielfunktion $f(x)$ auf der zulässigen Menge durch eine Minorante $\phi(x)$ beim Minimieren (Majorante beim Maximieren).

Definition 8.1 (Relaxation)
Ein Ersatzproblem $\min\{\phi(x) : x \in \Phi\}$ heißt Relaxation von $\min\{f(x) : x \in \mathcal{F}\}$, falls

$$\mathcal{F} \subseteq \Phi \qquad \text{und} \qquad \phi(x) \leq f(x) \text{ für alle } x \in \mathcal{F}. \tag{8.2}$$

Für Maximierungsprobleme ist das Ungleichungszeichen umzukehren. Ist das Ersatzproblem linear, semidefinit oder konvex, dann wird von einer LP(SDP, CP)-Relaxation gesprochen. Bezeichnungen nach Personen (Shor-Relaxation, Lagrange-Relaxation usw.), nach dem Vorgehen (Big-M-Relaxation[2]) oder nach den weggelassenen Restriktionen (Integer-Relaxation, Rangrelaxation) sind ebenso üblich.

Durch die Relaxation ergeben sich bei der Minimierung untere Schranken (auch duale Schranken genannt), denn $\phi_{\text{opt}} \leq f_{\text{opt}}$ gilt immer und $\phi_{\text{opt}} = f_{\text{opt}}$ manchmal, was von besonderem Interesse ist. Um möglichst enge duale Schranken zu erhalten, sollte Φ eine möglichst gute äußere Approximation von $\mathcal{F}$ sein.

Obere (primale) Schranken ergeben sich durch Einsetzen eines beliebigen zulässigen x in die Zielfunktion $f(x)$. Allerdings sind derartige Schranken ohne A-priori-Information oder

[2] Die Big-M-Relaxation ist für disjunktive Restriktionen geeignet. $\bigvee_{i=1}^{q}(g_i(x) \leq 0)$ ist äquivalent zu $g_i(x) \leq M(1-y_i); y_i \in \{0,1\}, \sum_{i=1}^{p} y_i = 1$, wobei $M \geq \max_{x\in\mathcal{X}} g_i(x)$ sein muss. Ein hinreichend großes M bewirkt also, dass bis auf eine Restriktion alle anderen redundant sind [474]. Die y_i lassen sich zudem durch $0 \leq y_i \leq 1$ relaxieren.

Heuristiken meist wertlos, da sie zu konservativ sind. Eine Alternative eröffnen vereinfachte Probleme, die durch Einschränkung des zulässigen Bereichs (Kontraktion, Anspannung, innere Approximation) und/oder leicht auswertbare Majoranten entstehen.

Definition 8.2 (Kontraktion)
Ein Ersatzproblem $\min\{\psi(x) : x \in \Psi\}$ heißt Kontraktion von $\min\{f(x) : x \in \mathcal{F}\}$, falls

$$\mathcal{F} \supseteq \Psi \qquad \text{und} \qquad \psi(x) \geq f(x) \text{ für alle } x \in \mathcal{F}. \tag{8.3}$$

Für Maximierungsprobleme ist das Ungleichungszeichen umzukehren.

Eine viel genutzte Form der Kontraktion ist die Elementfixierung, bei der r Komponenten von x festgehalten werden und somit nicht für die Minimierung verfügbar sind. Es gilt dann $\Psi = \mathcal{F} \cap \{x_{i_j} = x_{i_j}^{fix}; 1 \leq i_1 \leq \ldots \leq i_r \leq n\}$. Der Fall $r = n - 1$ ist kennzeichnend für Blockabstiegsverfahren. Ebenso kann die Suche entlang einer Abstiegsrichtung nach dem Majorisierungsprinzip [487] als spezielle Kontraktion interpretiert werden.

Aus den Definitionen 8.1 und 8.2 ergeben sich unmittelbar zahlreiche offensichtliche Regeln, die hier vom Autor in kompakter Form zusammengefasst werden:

$$\mathcal{X}_1 \subseteq \mathcal{X}_2 \Rightarrow \inf_{x \in \mathcal{X}_1} f(x) \geq \inf_{x \in \mathcal{X}_2} f(x) \tag{8.4a}$$

$$\mathcal{X}_1 \subseteq \mathcal{X}_2 \Rightarrow \sup_{x \in \mathcal{X}_1} f(x) \leq \sup_{x \in \mathcal{X}_2} f(x) \tag{8.4b}$$

$$\inf_{x \in \mathcal{X}_1 \cap \mathcal{X}_2} f(x) \geq \max\{\inf_{x \in \mathcal{X}_1} f(x), \inf_{x \in \mathcal{X}_2} f(x)\} \tag{8.4c}$$

$$\sup_{x \in \mathcal{X}_1 \cap \mathcal{X}_2} f(x) \leq \min\{\sup_{x \in \mathcal{X}_1} f(x), \sup_{x \in \mathcal{X}_2} f(x)\} \tag{8.4d}$$

$$\inf\{f(x) : \bigcap_{i=1}^{p} \mathcal{F}_i\} \geq \inf\{f(x) : \bigcap_{\substack{i=1 \\ i \neq j}}^{p} \mathcal{F}_i\} \tag{8.4e}$$

$$\sup\{f(x) : \bigcap_{i=1}^{p} \mathcal{F}_i\} \leq \sup\{f(x) : \bigcap_{\substack{i=1 \\ i \neq j}}^{p} \mathcal{F}_i\} \tag{8.4f}$$

$$\inf_{x \in \mathcal{X}}(f(x) + g(x)) \geq \inf_{x \in \mathcal{X}} f(x) + \inf_{x \in \mathcal{X}} g(x) \tag{8.4g}$$

$$\sup_{x \in \mathcal{X}}(f(x) + g(x)) \leq \sup_{x \in \mathcal{X}} f(x) + \sup_{x \in \mathcal{X}} g(x) \tag{8.4h}$$

$$f(x) \geq g(x) \Rightarrow \inf_{x \in \mathcal{X}} f(x) \geq \inf_{x \in \mathcal{X}} g(x) \tag{8.4i}$$

$$f(x) \geq g(x) \Rightarrow \sup_{x \in \mathcal{X}} f(x) \geq \sup_{x \in \mathcal{X}} g(x) \tag{8.4j}$$

$$f(x) \leq g(x) \Rightarrow \inf_{x \in \mathcal{X}} f(x) \leq \inf_{x \in \mathcal{X}} g(x) \tag{8.4k}$$

$$f(x) \leq g(x) \Rightarrow \sup_{x \in \mathcal{X}} f(x) \leq \sup_{x \in \mathcal{X}} g(x) \tag{8.4l}$$

Einzig für (8.4g) und (8.4h) ist die Relaxation nicht sofort offensichtlich, doch zeigt

$$\inf_{x\in\mathcal{X}}(f(x)+g(x)) = \inf_{\substack{x\in\mathcal{X}\\ y\in\mathcal{X}\\ x=y}}(f(x)+g(y)), \tag{8.5}$$

dass die rechte Seite von (8.4g) durch Weglassen von $x = y$ eine Relaxation auf $\mathcal{X}\times\mathcal{X}$ ist.

Beispiel 8.1 (Relaxation nichtkonvexer Restriktionen)
Mitunter treten konvexe und nichtkonvexe Funktionen gemeinsam auf, z. B.

$$\gamma_1 < g_1(\theta) < \gamma_2 \qquad g_1 \text{ nichtkonvexe Funktion} \tag{8.6a}$$
$$\gamma_3 < g_1(\theta) + g_2(\theta) < \gamma_4 \qquad g_2 \text{ konvexe Funktion.} \tag{8.6b}$$

Eine Elimination von $g_1(\theta)$ liefert schwächere, dafür aber konvexe Restriktionen

$$\gamma_1 + g_2(\theta) < \gamma_4 \qquad \text{konvex} \tag{8.7a}$$
$$\gamma_3 < \gamma_2 + g_2(\theta) \qquad \text{konvex.} \tag{8.7b}$$

Als eine Anwendung sei die Ableitung einer notwendigen Bedingung für die Schur-Stabilität einer (2×2)-Matrix genannt, für die $|\det A| < 1$ und $|\text{spur } A| < 1 + \det A$ notwendig und hinreichend ist. Hier ist $|\det A| < 1$ nichtkonvex. Mit der Abschätzung $1+\det A \le 1+|\det A|$, die für Abtastsysteme keineswegs ungünstig ist ($\det A \ge 0$ bei zweckmäßiger Wahl der Abtastzeit, vgl. Abschn. 2.2.10), verbleibt nach der Elimination die konvexe notwendige Stabilitätsbedingung $|\text{spur } A| < 2$.

8.1.2 Relaxation ohne Auswirkung

Besonders vorteilhaft für den Ingenieur sind Relaxationen, bei denen das vereinfachte Problem den gleichen Optimalwert annimmt, also $f_{\text{opt}} = \phi_{\text{opt}}$ gilt. Dann muss sich nicht mit einer genäherten Lösung begnügt werden und ein gegebenenfalls notwendiges iteratives Nachbessern der Lösung mit dem Startwert der genäherten Lösung entfällt. Wichtige Relaxationen mit der Eigenschaft $f_{\text{opt}} = \phi_{\text{opt}}$ werden deshalb nachfolgend angeführt:

1. Relaxation von redundanten Restriktionen
Formal handelt es sich beim Weglassen redundanter Restriktionen um eine Relaxation i. S. von Definition 8.1. Dennoch ist es unüblich, das entstehende Ersatzproblem als eine Relaxation zu bezeichnen, da es sich nicht wirklich entspannt hat.

2. Relaxation von Ungleichungen, die im Extremum inaktiv sind
Meist sind im Extremum des Originalproblems nicht alle Ungleichungen gleichzeitig aktiv. Formal könnten also bei der Relaxation alle in x_{opt} inaktiven Ungleichungen weggelassen

werden, ohne dass sich x_{opt} und f_{opt} ändern. Das Problem dabei ist, vorab zu wissen, welche Ungleichungen inaktiv sein werden. Erfahrung, A-priori-Wissen und Heuristiken helfen hier. Gestützt wird dies durch folgenden Satz, der sich direkt aus den Karush-Kuhn-Tucker-Bedingungen ableiten lässt [71].

Satz 8.1 (Relaxation inaktiver Ungleichungen, [71])
Ist x_{opt} globaler Minimierer von

$$f(x) \stackrel{!}{=} \text{Min} \qquad g_1(x) \leq 0, g_2(x) \leq 0_{p-1}, h(x) = 0_m \tag{8.8}$$

mit $g_1(x_{\text{opt}}) < 0$, dann ist x_{opt} lokaler, aber nicht notwendig globaler Minimierer von

$$f(x) \stackrel{!}{=} \text{Min} \qquad g_2(x) \leq 0_{p-1}, h(x) = 0_m \tag{8.9}$$

Die lokalen Minimierer des einfacheren Problems (8.9) sind somit Kandidaten für globale Minimierer von (8.8), sofern sie $g_1(x_{\text{loc}}) < 0$ erfüllen. Falls für keinen lokalen Minimierer die Bedingung $g_1(x_{\text{loc}}) < 0$ gilt, ist $g_1(x) = 0$ zwingend.

3. Relaxation von Gleichungen
In der Regel sind Relaxationen von Gleichungen nicht möglich. Eine Ausnahme bilden Restriktionen, die gewissermaßen die zulässige Menge „teilen“. So bewirkt $\det X = 1$, dass von den orthogonalen Matrizen nur die Drehungsmatrizen berücksichtigt werden. Da sich $\det X = 1$ unschön mit der Methode der Lagrange-Multiplikatoren behandeln lässt, wird diese Restriktion gern weggelassen. Als Preis dafür sind die lokalen Minimierer des reduzierten Problems zu betrachten, von denen es mindestens zwei gibt, vgl. Beispiel 4.1. Letztlich ist der richtige Minimierer auszuwählen.
Eine andere Ausnahme sind Symmetrierestriktionen, wenn der Gradient des freien Problems per se nur eine symmetrische Lösung zulässt.

Beispiel 8.2 (ML-Schätzer der Normalverteilungsparameter)
Seien $\mathbf{y_i}$; $i = 1, \ldots, N$ stochastisch unabhängige, identisch normalverteilte Zufallsvariablen mit $\mathbf{y}_i \sim N_n(\mu, \Sigma)$ mit $\text{rg}\Sigma = n$, dann folgt

$$f_{(\mathbf{y}_1,\ldots,\mathbf{y}_N)}(\xi_1, \ldots, \xi_N) = \frac{1}{((2\pi)^n \det \Sigma)^{N/2}} \exp\Big(-\frac{1}{2} \sum_{i=1}^{N} (\xi_i - \mu)^T \Sigma^{-1} (\xi_i - \mu) \Big) \tag{8.10a}$$

$$L(\mu, \Sigma) = \frac{1}{((2\pi)^n \det \Sigma)^{N/2}} \exp\Big(-\frac{1}{2} \sum_{i=1}^{N} (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \Big) \tag{8.10b}$$

$$l(\mu, \Sigma) = -\frac{Nn \ln(2\pi)}{2} - \frac{N}{2} \ln \det \Sigma - \frac{1}{2} \text{sp}(\Sigma^{-1} Z(\mu)) \tag{8.10c}$$

$$\text{mit } Z(\mu) = \sum_{i=1}^{N} (y_i - \mu)(y_i - \mu)^T \tag{8.10d}$$

Beim Maximieren der Likelihood-Funktion bzw. äquivalent von $l(\mu, \Sigma)$ sind an $\mu \in \mathbb{R}^n$ keine Restriktionen zu stellen, während für die Schätzung $\hat{\Sigma}_{ML}$ positive Definitheit zu fordern ist. Da die Behandlung einer solchen Restriktion ein wenig komplizierter ist, wird die Restriktion in der Hoffnung weggelassen, dass die freie Lösung von sich aus die Restriktion erfüllt. Aus den Optimalitätsbedingungen erster Ordnung folgt dann

$$\frac{\partial l(\mu, \Sigma)}{\partial \mu} = (\Sigma^{-1})^T (\sum_{i=1}^{N} y_i - N\mu) = 0_n$$
$$\frac{\partial l(\mu, \Sigma)}{\partial \Sigma} = -\frac{N}{2}(\Sigma^{-1})^T + \frac{1}{2}(\Sigma^{-1} Z(\mu) \Sigma^{-1})^T = 0_{n \times n},$$

was $\hat{\mu}_{ML} = \frac{1}{N} \sum_{i=1}^{N} y_i$ und $\hat{\Sigma}_{ML} = \frac{1}{N} Z(\hat{\mu}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{\mu})(y_i - \hat{\mu})^T$ liefert. $\hat{\Sigma}_{ML}$ ist mindestens nichtnegativ definit und für $N > n$ mit Wahrscheinlichkeit Eins positiv definit.

4. Relaxation generischer Restriktionen
Da Mengen mit generischen Eigenschaften (z. B. Vollrang, Steuerbarkeit) offen und dicht sind, schränken sie die zulässige Menge fast nicht ein. Umgekehrt ist klar, dass bei mageren Mengen (z. B. Rang-k-Bedingung) die Restriktion nicht weggelassen werden darf, da sonst fast immer ein Wert aus der Komplementärmenge angenommen wird.

5. Relaxation stochastisch fast sicherer Restriktionen
Hier greift die vorstehende Argumentation allerdings über Wahrscheinlichkeitsmaße.

6. Relaxation kompakter Mengen durch ihre konvexe Hülle

$$\max_{x \in \mathcal{F}} c^T x = \max_{x \in \mathrm{conv}\mathcal{F}} c^T x \qquad \mathcal{F} \text{ ist eine kompakte Menge} \tag{8.11}$$

Für fast jedes $c \in \mathbb{R}^n$ ist der Maximierer des relaxierten Problems ein Extrempunkt der konvexen Hülle conv$\mathcal{F}$. Als solcher ist er ein Element von $\mathcal{F}$ und zugleich Maximierer von $\mathcal{F}$. Der Beweis stützt sich auf das Ewald-Larman-Rogers-Theorem[3] [184].

Beispiel 8.3 (Relaxation bei Permutationsmatrizen)
Repräsentieren A und B gleiche oder ähnliche Daten in unterschiedlicher Anordnung, so kann dies aus dem Prokrustes-Problem

$$||AX - B||_F^2 \stackrel{!}{=} \text{Min} \qquad X \in \mathbb{R}^{n \times n} \text{ Permutationsmatrix} \tag{8.12}$$

erkannt werden. Elementare Umformungen führen auf

$$\mathrm{sp}(X^T B^T A) \stackrel{!}{=} \text{Max} \qquad X \in \mathbb{R}^{n \times n} \text{ Permutationsmatrix.} \tag{8.13}$$

[3] Satz: Für jeden konvexen Körper in $\mathbb{R}^n$ hat die Menge der Einheitsvektoren, die parallel zu einem Segment des Körperrandes liegen, ein endliches $(n-2)$-dimensionales Hausdorff-Maß [187]. Somit kann der Rand eines konvexen Körpers nicht Liniensegmente enthalten, die in alle Richtungen zeigen. Anschaulicher: Bezogen auf eine konvexe ebene Figur werden die durch c bestimmten Höhenlinien fast nie parallel zu einem Randsegment (Strecke) liegen, sodass fast nie ein gesamtes Segment Lösung der Maximierung ist.

Beide Probleme sind nicht konvex. Da die Permutationsmatrizen in den doppelstochastischen Matrizen[4] enthalten sind, kann das relaxierte Problem

$$\mathrm{sp}(X^T B^T A) \stackrel{!}{=} \mathrm{Max} \qquad \underbrace{X \in \mathbb{R}^{n\times n} : x_{ij} \geq 0, X1_n = 1_n, X^T 1_n = 1_n}_{\text{Menge der doppelstochastischen Matrizen}} \tag{8.14}$$

betrachtet werden. Dieses LP-Problem nimmt die Lösung an einer Ecke an. Da die Ecken der doppelstochastischen Matrizen aber gerade die Permutationsmatrizen sind[5], haben das Originalproblem und das relaxierte Problem die gleiche Lösung.

7. Lagrange-Relaxation ohne Dualitätslücke
Siehe Abschnitt 8.1.5.

8.1.3 Ausgewählte Relaxationen

In diesem Abschnitt wird das Vereinfachen (Modifizieren) von Problemen durch Probleme der linearen Optimierung und der semidefiniten Optimierung kurz skizziert, weshalb von LP-Relaxationen oder SDP-Relaxationen gesprochen wird. Etwas präziser wird die LP-Relaxation wie folgt definiert; die Definition für die SDP-Relaxation ist hierzu analog.

Definition 8.3 (LP-Relaxation)
Als LP-Relaxation (abgeleitet von Lineare Programmierung) wird ein Problem bezeichnet, dass $\mathcal{F}$ mit einem Φ überdeckt, das durch lineare Restriktionen in reellen Variablen beschrieben wird. Ist Φ ein Intervall, so wird auch von einer Intervallrelaxation und im Fall eines Hyperquaders von einer Box-Relaxation gesprochen.

Mittels LP-Relaxationen können unter anderem NP-schwere ganzzahlige lineare Optimierungsprobleme in reelle LP-Probleme überführt werden, welche sich in polynomialer Zeit lösen lassen. Nur selten erfüllt die reelle Lösung die Ganzzahligkeitsbedingung des ursprünglichen Problems. Aus der reellen Lösung und deren Extremum folgen Schranken und damit Mengeneingrenzungen für das ursprüngliche Problem. Allerdings ist die Annahme, wonach gerundete oder benachbarte ganzzahlige Lösungen Optimierer seien, im Allgemeinen falsch, wie das folgende Beispiel zeigt.

[4] Eine doppelstochastische Matrix ist entsprechend der Restriktion in (8.12) eine nichtnegative Matrix, deren Zeilen- und Spaltensumme gerade Eins ist. In den Anwendungen sind die Matrizenelemente zumeist Wahrscheinlichkeiten (also nichtnegative Größen) und die Summe-Eins-Bedingungen resultieren aus dem zweiten und dritten Kolmogorov-Axiom.

[5] Satz von Birkhoff: Die doppelstochastischen Matrizen sind die konvexe Hülle der Permutationsmatrizen [72]. Mit anderen Worten: Sie bilden eine kompakte konvexe Menge, das sog. Birkhoff-Polytop, dessen Extrempunkte (Punkte, die sich nicht als streng konvexe Kombination zweier Elemente der Menge darstellen lassen) die Permutationsmatrizen sind [529].

Beispiel 8.4 (Schlechte Relaxation)
Selbst wenn die Maxima von Original und Relaxation relativ dicht liegen, können deren Maximierer weit entfernt sein. So hat

$$5x_1 + 8x_2 \stackrel{!}{=} \text{Max} \qquad x_1 + x_2 \leq 6,\ 5x_1 + 9x_2 \leq 45;\ x_1, x_2 \in \mathbb{N}_{\geq} \tag{8.15}$$

das Optimum $f_{\text{opt}} = 40$ bei $(x_1, x_2)_{\text{opt}} = (0, 5)$, während die LP-Relaxation mit $x_1, x_2 \in \mathbb{R}$ den Wert $\phi_{\text{opt}} = 41.25$ bei $(x_1, x_2) = (2.25, 3.75)$ annimmt. Die gerundete Lösung liegt bei $(2, 4)$ im unzulässigen Gebiet, und die zulässige Nachbarlösung bei $(2, 3)$ liefert $f(2, 3) = 34$.

Anwendungen für Intervall- und Box-Relaxationen sind bevorzugt 2^n-elementige Mengenrestriktionen vom Typ $x \in \{a, b\}^n$. So steht $|x| = a$ für $x \in \{-a, a\}$ und wird durch $-a \leq x \leq a$ überdeckt. Eine Box-Relaxation der binären Zahlen $x \in \{0, 1\}^n$ ist $0_n \leq x \leq 1_n$.

Anmerkung 8.1 Für bestimmte endliche Mengen empfehlen sich von Fall zu Fall auch nichtlineare Relaxationen. So ist $x \in \{0, 1\}$ äquivalent zu $x(x-1) = x^2 - x = 0$, was durch eine quadratische Restriktion $x^2 - x \leq 0$ relaxiert werden kann. Die Trust-Region-Relaxation $||x||_2^2 = n$ überdeckt alle Vektoren aus $x \in \{+1, -1\}^n$.

Anmerkung 8.2 Bei geringer Kardinalität einer endlichen Menge ist die Lösung mehrerer Einzelprobleme für jedes Mengenelement gegenüber einer Relaxation zu favorisieren.

Zahlreiche Probleme lassen sich in eines mit quadratischer Zielfunktion und quadratischen Restriktionen (schließt lineare mit ein) umformen. Hierzu zählen Probleme mit polynomialen Ungleichungen (Splitting-Techniken), Probleme mit Binärrestriktionen (Trust-Region-Relaxation) aber auch geometrische Probleme wie Ellipsen-Schnittprobleme, um- und einbeschriebene Ellipsen (äußere bzw. innere Löwner-John-Ellipsoid). Da diese quadratischen Probleme schwer zu lösen sind, wenn sie nicht konvex sind, werden sie mit Blick auf eine spätere SDP- oder LMI-Formulierung umgeformt.

$$x^T A_0 x + 2b_0^T x + c_0 \stackrel{!}{=} \text{Min} \qquad x^T A_k x + 2b_k^T x + c_k \geq 0; k = 1, \ldots, p \tag{8.16a}$$

$$\text{spur}(A_0 Y) + 2b_0^T x + c_0 \stackrel{!}{=} \text{Min} \qquad \text{spur}(A_k Y) + 2b_k^T x + c_k \geq 0; k = 1, \ldots, p \tag{8.16b}$$
$$Y = xx^T$$

$$\text{spur}(A_0 Y) + 2b_0^T x + c_0 \stackrel{!}{=} \text{Min} \qquad \text{spur}(A_k Y) + 2b_k^T x + c_k \geq 0; k = 1, \ldots, p \tag{8.16c}$$
$$Y \succeq xx^T \qquad \text{Relaxation; konvexe Restriktion}$$

Wegen $Y = xx^T \Leftrightarrow Y \succeq xx^T, \text{rg} Y = 1$ handelt es sich in (8.16c) um eine Rangrelaxation. Weitere Umformungen mit dem Schur-Komplement-Formeln und der Shor-Prozedur führen auf ein SDP-Problem in LMI-Formulierung, weshalb die Relaxation auch Shor-Relaxation, SDP-Relaxation oder LMI-Relaxation des quadratischen Problems (8.16a) genannt wird. Wird (8.16c) in eine Maximierung umgeformt und einer Lagrange-Relaxation unterzogen, entsteht ebenfalls eine LMI-Formulierung, die den gleichen Zielfunktionswert aufweist [280].

8.1.4 Lagrange-Dualität

Dualität ist ein beliebtes Konzept in vielen Wissenschaftsbereichen (Geometrie, Algebra, Optimierung, Systemtheorie). Grob gesagt bedeutet es, dass Kenntnisse des primalen Objektes, Kenntnisse des dualen Objektes implizieren und umgekehrt. In der Regelungstechnik bilden Steuerbarkeit und Beobachtbarkeit ein solches Paar, in der Algebra sind es Normen oder Räume und in der Optimierung zwei „wesensverwandte“ Optimierungsaufgaben. Bekannte Dualitäten sind die Lagrange-Dualität, die Fenchel-Dualität [71] und die Wolfe-Dualität [631]. Nach einer Begriffsklärung werden die Vorteile der Umformung in duale Probleme genannt, um danach kurz in die Lagrange-Dualität einzuführen.

Definition 8.4 (Duales Problem)
Ein duales Problem ist eine Optimierungsaufgabe mit der Eigenschaft, dass ihr Extremum immer eine Schranke für ein Originalproblem, genannt primales Problem, ist.
Vereinfacht und mit anderen Worten: Zu einer Minimierung/Maximierung (primal) gehört eine Maximierung/Minimierung (dual), deren Maximum/Minimum eine untere/obere Schranke für das primale Problem ist.

Die Vorteile für eine derartige Problemmodifikation sind vielschichtig. So kann bei einem nichtkonvexen primalen Problem ein konvexes duales und damit einfacher zu lösendes Problem entstehen. Ferner tauscht sich die Rolle von Variablen und Restriktionen in dem Sinn, dass statt über 1000 Variablen mit 5 Restriktionen im dualen Problem über 5 Variablen mit 1000 Restriktionen zu optimieren ist. Zudem können die Restriktionen einfacher werden, s. z. B. den Wegfall von Gleichungsrestriktionen in Tabelle 8.1 auf Seite 301. Mit der Kenntnis einer guten Schranke aus dem dualen Problem ist bei NP-schweren primalen Problemen zumindest klar, ob ein Fortschreiten des numerischen Lösens des primalen Problems noch eine wesentliche Verbesserung liefern kann oder nicht.
Gelingt der Nachweis, dass die Schranke des dualen Problems dem Extremum des primalen Problems entspricht, dann lässt sich oft aus der dualen Lösung einfach auf die primale schließen. Wurden etwa im Fall der Lagrange-Dualität alle Restriktionen relaxiert (in die Lagrange-Funktion aufgenommen), so ist mit den nunmehr bekannten optimalen dualen Variablen ein freies Problem bezüglich der primalen Variablen zu lösen (u. U. ist eine geschlossene Lösung möglich). Über die Bedingung vom komplementären Schlupf werden die einzubeziehenden aktiven Ungleichungen angezeigt.

Nach dieser einleitenden Motivation für duale Probleme sei eine für beliebige Funktionen $K(x, y)$ gültige Minimax-Ungleichung von von-Neumann [71]

$$\sup_{y\in\mathcal{Y}} \underbrace{\inf_{x\in\mathcal{X}} K(x,y)}_{=K_*(y)} \leq \inf_{x\in\mathcal{X}} \underbrace{\sup_{y\in\mathcal{Y}} K(x,y)}_{=K^*(x)} \tag{8.17}$$

genannt[6], auf der die gängigen Dualitätszugänge beruhen. Eine der beiden Seiten ist dabei stets das primale Problem, die andere das duale. Beide beschränken einander. Indem für $K(x,y)$ eine Lagrange-Funktion $L(x,y)$ gewählt wird, ergeben sich die sogenannten Lagrange-dualen Probleme. Da bei der Lagrange-Funktion die y affin sind, ist das duale Problem stets konvex (minimieren über konvexe Mengen oder maximieren über konkave Mengen). Wird nunmehr berücksichtigt, dass jedes restringierte Minimierungsproblem zu einem Minimax-Problem über seiner Lagrange-Funktion äquivalent ist, vgl. exemplarisch

$$\inf_{\substack{x\in\mathcal{X}\\ g(x)\leq 0_p}} f(x) = \inf_{x\in\mathcal{X}} \left\{ \begin{array}{ll} f(x) & \text{für } g(x)\leq 0_p \\ \infty & \text{sonst} \end{array} \right\} = \inf_{x\in\mathcal{X}} \sup_{y\geq 0_p} \{f(x)+y^T g(x)\} = \inf_{x\in\mathcal{X}} \sup_{y\geq 0_p} L(x,y), \tag{8.18}$$

dann ist in diesem Fall die rechte Seite in (8.17) das primale und die linke das duale Problem.[7] Da die Ungleichung (8.17) für jedes $K(x,y)$ gilt, müssen hier nicht wie bei der Methode der Lagrange-Multiplikatoren alle Restriktionen aufgenommen werden. Es sind lediglich $\mathcal{X},\mathcal{Y}$ entsprechend anzupassen. Letztlich können die Lagrange-dualen Probleme als jeweils beste Lagrange-Relaxation aufgefasst werden. Für jedes zulässige x und beliebige y gilt nämlich

$$L_*(y) \leq \sup_y L_*(y) = Q_D \leq L(x,y) \leq f(x), \tag{8.19}$$

da $y_g^T h_1(x) = 0$ und $y_u^T g_1(x) \leq 0$. Somit ist $L_*(y)$ für jedes y eine Minorante (sehr einfache, da konstante Funktion) für f, und die beste Minorante (Relaxation) unter diesen ist gerade $\sup_y L_*(y)$. Im Gegensatz zum profanen Weglassen von Restriktionen bleiben die auf die beschriebene Weise relaxierten Restriktionen gewichtet in der Zielfunktion berücksichtigt.

Definition 8.5 (Lagrange-Dualität, [71])
Sei ein Originalproblem (primales Problem)

$$Q_P = \inf\{f(x) : g_1(x)\leq 0_{p_1}, h_1(x) = 0_{m_1};\ g_2(x)\leq 0_{p_2}, h_2(x) = 0_{m_2},\ x\in\mathcal{X}\subseteq\mathbb{R}^n\} \tag{8.20}$$

gegeben, dann bezeichnet

$$L(x,y) = f(x) + y_g^T h_1(x) + y_u^T g_1(x), \qquad y^T = (y_g^T, y_u^T) \tag{8.21}$$

die mit den zu relaxierenden Nebenbedingungen g_1, h_1 assoziierte Lagrange-Funktion mit den dualen Variable $y_g \in \mathbb{R}^{p_1}$ und $y_u \geq 0_{m_1}$. Ferner heißt

$$L_*(y) = \inf_{\substack{x\in\mathcal{X}\\ g_2(x)\leq 0_{p_2}\\ h_2(x)=0_{m_2}}} L(x,y) \tag{8.22}$$

[6] $\forall x,y : \inf_{x\in\mathcal{X}} K(x,y) \leq K(x,y) \Rightarrow \forall x : \sup_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} K(x,y) \leq \sup_{y\in\mathcal{Y}} K(x,y) \Rightarrow \sup_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} K(x,y) \leq \inf_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} K(x,y)$.

[7] Für Gleichungen und Maximierungen ergeben sich analoge Beziehungen.

die Lagrange-duale Funktion, mit deren Hilfe das duale Problem

$$Q_D = \sup\{L_*(y) : y_g \in \mathbb{R}^{p_1}, y_u \geq 0_{m_1}\} \tag{8.23}$$

formuliert wird. Die Differenz $Q_P - Q_D$ heißt Dualitätslücke. Handelt es sich beim Originalproblem um eine Maximierung, so ist die Lagrange-duale Funktion $L^*(y) = \sup_x L(x, y)$ durch Maximierung über x zu erhalten und das duale Problem wird zur Minimierung über $L^*(y)$. Da Q_D in diesem Fall eine obere Schranke für Q_P angibt, wird dann als Dualitätslücke $Q_D - Q_P$ definiert. Gilt $Q_P = Q_D$, so wird von starker Dualität zwischen primalem und dualem Problem gesprochen, andernfalls von schwacher Dualität[8].

Beispiel 8.5 (Lagrange-Dualität eines LP-Problems)
Für $c^T x \stackrel{!}{=} \text{Min}; Ax = b, x \in \mathbb{R}^n_{\geq}$ lautet die duale Funktion mit Relaxation der Gleichung

$$L_*(y) = \inf_{x \geq 0} L(x, y) = \inf_{x \geq 0}(c^T x + y^T(Ax - b)) = \left\{ \begin{array}{ll} -y^T b & \text{falls } c^T + y^T A \geq 0_n^T \\ -\infty & \text{sonst} \end{array} \right\}$$

und das duale Problem ist damit

$$\sup_{y \in \mathbb{R}^m} L_*(y) = \max\{-b^T y : A^T y + c \geq 0_n, y \in \mathbb{R}^m\} = \max\{b^T y : A^T y \leq c, y \in \mathbb{R}^m\}.$$

Wird zusätzlich auch $x \geq 0_n$ relaxiert, so ergibt sich über

$$\tilde{L}_*(y_g, y_u) = \inf_{x \geq 0} \tilde{L}(x, y_g, y_u) = \inf_{x \geq 0}(c^T x + y_g^T(Ax - b) - y_u^T x) = \left\{ \begin{array}{ll} -y_g^T b & \text{falls } A^T y_g + c \geq y_u \\ -\infty & \text{sonst} \end{array} \right\}$$

unter der Ausnutzung, dass der zulässige Bereich für y_g umso größer ist, je kleiner y_u ist, mit

$$\sup_{y_g \in \mathbb{R}^m, y_u \geq 0_n} \tilde{L}_*(y_g, y_u) = \sup_{y_g \in \mathbb{R}^m} \tilde{L}_*(y_g, 0_n) = \sup_{y \in \mathbb{R}^m} L_*(y)$$

in diesem Fall die gleiche duale Aufgabe.

Anmerkung 8.3 Das duale Problem stellt immer eine konvexe Optimierung dar (hier Maximierung über konkave Funktion), da $L_*(y)$ als Infimum einer Familie affiner Funktionen immer konkav ist.

Anmerkung 8.4 Ist das primale (duale) Problem unbeschränkt, so hat das duale (primale) Problem keine Lösung. So impliziert z. B. $\inf_x \sup_y L(x, y) = -\infty$ in (8.17) (unbeschränkte Minimierung), dass $\sup_y \sup_x L(x, y) = -\infty$ gilt, was wiederum $\mathcal{Y} = \emptyset$ erzwingt.

[8] Da die Fälle $Q_P = Q_D = +\infty$ bzw. $Q_P = Q_D = -\infty$ möglich sind, wird zur Definition der starken Dualität $Q_P = Q_D$ statt der Forderung „Dualitätslücke gleich Null“ herangezogen.

Anmerkung 8.5 Das duale Problem ist nicht intrinsisch. Anders ausgedrückt: Das duale Problem und sein Extremum sind weder eine Eigenschaft der primal zulässigen Menge noch der Zielfunktion selbst. Es hängt von der speziellen Formulierung ab. Äquivalente primale Probleme, die durch monotone Transformationen der Zielfunktion, das Einführen neuer Variablen, z. B. bei $(x_1-x_2)^2+(x_2-x_3)^2 \stackrel{!}{=} \text{Min}$ zu $(x_1-x_2)^2+(x_4-x_3)^2 \stackrel{!}{=} \text{Min}; x_2 = x_4$, oder die Hinzunahme redundanter Restriktionen entstehen, können also durchaus unterschiedliche Duale haben.
Die Hinzunahme redundanter Ungleichungsrestriktionen kann somit helfen, die Dualitätslücke zu verringern oder gar zu schließen. Diese Technik wird gern bei der ganzzahligen Optimierung genutzt und ist dort unter den Begriffen „gültige Schnittebene“ und „valid inequality“ bekannt. Noch faszinierender ist das Schließen der Dualitätslücke durch Hinzunahme einer redundanten Gleichungsrestriktion, s. $X^TX = I_n$ (Orthogonalität) und die redundante Restriktion $XX^T = I_n$ in [27].

Anmerkung 8.6 Bei einer Lagrange-Relaxation über endlichen Mengen $\mathcal{X}$ werden die Ungleichungsrestriktionen relaxiert, und es wird ein freies Problem über der endlichen Menge gelöst. Das geht mit speziellen Algorithmen (Kürzester Pfad, Maximaler Fluss, Minimaler Kostenfluss usw.) recht flott. Gegenüber der Integer-Relaxation mittels konvexer Hülle und Beibehaltung der Ungleichungen liefert die Lagrange-Relaxation bessere Schranken. Bei linearen Zielfunktionen und lineare Restriktionen sind beide Schranken gleich [71].

8.1.5 Starke Dualitätstheoreme

Bei Relaxationen sind jene Fälle mit starker Dualität ($Q_P = Q_D$) von besonderem Interesse. Noch wichtiger sind aber die Fälle, in denen zudem das duale Problem eine Lösung hat, aus der auf die des primalen geschlossen werden kann. Dann liegt nämlich nicht nur eine Problemmodifikation, sondern sogar eine Problemtransformation vor.

Da die dualen Variablen y in gleicher Weise in der Lagrange-Funktion auftreten wie die Lagrange-Multiplikatoren λ, μ bei der klassischen Optimierung, werden sie mitunter auch Lagrange-Multiplikatoren genannt, was aber unzweckmäßig ist. Ein Lagrange-Multiplikator zielt unter Differenzierbarkeitsannahmen auf ein lokales Minimum ab, eine duale Variable über die Lagrange-duale Funktion dagegen auf ein globales Minimum ohne eine Forderung nach Differenzierbarkeit. Für optimale duale Variable sollte deshalb der Begriff „Geometrischer Multiplikator“ verwendet werden.

Definition 8.6 (Geometrischer Multiplikator, [71])
Ein Vektor $\bar{y} = (\bar{y}_g^T, \bar{y}_u^T)^T$ mit $\bar{y}_u \geq 0_p$ heißt Geometrischer Multiplikator, wenn $Q_P = \inf_x L(x, \bar{y})$ gilt. $\bar{\mathcal{Y}}$ ist die Menge aller Geometrischen Multiplikatoren.

Es bestehen nun folgende Zusammenhänge:[9]

1. $Q_D = Q_P \Rightarrow \bar{\mathcal{Y}} = \mathcal{Y}_{\text{opt}}$ fundamentaler Zusammenhang
   In Worten: Bei starker Dualität sind die dualen Optimallösungen Geometrische Multiplikatoren und $y_{u,\text{opt}}^T g(x_{\text{opt}}) = 0$ gilt.
2. $Q_D < Q_P \Rightarrow \bar{\mathcal{Y}} = \emptyset$, d. h. sobald ein $\bar{y}$ existiert, ist $Q_D = Q_P$
3. $\bar{\mathcal{Y}}$ ist gleich der Menge der Lagrange-Multiplikatoren für konvexe eindeutige $\mathcal{C}^1$-Probleme
4. $\bar{y}$ hängt bei konvexen Problemen nicht von $x_{\text{opt}} \in \mathcal{X}_{\text{opt}}$ ab, wohl aber $(\lambda_{\text{opt}}, \mu_{\text{opt}})$
5. $\bar{y}$ kann existieren, ohne dass x_{opt} existiert, vgl. $\inf_{x \geq 0} 1/x$ mit $\bar{y} = 0$
6. $\bar{y}$ braucht nicht zu existieren, wohl aber x_{opt} /[10]
7. $\min_{x=0}(-x^2)$ hat ein $\lambda_{\text{opt}} = 0$, aber kein y_{opt} und kein $\bar{y}$
8. x_{opt} löst (8.20) $\Leftrightarrow x_{\text{opt}} = \underset{x \in \mathcal{X}}{\text{argmin}}\ L(x, \bar{y})$ und $\bar{y}_u^T g(x_{\text{opt}}) = 0$ und x_{opt} zulässig für (8.20)

Die Zulässigkeitsforderung in Punkt 8 der Zusammenhänge ist dabei essenziell, denn das Minimieren über $L(x, y_{\text{opt}})$ kann auch Lösungen liefern, die nicht das Originalproblem lösen, vgl. Beispiel 8.6. Das Beispiel zeigt zudem, das starke Dualität keine ausschließlich an konvexe Probleme geknüpfte Eigenschaft ist, wenngleich das Starke Dualitätstheorem für konvexe Probleme natürlich von zentraler Bedeutung ist. Es folgt direkt nach dem Beispiel.

Beispiel 8.6 (Nichtkonvexes Problem ohne Dualitätslücke)
Das Rayleigh-Quotiententheorem [299] liefert $\lambda_{\min}(A) = \min\{x^T A x : x^T x = 1\}$ für symmetrisches A, wobei alle auf Eins normierten Eigenvektoren zu $\lambda_{\min}(A)$ Lösungen sind. Das duale Problem lautet

$$\begin{aligned}\sup_{y \in \mathbb{R}} \inf_{x \in \mathbb{R}^n} (x^T A x - y(x^T x - 1)) &= \sup_{y \in \mathbb{R}} \left\{ \begin{array}{ll} y & \text{falls } A - yI_n \succeq 0_{n \times n} \\ -\infty & \text{sonst} \end{array} \right\} \\ &= \max\{y \in \mathbb{R} : A - yI_n \succeq 0_{n \times n}\} \\ &= \lambda_{\min}(A),\end{aligned}$$

wobei in der ersten Zeile das Infimum für $x_{\text{opt}} = 0_n$ im definiten Fall und zusätzlich für alle $x \in \mathcal{N}(A - yI_n)$ im semidefiniten Fall angenommen wird. Das duale Problem weist keine Dualitätslücke auf. Die implizite (versteckte) Restriktion $A - yI_n \succeq 0_{n \times n}$ erscheint über den „echten“ Definitionsbereich von $L_*(y)$. Korrespondierend zu $y_{\text{opt}} = \lambda_{\min}(A)$ lösen alle Eigenvektoren zu $\lambda_{\min}(A)$ und der Nullvektor das duale Problem, wobei aber der Nullvektor keine Lösung des primalen Problems ist.

Satz 8.2 (Starkes Dualitätstheorem – konvexe Restriktionen, [71])
$f(x) \stackrel{!}{=} \text{Min};\ x \in \mathcal{C} \subseteq \mathbb{R}^n,\ Ax = b,\ Cx \leq d,\ g_i(x) \leq 0;\ i = 1, \ldots, p$ mit f, g_i konvex (nicht

[9] Entnommen aus der Vorlesung (Lektion 18) von Bertsekas aus 2004:
dspace.mit.edu/bitstream/handle/1721.1/70523/6-253-spring-2004/contents/lecture-notes/lec_18.pdf

[10] $0 = \min\{x : x^2 \leq 0\}$ mit $x_{\text{opt}} = 0$ und $0 = \sup_{y \geq 0} \min_x x + yx^2$ mit $x_{\text{opt}} = -1/(2y)$ und $y_{\text{opt}} \to \infty$

notwendig differenzierbar) über konvexem $\mathcal{C}$, f_{opt} endlich und einem zulässigen $\tilde{x} \in \operatorname{ri}\mathcal{C}$[11] mit $g_i(\tilde{x}) < 0; i = 1, \ldots, q$ (Slater-Constraint-Qualification) hat keine Dualitätslücke zu seinem Lagrange-Dual, und es existiert mindestens ein Geometrischer Multiplikator.

Anmerkung 8.7 Ohne eine Constraint-Qualification können bei konvexen Problemen Dualitätslücken auftreten, vgl. Gegenbeispiel in [71]. Die Slater-Bedingung ist nur hinreichend. Schwächere sowie notwendige und hinreichende Bedingungen (i. d. R. mathematisch und algorithmisch weniger gut handhabbar) werden in [84] gegeben. Ohne g_i kann auf eine Constraint-Qualification verzichtet werden, wenn f_{opt} endlich ist und f über $\mathbb{R}^n$ konvex ist [71]. Für spezielle Problemklassen ergeben sich die nachfolgenden aufgeführten weitergehenden Aussagen. Ergänzend sei auf [420] für konvexe quadratische Probleme verwiesen.

Satz 8.3 (Allgemeiner Dualitätssatz für LP-Probleme, [153], [154])
Für primale (P) und duale (D) LP-Probleme in Tab. 8.1 gelten die folgenden Aussagen:

- P und D haben zulässige Lösungen ⇒ P und D haben Optimallösungen
- P und D haben zulässige Lösungen ⇒ Extrema sind gleich (nutze Satz 8.2)
- P (D) hat Optimallösung ⇒ D (P) hat Optimallösung
- P (D) hat keine zulässige Lösung ⇒ D (P) ist unbeschränkt oder D (P) hat keine Lösung
- P (D) ist unbeschränkt ⇒ D (P) hat keine zulässige Lösung (s. Anm. 8.4)

primales LP-Problem	duales LP-Problem
$\max\{c^T x : Ax \leq b, x \geq 0_n\}$	$\min\{b^T y : A^T y \geq c, y \geq 0_p\}$
$\min\{c^T x : Ax \geq b, x \geq 0_n\}$	$\max\{b^T y : A^T y \leq c, y \geq 0_p\}$
$\max\{c^T x : Ax = b, x \geq 0_n\}$	$\min\{b^T y : A^T y \geq c\}$
$\min\{c^T x : Ax = b, x \geq 0_n\}$	$\max\{b^T y : A^T y \leq c\}$
$\max\{c^T x : Ax \leq b\}$	$\min\{b^T y : A^T y = c, y \geq 0_p\}$
$\min\{c^T x : Ax \geq b\}$	$\max\{b^T y : A^T y = c, y \geq 0_p\}$

Tabelle 8.1: Primale und duale LP-Probleme [256]

[11] $\operatorname{ri}\mathcal{C}$ steht für das relative Innere der Menge $\mathcal{C}$, also für jene inneren Punkte von $\mathcal{C}$, die gleichzeitig $Ax = b$ erfüllen. Beispiel: Schneidet eine Ebene eine Kugel, so ist die Schnittfläche ein Kreis. Das Innere des Kreises ist dann das relative Innere.

Satz 8.4 (Generisch starke Dualität bei konvexen Kegelrestriktionen, [555])
Sei $\mathcal{C}$ ein properer Kegel[12], dann gilt für[13]

$$Q_P = \inf_x \{\langle c, x\rangle : Ax = b, Cx \succeq_{\mathcal{C}} d\} \tag{8.24a}$$

$$Q_D = \sup_{y_g, y_u} \{\langle b, y\rangle : A^* y_g + C^* y_u = c, y_u \succeq_{\mathcal{C}^\oplus} 0_n\} \tag{8.24b}$$

$Q_P = Q_D$ generisch[14]. Ferner liegt unter diversen Bedingungen an $(A, C, \mathcal{C})$ universelle Dualität vor, d. h., für beliebige b, d besteht dann (nicht nur generisch) starke Dualität.

Korollar: Dualität für SDP-Probleme:

$$Q_P = \inf_X \{\mathrm{sp}(CX) : \mathrm{sp}(A_i X) = b_i, X \succeq 0_{n\times n}\} \tag{8.25a}$$

$$Q_D = \sup_y \{b^T y : \underbrace{C - \sum_{i=1}^m y_i A_i}_{=Z} \succeq 0_{n\times n}\} \tag{8.25b}$$

mit $\mathrm{sp}(X_{\mathrm{opt}} Z_{\mathrm{opt}}) = 0 \Leftrightarrow X_{\mathrm{opt}} Z_{\mathrm{opt}} = Z_{\mathrm{opt}} X_{\mathrm{opt}} = 0_{n\times n}$. Wird $\mathrm{sp}(A_i X) = b_i; i = 1, \ldots, p$ kurz als $\mathrm{A}(x) = b$ mit dem Operator $\mathrm{A} : \mathcal{S}_n \to \mathbb{R}^m$ geschrieben, dann ist $\mathrm{A}^* : \mathbb{R}^m \to \mathcal{S}_n, y \mapsto \sum_{i=1}^m y_i A_i$ der adjungierte Operator. Zudem ist $\langle C, X\rangle = \mathrm{sp}(CX)$ das entsprechende Skalarprodukt auf den symmetrischen Matrizen und der Kegel $\mathcal{C}$ sind die nichtnegativ definiten Matrizen. Dieser Kegel ist selbstdual, also $\mathcal{C}^\oplus = \mathcal{C}$, womit die vorgenannte Dualität direkt abgelesen werden kann.

Anmerkung 8.8 Beachte: Die Dualitätsformulierung [476]

$$Q_P = \inf\{\langle c, x\rangle : x \in (\mathcal{L} + b) \cap \mathcal{C}\} \tag{8.26a}$$

$$Q_D = \inf\{\langle b, y\rangle : y \in (\mathcal{L}^\perp + c) \cap \mathcal{C}^\oplus\} \tag{8.26b}$$

weicht von der Lagrange-Dualität dahingehend ab, dass $Q_P + Q_D = \langle c, b\rangle$ gilt.

Nach dieser mathematisch geprägten Betrachtung sei hier nochmals an den Vorteil erinnert, den die Umformulierung bei starker Dualität bringt. Dadurch, dass sich die Rolle der Variablen tauscht, entspricht die Anzahl der Variablen des dualen Problems gerade der Anzahl der Gleichungsrestriktionen. Bei sehr wenigen Gleichungsrestriktionen ist also im dualen Problem unter Umständen über erheblich weniger Variablen zu optimieren.

[12] abgeschlossener, konvexer, fester (kein leeres Innere) und spitzer ($\mathcal{C} \cap (-\mathcal{C} = 0_n)$ Kegel

[13] Ein properer Kegel induziert eine Halbordnung $x \succeq_{\mathcal{C}} y \overset{\mathrm{def}}{\Leftrightarrow} x - y \in \mathcal{C}$. Mit A^* wird die adjungierte Matrix bezeichnet, d. h. $\forall x, y \in \mathcal{H}(\mathbb{R}) : \langle Ax, y\rangle = \langle x, A^* y\rangle$. Für das Standardskalarprodukt in $\mathbb{R}^n$ ist das gerade die Transponierte von A. $\mathcal{C}^\oplus = \{y \in \mathbb{R}^n : \langle y, x\rangle \geq 0, \forall x \in \mathcal{C}\}$ ist der positive Polarkegel.

[14] Generisch steht hier gleichermaßen für topologisch als auch metrisch generisch. Einen alternativen Beweis mit etwas stärkeren Aussagen liefert [504], indem sich auf das Ewald-Larman-Rogers-Theorem aus s. Abschn. 8.1.2 bezogen wird und indem das Hausdorff-Maß die Lebesgue-Maß-Null-Mengen besser quantifiziert.

8.2 Relaxations- und kontraktionsbasierte Algorithmen

In den nachfolgenden Abschnitten werden ausgewählte relaxations- und kontraktionsbasierte Iterationsalgorithmen vorgestellt. Dazu zählen (Übersichten in [387], [56]):

1. Zyklische Projektionsalgorithmen (Weglassen von Restriktionen)
2. Blockabstiegsverfahren (Festhalten aller bis auf eine Variable)
3. Alternierender Quadratmittelalgorithmus (Blockabstiegsverfahren für LS)
4. Quadratmittelalgorithmus mit iterativer Wichtung (Majorisierung der Zielfunktion)
5. Iterativer quadratischer Maximum-Likelihood-Algorithmus (Festhalten einer Variablen nach Anwendung der Spitting-Methode).

8.2.1 Zyklische Projektionsalgorithmen

Viele Identifikations- und Approximationsprobleme mit Restriktionen können in

$$||a - x||_{\mathcal{H}} \stackrel{!}{=} \text{Min} \qquad x \in \bigcap_{i=1}^{m} \mathcal{C}_i = \mathcal{C} \subset \mathcal{H}(\mathbb{R}) \tag{8.27}$$

mit skalarproduktinduzierter Norm umgeschrieben werden. So hat $||Ax - b||_2^2 \stackrel{!}{=} \text{Min}$ den gleichen Minimierer wie $||A^+b - x||_{A^TA} \stackrel{!}{=} \text{Min}$. Ähnliches gilt für matrixvariate Probleme. Für die wichtige Klasse (8.27) existieren seit langem Aussagen zu Relaxationsverfahren, die in jedem Schritt nur eine Restriktion betrachten und die Optimallösung x_{opt} durch eine Folge verketteter Projektionen mit zyklischer Wiederholung

$$x_{k+1} = P_{\mathcal{C}_m} P_{\mathcal{C}_{m-1}} \cdots P_{\mathcal{C}_1}(x_k) \tag{8.28}$$

ermitteln. Das Von-Neumann-Theorem für alternierende Projektionen (für $m = 2$) und dessen Verallgemeinerung durch Halperin (für $m \geq 2$) liefert eine solche Aussage.

Satz 8.5 (Halperin's Konvergenzsatz für zyklische Projektionen, [271])
Sind $\mathcal{C}_1, \ldots, \mathcal{C}_m$ abgeschlossene Unterräume in einem Hilbert-Raum $\mathcal{H}$ und $\mathcal{C} = \bigcap_{i=1}^{m} \mathcal{C}_i$, dann gilt für alle $x \in \mathcal{H}$

$$\lim_{k \to \infty} ||(P_{\mathcal{C}_m} P_{\mathcal{C}_{m-1}} \cdots P_{\mathcal{C}_1})^k(x) - \mathcal{P}_{\mathcal{C}}(x)|| = 0. \tag{8.29}$$

Die Konvergenz bleibt auch für den Schnitt affiner Unterräume (und damit auch Hyperebenen) mit $\mathcal{C} \neq \emptyset$ erhalten[15].

[15] Während der Schnitt von Unterräumen per se nie leer sein kann, da sie stets den Nullpunkt gemeinsam haben, ist im Fall der affinen Räume zusätzlich eine nichtleere Schnittmenge zu fordern.

Anmerkung 8.9 Zyklische Projektionsalgorithmen und deren Variationen (mit Beschleunigung [57], mit Modifikation der Projektionsreihenfolge, für konvexe Mengen usw.) konvergieren langsam. Obwohl Beispiele mit beliebig langsamer Konvergenz konstruiert werden können, rechtfertigt die Praxis in Verbindung mit schnellen Rechnern ihren Einsatz. Denn da für zahlreiche Einzelprojektionen kompakte Lösungsformeln existieren, ist so ein Algorithmus schnell programmiert.

Anmerkung 8.10 Die Konvergenzaussage gilt nicht für den Schnitt konvexer Mengen! Es folgt lediglich Konvergenz zu einem Punkt im Schnitt, sofern der Schnitt nicht leer ist. Alternierende Projektionen $y_k = P_{\mathcal{C}_2}(x_k)$ und $x_{k+1} = P_{\mathcal{C}_1}(y_k)$ liefern im Fall zweier Mengen mit leerem Schnitt $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$ Folgen $x_k \to x^* \in \mathcal{C}_1$ und $y_k \to y^* \in \mathcal{C}_2$ mit der Eigenschaft $||x^* - y^*||_{\mathcal{H}} = \text{dist}(\mathcal{C}_1, \mathcal{C}_2)$ [129]. Sie liefern somit einen einfachen Weg, um den Abstand zweier konvexer Mengen zu bestimmen.

Beispiel 8.7 (Zyklische Projektion versagt bei konvexen Mengen)
Für zwei Halbräume $\mathcal{C}_1 = \{(x_1, x_2) : x_2 \leq 0\}$ und $\mathcal{C}_2 = \{(x_1, x_2) : x_1 + x_2 \leq 0\}$ liefert die Projektion von $(2, 1)$ auf $\mathcal{C}_1$ den Punkt $(\sqrt{2}, 0)$ und die anschließende Projektion den Punkt (1,-1) im Schnitt beider Mengen. Die Optimallösung ist hingegen $(\frac{1}{2}, -\frac{1}{2})$.

Anmerkung 8.11 Die Konvergenzaussage gilt auch nicht für den in der Signalverarbeitung lange Zeit populären Cadzow-Algorithmus [115] zur strukturierten Niedrig-Rang-Reduktion, der alternierend auf affin-strukturierte Matrizen und rangreduzierende Matrizen projiziert. Aus einer Folge sich beständig verringernder Gütewerte darf nämlich nicht Konvergenz zur gesuchten Lösung geschlossen werden.

Beispiel 8.8 (Cadzow-Algorithmus versagt)
Gesucht sei die nächstgelegene Hankel-Matrix kleineren Rangs zu $\left[\begin{smallmatrix} 1&0&0 \\ 0&0&1 \\ 0&1&0 \end{smallmatrix}\right]$. Der Algorithmus liefert $\left[\begin{smallmatrix} 1&0&0 \\ 0&0&0 \\ 0&0&0 \end{smallmatrix}\right]$ statt $\left[\begin{smallmatrix} 0&0&0 \\ 0&0&1 \\ 0&1&0 \end{smallmatrix}\right]$. Das Beispiel bereitet auch den modernen STLN-Zugängen Schwierigkeiten [500].

Nach so viel Negativem soll nun der Dykstra-Algorithmus vorgestellt werden, der dank einer speziellen Korrektur Konvergenz zur Optimallösung im Schnitt konvexer Mengen liefert. Er erweitert so das Einsatzgebiet von Projektionsalgorithmen daher erheblich.

Satz 8.6 (Dykstra-Algorithmus funktioniert bei konvexen Mengen, [95][16])
Sind $\mathcal{C}_1, \ldots, \mathcal{C}_m$ abgeschlossene konvexe Mengen in einem Hilbert-Raum $\mathcal{H}$ und $\mathcal{C} = \bigcap_{i=1}^m \mathcal{C}_i \neq \emptyset$, dann liefert die Iteration in $k = 1, 2, \ldots$

$$q_1^{(0)} = \ldots = q_m^{(0)} = 0_{\mathcal{H}}, x_m^{(0)} = a \tag{8.30a}$$

$$\text{für } i = 1, \ldots, m \qquad w := x_{i-1}^{(k)} + q_i^{(k-1)} \qquad x_0^{(k)} = x_m^{(k-1)} \tag{8.30b}$$

$$x_i^{(k)} := P_{\mathcal{C}_i}(w) \tag{8.30c}$$

$$q_i^{(k)} = w - x_i^{(k)} \qquad \text{Dykstra-Korrekturterm} \tag{8.30d}$$

für jeden Startwert $a \in \mathcal{H}$ insgesamt m Folgen $x_i^{(k)}$, die allesamt gegen die Bestapproximation $x_{\text{opt}} = P_{\mathcal{C}}(a)$ von (8.27) streben.

Anmerkung 8.12 Die Dykstra-Korrektur kann für alle jene $\mathcal{C}_i$ entfallen, die affine Unterräume sind, denn da $q_i^{(k)} \perp \mathcal{C}_i$ für alle $q_i^{(k)}$ gilt, folgt ohnehin $P_{\mathcal{C}_i}(x_{i-1}^{(k)} + q_i^{(k-1)}) = P_{\mathcal{C}_i}(x_{i-1}^{(k)})$.

Anmerkung 8.13 Weiterführende Arbeiten sind [56] (Übersicht, Konvergenzergebnisse) und [97] (Erweiterung der Zielfunktionen; Bregman-Projektionen).

8.2.2 Blockabstiegsverfahren

Die Idee hinter diesem Verfahren ist wiederum, das Gesamtproblem über eine Folge einfacher zu lösender Teilprobleme anzugehen. Es eignet sich für stetige Zielfunktionen der Art

$$f(x_1, \ldots, x_n) \stackrel{!}{=} \text{Min} \qquad (x_1, \ldots, x_n) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n. \tag{8.31}$$

Definition 8.7 (Blockabstiegsverfahren, [387])
Ein Algorithmus für (8.31) mit den Startwerten $x_2^{(0)}, \ldots, x_n^{(0)}$ heißt Blockabstiegsverfahren[17], wenn er folgende Gestalt hat

$$x_1^{(k+1)} = \operatorname*{argmin}_{x_1 \in \mathcal{X}_1} f(x_1, x_2^{(k)}, \ldots, x_n^{(k)}) \qquad k = 0, 1, \ldots \tag{8.32a}$$

$$x_2^{(k+1)} = \operatorname*{argmin}_{x_2 \in \mathcal{X}_2} f(x_1^{(k+1)}, x_2, x_3^{(k)} \ldots, x_n^{(k)}) \tag{8.32b}$$

$$\ldots$$

$$x_n^{(k+1)} = \operatorname*{argmin}_{x_n \in \mathcal{X}_n} f(x_1^{(k+1)}, \ldots, x_{n-1}^{(k)}, x_n). \tag{8.32c}$$

[16] Die Arbeit verallgemeinert Dykstra's Idee [176], die sich auf den Schnitt konvexer Kegel bezieht. Der Algorithmus wurde unabhängig auch von Han [272] entwickelt und von ihm per Dualitätstheorie bewiesen.

