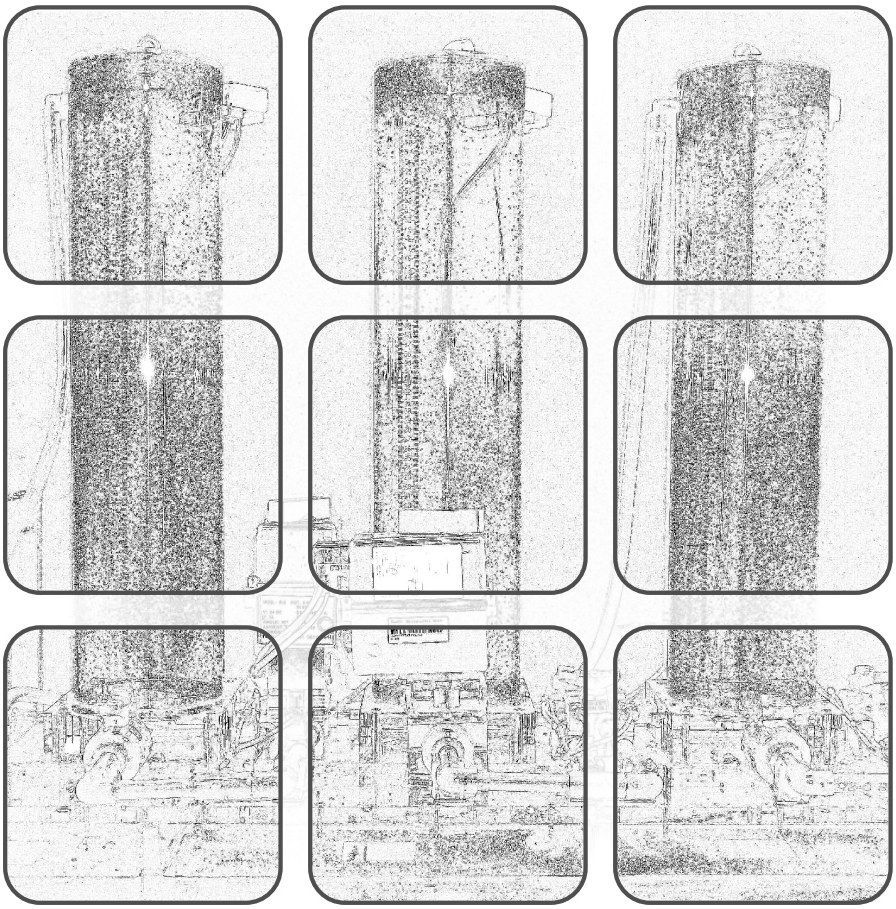
TECHNISCHE UNIVERSITÄT WIEN

AUTOMATION & CONTROL INSTITUTE
INSTITUT FÜR AUTOMATISIERUNGS- & REGELUNGSTECHNIK

OPTIMIERUNGSBASIERTE REGELUNGSMETHODEN

Vorlesung und Übung
Sommersemester 2024

Andreas Steinböck



Optimierungsbasierte Regelungsmethoden

Vorlesung und Übung
Sommersemester 2024

Andreas Steinböck

TU Wien
Institut für Automatisierungs- und Regelungstechnik
Gruppe für komplexe dynamische Systeme

Gußhausstraße 27–29
1040 Wien
Telefon: +43 1 58801 – 37615
Internet: https://www.acin.tuwien.ac.at

Inhaltsverzeichnis

1 Modellprädiktive Regelung

In diesem Abschnitt soll die Methode der modellprädiktiven Regelung kurz vorgestellt werden. Es werden dazu die Bestandteile und die Grundidee dieser Regelungsmethode erläutert und es werden Ansätze zum Nachweis der Stabilität von geschlossenen Regelkreisen diskutiert.

Der Begriff modellprädiktive Regelung wird im Englischen oft als *model predictive control (MPC)* bezeichnet. Mit MPC werden die folgenden Vorgehensweisen in Verbindung gebracht:

- Die Eingangsgröße $\mathbf{u}(t)$ eines dynamischen Systems wird basierend auf einem mit einem mathematischen Modell in die Zukunft prädizierten Systemverhalten gewählt.
- In vielen Fällen beruht die Wahl von $\mathbf{u}(t)$ auf der Lösung einer dynamischen Optimierungsaufgabe.
- Die Wahl von $\mathbf{u}(t)$ wird kontinuierlich oder zu diskreten Zeitpunkten wiederkehrend neu durchgeführt. So können am realen System aktuell gemessene oder beobachtete Größen in der Prädiktion berücksichtigt werden. Durch diese Rückkopplung wird der Regelkreis geschlossen.

Die modellbasierte Prädiktion des Systemverhaltens und die Formulierung der zu lösenden Optimierungsaufgabe erfolgt im Allgemeinen für einen Zeithorizont, der zum aktuellen Zeitpunkt t beginnt und in die Zukunft ragt. Für die wiederkehrende Wahl von $\mathbf{u}(t)$ muss dieser Horizont zeitlich fortgeschoben werden. Aus diesem Grund wird MPC im Englischen auch synonym als *receding horizon control (RHC)* oder *moving horizon control* bezeichnet [1.1].

Wesentliche Vorzüge von MPC sind:

- Der Methode wohnt insofern ein *intuitives, natürliches Handlungsmuster* inne, als in die Entscheidung über die aktuelle Stellgröße deren zukünftige Konsequenzen systematisch einfließen. Oft folgt auch menschliches Handeln (bewusst oder unbewusst) diesem Muster, z. B. beim Lenken eines Fahrzeuges im Straßenverkehr oder bei rollierenden Planungen in der Unternehmensführung.
- Mit der Methode können *bekannte zukünftige Ereignisse, Störungen oder Verläufe von Sollgrößen und Beschränkungen* systematisch berücksichtigt werden. Die Methode kann daher auch als *antizipatorischer* Ansatz verstanden werden.
- MPC ist eine Form der *optimalen Regelung*, denn die mit MPC berechnete Stellgröße geht meist aus der Lösung einer anwendungsspezifisch gestaltbaren, dynamischen Optimierungsaufgabe hervor. In diesem Fall kann es also (im Sinne dieser Optimierungsaufgabe) im aktuellen Zeithorizont keine bessere Stellgröße geben als jene, die MPC liefert.

- *Beschränkungen* von Eingangs-, Zustands- oder Ausgangsgrößen oder deren Zeitableitungen können systematisch berücksichtigt werden. D. h. das System kann nahe oder exakt an diesen Beschränkungen betrieben werden, was für die bestmögliche Erreichung der Regelungsziele förderlich sein kann.
- Es handelt sich um einen *relativ allgemeinen Ansatz*, der auf viele Klassen von dynamischen Systemen (z. B. auch Mehrgrößensysteme und Systeme mit Totzeitverhalten) anwendbar ist.
- Die Methode weist eine relativ gute *Übertragbarkeit* auf. Ist der Regelungsalgorithmus entwickelt und implementiert, so kann er mit relativ geringen Modifikationen und daher geringem Aufwand auf andere Systeme übertragen werden.

Während bei linearen dynamischen Systemen auch mit anderen Methoden, z. B. LQR-Regler (siehe [1.2]), ähnliche Regelungsziele erreicht werden können, sind die genannten Vorzüge in Summe im Bereich der nichtlinearen Regelung derzeit ein Alleinstellungsmerkmal von MPC. So nicht anders angeführt, wird hier der allgemeinere Fall der nichtlinearen modellprädiktiven Regelung (NMPC) behandelt.

Als Preis für die genannten Vorzüge sind zu nennen:

- MPC erfordert im Allgemeinen einen hohen *Rechenaufwand*, da meist eine dynamische Optimierungsaufgabe gelöst werden muss.
- Da die Lösung der dynamischen Optimierungsaufgabe grundsätzlich in *Echtzeit* erfolgen muss, sind je nach Systemdynamik eine effiziente Algorithmik und Programmierung sowie der Einsatz leistungsfähiger Rechner erforderlich.
- Der *Nachweis der Stabilität* des geschlossenen Kreises kann schwierig sein.
- Die erstmalige Entwicklung und Implementierung des Regelungsalgorithmus kann *aufwendig* sein.

Gängige Bücher zum Thema MPC sind z. B. [1.3–1.14], wobei in [1.5, 1.6] speziell auf NMPC eingegangen wird. Weitere Informationen finden sich auch in den Sammelbänden [1.15–1.19]. Darüber hinaus bieten die Übersichtsaufsätze [1.20–1.23] gute Einstiegspunkte in das Thema.

1.1 Bestandteile und Grundidee von MPC

1.1.1 Modell

Zur Prädiktion des zukünftigen Systemverhaltens wird ein mathematisches Modell verwendet. In vielen Fällen wird ein dynamisches Modell in Zustandsraumdarstellung herangezogen. Im zeitkontinuierlichen Fall lautet es

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)) \quad \forall\, t > 0\ , \quad \mathbf{x}(0) = \mathbf{x}_0 \tag{1.1}$$

mit den Zustandsgrößen $\mathbf{x}(t) \in \mathbb{R}^l$, dem Anfangszustand $\mathbf{x}_0$, den Eingangsgrößen $\mathbf{u}(t) \in \mathbb{R}^m$ und der Funktion $\mathbf{f} : \mathbb{R}^l \times \mathbb{R}^m \to \mathbb{R}^l$. Ziel der Regelung soll die Stabilisierung dieses Systems an einer Ruhelage sein. Für diese durch

$$\mathbf{0} = \mathbf{f}(\mathbf{x}_R, \mathbf{u}_R) \tag{1.2}$$

definierte Ruhelage soll $\mathbf{x}_R = \mathbf{0}$ und $\mathbf{u}_R = \mathbf{0}$ gelten[1]. Für MPC wird häufig auch die zeitdiskrete Systemdarstellung

$$\mathbf{x}_{k+1} = \mathbf{F}(\mathbf{x}_k, \mathbf{u}_k) \quad \forall\, k \geq 0 \tag{1.3}$$

mit dem Zeitgitter $t_k = kT_c$, dem Zeitindex $k = 0, 1, 2, \ldots$, der Abtastzeit T_c, der (exakten oder näherungsweisen) Repräsentation $\mathbf{x}_k$ des Zustandes $\mathbf{x}(t_k)$, dem Anfangszustand $\mathbf{x}_0$ und den Eingangsparametern $\mathbf{u}_k \in \mathbb{R}^M$ verwendet. Die Differenzengleichung (1.3) kann als Restriktion der (exakten oder näherungsweisen) Lösung von (1.1) auf das Zeitgitter interpretiert werden, wobei die Eingangsparameter $\mathbf{u}_k$ den Eingang $\mathbf{u}(t)$ im Intervall $[t_k, t_{k+1})$ definieren[2], d. h. es existiert eine Abbildung $\mathbf{U} : [0, T_c) \times \mathbb{R}^M \to \mathbb{R}^m$, so dass

$$\mathbf{u}(t) = \mathbf{U}(t - t_k, \mathbf{u}_k) \quad \forall\, t \in [t_k, t_{k+1}) \tag{1.4}$$

und $\mathbf{0} = \mathbf{U}(\tau, \mathbf{0})\ \forall\, \tau \in [0, T_c)$. Mit dieser Form der Eingangsparametrierung kann erreicht werden, dass der Suchraum der für MPC zu lösenden Optimierungsaufgabe eine finite und niedrige Dimension besitzt. Die durch (1.2) definierte Ruhelage des zeitkontinuierlichen Systems soll auch eine Ruhelage von (1.3) sein, d. h. es soll $\mathbf{0} = \mathbf{F}(\mathbf{0}, \mathbf{0})$ gelten.

Für lineare MPC werden häufig auch Übertragungsfunktionen, Impuls- oder Sprungantwortmodelle [1.24], polynomiale Modelle [1.25] oder andere aus der Systemidentifikation [1.2, 1.26] bekannte Modelldarstellungen verwendet. Ferner ist es möglich, durch Modellerweiterungen den Einfluss von Rauschen, Störungen [1.27] sowie Parameterunsicherheiten und -variationen [1.28] in MPC Formulierungen systematisch zu berücksichtigen.

Der Einfachheit halber werden in dieser Vorlesung die nominellen und ungestörten Modelle (1.1) und (1.3) verwendet. Man beachte, dass es sich hierbei um zeitinvariante Modelle handelt, was den Reglerentwurf und die Stabilitätsanalyse vereinfachen kann. Außerdem wird auf die Verwendung einer Ausgangsgleichung des dynamischen Systems verzichtet.

1.1.2 Horizonte

Wie in Abbildung 1.1a angedeutet, treten bei der Realisierung von MPC grundsätzlich zwei Zeithorizonte auf: ein *Prädiktionshorizont* $[t, t + T]$ (Englisch: *prediction horizon*) und ein *Steuerungshorizont* $[t, t + T_c]$ (Englisch: *control horizon*) mit $T \geq T_c > 0$. So nicht anders erwähnt, werden hier der Einfachheit halber Horizonte mit konstanter Länge T bzw. T_c verwendet. Das MPC Regelgesetz wird zum Zeitpunkt t ausgeführt.

Der Prädiktionshorizont $[t, t + T]$ ist jenes zukünftige Intervall für das das Systemverhalten basierend auf dem mathematischen Modell (1.1) und dem aktuellen Zustand

[1] Dies stellt keine Einschränkung dar, da andere Ruhelagen sofort in den Ursprung transformiert werden können.

[2] Im Falle einer einfachen Abtastregelung gilt natürlich $M = m$ und $\mathbf{u}(t) = \mathbf{u}_k\ \forall\, t \in [t_k, t_{k+1})$.

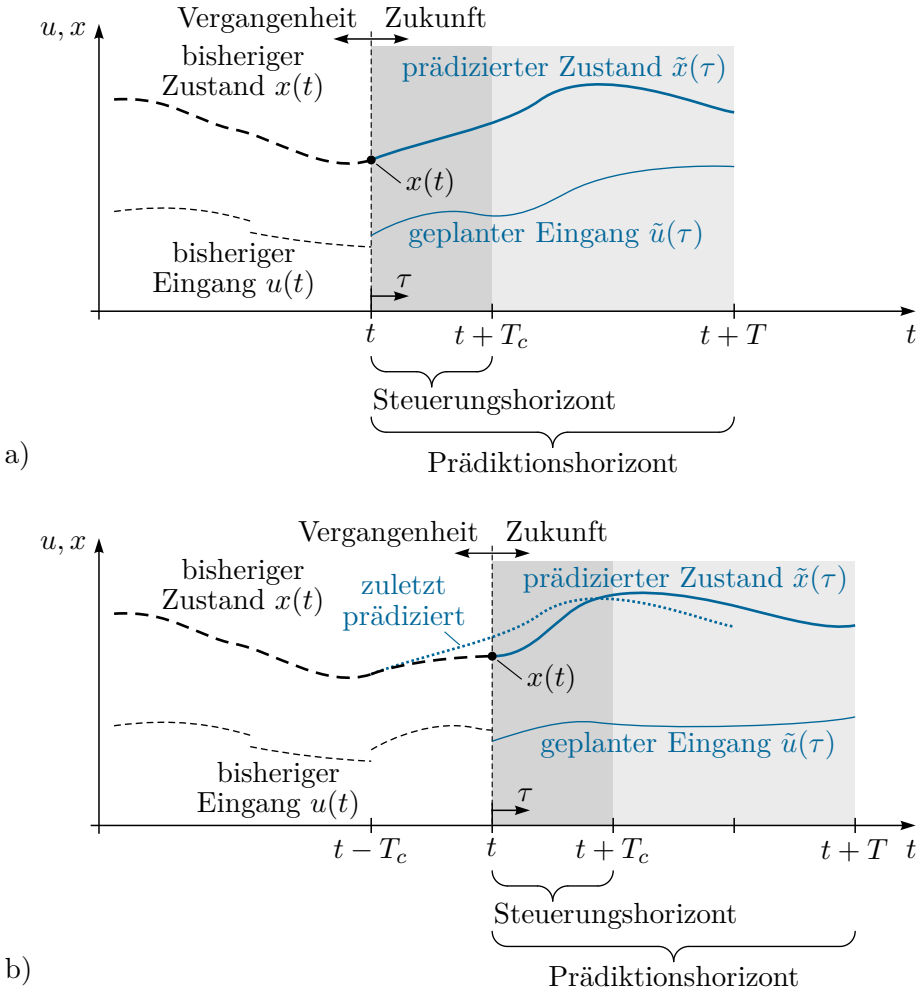


Abbildung 1.1: Horizonte und Signale einer MPC für ein System mit einem Eingang und einem Zustand, a) aktueller Horizont, b) nächstfolgender Horizont.

$\mathbf{x}(t)$ prädiziert wird. Zur Unterscheidung von den tatsächlich am System auftretenden Trajektorien $\mathbf{u}(t)$ und $\mathbf{x}(t)$ werden fortan die für das Intervall $\tau \in [0, T]$ geplante Eingangstrajektorie mit $\tilde{\mathbf{u}}(\tau)$ und die zugehörige mit (1.1) prädizierte Zustandstrajektorie mit $\tilde{\mathbf{x}}(\tau)$ bezeichnet, wobei für den Anfangszustand $\tilde{\mathbf{x}}(0) = \mathbf{x}(t)$ gilt. Zumeist wird für den Prädiktionshorizont eine lokale Zeitachse τ verwendet und am aktuellen Zeitpunkt t gilt $\tau = 0$.

Die geplante Eingangstrajektorie $\tilde{\mathbf{u}}(\tau)$ wird nun nicht im gesamten Prädiktionshorizont $[t, t+T]$ auf die Strecke aufgeschaltet sondern nur im ersten Abschnitt $[t, t+T_c]$ desselben. Dieser Abschnitt wird Steuerungshorizont genannt, denn offensichtlich ist der Regelkreis im Intervall $(t, t+T_c)$ *offen*.

Am Ende des Steuerungshorizonts muss das Regelgesetz (Lösung der Optimierungsaufgabe zur Bestimmung einer neuen Steuertrajektorie $\tilde{\mathbf{u}}(\tau)$) mit nun zeitlich verschobenen Horizonten (siehe Abbildung 1.1b) erneut ausgeführt werden. Genau dann fließt der aktuell

gemessene oder geschätzte Systemzustand $\mathbf{x}(t)$ in die Berechnung von $\tilde{\mathbf{u}}(\tau)$ ein, was den Regelkreis schließt.

Abbildung 1.1 zeigt das MPC Prinzip für ein System mit einem Eingang und einer skalaren Zustandsgröße anhand von zwei aufeinanderfolgenden Horizonten. In der Darstellung weicht die tatsächliche Zustandstrajektorie ($\mathbf{x}(\tau)$ für $\tau \in [t-T_c, t]$ in Abbildung 1.1b) während des ersten Steuerungshorizonts von der ursprünglich prädizierten Trajektorie ($\tilde{\mathbf{x}}(\tau)$ für $\tau \in [0, T_c]$ in Abbildung 1.1a) ab. Diese Abweichung im Steuerungshorizont kann beispielsweise durch Modellfehler oder unbekannte Störungen verursacht werden. Im übrigen Prädiktionshorizont (außerhalb des Steuerungshorizonts) treten solche Abweichungen im Allgemeinen immer auf, d. h. auch für das nominelle System ohne Modellfehler und Störungen. Der Grund hierfür ist, dass die prädizierten Zustandstrajektorien aus Optimierungsaufgaben mit jeweils unterschiedlichen Horizonten hervorgehen.

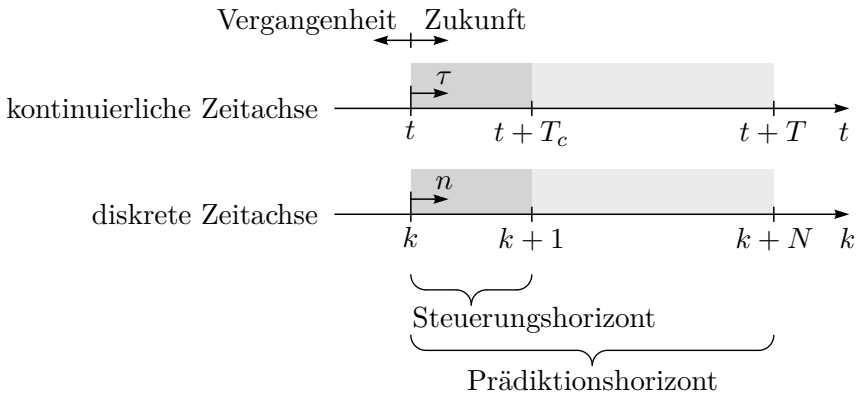


Abbildung 1.2: Horizonte bei zeitkontinuierlicher und zeitdiskreter Formulierung.

Wie in Abbildung 1.2 angedeutet, können analog zur zeitkontinuierlichen Formulierung auch im zeitdiskreten Fall ein Prädiktionshorizont $[t_k, t_{k+N}]$ und ein Steuerungshorizont $[t_k, t_{k+1}]$ verwendet werden. Der Prädiktionshorizont umfasse N Abtastintervalle und n sei ein lokaler Zeitindex, so dass $n = 0$ den Beginn t_k des Prädiktionshorizonts markiere. Die geplante Eingangsfolge $(\tilde{\mathbf{u}}_n)$ muss nur für die Gitterpunkte $n = 0, 1, \ldots, N-1$ (nicht aber für $n = N$) definiert werden, da $\tilde{\mathbf{u}}_N$ keinen Einfluss auf die (prädizierte) Zustandsfolge $(\tilde{\mathbf{x}}_n)$ im Prädiktionshorizont hat. Der Einfachheit halber soll hier der Steuerungshorizont immer genau die Länge T_c eines Abtastintervalls haben.

> *Bemerkung* 1.1. Die in diesem Skriptum verwendeten Definitionen der Begriffe *Prädiktionshorizont* und *Steuerungshorizont* orientieren sich an [1.6, 1.29]. Diese Begriffe werden in der Literatur (insbesondere im Englischen) aber häufig auch anders definiert. Beispielsweise wird als Steuerungshorizont (*control horizon*) meist jenes Zeitintervall verstanden auf dem der Eingang optimiert wird, während er danach konstant gehalten oder durch ein anders Stellgesetz vorgegeben wird [1.9, 1.10, 1.14, 1.19].

1.1.3 Beschränkungen

Bei vielen praktischen Anwendungen sind Eingangs-, Zustands- oder Ausgangsgrößen oder auch deren Zeitableitungen beschränkt. In dieser Vorlesung werden folgende Beschränkungen berücksichtigt:

$$\mathbf{x}(t) \in X \ , \quad \mathbf{u}(t) \in U \ , \quad \forall\, t \geq 0 \ . \tag{1.5}$$

Hierbei soll für die Menge X der zulässigen Zustände und die Menge U der zulässigen Eingänge

$$\{\mathbf{0}\} \subset X \subseteq \mathbb{R}^l \ , \quad \{\mathbf{0}\} \subset U \subseteq \mathbb{R}^m \tag{1.6}$$

gelten. Zusätzlich kann es für MPC Formulierungen erforderlich sein, am Ende des Prädiktionshorizontes restriktivere Zustandsbeschränkungen in der Form

$$\tilde{\mathbf{x}}(T) \in X_T \tag{1.7}$$

zu verwenden, wobei

$$\{\mathbf{0}\} \subseteq X_T \subseteq X \tag{1.8}$$

gelten soll. Man beachte, dass die Beschränkung (1.7) an den Prädiktionshorizont gebunden ist und daher mit diesem wiederkehrend zeitlich fortgeschoben wird.

Im zeitdiskreten Fall werden statt (1.5) und (1.7) die Beschränkungen

$$\mathbf{x}_k \in X \ , \quad \mathbf{u}_k \in U_d \ , \quad \forall\, k = 0, 1, 2, \ldots \tag{1.9a}$$

$$\tilde{\mathbf{x}}_N \in X_T \tag{1.9b}$$

verwendet, wobei die Menge U_d

$$\{\mathbf{0}\} \subset U_d \subseteq \mathbb{R}^M \tag{1.10}$$

und

$$\mathbf{u} \in U_d \quad \Rightarrow \quad \mathbf{U}(t, \mathbf{u}) \in U \quad \forall\, t \in [0, T_c) \tag{1.11}$$

erfüllen soll.

1.1.4 Skalares Gütemaß

Zur Beurteilung von zukünftigen Eingangstrajektorien und prädizierten Zustandstrajektorien wird im zeitkontinuierlichen Fall ein skalares Gütemaß (Gütefunktional) der Form

$$J_T(t, \mathbf{x}(t), \tilde{\mathbf{u}}(\,\cdot\,)) = C(\tilde{\mathbf{x}}(T)) + \int_0^T c(t+\tau, \tilde{\mathbf{x}}(\tau), \tilde{\mathbf{u}}(\tau))\, \mathrm{d}\tau \tag{1.12}$$

mit den Funktionen $C : \mathbb{R}^l \to \mathbb{R}_{\geq 0}$ und $c : \mathbb{R}_{\geq 0} \times \mathbb{R}^l \times \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ verwendet. Hierbei sei $C(\mathbf{x})$ positiv definit und $c(t, \mathbf{x}, \mathbf{u})$ sei positiv definit bezüglich der Argumente $\mathbf{x}$ und $\mathbf{u}$. Ziel bei der Wahl von $\tilde{\mathbf{u}}(\,\cdot\,)$ ist es, J_T zu minimieren. Die Ziele der Regelung fließen in die Gestaltung von C und c ein. Da im vorliegenden Fall die Ruhelage $\mathbf{x}_R = \mathbf{0}$ stabilisiert werden soll und gemäß (1.2) auch $\mathbf{u}_R = \mathbf{0}$ gilt, ist es sinnvoll C und c als positiv definite Funktionen vorzugeben. Die Gestaltung einer Gütefunktion für andere Regelungsziele, wie etwa die Stabilisierung um eine Solltrajektorie, wird z. B. in [1.6] besprochen.

Im zeitdiskreten Fall wird ein skalares Gütemaß (Gütefunktion) der Form

$$J_{d,N}(k, \mathbf{x}_k, (\tilde{\mathbf{u}}_n)) = D(\tilde{\mathbf{x}}_N) + \sum_{n=0}^{N-1} d_{k+n}(\tilde{\mathbf{x}}_n, \tilde{\mathbf{u}}_n) \tag{1.13}$$

mit den positiv definiten Funktionen $D : \mathbb{R}^l \to \mathbb{R}_{\geq 0}$ und $d_k : \mathbb{R}^l \times \mathbb{R}^M \to \mathbb{R}_{\geq 0}$ verwendet. Die Summe in (1.13) endet beim Index $N-1$, da $\tilde{\mathbf{x}}_N$ bereits vom Endkostenterm D erfasst wird und $\tilde{\mathbf{u}}_N$ keine Optimierungsgröße ist.

Wenn $\tilde{\mathbf{x}}_n = \tilde{\mathbf{x}}(nT_c)\ \forall n = 0, 1, \ldots, N$, $\tilde{\mathbf{u}}(\tau + nT_c) = \mathbf{U}(\tau, \tilde{\mathbf{u}}_n)\ \forall (\tau, n) \in [0, T_c) \times \{0, 1, \ldots, N-1\}$, $T = NT_c$, $D(\mathbf{x}) = C(\mathbf{x})$ und

$$d_{k+n}(\tilde{\mathbf{x}}(nT_c), \tilde{\mathbf{u}}_n) = \int_{nT_c}^{(n+1)T_c} c(t_k + \tau, \tilde{\mathbf{x}}(\tau), \tilde{\mathbf{u}}(\tau))\, \mathrm{d}\tau \quad \forall n = 0, 1, \ldots, N-1\ , \tag{1.14}$$

dann liefern (1.12) und (1.13) identische Werte. Dies ist eine naheliegende Möglichkeit zur Gestaltung von D und d ausgehend von C und c. Man beachte, dass eine Auswertung von (1.14) implizit auch die Berechnung der Lösung von (1.1) auf dem Intervall $\tau \in [nT_c, (n+1)T_c]$ erfordert.

1.1.5 Optimierung

Um im zeitkontinuierlichen Fall die Stellgröße $\tilde{\mathbf{u}}(\tau)$ im Prädiktionshorizont $\tau \in [0, T]$ zu wählen, wird die beschränkte dynamische Optimierungsaufgabe

$$\begin{aligned}
\tilde{\mathbf{u}}^*(\,\cdot\,) = \arg\min_{\tilde{\mathbf{u}}(\,\cdot\,)} \quad & J_T(t, \mathbf{x}(t), \tilde{\mathbf{u}}(\,\cdot\,)) = C(\tilde{\mathbf{x}}(T)) + \int_0^T c(t+\tau, \tilde{\mathbf{x}}(\tau), \tilde{\mathbf{u}}(\tau))\, \mathrm{d}\tau && \text{(1.15a)}\\
\text{u.B.v.} \quad & \dot{\tilde{\mathbf{x}}}(\tau) = \mathbf{f}(\tilde{\mathbf{x}}(\tau), \tilde{\mathbf{u}}(\tau))\ , \quad \tilde{\mathbf{x}}(0) = \mathbf{x}(t) && \text{(1.15b)}\\
& \tilde{\mathbf{x}}(\tau) \in X\ , \quad \tilde{\mathbf{u}}(\tau) \in U\ , \quad \forall \tau \in [0, T] && \text{(1.15c)}\\
& \tilde{\mathbf{x}}(T) \in X_T && \text{(1.15d)}
\end{aligned}$$

gelöst. In ihr sind das dynamische Modell, die Beschränkungen und das skalare Gütemaß aus den vorhergehenden Abschnitten zusammengefasst. Allgemein sei mit $J^*_{t|T}(\mathbf{x}(t)) = J_T(t, \mathbf{x}(t), \tilde{\mathbf{u}}^*(\,\cdot\,))$ der optimale Wert des Gütefunktionals in (1.15) für den Anfangszustand $\mathbf{x}(t)$ bezeichnet. Offensichtlich ist $J^*_{t|T}(\mathbf{x}(t))$ eine positiv definite Funktion.

Um im zeitdiskreten Fall die Stellgröße $\tilde{\mathbf{u}}_n$ im Prädiktionshorizont $n = 0, 1, \ldots, N-1$ zu wählen, wird die beschränkte statische Optimierungsaufgabe

$$\begin{aligned}
(\tilde{\mathbf{u}}^*_n) = \arg\min_{(\tilde{\mathbf{u}}_n)} \quad & J_{d,N}(k, \mathbf{x}_k, (\tilde{\mathbf{u}}_n)) = D(\tilde{\mathbf{x}}_N) + \sum_{n=0}^{N-1} d_{k+n}(\tilde{\mathbf{x}}_n, \tilde{\mathbf{u}}_n) && \text{(1.16a)}\\
\text{u.B.v.} \quad & \tilde{\mathbf{x}}_{n+1} = \mathbf{F}(\tilde{\mathbf{x}}_n, \tilde{\mathbf{u}}_n)\ , \quad \tilde{\mathbf{x}}_0 = \mathbf{x}_k && \text{(1.16b)}\\
& \tilde{\mathbf{x}}_n \in X\ , \quad \tilde{\mathbf{u}}_n \in U_d\ , \quad \forall n = 0, 1, \ldots, N-1 && \text{(1.16c)}\\
& \tilde{\mathbf{x}}_N \in X_T && \text{(1.16d)}
\end{aligned}$$

gelöst. Allgemein sei mit $J^*_{d,k|N}(\mathbf{x}_k) = J_{d,N}(k, \mathbf{x}_k, (\tilde{\mathbf{u}}^*_n))$ der optimale Wert der Gütefunktion in (1.16) für den Anfangszustand $\mathbf{x}_k$ bezeichnet. Offensichtlich ist $J^*_{d,k|N}(\mathbf{x}_k)$ eine positiv definite Funktion.

Die Lösung der Optimierungsaufgaben (1.15) und (1.16) ist im Allgemeinen mit hohem numerischem Aufwand verbunden und stellt meist eine zentrale Herausforderung bei MPC dar. Einige Lösungsmethoden wurden bereits in der Vorlesung *Optimierung* [1.30] besprochen. Weitere Methoden finden sich z. B. in den Büchern [1.31–1.42]. Es existieren darüber hinaus MPC Formulierungen, bei denen auf die Lösung oder zumindest die exakte Lösung von Optimierungsaufgaben verzichtet werden kann (siehe z. B. [1.43–1.46]). Diese Formulierungen werden gelegentlich als *suboptimal MPC* bezeichnet. Ein zu dieser Klasse gehörender, in [1.47] vorgeschlagener Ansatz wird in Abschnitt 1.2.5 diskutiert.

1.1.6 Annahmen

Es werden die folgenden Annahmen getroffen:

A1) Auf dem Gebiet $X \times U$ ist die Abbildung $\mathbf{f}$ aus (1.1) Lipschitz-stetig in ihren Argumenten [1.48].

A2) Auf den Gebieten X bzw. $X \times U$ sind die Kostenfunktionen C und c aus (1.12) Lipschitz-stetig in ihren Argumenten [1.48].

A3) Für die Kostenfunktionen C und c aus (1.12) existieren untere und obere Schranken der Form

$$\underline{C}||\mathbf{x}||_2^2 \leq C(\mathbf{x}) \qquad \leq \overline{C}||\mathbf{x}||_2^2 \tag{1.17a}$$

$$\underline{c}(||\mathbf{x}||_2^2 + ||\mathbf{u}||_2^2) \leq c(t, \mathbf{x}, \mathbf{u}) \leq \overline{c}(||\mathbf{x}||_2^2 + ||\mathbf{u}||_2^2) \quad \forall\, t \geq 0 \tag{1.17b}$$

mit geeignet gewählten Konstanten $\underline{C}$, $\overline{C}$, $\underline{c}$ und $\overline{c} \in \mathbb{R}_{>0}$.

A4) Auf dem Gebiet $X \times U_d$ ist die Abbildung $\mathbf{F}$ aus (1.3) Lipschitz-stetig in ihren Argumenten [1.48].

A5) Auf den Gebieten X bzw. $X \times U_d$ sind die Kostenfunktionen D und d_n aus (1.13) Lipschitz-stetig in ihren Argumenten [1.48].

A6) Für die Kostenfunktionen D und d_n aus (1.13) existieren untere und obere Schranken der Form

$$\underline{D}||\mathbf{x}||_2^2 \leq D(\mathbf{x}) \qquad \leq \overline{D}||\mathbf{x}||_2^2 \tag{1.18a}$$

$$\underline{d}(||\mathbf{x}||_2^2 + ||\mathbf{u}||_2^2) \leq d_k(\mathbf{x}, \mathbf{u}) \leq \overline{d}(||\mathbf{x}||_2^2 + ||\mathbf{u}||_2^2) \quad \forall\, k = 0, 1, 2, \ldots \tag{1.18b}$$

mit geeignet gewählten Konstanten $\underline{D}$, $\overline{D}$, $\underline{d}$ und $\overline{d} \in \mathbb{R}_{>0}$.

Aus der Annahme A1 folgt, dass das System (1.1) für jeden beschränkten Anfangszustand $\mathbf{x}_0 \in X$ und jede beschränkte Eingangstrajektorie $\mathbf{u}(t) \in U$ auf dem Intervall $[0, T]$ (für finites T) eine eindeutige und beschränkte Lösung besitzt. Aus der Annahme A4 folgt, dass das System (1.3) für jeden beschränkten Zustand $\mathbf{x}_k \in X$ und jeden beschränkten Eingang $\mathbf{u}_k \in U_d$ eine eindeutige und beschränkte Lösung $\mathbf{x}_{k+1}$ besitzt.

1.1.7 Regelgesetz

Die bisher genannten Bestandteile von MPC können nun zu einem Regelgesetz zusammengefasst werden. In der zeitkontinuierlichen Formulierung lautet es:

Zum Zeitpunkt t wird auf dem Prädiktionshorizont $[t, t+T]$ die beschränkte dynamische Optimierungsaufgabe (1.15) gelöst. Hierbei sei $\mathbf{x}(t)$ der gemessene oder geschätzte aktuelle Systemzustand. Während des Steuerungshorizonts $[t, t+T_c)$ wird die optimierte Stellgröße

$$\mathbf{u}(\tau) = \tilde{\mathbf{u}}^*(\tau - t) \quad \forall\, \tau \in [t, t+T_c) \tag{1.19}$$

mit $\tilde{\mathbf{u}}^*(\,\cdot\,)$ gemäß (1.15) auf die Strecke (1.1) aufgeschaltet. Zum Zeitpunkt $t+T_c$ wird das Regelgesetz von Neuem ausgeführt.

Im üblichen Fall $T_c > 0$ wird das Regelgesetz also zu zeitdiskreten Zeitpunkten ausgeführt (diskrete Regelung). Im Fall $T_c \to 0$ (kontinuierliche Regelung) gilt für die Stellgröße $\mathbf{u}(t) = \tilde{\mathbf{u}}^*(0)$ und (1.15) muss kontinuierlich gelöst werden.

In der zeitdiskreten Formulierung lautet das Regelgesetz wie folgt:

Zum Zeitpunkt t_k wird auf dem Prädiktionshorizont $k, \ldots, k+N$ die beschränkte Optimierungsaufgabe (1.16) gelöst. Hierbei sei $\mathbf{x}_k$ der gemessene oder geschätzte aktuelle Systemzustand. Für das zeitdiskrete System (1.3) gilt nun, dass zum Zeitpunkt t_k die optimierte Stellgröße

$$\mathbf{u}_k = \tilde{\mathbf{u}}_0^* \tag{1.20}$$

mit $(\tilde{\mathbf{u}}_n^*)$ gemäß (1.16) auf die Strecke aufgeschaltet wird. Für ein zeitkontinuierliches System (1.1) gilt, dass im Abtastintervall $[t_k, t_{k+1})$ (zugleich Steuerungshorizont) die Stellgröße

$$\mathbf{u}(\tau) = \mathbf{U}(\tau - t_k, \tilde{\mathbf{u}}_0^*) \quad \forall\, \tau \in [t_k, t_{k+1}) \tag{1.21}$$

mit $(\tilde{\mathbf{u}}_n^*)$ gemäß (1.16) auf die Strecke aufgeschaltet wird. Zum Zeitpunkt t_{k+1} wird das Regelgesetz von Neuem ausgeführt.

1.2 Stabilität

Es werden verschiedene Varianten zum Nachweis der Stabilität eines mit MPC geschlossenen Regelkreises diskutiert.

1.2.1 Prädiktionshorizont mit unendlicher Länge

Es wird die Stabilität des geschlossenen nominellen Regelkreises im Falle eines Prädiktionshorizontes mit unendlicher Länge analysiert. Die Ausführungen folgen im zeitkontinuierlichen Fall [1.49] und im zeitdiskreten Fall [1.6].

Für einen Prädiktionshorizont mit unendlicher Länge kann die zeitkontinuierliche

Optimierungsaufgabe (1.15) in die Form

$$\tilde{\mathbf{u}}^*(\,\cdot\,) = \arg\min_{\tilde{\mathbf{u}}(\,\cdot\,)} \quad J_\infty(t, \mathbf{x}(t), \tilde{\mathbf{u}}(\,\cdot\,)) = \int_0^\infty c(t+\tau, \tilde{\mathbf{x}}(\tau), \tilde{\mathbf{u}}(\tau))\,\mathrm{d}\tau \tag{1.22a}$$

$$\text{u.B.v.} \quad \dot{\tilde{\mathbf{x}}}(\tau) = \mathbf{f}(\tilde{\mathbf{x}}(\tau), \tilde{\mathbf{u}}(\tau))\ , \quad \tilde{\mathbf{x}}(0) = \mathbf{x}(t) \tag{1.22b}$$

$$\tilde{\mathbf{x}}(\tau) \in X\ , \quad \tilde{\mathbf{u}}(\tau) \in U\ , \quad \forall\,\tau \in [0, \infty) \tag{1.22c}$$

umgeschrieben werden. Dass hier die Beschränkung (1.15d) und ein allfälliger Endgewichtungsterm C keinen Einfluss auf die Lösung haben und somit entfallen können, folgt direkt aus dem Beweis des nachfolgenden Satzes.

Satz 1.1 (Stabilität bei Prädiktionshorizont mit unendlicher Länge, zeitkontinuierlich)**.** *Es seien die Annahmen A1-A3 erfüllt. $X_0 \subseteq X$ sei eine nichtleere Menge, genau so dass (1.22) für $\forall\,\mathbf{x}(t) \in X_0$ eine Lösung besitzt (nicht jedoch für $\forall\,\mathbf{x}(t) \in \mathbb{R}^l \backslash X_0$) und der zugehörige Wert des Gütefunktionals $J^*_{t|\infty}(\mathbf{x}(t)) < \infty$ erfüllt. Dann ist die Ruhelage $\mathbf{x}_R = \mathbf{0}$ des geregelten Systems (1.1) lokal asymptotisch stabil mit dem Einzugsbereich X_0.*

Beweis. Die Funktion $J^*_{t|\infty}(\mathbf{x}(t))$ ist positiv definit und ein Kandidat für eine Lyapunovfunktion. Da sie zeitabhängig ist, ist die Stabilitätsanalyse für ein nichtautonomes System vorzunehmen [1.48]. Es sind also zwei positiv definite, zeitinvariante Funktionen $\underline{J}^*_\infty(\mathbf{x}(t))$ und $\overline{J}^*_\infty(\mathbf{x}(t))$ zu suchen, die

$$\underline{J}^*_\infty(\mathbf{x}) \le J^*_{t|\infty}(\mathbf{x}) \le \overline{J}^*_\infty(\mathbf{x}) \quad \forall\,\mathbf{x} \in X_0, t \ge 0 \tag{1.23}$$

erfüllen. Offensichtlich eignen sich die beiden Ersatzprobleme

$$\underline{J}^*_\infty(\mathbf{x}(t)) = \min_{\underline{\mathbf{u}}(\,\cdot\,)} \quad \underline{J}_\infty(\mathbf{x}(t), \underline{\mathbf{u}}(\,\cdot\,)) = \int_0^\infty \underline{c}(||\underline{\mathbf{x}}(\tau)||_2^2 + ||\underline{\mathbf{u}}(\tau)||_2^2)\,\mathrm{d}\tau \tag{1.24a}$$

$$\text{u.B.v.} \quad \dot{\underline{\mathbf{x}}}(\tau) = \mathbf{f}(\underline{\mathbf{x}}(\tau), \underline{\mathbf{u}}(\tau))\ , \quad \underline{\mathbf{x}}(0) = \mathbf{x}(t) \tag{1.24b}$$

$$\underline{\mathbf{x}}(\tau) \in X\ , \quad \underline{\mathbf{u}}(\tau) \in U\ , \quad \forall\,\tau \in [0, \infty) \tag{1.24c}$$

und

$$\overline{J}^*_\infty(\mathbf{x}(t)) = \min_{\overline{\mathbf{u}}(\,\cdot\,)} \quad \overline{J}_\infty(\mathbf{x}(t), \overline{\mathbf{u}}(\,\cdot\,)) = \int_0^\infty \overline{c}(||\overline{\mathbf{x}}(\tau)||_2^2 + ||\overline{\mathbf{u}}(\tau)||_2^2)\,\mathrm{d}\tau \tag{1.25a}$$

$$\text{u.B.v.} \quad \dot{\overline{\mathbf{x}}}(\tau) = \mathbf{f}(\overline{\mathbf{x}}(\tau), \overline{\mathbf{u}}(\tau))\ , \quad \overline{\mathbf{x}}(0) = \mathbf{x}(t) \tag{1.25b}$$

$$\overline{\mathbf{x}}(\tau) \in X\ , \quad \overline{\mathbf{u}}(\tau) \in U\ , \quad \forall\,\tau \in [0, \infty)\ . \tag{1.25c}$$

Wegen (1.17b) und da die optimale Lösung $\tilde{\mathbf{u}}^*(\,\cdot\,)$ von (1.22) eine zulässige aber im Allgemeinen nicht optimale Lösung von (1.24) und (1.25) ist, gilt

$$J^*_{t|\infty}(\mathbf{x}(t)) < \infty \quad \Rightarrow \quad \overline{J}^*_\infty(\mathbf{x}(t)) \le \overline{J}_\infty(\mathbf{x}(t), \tilde{\mathbf{u}}^*(\,\cdot\,)) < \infty \tag{1.26}$$

und

$$\underline{J}^*_\infty(\mathbf{x}(t)) \le \underline{J}_\infty(\mathbf{x}(t), \tilde{\mathbf{u}}^*(\,\cdot\,)) \le J^*_{t|\infty}(\mathbf{x}(t))\ . \tag{1.27}$$

Da ferner die optimale Lösung $\overline{\mathbf{u}}^*(\,\cdot\,)$ von (1.25) eine zulässige aber im Allgemeinen nicht optimale Lösung von (1.22) ist, gilt

$$J^*_{t|\infty}(\mathbf{x}(t)) \leq J_\infty(t, \mathbf{x}(t), \overline{\mathbf{u}}^*(\,\cdot\,)) \leq \overline{J}^*_\infty(\mathbf{x}(t)) \ . \tag{1.28}$$

Damit ist (1.23) gezeigt.