[17] Blockabstiegsverfahren heißen mitunter auch Block-Relaxationsverfahren. Im Sinne der obigen Definition ist aber eher Block-Kontraktion angebracht. Die historische Bezeichnung als Relaxation rührt daher, dass mitunter Relaxation schlechthin als Vereinfachung und nicht als Entspannung gedeutet wird. Demgegenüber hat die hier getroffene Unterscheidung den Vorteil, dass sofort klar ist, ob das Verfahren eine Abschätzung nach unten oder oben liefert.

Beispiel 8.9 (Blockabstiegsverfahren)
$x_1^2 - x_1x_2 + x_2^2 \stackrel{!}{=} \text{Min}$ hat die Lösung $x_{\text{opt}} = (0,0)^T$. Die Zielfunktion ist für festes x_1 bzw. festes x_2 konvex, womit alle Teillösungen eindeutig sind. Mit $x_2^{(0)} = a$ ergeben sich die Iterierten aus den skalaren Teilproblemen zu $x_1^{(k)} = (\frac{1}{4})^k(2a)$ und $x_2^{(k)} = (\frac{1}{4})^k a$. Sie sind also Nullfolgen.

Die x_i können selbstverständlich auch Vektoren oder Matrizen sein. Von den Mengen $\mathcal{X}_i$ wird lediglich gefordert, dass für die Teilprobleme eine Lösung existiert, nicht notwendigerweise eine eindeutige. Die Mengen können durchaus endlich sein. Wichtig ist aber, dass die Variablen nicht wechselseitig über Restriktionen voneinander abhängen, da sonst die Iteration scheitern kann, wie das folgende Beispiel zeigt.

Beispiel 8.10 (Blockabstiegsverfahren scheitert bei nichtseparablen Mengen)
$x_1^2 + x_2^2 \stackrel{!}{=} \text{Min}; x_1 - x_2 = 1$ hat die Lösung $(\frac{1}{2}, -\frac{1}{2})$. Gestartet mit $x_2^{(0)} = 0$, bliebe $\mathcal{X}_1^{(1)} = \{1\}$ und damit wegen $x_1^{(1)} = 1$ ein Stopp bei einer falscher Lösung $(1,0)$.

Das nachfolgende Beispiel soll als Warnung dienen, dass aus einer Konvergenz der Iteration nicht auf Konvergenz zur richtigen Lösung geschlossen werden darf.

Beispiel 8.11 (Blockabstiegsverfahren liefert falsche Lösung, [1])
Es sei die Chebyshev-Approximation von $f(x) = x^2$ durch eine Gerade $g(x) = \theta_1 x + \theta_0$ für das Intervall [0,1] gesucht. Die Lösung für

$$\max_{x \in [0,1]} |x^2 - \theta_1 x - \theta_0| \stackrel{!}{=} \text{Min}$$

lautet $g_{\text{opt}}(x) = x + \frac{1}{8}$. Wird die Iteration mit $\theta_1^{(0)} = 0$ gestartet, folgt $\theta_0^{(1)} = \frac{1}{2}$ und mit festgehaltenem $\theta_0^{(1)} = \frac{1}{2}$ anschließend $\theta_1^{(1)} = 0$. Damit ist die Iteration zu Ende, Konvergenz lag vor, aber das Ergebnis $g(x) = \frac{1}{2}$ ist falsch.

Das Erschreckende an diesem Beispiel ist, dass die Zielfunktion in θ konvex ist. Allein deshalb müssten sich Blockabstiegsverfahren verbieten, wären da nicht ihre Einfachheit und die vielen erfolgreichen Anwendungen etwa in der Identifikation und Statistik. Der Zwiespalt rührt aus der unterschiedlichen Stärke in den Aussagen zur globalen und lokalen Konvergenz. Die globalen Aussagen greifen im Wesentlichen auf die Zangwill-Konvergenzbedingungen zurück und sind relativ schwach, s. [387] einschließlich weiterer Gegenbeispiele. Stärker sind die lokalen Aussagen, die auf einem Satz von Ostrowski [431] beruhen. Sie sind es letztlich, die die Rechtfertigung geben. Hinzu kommt, dass die Situation für eine Reihe spezieller Klassen nicht so dramatisch ist. Zwei für solche Klassen zugeschnittenen Verfahren, der ALS-Algorithmus und der IRLS-Algorithmus, werden wegen ihres häufigen Gebrauchs in den nachfolgenden beiden Abschnitten kurz diskutiert. Weitere Spezialverfahren finden sich in [387].

8.2.3 Alternierender Quadratmittelalgorithmus

Ein spezielles Blockabstiegsverfahren ist der Alternierende Quadratmittelalgorithmus (ALS-Algorithmus; Alternating Least Squares), der in jedem Schritt ein lineares LS-Problem löst. Die Klasse der separablen NLS-Probleme $||f(X,Y)||_F^2 \stackrel{!}{=} \text{Min}$, bei denen f bezüglich jeder Variablen affin ist, gestattet beispielsweise die Anwendung des ALS-Algorithmus

$$X^{(k+1)} = \arg\min_{X\in\mathcal{X}} ||f(X,Y^{(k)})||_F^2 \tag{8.33a}$$

$$Y^{(k+1)} = \arg\min_{Y\in\mathcal{Y}} ||f(X^{(k+1)},Y)||_F^2. \tag{8.33b}$$

Bei mehr als zwei Variablen heißt der Algorithmus zyklischer LS-Algorithmus. Eine Anwendung findet der ALS-Algorithmus bei der Identifikation von Wiener- und Hammerstein-Modellen, wie das folgende Beispiel zeigt.

Beispiel 8.12 (Identifikation von FIR-Hammerstein-Modellen)

$$y[k] = \sum_{j=1}^{p} b_{j-1}\Big(\sum_{i=1}^{n} c_i (u[k-j])^i\Big) + \varepsilon[k]; k = n, n+1, \ldots; b_0 = 1 \tag{8.34}$$

In der zugehörigen LS-Formulierung ist f affin in den unbekannten Variablen $x_i := b_i$ und $y_j := c_j$.

Sind für die möglicherweise restringierten Teilprobleme explizite Lösungen bekannt, ist der in den Gütewerten konvergente Algorithmus (monoton fallende, durch Null beschränkte Folge) schnell programmiert. Lokale Konvergenz ist gewährleistet, d. h., wird ein Startwert in der Nähe des globalen Minimierers gewählt, dann ist es unmöglich, nicht gegen diesen zu konvergieren. Begründung: x_{opt} ist ein Fixpunkt, die Iterierten bewegen sich immer in seine Richtung und die Folge der Gütewerte ist fallend. Die Konvergenz kann langsam sein. Globale Konvergenz liegt zwar meistens, aber eben nicht immer vor, wie das folgende einfache Beispiel, s. auch [586] für ein Gegenbeispiel mit einem Hammerstein-Modell.

Beispiel 8.13 (Global nichtkonvergente, aber lokal konvergente ALS)

$$\left\| \begin{matrix} x-2 \\ xy+4 \end{matrix} \right\|_2^2 \stackrel{!}{=} \text{Min} \Rightarrow x_{\text{opt}} = 2, y_{\text{opt}} = -2, Q_{\min} = 0$$

Es folgt $x_k = \dfrac{-4y_{k-1}+2}{y_{k-1}^2+1}$, $y_k = -\dfrac{4}{x_k} = \dfrac{2(y_{k-1}^2+1)}{2y_{k-1}-1}$. Für $y_0 = 1$ gilt $x_k \to 0$, $y_k \to \infty$, $x_k y_k \to -4$, und $Q_k \to 4$ (monoton fallend), d. h., die ALS konvergiert nicht gegen das Minimum, die Folge der Minimierer divergiert sogar. Für $y_0 \approx -2$ ergibt sich lokale Konvergenz.

Ursache für das Fehlschlagen der Konvergenz ist häufig eine Überparametrisierung, die entsteht, wenn X und Y keine natürlichen Unbekannten sind, sondern selbst über Reparametrisierung entstanden sind (z. B. Rangfaktorisierung). Zudem neigen die orthogonalen Suchrichtungen in $\mathcal{X} \times \mathcal{Y}$ zur Konvergenz gegen Sattelpunkte und ist die Konvergenzgeschwindigkeit im Vergleich zu anderen Verfahren geringer. Diesen Nachteilen steht die Einfachheit gegenüber, die bei konvexen Problemen, Problemen mit wenigen Minima oder bei gemischten Problemen (kontinuierliche und diskrete Variablen) zum Tragen kommt.

Eine Alternative zur ALS bietet insbesondere bei affinen und einigen einfachen konvexen Restriktionen das Gauß-Newton-Verfahren, das sich für

$$||A - BXCYD||_F^2 \stackrel{!}{=} \text{Min} \qquad X \in \mathcal{X}, Y \in \mathcal{Y} \tag{8.35}$$

wie folgt liest (ersetze X durch $X + \Delta X$, desgleichen für Y, vernachlässige $B\Delta X\, C\Delta Y D$)

$$(X_{k+1}, Y_{k+1}) = (X_k, Y_k) + \alpha_k \underset{\substack{X_k+\Delta X \in \mathcal{X} \\ Y_k+\Delta Y \in \mathcal{Y}}}{\operatorname{argmin}} ||A - BX_kCY_kD - BX_kC\Delta Y\, D - B\Delta X\, CY_kD||_F^2. \tag{8.36}$$

Bei bilinearen Problemen konvergiert gemeinhin die ungedämpften Version $\alpha_k = 1$ global.

8.2.4 Quadratmittelalgorithmus mit iterativer Wichtung

Der Quadratmittelalgorithmus mit iterativer Wichtung, kurz IRLS-Algorithmus (iteratively reweighted least squares), eignet sich für nichtquadratische Probleme der Art

$$\sum_{i=1}^{N} f(b_i - a_i^T x) \stackrel{!}{=} \text{Min}. \tag{8.37}$$

Diese lassen sich, durch eine Folge parameterlinearer quadratischer Probleme

$$x_{k+1} = \underset{x_k}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} w_i(x_k) \cdot (y_i - a_i^T x_k)^2 \quad \text{mit } w_i(x_k) = \frac{1}{z} \frac{\mathrm{d}f(z)}{\mathrm{d}z}\bigg|_{z=y_i-a_i^T x_k} < \infty \tag{8.38}$$

iterativ lösen, wenn f eine symmetrische konvexe Funktion ist und $\psi(z)/z = (df/dz)/z$ für $z > 0$ beschränkt ist und monoton fällt.

Die Funktion $f(x) = |x|^p$ hat für $1 < p < 2$ die geforderten Eigenschaft, weshalb die l_p-Approximation über $||Ax - b||_p^p \stackrel{!}{=} \text{Min}$ per IRLS berechnet werden kann. Zahlreiche Fehlerfunktionen für robuste M-Schätzer erfüllen ebenfalls die Voraussetzungen [439], weshalb der Algorithmus zur Lösung robuster Schätzprobleme gern eingesetzt wird.

Obwohl der IRLS-Algorithmus oft nur für freie Probleme angegeben wird, kann er ohne Schwierigkeiten auf restringierte Probleme erweitert werden, wenn die quadratischen Teilprobleme unter der Restriktion einfach zu lösen sind. Bei linearen Restriktionen ist das sicher der Fall, s. [125] für eine Anwendung (pro Iteration wird ein LSE-Problem gelöst).

Die Konvergenz der IRLS ist oft linear oder superlinear[18]. Sie geht ohne Modifikationen verloren, wenn die Forderung $w_i(x_k) < \infty$ verletzt wird. Das kann bei einigen Funktionen f wegen $\psi(z)/z$ bei Null- oder Quasinull-Residuen geschehen. Wohin die IRLS konvergiert, hängt von der Funktion f in (8.37) und bei Restriktionen von der zulässigen Menge ab. Durch das sukzessive Minimieren werden gemeinhin lokale Minimierer erreicht, bei konvexen Problemen ein globaler. Deshalb beziehen sich die wenigen verfügbaren Beweise auf den konvexen Fall, s. z. B. [77]. Sie beruhen vom Grundansatz auf einer konvexen Majorisierung (sichert absteigende Folge, schließt Multimodalität aus). Die Majorisierungstechnik wird ganz allgemein in [487] beschrieben und wurde vom Autor auch für einen neuen Konvergenzbeweis des Fuzzy-c-Means-Algorithmus [251] verwendet. Sie beruht darauf, dass der Minimierer einer majorisierenden Funktion (diese liegt komplett über der eigentlich zu optimierenden Funktion) zu einem Abstieg (kleineren Funktionswert) in der eigentlichen Funktion führt. Geschickte Umformungen, Abschätzungen über Ungleichungen und Monotoniebetrachtungen sind der Schlüssel zum Finden einer Majorante. Um diese Aussage zu stützen, wird hier im nachfolgenden Beispiel eine Majorante konstruiert.

Beispiel 8.14 (Majorisierung in der IRLS-Iteration)
Ist f eine symmetrische, konvexe, differenzierbare Funktion und $\varepsilon_i^{(k)} = b_i - a_i^T x^{(k)}$ das i-te Residuum im k-ten Schritt, dann wird die Zielfunktion $\sum_{i=1}^N f(b_i - a_i^T x)$ durch jede der nachstehenden Funktionen majorisiert

$$\mathrm{Maj}^{(k)}(x) = \frac{1}{2}\sum_{i=1}^{N} \frac{\psi(\varepsilon_i^{(k)})}{\varepsilon_i^{(k)}} \cdot (y_i - a_i^T x)^2 + \underbrace{\sum_{i=1}^{N} \left(f(\varepsilon_i^{(k)}) - \frac{1}{2}\varepsilon_i^{(k)}\psi(\varepsilon_i^{(k)}) \right)}_{=\text{konstant}}. \tag{8.39}$$

Um das zu verdeutlichen, wird gezeigt, dass jede der Teilfunktionen $f(y_i - a_i^T x) =: f(z)$ unterhalb der Funktion

$$\mathrm{maj}_i(z) = \frac{\psi(\varepsilon_i^{(k)})}{2\varepsilon_i^{(k)}} z^2 + \underbrace{f(\varepsilon_i^{(k)}) - \frac{1}{2}\varepsilon_i^{(k)}\psi(\varepsilon_i^{(k)})}_{=\text{konstant}}$$

liegt, d. h., dass $u_i(z) = \mathrm{maj}_i(z) - f(z) \geq 0$ gilt. Über $u_i'(z) = \frac{\mathrm{d}u_i(z)}{\mathrm{d}z} = \frac{\psi(\varepsilon_i^{(k)})}{\varepsilon_i^{(k)}} z - \psi(z)$ und mit den Eigenschaften von f folgt $u_i(\varepsilon_i^{(k)}) = u_i(-\varepsilon_i^{(k)}) = 0$ und $u_i'(\varepsilon_i^{(k)}) = u_i'(-\varepsilon_i^{(k)}) = 0$. Da $\psi(z)/z$ für $z > 0$ monoton fällt, fällt $u_i'(z)$ für $0 < z \leq \varepsilon_i^{(k)}$ und steigt für $z \geq \varepsilon_i^{(k)}$, was letztlich $u(z) \geq u(\varepsilon_i^{(k)}) = 0$ für $z \geq 0$ impliziert; desgleichen für $z \leq 0$ wegen der Symmetrie. Eine Summation über i liefert das gewünschte Resultat.

[18] Sei x_∞ der Konvergenzpunkt, dann hat ein Algorithmus eine lineare Konvergenzgeschwindigkeit, wenn ein c mit $0 < c < 1$ existiert und $||x_{k+1} - x_\infty|| \leq c||x_k - x_\infty||$ gilt. Er hat eine superlineare Konvergenz, wenn statt eines konstanten c eine Folge c_k mit $\lim_{k\to\infty} c_k =$ existiert. Der Abstand vom Konvergenzpunkt verringert sich also mit wachsendem k immer schneller.

Nachfolgend werden NLS-Probleme der Art

$$\sum_{i=1}^{N} \left(\frac{a_i^T x + b_i}{c_i^T x + d_i} - f_i \right)^2 \stackrel{!}{=} \text{Min} \qquad x \in \mathbb{R}^n \tag{8.40}$$

betrachtet, die äquivalent sind zu

$$Q(x) = \sum_{i=1}^{N} w_i(x) \cdot ((a_i^T - f_i c_i^T)x + (b_i - f_i d_i))^2 \stackrel{!}{=} \text{Min} \qquad \text{mit } w_i(x) = \frac{1}{(c_i^T x + d_i)^2}. \tag{8.41}$$

Diese Problem legen eine IRLS-Implementierung nahe, erfüllen aber die Voraussetzungen zu (8.38) nicht. Klar ist sofort, dass die Zielfunktionen $\tilde{Q}(x_k) = \sum_{i=1}^{N} w_i(x_k) \cdot ((a_i^T - f_i c_i^T)x + (b_i - f_i d_i))^2 + \mathit{Konstante}$ keine globalen Majoranten sein können, da der „quadratische“ Differenzterm $\sum_{i=1}^{N}(w_i(x_k) - w_i(x))[(a_i - f_i c_i)^T x]^2$ von $\tilde{Q}(x_k) - Q(x)$ negativ werden kann. Nichtsdestotrotz kann die IRLS-Folge der x_k konvergent sein, doch wird der Fixpunkt für x_k in der Regel nicht einmal ein lokaler Minimierer sein (Konvergenz zum globalen Minimierer ist wegen der Nichtkonvexität von $Q(x)$ ohnehin nicht zu erwarten), vgl. Fixpunktbedingung von $Q(x_k)$ und notwendigen Bedingung erster Ordnung für $Q(x)$.

Der ALS- und der IRLS-Algorithmus bilden eine Alternative zu Standardalgorithmen der nichtlinearen Optimierung, da sie einfach zu programmieren sind. Ein bisher weitgehend unbeachteter Vorteil resultiert aus ihrer Eigenschaft, eine Folge einfacher parameterlinearer LS-Probleme zu generieren und zu lösen. Dadurch eignen sie sich zur Entwicklung rekursiver LS-Algorithmen unter Restriktionen, bei denen die bekannten rekursiven LS-Algorithmen mit den iterativen Komponenten der Algorithmen kombiniert werden können.

8.2.5 Iterativer quadratischer Maximum-Likelihood-Algorithmus

STLS-Probleme [389], Frequenzschätzprobleme [471] und andere Probleme lassen sich in

$$x^T W(x) x \stackrel{!}{=} \text{Min} \qquad x^T x = 1 \tag{8.42}$$

mit einem für alle $x \in \mathbb{R}^n$ positiv definitem $W(x)$ umschreiben. Oft werden sie dann mit dem sog. IQML-Algorithmus (iterative quadratic maximum likelihood) gelöst:

$$x_{k+1} = \text{argmin}\{x^T W(x_k) x : x^T x = 1\} \tag{8.43a}$$

$$= \text{standardisierter Eigenvektor zu } \lambda_{\min}(W(x_k)) \tag{8.43b}$$

Die Idee dahinter ist zunächst eine Erweiterung von (8.42) auf

$$x^T W(y) x \stackrel{!}{=} \text{Min} \qquad x^T x = 1, y = x \tag{8.44}$$

und anschließend eine Kombination aus Weglassen der Restriktion $y = x$ und Fixieren $y = x_k$. Allerdings zeigt folgende Überlegung, dass der Algorithmus (8.43) gemeinhin nicht

gegen die Lösung von (8.44) konvergiert. Ist nämlich x_{fix} Fixpunkt der Iteration (8.43), dann ist x_{fix} ein Eigenvektor von $W(x_{\text{fix}})$ zum kleinsten Eigenwert (Rayleigh-Ritz-Ungleichung). Um zugleich Optimallösung zu sein, müsste x_{fix} die aus der Lagrange-Funktion $L(x, \lambda) = x^T W(x)x - \lambda(x^T x - 1)$ zu (8.42) folgende notwendige Bedingung (Ableiten bezüglich x)

$$2W(x)x + \begin{bmatrix} x^T \frac{\partial W(x)}{\partial x_1} x \\ \vdots \\ x^T \frac{\partial W(x)}{\partial x_n} x \end{bmatrix} = \lambda x \tag{8.45}$$

erfüllen. Die Optimallösung x_{opt} ist demnach anders als x_{fix} aber kein Eigenvektor von $W(x_{\text{opt}})$, es sei denn, der zweite Summand auf der linken Seite liegt parallel zum Eigenvektor. Da das im Allg. nicht der Fall sein wird, stimmen der Iterationsfixpunkt und die Optimallösung gemeinhin nicht überein.

Vom angegebenen IQML-Algorithmus gibt es diverse Modifikationen mit zahlreichen Anwendungen, s. die Literaturhinweise in der Arbeit [587], der auch das folgende Beispiel entnommen wurde.

Beispiel 8.15 (Zeit- und Ortsfrequenzschätzprobleme)
Solche Schätzprobleme führen auf Aufgaben vom Typ

$$\text{sp}[B(x)(B^H(x)B(x))^{-1}B^H(x)R] \stackrel{!}{=} \text{Min}, \tag{8.46}$$

wobei $B(x)$ Band-Toeplitz-Matrizen mit den Elementen von x sind, wobei x seinerseits für Polynomkoeffizienten mit einer Elementfixierung oder Eins-Normierung steht. R ist eine Schätzung der Kovarianzmatrix der Messdaten. Auch hier konvergiert der IQML-Algorithmus mit einer Wichtung $W(x_k) = (B^H(x_k)B(x_k))^{-1}$ i. Allg. nicht gegen den wahren Wert, was wegen der Verwandtschaft zu (8.42) analog zu (8.45) gezeigt werden kann, vgl. [587] für eine alternative Argumentation.

Obwohl der IQML-Algorithmus und seine Modifikationen gemeinhin nicht gegen die wahre Lösung konvergieren, sie also nur eine approximierte Lösung liefern, und obwohl sogar divergente Beispiele existieren[19], haben Algorithmen vom IQML-Typ eine empirische Rechtfertigung. Sie konvergieren fast immer, die Approximationen sind gut und die Algorithmen sind einfach zu programmieren. Dank leistungsfähiger Rechner empfiehlt es sich aber, diesen Algorithmentyp nur für Startschätzungen heranzuziehen, die dann mit klassischen nichtlinearen Optimierungsalgorithmen das jeweils originäre Problem lösen.

[19] Der Hinweis zur Divergenz bezieht sich auf eine Arbeit von He und Ren (1995) in [587]. Allerdings konnte diese Arbeit nicht recherchiert werden. In [336] wird von Divergenz der Steiglitz-McBride-Iteration berichtet (ohne Beispiel), die zur IQML-Iteration nach [449] äquivalent ist.

8.3 Matrixapproximationsprobleme

Standardschätzverfahren wie die LS und die TLS oder die Hauptkomponenten- und Faktoranalyse lassen Interpretationen zu, in denen ein nächstgelegener Vektor oder eine nächstgelegene Matrix gesucht ist (Formulierung als Projektionsproblem). Im Fall von Matrizen, die als Spezialfall die Vektoren einschließen, wird dann von Matrixapproximationsproblemen gesprochen. Anwendungen bei dynamischen Systemen betreffen beispielsweise die nächstgelegene Hurwitz-stabile Matrix oder nächstgelegene Matrix mit vorgegebenem Eigenwert und kommen meist a posteriori zum Einsatz, um eine Eigenschaft zu erzwingen, die durch das Schätzverfahren zunächst nicht erfüllt wurde. Im Rahmen einer umfangreichen Literaturrecherche wurden vom Autor mehr als 200 gelöste Matrixapproximationsprobleme zusammengestellt, von denen eine Auswahl als Literaturverweise in Kapitel 2 angegeben ist.
Das klassische Matrixapproximationsproblem wird dabei wie folgt definiert.

Definition 8.8 (Matrixapproximationsproblem)
Ein Problem der Art

$$||A - X|| \stackrel{!}{=} \text{Min} \qquad X \in \mathcal{M} \subset \mathbb{C}^{m \times n}, \tag{8.47}$$

bei dem alle Matrizen aus $\mathcal{M}$ eine bestimmte Eigenschaft besitzen, z. B. „symmetrisch“, wird als Matrixapproximationsproblem bezeichnet. Die Matrix A stammt nicht aus $\mathcal{M}$ (sonst $X_{\text{opt}} = A$ trivial). Alle Matrizen X_{opt}, für die das Minimum angenommen wird, heißen Approximanden oder Bestapproximationen, jeweils versehen mit der die Menge $\mathcal{M}$ kennzeichnenden Eigenschaft, z. B. „symmetrischer Approximand“.

Sind mehrere Eigenschaften, beschrieben durch r Teilmengen $\mathcal{E}_i$, gleichzeitig zu erfüllen, dann ist $\mathcal{M} = \bigcap_{i=1}^{r} \mathcal{M}_i$ zu wählen. Sollten die Mengen konvex sein, kann das Problem mit dem Dykstra-Algorithmus nach Satz 8.6 gelöst werden.

Die klassischen Matrixapproximationsprobleme lassen sich in unterschiedlicher Weise erweitern. Zum einen kann diejenige Matrix gesucht werden, die zu k Matrizen A_i den kleinsten Abstand hat. Anschaulich ist das diejenige Matrix in der Nähe des Mittelpunkts (s. auch Steiner-Weber-Problem [71]), die die geforderten Eigenschaften hat. In solchen Fällen wird von simultanen Matrixapproximationsproblemen gesprochen (Bsp. simultane Hauptkomponentenanalyse, parallele Faktoranalyse, individuelle Differenzenskalierung [349], [433]), da die optimale Matrix eben zugleich mehrere Matrizen im Kompromiss approximiert. Mögliche simultane Formulierungen sind

$$\sum_{i=1}^{k} ||A_i - X|| \stackrel{!}{=} \text{Min} \qquad \text{bzw.} \qquad \sum_{i=1}^{k} ||A_i - X||^2 \stackrel{!}{=} \text{Min} \qquad X \in \mathcal{M} \subset \mathbb{C}^{m \times n}. \tag{8.48}$$

Zum anderen kann statt einer einzelnen Matrix ein Paar oder ein Tupel von Matrizen betrachtet werden. Typische Anwendung liefern (verallgemeinerte) Zustandsraumdarstellungen

mit $\Sigma = (E, A, B, C, D)$ (E nur bei Deskriptorsystemen). Matrixapproximationen für nichtsteuerbare Paare, nicht positiv reelle Systeme, Nichtminimalphasensysteme werden in den Abschnitten in Kapitel 2 genannt und diskutiert. Als Tupel können auch die Koeffizientenmatrizen A_i von Matrixpolynomen $p(\xi) = \sum_{i=0}^{r} A_i \xi^i$ auftreten, wobei das Matrixbüschel $p(\lambda) = E\lambda - A$ von Deskriptorsystemen [111], [166] oder Polynommatrizen zweiten Grads aus der Mechanik $p(\lambda) = M\lambda^2 + B\lambda + C$ mit der Massenmatrix M, der Dämpfungsmatrix B und der Steifigkeitsmatrix C genannt seien, s. [339] für allgemeinere Probleme mit Eigenwertrestriktionen. Der Standardzugang zur Formulierung derartiger Probleme basiert auf der Blockmatrixdarstellung im folgenden Sinne

$$||(A, B) - (X_1, X_2)|| \stackrel{!}{=} \text{Min} \qquad (X_1, X_2) \in \mathcal{M}. \tag{8.49}$$

Eine Alternative für Matrizen und Vektoren, die sich ihrer Dimensionen wegen einer Blockformulierung entziehen, bietet eine spezielle Summenformulierung

$$\sqrt{||A - X_1||^2 + ||B - X_2||^2} \stackrel{!}{=} \text{Min} \qquad (X_1, X_2) \in \mathcal{M}. \tag{8.50}$$

Ein Beispiel ist der Abstand von Systemen bezüglich des E/A-Verhaltens, wenn für A der Vektor der Nennerpolynomkoeffizienten und B der der Zählerkoeffizienten verwendet wird [435]. Die Wurzel sichert im Übrigen, dass es sich bei der Zielfunktion um eine Metrik handelt; ohne sie wäre die Dreiecksungleichung verletzt.

Abhängig von der Verwendung des Modells interessieren nicht notwendig die nächstgelegenen Matrizen, sondern es ist nur der Abstand einer Matrix von einer Menge $\mathcal{M}$ mit bestimmten Eigenschaften gefragt. Kurzum, nicht der Minimierer, sondern das Minimum ist gesucht. Durch Substitution $X =: A + \Delta A$ entstehen sog. Abstandsprobleme.

Definition 8.9 (Abstandsproblem)
Probleme vom Typ

$$||\Delta A|| \stackrel{!}{=} \text{Inf} \qquad (A + \Delta A) \in \mathcal{M} \subset \mathbb{C}^{m \times n} \tag{8.51a}$$

$$\frac{||\Delta A||}{||A||} \stackrel{!}{=} \text{Inf} \qquad (A + \Delta A) \in \mathcal{M} \subset \mathbb{C}^{m \times n} \tag{8.51b}$$

heißen normweise bzw. relative Abstandsprobleme[20].

Ist die Teilmenge $\mathcal{M}$ abgeschlossen, so hat (8.47) eine Lösung. Für abgeschlossenes konvexes $\mathcal{M}$ und streng konvexe Normen (Frobenius-Norm, aber nicht Spektralnorm) ist die Lösung eindeutig. Diese Bedingungen sind hinreichend. Für nichtkonvexe Mengen, also etwa bei Rangapproximationsproblemen, können trotzdem eindeutige Lösungen existieren. Ist die

[20] Da $||A||$ fest ist, sind die Minimierer ΔA_{opt} beider Probleme gleich, sofern sie existieren.

Teilmenge $\mathcal{M}$ offen, existiert keine Lösung und das Minimum ist durch ein Infimum zu ersetzen. Solche Probleme können dennoch von Interesse sein, denn das Infimum gibt den Abstand zur zulässigen Menge an und ist somit ein Maß für das Verletztsein einer Eigenschaft oder die Reserven einer Eigenschaft. Zu beachten ist, dass es Mengen gibt, die weder offen noch abgeschlossen sind; dann kann eine Lösung existieren, muss aber nicht, vgl. Beispiel 4.2.

Das relative normweise Abstandsproblem lässt sich äquivalent zu (8.51b) auch als

$$\delta \stackrel{!}{=} \text{Inf} \qquad ||\triangle A|| \leq \delta ||A|| \text{ und } (A + \triangle A) \in \mathcal{M} \tag{8.52}$$

schreiben. In dieser Form ähnelt es dem sog. komponentenweisen Abstandsproblem. Im Gegensatz zum normweisen Abstandsproblem erfolgt beim komponentenweisen Abstandsproblem die Störung oder die Suche nach der nächstgelegenen Matrix nicht in alle Richtungen. So werden z. B. Nullelemente nicht gestört. Diese Forderung ist sinnvoll, wenn beispielsweise schwach besetzte Matrizen betrachtet werden, wie sie bei der Diskretisierung von partiellen Differenzialgleichungsmodellen entstehen. Um auch Änderungen in anderen Matrixelementen unterdrücken zu können, z. B. in Einselementen oder in a priori bekannten Elementen, und um eine Skalierung vornehmen zu können, kann zusätzlich eine entsprechende Wichtung eingesetzt werden, s. nachfolgende Definition auf der Rump [544] aufbauend die theoretischen Grundlagen und Anwendungen beschreibt.

Definition 8.10 (Komponentenweise Abstandsprobleme)
Standardversion:

$$\delta_{komp} \stackrel{!}{=} \text{Inf} \qquad |\triangle A| \leq \delta_{komp} |A| \text{ und } (A + \triangle A) \in \mathcal{M}, \text{ wobei } |A| \stackrel{\text{def}}{=} ((|a_{ij}|)). \tag{8.53}$$

gewichtete Version:

$$\delta_{komp}^{W} \stackrel{!}{=} \text{Inf} \qquad |\triangle A| \preceq \delta_{komp}^{W} |A \odot W| \text{ und } (A + \triangle A) \in \mathcal{M}; W \in \mathbb{R}^{n \times n} \tag{8.54}$$

$\odot$ bezeichnet dabei das Hadamard-Produkt (elementweises Produkt) [299].

Die Vielzahl der Möglichkeiten, Matrixapproximationsprobleme zu beschreiben, ist ein Indiz für deren praktische Bedeutung. Sie findet ihre Begründung in den unterschiedlichen Blickwinkeln und ist auch dem Schwierigkeitsgrad der mathematischen Behandlung geschuldet. So kann ein Problem in der einen Norm einfach zu lösen sein, in einer anderen Norm dagegen nicht. Prinzipiell streben die Publikationen jedoch immer geschlossene Lösungen an oder, wenn das nicht möglich ist, solche die den Suchraum signifikant verkleinern (1- bzw. 2-dimensionale Suche statt n^2-dimensionaler Suche).

Kapitel 9

Empfehlungen für die Identifikation

In diesem Kapitel werden wichtige Empfehlungen zur Identifikation unter Einbeziehung von Vorwissen zusammengestellt. Sie werden um grundlegende Empfehlungen ergänzt, die unabhängig von der Einbeziehung von Vorwissen sind. Anders als die vorangegangenen Kapitel, bei denen die Fragen „Welche Restriktionen sind einzubeziehen?“ und vor allem „Wie wird das gemacht?“ beantwortet wurden, geht es in diesem Kapitel um die Fragen „Worauf ist bei der Identifikation zu achten?“ und „Welche Konsequenzen ergeben sich aus den Ergebnissen der vorliegenden Arbeit?“ Die Darstellung ist thesenhaft gestaltet, um speziell Studenten und Neueinsteigern einige von Mathematik freie Ratschläge für die Modellbildung zu geben. Die Gliederung erfolgt nach Themenkomplexen und orientiert sich nur grob an der Reihenfolge der Kapitel dieser Arbeit.

9.1 Grundlegende Empfehlungen

Es gibt zahlreiche Bücher, die sich ausschließlich mit der Approximation oder der Regression befassen, und jene, die die Identifikation dynamischer Systeme als primären Inhalt haben. Da in dieser Arbeit die Klammer durch das Vorwissen und dessen Umsetzung mit Restriktionen gegeben ist, werden typische Anwendungen aus all den genannten Gebieten aufgeführt. Je nach Anwendungsgebiet und den Zielstellungen nach guter Anpassung, biasfreien und wenig streuenden Parametern bzw. technisch gesehen validen Parametern ergeben sich unterschiedliche Gütekriterien und Vorgehensweisen. Dessen ungeachtet gibt es einige grundlegenden Aspekte zu beachten, die unabhängig von der Zielsetzung und dem Umstand sind, ob Vorwissen einbezogen wird oder nicht. Eine Auswahl geben die nachfolgenden Thesen.

1. Problemspezifisches Gütekriterium wählen
Für die Auswahl einer geeigneten Methode zur Modellbildung ist zu beurteilen, um welche Art von Problem es sich handelt, d. h. ob eine Approximation oder eine Regression ausgeführt werden soll. Die Art der Problemstellung bestimmt nämlich entscheidend die Wahl des Gütekriteriums. Als hilfreich erweist sich dabei zunächst eine Einordnung entsprechend der Tabelle 9.1.[1]

Problemfeld	Modell	Störung
inverse Probleme	exakt	keine
Approximation	genähert	keine
Regression	exakt	signifikant
Approgression	genähert	signifikant

Tabelle 9.1: Einteilung der Modellbildungsprobleme

Kennzeichen der inversen Probleme ist, dass aus beobachteten oder gewünschten Wirkungen auf die Ursache geschlossen werden soll (Zustände, Quellen). Ortungsprobleme, Tomografieprobleme und auch das dem Regelungstechniker vertraute Beobachterproblem gehören hierzu. Gemeinhin sind inverse Probleme schwierig zu lösen. Restriktionen, Problemtransformationen und Kompromisskriterien, vor allem aber Problemmodifikationen durch Regularisierungsterme und Maßnahmen zum Erzwingen eindeutiger Lösungen helfen, die Probleme günstiger zu gestalten.

Die Tatsache, dass sowohl bei der Approximation als auch bei der Regression vielfach Quadratmittelprobleme gelöst werden, erklärt die in der Fußnote angesprochene begriffliche Unschärfe teilweise. Doch dabei wird ein BLUE-Schätzer (bester linearer biasfreier Schätzer) zur Regression über einen gänzlich anderen Ansatz abgeleitet als ein gewöhnlicher Quadratmittelausgleich zur Datenapproximation. Des Weiteren unterscheiden sich beide Problemklassen in der Interpretation des Ergebnisses (Modell nähert mit einer bestimmten Güte; Modell gilt mit einer bestimmten Signifikanz). Auch der verwendete mathematische Apparat ist ein anderer (Analysis, Algebra einerseits, Statistik andererseits). Das angesprochene Verschwimmen der Begriffe wird auch beim Kalman- und beim Kalman-Bucy-Filter (kontinuierliches

[1] Die vom Autor getroffene Unterteilung ist keineswegs Standard, da die Begriffe oftmals synonym verwendet werden. Das zeigen die Diskussionen auf Wikipedia unter dem Stichwort „Ausgleichsrechnung“, wo in Diskussionsbeiträgen zu Recht bemerkt wird: „Regression und Fit bzw. Methode der kleinsten Quadrate sind nicht synonym und unterscheiden sich in den Problemstellungen und den auszuwertenden Daten.“ Die hier bezüglich Approximation-Regression-Abgrenzung geführte Argumentation ähnelt der von Magnus und Neudecker [421]. Der nahezu unbekannte, aber sehr treffende Begriff der Approgression geht auf Bandemer [44] zurück.

Pendant) deutlich, der der Regression zugeordnet werden kann (exaktes lineares Modell, aber Störungen). Allzu oft findet sich die Formulierung „Der Zustandsvektor konvergiert gegen ...“. Dieser Aussage fehlt aber die Spezifizierung der Konvergenzart (in Wahrscheinlichkeit, im quadratischen Mittel, fast sicher), die anders als in der Analysis immer mit anzugeben ist.

Letztlich sei noch die Approgression genannt, die einen Kompromiss zwischen Approximation und Regression erreicht, indem in das Gütekriterium nicht nur die Fehlervarianz, sondern auch die Parameteranzahl einfließt. Das geschieht über sogenannte Informationskriterien [107].

Als Hilfestellung für den Leser gibt Bild 9.1 eine Übersicht zu Approximationsproblemen, während Tabelle 9.2 eine Kurzcharakterisierung von Regressionsmethoden vermittelt.

Erwähnt sei, dass Regressionsprobleme außer nach der verwendeten Methode auch modellspezifisch in parametrische und nichtparametrische Regression bzw. lineare und nichtlineare Regression oder fehlerspezifisch in nichtorthogonale und orthogonale Regression unterteilt werden. Da es bei der Regression letztlich darum geht, einen Schätzwert für die Modellparameter zu erhalten, handelt es sich um Punktschätzmethoden (im Gegensatz zu den in der Statistik ebenfalls verwendeten Bereichs- und Intervallschätzmethoden). Die Vorschrift, nach der die Störungen (aufgefasst als Zufallsvariablen) in eine Zufallsvariable für die unbekannten Modellparameter abgebildet werden, heißt Schätzfunktion. Das Bild dieser Funktion wird Schätzer genannt, während die konkrete Realisierung als Schätzwert bezeichnet wird. Die Methode, nach der die Schätzfunktion abgeleitet wird, heißt Schätzmethode oder hier Regressionsmethode.[2]

[2] Der Begriff der Schätzmethode ist weitreichender als der der Regressionsmethode, da bei der parametrischen Regression die Verteilungsfunktionen aus der Kombination von parametrischem Modell und Störungsmodell entstehen, während für Schätzmethoden die Herkunft und die Struktur der Verteilungsfunktionen völlig offen ist. Das erklärt auch, warum die Momentenmethode (Parameter einer Verteilung werden aus Momentenschätzwerten für die Verteilung ermittelt) oder die Minimum-Chi-Quadrat-Methode (für diskrete Parameter bevorzugt) nicht in der Tabelle 9.2 auftaucht.

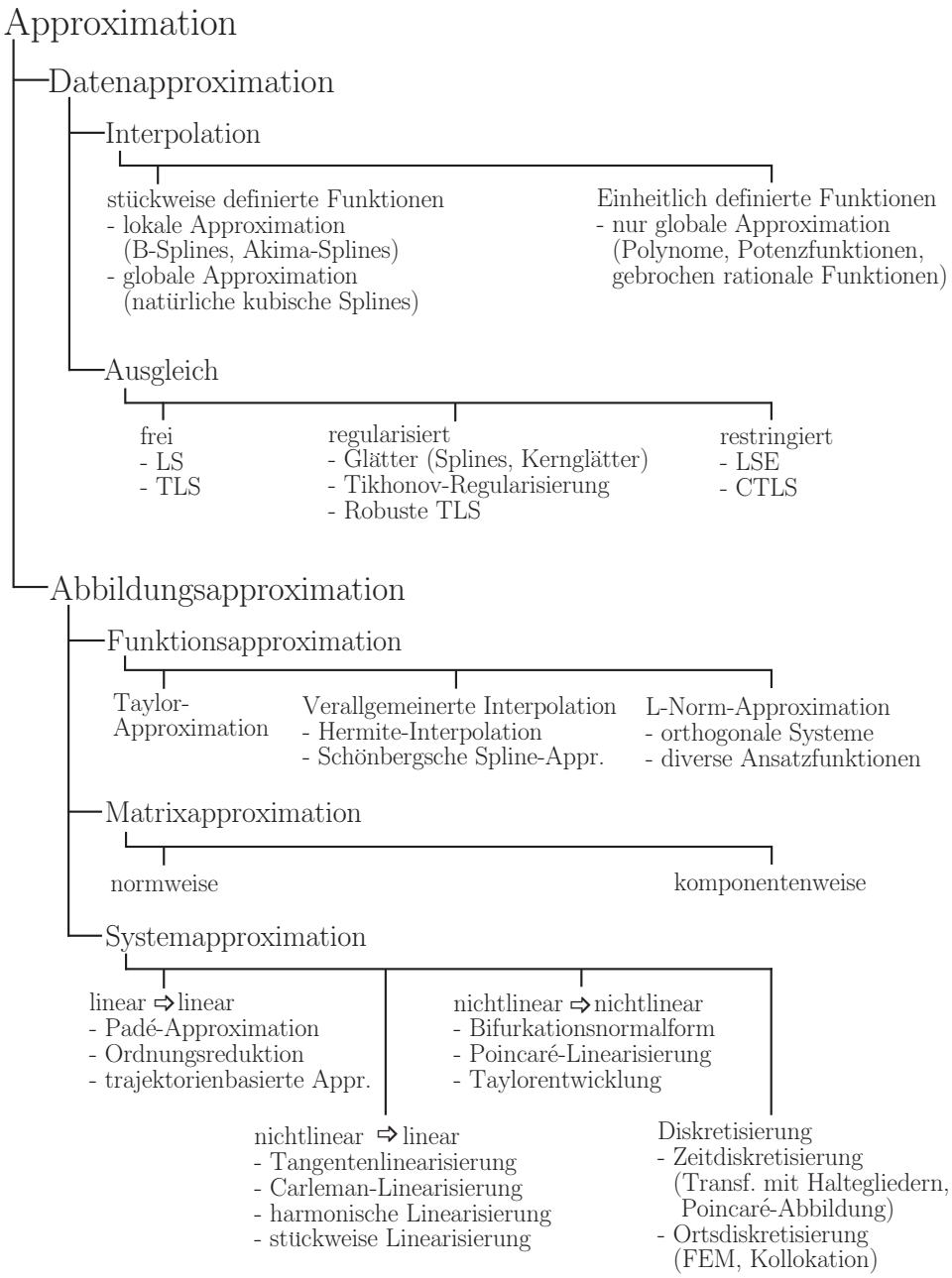

Bild 9.1: Übersicht zu Approximationszugängen

Methode	Kommentar
Bayes-Methode [44]	Vorwissen und Präferenzen über Parameter fließen über eine A-priori-Verteilung ein. Nur für einfache Modelle anwendbar, da schwierig.
Maximum-A-posteriori-Methode [44]	Modalwert der A-posteriori-Verteilung wird genutzt. Wie bei der Bayes-Methode wird eine A-priori-Verteilung benötigt. Median und Erwartungswert sind dem Modalwert oft überlegen. Bei nahezu gleichem Aufwand ist die Bayes-Methode zu bevorzugen.
Maximum-Likelihood-Methode [149]	Die Likelihood-Funktion $L(\theta\|x_1, x_2, \ldots, x_n) = f(x_1, x_2, \cdots, x_n; \theta)$ wird maximiert, die aus der Dichtefunktion folgt, wenn die Realisierungen der Zufallsvariablen in die Dichtefunktion eingesetzt werden. Die x_i und der Parametervektor θ tauschen ihre Rolle. Die ML-Methode liefert bei nichtinformativer A-priori-Verteilung (z. B. konstante Funktion) dieselben Ergebnisse wie die Bayes-Methode. ML-Schätzer haben gute statistische Eigenschaften.
Best-Methoden [421]	Bei diesen Methoden werden die Struktur des Schätzers (linear, affin, quadratisch) und/oder die Eigenschaften (biasfrei, kleinste Streuungen) vorgegeben. Der wohl bekannteste Schätzer ist der BLUE-Schätzer (best linear unbiased estimator).
Kleinste-Quadrat-Methode [421], [62]	Die Schätzfunktion entspricht der Lösungsformel des zugeordneten Quadratmittelproblems, mit dem Unterschied, dass die Abweichungen nunmehr als Zufallsvariablen interpretiert werden. Anders als bei den Best-Methoden, bei denen die Eigenschaften im Entwurf vorgegeben werden, müssen sie bei Kleinsten-Quadrat-Schätzern im Nachhinein ermittelt werden. Unter Linearitäts-, Normalverteilungs- und Unkorrelliertheitsannahmen liefern die Kleinste-Quadrat-Methode und die ML-Methode gleiche Ergebnisse.
Robuste Methoden [105]	Robuste Methoden zielen auf einen Kompromiss zwischen statistischen Eigenschaften und Robustheit. Sie erweisen sich als vorteilhaft, wenn wenige Daten verfügbar sind oder diese Ausreißern aufweisen. Häufig verwendet werden M-Schätzer (z. B. basierend auf der Huber-Funktion) und L-Schätzer (z. B. Median-Schätzer; Pendant zur l_1-Approximation).
Ridge-Methoden [55], [398]	Ridge-Regressionen werden angewendet, wenn betreffende Matrizen Kollinearitäten aufweisen. Dabei wird wie bei der Tikhonov-Regularisierung vorgegangen, wodurch die Streuung der Parameter auf Kosten eines Bias schrumpft.

Tabelle 9.2: Übersicht zu Regressionsmethoden

2. Black-Box-Modelle vermeiden
Black-Box-Modelle beschreiben Systeme durch Gleichungen, ohne die Komplexität und die dem System hinterliegende Physik zu berücksichtigen. Obwohl sich der Autor mehrfach mit der Modellierung durch Black-Box-Modelle intensiv befasste, überzeugten ihn die erzielten Ergebnisse nie richtig. Der Weg, auf eine theoretische Modellierung zu verzichten und schnell zu empirischen Lösungen zu gelangen, ist zwar verlockend, führt aber wegen der geringen Prozessdurchdringung nicht auf valide Ergebnisse. Die richtige Modellstruktur ist vielmehr der entscheidende Schlüssel zum Erfolg. Ein Gedankenexperiment soll das verdeutlichen: Angenommen ein System hat eine Totzeit in einem Rückkoppelzweig und die empirischen Ansätze kennen keine Totzeit oder lassen sie nicht in allen Zweigen zu, dann können die empirische Modelle zwar die gemessenen Daten unter Umständen gut anpassen, doch in der Regel zum Preis vieler Parameter. Darunter leidet die Robustheit des Modells, und es ist nicht sicher, wie weit System- und Modellverhalten bei anderen Eingangssignalen voneinander abweichen.

3. Vorwissen sammeln und sondieren
Um Vorwissen in die Modellbildung integrieren zu können, muss dieses Wissen zunächst erst einmal verfügbar sein. Auch hängt die Art des Vorwissens davon ab, ob ein dynamisches Modell entstehen soll oder ob ein Zusammenhang durch eine Funktion zu modellieren ist. Wenn beim dynamischen Modell die Modellstruktur und die Modellordnung (theoretische Analyse) geklärt sind, heißt es, die erwarteten Eigenschaften des Modells herauszuarbeiten. Als Anhaltspunkt hierfür können die Überschriften zu Kapitel 2 als eine Checkliste dienen. Selbstverständlich lassen sich weitere Eigenschaften hinzufügen. Desgleichen können die Überschriften von Kapitel 3 für Eigenschaften an Funktionen herangezogen werden.
Neben dem spezifischen Vorwissen über ein System oder eine Funktion sind je nach Anwendung auch physikalisch-technische Restriktionen an Parameter bekannt oder bei geometrischen Problemen Restriktionen zur Lage von Schwerpunkten oder Geraden sowie zur Ausrichtung von Objekten vorhanden.

4. Aktive Experimentation bevorzugen
Wenn es möglich ist, sollte eine aktive Experimentation vorgenommen werden. Bei aktiver Experimentation lässt sich nämlich das Testsignal designen und das System gezielt beeinflussen, während bei passiver Experimentation nur die protokollierten Prozesssignale verfügbar sind. Außerdem sollte bei aktiver Experimentation, wenn möglich (Stabilität, Sicherheit vorausgesetzt) an der offenen Strecke identifiziert werden. Wenn im geschlossenen Regelkreis identifiziert werden muss, ist die Stellgrößen-Regelgrößen-Beziehung gegenüber der Führungsgrößen-Regelgrößen-Beziehung zu bevorzugen. Kann nur über die Führungsgrößen-Regelgrößen-Beziehung identifiziert werden, sind Stellbegrenzungen zu vermeiden. Der Regler darf sonst nämlich nicht aus der Übertragungsfunktion des geschlossenen Kreises auf rein algebraische Weise herausgerechnet werden.

5. Messdaten überprüfen
Wichtig für eine erfolgreiche Modellbildung sind gute Prozessdaten. Es gilt also Datensegmente herauszusuchen, in denen
- eine prozesstypische Situation vorliegt,
- angemessene Prozessanregung stattfindet,
- keine Datenausreißer vorkommen,
- keine im Modell unberücksichtigten Einflüsse/Störungen auftreten und
- das Messrauschen gering ist.

Für niedrigdimensionale Systeme empfiehlt sich das, durch visuelle Beurteilung der Daten zu tun. Beim automatisierten Einsatz ist zunächst zu prüfen, welche der angesprochenen Aspekte das Computerprogramm zur Datenauswahl betrachtet. Typische Fehlerquellen in diesem Zusammenhang sind die Ergebnisverfälschung durch Ausreißer oder durch mangelnde Systemanregung.

Bei der passiven Experimentation ist das Datenmaterial nach sogenannten Eventsituationen zu untersuchen, in denen im Prozess etwas passiert. Dankbar für den Modellbildner, wenngleich unerwünscht für den Verfahrensingenieur, sind Ausfälle von Stellgliedern (Heizern, Pumpen). Da Regler ihrer Aufgabe entsprechend gegen den Modellbildner arbeiten – sie sorgen für Ruhe, der Modellbildner braucht Unruhe im Prozess – ist Datenmaterial aus der Inbetriebnahmephase meist mehr geeignet als jenes, wenn die Regelkreise bereits arbeiten. Sprungförmige Arbeitspunktwechsel sind ebenfalls interessant, Arbeitspunktwechsel über S-Kurven und Rampen sind dagegen weniger geeignet.

6. Ergebnisse visuell beurteilen
Bei geometrischen Problemen werden die Lösungen meist grafisch dargestellt, wohingegen bei Regressionen und Approximationen manchmal ausschließlich nach dem Gütewert beurteilt wird. Das ist gefährlich. So liefern Gleichungsfehlerkriterien besonders bei der Identifikation dynamischer Systeme trotz eines kleinen Zielfunktionswertes unter Umständen schlechte Modelle. Eine Simulation visualisiert solch schlechte Modelle sofort. Zu beachten ist aber, dass für die Simulation Anfangswerte benötigt werden, die Gleichungsfehlermethoden in der Regel nicht mitliefern. Das ist ein Argument, zumindest für niedrigdimensionale Offline-Modellbildungen dynamischer Systeme Ausgangsfehlermethoden zu verwenden.

Eine visuelle Beurteilung ist auch dann wichtig, wenn zwecks einfacherer Lösbarkeit auf Gütekriterien ausgewichen wird, die nicht dem natürlichen Fehlerempfinden entsprechen. Hierzu zählen in gewissem Sinn auch die Quadratfehlerkriterien. So wird bei der Anpassung von Punkten an eine Kurve intuitiv nicht das Quadrat, sondern vielmehr der Absolutbetrag, ja oft sogar der lotrechte Abstand beurteilt. Der Mensch denkt halt nicht quadratisch, was sich in vielen grafischen Darstellungen zur Geradenanpassung darin äußert, dass der Fehlereinfluss durch Striche statt durch die zugehörigen Fehlerquadrate kennzeichnet wird.

Zu den weniger intuitiven Fehlern zählen auch die Fehler der algebraischen (impliziten) Gleichung einer Kurve (z. B. Ellipsengleichung). Intuitiver sind in diesem Fall die geometrischen Fehler auf Basis der Parameterdarstellung der Kurve, die allerdings auf schwieriger auszuwertende Zielfunktionen führen. Wie bedeutend die Unterschiede in der Wahl der Gütekriterien für die praktische Anwendung letztlich sind, kann nicht allgemein gesagt werden, aber eine Visualisierung kann das weniger intuitive, leichter auswertbare Gütekriterium a posteriori rechtfertigen.

Abschließend sei aber vor einer alleinigen visuellen Beurteilung gewarnt. Schwierigkeiten wie große Parameterstreuungen oder -empfindlichkeiten können unerkannt bleiben. Als hilfreich erweisen sich dann die Parameterkovarianzmatrizen und Singulärwertuntersuchungen der Datenmatrizen. Zudem können lokale Nebenminima durch eine gute Modellanpassung ein globales Minimum vortäuschen, s. Beispiel 4.14. Diesem Problem kann durch mehrfache Optimierungsläufe mit unterschiedlichen Startwerten begegnet werden. Ein probates Mittel ist es auch, bei der Experimentation für eine hinreichende Anregung oder eine gleichmäßige Verteilung der Datenpunkte zu sorgen. Dadurch wird meist verhindert, dass das Nebenminimum etwa die gleiche Größe wie das globale Minimum hat, oder es wird erreicht, dass Nebenminima außerhalb des praktisch relevanten Parameterbereichs auftreten.

7. Parameterkovarianzen statt Fehlerkovarianzen betrachten
Bei Regressionsmodellen sollte – wenn möglich – die Parameterkovarianzmatrix mit angegeben werden. Es kann nämlich sein, dass die Gütewerte ein gutes Modell signalisieren, die Parameter aber hohe Varianzen haben. Das geschieht bei ungünstiger Verteilung der Daten, ungenügender Prozessanregung oder zu hohen Modellordnungen. Eine Hauptachsentransformation der Parameterkovarianzmatrix auf eine Diagonalmatrix der Eigenwerte lässt genauere Interpretationen zu. Die Eigenvektoren geben die Richtung der Hauptachsen an, die Kehrwerte der Wurzeln aus den Eigenwerten sind die Längen der zugehörigen Halbachsen des entsprechenden Ellipsoids. Damit signalisiert ein Eigenwert nahe Null eine große Streuung entlang der betreffenden Hauptachse. Gelegentlich sind solche großen Streuungen in spezielle Richtungen ein Indiz für eine bisher unbekannte Restriktion (lineare Abhängigkeit).

9.2 Arbeitspunkte und Statik bei der Identifikation

Die Identifikation linearer Modelle wird in der Literatur ausführlich behandelt. In der Praxis werden aber meist nur um einen Arbeitspunkt gültige lineare Modelle bestimmt. Dass dabei Gleichanteile zu eliminieren sind, die nicht im Zusammenhang mit den Kleinsignalen (Abweichungen um Arbeitspunkt) stehen, wird von Neueinsteigern auch schon mal vergessen. Hierauf und auf weitere Aspekte gehen die folgenden Thesen ein.

1. Klassisches Abziehen der Arbeitspunkte bevorzugen
Von den in Abschnitt 2.2 genannten Möglichkeiten Gleichanteile für die Identifikation linearisierter Modelle zu eliminieren, ist das Abziehen der Arbeitspunkte zu empfehlen, da es ein klares Bekenntnis zum Arbeitspunkt und zum Modellgültigkeitsbereich erfordert.

2. Arbeitspunkt zum Modell mit angeben
Wird das identifizierte Modell einer anderen Person übergeben, muss neben dem Modell auch der Arbeitspunkt und der Gültigkeitsbereich mit übermittelt werden. Fehlen diese Informationen, kann beispielsweise eine unbedarfte Simulation eines Ofenmodells, das für einen Arbeitspunkt von 1000°C gilt, dazu führen, dass der Ofen trotz Maximalleistung keine 500°C heiß wird. Die Ursache ist dann, dass bei 1000°C die Streckenverstärkung erheblich kleiner als bei niedrigen Temperaturen ist. Liegen hingegen die Informationen vor, wird der Arbeitspunkt zwangsläufig berücksichtigt – im Fall des Reglers enthält der Integratorwert die Heizleistung im Arbeitspunkt – und die Simulation sollte erfolgreich sein.

3. Statische Kennlinie vor der Dynamikidentifikation bestimmen
Bevor mit einer Erstellung dynamischer Modelle begonnen wird, sollte die statische Kennlinie oder das statische Kennfeld betrachtet werden. Hieraus ergeben sich Anhaltspunkte für den Modellgültigkeitsbereich. Ferner wird durch die Kenntnis der statischen Kennlinie gesichert, dass für die Identifikation linearer Modelle nur Arbeitspunktpaare (Eingangswert/Ausgangswert) auf der statischen Kennlinie gewählt werden.

4. Statikinformation bei Modellapproximation vererben
Angenommen ein komplexeres Modell sei bekannt und für den Reglerentwurf, eine Simulation oder eine andere Anwendung wird ein vereinfachtes Modell benötigt. Dann empfiehlt sich, dass das Approximationsverfahren, ggf. durch Hinzunahme von Restriktionen, die Statikinformation des komplexeren Modells übernimmt. Im Fall der Modellordnungsreduktion linearer Modelle heißt das, dass die statischen Verstärkungsmatrizen von komplexem und einfachem Modell übereinstimmen müssen. Ähnliches gilt, wenn ein nichtparametrisches Gewichtsfolgenmodell in ein parametrisches Übertragungsfunktionsmodell umgewandelt werden soll. Bei der Approximation eines nichtlinearen Modells durch ein einfacheres nichtlineares Modell sollten die statischen Kennlinien im Gültigkeitsbereich übereinstimmen. Außerdem sollten auch die Anzahl der Ruhelagen bzw. die Dimension der Ruhelagenmannigfaltigkeit erhalten bleiben.

5. Statikinformation zur Identifikation nichtlinearer Modelle nutzen
Bei niedrigparametrischen nichtlinearen Modellen gelingt unter Umständen eine Parameteridentifikation durch Identifikation mehrerer linearisierter Modelle, s. Abschnitt 2.1.3. Dadurch sinkt der Schwierigkeitsgrad.

6. Schätzung durch Statikrestriktionen verbessern
Das Einbeziehen von Vorwissen zur Statik bringt bei der Modellbildung in der Regel einen großen Gewinn. Das liegt daran, dass das Vorwissen zumeist auf Gleichungsrestriktionen führt, die ihrerseits den Parameterraum bei SISO-Systemen um eine Dimension verkleinern. Somit wird der Zufall auf den Messungen in weniger Parametern aggregiert, oder anders ausgedrückt, er wird besser weggemittelt.

Diese Aussage gilt allgemeiner für beliebige Gleichungsrestriktionen und aktive Ungleichungsrestriktionen (solche, bei denen die Lösung auf dem Rand liegt) [441], [233], [423].

7. Statikinformation bei Kennfeldern berücksichtigen
Gerade beim Verwenden empirischer Modellansätze ist darauf zu achten, dass das empirische Modell nicht gegen naturwissenschaftliche Gesetze verstößt. Die Statikinformation lässt sich aus Bilanzgleichungen gewinnen, aber auch durch Nullsetzen der Ableitungen bzw. Konstantsetzen der Signale bei zeitdiskreten Modellen.

8. Experimentation aus dem Arbeitspunkt heraus ausführen
Bei aktiver Experimentation empfiehlt sich eine Systemanregung aus dem Arbeitspunkt heraus. Neben der Tatsache, dass dann klar ist, für welchen Arbeitspunkt das Modell gilt, hat es den Vorteil, dass bei ausgangsfehlerorientierten Verfahren keine Systemanfangswerte geschätzt werden müssen.

9. Mittelwertfreie Signale bei aktiver Experimentation verwenden
Gegenüber Sprungsignalen führen Mäandersignale oder Pseudorauschternärsignale zu einer zweiseitigen Systemanregung. Zudem sind diese Signale mittelwertfrei, wodurch das Signal selbst keine Gleichwert- oder Arbeitspunktverschiebung hervorruft. Die Periodendauer der Signale sollte so gewählt werden, dass der für die Modellgültigkeit anvisierte Ausgangsbereich zu 95% ausgesteuert wird.

Aber: In der Praxis gibt es natürlich zahlreiche Beispiele, in denen sich eine mittelwertfreie Anregung verbietet. Angenommen der Zuckerabbau eines Menschen soll modelliert werden. Dann ist sicher eine impulsartige Anregung (eine bestimmte Anzahl Zuckerstücke wird gegessen) sinnvoll. Eine Sprunganregung scheidet in diesem Fall ebenfalls aus.

9.3 Identifikation dynamischer Modelle

Obwohl sich die Identifikation dynamischer Modelle praktisch zeitgleich mit der Regelungstechnik entwickelte, schließlich werden zum modellbasierten Entwurf dynamische Modelle benötigt, hat sie in der Lehre und Anwendung nie den Stellenwert wie die Regelungstechnik eingenommen. Vielfach beschränken sich die Lehrinhalte auf einfache Verfahren wie die Wendetangentenmethode oder diverse Quadratmittelverfahren auf der Basis von z-Modellen. Die

erhaltenen Modelle eignen sich aber nicht sonderlich für Zwei-Freiheitsgrade-Regelungen, da Wendetangentenmodelle relativ ungenau sind und z-Modelle den Makel der Nichtminimalphasigkeit tragen (Satz 2.18). Mangels guter Modelle kommen deshalb in der Praxis meist robuste Regelungslösungen in Betracht, bei denen zwangsläufig Performance verschenkt wird. In den letzten Jahren haben sich aber einige Änderungen vollzogen, die den Ingenieur heute in die Lage versetzen, gute Modelle mit vergleichsweise überschaubarem Aufwand zu erstellen. Eine erste neue Qualität ergibt sich aus der Möglichkeit, die theoretische Prozessanalyse computergestützt ausführen zu können, was besonders bei mechanischen, elektrischen und verfahrenstechnischen Strecken zum Tragen kommt. Als zweite neue Qualität seien die Fortschritte in der Computeralgebra genannt, mit deren Hilfe sich Gleichungen und Ungleichungen umstellen und vereinfachen lassen. Selbst das Lösen und Analysieren von Differenzialgleichungen ist nunmehr möglich. Als dritte neue Qualität ist die Verfügbarkeit leistungsfähiger numerischer Programme und der Portierbarkeit von Matlab-Lösungen in C-Programme zu nennen, die dem Ingenieur die praktische Umsetzung erleichtert und nicht mehr so tiefe Kenntnisse in Numerik und Optimierung voraussetzt wie zu früheren Zeiten. Die Herausforderungen liegen nunmehr in der Fähigkeit, die neuen Möglichkeiten zu nutzen, was das Beherrschen der Softwarewerkzeuge unabdingbar macht. Trotz dieser Fortschritte benötigt das Erstellen guter dynamischer Modelle immer noch eine gewisse Portion Erfahrung und Intuition. Beide werden benötigt, da Vorwissen über systemtheoretische Eigenschaften in vielen Fällen auf nicht einfach handhabbare Restriktionen führt. Um gerade Studenten und Neueinsteigern unschöne Erfahrungen zu ersparen, sind nachfolgend einige Empfehlungen und Erkenntnisse angeführt, die aus den praktischen Arbeiten des Autors und den theoretischen Darstellungen in dieser Arbeit stammen.

1. Zeitkontinuierliche Modellbeschreibungen bevorzugen
Zur Identifikation dynamischer Modelle werden in aller Regel abgetastete Signale verwendet, was eine Identifikation von z-Modellen nahelegt. Zudem überwiegt in Büchern die Beschreibung von Verfahren auf der Basis von Differenzengleichungen. Im Gegensatz dazu empfiehlt der Autor aber, zeitkontinuierliche Modelle sowohl im linearen als auch im nichtlinearen Fall zu verwenden. Als Begründung soll die moderne Literatur zur adaptiven und nichtlinearen Regelung dienen. In der Literatur zu adaptiven Reglern oder adaptiven Beobachtern mit linearen Modellen wird fast ausschließlich die Identifikation von zeitkontinuierlichen Modellen beschrieben. Der Grund liegt auf der Hand, denn für zeitkontinuierliche adaptive Regler ist der Entwurf und Stabilitätsnachweis schlichtweg einfacher als für rein zeitdiskrete adaptive Regler oder für hybride adaptive Regler mit zeitdiskretem Modell und kontinuierlichem Regler. Einen vergleichsweise großen Bogen um die zeitdiskreten Modelle und Regler tätigt auch die Literatur zu nichtlinearen Regelungen. Neben der schwierigeren mathematischen Behandlung hinsichtlich Entwurf und Stabilität kommt dort erschwerend hinzu, dass auch die zeitdiskreten nichtlinearen Abtastmodelle selbst erheblich komplizierter sein können als

ihre kontinuierlichen Pendants, siehe Beispiel 2.22. Neben den genannten Gründen gibt es weitere, die nicht dazu beitragen, dass sich zeitdiskrete Modelle aufdrängen. Hierzu zählen:

- Es gibt zeitdiskrete Modelle, die kein kontinuierliches Analagon haben. Als Beispiel sei ein zeitdiskretes Modell erster Ordnung mit negativ reellem Pol genannt, dessen Gewichtsfolge schwingt, was bei einem kontinuierlichen System erster Ordnung unmöglich ist, da dessen Gewichtsfunktion stets einen exponentiellen Verlauf hat.

- Minimalphasige zeitkontinuierliche LTI-Systeme mit einem Differenzgrad größer zwei sind als Abtastsysteme nach Satz 2.18 nichtminimalphasig. Das bedeutet aber, dass bei einem System dritter Ordnung ohne Zählerdynamik für das kontinuierliche Modell der Reglerentwurf, der Entwurf einer Vorsteuerung auf triviale Weise gelingt, während das für das zeitdiskrete Modell erheblich schwieriger ist.

- Der Differenzgrad vererbt sich vom kontinuierlichen nicht auf das zeitdiskrete Abtastmodell mit Halteglied. Er ist beim zeitdiskreten Modell in der Regel kleiner. Das hat einige Konsequenzen:
  - Trotz gleicher Modellordnung vielfach mehr zu identifizierende Parameter.
  - Mehr Parameter bedeuten eine schlechtere Schätzung.
  - Die Rückrechnung produziert messstörungsbedingt zusätzliche Zählerdynamik.
  - Der Entwurf einer Vorsteuerung beim Reglereinsatz wird erschwert.

- Der Einfluss der Abtastzeit bei der Identifikation ist auf zeitdiskrete Modelle größer als auf zeitkontinuierliche [247].

- Die Rückrechnung vom zeitdiskreten Modell in ein zeitkontinuierliches ist bei Systemen mit Totzeit nicht eindeutig, siehe Beispiel 4.13.

- Das Stabilitätsgebiet der zeitdiskreten Modelle ist das Innere des Einheitskreises, das der zeitkontinuierlichen die gesamte linke offene komplexe Halbebene. Daraus resultiert, dass zeitkontinuierliche Modelle durch Störungen auf den Messsignalen weniger dazu neigen, ohne stabilitätssichernde Restriktionen instabil zu werden. Die hohe Empfindlichkeit der Pole bezüglich der Nennerparameter unterstützt diese Aussage, siehe hierzu Beispiel 2.7.

- A-priori-Information zu Nullstellen des kontinuierlichen Modells lässt sich für Abtastmodelle mit Halteglied nicht nutzen, da zwischen den Nullstellen beider Modelle kein einfacher Zusammenhang existiert. Aber bei Verwendung von Tustin-Abtastmodellen besteht ein einfacher Zusammenhang (gegeben durch Bilineartransformation).

Anders als in der Regelungstechnik wiegen die Nachteile der zeitdiskreten Modelle in der Signal- und Datenanalyse weniger schwer, weshalb dort die zeitdiskreten Modelle wie Autokovarianzfolgen, diskrete Spektren, ARMA-Prädiktionsmodelle usw. verbreitet sind.

2. Zustandsvariablenfilterverfahren für Online-Identifikation nutzen
In den 1990er Jahren wurden viele Verfahren zur Online-Identifikation untersucht. Das entscheidende Problem bei der Identifikation von Differenzialgleichungen sind die Ableitungen. Das numerische Bilden der Ableitungen scheidet wegen der Störaufrauhung dabei aus. Also müssen zusätzliche Filter verwendet werden, die mit endlichem oder unendlichem Gedächtnis ausgelegt sein können. Untersuchungen des Autors in [247] ergaben, dass Filter mit unendlichem Gedächtnis Vorteile bieten. Die Wahl des Filtertyps ist dabei weniger kritisch; viel entscheidender ist die Wahl der Grenzfrequenz. Die Filterordnung muss mindestens der Systemordnung entsprechen. Letztlich sind die Zustandsvariablenfilter nichts anderes als die in den 1970er Jahren und in der modernen Literatur zu adaptiven Regelungen verwendeten Tiefpässe nur in geschickter Implementierung (Zustandsraum; unterschiedliche Halteglieddiskretisierungen für Eingangs- und Ausgangssignale). Weitere Hinweise, die die eindeutigkeitserzwingende Restriktion betreffen, finden sich in Abschnitt 4.4.4.

3. Zweckmäßige Koordinaten wählen
Gerade bei mechanischen Systemen gibt es eine Reihe von Alternativen, die Koordinaten für die Modellierung zu wählen. Je nach mechanischem Objekt, abhängig von kinematischer oder dynamischer Modellierung sowie dem Verwendungszweck (Identifikation, Simulation oder Systemanalyse) sind einige Darstellungen zweckmäßig oder eben nicht. Die Alternativen erstrecken sich von den generalisierten Koordinaten im Lagrange-Formalismus, den Orts- und Impulskoordinaten in Hamiltonschen Systemen, den Quaternionen für Drehbewegungen, der Denavit-Hartenberg-Notation in der Robotik, den Study-Parametern für die Vorwärtskinematik von Parallelrobotern, den Plücker-Koordinaten in der Liniengeometrie[3] für ungerichtete Geraden oder den dualen Vektoren[4] für gerichtete Geraden [305].

Die Wahl zweckmäßiger Zustandsraumkoordinaten (Zustandsgrößen) kann auch in der Regelungstechnik und der Identifikation helfen, die Arbeit zu vereinfachen. Im Abschnitt 2.3.2 werden hierfür einige weniger bekannte Normalformen angeführt. In diesem Zusammenhang sei auch auf die Vorteile von Deskriptordarstellungen verwiesen (im Linearen $E\dot{x} = Ax + Bu$,

[3] Im Gegensatz zur üblichen Raumpunktgeometrie, in der drei Koordinaten einen Punkt charakterisieren, bilden in der Liniengeometrie die Geraden der Ausgangspunkt jeglicher Betrachtungen. Ein Punkt ist als Schnittpunkt zweier Geraden darstellbar.

[4] Duale Vektoren basieren auf dualen Zahlen (nicht den binären Zahlen!), die ihrerseits ein algebraisches Objekt darstellen, das eng mit dem Begriff des Tangentialvektors verbunden ist. Die dualen Zahlen bilden eine zweidimensionale hyperkomplexe Algebra über $\mathbb{R}$ und sind mit den komplexen Zahlen vergleichbar, allerdings gilt für das nicht-reelle Element ε in $z = a + \varepsilon b$ nunmehr $\varepsilon^2 = 0$ statt wie bei komplexen Zahlen $\mathrm{i}^2 = -1$.

d. h. $\dot{x}$ ist nicht isoliert wie in der Kalman-Zustandsraumdarstellung). Sie entstehen in natürlicher Weise aus der Kombinationen von gewöhnlichen Differentialgleichungen und algebraischen Nebenbedingungen (Knoten- bzw. Maschenbedingungen in Stromkreisen, geometrische Zwangsbedingungen in Mechanik). Gegenüber Zustandsraumsystemen enthalten die Deskriptormodelle oft deutlich mehr Nullelemente und damit weniger zu bestimmende Parameter. Überdies haben die Parameter einen direkteren Einfluss, was eine bessere Topologie des Gütegebirges bewirkt, als wenn sie über komplizierte Parameterverknüpfungen nach Inversion von E einfließen. Letztlich bietet bei singulärem E die Aufspaltung in ein schnelles und langsames System Potenzial für Modellvereinfachungen.

4. Theoretische Modelle vereinfachen

Bekanntermaßen ist es schwieriger, hochparametrische Modelle zu identifizieren und zu regeln als niederparametrische. Deshalb sollte das theoretische Modell nach Möglichkeiten einer Vereinfachung untersucht werden. Gerade für Neueinsteiger ist es schwierig zu erkennen, welche Vereinfachungen sich anbieten. Nachfolgend wird eine kleine Auswahl diskutiert:

- Die kleine Winkelnäherung, bei der $\sin\theta \approx \theta$, $\tan\theta \approx \theta$ oder $\cos\theta \approx 1$ genutzt wird, ist vom mathematischen Pendel mit kleinen Ausschlägen gut bekannt. Sie gilt aber auch für kleine Differenzwinkel $\sin(\psi_1 - \psi_2) \approx \psi_1 - \psi_2$, wobei die Winkel ψ_1, ψ_2 gleichwohl groß sein können.

- Schnelle Dynamiken können, wenn die langsameren Dynamiken primär interessieren, durch ihre statische Verstärkung ersetzt werden. Das reduziert die Modellordnung zum Teil erheblich.

- Die Verwendung der Krümmung κ statt des Momentanradius r bei der Modellierung von Kurven bietet den Vorteil, dass die Gerade durch $\kappa = 0$ und nicht durch $r = \infty$ ausgedrückt werden kann. Das Beispiel lehrt, dass bei einem Modell immer auf eine stetige Änderung von Parametern zu achten ist und das unendliche große Werte als auch Singularitäten zu vermeiden sind. Eine geringe Krümmung bietet des Weiteren Potenzial zur Modellvereinfachung, denn häufig greift die Näherung $\frac{1}{1-d\kappa} \approx 1$.

- Unterlagerte schnelle Regelkreise können vielfach durch lineare Verzögerungsglieder erster Ordnung ersetzt werden. Nichtlinearitäten, gegen die der unterlagerte Regler ankämpft, als auch die Dynamik der unterlagerten Strecke brauchen dann nicht berücksichtigt werden. Als Verstärkung für das Verzögerungsglied ist $K = 1$ ein guter Anhaltspunkt (vorausgesetzt der Regelkreis hat keinen nennenswerten bleibenden Regelfehler).

- Eine entkoppelte Betrachtung kann ebenfalls helfen, die Komplexität zu verringern. Hierbei wird beispielsweise statt einer gekoppelten Zweigrößenstrecke nur eine Eingrößenstrecke betrachtet, wobei die Koppelwirkung sich dann durch einen zeitvariablen

Parameter äußert. Als Beispiel sei die entkoppelte Betrachtung von Längs- und Querdynamik eines Fahrzeugs in Standardfahrsituationen genannt. Besonders einfach wird es, wenn sich dieser Parameter relativ zur Strecke langsam ändert, vgl. Abschnitt 2.8. Vorteilhaft ist auch, wenn die einkoppelnde Größe messbar ist, da dann der zeitvariable Parameter bei der Modellerstellung bekannt ist.

- Wesentliche Vereinfachungen lassen sich auch erzielen, wenn komplizierte Dynamiken als Störungen aufgefasst werden. Die Idee stammt aus der Regelungstechnik, wo nichtlineare Dynamiken in der Mechanik (Reibungen, Coriolisterme) in einer Störung zusammengefasst werden und ein schneller Störgrößenbeobachter diese permanent schätzt, um sie alsdann per Störgrößenaufschaltung zu kompensieren.

- Eine aufwendige, aber erfolgversprechende Technik ist die Sensitivitäts- und Einflussanalyse. Gerade bei der computergestützten Erstellung von Modellen kommt es oft vor, dass die Modellgleichungen mehrere A4-Seiten füllen. Für solche Modelle experimentell die Parameter zu bestimmen, ist ein Unding. Da helfen auch Restriktionen nicht weiter. Dann ist es ratsam, sich zunächst eine Tabelle zu erstellen, in der Intervalle für die Prozessparameter notiert werden. Bei der Einflussanalyse wird dann untersucht, ob eine Änderung des Parameters durch Einsetzen der Intervallgrenzen einen signifikanten Einfluss auf das Systemverhalten hat. Bei mechanischen Systemen können so Teilsysteme mit schwachen translatorischen und rotatorischen Trägheiten erkannt und im Weiteren vernachlässigt werden. Bei der Sensitivitätsanalyse werden jene Parameter gesucht, auf die Ausgangs- oder interessierende Zustandsgrößen empfindlich reagieren. Diese Parameter gilt es im Modell auf jeden Fall zu berücksichtigen. Für sie sind auch Restriktionen besonders wichtig, da sie helfen können, die Schätzgüte zu erhöhen.

- Durch Kompensation von Eingangs- und Ausgangsnichtlinearitäten gelingt es mitunter, auf die Identifikation nichtlinearer Wiener- oder Hammersteinmodelle zu verzichten. Als Beispiel sei ein Gleichspannungsheizer an einem Ofen genannt. Statt des Spannungswertes, der wegen der Beziehung zwischen Spannung und Leistung quadratisch eingeht, wird als Eingangsgröße die Leistung gewählt. Somit kann das Ofenmodell in erster Näherung durch ein lineares Verzögerungsglied erster oder zweiter Ordnung, je nach Aufbau der Isolation, beschrieben werden. Andere Beispiele sind die Fahrzeuglenkung mit $u := \tan\delta$, wo mit dem Tangens des Lenkwinkels statt direkt mit dem Lenkwinkel gearbeitet wird. Desgleichen können nichtlineare Beziehungen zwischen Höhe und Volumen in ungleichmäßig geformten Behältern (Zylinder plus Kegel) kompensiert werden.

- Die Dimension der Ruhelagenmannigfaltigkeit bestimmt bei holonomen, also insbesondere bei linearen oder bei linearisierten hyperbolischen Systemen (keine Eigenwerte

auf imaginärer Achse), die Menge der unabhängig einregelbaren Größen. Mit einem SISO-LTI-System lässt sich demnach nur eine Größe einregeln. Als Klassiker sei der Doppelintegrator genannt, der beispielsweise eine reibungsfrei bewegte Masse nach dem Newtonschen Grundgesetz $F = m\ddot{x}$ beschreibt. Mit der Kraft als Eingang kann nur die Geschwindigkeit oder nur die Lage eingeregelt werden, aber nie beide gleichzeitig. Für die Modellbildung heißt das, dass die Zahl der Ein- und Ausgänge übereinstimmen muss. Ferner ergeben sich Schlussfolgerungen für die Wahl der Zustandsgrößen und Hinweise auf Modellvereinfachungspotenziale.

5. Vorwissen über Pole und Eigenwerte nutzen
Integratoren und Differenzierer im System, bekannte Pole in Teilsystemen, die bekannte Entwurfsdynamik eines unterlagerten Regelkreises oder bekannte Schwingungsmoden führen auf Gleichungsrestriktionen (meist sogar lineare). Dieses Vorwissen kann damit relativ einfach und zudem wirkungsvoll in der On- und Offline-Modellbildung genutzt werden. Schwieriger ist die Situation, wenn das Vorwissen weniger präzise durch Ungleichungen gegeben ist. Hierfür lassen sich dann als einfach zu handhabende Restriktionen oft nur notwendige Bedingungen formulieren. Im Abschnitt 2.2 wird gezeigt, wie die Umsetzung des Vorwissens erfolgt und was der Ingenieur zu beachten hat. Hervorzuheben ist der Unterabschnitt 2.2.10, der sich Restriktionen an die Pole widmet, die sich aus dem Abtasttheorem ableiten lassen.

6. Nach Diagonaldominanzen suchen
Ist ein LTI-System stabil, so liegen dessen Pole oder Eigenwerte in der linken offenen komplexen Halbebene. Eine Restriktion ist damit zwar schnell formuliert, ihre Handhabung aber ist sehr kompliziert. Die Situation ändert sich, wenn dass System eine Diagonaldominanz oder eine partielle Dominanz (nur ausgewählte Zustände betreffend) aufweist. In einem solchen Fall ergeben sich einfach zu handhabende lineare Ungleichungsrestriktionen an die Parameter, mit denen die Stabilität des Modells sichergestellt werden kann. Im Abschnitt 2.5.2 werden hierzu praktische Beispiele genannt. Selbst wenn keine Diagonaldominanz vorliegt, können Betrachtungen zur Diagonaldominanz helfen, die Stellgrößen-Regelgrößenzuordnung (E/A-Zuordnung) bei Mehrgrößenregelungen festzulegen. Hierzu können mehrere Modelle mit unterschiedlicher Zuordnung identifiziert werden und a posteriori wird diejenige Zuordnung und damit das Modell gewählt, das der Diagonaldominanz am nächsten kommt.

7. Stabilität in vielfältiger Weise sichern
Im Kapitel 2 nehmen Techniken einen breiten Raum ein, mit denen die Stabilität von Modellen garantiert werden kann, wenn das System stabil ist. Die Breite ist erforderlich, da sich Stabilität in den unterschiedlichen Darstellungsformen auch unterschiedlich äußert. Außerdem führen Stabilitätsrestriktionen in aller Regel auf nichtlineare und nichtkonvexe Beziehungen, was die Existenz prädestinierter Lösungszugänge nahezu ausschließt. Aus der Vielzahl möglicher Zugänge wurden erfolgversprechende aus der Literatur zusammengestellt, bewertet

und gegebenenfalls modifiziert. Die Zugänge reichen von Reparametrisierungen, dem Einsatz von nur notwendigen oder nur hinreichenden Bedingungen bis hin zu Formulierungen mit semidefiniten Restriktionen. Hervorzuheben sind in diesem Rahmen auch die Darstellungen zu Restriktionen, die sich durch die Betrachtung von Intervallsystemen ergeben, siehe Abschnitt 2.7. Dabei wird nicht nur ein Modell, sondern eine ganze Familie von Modellen betrachtet. Während diese Zugänge in der Regelungstheorie zur Stabilitätsanalyse und zum Stabilitätsnachweis in einigen Fällen erfolgreich eingesetzt werden, blieben sie in der Modellbildung bisher weitgehend unbeachtet. Einschränkend sei aber erwähnt, dass sie – wie in der Regelungstheorie auch – nur dann effektiv sind, wenn sich die Zahl der Intervallparameter auf wenige Parameter in einem Modell beschränkt oder die Intervalle hinreichend klein sind. Klar ist nämlich, dass sich mit der Anzahl der Intervallparameter und mit der Größe der Intervalle die Chance erhöht, dass in der Modellfamilie ein instabiles Modell liegt.

Hervorzuheben ist neben den klassischen Stabilitätskonzepten, dass in der Arbeit vorgestellte Konzept der praktischen Stabilität, welches in der Modellbildung anders als in der Regelungstechnik bisher keine Beachtung gefunden hat. Bei diesem Konzept wird im Gegensatz zu den klassischen Konzepten nicht nur das asymptotische Verhalten charakterisiert, sondern auch das transiente. Letztlich wird gezeigt, dass der Grad der Nichtnormalität einer Hurwitz-stabilen Systemmatrix wesentlich beeinflusst, ob sich die Norm des Zustandsvektors eines autonomen Systems zunächst nennenswert vergrößert – eventuell praktisch gefährlich bezüglich einer Zustandskomponente – um nach Erreichen eines Maximums asymptotisch gegen Null zu streben.

8. Instabile Modelle als Motivationsquelle ansehen
Ein instabiles geschätztes Modell für eine stabiles System erscheint zunächst als ein nutzloses Ergebnis, aber es erfordert auch, das Modell und/oder das Verfahren neu zu überdenken. Ideen hierfür ergeben sich aus einigen typischen Ursachen für die Instabilität:
- Kausalität verletzt, siehe Beispiel 2.28
- unberücksichtigte Messverzögerungen, Totzeiten
- Modellordnung zu hoch
- Modellstruktur falsch
- Kollinearitäten infolge von Rückkopplungen
- zu geringe Systemanregung (falsche oder zu wenige signifikante Frequenzen)
- schlecht gewählte Abtastzeit
- Ausreißer
- schlechtes Stör-Nutzsignalverhältnis

9. Passivität vererben und für die Regelung nutzen
Im Abschnitt 2.10 werden zunächst unterschiedliche Konzepte vorgestellt, und es wird erklärt, in welchen Zusammenhang die positiv reellen Funktionen zu den passiven Systemen

stehen und wie sich daraus Restriktionen ergeben. Wichtig für Neueinsteiger sind die Aussagen zum relativen Grad, den ein System haben muss, um überhaupt passiv zu sein und die Tatsache, dass das Produkt aus Ein- und Ausgangssignal physikalisch einer Leistung entsprechen sollte.
Passivität ist ähnlich wie die Stabilität schwierig zu handhaben. Diese Eigenschaft wird deshalb selten bei der klassischen Parameterschätzung – abgesehen von einfachen oder statischen Systemen – berücksichtigt. Anders ist die Situation bei der Modellreduktion, bei der die Passivitätseigenschaft auf das reduzierte Modell vererbt werden soll. Hierfür werden die entsprechenden Sätze und Literaturquellen bereitgestellt.
Für den Regelungstechniker ist die Passivität eines Systems eine schöne Eigenschaft, zumal einige Entwurfsverfahren sogar auf der Passifizierung durch Regelung oder durch dynamische Kompensatoren beruhen. Soll also ein Modell für die Regelung verwendet werden, dann kann unter Umständen durch geschickte Wahl der Stell- und Regelgrößen (Eingangs-/ Ausgangsgrößen der Strecke) ein passives Modell entstehen.

10. Minimalphasigkeit betrachten
Ob eine SISO-Strecke minimalphasig ist oder nicht, lässt sich experimentell anhand von Sprungexperimenten ermitteln. Aus dem Sprungexperiment ist eine Totzeit leicht ablesbar. Hat das System eine Totzeit, ist es stets nichtminimalphasig. Zudem kann die erkannte Totzeit bei einer Identifikation strukturell berücksichtigt werden, denn zur Identifikation sollten die Sprungsignale selbst nicht unbedingt verwendet werden, siehe Abschnitt 9.2 und Beispiel 4.14. Reagiert das System an der Stelle $t = 0$ mit einer entgegen der stationären Endlage gerichteten Bewegung, dann liegt eine ungerade Anzahl von Nullstellen auf der rechten komplexen Halbebene vor. Antwortet es auf einen Sprung zwar in Richtung des stationären Endwerts, zeigt dann aber mehrere Änderungen des Anstiegs, ohne dabei zu schwingen, dann liegt eine gerade Anzahl nichtminimalphasiger Nullstellen vor. Dieses Verhalten gilt gleichwohl im Nichtlinearen, obschon dann nicht von Nullstellen, sondern von instabiler Nulldynamik gesprochen wird. Aus der Anzahl der nichtminimalphasigen Nullstellen ergeben sich Vorzeichenrestriktionen, die sich aus den Vorzeichenregeln von Descartes ergeben.

Wesentlich komplizierter als im SISO-Fall sind die Zusammenhänge zur Minimalphasigkeit im MIMO-Fall und im Fall der nichtlinearen Systeme. Das schränkt das Formulieren handhabbarer Restriktionen ein. Deshalb sind die Ergebnisse und Aussagen des Abschnitts 2.13 mehr systemtheoretisch als modellbildungsmäßig orientiert. Für den Anwender heißt das, dass er zweckmäßigerweise sein Modell a posteriori auf Minimalphasigkeit untersucht. Liegt diese nicht vor, gestaltet sich der Reglerentwurf schwieriger. Dann kann es ratsam sein, die Wahl der Stell- und Regelgrößen und damit das Modell zu überdenken. Als Regel gilt: Je weiter Ein- und Ausgangsgrößen aus dynamischer Sicht auseinander sind, desto unwahrscheinlicher ist ein Nichtminimalphasensystem, denn eine Erhöhung des relativen Grads reduziert die Nullstellenanzahl und damit die Gefahr von Nichtminimalphasennullstellen.

11. Generische Eigenschaften a posteriori bewerten
Generische Eigenschaften sind vereinfacht gesprochen jene, die durch algebraische Gleichungsrestriktionen beschrieben werden. Im zweidimensionalen Raum ist beispielsweise eine Gerade eine solche generische Restriktion, wenn alle Punkte außerhalb der Geraden zulässig sind, die auf der Geraden aber nicht. Obwohl die Gerade selbst unendliche viele Punkte hat, sind die ausgeschlossenen Punkte aber eine magere Menge in der Ebene. Anschaulich bedeutet das, dass bei einer Identifikation fast nie ein Punkt (Modellparametersatz) auf der Geraden ermittelt wird.
Zu den generischen Systemeigenschaften gehören die Steuerbarkeit und die Beobachtbarkeit. Der Grund liegt bei beiden Eigenschaften in der Rangrestriktion, die durch das Kalman-Kriterium gegeben ist, das seinerseits eine endliche Menge algebraischer Gleichungsbedingungen enthält. Eine weitere Eigenschaft ist der Grenzstabilität, da die Menge der Systeme mit defektiven Eigenwerten auf der imaginären Achse (solche sind nicht Lyapunov-stabil) mager ist.
Der Vorteil generischer Eigenschaften ist es also, dass sie bei der Identifikation nicht berücksichtigt werden müssen, da sie ja fast immer von allein eingehalten werden. Generische Eigenschaften eröffnen aber ein Potenzial für die A-posteriori-Modellbewertung. So kann etwa der Abstand eines identifizierten steuerbaren Modells zu den nichtsteuerbaren Modellen bestimmt werden. Hieraus und aus der Analyse zum Grad der Steuerbarkeit einzelner Moden können Aussagen zur Wahl der Testsignale, zum bei der Regelung zu erwartenden Stellaufwand oder zur Vereinfachung (Vernachlässigung schlecht steuerbarer Moden) getroffen werden.

Aber: Obwohl defektive Eigenwerte zwar nicht generisch sind, heißt das nicht, dass sie in der Praxis nicht auftreten. Der Doppelintegrator in Zustandsraumdarstellung mit seinem defektiven Nulleigenwert ist ein entsprechendes Beispiel.

12. Naive Verallgemeinerungen vermeiden
Neueinsteiger in der Regelungstheorie und Modellbildung, aber auch langjährige Praktiker begehen nicht selten einen klassischen Fehler. Sie schließen aus der Theorie der LTI-Systeme durch Analogie auf Eigenschaften (Stabilität, Steuerbarkeit usw.) linearer zeitvariabler Systeme oder auch nichtlinearer Systeme. Oft mag dieser Schluss – wenn auch unzulässig – zum richtigen Ergebnis führen, dennoch ist das gefährlich. Im Kapitel 2 und im Anhang A.6.4 finden sich daher zahlreiche Gegenbeispiele, die den Leser diesbezüglich sensibilisieren sollen.

9.4 Identifikation von Funktionen

Die Anwendungspotenziale für Restriktionen an Funktionen in Kapitel 3 gehen weit über die klassische Identifikation parametrischer statischer Zusammenhänge und nichtparametrischer Stützwertemodelle von Folgen oder Spektren hinaus. So können die vorgestellten Techniken auch in der Bildverarbeitung, der Bahnplanung oder in nichttechnischen Anwendungsfeldern wie der Medizin oder Psychologie eingesetzt werden. Einige für technische Anwendungen wichtige Ergebnisse und Empfehlungen fassen die folgenden Thesen zusammen.

1. Information über die Funktion zusammentragen
Das Erstellen von Vorwissen zu einer Funktion geht am einfachsten, wenn wie bei der Kurvendiskussion vorgegangen wird. Einige hierfür hilfreiche Fragestellungen sind nachfolgend aufgeführt:
- Hat die Funktion spezielle Punkte, durch die sie geht? (Nullpunkt, Pi, y-Nullpunktoffset)
- Hat die Funktion Asymptoten?
- Hat die Funktion Singularitäten?
- Ist die Funktion stetig, nichtnegativ, monoton, konvex usw.?
- Ist die Funktion bezüglich einzelner Koordinaten monoton?
- Weist die Funktion Symmetrien auf? (gerade, ungerade, radialsymmetrisch, periodisch)
- Wie glatt ist die Funktion?
- Gibt es Integralrestriktionen? (Eins bei Dichten, Null bei periodischen Funktionen)
- Liegen Funktionswerte oder ihre Ableitungen in bekannten Ordinatenintervallen?
- Ist die Funktion unimodal oder sogar A-unimodal (s. Def. 3.4)?
- Welche Aussagen lassen sich über die Subniveaumengen treffen?
- Gelten Quadranten- oder Orthantenbeziehungen? (Sektorbedingungen)
- Gelten die zuvor herausgearbeiteten Eigenschaften intervallweise oder global?
- Soll die Funktion parametrisch oder nichtparametrisch beschrieben werden?
- Liefert die Physik Hinweise auf die Funktionenklasse? (Basislösungen)

Zu all diesen Fragen finden sich mögliche Antworten, Erklärungen in Kapitel 3, wobei die Zugänge, um das akquirierte Vorwissen letztlich mathematisch umzusetzen, im Mittelpunkt stehen.

2. Orthogonale Funktionensysteme verwenden
Beim Erstellen empirischer Modelle ist von vornherein oft nicht klar, wie hoch die Approximationsordnung zu wählen ist. Um dann bei Erhöhung der Ordnung eine möglichst große Reduktion des Fehlers zu erhalten, empfehlen sich orthogonale Funktionensysteme. Ähnlich der dem Ingenieur gut bekannten Fourier-Reihenanalyse, bei der die Orthogonalität von Sinus- und Kosinusfunktionen genutzt wird, gibt es auch orthogonale Polynomsysteme (z. B. Legendre-Polynome) [163].

Auch für normierte Räume (L^1, L^∞), die keine Hilbert-Räume sind und denen damit die Orthogonalität fehlt, gibt es geeignete Funktionensysteme. Welches System für die Anwendung das richtige ist, erfährt der Ingenieur in aller Regel nach einer kurzen Internetrecherche oder aus [568], [298], [506].

3. Polynomansätze nicht bei Singularitäten verwenden
Für Neueinsteiger sind Polynomansätze, da parameterlinear und mit Standardquadratmittelzugängen leicht handhabbar, oft das Mittel der ersten Wahl für viele Probleme. Dabei wird verkannt, dass sie für Funktionen mit Singularitäten ungeeignet sind, weil sie diese strukturell nicht abbilden können. Gebrochenrationale Funktionen oder andere Funktionen wie Tangens, Cotangens, Cotangens Hyperbolicus, Areatangens Hyperbolicus usw. sind dann die bessere Wahl, um das Verhalten nahe einer Singularität abzubilden.

4. Wahl des Funktionsansatzes nach Eigenvorgängen richten
Aus der Identifikation linearer Systeme ist bekannt, dass sich eine Lösung aus Eigenvorgängen und erzwungenen Vorgängen zusammensetzt. Auch kommen als Eigenvorgänge nur die zu wichtenden Funktionen $1, t, \mathrm{e}^{-at}, \sin(\omega t), \cos(\omega t), \mathrm{e}^{-at}\sin(\omega t), \mathrm{e}^{-at}\cos(\omega t)$ in Frage, wenn nichtgenerische Mehrfacheigenwerte/-pole außer Acht gelassen werden. Dieser Pool potenzieller Funktionen lässt sich durch theoretische Überlegungen eventuell weiter einschränken.

Obwohl für nichtlineare Systeme das Superpositionsprinzip nicht gilt, gelingt es bei einigen Systemen niedriger Ordnung und bei einfacher Systemanregung durch Sprünge, Impulse oder Sinusfunktionen, die Lösung in geschlossener Form anzugeben, vgl. hierzu auch Tabelle 7.3. Geschlossene Lösungen können des Weiteren durch Verwenden computeralgebraischer Löser oder durch Nachschlagen in Standardwerken [653], [333] erhalten werden. Letzteres empfiehlt sich für eine Reihe ein- und zweidimensionaler physikalischer Probleme, da in den Standardwerken auch Bedingungen für diverse Fallunterscheidungen angegeben werden. Letztlich stehen durch dieses Vorgehen – vorausgesetzt es ist erfolgreich – strukturell die richtigen Funktionsansätze zur Verfügung. Über den Umweg einer speziellen Lösung können somit einige nichtlineare System identifiziert werden, ohne auf numerische Lösungen der Differenzialgleichung zurückgreifen zu müssen.

5. Kennlinie nach Lösbarkeit der Differenzialgleichung auswählen
Statische Kennlinien von SISO-Systemen erscheinen bei einigen dynamischen Systemen als Absolutglied der Differenzialgleichung. Um die Vorteile geschlossener Lösungen zu nutzen – Identifikation der verbleibenden Parameter ist ohne Simulation der Differenzialgleichung möglich, Stabilitätsuntersuchung kann an Lösung statt indirekt erfolgen – ist es ratsam, strukturell solche Ansatzfunktionen zu wählen, mit denen direkt oder nach Substitution eine geschlossene Lösung der Differenzialgleichung gelingt. Für Kennlinien mit gesättigtem Anstieg werden solche Ansätze im Punkt 2 in Abschnitt 3.4.2 vorgestellt.

6. Translations- und Rotationsinvarianz beachten
Werden Funktionen zur Modellierung geometrischer Probleme verwendet, so ist insbesondere darauf zu achten, dass sich die entstehende Lösung in Relation zu Punkten nicht ändert, wenn die Punkte verschoben oder gedreht werden. Im Beispiel 4.16 wird gezeigt, dass das nicht der Fall sein muss, wenn Restriktionen ungeschickt gewählt werden. Auch eine ungeschickte Wahl der zu minimierenden Fehler kann ein Verhalten wie im Beispiel bewirken. Die Verwendung orthogonaler Abstandsfehler verhindert das. Zur Wahl geeigneter Normen bei der Kurvenanpassung und Funktionsapproximation sei auf [622] verwiesen.
Parametrische Darstellungen von Kurven und Flächen sind unter dem Gesichtspunkt der Invarianzen ebenfalls empfehlenswert (Abstände lassen sich teils einfacher berechnen; zusätzliche Restriktionen wie in Gleichungsdarstellungen entfallen durch angepasste Struktur der Parameterdarstellung), siehe auch Abschnitt 5.2.9. Für ebene Kurven empfiehlt sich darüber hinaus eine Parametrisierung über die Bogenlänge oder indirekt über die Krümmung bzw. Krümmungsänderung (Anwendung bei autonomen Fahrzeugen).
Eng verwandt mit der Invarianzproblematik ist die Frage nach der Allgemeingültigkeit eines Zugangs unter gedrehten Situationen. Am einfachsten lässt sich das anhand einer Geraden erklären. $y = mx + n$ ist als Modell nicht in der Lage, eine senkrechte Gerade darzustellen. Wird die Gerade dagegen in Hessescher Normalform $ax + by - d = 0$ mit $a^2 + b^2 = 1$ dargestellt, ist das kein Problem.

7. Semi-unendliche Probleme relaxieren
Durch intervallweise Restriktionen an die Nichtnegativität oder die Monotonie werden semiunendliche Probleme formuliert. „Semi-“, da die Zahl der Restriktionen unendlich ist (muss für jeden der unendlich vielen x-Werte im Intervall gelten), aber die Zahl der Parameter selbst endlich ist. Ein beliebte und wirkungsvolle Technik ist es, statt der unendlich vielen Restriktionen nur eine hohe Anzahl punktweiser (meist äquidistanter) Restriktionen zu betrachten. Das ist immer dann ratsam, wenn die punktweisen Restriktionen dabei linear sind und die Parameter ebenfalls linear im Funktionsansatz auftreten.

8. Autokovarianzfolgen zweckbezogen schätzen
Autokovarianzfolgen (AKF) spielen beim Erkennen von Zusammenhängen und als Zwischenmodell in der Signalanalyse eine wichtige Rolle. In der Literatur gibt es unterschiedliche Formeln, mit denen diese Folgen geschätzt werden können. Im Abschnitt 3.9 werden diese Formeln diskutiert. Ein entscheidendes Problem bei der Schätzung der Autokovarianzfolgen rührt aus ihrem Wesen: Sie werden eingesetzt, um Unkorreliertheiten zu erkennen. Unkorreliertheit bedeutet aber zwangsläufig das Nicht-stochastisch-unabhängig-sein. Damit ist aber die in der Statistik vielfach geforderte Annahme stochastischer Unabhängigkeit verletzt, was die ganze Sache komplizierter gestaltet. Zudem sinkt mit wachsender Zeitverschiebung die statistische Güte, weshalb zur Weiterverwendung in der Parameterschätzung nie die gesamte Folge herangezogen werden sollte. Letztlich ergeben sich aus der Forderung, Signale mit

den Eigenschaften einer identifizierten AKF über ein Filter mit endlichem Gedächtnis zu erzeugen, zusätzliche Restriktionen. Diese lassen sich aber elegant in lineare Matrixungleichungsrestriktionen umformen, sodass die entsprechenden Probleme effizient lösbar werden.

9.5 Eindeutigkeit und Identifizierbarkeit

Während in der Optimierungs- und der Approximationstheorie das Ermitteln eindeutiger Lösungen eine zentrale Stellung einnimmt, wird in der Schätztheorie und der experimentellen Modellbildung von der Identifizierbarkeit gesprochen. Letztere widmet sich der Frage, ob für ein parametrisiertes Modell mit einer bestimmten Methode (Gütekriterium) unter den vorliegenden Experimentationsbedingungen die „wahren“ Modellparameter bestimmt werden können. Im Kapitel 5 der Arbeit werden Fragen der Eindeutigkeit und der Identifizierbarkeit betrachtet. Hier werden einige Aussagen dazu kurz zusammengefasst.

1. Existenz eines Minimums prüfen
In der Modellbildung wird sich fast ausschließlich für die Minimierer und nicht für das Minimum interessiert. Deshalb sollten Formulierungen vermieden werden, die nur ein Infimum (größte untere Schranke) aufweisen, da diese keinen Wert im Definitionsbereich liefern, für den der kleinste Funktionswert angenommen wird. Zur Erinnerung: $f(x) = 1/x$ hat für $x > 0$ das Infimum Null, aber kein Minimum und damit keinen Minimierer. Überprüfen lässt sich eine Formulierung auf die Existenz eines Minimierers anhand der Bedingungen des Satzes 4.1. Letztlich sind dabei meistens vier Fälle kritisch: Erstens sind es Restriktionen, die durch strenge Ungleichungen beschrieben werden, die zudem scharf sind, in denen also eine Annäherung bis auf ein Epsilon gelingt. Zweitens sind es gleichungsrestriktionsbedingte Ausschlussmengen (z. B. durch Rangrestriktionen), die als Ungleichheitsrestriktionen geschrieben werden können und damit wie strenge Ungleichungen wirken. In beiden Fällen resultieren die Probleme aus der Tatsache, dass dann die betrachtete zulässige Menge der Variablen nicht abgeschlossen ist. Drittens sind es Richtungen im Unendlichen, in denen die Zielfunktion den Wert Null annimmt und wobei gleichzeitig eine asymptotische Annäherung an eine Restriktionsgrenze erfolgt. Viertens sind es Unbeschränktheiten im Inneren, die durch Singularitäten hervorgerufen werden.

Das Problem der Unbeschränktheiten nach unten tritt bei Formulierungen mit Normen nicht auf, da diese stets durch Null beschränkt sind. Das Problem mit den kritischen Richtungen nach Unendlich entfällt bei LS-Problemen, wenn spaltenreguläre Datenmatrizen vorliegen, denn dann strebt die Zielfunktion mit wachsender Norm des Parametervektors – also bezüglich aller Richtungen – nach Unendlich, weshalb im Unendlichen kein Infimum sein kann. Anschaulich heißt das, dass die Zielfunktion im Großen einem unendlichen Eierbecher ähnelt und nicht einem Trog. Das Problem mit der Annäherung an eine Ungleichung kann

vielfach auf eines nach Unendlich zurückgeführt werden, denn meist bedingt die Annäherung einzelner Parameter an die kritische Restriktion, dass andere Parameter nach Unendlich streben. Die angeführten drei Argumente sichern die Koerzitivität und über den Satz von Weierstraß die Existenz eines Minimums. Das Problem mit der Ungleichheit kann mit dem Argument magerer Mengen wegdiskutiert werden. Mit anderen Worten, das Problem hat dann eben generisch ein Minimum. Vorsicht ist in allen Fällen bei der Implementierung geboten. Eingebaute Rangtests sollten prüfen, ob die Voraussetzungen auch immer erfüllt sind, um unliebsame Programmabstürze zu vermeiden.

Eine wichtige Konsequenz aus dem Satz von Weierstraß ist, dass für Standard-LS-Probleme und Prokrustes-Probleme (LS-Probleme mit Restriktionen) Spaltenregularität der Datenmatrix die Existenz einer Lösung und gegebenenfalls sogar deren Eindeutigkeit sichert, vgl. hierzu die Sätze 4.10 bis 4.15.

Zu beachten ist aber, dass selbst bei einem Verletztsein der Spaltenregularität restringierte Probleme eine Lösung, ja sogar eine eindeutige Lösung haben können. Eine solche Situation liegt vor, wenn die Restriktionen die mit dem Spaltenrangverlust einhergehenden Koerzitivitätsprobleme oder Mehrdeutigkeiten kompensieren. Ist das der Fall, so führen beispielsweise bei linearen LS-Problemen mit linearen Restriktionen die Umformungen über die Eliminations- bzw. Nullraummethode auf neue, dann spaltenreguläre Probleme. Bei nichtlinearen Problemen gelten ähnliche Aussagen, wobei über die quadratische Zielfunktionsapproximation und die Taylor-linearisierten Restriktionen argumentiert werden kann.

2. Spaltenregularität der Datenmatrix sichern
Um der Forderung nach der Spaltenregularität gerecht zu werden, sind Überparametrisierungen zu vermeiden. Solche Überparametrisierungen entstehen bei Polkürzbarkeit (Beispiel 2.11) oder bei ungeschickter Formulierung in der MISO-Modellbildung (ähnlich der Polkürzbarkeit). Eine andere Quelle für den Verlust der Spaltenregularität ist eine ungenügende Systemanregung oder eine zu geringe Abtastzeit, sodass sich die Gleichungen kaum unterscheiden. Bei der Identifikation im geschlossenen Regelkreis kann auch der Regler einen Spaltenrangverlust verursachen. Das geschieht immer dann, wenn seine Ordnung kleiner als die der Strecke ist. Das Reglergesetz, das aus den Ausgangssignalen die neuen Stellsignale berechnet, macht dann die Eingangssignale linear abhängig. Es ist deshalb ratsam, für die Identifikation im geschlossenen Regelkreis abweichend vom üblichen PID-Regler mit einem Regler höherer Ordnung zu arbeiten.
Eine wichtige Konsequenz für den Praktiker sind Rückschlüsse auf die Daten, um Spaltenregularität zu garantieren. Datenkonstellationen, die einen Rangverlust bewirken, gilt es also von vornherein auszuschließen. Das Beschäftigen mit dieser Thematik kann somit auch neue Einsichten in das praktische Problem liefern. Probleme mit der Spaltenregularität entstehen beispielsweise in pathologischen Fällen der Identifikation von Kurven über algebraische

Gleichungen (z. B. Ellipsenfit), wenn etwa alle Punkte auf einer Gerade liegen. Solche Fälle sind für Offline-Anwendungen eher weniger tragisch, da sie schnell erkannt werden und eine Fehlermeldung oder ein Programmabsturz keine Folgen haben. Für Online- und Prozessbetriebsanwendungen sollte die Regularität der Datenmatrix also parallel immer mit überwacht werden.

Trotz des Einhaltens der Spaltenregularität ist bei der Identifikation von Polynomen auf die richtige Wahl des Polynomgrads zu achten. Angenommen ein Polynom zweiten Grads wird durch einen Ansatz dritter Ordnung bestimmt. Die zugeordnete Datenmatrix ist dann regulär, aber der führende Koeffizient wird zu Null oder nahe Null bestimmt. Das ist ja richtig so, aber das Problem ist dann, dass Wurzelberechnungen praktisch unbrauchbar sind (Division durch nahe Null, hohe Empfindlichkeit der Wurzeln).

3. Strenge Konvexität anstreben
Wie in der Arbeit bereits mehrfach angesprochen, hängt die Schwierigkeit eines Optimierungsproblems nicht davon ab, ob es linear oder nichtlinear ist, sondern ob es konvex oder nichtkonvex ist. Das liegt an den Eigenschaften konvexer Probleme:
- Subniveaumengen sind konvex.
- Menge der Minimierer ist konvex, wenn sie nicht leer ist.
- Lokale Minimierer sind gleichzeitig globale Minimierer.
- Festhängen von Suchverfahren in einem nichtglobalen Minimum entfällt.
- Analytische Berechnung des Minimierers erfordert nur Bedingung zur ersten Ableitung.
- Konvexität lässt sich einfach nachweisen (Ungleichungen, Monotonie).

All diese Aussagen lassen sich natürlich präzise in Sätze fassen, s. hierzu Abschnitt 4.2.1 und Abschnitt 6.6.4 mit Anmerkung 6.14.

Noch besser als Konvexität ist strenge Konvexität, da dann nur ein einziger Minimierer existiert. Auch hierfür werden in Abschnitt 4.2.1 Sätze bereitgestellt. Zudem werden in diesen Sätzen Abschwächungen hinsichtlich Quasi- und Pseudokonvexität betrachtet. Das ist auch wichtig, da Normen – auch wenn sie streng konvex genannt werden – keine streng konvexen Funktionen sind, wohl aber streng quasikonvexe Funktionen. Für den Ingenieur ergibt sich daraus, dass er aus Sicht der Eindeutigkeit Probleme in der euklidischen Norm oder der Frobenius-Norm formulieren sollte und nicht die Absolutsummennorm, die Maximalbetragsnorm oder die Spektralnorm wählen sollte.

Obwohl die Absolutsummennorm und die Maximalbetragsnorm keine streng konvexen Normen sind, kann ihr Einsatz dennoch angezeigt sein, da etwa die Minimierung der Summe der Abstände oder die Minimierung des Maximalabstandes dem technischen Problem besser angepasst sein kann. Auch bedeutet eine nicht streng konvexe Norm nicht zwingend Mehrdeutigkeiten, s. Abschnitt 4.1.2.

4. In Hilbert-Räumen arbeiten

Hilbert-Räume, d. h. Vektorräume, in denen ein Skalarprodukt erklärt ist, bieten viele Vorteile. Durch das Skalarprodukt wird eine streng konvexe Norm induziert, was der Eindeutigkeit dient, und über das Skalarprodukt lassen sich Winkel definieren. Von denen ist der rechte Winkel, der mit einem Skalarprodukt von Null korrespondiert, der wichtigste. Er spielt in den sogenannten Projektionssätzen, die den nächstgelegenen Punkt zu einer Menge spezifizieren, eine entscheidende Rolle. Der eigentliche Vorteil dieser Sätze, vgl. Satz 4.9, sind die notwendigen und hinreichenden Bedingungen, die ohne Ableitungen auskommen. Zudem gelten sie mit marginalen Abschwächungen auch in unendlichdimensionalen Hilbert-Räumen. Solche Räume treten beispielsweise bei der Funktionsapproximation im Sinne der L_2-Norm auf, werden aber in dieser Arbeit nicht weiter betrachtet.

Die gebräuchlichste Art für den Ingenieur, um eine Optimierung in einem Hilbert-Raum auszuführen, ist die Verwendung der euklidischen oder einer elliptischen Norm für vektorielle Variable und die Frobenius-Norm für Matrixvariable. Diese Normen implizieren zugehörige Skalarprodukte. Auf diesem Weg können Matrizenmengen, die lineare Unterräume formen (symmetrische, schiefsymmetrische, Toeplitz–Matrizen), in gleicher Weise behandelt werden.

5. Eindeutigkeit durch zusätzliche Restriktionen erzwingen

Es gibt eine Reihe von Problemen, die per se eine Menge äquivalenter Minimierer zulassen. Hierzu zählen Verhältnisprobleme, wo sich die Mehrdeutigkeit der Minimierer bezüglich des Minimums herauskürzt. Als Beispiel seien Datenmodelle genannt, bei denen die gesuchten Vektoren zum Beispiel eine optimale Trennung der Daten in Teilmengen gewährleisten sollen oder bei denen der lineare Unterraum maximaler Streuung bestimmt werden soll. Dazu zählen aber auch parameterlineare Ansätze, die auf impliziten Gleichungen fußen (z. B. Ellipsenfit). Hierbei bewirkt ein Dividieren der Gleichung wegen der Null auf der einen Seite, dass mehrere Parametersätze die gleiche Funktion beschreiben. In beiden vorgenannten Fällen kann die Eindeutigkeit durch zusätzliche Restriktionen, z. B. durch Normieren auf eine bestimmte Länge, meist Länge Eins, oder durch Fixieren eines Parameters auf Eins erreicht werden. Welche der beiden bevorzugten Techniken verwendet wird, hängt entscheidend vom resultierenden Lösungsaufwand ab. Welcher Parameter beim Einsfixieren allerdings auf Eins gesetzt wird, hängt hingegen von der Modellverwendung und gegebenenfalls von statistischen Aspekten ab, siehe hierzu auch Abschnitt 4.4.4.

Ein ganz anderer Weg, Eindeutigkeit zu erreichen, besteht darin, zunächst alle potenziellen Minimierer zu bestimmen. Insbesondere wenn sich diese Menge einfach beschreiben lässt, was bei konvexen Lösungsmengen häufig der Fall ist, kann in einem zweiten Schritt aus dieser Menge eine prädestinierte Lösung ausgewählt werden. Eine Möglichkeit ist, das normkleinste Element der Lösungsmenge zu nehmen, was die sog. Pseudonormallösung bei unterbestimmten LS-Problemen begründet. Einen alternativen, kaum bekannten Weg basierend auf Halbordnungsrelationen zeigt Beispiel 4.17.

6. Über Symmetrien nachdenken
Die Formulierung konvexer Probleme ist zwar wünschenswert, allerdings für eine Reihe von Problemen unmöglich. Hierzu zählen geometrischer Probleme, die allein aus Symmetriegründen mehrere isolierte Lösungen haben müssen. Dieser Nachteil ist aber zugleich ein Vorteil, denn er sagt, wo und wie viele solcher Minima zu erwarten sind. Wenn nämlich aus diversen Überlegungen klar ist, dass im ungestörten Fall mehrere Minimierer mit dem gleichen Minimum auftreten, dann werden diese Minimierer unter Störung nicht plötzlich verschwinden. Einer der Minimierer wird zum globalen Minimierer, der Rest zu lokalen Minimierern. Diese Information kann bei Suchverfahren zur Festlegung neuer Startwerte genutzt werden.

Mitunter sind diese Symmetrien nicht immer offensichtlich. Im Beispiel 4.14 basieren sie auf der Approximationsfähigkeit negativer Zählerkonstanten für Totzeiten. Das Beispiel liefert zudem einen Beleg dafür, dass Sprungsignale nur bedingt zur Identifikation geeignet sind. Diese Aussage wird auch durch Beispiel 4.15 gestützt.

7. Bei Offline-Identifikation Ausgangsfehlerverfahren bevorzugen
Obwohl aus Sicht der Konvexität alles für die Verwendung von Gleichungsfehleransätzen für die Identifikation dynamischer Modelle spricht, sollten insbesondere für niedrigparametrische lineare oder nichtlineare dynamische Modelle bei der Offline-Modellbildung bevorzugt ausgangsfehlerorientierte Gütekriterien verwendet werden. Dem Nachteil einer nichtkonvexen Problemformulierung, der damit einhergehenden Existenz lokaler Minima und des erforderlichen Einsatzes iterativer Suchverfahren stehen als Vorteil bessere statistische Eigenschaften der Schätzung und die selbststabilisierenden Eigenschaften der Suchverfahren gegenüber. Letztere können dazu führen, dass beispielsweise auf Stabilitätsrestriktionen an das Modell verzichtet werden kann. Das Suchverfahren verhindert gewissermaßen, dass das Modell instabil wird. Gleichungsfehlerorientierte Zugänge weisen diese Eigenschaft nicht auf.

Bei der Verwendung von Ausgangsfehlerverfahren für dynamische Modelle (zeitkontinuierliche Beschreibung wählen) werden die Anfangswerte für die Simulation zur Zielfunktionswertberechnung benötigt. Liegt keine Erregung aus dem Arbeitspunkt vor (z. B. Eventsituation bei passiver Experimentation), müssen die Anfangswerte als zusätzliche Parameter bestimmt werden. Vor dem naheliegenden Ansatz, den ersten Ausgangssignalwert und seine Ableitungen zu nehmen, sei gewarnt. Das Ausgangsverhalten kann nämlich sehr empfindlich bezüglich der Anfangswerte sein [247]. Als zweckmäßig erweist sich zudem, nicht vom Differenzialgleichungsmodell und dessen Anfangswerten auszugehen, da dann auch die Eingangssignalanfangswerte zu berücksichtigen sind. Besser sind geeignete Zustandsraumdarstellungen mit dem Anfangszustand als zusätzlichen Parametervektor.

8. Irrelevante lokale Minima durch Restriktionen ausschließen
Einige Probleme haben Minima, die praktisch bedeutungslos sind. Ein Beispiel sind Ausgangsfehlerzugänge, bei denen die Anfangswerte mit geschätzt werden müssen. Ein lokaler Minimierer ist dann als Anfangswert der Ausgangssignalmittelwert und ein Modell dessen Parameter Null oder Unendlich sind. Das Minimum nutzt die Optimalität des Mittelwertes bei der Least-Squares-Methode und ignoriert die Modelldynamik. Um diese uninteressante Lösung auszuschließen und vor allem um zu verhindern, dass die Suche gegen diesen Minimierer strebt, empfehlen sich sogenannte optimierungstechnische Restriktionen. Das sind Restriktionen, die in allen interessierenden Minima inaktiv sind. Bei Systemen mit Totzeit empfiehlt sich eine obere Schranke, um bei periodischer Anregung die Zahl der Subextrema zu reduzieren und um bei aperiodischen Vorgängen ähnliche Effekte wie bei unendlich großer Zeitkonstante zu vermeiden.

9.6 Tipps zum Problemlösen

In den Kapiteln 5 bis 8 werden Techniken vorgestellt, mit denen sich Parameterschätzprobleme vereinfachen und lösen lassen. Die Schwierigkeit für Neueinsteiger im Gegensatz zum Routinier besteht darin, in einem Problem das Potenzial für eine Vereinfachung oder gar einen zielgerichteten Lösungszugang zu erkennen. Die hohen mathematischen Anforderungen kommen erschwerend hinzu. In diesem Abschnitt werden deshalb Thesen formuliert, mit deren Hilfe das Vereinfachungspotenzial und erfolgversprechende Lösungszugänge erkannt werden können. Mitunter ergeben sich allein aus der Zuordnung des Problems und aus den Lösungsansätzen neue Suchbegriffe für die Recherche. Mit etwas Glück lässt sich so eine Lösung oder ein Lösungsverfahren finden, was dann weniger eigenes Rechnen erfordert. Merke: Die meisten eigenen Probleme hat so ähnlich fast immer schon jemand irgendwo gelöst.

1. Stringente mathematische Formulierung anstreben
Obwohl das Gegeben-Gesucht-Skizze-Prinzip durch die Ausbildung dem Ingenieur in Fleisch und Blut eingegangen sein sollte, ist immer wieder zu beobachten, dass Probleme nur halbherzig formuliert werden. Mögliche Fehler sind:
- unpräzises Formulieren der Zielfunktion
- vergessene versteckte Restriktionen (Rückrechnungs-, Definitionsbereichsbedingungen)
- fehlende Arbeitspunkte, Anfangswerte bei der Identifikation dynamischer Modelle
- fehlende Modellgültigkeitsbereiche
- fehlende Angaben zu den Daten (Anzahl, Qualität, interne Abhängigkeiten)

Zur stringenten mathematischen Formulierung gehört auch, dass das Problem in eine mathematische Notation umcodiert wird. Der Regelungstechniker macht das permanent, indem er u, y und x für Eingänge, Ausgänge und Zustände verwendet. In der Optimierung wird

mit x üblicherweise die Unbekannte benannt, während die Daten in Matrizen A, B oder Vektoren c, d zusammengefasst werden. Andere Fachdisziplinen belassen für geometrische Probleme die Bezeichner x, y und fassen die unbekannten Parameter beispielsweise in einem Vektor θ zusammen. Auf jeden Fall ist es zweckmäßig, wenn die physikalischen Bezeichner und Einheiten im mathematisch formulierten Problem nicht mehr auftauchen.