Gemäß dem Optimalitätsprinzip nach Bellman [1.2] gilt für das geregelte System

$$J^*_{t+T_c|\infty}(\mathbf{x}(t+T_c)) = J^*_{t|\infty}(\mathbf{x}(t)) - \int_t^{t+T_c} c(\tau, \mathbf{x}(\tau), \mathbf{u}(\tau))\,\mathrm{d}\tau \tag{1.29}$$

mit beliebigem $T_c \in [0, T]$. Die Optimierungsaufgabe (1.22) hat also auch zum Zeitpunkt $t+T_c$ eine Lösung, es gilt $\mathbf{x}(t+T_c) \in X_0$ und X_0 ist eine positiv invariante Menge. Mit (1.17b) folgt aus (1.29)

$$J^*_{t+T_c|\infty}(\mathbf{x}(t+T_c) \leq J^*_{t|\infty}(\mathbf{x}(t)) - \int_t^{t+T_c} \underline{c}||\mathbf{x}(\tau)||_2^2\,\mathrm{d}\tau \ . \tag{1.30}$$

Im Fall $T_c \to 0$ (kontinuierliche Regelung), welcher hier gemäß den in Abschnitt 1.1.6 getroffenen Stetigkeitsannahmen möglich ist, ergibt sich der Grenzwert

$$\begin{aligned} \dot{J}^*_{t|\infty}(\mathbf{x}(t)) &= \lim_{T_c \to 0} \frac{J^*_{t+T_c|\infty}(\mathbf{x}(t+T_c)) - J^*_{t|\infty}(\mathbf{x}(t))}{T_c} \\ &= -c(t, \mathbf{x}(t), \mathbf{u}(t)) \leq -\underline{c}||\mathbf{x}(t)||_2^2 \ . \end{aligned} \tag{1.31}$$

$J^*_{t|\infty}(\mathbf{x}(t))$ ist also eine Lyapunovfunktion und aus der direkten Methode nach Lyapunov [1.48] folgt die lokale asymptotische Stabilität der Ruhelage $\mathbf{x}_R = \mathbf{0}$.

Für $T_c \in (0, T]$ (diskrete Regelung mit $T_c > 0$) kann der Regelkreis als zeitdiskretes System auf dem Zeitgitter t, $t+T_c$, $t+2T_c$, ... aufgefasst werden und aus (1.30) folgt

$$J^*_{t+T_c|\infty}(\mathbf{x}(t+T_c)) - J^*_{t|\infty}(\mathbf{x}(t)) \leq -\int_t^{t+T_c} \underline{c}||\mathbf{x}(\tau)||_2^2\,\mathrm{d}\tau \quad \forall\, t \geq 0\ ,\ \mathbf{x}(t) \in X_0\ . \tag{1.32}$$

Die rechte Seite dieser Ungleichung ist aufgrund der Stetigkeit von $\mathbf{x}(t)$ negativ definit. $J^*_{t|\infty}(\mathbf{x}(t))$ ist also eine Lyapunovfunktion des gedachten zeitdiskreten Systems und aus der direkten Methode nach Lyapunov für zeitdiskrete Systeme [1.50] folgt die lokale asymptotische Stabilität der Ruhelage $\mathbf{x}_R = \mathbf{0}$. □

Für einen Prädiktionshorizont mit unendlicher Länge kann die zeitdiskrete Optimierungsaufgabe (1.16) in die Form

$$(\tilde{\mathbf{u}}^*_n) = \arg\min_{(\tilde{\mathbf{u}}_n)} \quad J_{d,\infty}(k, \mathbf{x}_k, (\tilde{\mathbf{u}}_n)) = \sum_{n=0}^{\infty} d_{k+n}(\tilde{\mathbf{x}}_n, \tilde{\mathbf{u}}_n) \tag{1.33a}$$

$$\text{u.B.v.} \quad \tilde{\mathbf{x}}_{n+1} = \mathbf{F}(\tilde{\mathbf{x}}_n, \tilde{\mathbf{u}}_n)\ , \quad \tilde{\mathbf{x}}_0 = \mathbf{x}_k \tag{1.33b}$$

$$\tilde{\mathbf{x}}_n \in X\ , \quad \tilde{\mathbf{u}}_n \in U_d\ , \quad \forall\, n = 0, 1, \ldots, \infty\ . \tag{1.33c}$$

umgeschrieben werden. Die Beschränkung (1.16d) und ein allfälliger Endgewichtungsterm D haben wiederum keinen Einfluss auf die Lösung und können daher entfallen.

Satz 1.2 (Stabilität bei Prädiktionshorizont mit unendlicher Länge, zeitdiskret)**.** *Es seien die Annahmen A4-A6 erfüllt. $X_0 \subseteq X$ sei eine nichtleere Menge, genau so dass (1.33) für $\forall \mathbf{x}_k \in X_0$ eine Lösung besitzt (nicht jedoch für $\forall \mathbf{x}_k \in \mathbb{R}^l \backslash X_0$) und der zugehörige Wert des Gütefunktionals $J^*_{d,k|\infty}(\mathbf{x}_k) < \infty$ erfüllt. Dann ist die Ruhelage $\mathbf{x}_R = \mathbf{0}$ des geregelten Systems (1.3) lokal asymptotisch stabil mit dem Einzugsbereich X_0.*

Beweis. In der nachfolgenden Aufgabe 1.1 wird gezeigt, dass positiv definite, zeitinvariante Funktionen $\underline{J}^*_{d,\infty}(\mathbf{x}_k)$ und $\overline{J}^*_{d,\infty}(\mathbf{x}_k)$ existieren, die

$$\underline{J}^*_{d,\infty}(\mathbf{x}_k) \leq J^*_{d,k|\infty}(\mathbf{x}_k) \leq \overline{J}^*_{d,\infty}(\mathbf{x}_k) \quad \forall \mathbf{x}_k \in X_0, k \in \mathbb{N}_0 \tag{1.34}$$

erfüllen.

Gemäß dem Optimalitätsprinzip nach Bellman [1.2] gilt für das geregelte System

$$J^*_{d,k+1|\infty}(\mathbf{x}_{k+1}) = J^*_{d,k|\infty}(\mathbf{x}_k) - d_k(\mathbf{x}_k, \mathbf{u}_k) \ . \tag{1.35}$$

Die Optimierungsaufgabe (1.33) hat also auch zum Zeitpunkt t_{k+1} eine Lösung, es gilt $\mathbf{x}_{k+1} \in X_0$ und X_0 ist eine positiv invariante Menge. Mit (1.18b) folgt aus (1.35)

$$J^*_{d,k+1|\infty}(\mathbf{x}_{k+1}) - J^*_{d,k|\infty}(\mathbf{x}_k) = -d_k(\mathbf{x}_k, \mathbf{u}_k) \leq -\underline{d}||\mathbf{x}_k||_2^2 \ . \tag{1.36}$$

$J^*_{d,k|\infty}(\mathbf{x}_k)$ stellt also eine Lyapunovfunktion dar und aus der direkten Methode nach Lyapunov für zeitdiskrete Systeme [1.50] folgt die lokale asymptotische Stabilität der Ruhelage $\mathbf{x}_R = \mathbf{0}$. □

Aufgabe 1.1. Zeigen Sie, dass positiv definite, zeitinvariante Funktionen $\underline{J}^*_{d,\infty}(\mathbf{x}_k)$ und $\overline{J}^*_{d,\infty}(\mathbf{x}_k)$ existieren, die

$$\underline{J}^*_{d,\infty}(\mathbf{x}_k) \leq J^*_{d,k|\infty}(\mathbf{x}_k) \leq \overline{J}^*_{d,\infty}(\mathbf{x}_k) \quad \forall \mathbf{x}_k \in X_0, k \in \mathbb{N}_0 \tag{1.37}$$

erfüllen.

Bei MPC in zeitkontinuierlicher Formulierung ist eine infinit-dimensionale Optimierungsaufgabe zu lösen. Dies gilt auch für die zeitdiskrete Formulierung von MPC, wenn ein unendlich langer Prädiktionshorizont verwendet wird. Der mit der Lösung einer infinit-dimensionalen Optimierungsaufgabe verbundene Rechenaufwand ist im Allgemeinen ebenfalls unbeschränkt, weshalb grundsätzlich nur näherungsweise Lösungen bestimmt werden können. Alle nachfolgenden Methoden zum Stabilitätsnachweis verwenden daher Prädiktionshorizonte mit endlicher Länge, so dass zumindest in der zeitdiskreten Formulierung nur eine finit-dimensionale Optimierungsaufgabe auftritt.

1.2.2 Endlicher Prädiktionshorizont mit vorgeschriebenem Endzustand

Es wird die Stabilität des geschlossenen nominellen Regelkreises im Falle eines endlichen Prädiktionshorizontes mit vorgeschriebenem Endzustand analysiert. Der zu erreichende

Endzustand entspricht der Ruhelage $\mathbf{x}_R = \mathbf{0}$. Die Ausführungen folgen im zeitkontinuierlichen Fall [1.49] und im zeitdiskreten Fall [1.6].

Für einen endlichen Prädiktionshorizont mit vorgeschriebenem Endzustand kann die zeitkontinuierliche Optimierungsaufgabe (1.15) in die Form

$$\tilde{\mathbf{u}}^*(\,\cdot\,) = \arg\min_{\tilde{\mathbf{u}}(\,\cdot\,)} \quad J_T(t, \mathbf{x}(t), \tilde{\mathbf{u}}(\,\cdot\,)) = \int_0^T c(t+\tau, \tilde{\mathbf{x}}(\tau), \tilde{\mathbf{u}}(\tau))\,\mathrm{d}\tau \tag{1.38a}$$

$$\text{u.B.v.} \quad \dot{\tilde{\mathbf{x}}}(\tau) = \mathbf{f}(\tilde{\mathbf{x}}(\tau), \tilde{\mathbf{u}}(\tau))\ , \quad \tilde{\mathbf{x}}(0) = \mathbf{x}(t) \tag{1.38b}$$

$$\tilde{\mathbf{x}}(\tau) \in X\ , \quad \tilde{\mathbf{u}}(\tau) \in U\ , \quad \forall\,\tau \in [0, T] \tag{1.38c}$$

$$\tilde{\mathbf{x}}(T) = \mathbf{0} \tag{1.38d}$$

umgeschrieben werden.

Satz 1.3 (Stabilität bei endlichem Prädiktionshorizont mit vorgeschriebenem Endzustand, zeitkontinuierlich)**.** *Es seien die Annahmen A1-A3 erfüllt. $X_0 \subseteq X$ sei eine nichtleere Menge, genau so dass (1.38) für $\forall\,\mathbf{x}(t) \in X_0$ eine Lösung besitzt (nicht jedoch für $\forall\,\mathbf{x}(t) \in \mathbb{R}^l \backslash X_0$). Dann ist die Ruhelage $\mathbf{x}_R = \mathbf{0}$ des geregelten Systems (1.1) lokal exponentiell stabil mit dem Einzugsbereich X_0.*

Beweis. Es sei $\tilde{\mathbf{u}}^*(\,\cdot\,)$ die optimale Lösung von (1.38) und $\tilde{\mathbf{x}}^*(\,\cdot\,)$ die zugehörige optimale Zustandstrajektorie. Ferner wird die zur Optimierungsaufgabe (1.38) gehörige Lösungsfunktion $\tilde{\mathbf{u}}^*(\tau) = \mathbf{k}(t+\tau, \tilde{\mathbf{x}}^*(\tau))$ definiert. Wegen der Annahmen A1-A3 ist $\mathbf{k} : \mathbb{R}_{\geq 0} \times X_0 \to U$ im Gebiet X_0 Lipschitz-stetig, d. h. es existiert eine Konstante $L_\mathbf{k}$, so dass $||\mathbf{k}(t, \mathbf{x}) - \mathbf{k}(t, \mathbf{y})||_2 \leq L_\mathbf{k} ||\mathbf{x} - \mathbf{y}||_2\ \forall\,\mathbf{x}, \mathbf{y} \in X_0$. Es gilt also auch

$$||\tilde{\mathbf{u}}^*(\tau)||_2 \leq L_\mathbf{k} ||\tilde{\mathbf{x}}^*(\tau)||_2 \quad \forall\,\tau \in [0, T]\ . \tag{1.39}$$

Weiters seien $L_\mathbf{x}$ und $L_\mathbf{u}$ die Lipschitzkonstanten der Funktion $\mathbf{f}$ aus (1.1) bezüglich ihres ersten und zweiten Arguments. Es gilt daher

$$||\dot{\tilde{\mathbf{x}}}^*(\tau)||_2 \leq ||\tilde{\mathbf{x}}^*(\tau)||_2 (L_\mathbf{x} + L_\mathbf{u} L_\mathbf{k}) \tag{1.40}$$

und folglich

$$||\mathbf{x}(t)||_2^2 \mathrm{e}^{-2(L_\mathbf{x} + L_\mathbf{u} L_\mathbf{k})\tau} \leq ||\tilde{\mathbf{x}}^*(\tau)||_2^2 \leq ||\mathbf{x}(t)||_2^2 \mathrm{e}^{2(L_\mathbf{x} + L_\mathbf{u} L_\mathbf{k})\tau} \quad \forall\,\tau \in [0, T]\ . \tag{1.41}$$

Mit den Ungleichungen (1.17b), (1.39) und (1.41) können sofort Faktoren $\overline{J} \geq \underline{J} > 0$ berechnet werden, so dass

$$\underline{J} ||\mathbf{x}(t)||_2^2 \leq J^*_{t|T}(\mathbf{x}(t)) \leq \overline{J} ||\mathbf{x}(t)||_2^2 < \infty \quad \forall\, t \geq 0\ ,\ \mathbf{x}(t) \in X_0 \tag{1.42}$$

erfüllt ist.

An die optimale Lösung $\tilde{\mathbf{u}}^*(\,\cdot\,)$ von (1.38) kann im Zeitintervall $(t+T, t+T+T_c]$ die zur Ruhelage $\mathbf{x}_R = \mathbf{0}$ gehörige Stellgröße $\mathbf{u}_R = \mathbf{0}$ angefügt werden. Die so erweiterte

Lösung soll mit

$$\tilde{\tilde{\mathbf{u}}}(\tau) = \begin{cases} \tilde{\mathbf{u}}^*(T_c + \tau) & \text{falls } \tau \in [-T_c, T - T_c] \\ \mathbf{0} & \text{falls } \tau \in (T - T_c, T] \end{cases} \tag{1.43}$$

und die zugehörige Zustandstrajektorie mit

$$\tilde{\tilde{\mathbf{x}}}(\tau) = \tilde{\mathbf{x}}^*(\min\{\tau + T_c, T\}) \ , \quad \forall\, \tau \in [-T_c, T] \tag{1.44}$$

bezeichnet werden. Wird dem MPC Regelgesetz (1.19) entsprechend im Steuerungshorizont $[t, t + T_c)$ die Stellgröße $\mathbf{u}(\tau) = \tilde{\mathbf{u}}^*(\tau - t)\ \forall\, \tau \in [t, t + T_c)$ aufgeschaltet, dann ist $\tilde{\tilde{\mathbf{u}}}(\,\cdot\,)$ eine zulässige aber im Allgemeinen nicht optimale Eingangstrajektorie der Optimierungsaufgabe (1.38) für den nachfolgenden Zeithorizont $[t + T_c, t + T + T_c]$ mit dem Anfangszustand $\mathbf{x}(t + T_c) = \tilde{\mathbf{x}}^*(T_c)$. Folglich gilt für die optimalen Werte des Gütefunktionals

$$\begin{aligned} J^*_{t+T_c|T}(\mathbf{x}(t + T_c)) &\leq \int_0^{T-T_c} c(t + T_c + \tau, \tilde{\tilde{\mathbf{x}}}(\tau), \tilde{\tilde{\mathbf{u}}}(\tau))\, \mathrm{d}\tau \\ &= J^*_{t|T}(\mathbf{x}(t)) - \int_t^{t+T_c} c(\tau, \mathbf{x}(\tau), \mathbf{u}(\tau))\, \mathrm{d}\tau \\ &\leq J^*_{t|T}(\mathbf{x}(t)) - \int_t^{t+T_c} \underline{c} ||\mathbf{x}(\tau)||_2^2\, \mathrm{d}\tau \ . \end{aligned} \tag{1.45}$$

Für $T_c \to 0$ (kontinuierliche Regelung) kann damit (ähnlich zum Beweis von Satz 1.1) der Grenzwert

$$\begin{aligned} \dot{J}^*_{t|T}(\mathbf{x}(t)) &= \lim_{T_c \to 0} \frac{J^*_{t+T_c|T}(\mathbf{x}(t + T_c)) - J^*_{t|T}(\mathbf{x}(t))}{T_c} \\ &= -c(t, \mathbf{x}(t), \mathbf{u}(t)) \leq -\underline{c} ||\mathbf{x}(t)||_2^2 \end{aligned} \tag{1.46}$$

berechnet werden. $J^*_{t|T}(\mathbf{x}(t))$ stellt also wieder eine Lyapunovfunktion dar und es folgt direkt die lokale exponentielle Stabilität der Ruhelage $\mathbf{x}_R = \mathbf{0}$ (vgl. [1.48]). Für $T_c \in (0, T]$ kann der Regelkreis als zeitdiskretes System auf dem Zeitgitter t, $t + T_c$, $t + 2T_c$, ... aufgefasst werden und mit den Ungleichungen (1.17b), (1.39) und (1.41) lässt sich eine Schranke $\Delta J > 0$ berechnen, so dass

$$J^*_{t+T_c|T}(\mathbf{x}(t + T_c)) - J^*_{t|T}(\mathbf{x}(t)) \leq -\Delta J ||\mathbf{x}(t)||_2^2 \quad \forall\, t \geq 0 \ , \ \mathbf{x}(t) \in X_0 \ . \tag{1.47}$$

$J^*_{t|T}(\mathbf{x}(t))$ ist also eine Lyapunovfunktion des gedachten zeitdiskreten Systems und es folgt direkt die lokale exponentielle Stabilität der Ruhelage $\mathbf{x}_R = \mathbf{0}$. □

Für einen endlichen Prädiktionshorizont mit vorgeschriebenem Endzustand kann die

zeitdiskrete Optimierungsaufgabe (1.16) in die Form

$$(\tilde{\mathbf{u}}_n^*) = \arg\min_{(\tilde{\mathbf{u}}_n)} \quad J_{d,N}(k, \mathbf{x}_k, (\tilde{\mathbf{u}}_n)) = \sum_{n=0}^{N-1} d_{k+n}(\tilde{\mathbf{x}}_n, \tilde{\mathbf{u}}_n) \tag{1.48a}$$

$$\text{u.B.v.} \quad \tilde{\mathbf{x}}_{n+1} = \mathbf{F}(\tilde{\mathbf{x}}_n, \tilde{\mathbf{u}}_n) \ , \quad \tilde{\mathbf{x}}_0 = \mathbf{x}_k \tag{1.48b}$$

$$\tilde{\mathbf{x}}_n \in X \ , \quad \tilde{\mathbf{u}}_n \in U_d \ , \quad \forall\, n = 0, 1, \ldots, N-1 \tag{1.48c}$$

$$\tilde{\mathbf{x}}_N = \mathbf{0} \tag{1.48d}$$

umgeschrieben werden.

Satz 1.4 (Stabilität bei endlichem Prädiktionshorizont mit vorgeschriebenem Endzustand, zeitdiskret)**.** *Es seien die Annahmen A4-A6 erfüllt. $X_0 \subseteq X$ sei eine nichtleere Menge, genau so dass (1.48) für $\forall\, \mathbf{x}_k \in X_0$ eine Lösung besitzt (nicht jedoch für $\forall\, \mathbf{x}_k \in \mathbb{R}^l \backslash X_0$). Dann ist die Ruhelage $\mathbf{x}_R = \mathbf{0}$ des geregelten Systems (1.3) lokal exponentiell stabil mit dem Einzugsbereich X_0.*

Beweis. Es sei $(\tilde{\mathbf{u}}_n^*)$ die optimale Lösung von (1.48) und $(\tilde{\mathbf{x}}_n^*)$ die zugehörige optimale Zustandsfolge. Wegen der Annahmen A4-A6 können analog zu den Ungleichungen (1.39) und (1.41) auch im zeitdiskreten Fall Abschätzungen für $(\tilde{\mathbf{u}}_n^*)$ und $(\tilde{\mathbf{x}}_n^*)$ bestimmt werden. Diese Abschätzungen erlauben die Berechnung von Faktoren $\overline{J} \geq \underline{J} > 0$, so dass

$$\underline{J} ||\mathbf{x}_k||_2^2 \leq J_{d,k|N}^*(\mathbf{x}_k) \leq \overline{J} ||\mathbf{x}_k||_2^2 < \infty \quad \forall\, \mathbf{x}_k \in X_0 \ , \ k \in \mathbb{N}_0 \tag{1.49}$$

erfüllt ist.

An die optimale Lösung $(\tilde{\mathbf{u}}_n^*)$ von (1.48) kann am Gitterpunkt $k+N$ die zur Ruhelage $\mathbf{x}_R = \mathbf{0}$ gehörige Stellgröße $\mathbf{u}_R = \mathbf{0}$ angefügt werden. Die so erweiterte Lösung soll mit

$$\tilde{\tilde{\mathbf{u}}}_n = \begin{cases} \tilde{\mathbf{u}}_{n+1}^* & \text{falls } n \in \{-1, 0, 1, \ldots, N-2\} \\ \mathbf{0} & \text{falls } n = N-1 \end{cases} \tag{1.50}$$

und die zugehörige Zustandsfolge mit

$$\tilde{\tilde{\mathbf{x}}}_n = \tilde{\mathbf{x}}_{\min\{n+1, N\}}^* \quad \forall\, n = -1, 0, 1, \ldots, N \tag{1.51}$$

bezeichnet werden. Wird dem MPC Regelgesetz (1.20) entsprechend zum Zeitpunkt t_k die Stellgröße $\mathbf{u}_k = \tilde{\mathbf{u}}_0^*$ aufgeschaltet, dann ist $(\tilde{\tilde{\mathbf{u}}}_n)$ eine zulässige aber im Allgemeinen nicht optimale Eingangsfolge der Optimierungsaufgabe (1.48) für den nachfolgenden Zeithorizont $k+1, \ldots, k+N+1$ mit dem Anfangszustand $\mathbf{x}_{k+1} = \tilde{\mathbf{x}}_1^*$. Folglich gilt

für die optimalen Werte der Gütefunktion

$$\begin{aligned} J^*_{d,k+1|N}(\mathbf{x}_{k+1}) &\leq \sum_{n=0}^{N-2} d_{k+1+n}(\tilde{\tilde{\mathbf{x}}}_n, \tilde{\tilde{\mathbf{u}}}_n) \\ &= J^*_{d,k|N}(\mathbf{x}_k) - d_k(\mathbf{x}_k, \mathbf{u}_k) \\ &\leq J^*_{d,k|N}(\mathbf{x}_k) - \underline{d}||\mathbf{x}_k||_2^2 \ . \end{aligned} \tag{1.52}$$

Aus (1.52) folgt

$$J^*_{d,k+1|N}(\mathbf{x}_{k+1}) - J^*_{d,k|N}(\mathbf{x}_k) \leq -\underline{d}||\mathbf{x}_k||_2^2 \ . \tag{1.53}$$

$J^*_{d,k|N}(\mathbf{x}_k)$ stellt also eine Lypunovfunktion für das zeitdiskrete System dar und es folgt direkt die lokale exponentielle Stabilität [1.50] der Ruhelage $\mathbf{x}_R = \mathbf{0}$. □

Die Bedingungen (1.38d) und (1.48d) erzwingen, dass der prädizierte Endzustand zum Zeitpunkt $t + T$ bzw. t_{k+N} im Ursprung liegt. Man beachte, dass dies im Allgemeinen nicht für die tatsächliche Zustandstrajektorie bzw. Zustandsfolge gilt. Diese nähern sich aber dem Ursprung zumindest exponentiell.

Die Verwendung eines endlichen Prädiktionshorizonts mit vorgeschriebenem Endzustand kann folgende Nachteile mit sich bringen: Die Methode ist nicht geeignet für Systeme, die zwar stabilisierbar aber nicht vollständig steuerbar sind. Bei der Lösung der Optimierungsaufgabe kann die exakte Einhaltung einer Gleichungsbedingung am Ende des Prädiktionshorizontes hohen Rechenaufwand erfordern. Bei kurzen Prädiktionshorizonten, welche aus Sicht des Rechenaufwands wünschenswert wären, kann das Vorschreiben eines Endzustandes zu kleinen Einzugsbereichen X_0 führen. Dies beeinträchtigt meist die Robustheit der Regelung. Bei den nachfolgenden Methoden zum Stabilitätsnachweis wird daher auf einen vorgeschriebenen Endzustand verzichtet.

1.2.3 Endlicher Prädiktionshorizont mit vorgeschriebenem Endgebiet und Endkostenterm

Es wird die Stabilität des geschlossenen nominellen Regelkreises im Falle eines endlichen Prädiktionshorizontes mit vorgeschriebenem Endgebiet und Endkostenterm analysiert. Die Ausführungen sind im zeitkontinuierlichen Fall an [1.49] und im zeitdiskreten Fall an [1.6] angelehnt.

Satz 1.5 (Stabilität bei endlichem Prädiktionshorizont mit vorgeschriebenem Endgebiet und Endkostenterm, zeitkontinuierlich)**.** *Es seien die Annahmen A1-A3 erfüllt. $X_0 \subseteq X$ sei eine nichtleere Menge, genau so dass (1.15) für $\forall \mathbf{x}(t) \in X_0$ eine Lösung besitzt (nicht jedoch für $\forall \mathbf{x}(t) \in \mathbb{R}^l \backslash X_0$). Es existiere ein Zustandsregelgesetz*

$$\mathbf{u}(t) = \boldsymbol{\kappa}(t, \mathbf{x}(t)) \tag{1.54}$$

mit $\boldsymbol{\kappa} : \mathbb{R} \times X_T \to U$*, so dass die Zustandstrajektorie* $\bar{\mathbf{x}}(\tau)$ *mit* $\tau \in [0, T_c]$ *des mit* $\boldsymbol{\kappa}(\tau+t, \bar{\mathbf{x}}(\tau))$ *geregelten Systems* (1.1) *für beliebige Anfangszustände* $\bar{\mathbf{x}}(0) = \mathbf{x}(t) \in X_T$

$$C(\bar{\mathbf{x}}(T_c)) - C(\mathbf{x}(t)) \leq -\int_0^{T_c} c(t+\tau, \bar{\mathbf{x}}(\tau), \boldsymbol{\kappa}(t+\tau, \bar{\mathbf{x}}(\tau)))\,\mathrm{d}\tau\ , \tag{1.55}$$

$\bar{\mathbf{x}}(\tau) \in X\ \forall \tau \in [0, T_c]$ *und* $\bar{\mathbf{x}}(T_c) \in X_T$ *erfüllt. Dann ist die Ruhelage* $\mathbf{x}_R = \mathbf{0}$ *des geregelten Systems* (1.1) *lokal exponentiell stabil mit dem Einzugsbereich* X_0*.*

Beweis. An die optimale Lösung $\tilde{\mathbf{u}}^*(\,\cdot\,)$ von (1.15), wobei $\tilde{\mathbf{x}}^*(\,\cdot\,)$ die zugehörige optimale Zustandstrajektorie sei, kann im Zeitintervall $(t+T, t+T_c+T]$ die Eingangstrajektorie des mit dem Zustandsregelgesetz (1.54) geregelten Systems angefügt werden. Die so erweiterte Lösung soll mit

$$\tilde{\tilde{\mathbf{u}}}(\tau) = \begin{cases} \tilde{\mathbf{u}}^*(T_c+\tau) & \text{falls } \tau \in [-T_c, T-T_c] \\ \boldsymbol{\kappa}(t+T_c+\tau, \bar{\mathbf{x}}(\tau)) & \text{falls } \tau \in (T-T_c, T] \end{cases} \tag{1.56}$$

und die zugehörige Zustandstrajektorie mit

$$\tilde{\tilde{\mathbf{x}}}(\tau) = \begin{cases} \tilde{\mathbf{x}}^*(T_c+\tau) & \text{falls } \tau \in [-T_c, T-T_c] \\ \bar{\mathbf{x}}(\tau) & \text{falls } \tau \in (T-T_c, T] \end{cases} \tag{1.57}$$

bezeichnet werden, wobei $\bar{\mathbf{x}}(\tau)$ mit $\tau \in [T-T_c, T]$ die Zustandstrajektorie des mit $\boldsymbol{\kappa}(t+T_c+\tau, \bar{\mathbf{x}}(\tau))$ geregelten Systems (1.1) für den Anfangszustand $\bar{\mathbf{x}}(T-T_c) = \tilde{\mathbf{x}}^*(T)$ ist. Wird dem MPC Regelgesetz (1.19) entsprechend im Steuerungshorizont $[t, t+T_c)$ die Stellgröße $\mathbf{u}(\tau) = \tilde{\mathbf{u}}^*(\tau - t)\ \forall \tau \in [t, t+T_c)$ aufgeschaltet, dann ist $\tilde{\tilde{\mathbf{u}}}(\,\cdot\,)$ eine zulässige aber im Allgemeinen nicht optimale Eingangstrajektorie der Optimierungsaufgabe (1.15) für den nachfolgenden Zeithorizont $[t+T_c, t+T_c+T]$ mit dem Anfangszustand $\mathbf{x}(t+T_c) = \tilde{\mathbf{x}}^*(T_c)$. Unter Berücksichtigung von (1.55) gilt folglich für die optimalen Werte des Gütefunktionals

$$\begin{aligned} J^*_{t+T_c|T}(\mathbf{x}(t+T_c)) &\leq C(\tilde{\tilde{\mathbf{x}}}(T-T_c)) + \int_0^{T-T_c} c(t+T_c+\tau, \tilde{\tilde{\mathbf{x}}}(\tau), \tilde{\tilde{\mathbf{u}}}(\tau))\,\mathrm{d}\tau \\ &= J^*_{t|T}(\mathbf{x}(t)) - \int_t^{t+T_c} c(\tau, \mathbf{x}(\tau), \mathbf{u}(\tau))\,\mathrm{d}\tau\ . \end{aligned} \tag{1.58}$$

Der Rest des Beweises erfolgt völlig analog zum Beweis von Satz 1.3. □

Bevor die zeitdiskrete Variante dieser Methode diskutiert wird, sollen kurz verschiedene Interpretationen des Ansatzes gegeben werden.

- Aus dem Zustandsregelgesetz (1.54) und der Ungleichung (1.55) folgt die Existenz einer Eingangstrajektorie, die sicherstellt, dass bei einer Verlängerung des Prädiktionshorizonts um T_c der Kostenzuwachs zufolge des integralen Kostenterms c durch eine Kostenreduktion zufolge des geänderten Endkostenterms C zumindest wettgemacht wird.

- Der Ansatz wird gelegentlich als *MPC mit quasi-unendlichem Horizont* bezeichnet [1.51], da gilt

$$C(\mathbf{x}(t)) \geq \int_t^\infty c(\tau, \mathbf{x}(\tau), \boldsymbol{\kappa}(\tau, \mathbf{x}(\tau)))\,\mathrm{d}\tau\ , \tag{1.59}$$

  sofern $\mathbf{x}(t) \in X_T$ und das System im Horizont $[t, \infty)$ mit dem Zustandsregelgesetz (1.54) betrieben wird. D. h. der Endkostenterm $C(\mathbf{x}(t))$ ist eine obere Schranke für den integralen Kostenterm in der gesamten Zukunft. Um (1.59) zu zeigen, kann man einfach (1.55) rekursiv einsetzen. Wegen (1.59) und $J^*_{t|T}(\mathbf{x}(t)) < \infty\ \forall\, \mathbf{x}(t) \in X_0$ kann der Stabilitätsbeweis auch analog zu jenem von Satz 1.1 für den unendlichen Prädiktionshorizont erfolgen.

- Man beachte, dass die Forderung (1.55) im Zusammenspiel mit den Annahmen A1 und A3 impliziert, dass der mit dem Zustandsregelgesetz $\boldsymbol{\kappa}(t, \mathbf{x}(t))$ und dem System (1.1) geschlossene Kreis im Sinne eines zeitdiskreten Systems mit dem Abtastintervall T_c lokal exponentiell stabil mit dem Einzugsbereich X_T sein muss. Allgemein kann der Entwurf eines stabilisierenden Regelgesetzes, das einen für die jeweilige Anwendung hinreichend großen Einzugsbereich erlaubt, schwierig sein. Dies kann grundsätzlich auch für den Entwurf des Zustandsregelgesetzes $\boldsymbol{\kappa}(t, \mathbf{x}(t))$, das den in Satz 1.5 genannten Anforderungen zu genügen hat und daher einen Einzugsbereich X_T sicherstellen muss, gelten. Wenn der Einzugsbereich X_T für die Anwendung ausreicht und auch sonst alle Regelungsziele bereits mit dem Zustandsregelgesetzes $\boldsymbol{\kappa}(t, \mathbf{x}(t))$ erreicht werden, kann auf die Verwendung von MPC verzichtet werden. Wenn jedoch X_T keine zufriedenstellende Ausdehnung besitzt, kann MPC als Stratgie verstanden werden um den vom Zustandsregler erreichten Einzugsbereich X_T auf X_0 auszuweiten.

- Mit der speziellen Wahl $X_T = \{\mathbf{0}\}$ erhält man sofort die in Abschnitt 1.2.2 beschriebene Formulierung. D. h. letztere ist ein Spezialfall des hier vorgestellten Ansatzes.

Satz 1.6 (Stabilität bei endlichem Prädiktionshorizont mit vorgeschriebenem Endgebiet und Endkostenterm, zeitdiskret)**.** *Es seien die Annahmen A4-A6 erfüllt. $X_0 \subseteq X$ sei eine nichtleere Menge, genau so dass (1.16) für $\forall\, \mathbf{x}_k \in X_0$ eine Lösung besitzt (nicht jedoch für $\forall\, \mathbf{x}_k \in \mathbb{R}^l \backslash X_0$). Es existiere ein Zustandsregelgesetz*

$$\mathbf{u}_k = \boldsymbol{\kappa}_k(\mathbf{x}_k) \tag{1.60}$$

mit $\boldsymbol{\kappa}_k : X_T \to U_d$, so dass für beliebige Anfangszustände $\mathbf{x}_k \in X_T$

$$D(\mathbf{F}(\mathbf{x}_k, \boldsymbol{\kappa}_k(\mathbf{x}_k))) - D(\mathbf{x}_k) \leq -d_k(\mathbf{x}_k, \boldsymbol{\kappa}_k(\mathbf{x}_k)) \tag{1.61}$$

und $\mathbf{F}(\mathbf{x}_k, \boldsymbol{\kappa}_k(\mathbf{x}_k)) \in X_T$ erfüllt sind. Dann ist die Ruhelage $\mathbf{x}_R = \mathbf{0}$ des geregelten Systems (1.3) lokal exponentiell stabil mit dem Einzugsbereich X_0.

Beweis. An die optimale Lösung $(\tilde{\mathbf{u}}_n^*)$ von (1.16), wobei $(\tilde{\mathbf{x}}_n^*)$ die zugehörige optimale Zustandsfolge sei, kann am Gitterpunkt $k+N$ die Stellgröße $\boldsymbol{\kappa}_{k+N}(\tilde{\mathbf{x}}_N^*)$ gemäß dem Zustandsregelgesetz (1.60) angefügt werden. Die so erweiterte Lösung soll mit

$$\tilde{\tilde{\mathbf{u}}}_n = \begin{cases} \tilde{\mathbf{u}}_{n+1}^* & \text{falls } n \in \{-1, 0, 1, \dots, N-2\} \\ \boldsymbol{\kappa}_{k+N}(\tilde{\mathbf{x}}_N^*) & \text{falls } n = N-1 \end{cases} \tag{1.62}$$

und die zugehörige Zustandsfolge mit

$$\tilde{\tilde{\mathbf{x}}}_n = \begin{cases} \tilde{\mathbf{x}}_{n+1}^* & \text{falls } n \in \{-1, 0, 1, \dots, N-1\} \\ F(\tilde{\mathbf{x}}_N^*, \boldsymbol{\kappa}_{k+N}(\tilde{\mathbf{x}}_N^*)) & \text{falls } n = N \end{cases} \tag{1.63}$$

bezeichnet werden. Wird dem MPC Regelgesetz (1.20) entsprechend zum Zeitpunkt t_k die Stellgröße $\mathbf{u}_k = \tilde{\mathbf{u}}_0^*$ aufgeschaltet, dann ist $(\tilde{\tilde{\mathbf{u}}}_n)$ eine zulässige aber im Allgemeinen nicht optimale Eingangsfolge der Optimierungsaufgabe (1.16) für den nachfolgenden Zeithorizont $k+1, \dots, k+N+1$ mit dem Anfangszustand $\mathbf{x}_{k+1} = \tilde{\mathbf{x}}_1^*$.

Unter Berücksichtigung von (1.61) gilt folglich für die optimalen Werte der Gütefunktion

$$\begin{aligned} J_{d,k+1|N}^*(\mathbf{x}_{k+1}) &\leq D(\tilde{\tilde{\mathbf{x}}}_{N-1}) + \sum_{n=0}^{N-2} d_{k+1+n}(\tilde{\tilde{\mathbf{x}}}_n, \tilde{\tilde{\mathbf{u}}}_n) \\ &= J_{d,k|N}^*(\mathbf{x}_k) - d_k(\mathbf{x}_k, \mathbf{u}_k) \ . \end{aligned} \tag{1.64}$$

Der Rest des Beweises erfolgt völlig analog zum Beweis von Satz 1.4. □

1.2.4 Endlicher Prädiktionshorizont mit Endkostenterm

Es wird ein zu der in Abschnitt 1.2.3 beschriebenen MPC Variante verwandter Ansatz vorgestellt, der kein vorgeschriebenes Endgebiet benötigt. Dies vereinfacht in der Regel die zugrunde liegende dynamische Optimierungsaufgabe erheblich (freier Endzustand). Der hier vorgestellte Ansatz wurde ursprünglich in [1.52] vorgeschlagen. Die nachfolgende Stabilitätsanalyse des geschlossenen nominellen Regelkreises ist an [1.14] angelehnt.

Für einen endlichen Prädiktionshorizont ohne vorgeschriebenem Endzustand entfällt in der zeitkontinuierlichen Optimierungsaufgabe (1.15) die Bedingung (1.15d). Die Form der prädizierten Zustandstrajektorien und damit das Konvergenzverhalten des Algorithmus können beeinflusst werden, indem der Endkostenterm C mit einem konstanten Faktor $\gamma > 0$ skaliert wird. Zu diesem Zweck kann (1.15) in die Form

$$\tilde{\mathbf{u}}^*(\,\cdot\,) = \arg\min_{\tilde{\mathbf{u}}(\,\cdot\,)} \quad J_T(t, \mathbf{x}(t), \tilde{\mathbf{u}}(\,\cdot\,)) = \gamma C(\tilde{\mathbf{x}}(T)) + \int_0^T c(t+\tau, \tilde{\mathbf{x}}(\tau), \tilde{\mathbf{u}}(\tau))\,\mathrm{d}\tau \tag{1.65a}$$

$$\text{u.B.v.} \quad \dot{\tilde{\mathbf{x}}}(\tau) = \mathbf{f}(\tilde{\mathbf{x}}(\tau), \tilde{\mathbf{u}}(\tau)) \ , \quad \tilde{\mathbf{x}}(0) = \mathbf{x}(t) \tag{1.65b}$$

$$\tilde{\mathbf{x}}(\tau) \in X \ , \quad \tilde{\mathbf{u}}(\tau) \in U \ , \quad \forall\, \tau \in [0, T] \tag{1.65c}$$

mit $\gamma > 0$ umgeschrieben werden. Die Annahme A3 gilt unverändert. Es wird von der Existenz eines Zustandsregelgesetzes $\boldsymbol{\kappa}(t, \mathbf{x}(t))$, welches das System für alle Anfangszu-

stände aus einem abgeschlossenen Gebiet X_Γ stabilisiert, ausgegangen. Hierbei sei X_Γ eine Niveaumenge des Endkostenterms $C(\tilde{\mathbf{x}}(T))$ mit dem Niveau $\Gamma > 0$, d. h.

$$X_\Gamma = \{\mathbf{x} \in X | C(\mathbf{x}) \leq \Gamma\} \supset \{\mathbf{0}\} \ . \tag{1.66}$$

Anstatt nun X_Γ als Endgebiet in der Optimierungsaufgabe vorzuschreiben, wird eine Niveaumenge $\underline{X}_0$ von $J^*_{t|T}(\mathbf{x}(t))$ gesucht, so dass die MPC für alle Anfangszustände aus $\underline{X}_0$ sicherstellt, dass die prädizierte Trajektorie $\tilde{\mathbf{x}}(\tau)$ innerhalb des Prädiktionshorizonts mit der Länge T in das Gebiet X_Γ einläuft.

Satz 1.7 (Stabilität bei endlichem Prädiktionshorizont mit Endkostenterm und ohne vorgeschriebenem Endgebiet, zeitkontinuierlich)**.** *Es seien die Annahmen A1-A3 erfüllt. Es existiere ein Zustandsregelgesetz*

$$\mathbf{u}(t) = \boldsymbol{\kappa}(t, \mathbf{x}(t)) \tag{1.67}$$

mit $\boldsymbol{\kappa} : \mathbb{R} \times X_\Gamma \to U$, einer zugehörigen Konstante $\Gamma > 0$ und einer Konstante $\gamma > 0$, so dass

$$\gamma \frac{\mathrm{d}C(\mathbf{x})}{\mathrm{d}\mathbf{x}} \mathbf{f}(\mathbf{x}, \boldsymbol{\kappa}(t, \mathbf{x})) \leq -c(t, \mathbf{x}, \boldsymbol{\kappa}(t, \mathbf{x})) \quad \forall\, \mathbf{x} \in X_\Gamma, t \geq 0 \tag{1.68}$$

erfüllt ist. Dann ist die Ruhelage $\mathbf{x}_R = \mathbf{0}$ des mit dem MPC Regelgesetz (1.19) entsprechend der Optimierungsaufgabe (1.65) geregelten Systems (1.1) lokal exponentiell stabil mit einem Einzugsbereich, der

$$\underline{X}_0 = \left\{ \mathbf{x} \in X \middle| J^*_{t|T}(\mathbf{x}) \leq \Gamma\left(\gamma + T\frac{\underline{c}}{\overline{C}}\right) \ \forall\, t \geq 0 \right\} \supset \{\mathbf{0}\} \tag{1.69}$$

beinhaltet.