2. Formulierungen in Matrixnotation anstreben
Der entscheidende Vorteil von Matrix- und Vektornotation ist deren Übersichtlichkeit. In Matlab kommt hinzu, dass Matrix-Vektoroperationen deutlich effektiver ausgeführt werden als die äquivalenten Formulierungen als Summen und deren Implementierung über Schleifen. Es sollte also geprüft werden, ob sich Summen durch Skalarprodukte darstellen lassen und ob sich durch Umordnen (Ausklammern), die Parameter in einem Vektor anordnen lassen.

3. Zweckmäßige Normen wählen
Tendenziell sollten streng konvexe Normen verwendet werden. Zudem sind bei additiven Termen in der Zielfunktion Normen unterschiedlichen Typus zu vermeiden, also entweder durchweg Hilbert-Raum-Normen (euklidische, elliptische, Frobenius-Norm) oder durchweg polyedrischen Normen (Manhattan-, Chebyshev-, Spaltensummen-, Zeilensummen-Norm) oder eben die Spektralnorm. Weiterhin ist er ratsam, wenn mehrere additive Terme in der Zielfunktion zu berücksichtigen sind, bei Hilbert-Raum-Normen die Summe der Normquadrate, bei polyedrischen Normen die Summe der Normen zu verwenden. Prinzipiell kann zwar das Summe-von-Normen-Problem (Weber-Problem) durch Umformung in semidefinite Probleme elegant gelöst werden, aber nicht so effizient (interessant bei hochdimensionalen Problemen) und auch nicht rekursiv wie ein Summe-von-Normquadraten-Problem.
Für bestimmte Eigenwert-Distanzprobleme empfiehlt sich die Spektralnorm gegenüber der Frobenius-Norm. Hierbei heben sich die komplizierten Zusammenhänge zwischen Matrixelementen und Eigenwerten und der komplizierte Zusammenhang zwischen Matrixelementen und maximalem Singulärwert in gewissen Grenzen gegenseitig auf.

4. Restriktionen analysieren
Ein wesentlicher Schritt bei der Problemlösung besteht darin, die „guten“ von den „schlechten“ Restriktionen zu unterscheiden. Lineare Gleichungs- und Ungleichungsrestriktionen zählen zu den guten Restriktionen, da sie einfach behandelbar sind. Bei den Ungleichungsrestriktionen sind jene gut, durch die für die Parameter konvexe Mengen formuliert werden. Hierzu zählen Kegelrestriktionen oder quadratische Restriktionen mit positiv definiter Matrix. Bei den Gleichungsfehlerrestriktionen sind jene gut, die kompakte Mengen beschreiben, da dann ein Extremum garantiert ist. Dabei muss durch die Restriktion nicht notwendigerweise eine zusammenhängende Menge beschrieben werden. Bei nicht zusammenhängenden Mengen, wie den orthogonalen Matrizen $X^T X = I_n$ – die Determinante von X kann +1 oder -1 sein – ist dann eben einfach über beide Teilmengen zu optimieren. Weniger proble-

matisch sind auch Gleichungsrestriktionen, bei denen eine konvexe Relaxation durch eine Ungleichungsrestriktion möglich ist (Kreis $||x||_2 = 1$ wird zur Kreisscheibe $||x||_2 \leq 1$), s. auch Abschnitt 8.1.2. Als schlechte Restriktionen seien Ganzzahligkeitsrestriktionen, oder-verknüpfte Restriktionen, aber auch ordnungs- und rangfixierende Restriktionen genannt. So ist die Forderung, wonach ein Polynom den Grad 3 haben soll, weit schwieriger als die Forderung, wonach es einen Grad ≤ 3 haben soll. Die Polynome bis zum Grad 3 bilden nämlich einen Vektorraum, die vom Grad 3 dagegen nicht. Ähnlich verhält es sich mit Rang- und Max-Rang-Restriktionen.

Neben der Einteilung der Restriktionen gilt es, die Menge der Restriktionen nach redundanten und singulären Restriktionen zu durchforsten, s. hierzu die Hinweise in den Abschnitten 5.1.3 und 5.1.4. Vergleichbar mit den singulären Ungleichungsrestriktionen, die zu Gleichungsrestriktionen entarten, können mehrere Gleichungsrestriktionen bei der Schnittmengenbildung auf isolierte Punkte oder Kurven führen. Diese gilt es zu erkennen, da sie einerseits beim Optimieren Schwierigkeiten bereiten, andererseits aber separat behandelt werden können bzw. müssen.

Bei der Analyse der Restriktionen empfiehlt es sich, jene zu kennzeichnen, für die eine Lösung in geschlossener Form oder eine einfache und schnelle iterative Lösung gelingt. Besitzt das Problem durchweg solche Restriktionen, dann kann es – da die Restriktionen simultan gelten müssen – dennoch schwierig zu lösen sein. Allerdings eröffnen die einfachen Teillösungen, Algorithmen auf der Basis zyklischer Projektionen zu konstruieren, vgl. Abschnitt 8.2.1.

Sind Zielfunktion und Restriktion von gleicher oder ähnlicher Struktur kann überlegt werden, ob statt der klassischen Formulierung auch eine Verhältnisformulierung der Anwendung gerecht wird. Die Restriktion ist dann weg und steht im Nenner der Zielfunktion. So sind die Bestimmung der maximalen Verformung einer Einheitssphäre und die Bestimmung des maximalen Streckungsfaktors einer beliebigen Sphäre äquivalente Aufgaben. Auch ist die Abstandsminimierung in logarithmischer Skalierung zu einer Verhältnisminimierung äquivalent, s. Beispiel 7.10. Verhältnisformulierungen finden unter anderem in der Datenanalyse (z. B. Diskriminanz-, Hauptkomponentenanalyse) Anwendung. Nachteilig für Neueinsteiger sind die ungewohnten Lösungszugänge, die größtenteils auf die vertrauten Ableitungszugänge verzichten und stattdessen Ungleichungstechniken (Beispiel 7.1, Abschnitt 5.3) oder SDP-Umformungen (Abschnitt 7.8) nutzen.

5. Suche nach freien und separablen Variablen

Freie Variablen, also jene, die keinen Restriktionen unterliegen, können vielfach zur Problemreduktion genutzt werden. Hierzu wird bezüglich der freien Variablen optimiert, um die erhaltene Lösung danach in die Zielfunktion einzusetzen, siehe Abschnitt 5.2.7. Im Problem

$$\left\| [A, 1_n] + \gamma [B, 0_n] \begin{bmatrix} \theta_{1:3} \\ \theta_4 \end{bmatrix} \right\|_2 \stackrel{!}{=} \text{Min}; \qquad \theta_{1:3}^T B \theta_{1:3} = 1 \tag{9.1}$$

sind θ_4 und γ freie Variablen. Da θ_4 zudem nicht mit γ verknüpft ist, lässt es sich leicht eliminieren. Eine Elimination von γ ist zwar theoretisch möglich, verkompliziert aber die Zielfunktion nur unnötig und bringt somit keinen Gewinn.

Das Verhältnis der Anzahl der freien Variablen zur Anzahl der Restriktionen gibt ein Indiz dafür, ob es sich lohnt, über eine Problemumformung basierend auf dem Dualitätsprinzip nachzudenken, vgl. Abschnitt 8.1.4 und 8.1.5. Durch Anwenden dieses Prinzips tauscht sich die Rolle der Variablen und Restriktionen, d. h., ein Problem mit vielen Variablen und wenigen Restriktionen wird zu einem mit wenigen Variablen und vielen Restriktionen. Unabhängig von den Variablen-Restriktionen-Beziehung erzeugt das Prinzip ein im Optimierungssinn duales Problem. So wird beispielsweise statt des kleinsten Abstands zwischen zwei Mengen der größte Abstand paralleler Stütztangenten gesucht, was natürlich äquivalent ist. Die Anwendung des Dualitätsprinzips kann sehr wirkungsvoll sein, ist aber für Neueinsteiger – abgesehen von gelösten Problemen – schwierig.

Vereinfachen lassen sich auch Probleme, in denen die Restriktionsmengen separabel sind, da dann mit Dekompositionstechniken, also dem Zerlegen in Teilprobleme, gearbeitet werden kann, siehe Abschnitt 7.3. Besonders leistungsstark sind Dekompositionszugänge, wenn neben den Restriktionsmengen auch die Zielfunktionen separabel in den Variablen sind. Weniger effektiv kann die Anwendung der Regel $\min\limits_{x_1\in\mathcal{X}_1,x_2\in\mathcal{X}_2} f(x_1,x_2) = \min\limits_{x_2\in\mathcal{X}_2}\{\min\limits_{x_1\in\mathcal{X}_1} f(x_1,x_2)\}$ sein. Sie erfordert eine Parametrisierung der Minimierer $x_{1,\min}$ durch x_2, die dann eingesetzt in $f(x_1,x_2)$ auf die neue Zielfunktion $\tilde{f}(x_{1,\min}(x_2),x_2)$ führt. Wie im vorherigen Fall von θ_4 und γ können dadurch einfacher oder schwieriger handhabbare Probleme entstehen.

Eine besondere Situation liegt vor, wenn die Variablen separabel in dem Sinn sind, dass eine Optimierung über der einen Variable bei festgehaltener anderer Variable einfach möglich ist. Das ist bei Quadratmittelproblemen und multilinearer Verknüpfung der Variablen der Fall, wie (9.1) bezüglich γ und θ. In einer solchen Situation bieten sich die Verfahren aus den Abschnitten 5.2.8, 8.2.2, 8.2.3 und 8.2.3 an.

6. Einsatz von Pseudoparametern prüfen
Eine beliebte Technik zur Problemvereinfachung besteht darin, nichtlineare Parameterverknüpfungen durch einen neuen Parameter zu ersetzen. Dabei sind zwei Fälle zu unterscheiden: Parameteranzahl bleibt gleich, s. Beispiel 7.4, oder sie erhöht sich. Bei einer Erhöhung sind zusätzliche Restriktionen zu berücksichtigen, vgl. Beispiel in Tabelle 7.2. Ein weiteres Beispiel ist das quadratische Eigenwertproblem $(\lambda^2 A_2 + \lambda A_1 + A_0)x = 0_n$, das mit dem Pseudoparameter $y = \lambda x$ (zusätzliche Restriktion) in $\lambda A_2 y + A_1 y + A_0 x = 0_n$ umgeschrieben werden kann und in Blocknotation zu einem verallgemeinerten Eigenwertproblem wird

$$\begin{bmatrix} 0_{n\times n} & I_n \\ -A_0 & -A_1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} I_n & 0_{n\times n} \\ 0_{n\times n} & A_2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{9.2}$$

7. Eigenwertprobleme erkennen
Quadratische Probleme mit quadratischen Restriktionen führen beim Einsatz der Lagrange-Multiplikator-Methode häufig auf Eigenwertprobleme (EWP). So liefert die Optimalitätsbedingung für $x^T Ax \stackrel{!}{=} \text{Min}$ unter $x^T Bx = 1$ ein verallgemeinertes EWP $(A-\lambda B)x = 0_n$ mit λ als Lagrange-Multiplikator oder im Sinne der Lösung als Eigenwert, s. [248] für Anwendungen und Lösungszugänge. Tabelle 9.3 benennt und charakterisiert die Eigenwertprobleme, womit das Erkennen und Zuordnen in künftigen Anwendungen erleichtert wird.

Art des Eigenwertproblems	Formulierung (für alle $x \neq 0_n$)
gewöhnlich [300]	$Ax = \lambda x$
verallgemeinert [300]	$Ax = \lambda Bx$ mit B nicht zwingend regulär
singulär [326], [179]	$(A - \lambda B)x = 0_m$; $A, B \in \mathbb{C}^{m\times n}, m \neq n$ oder $\det(A - \lambda B) \equiv 0$ bei $A, B \in \mathbb{C}^{n\times n}$
inhomogen [445]	$(A - \lambda I_n)x = b$
quadratisch [599]	$(\lambda^2 A_2 + \lambda A_1 + A_0)x = 0_n$
polynomial [598]	$\left(\sum\limits_{i=0}^{k} \lambda^i A_i\right)x = 0_n$
multiparametrisch [35], [466]	$A_1 x_1 = \lambda_1 B_{11} x_1 + \ldots + \lambda_k B_{1k} x_1; \quad x_1 \in \mathbb{R}^{n_1}$ $\vdots \qquad\qquad \vdots$ $A_k x_k = \lambda_1 B_{k1} x_k + \ldots + \lambda_k B_{kk} x_k; \quad x_k \in \mathbb{R}^{n_k}$
allgemein-multiparametrisch [102]	$\left(A - \sum\limits_{j=1}^{k} \lambda_j B_j\right)x = 0_n$
multivariat [136], [589], [590], [646][5]	$\begin{bmatrix} A_{11} & \ldots & A_{1q} \\ \vdots & & \vdots \\ A_{p1} & \ldots & A_{pq} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_q \end{bmatrix} = \begin{bmatrix} \lambda_1 x_1 \\ \vdots \\ \lambda_q x_q \end{bmatrix}$
nichtlinear[6]	$F(\lambda; A_1, \ldots, A_k)x = 0_n$
restringiert [561] Spezialfall: Pareto-EWP	$\mathcal{C} \ni x \perp Q(\lambda)x \in \mathcal{C}^{\oplus}$ mit $\mathcal{C}$ konvexer Kegel $x \geq 0_n \perp (A-\lambda I_n)x \geq 0_n$; $\quad Q(\lambda)$ Polynommatrix
Singulärwertproblem [300]	$Ax = \sigma y$ $A^T y = \sigma x$

Tabelle 9.3: Einteilung von Eigenwertproblemen

[5] Ein spezielles bivariates Problem, das bei der Identifikation endlicher Impulsantworten entsteht, wird in [Liu, Z.-Y.; Qian, J.; Xu, S.-F.: On the double eigenvalue problem] beschrieben.
`http://www.math.pku.edu.cn:8000/var/preprint/7229.pdf`

[6] Nichtlineare EWP entstehen u. a. bei der Untersuchung von Zustandsraumsystemen mit Totzeiten im Zustandsvektor. Die Determinante formuliert dann ein charakteristisches Quasipolynom [258].

Hinweise: Quadratische und polynomiale EWP lassen sich in verallgemeinerte EWP überführen (Linearisierung durch Dimensionserhöhung). Verallgemeinerte EWP können als gewöhnliche EWP formuliert werden, wenn entweder A oder B regulär ist. Inhomogene EWP können in quadratische EWP überführt werden [641]. Beim allgemein-multiparametrischen EWP ist A, anders als beim gewöhnlichen multiparametrischen EWP, keine Blockdiagonalmatrix, vgl. $A := \mathrm{Diag}(A_i)$, $B_j := \mathrm{Diag}(B_{ij})$, $x^T := [x_1^T, \ldots, x_k^T]$. Reguläre zweiparametrische (lineare und quadratische) EWP können mittels Kronecker-Produkt in verallgemeinerte EWP umgeformt werden [294].
Beachte: Mitunter fehlt für ein multiparametrisches EWP eine Gleichung. Diese kann aber aus Doppeleigenwertbedingungen oder aus hinreichenden Optimalitätsbedingungen folgen, s. [319][7].
Bei verallgemeinerten, quadratischen und nichtlinearen EWP können neben den endlichen Eigenwerten auch Eigenwerte im Unendlichen auftreten. Beim singulären EWP (Generisch sind quadratische Büschel $A-\lambda B$ regulär.) kommen noch die sog. unbestimmten Eigenwerte hinzu. Es ist auch möglich, dass singuläre EWP keine Eigenwerte besitzen, vgl. $A := [1,\ 2]^T$ und $B := [1,\ 3]^T$.

9. Anwendungsproblem als gelöstes Optimierungsproblem formulieren
Mit dieser These ist nicht gemeint „Wir haben eine Lösung und suchen ein Problem dazu.“, sondern die Erfahrung, dass das Wissen um gelöste Optimierungsprobleme ungeschickte Formulierungen vermeiden hilft. Ein Studium der Tabellen in Abschnitt A.4 sollte gerade für Neueinsteiger ein erster Schritt sein, um sich einen Überblick zu verschaffen.

Die Abschnitte dieses Kapitels stellen Ergebnisse, Empfehlungen und Tipps bereit, um bei der Identifikation von Modellen an sich und durch das Einbeziehen von Vorwissen erfolgreich zu sein. In den Thesen und in der Arbeit finden sich Hinweise dazu, wann strukturelle Umformungen und Parameterrestriktionen vorteilhaft sind, was durch Kompromisskriterien und Relaxation erreicht werden kann und dass das suboptimale zweistufige Vorgehen aus Weglassen und nachträglichem Sichern der Restriktion eine Lösungsoption ist. Dennoch bleibt die Identifikation guter Modelle eine große Herausforderung. Wie in anderen Wissenschaftsgebieten gilt auch hier: Erst das beständige selbständige Lösen von Problemen wird die eigene Erfahrung verbessern und die Intuition dafür stärken, was, wie und wann am besten funktioniert.

[7] Nichtiterative Lösung in [Muhič, A.; Plestenjak, B.: A method for computing all values λ such that $A+\lambda B$ has a multiple eigenvalue. Preprint (2013)]
`http://www-lp.fmf.uni-lj.si/plestenjak/Papers/PreprintDouble.pdf`.

Kapitel 10

Zusammenfassung

Die vorliegende Arbeit stellt Techniken, Ansätze und Übersichten bereit, um Vorwissen an Systeme und Funktionen bei der Modellierung einzubeziehen. Gegenstand der Arbeit sind die mathematische Beschreibung von Vorwissen in Form parametrischer und struktureller Restriktionen an das Modell und die Darstellung von Methoden zur Vereinfachung restringierter Optimierungsprobleme.

Mit dieser Arbeit wird erstmalig ein breiter Überblick zum Aufstellen und zur mathematischen Formulierung von Restriktionen gegeben. Für den Ingenieur sind dabei insbesondere die systemtheoretischen Restriktionen an dynamische Modelle (Kapitel 2) und an Funktionen (Kapitel 3) von Bedeutung. Ferner sind fast immer auch Restriktionen zu stellen, die eine Überparametrierung verhindern, Identifizierbarkeit sichern oder ein Problem eindeutig machen (Kapitel 4). Insofern liefern die Kapitel 2 bis 4 das Rüstzeug für eine wichtige Phase bei der Modellbildung, nämlich die Aggregation von Vorwissen und die mathematische Formulierung von Vorwissen. Diese Phase wird ausgeführt, wenn die Systemgrenzen und der Modelltyp (parametrisch bzw. nichtparametrisch, Zustandsraum bzw. Übertragungsfunktion usw.) festliegen. Eine Freiheit, die dann für die Modellbildung noch besteht, betrifft die Entscheidung, ob das Vorwissen über parametrische oder strukturelle Restriktionen umgesetzt werden soll. Hierfür gibt es zwar keine Regel, da das vom umzusetzenden Vorwissen abhängt, aber für viele wichtige Eigenschaften finden sich Lösungsansätze in dieser Arbeit. Die Ansätze können im Normalfall auch auf andere, nicht angeführte Eigenschaften übertragen werden. Ist das Vorwissen vage, wie etwa im Fall der Forderung nach einem möglichst glatten Verlauf einer Funktion, dann kann eine einzelne Restriktion viel zu streng oder auch viel zu schwach sein. In einem solchen Fall bleibt die Möglichkeit, die Restriktion durch einen zusätzlichen Parameter in ihrer Stärke zu regulieren und das Problem für unterschiedlich starke Restriktionen zu lösen. Die einzelnen Lösungen sind dann a posteriori zu beurteilen. Alternativ kann ein Kompromisszugang gewählt werden, bei welchem die Forderung über einen zusätzlichen

Term in die Zielfunktion aufgenommen wird. Durch einen Gewichtungsparameter lässt sich der Einfluss der Restriktion variieren.

Da sich die Gliederung der Arbeit in den Kapiteln 2 und 3 nach dem jeweiligen Vorwissen richtet, ermöglicht sie dem Leser auch ohne Kenntnis der kompletten Arbeit, Lösungen für seine umzusetzenden Systemeigenschaften zu erarbeiten. Zahlreiche Beispiele, vor allem aber auch Gegenbeispiele zu naheliegenden, jedoch falschen Schlussfolgerungen unterstützen ihn dabei. Insofern sind die beiden Kapitel auch losgelöst von der Modellbildung für Leser mit stärker systemtheoretischem Fokus von Interesse. Ergänzt werden die eher der Plausibilisierung dienenden Beispiele durch eigene praktische Modellierungsanwendungen, die es im Rahmen von Regelungsaufgaben für Mehrzonenöfen, verfahrenstechnische Anlagen und autonome Fahrzeuge zu lösen galt.

Während das Aggregieren des Vorwissens zunächst unabhängig von der Wahl der Zielfunktion ist, muss beim Formulieren des Vorwissens die Zielfunktion im Blick behalten werden. Aber nicht nur sie, auch die verfügbare Software, die Anwendungsplattform (Mikrocontroller, Stand-alone-Rechner) und andere Aspekte haben Einfluss auf die Formulierung. Zielfunktion und Restriktionen bestimmen nämlich gemeinsam den algorithmischen Aufwand und entscheiden auch darüber, ob mehrere lokale Minima auftreten und ob eine eindeutige globale Lösung existiert. Allzu oft wird dieses Zusammenspiel vernachlässigt, weil gehofft wird, dass leistungsfähige Rechner und Standardoptimierungsalgorithmen das entstehende Problem schon angemessen schnell lösen. Für Mikrocontrolleranwendungen und auch für mittelgroße Probleme (ab zehn Parameter, manchmal auch schon ab fünf Parameter) gilt das aber nicht zwingend. Die Algorithmen finden zwar relativ schnell einen Minimierer, ob dieser aber der globale ist, lassen sie offen. In der Arbeit wird deshalb der konvexen oder besser noch streng konvexen Formulierung bzw. Umformulierung von Problemen der Vorrang gegeben (u. a. in den Abschnitten 4.1.2, 4.2.1, 7.8). Erreichbar ist das unter anderem durch Normen über parameterlineare Funktionen und bei linearen oder semidefiniten Restriktionen.

Als Konsequenz aus der Forderung nach Eindeutigkeit und nach schnellen und numerisch stabilen Algorithmen werden zumeist Quadratmittelansätze mit linearen Gleichungs- und Ungleichungsrestriktionen präferiert. Doch hier sollte genau geprüft werden, ob die quadratische Zielfunktion wirklich problemadäquat ist. Bei Approximationsproblemen ist beispielsweise die Absolutfehlersumme eine alternative Zielfunktion, da sie nicht nur anschaulicher ist, sondern auch robustere Ergebnisse impliziert. Aus diesem Grund heraus werden sowohl in einigen Abschnitten als auch im Anhang wiederholt nichtquadratische Zielfunktionen und deren Umformung betrachtet, wobei die Darstellungen selbst kurz gehalten werden und durch Literatur für vertiefende Betrachtungen ergänzt werden.

Obwohl sich die angesprochenen Konflikte aus zu langsamer Konvergenz, Multimodalität und Nichteindeutigkeit durch konvexe Formulierungen weitgehend auflösen lassen, muss für

den Fall, dass keine konvexe Formulierung gelingt, nach Alternativen gesucht werden. Das ist leider bei vielen Restriktionen für LTI-Systeme so, da Wurzeln, Pole, Nullstellen und Eigenwerte in komplizierter Weise von den Polynomkoeffizienten bzw. Matrixelementen abhängen. Daher gibt es auch keinen universellen Zugang, sondern viele problemspezifische Lösungen, die direkt im jeweiligen Abschnitt diskutiert werden. Dennoch lassen sich einige wiederkehrende Lösungsprinzipien ableiten, wie Relaxation, Parametertransformation, Dekomposition, Matrixapproximation und Generizität. Diese und weitere Prinzipien werden losgelöst von der Anwendung in den Kapiteln 5 bis 8 näher untersucht. Jedes dieser Kapitel beschreibt eine Methodenklasse, die die Probleme auf unterschiedliche Weise vereinfacht und Möglichkeiten eröffnet, einen Teil der Konflikte zu bewältigen. Konkret werden die Reduktionsmethoden (Kapitel 5), die Erweiterungsmethoden (Kapitel 6), die Transformationsmethoden (Kapitel 7) und die Modifikationsmethoden (Kapitel 8) vorgestellt. Auch für diesen Teil der Arbeit wurde eine weitgehende separate Lesbarkeit der einzelnen Abschnitte angestrebt. Hierfür erfolgt eine kurze Darstellung der theoretischen Grundlagen, die durch Beispiele untersetzt und Anmerkungen ergänzt wird. Als Beispiele für potenzielle Umformungen werden restringierte Optimierungsprobleme zur Approximation, Regression und Ordnungsreduktion genutzt, um die Breite der Einsatzmöglichkeiten aufzuzeigen. Dabei gilt eine Umformung als erfolgreich, wenn eine Aufgabe in der neuen Darstellung effizienter lösbar ist.

Als Konsequenz zum Teil 2 der Arbeit und aus den eigenen Erfahrungen beim Lösen von Modellierungsaufgaben ist es für die erfolgreiche Anwendung der beschriebenen Techniken wichtig zu wissen, in welche Richtungen die Umformungen erfolgen sollen. Das ist aber nur möglich, wenn das Ziel der Umformungen bekannt oder zumindest intuitiv klar ist. Eine Kenntnis der gut lösbaren restringierten Optimierungsprobleme und der verfügbaren Software (Standard, freie Downloads) ist somit unabdingbar. Die Zusammenstellung häufig auftretender Probleme im Anhang hilft hier. Nach der Festlegung des Ziels muss das jeweilige Problem analysiert werden. Hierbei sind Fragen zu beantworten wie: „Gibt es zuviele Variablen oder zuviele Restriktionen?“, „Gibt es lineare Restriktionen, die sich zur Elimination von Variablen eignen?“ oder „Welche Restriktionen sind besonders schwierig zu handhaben (Rangrestriktionen, multivariate Polynomrestriktionen)?“. Diese Analyse führt dann letztlich auf die entsprechende Methodenklasse und schlussendlich zur Wahl einer konkreten Umformungsmethode. Gleichwohl ein wenig Glück bei den Entscheidungen gehört dazu, wobei allerdings zu erwähnen ist, dass selbst Spezialisten keineswegs immer die gleichen Entscheidungen treffen.

In der Kurzbeschreibung zur Methodik, wie Vorwissen in die Modellbildung über Restriktionen einzubeziehen ist, wird deutlich, dass und vor allem wie die in der Einleitung angeführten Ziele erfüllt werden. Ergänzend seien einige Besonderheiten und ausgewählte Ergebnisse der Arbeit genannt. Hierzu zählt der weitgehende Verzicht auf Beweise, um dem lösungsorientierten Ingenieur eine stärkere Konzentration auf die Satzaussagen zu ermöglichen und

eine größere thematische Breite abdecken zu können. Der Kompromiss zwischen thematischer Breite und erforderlicher Tiefe wird durch die Zweiteilung, die Gliederung und durch Formelzusammenstellungen (u. a. Zeit-Frequenzbereichsbeziehungen, Dekompositionsregeln, Umformungen von LMI-Restriktionen, Stabilitätskonzepte) und Übersichten (u. a. Nomenklatur der Restriktionen und Optimierungsprobleme, Umformung parameternichtlinearer in parameterlineare Probleme, Implikationen zur Stabilität, Bifurkationen) geschlossen. Während sich dabei die eigenschaftsbezogene Gliederung im Teil 1 in Werken zur Systemtheorie wiederfindet, ist die hier gewählte methodenorientierte Gliederung im Teil 2 in der Optimierung unüblich. Dort wird bevorzugt hinsichtlich der Optimierungsprobleme (linear, nichtlinear; konvex, nichtkonvex) oder in Theorie und numerische Umsetzung gegliedert.

Eine weitere Besonderheit dieser Arbeit sind die Modifikationsmethoden, die in der Optimierungsliteratur weitgehend unbeachtet bleiben. Das liegt daran, dass die betreffende Community stillschweigend beschlossen hat, dass Verfahren, die nicht wenigstens quadratisch konvergieren ebenso uninteressant sind wie jene, die nur suboptimale Lösungen liefern. Anwender aus der Statistik, der Signalmodellierung, der Regelungstechnik sehen das weniger streng und erzielen mit den Methoden gute praktische Lösungen. So kann beispielsweise eine lineare Konvergenz je nach Problemordnung völlig ausreichen, zumal die Algorithmen vielfach deutlich einfacher sind, was die Implementierung auf einem Mikrocontroller vielleicht überhaupt erst gestattet. Auch die Suboptimalität ist in bestimmten Anwendungen kein Hindernis, da beispielsweise ein Regler ohnehin mit Modellfehlern umgehen muss. Der Tausch von Optimalität gegen Suboptimalität mit dem Ziel eines Rechenzeit- oder Eindeutigkeitsgewinns ist durchaus ein probates Mittel.

In Standardwerken zur Modellbildung spielen die Matrixapproximationen bisher kaum eine Rolle, wird vom Eckart-Young-Mirsky-Theorem [299] zur Herleitung der totalen Least-Squares-Methode einmal abgesehen. In dieser Arbeit hingegen werden sie immer wieder herangezogen. Zum einen gestatten sie Vereinfachungen im Sinne einer Problemmodifikation, zum anderen dienen sie der Modellbewertung, indem sie Ja/Nein-Eigenschaften verstetigen und quantifizieren. Sie eröffnen zudem eine geometrische Sicht auf bestimmte Probleme.

Eine weitere Abweichung zu herkömmlichen Darstellungen betrifft die Matrixableitungskalküle. Diese werden besonders in der Ingenieur- aber auch Statistikliteratur [421], wo sie ebenso als Werkzeug verwendet werden, als reine Schemata der partiellen Ableitungen angesehen. Solange es nur um das Ableiten und Nullsetzen zum Bestimmen der stationären Lösungspunkte geht, mag diese Betrachtung ausreichen. In dieser Arbeit wird aber auf den direkten Bezug der Schemata zur Fréchet-Ableitung hingewiesen. Darauf aufbauend wird ein neues Kalkül zur Ableitung vorgestellt, das angewendet werden kann, wenn zwischen den Matrixelementen affine Abhängigkeiten bestehen.

Die angeführten Besonderheiten in der Gestaltung und der Betrachtungsweise untermauern den Kompromiss aus thematischer Breite und erforderlicher Tiefe. Letztendlich bleibt es aber ein Kompromiss allein schon wegen des zu begrenzenden Umfangs der Arbeit. Statistische Aspekte oder strukturelle Restriktionen an Grade, Indizes, Ordnungen bleiben offen. Auch das Erstellen und Behandeln von Restriktionen bei der Modellierung nichtlinearer Systeme fällt vergleichsweise kurz aus. Für nichtlineare Systeme sind nämlich im Gegensatz zur linearen System- und Regelungstheorie die Methoden zur Identifikation und erst recht die mit Berücksichtigung von Vorwissen nicht so weit entwickelt. Zudem verhindert allein die facettenreiche Stabilitätseigenschaft im weiten Umfang generalisierbare Lösungsmethoden. Es ist deshalb zu erwarten, dass sich neue Zugänge in den kommenden Jahren vordergründig auf konkrete Anwendungen und enge Systemklassen beschränken werden. Das trifft auch auf die Einsatzmöglichkeiten des bereits bewährten Verbundbeobachterzugangs zu, bei dem Zustände und Parameter gleichermaßen bestimmt werden. Sowohl seine Konstruktion als auch der Stabilitätsnachweis, gegebenenfalls mit Angabe des Einzugsbereichs, stellen keine einfachen Aufgaben dar.

Wesentlich positiver wird sich der Einsatz numerischer Suchverfahren gestalten, die für die Modellierung nichtlinearer Systeme und zur Lösung nichtkonvexer Probleme ganz allgemein eingesetzt werden können. Das liegt vor allem an der wachsenden Rechnerleistung, der den Algorithmen eigenen Selbststabilisierung durch die permanente Gütewertverbesserung und der einfachen Anpassbarkeit an unterschiedliche Problemstellungen. Dem Nachteil eines nicht vollständigen Verständnisses über das Lösungsverhalten, also zum Beispiel über existierende Nebenminima, kann durch numerische Großzahlversuche begegnet werden.

Ein weiterer Trend der kommenden Jahre wird der verstärkte Einsatz der semidefiniten Optimierung in der Modellbildung sein, wobei insbesondere die Anzahl praktischer Anwendungen zunehmen wird. Das liegt an der mittlerweile gut verfügbaren Software, aber auch daran, dass diese Thematik zunehmend in die universitäre Ingenieurausbildung Einzug hält.

Zusammenfassend ist festzuhalten: Das Einbeziehen von Vorwissen über Parameterrestriktionen belohnt den Anwender mit besseren Modellen, verlangt ihm aber weitreichende mathematische Kenntnisse ab.

Anhang

A.1 Anmerkungen zum Bezeichnerapparat

Entsprechend dem fachübergreifenden Charakter der Arbeit wurde sich bei der Variablenbezeichnung an den in den jeweiligen Wissensgebieten gebräuchlichen Darstellungen orientiert. So ist x bei systemtheoretischen Betrachtungen die Zustandsvariable, bei optimierungstechnischen Betrachtungen hingegen die Unbekannte, bei der Datenapproximation wiederum das Argument von f. Die jeweilige Bedeutung erschließt sich dabei aus dem Kontext.
Mit Großbuchstaben $A, B, C, \ldots$ werden i. Allg. Matrizen bezeichnet; eine Ausnahme bilden die statische Verstärkung K oder die Lipschitz-Konstante L. Als Standardmatrizen treten die Nullmatrix $0_{m\times n}$, die Einsmatrix $1_{m\times n}$ und die Einheitsmatrix I_n auf, wobei die Indizes die jeweiligen Dimension angeben. A^T meint die transponierte, A^H die konjugiert transponierte Matrix. Kleine Buchstaben a, b, c bezeichnen in aller Regel Vektoren, gelegentlich aber auch skalare Variablen. Griechische Buchstaben stehen für Winkel, reelle Variablen oder in einigen Fällen auch für komplexe Variablen. Zufallsvariablen erscheinen im Fettdruck. Mengen werden in kalligrafischer Schrift gesetzt, also $\mathcal{A}, \mathcal{B}, \ldots$ Für die Zahlenbereiche werden die Symbole $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}$ und $\mathbb{C}$ gewählt. $\mathbb{R}^n$ bzw. $\mathbb{R}^{m\times n}$ steht für die betreffenden Vektor- bzw. Matrixerweiterungen und $\mathbb{R}_>$, $\mathbb{R}_\geq$ bzw. $\mathbb{R}^n_\geq$ für die positiven, die nichtnegativen Zahlen bzw. den nichtnegativen Orthanten. Eine allgemeine konvexe Menge wird gemeinhin mit $\mathcal{C}$, ein Vektorraum mit $\mathcal{V}$ und ein reeller Hilbert-Raum mit $\mathcal{H}(\mathbb{R})$ bezeichnet. $\mathcal{L}^\perp$ bezeichnet das orthogonale Komplement der Menge $\mathcal{L}$. Für die stetigen, einmal stetig differenzierbaren, k-mal stetig differenzierbaren, die glatten bzw. die analytischen Funktionen werden die Symbole $\mathcal{C}^0, \mathcal{C}^1, \mathcal{C}^k, \mathcal{C}^\infty$ bzw. $\mathcal{C}^\omega$ verwendet. Die Mengen $\mathcal{S}_n, \mathcal{S}^\geq_n, \mathcal{S}^>_n$ kennzeichnen die symmetrischen, nichtnegativ definiten bzw. positiv definiten Matrizen. $\mathcal{N}(A)$ und $\mathcal{R}(A)$ spezifizieren den Null- und Bildraum einer Matrix; für Funktionssymbole und Normen gelten die in der Mathematik üblichen Standards. Weniger gebräuchliche Symbole werden im Text oder in einer Fußnote erklärt.
Der Vergleich von Vektoren erfolgt elementweise (natürlichen Halbordnung), d. h. $x \geq y$ meint $x_i \geq y_i; i = 1,\ldots, n$. Symmetrische Matrizen werden im Sinne der Löwner-Halbordnung verglichen, d. h. $A \succeq B$ meint $A - B \in \mathcal{S}^\geq_n$.
Abweichend von der üblichen Art der Formulierung einer Minimierungsaufgabe gemäß

minimiere	x^2
unter der Nebenbedingung	$x < 4$

wird hier die kompakte Darstellung $x^2 \stackrel{!}{=} \text{Min}; x < 4$ gewählt. $\stackrel{!}{=}$ ist als eine Aufforderung zum Lösen der Aufgabe zu verstehen und kann beim Lesen durch „soll sein“ ersetzt werden.

A.2 Abkürzungen

Abkürzungen für Methoden sind im Abschnitt A.4 inklusive der mathematische Charakterisierung zusammengestellt. Alle Abkürzungen, die sich auf Stabilitätseigenschaften beziehen, finden sich in Abschnitt A.6.2. Die im Text mehrfach verwendeten Abkürzungen enthält die folgende Übersicht, während Abkürzungen, die nur in einem Abschnitt vorkommen, in diesem benannt und erklärt werden.

AKF	Autokovarianzfunktion
ALS	Alternating Least Squares
AR	Autoregressive
E/A	Eingangs-/Ausgangs-
F-, G-	Fréchet-(differenzierbar; Ableitung), dgl. für Gâteaux-
IQML	Iterative Quadratic Maximum Likelihood
IRLS	Iteratively Reweighted Least Squares
KYP	Kalman-Yakubovich-Popov
LMI	Linear Matrix Inequality
LS	Least Squares
LTI	Linear Time-invariant
LTV	Linear Time-variant
MA	Moving-Average
MIMO	Multiple Input Multiple Output
MISO	Multiple Input Single Output
ML	Maximum Likelihood
MSE	Mean Squared Error
PR	Positive Real
QR	Matrixprodukt QR mit Q orthogonal und R obere Dreiecksmatrix
SISO	Single Input Single Output
SPR	Strictly Positive Real
SQP	Sequential Quadratic Programming
SVD	Singular Value Decomposition

A.3 Begriffsbestimmung für Optimierungsprobleme

Tabelle A.1 fasst Begriffe der Optimierung zusammen [71], [320], [419], wobei lediglich die Unterscheidung in explizite und implizite zulässige Menge nicht allgemeiner Standard ist. Sie erweist sich aber als nützlich, da die implizite zulässige Menge (Schnitt der Definitionsbereiche, bezeichnet mit $\operatorname{dom} f$) die sog. impliziten Restriktionen formuliert (s. Abschn. A.5). Diese zusätzlichen Einschränkungen ergeben sich beispielsweise aus der Nichtnegativität von Radikanden und werden von unerfahrenen Anwendern gern übersehen.

Bezeichnung	Kürzel	Bedingung
explizite zulässige Menge	$\mathcal{F}_{\text{expl}}$	$\mathcal{F}_{\text{expl}} = \{x \in \mathcal{V} : g(x) \leq 0_p, h(x) = 0_m\}$
implizite zulässige Menge	$\mathcal{F}_{\text{impl}}$	$\mathcal{F}_{\text{impl}} = \operatorname{dom} f \cap \bigcap_{i=1}^{p} \operatorname{dom} g_i \cap \bigcap_{i=1}^{m} \operatorname{dom} h_i$
zulässige Menge	$\mathcal{F}$	$\mathcal{F} = \mathcal{F}_{\text{expl}} \cap \mathcal{F}_{\text{impl}}$
Supremum von f auf $\mathcal{F}$ kleinste obere Schranke	$\sup\limits_{x\in\mathcal{F}} f(x)$	$\sup\limits_{x\in\mathcal{F}} f(x) \stackrel{\text{def}}{=} \min\limits_{y\in\mathbb{R}\cup\{\infty\}} \{y : y \geq f(x); x \in \mathcal{F}\}$ $\sup\limits_{x\in\emptyset} f(x) \stackrel{\text{def}}{=} -\infty$
Infimum von f auf $\mathcal{F}$ größte untere Schranke	$\inf\limits_{x\in\mathcal{F}} f(x)$	$\inf\limits_{x\in\mathcal{F}} f(x) \stackrel{\text{def}}{=} \max\limits_{y\in\mathbb{R}\cup\{-\infty\}} \{y : y \leq f(x); x \in \mathcal{F}\}$ $\inf\limits_{x\in\emptyset} f(x) \stackrel{\text{def}}{=} +\infty$
globales Minimum von f absolutes Minimum	$\min\limits_{x\in\mathcal{F}} f(x)$	$\forall x \in \mathcal{F} : f(x) \geq \min\limits_{x\in\mathcal{F}} f(x) = f_{\min}$ und $\min\limits_{x\in\mathcal{F}} f(x) \in \mathcal{WB}(f)$
lokales Minimum von f relatives Minimum[1]	$f(x_{\text{loc}})$	$\exists \mathcal{U}(x_{\text{loc}}) \subset \mathcal{V}$, $\forall x \in \mathcal{U}(x_{\text{loc}}) \cap \mathcal{F} : f(x) \geq f(x_{\text{loc}})$
strenges lokales Minimum isoliertes lokales Minimum		$\exists \mathcal{U}(x_{\text{loc}}) \subset \mathcal{V}$, $\forall x \in \mathcal{U}(x_{\text{loc}}) \cap \mathcal{F} \backslash \{x_{\text{loc}}\} : f(x) > f(x_{\text{loc}})$
Extremum	$f_{\min}, f_{\max}$	$f_{\min} = f(x_{\text{opt}}) = \min\limits_{x\in\mathcal{F}} f(x)$ $f_{\max} = f(x_{\text{opt}}) = \max\limits_{x\in\mathcal{F}} f(x)$
Minimierer von f	$\arg\min\limits_{x\in\mathcal{F}} f(x)$	$f(\arg\min\limits_{x\in\mathcal{F}} f(x)) = \min\limits_{x\in\mathcal{F}} f(x)$
Lösung, Optimallösung, Optimalpunkt	x_{opt}	$x_{\text{opt}} = \arg\min\limits_{x\in\mathcal{F}} f(x)$ $x_{\text{opt}} = \arg\max\limits_{x\in\mathcal{F}} f(x)$
Lösungsmenge	$\mathcal{X}_{\text{opt}}$	$\mathcal{X}_{\text{opt}} = \{x_{\text{opt}} \in \mathcal{F} : f(x_{\text{opt}}) = \min\limits_{x\in\mathcal{F}} f(x)\}$ $\mathcal{X}_{\text{opt}} = \{x_{\text{opt}} \in \mathcal{F} : f(x_{\text{opt}}) = \max\limits_{x\in\mathcal{F}} f(x)\}$
ε-approximative Lösung	x_ε	$f(x_\varepsilon) \leq f(x_{\text{opt}})$ und $\lvert h(x_\varepsilon)\rvert \leq \varepsilon 1_m$ und $g(x_\varepsilon) \leq \varepsilon 1_n$

Tabelle A.1: Elementare Grundbegriffe aus der Optimierung

[1] Jeder isolierte Punkt ist sowohl ein lokaler Minimierer als auch ein lokaler Maximierer.

In Tabelle A.2 werden Standardbezeichnungen für Optimierungsprobleme aus [71], [320], [419] kompakt zusammengefasst und in Beispiel A.1 erläutert. Der Vollständigkeit halber sind die unendlichdimensionalen Probleme mit aufgeführt, obwohl sie in dieser Arbeit nicht betrachtet werden. Unendlichdimensionale Vektorräume treten beispielsweise bei der nichtparametrischen Modellierung auf, wenn bezüglich einer Funktion zu optimieren ist.
Der Begriff der semiunendlichen Probleme gehört nicht zum Standardrepertoire. Er kennzeichnet den Umstand, dass eine endliche Anzahl von Optimierungsvariablen eine unendliche Anzahl von Restriktionen einhalten soll. Ein Beispiel ist die Bestimmung der Polynomkoeffizienten unter der Nichtnegativität für alle, unendlich vielen x. Ein weiteres Beispiel ist die Schätzung einer endlichen Autokovarianzfolge, die eine nichtnegative Spektraldichte entlang des Einheitskreises aufweisen muss. Beide Beispiele werden in der Arbeit betrachtet, und es werden Wege zur Umgehung der unendlich vielen Restriktionen aufgezeigt.

Bezeichnung	**Erklärung**
freies Problem	Problem ohne Restriktionen
restringiertes Problem	Problem mit Restriktionen
zulässiges Problem	$\mathcal{F} \neq \emptyset$
unzulässiges Problem	$\mathcal{F} = \emptyset,\ \min\limits_{x\in\mathcal{F}} f(x) \stackrel{\text{def}}{=} \infty$
Zulässigkeitsproblem	Finde ein $x \in \mathcal{F}$ oder zeige, dass $\mathcal{F} = \emptyset$.
beschränktes Problem	$\mathcal{F} \neq \emptyset,\ \inf\limits_{x\in\mathcal{F}} f(x)$ ist endlich.
unbeschränktes Problem	$\mathcal{F} \neq \emptyset,\ \inf\limits_{x\in\mathcal{F}} f(x) = -\infty$
lösbares Problem	$\mathcal{X}_{\text{opt}} \neq \emptyset$ Klassifikation nach Kardinalität von $\mathcal{X}_{\text{opt}}$ (eindeutig; mehrdeutig endlich, mehrdeutig unendlich)
nicht lösbares Problem	$\mathcal{X}_{\text{opt}} = \emptyset$
äquivalentes Problem	$\mathcal{F}$, $\mathcal{X}_{\text{opt}}$, f_{opt} sind gleich bzw. ergeben sich wechselseitig.
endlichdimensionales Problem	$\mathcal{F} \subseteq \mathcal{V}, \dim \mathcal{V} < \infty$ mit endlich vielen Restriktionen
unendlichdimensionales Pbl.	$\mathcal{F} \subseteq \mathcal{V}, \dim \mathcal{V} = \infty$
semiunendliches Problem	$\mathcal{F} \subseteq \mathcal{V}, \dim \mathcal{V} < \infty$ mit unendlich vielen Restriktionen

Tabelle A.2: Klassifikation der Optimierungsprobleme

Beispiel A.1 (Erläuterung zur Problemcharakterisierung)
Jedes unzulässige Problem ist nicht lösbar, aber nicht umgekehrt. So ist $1/x \stackrel{!}{=} \text{Min}$ ist auf $\mathbb{R}_>$ nicht lösbar, aber zulässig und beschränkt.
$x^3 - 3x \stackrel{!}{=} \text{Min}$ ist auf $\mathbb{R}$ unbeschränkt, nicht lösbar, hat aber einen lokalen Minimierer.
$x_1 \stackrel{!}{=} \text{Min}; -1 \leq x_1 \leq 1, x_2 \leq 1$ ist beschränkt, die Lösungsmenge ist aber unbeschränkt.
Ein zulässiges, nicht lösbares Problem kann beschränkt oder unbeschränkt sein.

A.4 Nomenklatur von Optimierungsproblemen

In den folgenden Tabellen werden vom Autor Optimierungsprobleme kompakt zusammengestellt, die sich während seiner Arbeiten zur Modellierung ergaben und bei Literaturrecherchen als interessant und nützlich erachtet wurden. Die Übersichten zeigen dem Ingenieur die Unterschiede in den Zielfunktionen und/oder Restriktionen und ermöglichen die Zuordnung einer konkreten Aufgabe zu einer Problemklasse. Durch gezielte Suche nach dem englischen Fachbegriff können schnell Aussagen zur Existenz und Eindeutigkeit der Lösungen, zu geschlossenen Lösungen oder zugeschnittenen Algorithmen gefunden werden.

Ordinary Least Squares	LS, OLS	$\|\|Ax-b\|\|_2^2 \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n$
Weighted Least Squares	WLS	$\|\|D(Ax-b)\|\|_2^2 \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n; D \in \mathcal{D}_m^>$
Singular Generalized LS	SGLS	$\|\|u\|\|_2^2 \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n, u \in \mathbb{R}^k : b = Ax + Bu,$ $BB^T = W \in \mathcal{S}_n^\geq, \text{rg}W = k$
Matrix Least Squares	MLS	$\|\|AXC - B\|\|_F^2 \stackrel{!}{=} \text{Min} \quad X \in \mathbb{R}^{m \times n}$
Equality Constrained LS	LSE	$\|\|Ax-b\|\|_2^2 \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : Cx = d; C \in \mathbb{R}^{p \times n}$
Minimum Norm with Linear Equation System	MN-LES MN_W-LES	$\|\|x\|\|_2 \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : Cx = d$ $\|\|x\|\|_W \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : Cx = d$
Minimum Norm LS	MN-LS	$\|\|x\|\|_2 \stackrel{!}{=} \text{Min} \quad x \in \text{argmin}\, \|\|Ax-b\|\|_2$
Partial Least Squares	PLS	$\|\|Ax-b\|\|_2^2 \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : x = Wp; W \in \mathbb{R}^{n \times s}$ $W = (A^Tb, A^TAA^Tb, \ldots, (A^TA)^{s-1}A^Tb)$
Principal Component Regression	PCR	$\|\|Ax-b\|\|_2^2 \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : x = V_s p; V \in \mathcal{O}_{n,s}$ $A = (U_s, U_{m-s})\text{diag}(\Sigma_{s\times s}, \Sigma_{(m-s)\times(n-s)})(V_s, V_{n-s})^T$
Tikhonov Regularized LS	Ti-LS	$\|\|Ax-b\|\|_2^2 + \alpha\|\|Lx\|\|_2^2 \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n, \alpha \in \mathbb{R}^>$
Inequality Constr. LS	LSI	$\|\|Ax-b\|\|_2^2 \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : l \leq Cx \leq u$
Box constrained LS	BLS	$\|\|Ax-b\|\|_2^2 \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : l \leq x \leq u$
NonNegative LS	NNLS	$\|\|Ax-b\|\|_2^2 \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : x \geq 0$
Linear Inequality LS	LILS	$\|\|\max\{Ax-b, 0_m\}\|\|_2^2 \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n$ komponentenweise max-operation liefert Fehlerbeitrag, wenn $(Ax \leq b)_i$ nicht gilt

Tabelle A.3: Nomenklatur linearer LS-orientierter Kriterien

Robust LS	RLS	$\max_{\|\|\Delta A\|\|_F \le r_A, \|\|\Delta b\|\|_2 \le r_b} \|\|(A+\Delta A)x-(b+\Delta b)\|\|_2 \stackrel{!}{=} \text{Min}; x \in \mathbb{R}^n$ r_A, r_B a priori bekannte Schranken $\max_{\|\|(\Delta A \vdots \Delta b)\|\|_F \le r} \|\|(A+\Delta A)x-(b+\Delta b)\|\|_2 \stackrel{!}{=} \text{Min}; x \in \mathbb{R}^n$ r a priori bekannte Schranke
Nonlinear Least Squares	NLS	$\sum_{i=1}^{N}(\phi(x,a_i)-b_i)^2 \stackrel{!}{=} \text{Min} \quad x \in \mathbb{C}^n$ explizite Form $\sum_{i=1}^{N}\varepsilon^2(x,a_i,b_i) \stackrel{!}{=} \text{Min} \quad x \in \mathbb{C}^n$ implizite Form
Constrained NLS	CNLS	$\sum_{i=1}^{N}(\phi(x,a_i)-b_i)^2 \stackrel{!}{=} \text{Min} \quad h(x)=0_m, g(x) \le 0_r$
LS with a Quadratic Inequality	LSQI	$\|\|Ax-b\|\|_2^2 \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : \|\|Cx-d\|\|_2 \le \gamma$ $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{p\times n}, \gamma \in \mathbb{R}^{>}, m+p \ge n \ge p$
LS over Sphere	LSS	$\|\|Ax-b\|\|_2^2 \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : \|\|x\|\|_2 \le \gamma; \gamma \in \mathbb{R}^{>}$
Orthogonal Projection	OP	$\|\|x-b\|\|_2 \stackrel{!}{=} \text{Min} \quad x \in \mathcal{M} \subseteq \mathbb{R}^n$ $\mathcal{M}$ hat bestimmte Eigenschaften
LS with Variable Metric	LSVM	$\|\|Ax_1-b\|\|_{W(x_2)}^2 \stackrel{!}{=} \text{Min} \quad x_1 \in \mathbb{R}^n, x_2 \in \mathbb{R}^p$
Procrustian Problems	PP	$\|\|A-XBY\|\|_F \stackrel{!}{=} \text{Min} \quad X \in \mathcal{M}_1, Y \in \mathcal{M}_2$ $\mathcal{M}_1, \mathcal{M}_2$ haben bestimmte Eigenschaften $\mathcal{M}_1, \mathcal{M}_2$ sind abgeschlossen
Normalized Generalized Block-Rayleigh-Quotient	NGBRQ	$\text{spur}(X^TAX) \stackrel{!}{=} \text{Min} \quad X \in \mathbb{C}^{m\times n} : X^TBX = I_n$
Linear Matrix Inequality Least Squares	LS-LMI	$\|\|AXC-B\|\|_F^2 \stackrel{!}{=} \text{Min} X \in \mathbb{R}^{m\times n} : L \preceq F(X) \preceq U$ F ist eine lineare Abbildung
Least Median of Squares	LMS_w	$\varepsilon_w^2(x) \stackrel{!}{=} \text{Min} \quad x \in \mathcal{X} \subseteq \mathbb{R}^n$ geordnet gemäß $\varepsilon_1^2 \le \varepsilon_2^2 \le \ldots \le \varepsilon_N^2$
Least Trimmed Squares	LTS_w	$\sum_{i=1}^{w}\varepsilon_i^2(x) \stackrel{!}{=} \text{Min} \quad x \in \mathcal{X} \subseteq \mathbb{R}^n$ geordnet gemäß $\varepsilon_1^2 \le \varepsilon_2^2 \le \ldots \le \varepsilon_N^2$
Dead-Zone LS	DZ-LS	$\sum_{i=1}^{N} l_q(\varepsilon_i(x)) \stackrel{!}{=} \text{Min}; l_q(\varepsilon) = \begin{cases} \frac{1}{2}(x-c)^2 & x > c \\ 0 & \|x\| \le c \\ \frac{1}{2}(x+c)^2 & x < -c \end{cases}$

Tabelle A.4: Nomenklatur nichtlinearer LS-orientierten Kriterien

Minimum Norm problem	MN_p	$\|\|x\|\|_p \stackrel{!}{=} \mathrm{Min} \quad x \in \mathbb{R}^n : Ax = b; 1 \leq p \underset{(=)}{\leq} \infty$
l_p-approximation	l_p	$\|\|Ax - b\|\|_p \stackrel{!}{=} \mathrm{Min} \quad x \in \mathbb{R}^n; A \in \mathbb{R}^{m \times n}$
log-l_∞-approximation	log-l_∞	$\max_i \|\log(a_i^T x) - \log b_i\| \stackrel{!}{=} \mathrm{Min}, x \in \mathbb{R}^n; A \in \mathbb{R}^{m \times n}$
MN_q-l_p-approximation	MN_q-l_p	$\|\|x\|\|_q \stackrel{!}{=} \mathrm{Min} \quad x \in \mathrm{argmin}\{\|\|Ax - b\|\|_p\}$
Total-l_p-approximation	T-l_p	$\|\|[A, b]x\|\|_p \stackrel{!}{=} \mathrm{Min} \quad x :\in \mathbb{R}^n : \|\|x\|\|_q = 1$
Structured Nonlinear Total Least Norm	SNTLN_p	$\left\| \begin{smallmatrix} \Delta b \\ D\Delta z \end{smallmatrix} \right\|_p \stackrel{!}{=} \mathrm{Min} \quad A(z + \Delta z)x = b + \Delta b$ $A(z)$ hat eine nichtlineare Struktur
Robust TLN	RTLN	$\max_{\|\|\Delta A\|\|_\rho \leq r_A, \|\|\Delta b\|\|_\alpha \leq r_b} \|\|(A + \Delta A)x - (b + \Delta b)\|\|_\alpha \stackrel{!}{=} \mathrm{Min}\ x \in \mathbb{R}^n$ $\|.\|_\rho$ ist separabel bezüglich $\|.\|_\alpha$
M-Estimator	ME	$\sum_{i=1}^{N} \rho(\varepsilon_i(x)/\sigma_i) \stackrel{!}{=} \mathrm{Min} \quad x \in \mathcal{X} \subseteq \mathbb{R}^n$
Set-Membership Identification	SMI	$\sum_{i=1}^{N} l_c(\varepsilon_i(x)) \stackrel{!}{=} \mathrm{Min};\ l_c(\varepsilon) = \begin{cases} 0 & \|\varepsilon\| \leq c \\ \infty & \|\varepsilon\| > c \end{cases}$ equivalent: $x_{\mathrm{opt}} \in \{x : \|\varepsilon_i(x)\| \leq c\}$
Relative Component-wise Distance Problem	RCDP	$\delta \stackrel{!}{=} \mathrm{Min} \quad \|\Delta A\| \leq \delta \|A\|, (A + \Delta A) \in \mathcal{M}$ $\mathcal{M}$ hat bestimmte Eigenschaften
Determinant Maximization	MaxDet	$\ln \det G(x) - c^T x \stackrel{!}{=} \mathrm{Max} \quad x \in \mathbb{R}^n : F(x) \succeq 0_{p \times p}$ $G(x) = G_0 + \sum_{i=1}^{n} x_i G_i \succ 0_{q \times q}, G_i^T = G_i$ $F(x) = F_0 + \sum_{i=1}^{n} x_i F_i, F_i^T = F_i$
Matrix Approximation Problems	MAP	$\|\|A - X\|\| \stackrel{!}{=} \mathrm{Min} \quad X \in \mathcal{M} \subseteq \mathbb{C}^{m \times n}$ $\mathcal{M}$ hat bestimmte Eigenschaften

Tabelle A.5: Nomenklatur ausgewählter nichtquadratischer Kriterien

Total LS distance	TLS	$\|\|[\Delta A, \Delta b]\|\|_F \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : (A + \Delta A)x = b + \Delta b$
TLS with normed x	TLSN_W	$\|\|[\Delta A, \Delta b]\|\|_F \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^{n+1} : \ \|\|x\|\|_W = 1$ $([A, b] + [\Delta A, \Delta b])x = 0_m$
Tikhonov regularized TLS	Ti-TLS	$\|\|[\Delta A, \Delta b]\|\|_F \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : \|\|Lx\|\|_2 \leq \delta$ $(A + \Delta A)x = b + \Delta b$
Truncated TLS distance	TTLS	$\|\|[\Delta A, \Delta b]\|\|_F \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : (A + \Delta A)x = b + \Delta b$ $\text{rg}(A + \Delta A, b + \Delta b) = s < n$
Data LS distance	DLS	$\|\|\Delta A\|\|_F \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : (A + \Delta A)x = b$
Weighted TLS distance	WTLS	$\|\|[\Delta A, \Delta b]D\|\|_F \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : (A + \Delta A)x = b + \Delta b$ $A \in \mathbb{R}^{m \times n}, D \in \mathcal{D}^{>}_{n+1}$
Complete WTLS	CWTLS	$\|\|D_1[\Delta A, \Delta B]D_2\|\|_F \stackrel{!}{=} \text{Min} \quad D_1 \in \mathcal{D}^{>}_m, D_2 \in \mathcal{D}^{>}_{n+k}$ $X \in \mathbb{R}^{n \times n} : (A + \Delta A)X = B + \Delta B; \ A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{m \times k}$
Fixed Column TLS	FCTLS	$\|\|\Delta A_2\|\|_F \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^{n+1} : \|\|x\|\|_2 = 1$ $(A_1, A_2 + \Delta A_2)x = 0$
TLS with Equalities	TLSE	$\|\|[\Delta A, \Delta b]\|\|_F \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : (A + \Delta A)x = b + \Delta b$ $Cx = d; C \in \mathbb{R}^{p \times n}$
Principal Component Analysis	PCA	$\sum_{i=1}^{n} \|\|a_i - x_i\|\|_2^2 \stackrel{!}{=} \text{Min} \quad x_i \in \mathcal{B}_k \subset \mathbb{R}^m$ $\mathcal{B}_k$ beliebige k-dimensionale Hyperebene
Generalized TLS	GTLS	$\|\|E\|\|_F \stackrel{!}{=} \text{Min} \quad X \in \mathbb{R}^{m \times n}, A \in \mathbb{R}^{p \times m}, B \in \mathbb{R}^{p \times q}$ $\left[[A, B] + E(\text{cov}\boldsymbol{\varepsilon})^{1/2}\right] \begin{bmatrix} X \\ -I_q \end{bmatrix} = 0_{p \times n}$
Nonlinear TLS Orthogonal Regression	NTLS OR	$\sum_{i=1}^{N} \|\|\Delta z_i\|\|_2^2 \stackrel{!}{=} \text{Min} \quad f(z_i + \Delta z_i, x) = 0; i = 1, \ldots, N$ $z_i, \Delta z_i \in \mathbb{R}^p, x \in \mathbb{R}^n$

Tabelle A.6: Nomenklatur TLS-orientierter Kriterien

Affine Structured TLS	ASTLS	$\|\|[\triangle A, \triangle b]\|\|_F \stackrel{!}{=} \text{Min} \quad x \in \mathbb{C}^n : (A + \triangle A)x = b + \triangle b$ $(\triangle A, \triangle b)$ hat die Struktur von (A, b)
Constrained TLS	CTLS	$f^T D f \stackrel{!}{=} \text{Min}\, x \in \mathbb{R}^n, f \in \mathbb{R}^k; D \in \mathcal{D}_k^{\geq}$ $(A + [F_1 f, \ldots, F_n f])x = b + F_{n+1} f$ f Rauschvektor; F_i binäre Matrizen
Structured TLS	STLS	$\sum_{i=1}^{k} w_i (t_i - s_i)^2 \stackrel{!}{=} \text{Min} \qquad t_i \in \mathbb{R}, i = 1, \ldots, k$ $\left(T_0 + \sum_{i=1}^{k} t_i T_i\right) y = 0_{n+1}, \|\|y\|\|_2 = 1; T_i \in \mathbb{R}^{m \times (n+1)}$ s_i sind Elemente von $(A, b); T_i$ binäre Matrizen
Structured Total Least Norm	STLN_2	$\left\|\left[\begin{array}{c} D_\alpha \alpha \\ D_\beta \beta \end{array}\right]\right\|_2 \stackrel{!}{=} \text{Min}\, x \in \mathbb{R}^n, \alpha \in \mathbb{R}^p, \beta \in \mathbb{R}^q; D_\cdot \in \mathcal{D}_p^{\geq}$ $Ax + X\alpha = b + F\beta + G\alpha$ F, G binäre Matrizen

Tabelle A.7: Nomenklatur von TLS-Kriterien mit affinen Nebenbedingungen

Linear Programming	LP	Allgemeine Form $c^T x \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : \; Ax = b; A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$ $Cx \leq d; C \in \mathbb{R}^{p \times n}, d \in \mathbb{R}^p$
		Standardform[2] $c^T x \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : Ax = b, x \geq 0_n$
		Ungleichungsform $c^T x \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : Ax \geq b$
		Kanonische Form $c^T x \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : Ax \geq b, x \geq 0_n$
Second-order Cone Progr.	SOCP	$c^T x \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : \|\|C_i x + d_i\|\| \leq e_i^T x + f_i$ $C_i \in \mathbb{R}^{n_i \times m}, d_i \in \mathbb{R}^{n_i}, e_i \in \mathbb{R}^m,$ $f_i \in \mathbb{R}; i = 1, \ldots, p$
Semidefinite Programming	SDP	Kanonische Form (LMI-Form) $c^T x \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : F(x) \succeq 0_{p \times p}$
		Standardform $c^T x \stackrel{!}{=} \text{Min} \quad Ax = b, x \in \mathcal{C}$ $\mathcal{C}$ ist zu den semidefiniten Matrizen isomorph
Convex Programming	CP	Allgemeine Form[3] $f(x) \stackrel{!}{=} \text{Min} \quad x \in \mathcal{C} : h(x) = 0_m, g(x) \leq 0_p$ $\mathcal{C}$ ist eine konvexe Menge; h_j affin, f, g_j konvex
		Standardform $c^T x \stackrel{!}{=} \text{Min} \quad x \in \mathbb{R}^n : Ax = b, g(x) \leq 0_p$
		Epigraph-Form $x_{n+1} \stackrel{!}{=} \text{Min} \quad \begin{bmatrix} x \\ x_{n+1} \end{bmatrix} \in \mathbb{R}^{n+1} :$ $Ax = b, g(x) \leq 0_p, f(x) - x_{n+1} \leq 0$

Tabelle A.8: Nomenklatur einiger konvexer Optimierungsprobleme

[2] Mitunter heißt die kanonische Form auch Standardform und die Standardform dann Slackform. Hier soll die Standardform folgende Merkmale haben: lineare Zielfunktion, lineare Gleichungsrestriktionen und eine Bedingung $x \in \mathcal{C}$. Die einzelnen Probleme unterscheiden sich dann in der Menge $\mathcal{C}$. Es gilt folgende Hierarchie LP $\subset$ SOCP $\subset$ SDP $\subset$ CP.