Beweis. Es sei $\tilde{\mathbf{u}}^*(\,\cdot\,)$ die optimale Lösung von (1.65) und $\tilde{\mathbf{x}}^*(\,\cdot\,)$ die zugehörige optimale Zustandstrajektorie, wobei $\tilde{\mathbf{x}}^*(0) = \mathbf{x}(t) \in \underline{X}_0$ gelten soll. Aus (1.17a) folgt die Implikation

$$||\mathbf{x}||_2^2 \leq \frac{\Gamma}{\overline{C}} \quad \Rightarrow \quad \mathbf{x} \in X_\Gamma \ . \tag{1.70}$$

Wegen der Ungleichung

$$\begin{aligned} \Gamma\left(\gamma + T\frac{\underline{c}}{\overline{C}}\right) \geq J^*_{t|T}(\mathbf{x}(t)) &= \gamma C(\tilde{\mathbf{x}}^*(T)) + \int_0^T c(t+\tau, \tilde{\mathbf{x}}^*(\tau), \tilde{\mathbf{u}}^*(\tau))\,\mathrm{d}\tau \\ &\geq \gamma C(\tilde{\mathbf{x}}^*(T)) + \int_0^T \underline{c}||\tilde{\mathbf{x}}^*(\tau)||_2^2\,\mathrm{d}\tau \end{aligned} \tag{1.71}$$

(vgl. (1.17b)) muss es also einen Zeitpunkt $\bar{\tau} \in [0, T]$ geben, so dass $\tilde{\mathbf{x}}^*(\bar{\tau}) \in X_\Gamma$. Es sei nun $\bar{\mathbf{x}}(\tau)$ mit $\tau \in [\bar{\tau}, T]$ die Zustandstrajektorie des mit $\boldsymbol{\kappa}(t+\tau, \bar{\mathbf{x}}(\tau))$ geregelten Systems (1.1) für den Anfangszustand $\bar{\mathbf{x}}(\bar{\tau}) = \tilde{\mathbf{x}}^*(\bar{\tau})$. Wegen (1.68) und der

Suboptimalität der Eingangstrajektorie $\boldsymbol{\kappa}(t+\tau, \bar{\mathbf{x}}(\tau))$ für $\tau \in [\bar{\tau}, T]$ gilt

$$\begin{aligned} \gamma C(\tilde{\mathbf{x}}^*(\bar{\tau})) &\geq \gamma C(\bar{\mathbf{x}}(T)) + \int_{\bar{\tau}}^{T} c(t+\tau, \bar{\mathbf{x}}(\tau), \boldsymbol{\kappa}(t+\tau, \bar{\mathbf{x}}(\tau)))\,\mathrm{d}\tau \\ &\geq \gamma C(\tilde{\mathbf{x}}^*(T)) + \int_{\bar{\tau}}^{T} c(t+\tau, \tilde{\mathbf{x}}^*(\tau), \tilde{\mathbf{u}}^*(\tau))\,\mathrm{d}\tau \geq \gamma C(\tilde{\mathbf{x}}^*(T)) \ . \end{aligned} \tag{1.72}$$

Daraus folgt

$$\Gamma \geq C(\tilde{\mathbf{x}}^*(\bar{\tau})) \geq C(\tilde{\mathbf{x}}^*(T)) \tag{1.73}$$

und somit $\tilde{\mathbf{x}}^*(T) \in X_\Gamma$. An die optimale Lösung $\tilde{\mathbf{u}}^*(\,\cdot\,)$ von (1.65) kann folglich im Zeitintervall $(t+T, t+T_c+T]$ die Eingangstrajektorie des mit dem Zustandsregelgesetz (1.67) geregelten Systems angefügt werden. Die so erweiterte Lösung soll mit

$$\tilde{\tilde{\mathbf{u}}}(\tau) = \begin{cases} \tilde{\mathbf{u}}^*(T_c+\tau) & \text{falls } \tau \in [-T_c, T-T_c] \\ \boldsymbol{\kappa}(t+T_c+\tau, \bar{\tilde{\mathbf{x}}}(\tau)) & \text{falls } \tau \in (T-T_c, T] \end{cases} \tag{1.74}$$

und die zugehörige Zustandstrajektorie mit

$$\tilde{\tilde{\mathbf{x}}}(\tau) = \begin{cases} \tilde{\mathbf{x}}^*(T_c+\tau) & \text{falls } \tau \in [-T_c, T-T_c] \\ \bar{\tilde{\mathbf{x}}}(\tau) & \text{falls } \tau \in (T-T_c, T] \end{cases} \tag{1.75}$$

bezeichnet werden, wobei $\bar{\tilde{\mathbf{x}}}(\tau)$ mit $\tau \in [T-T_c, T]$ die Zustandstrajektorie des mit $\boldsymbol{\kappa}(t+T_c+\tau, \bar{\tilde{\mathbf{x}}}(\tau))$ geregelten Systems (1.1) für den Anfangszustand $\bar{\tilde{\mathbf{x}}}(T-T_c) = \tilde{\mathbf{x}}^*(T)$ ist. Wird dem MPC Regelgesetz (1.19) entsprechend im Steuerungshorizont $[t, t+T_c)$ die Stellgröße $\mathbf{u}(\tau) = \tilde{\mathbf{u}}^*(\tau - t)\ \forall\, \tau \in [t, t+T_c)$ aufgeschaltet, dann ist $\tilde{\tilde{\mathbf{u}}}(\,\cdot\,)$ eine zulässige aber im Allgemeinen nicht optimale Eingangstrajektorie der Optimierungsaufgabe (1.65) für den nachfolgenden Zeithorizont $[t+T_c, t+T_c+T]$ mit dem Anfangszustand $\mathbf{x}(t+T_c) = \tilde{\mathbf{x}}^*(T_c)$. Unter Berücksichtigung von (1.68) gilt folglich für die optimalen Werte des Gütefunktionals

$$\begin{aligned} J^*_{t+T_c|T}(\mathbf{x}(t+T_c)) &\leq \gamma C(\tilde{\tilde{\mathbf{x}}}(T-T_c)) + \int_0^{T-T_c} c(t+T_c+\tau, \tilde{\tilde{\mathbf{x}}}(\tau), \tilde{\tilde{\mathbf{u}}}(\tau))\,\mathrm{d}\tau \\ &= J^*_{t|T}(\mathbf{x}(t)) - \int_t^{t+T_c} c(\tau, \mathbf{x}(\tau), \mathbf{u}(\tau))\,\mathrm{d}\tau \ . \end{aligned} \tag{1.76}$$

Daraus folgt nun noch

$$\Gamma\Big(\gamma + T\frac{c}{C}\Big) \geq J^*_{t|T}(\mathbf{x}(t)) \geq J^*_{t+T_c|T}(\mathbf{x}(t+T_c)) \tag{1.77}$$

und somit $\mathbf{x}(t+T_c) \in \underline{X}_0$. Der Rest des Beweises erfolgt völlig analog zum Beweis von Satz 1.3. □

Anhand der Definition (1.69) wird klar, welche Parameter des Entwurfs zu verändern sind um einen größeren Einzugsbereich zu erhalten. Für eine möglichst große Menge $\underline{X}_0$ sollte der Niveauwert Γ maximal sein. Γ wird jedoch meist durch das verwendete Zustandsregelgesetz $\boldsymbol{\kappa}(t, \bar{\mathbf{x}}(t))$ limitiert. Die Länge T des Prädiktionshorizonts wirkt sich direkt proportional auf

das Niveau von $\underline{X}_0$ aus. Lange Prädiktionshorizonte sind aufgrund des damit verbundenen numerischen Aufwands beim Lösen der zugrundeliegenden Optimierungsaufgabe oft nicht erwünscht. Eine Erhöhung des Verstärkungsfaktors γ bewirkt eine Erhöhung des Niveaus von $\underline{X}_0$ und eine vereinfachte Einhaltung der Bedingung (1.68). Natürlich verändert eine Erhöhung von γ auch das Regelverhalten, da der integrale Kostenterm c gegenüber dem Endkostenterm γC an Bedeutung verliert. Es ist ferner zu beachten, dass eine Änderung von T oder γ sich auch auf den Term $J^*_{t|T}(\mathbf{x})$ in (1.69) auswirkt.

Für einen endlichen Prädiktionshorizont ohne vorgeschriebenem Endzustand und mit erhöhter Gewichtung des Endkostenterms kann die zeitdiskrete Optimierungsaufgabe (1.16) in die Form

$$(\tilde{\mathbf{u}}_n^*) = \arg\min_{(\tilde{\mathbf{u}}_n)} \quad J_{d,N}(k, \mathbf{x}_k, (\tilde{\mathbf{u}}_n)) = \gamma D(\tilde{\mathbf{x}}_N) + \sum_{n=0}^{N-1} d_{k+n}(\tilde{\mathbf{x}}_n, \tilde{\mathbf{u}}_n) \tag{1.78a}$$

$$\text{u.B.v.} \quad \tilde{\mathbf{x}}_{n+1} = \mathbf{F}(\tilde{\mathbf{x}}_n, \tilde{\mathbf{u}}_n) \ , \quad \tilde{\mathbf{x}}_0 = \mathbf{x}_k \tag{1.78b}$$

$$\tilde{\mathbf{x}}_n \in X \ , \quad \tilde{\mathbf{u}}_n \in U_d \ , \quad \forall\, n = 0, 1, \ldots, N-1 \tag{1.78c}$$

umgeschrieben werden, wobei wieder $\gamma > 0$ gilt. Die Annahme A6 gilt unverändert. Es wird wieder eine Niveaumenge

$$X_\Gamma = \{\mathbf{x} \in X | D(\mathbf{x}) \leq \Gamma\} \supset \{\mathbf{0}\} \tag{1.79}$$

definiert.

Satz 1.8 (Stabilität bei endlichem Prädiktionshorizont mit Endkostenterm und ohne vorgeschriebenem Endgebiet, zeitdiskret)**.** *Es seien die Annahmen A4-A6 erfüllt. Es existiere ein Zustandsregelgesetz*

$$\mathbf{u}_k = \boldsymbol{\kappa}_k(\mathbf{x}_k) \tag{1.80}$$

mit $\boldsymbol{\kappa}_k : X_\Gamma \to U_d$, einer zugehörigen Konstante $\Gamma > 0$ und einer Konstante $\gamma > 0$, so dass

$$\gamma D(\mathbf{F}(\mathbf{x}_k, \boldsymbol{\kappa}_k(\mathbf{x}_k))) - \gamma D(\mathbf{x}_k) \leq -d_k(\mathbf{x}_k, \boldsymbol{\kappa}_k(\mathbf{x}_k)) \quad \forall\, \mathbf{x}_k \in X_\Gamma, k \in \mathbb{N}_0 \tag{1.81}$$

erfüllt ist. Dann ist die Ruhelage $\mathbf{x}_R = \mathbf{0}$ des mit dem MPC Regelgesetz (1.20) entsprechend der Optimierungsaufgabe (1.78) geregelten Systems (1.3) lokal exponentiell stabil mit einem Einzugsbereich, der

$$\underline{X}_0 = \left\{ \mathbf{x} \in X \middle| J^*_{d,k|N}(\mathbf{x}) \leq \Gamma\left(\gamma + N\frac{d}{\overline{D}}\right) \ \forall\, k \in \mathbb{N}_0 \right\} \supset \{\mathbf{0}\} \tag{1.82}$$

beinhaltet.

Beweis. Es sei $(\tilde{\mathbf{u}}_n^*)$ die optimale Lösung von (1.78) und $(\tilde{\mathbf{x}}_n^*)$ die zugehörige optimale Zustandsfolge, wobei $\tilde{\mathbf{x}}^*(0) = \mathbf{x}_k \in \underline{X}_0$ gelten soll. Aus (1.18a) folgt die Implikation

$$||\mathbf{x}||_2^2 \leq \frac{\Gamma}{\overline{D}} \quad \Rightarrow \quad \mathbf{x} \in X_\Gamma \ . \tag{1.83}$$

Wegen der Ungleichung

$$\begin{aligned}\Gamma\Big(\gamma + N\frac{\underline{d}}{\overline{D}}\Big) \geq J^*_{d,k|N}(\mathbf{x}_k) &= \gamma D(\tilde{\mathbf{x}}^*_N) + \sum_{n=0}^{N-1} d_{k+n}(\tilde{\mathbf{x}}^*_n, \tilde{\mathbf{u}}^*_n) \\ &\geq \gamma D(\tilde{\mathbf{x}}^*_N) + \sum_{n=0}^{N-1} \underline{d}||\tilde{\mathbf{x}}^*_n||_2^2\end{aligned} \tag{1.84}$$

(vgl. (1.18b)) muss es also einen Zeitindex $\bar{n} \in \{0, 1, \ldots, N\}$ geben, so dass $\tilde{\mathbf{x}}^*_{\bar{n}} \in X_\Gamma$. Es sei nun $(\bar{x}_n)$ mit $n = \bar{n}, \bar{n}+1, \ldots, N$ die Zustandsfolge des mit $\boldsymbol{\kappa}_{k+n}(\bar{\mathbf{x}}_n)$ geregelten Systems (1.3) für den Anfangszustand $\bar{\mathbf{x}}_{\bar{n}} = \tilde{\mathbf{x}}^*_{\bar{n}}$. Wegen (1.81) und der Suboptimalität der Eingangsfolge $(\boldsymbol{\kappa}_{k+n}(\bar{\mathbf{x}}_n))$ für $n = \bar{n}, \bar{n}+1, \ldots, N-1$ gilt

$$\begin{aligned}\gamma D(\tilde{\mathbf{x}}^*_{\bar{n}}) &\geq \gamma D(\bar{\mathbf{x}}_N) + \sum_{n=\bar{n}}^{N-1} d_{k+n}(\bar{\mathbf{x}}_n, \boldsymbol{\kappa}_{k+n}(\bar{\mathbf{x}}_n)) \\ &\geq \gamma D(\tilde{\mathbf{x}}^*_N) + \sum_{n=\bar{n}}^{N-1} d_{k+n}(\tilde{\mathbf{x}}^*_n, \tilde{\mathbf{u}}^*_n) \geq \gamma D(\tilde{\mathbf{x}}^*_N)\ .\end{aligned} \tag{1.85}$$

Daraus folgt

$$\Gamma \geq D(\tilde{\mathbf{x}}^*_{\bar{n}}) \geq D(\tilde{\mathbf{x}}^*_N) \tag{1.86}$$

und somit $\tilde{\mathbf{x}}^*_N \in X_\Gamma$. An die optimale Lösung $(\tilde{\mathbf{u}}^*_n)$ von (1.78) kann folglich am Gitterpunkt $k+N$ die Stellgröße $\boldsymbol{\kappa}_{k+N}(\tilde{\mathbf{x}}^*_{k+N})$ gemäß dem Zustandsregelgesetz (1.80) angefügt werden. Die so erweiterte Lösung soll mit

$$\tilde{\tilde{\mathbf{u}}}_n = \begin{cases}\tilde{\mathbf{u}}^*_{n+1} & \text{falls } n \in \{-1, 0, 1, \ldots, N-2\} \\ \boldsymbol{\kappa}_{k+N}(\tilde{\mathbf{x}}^*_N) & \text{falls } n = N-1\end{cases} \tag{1.87}$$

und die zugehörige Zustandsfolge mit

$$\tilde{\tilde{\mathbf{x}}}_n = \begin{cases}\tilde{\mathbf{x}}^*_{n+1} & \text{falls } n \in \{-1, 0, 1, \ldots, N-1\} \\ F(\tilde{\mathbf{x}}^*_N, \boldsymbol{\kappa}_{k+N}(\tilde{\mathbf{x}}^*_N)) & \text{falls } n = N\end{cases} \tag{1.88}$$

bezeichnet werden. Wird dem MPC Regelgesetz (1.20) entsprechend zum Zeitpunkt t_k die Stellgröße $\mathbf{u}_k = \tilde{\mathbf{u}}^*_0$ aufgeschaltet, dann ist $(\tilde{\tilde{\mathbf{u}}}_n)$ eine zulässige aber im Allgemeinen nicht optimale Eingangstrajektorie der Optimierungsaufgabe (1.78) für den nachfolgenden Zeithorizont $k+1, \ldots, k+N+1$ mit dem Anfangszustand $\mathbf{x}_{k+1} = \tilde{\mathbf{x}}^*_1$. Unter Berücksichtigung von (1.81) gilt folglich für die optimalen Werte der Gütefunktion

$$\begin{aligned}J^*_{d,k+1|N}(\mathbf{x}_{k+1}) &\leq \gamma D(\tilde{\tilde{\mathbf{x}}}_{N-1}) + \sum_{n=0}^{N-2} d_{k+1+n}(\tilde{\tilde{\mathbf{x}}}_n, \tilde{\tilde{\mathbf{u}}}_n) \\ &= J^*_{d,k|N}(\mathbf{x}_k) - d_k(\mathbf{x}_k, \mathbf{u}_k)\ .\end{aligned} \tag{1.89}$$

Daraus folgt nun noch

$$\Gamma\Big(\gamma + N\frac{d}{\overline{D}}\Big) \geq J^*_{d,k|N}(\mathbf{x}_k) \geq J^*_{d,k+1|N}(\mathbf{x}_{k+1}) \tag{1.90}$$

und somit $\mathbf{x}_{k+1} \in \underline{X}_0$. Der Rest des Beweises erfolgt völlig analog zum Beweis von Satz 1.4. □

In diesem Abschnitt wurde eine MPC Variante besprochen, die zwar kein vorgeschriebenes Endgebiet aber einen Endkostenterm benötigt. Es wurde beobachtet, dass zwischen der Gewichtung γ des Endkostenterms, der Horizontlänge T bzw. N und der Größe der Menge $\underline{X}_0$, die im Einzugsbereich liegt, ein Zusammenhang besteht. Da der Endkostenterm unerwünschte Auswirkungen auf das Verhalten des geschlossenen Regelkreises haben kann, wurden MPC Varianten entwickelt, die ohne einen Endkostenterm auskommen. Nachdem MPC mit unendlich langem Prädiktionshorizont ebenfalls weder einen Endkostenterm noch einen vorgeschriebenes Endgebiet benötigt, ist es nicht verwunderlich, dass eine Mindestlänge für den Prädiktionshorizont existiert, so dass auch bei Verzicht auf einen Endkostenterm und ein vorgeschriebenes Endgebiet exponentielle oder zumindest asymptotische Stabilität nachgewiesen werden kann. Dies wurde für den zeitkontinuierlichen Fall in [1.53] und für den zeitdiskreten Fall in [1.54] gezeigt und wird in dieser Vorlesung nicht weiter besprochen.

1.2.5 Endlicher Prädiktionshorizont mit vorgeschriebenem Endgebiet

Es wird ein zu der in Abschnitt 1.2.3 beschriebenen MPC Variante verwandter Ansatz vorgestellt, der ebenfalls ein vorgeschriebenes Endgebiet benötigt aber keinen Endkostenterm C. Die Länge T des Prädiktionshorizonts ist nun keine feste Größe mehr und wird durch die Optimierungsvariable $\tilde{T}$ ersetzt, d. h. es werden Prädiktionshorizonte mit variabler Länge zugelassen. Die Idee für diesen Ansatz entstammt [1.47][3]. Der Einfachheit halber wird hier nur die zeitkontinuierliche Variante dieses MPC Ansatzes vorgestellt. Die Umsetzung der gleichen Idee in zeitdiskreter Form ist einfach möglich, erfordert aber weitere Überlegungen, denn wegen der zusätzlichen Optimierungsvariable $\tilde{N} \in \mathbb{N}_{>0}$ für die variable Horizontlänge tritt eine gemischt-ganzzahlige Optimierungsaufgabe auf.

Mit der zusätzlichen Optimierungsvariable $\tilde{T}$ kann die zeitkontinuierliche Optimierungsaufgabe (1.15) in die Form

$$(\tilde{\mathbf{u}}^*(\,\cdot\,), \tilde{T}^*) = \arg\min_{(\tilde{\mathbf{u}}(\,\cdot\,), \tilde{T})} \quad J_{\tilde{T}}(t, \mathbf{x}(t), \tilde{\mathbf{u}}(\,\cdot\,)) = \int_0^{\tilde{T}} c(t+\tau, \tilde{\mathbf{x}}(\tau), \tilde{\mathbf{u}}(\tau))\,\mathrm{d}\tau \tag{1.91a}$$

$$\text{u.B.v.} \quad \dot{\tilde{\mathbf{x}}}(\tau) = \mathbf{f}(\tilde{\mathbf{x}}(\tau), \tilde{\mathbf{u}}(\tau))\ , \quad \tilde{\mathbf{x}}(0) = \mathbf{x}(t) \tag{1.91b}$$

$$\tilde{\mathbf{x}}(\tau) \in X\ , \quad \tilde{\mathbf{u}}(\tau) \in U\ , \quad \forall\, \tau \in [0, \tilde{T}] \tag{1.91c}$$

$$\tilde{\mathbf{x}}(\tilde{T}) \in X_T \tag{1.91d}$$

$$\tilde{T} \in (0, \infty) \tag{1.91e}$$

[3] Ein anderer Ansatz, der auch bei Prädiktionshorizonten mit fester Länge keinen Endkostenterm benötigt, findet sich in [1.43].

umgeschrieben werden. Ferner wird ein Zustandsregelgesetz

$$\mathbf{u}(t) = \boldsymbol{\kappa}(t, \mathbf{x}(t)) \tag{1.92}$$

mit $\boldsymbol{\kappa} : \mathbb{R} \times X_T \to U$ benötigt, das folgenden Anforderungen genügt: Das mit dem Zustandsregelgesetz $\boldsymbol{\kappa}(\tau + t, \bar{\mathbf{x}}(\tau))$ geregelte System (1.1) ist lokal asymptotisch (exponentiell) stabil mit dem Einzugsbereich X_T. Für die mit $\bar{\mathbf{x}}(\tau)$ bezeichneten Zustandstrajektorien dieses geregelten Systems gilt für beliebige Anfangszustände $\bar{\mathbf{x}}(0) = \mathbf{x}(t) \in X_T$, dass $\bar{\mathbf{x}}(\tau) \in X \ \forall \tau \geq 0$.

Der in [1.47] vorgeschlagene MPC Algorithmus wird als *Zwei-Phasen MPC* (Englisch: *dual-mode MPC*) bezeichnet und ist in Tabelle 1.1 zusammengefasst. Die Schritte 1 und 2 realisieren eine permanent auszuführende Schleife. Schritt 3 wird wiederkehrend auf einem Zeitgitter mit dem Abtastintervall T_c ausgeführt.

Initialisierung	Wenn $\mathbf{x}(t) \notin X_T$, wähle ein Paar $(\tilde{\mathbf{u}}(\,\cdot\,), \tilde{T})$, so dass (1.91b)-(1.91e) und $J_{\tilde{T}}(t, \mathbf{x}(t), \tilde{\mathbf{u}}(\,\cdot\,)) < \infty$ erfüllt sind, und setze $t_c = t$.
Schritt 1	Wenn $\mathbf{x}(t) \in X_T$, gehe zu **Ende**.
Schritt 2	Wenn $t < t_c + T_c$, verwende $\mathbf{u}(t) = \tilde{\mathbf{u}}(t - t_c)$ und gehe zu **Schritt 1**.
Schritt 3	Setze $\tilde{\tilde{T}} = \tilde{T} - T_c$ und $\tilde{\tilde{\mathbf{u}}}(\tau) = \tilde{\mathbf{u}}(\tau + T_c)\ \forall \tau \in [0, \tilde{\tilde{T}}]$. Wähle ein Paar $(\tilde{\mathbf{u}}(\,\cdot\,), \tilde{T})$, so dass (1.91b)-(1.91e) und $$J_{\tilde{T}}(t, \mathbf{x}(t), \tilde{\mathbf{u}}(\,\cdot\,)) \leq J_{\tilde{\tilde{T}}}(t, \mathbf{x}(t), \tilde{\tilde{\mathbf{u}}}(\,\cdot\,)) \tag{1.93}$$ erfüllt sind, setze $t_c = t$ und gehe zu **Schritt 1**.
Ende	Verwende fortan das asymptotisch (exponentiell) stabilisierende Regelgesetz (1.92).

Tabelle 1.1: Zwei-Phasen MPC gemäß [1.47].

Der Regler gemäß Tabelle 1.1 weist einige Eigenschaften auf, die ihn von den bisher beschriebenen MPC Formulierungen unterscheiden:

- Es ist zu keinem Zeitpunkt die Lösung einer Optimierungsaufgabe nötig. Es werden lediglich Paare $(\tilde{\mathbf{u}}(\,\cdot\,), \tilde{T})$ benötigt, die im Sinne von (1.91) zulässig sind und die Bedingung $J_{\tilde{T}}(t, \mathbf{x}(t), \tilde{\mathbf{u}}(\,\cdot\,)) < \infty$ erfüllen. Man spricht daher auch von einer *suboptimalen* MPC Formulierung.

- Das Finden eines Paares $(\tilde{\mathbf{u}}(\,\cdot\,), \tilde{T})$ in Schritt 3 des Algorithmus ist immer möglich. Trivialerweise kann auch $(\tilde{\mathbf{u}}(\,\cdot\,), \tilde{T}) = (\tilde{\tilde{\mathbf{u}}}(\,\cdot\,), \tilde{\tilde{T}})$ verwendet werden, was aber zu keiner Verbesserung im Sinne der Gütefunktion führt.

- Sobald der Zustand $\mathbf{x}(t)$ das Gebiet X_T erreicht, wird auf das Zustandsregelgesetz (1.92) umgeschaltet. Ab diesem Umschaltzeitpunkt erfolgt keine weitere Optimierung der Stellgröße im Sinne des Gütefunktionals (1.91a). Dieser eventuell nachteilige

Umstand sollte auch bei der Festlegung von X_T berücksichtigt werden. Wie der nachfolgende Satz zeigt, wird dieser Umschaltzeitpunkt in endlicher Zeit erreicht.

- Auch der Fall $\tilde{T} < T_c$ ist hier zulässig. Er impliziert, dass innerhalb des aktuellen Abtastintervalls auf das Regelgesetz (1.92) umgeschaltet wird und der Algorithmus terminiert.

Satz 1.9 (Stabilität der Zwei-Phasen MPC, zeitkontinuierlich)**.** *Es seien die Annahmen A1 und A3 erfüllt. Es existiere ein Zustandsregelgesetz*

$$\mathbf{u}(t) = \boldsymbol{\kappa}(t, \mathbf{x}(t)) \tag{1.94}$$

mit $\boldsymbol{\kappa} : \mathbb{R} \times X_T \to U$, so dass das damit geregelte System (1.1) lokal asymptotisch (exponentiell) stabil mit dem Einzugsbereich X_T ist. Es gelte $T_c > 0$ und es existiere ein $\alpha > 0$, so dass $\{\mathbf{x} \in X \mid ||\mathbf{x}||_2^2 \leq \alpha\} \subseteq X_T$. $X_0 \subseteq X$ sei eine nichtleere Menge, genau so dass (1.91) für $\forall\, \mathbf{x}(t) \in X_0 \backslash X_T$ eine Lösung besitzt (nicht jedoch für $\forall\, \mathbf{x}(t) \in \mathbb{R}^l \backslash X_0$). Dann ist die Ruhelage $\mathbf{x}_R = \mathbf{0}$ des mit der Zwei-Phasen MPC gemäß Tabelle 1.1 geregelten Systems (1.1) lokal asymptotisch (exponentiell) stabil mit dem Einzugsbereich X_0 und die Zustandstrajektorie erreicht die Menge X_T in endlicher Zeit.

Beweis. Wenn für den Anfangszustand $\mathbf{x}(t) \in X_T$ gilt, ist nichts zu zeigen. Es müssen daher nur noch die Fälle $\mathbf{x}(t) \in X_0 \backslash X_T$ untersucht werden. Falls $\tilde{T} \leq T_c$, so erreicht die Zustandstrajektorie das Gebiet X_T innerhalb des aktuellen Abtastintervalls und es ist nichts weiter zu zeigen.

Für die übrigen Fälle $\tilde{T} > T_c$ wird das System (1.1) im Abtastintervall (Steuerungshorizont) $[t, t + T_c]$ mit der geplanten Trajektorie $\tilde{\mathbf{u}}(\,\cdot\,)$ betrieben und es gilt im betrachteten nominellen Fall $\mathbf{x}(t + T_c) = \tilde{\mathbf{x}}(T_c)$. Ferner gilt wegen $T_c > 0$, (1.17b), $\mathbf{x} \notin X_T \Rightarrow ||\mathbf{x}||_2^2 > \alpha$ und $\tilde{\mathbf{x}}(\tau) \notin X_T\ \forall\, \tau \in [0, T_c]$, dass

$$\begin{aligned} J_{\tilde{\tilde{T}}}(t + T_c, \mathbf{x}(t + T_c), \tilde{\tilde{\mathbf{u}}}(\,\cdot\,)) &= J_{\tilde{T}}(t, \mathbf{x}(t), \tilde{\mathbf{u}}(\,\cdot\,)) - \int_0^{T_c} c(t + \tau, \tilde{\mathbf{x}}(\tau), \tilde{\mathbf{u}}(\tau))\, \mathrm{d}\tau\ , \\ &\leq J_{\tilde{T}}(t, \mathbf{x}(t), \tilde{\mathbf{u}}(\,\cdot\,)) - T_c \underline{c} \alpha\ , \end{aligned} \tag{1.95}$$

wobei $\tilde{\tilde{T}}$ und $\tilde{\tilde{\mathbf{u}}}(\,\cdot\,)$ in Schritt 3 des Algorithmus (siehe Tabelle 1.1) definiert sind. Es sei nun das Paar $(\tilde{\mathbf{u}}(\,\cdot\,), \tilde{T})$ jenes, das zum Zeitpunkt t in Schritt 3 des Algorithmus gewählt wurde, und $(\tilde{\mathbf{u}}'(\,\cdot\,), \tilde{T}')$ jenes, das zum Zeitpunkt $t + T_c$ gewählt wurde. In Schritt 3 des Algorithmus kann (1.93) trivial durch die mögliche Wahl $(\tilde{\mathbf{u}}(\,\cdot\,), \tilde{T}) = (\tilde{\tilde{\mathbf{u}}}(\,\cdot\,), \tilde{\tilde{T}})$ erfüllt werden. Aus (1.93) und (1.95) folgt

$$J_{\tilde{T}'}(t + T_c, \mathbf{x}(t + T_c), \tilde{\mathbf{u}}'(\,\cdot\,)) \leq J_{\tilde{T}}(t, \mathbf{x}(t), \tilde{\mathbf{u}}(\,\cdot\,)) - T_c \underline{c} \alpha\ . \tag{1.96}$$

Es gilt

$$J_{\tilde{T}}(t, \mathbf{x}(t), \tilde{\mathbf{u}}(\,\cdot\,)) \leq T_c \underline{c} \alpha \quad \Rightarrow \quad \exists \tau \in [0, T_c] \colon \tilde{\mathbf{x}}(\tau) \in X_T\ . \tag{1.97}$$

Aus einem Anfangswert $J_{\tilde{T}}(t, \mathbf{x}(t), \tilde{\mathbf{u}}(\,\cdot\,)) < \infty$ des Gütefunktionals am Beginn der Regelung (vgl. Tabelle 1.1), der Mindestreduktion $T_c \underline{c} \alpha$ des Gütefunktionals je Abtastschritt gemäß (1.96) und der Abbruchbedingung (1.97) kann sofort eine obere Schranke für jene Zeitspanne berechnet werden, die vergeht bis die Zustandstrajektorie das Gebiet X_T erreicht. □

1.3 Implementierung

Es werden kurz zwei Aspekte der Implementierung betrachtet.

1.3.1 Entwurf eines stabilisierenden Zustandsreglers für ein Endgebiet

In den Abschnitten 1.2.3-1.2.5 wurde für ein vorgeschriebenes oder automatisch erreichtes Endgebiet die Existenz eines stabilisierenden Zustandsreglers vorausgesetzt. Zusätzlich wurde gefordert, dass der mit diesem Regler geschlossene Regelkreis zu einer gewissen Mindestreduktion des Endkostenterms C bzw. D führt (siehe z. B. die Bedingungen (1.55) und (1.61)). Zum Entwurf eines solchen Zustandsregelgesetzes können zahlreiche der in den Vorlesungen *Regelungssysteme* [1.2] und *Nichtlineare dynamische Systeme und Regelung* [1.48] besprochenen Methoden verwendet werden. Im Falle einer stabilisierbaren, linearen, zeitinvarianten Strecke genügt wegen (1.17) bzw. (1.18) meist schon der Entwurf eines LQR-Reglers (siehe [1.2]). Für den Fall einer nichtlinearen Strecke (1.1), deren Linearisierung an der Ruhelage $\mathbf{x}_R = \mathbf{0}$, $\mathbf{u}_R = \mathbf{0}$ stabilisierbar ist, wurde in [1.51] eine Methode zur Konstruktion eines linearen Zustandsreglers vorgeschlagen, der bei quadratischem Gütefunktional die geforderte Mindestreduktion des Endkostenterms lokal sicherstellt. Diese Methode wird hier kurz vorgestellt. Es sei

$$\dot{\boldsymbol{\xi}} = \mathbf{A}\boldsymbol{\xi} + \mathbf{B}\mathbf{u} \tag{1.98}$$

die Linearisierung von (1.1) an der Ruhelage $\mathbf{x}_R = \mathbf{0}$, $\mathbf{u}_R = \mathbf{0}$. Das System (1.98) sei stabilisierbar. Die Funktion $\lambda_{\max}(\,\cdot\,)$ liefere den größten Realteil der Eigenwerte einer Matrix.

Lemma 1.1 (Konstruktion eines linearen Regelgesetzes für das Endgebiet)**.** *Es sei*

$$\mathbf{u} = \mathbf{K}\boldsymbol{\xi} \tag{1.99}$$

ein Zustandsregelgesetz für das System (1.98), so dass $\mathbf{A}+\mathbf{BK}$ *eine Hurwitzmatrix ist. Mit einer Konstante* $\kappa \in [0, -\lambda_{\max}(\mathbf{A}+\mathbf{BK}))$ *und den symmetrisch, positiv definiten Matrizen* $\mathbf{Q} \in \mathbb{R}^{l\times l}$ *und* $\mathbf{R} \in \mathbb{R}^{m\times m}$ *hat die Lyapunov-Gleichung*

$$(\mathbf{A}+\mathbf{BK}+\kappa\mathbf{E})^{\mathrm{T}}\mathbf{P} + \mathbf{P}(\mathbf{A}+\mathbf{BK}+\kappa\mathbf{E}) + \mathbf{Q} + \mathbf{K}^{\mathrm{T}}\mathbf{R}\mathbf{K} = \mathbf{0} \tag{1.100}$$

daher eine eindeutige und symmetrisch, positiv definite Lösung $\mathbf{P}$*. Ferner existiert eine Konstante* $\Gamma \in \mathbb{R}_{>0}$*, so dass auf dem Gebiet*

$$X_\Gamma = \{\mathbf{x} \in X | \mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{x} \le \Gamma\} \supset \{\mathbf{0}\} \tag{1.101}$$

Folgendes gilt:

- *Der Regler hält Stellgrößenbeschränkungen der Art* $\mathbf{K}\boldsymbol{\xi} \in U \ \forall\, \boldsymbol{\xi} \in X_\Gamma$ *ein.*
- X_Γ *ist eine positiv invariante Menge des mit* $\mathbf{u} = \mathbf{K}\mathbf{x}$ *geregelten nichtlinearen Systems* (1.1).
- *Die Trajektorien* $\bar{\mathbf{x}}(\tau)$ *und* $\bar{\mathbf{u}}(\tau)$ *mit* $\tau \geq 0$ *des gemäß* $\bar{\mathbf{u}}(\tau) = \mathbf{K}\bar{\mathbf{x}}(\tau)$ *geregelten nichtlinearen Systems* (1.1) *erfüllen für jeden Anfangszustand* $\bar{\mathbf{x}}(0) = \mathbf{x}(t) \in X_\Gamma$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\bar{\mathbf{x}}^{\mathrm{T}}(t)\mathbf{P}\bar{\mathbf{x}}(t) &\leq \bar{\mathbf{x}}^{\mathrm{T}}(t)((\mathbf{A}+\mathbf{B}\mathbf{K}+\kappa\mathbf{E})^{\mathrm{T}}\mathbf{P}+\mathbf{P}(\mathbf{A}+\mathbf{B}\mathbf{K}+\kappa\mathbf{E}))\bar{\mathbf{x}}(t)\\ &= -\bar{\mathbf{x}}^{\mathrm{T}}(t)(\mathbf{Q}+\mathbf{K}^{\mathrm{T}}\mathbf{R}\mathbf{K})\bar{\mathbf{x}}(t)\end{aligned} \tag{1.102}$$

und folglich auch

$$\mathbf{x}^{\mathrm{T}}(t)\mathbf{P}\mathbf{x}(t) \geq \int_0^\infty \bar{\mathbf{x}}^{\mathrm{T}}(\tau)\mathbf{Q}\bar{\mathbf{x}}(\tau)+\bar{\mathbf{u}}^{\mathrm{T}}(\tau)\mathbf{R}\bar{\mathbf{u}}(\tau)\,\mathrm{d}\tau\ . \tag{1.103}$$

Der Nachweis dieses Lemmas ist in [1.51] zu finden. Aus dem Lemma folgt direkt, dass im Falle $C(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{x}$ und $c(t,\mathbf{x},\mathbf{u}) = \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}+\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u}$ das Regelgesetz (1.99) die Anforderungen der Sätze 1.5, 1.7 und 1.9 erfüllt. Der zeitdiskrete Fall kann analog zu Lemma 1.1 behandelt werden.

1.3.2 Rechenzeit zur Ausführung des Regelgesetzes

Bislang wurde angenommen, dass genau am Beginn des Prädiktionshorizontes, also zum Zeitpunkt t bzw. t_k die dynamische Optimierungsaufgabe, z. B. (1.15) oder (1.16), instantan (also mit Rechenzeit Null) exakt oder zumindest suboptimal gelöst werden kann. Dies wäre nötig, da der Anfangszustand $\mathbf{x}(t)$ oder $\mathbf{x}_k$ nicht vor diesem Zeitpunkt bekannt ist und ab diesem Zeitpunkt aber bereits die neu berechnete Stellgröße aufgeschaltet werden muss (siehe auch Abschnitt 1.1.2). Da die Lösung des Optimierungsproblems praktisch immer eine von Null verschiedene Rechenzeit beansprucht, ist diese bisherige Annahme nicht realistisch.

Praktisch kann bei der Implementierung wie folgt vorgegangen werden, um bereits das vorhergehende Abtastintervall $(t-T_c, t)$ oder (t_{k-1}, t_k) für Berechnungen zu nutzen. Zum Abtastzeitpunkt $t-T_c$ bzw. t_{k-1} werden *Messungen* vorgenommen, aus denen der Systemzustand $\mathbf{x}(t-T_c)$ oder $\mathbf{x}_{k-1}$ unmittelbar folgt oder geschätzt werden kann. Diese *Schätzung* kann natürlich Rechenzeit in Anspruch nehmen. Anschließend wird ausgehend von $\mathbf{x}(t-T_c)$ oder $\mathbf{x}_{k-1}$ das dynamische Modell (1.1) oder (1.3) mit der bereits (früher festgelegten und daher) bekannten Stellgröße für den Steuerungshorizont $[t-T_c, t]$ oder $[t_{k-1}, t_k]$ integriert, um den zukünftigen Zustand $\mathbf{x}(t)$ oder $\mathbf{x}_k$ zu prädizieren. Diese (möglichst genaue) *Simulation* kann natürlich Rechenzeit in Anspruch nehmen. Anschließend wird die *Lösung* der dynamischen Optimierungsaufgabe (1.15) oder (1.16) für den prädizierten Anfangszustand $\mathbf{x}(t)$ oder $\mathbf{x}_k$ berechnet. Diese Rechnung muss vor dem Zeitpunkt t bzw. t_k abgeschlossen sein. Die Implementierung wird daher als *echtzeitfähig* bezeichnet, wenn innerhalb der zur Verfügung stehenden Abtastzeit T_c nacheinander

die Auswertung der Messung, die besagte Schätzung des Anfangszustandes, die besagte Simulation und die Lösung der dynamischen Optimierungsaufgabe berechnet werden können.

1.3.3 Methoden zur Lösung von Optimalsteuerungsaufgaben

Die Lösung einer dynamischen Optimierungsaufgabe, z. B. (1.15) oder (1.16), stellt meist eine zentrale Herausforderung bei der echtzeitfähigen Implementierung einer MPC dar. Ob eine solche Lösung in der zur Verfügung stehenden Zeit gelingt und mit welcher Methode dies erfolgen soll, hängt auch wesentlich von der Problemformulierung (Wahl des Gütefunktionals oder der Gütefunktion, Festlegung von Beschränkungen, etc.) ab. Es werden kurz einige gängige Lösungsstrategien besprochen.

Zeitkontinuierliche Formulierungen müssen im Regelfall in eine zeitdiskrete Formulierung transformiert werden, um auf einem Rechner implementiert zu werden. Hierbei wird zwischen *direkten* und *indirekten* Verfahren unterschieden:

- Wie in der Vorlesung *Optimierung* [1.30] ausführlich besprochen, wird die dynamische Optimierungsaufgabe bei *indirekten Verfahren* zunächst mittels Variationsrechnung oder dem Minimumsprinzip von Pontryagin in zeitkontinuierliche Optimalitätsbedingungen in Form eines Randwertproblems umgeschrieben. Das erhaltene Randwertproblem kann meist mit gängigen numerischen Methoden [1.55–1.57], wie z. B. Einfach-Schießverfahren, Mehrfach-Schießverfahren und Kollokationsverfahren, gelöst werden. Gelegentlich ist eine Lösung nicht möglich, z. B. wenn adjungierte Variablen zufolge von Zustandsbeschränkungen unstetig sind.

- Bei *direkten Verfahren* wird die zeitkontinuierliche Optimierungsaufgabe mittels numerischer Integrationsverfahren [1.55–1.57] direkt diskretisiert. Dies betrifft vor allem integrale Kostenanteile (vgl. (1.15a)) im Gütefunktional und Nebenbedingungen in Form von Differenzialgleichungen (vgl. (1.15b)). Das Resultat lässt sich im Allgemeinen in der Form (1.16) darstellen. Werden implizite Zeitintegrationsverfahren verwendet, so ist (1.16b) durch eine implizite Differenzengleichung zu ersetzen. Die gewählte Diskretisierungsschrittweite sowie die Art der Parametrierung der Eingangsgröße (vgl. (1.4)) haben einen direkten Einfluss auf die erzielte Genauigkeit und auf die Dimension der resultierenden zeitdiskreten Optimierungsaufgabe. In [1.6] wird studiert welche Auswirkungen Diskretisierungsfehler auf die Stabilität und Konvergenz von MPC haben können.

Zeitdiskrete Formulierungen können (außer bei unendlich langem Prädiktionshorizont) als finit-dimensionale, statische Optimierungsaufgaben aufgefasst werden. D. h. sie können in der Form

$$\min_{\mathbf{z}\in\mathbb{R}^o} \quad J(\mathbf{z}) \tag{1.104a}$$

$$\text{u.B.v.} \quad \mathbf{g}(\mathbf{z}) = \mathbf{0} \tag{1.104b}$$

$$\mathbf{h}(\mathbf{z}) \leq \mathbf{0} \tag{1.104c}$$

angeschrieben werden. Hierbei hat der Suchraum die Dimension o, $J : \mathbb{R}^o \to \mathbb{R}_{\geq 0}$ ist eine entsprechende Kostenfunktion und alle Beschränkungen sind in $\mathbf{g} : \mathbb{R}^o \to \mathbb{R}^p$ und $\mathbf{h} : \mathbb{R}^o \to \mathbb{R}^q$ zusammengefasst.

Zur Lösung der Standardaufgabe (1.104) stehen zahlreiche numerische Algorithmen zur Verfügung (siehe [1.31–1.42]). Viele wurden auch bereits in der Vorlesung *Optimierung* [1.30] besprochen, so dass hier auf eine weitere Vertiefung verzichtet werden kann. Es verbleibt die Frage, wie die zeitdiskrete Optimierungsaufgabe (1.16) bzw. ihre Abwandlungen (1.33), (1.48) und (1.78) in die Form (1.104) umgeschrieben werden. Nachfolgende aus [1.6] entnommene Varianten zeigen, dass die Antwort auf diese Frage keineswegs eindeutig ist.

- **Volldiskretisierung:** In der Optimierungsaufgabe (1.16) sind die Größen $\tilde{\mathbf{x}}_0$, $\tilde{\mathbf{x}}_1$, …, $\tilde{\mathbf{x}}_N$ und $\tilde{\mathbf{u}}_0$, $\tilde{\mathbf{u}}_1$, …, $\tilde{\mathbf{u}}_{N-1}$ unbekannt. Es ist daher naheliegend, sie alle im Vektor der Optimierungsvariablen

$$\mathbf{z} = \begin{bmatrix} \tilde{\mathbf{x}}_0^{\mathrm{T}} & \tilde{\mathbf{x}}_1^{\mathrm{T}} & \cdots & \tilde{\mathbf{x}}_N^{\mathrm{T}} & \tilde{\mathbf{u}}_0^{\mathrm{T}} & \tilde{\mathbf{u}}_1^{\mathrm{T}} & \cdots & \tilde{\mathbf{u}}_{N-1}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \tag{1.105}$$

  zusammenzufassen. Damit hat die Optimierungsaufgabe die Dimension $o = (N+1)l + NM$. Die Funktionen J, $\mathbf{g}$ und $\mathbf{h}$ folgen direkt aus (1.16). Diese Variante hat den Nachteil, dass sie im Allgemeinen zu einer hochdimensionalen Optimierungsaufgabe führt. Sie hat den Vorteil, dass Ableitungen von J, $\mathbf{g}$ und $\mathbf{h}$ bezüglich $\mathbf{z}$, wie sie oft von numerischen Lösungsverfahren benötigt werden, sehr einfach analytisch berechnet werden können. Man beachte, dass die unbekannten Zustände $\tilde{\mathbf{x}}_n$ und die gesuchten Eingangsparameter $\tilde{\mathbf{u}}_n$ grundsätzlich mit der gleichen Genauigkeit bestimmt werden, was insbesondere dann von Bedeutung ist, wenn ein iteratives Lösungsverfahren in einer MPC Implementierung frühzeitig abgebrochen wird.

- **Unterlagerte Zeitintegration:** Diese Variante wird auch Teildiskretisierung oder reduzierte Diskretisierung genannt [1.58]. Entsprechend (1.16b) lassen sich die Zustandsgrößen $\tilde{\mathbf{x}}_0$, $\tilde{\mathbf{x}}_1$, …, $\tilde{\mathbf{x}}_N$ (zumindest formal) als Funktionen der Eingangsparameter $\tilde{\mathbf{u}}_0$, $\tilde{\mathbf{u}}_1$, …, $\tilde{\mathbf{u}}_{N-1}$ und des bekannten Anfangszustands $\mathbf{x}_k$ ausdrücken. D. h. es existiert eine eindeutige Abbildung

$$\begin{bmatrix} \mathbf{x}_k^{\mathrm{T}} & \tilde{\mathbf{u}}_0^{\mathrm{T}} & \tilde{\mathbf{u}}_1^{\mathrm{T}} & \cdots & \tilde{\mathbf{u}}_{N-1}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \mapsto \begin{bmatrix} \tilde{\mathbf{x}}_0^{\mathrm{T}} & \tilde{\mathbf{x}}_1^{\mathrm{T}} & \cdots & \tilde{\mathbf{x}}_N^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}, \tag{1.106}$$

  mit der alle Zustandsvektoren $\tilde{\mathbf{x}}_n$ in (1.16) eliminiert werden können. Wird die verbleibende Optimierungsaufgabe in die Form (1.104) umgeschrieben, so lauten die Optimierungsvariablen

$$\mathbf{z} = \begin{bmatrix} \tilde{\mathbf{u}}_0^{\mathrm{T}} & \tilde{\mathbf{u}}_1^{\mathrm{T}} & \cdots & \tilde{\mathbf{u}}_{N-1}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}. \tag{1.107}$$

  Damit hat die Optimierungsaufgabe die Dimension $o = NM$. Die Funktionen J, $\mathbf{g}$ und $\mathbf{h}$ folgen wieder direkt aus (1.16), wobei die in (1.106) verwendeten Gleichungsbeschränkungen (1.16b) nun nicht mehr in $\mathbf{g}$ aufzunehmen sind. Diese Variante hat den Vorteil, dass sie im Allgemeinen zu einer niedrigdimensionalen Optimierungsaufgabe führt. Sie hat aber den Nachteil, dass totale Ableitungen von J, $\mathbf{g}$ und $\mathbf{h}$

bezüglich $\mathbf{z}$ grundsätzlich auch das meist numerisch sensitive Nachdifferenzieren der Abbildung (1.106) erfordern. Bei der Implementierung von MPC Algorithmen wird (1.106) oft als unterlagerte Zeitintegrationsroutine realisiert. Daher ist es auch möglich, die Genauigkeit der Zeitintegration des dynamischen Systems unabhängig von einer Genauigkeitsanforderung oder einem allfälligen frühzeitigen Abbruch des iterativen Optimierungsverfahrens (suboptimale MPC) vorzugeben.