[3] Beachte: Die Darstellungsformen unterstellen durchweg, dass kein $g_j(x) \leq 0$ singulär ist.
$x_1^2 + x_2^2 \stackrel{!}{=} \text{Min}; g(x) = x_1/(1 + x_2^2) \leq 0, h(x) = (x_1 + x_2)^2 = 0$ mit $\mathcal{F} = \{(x_1, x_2) : x_1 = -x_2 \leq 0\}$ ist ein konvexes Problem, das nicht in der allgemeinen Form vorliegt. Die hierzu äquivalente Form lautet $x_1^2 + x_2^2 \stackrel{!}{=} \text{Min}; x_1 \leq 0, x_1 + x_2 = 0$.

A.5 Begriffsbestimmung für Restriktionen

Die Tabellen A.9 und A.10 geben eine Übersicht zur Charakterisierung von Restriktionen hinsichtlich ihrer mathematischen Eigenschaften, wobei einige Begriffe selbsterklärend sind und andere zum mathematischen Standardrepertoire gehören [71], [320], [419]. In Tabelle A.11 hat der Autor ergänzend einige Charakterisierungen von Restriktionen aus funktioneller Sicht zusammengestellt. Einfache bzw. typische Beispiele unterstützen die Begriffserklärungen.

Bezeichnung	**Erklärung**
Gleichungsrestriktion	Restriktion der Art $h_i(x) = 0$
Ungleichungsrestriktion[4]	Restriktion der Art $g_i(x) \leq 0$ oder $g_i(x) < 0$
Ungleichheitsrestriktion (engl.: inequation)	Restriktion der Art $g_i(x) \neq 0$ $g_i(x) \neq 0 \Leftrightarrow g_i(x) < 0, g_i(x) > 0$
Operatorrestriktionen	Restriktionen, die durch Operatoren formuliert werden.
funktionale Restriktion	x_i hängt von anderen x_j ab, z. B. $3x_1 + 2x_2 = 5$
nichtfunktionale Restriktion	x_i hängt nicht von anderen x_j ab, z. B. $x_i \leq 5$
aktive Restriktion	$g_i(x) \leq 0$ heißt aktiv in einem Punkt (oder straff bez. eines Punkts) $x_0 \in \mathcal{F}$, wenn $g_i(x_0) = 0$ gilt.
inaktive Restriktion	$g_i(x) \leq 0$ heißt inaktiv in einem Punkt (oder locker bez. eines Punkts) $x_0 \in \mathcal{F}$, wenn $g_i(x_0) < 0$ gilt.
reguläre Restriktion	Ungleichungsrestriktion, die für mindestens ein x streng ist.
singuläre Restriktion	Ungleichungsrestriktion, die für kein $x \in \mathcal{F}$ streng ist. Sie ist also per se aktiv und formuliert implizit eine Gleichungsrestriktion, z. B. $x^2 \leq 0$.
konsistente Restriktionen	Wenigstens ein x im Definitionsbereich von f erfüllt $h(x) = 0_m, g(x) \leq 0_p$.
inkonsistente Restriktionen	Kein x im Definitionsbereich von f erfüllt $h(x) = 0_m$, $g(x) \leq 0_p$.

Tabelle A.9: Begriffsbestimmung für Restriktionen (Teil 1)

[4] Charakterisierungen von Ungleichungen (engl.: inequalities):

Standardform	rechte Seite der Ungleichung ist Null
strenge Ungleichung	Relationszeichen $<, >$ wird benutzt
nicht-strenge Ungleichung	Relationszeichen $\leq, \geq$ wird benutzt
scharfe Ungleichung	Gleichheit gilt für mindestens ein x bei nicht-strengen Ungleichungen $\forall \varepsilon > 0 : \exists x \in \mathcal{U}_\varepsilon(0)$ bei strengen Ungleichungen in Standardform
schwächere Ungleichung	g_1 ist schwächer als g_2, wenn $\{x : g_2(x) \leq 0\} \subset \{x : g_1(x) \leq 0\}$

Bezeichnung	**Erklärung**
redundante Restriktion	Restriktion, die durch andere Restriktionen impliziert wird. Ein Streichen einer redundanten Restriktion ändert $\mathcal{F}$ nicht. streng redundant: $g_i(x) < 0$ für alle $x \in \mathcal{F}$ schwach redundant: $g_i(x) \leq 0$ für ein x oder $h_i(x) = 0$
explizite Restriktionen	Alle Restriktionen, die durch Gleichungen, Ungleichungen und/oder Operatorbeziehungen im Optimierungsproblem formuliert sind.
implizite Restriktion	Restriktion, die nicht durch f, g, h explizit aufgeführt wird. Sie resultiert aus dem Definitionsbereichen von f, g, h bzw. aus Resubstitutionsforderungen bei Parametersubstitutionen.
separable Restriktionen	Ist $\begin{bmatrix} x_A \\ x_B \end{bmatrix}$ eine permutierte Darstellung von x und gilt $\mathcal{F}_x = \mathcal{F}_{x_A} \times \mathcal{F}_{x_B}$, dann heißen die Restriktionen separabel bez. x_A und x_B.
disjunktive Restriktionen	$x \in \cup_{i=1}^{q} \mathcal{F}_i; \mathcal{F}_i \subset \mathcal{V}$ oder äquivalent $\bigvee_{i=1}^{q} (x \in \mathcal{F}_i)$
algebraische Restriktion	Gleichungen oder Ungleichungen, die keine Ableitungen oder Integrale enthalten.
differenzielle Restriktion	Üblicherweise eine Differenzialgleichung als Restriktion; tritt bei der Optimalsteuerung auf, wenn das System als Restriktion fungiert.
integrale Restriktion	Integralbeziehung als Restriktion
stochastische Restriktion	Restriktion, die in Wahrscheinlichkeit gelten muss. Sie ist eine spezielle Operatorrestriktion. In stochastischen Restriktionen sind meist Parameter und Daten miteinander verknüpft.

Tabelle A.10: Begriffsbestimmung für Restriktionen (Teil 2)

Beispiel A.2 (Implizite Restriktionen)
Quellen für implizierte Restriktionen sind Funktionen, die nicht über dem gesamten $\mathbb{R}^n$ definiert sind. So enthalten $\ln x$ bzw. $f(x_1, x_2) = \sqrt{x_1^3 + x_2^3}$ die implizite Restriktion $x > 0$ bzw. $x_1^3 + x_2^3 \geq 0$. Implizite Restriktionen ergeben sich in Lagrange-dualen Problemen für die äußere Optimierung aus der inneren Optimierung, s. Beispiel 8.6. Das Vernachlässigen impliziter Restriktionen kann falschen Ergebnissen führen, s. Beispiel 5.14. Ergänzend sei ein Klassiker für eine implizite Restriktion genannt: die ML-Schätzung des Gleichverteilungsparameters.

Aus $f_{\boldsymbol{x}_i}(\xi_i) = \begin{cases} 1/\theta & 0 \le \xi_i \le \theta \\ 0 & \xi_i < 0 \vee \xi_i > \theta \end{cases}$ ergibt sich für n unabhängige, identisch verteilte $\boldsymbol{x}_i$

$f_{\boldsymbol{x}}(\xi) = \begin{cases} 1/\theta^n & \xi \in \mathcal{D} = \{\xi : 0 \le \xi \le \theta\} \\ 0 & \xi \notin \mathcal{D} \end{cases}$ und $L(\theta; x) = \begin{cases} 1/\theta^n & \theta \ge \max_i x_i \\ 0 & \theta < \max_i x_i \end{cases}$.

Formales Differenzieren liefert wegen $\frac{\mathrm{d}L}{\mathrm{d}\theta} = -n/\theta^{n+1} < 0$ kein Maximum. Aus der Restriktion $\theta \ge \max\limits_{1 \le i \le n} x_i$ und dem monotonen Fallen von $L(\theta; x)$ auf $\theta \ge \max\limits_{1 \le i \le n} x_i$ folgt $\hat{\theta}_{ML} = \max\limits_{1 \le i \le n} x_i$.

Beispiel A.3 (Disjunktive Restriktionen)
Restriktionen aus Fuzzy-Modellen (*IF* $(x \in \mathcal{F}_1)$ *THEN* $(x \in \mathcal{F}_2)$ *ELSE* $(x \in \mathcal{F}_3)$), äquivalent $(x \in \mathcal{F}_1 \cap \mathcal{F}_2) \vee (x \notin \mathcal{F}_1, x \in \mathcal{F}_3)$, und Restriktionen bei stückweise definierten Funktionen $f(x) \le \begin{cases} 1 & x \in (-\infty, 0] \\ 2 & x \in (0, \infty) \end{cases}$, äquivalent $(x \in (-\infty, 0], f(x) \le 1) \vee (x \in (0, \infty], f(x) \le 2)$, zählen zu den disjunktiven (oder-verknüpften) Restriktionen. Ihre Behandlung erfolgt über Fallunterscheidungen, Umschreiben in eine einzelne Restriktion oder Relaxationen.

Beispiel A.4 (Stochastische Restriktion)
$\mathrm{Prob}(\mathbf{a}^T x \le b) \ge \eta$ mit $\mathbf{a} \sim N(\mu_{\mathbf{a}}, \Sigma_{\mathbf{a}})$ ist eine stochastische Restriktion. Ähnliche Restriktionen entstehen, wenn als Operator der Erwartungswert fungiert. Um stochastische Restriktionen der Optimierung zugänglich zu machen, müssen sie umgeformt werden. Im Beispiel führt dies auf $\mu_{\mathbf{a}}^T x + \Phi^{-1}(\eta) ||\Sigma_{\mathbf{a}}^{1/2} x||_2 \le b$, also eine Kegelrestriktion zweiter Ordnung (SOC-Restriktion) [287].

Beispiel A.5 (Physikalisch-technische Restriktionen)
Geometrische Überlegungen liefern für ein sich auf einer Ebene bewegendes Fahrzeug die Ebenenbedingung $c^T x - d = 0$ und für einen Satelliten die Drehspiegelungsbedingung $A^T A = I$ (orthogonal) und $\det A = 1$. Aus den Erhaltungssätzen folgt etwa für Mehrzonenöfen oder die medizinische Substanzverfolgung mittels Radioisotopen, dass die Systemmatrix diagonaldominant sein muss. Grenzen ergeben sich allein schon durch Vorzeichenbedingungen (Konzentrationen, spezifische Wärmekapazität) von Prozesskoeffizienten, die in Bedingungen für die Parameter umgerechnet werden können (Intervallrechnung).
Die physikalisch-technischen Restriktionen sind ihrerseits nicht auf Parameter beschränkt, sondern können sich auch auf Signale beziehen. Eine zu schätzende Temperatur kann in bestimmten Prozessen nur sinken (keine Energiezufuhr). Ähnliches gilt für den Durchfluss, wenn hinsichtlich des Zusetzens eines Rohres fehlerdetektiert werden soll.

Beispiel A.6 (Temporäre Restriktionen)
In SQP-Algorithmen[5] werden Trust-Region-Ellipsoide [71] eingebaut, die dem Fortschritt des Algorithmus entsprechend angepasst werden und die im Grenzpunkt inaktiv sind.

[5] Es wird eine Folge quadratischer Ersatzprobleme gelöst, die durch eine quadratische Approximation der Zielfunktion und Linearisieren der Restriktionen in jedem Iterationsschritt entstehen.

Bezeichnung	**Erklärung**
physikalisch-technische Restriktion	Restriktion, die aus der Geometrie, Erhaltungssätzen, physikalisch-technischen Grenzen usw. folgt.
magere Restriktion	Restriktion, die nur eine magere Teilmenge ausschließt; kann häufig entfallen, da sie generisch erfüllt wird; z. B. Restriktion für Steuerbarkeit eines LTI-Systems
eindeutigkeitserzwingende Restriktion	Restriktion, die dem Problem zugefügt wird, um eine prädestinierte Lösung aus einer mehrelementigen Lösungsmenge zu erhalten.
temporäre Restriktion	Ungleichungsrestriktion, die $\mathcal{F}$ in einem iterativen Algorithmus zeitlich befristet eingrenzt, im Minimum aber stets inaktiv ist.
optimierungstechnische Restriktion	Ungleichungsrestriktion, die dem Problem a priori hinzugefügt wird, um uninteressante oder entartete Subminima auszuschließen und somit das Konvergenzverhalten zu verbessern; sie ist im Minimum inaktiv.
zu vererbende Restriktion	Restriktion oder strukturelle Maßnahme, um bestimmte Systemeigenschaften bei Modellapproximationen zu erhalten (Stabilität, Passivität, Monotonie, usw.)
methodenbedingte Restriktion	Verheftungsrestriktion der Interpolation; Regularisierungsrestriktion; Fehlerkopplungsbedingungen bei TLS, CTLS usw.
schätzungsbedingte Restriktion	Restriktion, die insbesondere aus der Forderung nach Erwartungstreue, Konsistenz, Effizienz oder auch Symmetrie, Definitheit resultieren.

Tabelle A.11: Begriffsbestimmung für Restriktionen (Teil 3)

Beispiel A.7 (Schätzungsbedingte Restriktionen)
Für einen affinen Schätzer $\widehat{W\boldsymbol{\theta}} = A\boldsymbol{y} + c$ einer Lineartransformation von $\hat{\boldsymbol{\theta}}$ folgen aus der Forderung nach Erwartungstreue

$$\mathrm{E}\{\widehat{W\boldsymbol{\theta}}\} = W\theta = AX\theta + c \qquad \forall\theta$$

die Restriktionen

$$AX = W \qquad \text{und} \qquad c = 0.$$

Für den besten quadratischen positiven Schätzer für die Varianz, d. h. $\hat{\sigma}^2 = y^T Ay$ folgt aus der Positivität

$$A = C^T C.$$

A.6 Stabilität

A.6.1 Formelnotation für Stabilitätsdefinitionen

Symbol	Name, ggf. Definition
Hom	Homöomorphismus: stetige, bijektive Abbildung mit stetiger Inverse
$\mathcal{K}$	Klasse-$\mathcal{K}$-Funktion (Kamke-Massera-Funktion): $\gamma \in \mathcal{C}^0 : [0,a) \to \mathbb{R}_{\geq}$, streng wachsend, $\gamma(0) = 0$ Beispiel: $\gamma_1(r) = \ln \frac{1}{1-r}$ mit $a = 1$; $\gamma_2(r) = 1 - \mathrm{e}^{-r}$ (beschränkt) Regeln: $\gamma_1(.) + \gamma_2(.), c\gamma(.); c > 0, \gamma_1(\gamma_2(.)), \max\{\gamma_1, \gamma_2\}, \min\{\gamma_1, \gamma_2\} \in \mathcal{K}$ $\gamma \in \mathcal{K}$ auf $[0,a) \Rightarrow \exists! \gamma^{-1} \in \mathcal{K}$ auf $[0, \lim_{r \to a} \gamma(r))$ Sofern a nicht spezifiziert ist, wird $a = \infty$, also $[0,\infty) = \mathbb{R}_{\geq}$ unterstellt.
$\mathcal{K}_\infty$	Klasse-$\mathcal{K}_\infty$-Funktion: $\gamma \in \mathcal{K} : \mathbb{R}_{\geq} \to \mathbb{R}_{\geq}$, $\lim_{r \to \infty} \gamma(r) = \infty$ (unbeschränkt) Beispiel: $\alpha(r) = r^c, c > 0$
$\mathcal{KL}$	$\beta : \mathbb{R}_{\geq} \times \mathbb{R}_{\geq} \to \mathbb{R}_{\geq}$ mit $\forall r \geq 0 : \beta(r,.) \in \mathcal{K}$ und $\forall t \geq 0 : \beta(.,t) \in \mathcal{L}$ Beispiel: $\beta_1(r,t) = r\mathrm{e}^{-t}, \beta_2(r,t) = \sqrt{\frac{r^2}{1+2tr^2}}$ /[6]
$\mathcal{L}$	Klasse-$\mathcal{L}$-Funktion: $\varphi : \mathbb{R}_{\geq} \to \mathbb{R}_{\geq}$, monoton fallend, $\lim_{t \to \infty} \varphi(t) = 0$ Beispiel: $\varphi(t) = \mathrm{e}^{-\lambda t}, \lambda > 0$
$\mathcal{L}^p; 1 \leq p \leq \infty$	p-fach Lebesgue-integrierbare Funktionen (Lebesgue-Räume)
$\mathcal{L}^\infty$	im Wesentlichen beschränkte Funktionen, d. h. die Werte Unendlich werden auf einer Menge vom Maß Null angenommen
$\mathcal{L}_n^p$	Menge der n-dimensionalen Vektoren mit Elementen aus $\mathcal{L}^p$
$\|\|x(t)\|\|_{\mathcal{L}_n^p}$	$\mathcal{L}_n^p$-Norm: $\|\|x(t)\|\|_{\mathcal{L}_n^p} \stackrel{\text{def}}{=} \begin{cases} (\int_0^\infty \|\|x(t)\|\|_2^p \mathrm{d}t)^{1/p} & p \in [1,\infty) \\ \operatorname{ess\,sup}_{0 \leq t} \|\|x(t)\|\|_2 & p = \infty \end{cases}$
$\mathcal{L}^{pe}; 1 \leq p \leq \infty$	lokal p-fach Lebesgue-integrierbare Funktionen
$\mathcal{L}_n^{\infty e}$	messbare Funktionen $x : [0,\infty) \to (\mathbb{R} \cup \infty)^n$, die lokal, d. h. auf jedem endlichen Teilintervall von $[0,\infty)$, im Wesentlichen beschränkt sind
$\|\|x_{[0,t]}\|\|_{\mathcal{L}_n^{pe}}$	abgeschnittene $\mathcal{L}_n^p$-Norm: $\|\|x_{[0,t]}\|\|_{\mathcal{L}_n^{pe}} \stackrel{\text{def}}{=} \begin{cases} \left(\int_0^t \|\|x(\tau)\|\|_2^p \mathrm{d}\tau\right)^{1/p} & p \in [1,\infty) \\ \operatorname{ess\,sup}_{0 \leq \tau \leq t} \|\|x(\tau)\|\|_2 & p = \infty \end{cases}$
$\mathcal{N}$	Klasse-$\mathcal{N}$-Funktion: $\gamma : \mathbb{R}_{\geq} \to \mathbb{R}_{\geq}$ stetig, nicht fallend
$\mathcal{N}_0$	Klasse-$\mathcal{N}_0$-Funktion: $\gamma \in \mathcal{N}, \gamma(0) = 0$
$x_{[0,t]}$	in t abgeschnittene Signale: $x_{[0,t]} \stackrel{\text{def}}{=} \begin{cases} x(\tau) & \tau \in [0,t] \\ 0 & \tau \in (t,\infty) \end{cases}$

[6] Statt $\mathcal{K}_\infty\mathcal{L}$ wird meist $\mathcal{KL}_\infty$ geschrieben [367]. Zudem werden globale Konzepte vielfach nur mittels $\mathcal{KL}$ definiert, denn: Sei $\beta_0(r) = \beta(r,0)$, so ist $\beta_0 \in \mathcal{K}_\infty$, da $\beta(r,0) \geq r, \forall r \leq \|\|x_0\|\|$ und im globalen Fall $\|\|x_0\|\|$ beliebig groß sein kann. Mithin wird in globalen Stabilitätsdefinitionen die Existenz einer $\mathcal{KL}$-Funktion gefordert und diese enthält $\mathcal{KL}_\infty$ per se. Wird die Globalität an $\mathcal{M}$ statt an $\mathbb{R}^n$ festgemacht, besteht im Fall der Beschränktheit von $\mathcal{M}$ sogar die Chance, eine echte $\mathcal{KL}$-Funktion statt eine $\mathcal{KL}_\infty$ verwenden zu können.

A.6.2 Stabilitätskonzepte

Stabilitätskonzepte für *Ruhelage*[7] $x_e = 0_n$ in t_0 von $\dot{x} = f(x,t); f(0_n,t) \equiv 0_n, \forall t \geq t_0$	
lokale Konzepte: (L) für lokal	
Lagrange-Stabilität in t_0	$\forall \delta > 0,\ \exists \varepsilon(\delta, t_0) > 0 : \|\|x_0\|\| < \delta \Rightarrow \|\|x(t)\|\| < \varepsilon < \infty$
Lyapunov-Stabilität[8] in t_0 (LS)	$\forall \varepsilon > 0,\ \exists \delta(\varepsilon, t_0) > 0 : \|\|x_0\|\| < \delta \Rightarrow \|\|x(t)\|\| < \varepsilon, \forall t \geq t_0$
totale Stabilität (TS) Stabilität unter Dauerstörung[9]	$\forall t_0 \geq 0, \forall \varepsilon > 0, \exists \delta_1(\varepsilon) > 0, \delta_2(\varepsilon) > 0 :$ $\|\|x_0\|\| < \delta_1, \|\|\Delta f(x,t)\|\| < \delta_2 \Rightarrow \|\|x_{f+\Delta f}(t; x_0, t_0)\|\| < \varepsilon$
Attraktivität in t_0 (LATT)	$\exists \delta(t_0) > 0 : \|\|x_0\|\| < \delta \Rightarrow \lim_{t\to\infty} x(t) = 0$
asymptotische Stab. in t_0 (LAS)	Lyapunov-Stabilität in t_0 plus Attraktivität in t_0 (nur für isolierte Ruhelagen sinnvoll)
gleichmäßige lokale Konzepte: Die Variablen δ und T müssen unabhängig von t_0 sein.	
gleichmäßige Lyapunov-Stabilität (ULS); U für uniform	$\forall \varepsilon > 0,\ \exists \delta(\varepsilon) > 0 : \|\|x_0\|\| < \delta \Rightarrow \|\|x(t)\|\| < \varepsilon, \forall t \geq t_0 \geq 0$ $\exists \gamma \in \mathcal{K}, \delta > 0 : \|\|x_0\|\| < \delta \Rightarrow \|\|x(t)\|\| \leq \gamma(\|\|x_0\|\|)$
gleichm. Attraktivität (ULATT)	$\forall c > 0, \exists T(c) > 0 : \|\|x(t)\|\| \leq c\|\|x_0\|\|,\ \forall t \geq t_0 + T, t_0 \geq 0$
ULAS (=ULS+ULATT)	$\exists \beta \in \mathcal{KL}, \delta > 0 : \|\|x_0\|\| < \delta \Rightarrow \|\|x(t)\|\| \leq \beta(\|\|x_0\|\|, t - t_0)$
gleichm. exponentielle S. (ULES)	$\exists \delta, \alpha, c > 0 : \|\|x_0\|\| < \delta \Rightarrow \|\|x(t)\|\| \leq c\,\mathrm{e}^{-\alpha(t-t_0)}\|\|x_0\|\|$ Anm.: $\alpha = 0$ kennzeichnet die Krasovskii-Stabilität.
regionale Konzepte: Angabe des Einzugsbereichs Ω; $x_e = 0$ ist einzige Ruhelage in Ω.	
AS im Großen	Lyapunov-Stabilität plus $x_0 \in \Omega \Rightarrow \lim_{t\to\infty} x(t) = 0$
globale Konzepte: Die x_0-Prämisse der lokalen Konzepte wird durch $\forall x_0 \in \mathcal{M} \subseteq \mathbb{R}^n$ ersetzt und $x_e = 0$ ist einzige Ruhelage in zulässiger Anfangswertmenge $\mathcal{M}$.[10] (G) für global	
Ultimative Beschränktheit Levinson-Dissipativität (L-DISS)	$\forall x_0 \in \mathcal{M} \subseteq \mathbb{R}^n, \exists r > 0 : \overline{\lim_{t\to\infty}} \|\|x(t)\|\| \leq r$ $\forall x_0, \exists r > 0, \exists t_1 \geq t_0 + T(x_0, t_0) : \|\|x(t)\|\| \leq r,\ \forall t \geq t_1$
Globale Attraktivität (GATT)	$\forall x_0 \in \mathcal{M} \subseteq \mathbb{R}^n : \lim_{t\to\infty} \|\|x(t)\|\| = 0$
Globale asymptotische Stabilität	GAS=LS+GATT
gleichmäßige globale Stab. (UGS)	$\forall x_0, \exists \gamma \in \mathcal{K}_\infty : \|\|x(t)\|\| \leq \gamma(\|\|x_0\|\|), \forall t \geq t_0 \geq 0$
gleichmäßige GATT (UGATT)	$\forall x_0, \exists r, T(\varepsilon) > 0 : \|\|x_0\|\| < r \Rightarrow \|\|x(t)\|\| < \varepsilon, \forall t \geq t_0 + T$
gleichmäßige GAS (UGAS)	$\forall x_0,\ \exists \beta \in \mathcal{KL}_\infty : \|\|x(t)\|\| \leq \beta(\|\|x_0\|\|, t - t_0), \forall t \geq t_0 \geq 0$

Tabelle A.12: Interne Stabilitätskonzepte für Standardruhelage

[7] Der Ruhelagenbegriff (konstante Lösung für $t \geq t_0$) kann durch Einbeziehen von Unendlich sowohl bzgl. der Zeit, z. B. über $0_n = \lim_{t\to\infty} f(x_e, t)$ bzw. über Transformation $t := t/(t+1)$ von $[0, \infty)$ auf $[0, 1]$ [542], als auch bzgl. des Zustands etwa durch hemisphärische Transformation [343] erweitert werden.

[8] Die ε-δ-Definition kann als Stetigkeit der Abbildung $x : \mathbb{R}^n \times \mathbb{R}_\geq \to \mathcal{C}^0([0,\infty), \mathbb{R}^n), (x_0, t_0) \mapsto x(t)$, kurzum als Stetigkeit von den Anfangswerten, interpretiert werden.

[9] Das Konzept der totalen Stabilität geht auf Duboshin (1940) [169] und Malkin (1944) [426] zurück.

[10] Global meint nicht zwingend $\mathbb{R}^n$, sondern nur den gesamten Gültigkeitsbereich der Anfangswerte.

Anmerkung A.1 Ist $x(t;x_0,t_0)$ eine Lösung (häufig eine Ruhelage $x(t;x_0,t_0) \equiv x_e$ für $t \geq t_0$), so kann deren Stabilität anhand des über $\tilde{t} = t - t_0$ und $\tilde{x}(\tilde{t}) = x(\tilde{t}) - x(\tilde{t}+t_0;x_0,t_0)$ transformierten Systems (sog. Gleichung der gestörten Bewegung)

$$\dot{\tilde{x}} = f(\tilde{x}+x(\tilde{t}+t_0;x_0,t_0),\tilde{t}+t_0) - f(x(\tilde{t}+t_0;x_0,t_0),\tilde{t}+t_0) =: \tilde{f}(\tilde{x},\tilde{t}); \quad \tilde{f}(0_n,\tilde{t}) \equiv 0_n, \forall \tilde{t} \geq 0 \tag{A.1}$$

und dessen Ruhelage $\tilde{x} = 0_n$ in $\tilde{t} = 0$ untersucht werden. Da nichtkonstante Lösungen auch Bewegungen genannt werden, wird mitunter von Bewegungsstabilität gesprochen.

Stabilitätskonzepte für *invariante Mengen*[11] $\mathcal{S}$ von Systemen $\dot{x} = f(x,t)$	
Lyapunov-Stabilität	$\forall \varepsilon > 0,\ \exists \delta > 0 : \operatorname{dist}(x_0,\mathcal{S}) < \delta \Rightarrow \operatorname{dist}(x(t),\mathcal{S}) < \varepsilon$
Ersetze in Tab. A.12 Norm durch Distanzfunktion zwecks Definition anderer Konzepte.	

Stabilitätskonzepte für *Bewegungen* und *Trajektorien* von Systemen $\dot{x} = f(x,t)$	
Lyapunov-Bewegungsstabilität	$\forall \varepsilon > 0, \exists \delta > 0 : \|\|x_0 - x_1\|\| < \delta \Rightarrow \|\|x(t;x_0) - x(t;x_1)\|\| < \varepsilon$
Zhukovskii-Stabilität[12] starke orbitale Stabilität Stabilität unter Zeitskalierung	$\forall \varepsilon > 0, \exists \delta > 0, \exists \varphi \in \operatorname{Hom} : \mathbb{R}_{\geq} \to \mathbb{R}_{\geq} :$ $\|\|x_0 - x_1\|\| < \delta \Rightarrow \|\|x(\varphi(t);x_0) - x(t;x_1)\|\| < \varepsilon,\ \forall t \geq t_0$
Poincaré-Stabilität orbitale Stabilität[13] bahnstabil, pfadstabil	$\mathcal{T} \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : x = x(t;x_0), t \geq 0\}$ ist eine Trajektorie $\forall \varepsilon > 0,\ \exists \delta > 0 : \|\|x_0 - x_1\|\| < \delta \Rightarrow \operatorname{dist}(\mathcal{T}, x(t;x_1)) < \varepsilon$

Tabelle A.13: Interne Stabilitätskonzepte für Mengen, Bewegungen und Trajektorien

Anmerkung A.2 Stabilität einer Ruhelage meint strenggenommen immer Stabilität der Lösung $x(t) \equiv x_e$. Hat ein System nur eine einzige Ruhelage, so wird gelegentlich lax von der Stabilität des Systems gesprochen.

[11] Eine Menge $\mathcal{S}$ heißt invariante Menge eines Systems, wenn jede Trajektorie, die in $\mathcal{S}$ startet, für alle Zeit in $\mathcal{S}$ bleibt. Beispielsweise sind Ruhelagen, deren Einzugsbereiche oder Orbits invariante Mengen.

[12] Zhukovskii-Stabilität ist nichts anderes als Lyapunov-Stabilität von (zeit)reparametrisierten Lösungen. Zhukovskii entwickelte sein Konzept 1882, also 10 Jahre vor Lyapunov. Fast vergessen, wurde es in jüngster Zeit für die Untersuchung seltsamer Attraktoren wiederbelebt [391].

[13] Obwohl die Bezeichnungen Poincaré-Stabilität und orbitale Stabilität häufig synonym gebraucht werden, empfiehlt es sich, von orbitaler Stabilität nur zu sprechen, wenn es sich um eine abgeschlossene Kurve, also um einen Orbit, handelt. Die Bezeichnung Poincaré-Stabilität schließt gleichermaßen nicht abgeschlossene Kurven mit ein, wenngleich Poincaré nur periodische Orbits betrachtete.

Anmerkung A.3 Die Stabilitätstheoreme für Ruhelagen lassen sich formal auf kompakte Mengen erweitern; für nicht-kompakten Mengen bedarf es größerer Sorgfalt. Stabilität kann auf beliebige Mengen erweitert werden, auch auf solche, die keine Lösungen bzw. Trajektorien der Differenzialgleichung sind bzw. enthalten, z. B. die sog. Grenzmengen [268], [37].

Stabilitätskonzepte für Systeme $\dot{x} = f(x,u,t)$	
Asymptotische Verstärkung (AG) asymptotic gain [576]	$\forall x_0 \in \mathcal{M} \subseteq \mathbb{R}^n, u \in \mathcal{L}_m^\infty, \exists \gamma \in \mathcal{N}_0$: $\limsup_{t\to\infty} \|\|x(t)\|\| \leq \gamma(\|\|u\|\|_{\mathcal{L}_m^\infty})$ $\gamma(.)$ heißt nichtlineare Verstärkung
CICS-Eigenschaft [576] converging input converging state	$\forall x_0 \in \mathcal{M} \subseteq \mathbb{R}^n : \lim_{t\to\infty} u(t) = 0 \Rightarrow \lim_{t\to\infty} x(t) = 0$
CIBS-Eigenschaft converging input bounded state	$\forall x_0, \exists c < \infty : \lim_{t\to\infty} u(t) = 0 \Rightarrow \|\|x(t)\|\| < c, \forall t \geq t_0$
UGAS, wenn $f(0_n,u,t) \equiv 0_n$	$\forall x_0 \in \mathcal{M} \subseteq \mathbb{R}^n, \exists \beta \in \mathcal{KL} : \|\|x(t)\|\| \leq \beta(\|\|x_0\|\|, t-t_0)$
Interpretierbar als robuste Stabilität bez. Störungen (Parameterunsicherheiten), die die Ruhelage nicht ändern.	
Eingangs-Zustands-Stabilität input-to-state stability (ISS) ist ein globales Konzept [367]	$\forall x_0 \in \mathcal{M} \subseteq \mathbb{R}^n, u \in \mathcal{L}_m^{\infty e}, \exists \beta \in \mathcal{KL}, \gamma \in \mathcal{K}$: $\|\|x(t)\|\| \leq \max\{\beta(\|\|x_0\|\|, t-t_0), \gamma(\|\|u_{[t_0,t]}\|\|_{\mathcal{L}_m^{\infty e}})\}$ äquivalent: $\forall x_0 \in \mathcal{M} \subseteq \mathbb{R}^n, \exists \bar{\beta} \in \mathcal{KL}, \bar{\gamma} \in \mathcal{K}$: $\|\|x(t)\|\| \leq \bar{\beta}(\|\|x_0\|\|, t-t_0) + \bar{\gamma}(\|\|u_{[t_0,t]}\|\|_{\mathcal{L}_m^{\infty e}}), \forall t \geq t_0$
Erweiterbar auf attraktive invariante, nicht notwendig kompakte Mengen $\mathcal{S}$, indem $\|\|x(t)\|\|$, $\|\|x_0\|\|$ durch $\mathrm{dist}(x(t),\mathcal{S})$, $\mathrm{dist}(x_0,\mathcal{S})$ ersetzt werden (SET-ISS) [580].	
integral ISS (IISS) [25]	$\exists \beta \in \mathcal{KL}, \gamma_1, \gamma_2 \in \mathcal{K}$: $\|\|x(t)\|\| \leq \beta(\|\|x_0\|\|, t-t_0) + \gamma_1\left(\int_{t_0}^t \gamma_2(\|\|u_{[t_0,t]}\|\|_{\mathcal{L}_m^{\infty e}})\,\mathrm{d}\tau\right)$
BIBS-Stabilität [576] bounded-input-bounded-state	$\forall x_0 \in \mathcal{M} \subseteq \mathbb{R}^n, u \in \mathcal{L}_m^\infty, \exists N(M,x_0,t_0) < \infty$: $\operatorname{ess\,sup}_{t\geq t_0} \|\|u(t)\|\| \leq M < \infty \Rightarrow \operatorname{ess\,sup}_{t\geq t_0} \|\|x(t)\|\| \leq N$
UBIBS-Stabilität	$\forall x_0 \in \mathcal{M} \subseteq \mathbb{R}^n, u \in \mathcal{L}_m^\infty, \exists \sigma_1, \sigma_2 \in \mathcal{N}$: $\|\|x(t)\|\|_{\mathcal{L}_n^\infty} \leq \max\{\sigma_1(\|\|x_0\|\|), \sigma_2(\|\|u(t)\|\|_{\mathcal{L}_m^\infty})\}$

Tabelle A.14: Externe Stabilitätskonzepte für Systeme ohne Ausgang

Stabilitätskonzepte für Systeme $\dot{x} = f(x,u,t); y = h(x,t)$	
$\mathcal{L}^q$-Stabilität; BIBO für $q = \infty$ (Zames-Operatorstabilität)	$\forall x_0 \in \mathcal{M} \subseteq \mathbb{R}^n, t_0 \geq 0 : u \in \mathcal{L}_m^q \Rightarrow y \in \mathcal{L}_p^q$
UBIBO-Stabilität[14]	$\forall u \in \mathcal{L}_m^\infty, \exists c,d < \infty : \operatorname{ess\,sup}_{t \geq t_0} \|\|y(t)\| \| \leq c \operatorname{ess\,sup}_{t \geq t_0} \|\|u(t)\| \| + d$ äquivalent für LTI [577]: $\|\|G(t)\|\|_{\mathcal{L}^1} < \infty$
gleichmäßige $\mathcal{L}^{pe}$-$\mathcal{L}^{qe}$-Stabilität	$\forall u \in \mathcal{L}_m^{pe}, \exists c,d < \infty : \|\|y_{[t_0,t]}\|\|_{\mathcal{L}_p^{qe}} \leq c \|\|u_{[t_0,t]}\|\|_{\mathcal{L}_m^{pe}} + d$ c ist gleichmäßige endliche $\mathcal{L}^{pe}$-$\mathcal{L}^{qe}$-Verstärkung[15]
E/A-Stabilität [575] input-to-output stability (IOS)	$\forall x_0 \in \mathcal{M} \subseteq \mathbb{R}^n, t_0 \geq 0, \forall u \in \mathcal{L}_m^{\infty e}, \exists \beta \in \mathcal{KL}, \gamma \in \mathcal{K} :$ $\|\|y(t)\| \| \leq \beta(\|\|u_{[t_0,T]}\|\|_{\mathcal{L}_m^{\infty e}}, t-T) + \gamma(\|\|u_{[T,\infty)}\|\|_{\mathcal{L}_m^{\infty e}})$ für fast alle Paare $T < t$ mit $t_0 \leq T \leq t$
E-A/Z-Stabilität [401] (IOSS)	$\forall x_0 \in \mathcal{M} \subseteq \mathbb{R}^n, t_0 \geq 0, \forall u \in \mathcal{L}_m^{\infty e}, \exists \beta \in \mathcal{KL}, \gamma_1, \gamma_2 \in \mathcal{K}_\infty :$ $\|\|x(t)\| \| \leq \beta(\|\|x_0\|\|, t-t_0) + \gamma_1(\|\|u_{[t_0,t]}\|\|_{\mathcal{L}_m^{\infty e}}) + \gamma_2(\|\|y_{[t_0,t]}\|\|_{\mathcal{L}_p^{\infty e}})$
A/Z-Stabilität [583] output-to-state stability (OSS), wenn $u \equiv 0_m$	$\forall x_0 \in \mathcal{M} \subseteq \mathbb{R}^n, t_0 \geq 0, \exists \beta \in \mathcal{KL}, \gamma \in \mathcal{K}_\infty :$ $\|\|x(t)\| \| \leq \beta(\|\|x_0\|\|, t-t_0) + \gamma(\|\|y_{[t_0,t]}\|\|_{\mathcal{L}_p^{\infty e}})$ für LTI-Systeme äquivalent zur Detektierbarkeit
A/E-Z-Stabilität output-to-input-and-state stability (OISS) [401]	$\forall x_0 \in \mathcal{M} \subseteq \mathbb{R}^n, \exists k \in \mathbb{N}^>, \beta \in \mathcal{KL}, \gamma \in \mathcal{K}_\infty, \forall u \in \mathcal{C}^{k-1} :$ $\left\| \begin{bmatrix} u(t) \\ x(t) \end{bmatrix} \right\| \leq \beta(\|\|x_0\|\|, t-t_0) + \gamma \left(\left\| \begin{bmatrix} y_{[t_0,t]} \\ \vdots \\ y^{(k)}_{[t_0,t]} \end{bmatrix} \right\|_{\mathcal{L}^{\infty e}_{p(k+1)}} \right)$ Linksstabilität für MIMO-LTI-Systeme Minimalphasigkeit für SISO-LTI-Systeme
Hyperstabilität [20]	$\forall u$ mit $\int_0^t u^T(\tau) y(\tau) \mathrm{d}\tau \leq \delta \|\|x_0\|\| \sup_{0 \leq \tau \leq t} \|\|x(\tau)\|\|, \ \forall t \geq 0$ gilt: $\exists k, \delta > 0 : \ \|\|x(t)\| \| \leq k(\|\|x_0\|\| + \delta)$
asymptotische Hyperstabilität	hyperstabil und $x = 0_n$ ist GAS für $u \equiv 0_m$
Nichtexpansivität Kontraktivität bei strenger Ungl.	$\forall x(t,0_n), \forall t \geq t_0 : \int_{t_0}^t \|\|y(\tau)\|\|_2^2 \mathrm{d}\tau \leq \int_{t_0}^t \|\|u(\tau)\|\|_2^2 \mathrm{d}\tau$ äquivalent für $G(s) : \|\|G(s)\|\|_\infty = \sup_{\Re e\, s > 0} \sigma_{\max}(G(s)) \leq 1$

Tabelle A.15: Externe Stabilitätskonzepte für Systeme mit Ausgang

[14] Werden nur messbare beschränkte Signale zugelassen, kann das Wesentliche Supremum durch das gewöhnliche Supremum ersetzt werden. UBIBO-Stabilität ist gleichbedeutend mit der Beschränktheit des Übertragungsoperators. Im Linearen ist $d = 0$, wenn zusätzlich $x_0 = 0$ angenommen wird; das kleinste c ist dann die Operatornorm des Übertragungsoperators. Im Nichtlinearen ist die additive Konstante auch für $x_0 = 0$ vonnöten, wie $\dot{y} = -y + 1 + u$ zeigt.

[15] Damit $h : \mathcal{L}_n^{\infty e} \to \mathcal{L}_p^{\infty e}, x \mapsto y$ gilt, muss ohnehin $\sup\{h(x) : \|\|x\|\| \leq a\} < \infty, \forall a > 0$ sein. Mit der Zusatzbedingung $h(0_n) = 0$ folgen die $\mathcal{K}$-beschränkte Funktionen, d. h. $h(0_n) = 0$, $\exists \alpha \in \mathcal{K} : \|\|h(x)\| \| \leq \alpha(\|\|x\|\|)$. ISS plus $\mathcal{K}$-beschränktes h ergibt IOS [575].

A.6.3 Implikationen zwischen den Konzepten

Bei LTI-Systemen sind für fast alle Systeme die jeweiligen Stabilitätsdefinitionen zueinander äquivalent (Ausnahme: grenzstabile und neutral stabile Systeme). LTV-Systeme und nichtlineare Systeme können dagegen in dem einen Sinne stabil sein, in einem anderen aber nicht. Implikationen und ggf. Gegenbeispiele helfen das zu untermauern, s. auch [268].

GAS = UGAS	für $\dot{x} = f(x)$ und für $\dot{x} = f(x,t)$ mit $f(x,t+T) = f(x,t)$ $\dot{x} = (6t\sin t - 2t)x$ ist nur GAS, da nicht UGS [345] $\dot{x} = -\frac{x}{t+1}$; $x(t) = \frac{1+t_0}{1+t}x(t_0)$ ist nur GAS, da nicht UGATT [345]
UGAS$\not\Rightarrow$ UGES	$\dot{x} = -x^3$
ISS$\Rightarrow$ UGAS	klar, vgl. Definitionen mit $u \equiv 0$ Äquivalenz für Hurwitz-stabile LTI-Systeme $\bar{\gamma}(\|\|u\|\|_{\mathcal{L}^{\infty e}}) = \|\|B\|\|_2 \int_0^\infty \|\|\mathrm{e}^{A\tau}\|\|_2 \mathrm{d}\tau \, \|\|u\|\|_{\mathcal{L}^{\infty e}}, \bar{\beta}(\|\|x_0\|\|, t) = \|\|\mathrm{e}^{At}\|\|_2 \|\|x_0\|\|$ Beispiel für ISS: $\dot{x} = -ax^k + bx^p\varphi(u), k, p \in \mathbb{N}, p < k, k$ ungerade, $\varphi \in \mathcal{C}^1, \varphi(0) = 0$
UGAS$\not\Rightarrow$ISS	$\dot{x} = -\mathrm{sat}(x) + u$; $u \equiv 2$, $x_0 = 1$, $x(t) = 1 + t$ auch unter Hinzunahme der absoluten Integrabilität von u nicht, denn es existieren UGAS-Systeme $\dot{x} = f(x) + u$ in $\mathbb{R}^2$ mit der Eigenschaft $\forall \varepsilon > 0, \exists u : \|\|u\|\|_{\mathcal{L}^1} < \varepsilon : x(t) \to \infty$ [579]
ISS$\Rightarrow$IISS	nicht umgekehrt, vgl. $\dot{x} = -\arctan x + u$ mit $x_0 = 1, u \equiv \frac{\pi}{2}$
IISS$\Rightarrow$ UGAS	klar, vgl. Definitionen mit $u \equiv 0$
CICS$\not\Rightarrow$ISS	$\dot{x} = -x + xu$, aber ISS$\Rightarrow$CICS
ISS$\Rightarrow$BIBS, CIBS	für zahlreiche weitere Implikationen zu ISS s. [582]
UBIBS$\not\Rightarrow$ISS	$\dot{x} = -(\cos u)^2 x; u \equiv \frac{\pi}{2}$ (System ist neben UBIBS auch UGAS.)
IOS$\Rightarrow$UBIBO UBIBO$\not\Rightarrow$IOS	IOS ist mehr, da zusätzlich $y \to 0$, wenn $u \to 0$ gelten muss [575] $y = u + 1$
UGAS$\not\Rightarrow$CICS	$\dot{x} = -x + x^2 u; u(t) = \mathrm{e}^{-t}1(t), x(t) = \frac{2x_0}{(2-x_0)\mathrm{e}^t + x_0\mathrm{e}^{-t}} \to \infty$ für $x_0 > 2$ Implikation gilt auch für vorwärts vollst. Systeme nicht [198] UGAS ist notwendig für CICS, d. h. CICS$\Rightarrow$ UGAS für LTI-Systeme gilt UGAS$\Rightarrow$CICS
UGAS$\not\Rightarrow$CIBS	$\dot{x} = -x + (x^2+1)u; u(t) = (2t+2)^{-1/2}, x_0 = \sqrt{2}, x(t) = \sqrt{2t+2}$
UBIBO$\Rightarrow$CICO	[577]
UGAS$\not\Rightarrow$UBIBO	$\dot{y} = -y + (y^2+1)u; u \equiv 1$ Für LTI-Systeme gilt UGAS$\Rightarrow$UBIBO, [577]
UBIBO$\not\Rightarrow$ UGAS	$\dot{x}_1 = x_1, \dot{x}_2 = -x_2 + u, y = x_2$ LTI: UBIBO + Minimalität $\Rightarrow$ UGAS

Tabelle A.16: Ausgewählte Implikationen zur Stabilität (Teil 1)

UGES⇒$\mathcal{L}^q$; $q \in [1,\infty]$	unter diversen Abschätzungen für $f(.)$ und $h(.)$ [345]
ISS⇒$\mathcal{L}^\infty$	unter diversen Abschätzungen für $f(.)$ und $h(.)$ [345]
ULES⇒s$\mathcal{L}^q$; $q \in [1,\infty]$	Der Vorsatz „s“ steht für „small-signal“. unter diversen Abschätzungen für $f(.)$ und $h(.)$ [345]
ULAS⇒sUBIBO	Der Vorsatz „s“ steht für „small-signal“. unter diversen Abschätzungen für $f(.)$ und $h(.)$ [345]
ULES⇒ ULAS	Für LTV-Systeme gilt Äquivalenz [543].
$\dot{x} = (\sin\ln(t+1) + \cos\ln(t+1) - a)x$; $x(t) = \mathrm{e}^{(t+1)\sin\ln(t+1) - (t_0+1)\sin\ln(t_0+1) - a(t-t_0)}x(t_0)$ instabil für $a < 1$; LS, nicht ULS, nicht LAS für $a = 1$; LAS und ULS, nicht ULAS für $a = \sqrt{2}$; ULAS für $a > \sqrt{2}$	
TS⇏ ULAS	$\dot{x} = \begin{cases} -x^2 \sin\frac{1}{x} & x \neq 0 \\ 0 & x = 0, \end{cases}$ aber, Implikation gilt für $\dot{x} = Ax$ Umkehrung: ULAS + f gleichm. lok. Lipschitz[16]⇒ TS [442], [340]
$\mathcal{L}^p$-$\mathcal{L}^q$-Stabilität	Für LTI-Systeme siehe [311], [168].
GATT⇏ GAS	Vinograd-Beispiel: instabil, aber global attraktiv[17] $\dot{x}_1 = \frac{x_1^2(x_2 - x_1) + x_2^5}{(x_1^2 + x_2^2)(1 + (x_1^2 + x_2^2)^2)}$ $\dot{x}_1 = 0$ für $x_1 = x_2 = 0$ $\dot{x}_2 = \frac{x_2^2(x_2 - 2x_1)}{(x_1^2 + x_2^2)(1 + (x_1^2 + x_2^2)^2)}$ $\dot{x}_2 = 0$ für $x_1 = x_2 = 0$ Für LTV-Systeme sind GATT und GAS äquivalent.
Ultimative Beschränktheit ⇒ Lagrange-Stabilität, nicht umgekehrt	
Lyapunov-Stabilität ⇒ Zhukovskii-Stabilität ⇒ Poincaré-Stabilität [198] Lyapunov-, Zhukovskii- und Poincaré-Stabilität sind äquivalent für Ruhelagen Zhukovskii-, Poincaré-Stabilität sind äquivalent für periodische Lösungen, wenn $f \in \mathcal{C}^1$ Für Unterschiede s. [198] und [391]	
$x(t) = t + x_0$ zu $\dot{x} = 1$ ist Lyapunov-stabil, aber nicht Lagrange-stabil.	
Für minimale LTI-Systeme sind Hyperstabilität und Passivität bzw. asympt. Hyperstabilität und strenge Passivität äquivalent [20].[18] Für $u \equiv 0_m$ folgt LS bzw. GAS. Asymptotische Hyperstabilität ⇒ BIBO-Stabilität	
Nichtexpansivität ⇒ Lyapunov-Stabilität Kontraktivität ⇒ GAS	

Tabelle A.17: Ausgewählte Implikationen zur Stabilität (Teil 2)

[16] Gleichmäßig lokal Lipschitz-stetig meint lokal Lipschitz-stetig unabhängig von t.

[17] Etwas einfacher ist $\dot{x}_1 = x_1^2 - x_2^2, \dot{x}_2 = 2x_1x_2$, bei dem der Punkt Unendlich in $\mathbb{R}^2$ eingeschlossen wird und $+\infty, -\infty$ als ein Punkt betrachtet werden (Alexandrov-Kompaktifizierung des $\mathbb{R}^2$) [548].

[18] Statt der Begriffspaare „Hyperstabilität, asymptotische Hyperstabilität“ werden gelegentlich die Paare „allgemeine Hyperstabilität, Hyperstabilität“ bzw. „schwache Hyperstabilität, Hyperstabilität“ verwendet.

A.6.4 Stabilitätssätze

Dieser Abschnitt enthält nützliche Hinweise zur Stabilitätsüberprüfung. Gegenbeispiele warnen vor Fehlschlüssen. Die Sätze werden dabei in die folgenden elf Gruppen eingeteilt:

1. Stabilitätscharakterisierungen für LTV-Systeme
2. Lyapunov-basierte Sätze
3. Ableitungsbasierte Sätze für autonome Systeme
4. Ableitungsbasierte Sätze für nichtautonome Systeme
5. Ableitungsbasierte Sätze für nichtkonstante Lösungen
6. Ableitungsbasierte Sätze für $\mathcal{L}^q$-Stabilität
7. Sätze für pseudolineare Systeme
8. Sätze über verschwindende Störterme
9. Sätze für gekoppelte Systeme
10. Abschätzungstheoreme
11. Fixpunkttheoreme

1. Stabilitätscharakterisierungen für LTV-Systeme

Für LTV-Systeme mit $x = 0_n$ als einzige Ruhelage sind alle lokalen Stabilitätsaussagen zugleich global gültig, da $x(t) = \Phi(t, t_0)x(t_0)$ linear vom Anfangszustand abhängt. Somit kann die Stabilität äquivalent über die Transitionsmatrix $\Phi(t, t_0)$ beschrieben werden. Die Aussagen aus [614] sind hier in Tabelle A.18 kompakt zusammengefasst, wobei induzierte Normen[19] zu verwenden sind. Einfacher zu handhabende Charakterisierungen sind für spezielle Systemklassen in Abschnitt 2.8 aufgeführt.

LS	$\sup\limits_{t \geq t_0} \|\|\Phi(t, t_0)\|\| < \infty$
LAS	$\lim\limits_{t \to \infty} \|\|\Phi(t, t_0)\|\| = 0$
LES	$\exists c \geq 1, \exists \lambda > 0, \forall t \geq 0 : \|\|\Phi(t, 0)\|\| \leq c\, \mathrm{e}^{-\lambda t}$ $\Leftrightarrow \varrho_L < 0$ (s. Fußnote 32, [152], [290])
ULS	$\sup\limits_{t_0 \geq 0} \sup\limits_{t \geq t_0} \|\|\Phi(t, t_0)\|\| < \infty$
ULAS = ULES	$\sup\limits_{t_0 \geq 0} \sup\limits_{t \geq t_0} \|\|\Phi(t, t_0)\|\| < \infty$ und $\lim\limits_{t \to \infty} \sup\limits_{t_0 \geq 0} \|\|\Phi(t, t_0)\|\| = 0$ $\exists c \geq 1, \exists \lambda > 0, \forall t \geq t_0 \geq 0 : \|\|\Phi(t, t_0)\|\| \leq c\, \mathrm{e}^{-\lambda (t - t_0)}$ $\Leftrightarrow \varrho_B < 0$ (s. Anm. A.6, [152], [290])

Tabelle A.18: Stabilitätscharakterisierungen für LTV-Systeme

[19] Eine induzierte Norm (Operatornorm, natürliche Matrixnorm) wird durch eine Vektornorm $\|\,.\,\|_{\mathcal{V}}$ über $\|A\|_{\mathcal{V}} \stackrel{\text{def}}{=} \max_{x \neq 0_n} \frac{\|Ax\|_{\mathcal{V}}}{\|x\|_{\mathcal{V}}}$ erzeugt. So werden Spaltensummen-, Spektral- und Zeilensummennorm jeweils durch die l_1-, l_2- und l_∞-Norm induziert; die Frobenius-Norm ist hingegen keine induzierte Norm.

2. Lyapunov-basierte Sätze
Ohne Kenntnis der Lösungen können Stabilitätseigenschaften von Ruhelagen mit dem Lyapunov-Theorem [614], [548], dem LaSalle-Yoshizawa-Theorem [367] oder Modifikationen dieser Theoreme nachgewiesen werden, siehe hierzu Tabelle A.19[20]. Leider geben die Theoreme keine Auskunft darüber, wie entsprechende Lyapunov-Funktionen zu finden sind. Bewährte Methoden hierfür sind die Variable Gradientenmethode (s. auch Ingwerson-Variante [262], [501]), die Krasovskii-Methode [548] und die Chataev-Methode [546]. Weitere Methoden werden in [493], [332] beschrieben. Als nützlich erweist sich in diesem Zusammenhang auch die Eigenschaft, dass die Menge der Lyapunov-Funktionen für eine asymptotisch stabile Ruhelage konvex ist.

Neben den genannten analytischen Zugängen zur Konstruktion von Lyapunov-Funktionen gibt es viele algorithmische Methoden. Für Polynomvektorfelder bieten sich Sum-of-squares-Typ-Ansätze für die Lyapunov-Funktion an. Dabei sind SDP-Zulässigkeitsprobleme zu lösen. Modifikationen ermöglichen in einigen Fällen ($\cos x$, $\sin x$, $\sqrt[n]{x}$, e^x, spezielle Sättigungsfunktionen und rationale Funktionen ohne reelle Pole) die Erweiterung auf nicht-polynomiale Vektorfelder [498], [499]. Ebenfalls auf SDP-Probleme führt die Methode in [10], die ein Lyapunov-Theorem mit höheren Ableitungen nutzt, oder die Methode in [205], die parameterabhängige Lyapunov-Funktionen verwendet.
Einen weiteren Zugang liefert die Carleman-Linearisierung [361]. Für das höherdimensionale approximierende lineare System $\dot{z} = Az$ zu $\dot{x} = f(x)$ kann, wenn $A^TP + PA = -I$ eine positiv definite Lösung P hat, eine Lyapunov-Funktion $V(z) = z^TPz$ mit $z = \phi(x)$ (entsprechend der Approximation) erstellt werden. Umgekehrt sichert das McCann-Theorem, wonach jedes UGAS-System topologisch äquivalent zu einem $2n$-dimensionalen linearen System ist, dass zumindest die unendlichdimensionale lineare Approximation eine (unendlichdimensionale) positiv definite Lösung P hat; endliche Approximationen haben ggf. nur endliche Einzugsbereiche [281].
Vektor-Lyapunov-Funktionen empfehlen sich für große, gering gekoppelte Systeme [380]. Statt einer einzigen Lyapunov-Funktion werden mehrere über eine positive Wichtung kombiniert. An die Stelle des Nachweises der Negativität der Ableitung über die Löwner-Halbordnung tritt der Nachweis der Negativität im Sinne der natürliche Halbordnung für Vektoren, was die Konstruktion der Lyapunov-Funktionen erleichtert.