- **Mehrfach-Schießverfahren:** Im Mehrfach-Schießverfahren werden die Grundideen der *Volldiskretisierung* und der *unterlagerten Zeitintegration* vereint. Es werden daher nicht alle sondern nur einige der unbekannten Zustandsvektoren $\tilde{\mathbf{x}}_0, \tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_N$ zu den Optimierungsvariablen $\mathbf{z}$ hinzugenommen. Werden die Zustandsvektoren $\tilde{\mathbf{x}}_{n_i}$ mit $n_i \in \{0, 1, \ldots, N\}$ und $0 \leq n_1 < n_2 < \ldots < n_I \leq N$ als Optimierungsvariablen in den Vektor

$$\mathbf{z} = \begin{bmatrix} \tilde{\mathbf{x}}_{n_1}^{\mathrm{T}} & \tilde{\mathbf{x}}_{n_2}^{\mathrm{T}} & \cdots & \tilde{\mathbf{x}}_{n_I}^{\mathrm{T}} & \tilde{\mathbf{u}}_0^{\mathrm{T}} & \tilde{\mathbf{u}}_1^{\mathrm{T}} & \cdots & \tilde{\mathbf{u}}_{N-1}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \tag{1.108}$$

aufgenommen, so hat dieser die Dimension $o = Il + NM$ und von den Gleichungsbeschränkungen (1.16b) sind nur jene in $\mathbf{g}$ aufzunehmen, die auf ihrer linken Seite ein unbekanntes $\tilde{\mathbf{x}}_{n_i}$ stehen haben.

1.4 Literatur

[1.1] F. Allgöwer, T. Badgwell, J. Qin, J. Rawlings und S. Wright, „Nonlinear Predictive Control and Moving Horizon Estimation: An Introductory Overview," in *Advances in Control*, Springer London, 1999, S. 391–449.

[1.2] W. Kemmetmüller, *Skriptum zur VO Regelungssysteme (WS 2023/2024)*, Institut für Automatisierungs- und Regelungstechnik, TU Wien, 2023. Adresse: https://www.acin.tuwien.ac.at/master/regelungssysteme/.

[1.3] M. Alamir, *A Pragmatic Story of Model Predictive Control: Self-Contained Algorithms and Case-Studies*. CreateSpace Independent Publishing Platform, 2013.

[1.4] J.A. Rossiter, *Model-Based Predictive Control: A Practical Approach*. Boca Raton, Florida: CRC Press, 2003.

[1.5] M. Alamir, *Stabilization of Nonlinear Systems Using Receding-horizon Control Schemes: A Parametrized Approach for Fast Systems* (Lecture Notes in Control and Information Sciences). Springer, 2006, Bd. 339.

[1.6] L. Grüne und J. Pannek, *Nonlinear Model Predictive Control*, 2. Aufl. London: Springer, 2017.

[1.7] F. Borrelli, A. Bemporad und M. Morari, *Predictive Control for Linear and Hybrid Systems*. Cambridge University Press, 2015, in Druck.

[1.8] E.F. Camacho und C. Bordons, *Model Predictive Control* (Advanced Textbooks in Control and Signal Processing), 2. Aufl. Springer-Verlag, 2004.

[1.9] R. Dittmar und B.M. Pfeiffer, *Modellbasierte prädiktive Regelung: Eine Einführung für Ingenieure*. München: Oldenbourg, 2004.

[1.10] B. Kouvaritakis und M. Cannon, *Model Predictive Control: Classical, Robust and Stochastic*. Cham: Springer, 2016.

[1.11] W. Kwon und S. Han, *Receding Horizon Control* (Advanced Textbooks in Control and Signal Processing). Springer, 2005.

[1.12] J.M. Maciejowski, *Predictive Control with Constraints*. Prentice Hall, 2002.

[1.13] J. Maestre und R. Negenborn, Hrsg., *Distributed Model Predictive Control Made Easy* (Intelligent Systems, Control and Automation: Science and Engineering). Springer, 2014, Bd. 69.

[1.14] J.B. Rawlings, D.Q. Mayne und M.M. Diehl, *Model Predictive Control: Theory, Computation, and Design*, 2. Aufl. Madison, Wisconsin: Nob Hill Publishing, 2017.

[1.15] F. Allgöwer und A. Zheng, Hrsg., *Nonlinear Model Predictive Control*, Bd. 26, Progress in Systems and Control Theory, Basel: Birkhäuser, 2000.

[1.16] R. Findeisen, F. Allgöwer und L.B. Biegler, Hrsg., *Assessment and Future Directions of Nonlinear Model Predictive Control*, Bd. 358, Lecture Notes in Control and Information Sciences, Berlin: Springer, 2007.

[1.17] B. Kouvaritakis und M. Cannon, Hrsg., *Nonlinear Predictive Control: Theory and Practice*, Bd. 61, IEE Control Engineering Series, Herts, UK: Institution of Electrical Engineers, 2001.

[1.18] L. Magni, D.M. Raimondo und F. Allgöwer, Hrsg., *Nonlinear Model Predictive Control*, Bd. 384, Lecture Notes in Control and Information Science, Berlin, Heidelberg: Springer, 2009.

[1.19] S. Raković und W. Levine, Hrsg., *Handbook of Model Predictive Control*. Cham: Birkhäuser, 2019.

[1.20] C.E. Garcia, D.M. Prett und M. Morari, „Model Predictive Control: Theory and Practice - a Survey," *Automatica*, Jg. 25, Nr. 3, S. 335–348, 1989.

[1.21] D.Q. Mayne, J.B. Rawlings, C.V. Rao und P.O.M. Scokaert, „Constrained model predictive control: Stability and optimality," *Automatica*, Jg. 36, Nr. 6, S. 789–814, 2000.

[1.22] S. Qin und T. Badgwell, „A survey of industrial model predictive control technology," *Control Engineering Practice*, Jg. 11, Nr. 7, S. 733–764, 2003.

[1.23] D. Mayne, „Model predictive control: Recent developments and future promise," *Automatica*, Jg. 50, Nr. 12, S. 2967–2986, 2014.

[1.24] J. Richalet, A. Rault, J. Testud und J. Papon, „Model predictive heuristic control: Applications to industrial processes," *Automatica*, Jg. 14, Nr. 5, S. 413–428, 1978.

[1.25] D. Clarke, C. Mohtadi und P. Tuffs, „Generalized predictive control: Part I. The basic algorithm," *Automatica*, Jg. 23, Nr. 2, S. 137–148, 1987.

[1.26] L. Ljung, *System Identification: Theory for the User*, 2. Aufl. Upper Saddle River, New Jersey: Prentice Hall, 1999.

[1.27] P. Scokaert und D. Mayne, „Min-max feedback model predictive control for constrained linear systems," *IEEE Transactions on Automatic Control*, Jg. 43, Nr. 8, S. 1136–1142, Aug. 1998.

[1.28] K. Tae-Hyoung, H. Fukushima und T. Sugie, „Robust adaptive model predictive control based on comparison model," in *Proceedings of the 43rd IEEE Conference on Decision and Control (CDC)*, Bd. 2, Dez. 2004, S. 2041–2046.

[1.29] R. Sanfelice, „Handbook of Model Predictive Control," in Cham: Birkhäuser, 2019, Kap. Hybrid Model Predictive Control.

[1.30] A. Steinböck, *Skriptum zur VU Optimierung (WS 2023/2024)*, Institut für Automatisierungs- und Regelungstechnik, TU Wien, 2023. Adresse: https://www.acin.tuwien.ac.at/master/optimierung/.

[1.31] D.P. Bertsekas, *Nonlinear Programming*, 2. Aufl. Belmont, Massachusetts: Athena Scientific, 1999.

[1.32] E. Polak, *Optimization: Algorithms and Consistent Approximations* (Applied Mathematical Sciences 124). New York: Springer, 1997.

[1.33] J.L. Speyer und D.H. Jacobson, *Primer on Optimal Control Theory* (Advances in Design and Control). Philadelphia: Siam, 2010.

[1.34] K.L. Teo, C.J. Goh und K.H. Wong, *A Unified Computational Approach to Optimal Control Problems* (Pitman Monographs and Surveys in Pure and Applied Mathematics). New York: John Wiley & Sons, 1991.

[1.35] J. T. Betts, *Practical Methods for Optimal Control Using Nonlinear Programming* (Advances in Design and Control). Philadelphia, USA: SIAM - Society for Industrial und Applied Mathematics, 2001.

[1.36] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal und C.A. Sagastizábal, *Numerical Optimization - Theoretical and Practical Aspects*, 2. Aufl. Berlin, Heidelberg: Springer, 2006.

[1.37] S. Boyd und L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004.

[1.38] A.E. Bryson und Y.-C. Ho, *Applied Optimal Control*. New York: John Wiley & Sons, 1975.

[1.39] P.E. Gill, W. Murray und M.H. Wright, *Practical Optimization*. London: Academic Press, 1981.

[1.40] C.T. Kelley, *Iterative Methods for Optimization* (Frontiers in Applied Mathematics 18). Philadelphia: Society for Industrial und Applied Mathematics, 1999.

[1.41] D.G. Luenberger und Y. Ye, *Linear and Nonlinear Programming* (International Series in Operations Research and Management Science), 3. Aufl. New York: Springer, 2008.

[1.42] J. Nocedal und S.J. Wright, *Numerical Optimization* (Springer Series in Operations Research), 2. Aufl. New York: Springer, 2006.

[1.43] P. Scokaert, D. Mayne und J. Rawlings, „Suboptimal model predictive control (feasibility implies stability)," *IEEE Transactions on Automatic Control*, Jg. 44, Nr. 3, S. 648–654, März 1999.

[1.44] K. Graichen und A. Kugi, „Stability of incremental model predictive control without terminal constraints," *IEEE Transactions on Automatic Control*, Jg. 55, Nr. 11, S. 2576–2580, 2010.

[1.45] M. Diehl, R. Findeisen, F. Allgöwer, H.G. Bock und J.P. Schlöder, „Nominal stability of real-time iteration scheme for nonlinear model predictive control," *IEE Proceedings Control Theory and Applications*, Jg. 152, Nr. 3, S. 296–308, Mai 2005.

[1.46] D. DeHaan und M. Guay, „A real-time framework for model-predictive control of continuous-time nonlinear systems," *IEEE Transactions on Automatic Control*, Jg. 52, Nr. 11, S. 2047–2057, Nov. 2007.

[1.47] H. Michalska und D.Q. Mayne, „Robust receding horizon control of constrained nonlinear systems," *IEEE Transactions on Automatic Control*, Jg. 38, Nr. 11, S. 1623–1633, 1993.

[1.48] A. Kugi, *Skriptum zur VO Nichtlineare dynamische Systeme und Regelung (SS 2023)*, Institut für Automatisierungs- und Regelungstechnik, TU Wien, 2023. Adresse: https://www.acin.tuwien.ac.at/master/nichtlineare-dynamische-systeme-und-regelung/.

[1.49] K. Graichen, *Skriptum zur VO Methoden der Optimierung und optimalen Steuerung (WS 2014/2015)*, Institut für Mess-, Regel- und Mikrotechnik, Universität Ulm, 2014.

[1.50] M. Vidyasagar, *Nonlinear Systems Analysis* (Classics in Applied Mathematics 42), 2. Aufl. Philadelphia: SIAM, 1992.

[1.51] H. Chen und F. Allgöwer, „A quasi-infinite horizon nonlinear model predictive control scheme with guaranteed stability," *Automatica*, Jg. 34, Nr. 10, S. 1205–1217, 1998.

[1.52] D. Limon, T. Alamo, F. Salas und E. Camacho, „On the stability of constrained MPC without terminal constraint," *IEEE Transactions on Automatic Control*, Jg. 51, Nr. 5, S. 832–836, Mai 2006.

[1.53] A. Jadbabaie und J. Hauser, „On the Stability of Receding Horizon Control With a General Terminal Cost," *IEEE Transactions on Automatic Control*, Jg. 50, Nr. 5, S. 674–678, 2005.

[1.54] A. Boccia, L. Grüne und K. Worthmann, „Stability and feasibility of state constrained MPC without stabilizing terminal constraints," *Systems & Control Letters*, Jg. 72, Nr. 0, S. 14–21, 2014.

[1.55] H.R. Schwarz und N. Köckler, *Numerische Mathematik*, 6. Aufl. Wiesbaden: B.G. Teubner, 2006.

[1.56] M. Hermann, *Numerik gewöhnlicher Differentialgleichungen: Anfangs- und Randwertprobleme*. München: Oldenbourg, 2004.

[1.57] J. Stoer und R. Bulirsch, *Introduction to Numerical Analysis* (Texts in Applied Mathematics 12), 3. Aufl. New York, Berlin: Springer, 2002.

[1.58] M. Gerdts, *Optimal Control of ODEs and DAEs*. Berlin, Boston: De Gruyter, 2012.

2 Zustandsschätzung auf bewegten Horizonten

In diesem Abschnitt wird die Methode der Zustandsschätzung auf bewegten Horizonten kurz vorgestellt. Es werden dazu die Bestandteile dieser Schätzmethode und Möglichkeiten zur Berücksichtigung von Informationen, die vor dem aktuellen Horizont gesammelt wurden, diskutiert. Abschließend wird eine wahrscheinlichkeitstheoretische Interpretation der Methode gegeben und skizziert, wie neben Zuständen auch Systemparameter geschätzt werden können.

Der Begriff Schätzung auf bewegtem Horizont wird im Englischen oft als *moving horizon estimation (MHE)* oder *receding horizon estimation* bezeichnet. Hier wird der allgemeine Fall der *nichtlinearen modellbasierten* Zustandsschätzung auf bewegten Horizonten behandelt. Mit MHE werden die folgenden Vorgehensweisen in Verbindung gebracht:

- Für einen bestimmten Zeithorizont werden der Anfangszustand eines dynamischen Systems und die auf das System wirkenden Störungen geschätzt.
- Für die Schätzung werden Messwerte des Systemausgangs, ein mathematisches Modell zur Beschreibung des Systemverhaltens und meist Informationen aus früheren Schätzungen verwendet. Die Schätzung beruht auf der Minimierung der Diskrepanz zwischen den Messwerten des Systemausgangs und den mit dem Modell berechneten Ausgangswerten.
- Im Allgemeinen erfordert diese Schätzung die Lösung einer dynamischen Optimierungsaufgabe.
- Die Schätzung wird zu diskreten Zeitpunkten wiederkehrend durchgeführt.

Die modellbasierte Berechnung des Systemverhaltens und die Formulierung der zu lösenden Optimierungsaufgabe erfolgt im Allgemeinen für einen Zeithorizont, der in der Vergangenheit beginnt und zum aktuellen Zeitpunkt endet. Für die wiederkehrende Schätzung muss dieser Horizont zeitlich fortgeschoben werden.

MHE eignet sich gut zur Zustandsschätzung bei Systemen, die

- Beschränkungen unterworfen sind,
- nichtlinear sind,
- nur selten oder zu unregelmäßigen Zeitpunkten Messungen zulassen und
- Störungen oder Rauschen mit nicht näher spezifizierten stochastischen Eigenschaften ausgesetzt sind.

Als potentielle mit MHE verbundene Schwierigkeiten sind zu nennen:

- Die Methode erfordert mitunter einen hohen *Rechenaufwand*, da eine dynamische Optimierungsaufgabe gelöst werden muss.
- Die erstmalige Entwicklung und Implementierung des Schätzverfahrens samt einem Lösungsalgorithmus für die zugehörige dynamische Optimierungsaufgabe kann *aufwendig* sein.

Einen Überblick über die Methode und das Forschungsgebiet MHE bieten die Arbeiten [2.1–2.4]. Darüber hinaus bieten die Beiträge [2.5–2.8] gute Einstiegspunkte in das Thema. In [2.7, 2.9] wird MHE mit dem Extended Kalman-Filter [2.10] verglichen. Das vorliegende Skriptum orientiert sich an [2.11, 2.12].

2.1 Bestandteile von MHE

2.1.1 Modell

Zur Beschreibung des Systemverhaltens dient grundsätzlich ein mathematisches Modell, dessen Eingänge zum Teil Zufallsvariablen sein können. In diesem Abschnitt wird das folgende zeitvariante zeitdiskrete dynamische Modell in Zustandsraumdarstellung verwendet:

$$\mathbf{x}_{k+1} = \mathbf{f}_k(\mathbf{x}_k, \mathbf{w}_k) \qquad \forall\, k \in \mathbb{N}_0 \tag{2.1a}$$

$$\mathbf{y}_k = \mathbf{h}_k(\mathbf{x}_k) + \mathbf{v}_k \qquad \forall\, k \in \mathbb{N}_0\ . \tag{2.1b}$$

Hierbei ist $k \in \mathbb{N}_0$ der Zeitindex, $t_k \in \mathbb{R}_{\geq 0}$ die zugehörige Zeit, $\mathbf{x}_k \in \mathbb{R}^n$ der exakte Systemzustand, $\mathbf{x}_0$ der exakte Anfangszustand, $\mathbf{w}_k \in \mathbb{R}^p$ eine Prozessstörung (Prozessrauschen), $\mathbf{v}_k \in \mathbb{R}^q$ eine Messstörung (Messrauschen) und $\mathbf{y}_k \in \mathbb{R}^q$ der Messwert des Systemausgangs. Allfällige zusätzliche Systemeingänge, z. B. Stelleingänge, seien bekannt und bereits in $\mathbf{f}_k$ enthalten. Das Modell (2.1) kann aus der Abtastung eines zeitkontinuierlichen dynamischen Systems hervorgehen, wobei die Abtastzeit variabel sein kann.

Im Allgemeinen entstammen die Störungen $\mathbf{w}_k$ und $\mathbf{v}_k$ stationären stochastischen Prozessen [2.10]. Der Einfachheit halber werden sie hier als unbekannte, beschränkte, mittelwertfreie Zufallsvariablen aufgefasst. Auch der Anfangszustand $\mathbf{x}_0$ sei eine unbekannte, beschränkte Zufallsvariable.

Es ist nun zu unterscheiden zwischen den Systemvariablen ($\mathbf{x}_k$, $\mathbf{w}_k$, $\mathbf{v}_k$, $\mathbf{y}_k$), den korrespondierenden Größen ($\tilde{\mathbf{x}}_k$, $\tilde{\mathbf{w}}_k$, $\tilde{\mathbf{v}}_k$, $\mathbf{y}_k$) in einer Schätzaufgabe und deren (optimalen) Schätzwerten ($\hat{\mathbf{x}}_k$, $\hat{\mathbf{w}}_k$, $\hat{\mathbf{v}}_k$, $\mathbf{y}_k$), die als Lösung der Schätzaufgabe von einem Beobachter bestimmt werden. Diese Variablen und ihre Zusammenhänge sind in Tabelle 2.1 zusammengefasst. Es ist zu beachten, dass die gemessene Ausgangsgröße $\mathbf{y}_k$ im realen System und in der Schätzaufgabe identisch ist.

2.1.2 Horizont

In den Abschnitten 2.3 und 2.4 wird erläutert, warum zur Zustandsschätzung im Allgemeinen nicht die gesamte Information, die in allen in der Vergangenheit aufgetretenen

	Systemvariablen	Variablen im Schätzproblem	Optimale Schätzwerte
Zustand	$\mathbf{x}_k$	$\tilde{\mathbf{x}}_k$	$\hat{\mathbf{x}}_k$
Prozessstörung	$\mathbf{w}_k$	$\tilde{\mathbf{w}}_k$	$\hat{\mathbf{w}}_k$
Systemdynamik	$\mathbf{x}_{k+1} = \mathbf{f}_k(\mathbf{x}_k, \mathbf{w}_k)$	$\tilde{\mathbf{x}}_{k+1} = \mathbf{f}_k(\tilde{\mathbf{x}}_k, \tilde{\mathbf{w}}_k)$	$\hat{\mathbf{x}}_{k+1} = \mathbf{f}_k(\hat{\mathbf{x}}_k, \hat{\mathbf{w}}_k)$
Messstörung	$\mathbf{v}_k$	$\tilde{\mathbf{v}}_k$	$\hat{\mathbf{v}}_k$
Nominelle Ausgangsgröße	$\mathbf{h}_k(\mathbf{x}_k)$	$\mathbf{h}_k(\tilde{\mathbf{x}}_k)$	$\mathbf{h}_k(\hat{\mathbf{x}}_k)$
Gemessene Ausgangsgröße	$\mathbf{y}_k = \mathbf{h}_k(\mathbf{x}_k) + \mathbf{v}_k$	$\mathbf{y}_k = \mathbf{h}_k(\tilde{\mathbf{x}}_k) + \tilde{\mathbf{v}}_k$	$\mathbf{y}_k = \mathbf{h}_k(\hat{\mathbf{x}}_k) + \hat{\mathbf{v}}_k$

Tabelle 2.1: Variablen des Systems und des Beobachters.

Messwerten $\mathbf{y}_k$ enthalten ist, exakt ausgenützt werden kann. Aus diesem Grund werden bei MHE nur Messwerte aus einem Zeithorizont verwendet, der in der Vergangenheit beginnt, zum aktuellen Zeitpunkt endet und eine finite Länge besitzt. Informationen aus Messwerten, die vor dem aktuellen Horizont generiert wurden, können aber näherungsweise in der aktuellen Schätzung berücksichtigt werden. Um neu hinzukommende Messwerte zu berücksichtigen, wird der Horizont zeitlich fortbewegt und die Schätzung wiederholt. D. h. mit jedem Abtastschritt wird ein neuer Messwert zur Schätzung hinzugenommen und der älteste bisher berücksichtigte Messwert verworfen.

Es soll nun $N \geq 1$ die Länge des Horizonts und K der aktuelle Zeitindex sein, d. h. t_K ist der aktuelle Zeitpunkt. Es werden die Messwerte $\mathbf{y}_{K-N}, \mathbf{y}_{K-N+1}, \ldots, \mathbf{y}_{K-1}$ zur Schätzung herangezogen. Sie werden in der Folge $\mathbf{y}_{K-N|K-1} = (\mathbf{y}_{K-N}, \ldots, \mathbf{y}_{K-1})$ zusammengefasst. Der aktuelle Messwert $\mathbf{y}_K$ soll keine Berücksichtigung mehr finden, damit das Zeitintervall (t_{K-1}, t_K) zur Berechnung der Lösung der Schätzaufgabe zur Verfügung steht. Wird unter Verwendung der Messwerte $\mathbf{y}_{K-N|K-1}$ der Zustand $\mathbf{x}_k$ geschätzt, so handelt es sich im Fall $k \geq K$ strenggenommen um eine *Prädiktion*, im Fall $k = K-1$ um eine *Filterung* und im Fall $k \in \{K-N, K-N+1, \ldots, K-2\}$ um eine *Glättung*. Die hier vorgestellte MHE Variante vereint grundsätzlich diese Vorgehensweisen:

> Zum Zeitpunkt K werden die Zustandsgröße $\mathbf{x}_{K-N}$ und die zu einer Folge zusammengefassten Prozessstörungen $\mathbf{w}_{K-N|K-1} = (\mathbf{w}_{K-N}, \ldots, \mathbf{w}_{K-1})$ geschätzt. Daraus können mithilfe von (2.1a) sofort Schätzwerte für die Zustände $\mathbf{x}_{K-N+1}$, $\mathbf{x}_{K-N+2}$, $\ldots$, $\mathbf{x}_K$ berechnet werden.

Für eine kompaktere Schreibweise soll nun $\check{\mathbf{x}}_k(l, \mathbf{x}_l, \mathbf{w}_{l|k-1})$ mit $l \leq k$ die Lösung von (2.1a) für den Anfangszustand $\mathbf{x}_l$ zum Zeitpunkt t_l und die Störfolge $\mathbf{w}_{l|k-1}$ sein. Im speziellen Fall $l = k$ sei $\mathbf{w}_{l|k-1}$ eine leere Folge und $\check{\mathbf{x}}_k(l, \mathbf{x}_l, \mathbf{w}_{l|k-1}) = \mathbf{x}_l$. In ähnlicher Weise wird auch für die Folge der Messstörungen die abgekürzte Schreibweise $\mathbf{v}_{K-N|K-1} = (\mathbf{v}_{K-N}, \ldots, \mathbf{v}_{K-1})$ verwendet.

Wird der MHE Beobachter neu eingeschaltet, so gilt zunächst für die Horizontlänge $N = K$, d. h. alle verfügbaren Messwerte werden genützt. Bei MHE wird ab einem gewissen Zeitindex N festgehalten und es werden fortan nur noch die Messwerte aus einem Horizont mit fester, finiter Länge N zur Schätzung verwendet.

2.1.3 Beschränkungen

Bei vielen praktischen Anwendungen sind Stör-, Zustands- oder Ausgangsgrößen beschränkt. Hier werden die folgenden Beschränkungen berücksichtigt:

$$\mathbf{x}_k \in X_k \ , \quad \mathbf{w}_k \in W_k \ , \quad \mathbf{v}_k \in V_k \ , \quad \forall\, k \in \mathbb{N}_0 \ . \tag{2.2}$$

Hierbei seien X_k, W_k und V_k zeitvariante Mengen, die

$$X_k \subseteq \mathbb{R}^n \ , \quad \mathbf{0} \in W_k \subseteq \mathbb{R}^p \ , \quad \mathbf{0} \in V_k \subseteq \mathbb{R}^q \tag{2.3}$$

erfüllen. Die Mengen W_k und V_k bieten eine einfache Möglichkeit, die Beschränktheit von Störungen zu modellieren, die sich z. B. auch in Form von lokal verschwindenden Wahrscheinlichkeitsdichten manifestiert. Bei der Interpretation und Festlegung der Menge X_k ist Sorgfalt angebracht: Anders als ein Regler, kann ein Zustandsschätzer die Einhaltung von Schranken in der realen Strecke nicht (direkt) beeinflussen. Ist ein realer Systemzustand tatsächlichen physikalischen Schranken ausgesetzt, so sollten diese impliziert durch das Modell (2.1) abgebildet sein. Ist dies, z. B. aufgrund von Modellungenauigkeiten, nicht der Fall, so können solche Schranken durch eine entsprechende Wahl von X_k erzwungen werden. Durch die Festlegung von X_k können also Modellfehler kompensiert oder Modellvereinfachungen erzielt werden. Eine falsche Wahl der Schranken X_k kann aber zu (unphysikalischen) Akausalitäten führen [2.1] oder die Konvergenz von Schätzwerten gegen ihre wahren Werte verhindern [2.2].

2.1.4 Skalares Gütemaß

Zur Beurteilung der Qualität von Schätzwerten $\tilde{\mathbf{x}}_{K-N}$ und $\tilde{\mathbf{w}}_{K-N|K-1}$ wird ein skalares Gütemaß der Form

$$J_{K|N}(\tilde{\mathbf{x}}_{K-N}, \tilde{\mathbf{w}}_{K-N|K-1}) = B_{K-N}(\tilde{\mathbf{x}}_{K-N}) + \sum_{k=K-N}^{K-1} b_k(\tilde{\mathbf{w}}_k, \tilde{\mathbf{v}}_k) \tag{2.4}$$

mit den Funktionen $B_k : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ und $b_k : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}_{\geq 0}$ minimiert. Die Anfangskostenfunktion $B_{K-N} \geq 0$ enthält bekannte oder zu früheren Zeitpunkten geschätzte Informationen über den Anfangszustand $\mathbf{x}_{K-N}$. Da von mittelwertfreien Störgrößen ausgegangen wird, soll $b_k(\mathbf{w}, \mathbf{v})$ positiv definit bezüglich beider Argumente $\mathbf{w}$ und $\mathbf{v}$ sein. Gemäß Tabelle 2.1 werden die Schätzwerte $\tilde{\mathbf{v}}_k$ für $\forall\, k = K-N, \ldots, K-1$ mittels

$$\tilde{\mathbf{v}}_k = \mathbf{y}_k - \mathbf{h}_k(\check{\mathbf{x}}_k(K-N, \tilde{\mathbf{x}}_{K-N}, \tilde{\mathbf{w}}_{K-N|k-1})) \tag{2.5}$$

berechnet.

2.1.5 Optimierung

Zur Bestimmung der optimalen Schätzgrößen wird die beschränkte statische Optimierungsaufgabe

$$
\begin{aligned}
&(\hat{\mathbf{x}}_{K-N}, \hat{\mathbf{w}}_{K-N|K-1}) = \\
&\quad \underset{\substack{(\tilde{\mathbf{x}}_{K-N},\\ \tilde{\mathbf{w}}_{K-N|K-1})}}{\arg\min} \quad J_{K|N}(\tilde{\mathbf{x}}_{K-N}, \tilde{\mathbf{w}}_{K-N|K-1}) && && \text{(2.6a)}\\
&\qquad \text{u.B.v.} \quad \tilde{\mathbf{x}}_{k+1} = \mathbf{f}_k(\tilde{\mathbf{x}}_k, \tilde{\mathbf{w}}_k) && \forall\, k = K-N, \dots, K-1 && \text{(2.6b)}\\
&\qquad\qquad \tilde{\mathbf{v}}_k = \mathbf{y}_k - \mathbf{h}_k(\tilde{\mathbf{x}}_k) && \forall\, k = K-N, \dots, K-1 && \text{(2.6c)}\\
&\qquad\qquad \tilde{\mathbf{x}}_k \in X_k && \forall\, k = K-N, \dots, K && \text{(2.6d)}\\
&\qquad\qquad \tilde{\mathbf{w}}_k \in W_k\ , \quad \tilde{\mathbf{v}}_k \in V_k\ , && \forall\, k = K-N, \dots, K-1 && \text{(2.6e)}
\end{aligned}
$$

gelöst. In ihr sind das dynamische Modell, die Beschränkungen und das skalare Gütemaß aus den vorhergehenden Abschnitten zusammengefasst. Der optimale Wert der Kostenfunktion von (2.6) wird in der Form $\hat{J}_{K|N} = J_{K|N}(\hat{\mathbf{x}}_{K-N}, \hat{\mathbf{w}}_{K-N|K-1})$ abgekürzt.

Die Lösung der Optimierungsaufgabe (2.6) ist im Allgemeinen mit hohem numerischem Aufwand verbunden, was eine der zentralen Herausforderungen bei MHE Beobachtern sein kann. Da (2.6) strukturell ähnlich zu den bei der modellprädiktiven Regelung auftretenden Optimierungsproblemen ist, wird auf die in Abschnitt 1 angeführten Literaturverweise für Lösungsmethoden verwiesen.

2.1.6 Annahmen

Definition 2.1 (Vergleichsfunktionen)**.** Es werden die folgenden Klassen von Funktionen definiert:

$$
\begin{aligned}
\mathcal{K} &= \{\sigma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0} \,|\, \sigma(0) = 0 \text{ und } \sigma \text{ ist stetig und streng monoton steigend}\}\\
\mathcal{K}_\infty &= \{\sigma \in \mathcal{K} \,|\, \lim_{t\to\infty} \sigma(t) = \infty\}\\
\mathcal{L} &= \{\gamma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0} \,|\, \lim_{t\to\infty} \gamma(t) = 0 \text{ und } \gamma \text{ ist stetig und nicht steigend}\}\\
\mathcal{KL} &= \{\beta : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0} \,|\, \beta(\,\cdot\,, t) \in \mathcal{K} \text{ und } \beta(s, \,\cdot\,) \in \mathcal{L}\}
\end{aligned}
$$

Abbildung 2.1 zeigt ein Beispiel einer Funktion der Klasse $\mathcal{KL}$.

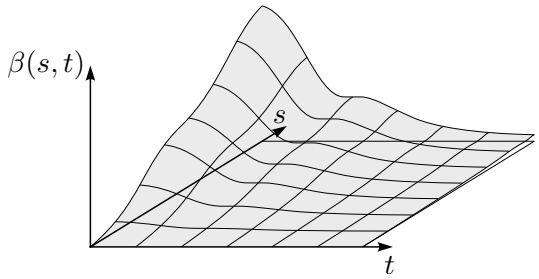


Abbildung 2.1: Beispiel für eine Funktion der Klasse $\mathcal{KL}$.

Es werden die folgenden Annahmen getroffen:

A1) Für jeden beschränkten Zustand $\mathbf{x}_k \in X_k$ und jede beschränkte Störung $\mathbf{w}_k \in W_k$ besitze (2.1a) eine eindeutige und beschränkte Lösung $\mathbf{x}_{k+1}$. Die Funktion $\mathbf{f}_k$ ist Lipschitz-stetig.

A2) Für jeden beschränkten Zustand $\mathbf{x}_k \in X_k$ und jede beschränkte Störung $\mathbf{v}_k \in V_k$ besitze (2.1b) eine eindeutige und beschränkte Lösung $\mathbf{y}_k$. Die Funktion $\mathbf{h}_k$ ist stetig.

A3) Die Störungen $\mathbf{w}_k$ und $\mathbf{v}_k$ mit $k \in \mathbb{N}_0$ sind beschränkt und es gilt

$$\lim_{k\to\infty} \mathbf{w}_k = \mathbf{0} \ , \qquad \lim_{k\to\infty} \mathbf{v}_k = \mathbf{0} \ . \tag{2.7}$$

Das folgende Lemma ist eine Konsequenz der Annahme A3. Der Beweis dazu findet sich in [2.13].

Lemma 2.1 (Beschränktheit der Summe von konvergenten Störungen)**.** *Ist die Annahme A3 erfüllt, dann existiert eine Funktion $\overline{\gamma}_b \in \mathcal{K}_\infty$, so dass*

$$\sum_{k=0}^{\infty} \overline{\gamma}_b(||\mathbf{w}_k||_2 + ||\mathbf{v}_k||_2) < \infty \ . \tag{2.8}$$

A4) Der Vektor $\bar{\mathbf{x}}_k$ soll nun entweder dem bekannten Zustand $\mathbf{x}_k$ oder einem früher gefundenen Schätzwert für $\mathbf{x}_k$ entsprechen. Fortan wird $\bar{\mathbf{x}}_k$ daher als *A-priori-Schätzung des Zustands* $\mathbf{x}_k$ bezeichnet. Einem MHE Beobachter mit dem Horizont $K-N, \dots, K$ steht der Wert $\bar{\mathbf{x}}_{K-N}$ zur Verfügung, im Allgemeinen jedoch nicht der tatsächliche Zustand $\mathbf{x}_{K-N}$.

A5) Die Kostenfunktionen B_k und b_k aus (2.4) sind stetig und für beliebige $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{w} \in \mathbb{R}^p$ und $\mathbf{v} \in \mathbb{R}^q$ durch

$$\underline{\gamma}_B(||\mathbf{x} - \bar{\mathbf{x}}_k||_2) \le B_k(\mathbf{x}) - B_k(\bar{\mathbf{x}}_k) \le \overline{\gamma}_B(||\mathbf{x} - \bar{\mathbf{x}}_k||_2) \qquad \forall\, k \in \mathbb{N}_0 \tag{2.9a}$$

$$\underline{\gamma}_b(||\mathbf{w}||_2 + ||\mathbf{v}||_2) \le b_k(\mathbf{w}, \mathbf{v}) \qquad \le \overline{\gamma}_b(||\mathbf{w}||_2 + ||\mathbf{v}||_2) \qquad \forall\, k \in \mathbb{N}_0 \tag{2.9b}$$

beschränkt, wobei $\underline{\gamma}_B$, $\overline{\gamma}_B$, $\underline{\gamma}_b$, $\overline{\gamma}_b \in \mathcal{K}_\infty$ und $\overline{\gamma}_b$ in Lemma 2.1 definiert wurde.

Aus (2.9a) folgt natürlich $B_k(\mathbf{x}) \ge B_k(\bar{\mathbf{x}}_k)$, wobei das Gleichheitszeichen nur im Fall $\mathbf{x} = \bar{\mathbf{x}}_k$ gilt. Bei bekanntem $B_k(\mathbf{x})$ kann $\bar{\mathbf{x}}_k$ durch Minimierung von $B_k(\,\cdot\,)$ berechnet werden. Ist die Annahme A3 erfüllt, so folgt aus (2.8) und (2.9b) die Ungleichung

$$\sum_{k=0}^{\infty} b_k(\mathbf{w}_k, \mathbf{v}_k) < \infty \ . \tag{2.10}$$

2.2 Stabilität von Zustandsschätzern

Bevor verschiedene Varianten von Zustandsbeobachtern basierend auf der Optimierungsaufgabe (2.6) besprochen werden, wird der Begriff der *Stabilität des Schätzfehlers* in Anlehnung an [2.2] definiert. Ist das System (2.1) keinen Störungen $\mathbf{w}_k$ und $\mathbf{v}_k$ ausgesetzt und ist der Anfangszustand exakt bekannt, d. h. $\bar{\mathbf{x}}_{K-N} = \mathbf{x}_{K-N}$, dann gilt für die optimale Lösung von (2.6) mit $N \le K$ natürlich

$$\hat{\mathbf{x}}_{K-N} = \mathbf{x}_{K-N} \tag{2.11a}$$

$$\hat{\mathbf{w}}_{K-N|K-1} = (\mathbf{0}, \dots, \mathbf{0}) \tag{2.11b}$$

$$\hat{\mathbf{v}}_k = \mathbf{y}_k - \mathbf{h}_k(\check{\mathbf{x}}_k(K-N, \hat{\mathbf{x}}_{K-N}, (\mathbf{0}, \dots, \mathbf{0})) = \mathbf{0} \qquad \forall\, k = K-N, \dots, K-1 \tag{2.11c}$$

und $\hat{J}_{K|N} = B_{K-N}(\mathbf{x}_{K-N})$. Abweichungen von diesem Idealzustand können verschiedene Ursachen haben:

- Ungenaue Informationen über den Anfangszustand, $\bar{\mathbf{x}}_{K-N} \neq \mathbf{x}_{K-N}$
- Nicht verschwindende Prozessstörungen, $\mathbf{w}_k \neq \mathbf{0}$
- Nicht verschwindende Messstörungen, $\mathbf{v}_k \neq \mathbf{0}$

Die folgenden Stabilitätsdefinitionen klären, wie der Schätzfehler auf derartige Abweichungen reagiert.

Definition 2.2 (Nominell global asymptotisch stabiler Schätzfehler (NGAS))**.** Eine Zustandsschätzung für das System (2.1) beginnend zum Zeitpunkt $k \geq 0$ ist *nominell global asymptotisch stabil*, wenn im Falle von verschwindenden Störungen, d. h. $\mathbf{w}_l = \mathbf{0}$ und $\mathbf{v}_l = \mathbf{0}$ für $\forall\, l = k, \ldots, K-1$ mit $k \leq K$, eine Funktion $\beta(\,\cdot\,,\cdot\,) \in \mathcal{KL}$ existiert, so dass für beliebige $\mathbf{x}_k$, $\bar{\mathbf{x}}_k \in \mathbb{R}^n$ gemäß Annahme A4 und für $\forall\, K \in \mathbb{N}_0$

$$||\underbrace{\check{\mathbf{x}}_K(k, \mathbf{x}_k, (\mathbf{0}, \ldots, \mathbf{0}))}_{=\,\mathbf{x}_K} - \check{\mathbf{x}}_K(k, \hat{\mathbf{x}}_k, \hat{\mathbf{w}}_{k|K-1})||_2 \leq \beta(||\mathbf{x}_k - \bar{\mathbf{x}}_k||_2, K-k) \tag{2.12}$$

erfüllt ist.

Definition 2.3 (Robust global asymptotisch stabiler Schätzfehler (RGAS))**.** Eine Zustandsschätzung für das System (2.1) beginnend zum Zeitpunkt $k \geq 0$ ist *robust global asymptotisch stabil*, wenn bei gegebenen Funktionen $\overline{\gamma}_B(\,\cdot\,)$, $\overline{\gamma}_b(\,\cdot\,) \in \mathcal{K}_\infty$ für jedes $\varepsilon > 0$ ein $\delta(\varepsilon) > 0$ existiert, so dass für beliebige $\mathbf{x}_k$, $\bar{\mathbf{x}}_k \in \mathbb{R}^n$ gemäß Annahme A4 und beliebige Störungen $\mathbf{w}_l$ und $\mathbf{v}_l$ für $\forall\, l = k, \ldots, K-1$, die der Annahme A3 genügen,

$$\begin{aligned} &\overline{\gamma}_B(||\mathbf{x}_k - \bar{\mathbf{x}}_k||_2) + \sum_{l=k}^{\infty} \overline{\gamma}_b(||\mathbf{w}_l||_2 + ||\mathbf{v}_l||_2) \leq \delta(\varepsilon) \quad \Rightarrow \\ &\qquad ||\underbrace{\check{\mathbf{x}}_K(k, \mathbf{x}_k, \mathbf{w}_{k|K-1})}_{=\,\mathbf{x}_K} - \check{\mathbf{x}}_K(k, \hat{\mathbf{x}}_k, \hat{\mathbf{w}}_{k|K-1})||_2 \leq \varepsilon \qquad \forall\, K \in \mathbb{N}_0 \end{aligned} \tag{2.13}$$

mit $K \geq k$ erfüllt ist und darüber hinaus

$$\lim_{K\to\infty} (\underbrace{\check{\mathbf{x}}_K(k, \mathbf{x}_k, \mathbf{w}_{k|K-1})}_{=\,\mathbf{x}_K} - \check{\mathbf{x}}_K(k, \hat{\mathbf{x}}_k, \hat{\mathbf{w}}_{k|K-1})) = \mathbf{0} \tag{2.14}$$

gilt.

Die Definition 2.3 verlangt also, dass der Beobachtungsfehler beschränkt bleibt und für $K \to \infty$ gegen $\mathbf{0}$ konvergiert. Jeder RGAS Schätzer ist auch NGAS.

2.3 Zustandsschätzung mit vollständiger Information

Wird die Zustandsschätzung mit der Horizontlänge $N = K$ durchgeführt, so spricht man von *Schätzung mit vollständiger Information*. Diese Beobachtervariante wird hier kurz studiert, da sie aus theoretischer Sicht sehr gute Stabilitäts- und Konvergenzeigenschaften besitzt. Ihr Nachteil ist, dass der Rechenaufwand im Allgemeinen zumindest proportional zu K und daher ohne Schranken wächst.

Im vorliegenden Abschnitt wird von einer Zustandsschätzung durch Lösen der Optimierungsaufgabe (2.6) mit $N = K$ ausgegangen. Sind das Modell (2.1) und die Beschränkungen (2.2) bekannt, so sind beim Entwurf des Schätzers lediglich die Kostenfunktionen B_0 und b_k für $\forall\, k = 0, \ldots, K-1$ zu wählen.

Es stellt sich nun die Frage für welche Klasse von Systemen der Fehler einer Zustandsschätzung mit $N = K$ überhaupt stabil sein kann. Ähnlich zur *Detektierbarkeit* bei linearen Systemen kann für nichtlineare zeitdiskrete Systeme die Eigenschaft der *inkrementellen Eingangs/Ausgangs-Zustands-Stabilität* (siehe [2.12, 2.14]) als notwendige Voraussetzung für einen stabilen Beobachtungsfehler verwendet werden.

Definition 2.4 (Inkrementelle Eingangs/Ausgangs-Zustands-Stabilität (IIOSS))**.** Das System (2.1) ist *inkrementell Eingangs/Ausgangs-Zustands-stabil*, wenn Funktionen $\beta(\,\cdot\,,\cdot\,) \in \mathcal{KL}$ und $\gamma_1(\,\cdot\,)$, $\gamma_2(\,\cdot\,) \in \mathcal{K}$ existieren, so dass für beliebige Anfangszustände $\mathbf{x}_a$, $\mathbf{x}_b$ zu einem Zeitpunkt $k \in \mathbb{N}_{\geq 0}$, beliebige Störfolgen $\mathbf{w}_{a,k|K-1}$, $\mathbf{w}_{b,k|K-1}$ und für $\forall\, k \in \mathbb{N}_0$ mit $k \leq K$

$$\begin{aligned} &||\check{\mathbf{x}}_K(k, \mathbf{x}_a, \mathbf{w}_{a,k|K-1}) - \check{\mathbf{x}}_K(k, \mathbf{x}_b, \mathbf{w}_{b,k|K-1})||_2 \\ &\quad \leq \beta(||\mathbf{x}_a - \mathbf{x}_b||_2, K-k) + \gamma_1\Big(\max_{l=k,\ldots,K-1}\{||\mathbf{w}_{a,l} - \mathbf{w}_{b,l}||_2\}\Big) \\ &\qquad + \gamma_2\Big(\max_{l=k,\ldots,K-1}\{||\mathbf{h}_l(\check{\mathbf{x}}_l(k, \mathbf{x}_a, \mathbf{w}_{a,k|l-1})) - \mathbf{h}_l(\check{\mathbf{x}}_l(k, \mathbf{x}_b, \mathbf{w}_{b,k|l-1}))||_2\}\Big) \end{aligned} \tag{2.15}$$

gilt.

In dieser Definition steht der Begriff *inkrementell* für den Vergleich zweier beliebiger Zustandsfolgen. Die *Eingangs/Ausgangs-Zustands-Stabilität* (IOSS) [2.14] ist eine schwächere Eigenschaft, betrachtet nur eine einzelne Zustandsfolge und gibt an, wann diese Zustandsfolge in den Ursprung konvergiert. Für Systeme mit $\mathbf{0} = \mathbf{f}_k(\mathbf{0}, \mathbf{0})$ und $\mathbf{0} = \mathbf{h}_k(\mathbf{0})$ impliziert die Eigenschaft IIOSS die Eigenschaft IOSS [2.14]. Die Umkehrung gilt nicht. Für lineare Systeme sind IIOSS und IOSS äquivalente Eigenschaften.

Lemma 2.2 (Konvergenz des Zustandes von IIOSS-Systemen)**.** *Wenn das System (2.1) die IIOSS-Eigenschaft gemäß Definition 2.4 besitzt sowie für* $\lim_{K\to\infty}(\mathbf{w}_{a,K} - \mathbf{w}_{b,K}) = \mathbf{0}$ *und beliebige Anfangszustände* $\mathbf{x}_a$, $\mathbf{x}_b$ *zum Zeitpunkt* $k \in \mathbb{N}_0$ $\lim_{K\to\infty}(\mathbf{h}_K(\check{\mathbf{x}}_K(k, \mathbf{x}_a, \mathbf{w}_{a,k|K-1})) - \mathbf{h}_K(\check{\mathbf{x}}_K(k, \mathbf{x}_b, \mathbf{w}_{b,k|K-1}))) = \mathbf{0}$ *mit* $K \geq k$ *gilt, dann folgt daraus*

$$\lim_{K\to\infty}(\check{\mathbf{x}}_K(k, \mathbf{x}_a, \mathbf{w}_{a,k|K-1}) - \check{\mathbf{x}}_K(k, \mathbf{x}_b, \mathbf{w}_{b,k|K-1})) = \mathbf{0}\ . \tag{2.16}$$

Aufgabe 2.1. Beweisen Sie Lemma 2.2.

Satz 2.1 (RGAS der Zustandsschätzung mit vollständiger Information)**.** *Für ein System (2.1), das die IIOSS-Eigenschaft besitzt, und Störungen, die der Annahme A3 genügen, führt der Zustandsschätzer (2.6) mit $N = K$ (Schätzung mit vollständiger Information) und Kostenfunktionen B_0 und b_k gemäß der Annahme A5 zu einem robust global asymptotisch stabilen Schätzfehler.*

Beweis. Es sei $(\tilde{\mathbf{x}}_0, \tilde{\mathbf{w}}_{0|\infty}) = (\mathbf{x}_0, \mathbf{w}_{0|\infty})$ eine zulässige, wenngleich nicht notwendigerweise optimale und bekannte Lösung der Optimierungsaufgabe (2.6) für $N = K \to \infty$. Aus Lemma 2.1 und der Annahme A5 erhält man

$$\begin{aligned} &J_{\infty|\infty}(\tilde{\mathbf{x}}_0,\tilde{\mathbf{w}}_{0|\infty}) \\ &\qquad = B_0(\mathbf{x}_0) + \sum_{k=0}^{\infty} b_k(\mathbf{w}_k, \mathbf{v}_k) - B_0(\bar{\mathbf{x}}_0) + B_0(\bar{\mathbf{x}}_0) \\ &\qquad \leq \overline{\gamma}_B(||\mathbf{x}_0 - \bar{\mathbf{x}}_0||_2) + \sum_{k=0}^{\infty} \overline{\gamma}_b(||\mathbf{w}_k||_2 + ||\mathbf{v}_k||_2) + B_0(\bar{\mathbf{x}}_0) = \overline{J} \ , \end{aligned} \tag{2.17}$$

wobei $\overline{J}$ eine Konstante ist. Daraus folgt direkt

$$\hat{J}_{K|K} \leq \overline{J} \qquad \forall\, K \in \mathbb{N}_0 \ . \tag{2.18}$$

Diese Beschränktheit des optimalen Gütefunktionswert $\hat{J}_{K|K}$ und die Annahme A5 implizieren die Existenz einer eindeutigen Lösung der Optimierungsaufgabe (2.6) mit $N = K$ für $\forall\, K \in \mathbb{N}_0$.

Es sei $(\hat{\mathbf{x}}_0, \hat{\mathbf{w}}_{0|K-1})$ das zu einem beliebigen Zeitpunkt K gefundene optimale Schätzergebnis. Sicher ist dieses Schätzergebnis auch eine zulässige aber im Allgemeinen nicht optimale Lösung der Optimierungsaufgabe (2.6) mit $N = K - 1$ zum Zeitpunkt $K - 1$. Es gilt daher

$$\begin{aligned} &J_{K-1|K-1}(\hat{\mathbf{x}}_0, \hat{\mathbf{w}}_{0|K-2}) \\ &\qquad = \hat{J}_{K|K} - b_{K-1}(\hat{\mathbf{w}}_{K-1}, \underbrace{\mathbf{y}_{K-1} - \mathbf{h}_{K-1}(\check{\mathbf{x}}_{K-1}(0, \hat{\mathbf{x}}_0, \hat{\mathbf{w}}_{0|K-2}))}_{= \hat{\mathbf{v}}_{K-1}}) \end{aligned} \tag{2.19}$$

und folglich

$$\hat{J}_{K|K} \geq \hat{J}_{K-1|K-1} + b_{K-1}(\hat{\mathbf{w}}_{K-1}, \hat{\mathbf{v}}_{K-1}) \ . \tag{2.20}$$

Gemäß (2.18) und (2.20) ist die Folge $(\hat{J}_{K|K})$ nach oben durch $\overline{J}$ beschränkt und nicht fallend. Aus (2.20), das für beliebige K gilt, folgt daher

$$\lim_{K\to\infty} b_K(\hat{\mathbf{w}}_K, \hat{\mathbf{v}}_K) = 0 \ , \tag{2.21}$$

was wegen (2.9b) und (2.5)

$$\lim_{K\to\infty} \hat{\mathbf{w}}_K = \mathbf{0} \tag{2.22a}$$

$$\lim_{K\to\infty} \hat{\mathbf{v}}_K = \lim_{K\to\infty} (\mathbf{y}_K - \mathbf{h}_K(\check{\mathbf{x}}_K(0, \hat{\mathbf{x}}_0, \hat{\mathbf{w}}_{0|K-1}))) = \mathbf{0} \tag{2.22b}$$

impliziert. Weil gemäß Annahme A3 auch

$$\lim_{K\to\infty} \mathbf{w}_K = \mathbf{0} \tag{2.23a}$$

$$\lim_{K\to\infty} \mathbf{v}_K = \lim_{K\to\infty} (\mathbf{y}_K - \mathbf{h}_K(\underbrace{\check{\mathbf{x}}_K(0, \mathbf{x}_0, \mathbf{w}_{0|K-1})}_{=\mathbf{x}_K})) = \mathbf{0} \tag{2.23b}$$

gilt, erhält man

$$\lim_{K\to\infty} (\hat{\mathbf{w}}_K - \mathbf{w}_K) = \mathbf{0} \tag{2.24a}$$

$$\lim_{K\to\infty} (\mathbf{h}_K(\check{\mathbf{x}}_K(0, \hat{\mathbf{x}}_0, \hat{\mathbf{w}}_{0|K-1})) - \mathbf{h}_K(\mathbf{x}_K)) = \mathbf{0}\ . \tag{2.24b}$$

Da das System die IIOSS-Eigenschaft besitzt, ist die Existenz von Funktionen $\beta(\cdot\,,\cdot) \in \mathcal{KL}$ und $\gamma_1(\cdot)$, $\gamma_2(\cdot) \in \mathcal{K}$ gesichert, so dass

$$\begin{aligned} &||\check{\mathbf{x}}_K(k, \hat{\mathbf{x}}_k, \hat{\mathbf{w}}_{k|K-1}) - \underbrace{\check{\mathbf{x}}_K(k, \mathbf{x}_k, \mathbf{w}_{k|K-1})}_{=\mathbf{x}_K}||_2 \\ &\quad \le \beta(||\hat{\mathbf{x}}_k - \mathbf{x}_k||_2, K-k) + \gamma_1\Big(\max_{l=k,\dots,K-1}\{||\hat{\mathbf{w}}_l - \mathbf{w}_l||_2\}\Big) \\ &\qquad + \gamma_2\Big(\max_{l=k,\dots,K-1}\{||\mathbf{h}_l(\check{\mathbf{x}}_l(k, \hat{\mathbf{x}}_k, \hat{\mathbf{w}}_{k|l-1})) - \mathbf{h}_l(\underbrace{\check{\mathbf{x}}_l(k, \mathbf{x}_k, \mathbf{w}_{k|l-1})}_{=\mathbf{x}_l})||_2\}\Big) \end{aligned} \tag{2.25}$$

für $\forall\, k, K \in \mathbb{N}_0$ mit $k \le K$ gilt. Analog zu Lemma 2.2 folgt aus (2.24) und (2.25)

$$\lim_{K\to\infty} (\check{\mathbf{x}}_K(k, \hat{\mathbf{x}}_k, \hat{\mathbf{w}}_{k|K-1}) - \underbrace{\check{\mathbf{x}}_K(k, \mathbf{x}_k, \mathbf{w}_{k|K-1})}_{=\mathbf{x}_K}) = \mathbf{0} \tag{2.26}$$

womit die in (2.14) geforderte Konvergenz des Schätzfehlers gezeigt ist.

Wählt man nun ein festes $\delta \ge 0$ und fordert $\overline{J} = \delta$ für $\overline{J}$ aus (2.17), so folgt unter Berücksichtigung von (2.9a) und $B_0(\bar{\mathbf{x}}_0) \ge 0$ aus (2.17) und (2.18), dass

$$\overline{\gamma}_B(||\mathbf{x}_0 - \bar{\mathbf{x}}_0||_2) \le \delta\ , \qquad \underline{\gamma}_B(||\hat{\mathbf{x}}_0 - \bar{\mathbf{x}}_0||_2) \le \delta\ , \tag{2.27}$$

wobei hier $\hat{\mathbf{x}}_0$ der zu einem beliebigen Zeitpunkt K ermittelte optimale Schätzwert des Anfangszustands $\mathbf{x}_0$ ist. Umformung von (2.27) führt mithilfe der Dreiecksungleichung auf

$$||\hat{\mathbf{x}}_0 - \mathbf{x}_0||_2 \le \overline{\gamma}_B^{-1}(\delta) + \underline{\gamma}_B^{-1}(\delta)\ , \tag{2.28}$$

womit die Beschränktheit des Anfangsschätzfehlers für Schätzungen zu beliebigen Zeitpunkten K gesichert ist. Mit der Forderung $\overline{J} = \delta$ folgt unter Berücksichtigung von (2.9b) und $B_0(\bar{\mathbf{x}}_0) \geq 0$ aus (2.17) und (2.18) ferner, dass

$$\overline{\gamma}_b(||\mathbf{w}_l||_2) \leq \delta \ , \qquad \underline{\gamma}_b(||\hat{\mathbf{w}}_l||_2) \leq \delta \ , \qquad \forall\, l \in \mathbb{N}_0 \tag{2.29a}$$

$$\overline{\gamma}_b(||\mathbf{v}_l||_2) \leq \delta \ , \qquad \underline{\gamma}_b(||\hat{\mathbf{v}}_l||_2) \leq \delta \ , \qquad \forall\, l \in \mathbb{N}_0 \ , \tag{2.29b}$$

wobei hier $\hat{\mathbf{w}}_l$ und $\hat{\mathbf{v}}_l$ die zu einem beliebigen Zeitpunkt $K > l$ berechneten optimalen Schätzwerte für die Störungen $\mathbf{w}_l$ und $\mathbf{v}_l$ sind. Unter Verwendung von (2.5) und der Dreiecksungleichung lässt sich (2.29) in die Form

$$||\hat{\mathbf{w}}_l - \mathbf{w}_l||_2 \leq \overline{\gamma}_b^{-1}(\delta) + \underline{\gamma}_b^{-1}(\delta) \qquad \forall\, l \in \mathbb{N}_0 \tag{2.30a}$$

$$\begin{aligned} ||\hat{\mathbf{v}}_l - \mathbf{v}_l||_2 &= ||\mathbf{h}_l(\check{\mathbf{x}}_l(0, \hat{\mathbf{x}}_0, \hat{\mathbf{w}}_{0|l-1})) - \mathbf{h}_l(\mathbf{x}_l)||_2 \\ &\leq \overline{\gamma}_b^{-1}(\delta) + \underline{\gamma}_b^{-1}(\delta) \qquad \forall\, l \in \mathbb{N}_0 \end{aligned} \tag{2.30b}$$

umschreiben. Die Abschätzungen (2.28) und (2.30) können nun gemeinsam mit $k = 0$ in (2.25) eingesetzt werden und man erhält unter Berücksichtigung von $\beta(\,\cdot\,, l) \leq \beta(\,\cdot\,, 0)\ \forall\, l \geq 0$

$$\begin{aligned} ||\check{\mathbf{x}}_K(0, \hat{\mathbf{x}}_0, \hat{\mathbf{w}}_{0|K-1}) - \mathbf{x}_K||_2 \leq\ & \beta(\overline{\gamma}_B^{-1}(\delta) + \underline{\gamma}_B^{-1}(\delta), 0) + \gamma_1(\overline{\gamma}_b^{-1}(\delta) + \underline{\gamma}_b^{-1}(\delta)) \\ & + \gamma_2(\overline{\gamma}_b^{-1}(\delta) + \underline{\gamma}_b^{-1}(\delta)) \qquad \forall\, K \in \mathbb{N}_0 \ , \end{aligned} \tag{2.31}$$

wobei K wieder der Zeitpunkt der Schätzung ist. Die gesamte rechte Seite von (2.31) ist eine $\mathcal{K}$-Funktion in δ und daher nach δ auflösbar. Es existiert also zu jedem gegebenen $\varepsilon \geq 0$ ein $\delta(\varepsilon) \geq 0$ wobei $\delta(0) = 0$, so dass

$$||\check{\mathbf{x}}_K(0, \hat{\mathbf{x}}_0, \hat{\mathbf{w}}_{0|K-1}) - \mathbf{x}_K||_2 \leq \varepsilon \qquad \forall\, K \in \mathbb{N}_0 \ . \tag{2.32}$$

Damit ist auch die Existenz einer Beziehung $\delta(\varepsilon)$, die die Erfüllung der Implikation (2.13) und somit die Beschränktheit des Schätzfehlers sichert, gezeigt. □

2.4 Zustandsschätzung auf bewegtem Horizont

Erfolgt zum Zeitindex K eine MHE Zustandsschätzung durch Lösen der Optimierungsaufgabe (2.6), so werden Messwerte $\mathbf{y}_{K-N|K-1}$ aus dem aktuellen Zeithorizont $K - N$, …, K verwendet. Dieser umfasst $N \in \mathbb{N}_0$ Abtastintervalle, wobei N eine feste finite Zahl ist. Im vorliegenden Abschnitt wird nur noch der Fall $K > N$ explizit betrachtet. Im Fall $K \leq N$, also wenn der Beobachter neu eingeschaltet wird, verwendet man einfach die Zustandsschätzung mit vollständiger Information gemäß Abschnitt 2.3.

Sind das Modell (2.1) und die Beschränkungen (2.2) bekannt, so sind beim Entwurf des MHE Beobachters lediglich die Horizontlänge N sowie die Kostenfunktionen B_{K-N} und b_k für $\forall\, k = K - N, \ldots, K - 1$ zu wählen. Diese Wahl ist natürlich entscheidend für

die Stabilität und Konvergenz des Schätzfehlers. Ein Vergleich der Gütefunktion

$$J_{K|K}(\tilde{\mathbf{x}}_0, \tilde{\mathbf{w}}_{0|K-1}) = B_0(\tilde{\mathbf{x}}_0) + \sum_{k=0}^{K-1} b_k(\tilde{\mathbf{w}}_k, \tilde{\mathbf{v}}_k) \tag{2.33}$$

aus der Zustandsschätzung mit vollständiger Information mit der Gütefunktion

$$J_{K|N}(\tilde{\mathbf{x}}_{K-N}, \tilde{\mathbf{w}}_{K-N|K-1}) = B_{K-N}(\tilde{\mathbf{x}}_{K-N}) + \sum_{k=K-N}^{K-1} b_k(\tilde{\mathbf{w}}_k, \tilde{\mathbf{v}}_k) \tag{2.34}$$

für MHE zeigt, dass ein wesentlicher Punkt des Entwurfes die Einbeziehung vergangener Informationen und Schätzergebnisse in den Anfangskostenterm $B_k(\,\cdot\,)$ (meist mit $k = K - N$) ist. $B_k(\tilde{\mathbf{x}}_k)$ soll daher Abweichungen zwischen $\tilde{\mathbf{x}}_k$ und der A-priori-Schätzung $\bar{\mathbf{x}}_k$ bestrafen. Weist der Schätzwert $\bar{\mathbf{x}}_k$ eine hohe (geringe) *statistische Zuverlässigkeit* auf, so sollte die Abweichung $\tilde{\mathbf{x}}_k - \bar{\mathbf{x}}_k$ durch $B_k(\,\cdot\,)$ entsprechend hoch (gering) bestraft werden [2.8].

2.4.1 Anfangskostenterm für vollständige Information

Eine naheliegende Frage ist nun, wie der Anfangskostenterm $B_k(\,\cdot\,)$ zu wählen ist, damit der MHE Beobachter und die Zustandsschätzung mit vollständiger Information (siehe Abschnitt 2.3) die gleichen Ergebnisse liefern. Um diese Frage zu beantworten (siehe [2.8]), wird zunächst (2.33) in die Form

$$J_{K|K}(\tilde{\mathbf{x}}_0, \tilde{\mathbf{w}}_{0|K-1}) = \underbrace{B_0(\tilde{\mathbf{x}}_0) + \sum_{k=0}^{K-N-1} b_k(\tilde{\mathbf{w}}_k, \tilde{\mathbf{v}}_k)}_{\text{Anfangskostenterm}} + \sum_{k=K-N}^{K-1} b_k(\tilde{\mathbf{w}}_k, \tilde{\mathbf{v}}_k) \tag{2.35}$$

umgeschrieben. Offensichtlich hängt die letzte Summe in (2.35) nur von $\tilde{\mathbf{x}}_{K-N}$, $\tilde{\mathbf{w}}_{K-N|K-1}$ und den gegebenen Messwerten $\mathbf{y}_{K-N|K-1}$ ab. Die Unabhängigkeit des Systemverhaltens im Horizont $K-N, \ldots, K$ von Zuständen vor dem Zeitpunkt $K-N$ wird auch als *Markov-Eigenschaft* des zugrunde liegenden stochastischen Prozesses bezeichnet. Es kann daher der als *Anfangskostenterm* bezeichnete Term in (2.35) in ein separates Optimierungsproblem als Funktion von $\tilde{\mathbf{x}}_{K-N}$ ausgelagert werden. Dieses lautet für einen allgemeinen Zeitpunkt $k \in \mathbb{N}_0$

$$\begin{aligned}
\hat{J}^B_{k|k}(\mathbf{x}) = \min_{(\tilde{\mathbf{x}}_0, \tilde{\mathbf{w}}_{0|k-1})} \quad & J_{k|k}(\tilde{\mathbf{x}}_0, \tilde{\mathbf{w}}_{0|k-1}) && && \text{(2.36a)}\\
\text{u.B.v.} \quad & \tilde{\mathbf{x}}_{l+1} = \mathbf{f}_l(\tilde{\mathbf{x}}_l, \tilde{\mathbf{w}}_l) && \forall\, l = 0, \ldots, k-1 && \text{(2.36b)}\\
& \tilde{\mathbf{v}}_l = \mathbf{y}_l - \mathbf{h}_l(\tilde{\mathbf{x}}_l) && \forall\, l = 0, \ldots, k-1 && \text{(2.36c)}\\
& \tilde{\mathbf{x}}_l \in X_l && \forall\, l = 0, \ldots, k-1 && \text{(2.36d)}\\
& \tilde{\mathbf{w}}_l \in W_l\,, \quad \tilde{\mathbf{v}}_l \in V_l\,, && \forall\, l = 0, \ldots, k-1 && \text{(2.36e)}\\
& \tilde{\mathbf{x}}_k = \mathbf{x}\,. && && \text{(2.36f)}
\end{aligned}$$

Die Funktion $\hat{J}^B_{k|k}(\mathbf{x})$ liefert die minimalen Kosten, die entstehen wenn die Zustandsfolge zum Zeitpunkt k am Punkt $\mathbf{x} \in X_k$ ankommt. Die Funktion wird daher auch als

Ankunftskostenterm (Englisch: *arrival cost*) bezeichnet (vgl. [2.2]). Sie sorgt für die Berücksichtigung von Informationen, die vor dem Zeitpunkt k gesammelt wurden (z. B. die Messwerte $\mathbf{y}_{0|k-1}$) und daher nicht explizit in einer MHE Optimierungsaufgabe auftreten, deren Horizont zum Zeitpunkt k beginnt. Der einzige Unterschied zwischen (2.36) und der ursprünglichen Optimierungsaufgabe (2.6) ist die Endbedingung (2.36f). Offensichtlich gilt aufgrund von dieser Endbedingung

$$\hat{J}^B_{k|k}(\mathbf{x}) \geq \hat{J}_{k|k} \qquad \forall\, \mathbf{x} \in X_k, k \in \mathbb{N}_0 \ . \tag{2.37}$$

Lemma 2.3 (Bedingung für Äquivalenz zwischen MHE und Zustandsschätzung mit vollständiger Information)**.** *Wenn die MHE Optimierungsaufgabe* (2.6) *für* $K \geq N$ *unter Verwendung des speziellen Anfangskostenterms*

$$B_{K-N}(\tilde{\mathbf{x}}_{K-N}) = \hat{J}^B_{K-N|K-N}(\tilde{\mathbf{x}}_{K-N}) \tag{2.38}$$

mit $\hat{J}^B_{K-N|K-N}(\,\cdot\,)$ *gemäß* (2.36) *formuliert und gelöst wird, so stimmt ihre Lösung im Intervall* $K-N, \dots, K$ *mit jener der Zustandsschätzung mit vollständiger Information gemäß Abschnitt 2.3 exakt überein.*

Der Beweis dieses Lemmas folgt konstruktiv aus dem Vorangegangen.

Es gilt auch hier, dass $\bar{\mathbf{x}}_{K-N}$ durch Minimierung von $B_{K-N}(\,\cdot\,)$ berechnet werden kann. Das entspricht einer Lösung der Optimierungsaufgabe (2.36) ohne die Endbedingung (2.36f), was wiederum auf die ursprüngliche Optimierungsaufgabe (2.6) und $\check{\mathbf{x}}_{K-N}(0, \hat{\mathbf{x}}_0, \hat{\mathbf{w}}_{0|K-N-1})$ führt.

Lemma 2.3 zeigt, dass eine MHE Formulierung gefunden werden kann, die zur Zustandsschätzung mit vollständiger Information äquivalent ist und daher die gleichen guten Stabilitäts- und Konvergenzeigenschaften besitzt. Im Allgemeinen ist mit dieser MHE Formulierung aber auch der Rechenaufwand äquivalent zu jenem der Zustandsschätzung mit vollständiger Information. Der hohe Rechenaufwand ist dem Anfangskostenterm (2.38) geschuldet und limitiert den praktischen Nutzen dieser MHE Formulierung. Im Folgenden werden daher die Stabilitäts- und Konvergenzeigenschaften von MHE Varianten mit alternativen Anfangskostentermen $B_k(\,\cdot\,)$ untersucht.

2.4.2 Kein Anfangskostenterm

In der MHE Formulierung kann auch einfach auf den Anfangskostenterm B_{K-N} verzichtet werden, d. h. $B_k(\,\cdot\,) = 0\ \forall\, k \in \mathbb{N}_0$ und die Optimierungsaufgabe lautet

$$\hat{J}^0_{K|N} = \min_{\substack{(\tilde{\mathbf{x}}_{K-N},\\ \tilde{\mathbf{w}}_{K-N|K-1})}} J_{K|N}(\tilde{\mathbf{x}}_{K-N}, \tilde{\mathbf{w}}_{K-N|K-1}) = \sum_{k=K-N}^{K-1} b_k(\tilde{\mathbf{w}}_k, \tilde{\mathbf{v}}_k) \tag{2.39a}$$

$$\text{u.B.v.} \quad \tilde{\mathbf{x}}_{k+1} = \mathbf{f}_k(\tilde{\mathbf{x}}_k, \tilde{\mathbf{w}}_k) \qquad \forall\, k = K-N, \dots, K-1 \tag{2.39b}$$

$$\tilde{\mathbf{v}}_k = \mathbf{y}_k - \mathbf{h}_k(\tilde{\mathbf{x}}_k) \qquad \forall\, k = K-N, \dots, K-1 \tag{2.39c}$$

$$\tilde{\mathbf{x}}_k \in X_k \qquad \forall\, k = K-N, \dots, K \tag{2.39d}$$

$$\tilde{\mathbf{w}}_k \in W_k\ , \quad \tilde{\mathbf{v}}_k \in V_k\ , \qquad \forall\, k = K-N, \dots, K-1\ . \tag{2.39e}$$

Allerdings reicht dann die IIOSS-Eigenschaft des Systems (2.1) nicht mehr aus, um die Existenz eines optimalen MHE Schätzwertes und die (asymptotische) Stabilität des Schätzfehlers zu garantieren. Wegen der Wahl $B_k(\,\cdot\,) = 0$ ist die Bedingung (2.9a) aus Annahme A5 in diesem Abschnitt nicht erfüllbar. Aus diesem Grund ist die Existenz einer beschränkten optimalen Lösung der MHE Optimierungsaufgabe in diesem Fall nicht mehr gesichert. Gemäß [2.2] ist die nachfolgend definierte Beobachtbarkeitseigenschaft des Systems (2.1), welche stärker ist als die IIOSS-Eigenschaft, für die Existenz und Eindeutigkeit der Lösung von (2.39) hinreichend.

Definition 2.5 (Beobachtbarkeit des Anfangszustandes)**.** Der *Anfangszustand* des Systems (2.1) ist *beobachtbar*, wenn eine Zahl $\underline{N} \in \mathbb{N}$ und Funktionen $\gamma_1(\,\cdot\,), \gamma_2(\,\cdot\,) \in \mathcal{K}$ existieren, so dass für beliebige Anfangszustände $\mathbf{x}_a$, $\mathbf{x}_b$ zum Zeitpunkt $K-N$, beliebige Störfolgen $\mathbf{w}_{a,K-N|K-1}$, $\mathbf{w}_{b,K-N|K-1}$ und $\forall\, K, N \in \mathbb{N}_0$ mit $K \geq N \geq \underline{N}$

$$\begin{aligned} ||\mathbf{x}_a - \mathbf{x}_b||_2 \leq\ & \gamma_1\Big(\max_{k=K-N,\dots,K-1} \{||\mathbf{w}_{a,k} - \mathbf{w}_{b,k}||_2\} \Big) \\ & + \gamma_2\Big(\max_{k=K-N,\dots,K-1} \{||\mathbf{h}_k(\check{\mathbf{x}}_k(K-N, \mathbf{x}_a, \mathbf{w}_{a,K-N|k-1})) \\ & \qquad - \mathbf{h}_k(\check{\mathbf{x}}_k(K-N, \mathbf{x}_b, \mathbf{w}_{b,K-N|k-1}))||_2\} \Big) \end{aligned} \tag{2.40}$$

gilt.

Definition 2.6 (Beobachtbarkeit des Endzustandes)**.** Der *Endzustand* des Systems (2.1) ist *beobachtbar*, wenn eine Zahl $\underline{N} \in \mathbb{N}$ und Funktionen $\bar{\gamma}_1(\,\cdot\,), \bar{\gamma}_2(\,\cdot\,) \in \mathcal{K}$ existieren, so dass für beliebige Anfangszustände $\mathbf{x}_a$, $\mathbf{x}_b$ zum Zeitpunkt $K-N$, beliebige Störfolgen $\mathbf{w}_{a,K-N|K-1}$, $\mathbf{w}_{b,K-N|K-1}$ und $\forall\, K, N \in \mathbb{N}_0$ mit $K \geq N \geq \underline{N}$

$$\begin{aligned} &||\check{\mathbf{x}}_K(K-N, \mathbf{x}_a, \mathbf{w}_{a,K-N|K-1}) - \check{\mathbf{x}}_K(K-N, \mathbf{x}_b, \mathbf{w}_{b,K-N|K-1})||_2 \\ &\quad \leq \bar{\gamma}_1\Big(\max_{k=K-N,\dots,K-1} \{||\mathbf{w}_{a,k} - \mathbf{w}_{b,k}||_2\} \Big) \\ &\qquad + \bar{\gamma}_2\Big(\max_{k=K-N,\dots,K-1} \{||\mathbf{h}_k(\check{\mathbf{x}}_k(K-N, \mathbf{x}_a, \mathbf{w}_{a,K-N|k-1})) \\ &\qquad\qquad - \mathbf{h}_k(\check{\mathbf{x}}_k(K-N, \mathbf{x}_b, \mathbf{w}_{b,K-N|k-1}))||_2\} \Big) \end{aligned} \tag{2.41}$$

gilt.

Bemerkung 2.1. Die Eigenschaft *Beobachbarkeit des Anfangszustandes* sichert also bei gleichen Störungen $\mathbf{w}_{a,K-N|K-1} = \mathbf{w}_{b,K-N|K-1}$ und $\mathbf{v}_{a,K-N|K-1} = \mathbf{v}_{b,K-N|K-1}$ und gleichen Messwerten $\mathbf{y}_{a,K-N|K-1} = \mathbf{y}_{b,K-N|K-1}$ für $\forall\, N \geq \underline{N}$ die Äquivalenz der Anfangszustände, d. h. $\mathbf{x}_a = \mathbf{x}_b$. Die Eigenschaft *Beobachbarkeit des Endzustandes* hingegen sichert bei gleichen Störungen $\mathbf{w}_{a,K-N|K-1} = \mathbf{w}_{b,K-N|K-1}$ und $\mathbf{v}_{a,K-N|K-1} = \mathbf{v}_{b,K-N|K-1}$ und gleichen Messwerten $\mathbf{y}_{a,K-N|K-1} = \mathbf{y}_{b,K-N|K-1}$ für $\forall\, N \geq \underline{N}$ für beliebige Anfangszustände $\mathbf{x}_a$, $\mathbf{x}_b$ die Äquivalenz der Endzustände, d. h. $\check{\mathbf{x}}_K(K-N, \mathbf{x}_a, \mathbf{w}_{a,K-N|K-1}) = \check{\mathbf{x}}_K(K-N, \mathbf{x}_b, \mathbf{w}_{b,K-N|K-1})$. Wegen der Annahme A1 über die Lipschitz-Stetigkeit der Funktion $\mathbf{f}_k$ impliziert die Beobachbarkeit des Anfangszustandes die Beobachbarkeit des Endzustandes. Ein Vergleich der Definitionen 2.4 und 2.6 zeigt ferner, dass (abgesehen von der Forderung $N \geq \underline{N}$) die Beobachbarkeit des Endzustandes die Eigenschaft IIOSS impliziert.

Satz 2.2 (RGAS der MHE Zustandsschätzung ohne Anfangskostenterm)**.** *Für ein System* (2.1)*, dessen Anfangszustand beobachtbar im Sinne der Definition 2.5 ist, mit Störungen, die der Annahme A3 genügen, führt der MHE Zustandsschätzer* (2.39) *(Schätzung ohne Anfangskostenterm) mit festem N und b_k gemäß der Annahme A5 zu einem robust global asymptotisch stabilen Schätzfehler.*

Beweis. Der Beweis erfolgt ähnlich zum Beweis des Satzes 2.1. Für die Optimierungsaufgabe (2.39) sei nun $(\tilde{\mathbf{x}}_{K-N}, \tilde{\mathbf{w}}_{K-N|K-1}) = (\mathbf{x}_{K-N}, \mathbf{w}_{K-N|K-1})$ eine zulässige, wenngleich nicht notwendigerweise optimale und bekannte Lösung. Folglich gilt mit (2.9b) und dem in (2.17) definierten Wert $\overline{J}$

$$\begin{aligned}\sum_{k=K-N}^{K-1} \underline{\gamma}_b(||\hat{\mathbf{w}}_k||_2 + ||\hat{\mathbf{v}}_k||_2) \leq \hat{J}^0_{K|N} \leq J_{K|N}(\tilde{\mathbf{x}}_{K-N}, \tilde{\mathbf{w}}_{K-N|K-1}) \\ = \sum_{k=K-N}^{K-1} b_k(\mathbf{w}_k, \mathbf{v}_k) \leq \sum_{k=K-N}^{K-1} \overline{\gamma}_b(||\mathbf{w}_k||_2 + ||\mathbf{v}_k||_2) \leq \overline{J}\ .\end{aligned} \tag{2.42}$$

Um aus dieser Beschränktheit des optimalen Gütefunktionswertes $\hat{J}^0_{K|N}$ auf die Existenz einer Lösung $(\hat{\mathbf{x}}_{K-N}, \hat{\mathbf{w}}_{K-N|K-1})$ der Optimierungsaufgabe (2.39) mit finitem $\hat{\mathbf{x}}_{K-N}$ schließen zu können, wird die Eigenschaft Beobachtbarkeit des Anfangszustandes benötigt. Unter Verwendung der Annahme A3 folgen aus (2.42) die Konvergenzresultate

$$\lim_{K\to\infty} \hat{J}^0_{K|N} = 0 \tag{2.43a}$$

$$\lim_{K\to\infty} \hat{\mathbf{w}}_{K-k} = \mathbf{0} \qquad \forall\, k = 1, \ldots, N \tag{2.43b}$$

$$\begin{aligned}\lim_{K\to\infty} \hat{\mathbf{v}}_{K-k} &= \lim_{K\to\infty} (\mathbf{y}_{K-k} - \mathbf{h}_{K-k}(\check{\mathbf{x}}_{K-k}(K-N, \hat{\mathbf{x}}_{K-N}, \hat{\mathbf{w}}_{K-N|K-k-1}))) \\ &= \mathbf{0} \qquad \forall\, k = 1, \ldots, N\ ,\end{aligned} \tag{2.43c}$$

wobei in (2.43c) wieder (2.5) verwendet wurde. Weil gemäß der Annahme A3 für

einen finiten Wert N auch

$$\lim_{K\to\infty} \mathbf{w}_{K-k} = \mathbf{0} \qquad \forall\, k = 1,\dots,N \tag{2.44a}$$

$$\lim_{K\to\infty} \mathbf{v}_{K-k} = \lim_{K\to\infty} (\mathbf{y}_{K-k} - \mathbf{h}_{K-k}(\mathbf{x}_{K-k})) = \mathbf{0} \qquad \forall\, k = 1,\dots,N \tag{2.44b}$$

gilt, erhält man

$$\lim_{K\to\infty} (\hat{\mathbf{w}}_{K-k} - \mathbf{w}_{K-k}) = \mathbf{0} \qquad \forall\, k = 1,\dots,N \tag{2.45a}$$

$$\begin{aligned}\lim_{K\to\infty} (\mathbf{h}_{K-k}(\check{\mathbf{x}}_{K-k}(K-N, \hat{\mathbf{x}}_{K-N}, \hat{\mathbf{w}}_{K-N|K-k-1})) - \mathbf{h}_{K-k}(\mathbf{x}_{K-k}))\\ = \mathbf{0} \qquad \forall\, k = 1,\dots,N\ .\end{aligned} \tag{2.45b}$$

Da das System die Eigenschaft Beobachtbarkeit des Anfangszustandes besitzt und diese die Beobachtbarkeit des Endzustandes impliziert, ist die Existenz von Funktionen $\bar{\gamma}_1(\,\cdot\,)$, $\bar{\gamma}_2(\,\cdot\,) \in \mathcal{K}$ und einer Zahl $\underline{N} \in \mathbb{N}$ gesichert, so dass

$$\begin{aligned}&||\check{\mathbf{x}}_K(K-N, \hat{\mathbf{x}}_{K-N}, \hat{\mathbf{w}}_{K-N|K-1}) - \mathbf{x}_K||_2\\ &\le \bar{\gamma}_1\Big(\max_{k=K-N,\dots,K-1}\{||\hat{\mathbf{w}}_k - \mathbf{w}_k||_2\}\Big)\\ &\quad + \bar{\gamma}_2\Big(\max_{k=K-N,\dots,K-1}\{||\mathbf{h}_k(\check{\mathbf{x}}_k(K-N, \hat{\mathbf{x}}_{K-N}, \hat{\mathbf{w}}_{K-N|k-1})) - \mathbf{h}_k(\mathbf{x}_k)||_2\}\Big)\end{aligned} \tag{2.46}$$

für $\forall\, K, N \in \mathbb{N}_0$ mit $K \ge N \ge \underline{N}$ und finitem N gilt. Aus (2.45) und (2.46) folgt für finites $N \ge \underline{N}$

$$\lim_{K\to\infty} (\check{\mathbf{x}}_K(K-N, \hat{\mathbf{x}}_{K-N}, \hat{\mathbf{w}}_{K-N|K-1}) - \mathbf{x}_K) = \mathbf{0}\ , \tag{2.47}$$

womit die in (2.14) geforderte Konvergenz des Schätzfehlers gezeigt ist.

Wählt man nun ein festes $\delta \ge 0$ und fordert $\overline{J} = \delta$ für $\overline{J}$ aus (2.17) bzw. (2.42), so folgt in einer zum Beweis des Satzes 2.1 analogen Weise, dass

$$||\hat{\mathbf{w}}_k - \mathbf{w}_k||_2 \le \overline{\gamma}_b^{-1}(\delta) + \underline{\gamma}_b^{-1}(\delta) \qquad \forall\, k = K-N,\dots,K-1 \tag{2.48a}$$

$$\begin{aligned}||\mathbf{h}_k(\check{\mathbf{x}}_k(K-N, \hat{\mathbf{x}}_{K-N}, \hat{\mathbf{w}}_{K-N|k-1})) - \mathbf{h}_k(\mathbf{x}_k)||_2 \le \overline{\gamma}_b^{-1}(\delta) + \underline{\gamma}_b^{-1}(\delta)\\ \forall\, k = K-N,\dots,K-1\end{aligned} \tag{2.48b}$$

gilt. Einsetzen der Abschätzung (2.48) in (2.46) liefert

$$\begin{aligned}&||\check{\mathbf{x}}_K(K-N, \hat{\mathbf{x}}_{K-N}, \hat{\mathbf{w}}_{K-N|K-1}) - \mathbf{x}_K||_2\\ &\qquad \le \bar{\gamma}_1(\overline{\gamma}_b^{-1}(\delta) + \underline{\gamma}_b^{-1}(\delta)) + \bar{\gamma}_2(\overline{\gamma}_b^{-1}(\delta) + \underline{\gamma}_b^{-1}(\delta))\end{aligned} \tag{2.49}$$

für $\forall\, K, N \in \mathbb{N}_0$ mit $K \geq N \geq \underline{N}$ und finitem N. Die gesamte rechte Seite von (2.49) ist eine $\mathcal{K}$-Funktion in δ und daher nach δ auflösbar. Es existiert also zu jedem gegebenen $\varepsilon \geq 0$ ein $\delta(\varepsilon) \geq 0$ wobei $\delta(0) = 0$, so dass

$$||\check{\mathbf{x}}_K(K-N, \hat{\mathbf{x}}_{K-N}, \hat{\mathbf{w}}_{K-N|K-1}) - \mathbf{x}_K||_2 \leq \varepsilon \tag{2.50}$$

für $\forall\, K, N \in \mathbb{N}_0$ mit $K \geq N \geq \underline{N}$ und finitem N. Damit ist auch die Existenz einer Beziehung $\delta(\varepsilon)$, die die Erfüllung der Implikation (2.13) und somit die Beschränktheit des Schätzfehlers sichert, gezeigt. □

Nachteilig beim MHE Zustandsbeobachter ohne Anfangskostenterm ist, dass die Strecke beobachtbar gemäß Definition 2.5 sein muss und dass dies eine möglicherweise große Mindesthorizontlänge $\underline{N}$ erfordert. Nachfolgend werden daher Bedingungen für einen von Null verschiedenen Anfangskostenterm formuliert, so dass diese Beobachtbarkeitsbedingung nicht mehr erforderlich ist.

2.4.3 Approximation der Ankunftskosten

Der vorliegende Abschnitt beschreibt eine MHE Formulierung mit einem Anfangskostenterm $B_k(\,\cdot\,)$, der selbst aus der Betrachtung eines finiten mitbewegten Horizonts hervorgeht. Es wird versucht, die guten Stabilitäts- und Konvergenzeigenschaften der Zustandsschätzung mit vollständiger Information (siehe die Abschnitte 2.3 und 2.4.1) zu erreichen und trotzdem den Rechenaufwand in einem vertretbaren Bereich zu halten.

Es wird zunächst in Anlehnung an den Ankunftskostenterm (2.36) für vollständige Information, der sogenannte *MHE Ankunftskostenterm*

$$\begin{aligned}
\hat{J}^B_{k|N}(\mathbf{x}) = \min_{\substack{(\tilde{\mathbf{x}}_{k-N},\\ \tilde{\mathbf{w}}_{k-N|k-1})}} \quad & J_{k|N}(\tilde{\mathbf{x}}_{k-N}, \tilde{\mathbf{w}}_{k-N|k-1}) && && \text{(2.51a)}\\
\text{u.B.v.} \quad & \tilde{\mathbf{x}}_{l+1} = \mathbf{f}_l(\tilde{\mathbf{x}}_l, \tilde{\mathbf{w}}_l) && \forall\, l = k-N, \ldots, k-1 && \text{(2.51b)}\\
& \tilde{\mathbf{v}}_l = \mathbf{y}_l - \mathbf{h}_l(\tilde{\mathbf{x}}_l) && \forall\, l = k-N, \ldots, k-1 && \text{(2.51c)}\\
& \tilde{\mathbf{x}}_l \in X_l && \forall\, l = k-N, \ldots, k-1 && \text{(2.51d)}\\
& \tilde{\mathbf{w}}_l \in W_l\ , \qquad \tilde{\mathbf{v}}_l \in V_l\ , && \forall\, l = k-N, \ldots, k-1 && \text{(2.51e)}\\
& \tilde{\mathbf{x}}_k = \mathbf{x} && && \text{(2.51f)}
\end{aligned}$$

definiert, der den finiten Horizont $k-N, \ldots, k$ mit $k > N$ abdeckt. Für die Gütefunktion gilt in üblicher Art

$$J_{k|N}(\tilde{\mathbf{x}}_{k-N}, \tilde{\mathbf{w}}_{k-N|k-1}) = B_{k-N}(\tilde{\mathbf{x}}_{k-N}) + \sum_{l=k-N}^{k-1} b_l(\tilde{\mathbf{w}}_l, \tilde{\mathbf{v}}_l)\ . \tag{2.52}$$

Für den Fall $k \leq N$ wird $\hat{J}^B_{k|N}(\,\cdot\,) = \hat{J}^B_{k|k}(\,\cdot\,)$ mit $\hat{J}^B_{k|k}(\,\cdot\,)$ gemäß (2.36) verwendet.

Mit rekursiven Definitionen $B_k(\,\cdot\,) = \hat{J}^B_{k|N}(\,\cdot\,)$, $B_{k-N}(\,\cdot\,) = \hat{J}^B_{k-N|N}(\,\cdot\,)$, ... würde man also wieder den *Ankunftskostenterm für vollständige Information* gemäß (2.36) erhalten. Es stellt sich daher die Frage, wie die Funktion $\hat{J}^B_{k|N}(\,\cdot\,)$ sinnvoll durch $B_k(\,\cdot\,)$ approximiert werden kann. Es kann gezeigt werden (siehe [2.2]), dass für Stabilität und Konvergenz des

MHE Schätzfehlers die Einhaltung der folgenden Ungleichungen von Bedeutung ist:

$$\hat{J}^B_{k|N}(\mathbf{x}) \geq B_k(\mathbf{x}) \geq \hat{J}_{k|N} + \underline{\gamma}_B(||\mathbf{x} - \bar{\mathbf{x}}_k||_2) \qquad \forall\, \mathbf{x} \in X_k, k > N \ . \tag{2.53}$$

Hierbei ist $\underline{\gamma}_B(\,\cdot\,) \in \mathcal{K}_\infty$ die bereits in Annahme A5 verwendete Vergleichsfunktion und

$$\bar{\mathbf{x}}_k = \arg\min_{\mathbf{x} \in X_k} \hat{J}^B_{k|N}(\mathbf{x}) = \check{\mathbf{x}}_k(k-N, \hat{\mathbf{x}}_{k-N}, \hat{\mathbf{w}}_{k-N|k-1}) \tag{2.54}$$

die A-priori-Schätzung gemäß Annahme A4, wobei $(\hat{\mathbf{x}}_{k-N}, \hat{\mathbf{w}}_{k-N|k-1})$ natürlich einfach die Lösung der ursprünglichen Optimierungsaufgabe (2.6) (ohne die Endgleichungsbeschränkung (2.51f)) ist und

$$\hat{J}_{k|N} = B_k(\bar{\mathbf{x}}_k) = \min_{\mathbf{x} \in X_k} \hat{J}^B_{k|N}(\mathbf{x}) \tag{2.55}$$

der zugehörige optimale Gütefunktionswert. Für einen skalaren Zustand x ist in Abbildung 2.2 ein Beispiel für $B_k(\,\cdot\,)$ dargestellt, welches (2.53) bis (2.55) erfüllt.