Die Forderung nach stetiger Differenzierbarkeit der Lyapunov-Funktion lässt sich durch Verwenden der Dini-Rechtsableitung (Molchanov-Pyatinskii-Theorem) [510] abschwächen. Somit können stückweise Lyapunov-Funktionen genutzt werden, z. B. $V(x) = \max\{V_1(x), V_2(x)\}$ [638], um in Verbindung mit LMI-Formulierungen die Einzugsbereiche zu vergrößern. Die

[20] In Tab. A.19 wird aus Platzgründen für die Ableitung der Lyapunov-Funktion $V(x,t)$ entlang der Lösungen $\dot{V}(x,t) := \frac{\mathrm{d}}{\mathrm{d}t}V(x(t;x_0,t_0))|_{t=t_0+0;x_0=x}$ geschrieben, obwohl $L_fV(x,t)$ prägnanter ist.

	Eigenschaft	Bedingung an V	Bedingung an $\dot{V}$
		lokal: $\forall x \in \mathcal{U}_\varepsilon(0_n)\backslash 0_n, \forall t \geq t_0$:	
a	$\dot{x} = f(x)$	$V(0_n) = 0, V \in C^1$	
b	$\dot{x} = f(x,t)$	$V(0_n,t) \equiv 0, V \in C^1$	
1b	stabil	$0 < W_1(x) \leq V(x,t)$	$\dot{V}(x,t) \leq 0$ Lyapunov [345]
2a	gleichmäßig	$0 < V(x)$	$\dot{V}(x) \leq 0$ Lyapunov [345]
2b	stabil	$0 < W_1(x) \leq V(x,t) \leq W_2(x)$	$\dot{V}(x,t) \leq 0$ Malkin [345]
3b	asympt. stabil	$\underbrace{0 < W_1(x) \leq V(x,t)}_{\triangleq \inf_t V(x,t) > 0} \leq W_2(x)$	$\dot{V}(x,t) \leq -g(t)W_3(x) < 0$ $\int_{t_0}^{\infty} g(t)\mathrm{d}t = \infty$ Hahn [268]
		$W_i(0_n) = 0$, $W_i(x) > 0$ $W_i \in C^0$	$\dot{V}(x,t) \leq h(V(x,t),t), h(0,t) \equiv 0$ $\dot{v} = h(v,t)$ ist GAS; $h \in C^0$ comparison theorem [268], [399]
		$0 \leq V(x,t)$; $\|\|f(x,t)\|\| \leq c$	$\dot{V}(x,t) \leq -W_3(x) < 0$ Marachkov [380]
		$0 \leq V(x,t)$; $0 \leq \tilde{V}(x,t)$	$\dot{V}(x,t) \leq -\alpha(\tilde{V}(x,t)); \alpha \in \mathcal{K}$ Salvadori $\dot{\tilde{V}}(x,t) \leq c$ o. $-\dot{\tilde{V}}(x,t) \leq c$ [519]
		$w(x,t)W_1(x) \leq V(x,t)$ $\lim_{t\to\infty} w(t,x) = \infty$	$\dot{V}(x,t) \leq 0$ [399]
4a	gleichmäßig asymp. stabil	$0 < V(x)$	$\dot{V}(x) < 0$ Lyapunov [345]
			$\dot{V}(x) \leq 0$ Barbashin-Krasovskii [345]
			$\nexists x(t) \in \{x : \dot{V}(x) = 0\} : x(t) \not\equiv 0_n$ [358]
			$V^{(2k+1)}(x) < 0 \; \forall x \neq 0 : \dot{V}(x) = 0$ $V^{(i)}(x) = 0$ für $i = 2,3,\ldots,2k$ [467]
		$0 \leq V(x)$	$\dot{V}(x) \leq 0$; UAS bez. $\{x : \dot{V}(x) = 0\}$ Kalitine [122]
4b		$0 < W_1(x) \leq V(x,t) \leq W_2(x)$	$\dot{V}(x,t) \leq -W_3(x)$ LaSalle-Yoshizawa
			$\exists T : V(x(t+T), t+T) - V(x(t),t) \leq -W_3(x(t))$ Narendra-Annaswamy [472], [7]
		$0 < W_1(x) \leq V_1(x,t) \leq W_2(x)$ $\max\{\|V_2(x,t)\|, \|f(x,t)\|\} \leq c$	$\dot{V}_1(x,t) \leq -W_4(x) \leq 0$ Matrosov $\dot{V}_2(x,t) \gneqq 0$ auf $\{x : W_4(x) = 0\}$ [425]
5	gleichmäßig exp. stabil	$k_1\|\|x\|\|^2 \leq V(x,t) \leq k_2\|\|x\|\|^2$ $\|\|\nabla_x V(x,t)\|\| \leq k_4\|\|x\|\|$	$\dot{V}(x,t) \leq -k_3\|\|x\|\|^2$; $k_i > 0$ Krasovskii [548]
6	instabil	$V(x,t) \leq W_2(x)$	$\dot{V}(x,t) \geq W_3(x) > 0$ Lyapunov [614]
		$\exists x_0 \in \mathcal{U}_\delta\backslash 0_n : V(x_0,t_0) > 0$	$\dot{V}(x,t) = k_3 V(x,t) + w(x,t)$ $w(x,t) \geq 0$, $k_3 > 0$ Lyapunov [614]
7	Globalität	$\forall x \neq 0_n, \forall t \geq t_0$: V-Bedingung	$\forall x \neq 0_n, \forall t \geq t_0$: $\dot{V}$-Bedingung
		zzgl. $\lim_{\|\|x\|\|\to\infty} V(x) = \infty$ bzw. $\lim_{\|\|x\|\|\to\infty} W_1(x) = \infty$	[614]

Tabelle A.19: Stabilitätstheoreme vom Lyapunov-Typ

Forderung nach negativer Definitheit der Ableitung der Lyapunov-Funktion lässt sich ebenfalls abschwächen, was allerdings die Hinzunahme anderer Bedingungen erfordert, s. das LaSalle-Theorem [345] oder die Theoreme in Tabelle A.19. Letztlich ist im AS-Fall zu zeigen, dass die Menge der Punkte, in denen die Ableitung der Lyapunov-Funktion verschwindet, keine invariante Menge außer dem Nullpunkt enthält. Eine Erweiterung des Invarianzprinzips auf periodische Systeme gelingt formal, wohingegen der allgemeine zeitvariante Fall schwieriger ist, da bei diesen Systemen positive Grenzmengen nicht notwendigerweise invariant sind. Entsprechende Sätze fußen deshalb auf dem Barbălat-Lemma[21] [548], der Methode der Grenzgleichungen [31], [386], dem Matrosov-Theorem [444], [413] oder für eine spezielle Klasse auf einem Ergebnis vom Zentrumsmannigfaltigkeit-Typ [43].
Analog zu den Stabilitätstheoremen vom Lyapunov-Typ gibt es auch Theoreme für den Nachweis der Instabilität, s. Tabelle A.19 oder das Chataev-Theorem [548].
Alle vorgenannten Theoreme sind nur hinreichend. Es gibt lediglich Umkehrtheoreme (Persidskii für LS; Massera für LAS, LES; Corless für ULES; Zubov für ULS; Kurcveĭl für ULAS[22]) [595], [455], [37], die die Existenz einer Lyapunov-Funktion garantieren.

Lässt sich keine globalgültige Lyapunov-Funktion finden, so ist die Ruhelage möglicherweise nicht alleinige invariante Menge im Zustandsraum. Andere Ruhelagen lassen sich per Ruhelagenanalyse schnell ausschließen. Periodische Orbits oder gar kompliziertere Mengen sind schwieriger zu detektieren. Speziell für den Nachweis eines periodischen Orbits in $\mathbb{R}^2$ gibt es das Poincaré-Bendixson-Theorem [292]; zum Ausschluss das Bendixson- und das Dulac-Theorem [265], s. [186] für eine Übersicht zu Sätzen für $n > 2$.

Erweiterungen: Stillschweigend wurde bisher die Eindeutigkeit der Lösungen vorausgesetzt. Das Konzept der globalen starken Lyapunov-Stabilität (GSS) erweitert GAS, indem alle (nicht notwendig eindeutigen) Lösungen nach 0_n konvergieren müssen. Hierfür existiert ein Satz von Kurcveĭl [370], [522]. Im dualen Lyapunov-Konzept nach Rantzer [526] geht es hingegen um die Konvergenz fast aller Lösungen gegen eine Nullruhelage.
Zwischen Stabilität und asymptotischer Stabilität ist die Semistabilität einzuordnen. Sie findet Anwendung bei Systemen mit einem Ruhelagenkontinuum und fordert, dass jede Lösung, die in einer Umgebung einer stabilen Ruhelage startet, zu einer (möglicherweise anderen) stabilen Ruhelage strebt, die ihrerseits stetig von den Anfangswerten abhängt.
Die Konzepte der semiglobalen bzw. praktischen asymptotischen Stabilität sind zweckmäßige Abschwächungen. Im ersten Fall wird analysiert, ob ein parametrisiertes System (mit Einstellparameter) eine beliebige Ausdehnung des Einzugsbereichs gestattet; im zweiten, ob eine Konvergenz beliebig nah zur Ruhelage gelingt.

---

[21] Zugänge diesen Typs zeigen gemeinhin GAS, aber nicht UGAS.

[22] Für $f \in \mathcal{C}^0$ impliziert UAS die Existenz einer glatten Lyapunov-Funktion; Lyapunov-Stabilität impliziert dies aber nicht.

Im Konzept der bedingten Stabilität wird die zulässige Variation der Anfangswerte auf eine Menge $\mathcal{S}$ eingeschränkt. Danach ist die instabile Ruhelage $x = 0_2$ von $\dot{x}_1 = x_1; \dot{x}_2 = -x_2$ auf $\mathcal{S} = 0 \times \mathbb{R}$ asymptotisch stabil [400]. Bezieht sich $\mathcal{S}$ direkt auf eine Teilmenge der Zustandskoordinaten, so wird auch von partieller Stabilität gesprochen, die mit dem Rumyantsev-Theorem behandelt werden kann [545]. Anwendung finden diese Konzepte, wenn bestimmte Zustände praktisch irrelevant sind (z. B. Integratoren bei Transformation auf Zeitinvarianz oder bei Pfadregelproblemen, nichtverschwindende Parameterfehler bei aber asymptotisch stabilen Regelfehlern in adaptiven Systemen). Letztlich sei das Konzept der Stabilität bezüglich einer Funktion genannt, das bereits von Lyapunov eingeführt wurde und Bedeutung für die Zustands-Ausgangs-Darstellung hat. Die betrachtete Funktion ist dann gerade $y = h(x)$.

Andersartige Erweiterungen beziehen sich wegen unstetiger rechter Seiten auf verallgemeinerte Lösungen von $\dot{x} = f(x,t); x(t_0) = x_0$ [119], z. B. geschrieben als Integralgleichung

$$x(t) = x_0 + \int_{t_0}^{t} f(x(\tau), \tau)\, \mathrm{d}\tau. \tag{A.2}$$

Der Vorteil der Darstellung (A.2) ist, dass abzählbar viele Unstetigkeiten von f beim Integrieren keine Rolle spielen. Da die Integration zudem glättet, sind die über (A.2) definierten Carathéodory-Lösungen absolutstetig und erfüllen die Differenzialgleichung fast überall. Eine Alternative sind die Euler-Lösungen, die auf einer Glättung der rechten Seite beruhen (stückweise affine Approximationen von f). Letztlich seien noch die Krasovskii- bzw. Filippov-Lösungen genannt, die das Prinzip des Differenzialeinschlusses nutzen (mengenwertige Betrachtung an $\dot{x} \in F(x,t)$). Die Filippov-Lösungen finden bei der Modellierung von Haft- und Gleitreibung sowie in Sliding-Mode-Reglern Anwendung.

Die Stabilitätsbetrachtung für nichtlineare differenzial-algebraischen Systeme erfolgt analog zum Linearen [67], [206] meist durch eine Aufspaltung der Komponenten in jene, die der Dynamik unterliegen, und jene, die durch die algebraische Gleichung fixiert werden [453]. Als hilfreiches Werkzeug in der Lyapunov-Theorie aber auch bei der Exakten Linearisierung erweist sich dabei für Systeme in semi-expliziter Form $\dot{x} = f(x,z)$ und $0 = g(x,z)$ die M-Ableitung

$$M_f V(x,z) = \left(\frac{\partial V}{\partial x^T} - \frac{\partial V}{\partial z^T} \left(\frac{\partial g}{\partial z^T} \right)^{-1} \frac{\partial g}{\partial x^T} \right) \cdot f(x,z), \tag{A.3}$$

die letztlich die Rolle der Lie-Ableitung $L_f V(x)$ übernimmt [620].

Bei Systemen mit Totzeit sind statt Lyapunov-Funktionen sog. Lyapunov-Krasovskii-Funktionale zu verwenden, da numehr Anfangswertfunktionen auftreten. Der einhergehenden schwierigeren Behandlung kann häufig über das Razumikhin-Theorem ausgewichen werden, welches Funktionen und Bedingungen nutzt, die denen vom Lyapunov-Typ ähneln [258].

Einzugsbereich: Lokale asymptotische Stabilität ist eine für Anwendungen unzureichende Eigenschaft. Vielmehr ist regionale Stabilität vonnöten, die die asymptotische Stabilität unter Angabe des exakten Einzugsbereichs $\Omega = \{x_0 \in \mathbb{R}^n : \lim_{t\to\infty} x(t; x_0, t_0) \to 0_n\}$ oder einer

Approximation bezeichnet [367]. Ω ist eine offene, einfach zusammenhängende[23], invariante Menge. Der Rand von Ω ist ebenfalls eine invariante Menge, die von Trajektorien geformt wird [37], [595][24]. Eine Lyapunov-Funktion $V_m(x)$, die den exakten Einzugsbereich liefert, heißt maximale Lyapunov-Funktion [609].[25] Sie erfüllt zu den Eigenschaften einer Lyapunov-Funktion zusätzlich $\lim_{x \to \mathrm{bd}\Omega} V(x) = \infty$. Diese Eigenschaft ist das Pendant zur globalen Unbeschränktheit bei $\Omega = \mathbb{R}^n$. Bis auf wenige Ausnahmen lassen sich $V_m(x)$ und letztlich auch Ω nicht geschlossen berechnen. Zum einen gibt es Einzugsbereiche von fraktaler Struktur, wenn beispielsweise mehrere Attraktoren im Spiel sind, zum anderen erfordert $V_m(x)$, eine partielle Differenzialgleichung zu lösen (Zubov-Theorem [652], [268])

$$\frac{\partial v(x)}{\partial x^T} f(x) = -h(x)(1 - v(x)). \tag{A.4}$$

Dabei ist $h(x)$ eine positiv definite Funktion, die so gewählt wird, dass sich (A.4) gut lösen lässt. Das Urbild von $v(x)$ ist der exakte Einzugsbereich $\Omega = \{x \in \mathbb{R}^n : 0 \leq v(x) < 1\}$, und es gilt $V_m(x) = -\ln(1 - v(x))$. Nach dem Zubov-Theorem existieren also maximale Lyapunov-Funktionen, die ihrerseits von der Wahl von $h(x)$ abhängen. Sollte f nicht die Existenz einer Lösung für alle x_0 garantieren (etwa bei endlicher Fluchtzeit) oder f auf $\mathbb{R}^n$ nicht überall stetig, definiert und beschränkt sein, ist die rechte Seite in (A.4) um den Faktor $\sqrt{1 + ||f(x)||^2}$ zu ergänzen, der aus einer Zeitskalentransformation stammt [257]. Letztlich ergeben sich vereinfachte Berechnungen für $V_m(x)$, wenn Ω bekannt ist [652].

Da (A.4) nur selten geschlossen lösbar ist, wird auf Reihenentwicklungen [609] oder Modifikationen der Differenzialgleichung [217] ausgewichen. Alternativ wird der Einzugsbereich direkt über parametrisierte Lyapunov-Funktionen per Optimierung bestimmt. Dabei wird ausgenutzt, dass die offenen Subniveaumengen $\mathcal{L}_\alpha^- = \{x \in \mathbb{R}^n : V(x) < \alpha\}$ mit $\forall x \in \mathcal{L}_\alpha^- \backslash 0_n : \dot{V}(x) < 0$, innere Approximationen des Einzugsbereichs darstellen, wenn sie einfach zusammenhängend und beschränkt sind und natürlich den Nullpunkt umschließen. Bewährt haben sich jene Verfahren, die auf SDP-Formulierungen führen [131]. Eine gänzlich andere Idee verfolgen Henrion und Korda [278], die die Einzugsbereichsberechnung in ein unendlichdimensionales lineares Optimierungsproblem auf dem Kegel der Borel-Maße umformulieren. Dank endlichdimensionaler LMI-Relaxationen können somit gute Approximationen des Einzugsbereichs erhalten werden.

Einzugsbereiche für nichtautonome Systeme ohne Eingänge können über den Umweg ordnungserhaltender[26] oder ordnungserhöhender Autonomisierungen erhalten werden. Bei einer

[23] Eine wegzusammenhängende Menge heißt einfach zusammenhängend, wenn jede geschlossene Kurve stetig zu einem Punkt geschrumpft werden kann. Im $\mathbb{R}^2$ darf die Menge keine „Löcher“ haben.

[24] Der Rand kann auch vollständig aus Ruhelagen bestehen, vgl. $\dot{x}_i = (-a + |x_1| + |x_2|)x_i$ [257].

[25] Für $\dot{x}_1 = -x_1 + 2x_1^2 x_2$; $\dot{x}_2 = -x_2$ gilt $\Omega = \{x \in \mathbb{R}^2 : x_1 x_2 < 1\}$ und $V_m(x) = \frac{x_1^2}{1 - x_1 x_2} + x_2^2$ [268].

[26] Für $\dot{x} = A(t)x$ bewerkstelligt $x = T(t)z$ mit $\dot{T}(t) = A(t)T(t) - T(t)\tilde{A}$ die Transformation auf das autonome System $\dot{z} = \tilde{A}z$, siehe [362] für Hinweise zur Konstruktion von $T(t)$. Zur Autonomisierung linearer und nichtlinearer Differenzialgleichungen empfehlen sich Kummer-Liouville-Transformationen [69].

Ordnungserhöhung sollte dabei keine klassische Integratorerweiterung gewählt werden[27], sondern eine Zeittransformation vom Fowler-Typ (z. B. $t = -\ln \tau^2$), sodass auch der erweiterte Systemzustand τ asymptotisch stabil ist [219]. Erweiterungen auf nichtautonome Systeme unter Störungen oder mit Eingängen (Konstruktion von Control-Lyapunov-Funktionen) werden in [116] betrachtet.

3. Ableitungsbasierte Sätze für autonome Systeme
ULES kann für $\dot{x} = f(x); f(0_n) = 0_n$ und stetig differenzierbares f in $\mathcal{U}_\varepsilon(0_n)$[28] nach Lyapunovs indirekter Methode [614] im endlichdimensionalen Fall über die Linearisierung

$$\Delta\dot{x} = A\Delta x \qquad \text{mit } A = \left.\frac{\partial f}{\partial x^T}\right|_{x=0_n} = \mathrm{D}f(0_n) \tag{A.5}$$

geschlossen werden, wenn A Hurwitz-stabil ist.[29] In diesem Fall gilt auch die Umkehrung, d. h. ULES von $\dot{x} = f(x)$ impliziert ein Hurwitz-stabiles A [548]. Das Hartman-Grobman-Theorem [548], [596], [275] liefert anhand von A bei hyperbolischen Ruhelagen (keine Eigenwerte auf imaginärer Achse) noch weitergehende Aussagen über das Verhalten um die Ruhelage. Im Fall nicht-hyperbolischer Ruhelagen ist keine Aussage möglich, vgl. $\dot{x} = x^3$ mit instabiler und $\dot{x} = -x^3$ mit stabiler Ruhelage, aber gemeinsamer Ruhelagenlinearisierung $\Delta\dot{x} = 0$. In solchen Situationen hilft die Methode der Zentrumsmannigfaltigkeit weiter [548].

Globale Aussagen lassen sich aus nichtkonstanten Jacobi-Matrizen nicht ableiten, selbst dann nicht, wenn sie für alle x Hurwitz-stabil sind.

Markus-Yamabe-Vermutung [437], (1960):
$\dot{x} = f(x)$, $f(0_n) = 0_n$ mit $f \in \mathcal{C}^1$ ist GAS, wenn $\Re e\, \lambda_i\,(\mathrm{D}f(x)) < 0$ für alle x gilt.

Beispiel A.8 (Gegenbeispiel zur Markus-Yamabe-Vermutung [138], (1997))

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix} = \begin{bmatrix}-x_1 + x_3(x_1 + x_2x_3)^2\\ -x_2 - (x_1 + x_2x_3)^2\\ -x_3\end{bmatrix}, \quad \mathrm{D}f(x) = \begin{bmatrix}-1 + x_3p(x) & x_3^2p(x) & x_2x_3p(x) + \frac{1}{4}p^2(x)\\ -p(x) & -1 - x_3p(x) & -x_3p(x)\\ 0 & 0 & -1\end{bmatrix}$$

mit $p(x) = 2(x_1 + x_2x_3)$ zeigt, dass $\lambda_{1,2,3}(\mathrm{D}f(x)) = -1$ für alle $x \in \mathbb{R}^n$ gilt. Die Lösung $x_1(t) = 18\mathrm{e}^t$, $x_2(t) = -12\mathrm{e}^{2t}$, $x_3(t) = \mathrm{e}^{-t}$ ist unbeschränkt.

Anmerkung A.4 Die Markus-Yamabe-Vermutung ist für $n = 1$ und $n = 2$ richtig, wobei Beweise für den letzteren Fall unabhängig in [189], [227] und [263] erbracht wurden. Sie ist

[27] Bei klassischer Integratorerweiterung erfolgt die Zeittransformation a posteriori durch das Einbeziehen von Unendlich als Ruhelage, siehe dazu Fußnote 7.

[28] Mitunter wird die Forderung, dass f stetig differenzierbar sein muss, vergessen. Weder die Existenz der Jacobi-Matrix (Schema der partiellen Ableitungen) noch die Existenz der Fréchet-Ableitung reichen aus, da diese Punktkonzepte keine Lipschitz-Stetigkeit in der Umgebung implizieren.

[29] Ein analoges Theorem gilt für unendlichdimensionale Systeme [118].

auch richtig, wenn $\mathrm{D}f(x)+(\mathrm{D}f(x))^T$ eine Hurwitz-stabile Matrix ist [274]. Selbstverständlich ist sie nur hinreichend, vgl. $\dot{x} = -x|x|$. Überdies können die Eigenwerte der Jacobi-Matrix in bestimmten Gebieten sogar positiv sein. So sind sie für das GAS-System $\dot{x} = -4x|x-1| - x$ im Intervall $(\frac{5}{8}, 1)$ positiv. Das zeigt einmal mehr, dass die Eigenwerte der Jacobi-Matrix zunächst nur etwas darüber aussagen, wie sich benachbarte Punkte im Phasenraum relativ zueinander(!) bewegen.

Anmerkung A.5 Das Pendant zur Marcus-Yamabe-Vermutung ist im Diskreten die LaSalle-Vermutung [374], wonach aus einer für alle x Schur-stabilen Jacobi-Matrix auf GAS geschlossen werden kann. Auch diese Vermutung ist falsch [440], und zwar sogar für $n \geq 2$. Eingeschränkt auf Polynomvektorfelder gilt sie für $n \leq 2$, nicht aber für $n > 2$. Eine allgemeingültige Ausnahme bilden radiale $\mathrm{D}f(x)$ (Spektralradius und -norm sind gleich). Für Polynome beliebiger Ordnung gilt die Verschärfung der LaSalle-Vermutung, wonach Schur-Stabilität von $\mathrm{absD}f(x)$ für alle x GAS impliziert [139].

4. Ableitungsbasierte Sätze für nichtautonome Systeme

Beim Anwenden von Lyapunovs indirekter Methode auf Systeme $\dot{x} = f(x,t); f(0_n,t) \equiv 0_n$ ist Sorgfalt geboten. Das beginnt mit einer „gleichmäßigen“ Linearisierung[30]

$$\Delta\dot{x} = A(t)\Delta x \qquad \text{mit} \qquad A(t) = \left.\frac{\partial f(x,t)}{\partial x^T}\right|_{x=0_n} \text{ und } \lim_{||x||\to 0} \sup_{t>0} \frac{||f(x,t) - A(t)x||}{||x||} = 0. \quad \text{(A.6)}$$

Die Bedingung (A.6) verhindert ein unbeschränktes Wachstum des Restterms und somit seine Dominanz über den Linearteil. Ohne (A.6) entstünden Fälle wie $\dot{x} = -x + \mathrm{e}^t x^2$ mit der Lösung $x(t) = 1/(\mathrm{e}^{t-t_0}\frac{1}{x(t_0)} - (t-t_0)\mathrm{e}^t)$ (endliche Fluchtzeit $t = \frac{1}{x_0}$ für $t_0 = 0$), in denen trotz stabilem Linearteil die Lösung instabil ist. Natürlich kann der Rest auch konform mit dem Linearteil wirken, z. B. bei $\dot{x} = -x - tx^3$, weshalb (A.6) für abgeleitete Stabilitätsaussagen nur hinreichend ist. Eine Abschwächung von (A.6) etwa auf eine gleichmäßige Lipschitz-Bedingung an f gelingt nicht [41].

Die angesprochene Sorgfalt ist ein weiteres Mal vonnöten, wenn nämlich aus der Stabilität der Linearisierung $\Delta\dot{x} = A(t)\Delta x$ nach (A.6) mit $\sup_{t\in[0,\infty)} |A(t)| < \infty$ auf die des Originals $\dot{x} = f(x,t)$ geschlossen werden soll. So impliziert gleichmäßige asymptotische Stabilität der Ruhelage $\Delta x = 0_n$, dass $x(t) \equiv 0_n$ gleichmäßig lokal asymptotisch stabil ist [548]. Umgekehrt kann aus ULES (ULAS) von $x(t) \equiv 0_n$ auf UGES (UGAS) der Linearisierung geschlossen werden [548], [203]. Ferner sichert exponentielle Dichotomie[31], dass die Stabilitätseigenschaft

[30] Die Bedingung besagt, dass der Rest schneller als x und zudem unabhängig von t nach null strebt. Sie kann mit $g(x,t) = f(x,t) - A(t)x$ auch als $\forall t : ||g(x,t)|| \leq h(x)||x||, \lim_{x\to 0_n} h(x) = 0$ oder als $\forall t \geq 0, x \in \mathcal{U}_\varepsilon(0_n) : ||g(x,t))|| \leq c||x||^p, c > 0, p > 1$ geschrieben werden.

[31] Exponentielle Dichotomie stellt die Erweiterung der hyperbolischen Ruhelagen auf nichtautonome Systeme dar. Sie fordert eine Aufspaltung in eine lineare Mannigfaltigkeit von Lösungen, die exponentiell nach null, und eine, die exponentiell nach unendlich strebt [143].

der Nulllösungen $\Delta x(t) \equiv 0_n$ und $x(t) \equiv 0_n$ gleichlautend ist [143]. Auf die Gleichmäßigkeit bzw. die exponentielle Dichotomie kann ohne zusätzliche Einschränkungen nicht verzichtet werden! Eine solche Einschränkung ist die auf Lyapunov zurückgehende Forderung nach einer regulären linearen Approximation. Ist sie erfüllt, darf asymptotische Stabilität bzgl. $x(t)$ gefolgert werden, wenn der größte Lyapunov-Exponent ϱ_L des Δx-Systems negativ ist. Die Forderung nach Regularität i. S. von Lyapunov ist dabei substanziell, wie das legendäre Perron-Beispiel A.9 belegt.

Beispiel A.9 (Perron-Effekt [505], (1930))
Als Perron-Effekt wird der Vorzeichenwechsel des größten Lyapunov-Exponenten[32] des Originalsystems (A.7) gegenüber dem seiner Linearisierung (A.8) bezeichnet, wodurch beide ein lokal unterschiedliches Stabilitätsverhalten aufweisen. Das System

$$\dot{x}_1 = -a x_1 \tag{A.7a}$$
$$\dot{x}_2 = (\sin\ln(t+1) + \cos\ln(t+1) - 2a)x_2 + x_1^2 \tag{A.7b}$$

mit $1 < 2a < 1 + \frac{1}{2}\mathrm{e}^{-\pi}$ hat die Lösung

$$x_1(t) = \mathrm{e}^{-a(t-t_0)} x_1(t_0)$$
$$x_2(t) = \mathrm{e}^{(t+1)\sin\ln(t+1) - (t_0+1)\sin\ln(t_0+1) - 2a(t-t_0)} \left(x_2(t_0) + x_1^2(t_0) \int_{t_0}^{t} \mathrm{e}^{-(\tau+1)\sin\ln(\tau+1)} \mathrm{d}\tau \right),$$

aus der die Lyapunov-Exponenten $LE_1 = 1 + \frac{1}{2}\mathrm{e}^{-\pi} - 2a > 0$ und $LE_2 = -a$ berechenbar sind [393]. Dabei signalisiert LE_1 Instabilität. Die Linearisierung von (A.7) lautet

$$\Delta\dot{x}_1 = -a\Delta x_1 \tag{A.8a}$$
$$\Delta\dot{x}_2 = (\sin\ln(t+1) + \cos\ln(t+1) - 2a)\Delta x_2. \tag{A.8b}$$

$$\Delta x(t) = \Phi(t,t_0)\Delta x(t_0) = \begin{bmatrix} \mathrm{e}^{-a(t-t_0)} & 0 \\ 0 & \mathrm{e}^{(t+1)\sin\ln(t+1) - (t_0+1)\sin\ln(t_0+1) - 2a(t-t_0)} \end{bmatrix} \begin{bmatrix} \Delta x_1(t_0) \\ \Delta x_2(t_0) \end{bmatrix}$$

liefert direkt $LE_1 = 1 - 2a < 0$ und $LE_2 = -a$, weshalb $\Delta x(t) \equiv 0_2$ asymptotisch stabil ist, allerdings nicht UGAS.[33]

---

[32] Beachte: Die Begriffsbezeichnungen sind in der Literatur nicht einheitlich. Sei $z(t)$ eine Lösung von $\dot{z} = A(t)z$, so heißt $CE(z(t)) = \limsup_{t\to\infty} \frac{1}{t} \ln ||z(t)||_2$ charakteristischer Exponent dieser Lösung (Der negative CE entspricht der von Lyapunov eingeführten charakteristischen Zahl.). Als Lyapunov-Exponenten werden die Zahlen $LE_i = \limsup_{t\to\infty} \frac{1}{t-t_0} \ln \sigma_i(\Phi(t,t_0))$ bezeichnet, wobei $\sigma_i(.)$ die Singulärwerte der Fundamentalmatrix $\Phi(t,t_0)$ sind. Für LTV-Systeme stimmen der größte charakteristische Exponent $\varrho_C = \sup_{z_0\in\mathbb{R}^n} CE(z(t;\Delta z_0))$ und der größte Lyapunov-Exponent $\varrho_L = LE_1$ überein. Regularität eines LTV-Systems liegt vor, wenn $\sum_{i=1}^{n} CE_i(z_i(t)) = \liminf_{t\to\infty} \frac{1}{t} \ln|\det\Phi(t,0)|$ für ein normales Fundamentalsystem $\{z_i(t); i = 1,\ldots,n\}$ gilt. Die Lyapunov-Exponenten für ein LTI-System sind die Realteile der Eigenwerte und für ein periodisches LTV-System die Realteile der Floquet-Exponenten.

[33] Wegen $||\Delta x(t)||_2 \le c\,\mathrm{e}^{-\lambda(t-t_0)}||\Delta x(t_0)||_2$ gilt ULES und damit auch ULAS. Mit $e^{(t+1)\sin\ln(t+1)} \le \mathrm{e}^{t+1}$, $\mathrm{e}^{-(t_0+1)\sin\ln(t_0+1)} \le \mathrm{e}^{t_0+1}$ folgt $|\Delta x_2(t)| \le \mathrm{e}^{2(t_0+1)}\mathrm{e}^{(1-2a)(t-t_0)}|\Delta x_2(t_0)|$, was zeigt, dass $c = \mathrm{e}^{2(t_0+1)}$ nicht unabhängig von t_0 ist.

Der Perron-Effekt kann auch entgegengesetzt wirken, also mit instabilem Δx-, aber stabilem x-System [393]. Das widerlegt die Aussage, wonach eine Lösung bzw. ein System instabil sei, sofern der größte Δx-System-Lyapunov-Exponent positiv ist.

Anmerkung A.6 Wird statt des größten Lyapunov-Exponenten der Bohl-Exponent

$$\varrho_B = \sup_{\Delta x(t_0)\neq 0_n} \limsup_{\substack{t\to\infty \\ t-\tau\to\infty}} \frac{\ln ||\Phi(t,t_0)\Delta x(t_0)||_2 - \ln ||\Phi(\tau,t_0)\Delta x(t_0)||_2}{t-\tau}$$

von $\Delta\dot{x} = A(t)\Delta x$ herangezogen, so kann aus dessen Negativität auf ULES von $\dot{x} = f(x,t)$ geschlossen werden [152]. Für Beispiel A.9 ist $\varrho_B = \sqrt{2} - 2a$. Die Stabilitätsanforderung an a aus $\varrho_B < 0$ ist dabei klar konservativer als die des x-Systems mit $\varrho_L = 1 + \frac{1}{2}\mathrm{e}^{-\pi} - 2a < 0$. Es gilt $\varrho_L \leq \varrho_B \leq M$ mit $||A(t)||_2 \leq M$ und $\varrho_L = \varrho_B = \max_i \Re e\, \lambda_i(A)$ für $A(t) \equiv A$.
Mit dem Bohl-Exponent lässt sich eine notwendige Bedingung für chaotischer Lösungen formulieren, während deren Beschränktheit und $\varrho_L > 0$ dafür hinreichend ist [198].

Anmerkung A.7 Ist $A(t)$ kinematisch ähnlich zu einer konstanten Hurwitz-stabilen Matrix B, d. h. $\exists P(.) : -P^{-1}(\dot{P} - AP) = B$, so sichert $\varrho_L < 0$ des Δx-Systems ULES [436]. Weiterhin impliziert $\varrho_L = \varrho_C < 0$ für Systeme erster Ordnung LAS des Originals [393].

Anmerkung A.8 Lyapunovs Satz zur Stabilität durch eine Erste-Ordnung-Approximation (synonym für indirekte Methode) wird durch das Chataev-Malkin-Massera-Theorem [371] verfeinert und das Leonov-Theorem (2008) [391] ergänzt. Letztlich geht es in beiden Theoremen um Zusammenhänge zwischen der exponentiellen Wachstumsrate des Restterms und der Determinante der Cauchy-Matrix. Als Übersichtsarbeiten und Ergänzungen (Instabilitätstheoreme, diskrete Systeme, Beispiele) seien [393], [53] und [150] genannt.

Gänzlich anders ist das Kelemen-Theorem [344], das sich auf Systeme $\dot{x} = f(x,u)$ bezieht. Es weist große Ähnlichkeit mit den Sätzen für langsam-zeitvariante Systeme auf, s. Abschn. 2.8. Danach sichert eine hinreichend langsame Änderung von u Stabilität bezüglich einer Familie asymptotisch stabiler Ruhelagen, womit Stabilität einiger Gain-Scheduling-Strategien oder der Feedforward-Linearisierung [238] begründet werden kann.

5. Ableitungsbasierte Sätze für nichtkonstante Lösungen
Die Stabilität von Lösungen $x(t;x_0,t_0)$ kann auf zwei Wegen untersucht werden. Der erste Weg orientiert sich am transformierten System (A.1) und untersucht die Ruhelage $\tilde{x} = 0_n$ von $\dot{\tilde{x}} = \tilde{f}(\tilde{x},t); \tilde{f}(0_n,t) \equiv 0_n$ mit den Methoden von Punkt 4. Die Stabilitätseigenschaften von $\tilde{x}(t) \equiv 0_n$ übertragen sich auf $x(t;x_0,t_0)$. Der zweite Weg bezieht sich nicht auf das transformierte System, sondern direkt auf $\dot{x} = f(x,t)$. Um in den Betrachtungen die unschöne Abhängigkeit von t_0 loszuwerden (bei Lyapunov-Stabilität wird über die Anfangswerte, nicht aber über die Anfangszeit variiert), erfolgt zunächst eine Zeittransformation $t := t - t_0$, die im zeitinvarianten Fall natürlich gegenstandslos ist. Somit kann ohne Einschränkungen von $\dot{x} = f(x,t); x(0) = x_0$ bzgl. einer Lösung $x(t;x_0)$ ausgegangen werden.

Satz A.1 (Stabilität nichtkonstanter Lösungen, [390])
Für $f \in \mathcal{C}^2$ folgt Lyapunov-Stabilität einer Lösung $x(t;x_0)$ aus $A(t;\bar{x}_0) = \mathrm{D}f(x(t;\bar{x}_0),t)$, wenn $||\Phi(t,0;\bar{x}_0)||_2 \leq \gamma(t) < \infty$; $\bar{x}_0 \in \mathcal{U}_\varepsilon(x_0)$ gilt. $\lim\limits_{t\to\infty} \gamma(t) \to 0$ impliziert dann asymptotische Stabilität von $x(t;x_0)$.

Gegenüber Lyapunovs indirekter Methode wird im Satz A.1 Gleichmäßigkeit bzgl. x_0 gefordert. Ferner verhindert das Umgebungskonzept den Perron-Effekt, denn nur Lösungen auf dem Rand des Flusses, der stabil in der ersten Approximation ist, können den Perron-Effekt hervorrufen. Klar ist auch, dass $\forall \bar{x}_0 \in \mathcal{U}_\varepsilon(x_0) : \varrho_L(\bar{x}_0) < 0$ asymptotische Stabilität sichert. Die Notwendigkeit, Lösungen des nichtlinearen Systems in der Umgebung zu betrachten, zeigt sich anhand der instabilen Lösungen in der Umgebung von $x_0 = 0_2$ in (A.7). Regularität der Linearisierung erübrigt diesen Aufwand in Weg 1. Gleichmäßige (asymptotische) Stabilität der Nulllösung der Linearisierung reduziert in beiden Wegen ebenso den Aufwand und lässt zudem den Schluss auf gleichmäßige (asymptotische) Stabilität der Lösung zu.

Beispiel A.10 (Linearisierung um nichtkonstante Lösungen)
$\dot{x} = x^2 - 1$ hat die instabile Ruhelage $x_1(t) \equiv 1$, die stabile Ruhelage $x_2(t) \equiv -1$ und die Lösung $x_3(t) = \frac{x_0 - \tanh t}{1 - x_0 \tanh t}$ für $|x_0| \neq 1$. Untersucht werden soll $x_3(t;0) = -\tanh t$.
Im ersten Weg wird nur um $x_3(t;0)$ linearisiert und mit Regularität argumentiert; im zweiten Weg wird um alle Umgebungslösungen linearisiert.
Erster Weg: $\dot{\tilde{x}} = -2(\tanh t)\tilde{x} + \tilde{x}^2$ folgt über $\tilde{x} = x + \tanh t$ und liefert die Linearisierung $\Delta\dot{x} = -2(\tanh t)\Delta x$. Über $a(t) = -2\tanh t < 0$ kann asymptotische Stabilität von $\Delta x \equiv 0$ und wegen der Regularität bei Systemen erster Ordnung asymptotische Stabilität von $\tilde{x}(t) \equiv 0$ und letztlich von $x_3(t) = \tanh t$ geschlossen werden.
Zweiter Weg: Die Linearisierung bzgl. aller Umgebungslösungen lautet $\Delta\dot{x} = 2\frac{\Delta x_0 - \tanh t}{1 - \Delta x_0 \tanh t}\Delta x$ mit $\Delta x_0 \in \mathcal{U}_\varepsilon(0)$; $\varepsilon < 1$. Alle Systeme sind wegen $a(t;\Delta x_0) < 0$ asymptotisch stabil. Somit ist $x_3(t) = \tanh t$ eine asymptotisch stabile Lösung. Analog wird argumentiert, wenn $|\Phi(t,0;\Delta x_0)|$ (ungleich aufwändiger) entsprechend Satz A.1 betrachtet wird.
Die Lösung von $\Delta\dot{x} = -2(\tanh t)\Delta x$ lautet $z(t) = \frac{\cosh^2 t_0}{\cosh^2 t}\Delta x(t_0)$ und zeigt, dass die Lösung sogar gleichmäßig stabil ist, denn es gilt $\sup_{t_0 \geq 0} \sup_{t \geq t_0} |\frac{\cosh^2 t_0}{\cosh^2 t}| = 1 < \infty$. Allerdings ist die Lösung nicht gleichmäßig asymptotische stabil, da ihr hierzu die gleichmäßige Attraktivität fehlt. Sie ist halt nur attraktiv, was übrigens auch direkt aus $x_3(t,x_0) \to -1$ folgt.

6. Ableitungsbasierte Sätze für $\mathcal{L}^q$-Stabilität
Für Systeme $\dot{x} = f(x,u,t)$ und $y = h(x,u,t)$ darf aus der lokalen Ruhelagenstabilität, der Beschränktheit und Gleichmäßigkeit bzgl. t der Jacobi-Matrizen $\frac{\partial f}{\partial x^T}$ und $\frac{\partial f}{\partial u^T}$ sowie einer Abschätzung für h auf lokale (besser Kleinsignal) $\mathcal{L}^q$-Stabilität geschlossen werden. Zwei Sätze finden sich für ULAS bzw. ULES in [345].

7. Sätze für pseudolineare Systeme[34] $\dot{x} = A(x)x$

Die Forderung, dass alle Eigenwerte von $A(x)$ für alle x in der linken offenen komplexen Halbebene liegen, ist weder notwendig noch hinreichend, vgl. Beispiel A.11. Auch die Hinzunahme einer exponentiellen Beschränktheitsbedingung zur Hurwitz-Bedingung, wie sie von Langson und Alleyne (2002) [383] propagiert wird, erweist sich als nicht ausreichend für GAS, wie das Gegenbeispiel in [465] zeigt. Lediglich ULES kann geschlossen werden, wenn $A(0_n)$ Hurwitz-stabil ist, was sich direkt über $\frac{\partial A(x)}{\partial x^T}|_{x=0_n} = ((\frac{\partial A(x)}{\partial x_j}x + A(x)e_j))|_{x=0_n} = A(0_n)$ mit der indirekten Methode von Lyapunov begründen lässt.

Beispiel A.11 (Stabilitätsproblematik bei pseudolinearen Systemen)
Das System

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -x_1^2 + \frac{1}{2}x_1x_2 & 0 \\ 0 & -(1+x_2^2) + x_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

ist UGES [45], obwohl die Eigenwerte von $A(x)$ nicht überall negativ sind. Für

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & x_1^2 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \text{mit} \qquad \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

sind $\lambda_{1,2}(A(x)) \equiv -1$, was GAS nahe legen könnte; doch die Lösung $x_1(t) = \frac{2x_2(t)}{x_2^2(t)-2}$ und $x_2(t) = 2\mathrm{e}^{-t}$ entartet in $t = \ln\sqrt{2}$ [604]. Die Nichtanwendbarkeit der Eigenwertbetrachtung zeigt sich auch in der äquivalenten Darstellung für das vorstehende System

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 + g_1(x_1,x_2)x_2 & x_1^2 - g_1(x_1,x_2)x_1 \\ g_2(x_1,x_2)x_2 & -1 - g_2(x_1,x_2)x_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

in der $g_1, g_2 \in \mathcal{C}^1$ frei wählbar sind. Für $g_1(x) \equiv 1$ und $g_2(x) \equiv 0$ bleibt ein Eigenwert des jetzt neuen $A(x)$, aber immer noch gleichen $\dot{x} = f(x)$, bei -1, während der andere bei $-1+x_2$ liegt und für den gewählten Anfangswert $x_2(0) = 2$ instabil ist.

Pseudolineare Systeme erster Ordnung $\dot{x} = a(x)x$ sind genau dann GAS, wenn $a(x) < 0$ für alle $x \in \mathbb{R}\backslash\{0\}$ gilt. Das ist nämlich gleichbedeutend mit $a(x) = f(x)/x < 0$ bzw. äquivalent $\dot{V}(x) = f(x)x < 0$ für $V(x) = \frac{1}{2}x^2$. Ferner sind pseudolineare Systeme GAS, wenn alle Eigenwerte von $A(x) + A^T(x)$ in der linken offenen Halbebene liegen. Für Systeme mit einer speziellen Diagonaldominanz wird in [45] ein Stabilitätstheorem angegeben. Darüber hinaus findet sich für zustandsabhängige Schaltsysteme $A(x) \in \{A_1, \ldots, A_k\}$ in [45] ein i. Allg. konservatives Kriterium, das die Verweilzeiten und Abklingraten in den stabilen und instabilen Bereichen $\Omega_1, \ldots, \Omega_k$ in Beziehung setzt. Der Existenz quadratischer Lyapunov-Funktionen für zustandsabhängige Schaltsysteme widmet sich [244].

[34] Andere Begriffe für Systeme $\dot{x} = A(x)x + B(x)u; y = C(x)x + D(x)u$ sind „linearähnliche faktorisierte Systeme“, „Zustandsabhängige-Koeffizienten-Parametrisierung“, „Erweiterte Linearisierung“.

Für pseudolineare Differenzialgleichungen $x^{(n)} + f_{n-1}x^{(n-1)} + \ldots + f_1 x = 0$ mit den Funktionen $f_i = f_i(x^{(n-1)}, \ldots, \dot{x}, x)$ versagt die verallgemeinerte Hurwitz-Bedingung ebenfalls als Mittel zum Stabilitätsnachweis. Die Gültigkeit der Hurwitz-Bedingungen an die f_i für alle x garantiert demnach nicht Stabilität, wie es fälschlicherweise in [210] behauptet wird, vgl. hierzu $\ddot{x} + \dot{x} + f(\dot{x}, x)x = 0$ mit $x(0) = 0$, $\dot{x}(0) = 1$ und $f(\dot{x}, x) = \begin{cases} 10^{-3} & x\dot{x} \geq 0 \\ 10^{3} & x\dot{x} < 0 \end{cases}$ aus [630].

8. Sätze über verschwindende Störterme

Die folgenden Sätze beziehen sich auf Systeme, deren Ruhelage in einem bestimmten Sinne stabil ist. Die Sätze beantworten, ob und bis zu welchem Maße ein Störterm, der in der betreffenden Ruhelage verschwindet, zulässig ist, um das Stabilitätsverhalten nicht zu ändern. Den Schwerpunkt bilden die lineardominanten Systeme, bei denen das LTI-Verhalten den Einfluss nichtlinearer oder zeitvarianter Störterme dominiert.

Klassische Lyapunov-Abschätzung

$$\dot{x} = Ax + g(x) \quad \text{mit } ||g(x)||_2 \leq \gamma ||x||_2, \forall x \in \Omega, 0 < \gamma < \tfrac{1}{2\lambda_{\max}(P)}, g(0_n) = 0_n, g \in \mathcal{C}^0 \quad \text{(A.9)}$$

Hurwitz-Stabilität von A und $PA + A^T P = -I_n$ impliziert ULES auf Ω. Umgekehrt liegt ULES für

$$\dot{x} = Ax + G(x)x \quad \lim_{x \to 0_n} G(x) = 0_{n \times n}; \forall L > 0, \exists r_1 > 0 : ||G(x)||_2 \leq L, \forall ||x||_2 < r_1 \quad \text{(A.10)}$$

nur vor, wenn A Hurwitz-stabil ist.

Abschätzung mittels Stabilitätsradius

$$\dot{x} = Ax + g(x) \qquad \text{mit } ||g(\zeta)||_2 < r_{\mathbb{C}}^{-}(A) \cdot ||\zeta||_2, \forall \zeta \in \mathbb{C}^n \backslash \{0_n\} \text{ und } g(0_n) = 0_n, \quad \text{(A.11)}$$

wobei g lokal Lipschitz-stetig, differenzierbar in 0_n und von endlicher Verstärkung ist.
Hurwitz-Stabilität von A impliziert UGAS [289]. Gilt $\sup_\zeta \frac{||g(\zeta)||_2}{||\zeta||_2} < r_{\mathbb{C}}^{-}(A)$, folgt sogar UGES. Lokale Versionen mit δ als lokaler Lipschitz-Konstante über Ω und $0 < \delta \leq r_{\mathbb{C}}^{-}(A)$ für LAS und mit $0 < \delta < r_{\mathbb{C}}^{-}(A)$ für ULES [290] ergänzen diese Familie von Sätzen. Die letztgenannte Bedingung ist weniger einschränkend als (A.9), vgl. Abschn. 5.6 in [290].
Ist g reell – was üblicherweise der Fall ist – kann g durch $g(\zeta) := g(\Re e\, \zeta)$ fortgesetzt werden und die Resultate gelten dann für die fortgesetzten Funktionen. Die Bedingungen sind hinreichend; für ein konkretes g mit $\sup_x \frac{||g(x)||_2}{||x||_2} \geq r_{\mathbb{C}}^{-}(A)$ kann durchaus Stabilität gelten.

Lyapunov-Malkin-Theorem

$$\dot{x_1} = Ax_1 + g(x_1, x_2) \qquad g(0_{n_1}, x_2) = 0_{n_1}, \quad \text{(A.12a)}$$

$$\dot{x_2} = f_2(x_1, x_2) \qquad f_2(0_{n_1}, x_2) = 0_{n_2} \quad \text{(A.12b)}$$

Wenn A Hurwitz-stabil ist, dann ist $(0_{n_1}, 0_{n_2})$ stabil und asymptotisch stabil bezüglich x_1. Für Trajektorien, die nahe $(0_{n_1}, 0_{n_2})$ starten, gilt $\lim_{t \to \infty} x_1(t) = 0_n$ und $\lim_{t \to \infty} x_2(t) = c$ [82].

Asymptotisch zeitinvariante Systeme

$$\dot{x} = Ax + g(x,t) \qquad \text{mit } \lim_{t\to\infty} g(x,t) = 0_n \tag{A.13}$$

Die Folgerung, wonach GAS der Lösung von (A.13) für Hurwitz-stabiles A vorliegt, ist falsch, vgl. $\dot{x} = -x + 2\mathrm{e}^{-t}x^2; x(0) = 1$ mit $x(t) = \mathrm{e}^t$.

Lyapunov-Abschätzung für zeitvarianten Störterm

$$\dot{x} = Ax + g(x,t) \qquad \text{mit } ||g(x,t)||_2 \leq h(t)||x||_2,\ \int_0^\infty h(t)\,\mathrm{d}t < \infty,\ g \in \mathcal{C}^1 \tag{A.14}$$

Lyapunov-stabiles A sichert Lyapunov-Stabilität in (A.14) [59]. Hurwitz-Stabilität von A sichert UAS [350]. Wird A durch $A(t)$ ersetzt, so darf aus UAS von $\dot{x} = A(t)x$ weiterhin auf UAS des gestörten Systems geschlossen werden [592].

Abschätzung für speziellen asymptotisch verschwindenden Störterm

$$\dot{x} = (A(t) + G(t))x \qquad \text{mit } \forall t_0 \geq 0, \exists \beta < \infty : \int_{t_0}^\infty ||G(t)||\mathrm{d}t \leq \beta \tag{A.15}$$

UGES bzw. UGS für $\dot{x} = A(t)x$ sichert UGES bzw. UGS von (A.15) [543].

Abschätzung mit quadratischem Lyapunov-Kandidaten

$$\begin{aligned}\dot{x} = f(x)+g(x,t) \qquad & \text{mit } ||g(x,t)||_2 \leq \gamma||x||_2,\ \forall x \in \mathcal{B}(0_n,r), 0 < \gamma \leq c_3/c_4, \forall t \geq 0 \\ & \text{und } c_1||x||_2^2 \leq V(x,t) \leq c_2||x||_2^2, \dot{V}(x,t) \leq -c_3||x||_2^2, ||\tfrac{\partial V}{\partial x}||_2 \leq c_4||x||_2\end{aligned} \tag{A.16}$$

ULES für $\dot{x} = f(x)$ auf $\mathcal{B}(0_n, r)$ sichert ULES von (A.16) [345]. Die Abschätzung ist häufig konservativ. Außerdem kann aus LAS für $\dot{x} = f(x)$ LAS von (A.16) gefolgert werden, wenn $V(x,t)$ eine Lyapunov-Funktion vom quadratischen Typ ist, also $\dot{V}(x,t) \leq -c_3\phi^2(x)$, $||\frac{\partial V}{\partial x}||_2 \leq c_4\phi(x)$ mit $||g(x,t)||_2 \leq \gamma\phi(x)$ und $0 < \gamma \leq c_3/c_4$.
Die ULES-Bedingung ist wichtig, denn $x(t) \equiv 0$ von $\dot{x} = -x^3 + cx$ mit $c > 0$ ist instabil, obwohl $\dot{x} = f(x) = -x^3$ sogar GAS ist. Die LAS-Bedingung greift mit $V(x) = x^4, \phi(x) = x^3, \gamma = 1$ wegen $|cx| \not\leq 1x^3$ in $\mathcal{B}(0_n, \varepsilon)$ nicht.

Die vorgenannten Sätze geben quantitative Abschätzungen für die Einhaltung der Stabilität unter Störungen der Differenzialgleichung. Qualitativ gesehen ist nur ULES robust gegenüber hinreichend kleinen Störungen. Dabei ist die Gleichmäßigkeit von Bedeutung, denn es gibt Systeme, die UGS sind und die Eigenschaft der globalen exponentiellen Attraktivität (Konvergenz) besitzen, die aber nicht total stabil sind. Ursache ist, dass bei globaler exponentieller Attraktivität die Parameter der Vergleichsfunktion nicht unabhängig von x_0 und t_0 sind.

9. Sätze für gekoppelte Systeme

Viele größere Systeme weisen spezielle Kopplungsstrukturen auf, z. B. eine Kaskadenform. Dann genügt es, meist einfache Bedingungen zu ergänzen, um für die Teilsysteme Stabilität nachzuweisen. Nachfolgend sind einige wichtige Beziehungen zusammengefasst.

Kaskadensysteme vom Typ:

$$\dot{x}_1 = f_1(x_1, x_2) \qquad f_1(0_{n_1}, 0_{n_2}) = 0_{n_1} \qquad f_1 \text{ lokal Lipschitz-stetig} \tag{A.17a}$$

$$\dot{x}_2 = f_2(x_2) \qquad f_2(0_{n_2}) = 0_{n_2} \qquad f_2 \text{ lokal Lipschitz-stetig} \tag{A.17b}$$

Implikationen:
- ULS für $\dot{x}_2 = f_2(x_2)$ und ULAS für $\dot{x}_1 = f_1(x_1, 0_{n_2}) \Rightarrow$ (0,0)-ULS gesamt [595]
- ULAS für $\dot{x}_2 = f_2(x_2)$ und ULAS für $\dot{x}_1 = f_1(x_1, 0_{n_2}) \Leftrightarrow$ (0,0)-ULAS gesamt [595]
  Gilt nicht für UGAS, vgl. $\dot{x}_1 = -x_1 + x_1^2 x_2, \dot{x}_2 = -x_2$ und $x_1(0) \cdot x_2(0) > 2$.
  Aber, wenn beide UGAS und f_1 global Lipschitz oder f_1 CIBS, dann (0,0)-UGAS [574]
- UGAS für $\dot{x}_2 = f_2(x_2)$ und ISS für $\dot{x}_1 = f_1(x_1, x_2) \Rightarrow$ (0,0)-UGAS gesamt [29]
- UGAS + ULES für f_2 und IISS für f_1, wenn affin in $x_2 \Rightarrow$ (0,0)-UGAS gesamt [29]
- UGAS + ULES für f_2 und $\dot{x}_1 = f_{11}(x_1) + f_{12}(x_1, x_2)$ mit $f_{12}(x_1, 0_{n_2}) = 0_{n_1}$ siehe [448]

Kaskadensysteme vom Typ:

$$\dot{x}_1 = f_1(x_1, u) \qquad f_1(0_{n_1}, 0) = 0_{n_2} \qquad f_1 \text{ lokal Lipschitz-stetig} \tag{A.18a}$$

$$\dot{x}_2 = f_2(x_1, x_2) \qquad f_2(0_{n_1}, 0_{n_2}) = 0_{n_2} \qquad f_2 \text{ lokal Lipschitz-stetig} \tag{A.18b}$$

- ISS für f_1 und ISS für f_2 mit x_1 als Eingang $\Rightarrow$ ISS gesamt [581]

Small-Gain-Theorem

$$\dot{x}_1 = f_1(x_1, x_2) \qquad f_1(0_{n_1}, 0_{n_2}) = 0_{n_1} \qquad f_1 \text{ lokal Lipschitz-stetig} \tag{A.19a}$$

$$\dot{x}_2 = f_2(x_1, x_2) \qquad f_2(0_{n_1}, 0_{n_2}) = 0_{n_2} \qquad f_2 \text{ lokal Lipschitz-stetig} \tag{A.19b}$$

ISS für f_1 mit x_2 als Eingang und Verstärkung $\gamma_1 \in \mathcal{K}_\infty$ sowie ISS für f_2 bzgl. x_1 und $\gamma_2 \in \mathcal{K}_\infty$ liefert UGAS, wenn die Small-Gain-Bedingung $\gamma_1(\gamma_2(r)) \leq r$ erfüllt ist. Werden f_1 und/oder f_2 mit Eingängen versehen, folgt unter einer analogen Bedingung ISS [323].

Zwei-Zeitskalen-Systeme (singulär gestörte Systeme)

$$\dot{x}_1 = f_1(x_1, x_2, t; \varepsilon) \qquad x_1(t_0) = x_{1;\varepsilon} \tag{A.20a}$$

$$\varepsilon \dot{x}_2 = f_2(x_1, x_2, t; \varepsilon) \qquad x_2(t_0) = x_{2;\varepsilon} \tag{A.20b}$$

und $0 < \varepsilon \ll 1$ hat mit x_1 langsame und mit x_2 schnelle Komponenten. Durch die Zeittransformation $\tau = (t - t_0)/\varepsilon$ lautet die Differenzialgleichung bezüglich τ

$$x_1' = \varepsilon f_1(x_1, x_2, t_0 + \varepsilon\tau; \varepsilon) \qquad x_1(0) = x_{1;\varepsilon} \tag{A.21a}$$

$$x_2' = f_2(x_1, x_2, t_0 + \varepsilon\tau; \varepsilon) \qquad x_2(0) = x_{2;\varepsilon}. \tag{A.21b}$$

Mit $\varepsilon = 0$ kann die schnelle Dynamik $x_2' = f_2(x_{1;0}, x_2, t_0; 0); x_2(0) = x_{2;0}$ untersucht werden. Die durch $0 = f_2(x_1, x_2, t; 0)$ definierte Menge heißt langsame Mannigfaltigkeit. Sie kann aus mehreren Teilmengen bestehen. Sei $x_2 = \phi(x_1, t)$ eine Beschreibung einer Teilmenge der Mannigfaltigkeit, so reduziert sich die langsame Dynamik auf

$$\dot{x}_1 = f_1(x_1, \phi(x_1, t), t; 0) \qquad x_1(t_0) = x_{1;\varepsilon}. \tag{A.22}$$

Exponentielle Stabilität beider Teildynamiken impliziert lokale exponentielle Stabilität der Ruhelagen von (A.20) [355]. Asymptotische Stabilität (da nicht robust bezüglich Störungen) reicht nicht, vgl. $\dot{x}_1 = -x_1^3 + \varepsilon x_1$ und $\varepsilon \dot{x}_2 = -x_2$ mit x_1 GAS für $\varepsilon = 0$, aber instabiler Ruhelage des Originalsystems. Globale exponentielle Stabilität beider Systeme impliziert nicht globale Stabilität, vgl. $\dot{x}_1 = -x_1 + x_1^2 x_2$ und $\varepsilon \dot{x}_2 = -x_2$ mit $\dot{x}_1 = -x_1$ und $x_2' = -x_2$.

10. Abschätzungstheoreme

Abschätzungstheoreme (s. auch comparison theorem, comparison principle) eröffnen die Möglichkeit [333], qualitative Eigenschaften einer Differenzialgleichung durch Vergleich mit einer anderen, einfacher lösbaren Differenzialgleichung abzuleiten. So liefert das Sturmsche Theorem Aussagen zur Nullstellenanzahl von $\ddot{x} + a_0(t)x = 0$, während Differenzialungleichungen analog zum Majoranten- bzw. Minoranten-Prinzip von Folgen helfen, auf Stabilität bzw. Instabilität zu schließen. Erfüllen die beiden skalaren Systeme

$$\dot{x}_1 = f_1(t, x_1) \qquad x_1(t_0) = x_{10} \tag{A.23a}$$

$$\dot{x}_2 = f_2(t, x_2) \qquad x_2(t_0) = x_{20} \tag{A.23b}$$

$$f_1(t, x) \leq f_2(t, x) \quad \forall t \geq t_0 \qquad x_{10} \leq x_{20} \tag{A.23c}$$

und hat eines der Systeme eine eindeutige Lösung, so gilt $x_1(t) \leq x_2(t)$ für $t \geq t_0$ [333]. Bei Nichtnegativität der Lösungen folgt dann $x_2(t) \to 0 \Rightarrow x_1(t) \to 0$. Theoreme dieses Typs erweisen sich als nützlich, wenn die Ableitung eines Lyapunov-Kandidaten $V(x)$ das Vorzeichen wechselt (das Lyapunov-Theorem also nicht mehr greift). Gelingt eine Abschätzung $\dot{V}(x(t)) \leq a(t)\psi(V)$, so kann aus dem Stabilitätsverhalten von $v(t) \equiv 0$ des skalaren Systems $\dot{v} = a(t)\psi(v)$ auf das von $x(t) \equiv 0_n$ geschlossen werden [379], [399].