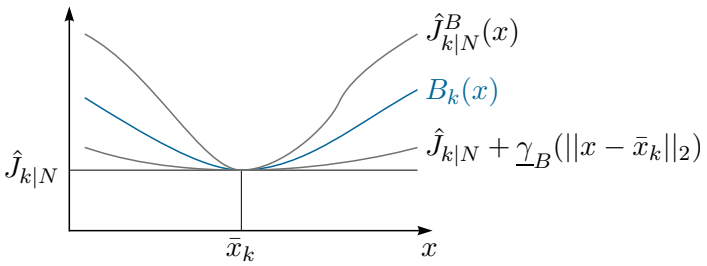


Abbildung 2.2: Beispiel für einen möglichen Anfangskostenterm $B_k(x)$, der (2.53) und (2.55) erfüllt.

Eine Voraussetzung für die Konvergenz der MHE Zustandsschätzung ist die im folgenden Satz postulierte Beschränktheit des Anfangskostenterms.

Satz 2.3 (Beschränktheit der MHE Gütefunktionswerte)**.** *Wird die Ungleichung* (2.53) *eingehalten, so gilt*

$$\hat{J}^B_{k|N}(\mathbf{x}) \leq \hat{J}^B_{k|k}(\mathbf{x}) \ , \qquad \hat{J}_{k|N} \leq \hat{J}_{k|k} \ , \qquad \forall\, \mathbf{x} \in X_k, k \geq 0 \ . \tag{2.56}$$

Aufgabe 2.2. Beweisen Sie Satz 2.3. Dies gelingt besonders einfach mittels vollständiger Induktion.

Definition 2.7 (MHE Detektierbarkeit)**.** Das System (2.1) ist *MHE detektierbar*, wenn das erweiterte Modell

$$\mathbf{x}_{k+1} = \mathbf{f}_k(\mathbf{x}_k, \mathbf{w}_{1,k}) + \mathbf{w}_{2,k} \qquad \forall\, k \in \mathbb{N}_0 \tag{2.57a}$$

$$\mathbf{y}_k = \mathbf{h}_k(\mathbf{x}_k) + \mathbf{v}_k \qquad \forall\, k \in \mathbb{N}_0 \tag{2.57b}$$

mit der erweiterten Prozessstörung $\mathbf{w}_k = [\mathbf{w}_{1,k}^{\mathrm{T}}, \mathbf{w}_{2,k}^{\mathrm{T}}]^{\mathrm{T}}$ die IIOSS-Eigenschaft gemäß Definition 2.4 besitzt.

Bemerkung 2.2. Die Eigenschaft *MHE Detektierbarkeit* ist geringfügig restriktiver als die Eigenschaft IIOSS aber weniger restriktiv als die Eigenschaft Beobachtbarkeit des Endzustands gemäß Definition 2.6. Bei vielen IIOSS Systemen tritt bereits die ursprüngliche Störung $\mathbf{w}_k$ in (2.1a) nur additiv auf. Diese Systeme sind natürlich automatisch MHE detektierbar.

Satz 2.4 (RGAS des MHE Zustandsbeobachters mit Approximation der MHE Ankunftskosten)**.** *Für ein System* (2.1)*, das MHE detektierbar im Sinne der Definition 2.7 ist, mit Störungen, die der Annahme A3 genügen, führt der MHE Zustandsschätzer* (2.6) *mit festem N sowie einem Anfangskostenterm B_k, der die Ungleichung* (2.53) *erfüllt (Schätzung mit Approximation der MHE Ankunftskosten), und einer Kostenfunktion b_k gemäß der Annahme A5 zu einem robust global asymptotisch stabilen Schätzfehler.*

Der Beweis dieses Satzes ist in [2.2] zu finden. Mit der Ungleichung (2.53) kennt man nun Bedingungen, die der Anfangskostenterm $B_k(\,\cdot\,)$ erfüllen muss, um einen RGAS Beobachter zu erhalten. Der tatsächliche Entwurf der Funktion $B_k(\,\cdot\,)$ bleibt aber eine anspruchsvolle Aufgabe [2.11].

2.5 Maximum-a-posteriori Zustandsschätzung

Wird das System (2.1) als stochastischer Prozess aufgefasst, so liefert die Wahrscheinlichkeitstheorie konkrete Hinweise auf eine sinnvolle Wahl der Funktionen $B_k(\,\cdot\,)$ und $b_k(\,\cdot\,,\,\cdot\,)$. Sind bestimmte Wahrscheinlichkeitsdichtefunktionen bekannt, so kann mit dem *Maximum-a-posteriori (MAP) Schätzer* sogar ein optimaler Beobachter konstruiert werden. Die nachfolgenden Ausführungen basieren zum Teil auf [2.7, 2.15]. Einige Grundlagen der Stochastik wurden auch bereits in der Vorlesung *Regelungssysteme* [2.10] besprochen. Zunächst wird der Begriff des Modalwerts definiert:

Definition 2.8 (Modalwert einer Wahrscheinlichkeitsdichtefunktion)**.** Für eine stetige Wahrscheinlichkeitsdichtefunktion $P_{\mathbf{z}}(\mathbf{z})$ ist der Modalwert $\hat{\mathbf{z}}$ definiert als

$$\hat{\mathbf{z}} = \arg\max_{\mathbf{z}} P_{\mathbf{z}}(\mathbf{z}) \ . \tag{2.58}$$

Bei der Realisierung einer Zufallsvariable $\mathbf{z}$ tritt also der Modalwert $\hat{\mathbf{z}}$ mit der höchsten Wahrscheinlichkeit auf. Man beachte, dass (2.58) immer eine Lösung besitzt, diese muss aber nicht eindeutig sein.

Gemäß dem Modell (2.1) liegen statistische Zusammenhänge zwischen den Messgrößen $\mathbf{y}_{0|K-1}$ und den unbekannten Größen $\mathbf{x}_0$ und $\mathbf{w}_{0|K-1}$ vor. Für das Schätzproblem mit vollständiger Information können diese Zusammenhänge durch die *bedingte Wahrscheinlichkeitsdichte*

$$P(\mathbf{x}_0, \mathbf{w}_{0|K-1} | \mathbf{y}_{0|K-1}) \tag{2.59}$$

beschrieben werden.

Definition 2.9 (Maximum-a-posteriori (MAP) Zustandsschätzer)**.** Ein MAP Zustandsschätzer für das System (2.1) liefert für gegebene Messwerte $\mathbf{y}_{0|K-1}$, d. h. bei Berücksichtigung der vollständigen Information, als Schätzwert für die unbekannten Größen $\mathbf{x}_0$ und $\mathbf{w}_{0|K-1}$ zum Zeitpunkt $K \in \mathbb{N}_0$ den Modalwert

$$(\hat{\mathbf{x}}_0, \hat{\mathbf{w}}_{0|K-1}) = \underset{(\tilde{\mathbf{x}}_0, \tilde{\mathbf{w}}_{0|K-1})}{\arg\max} P(\tilde{\mathbf{x}}_0, \tilde{\mathbf{w}}_{0|K-1} | \mathbf{y}_{0|K-1}) \ . \tag{2.60}$$

Diese Optimierungsaufgabe kann um die Beschränkungen (2.2) erweitert werden. Derartige Beschränkungen können aber auch bereits in der Formulierung von (2.59) berücksichtigt werden, indem die Wahrscheinlichkeitsdichte für nicht zulässige Werte auf Null gesetzt wird. Der Einfachheit halber wird in diesem Abschnitt auf Beschränkungen der Art (2.2) verzichtet.

Im nächsten Schritt soll die bedingte Wahrscheinlichkeitdichte (2.59) konkreter formuliert werden. Da es sich bei $\mathbf{x}_0$ und $\mathbf{w}_{0|K-1}$ um stochastisch unabhängige Zufallsvariablen handelt, gilt

$$P(\mathbf{x}_0, \mathbf{w}_{0|K-1}) = P_{\mathbf{x}_0}(\mathbf{x}_0) \prod_{k=0}^{K-1} P_{\mathbf{w}_k}(\mathbf{w}_k) \ . \tag{2.61}$$

Berücksichtigt man (2.1) und die stochastische Unabhängigkeit der Zufallsvariablen $\mathbf{v}_{0|K-1}$, so folgt für die bedingte Wahrscheinlichkeit der Messgrößen

$$P(\mathbf{y}_{0|K-1} | \mathbf{x}_0, \mathbf{w}_{0|K-1}) = \prod_{k=0}^{K-1} P_{\mathbf{v}_k}(\mathbf{y}_k - \mathbf{h}_k(\check{\mathbf{x}}_k(0, \mathbf{x}_0, \mathbf{w}_{0|k-1}))) = \prod_{k=0}^{K-1} P_{\mathbf{v}_k}(\mathbf{v}_k) \ . \tag{2.62}$$

Mit dem Satz von Bayes folgt aus (2.61) und (2.62) die bedingte Wahrscheinlichkeitsdichte

$$\begin{aligned} P(\mathbf{x}_0, \mathbf{w}_{0|K-1} | \mathbf{y}_{0|K-1}) &= \frac{P(\mathbf{y}_{0|K-1} | \mathbf{x}_0, \mathbf{w}_{0|K-1}) P(\mathbf{x}_0, \mathbf{w}_{0|K-1})}{P(\mathbf{y}_{0|K-1})} \\ &= \frac{P_{\mathbf{x}_0}(\mathbf{x}_0) \prod_{k=0}^{K-1} P_{\mathbf{w}_k}(\mathbf{w}_k) P_{\mathbf{v}_k}(\mathbf{v}_k)}{P(\mathbf{y}_{0|K-1})} \ , \end{aligned} \tag{2.63}$$

welche direkt in das Schätzgesetz (2.60) eingesetzt werden könnte. Berücksichtigt man (2.1b), die Monotonizität der Logarithmusfunktion $\ln(\,\cdot\,)$ und dass für gegebene Messwerte $\mathbf{y}_{0|K-1}$ der Nenner von (2.63) eine feste Zahl ist, so kann die vereinfachte Optimierungsaufgabe

$$(\hat{\mathbf{x}}_0, \hat{\mathbf{w}}_{0|K-1}) = \underset{(\tilde{\mathbf{x}}_0, \tilde{\mathbf{w}}_{0|K-1})}{\arg\min} \left(B_0(\tilde{\mathbf{x}}_0) + \sum_{k=0}^{K-1} b_k(\tilde{\mathbf{w}}_k, \mathbf{y}_k - \mathbf{h}_k(\check{\mathbf{x}}_k(0, \tilde{\mathbf{x}}_0, \tilde{\mathbf{w}}_{0|k-1}))) \right) \tag{2.64}$$

mit den Kostenfunktionen

$$B_0(\tilde{\mathbf{x}}_0) = -\ln(P_{\mathbf{x}_0}(\tilde{\mathbf{x}}_0)) \tag{2.65a}$$

$$b_k(\tilde{\mathbf{w}}_k, \tilde{\mathbf{v}}_k) = -\ln(P_{\mathbf{w}_k}(\tilde{\mathbf{w}}_k)) - \ln(P_{\mathbf{v}_k}(\tilde{\mathbf{v}}_k)) \tag{2.65b}$$

formuliert werden. Offensichtlich liefert sie das gleiche Ergebnis wie (2.60).

Da (2.64) der ursprünglichen Optimierungsaufgabe (2.6) (ohne Beschränkungen) mit $N = K$ für die Zustandsschätzung mit vollständiger Information entspricht, ist mit (2.65) die aus wahrscheinlichkeitstheoretischer Sicht optimale Formulierung der Kostenfunktionen $B_0(\,\cdot\,)$ und $b_k(\,\cdot\,,\,\cdot\,)$ gefunden.

Analog zu Definition 2.9 kann natürlich sofort ein MHE MAP Zustandsschätzer formuliert werden. Es tritt dann statt (2.59) die bedingte Wahrscheinlichkeitsdichte $P(\mathbf{x}_{K-N}, \mathbf{w}_{K-N|K-1}|\mathbf{y}_{K-N|K-1})$ mit $K > N$ und festem N auf. Bei der Formulierung der Wahrscheinlichkeitsdichte von $\mathbf{x}_{K-N}$ sind bei Verwertung der vollständigen Information auch die Messwerte $\mathbf{y}_{0|K-N-1}$ zu berücksichtigen. In diesem Fall lautet die der Optimierungsaufgabe zugrunde liegende bedingte Wahrscheinlichkeitsdichte also

$$\begin{aligned} &P(\mathbf{x}_{K-N}, \mathbf{w}_{K-N|K-1}|\mathbf{y}_{0|K-1}) \\ &\qquad = P(\mathbf{x}_{K-N}|\mathbf{y}_{0|K-N-1}) P(\mathbf{w}_{K-N|K-1}|\mathbf{x}_{K-N}, \mathbf{y}_{K-N|K-1}) \,. \end{aligned} \tag{2.66}$$

Die hier vorgenommene Aufspaltung in zwei Faktoren ist wegen der Markov-Eigenschaft des stochastischen Prozesses zulässig.

Die Wahrscheinlichkeitsdichte $P(\mathbf{x}_{K-N}|\mathbf{y}_{0|K-N-1})$ modelliert alle Informationen, die vor dem Zeitpunkt $K-N$ gesammelt wurden. Sie kann gemäß dem Satz der totalen Wahrscheinlichkeit durch Integration der Wahrscheinlichkeitsdichte $P(\mathbf{x}_0, \mathbf{w}_{0|K-N-1}|\mathbf{y}_{0|K-N-1})$ über die Variablen $\mathbf{x}_0$, $\mathbf{w}_0$, $\mathbf{w}_1$, …, $\mathbf{w}_{K-N-1}$ unter Berücksichtigung der Bedingung $\mathbf{x}_{K-N} = \check{\mathbf{x}}_{K-N}(0, \mathbf{x}_0, \mathbf{w}_{0|K-N-1})$ berechnet werden. Natürlich korrespondiert $P(\mathbf{x}_{K-N}|\mathbf{y}_{0|K-N-1})$ mit dem Ankunftskostenterm (2.36) für vollständige Information. In vielen praktischen Fällen wird man aufgrund des Rechenaufwands $P(\mathbf{x}_{K-N}|\mathbf{y}_{0|K-N-1})$ durch eine leichter berechenbare Approximation ersetzen. Wird z. B. von einer gleichverteilten Zufallsvariable $\mathbf{x}_{K-N}$ ohne Berücksichtigung der Messwerte $\mathbf{y}_{0|K-N-1}$ ausgegangen, so ist $P(\mathbf{x}_{K-N}|\mathbf{y}_{0|K-N-1})$ konstant bezüglich all seiner Argumente und man erhält eine MHE Zustandsschätzung ohne Anfangskostenterm (vgl. Abschnitt 2.4.2). In [2.16] werden verschiedene weitere Möglichkeiten zur Approximation von $P(\mathbf{x}_{K-N}|\mathbf{y}_{0|K-N-1})$ mittels Filter (extended Kalman-Filter, unscented Kalman-Filter, Partikelfilter, etc.), die entlang der mit MHE geschätzten Folge $\hat{\mathbf{x}}_0$, $\hat{\mathbf{x}}_1$, …, $\hat{\mathbf{x}}_{K-N}$ mitgerechnet werden, vorgestellt.

Fasst man nun für gegebene, feste Messwerte $\mathbf{y}_{0|K-N-1}$ den Zustand $\mathbf{x}_{K-N}$ als Zufallsvariable mit der Wahrscheinlichkeitsdichte

$$P_{\mathbf{x}_{K-N}}(\mathbf{x}_{K-N}) = P(\mathbf{x}_{K-N}|\mathbf{y}_{0|K-N-1}) \tag{2.67}$$

auf, so erhält man in völlig analoger Weise zur Herleitung von (2.64)

$$\begin{aligned} &(\hat{\mathbf{x}}_{K-N}, \hat{\mathbf{w}}_{K-N|K-1}) \\ &\qquad = \underset{\substack{(\tilde{\mathbf{x}}_{K-N},\\ \tilde{\mathbf{w}}_{K-N|K-1})}}{\arg\min} \Bigg(B_{K-N}(\tilde{\mathbf{x}}_{K-N}) \\ &\qquad\quad + \sum_{k=K-N}^{K-1} b_k(\tilde{\mathbf{w}}_k, \mathbf{y}_k - \mathbf{h}_k(\check{\mathbf{x}}_k(K-N, \tilde{\mathbf{x}}_{K-N}, \tilde{\mathbf{w}}_{K-N|k-1}))) \Bigg) \end{aligned} \tag{2.68}$$

mit den aus wahrscheinlichkeitstheoretischer Sicht optimalen Kostenfunktionen

$$B_k(\tilde{\mathbf{x}}_k) = -\ln(P_{\mathbf{x}_k}(\tilde{\mathbf{x}}_k)) = -\ln(P(\tilde{\mathbf{x}}_k|\mathbf{y}_{0|k-1})) \tag{2.69a}$$
$$b_k(\tilde{\mathbf{w}}_k, \tilde{\mathbf{v}}_k) = -\ln(P_{\mathbf{w}_k}(\tilde{\mathbf{w}}_k)) - \ln(P_{\mathbf{v}_k}(\tilde{\mathbf{v}}_k)) \ . \tag{2.69b}$$

2.6 Zustands- und Parameterschätzung

In vielen Aufgabenstellungen sollen nicht nur die Systemzustände $\mathbf{x}_k$ sondern auch unbekannte Systemparameter $\boldsymbol{\rho} \in R \subseteq \mathbb{R}^r$ geschätzt werden [2.3, 2.17], wobei R die Menge der zulässigen Systemparameter ist. Nachfolgend wird skizziert, wie dies mit den bisher in diesem Abschnitt beschriebenen Verfahren einfach möglich ist.

Ist bekannt, dass die Parameter $\boldsymbol{\rho}$ zumindest innerhalb des Schätzhorizonts $K-N, \dots, K$ konstant sind, so kann das Modell (2.1) in der Form

$$\mathbf{x}_{k+1} = \mathbf{f}_k(\mathbf{x}_k, \mathbf{w}_k, \boldsymbol{\rho}) \qquad \forall\, k \in \mathbb{N}_0 \tag{2.70a}$$
$$\mathbf{y}_k = \mathbf{h}_k(\mathbf{x}_k, \boldsymbol{\rho}) + \mathbf{v}_k \qquad \forall\, k \in \mathbb{N}_0 \tag{2.70b}$$

angeschrieben werden. Zusätzlich zu $\tilde{\mathbf{x}}_{K-N}$ und $\tilde{\mathbf{w}}_{K-N|K-1}$ wird dann in der Optimierungsaufgabe $\tilde{\boldsymbol{\rho}} \in R$ als (beschränkte) Optimierungsvariable verwendet. Die Kostenfunktionen B_k und b_k sind entsprechend anzupassen. Mit jedem Aufruf des MHE Schätzers wird ein neuer Schätzwert $\hat{\boldsymbol{\rho}}$ für $\boldsymbol{\rho}$ ermittelt.

Ist hingegen bekannt, dass die Parameter $\boldsymbol{\rho}$ einer (zumeist langsamen) Dynamik unterliegen und ist die Länge N des Schätzhorizonts so groß, dass dynamische Änderungen von $\boldsymbol{\rho}$ relevant sein können, so kann $\boldsymbol{\rho}_k \in R_k \subseteq \mathbb{R}^r$ als weiterer Systemzustand aufgefasst werden und das Modell (2.70) in die Form

$$\mathbf{x}_{k+1} = \mathbf{f}_k(\mathbf{x}_k, \mathbf{w}_k, \boldsymbol{\rho}_k) \qquad \forall\, k \in \mathbb{N}_0 \tag{2.71a}$$
$$\boldsymbol{\rho}_{k+1} = \mathbf{g}_k(\mathbf{x}_k, \mathbf{w}_k, \boldsymbol{\rho}_k) \qquad \forall\, k \in \mathbb{N}_0 \tag{2.71b}$$
$$\mathbf{y}_k = \mathbf{h}_k(\mathbf{x}_k, \boldsymbol{\rho}_k) + \mathbf{v}_k \qquad \forall\, k \in \mathbb{N}_0 \tag{2.71c}$$

umgeschrieben werden. Die Funktion $\mathbf{g}_k$ ist nun entweder ein bekanntes Modell der Parameteränderungen oder, falls ein solches nicht vorhanden ist, ein *random-walk Modell* der Form

$$\mathbf{g}_k(\mathbf{x}_k, \mathbf{w}_k, \boldsymbol{\rho}_k) = \boldsymbol{\rho}_k + \mathbf{w}_k^{\rho} \tag{2.72}$$

mit zufälligen Störungen $\mathbf{w}_k^{\rho}$, die Teil der Prozessstörung $\mathbf{w}_k$ sind. Je nach Art der Parameteränderungen kann $\mathbf{w}_k^{\rho}$ stark beschränkt oder in der Kostenfunktion b_k stark bestraft werden.

2.7 Literatur

[2.1] C. Rao, „Moving Horizon Strategies for the Constrained Monitoring and Control of Nonlinear Discrete-Time Systems,“ Diss., University of Wisconsin-Madison, Wisconsin, 2000.

[2.2] J.B. Rawlings, D.Q. Mayne und M.M. Diehl, *Model Predictive Control: Theory, Computation, and Design*, 2. Aufl. Madison, Wisconsin: Nob Hill Publishing, 2017.

[2.3] N. Haverbeke, „Efficient Numerical Methods for Moving Horizon Estimation,“ Diss., Katholieke Universiteit Leuven, Heverlee, Belgium, 2011.

[2.4] P. Philipp, „Centralized and Distributed Moving Horizon Strategies for State Estimation of Networked Control Systems,“ Diss., Technische Universität München, München, 2014.

[2.5] F. Allgöwer, T. Badgwell, J. Qin, J. Rawlings und S. Wright, „Nonlinear Predictive Control and Moving Horizon Estimation: An Introductory Overview,“ in *Advances in Control*, Springer London, 1999, S. 391–449.

[2.6] M. Alamir, „Nonlinear Moving Horizon Observers: Theory and Real-Time Implementation,“ in *Nonlinear Observers and Applications*, Ser. Lecture Notes in Control and Information Sciences, Bd. 363, Berlin Heidelberg: Springer, 2007, S. 139–179.

[2.7] E. Haseltine und J. Rawlings, „Critical evaluation of extended Kalman filtering and moving-horizon estimation,“ *Industrial and Engineering Chemistry Research*, Jg. 44, Nr. 8, S. 2451–2460, 2005.

[2.8] C. Rao, J. Rawlings und D. Mayne, „Constrained state estimation for nonlinear discrete-time systems: stability and moving horizon approximations,“ *IEEE Transactions on Automatic Control*, Jg. 48, Nr. 2, S. 246–258, Feb. 2003.

[2.9] D. Simon, „Kalman filtering with state constraints: a survey of linear and nonlinear algorithms,“ *Control Theory Applications, IET*, Jg. 4, Nr. 8, S. 1303–1318, 2010.

[2.10] W. Kemmetmüller, *Skriptum zur VO Regelungssysteme (WS 2023/2024)*, Institut für Automatisierungs- und Regelungstechnik, TU Wien, 2023. Adresse: https://www.acin.tuwien.ac.at/master/regelungssysteme/.

[2.11] J.B. Rawlings und D.Q. Mayne, *Model Predictive Control: Theory and Design*. Madison, Wisconsin: Nob Hill Publishing, 2009.

[2.12] J. Rawlings und L. Ji, „Optimization-based state estimation: Current status and some new results,“ *Journal of Process Control*, Jg. 22, Nr. 8, S. 1439–1444, 2012.

[2.13] E. Sontag, „Comments on integral variants of ISS,“ *Systems and Control Letters*, Jg. 34, S. 93–100, 1998.

[2.14] E. Sontag und Y. Wang, „Output-to-state stability and detectability of nonlinear systems,“ *Systems and Control Letters*, Jg. 29, S. 279–290, 1997.

[2.15] C. Rao und J. Rawlings, „Nonlinear Moving Horizon State Estimation,“ in *Nonlinear Model Predictive Control*, Ser. Progress in Systems and Control Theory, Bd. 26, Birkhäuser Basel, 2000, S. 45–69.

[2.16] S. Ungarala, „Computing arrival cost parameters in moving horizon estimation using sampling based filters,“ *Journal of Process Control*, Jg. 19, Nr. 9, S. 1576–1588, 2009.

[2.17] T. Johansen, „Introduction to Nonlinear Model Predictive Control and Moving Horizon Estimation,“ in *Selected Topics on Constrained and Nonlinear Control*, Bratislav, Trondheim: STU/NTNU, 2011, S. 1–53.

3 Optimierungsbasierte Schätzung

3.1 Parameterschätzung für ein lineares Modell

3.1.1 Der reguläre Fall

Von einem System sind die zu einem Vektor zusammengefassten Ausgangswerte $\mathbf{y} \in \mathbb{R}^m$ (z. B. aus Messungen) verfügbar. Für sie wird ein lineares Modell der Form

$$\mathbf{y} = \mathbf{S}\mathbf{p} + \mathbf{v} \tag{3.1}$$

mit unbekannten Systemparametern $\mathbf{p} \in \mathbb{R}^n$ angenommen. Hierbei ist die deterministische spaltenreguläre Matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$ (Datenmatrix) bekannt und $\mathbf{v}$ ist eine stochastische Störung (Zufallszahl) mit Erwartungswert $\mathrm{E}(\mathbf{v}) = \mathbf{0}$ und bekannter, symmetrischer, positiv definierter Kovarianzmatrix $\mathrm{E}(\mathbf{v}\mathbf{v}^{\mathrm{T}}) = \mathbf{Q} > 0$. Die Verteilung von $\mathbf{v}$ kann beliebig und unbekannt sein. Natürlich kann sich die Abbildung (3.1) aus Messungen des Ausgangs eines linearen dynamischen Systems ergeben (siehe dazu die Beispiele 3.1 und 3.2 am Ende dieses Abschnitts). Der Vektor $\mathbf{p}$ könnte dann neben Systemparametern auch den unbekannten Anfangszustand des Systems enthalten.

Der nun folgende Entwurf eines optimalen Schätzers ist an [3.1] angelehnt. Alternative Herleitungen finden sich in [3.2–3.4]. Die unbekannten Parameter $\mathbf{p}$ sollen von einem *linearen* Schätzer der Form

$$\hat{\mathbf{p}} = \mathbf{K}\mathbf{y} \tag{3.2}$$

mit einer noch zu bestimmenden Matrix $\mathbf{K} \in \mathbb{R}^{n \times m}$ *erwartungstreu* und *mit minimaler Varianz* des Parameterschätzfehlers geschätzt werden. Es soll also

$$\mathrm{E}(\hat{\mathbf{p}}) = \mathbf{p} \tag{3.3}$$

gelten und die Summe der Einzelvarianzen

$$\sum_{i=1}^{n} \mathrm{E}\left((\hat{p}_i - p_i)^2\right) = \mathrm{E}\left(\|\hat{\mathbf{p}} - \mathbf{p}\|_2^2\right) = \mathrm{spur}\left(\mathrm{E}\left((\hat{\mathbf{p}} - \mathbf{p})(\hat{\mathbf{p}} - \mathbf{p})^{\mathrm{T}}\right)\right) \tag{3.4}$$

soll minimal sein. Wegen

$$\mathrm{E}(\hat{\mathbf{p}}) = \mathrm{E}(\mathbf{K}\mathbf{y}) = \mathrm{E}(\mathbf{K}(\mathbf{S}\mathbf{p} + \mathbf{v})) = \mathbf{K}(\mathbf{S}\mathbf{p} + \mathrm{E}(\mathbf{v})) = \mathbf{K}\mathbf{S}\mathbf{p} \tag{3.5}$$

muss für die Erwartungstreue des Schätzers

$$\mathbf{K}\mathbf{S} = \mathbf{E} \tag{3.6}$$

erfüllt sein. Berücksichtigt man dies in (3.4), so folgt

$$\begin{aligned} &\mathrm{E}\left(\|\hat{\mathbf{p}}-\mathbf{p}\|_2^2\right)=\mathrm{E}\left(\|\mathbf{K}(\mathbf{S}\mathbf{p}+\mathbf{v})-\mathbf{p}\|_2^2\right)=\mathrm{E}\left(\|\mathbf{K}\mathbf{v}\|_2^2\right)=\mathrm{E}\left(\mathbf{v}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}\mathbf{K}\mathbf{v}\right)\\ &=\mathrm{E}\left(\operatorname{spur}\left(\mathbf{K}\mathbf{v}\mathbf{v}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}\right)\right)=\operatorname{spur}\left(\mathbf{K}\mathbf{Q}\mathbf{K}^{\mathrm{T}}\right)=\sum_{i,j,k}K_{ij}Q_{jk}K_{ik}=(\mathbf{K}\mathbf{Q}):\mathbf{K} \end{aligned} \tag{3.7}$$

mit $\mathbf{K}=[K_{ij}]$, $\mathbf{Q}=[Q_{ij}]$ und dem doppelt verjüngenden Produkt $\mathbf{A}:\mathbf{B}=\sum_{i,j}A_{ij}B_{ij}$ zweier Matrizen $\mathbf{A}$ und $\mathbf{B}$. Zur Bestimmung von $\mathbf{K}$ wird daher die Optimierungsaufgabe

$$\min_{\mathbf{K}\in\mathbb{R}^{n\times m}} \quad (\mathbf{K}\mathbf{Q}):\mathbf{K} \tag{3.8a}$$

$$\text{u.B.v.} \quad \mathbf{K}\mathbf{S}-\mathbf{E}=\mathbf{0} \tag{3.8b}$$

formuliert. Die zugehörige Lagrangefunktion

$$\begin{aligned} L(\mathbf{K},\mathbf{\Lambda})&=(\mathbf{K}\mathbf{Q}):\mathbf{K}+(\mathbf{K}\mathbf{S}-\mathbf{E}):\mathbf{\Lambda}\\ &=\sum_{i,j,k}K_{ij}Q_{jk}K_{ik}+\sum_{i,j,k}K_{ij}S_{jk}\Lambda_{ik}-\sum_{i,j}\delta_{ij}\Lambda_{ij} \end{aligned} \tag{3.9}$$

mit $\mathbf{S}=[S_{ij}]$, den Lagrange-Multiplikatoren $\mathbf{\Lambda}=[\Lambda_{ij}]$ und dem Kronecker-Symbol

$$\delta_{ij}=\begin{cases}1 & \text{für } i=j\\ 0 & \text{für } i\neq j\end{cases} \tag{3.10}$$

führt auf die KKT-Bedingungen (erster Ordnung)

$$\frac{\partial L(\mathbf{K},\mathbf{\Lambda})}{\partial K_{ij}}=\sum_k(Q_{jk}K_{ik}+K_{ik}Q_{kj})+\sum_k S_{jk}\Lambda_{ik}=0 \qquad \forall\, i,j \tag{3.11a}$$

$$\frac{\partial L(\mathbf{K},\mathbf{\Lambda})}{\partial \Lambda_{ij}}=\sum_k K_{ik}S_{kj}-\delta_{ij}=0 \qquad \forall\, i,j\ . \tag{3.11b}$$

In Matrixschreibweise lauten diese

$$\frac{\partial L(\mathbf{K},\mathbf{\Lambda})}{\partial \mathbf{K}}=2\mathbf{K}\mathbf{Q}+\mathbf{\Lambda}\mathbf{S}^{\mathrm{T}}=\mathbf{0} \tag{3.12a}$$

$$\frac{\partial L(\mathbf{K},\mathbf{\Lambda})}{\partial \mathbf{\Lambda}}=\mathbf{K}\mathbf{S}-\mathbf{E}=\mathbf{0}\ . \tag{3.12b}$$

Wird (3.12a) rechtsseitig mit $\mathbf{Q}^{-1}\mathbf{S}$ multipliziert, so ergibt sich unter Verwendung von (3.12b)

$$\mathbf{\Lambda}=-2(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1}\ . \tag{3.13}$$

Nach Einsetzen in (3.12a) folgt daher

$$\mathbf{K}=(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\ . \tag{3.14}$$

Zum Nachweis der strikten Optimalität von $\mathbf{K}$ gemäß (3.14) muss noch gezeigt werden, dass jede andere Matrix $\tilde{\mathbf{K}}$, die $\tilde{\mathbf{K}} \neq \mathbf{K}$ und die Nebenbedingung (3.8b) erfüllt, zu einem schlechteren Kostenfunktionswert (3.8a) führt. Dieser Nachweis folgt unter Beachtung der Identitäten $(\mathbf{KQ}) : \mathbf{K} = \text{spur}(\mathbf{KQK}^{\mathrm{T}})$, $\tilde{\mathbf{K}}\mathbf{S} = \mathbf{E}$, $\mathbf{Q} > 0$, $\tilde{\mathbf{K}} \neq \mathbf{K}$ und $\mathbf{KQK}^{\mathrm{T}} = (\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1}$ aus der Ungleichung

$$\begin{aligned} 0 &< \text{spur}\left((\tilde{\mathbf{K}} - \mathbf{K})\mathbf{Q}(\tilde{\mathbf{K}} - \mathbf{K})^{\mathrm{T}}\right) \\ &= \text{spur}\left(\tilde{\mathbf{K}}\mathbf{Q}\tilde{\mathbf{K}}^{\mathrm{T}}\right) - 2\,\text{spur}\left(\mathbf{KQ}\tilde{\mathbf{K}}^{\mathrm{T}}\right) + \text{spur}\left(\mathbf{KQK}^{\mathrm{T}}\right) \\ &= \text{spur}\left(\tilde{\mathbf{K}}\mathbf{Q}\tilde{\mathbf{K}}^{\mathrm{T}}\right) - 2\,\text{spur}\left((\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1}\underbrace{\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{Q}\tilde{\mathbf{K}}^{\mathrm{T}}}_{\mathbf{E}}\right) + \text{spur}\left(\mathbf{KQK}^{\mathrm{T}}\right) \\ &= \text{spur}\left(\tilde{\mathbf{K}}\mathbf{Q}\tilde{\mathbf{K}}^{\mathrm{T}}\right) - \text{spur}\left(\mathbf{KQK}^{\mathrm{T}}\right) . \end{aligned} \tag{3.15}$$

Der optimale lineare erwartungstreue Schätzer lautet also

$$\hat{\mathbf{p}} = (\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{y} . \tag{3.16}$$

Er wird auch als BLUE (best linear unbiased estimator) oder *Gauß-Markov-Schätzer* bezeichnet. Genügt $\mathbf{v}$ außerdem einer Normalverteilung, so kann gezeigt werden, dass es im Sinne des Minimum-Varianz-Kriteriums keinen besseren (linearen oder nichtlinearen) erwartungstreuen Schätzer als (3.16) gibt [3.1, 3.5].

Mit dem Schätzer (3.16) lautet die Kovarianzmatrix des Parameterschätzfehlers

$$\mathrm{E}\left((\hat{\mathbf{p}} - \mathbf{p})(\hat{\mathbf{p}} - \mathbf{p})^{\mathrm{T}}\right) = \mathbf{KQK}^{\mathrm{T}} = (\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1} > 0 . \tag{3.17}$$

Sie vergrößert sich also tendenziell mit wachsenden Werten für $\mathbf{Q}$ und sie verkleinert sich tendenziell mit wachsenden Werten für $\mathbf{S}$ und m (Anzahl der Messwerte). Ihre positive Definitheit folgt aus der Zeilenregularität von $\mathbf{K}$ und der Regularität von $\mathbf{Q}$.

Aufgabe 3.1 (Gewichtete lineare Least-Squares Methode). Zeigen Sie, dass (3.16) die eindeutige Lösung $\tilde{\mathbf{p}}^*$ des quadratischen Optimierungsproblems

$$\min_{\tilde{\mathbf{p}} \in \mathbb{R}^n} \quad (\mathbf{S}\tilde{\mathbf{p}} - \mathbf{y})^{\mathrm{T}}\mathbf{Q}^{-1}(\mathbf{S}\tilde{\mathbf{p}} - \mathbf{y}) \tag{3.18}$$

ist, welche auch als *gewichtete lineare Least-Squares Methode* oder *gewichtete lineare Regression* bezeichnet wird. Berechnen Sie auch die Hessematrix der Kostenfunktion in (3.18) und vergleichen Sie diese mit der Kovarianzmatrix des Parameterschätzfehlers gemäß (3.17).

Beispiel 3.1 (Anfangszustand eines autonomen, zeitdiskreten LTI-Systems). Es soll basierend auf gemessenen Ausgangswerten $\mathbf{y}_k$ mit $k = 0, 1, \ldots, N-1$ ein BLUE Schätzer für den Anfangszustand $\mathbf{x}_0$ des autonomen, zeitdiskreten, linearen, zeitinvarianten Systems

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k \tag{3.19a}$$
$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{v}_k \tag{3.19b}$$

entworfen werden. Die zufälligen Messstörungen $\mathbf{v}_k$ sind durch

$$\mathrm{E}(\mathbf{v}_k) = \mathbf{0} \ , \qquad \mathrm{E}(\mathbf{v}_k \mathbf{v}_j^{\mathrm{T}}) = \mathbf{Q}_k \delta_{kj} \tag{3.20}$$

mit $\mathbf{Q}_k > 0$ charakterisiert.

Für die Zustandsfolge ergibt sich die Lösung

$$\mathbf{x}_k = \mathbf{A}^k \mathbf{x}_0 \ , \tag{3.21}$$

so dass für den Ausgang

$$\mathbf{y}_k = \mathbf{C}\mathbf{A}^k \mathbf{x}_0 + \mathbf{v}_k \tag{3.22}$$

folgt. Werden alle gemessenen Ausgangswerte in einem Vektor assembliert, so ergibt sich

$$\underbrace{\begin{bmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_{N-1} \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} \mathbf{C} \\ \mathbf{C}\mathbf{A} \\ \vdots \\ \mathbf{C}\mathbf{A}^{N-1} \end{bmatrix}}_{\mathbf{S}} \mathbf{x}_0 + \underbrace{\begin{bmatrix} \mathbf{v}_0 \\ \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_{N-1} \end{bmatrix}}_{\mathbf{v}} \ . \tag{3.23}$$

Mit $\mathbf{p} = \mathbf{x}_0$ und der Blockdiagonalmatrix

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_0 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{Q}_{N-1} \end{bmatrix} \tag{3.24}$$

entspricht dies natürlich genau dem linearen Modell (3.1). Der Anfangszustand $\mathbf{x}_0$ kann daher mit dem BLUE Schätzer (3.16) geschätzt werden, wenn $\mathbf{S}$ spaltenregulär ist. Ob dies der Fall ist, hängt auch von N, also der Anzahl der verfügbaren Messungen ab. Einsetzen von $\mathbf{S}$ und $\mathbf{Q}$ in (3.16) und Ausmultiplizieren liefern

$$\hat{\mathbf{x}}_0 = \mathbf{G}^{-1} \sum_{k=0}^{N-1} (\mathbf{A}^{\mathrm{T}})^k \mathbf{C}^{\mathrm{T}} \mathbf{Q}_k^{-1} \mathbf{y}_k \tag{3.25}$$

mit der Abkürzung

$$\mathbf{G} = \sum_{k=0}^{N-1} (\mathbf{A}^{\mathrm{T}})^k \mathbf{C}^{\mathrm{T}} \mathbf{Q}_k^{-1} \mathbf{C}\mathbf{A}^k \ . \tag{3.26}$$

$\mathbf{G}$ wird (insbesondere für $\mathbf{Q}_k = \mathbf{E}$) als *Gramsche Matrix* (observability Gramian) bezeichnet. Ihre Regularität ist eine Voraussetzung für die eindeutige Schätzbarkeit (vollständige Beobachtbarkeit) von $\mathbf{x}_0$. Für die Kovarianzmatrix des Parameterschätzfehlers gilt gemäß (3.17)

$$\mathrm{E}\left((\hat{\mathbf{x}}_0 - \mathbf{x}_0)(\hat{\mathbf{x}}_0 - \mathbf{x}_0)^{\mathrm{T}}\right) = \mathbf{G}^{-1} > 0 \ . \tag{3.27}$$

Beispiel 3.2 (Anfangszustand eines autonomen, zeitkontinuierlichen LTI-Systems). Analog zu Beispiel 3.1 soll nun der zeitkontinuierliche Fall betrachtet werden. Es wird basierend auf gemessenen Ausgangswerten $\mathbf{y}(t)$ mit $t \in [0, T]$ ein BLUE Schätzer für den Anfangszustand $\mathbf{x}_0$ des autonomen, zeitkontinuierlichen, linearen, zeitinvarianten Systems

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) \ , \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{3.28a}$$
$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{v}(t) \tag{3.28b}$$

entworfen. Die zufälligen Messstörungen $\mathbf{v}(t)$ sind durch

$$\mathrm{E}(\mathbf{v}(t)) = \mathbf{0} \ , \qquad \mathrm{E}(\mathbf{v}(t)\mathbf{v}^{\mathrm{T}}(\tau)) = \mathbf{Q}(t)\delta(t-\tau) \tag{3.29}$$

mit $\mathbf{Q}(t) > 0$ und der Dirac Delta-Funktion $\delta(t)$ charakterisiert.

Für die Zustandstrajektorie ergibt sich die Lösung

$$\mathbf{x}(t) = \exp(\mathbf{A}t)\mathbf{x}_0 \ , \tag{3.30}$$

so dass für den Ausgang

$$\mathbf{y}(t) = \mathbf{C}\exp(\mathbf{A}t)\mathbf{x}_0 + \mathbf{v}(t) \tag{3.31}$$

folgt.

Anders als in Beispiel 3.1 können nun nicht alle gemessenen Ausgangswerte $\mathbf{y}(t)$ mit $t \in [0, T]$ in einem gemeinsamen Vektor assembliert werden. Es wird daher analog zum finit-dimensionalen Schätzerentwurf (3.16) ein neuer linearer Schätzer für $\mathbf{x}_0$ entwickelt. Allgemein kann dazu die lineare Abbildung

$$\hat{\mathbf{x}}_0 = \int_0^T \mathbf{K}(t)\mathbf{y}(t)\,\mathrm{d}t \tag{3.32}$$

mit der noch unbekannten Matrix $\mathbf{K}(t)$ angesetzt werden. Damit dieser Schätzer erwartungstreu ist, muss

$$\begin{aligned} \mathrm{E}(\hat{\mathbf{x}}_0) &= \mathrm{E}\left(\int_0^T \mathbf{K}(t)(\mathbf{C}\exp(\mathbf{A}t)\mathbf{x}_0 + \mathbf{v}(t))\,\mathrm{d}t\right) \\ &= \underbrace{\int_0^T \mathbf{K}(t)\mathbf{C}\exp(\mathbf{A}t)\,\mathrm{d}t}_{\mathbf{E}}\,\mathbf{x}_0 = \mathbf{x}_0 \end{aligned} \tag{3.33}$$

erfüllt sein. Für die Kovarianzmatrix des Schätzfehlers $\hat{\mathbf{x}}_0 - \mathbf{x}_0$ gilt unter Berücksich-

tigung von (3.29) und (3.33)

$$\begin{aligned}
&\mathrm{E}((\hat{\mathbf{x}}_0 - \mathbf{x}_0)(\hat{\mathbf{x}}_0 - \mathbf{x}_0)^{\mathrm{T}}) \\
&= \mathrm{E}\left(\left(\int_0^T \mathbf{K}(t)(\mathbf{C}\exp(\mathbf{A}t)\mathbf{x}_0 + \mathbf{v}(t))\,\mathrm{d}t - \mathbf{x}_0\right)\right. \\
&\qquad \left.\left(\int_0^T \mathbf{K}(\tau)(\mathbf{C}\exp(\mathbf{A}\tau)\mathbf{x}_0 + \mathbf{v}(\tau))\,\mathrm{d}\tau - \mathbf{x}_0\right)^{\mathrm{T}}\right) \\
&= \mathrm{E}\left(\int_0^T \mathbf{K}(t)\mathbf{v}(t)\,\mathrm{d}t \int_0^T \mathbf{v}^{\mathrm{T}}(\tau)\mathbf{K}^{\mathrm{T}}(\tau)\,\mathrm{d}\tau\right) \\
&= \mathrm{E}\left(\int_0^T \int_0^T \mathbf{K}(t)\mathbf{v}(t)\mathbf{v}^{\mathrm{T}}(\tau)\mathbf{K}^{\mathrm{T}}(\tau)\,\mathrm{d}\tau\,\mathrm{d}t\right) \\
&= \int_0^T \int_0^T \mathbf{K}(t)\,\mathrm{E}(\mathbf{v}(t)\mathbf{v}^{\mathrm{T}}(\tau))\mathbf{K}^{\mathrm{T}}(\tau)\,\mathrm{d}\tau\,\mathrm{d}t \\
&= \int_0^T \mathbf{K}(t)\mathbf{Q}(t)\mathbf{K}^{\mathrm{T}}(t)\,\mathrm{d}t\ .
\end{aligned} \tag{3.34}$$

Für die aufsummierten Einzelvarianzen der Schätzfehler $\hat{\mathbf{x}}_0 - \mathbf{x}_0$ ergibt sich damit

$$\begin{aligned}
&\mathrm{E}((\hat{\mathbf{x}}_0 - \mathbf{x}_0)^{\mathrm{T}}(\hat{\mathbf{x}}_0 - \mathbf{x}_0)) = \mathrm{E}(\mathrm{spur}((\hat{\mathbf{x}}_0 - \mathbf{x}_0)(\hat{\mathbf{x}}_0 - \mathbf{x}_0)^{\mathrm{T}})) \\
&= \int_0^T \mathrm{spur}(\mathbf{K}(t)\mathbf{Q}(t)\mathbf{K}^{\mathrm{T}}(t))\,\mathrm{d}t = \int_0^T (\mathbf{K}(t)\mathbf{Q}(t)) : \mathbf{K}(t)\,\mathrm{d}t\ .
\end{aligned} \tag{3.35}$$

Der Schätzer soll diesen Wert minimieren. Zur Bestimmung von $\mathbf{K}(t)$ muss daher die dynamische Optimierungsaufgabe

$$\min_{\mathbf{K}(t)} \quad \int_0^T (\mathbf{K}(t)\mathbf{Q}(t)) : \mathbf{K}(t)\,\mathrm{d}t \tag{3.36a}$$

$$\text{u.B.v.} \quad \dot{\mathbf{Z}}(t) = \mathbf{K}(t)\mathbf{C}\exp(\mathbf{A}t) \tag{3.36b}$$

$$\mathbf{Z}(0) = \mathbf{0} \tag{3.36c}$$

$$\mathbf{Z}(T) = \mathbf{E} \tag{3.36d}$$

gelöst werden, wobei die durch (3.33) definierten isoperimetrischen Beschränkungen durch das Randwertproblem (3.36b)-(3.36d) ersetzt wurden. Die zugehörige Hamiltonfunktion lautet

$$H(t, \mathbf{Z}, \mathbf{K}, \mathbf{\Lambda}) = (\mathbf{K}(t)\mathbf{Q}(t)) : \mathbf{K}(t) + (\mathbf{K}(t)\mathbf{C}\exp(\mathbf{A}t)) : \mathbf{\Lambda}(t)\ , \tag{3.37}$$

so dass sich die Optimalitätsbedingungen (erster Ordnung)

$$\dot{\mathbf{Z}}(t) = \frac{\partial H}{\partial \mathbf{\Lambda}} = \mathbf{K}(t)\mathbf{C}\exp(\mathbf{A}t) \tag{3.38a}$$

$$\dot{\mathbf{\Lambda}}(t) = -\frac{\partial H}{\partial \mathbf{Z}} = \mathbf{0} \tag{3.38b}$$

$$\mathbf{0} = \frac{\partial H}{\partial \mathbf{K}} = 2\mathbf{K}(t)\mathbf{Q}(t) + \mathbf{\Lambda}(t)\exp(\mathbf{A}^{\mathrm{T}}t)\mathbf{C}^{\mathrm{T}} \tag{3.38c}$$

ergeben. Daraus folgt

$$\mathbf{\Lambda}(t) = \underset{t}{\text{konst.}} = \mathbf{\Lambda} \tag{3.39a}$$

$$\mathbf{K}(t) = -\frac{1}{2}\mathbf{\Lambda}\exp(\mathbf{A}^{\mathrm{T}}t)\mathbf{C}^{\mathrm{T}}\mathbf{Q}^{-1}(t) \tag{3.39b}$$

Einsetzen von (3.39b) in (3.33) führt auf

$$\mathbf{\Lambda} = -2\left(\int_0^T \exp(\mathbf{A}^{\mathrm{T}}t)\mathbf{C}^{\mathrm{T}}\mathbf{Q}^{-1}(t)\mathbf{C}\exp(\mathbf{A}t)\,\mathrm{d}t\right)^{-1} \tag{3.40}$$

und schließlich den Schätzer

$$\hat{\mathbf{x}}_0 = \mathbf{G}^{-1}\int_0^T \exp(\mathbf{A}^{\mathrm{T}}t)\mathbf{C}^{\mathrm{T}}\mathbf{Q}^{-1}(t)\mathbf{y}(t)\,\mathrm{d}t \tag{3.41}$$

mit der Abkürzung

$$\mathbf{G} = \int_0^T \exp(\mathbf{A}^{\mathrm{T}}t)\mathbf{C}^{\mathrm{T}}\mathbf{Q}^{-1}(t)\mathbf{C}\exp(\mathbf{A}t)\,\mathrm{d}t\ . \tag{3.42}$$

$\mathbf{G}$ ist (insbesondere für $\mathbf{Q}(t) = \mathbf{E}$) auch als *Gramsche Matrix* (observability Gramian) bekannt. Ihre Regularität ist eine Voraussetzung für die eindeutige Schätzbarkeit (vollständige Beobachtbarkeit) von $\mathbf{x}_0$. Einsetzen von (3.39b) und (3.40) in (3.34) liefert für die Kovarianzmatrix des Schätzfehlers

$$\mathrm{E}\left((\hat{\mathbf{x}}_0 - \mathbf{x}_0)(\hat{\mathbf{x}}_0 - \mathbf{x}_0)^{\mathrm{T}}\right) = \mathbf{G}^{-1} > 0\ . \tag{3.43}$$

3.1.2 Der Fall einer nicht spaltenregulären Datenmatrix

Ist im Modell (3.1) die Datenmatrix $\mathbf{S} \in \mathbb{R}^{m\times n}$ nicht spaltenregulär, so existiert für die unbekannten Parameter $\mathbf{p}$ kein linearer erwartungstreuer Schätzer, da die Bedingung (3.6) nicht erfüllbar ist. In diesem Fall liegt eine *Redundanz der Parameter* $\mathbf{p}$ im Modellansatz (3.1) vor. Diese Redundanz äußert sich dadurch, dass unterschiedliche Parameterwerte $\mathbf{p}_1 \neq \mathbf{p}_2$ zu gleichen Erwartungswerten $\mathbf{S}\mathbf{p}_1 = \mathbf{S}\mathbf{p}_2$ des Systemausgangs $\mathbf{y}$ führen können. Es gilt dann

$$\mathbf{S}(\mathbf{p}_1 - \mathbf{p}_2) = \mathbf{0} \tag{3.44}$$

und $\mathbf{p}_1$ und $\mathbf{p}_2$ sind am Ausgang nicht unterscheidbar. Eine Möglichkeit dieses Problem zu umgehen ist eine geänderte Wahl des Modells (3.1), so dass die Datenmatrix $\mathbf{S}$ spaltenregulär ist. Eine zweite Möglichkeit mit dem Problem umzugehen besteht darin, nicht die Werte von $\mathbf{p}$ selbst, sondern die Werte einer linearen Abbildung $\mathbf{Tp}$ von $\mathbf{p}$ mit $\mathbf{T} \in \mathbb{R}^{o\times n}$ zu schätzen. Gemäß (3.44) sind alle Parameterwerte im Kern($\mathbf{S}$) der Matrix $\mathbf{S}$ nicht unterscheidbar und daher nicht erwartungstreu schätzbar. Da $\mathbb{R}^n = \text{Kern}(\mathbf{S}) \oplus \text{Bild}(\mathbf{S}^\mathrm{T})$ gilt, sind im Umkehrschluss Parameterwerte genau dann erwartungstreu schätzbar, wenn sie im Bild($\mathbf{S}^\mathrm{T}$) liegen. Folglich stellen die Einschränkungen

$$\text{Kern}(\mathbf{S}) \subseteq \text{Kern}(\mathbf{T}) \tag{3.45a}$$

$$\text{Bild}(\mathbf{T}^\mathrm{T}) \subseteq \text{Bild}(\mathbf{S}^\mathrm{T}) \tag{3.45b}$$

für die Wahl von $\mathbf{T}$ die erwartungstreue Schätzbarkeit der Werte von $\mathbf{Tp}$ sicher. Im nächsten Schritt soll daher ein linearer Schätzer

$$\widehat{\mathbf{Tp}} = \mathbf{Ky} \tag{3.46}$$

für die Werte von $\mathbf{Tp}$ mit gegebenem $\mathbf{T}$ gemäß (3.45) entworfen werden, der erwartungstreu ist und minimale Varianz des Schätzfehlers erreicht. Für Erwartungstreue muss wegen $\mathrm{E}(\widehat{\mathbf{Tp}}) = \mathbf{KSp}$ die Bedingung

$$\mathbf{KS} = \mathbf{T} \tag{3.47}$$

erfüllt sein. Sie kann, wie es sein muss, nur erfüllt werden, wenn (3.45b) gilt. Unter Berücksichtigung von (3.47) gilt für die Varianz des Schätzfehlers (Summe der Einzelvarianzen)

$$\mathrm{E}\left(\left\|\widehat{\mathbf{Tp}} - \mathbf{Tp}\right\|_2^2\right) = \text{spur}\left(\mathbf{KQK}^\mathrm{T}\right) = (\mathbf{KQ}) : \mathbf{K} \ . \tag{3.48}$$

Zur Bestimmung von $\mathbf{K}$ ist daher die Optimierungsaufgabe

$$\min_{\mathbf{K}\in\mathbb{R}^{o\times m}} \quad (\mathbf{KQ}) : \mathbf{K} \tag{3.49a}$$

$$\text{u.B.v.} \quad \mathbf{KS} - \mathbf{T} = \mathbf{0} \tag{3.49b}$$

zu lösen. Es ist zu beachten, dass hier die Nebenbedingungen (3.49b) wegen der fehlenden Spaltenregularität von $\mathbf{S}$ nicht funktional unabhängig sind, d. h. die LICQ Bedingung ist nicht erfüllt. Folglich können Lagrange-Multiplikatoren aus den KKT-Bedingungen zwar berechnet werden, aber sie sind nicht eindeutig. Abgesehen davon kann die optimale Lösung von (3.49) analog zu Abschnitt 3.1.1 berechnet werden und lautet

$$\mathbf{K} = \mathbf{T}(\mathbf{S}^\mathrm{T}\mathbf{Q}^{-1}\mathbf{S})^\dagger\mathbf{S}^\mathrm{T}\mathbf{Q}^{-1} \ , \tag{3.50}$$

wobei $(\,\cdot\,)^\dagger$ die *Pseudoinverse* einer Matrix liefert[1]. Der optimale lineare erwartungstreue Schätzer lautet also

[1] Die *Pseudoinverse* (Moore-Penrose-Inverse) $\mathbf{A}^\dagger \in \mathbb{R}^{n\times m}$ einer Matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ ist durch die Bedingungen

$$\mathbf{AA}^\dagger\mathbf{A} = \mathbf{A} \ , \qquad \mathbf{A}^\dagger\mathbf{AA}^\dagger = \mathbf{A}^\dagger \ , \qquad (\mathbf{AA}^\dagger)^\mathrm{T} = \mathbf{AA}^\dagger \ , \qquad (\mathbf{A}^\dagger\mathbf{A})^\mathrm{T} = \mathbf{A}^\dagger\mathbf{A}$$

definiert. Die Pseudoinverse kann mit Hilfe der Singulärwertzerlegung berechnet werden. Für $\mathbf{A} \in \mathbb{R}^{m\times n}$ mit $m \geq n$ lautet die Singulärwertzerlegung

$$\mathbf{A} = \mathbf{U\Sigma V}^\mathrm{T}$$

$$\widehat{\mathbf{Tp}} = \mathbf{T}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{\dagger}\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{y} \tag{3.51}$$

und für die Kovarianzmatrix des Schätzfehlers gilt

$$\mathrm{E}\left((\widehat{\mathbf{Tp}} - \mathbf{Tp})(\widehat{\mathbf{Tp}} - \mathbf{Tp})^{\mathrm{T}}\right) = \mathbf{K}\mathbf{Q}\mathbf{K}^{\mathrm{T}} = \mathbf{T}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{\dagger}\mathbf{T}^{\mathrm{T}} \ . \tag{3.52}$$

Sie ist für zeilenreguläres $\mathbf{T}$ positiv definit.

Aufgabe 3.2. Zeigen Sie, dass der Schätzer (3.51) erwartungstreu ist und unter allen linearen erwartungstreuen Schätzern die minimale Varianz des Schätzfehlers erreicht.

Aufgabe 3.3 (Gewichtete lineare Least-Squares Methode bei nicht spaltenregulärer Datenmatrix). Zeigen Sie, dass (3.51) dem Ausdruck

$$\mathbf{T}\,\arg\min_{\tilde{\mathbf{p}}\in\mathbb{R}^n} (\mathbf{S}\tilde{\mathbf{p}} - \mathbf{y})^{\mathrm{T}}\mathbf{Q}^{-1}(\mathbf{S}\tilde{\mathbf{p}} - \mathbf{y}) \tag{3.53}$$

entspricht, also dass $\widehat{\mathbf{Tp}} = \mathbf{T}\tilde{\mathbf{p}}^*$ gilt, wobei $\tilde{\mathbf{p}}^*$ die Lösung des quadratischen Optimierungsproblems in (3.53) ist. Untersuchen Sie auch die Eindeutigkeit dieser Lösung. Berechnen Sie ferner die Hessematrix zu diesem Optimierungsproblem und vergleichen Sie diese mit der Kovarianzmatrix des Schätzfehlers gemäß (3.52).

Lösung von Aufgabe 3.3. Das Optimierungsproblem in (3.53) kann in die Form

$$\min_{\tilde{\mathbf{p}}\in\mathbb{R}^n} \quad \tilde{\mathbf{p}}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}\tilde{\mathbf{p}} - 2\mathbf{y}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}\tilde{\mathbf{p}} + \mathbf{y}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{y} \tag{3.54}$$

umgeschrieben werden. Die zugehörige Hessematrix lautet daher $\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}$. Sie ist positiv semidefinit, da $\mathbf{S}$ nicht spaltenregulär ist. Folglich kann die Lösung von (3.54) nicht eindeutig sein. Optimale Lösungen existieren aber sicher, da

$$\mathrm{Kern}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}) \subseteq \mathrm{Kern}(\mathbf{y}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}) \tag{3.55}$$

und folglich

$$2\mathbf{y}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}\tilde{\mathbf{p}} > 0 \quad \Rightarrow \quad \tilde{\mathbf{p}}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}\tilde{\mathbf{p}} > 0 \tag{3.56}$$

---

mit der $m \times n$ Matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \mathrm{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_n\} \\ \mathbf{0} \end{bmatrix}$$

und orthogonalen Matrizen $\mathbf{U} \in \mathbb{R}^{m\times m}$ und $\mathbf{V} \in \mathbb{R}^{n\times n}$. Die Pseudoinverse folgt dann in der Form

$$\mathbf{A}^{\dagger} = \mathbf{V}\mathbf{\Sigma}^{\dagger}\mathbf{U}^{\mathrm{T}}$$

mit der $n \times m$ Matrix

$$\mathbf{\Sigma}^{\dagger} = \begin{bmatrix} \mathrm{diag}\{\sigma_1^{\dagger}, \sigma_2^{\dagger}, \ldots, \sigma_n^{\dagger}\} & \mathbf{0} \end{bmatrix}$$

und

$$\sigma_i^{\dagger} = \begin{cases} 0 & \text{für } \sigma_i = 0 \\ \sigma_i^{-1} & \text{sonst .} \end{cases}$$

Analoges gilt für den Fall $m \leq n$.

gilt. Die notwendige Optimalitätsbedingung erster Ordnung für (3.54) lautet

$$2\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}\tilde{\mathbf{p}}^{*} - 2\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{y} = \mathbf{0}\ . \tag{3.57}$$

Sie ist auch hinreichend für ein Optimum, da die Optimierungsaufgabe (3.54) konvex ist. Es gelten nun die Beziehungen

$$\mathrm{Kern}(\mathbf{S}) = \mathrm{Kern}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}) \tag{3.58a}$$

$$\mathrm{Bild}(\mathbf{S}^{\mathrm{T}}) = \mathrm{Bild}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}) = \mathrm{Bild}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})\ . \tag{3.58b}$$

Aufgabe 3.4. Zeigen Sie die Gültigkeit dieser Beziehungen.

Lösung von Aufgabe 3.4. Gleichung (3.58a) und $\mathrm{Bild}(\mathbf{S}^{\mathrm{T}}) = \mathrm{Bild}\,\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1})$ folgen aus der Regularität von $\mathbf{Q}^{-1}$. Wegen (3.58a) und $\mathbb{R}^n = \mathrm{Kern}(\mathbf{S}) \oplus \mathrm{Bild}\,\mathbf{S}^{\mathrm{T}}) = \mathrm{Kern}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}) \oplus \mathrm{Bild}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})$ gilt $\mathrm{Bild}(\mathbf{S}^{\mathrm{T}}) = \mathrm{Bild}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})$.

Wegen (3.58b) ist die Gleichung (3.57) (unabhängig von $\mathbf{y}$) konsistent, d. h. es existieren stets Lösungen $\tilde{\mathbf{p}}^{*}$. Jene Anteile von $\tilde{\mathbf{p}}^{*}$ die im $\mathrm{Kern}(\mathbf{S})$ liegen haben keine Auswirkungen auf (3.54) und (3.57). Dies hat zwei Konsequenzen: a) Eine Lösung $\tilde{\mathbf{p}}^{*}$ des quadratischen Optimierungsproblems in (3.53) kann nicht eindeutig sein. b) Von Interesse ist insbesondere jener Anteil von $\tilde{\mathbf{p}}^{*}$ der nicht im $\mathrm{Kern}(\mathbf{S})$ liegt.

$(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{\dagger}\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}$ ist eine orthogonale Projektionsmatrix, welche jeden Vektor des $\mathbb{R}^n$ in den Zeilenraum der Matrix $\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}$ projiziert. Dieser entspricht wegen der Symmetrie von $\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}$ auch ihrem Spaltenraum und damit dem $\mathrm{Bild}(\mathbf{S}^{\mathrm{T}})$. Folglich ist $\mathbf{E} - (\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{\dagger}\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}$ eine orthogonale Projektionsmatrix, welche jeden Vektor des $\mathbb{R}^n$ in den Orthogonalraum von $\mathrm{Bild}(\mathbf{S}^{\mathrm{T}})$, also in den $\mathrm{Kern}(\mathbf{S})$ projiziert. Mit diesen Projektionsmatrizen kann jede Lösung $\tilde{\mathbf{p}}^{*}$ von (3.57) wie folgt orthogonal zerlegt werden.

$$\tilde{\mathbf{p}}^{*} = \underbrace{(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{\dagger}\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}\tilde{\mathbf{p}}^{*}}_{\tilde{\mathbf{p}}_1^{*}} + \underbrace{(\mathbf{E} - (\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{\dagger}\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})\tilde{\mathbf{p}}^{*}}_{\tilde{\mathbf{p}}_0^{*}} \tag{3.59}$$

Es gilt dann $\tilde{\mathbf{p}}_1^{*} \in \mathrm{Bild}(\mathbf{S}^{\mathrm{T}})$ und $\tilde{\mathbf{p}}_0^{*} \in \mathrm{Kern}(\mathbf{S}) \subseteq \mathrm{Kern}(\mathbf{T})$. Natürlich ist $\tilde{\mathbf{p}}_1^{*}$ eine partikuläre Lösung von (3.57) und $\tilde{\mathbf{p}}_0^{*}$ eine homogene Lösung. Der Wert $\tilde{\mathbf{p}}_1^{*}$ ist eindeutig, da die Projektion der Optimierungsaufgabe (3.54) in das $\mathrm{Bild}(\mathbf{S}^{\mathrm{T}})$ strikt konvex ist. Folglich ist auch der Schätzwert $\widehat{\mathbf{Tp}} = \mathbf{T}\tilde{\mathbf{p}}^{*} = \mathbf{T}\tilde{\mathbf{p}}_1^{*}$ eindeutig.

Aufgabe 3.5. Zeigen Sie, dass für jedes $\tilde{\mathbf{p}}^{*}$ welches (3.57) erfüllt, der Schätzwert $\mathbf{T}\tilde{\mathbf{p}}^{*}$ erwartungstreu ist und für die Kovarianzmatrix seines Schätzfehlers

$$\mathrm{E}\left((\mathbf{T}\tilde{\mathbf{p}}^{*} - \mathbf{Tp})(\mathbf{T}\tilde{\mathbf{p}}^{*} - \mathbf{Tp})^{\mathrm{T}}\right) = \mathbf{T}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{\dagger}\mathbf{T}^{\mathrm{T}} \tag{3.60}$$

gilt.

Diese Kovarianzmatrix enthält also die Pseudoinverse der Hessematrix $\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}$ aus der Optimierungsaufgabe (3.54).

Eine mögliche Lösung von (3.57) lautet

$$\tilde{\mathbf{p}}^* = (\mathbf{S}^\mathrm{T}\mathbf{Q}^{-1}\mathbf{S})^\dagger \mathbf{S}^\mathrm{T}\mathbf{Q}^{-1}\mathbf{y} \; . \tag{3.61}$$

Um dies zu zeigen, wird der Ausdruck (3.61) in (3.57) eingesetzt, was

$$2\mathbf{S}^\mathrm{T}\mathbf{Q}^{-1}\mathbf{S}(\mathbf{S}^\mathrm{T}\mathbf{Q}^{-1}\mathbf{S})^\dagger \mathbf{S}^\mathrm{T}\mathbf{Q}^{-1}\mathbf{y} - 2\mathbf{S}^\mathrm{T}\mathbf{Q}^{-1}\mathbf{y} = \mathbf{0} \tag{3.62}$$

liefert. Die hier auftretende orthogonale Projektionsmatrix $\mathbf{S}^\mathrm{T}\mathbf{Q}^{-1}\mathbf{S}(\mathbf{S}^\mathrm{T}\mathbf{Q}^{-1}\mathbf{S})^\dagger$ projiziert jeden Vektor des $\mathbb{R}^n$ in den Spaltenraum der Matrix $\mathbf{S}^\mathrm{T}\mathbf{Q}^{-1}\mathbf{S}$, welcher dem Bild($\mathbf{S}^\mathrm{T}$) entspricht. Aus $\mathbf{S}^\mathrm{T}\mathbf{Q}^{-1}\mathbf{y} \in$ Bild($\mathbf{S}^\mathrm{T}$) folgt daher direkt die Gültigkeit der Gleichung (3.62).

Unterzieht man den Ausdruck (3.61) einer orthogonalen Zerlegung gemäß (3.59), so zeigt sich noch

$$\tilde{\mathbf{p}}_1^* = (\mathbf{S}^\mathrm{T}\mathbf{Q}^{-1}\mathbf{S})^\dagger \mathbf{S}^\mathrm{T}\mathbf{Q}^{-1}\mathbf{S}\tilde{\mathbf{p}}^* = (\mathbf{S}^\mathrm{T}\mathbf{Q}^{-1}\mathbf{S})^\dagger \mathbf{S}^\mathrm{T}\mathbf{Q}^{-1}\mathbf{y} = \tilde{\mathbf{p}}^* \tag{3.63a}$$

$$\tilde{\mathbf{p}}_0^* = (\mathbf{E} - (\mathbf{S}^\mathrm{T}\mathbf{Q}^{-1}\mathbf{S})^\dagger \mathbf{S}^\mathrm{T}\mathbf{Q}^{-1}\mathbf{S})\tilde{\mathbf{p}}^* = \mathbf{0} \; . \tag{3.63b}$$

Das heißt, für $\tilde{\mathbf{p}}^*$ gemäß (3.61) gilt bereits $\tilde{\mathbf{p}}^* \in$ Bild($\mathbf{S}^\mathrm{T}$).

3.1.3 Der singuläre Fall

In diesem Abschnitt wird im Modell (3.1) die Voraussetzung einer positiv definiten Kovarianzmatrix $\mathrm{E}(\mathbf{v}\mathbf{v}^\mathrm{T}) = \mathbf{Q}$ aufgegeben. $\mathbf{Q}$ muss daher nur noch positiv semidefinit sein. Ihr Rang wird mit $q = \mathrm{rang}(\mathbf{Q})$ bezeichnet und es gilt $q < m$. Aus dem nachfolgenden Lemma 3.1 folgt, dass $\mathbf{v}$ in diesem Fall auf einen Unterraum des $\mathbb{R}^m$ beschränkt ist.

Lemma 3.1 (Raum der stochastischen Störung). *Für eine stochastische Störung* $\mathbf{v}$ *mit Erwartungswert* $\mathrm{E}(\mathbf{v}) = \mathbf{0}$ *und Kovarianzmatrix* $\mathrm{E}(\mathbf{v}\mathbf{v}^\mathrm{T}) = \mathbf{Q}$ *gilt*

$$\mathbf{v} \in \mathrm{Bild}(\mathbf{Q}) \; . \tag{3.64}$$

Beweis. Die Störung $\mathbf{v}$ wird in der Form

$$\mathbf{v} = \underbrace{\mathbf{Q}\mathbf{Q}^\dagger\mathbf{v}}_{\mathbf{v}_1} + \underbrace{(\mathbf{E} - \mathbf{Q}\mathbf{Q}^\dagger)\mathbf{v}}_{\mathbf{v}_2} \tag{3.65}$$

orthogonal zerlegt. Die orthogonale Projektionsmatrix $\mathbf{Q}\mathbf{Q}^\dagger$ projiziert jeden Vektor des $\mathbb{R}^m$ in den Spaltenraum von $\mathbf{Q}$. Offensichtlich gilt

$$\mathrm{E}(\mathbf{v}_2) = \mathbf{0} \; . \tag{3.66}$$

Außerdem ergibt sich

$$\begin{aligned} \mathrm{E}(\mathbf{v}_2\mathbf{v}_2^{\mathrm{T}}) &= \mathrm{E}((\mathbf{E}-\mathbf{Q}\mathbf{Q}^{\dagger})\mathbf{v}\mathbf{v}^{\mathrm{T}}(\mathbf{E}-\mathbf{Q}\mathbf{Q}^{\dagger})^{\mathrm{T}}) \\ &= (\mathbf{E}-\mathbf{Q}\mathbf{Q}^{\dagger})\mathbf{Q}(\mathbf{E}-\mathbf{Q}\mathbf{Q}^{\dagger})^{\mathrm{T}} \\ &= (\mathbf{Q}-\mathbf{Q}\mathbf{Q}^{\dagger}\mathbf{Q})(\mathbf{E}-\mathbf{Q}\mathbf{Q}^{\dagger})^{\mathrm{T}} = \mathbf{0}\ . \end{aligned} \tag{3.67}$$

Daraus folgt $\mathbf{v}_2 = \mathbf{0}$ und (3.65) reduziert sich zu $\mathbf{v} = \mathbf{Q}\mathbf{Q}^{\dagger}\mathbf{v}$. Dies beweist die Beziehung (3.64). □

Weiterhin wird keine Spaltenregularität der Datenmatrix $\mathbf{S}$ im Modell (3.1) gefordert. Dementsprechend gilt analog zu Abschnitt 3.1.2, dass $\mathbf{p}$ nicht erwartungstreu geschätzt werden kann, sondern nur die Werte von $\mathbf{T}\mathbf{p}$ mit gegebenem $\mathbf{T} \in \mathbb{R}^{o\times n}$ welches (3.45) erfüllt. Es soll daher wieder ein Schätzer $\widehat{\mathbf{T}\mathbf{p}}$ für $\mathbf{T}\mathbf{p}$ entworfen werden.

Ursachen für eine singuläre Kovarianzmatrix

Eine Singularität von $\mathbf{Q}$ kann beispielsweise in folgenden Fällen auftreten.

- Es existiert eine konstante Matrix $\mathbf{B} \neq \mathbf{0}$ so, dass jede Realisierung der stochastischen Störung $\mathbf{v}$
$$\mathbf{0} = \mathbf{B}\mathbf{v} \tag{3.68}$$
erfüllt. Praktisch kann dies folgende Ursachen haben:
  - Einzelne Komponenten der stochastischen Störung $\mathbf{v}$ sind streng Null, was beispielsweise bei störungsfreien (exakten) Messungen zutrifft.
  - Einzelne Komponenten der stochastischen Störung $\mathbf{v}$ weisen eine lineare Abhängigkeit auf, was beispielsweise bei Messungen mit perfekt korrelierten Störungen zutrifft.

  Wegen (3.64) lautet eine Möglichkeit zur Berechnung einer Matrix $\mathbf{B}$ bei bekanntem $\mathbf{Q}$ (siehe [3.3])
$$\mathbf{B} = \mathbf{E} - \mathbf{Q}\mathbf{Q}^{\dagger}\ . \tag{3.69}$$

- Wenn die Systemparameter $\mathbf{p}$ eine lineare Gleichung der Art
$$\mathbf{a} = \mathbf{A}\mathbf{p} \tag{3.70}$$
mit konstanten Ausdrücken $\mathbf{a}$ und $\mathbf{A}$ erfüllen müssen, so kann dies berücksichtigt werden, indem ein erweitertes Modell der Form
$$\underbrace{\begin{bmatrix}\mathbf{y}\\ \mathbf{a}\end{bmatrix}}_{\tilde{\mathbf{y}}} = \underbrace{\begin{bmatrix}\mathbf{S}\\ \mathbf{A}\end{bmatrix}}_{\tilde{\mathbf{S}}}\mathbf{p} + \underbrace{\begin{bmatrix}\mathbf{v}\\ \mathbf{0}\end{bmatrix}}_{\tilde{\mathbf{v}}} \tag{3.71}$$
statt (3.1) verwendet wird. Die Kovarianzmatrix $\mathrm{E}(\tilde{\mathbf{v}}\tilde{\mathbf{v}}^{\mathrm{T}}) = \begin{bmatrix}\mathbf{Q} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$ ist singulär.

- Wenn der Systemausgang $\mathbf{y}$ eine lineare Gleichung der Art
$$\mathbf{a} = \mathbf{A}\mathbf{y} \tag{3.72}$$

mit konstanten Ausdrücken $\mathbf{a}$ und $\mathbf{A}$ erfüllen muss, so gilt nach Einsetzen von (3.1)

$$\mathbf{a} = \mathbf{A}(\mathbf{S}\mathbf{p} + \mathbf{v}) \ . \tag{3.73}$$

Wird der Erwartungswert dieser Gleichung berechnet und $\mathrm{E}(\mathbf{v}) = \mathbf{0}$ berücksichtigt, so führt dies auf die Bedingung

$$\mathbf{a} = \mathbf{A}\mathbf{S}\mathbf{p} \ , \tag{3.74}$$

welche äquivalent zu (3.70) behandelt werden kann. Wird (3.74) von (3.73) abgezogen, so ergibt sich die weitere Bedingung

$$\mathbf{0} = \mathbf{A}\mathbf{v} \ , \tag{3.75}$$

welche äquivalent zu (3.68) interpretiert werden kann.

Die genannten Fälle können auch näherungsweise auftreten. Dann ist $\mathbf{Q}$ zwar nicht streng singulär aber möglicherweise numerisch schlecht konditioniert (großes Verhältnis aus maximalem zu minimalem Eigenwert von $\mathbf{Q}$, d. h. $\lambda_{\max}/\lambda_{\min} \gg 1$). Auch dann kann es sinnvoll sein, einen der nachfolgend beschriebenen Schätzer zu verwenden [3.6].

Definition 3.1 (Konsistenz der Messwerte)**.** Die Ausgangswerte (Messwerte) $\mathbf{y}$ werden als konsistent mit dem Modell (3.1) bezeichnet, wenn sie

$$\mathbf{y} \in \mathrm{Bild}\Big(\begin{bmatrix}\mathbf{S} & \mathbf{Q}\end{bmatrix}\Big) \tag{3.76}$$

erfüllen.

In den Abschnitten 3.1.1 und 3.1.2 war die Bedingung (3.76) wegen der Regularität von $\mathbf{Q}$ trivial erfüllt, d. h. beliebige Ausgangswerte $\mathbf{y}$ waren konsistent. Treten im aktuellen Abschnitt Ausgangswerte $\mathbf{y}$ auf die (3.76) (signifikant) verletzen, also nicht konsistent sind, so sollte das Modell (3.1) verbessert werden. Dies kann z. B. durch eine andere Wahl von $\mathbf{S}$ oder $\mathbf{Q}$ erfolgen.

Schätzerentwurf

Es soll ein linearer Schätzer der Form

$$\widehat{\mathbf{T}\mathbf{p}} = \mathbf{K}\mathbf{y} \tag{3.77}$$

für $\mathbf{T}\mathbf{p}$ entworfen werden, der erwartungstreu ist und minimale Varianz des Schätzfehlers sicherstellt. Für Erwartungstreue muss wegen $\mathrm{E}(\widehat{\mathbf{T}\mathbf{p}}) = \mathbf{K}\mathbf{S}\mathbf{p}$ wieder die Bedingung (3.47) erfüllt sein. Auch in diesem Fall gilt für die Varianz des Schätzfehlers (3.48). Zur Bestimmung von $\mathbf{K}$ ist daher wieder die Optimierungsaufgabe

$$\min_{\mathbf{K}\in\mathbb{R}^{o\times m}} \quad (\mathbf{K}\mathbf{Q}) : \mathbf{K} \tag{3.78a}$$

$$\text{u.B.v.} \quad \mathbf{K}\mathbf{S} - \mathbf{T} = \mathbf{0} \tag{3.78b}$$

zu lösen. Die zugehörigen KKT-Bedingungen (erster Ordnung) lauten

$$2\mathbf{K}\mathbf{Q} + \mathbf{\Lambda}\mathbf{S}^{\mathrm{T}} = \mathbf{0} \tag{3.79a}$$

$$\mathbf{K}\mathbf{S} - \mathbf{T} = \mathbf{0} \ , \tag{3.79b}$$

bzw. in zusammengefasster Schreibweise

$$\begin{bmatrix}\mathbf{K} & \mathbf{\Lambda}\end{bmatrix}\begin{bmatrix}2\mathbf{Q} & \mathbf{S}\\ \mathbf{S}^{\mathrm{T}} & \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{0} & \mathbf{T}\end{bmatrix} . \tag{3.80}$$

Es existieren Lösungen für $\begin{bmatrix}\mathbf{K} & \mathbf{\Lambda}\end{bmatrix}$ sofern (3.45) erfüllt ist. Dann gilt

$$\mathrm{Bild}\left(\begin{bmatrix}\mathbf{0} & \mathbf{T}\end{bmatrix}^{\mathrm{T}}\right) \subseteq \mathrm{Bild}\left(\begin{bmatrix}2\mathbf{Q} & \mathbf{S}\\ \mathbf{S}^{\mathrm{T}} & \mathbf{0}\end{bmatrix}\right) . \tag{3.81}$$

Wegen der Singularität von $\mathbf{Q}$ kann $\mathbf{K}$ im Allgemeinen nicht eindeutig berechnet werden. Eine mögliche Lösung von (3.80) lautet

$$\begin{bmatrix}\mathbf{K} & \mathbf{\Lambda}\end{bmatrix} = \begin{bmatrix}\mathbf{0} & \mathbf{T}\end{bmatrix}\begin{bmatrix}2\mathbf{Q} & \mathbf{S}\\ \mathbf{S}^{\mathrm{T}} & \mathbf{0}\end{bmatrix}^{\dagger} . \tag{3.82}$$

Aufgabe 3.6. Zeigen Sie, dass (3.82) die Gleichung (3.80) erfüllt. Berücksichtigen Sie dabei (3.81).

Jede Matrix $\mathbf{K}$, welche (3.80) erfüllt, ist optimal im Sinne der Optimierungsaufgabe (3.78). Zum Nachweis dieser Aussage muss gezeigt werden, dass keine andere Matrix $\tilde{\mathbf{K}}$, die $\tilde{\mathbf{K}} \neq \mathbf{K}$ und die Nebenbedingung (3.78b) erfüllt, zu einem besseren Kostenfunktionswert (3.78a) führt. Dieser Nachweis folgt unter Beachtung der Identitäten $\mathrm{spur}\,(\mathbf{KQK}^{\mathrm{T}}) = (\mathbf{KQ}) : \mathbf{K}$, (3.79a), $\tilde{\mathbf{K}}\mathbf{S} = \mathbf{T} = \mathbf{KS}$, $\mathbf{Q} \geq 0$ und $\tilde{\mathbf{K}} \neq \mathbf{K}$ aus der Ungleichung

$$\begin{aligned}
0 &\leq \mathrm{spur}\,((\tilde{\mathbf{K}} - \mathbf{K})\mathbf{Q}(\tilde{\mathbf{K}} - \mathbf{K})^{\mathrm{T}})\\
&= \mathrm{spur}\,(\tilde{\mathbf{K}}\mathbf{Q}\tilde{\mathbf{K}}^{\mathrm{T}}) - 2\,\mathrm{spur}\,(\mathbf{KQ}\tilde{\mathbf{K}}^{\mathrm{T}}) + \mathrm{spur}\,(\mathbf{KQK}^{\mathrm{T}})\\
&= \mathrm{spur}\,(\tilde{\mathbf{K}}\mathbf{Q}\tilde{\mathbf{K}}^{\mathrm{T}}) + \mathrm{spur}\,(\mathbf{\Lambda}\mathbf{S}^{\mathrm{T}}\tilde{\mathbf{K}}^{\mathrm{T}}) + \mathrm{spur}\,(\mathbf{KQK}^{\mathrm{T}})\\
&= \mathrm{spur}\,(\tilde{\mathbf{K}}\mathbf{Q}\tilde{\mathbf{K}}^{\mathrm{T}}) + \mathrm{spur}\,(\mathbf{\Lambda}\mathbf{T}^{\mathrm{T}}) + \mathrm{spur}\,(\mathbf{KQK}^{\mathrm{T}})\\
&= \mathrm{spur}\,(\tilde{\mathbf{K}}\mathbf{Q}\tilde{\mathbf{K}}^{\mathrm{T}}) + \mathrm{spur}\,(\mathbf{\Lambda}\mathbf{S}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}) + \mathrm{spur}\,(\mathbf{KQK}^{\mathrm{T}})\\
&= \mathrm{spur}\,(\tilde{\mathbf{K}}\mathbf{Q}\tilde{\mathbf{K}}^{\mathrm{T}}) - 2\,\mathrm{spur}\,(\mathbf{KQK}^{\mathrm{T}}) + \mathrm{spur}\,(\mathbf{KQK}^{\mathrm{T}})\\
&= \mathrm{spur}\,(\tilde{\mathbf{K}}\mathbf{Q}\tilde{\mathbf{K}}^{\mathrm{T}}) - \mathrm{spur}\,(\mathbf{KQK}^{\mathrm{T}}) .
\end{aligned} \tag{3.83}$$

Wegen der Singularität von $\mathbf{Q}$ bzw. der im Allgemeinen nicht vorhandenen Eindeutigkeit von $\mathbf{K}$ kann $\mathbf{K}$ gemäß (3.80) nicht strikt optimal sein.

Mit einem erwartungstreuen Schätzer (3.77) lautet die Kovarianzmatrix des Parameterschätzfehlers

$$\mathrm{E}\,((\widehat{\mathbf{Tp}} - \mathbf{Tp})(\widehat{\mathbf{Tp}} - \mathbf{Tp})^{\mathrm{T}}) = \mathbf{KQK}^{\mathrm{T}} \geq 0 . \tag{3.84}$$

Im Falle einer regulären Kovarianzmatrix $\mathbf{Q}$ konnte in den Aufgaben 3.1 und 3.3 mit der gewichteten linearen Least-Squares Methode jeweils eine unbeschränkte Optimierungsaufgabe formuliert werden, deren Lösung zum exakt gleichen optimalen linearen erwartungstreuen Schätzer für $\mathbf{p}$ bzw. $\mathbf{Tp}$ führt. Als Gewichtungsmatrix in der Optimierungsaufgabe wurde jeweils $\mathbf{Q}^{-1}$ verwendet. Im aktuellen Fall einer singulären Kovarianzmatrix $\mathbf{Q}$ kann keine solche unbeschränkte Optimierungsaufgabe formuliert werden.

Transformation des Modells

Als Alternative zum obigen Schätzerentwurf wird eine Transformation des Modells (3.1) vorgestellt, die ein reduziertes Modell mit regulärer Kovarianzmatrix der Störung liefert. Dies führt im Allgemeinen nicht auf einen Schätzer mit $\mathbf{K}$ gemäß (3.82).

Im Falle einer singulären Kovarianzmatrix $\mathbf{Q}$ enthält der Vektor $\mathbf{v}$ der stochastischen Störung redundante Information. Um diese Redundanz noch klarer darzustellen und schließlich zu beseitigen, wird das Modell (3.1) in einen stochastischen und einen deterministischen Teil transformiert. Man wählt dazu eine reguläre Matrix $\begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix} \in \mathbb{R}^{m\times m}$ so, dass

$$\mathbf{V}_1^{\mathrm{T}} \in \mathrm{Bild}(\mathbf{Q}) \tag{3.85a}$$
$$\mathbf{V}_2^{\mathrm{T}} \in \mathrm{Kern}(\mathbf{Q}) \tag{3.85b}$$

gilt. Eine solche Matrix kann z. B. erstellt werden, indem ihre Zeilen die aus der Singulärwertzerlegung von $\mathbf{Q} = \mathbf{U}\,\mathrm{diag}\{\sigma_1, \sigma_2, \ldots\}\mathbf{U}^{\mathrm{T}}$ folgenden Singulärvektoren enthalten, d. h. $\begin{bmatrix}\mathbf{V}_1^{\mathrm{T}} & \mathbf{V}_2^{\mathrm{T}}\end{bmatrix} = \mathbf{U}$. Wird das Modell (3.1) linksseitig einmal mit $\mathbf{V}_1$ und einmal mit $\mathbf{V}_2$ multipliziert, so ergibt sich das transformierte Modell

$$\mathbf{V}_1\mathbf{y} = \mathbf{V}_1\mathbf{S}\mathbf{p} + \mathbf{V}_1\mathbf{v} \tag{3.86a}$$
$$\mathbf{V}_2\mathbf{y} = \mathbf{V}_2\mathbf{S}\mathbf{p}\ , \tag{3.86b}$$

wobei (3.86a) den stochastischen und (3.86b) den deterministischen Teil enthält. Die stochastische Störung $\tilde{\mathbf{v}} = \mathbf{V}_1\mathbf{v}$ hat die Dimension q. Ihre Kovarianzmatrix lautet

$$E(\tilde{\mathbf{v}}\tilde{\mathbf{v}}^{\mathrm{T}}) = \mathrm{E}(\mathbf{V}_1\mathbf{v}\mathbf{v}^{\mathrm{T}}\mathbf{V}_1^{\mathrm{T}}) = \mathbf{V}_1\mathbf{Q}\mathbf{V}_1^{\mathrm{T}} \tag{3.87}$$

und ist wegen (3.85a) regulär. Damit ist es gelungen die eingangs erwähnte Redundanz in $\mathbf{v}$ zu beseitigen.

> *Bemerkung* 3.1 (Beschränkte gewichtete lineare Least-Squares Methode). Es kann nun die beschränkte gewichtete lineare Least-Squares Optimierungsaufgabe
>
> $$\widehat{\mathbf{T}\mathbf{p}} = \mathbf{T}\arg\min_{\tilde{\mathbf{p}}\in\mathbb{R}^n} (\mathbf{S}\tilde{\mathbf{p}} - \mathbf{y})^{\mathrm{T}}\mathbf{V}_1^{\mathrm{T}}(\mathbf{V}_1\mathbf{Q}\mathbf{V}_1^{\mathrm{T}})^{-1}\mathbf{V}_1(\mathbf{S}\tilde{\mathbf{p}} - \mathbf{y}) \tag{3.88a}$$
> $$\text{u.B.v. } \mathbf{V}_2(\mathbf{y} - \mathbf{S}\tilde{\mathbf{p}}) = \mathbf{0} \tag{3.88b}$$
>
> formuliert und gelöst werden. Die zugehörige Hessematrix lautet $\mathbf{S}^{\mathrm{T}}\mathbf{V}_1^{\mathrm{T}}(\mathbf{V}_1\mathbf{Q}\mathbf{V}_1^{\mathrm{T}})^{-1}\mathbf{V}_1\mathbf{S}$ und ist im Allgemeinen nur positiv semidefinit, weshalb die Lösung von (3.88) nicht eindeutig sein kann. Optimale Lösungen existieren aber sicher, was analog zur Lösung der Aufgabe 3.3 gezeigt kann. Im Weiteren wird ein alternativer Weg zum Entwurf eines Schätzers $\widehat{\mathbf{T}\mathbf{p}}$ (basierend auf bisherigen Ergebnissen und ohne eine weitere Optimierungsaufgabe) beschrieben.