11. Fixpunkttheoreme

Besonders für den Stabilitätsnachweis in zeitdiskreten Systemen eignen sich Fixpunkttheoreme. Die bekanntesten sind die von Banach, Brouwer, Schauder, Tarski und Krasnoselskii. Ihre Anwendung ist dabei nicht nur auf zeitdiskrete Systeme beschränkt, wie in [109] gezeigt wird. Zu beachten ist, dass die Kontraktionseigenschaft von der gewählten Norm abhängt. So ist $f(x_1, x_2) = [0.99x_1, 0.49(x_1 + x_2)]^T$ in $||.||_\infty$ kontraktiv, in $||.||_2$ aber nicht.

Literaturverzeichnis

[1] *Abatzoglou, T.; O'Donnell, B.:* Minimization by coordinate descent. J. of Optimization Theory and Applications 36 (1982), 163–174.

[2] *Aboky, C.; Sallet, G.; Vivalda, J.-C.:* Observers for Lipschitz non-linear systems. Int. J. Control 75 (2002) 3, 204–212.

[3] *Ackermann, J.:* Does it suffice to check a subset of multilinear parameters in robustness analysis? IEEE Transactions on Automatic Control 37 (1992), 487–488.

[4] *Adamy, J.:* Nichtlineare Regelungen. Berlin: Springer-Verlag, 2009.

[5] *Ademola, T.A.; Arawomo, P.O.:* Stability and ultimate boundedness of solutions to certain third-order differential equations. Applied Mathematics E-Notes 10 (2010), 61–69.

[6] *Aeyels, D.:* Generic observability of differentiable systems. SIAM J. Control Optim. 19 (1981) 5, 595–603.

[7] *Aeyels, D.; Peuteman, J.:* A new asymptotic stability criterion for non-linear time-variant differential equations. IEEE Trans. on Automatic Control 43 (1998), 968–971.

[8] *Aguirre, B.; Suárez, R.:* Algebraic test for the Hurwitz stability of a given segment of polynomials. Boletin de la Sociedad Matematica Mexicana 12 (2006) 2, 261–275.

[9] *Aguirre, L.A.; Lopes, R.A.M.; Amaral, G.F.V.; Letellier, C.:* Constraining the topology of neural networks to ensure dynamics with symmetry properties. Phys. Rev. E 69 (2004), online publication 026701.

[10] *Ahmadi, A.A.; Parrilo, P.A.:* On higher order derivatives of Lyapunov functions. Proc. American Control Conference, 2011.

[11] *Ahn, S.J.; Rauh, W.; Warnecke, H.-J.:* Least-squares orthogonal distances fitting of circle, sphere, ellipse, hyperbola, and parabola. Pattern Recognition 34 (2001), 2283–2303.

[12] *Ahn, Y.J.:* Constrained Jacobi polynomial and constrained Chebychev polynomial. Commun. Korean Math. Soc. 23 (2008) 2, 279–284.

[13] *Alam, R.:* On the construction of nearest defective matrices to a normal matrix. Linear Algebra and its Applications 395 (2005), 367–370.

[14] *Algaba, A.; Merino, M.; Rodríguez-Luis, A.J.:* Takens Bogdanov bifurcations of periodic orbits and Arnold's tongues in a three-dimensional electronic model. International Journal of Bifurcation and Chaos 11 (2001) 2, 513–531.

[15] *Aling, H.; Schumacher, J.M.:* A nine-fold canonical decomposition for linear systems. Int. J. Control 39 (1984) 4, 779–805.

[16] *Almasy, G.A.; Sztano, T.:* Empirical models satisfying balance equations. Contribution des calculateurs electroniques au developpement du genie chimique et de la chimie industrielle/ Societe de Chimie Industrielle. – Paris, Vol. C. Reacteurs et ateliers (1978), 1–5.

[17] *Amato, F.; Celentano, G.; Garofalo, F.:* New sufficient conditions for the stability of slowly varying linear systems. IEEE Transaction on Automatic Control 38 (1993) 9, 1409–1411.

[18] *Amitsur, A.C.:* Differential polynomials and division algebras. Ann. Math. 59 (1954), 245–278.

[19] *Anda, A.A.; Park, H.:* Self-scaling fast rotations for stiff and equality-constrained linear least squares problems. Linear Algebra and its Applications 234 (1996), 137–161.

[20] *Anderson, B.D.O.:* A simplified viewpoint of hyperstability. IEEE Transactions on Automatic Control 13 (1968), 292–294.

[21] *Anderson, B.D.O.; Jury, E.I.; Mansour, M.:* On robust Hurwitz polynomials. IEEE Transactions on Automatic Control 32 (1987) 10, 909–913.

[22] *Anderson, T.W.:* Statistical inference for covariance matrices with linear structure. In: Multivariate Analysis II, Proceedings of the Second International Symposium on Multivariate Analysis, Dayton, ohio, 1968 (Hrsg.: Krishnaiah, P.), New York: Academic Press 1969, 55–66.

[23] *Andersson, L.-E.; Elfving, T.:* A constrained Procrustes problem. SIAM J. on Matrix Analysis and Applications 18 (1997) 1, 124–139.

[24] *Andrievsky, B.R.; Churilov, A.N.; Fradkov, A.L.:* Feedback Kalman-Yakubovich lemma and its applications to adaptive control. Proc. Decision and Control, Kobe, Japan, 1996, 4537–4542.

[25] *Angeli, D.; Sontag, E.D.; Wang, Y.:* A characterization of integral input to state stability. IEEE Transactions on Automatic Control 45 (2000) 6, 1082–1097.

[26] *d' Angio, L.; Audoly, S.; Bellu, G.; Saccomani, M.P.; Cobelli, C.:* Structural identifiability of nonlinear systems: Algorithms based on differential ideals. Proc. 10th IFAC/IFORS Symposium on System Identification, Copenhagen, Denmark, volume 3, 1994, 13–18.

[27] *Anstreicher, K.; Wolkowicz, H.:* On Lagrangian relaxation of quadratic matrix constraints. SIAM J. Matrix Anal. Appl. 22 (2000) 1, 41–55.

[28] *Appell, J.:* Analysis in Beispielen und Gegenbeispielen. Berlin: Springer-Verlag, 2009.

[29] *Arcak, M.; Angeli, D.; Sontag, E.D.:* A unifying integral ISS framework for stability of nonlinear cascades. SIAM J. Control Optim. 40 (2002) 6, 1888–1904.

[30] *Arrowsmith, D.K.; Place, C.M.:* An introduction to dynamical systems. Cambridge University Press, 1990.

[31] *Artstein, Z.:* Uniform asymptotic stability via the limiting equations. J. Diff. Equat. 27 (1978), 172–189.

[32] *Åström, K.J.; Hagander, P.; Sternby, J.:* Zeros of sampled systems. Automatica 20 (1984) 1, 31–38.

[33] *Åström, K.J.; Murray, R.M.:* Feedback systems. Princeton University Press, 2008.

[34] *Åström, K.J.; Wittenmark, B.:* Adaptive control, 2nd edition. Pearson Education, 1996.

[35] *Atkinson, F.V.:* Multiparameter eigenvalue problems. New York: Academic Press, 1972.

[36] *Avdeenko, T.; Kargin, S.:* New results on global identifiability of linear state space models. Proc. 13th IFAC/IFORS Symposium on System Identification, Rotterdam, The Netherlands, 2003, 725–730.

[37] *Bacciotti, A.; Rosier, L.:* Liapunov function and stability in control theory. Berlin, Heidelberg: Springer-Verlag, 2005.

[38] *Back, A.D.; Tsoi, A.C.; Horne, B.G.; Giles, C.L.:* Alternative discrete-time operators and their application to nonlinear models. Technical report, CS-TR-3738, Institute for Advanced Computer Studies, University of Maryland, 1997.

[39] *Bai, Z.; Feldmann, P.; Freund, R.:* How to make theoretically passive reduced-order models passive in practice. Proc. IEEE Custom Integrated Circuits Conference, 1998, 207–210.

[40] *Baier, R.; Farkhi, E.:* The directed subdifferential of DC functions. Proc. on the Conference on Nonlinear Analysis and Optimization in celebration of Alex Ioffe's 70th and Simeon Reich's 60th birthdays, June 18-24, Technion, Haifa, Israel, 2008.

[41] *Baker, R.A.:* Lyapunov's first method applied to time-varying systems. IEEE Transactions on Automatic Control 15 (1970) 1, 143–144.

[42] *Balakrishnan, V.; Boyd, S.:* Existence and uniqueness of optimal matrix scalings. SIAM J. Matrix Anal. Appl. 16 (1995) 1, 29–39.

[43] *Balan, R.:* An extension of Barbashin-Krasovskii-Lasalle theorem to a class of nonautonomous systems. Nonlinear Dynamic System Theory 8 (2008) 3, 255–268.

[44] *Bandemer, H.:* Theorie und Anwendung der optimalen Versuchsplanung. I. Handbuch zur Theorie. Berlin: Akademie-Verlag, 1977.

[45] *Banks, S.P.; Al-Jurani, S.K.:* Pseudo-linear systems, Lie algebras, and stability. IMA J. of Mathematical Control & Information 13 (1996), 385–401.

[46] *Bao, J.; Lee, P.L.:* Process control – The passive systems approach. London: Springer-Verlag, 2007.

[47] *Barabanov, N.E.:* About the problem of Kalman (in russian). Sib. Mat. J. 29 (1988) 3, 3–11.

[48] *Baratchart, L.:* Existence and generic properties for l^2 approximants of linear systems. I.M.A. Journal of Math. Control and Identification 3 (1986), 89–101.

[49] *Barbashin, E.A.:* Lyapunov functions. Moscow: Nauka, 1970.

[50] *Barlow, R.E.; Bartholomew, D.J.; Bremner, J.M.; Brunk, H.D.:* Statistical inference under order restrictions – The theory and application of isotonic regression. New York: John Wiley & Sons, 1972.

[51] *Barmish, B.R.:* New tools for robustness of linear systems. New York: Macmillan Publishing Company, 1994.

[52] *Barmish, B.R.; Hollot, C.V.:* Counter-example to a recent result on the stability of interval matrices by S.Białas. Int. J. Control 39 (1984) 5, 1103–1104.

[53] *Barreira, L.; Valls, C.:* Stability in nonautonomous dynamics: A survey of recent results. São Paulo J. Math. Sci. 1 (2007), 133–174.

[54] *Bartlett, A.C.; Hollot, C.V.; Huang, L.:* Root locations of an entire polytope of polynomials: It suffices to check the edges. Mathematics of Control, Signals, and Systems 1 (1988), 61–71.

[55] *Batah, F.S.M.; Gore, S.D.:* Ridge regression estimator: combining unbiased and ordinary ridge regression of estimation. Surveys in Mathematics and its Applications 4 (2009), 99–109.

[56] *Bauschke, H.H.; Borwein, J.M.; Lewis, A.S.:* On the method of cyclic projections for convex sets in Hilbert space. Contemporary Mathematics 204 (1997), 1–38.

[57] *Bauschke, H.H.; Deutsch, F.; Hundal, H.; Park, S.-H.:* Accelerating the convergence of the method of alternating projections. Transactions of the American Mathematical Society 355 (2003) 9, 3433–3461.

[58] *Beliakov, G.:* Monotonicity preserving approximation of multivariate scattered data. BIT Numerical Mathematics 45 (2005), 653–677.

[59] *Bellman, R.:* Stability theory of differential equations. New York: McGraw-Hill, 1953.

[60] *Bellman, R.; Åström, K.J.:* On structural identifiability. Mathematical Biosciences 7 (1970), 329–339.

[61] *Bellman, R.; Fan, K.:* On systems of linear inequalities in Hermitian matrix variables. Convexity 7 (1963), 1–11.

[62] *Ben-Israel, A.; Greville, T.N.E.:* Generalized inverses: Theory and applications. New York: John Wiley & Sons, 1974.

[63] *Ben-Tal, A.; Nemirovskii, A.:* Lectures on modern convex optimization. Philadelphia: SIAM, 2001.

[64] *Benner, P.; Chu, D.:* A new test for passivity of descriptor systems. Oberwolfach Report 2 (2005) 1.

[65] *Benninger, N.F.; Rivoir, J.:* Ein neues konsistentes Maß zur Beurteilung der Steuerbarkeit in linearen, zeitinvarianten Systemen. Automatisierungstechnik 34 (1986), 473–479.

[66] *Berger, T.; Ilchmann, A.:* Zero dynamics of time-varying linear systems. Technical report, TU Ilmenau, Institut für Mathematik; Online-Veröffentlichung, 2010.

[67] *Berger, T.; Ilchmann, A.:* On stability of time-varying linear differential-algebraic equations. Int. J. of Control 86 (2013) 6, 1060–1076.

[68] *ten Berge, J.M.F.:* Least squares optimization in multivariate analysis. DSWO Press, Leiden University, The Netherlands, 1993.

[69] *Berkovich, L.M.:* Method of factorization of ordinary differential operators and some of its applications. Applicable Analysis and Discrete Mathematics 1 (2007), 122–149.

[70] *Bernzen, W.:* Nichtlineare Approximation nichtlinearer Systeme durch lineare Identifikation und Modellkombination. Technical report, Forschungsbericht 18/95, Gerhard-Mercator-Universität – GH Duisburg, Mess-, Steuer- und Regelungstechnik, 1995.

[71] *Bertsekas, D.P.:* Nonlinear programming, 2nd edition. Belmont: Athena Scientific, 1999.

[72] *Bhatia, R.:* Matrix analysis. New York, Berlin, Heidelberg: Springer-Verlag, 1996.

[73] *Bhattacharyya, S.P.; Chapellat, H.; Keel, L.H.:* Robust control: The parametric approach. Prentice Hall PTR Upper Saddle River, NJ, USA, 1995.

[74] *Białas, S.:* A necessary and sufficient condition for the stability of convex combinations of stable polynomial or matrices. Bulletin of the Polish Academy of Sciences, Technical Sciences 33 (1985), 473–480.

[75] *Białas, S.:* A necessary and sufficient condition for the stability of convex combinations of polynomials. Control and Cybernetics 33 (2004) 4, 589–597.

[76] *Bialas, S.; Góra, M.:* A few results concerning the Hurwitz stability of polytopes of complex polynomials. Linear Algebra and its Applications 436 (2012), 1177–1188.

[77] *Bissantz, N.; Dümbgen, L.; Munk, A.; Stratmann, B.:* Convergence analysis of generalized iteratively reweighted least squares algorithms on convex function spaces. SIAM J. Optimization 19 (2009) 4, 1828–1845.

[78] *Björck, Å.:* Least squares methods. In: Handbook of Numerical Analysis (Hrsg.: Ciarlet, P.G.; Lions, J.L.), volume I. Solution of Equations in R^n. Part 1., Elsevier/North Holland, Amsterdam 1990, 466–647.

[79] *Björck, Å.:* Numerical methods for least squares problems. SIAM, Philadelphia, PA, 1996.

[80] *Blanco, A.M.; Bandoni, J.A.:* Eigenvalue and singular value optimization. Mecanica Computational 22 (2003), 1258–1272.

[81] *Blanco, A.M.; Bandoni, J.A.:* Optimal design of stable processes. Latin American Applied Research 33 (2003) 2, 123–128.

[82] *Bloch, A.M.; Crouch, P.; Baillieul, J.; Marsden, J.:* Nonholonomic mechanics and control. New York: Springer-Verlag, 2003.

[83] *Bochner, S.:* Lecture on Fourier integrals. Princeton University Press, 1959.

[84] *Boţ, R.I.; Grad, S.-M.; Wanka, G.:* On strong and total Lagrange duality for convex optimization problems. J. Math. Anal. Appl. 337 (2008), 1315–1325.

[85] *Böning, W.:* Analytische Darstellung der Kennlinien nichtlinearer Zweipole. Electrical Engineering 45 (1960) 4, 265–278.

[86] *Böning, W.:* Der Einschaltvorgang der Spule mit gekrümmter magnetischer Kennlinie und ohmschem Widerstand an Wechselspannung. Electrical Engineering 50 (1965) 3, 171–183.

[87] *Bookstein, F.L.:* Fitting conic sections to scattered data. Computer Graphics and Image Processing 9 (1979), 56–71.

[88] *de Boor, C.:* Calculating with B-splines. J. of Approximation Theory 6 (1972), 50–62.

[89] *Boros, T.; Rózsa, P.; Mironovskij, L.; Mihajlov, N.:* A uniform algorithm for the transformation of multivariable systems into canonical forms. Linear Algebra and its Applications 147 (1991), 441–467.

[90] *Borwein, J.M.; Lewis, A.S.:* Convex analysis and nonlinear optimization: Theory and examples. New York: Springer Science and Business Media, 2006.

[91] *Boyd, S.; Balakrishnan, V.:* A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its l_∞-norm. Systems and Control Letters 15 (1990), 1–7.

[92] *Boyd, S.; El Ghaoui, L.; Féron, E.; Balakrishnan, V.:* Linear matrix inequalities in system and control theory. SIAM Studies in Applied Mathematics, Vol.15, Philadelphia, Pennsylvania, 1994.

[93] *Boyd, S.; Vandenberghe, L.:* Convex optimization. Cambridge: Cambridge University Press, 2004.

[94] *Boyd, S.; Yang, Q.:* Structured and simultaneous Lyapunov functions for system stability problems. International Journal of Control 49 (1989) 6, 2215–2240.

[95] *Boyle, J.P.; Dykstra, R.L.:* A method for finding projections onto the intersection of convex sets in Hilbert space. Proc. Advances in Order Restricted Statistical Inference (Hrsg.: Dykstra, R.; Robertson, T.; Wright, F.), volume 37, Springer-Verlag; Lecture Notes in Statistics. 1985, 28–47.

[96] *Brandwood, D.H.:* A complex gradient operator and its application in adaptive array theory. Communications, Radar and Signal Processing 130 (1983) 1, 11–16.

[97] *Bregman, L.M.; Censor, Y.; Reich, S.:* Dykstra's algorithm as the nonlinear extension of Bregman's optimization method. Journal of Convex Analysis 6 (1999) 2, 319–333.

[98] *Bretthauer, G.:* Identifikation rückgekoppelter Mehrgrößensysteme im Frequenzbereich – Einheitliche Darstellung und Vergleich der Verfahren. Habilitation, TU Dresden, 1983.

[99] *Bretthauer, G.; Kaufman, M.:* Identifiability of linear closed-loop systems. Proc. Control, Systems, Robotics and Automation. EOLSS, 2003.

[100] *Brien, A.; Vandenberghe, L.:* Handling nonnegative constraints in spectral estimation. Proc. 34th Asilomar Conference on Signals, Systems, and Computers, 2000, 202–206.

[101] *Bronstein, I.N.; Semendjajew, K.A.; Musiol, G.and Mühlig, H.:* Taschenbuch der Mathematik. 2. Aufl., Thun und Frankfurt a. M.: Verlag Harri Deutsch, 1995.

[102] *Browne, P.J.; Sleeman, B.D.:* A numerical technique for multiparameter eigenvalue problems. IMA J. of Numerical Analysis 2 (1982), 451–457.

[103] *Buchberger, B.:* Ein Algorithmus zum Auffinden der Basiselemente des Restklassenringes nach einem nulldimensionalen Polynomideal. PhD thesis, Universität Innsbruck, 1965.

[104] *Buchta, H.:* Zweistufiges Verfahren zur Identifikation linearer dynamischer Systeme vom SISO-Typ. msr 30 (1987) 8, 362–366.

[105] *Bunke, H.; Bunke, O.:* Nonlinear regression, functional relations and robust methods. New York: John Wiley & Sons, 1989.

[106] *Burg, J.P.; Luenberger, D.G.; Wenger, D.L.:* Estimation of structured covariance matrices. Proc. IEEE 40 (1982) 9, 963–974.

[107] *Burnham, K.P.; Anderson, D.R.:* Model selection and multimodel inference: A practical information-theoretic approach. New York: Springer-Verlag, 2002.

[108] *Burns, J.A.; Peichl, G.H.:* Control system radii and robustness under approximation. Nonconvex Optimization and its Applications 81 (2006), 25–62.

[109] *Burton, T.A.; Furumochi, T.:* A note on stability by Schauder's theorem. Funkcialaj Ekvacioj 44 (2002), 73–82.

[110] *Byers, R.:* A bisection method for measuring the distance of a stable matrix to the unstable matrices. SIAM J. Sci. Stat. Comput. 9 (1988), 875–881.

[111] *Byers, R.; He, C.; Mehrmann, V.:* Where is the nearest non-regular pencil? Linear Algebra and its Applications 285 (1998), 81–105.

[112] *Byrnes, C.I.; Isidori, A.:* Local stabilization of minimum phase nonlinear systems. System and Control Letters 11 (1988), 9–17.

[113] *Byrnes, C.I.; Isidori, A.; Willems, J.C.:* Passivity, feedback equivalence and the global stabilization of minimum phase nonlinear systems. IEEE Transactions on Automatic Control 10 (1991), 1122–1137.

[114] *Byrnes, C.I.; Landau, H.J.; Lindquist, A.:* On the well-posedness of the rational covariance extension problem. In: Current and Future Directions in Applied Mathematics, Birkhäuser-Verlag 1997, 83–106.

[115] *Cadzow, J.A.:* Signal enhancement – A composite property mapping algorithm. IEEE Transactions on Acoustics, Speech, and Signal Processing 36 (1988) 1, 49–62.

[116] *Camilli, F.; Grüne, L.; Wirth, F.:* Calculating the domain of attraction: Zubov's method and extensions. Proc. of the Stability and Control Processes Conference, 2005, 27–36.

[117] *Camlibel, M.K.; Frasca, R.:* Extension of Kalman-Yakubovich-Popov lemma to descriptor systems. Proc. Decision and Control, New Orleans, USA, 2007, 1094–1099.

[118] *Ceragioli, F.M.:* The linearization method of stability theory in Banach spaces. Rend. Sem. Mat. Univ. Pol. Torino 54 (1996) 2, 167–177.

[119] *Ceragioli, F.M.:* Discontinuous ordinary differential equations and stabilization. PhD thesis, Universitá degli Studi di Firenze – Dipartimento di Matematica Ulisse Dini, 1999.

[120] *Cernikov, S.N.:* Contraction of finite systems of linear inequalities. Soviet Mathematics Doklady 4 (1963) 5, 1520–1524.

[121] *Cesari, L.:* Asymptotic behavior and stability problems in ordinary differential equations. Berlin, Göttingen, Heidelberg: Springer-Verlag, 1959.

[122] *Chabour, R.; Kalitine, B.:* Semi-definite Lyapunov functions: stability and stabilizability. Technical report, Universite de Metz, 2002.

[123] *Chandrasekaran, S.; Golub, G.H.; Gu, M.; Sayed, A.H.:* Efficient algorithms for least squares type problems with bounded uncertainties. In: Recent advances in total least squares techniques and error-in-variables modelling (Hrsg.: van Huffel, S.), Philadelphia: Society for Industrial and Applied Mathematics 1997, 171–180.

[124] *Chandrasekaran, S.; Golub, G.H.; Gu, M.; Sayed, A.H.:* An efficient algorithm for a bounded errors-in-variables model. SIAM J. Matrix Anal. Appl. 20 (1999) 4, 839–859.

[125] *Chartrand, R.; Yin, W.:* Iteratively reweighted algorithms for compressive sensing. Proc. 33rd International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2008.

[126] *Chen, C.F.; Chu, H.:* A matrix for evaluating Schwarz's form. IEEE Transactions on Automatic Control 11 (1966), 303–305.

[127] *Chen, H.; van Huffel, S.; Dowling, E.M.; DeGroat, R.:* TLS based methods for exponential parameter estimation. In: Recent advances in total least squares techniques and error-in-variables modelling (Hrsg.: van Huffel, S.), Philadelphia: Society for Industrial and Applied Mathematics 1997, 295–305.

[128] *Chen, Y.; McInroy, J.E.:* Estimation of symmetric positive-definite matrices from imperfect measurements. IEEE Trans. on Automatic Control 47 (2002) 10, 1721–1725.

[129] *Cheney, W.; Goldstein, A.:* Proximity maps for convex sets. Proceedings of the AMS 10 (1959), 448–450.

[130] *Chesi, G.:* LMI techniques for optimization over polynomials in control: A survey. IEEE Transactions on Automatic Control 55 (2010) 11, 2500–2510.

[131] *Chesi, G.:* Domain of attraction. Analysis and control via sos programming. London: Springer-Verlag, 2011.

[132] *Chin, P.; Corless, R.M.; Corliss, G.F.:* Optimization strategies for the approximated GCD problem. Proc. ISSAC, ACM Press, 1998, 228–235.

[133] *Chou, C.T.; Maciejowski, J.M.:* System identification using balanced parametrizations. IEEE Transactions on Automatic Control 42 (1997) 7, 956–974.

[134] *Chu, M.T.; Driessel, K.R.:* The projected gradient method for least squares matrix approximations with spectral constraints. SIAM J. Numer. Anal. 27 (1990) 4, 1050–1060.

[135] *Chu, M.T.; Trendafilov, N.T.:* The orthogonality constrained regression revisited. Journal of Computational & Graphical Statistics 10 (2001) 4, 746–771.

[136] *Chu, M.T.; Watterson, J.L.:* On a multivariate eigenvalue problem. part I: Algebraic theory and a power method. SIAM J. Sci. Comput. 14 (1993), 1089–1106.

[137] *Chui, N.L.C.; Maciejowski, J.M.:* Realization of stable models with subspace methods. Proc. 13th IFAC Triennial World Congress, San Francisco, California, 1996, 323–332.

[138] *Cima, A.; van den Essen, A.; Gasull, A.; Hubbers, E.; Mañosas, F.:* A polynomial counterexample to the Markus-Yamabe conjecture. Adv. Math. 131 (1997), 453–457.

[139] *Cima, A.; Gasull, A.; Mañosas, F.:* A note on LaSalle's problem. Annales Polonici Mathematici LXXVI (2001), 33–46.

[140] *Clotet, J.; García-Planas, M.I.; Magret, M.D.:* Estimating distances from quadruples satisfying stability properties to quadruples not satisfying them. Linear Algebra and its Applications 332-334 (2001), 541–567.

[141] *Co, T.; Zhong, K.:* Lie transformations for parameter estimation of continuous-time nonlinear systems. Proc. American Control Conference, Albuquerque, New Mexico, 1997, 3053–3057.

[142] *Cohen, J.E.; Rothblum, U.G.:* Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. Linear Algebra and its Applications 190 (1993), 149–168.

[143] *Coppel, W.A.:* Dichotomies and stability theory. In: Lecture Notes in Mathematics, Vol. 629, Berlin: Springer-Verlag 1978, 160–162.

[144] *Correa, M.V.; Aguirre, L.A.; Saldanha, R.R.:* Using steady-state priori knowledge to constrain parameter estimates in nonlinear system identification. IEEE Transactions on Circuits and Systems 49 (2002) 9, 1376–1381.

[145] *Cox, A.J.; Higham, N.J.:* Stability of Householder QR factorization for weighted least squares problems. Numerical Analysis 1997. Proc. 17th Dundee Biennial Conference, Pitman Research Notes in Mathematics 380, 1998, 57–83.

[146] *Cox, A.J.; Higham, N.J.:* Backward error bounds for constrained least squares problems. BIT 39 (1999) 1, 34–50.

[147] *Cox, A.J.; Higham, N.J.:* Row-wise backward stable elimination methods for the equality constrained least squares problem. SIAM J. Matrix Anal. Appl. 21 (1999) 1, 313–326.

[148] *Cox, M.G.:* The numerical evaluation of B-splines. J. Inst. Maths. Applics. 10 (1972), 134–149.

[149] *Cramer, J.S.:* Econometric Applications of Maximum Likelihood Methods. Cambridge University Press, 1986.

[150] *Czornik, A.:* On the regularity of discrete linear systems. Linear Algebra and its Applications 432 (2010), 2745–2753.

[151] *Dahlbom, U.:* Variance estimates based on knowledge of monotonicity and concavity properties. Technical report, Research Report 1998:7, Department of Statistics, Göteborg University, Sweden, 1998.

[152] *Daletskii, Y.L.; Krein, M.G.:* Stability of solutions of differential equations in Banach spaces. AMS, Providence, R.I., 1974.

[153] *Dantzig, G.B.; Thapa, M.N.:* Linear programming 1: Introduction. Berlin: Springer-Verlag, 1997.

[154] *Dantzig, G.B.; Thapa, M.N.:* Linear Programming 2: Theory and Extensions. Berlin: Springer-Verlag, 2003.

[155] *Dasgupta, S.; Gevers, M.; Bastin, G.; Campion, G.; Chen, L.:* Identifiability of scalar linearly parametrized polynomial systems. Proc. 9th IFAC/IFORS Symposium on System Identification, Budapest, Hungary, 1991, 374–378.

[156] *David, B.; Bastin, G.:* An estimator of the inverse covariance matrix and its applications to ML parameter estimation in dynamical systems. Automatica 37 (2001), 99–106.

[157] *Davidson, T.N.; Luo, Z.-Q.; Sturm, J.F.:* Linear matrix inequality formulation of spectral mask constraints. IEEE Transactions on Signal Processing 50 (2002) 11, 2702–2715.

[158] *Davison, E.J.; Wang, S.H.:* Properties and calculation of transmission zeros of linear multivariable systems. Automatica 10 (1974), 643–658.

[159] *Dax, A.:* The distance between two convex sets. Linear Algebra and its Applications 416 (2006), 184–213.

[160] *Demmel, J.W.:* A counterexample for two conjectures about stability. IEEE Trans. on Automatic Control 32 (1987), 34 242.

[161] *Demmel, J.W.; Edelman, A.:* The dimension of matrices (matrix pencils) with given Jordan (Kronecker) canonical forms. Linear Algebra and its Applications 230 (1995), 61–87.

[162] *Denis-Vidal, L.; Joly-Blanchard, G.:* Equivalence and identifiability analysis of uncontrolled nonlinear dynamical systems. Proc. MTNS'2000 (B50), Perpignan, France, 2000.

[163] *Deutsch, F.R.:* Best approximation in inner product spaces. CMS Books in Mathematics, New York: Springer-Verlag, 2001.

[164] *Dias, R.:* A note on maximum likelihood density estimation using a proxy of the Kullback-Leibler distance. Statistics and Probability Letters 13 (2000) 2, 1–10.

[165] *Dobson, I.:* Distance to bifurcation in multidimensional parameter space: Margin sensitivity and closest bifurcations. In: Bifurcation control – Theory and applications (Hrsg.: Chen, G.; Hill, D.J.; Yu, X.), Berlin: Springer-Verlag, LNCIS 293, 2003.

[166] *van Dooren, P.; Vermaut, V.:* On stability radii of generalized eigenvalue problems. Proc. European Conference on Control, 1997.

[167] *Douglas, R.; Shapiro, H.; Shields, A.:* Cyclic vectors and invariant subspaces for the backward shift operator. Annales de l'Institut Fourier, Grenoble 20 (1970), 37–76.

[168] *Doyle, J.; Francis, B.; Tannenbaum, A.:* Feedback control theory. Macmillan Publishing, 1990.

[169] *Duboshin, G.N.:* The problem of stability of motion under persistently acting perturbations. Trudy Gos. Astr. Inst. Sternberg 14 (1940) 1.

[170] *Duc, L.H.; Ilchmann, A.; Siegmund, S.; Taraba, P.:* On stability of linear time-varying second-order differential equations. Quart. Appl. Math. 64 (2006), 137–151.

[171] *Dullerud, G.E.; Paganini, F.G.:* A course in robust control theory: A convex approach. New York: Springer-Verlag, Texts in Applied Mathematics 36, 1999.

[172] *Dumitrescu, B.:* On convex stability domain and optimization of IIR filters. Proc. EUSIPCO, Toulouse, France, volume 2, 2002, 191–194.

[173] *Dumitrescu, B.; Tăbuş, I.; Stoica, P.:* On the parametrization of positive real sequences and MA parameter estimation. IEEE Transactions on Signal Processing 49 (2001) 11, 2630–2639.

[174] *Durgaprasad, G.; Rao, G.P.; Patra, A.; Mukhopadhyay, S.:* Indirect methods of estimation of parameters of discrete-time models. Proc. 9th IFAC/IFORS Symp. on Ident. and Sys. Par. Est., Budapest, Hungary, 1991.

[175] *Dwyer, P.S.; MacPhail, M.S.:* Symbolic matrix derivatives. Ann. Math. Statist. 19 (1948) 4, 517–534.

[176] *Dykstra, R.L.:* An algorithm for restricted least squares regression. J. of the Am. Stat. Ass. 78 (1983), 837–842.

[177] *Ebenbauer, C.; Allgöwer, F.:* Minimum-phase property of nonlinear systems in terms of a dissipation inequality. Proc. American Control Conference, Bosten, Massachusetts, 2004, 1737–1742.

[178] *Edelman, A.; Arias, T.A.; Smith, S.T.:* The geometry of algorithms with orthogonality constraints. SIAM J. on Matrix Analysis and Applications 20 (1998) 2, 303–353.

[179] *Edelman, A.; Elmroth, E.; Kågström, B.:* A geometric approach to perturbation theory of matrices and matrix pencils. Part I: Versal deformations. SIAM J. Matrix Anal. Appl. 18 (1997) 3, 653–692.

[180] *Eising, R.:* Between controllable and uncontrollable. Systems and Control Letters 4 (1984), 263–264.

[181] *Elsner, L.; Ikramov, K.D.:* Normal matrices: An update. Linear Algebra and its Applications 285 (1998), 291–303.

[182] *Emiris, I.Z.; Gallico, A.; Lombardi, H.:* Certified approximate univariate GCDs. J. Pure and Applied Algebra, Special Issue on Algorithms for Algebra 117-118 (1997), 229–251.

[183] *Espinoza, M.; Suykens, J.A.K.; de Moor, B.:* Imposing symmetry in least squares support vector machines regression. Proc. Decision and Control, Seville, Spain, 2005, 5716–5721.

[184] *Ewald, G.; Larman, D.G.; Rogers, C.A.:* The directions of a line segment of the r-dimensional balls of the boundary of a convex body in Euclidean space. Mathematika 17 (1970), 1–20.

[185] *Farina, L.; Rinaldi, S.:* Positive linear systems: Theory and applications. New York: John Wiley & Sons, 2000.

[186] *Feckan, M.:* A generalization of Bendixson's criterion. American Mathematical Society 129 (2001), 3395–3399.

[187] *Federer, H.:* Geometric measure theory. Berlin: Springer-Verlag, 1969.

[188] *Feldmann, S.; Lang, P.:* A least squares approach to reduce stable discrete linear systems preserving their stability. Linear Algebra and its Applications 381 (2004), 141–163.

[189] *Feßler, R.:* A proof of the two-dimensional Markus-Yamabe stability conjecture and a generalization. Ann. Polon. Math. 62 (1995), 45–75.

[190] *Fernández, L.A.:* On the limits of the Lagrange multiplier rule. SIAM Review 39 (1997) 2, 292–297.

[191] *Filatov, N.M.; Unbehauen, H.:* Adaptive dual control. Berlin: Springer-Verlag, 2004.

[192] *Filippov, A.F.:* Differential equations with discontinuous righthand sides. Dordrecht: Kluwer Academic Publishers, 1988.

[193] *Fitzgibbon, A.; Pilu, M.; Fisher, R.B.:* Direct least squares fitting of ellipses. IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (1999) 5, 476–480.

[194] *Fletcher, R.:* Practical methods of optimization. New York: John Wiley & Sons, 1987.

[195] *Flett, T.M.:* Differential analysis. Cambridge: Cambridge University Press, 1980.

[196] *Föllinger, O.:* Lineare Abtastsysteme. München: R. Oldenbourg Verlag, 1974.

[197] *Föllinger, O.:* Regelungstechnik. Heidelberg: Hüthig Buch Verlag, 1992.

[198] *Fradkov, A.L.; Pogromsky, A.Y.:* Introduction to control of oscillations and chaos. Series A, Vol. 35, World Scientific, 1998.

[199] *Frank, E.:* On the zeros of polynomials with complex coefficients. Bull. Amer. Math. Soc. 52 (1946), 144–158.

[200] *Frenk, J.B.G.; Kassay, G.; Kolumbán, J.:* On equivalent results in minimax theory. European Journal of Operational Research 157 (2004), 46–58.

[201] *Friedland, S.; Torokhti, A.:* Generalized rank-constrained matrix approximation. SIAM J. Matrix Anal. Appl. 29 (2007) 2, 656–659.

[202] *Fritsch, F.N.; Carlson, R.E.:* Monotone piecewise cubic interpolation. SIAM J. Numer. Anal. 17 (1980) 2, 238–246.

[203] *Fromion, V.; Monaco, S.; Normand-Cyrot, D.:* A link between input-output stability and Lyapunov stability. Systems & Control Letters 27 (1996), 243–248.

[204] *Fu, M.; Olbrot, A.W.; Polis, M.P.:* Robust stability for time-delay systems: The edge theorem and graphical tests. IEEE Trans. on Automatic Control 34 (1989) 8, 813–820.

[205] *Gahinet, P.; Apkarian, P.; Chilali, M.:* Parameter-dependent Lyapunov functions for real parametric uncertainty. IEEE Trans. on Automatic Control 41 (1996), 436–442.

[206] *Gahinet, P.; Nemirovski, A.; Laub, A.; Chilali, M.:* LMI control toolbox. User's guide. The MathWorks, Inc., 1995. www.mathworks.co.uk/help/releases/R13sp2/pdf_doc/lmi/lmi.pdf.

[207] *Gallant, A.R.:* Nonlinear statistical models. New York: John Wiley & Sons, 1987.

[208] *Gander, W.; Golub, G.H.; Strebel, R.:* Least-squares fitting of circles and ellipses. BIT 43 (1994), 558–578.

[209] *Gantmacher, F.R.:* The theory of matrices, vol. I,II (trans. K.A. Hirsch). New York: Chelsea, 1960.

[210] *Garg, D.P.; Shanidze, Z.G.; Rondell, E.G.:* Global stability of solutions of non-linear control systems. Int. J. Systems Sci. 20 (1989) 10, 1909–1924.

[211] *Gasparyan, O.N.:* Linear and nonlinear multivariable feedback control: A classical approach. New York: John Wiley & Sons, 2008.

[212] *von z. Gathen, J.; Gerhard, J.:* Modern Computer Algebra. Cambridge University Press, 1999.

[213] *Geiger, C.; Kanzow, C.:* Theorie und Numerik restringierter Optimierungsaufgaben. Berlin: Springer-Verlag, 2002.

[214] *Gelbaum, B.R.; Olmsted, J.M.H.:* Counterexamples in analysis. Dover Books, 2003.

[215] *Gel'fand, I.M.; Shilov, G.E.:* Generalized Functions, Vol. 1: Properties and Operations. New York: Academic Press, 1964.

[216] *Gellert, W.; Küstner, H.; Hellwich, M.; Kästner, H.:* Kleine Enzyklopädie Mathematik. VEB Bibliographisches Institut Leipzig, 1979.

[217] *Genesio, R.; Tartaglia, M.; Vicino, A.:* On the estimation of asymptotic stability regions: State of the art and new proposals. IEEE Transactions on Automatic Control 30 (1985) 8, 747–755.

[218] *van Gestel, T.; Suykens, J.; de Moor, B.; van Overschee, P.:* Identification of stable and positive real models using matrix inequalities in subspace methods. Technical report, Internal Report 99-30, ESAT-SISTA, K.U.Leuven, Belgium, 1999.

[219] *Giesl, P.; Wendland, H.:* Numerical determination of the basin of attraction for exponentially asymptotically autonomous dynamical systems. Nonlinear Analysis 74 (2011), 3191–3202.

[220] *Gil', M.I.:* On one class of absolute stable systems. Sov. Phys. Dokladi 280 (1983) 4, 811–815.

[221] *Gil', M.I.:* The input-output version of Aizerman's conjecture. International Journal of Robust and Nonlinear Control 8 (1998), 1219–1226.

[222] *Giraud, B.G.:* Constrained orthogonal polynomials. J. Phys. A: Math. Gen. 38 (2005), 7299–7311.

[223] *Glad, S.T.:* Identifiability with constraints. Proc. 4th IFAC Symposium, Enschede, The Netherlands, 1998, 437–440.

[224] *Glad, T.:* Step response of nonlinear non-minimum phase systems. Proc. Nonlinear Control Systems 2004, Stuttgart, Germany (Hrsg.: Allgöwer, F.; Zeitz, M.), volume 3, Elsevier Science Ltd. 2005, 1165–1169.

[225] *Glenberg, A.M.:* Learning from data: An introduction to statistical reasoning. Mahwah: Lawrence Earlbaum Associates, 1996.

[226] *Glunt, W.; Hayden, T.L.; Reams, R.:* The nearest doubly stochastic matrix to a real matrix with the same first moment. Numerical Linear Algebra with Applications 5 (1998), 475–482.

[227] *Glutsyuk, A.:* The complete solution of the Jacobian problem for vector fields on the plane. Comm. Moscow Math. Soc., Russian Math. Surveys 49 (1994), 185–186.

[228] *Goethals, I.; van Gestel, T.; Suykens, J.; van Dooren, P.; de Moor, B.:* Identification of positive real models in subspace identification by using regularization. IEEE Transactions on Automatic Control 48 (2003) 10, 1843–1847.

[229] *Golomb, M.:* Zur Theorie der nichtlinearen Integralgleichungen, Integralgleichungssysteme und allgemeinen Funktionalgleichungen. Math. Z. 39 (1935), 45–75.

[230] *Golub, G.H.; van Loan, C.F.:* Matrix computations. Baltimore: The John Hopkins University Press, 1983.

[231] *Golub, G.H.; Pereyra, V.:* The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. SIAM Journal of Numerical Analysis 10 (1973), 413–432.

[232] *Goodwin, G.C.; Sin, K.S.:* Adaptive filtering, prediction and control. Prentice Hall, 1984.

[233] *Gorman, J.D.; Hero, A.O.:* On the application of Cramer-Rao-type lower bounds for constrained estimation. Proc. Int. Conference on Acoustic, 1991, 1333–1336.

[234] *Gowda, M.S.:* Minimising quadratic functionals over closed convex cones. Bull. Austral. Math. Soc. 39 (1989), 15–20.

[235] *Gower, J.C.:* Multivariate analysis: ordination multidimensional scaling and allied topics. In: Handbook of applicable mathematics, Volume VI: Statistics, Part B, Chichester: John Wiley & Sons 1984, 727–781.

[236] *Gracia, J.M.:* Nearest matrix with two prescribed eigenvalues. Linear Algebra nd its Applications 401 (2005), 277–294.

[237] *Gracia, J.M.; de Hoyos, I.; Valesco, F.E.:* Safety neighbourhoods for the invariants of the matrix similarity. Linear and Multilinear Algebra 46 (1999) 1-2, 25–49.

[238] *Graichen, K.; Hagenmeyer, V.; Zeitz, M.:* A new approach to inversion-based feedforward control design for nonlinear systems. Automatica 41 (2005) 12, 2033–2041.

[239] *Graillat, S.:* A note on a nearest polynomial with a given root. ACM SIGSAM Bulletin 39 (2005) 2, 53–60.

[240] *Graillat, S.; Langlois, P.:* A comparison of real and complex pseudozero sets for polynomials with real coefficients. Proc. RNC-6, Real Numbers and Computer Conference, Schloss Dagstuhl, Germany (Hrsg.: Frougny, C. et al.), Nov. 2004, 103–112.

[241] *Graillat, S.; Langlois, P.:* Computation of stability radius for polynomials. Technical Report 31, Laboratoire MANO, 2004.

[242] *Grandine, T.A.; et.al.:* Generating surface lofts to scattered data. Technical report, Engineering Computing and Analysis ECA-TR-157, Boeing Computer Services, Seattle, WA, 1991.

[243] *Grewal, M.S.; Glover, K.:* Identifiability of linear and nonlinear dynamical systems. IEEE Transactions on Automatic Control 21 (1976), 833–837.

[244] *Griggs, W.M.; King, C.K.; Shorton, R.N.; Mason, O.; Wulff, K.:* Quadratic Lyapunov functions for systems with state-dependent switching. Linear Algebra and its Applications 433 (2010), 52–63.

[245] *Groeneboom, P.:* Estimating a monotone density. Proc. Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II (Hrsg.: Le Cam, L.; Olshen, R.), 1985.

[246] *Groeneboom, P.; Jongbloed, G.; Wellner, J.A.:* Estimation of a convex function: Characterizations and asymptotic theory. The Annals of Statistics 29 (2001), 1653–1698.

[247] *Gröll, L.:* Modellbildung für kontinuierliche Systeme mittels direkter Identifikation. PhD thesis, TU Dresden, 1995.

[248] *Gröll, L.:* Parameterrestriktionen bei der Identifikation am Beispiel des LSQ-Problems. Automatisierungstechnik 52 (2004) 1, 46–53.

[249] *Gröll, L.:* Integration von Vorwissen über Eigenschaften von Funktionen bei der Modellbildung. Automatisierungstechnik 60 (2012), 405–416.

[250] *Gröll, L.; Irle, P.:* An observer-based approach for the projection onto a 2d-curve under movement. Robotics and Autonomous Systems 59 (2011), 256–261.

[251] *Gröll, L.; Jäkel, J.:* A new convergence proof of fuzzy c-means. IEEE Trans. on Fuzzy Systems 13 (2005) 5, 717–720.

[252] *Gröll, L.; Janda, O.:* Analytische Zugänge für die bakengestützte Positionsbestimmung automatisch geführter Fahrzeuge. Proc. GMA-Kongreß 96, Meß- und Automatisierungstechnik, VDI Berichte 1282, 1996, 377–386.

[253] *Gröll, L.; Kapp, A.:* Effect of fast motion on range images acquired by lidar scanners for automotive applications. IEEE Transactions on Signal Processing 55 (2007), 2945–2953.

[254] *Gröll, L.; Matthes, J.:* Quellenlokalisierung mit räumlich verteilten, punktweisen Konzentrationsmessungen. Proc. GMA-Fachausschuss 1.30: Modellbildung, Identifikation und Simulation in der Automatisierungstechnik; Workshop in Bosen, 2005, 89–104.

[255] *Gröll, L.; Schilasky, R.; Bretthauer, B.:* Identifikation von Mehrzonenöfen. Proc. Fachtagung „Moderne Methoden des Regelungs- und Steuerungsentwurfs“, 20.-21.3.1997 Magdeburg, 50–57.

[256] *Grötschel, M.:* Lineare Optimierung, 2003. http://www.zib.de/groetschel/teaching/skriptADMII.pdf.

[257] *Gruyitch, L.; Richard, J.-P.; Borne, P.; Gentina, J.-C.:* Stability domains. Chapman & Hall, 2004.

[258] *Gu, K.; Kharitonov, V.L.; Chen, J.:* Stability of time-delay systems. Boston: Birkhäuser, 2003.

[259] *Gu, M.; Mengi, E.; Overton, M.L.; Xia, J.; Zhu, J.:* Fast methods for estimating the distance to uncontrollability. SIAM J. Mat. Anal. Appl. 28 2, 477–502.

[260] *Guckenheimer, J.; Holmes, P.:* Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. New York: Springer-Verlag, 1983.

[261] *Guo, L.; Tomizuka, M.:* Parameter identification with derivative shift operator parametrization. Automatica 35 (1999) 6, 1073–1080.

[262] *Gurel, O.; Lapidus, L.:* A guide to the generation of Lyapunov functions. Industrial and Engineering Chemistry 61 (1969) 3, 30–41.

[263] *Gutierrez, C.:* A solution of the bidimensional global asymptotic stability conjecture. Ann. Inst. H. Poincaré Anal. Non Linéaire 12 (1995), 627–671.

[264] *Haas, V.:* A stability result for a third order nonlinear differential equation. London Math. Soc. 40 (1965), 31–33.

[265] *Haddad, W.M.; Chellaboina, V.S.:* Nonlinear dynamical systems and control. Princeton University Press, 2008.

[266] *Hadjiasavvas, N.:* Generalized convexity, generalized monotonicity and nonsmooth analysis. In: Handbook of generalized convexity and monotonicity (Hrsg.: Hadjiasavvas, N.; Komlósi, S.; Schaible, S.), New York: Springer-Verlag 2005, 465–500.

[267] *Hadjiasavvas, N.; Schaible, S.:* Generalized monotone multi-valued maps. In: Encyclopedia of optimization, Vol. 2 (Hrsg.: Floudas, C.; Pardalos, P.), Dordrecht: Kluwer Academic Publishers 2001, 224–229.

[268] *Hahn, W.:* Stability of motion. New York: Springer, 1963.

[269] *Hakimi-M, M.; Khaloozadeh, H.:* Revision on the frequency domain conditions for strict positive realness. Int. J. of Control, Automation, and Systems 5 (2007) 1, 1–7.

[270] *Halkin, H.:* Implicit functions and optimization problems without continuous differentiability of the data. SIAM J. Control Optim. 12 (1974), 229–236.

[271] *Halperin, I.:* The product of projection operators. Acta Sci. Math. 23 (1962), 96–99.

[272] *Han, S.-P.:* A successive projection method. Mathematical Programming 40 (1988), 1–14.

[273] *Hanson, D.L.; Pledger, G.:* Consistency in concave regression. Ann. Statist. 4 (1976), 1038–1050.

[274] *Hartman, P.:* On stability in the large for systems of ordinary differential equations. Canad. J. Math. 13 (1961), 480–492.

[275] *Hartman, P.:* Ordinary differential equations. Baltimore: Hartmann, 1973.

[276] *Hautus, M.L.J.:* Controllability and observability conditions of linear autonomous systems. Indagationes Mathematicae 31 (1969), 443–448.

[277] *He, C.; Watson, G.A.:* An algorithm for computing the distance to instability. SIAM J. Matrix Anal. Appl. 20 (1998) 1, 101–116.

[278] *Henrion, D.; Korda, M.:* Convex computation of the region of attraction of polynomial control systems. IEEE Trans. on Automatic Control 59 (2014) 2, 297–312.

[279] *Henrion, D.; Lasserre, J.B.:* Solving global optimization problems over polynomials with GloptiPoly 2.1. Proc. International Workshop on Global Constrained Optimization and Constraint Satisfaction (Cocos'02), Sophia Antipolis, France, October 2-4, 2002.

[280] *Henrion, D.; Tarbouriech, S.; Arzelier, D.:* LMI approximations for the radius of the intersection of ellipsoids: A survey. Journal of Optimization Theory and Applications 108 (2001) 1, 1–28.

[281] *Hernandez, C.N.; Banks, S.P.:* A generalization of Lyapunov's equation to nonlinear systems. Proc. 6th IFAC Symposium in Stuttgart, Nonlinear Control Systems, 2004, 745–750.

[282] *Hertz, D.:* The extreme eigenvalues and stability of real symmetric interval matrices. IEEE Transactions on Automatic Control 37 (1992) 4, 532–535.

[283] *Higham, N.J.:* Computing a nearest symmetric positive semidefinite matrix. Linear Algebra and its Applications 103 (1988), 103–118.

[284] *Higham, N.J.:* The symmetric Procrustes problem. BIT 28 (1988), 133–143.

[285] *Higham, N.J.:* Computing the nearest correlation matrix. Technical report, University of Manchester, Departments of Mathematics, Numerical Analysis Report 369, 2000.

[286] *Hilbert, D.:* Über die Darstellung definiter Formen als Summe von Formenquadraten. Math. Ann. 32 (1888), 342–350.

[287] *Hindi, H.:* A tutorial on convex optimization. Proc. American Control Conference, Boston, Massachusetts, 2004, 3252–3265.

[288] *Hinrichsen, D.; Plischke, E.; Pritchard, A.J.:* On the transient behaviour of stable linear systems. Proc. 6th ECC, Porto, CD-ROM, 2001.

[289] *Hinrichsen, D.; Pritchard, A.J.:* Destabilization by output feedback. Differential and Integral Equations 5 (1992), 357–386.

[290] *Hinrichsen, D.; Pritchard, A.J.:* Mathematical system theory I. Berlin: Springer-Verlag, 2005.

[291] *Hinrichsen, D.; Son, N.K.:* The complex stability radius of discrete-time system. Proc. 28th Conference on Decision and Control, Tampa, Florida, 1989.

[292] *Hirsch, M.W.; Smale, S.; Devaney, R.L.:* Differential equations, dynamical systems, an introduction to chaos. Academic Press, 2004.

[293] *Hoagg, J.B.; Bernstein, D.S.:* Nonminimum-phase zero. IEEE Control Systems Magazine 45 (2007), 45–57.

[294] *Hochstenbach, M.E.; Muhič, A.; Plestenjak, B.:* On linearizations of the quadratic two-parameter eigenvalue problems. Linear Algebra and its Applications 436 (2012), 2725–2743.

[295] *van den Hof, J.M.:* Realization of positive linear systems. Linear Algebra and its Applications 256 (1997), 287–308.

[296] *van den Hof, J.M.:* Structural identifiability of linear compartmental systems. IEEE Transactions on Automatic Control 43 (1998) 6, 800–818.

[297] *Hollot, C.V.:* On Markov's theorem: Its like Kharitonov's but twice as nice. Proc. 27th CDC, Austin, Texas, 1988, 515–518.

[298] *Holmes, R.B.:* A course on optimization and best approximation. Berlin: Springer-Verlag, 1972.

[299] *Horn, R.A.; Johnson, C.R.:* Matrix analysis. Cambridge: Cambridge University Press, 1985.

[300] *Horn, R.A.; Johnson, C.R.:* Topics in matrix analysis. Cambridge: Cambridge University Press, 1999.

[301] *Hu, G.; Davison, E.J.:* Real controllability/stabilizability radius of LTI systems. IEEE Transactions on Automatic Control 49 (2004) 2, 254–257.

[302] *Hubbard, J.H.; West, B.H.:* Differential equations: A dynamical systems approach. New York: Springer-Verlag, 1995.

[303] *van Huffel, S.; Vandewalle, J.:* The total least squares problem: Computational aspects and analysis. Philadelphia: SIAM, 1991.

[304] *Hui, S.; Lillo, W.E.; Żak, S.H.:* Solving minimum norm problems using penalty functions and the gradient method. Automatica 31 (1995) 1, 115–124.

[305] *Husty, M.; Karger, A.; Sachs, H.; Steinhilper, W.:* Kinematik und Robotik. Berlin: Springer-Verlag, 1997.

[306] *Hynkova, H.:* Analyse von Zugängen zur Längsregelung autonomer Fahrzeuge. Diplomarbeit, Institut für Angewandte Informatik / Automatisierungstechnik am Karlsruher Institut für Technologie, 2007.

[307] *Ilchmann, A.:* Contributions to time-varying linear control systems. Ammersbek b. Hamburg: Verlag an der Lottbek, 1989.

[308] *Ilchmann, A.:* Algebraic theory of time-varying linear systems: A survey. In: Selected plenaries, milestones and surveys, Eds.: P. Horacek, M. Simandl, P. Zitek; Prague, Czech Republic 2005, 312–318.

[309] *Imbert, J.-L.:* Fourier's elimination: Which to choose? Proc. PPCP, 1993, 117–129.

[310] *Imbert, J.L.:* About redundant inequalities generated by Fourier's algorithm. Proc. Fourth International Conference on Artificial Intelligence, AIMSA'90, Varna, Bulgaria, North Holland. 1990, 117–127.

[311] *Ioannou, P.A.; Sun, J.:* Robust adaptive control. New Jersey: Prentice-Hall, 1995.

[312] *Isermann, R.:* Schätzung physikalischer Parameter für dynamische Prozesse. Automatisierungstechnik 39 (1991) 9/10, 323–328, 371–375.

[313] *Isermann, R.:* Identifikation dynamischer Systeme: Besondere Methoden, Anwendungen. Berlin: Springer-Verlag, 1992.

[314] *Isermann, R.:* Identifikation dynamischer Systeme: Grundlegende Methoden. Berlin: Springer-Verlag, 1992.

[315] *Isidori, A.:* Nonlinear control systems. New York: Springer-Verlag, 1995.

[316] *Jäkel, J.; Gröll, L.; Mikut, R.:* Bewertungsmaße zum Generieren von Fuzzy-Regeln unter Beachtung linguistisch motivierter Restriktionen. Proc. 8. Workshop Fuzzy Control des GMA-FA 5.22, 1998, 15–28.

[317] *Jang, S.S.; Joseph, B.; Mukai, H.:* Comparison of two approaches to on-line parameter and state estimation of nonlinear systems. Industrial & Engineering Chemistry Process Design and Development 25 (1986) 3, 809–814.

[318] *Jansson, M.; Ottersten, B.:* Structured covariance matrix estimation: A parametric approach. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000, 3172–3175.

[319] *Jarlebring, E.; Kvaal, S.; Michiels, W.:* Computing all pairs (λ, μ) such that λ is a double eigenvalue of $A + \mu B$. SIAM J. Matrix Analysis & Applications 32 (2011) 3, 902–927.

[320] *Jarre, F.; Stoer, J.:* Optimierung. Berlin: Springer-Verlag, 2004.

[321] *Jaulin, L.; Kieffer, M.; Didrit, O.; Walter, E.:* Applied interval analysis. London: Springer-Verlag, 2003.

[322] *Jeyakumar, V.; Luc, D.T.:* Approximate Jacobian matrices for nonsmooth continuous maps and C^1-optimization. SIAM Journal on Control and Optimization 36 (1998) 5, 1815–1832.

[323] *Jiang, Z.-P.; Teel, A.; Praly, L.:* Small-gain theorem for ISS systems and applications. Math. of Control, Signals, and Systems 7 (1994), 104–130.

[324] *Johnes, R.H.:* Maximum likelihood fitting of ARMA models to time series with missing observations. Technometrics 20 (1980), 389–395.

[325] *Kagiyama, M.; Kajiya, F.; Hoki, N.; Imamura, M.; Tomonaga, G.:* Identification of parameters in compartmental system when the structure is given beforehand. Systems, Computers, Controls 11 (1980) 1, 82–90.

[326] *Kågström, B.:* Singular matrix pencils. In: Templates for the solution of algebraic eigenvalue problems, SIAM 2000, 260–276.

[327] *Kahan, W.:* Numerical linear algebra. Canadian Math. Bull. 9 (1966), 757–801.

[328] *Kailath, T.:* Linear systems. Prentice Hall, 1980.

[329] *Kakutani, S.:* A generalization of Brouwer's fixed point theorem. Duke Mathematical Journal 8 (1941) 3, 457–459.

[330] *Kale, A.A.; Tits, A.L.:* On Kharitonov's theorem without invariant degree assumption. Automatica 36 (2000) 7, 1075–1076.

[331] *Kalman, R.E.:* On the general theory of control systems. Proc. First IFAC Congress, Moscow, 1960, volume 1, 1961, 481–492.

[332] *Kamaleddin, S.; Nikravesh, Y.:* Nonlinear systems stability analysis. CRC Press, 2013.

[333] *Kamke, E.:* Differentialgleichungen: Lösungsmethoden und Lösungen. Chelsea Pub Co, 1974.

[334] *Kanungo, T.; Gay, D.M.; Haralick, R.M.:* Constrained monotone regression of ROC curves and histograms using splines and polynomials. Proc. International Conference on Image Processing, Washington, DC, Oct. 23-26, 1995 (Hrsg.: IEEE Signal Processing Society, Los Alamitos, C.), 1995, 292–295.

[335] *Kappos, E.:* A global geometrical input-output linearization theory. IMA Journal of Mathematical Control & Information 9 (1992), 1–21.

[336] *Karjalainen, M.; Esquef, P.A.A.; Antsalo, P.; Mäkivirta, A.; Välimäki, V.:* AR/ARMA analysis and modeling of modes in resonant and reverberant systems. Proc. 112th AES Convention, Convention Paper 5590, 2002.

[337] *Karmarkar, N.:* A new polynomial time algorithm for linear programming. Combinatorica 4 (1984), 373–395.

[338] *Karow, M.; Kressner, D.:* On the structured distance to uncontrollability. Technical report, Eidgenössische Technische Hochschule; Zürich, Switzerland, 2008.

[339] *Karow, M.; Mengi, E.:* Matrix polynomials with specified eigenvalues. Numerical Analysis (2013).

[340] *Kato, J.:* Uniformly asymptotic stability and total stability. Tohoku Math. J. 22 (1970), 254–269.

[341] *Kaufman, L.:* A variable projection method for solving separable nonlinear least squares problems. BIT Numerical Mathematics 15 (1975), 49–57.

[342] *Kaufman, L.; Sylvester, G.:* Separable nonlinear least squares with multiple right-hand sides. SIAM J. Matrix Anal. Appl. 13 (1992), 68–89.

[343] *Keith, W.L.; Rand, R.H.:* Dynamic of a system exhibiting the global bifurcation of a limit circle at infinity. Int. J. Non-Linear Mechanics 20 (1985) 4, 325–338.

[344] *Kelemen, M.:* A stability property. IEEE Trans. on Automatic Control 31 (1986), 766–768.

[345] *Khalil, H.K.:* Nonlinear systems. Prentice Hall, 1996.

[346] *Kharitonov, V.L.:* Asymptotic stability of an equilibrium position of a family of systems of linear differential equations (in russian). Differensial'nye Uravenya 14 (1978), 2086–2088.

[347] *Kielbasinski, A.; Schwetlick, H.:* Numerische lineare Algebra. Berlin: Deutscher Verlag der Wissenschaften, 1988.

[348] *Kiendl, H.:* Totale Stabilität von linearen Regelungssystemen bei ungenau bekannten Parametern der Regelstrecke. Automatisierungstechnik 33 (1985) 12, 379–386.

[349] *Kiers, H.A.L.; ten Berge, J.M.F.:* Alternating least squares algorithms for simultaneous component analysis with equal component weight matrices in two or more populations. Psychometrika 54 (1989), 467–473.

[350] *Kiet, T.T.; Phat, V.N.:* Lyapunov stability of nonlinear time-varying differential equations. Acta Mathematica Vietnamica 25 (2000) 2, 231–249.

[351] *van der Kloet, P.; Neerhoff, F.L.:* Dynamic eigenvalues for scalar linear time-varying systems. Proc. International Symposium on Mathematical Theory of Networks and Systems, Notre Dame, Indiana, U.S.A., 12 - 16 Aug, 2002, 1–8. citeseer.ist.psu.edu/567138.html.

[352] *Koditschek, D.E.; Narendra, K.S.:* The stability of second-order quadratic differential equations. IEEE Transactions on Automatic Control 27 (1982) 4, 783–798.

[353] *Kohler, D.A.:* Projection of convex polyhedral sets. PhD thesis, University of California, Berkeley, 1967.

[354] *Kojima, M.; Tunçel, L.:* Cones of matrices and successive convex relaxations of nonconvex sets. SIAM J. Optim. 10 (2000) 3, 750–778.

[355] *Kokotović, P.; Khalil, H.K.; O'Reilly, J.:* Singular Perturbation Methods in Control. SIAM, 1999.

[356] *Korn, U.; Wilfert, H.-H.:* Mehrgrößenregelungen. Berlin: VEB Verlag Technik, 1982.

[357] *Kottenstette, N.; Antsaklis, P.J.:* Relationships between positive real, passive dissipative and positive systems. Proc. American Control Conference, Baltimore, MD, USA, 2010.

[358] *Kovalev, A.M.:* The construction of Lyapunov functions with sign-definite derivative for systems satisfying the Barbashin-Krasovskii theorem. Journal of Applied Mathematics and Mechanics 72 (2008), 164–168.

[359] *Kočvara, M.; Stingl, M.:* PENNON – A code for convex nonlinear and semidefinite programming. Optimization Methods and Software 8 (2003), 317–333. http://www.penopt.com.

[360] *Kowalczuk, Z.; Kozłowski, J.:* Continuous-time approaches to identification of continuous-time systems. Automatica 36 (2000), 1229–1236.

[361] *Kowalski, K.; Steeb, W.:* Nonlinear dynamical systems and Carleman linearization. Singapore: World Scientific Publishing, 1991.

[362] *Kozlov, V.V.; Furta, S.D.:* Asymptotic solutions of strongly nonlinear systems of differential equations. New York: Springer-Verlag, 2013.

[363] *Kreil, W.F.:* Stabilitätsverfahren für lineare zeitvariable Systeme. Fortschrittsberichte der VDI Zeitschriften, Reihe 8, Nr. 16, Düsseldorf: VDI-Verlag, 1974.

[364] *Kressner, D.:* Deflation in Krylov subspace methods and distance to uncontrollability. Annali dell'Universita di Ferrara 53 (2007) 2, 309–318.

[365] *Krislock, N.G.B.:* Numerical solution of semidefinite constrained least squares problems. PhD thesis, The University of British Columbia, 2003. www.math.uregina.ca/ krislock/.

[366] *Krokavec, D.; Filasová, A.:* Equivalent representations of bounded real lemma. Proc. 18th Int. Confr. on Process Control, 2011.

[367] *Krstić, M.; Kanellakopoulos, I.; Kokotović, P.:* Nonlinear and adaptive control design. New York: John Wiley & Sons, 1995.

[368] *Kruger, C.J.:* Constrained cubic spline interpolation, 2003. http://www.korf.co.uk/spline.pdf.

[369] *Kuhnen, K.:* Kompensation komplexer gedächtnisbehafteter Nichtlinearitäten in Systemen mit aktiven Materialien. Habilitationsschrift, Universität des Saarlandes, Saarbrücken, Shaker Verlag Aachen, 2008.

[370] *Kurcveĭl, J.:* On the inversion of Lyapunov's second theorem on the stability of motion. American Mathematical Society Translations, Series 2 24 (1956), 19–77.

[371] *Kuznetsov, N.V.:* Stability and oscillations of dynamical systems. PhD thesis, University of Jyväskylä, 2008.

[372] *Kuznetsov, Y.:* Elements of Applied Bifurcation Theory. Springer: Applied Mathematical Sciences, Vol. 112, 2004.

[373] *Kwatny, H.G.; Chang, B.-C.; Wang, S.-P.:* Static bifurcation in mechanical control systems. In: Bifurcation control – Theory and applications (Hrsg.: Chen, G.; Hill, D.J.; Yu, X.), Berlin: Springer-Verlag 2003.

[374] *La Salle, J.P.:* The stability of dynamical systems. Society for Industrial and Applied Mathematics, Philadelphia, 1976.

[375] *La Salle, J.P.:* The stability and control of discrete processes. Berlin: Springer-Verlag, 1986.

[376] *La Torre, D.:* Necessary optimality conditions for nonsmooth optimization problems. Technical report, UNIMI Department of Economics No. 21.2002, 2002.

[377] *Laffey, T.J.; Šmigoc, H.:* Common Lyapunov solutions for two matrices whose difference has rank one. Linear Algebra and its Applications 431 (2009), 228–240.

[378] *Lagerberg, A.:* A literature survey on control of automotive powertrains with backlash. Technical report, Control and Automation Laboratory, Department of Signals and Systems, Chalmers University of Technology, 2001.

[379] *Lakshmikantham, V.; Leela, S.:* Differential and integral inequalities; theory and applications. London: Academic Press, 1969.

[380] *Lakshmikantham, V.; Matrosov, V.M.; Sivasundaram, S.:* Vector Lyapunov functions and stability analysis of nonlinear systems. Kluwer Academic Publishers, 1991.

[381] *Lam, S.; Davison, E.J.:* The transmission zero at s radius and the minimum phase radius of LTI systems. Proc. 17th IFAC World Congress, 2008, 6371–6376.

[382] *Lancaster, P.; Tismenetsky, M.:* The theory of matrices. San Diego: Academic Press, 1985.

[383] *Langson, W.; Alleyne, A.:* A stability result with application to nonlinear regulation. J. Dyn. Syst. Meas. Control 124 (2002), 452–456.

[384] *Lasserre, J.B.:* Polynomials with all zeros real and in a prescribed interval. Journal of Algebraic Combinatorics 16 (2002) 3, 231–237.

[385] *Lauritzen, N.:* Lectures on convex sets. Technical report, Aarhus University, 2010.

[386] *Lee, T.C.; Liaw, D.-C.; Chen, B.S.:* A general invariance principle for nonlinear time-varying systems. IEEE Transactions on Automatic Control 46 (2001) 12, 1989–1993.

[387] *de Leeuw, J.:* Block-relaxation algorithms in statistics. In: Information Systems and Data Analysis, Springer Berlin Heidelberg 1994, 308–324.

[388] *Lemmen, M.; Jelali, M.:* Differentialgeometrische Steuer- und Beobachtbarkeitsanalyse nichtlinearer Systeme. Technical Report 8/96, Gerhard-Mercator-Universität-GH Duisburg, Meß-, Steuer- und Regelungstechnik, 1996.

[389] *Lemmerling, P.; Vanhamme, L.; van Huffel, S.; de Moor, B.:* IQML-like algorithms for solving structured total least squares problems: A unified view. Signal Processing 81 (2001), 1935–1945.

[390] *Leonov, G.A.:* On stability in the first approximation. J. of Appl. Mathem. and Mech. 62 4, 511–517.

[391] *Leonov, G.A.:* Strange attractors and classical stability theory. St. Petersburg University Press, 2008.

[392] *Leonov, G.A.; Bragin, V.O.; Kuznetsov, N.V.:* Algorithm for constructing counterexamples to the Kalman problem. Doklady Mathematics 82 (2010) 1, 540–542.

[393] *Leonov, G.A.; Kuznetsov, N.V.:* Time-varying linearization and the Perron effects. International Journal of Bifurcation and Chaos 17 (2007) 4, 1079–1107.

[394] *Lev-Ari, H.; Bistritz, Y.; Kailath, T.:* Generalized Bezoutians and families of efficient zero-location procedures. IEEE Transactions on Circuits and Systems 38 (1991) 2, 170–186.

[395] *Li, H.; Stoica, P.; Li, J.:* Computationally efficient maximum likelihood estimation of structured covariance matrices. IEEE Transactions on Signal Processing 47 (1999) 5, 1314–1323.

[396] *Li, R.:* Test positive realness of a general transfer function matrix. Technical report, University of Kentucky, 2000.

[397] *Li, X.; Zhang, Q.; Su, H.:* An adaptive observer for joint estimation of states and parameters in both state and output equations. International Journal of Adaptive Control and Signal Processing 25 (2011) 9, 831–842.

[398] *Li, Y.; Yang, H.:* A new Liu-type estimator in linear regression model. Statistical Papers 53 (2012) 2, 427–437.

[399] *Liao, X.; Wang, L.Q.; Yu, P.:* Stability of dynamical systems. Amsterdam: Elsevier, 2007.

[400] *Liao, X.; Yu, P.:* Absolute stability of nonlinear control systems. Springer Science, 2008.

[401] *Liberzon, D.; Morse, A.S.; Sontag, E.D.:* A new definition of the minimum-phase property for nonlinear systems, with an application to adaptive control. Proc. Decision and Control, Sydney, Australia, 2000, 2106–2111.

[402] *Lippert, R.A.:* Fixing two eigenvalues by a minimal perturbation. Linear Algebra and its Applications 406 (2005), 177–200.

[403] *Lippert, R.A.:* Fixing multiple eigenvalues by a minimal perturbation. Linear Algebra and its Applications 432 (2010), 1785–1817.

[404] *Lippert, R.A.; Edelman, A.:* The computation and sensitivity of double eigenvalues. In: Lecture Notes in Pure and Applied Mathematics, Advances in Computational Mathematics 202, Dekker 1998, 353–393.

[405] *Litz, L.:* Reduktion der Ordnung linearer Zustandsraummodelle mittels modaler Verfahren. Stuttgart: Hochschulverlag, 1979.

[406] *Ljung, L.:* System Identification – Theory for the user. Upper Saddle River: PTR Prentice Hall, 1999.

[407] *van Loan, C.:* How near is a stable matrix to an unstable. Contemporary Mathematics, Linear algebra and its role in system theory 47 (1985), 465–478.

[408] *van Loan, C.:* On the method of weighting for equality-constrained least-squares problems. SIAM J. Numer. Anal. 22 (1985) 5, 851–864.

[409] *van Loan, C.F.:* The ubiquitous Kronecker product. Journal of Computational and Applied Mathematics 123 (2000), 85–100.

[410] *van Loan, C.F.; Pitsianis, N.P.:* Approximation with Kronecker products. Proc. Linear Algebra for Large Scale and Real Time Applications, Kluwer Publications, 1993, 293–314.

[411] *Lobo, M.S.; Vandenberghe, L.; Boyd, S.; Lebret, H.:* Applications of second-order cone programming. Linear Algebra and its Applications 284 (1998), 193–228.

[412] *López-Valcarce, R.; Dasgupta, S.:* A new proof for the stability of equation-error models. IEEE Signal Processing Letters 6 (1999) 6, 148–150.

[413] *Loría, A.; Panteley, E.; Popović, D.; Teel, A.R.:* A nested Matrosov theorem and persistency of excitation for uniform convergence in stable nonautonomous systems. IEEE Transactions on Automatic Control 50 (2005) 2, 183–198.

[414] *Lückel, J.; Müller, P.C.:* Analyse von Steuerbarkeits-, Beobachtbarkeits- und Störbarkeitsstrukturen linearer, zeitinvarianter Systeme. Regelungstechnik 23 (1975), 357–362.

[415] *Ludyk, G.:* Stability of time-variant discrete-time systems. Braunschweig, Wiesbaden: Vieweg, 1985.

[416] *Ludyk, G.:* CAE von dynamischen Systemen. New York, Berlin, Heidelberg: Springer-Verlag, 1990.

[417] *Luenberger, D.G.:* Optimization by vector space methods. New York: John Wiley & Sons, 1969.

[418] *Luenberger, D.G.:* Introduction to dynamic systems. New York: John Wiley & Sons, 1979.

[419] *Luenberger, D.G.; Ye, Y.:* Linear and nonlinear programming. New York: Springer-Verlag, 2008.

[420] *Luo, Z.-Q.; Zhang, S.:* On the extension of Frank-Wolfe theorem. Technical report, Erasmus University Rotterdam, The Netherlands, 1997.

[421] *Magnus, J.R.; Neudecker, H.:* Matrix differential calculus with applications in statistics and econometrics. New York: John Wiley & Sons Ltd., 1988.

[422] *Magnus, J.R.; Neudecker, H.:* Matrix differential calculus with applications in statistics and econometrics. Chichester: John Wiley & Sons, 1988.

[423] *Mahata, K.; Söderström, T.:* Improved estimation performance using known linear constraints. Automatica 40 (2004), 1307–1318.

[424] *Malek-Zavarei, M.:* The stability of linear time-varying systems. Int. J. Control 27 (1978) 5, 809–815.

[425] *Malisoff, M.; Mazenc, F.:* Construction of strict Lyapunov functions. London: Springer-Verlag, 2009.

[426] *Malkin, I.G.:* Stability in the case constantly acting disturbances. Prikl. Mat. Mekh. 8 (1944), 241–245.

[427] *Malyshev, A.N.:* On Wilkinson's problem. Technical report, Dept. of Informatics, University of Bergen, Norway, 1997.

[428] *Mangasarian, O.L.:* Pseudo-convex functions. SIAM Series A 3 (1965) 2, 281–290.

[429] *Manton, J.H.; Mahony, R.; Hua, Y.:* The geometry of weighted low rank approximations. IEEE Transactions on Signal Processing 51 (2003) 2, 500–514.

[430] *Maodong, Y.:* Near minimax polynomial approximation. Appl. Math.-JCU 13B (1998), 117–122.

[431] *Marcus, M.; Minc, H.:* A survey of matrix theory and matrix inequalities. New York: Dover Publications, 1992.

[432] *Marden, M.:* Geometry of polynomials. In: Mathematical surveys, Vol. 3, American Mathematical Society 1966, 194–206.

[433] *Mardia, K.V.; Kent, J.T.; Bibby, J.M.:* Multivariate Analysis. London: Academic Press, 1995.

[434] *Marí, J.; Stoica, P.; McKelvey, T.:* Vector ARMA estimation: A reliable subspace approach. IEEE Transactions on Signal Processing 48 (2000) 7, 2092–2104.

[435] *Markovsky, I.:* Low rank approximation. London: Springer-Verlag, 2012.

[436] *Markus, L.:* Continuous matrices and the stability of differential systems. Mathematische Zeitschrift 62 (1955), 310–319.

[437] *Markus, L.; Yamabe, H.:* Global stability criteria for differential systems. Osaka Math. J. 12 (1960), 305–317.

[438] *Marmorat, J.-P.; Olivi, M.; Hanzon, B.; Peeters, R.L.M.:* Matrix rational H^2 approximation: A state-space approach using Schur parameters. Proc. CDC02, Las-Vegas, Nevada, USA, 2002, 4244–4249.

[439] *Maronna, R.A.; Martin, D.R.; Yohai, V.J.:* Robust statistics: Theory and methods. New York: Wiley, 2006.

[440] *Martelli, M.:* Global stability of stationary states of discrete dynamical systems. Ann. Sci. Québec 22 (1998) 2, 201–212.

[441] *Marzetta, Th.L.:* A simple derivation of the constrained multiple parameter Cramér-Rao bound. IEEE Transactions on Signal Processing 41 (1993) 6, 2247–2249.

[442] *Massera, J.L.:* Contributions to stability theory. Annals of Mathematics 64 (1956) 1, 182–206.

[443] *Mathias, R.:* A chain rule for matrix functions and applications. SIAM J. Matrix Anal. Appl. 17 (1996), 610–620.

[444] *Matrosov, V.M.:* On the stability of motion. J. Appl. Math. Mech. 26 (1962), 1337–1353.

[445] *Mattheij, R.M.M.; Söderlind, G.:* On inhomogeneous eigenvalue problems. Linear Algebra and its Applications 88 (1987), 507–531.

[446] *Matthes, J.; Gröll, L.; Keller, H.B.:* Source localization based on pointwise concentration measurements. Sensors and Actuators A 115 (2004), 32–37.

[447] *Mazenc, F.; Malisoff, M.:* Lyapunov function constructions for slowly time-varying systems. Proc. Decision and Control, San Diego, California, 2006, 5108–5113.

[448] *Mazenc, F.; Sepulchre, R.; Jankovic, M.:* Lyapunov functions for stable cascades and applications to global stabilization. Proc. Decision and Control, San Diego, California, December 1997, 2843–2846.

[449] *McClellan, J.H.; Lee, D.:* Exact equivalence of the Steiglitz-McBride iteration and IQML. IEEE Transactions on Signal Processing 39 (1991) 2, 509–512.

[450] *McGinnie, B.P.:* A balanced view of system identification. PhD thesis, Cambridge University (U.K.), 1994.

[451] *Mead, J.L.; Renaut, R.A.:* Least squares problems with inequality constraints as quadratic constraints. Linear Algebra and its Applications 432 (2010), 1936–1949.

[452] *Mengi, E.:* On the estimation of the distance to uncontrollability for higher order systems. SIAM J. of Matrix Analysis and Appl. 30 (2008) 1, 154–172.

[453] *Menrath, M.:* Stability criteria for nonlinear fully implicit differential-algebraic systems. PhD thesis, Universität Köln, 2011.

[454] *Meyer, C.D.:* Matrix analysis and applied linear algebra. Philadelphia: SIAM, 2000.

[455] *Michel, A.N.; Hou, L.; Liu, D.:* Stability of dynamical systems. Continuous, discontinuous and discrete systems. Birkhäuser, 2003.

[456] *Michler, O.:* Analyse linearer Regelungssysteme mit Unbestimmtheiten. Aachen: Shaker Verlag, 2000.

[457] *Minnichelli, R.; Anagnost, J.; Desoer, C.:* An elementary proof of Kharitonov's theorem with extensions. IEEE Transaction on Automatic Control 34 (1989), 995–998.

[458] *Mittal, S.K.; Chandra, D.; Dwivedi, B.:* The effect of time-moments and Markov-parameters on reduced-order modeling. ARPN Journal on Engineering and Applied Sciences 4 (2009) 5, 8–14.

[459] *Müller, P.H.:* Lexikon der Stochastik, 5. Auflage. Berlin: Akademie-Verlag, 1991.

[460] *de Moor, B.; Gevers, M.; Goodwin, G.:* Overbiased, underbiased and unbiased estimation of transfer functions. Proc. 9th IFAC/IFORS Symposium on System Identification, Budapest, Hungary, 1991, 946–951.

[461] *Moroşanu, G.; Vladimirescu, C.:* Stability for a nonlinear second order ODE. Funkcialaj Ekvacioj 48 (2005), 49–56.

[462] *Morse, A.S.:* Towards a unified theory of parameter adaptive control, part II: Certainty equivalence and implizit tuning. IEEE Transactions on Automatic Control 37 (1992), 15–29.

[463] *Moses, R.L.; Liu, D.:* Determining the closest stable polynomial to an unstable one. IEEE Transactions on Signal Processing 39 (1991) 4, 901–906.

[464] *Motzkin, T.S.:* The arithmetic-geometric inequality. In: Inequalities (Proc. Sympos. Wright-Patterson Air Force Base, Ohio, 1965), New York: Academic Press 1967, 205–224.

[465] *Muhammad, S.; van der Woude, J.:* A counter example to a recent result on the stability of non-linear systems. IMA Journal of Mathematical Control and Information 26 (2009), 319–323.

[466] *Muhič, A.; Plestenjak, B.:* On the singular two-parameter eigenvalue problem. Electron. J. Linear Algebra 18 (2009), 420–437.

[467] *Mukherjee, R.; Chen, D.:* Asymptotic stability theorem for autonomous systems. J. of Guidance, Control, and Dynamics 16 (1993) 5, 961–963.

[468] *Mukhopadhyay, S.; Patra, A.; Rao, G.P.:* New class of discrete time models for continuous time systems. International Journal of Control 55 (1992) 5, 1161–1187.

[469] *Müllhaupt, P.:* Analysis and control of underactuated mechanical nonminimum-phase systems. PhD thesis, École Polytechnique Fédérale de Lausanne, 1999.

[470] *Myers, R.H.:* Classical and modern regression with applications. Belmont: Duxbury Press, 1990.

[471] *Nagesha, V.; Kay, S.:* On frequency estimation with the IQML algorithm. Transactions on Signal Processing 42 (1994) 9, 2509–2513.

[472] *Narendra, K.S.; Annaswamy, A.M.:* Persistent excitation in adaptive systems. Int. J. of Control 45 (1987), 127–160.

[473] *Narendra, K.S.; Taylor, J.H.:* Frequency domain criteria for absolute stability. New York, London: Academic Press, 1973.

[474] *Nemhauser, G.L.; Wolsey, L.A.:* Integer and combinatorial optimization. New York: Wiley, 1988.

[475] *Nesterov, Y.:* Introductory lectures on convex optimization. Dordrecht: Kluwer Academic Press, 2004.

[476] *Nesterov, Y.; Nemirovskii, A.:* Interior-point polynomial algorithms in convex programming. SIAM Studies in Applied Mathematics, Vol. 13, 2001.

[477] *Neter, J.; Kutner, M.; Nachtsheim, C.; Wasserman, W.:* Applied linear statistical models. 4th edition. Chicago: Irwin, 1996.

[478] *Neubacher, A.:* Another elementary proof of Kharitonov's theorem. Technical report, Bericht 97-19, Johannes Kepler Universität Linz, 1997.

[479] *Nijmeijer, H.; van der Schaft, A.J.:* Nonlinear dynamical control systems. Berlin: Springer-Verlag, 1990.

[480] *Nikolaev, Y.P.:* The set of stable polynomials of linear discrete systems: its geometry. Automation and Remote Control 63 (2002) 7, 1080–1088.

[481] *Nobile, A.G.; Ricciardi, L.M.; Sacerdote, L.:* On Gompertz growth model and related difference equation. Biological Cybernetics 42 (1982), 221–229.

[482] *Nussenzveig, H.M.:* Causality and dispersion relations. New York: Academic Press, 1972.

[483] *Nutter, F.W.:* Quantifying the temporal dynamics of plant virus epidemics: A review. Crop Protection 16 (1997) 7, 603–618.

[484] *Ober, R.J.:* Balanced parametrization of classes of linear systems. SIAM J. Contr. Optim. 29 (1991), 1251–1287.

[485] *Oja, P.:* Comonotone adaptive interpolating splines. BIT 42 (2002), 842–855.

[486] *de Oliveira, M.C.:* A robust version of the elimination lemma. Proc. 16th IFAC World Congress, volume 16, 2005.

[487] *Ortega, J.M.; Rheinboldt, W.C.:* Iterative solution of nonlinear equations in several variables. New York: Academic Press, 1970.

[488] *Osborne, E.E.:* On pre-conditioning of matrices. Journal of the ACM 7 (1960) 4, 338–345.

[489] *Osborne, M.R.:* Some special nonlinear least squares problems. SIAM Journal on Numerical Analysis 12 (1975) 4, 571–592.

[490] *van Overschee, P.; de Moor, B.:* Subspace identification for linear systems – Theory, implementation, applications. Kluwer Academic Publishers, 1996.

[491] *Overton, M.L.; van Dooren, P.:* On computing the complex passivity radius. Proc. 44th IEEE CDC and ECC '05, 2005, 7960–7964.

[492] *Padmanabhan, P.; Hollot, C.V.:* Stability of interval matrices. Proc. IFAC 12th Triennial World Congress, Sydney, Australia, 1993, 379–380.

[493] *Pai, M.A.:* Power system stability: Analysis by the direct method of Lyapunov. North-Holland, 1981.

[494] *Paige, C.C.:* Computer solution and perturbation analysis of generalized least squares problems. Math. Comp. 33 (1979), 171–184.

[495] *Paige, C.C.:* Properties of numerical algorithms related to computing controllability. IEEE Transactions on Automatic Control 26 (1981) 1, 130–138.

[496] *Palusinski, O.; Stern, P.; Wall, E.; Moe, M.:* Comments on "An energy metric algorithm for the generation of Liapunov functions". IEEE Transactions on Automatic Control 14 (1969) 1, 110–111.

[497] *Paoletti, S.; Juloski, A.L.; Ferrari-Trecate, G.; Vidal, R.:* Identification of hybrid systems: A tutorial. Eur. J. Control 513 (2007), 242–260.

[498] *Papachristodoulou, A.; Prajna, S.:* On the construction of Lyapunov functions using the sum of squares decomposition. Proc. Decision and Control, Las Vegas, Nevada, 2002, 3482–3487.

[499] *Papachristodoulou, A.; Prajna, S.:* Analysis of non-polynomial systems using the sum of squares decomposition. In: Positive polynomials in control, Berlin: Springer-Verlag 2005, 23–43.

[500] *Park, H.; Zhang, L.; Ben Rosen, J.:* Low rank approximation of a Hankel matrix by structured total least norm. BIT 39 (1999) 4, 757–779.

[501] *Parks, P.C.:* A.M. Lyapunov's stability theory – 100 years on. IMA Journal of Mathematical Control and Information 9 (1992), 275–303.

[502] *Parks, T.A.:* Reducible nonlinear programming problems. Technical report, Department of Mathematical Sciences. Rice University, Houston, Texas, Author's Ph.D. thesis, 1985.

[503] *Parring, A.-M.:* About the concept of the matrix derivative. Linear Algebra and its Applications 176 (1992), 223–235.

[504] *Pataki, G.; Tunçel, L.:* On the generic properties of convex optimization problems in conic form. Math. Program., Ser. A 89 (2001), 449–457.

[505] *Perron, O.:* Die Stabilitätsfrage bei Differentialgleichungen. Math. Zeitschrift 32 (1930), 703–728.

[506] *Phillips, G.M.:* Interpolation and approximation by polynomials. New York: Springer-Verlag, 2003.

[507] *de Pierro, A.R.; Wei, M.:* Some new properties of the equality constrained and weighted least squares problem. Linear Algebra and its Applications 320 (2000), 145–165.

[508] *Pintelon, R.; Schoukens, J.:* System Identification: A Frequency Domain Approach. New York: IEEE Press, 2001.

[509] *Plaschko, P.; Brod, K.:* Nichtlineare Dynamik, Bifurkation und Chaotische Systeme. Braunschweig: Friedr. Vieweg & Sohn Verlagsgesellschaft mbH, 1995.

[510] *Pola'nski, A.:* Lyapunov function construction by linear programming. IEEE Transactions on Automatic Control 42 (1997) 7, 1013–1016.

[511] *Pólya, G.; Szegö, G.:* Problems and theorems in analysis II. New York, Berlin, Heidelberg: Springer-Verlag, 1976.

[512] *Polyanin, A.D.; Zaitsev, V.F.:* Handbook of exact solutions for ordinary differential equations. New York: CRC Press, 1995.

[513] *Ponstein, J.:* Seven kinds of convexity. SIAM Rev. 9 (1967), 115–119.

[514] *Powers, V.; B., R.:* Polynomials that are positive on an interval. Trans. Amer. Math. Soc. 352 (2000), 4677–4692.

[515] *Powers, V.; Wörnmann, T.:* An algorithm for sums of squares of real polynomials. J. of Pure and Applied Algebra 127 (1998) 1, 99–104.

[516] *Pradeep, S.; Shrivastava, S.K.:* Some recent results on the stability of linear time varying systems. Sudhana 13 (1988), 157–167.

[517] *Price, M.G.; Cook, G.:* Identification/observation using an extended Lunberger observer. IEEE Trans. Industrial Electronics 29 (1982) 4.

[518] *Pritchard, A.J.:* Stability and stabilization of second-order systems. J. Inst. Maths. Appl. 7 (1971), 348–360.

[519] *Pucci, P.; Serrin, J.:* Remarks on Lyapunov stability. Differential and integral equations: An international journal for theory & application 8 (1995) 6, 1265–1278.

[520] *Pugh, A.C.; Ratcliffe, P.A.:* On the zeros and poles of a rational matrix. Int. J. Control 30 (1979) 2, 213–226.

[521] *Qui, L.; Bernhardsson, B.; Rantzer, A.; Davison, E.J.; Young, P.M.; Doyle, J.C.:* A formula for computation of the real stability radius. Automatica 31 (1995) 6, 879–890.

[522] *Quian, C.; Lin, W.:* Global stabilization of nonlinear systems: A continuous feedback framework. Proc. Nonlinear and adaptive control (Hrsg.: Zinober, A.; Owens, D.), volume 281, Springer-Verlag; LNCIS 281. 2003, 295–315.

[523] *Ramachandran, V.; Gargour, C.S.:* An implementation of a stability test of 1-d discrete system based on Schussler's theorem and some consequent coefficient conditions. J. Franklin Inst. 317 (1984) 5, 341–358.

[524] *Ramsay, J.O.:* Estimating smooth monotone functions. J. Royal Statistical Society B 60 (1998), 365–375.

[525] *Rantzer, A.:* On the Kalman-Yakubovich-Popov lemma. Systems & Control Letters 28 (1996), 7–10.

[526] *Rantzer, A.:* A dual to Lyapunov's stability theorem. Systems & Control Letters 42 (2001), 161–168.

[527] *Rao, C.R.:* Matrix approximations and reduction of dimensionality in multivariate statistical analysis. Proc. Multivariate Analysis-V, 1980, 3–22.

[528] *Rao, C.R.; Mitra, S.K.:* Generalized inverse of matrices and its applications. New York: John Wiley & Sons, 1971.

[529] *Rao, C.R.; Rao, M.B.:* Matrix algebra and its applications to statistics and econometrics. Singapore: World Scientific, 1998.

[530] *Rao, C.R.; Toutenburg, H.:* Linear models. Berlin: Springer-Verlag, 2nd edition, 1999.

[531] *Read, N.K.; Ray, W.H.:* Application of nonlinear dynamic analysis to the identification and control of nonlinear systems – I. Simple dynamics, II. More complex dynamics, III. n-dimensional systems. J. Process Control 8 (1998), 1–46.

[532] *Reinschke, K.:* Lineare Regelungs- und Steuerungstheorie. Berlin: Springer-Verlag, 2006.

[533] *Reis, T.; Stykel, T.:* Passivity-preserving model reduction of differential-algebraic equations in circuit simulation. Proc. Appl. Math. Mech. 7 (2007), 1021601–1021602.

[534] *Rinehart, R.F.:* The derivative of a matrix function. Proc. Amer. Math. Soc. 7 (1956), 2–5.

[535] *Rizzi, P.A.:* Microwave engineering: Passive circuits. Prentice Hall, 1988.

[536] *Röbenack, K.:* Structure matters – some notes on high gain observer design for nonlinear systems. Proc. Proc. of the 9th International Multi-Conference on Systems, Signals and Devices, Chemnitz, Germany 2012.

[537] *Robertson, T.; Wright, F.T.; Dykstra, R.L.:* Order restricted statistical inference. New York: John Wiley & Sons, 1988.

[538] *Rockafellar, R.T.:* Convex analysis. Princeton: Princeton University Press, 1970.

[539] *Rockafellar, R.T.:* Lagrange multipliers and optimality. SIAM Review 35 (1993) 2, 183–238.

[540] *Rogers, G.S.:* Matrix derivatives. Lecture Notes in Statistics, New York: Marcel Dekker Inc., 1980.

[541] *Rohn, J.:* Positive definiteness and stability of interval matrices. SIAM J. on Matrix Analysis and Applications 15 (1994) 1, 175–184.

[542] *Roussel, M.R.:* Chemistry 5850: Nonlinear dynamics, lecture 8, 2005. http://people.uleth.ca/ roussel/nld/.

[543] *Rugh, W.J.:* Linear system theory. Prentice Hall, 2nd edition, 1996.

[544] *Rump, S.M.:* Structured perturbations part II: componentwise distances. SIAM J. Matrix Analysis and Applications 25 (2003) 1, 31–56.

[545] *Rumyantsev, V.V.:* On stability of motions with respect to part of variables. Vestnik of Moscow University, Serie Math. and Mech. 4 (1957), 9–16.

[546] *Rumyantsev, V.V.:* A comparison of three methods of constructing Lyapunov functions. J. Appl. Maths. Mechs. 59 (1995) 6, 873–877.

[547] *Saccomani, M.P.:* Linearization in the parameters via differential algebra techniques. Proc. 13th IFAC-Symposium on System Identification, Rotterdam, Netherlands, 2003, 1246–1251.

[548] *Sastry, S.:* Nonlinear systems. New York: Springer-Verlag, 1999.

[549] *Scharf, L.L.:* Statistical signal processing. Addison-Wesley Publishing Company, 1991.

[550] *Schlittgen, R.; Streitberg, B.H.J.:* Zeitreihenanalyse. München, Wien: Oldenbourg, 2001.

[551] *Schmid, R.; Pandey, A.:* The role of nonminimum phase zeros in the transient response of multivariable systems. Proc. 50th CDC-EDC, 2011, 471–475.

[552] *Schrader, C.B.; Sain, M.K.:* Research on system zeros: A survey. Proc. Decision and Control, Austin, Texas, 1988, 890–901.

[553] *Schuermans, M.; Lemmerling, P.; van Huffel, S.:* Structured weighted low rank approximation. Numerical Linear Algebra with Applications 11 (2003), 609–618.

[554] *Schuermans, M.; Markovsky, I.; Wentzell, P.D.; van Huffel, S.:* On the equivalence between total least squares and maximum likelihood PCA. Analytica Chimica Acta 544 (2005), 254–267.

[555] *Schurr, S.P.; Tits, A.L.; O'Leary, D.P.:* Universal duality in conic convex optimization. Math. Program. Ser. A 109 (2007), 68–88.

[556] *Schut, G.H.:* Construction of orthogonal matrices and their application in analytical photogrammetry. Photogrammetria 15 (1959), 149–162.

[557] *Schwartz, L.:* Cours d'analysis, 2. Paris: Hermann, 1967.

[558] *Schwarz, H.R.:* Ein Verfahren zur Stabilitätsfrage bei Matrizen-Eigenwertproblemen. Z. Angew. Math. Phys. 7 (1956), 473–500.

[559] *Schwetlick, H.; Schütze, T.:* Least squares approximation by splines with free knots. BIT Numerical Mathematics 35 (1995) 3, 361–384.

[560] *Searl, S.R.:* Linear models. New York: John Wiley & Sons, 1971.

[561] *Seeger, A.; Pinto da Costa, A.:* Cone-constrained eigenvalue problems: theory and algorithms. Comput. Optim. Appl. 45 (2010) 1, 25–57.

[562] *Sezer, M.E.; Šiljak, D.D.:* On stability of interval matrices. IEEE Trans. Automat. Contr. 39 (1994), 368–371.

[563] *Shamma, J.; Athans, M.:* Gain scheduling: Potential hazards and possible remedies. IEEE Control Systems Magazine 12 (1992), 101–107.

[564] *Shankwitz, C.; Georgiou, T.T.:* Maximum entropie identification and min-max optimal prediction. Proc. 30th Conference on Decision and Control, Brighton, England, 1991, 617–622.

[565] *Shilnikov, L.P.; Shilnikov, A.P.; Turaev, D.V.; Chua, L.O.:* Methods of qualitative theory in nonlinear dynamics. Part II. World Scientific Publishing Co., 2001.

[566] *Shilov, G.E.:* Generalized functions and partial differential equations. New York: Gordon and Breach, 1968.

[567] *Silveira, L.M.; Kamon, M.; Elfadel, I.; White, J.:* A coordinate-transformed Arnoldi algorithm for generating guaranteed stable reduced-order models of RLC circuits. Proc. IEEE Int. Conference on Computer-Aided Design, 1996, 288–294.

[568] *Singer, I.:* Best approximation in normed linear spaces by elements of linear subspaces. Berlin: Springer-Verlag, 1970.

[569] *Sion, M.:* On general minimax theorems. Pacific J. Math. 8 (1958), 171–176.

[570] *Slotine, J.J.E.; Li, W.:* Applied nonlinear control. Englewood Cliffs, NJ: Prentice Hall, 1991.

[571] *Söderström, T.; Stoica, P.:* On the stability of dynamic models obtained by least squares identification. IEEE Transactions on Automatic Control 26 (1981), 575–577.

[572] *Söderström, T.; Stoica, P.:* System identification. New York: Prentice Hall, 1989.

[573] *Sontag, E.D.:* On the observability of polynomial systems, I: Finite time problems. SIAM J. Control and Optimization 17 (1979) 1, 139–151.

[574] *Sontag, E.D.:* Remarks on stabilization and input-to-state stability. Proc. Decision and Control, Tampa, Florida, 1989, 1376–1378.

[575] *Sontag, E.D.:* Smooth stabilization implies coprime factorization. IEEE Trans. Automatic Control 34 (1989), 435–443.

[576] *Sontag, E.D.:* On the input-to-state stability property. European J. Control 1 (1995), 24–36.

[577] *Sontag, E.D.:* Mathematical control theory. New York: Springer-Verlag, 1998.

[578] *Sontag, E.D.:* Input to state stability: basic concepts and results. In: Nonlinear and optimal control theory. Lecture Notes in Mathematics, Berlin: Springer-Verlag 2008, 163–220.

[579] *Sontag, E.D.; Krichman, M.:* An example of a GAS system which can be destabilized by an integrable perturbation. IEEE Transactions on Automatic Control 48 (2003) 6, 1046–1049.

[580] *Sontag, E.D.; Lin, Y.:* Stabilization with respect to noncompact sets: Lyapunov characterizations and effect of bounded inputs. Proc. Nonlinear Control System Design Symp., Bordeaux, IFAC Publications (Hrsg.: Fliess, M.), 1992, 9–14.

[581] *Sontag, E.D.; Wang, Y.:* On characterization of the input-to-state stability property. Systems and Control Letters 24 (1995), 351–359.

[582] *Sontag, E.D.; Wang, Y.:* New characterization to input to state stability. IEEE Transactions on Automatic Control 41 (1996) 9, 1283–1294.

[583] *Sontag, E.D.; Wang, Y.:* Output-to-state stability and detectability of nonlinear systems. Systems and Control Letters 29 (1997), 279–290.

[584] *Sreedhar, J.; van Dooren, P.; Tits, A.L.:* A fast algorithm to compute the real structured stability radius. Proc. Conference on Centennial Hurwitz on Stability Theory, Ticino, Switzerland, May 21-26, 1995.

[585] *Stickel, E.:* On the Fréchet derivative of matrix functions. Linear Algebra and its Application 91 (1987), 83–88.

[586] *Stoica, P.:* On the convergence of an iterative algorithm used for Hammerstein system identification. IEEE Trans. on Automatic Control 26 (1981) 4, 967–969.

[587] *Stoica, P.; Li, J.; Söderström, T.:* On the inconsistency of IQML. IEEE Transactions on Signal Processing 56 (1997) 2, 185–190.

[588] *Stoica, P.; Moses, R.:* On the unit circle problem: The Schur-Cohn procedure revisited. Technical report, SAMPL-87-06, Department of Electrical Engineering, The Ohio State University, Columbus, 1987.

[589] *Sun, J.G.:* An algorithm for the solution of multiparameter eigenvalue problems (I). J. Comput. Math. 8 (1986) 2, 137–149.

[590] *Sun, J.G.:* An algorithm for the solution of multiparameter eigenvalue problems (II). J. Comput. Math. 8 (1986) 4, 354–363.

[591] *Suykens, J.A.K.; Vandewalle, J.P.L.; de Moor, B.:* Artificial neural networks for modelling and control of non-linear systems. Kluwer Academic Publisher, 1996.

[592] *Taniguchi, T.:* Stability theorems of perturbed linear ordinary differential equations. J. of Mathematical Analysis and Applications 149 (1990), 583–598.

[593] *Tao, G.:* Adaptive control design and analysis. New York: John Wiley & Sons Ltd., 2003.

[594] *Tao, G.; Ioannou, P.A.:* Strictly positive real matrices and the Lefschetz-Kalman-Yakubovich lemma. IEEE Transactions on Automatic Control 33 (1988), 1183–1185.

[595] *Terrell, W.J.:* Stability and stabilization. Princeton University Press, 2009.

[596] *Teschl, G.:* Ordinary Differential Equations and Dynamical Systems. Amer. Math. Soc., Graduate Studies in Mathematics, Vol. 140, 2012.

[597] *Tikhonov, A.N.; Arsenin, V.Y.:* Solution of ill-posed problems. Washington, DC: Winston, 1977.

[598] *Tisseur, F.; Higham, N.J.:* More on pseudospectra for polynomial eigenvalue problems and application in control theory. SIAM Journal on Matrix Analysis and Applications 23 (2001) 1, 187–208.

[599] *Tisseur, F.; Meerbergen, K.:* The quadratic eigenvalue problem. SIAM Rev. 43 (2001) 2, 235–286.

[600] *Tokarzewski, J.:* System zeros analysis via Moore-Penrose pseudoinverse and SVD of the first nonzero Markov parameter. IEEE Trans. on Automatic Control 43 (1998) 9.

[601] *Tong, Y.L.:* The multivariate normal distribution. New York: Springer-Verlag, 1990.

[602] *Tsimbinos, J.; Lever, K.V.:* Sampling frequency requirements for identification and compensation of nonlinear systems. Proc. ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing. Adelaide, SA, Australia. IEEE Signal Process. Soc., 1994, III/513–516.

[603] *Tsing, N.-K.:* Convexity of the largest singular value of $e^D M e^{-D}$: A convexity lemma. IEEE Transactions on Automatic Control 35 (1990), 748–749.

[604] *Tsiotras, P.; Corless, M.; Rotea, M.:* Counter-example to a recent result on the stability of nonlinear systems. IMA Journal of Mathematical Control and Information 13 (1996), 129–130.

[605] *Tulleken, H.J.A.F.:* Grey-box modelling and identification using physical knowledge and bayesian techniques. Automatica 29 (1993) 2, 285–308.

[606] *Tuy, H.:* Minimax theorems revisited. Acta Mathematica Vietnamica 29 (2004) 3, 217–229.

[607] *Unbehauen, R.:* Systemtheorie. München: Oldenbourg-Verlag, 1993.

[608] *Vandenberghe, L.; Boyd, S.; Wu, S.-P.:* Determinant maximization with linear matrix inequality constraints. SIAM Journal on Matrix Analysis and Applications 19 (1998) 2, 499–533.

[609] *Vannelli, A.; Vidyasagar, M.:* Maximal Lyapunov functions and domains of attraction for autonomous nonlinear systems. Automatica 21 (1985) 1, 69–80.

[610] *Vardulakis, A.I.G.; Limebeer, D.N.J.; Karcanias, N.:* Structure and Smith-McMillan form of a rational matrix at infinity. Int. J. of Control 35 (1982), 701–725.

[611] *Varfolomeev, A.G.:* Zur Ableitung skalarer Funktionen nach einem symmetrischen Matrixargument (in Russisch). Technical report, Universität von Petrosavod, 1995.

[612] *Verboven, P.:* Frequency-domain system identification for modal analysis. PhD thesis, Vrije Universiteit Brussel, 2002.

[613] *Verdult, V.; Verhaegen, M.; Chou, C.T.:* Identification of MIMO bilinear state space models using separable least squares. Proc. American Control Conference, San Diego, California, 1999, 838–842.

[614] *Vidyasagar, M.:* Nonlinear systems analysis. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[615] *Vidyasagar, M.:* Control system synthesis. Morgan & Claypool, 2011.

[616] *Volkmann, P.:* Gewöhnliche differentialungleichungen mit quasimonoton wachsenden funktionen in topologischen vektorräumen. Math. Z. 127 (1972), 157–164.

[617] *Wall, H.S.:* Polynomials whose zeros have negative real parts. Amer. Math. Monthly 52 (1945), 308–322.

[618] *Walter, R.:* Real and complex analysis. McGraw-Hill, 1987.

[619] *Wang, D.; Li, J.; Huang, M.; Jiang, Y.:* Unique normal form of Bogdanov-Takens singularities. J. of Differential equations 163 (2000), 223–238.

[620] *Wang, J.; Chen, C.:* Exact linearization of nonlinear differential-algebraic systems. Proc. Int. Conf. on Information Technology and Information Networks, Beijing, volume 4, 2001, 284–290.

[621] *Wang, L.; Wang, Z.; Yu, W.:* Stability of polytopic polynomial matrices. Proc. American Control Conference, Arlington, Virginia, 2001, 4695–4696.

[622] *Watson, G.A.:* Choice of norms for data fitting and function approximation. Acta Numerica 7 (1998), 337–377.

[623] *Watson, G.A.:* Least squares fitting of parametric surfaces to measured data. ANZIAM Journal 42 (2000), 68–95.

[624] *Watson, G.A.:* Data fitting problems with bounded uncertainties in the data. SIAM J. Matrix Analysis and Applications 22 (2001), 1274–1293.

[625] *Weinmann, A.:* Uncertain models and robust control. Berlin: Springer-Verlag, 1991.

[626] *Werling, M.:* Ein neues Konzept für die Trajektoriengenerierung und -stabilisierung in zeitkritischen Verkehrsszenarien. Dissertation. KIT Scientific Publishing: Schriftenreihe des Instituts für Angewandte Informatik / Automatisierungstechnik am Karlsruher Institut für Technologie, Band 34, 2011.

[627] *Werling, M.; Gröll, L.; Bretthauer, G.:* Invariant trajectory tracking with a full-size autonomous road vehicle. IEEE Transactions on Robotics 26 (2010), 758–765.

[628] *Wernstedt, J.:* Experimentelle Prozeßanalyse. Berlin: Verlag Technik, 1989.

[629] *Willems, J.C.:* Dissipative dynamical systems – Part II: Linear systems with quadratic supply rate. Berlin: Springer-Verlag, 1972.

[630] *Willems, J.L.; Aeyels, D.:* Comments on "Global stability of solutions of non-linear control systems". Int. J. Systems Sci. 22 (1991) 2, 443–446.

[631] *Wolfe, P.:* A duality theorem for non-linear programming. Quarterly of Applied Mathematics 19 (1961), 239–244.

[632] *Wong, N.; Balakrishnan, V.:* Fast positive real balanced truncation via quadratic alternating direction implicit iteration. Transactions on Computer-aided Design of integrated circuits and systems 26 (2007) 9, 1725–1731.

[633] *Woodgate, K.G.:* Least squares solution of $f = pg$ over positive semidefinite symmetric p. Linear Algebra and its Applications 245 (1996), 171–190.

[634] *Woodgate, K.G.:* Efficient stiffness matrix estimation for elastic structures. Computers & Structures 69 (1998), 79–84.

[635] *Wu, M.Y.:* A note on stability of linear time-varying systems. IEEE Trans. Automat. Control 19 (1974), 162.

[636] *Wu, S.-P.; Boyd, S.; Vandenberghe, L.:* FIR filter design via semidefinite programming and spectral factorization. Proc. Decision and Control, Kobe, Japan, 1996, 271–276.

[637] *Wyatt, J.L.; Chua, L.O.; Gannett, J.W.; Göknar, I.C.; Green, D.N.:* Energy concepts in the state-space theory of nonlinear n-ports: Part I – Passivity. IEEE Transactions on Circuits and Systems 28 (1981) 1, 48–61.

[638] *Xie, L.; Shishkin, S.; Fu, M.:* Piecewise Lyapunov functions for robust stability of linear time-varying systems. Systems & Control Letters 31 (1997), 165–171.

[639] *Xu, X.-J.; Mei, F.-X.:* First integrals and stability of second-order differential equations. Chinese Physics 15 (2006) 6, 1134–1136.

[640] *Yakubovich, V.A.:* Matrix inequalities method in stability theory for nonlinear control systems: I. Absolute stability of forced vibrations. Aut. Remote Control 7 (1964), 905–917.

[641] *Ying, L.; Dayong, C.:* Inhomogeneous eigenvalue problems. Tsinghua Science and Technology 3 (1998) 4, 1260–1264.

[642] *Yu, W.:* Uplink-downlink duality via minimax duality. IEEE Trans. on Information Theory 52 (2006) 2, 361–374.

[643] *Yu, X.; Chen, Z.:* Identification of relative degree. Proc. American Control Conference, Philadelphia, Pennsylvania, 1998, 1928–1932.

[644] *Zeitz, M.:* Minimalphasigkeit – keine relevante Eigenschaft für die Regelungstechnik. Automatisierungstechnik 62 (2014) 1, 3–10.

[645] *Zhang, F.:* Matrix theory – Basic results and techniques. New York, Berlin, Heidelberg: Springer-Verlag, 1999.

[646] *Zhang, L.-H.; Chu, M.T.:* Computing absolute maximum correlation. IMA J. of Numerical Analysis 32 (2012), 163–184.

[647] *Zhou, K.; Doyle, J.C.:* Essentials of robust control. Prentice Hall, 1998.

[648] *Zhu, F.-L.; Ding, X.-H.:* The design of reduced-order observer for systems with monotone nonlinearities. Acta Automatica Sinica 33 (2007) 12, 129 1293.

[649] *Zhu, J.; Johnson, C.D.:* Unified canonical forms for matrices over differential ring. Linear Algebra and its Applications 147 (1991), 201–248.

[650] *Zhu, Y.M.:* Generalised sampling theorem. IEEE Trans. on Circuits and Systems: Analog and Digital Signal Processing 39 (1992) 8, 587–588.

[651] *Zi̧etak, K.:* Strict spectral approximation of a matrix and some related problems. Applicationes Mathematicae 24 (1997), 267–280.

[652] *Zubov, V.I.:* Methods of A. M. Lyapunov and their application. Groningen: P. Noordhoff, 1964.

[653] *Zwillinger, D.:* Handbook of Differential Equations 3rd edition. Academic Press, Boston, MA, 1997.

Bereits veröffentlicht wurden in der Schriftenreihe des Instituts für Angewandte Informatik / Automatisierungstechnik bei KIT Scientific Publishing

1 BECK, S.
Ein Konzept zur automatischen Lösung von Entscheidungsproblemen bei Unsicherheit mittels der Theorie der unscharfen Mengen und der Evidenztheorie, 2005

2 MARTIN, J.
Ein Beitrag zur Integration von Sensoren in eine anthropomorphe künstliche Hand mit flexiblen Fluidaktoren, 2004

3 TRAICHEL, A.
Neue Verfahren zur Modellierung nichtlinearer thermodynamischer Prozesse in einem Druckbehälter mit siedendem Wasser-Dampf Gemisch bei negativen Drucktransienten, 2005

4 LOOSE, T.
Konzept für eine modellgestützte Diagnostik mittels Data Mining am Beispiel der Bewegungsanalyse, 2004

5 MATTHES, J.
Eine neue Methode zur Quellenlokalisierung auf der Basis räumlich verteilter, punktweiser Konzentrationsmessungen, 2004

6 MIKUT, R.; Reischl, M. (HRSG.)
Proceedings – 14. Workshop Fuzzy-Systeme und Computational Intelligence
Dortmund, 10. - 12. November 2004

7 ZIPSER, S.
Beitrag zur modellbasierten Regelung von Verbrennungsprozessen, 2004

8 STADLER, A.
Ein Beitrag zur Ableitung regelbasierter Modelle aus Zeitreihen, 2005

9 MIKUT, R.; REISCHL, M. (HRSG.)
Proceedings – 15. Workshop Computational Intelligence
Dortmund, 16. - 18. November 2005

10 BÄR, M.
µFEMOS – Mikro-Fertigungstechniken für hybride mikrooptische Sensoren, 2005

11 SCHAUDEL, F.
Entropie- und Störungssensitivität als neues Kriterium zum Vergleich verschiedener Entscheidungskalküle, 2006

12 SCHABLOWSKI-TRAUTMANN, M.
Konzept zur Analyse der Lokomotion auf dem Laufband bei inkompletter Querschnittlähmung mit Verfahren der nichtlinearen Dynamik, 2006

13 REISCHL, M.
Ein Verfahren zum automatischen Entwurf von Mensch-Maschine-Schnittstellen am Beispiel myoelektrischer Handprothesen, 2006

14 KOKER, T.
Konzeption und Realisierung einer neuen Prozesskette zur Integration von Kohlenstoff-Nanoröhren über Handhabung in technische Anwendungen, 2007

15 MIKUT, R.; REISCHL, M. (HRSG.)
Proceedings – 16. Workshop Computational Intelligence
Dortmund, 29. November - 1. Dezember 2006

16 LI, S.
Entwicklung eines Verfahrens zur Automatisierung der CAD/CAM-Kette in der Einzelfertigung am Beispiel von Mauerwerksteinen, 2007

17 BERGEMANN, M.
Neues mechatronisches System für die Wiederherstellung der Akkommodationsfähigkeit des menschlichen Auges, 2007

18 HEINTZ, R.
Neues Verfahren zur invarianten Objekterkennung und -lokalisierung auf der Basis lokaler Merkmale, 2007

19 RUCHTER, M.
A New Concept for Mobile Environmental Education, 2007

20 MIKUT, R.; Reischl, M. (HRSG.)
Proceedings – 17. Workshop Computational Intelligence
Dortmund, 5. - 7. Dezember 2007

21 LEHMANN, A.
Neues Konzept zur Planung, Ausführung und Überwachung von Roboteraufgaben mit hierarchischen Petri-Netzen, 2008

22 MIKUT, R.
Data Mining in der Medizin und Medizintechnik, 2008

23 KLINK, S.
Neues System zur Erfassung des Akkommodationsbedarfs im menschlichen Auge, 2008

24 MIKUT, R.; REISCHL, M. (HRSG.)
Proceedings – 18. Workshop Computational Intelligence
Dortmund, 3. - 5. Dezember 2008

25 WANG, L.
Virtual environments for grid computing, 2009

26 BURMEISTER, O.
Entwicklung von Klassifikatoren zur Analyse und Interpretation zeitvarianter Signale und deren Anwendung auf Biosignale, 2009

27 DICKERHOF, M.
Ein neues Konzept für das bedarfsgerechte Informations- und Wissensmanagement in Unternehmenskooperationen der Multimaterial-Mikrosystemtechnik, 2009

28 MACK, G.
Eine neue Methodik zur modellbasierten Bestimmung dynamischer Betriebslasten im mechatronischen Fahrwerkentwicklungsprozess, 2009

29 HOFFMANN, F.; HÜLLERMEIER, E. (HRSG.)
Proceedings – 19. Workshop Computational Intelligence Dortmund, 2. - 4. Dezember 2009

30 GRAUER, M.
Neue Methodik zur Planung globaler Produktionsverbünde unter BerücKsichtigung der Einflussgrößen Produktdesign, Prozessgestaltung und Standortentscheidung, 2009

31 SCHINDLER, A.
Neue Konzeption und erstmalige Realisierung eines aktiven Fahrwerks mit Preview-Strategie, 2009

32 BLUME, C.; JAKOB, W.
GLEAN. General Learning Evolutionary Algorithm and Method Ein Evolutionärer Algorithmus und seine Anwendungen, 2009

33 HOFFMANN, F.; HÜLLERMEIER, E. (HRSG.)
Proceedings – 20. Workshop Computational Intelligence Dortmund, 1. - 3. Dezember 2010

34 WERLING, M.
Ein neues Konzept für die Trajektoriengenerierung und -stabilisierung in zeitkritischen Verkehrsszenarien, 2011

35 KÖVARI, L.
Konzeption und Realisierung eines neuen Systems zur produktbegleitenden virtuellen Inbetriebnahme komplexer Förderanlagen, 2011

36 GSPANN, T. S.
Ein neues Konzept für die Anwendung von einwandigen Kohlenstoff-nanoröhren für die pH-Sensorik, 2011

37 LUTZ, R.
Neues Konzept zur 2D- und 3D-Visualisierung kontinuierlicher, multidimensionaler, meteorologischer Satellitendaten, 2011

38 BOLL, M.-T.
Ein neues Konzept zur automatisierten Bewertung von Fertigkeiten in der minimal invasiven Chirurgie für Virtual Reality Simulatoren in Grid-Umgebungen, 2011

39 GRUBE, M.
Ein neues Konzept zur Diagnose elektrochemischer Sensoren am Beispiel von pH-Glaselektroden, 2011

40 HOFFMANN, F.; Hüllermeier, E. (HRSG.)
Proceedings – 21. Workshop Computational Intelligence Dortmund, 1. - 2. Dezember 2011

41 KAUFMANN, M.
Ein Beitrag zur Informationsverarbeitung in mechatronischen Systemen, 2012

42 NAGEL, J.
Neues Konzept für die bedarfsgerechte Energieversorgung des Künstlichen Akkommodationssystems, 2012

43 RHEINSCHMITT, L.
Erstmaliger Gesamtentwurf und Realisierung der Systemintegration für das Künstliche Akkommodationssystem, 2012

44 BRÜCKNER, B. W.
Neue Methodik zur Modellierung und zum Entwurf keramischer Aktorelemente, 2012

45 HOFFMANN, F.; Hüllermeier, E. (HRSG.)
Proceedings – 22. Workshop Computational Intelligence Dortmund, 6. - 7. Dezember 2012

46 HOFFMANN, F.; Hüllermeier, E. (HRSG.)
Proceedings – 23. Workshop Computational Intelligence Dortmund, 5. - 6. Dezember 2013

47 SCHILL, O.
Konzept zur automatisierten Anpassung der neuronalen Schnittstellen bei nichtinvasiven Neuroprothesen, 2014

48 BAUER, C.
Neues Konzept zur Bewegungsanalyse und -synthese für Humanoide Roboter basierend auf Vorbildern aus der Biologie, 2014

49 WAIBEL, P.
Konzeption von Verfahren zur kamerabasierten Analyse und Optimierung von Drehrohrofenprozessen, 2014

50 HOFFMANN, F.; HÜLLERMEIER, E. (HRSG.)
Proceedings. 24. Workshop Computational Intelligence, Dortmund, 27. - 28. November 2014

51 EICHIN, D.
Modellbasiertes Konzept zur vollautomatisierten Montageendprüfung von asynchron angetriebenen Getriebemotoren im lastlosen Zustand, 2015

52 GRÖLL, L.
Methodik zur Integration von Vorwissen in die Modellbildung, 2015

Die Schriften sind als PDF frei verfügbar, eine Nachbestellung der Printversion ist möglich.
Nähere Informationen unter www.ksp.kit.edu.

Das Einbeziehen von Vorwissen in die Modellbildung liefert validere Modelle für technische Systeme und für Zusammenhänge, die durch Funktionen beschrieben werden. Welches Vorwissen dabei in der Praxis vorliegen kann und wie es mathematisch über Restriktionen zu formulieren ist, wird im ersten Teil der Arbeit erörtert. Der zweite Teil widmet sich der Behandlung der entstehenden restringierten Optimierungsprobleme, indem er in vier Zugängen Wege aufzeigt, die die Probleme für Standardlöser aufbereiten. Ein Leitfaden für die experimentelle Modellbildung mit Vorwissen ergänzt die Betrachtungen. Zahlreiche Beispiele vertiefen die Aussagen, während 32 Tabellen und eine Zusammenstellung zur Stabilitätsthematik dem Überblick dienen.

„Auch wenn es schwierig sein kann, Vorwissen zu integrieren, so wird die Mühe dafür meist stärker belohnt als jene für den Einsatz komplexerer Identifikationsmethoden.“

ISSN 1614-5267 | ISBN 978-3-7315-0303-3