Um bei der Schätzung die Zwangsbedingung (3.86b) automatisch zu berücksichtigen, kann der Parametervektor $\mathbf{p}$ in der Form

$$\mathbf{p} = (\mathbf{V}_2\mathbf{S})^\dagger \mathbf{V}_2\mathbf{y} + (\mathbf{E} - (\mathbf{V}_2\mathbf{S})^\dagger \mathbf{V}_2\mathbf{S})\tilde{\mathbf{p}} \tag{3.89}$$

mit neuen Parametern $\tilde{\mathbf{p}} \in \mathbb{R}^n$ dargestellt werden. Der erste Summand ist eine partikuläre Lösung von (3.86b), wie leicht durch Einsetzen in (3.86b) und Berücksichtigung von (3.1), Lemma 3.1 und (3.85b) gezeigt werden kann. Der zweite Summand ist eine homogene Lösung von (3.86b), was sofort durch Einsetzen in den homogenen Teil von (3.86b) gezeigt werden kann. Unter Verwendung von (3.89) kann das Modell (3.86) auf die Form

$$\underbrace{\mathbf{V}_1(\mathbf{E} - \mathbf{S}(\mathbf{V}_2\mathbf{S})^\dagger \mathbf{V}_2)\mathbf{y}}_{\tilde{\mathbf{y}}} = \underbrace{\mathbf{V}_1\mathbf{S}(\mathbf{E} - (\mathbf{V}_2\mathbf{S})^\dagger \mathbf{V}_2\mathbf{S})}_{\tilde{\mathbf{S}}}\tilde{\mathbf{p}} + \underbrace{\mathbf{V}_1\mathbf{v}}_{\tilde{\mathbf{v}}} \tag{3.90}$$

reduziert werden. Wegen (3.85b) und $\mathbf{V}_2\mathbf{S}(\mathbf{E} - (\mathbf{V}_2\mathbf{S})^\dagger \mathbf{V}_2\mathbf{S}) = \mathbf{0}$ gilt

$$\text{Bild}(\mathbf{S}(\mathbf{E} - (\mathbf{V}_2\mathbf{S})^\dagger \mathbf{V}_2\mathbf{S})) \subseteq \text{Bild}(\mathbf{Q}) \ . \tag{3.91}$$

Daraus folgt gemeinsam mit (3.85a)

$$\text{Kern}(\mathbf{V}_1\mathbf{S}(\mathbf{E} - (\mathbf{V}_2\mathbf{S})^\dagger \mathbf{V}_2\mathbf{S})) = \text{Kern}(\mathbf{S}(\mathbf{E} - (\mathbf{V}_2\mathbf{S})^\dagger \mathbf{V}_2\mathbf{S})) \tag{3.92}$$

und schließlich wegen (3.45a)

$$\text{Kern}(\underbrace{\mathbf{V}_1\mathbf{S}(\mathbf{E} - (\mathbf{V}_2\mathbf{S})^\dagger \mathbf{V}_2\mathbf{S})}_{\tilde{\mathbf{S}}}) \subseteq \text{Kern}(\underbrace{\mathbf{T}(\mathbf{E} - (\mathbf{V}_2\mathbf{S})^\dagger \mathbf{V}_2\mathbf{S})}_{\tilde{\mathbf{T}}}) \ . \tag{3.93}$$

Damit ist sichergestellt, dass mit dem Modell (3.90) und der Methode aus Abschnitt 3.1.2 ein eindeutiger Schätzwert $\widehat{\tilde{\mathbf{T}}\tilde{\mathbf{p}}}$ für $\tilde{\mathbf{T}}\tilde{\mathbf{p}} = \mathbf{T}(\mathbf{E} - (\mathbf{V}_2\mathbf{S})^\dagger \mathbf{V}_2\mathbf{S})\tilde{\mathbf{p}}$ berechnet werden kann. Unter Verwendung von (3.89) folgt daraus schließlich der gesuchte Schätzer

$$\widehat{\mathbf{T}\mathbf{p}} = \mathbf{T}(\mathbf{V}_2\mathbf{S})^\dagger \mathbf{V}_2\mathbf{y} + \widehat{\tilde{\mathbf{T}}\tilde{\mathbf{p}}} \ . \tag{3.94}$$

3.2 Parameterschätzung für ein nichtlineares Modell

Von einem System sind die zu einem Vektor zusammengefassten Ausgangswerte $\mathbf{y} \in \mathbb{R}^m$ (z. B. aus Messungen) verfügbar. Für sie wird ein nichtlineares Modell der Form

$$\mathbf{y} = \mathbf{h}(\mathbf{p}) + \mathbf{v} \tag{3.95}$$

mit unbekannten Systemparametern $\mathbf{p} \in \mathbb{R}^n$ angenommen. Hierbei ist die deterministische Abbildung $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^m$ bekannt und $\mathbf{v}$ ist eine stochastische Störung (Zufallszahl) mit Erwartungswert $\text{E}(\mathbf{v}) = \mathbf{0}$ und bekannter, symmetrischer, positiv semidefinierter Kovarianzmatrix $\text{E}(\mathbf{v}\mathbf{v}^\text{T}) = \mathbf{Q} \geq 0$. Die Verteilung von $\mathbf{v}$ kann beliebig und unbekannt sein.

Es soll ein Schätzer $\hat{\mathbf{p}}$ für die unbekannten Parameter $\mathbf{p}$ entworfen werden. Vorbereitend dafür werden zunächst Möglichkeiten zur Berechnung der Ableitung von $\mathbf{h}$ bezüglich $\mathbf{p}$ vorgestellt.

Berechnung der Jacobi-Matrix $(\nabla \mathbf{h})(\mathbf{p})$

Die Berechnung der Jacobi-Matrix

$$(\nabla \mathbf{h})(\mathbf{p}) = \left(\frac{\partial \mathbf{h}}{\partial \mathbf{p}}\right)^{\mathrm{T}} = \begin{bmatrix} \frac{\partial h_1}{\partial p_1} & \dots & \frac{\partial h_m}{\partial p_1} \\ \vdots & & \vdots \\ \frac{\partial h_1}{\partial p_n} & \dots & \frac{\partial h_m}{\partial p_n} \end{bmatrix} \tag{3.96}$$

kann z. B. mit den im Skriptum Optimierung [3.7] erläuterten Verfahren analytisch oder näherungsweise numerisch erfolgen. Ergänzend werden hier für den Fall, dass die Abbildung $\mathbf{h}(\mathbf{p})$ die Lösung eines zeitkontinuierlichen oder zeitdiskreten dynamischen Systems beinhaltet, mögliche Berechnungswege für $(\nabla \mathbf{h})(\mathbf{p})$ vorgestellt.

- Setzt sich $\mathbf{y} \in \mathbb{R}^m$ aus den zu diskreten Zeitpunkten t_k mit $k = 0, \dots, N-1$ aufgenommenen Messungen $\bar{\mathbf{y}}(t_k)$ eines Systemausgangs zusammen und wird das System durch das *zeitkontinuierliche dynamische Modell*

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{p}) \qquad t \geq t_0 \tag{3.97a}$$
$$\mathbf{x}(t_0) = \mathbf{x}_0(\mathbf{p}) \tag{3.97b}$$
$$\bar{\mathbf{y}}(t) = \bar{\mathbf{h}}(t, \mathbf{x}(t), \mathbf{p}) + \bar{\mathbf{v}}(t) \qquad t \geq t_0 \tag{3.97c}$$

  beschrieben, so gilt

$$\mathbf{y} = \left[\bar{\mathbf{y}}^{\mathrm{T}}(t_0) \quad \dots \quad \bar{\mathbf{y}}^{\mathrm{T}}(t_{N-1})\right]^{\mathrm{T}} \tag{3.98a}$$
$$\mathbf{h}(\mathbf{p}) = \left[\bar{\mathbf{h}}^{\mathrm{T}}(t_0, \mathbf{x}(t_0), \mathbf{p}) \quad \dots \quad \bar{\mathbf{h}}^{\mathrm{T}}(t_{N-1}, \mathbf{x}(t_{N-1}), \mathbf{p})\right]^{\mathrm{T}}. \tag{3.98b}$$

  Nach Berechnung der Zustandstrajektorie $\mathbf{x}(t)$ kann die Sensitivität

$$\mathbf{X}(t) := \frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}\mathbf{p}} \tag{3.99}$$

  von $\mathbf{x}(t)$ bezüglich $\mathbf{p}$ durch Lösung der Anfangswertaufgabe (Sensitivitätsdifferentialgleichung)

$$\dot{\mathbf{X}}(t) = \frac{\partial \mathbf{f}(t, \mathbf{x}, \mathbf{p})}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{x}(t)} \mathbf{X}(t) + \frac{\partial \mathbf{f}(t, \mathbf{x}(t), \mathbf{p})}{\partial \mathbf{p}} \qquad t \geq t_0 \tag{3.100a}$$
$$\mathbf{X}(t_0) = \frac{\mathrm{d}\mathbf{x}_0(\mathbf{p})}{\mathrm{d}\mathbf{p}} \tag{3.100b}$$

  ermittelt werden. Damit kann die Ableitung

$$\frac{\mathrm{d}\bar{\mathbf{h}}(t, \mathbf{x}(t), \mathbf{p})}{\mathrm{d}\mathbf{p}} = \frac{\partial \bar{\mathbf{h}}(t, \mathbf{x}, \mathbf{p})}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{x}(t)} \mathbf{X}(t) + \frac{\partial \bar{\mathbf{h}}(t, \mathbf{x}(t), \mathbf{p})}{\partial \mathbf{p}} \tag{3.101}$$

  zu den Messzeitpunkten $t = t_k$ ausgewertet und zur Jacobi-Matrix

$$(\nabla \mathbf{h})(\mathbf{p}) = \left[\left(\frac{\mathrm{d}\bar{\mathbf{h}}(t_0, \mathbf{x}(t_0), \mathbf{p})}{\mathrm{d}\mathbf{p}}\right)^{\mathrm{T}} \quad \dots \quad \left(\frac{\mathrm{d}\bar{\mathbf{h}}(t_{N-1}, \mathbf{x}(t_{N-1}), \mathbf{p})}{\mathrm{d}\mathbf{p}}\right)^{\mathrm{T}}\right] \tag{3.102}$$

assembliert werden. Für eine speichereffiziente Implementierung sollten die Integration von (3.97a) und (3.100a) und die Assemblierung von (3.102) synchron durchgeführt werden. Dann kann auf eine Speicherung der Trajektorien $\mathbf{x}(t)$ und $\mathbf{X}(t)$ verzichtet werden.

- Setzt sich $\mathbf{y} \in \mathbb{R}^m$ aus den zeitdiskreten Messungen $\bar{\mathbf{y}}_k$ mit $k = 0, \dots, N-1$ eines Systemausgangs zusammen und wird das System durch das *zeitdiskrete dynamische Modell*

$$\mathbf{x}_{k+1} = \mathbf{f}_k(\mathbf{x}_k, \mathbf{p}) \qquad k \geq 0 \tag{3.103a}$$
$$\mathbf{x}_0 = \mathbf{x}_0(\mathbf{p}) \tag{3.103b}$$
$$\bar{\mathbf{y}}_k = \bar{\mathbf{h}}_k(\mathbf{x}_k, \mathbf{p}) + \bar{\mathbf{v}}_k \qquad k \geq 0 \tag{3.103c}$$

beschrieben, so gilt

$$\mathbf{y} = \begin{bmatrix} \bar{\mathbf{y}}_0^{\mathrm{T}} & \dots & \bar{\mathbf{y}}_{N-1}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \tag{3.104a}$$
$$\mathbf{h}(\mathbf{p}) = \begin{bmatrix} \bar{\mathbf{h}}_0^{\mathrm{T}}(\mathbf{x}_0, \mathbf{p}) & \dots & \bar{\mathbf{h}}_{N-1}^{\mathrm{T}}(\mathbf{x}_{N-1}, \mathbf{p}) \end{bmatrix}^{\mathrm{T}}. \tag{3.104b}$$

In diesem Fall kann $(\nabla \mathbf{h})(\mathbf{p})$ direkt durch Nachdifferenzieren (Anwendung der Kettenregel) berechnet werden. Zur übersichtlichen Darstellung kann (analog zum zeitkontinuierlichen Fall) auch hier die Sensitivität

$$\mathbf{X}_k := \frac{\mathrm{d}\mathbf{x}_k}{\mathrm{d}\mathbf{p}} \tag{3.105}$$

definiert und mittels der Differenzengleichung

$$\mathbf{X}_{k+1} = \frac{\partial \mathbf{f}_k(\mathbf{x}_k, \mathbf{p})}{\partial \mathbf{x}_k} \mathbf{X}_k + \frac{\partial \mathbf{f}_k(\mathbf{x}_k, \mathbf{p})}{\partial \mathbf{p}} \qquad k \geq 0 \tag{3.106a}$$
$$\mathbf{X}_0 = \frac{\mathrm{d}\mathbf{x}_0(\mathbf{p})}{\mathrm{d}\mathbf{p}} \tag{3.106b}$$

berechnet werden. Damit kann die Ableitung

$$\frac{\mathrm{d}\bar{\mathbf{h}}_k(\mathbf{x}_k, \mathbf{p})}{\mathrm{d}\mathbf{p}} = \frac{\partial \bar{\mathbf{h}}_k(\mathbf{x}_k, \mathbf{p})}{\partial \mathbf{x}_k} \mathbf{X}_k + \frac{\partial \bar{\mathbf{h}}_{N-1}(\mathbf{x}_{N-1}, \mathbf{p})}{\partial \mathbf{p}} \tag{3.107}$$

ausgewertet und zur Jacobi-Matrix

$$(\nabla \mathbf{h})(\mathbf{p}) = \begin{bmatrix} \left(\frac{\mathrm{d}\bar{\mathbf{h}}_0(\mathbf{x}_0, \mathbf{p})}{\mathrm{d}\mathbf{p}}\right)^{\mathrm{T}} & \dots & \left(\frac{\mathrm{d}\bar{\mathbf{h}}_{N-1}(\mathbf{x}_{N-1}, \mathbf{p})}{\mathrm{d}\mathbf{p}}\right)^{\mathrm{T}} \end{bmatrix} \tag{3.108}$$

assembliert werden.

3.2.1 Der reguläre Fall

Für das Modell (3.95) mit einer bekannten, symmetrischen, positiv definiten Kovarianzmatrix $\mathrm{E}(\mathbf{v}\mathbf{v}^{\mathrm{T}}) = \mathbf{Q} > 0$ soll nun ein Schätzer

$$\hat{\mathbf{p}} = \mathbf{k}(\mathbf{y}) \tag{3.109}$$

für die unbekannten Parameter $\mathbf{p}$ entworfen werden. Ohne weitere Kenntnisse über die Verteilung von $\mathbf{v}$ kann hier wegen der Nichtlinearität der Funktion $\mathbf{h}$ im Allgemeinen kein Schätzer $\mathbf{k}(\,\cdot\,)$ konstruiert werden, der stochastische Eigenschaften wie Erwartungstreue

$$\mathrm{E}(\hat{\mathbf{p}}) = \mathrm{E}(\mathbf{k}(\mathbf{h}(\mathbf{p}) + \mathbf{v})) = \mathbf{p} \tag{3.110}$$

und minimale Varianz des Parameterschätzfehlers

$$\min_{\mathbf{k}(\,\cdot\,)} \quad \mathrm{E}(\|\hat{\mathbf{p}} - \mathbf{p}\|_2^2) = \mathrm{E}(\|\mathbf{k}(\mathbf{h}(\mathbf{p}) + \mathbf{v}) - \mathbf{p}\|_2^2) \tag{3.111}$$

sicherstellt. Selbst wenn die Verteilung von $\mathbf{v}$ bekannt ist, ist die Konstruktion eines solchen Schätzers aus zumindest zwei Gründen eine schwierige Aufgabe: a) Die durch (3.110) beschränkte Optimierungsaufgabe (3.111) ist infinit-dimensional. b) In (3.110) und (3.111) sind Erwartungswerte von nichtlinear transformierten Zufallsgrößen zu berechnen (siehe z. B. [3.8–3.10]). Wahrscheinlichkeitstheoretische Überlegungen dazu werden in Abschnitt 3.2.4 angestellt. Ohne diese Überlegungen kann zunächst ein Schätzer mit Hilfe der *gewichteten Least-Squares Methode* entworfen werden.

Für das lineare Modell (3.1) hat sich in Aufgabe 3.1 gezeigt, dass die gewichtete lineare Least-Squares Methode den gleichen erwartungstreuen linearen Schätzer liefert wie die Minimierung der Varianz des Parameterschätzfehlers. Für das nichtlineare Modell (3.95) gilt dieser Zusammenhang im Allgemeinen nicht. Basierend auf den Ergebnissen der Aufgabe 3.1 erscheint es aber sinnvoll, hier die gewichtete nichtlineare Least-Squares Optimierungsaufgabe

$$\hat{\mathbf{p}} = \mathbf{k}(\mathbf{y}) = \arg\min_{\tilde{\mathbf{p}} \in \mathbb{R}^n} (\mathbf{h}(\tilde{\mathbf{p}}) - \mathbf{y})^{\mathrm{T}} \mathbf{W} (\mathbf{h}(\tilde{\mathbf{p}}) - \mathbf{y}) \tag{3.112}$$

mit der Gewichtungsmatrix $\mathbf{W} = \mathbf{Q}^{-1} \geq 0$ zu lösen. Grundsätzlich muss $\mathbf{W}$ symmetrisch und positiv semidefinit sein.

Die Optimierungsaufgabe (3.112) kann z. B. iterativ mit der Newton-Methode oder der Gauss-Newton-Methode gelöst werden (siehe [3.7]). Ein Iterationsschritt der Gauss-Newton-Methode lautet

$$\tilde{\mathbf{p}}_{k+1} = \tilde{\mathbf{p}}_k - \Big((\nabla \mathbf{h})(\tilde{\mathbf{p}}_k) \mathbf{W} (\nabla \mathbf{h})^{\mathrm{T}}(\tilde{\mathbf{p}}_k) \Big)^{-1} (\nabla \mathbf{h})(\tilde{\mathbf{p}}_k) \mathbf{W} (\mathbf{h}(\tilde{\mathbf{p}}_k) - \mathbf{y}) \ . \tag{3.113}$$

Damit die Matrix $(\nabla \mathbf{h})(\tilde{\mathbf{p}}_k) \mathbf{W} (\nabla \mathbf{h})^{\mathrm{T}}(\tilde{\mathbf{p}}_k)$ invertierbar ist, muss $(\nabla \mathbf{h})(\tilde{\mathbf{p}}_k)$ zeilenregulär sein.

Für den Spezialfall $\mathbf{h}(\mathbf{p}) = \mathbf{S}\mathbf{p}$, d. h. bei einem linearen Modell, liefern die Newton-Methode und die Gauss-Newton-Methode identische Iterationsvorschriften und konvergieren in einem Schritt. Es liegt also dann eine analytische Lösung vor und es gilt $(\nabla \mathbf{h})(\mathbf{p}) = \mathbf{S}^{\mathrm{T}}$, womit sich (3.113) zum Schätzer (3.16) vereinfacht.

Es zeigt sich also, dass ein lineares Modell erhebliche Vorteile gegenüber einem nichtlinearen Modell besitzt. Nachfolgend wird eine Möglichkeit gezeigt, wie diese Vorteile zumindest teilweise genützt werden können, wenn manche der zu schätzenden Parameter linear auftreten.

Separierung in linear und nichtlinear auftretende Parameter

Häufig kann ein nichtlineares Modell (3.95) in der Form

$$\mathbf{y} = \mathbf{h}_0(\mathbf{p}_{\text{nlin}}) + \mathbf{H}(\mathbf{p}_{\text{nlin}})\mathbf{p}_{\text{lin}} + \mathbf{v} \tag{3.114}$$

dargestellt werden, d. h. mit separierten linearen Parametern $\mathbf{p}_{\text{lin}}$ und nichtlinearen Parametern $\mathbf{p}_{\text{nlin}}$, welche im Vektor $\mathbf{p} = \begin{bmatrix}\mathbf{p}_{\text{nlin}}^{\mathrm{T}} & \mathbf{p}_{\text{lin}}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}$ zusammengefasst werden. Damit lautet die Optimierungsaufgabe in (3.112)

$$\min_{\tilde{\mathbf{p}}\in\mathbb{R}^n} \quad (\mathbf{h}_0(\tilde{\mathbf{p}}_{\text{nlin}}) + \mathbf{H}(\tilde{\mathbf{p}}_{\text{nlin}})\tilde{\mathbf{p}}_{\text{lin}} - \mathbf{y})^{\mathrm{T}}\mathbf{W}(\mathbf{h}_0(\tilde{\mathbf{p}}_{\text{nlin}}) + \mathbf{H}(\tilde{\mathbf{p}}_{\text{nlin}})\tilde{\mathbf{p}}_{\text{lin}} - \mathbf{y}) \ . \tag{3.115}$$

Analog zu Abschnitt 3.1.1 lautet die analytische Lösung für $\mathbf{p}_{\text{lin}}$ daher

$$\hat{\mathbf{p}}_{\text{lin}} = \left(\mathbf{H}^{\mathrm{T}}(\tilde{\mathbf{p}}_{\text{nlin}})\mathbf{W}\mathbf{H}(\tilde{\mathbf{p}}_{\text{nlin}})\right)^{-1}\mathbf{H}^{\mathrm{T}}(\tilde{\mathbf{p}}_{\text{nlin}})\mathbf{W}(\mathbf{y} - \mathbf{h}_0(\tilde{\mathbf{p}}_{\text{nlin}})) \ . \tag{3.116}$$

Einsetzen dieser Lösung in (3.115) liefert die nichtlineare Optimierungsaufgabe

$$\begin{aligned}\min_{\tilde{\mathbf{p}}_{\text{nlin}}} \quad & (\mathbf{h}_0(\tilde{\mathbf{p}}_{\text{nlin}}) - \mathbf{y})^{\mathrm{T}}\left(\mathbf{W} - \mathbf{W}\mathbf{H}(\tilde{\mathbf{p}}_{\text{nlin}})\left(\mathbf{H}^{\mathrm{T}}(\tilde{\mathbf{p}}_{\text{nlin}})\mathbf{W}\mathbf{H}(\tilde{\mathbf{p}}_{\text{nlin}})\right)^{-1}\mathbf{H}^{\mathrm{T}}(\tilde{\mathbf{p}}_{\text{nlin}})\mathbf{W}\right) \\ & (\mathbf{h}_0(\tilde{\mathbf{p}}_{\text{nlin}}) - \mathbf{y}) \ ,\end{aligned} \tag{3.117}$$

welche aufgrund ihrer reduzierten Dimension häufig mit geringerem Rechenaufwand als (3.115) gelöst werden kann. Rückeinsetzen der Lösung $\hat{\mathbf{p}}_{\text{nlin}}$ aus (3.117) in (3.116) liefert den Wert $\hat{\mathbf{p}}_{\text{lin}}$.

Transformation auf lineares Modell

In seltenen Fällen kann ein (lokaler) Diffeomorphismus (bijektive, stetig differenzierbare Abbildung, deren Umkehrabbildung auch stetig differenzierbar ist) $\mathbf{p} = \mathbf{g}(\mathbf{p}_{\text{lin}})$ so gefunden werden, dass

$$\mathbf{h}(\mathbf{g}(\mathbf{p}_{\text{lin}})) = \mathbf{S}\mathbf{p}_{\text{lin}} \tag{3.118}$$

gilt. Dann kann das nichtlineare Modell (3.95) auf die lineare Form

$$\mathbf{y} = \mathbf{S}\mathbf{p}_{\text{lin}} + \mathbf{v} \tag{3.119}$$

vereinfacht werden. Gelingt dies nicht, so kann möglicherweise zumindest ein (lokaler) Diffeomorphismus $\mathbf{p} = \mathbf{g}(\mathbf{p}_{\text{nlin}}, \mathbf{p}_{\text{lin}})$ gefunden werden, mit dem sich das nichtlineare Modell (3.95) auf die separierbare Form (3.114) vereinfachen lässt.

> *Beispiel* 3.3 (Schätzung unbekannter harmonischer Signale). Ein skalares Ausgangssignal wird gemessen, mit einer Periodendauer von 1 s abgetastet und im Vektor $\mathbf{y} \in \mathbb{R}^m$ zusammengefasst, d. h. $\mathbf{y} = [y_k]_{k=0,\dots,m-1}$. Diese Messwerte beinhalten eine stochastische Störung $\mathbf{v}$ mit Erwartungswert $\mathrm{E}(\mathbf{v}) = \mathbf{0}$ und symmetrischer, positiv definierter Kovarianzmatrix $\mathrm{E}(\mathbf{v}\mathbf{v}^{\mathrm{T}}) = q\mathbf{E} \geq \mathbf{0}$, wobei $q > 0$ unbekannt ist. Die Messwerte sind in Abbildung 3.1 dargestellt.

Bemerkung 3.2. Die in Abbildung 3.1 gezeigten Werte wurden algorithmisch generiert. Dabei wurde zum eigentlichen Signal eine im Intervall $[-2.5, 2.5]$ gleichverteilte zufällige Störung addiert. Das Signal-Rausch-Verhältnis beträgt etwa 0 dB.

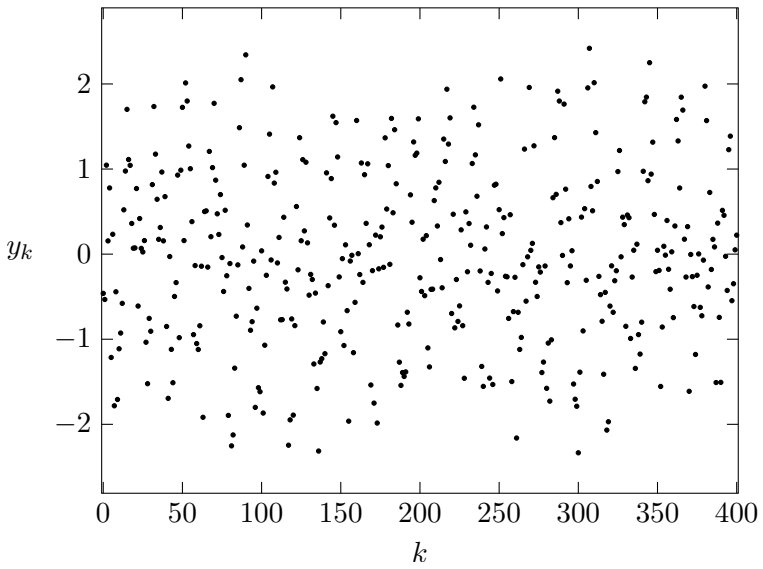


Abbildung 3.1: Messwerte.

Es ist bekannt, dass das ungestörte Ausgangssignal durch Addition zweier harmonischer Signale entstand. Es sollen alle Parameter (Amplitude, Frequenz und Phase) dieser harmonischen Signale geschätzt werden. Es kann daher das Modell

$$y_k = A_1 \cos(2\pi f_1 k + \varphi_1) + A_2 \cos(2\pi f_2 k + \varphi_2) + v_k \tag{3.120}$$

verwendet werden. Von den zu schätzenden Größen $\mathbf{p} = \begin{bmatrix} A_1 & A_2 & \varphi_1 & \varphi_2 & f_1 & f_2 \end{bmatrix}^{\mathrm{T}}$ treten hier nur A_1 und A_2 in linearer Form auf. Eine Transformation auf ein lineares Modell ist nicht möglich. Mit dem lokalen Diffeomorphismus

$$\mathbf{p} = \begin{bmatrix} A_1 \\ A_2 \\ \varphi_1 \\ \varphi_2 \\ f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} \sqrt{p_{\mathrm{lin},1}^2 + p_{\mathrm{lin},2}^2} \\ \sqrt{p_{\mathrm{lin},3}^2 + p_{\mathrm{lin},4}^2} \\ \arctan\left(\frac{-p_{\mathrm{lin},2}}{p_{\mathrm{lin},1}}\right) \\ \arctan\left(\frac{-p_{\mathrm{lin},4}}{p_{\mathrm{lin},3}}\right) \\ p_{\mathrm{nlin},1} \\ p_{\mathrm{nlin},2} \end{bmatrix}, \tag{3.121a}$$

dessen Umkehrabbildung

$$\begin{bmatrix}\mathbf{p}_{\text{nlin}}\\ \mathbf{p}_{\text{lin}}\end{bmatrix} = \begin{bmatrix}p_{\text{nlin},1}\\ p_{\text{nlin},2}\\ p_{\text{lin},1}\\ p_{\text{lin},2}\\ p_{\text{lin},3}\\ p_{\text{lin},4}\end{bmatrix} = \begin{bmatrix}f_1\\ f_2\\ A_1\cos(\varphi_1)\\ -A_1\sin(\varphi_1)\\ A_2\cos(\varphi_2)\\ -A_2\sin(\varphi_2)\end{bmatrix} \tag{3.121b}$$

lautet, kann das Modell aber zumindest in die separierbare Form

$$\begin{aligned} y_k = {} & p_{\text{lin},1}\cos(2\pi p_{\text{nlin},1}k) + p_{\text{lin},2}\sin(2\pi p_{\text{nlin},1}k)\\ & + p_{\text{lin},3}\cos(2\pi p_{\text{nlin},2}k) + p_{\text{lin},4}\sin(2\pi p_{\text{nlin},2}k) + v_k \end{aligned} \tag{3.122}$$

und somit die Darstellung (3.114) mit

$$\mathbf{h}_0(\mathbf{p}_{\text{nlin}}) = \mathbf{0} \tag{3.123a}$$

$$\mathbf{H}(\mathbf{p}_{\text{nlin}}) = \begin{bmatrix}\mathbf{H}_0(\mathbf{p}_{\text{nlin}})\\ \mathbf{H}_1(\mathbf{p}_{\text{nlin}})\\ \vdots\\ \mathbf{H}_{m-1}(\mathbf{p}_{\text{nlin}})\end{bmatrix} \tag{3.123b}$$

$$\mathbf{H}_k(\mathbf{p}_{\text{nlin}}) = \begin{bmatrix}\cos(2\pi p_{\text{nlin},1}k) & \sin(2\pi p_{\text{nlin},1}k) & \cos(2\pi p_{\text{nlin},2}k) & \sin(2\pi p_{\text{nlin},2}k)\end{bmatrix} \tag{3.123c}$$

transformiert werden. In der Optimierungsaufgabe (3.115) kann $\mathbf{W} = \mathbf{E}/q$ verwendet werden, wobei der unbekannte Parameter q in der Lösung

$$\hat{\mathbf{p}}_{\text{lin}} = \Big(\mathbf{H}^{\mathrm{T}}(\tilde{\mathbf{p}}_{\text{nlin}})\mathbf{H}(\tilde{\mathbf{p}}_{\text{nlin}})\Big)^{-1}\mathbf{H}^{\mathrm{T}}(\tilde{\mathbf{p}}_{\text{nlin}})\mathbf{y}\ , \tag{3.124}$$

vgl. (3.116), durch Kürzung wegfällt. Die Optimierungsaufgabe (3.117) vereinfacht sich wegen (3.123a) zu

$$\min_{\tilde{\mathbf{p}}_{\text{nlin}}} \quad \underbrace{\frac{1}{q}\mathbf{y}^{\mathrm{T}}\Big(\mathbf{E} - \mathbf{H}(\tilde{\mathbf{p}}_{\text{nlin}})\Big(\mathbf{H}^{\mathrm{T}}(\tilde{\mathbf{p}}_{\text{nlin}})\mathbf{H}(\tilde{\mathbf{p}}_{\text{nlin}})\Big)^{-1}\mathbf{H}^{\mathrm{T}}(\tilde{\mathbf{p}}_{\text{nlin}})\Big)\mathbf{y}}_{J(\tilde{\mathbf{p}}_{\text{nlin}})}\ , \tag{3.125}$$

wobei ohne Einfluss auf die Lösung $q = 1$ verwendet werden kann. Natürlich könnte in (3.125) auch der konstante Term $\mathbf{y}^{\mathrm{T}}\mathbf{y}/q$ fortgelassen werden. Für die Messwerte aus Abbildung 3.1 ist $J(\mathbf{p}_{\text{nlin}})$ in Abbildung 3.2 dargestellt. Da die Parameter f_1 und f_2 in J symmetrisch auftreten, reicht es den dargestellten Bereich $f_1 \geq f_2$ zu betrachten. Die Optimierungsaufgabe (3.125) hat in diesem Bereich ein eindeutiges

Minimum (siehe Abbildung 3.2) und kann numerisch, z. B. mit der MATLAB-Funktion fminunc, gelöst werden. Anschließend wird $\hat{\mathbf{p}}_{\text{lin}}$ gemäß (3.124) berechnet.

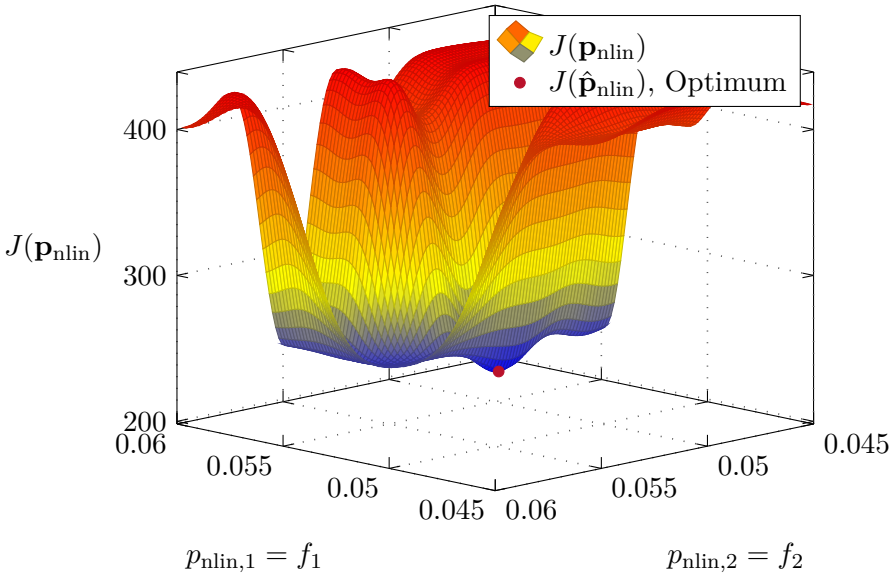


Abbildung 3.2: Kostenfunktion.

Parameter	Wahrer Wert	Geschätzter Wert
A_1	1	0.962
A_2	0.3	0.320
φ_1	0.5	0.508
φ_2	-2.5	-2.402
f_1	0.055 Hz	0.055 02 Hz
f_2	0.05 Hz	0.049 82 Hz

Tabelle 3.1: Wahre und geschätzte Parameterwerte.

Die so geschätzten Parameterwerte sind in Tabelle 3.1 mit jenen verglichen, die zur Generierung der Werte y_k verwendet wurden. Abbildung 3.3 zeigt die Messwerte und den geschätzten Verlauf des unverrauschten Signals.

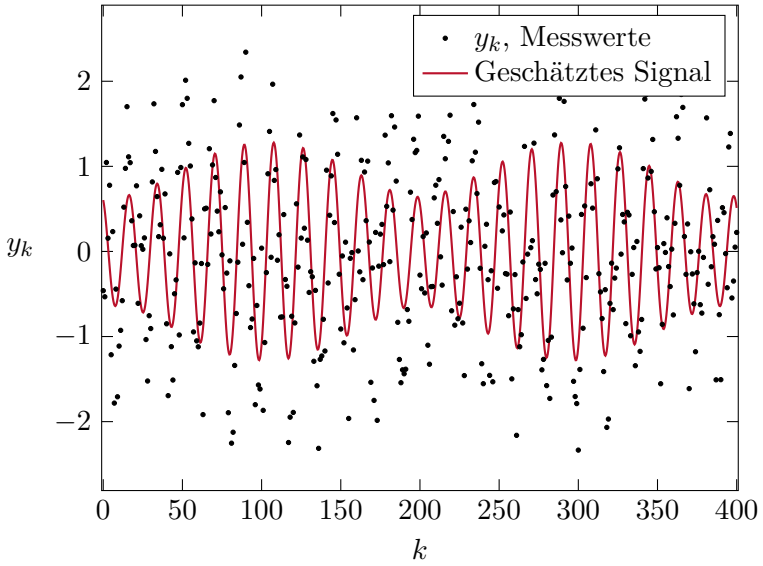


Abbildung 3.3: Messwerte und geschätzter Verlauf des unverrauschten Signals.

Bemerkung 3.3. Ohne die Transformation (3.121) wäre die Schätzaufgabe nur schwer lösbar, da in einem 4-dimensionalen Suchraum (mit zahlreichen lokalen Minima) für die nichtlinear auftretenden Parameter φ_1, φ_2, f_1 und f_2 nach einem globalen Optimum gesucht werden müsste.

3.2.2 Der singuläre Fall

Ist $\mathbf{Q}$ singulär, so kann dies analoge Ursachen haben wie jene, die im Abschnitt 3.1.3 für den linearen Fall angegeben wurden. Ist $\mathbf{Q}$ singulär, so kann in der Least-Squares Optimierungsaufgabe (3.112) die Gewichtungsmatrix $\mathbf{W} = (\mathbf{Q} + \varepsilon\mathbf{E})^{-1}$ mit einem geeigneten Parameter $\varepsilon > 0$ verwendet werden.

Alternativ dazu kann (analog zu Abschnitt 3.1.3) das Modell (3.95) so transformiert werden, dass sich ein reduziertes Modell mit regulärer Kovarianzmatrix der Störung ergibt. Man wählt eine reguläre Matrix $\begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix} \in \mathbb{R}^{m\times m}$ so, dass

$$\mathbf{V}_1^{\mathrm{T}} \in \mathrm{Bild}(\mathbf{Q}) \tag{3.126a}$$
$$\mathbf{V}_2^{\mathrm{T}} \in \mathrm{Kern}(\mathbf{Q}) \tag{3.126b}$$

erfüllt sind. Wird das Modell (3.95) linksseitig einmal mit $\mathbf{V}_1$ und einmal mit $\mathbf{V}_2$ multipliziert, so ergibt sich das transformierte Modell

$$\mathbf{V}_1\mathbf{y} = \mathbf{V}_1\mathbf{h}(\mathbf{p}) + \mathbf{V}_1\mathbf{v} \tag{3.127a}$$
$$\mathbf{V}_2\mathbf{y} = \mathbf{V}_2\mathbf{h}(\mathbf{p}) \ , \tag{3.127b}$$

wobei (3.127a) den stochastischen und (3.127b) den deterministischen Teil des Modells enthält. Die Kovarianzmatrix der stochastischen Störung $\mathbf{V}_1\mathbf{v}$ lautet

$$\mathrm{E}(\mathbf{V}_1\mathbf{v}\mathbf{v}^{\mathrm{T}}\mathbf{V}_1^{\mathrm{T}}) = \mathbf{V}_1\mathbf{Q}\mathbf{V}_1^{\mathrm{T}} \tag{3.128}$$

und ist wegen (3.126a) regulär. Damit kann nun (analog zu Bemerkung 3.1) die beschränkte gewichtete nichtlineare Least-Squares Optimierungsaufgabe

$$\hat{\mathbf{p}} = \mathbf{k}(\mathbf{y}) = \arg\min_{\tilde{\mathbf{p}}\in\mathbb{R}^n} (\mathbf{h}(\tilde{\mathbf{p}}) - \mathbf{y})^{\mathrm{T}}\mathbf{W}(\mathbf{h}(\tilde{\mathbf{p}}) - \mathbf{y}) \tag{3.129a}$$

$$\text{u.B.v. } \mathbf{V}_2\mathbf{y} = \mathbf{V}_2\mathbf{h}(\tilde{\mathbf{p}}) \tag{3.129b}$$

mit der Gewichtungsmatrix $\mathbf{W} = \mathbf{V}_1^{\mathrm{T}}(\mathbf{V}_1\mathbf{Q}\mathbf{V}_1^{\mathrm{T}})^{-1}\mathbf{V}_1 \geq 0$ formuliert und gelöst werden.

3.2.3 Der kollineare Fall

Ähnlich zu Abschnitt 3.1.2 für ein lineares Modell, ist die Schätzung von $\mathbf{p}$ im Modell (3.95) nicht eindeutig oder nur unzuverlässig möglich, wenn unterschiedliche Parameterwerte $\mathbf{p}_1 \neq \mathbf{p}_2$ zu gleichen Erwartungswerten $\mathbf{h}(\mathbf{p}_1) = \mathbf{h}(\mathbf{p}_2)$ oder ähnlichen Erwartungswerten $\mathbf{h}(\mathbf{p}_1) \approx \mathbf{h}(\mathbf{p}_2)$ des Systemausgangs $\mathbf{y}$ führen können.

Sensitivitätsmatrix

Zur Beurteilung, ob dieses Problem vorliegt, kann die *Sensitivitätsmatrix*

$$\mathbf{S}(\mathbf{p}) = (\nabla\mathbf{h})^{\mathrm{T}}(\mathbf{p}) \tag{3.130}$$

verwendet werden. Aus einer Taylorreihenentwicklung am Punkt $\mathbf{p}_1$ folgt

$$\mathbf{h}(\mathbf{p}_2) = \mathbf{h}(\mathbf{p}_1) + \mathbf{S}(\mathbf{p}_1)(\mathbf{p}_2 - \mathbf{p}_1) + \mathcal{O}(\|\mathbf{p}_2 - \mathbf{p}_1\|_2^2) \ . \tag{3.131}$$

Damit nun (bei Vernachlässigung des Restterms in (3.131)) für $\mathbf{p}_1 \neq \mathbf{p}_2$ jedenfalls $\mathbf{h}(\mathbf{p}_1) \neq \mathbf{h}(\mathbf{p}_2)$, oder äquivalent $\|\mathbf{h}(\mathbf{p}_1) - \mathbf{h}(\mathbf{p}_2)\|_2 \neq 0$ gilt, muss $\mathbf{S}(\mathbf{p}_1)$ spaltenregulär sein, d. h. $\mathrm{rang}(\mathbf{S}(\mathbf{p}_1)) = n$. Selbst wenn $\mathbf{S}(\mathbf{p}_1)$ spaltenregulär ist, besteht noch immer die Gefahr von ähnlichen Erwartungswerten $\mathbf{h}(\mathbf{p}_1) \approx \mathbf{h}(\mathbf{p}_2)$ ($\|\mathbf{h}(\mathbf{p}_1) - \mathbf{h}(\mathbf{p}_2)\|_2 \approx 0$) selbst bei signifikant unterschiedlichen Werten $\mathbf{p}_1 \neq \mathbf{p}_2$. Dieser Fall tritt genau dann ein, wenn aus den Spalten von $\mathbf{S}(\mathbf{p})$ eine Linearkombination ähnlich dem Vektor $\mathbf{0}$ gebildet werden kann, und wird als *Kollinearität* der Spalten von $\mathbf{S}(\mathbf{p})$ bezeichnet [3.11]. Für ein nichtlineares Modell hängt die Kollinearitätseigenschaft im Allgemeinen auch von den Parameterwerten $\mathbf{p}$ ab. Für eine kompaktere Notation wird fortan $\mathbf{p}$ nicht mehr explizit als Argument angegeben, d. h. $\mathbf{S} = \mathbf{S}(\mathbf{p})$.

Prüfung auf Kollinearität

Um eine mögliche Kollinearität einer Matrix $\mathbf{S}$ quantitativ zu bewerten, werden zunächst die Spalten $\mathbf{S}_j = \partial\mathbf{h}/\partial p_j$ von $\mathbf{S} = [S_{ij}] = [\partial h_i/\partial p_j]$ mit $j = 1, \ldots, n$ normiert, so dass sich

$$\tilde{\mathbf{S}} = [\tilde{S}_{ij}] = \left[\frac{S_{ij}}{\|\mathbf{S}_j\|_2}\right] = \mathbf{S}\,\mathrm{diag}\{\|\mathbf{S}_1\|_2, \|\mathbf{S}_2\|_2, \ldots, \|\mathbf{S}_n\|_2, \}^{-1} \tag{3.132}$$

ergibt. Hierbei wird davon ausgegangen, dass $\mathbf{S}$ keine Nullspalten enthält. Der in [3.12] vorgeschlagene Wert

$$\rho = \min_{\mathbf{r}\in\mathbb{R}^n}\left(\frac{\|\tilde{\mathbf{S}}\mathbf{r}\|_2}{\|\mathbf{r}\|_2}\right) = \min_{\substack{\mathbf{r}\in\mathbb{R}^n,\\ \|\mathbf{r}\|_2=1}} \|\tilde{\mathbf{S}}\mathbf{r}\|_2 = \sqrt{\lambda_{\min}(\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}})} \;\in\; [0,1] \tag{3.133}$$

verkleinert sich mit zunehmender Kollinearität der Spalten $\mathbf{S}_j$ von $\mathbf{S}$. Die Funktion $\lambda_{\min}(\,\cdot\,)$ liefert den kleinsten Eigenwert einer symmetrischen Matrix. Die übergebene Matrix $\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}$ ist symmetrisch und positiv semi-definit. Ihre Einträge sind auf das Intervall $[-1, 1]$ beschränkt und enthalten die sogenannten *Kosinus-Ähnlichkeiten* (siehe [3.13]) der Spalten $\mathbf{S}_j$. Alle Einträge in der Hauptdiagonale von $\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}$ haben folglich den Wert 1.

Perfekte Kollinearität tritt im Fall $\rho = 0$, also $\mathrm{rang}(\mathbf{S}) < n$ auf. Im Gegensatz dazu gilt $\rho = 1$ bei bestmöglicher Unabhängigkeit (Orthogonalität) der Spalten $\mathbf{S}_j$. In (3.133) wird die Matrix $\tilde{\mathbf{S}}$ mit normierten Spalten (und nicht $\mathbf{S}$) verwendet, um eine gute Interpretierbarkeit und Vergleichbarkeit verschiedener Werte ρ zu ermöglichen.

Gemäß [3.12] kann der Wert ρ wie folgt interpretiert werden: Ein Änderung an den Messwerten $\mathbf{y}$ bzw. an den Erwartungswerten $\mathbf{h}(\mathbf{p})$, die durch eine Änderung eines Parameters p_j hervorgerufen wird, kann (in linearer Näherung) bis auf einen Anteil ρ durch eine Änderung der übrigen Parameter p_k mit $k \neq j$ kompensiert werden. Der Kehrwert von ρ wird in der Literatur auch als *Kollinearitätsindex* bezeichnet [3.12, 3.14]. Es gilt natürlich $1/\rho \geq 1$. Übersteigt $1/\rho$ einen bestimmten Schwellwert, so geht man von Kollinearität und daher schlechter Schätzbarkeit der Parameter $\mathbf{p}$ aus. In [3.12] wird dieser Schwellwert für $1/\rho$ beispielsweise mit 20 angegeben.

$$\frac{1}{\rho} = \frac{1}{\sqrt{\lambda_{\min}(\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}})}} > 20 \quad \Rightarrow \quad \text{Kollinearität} \tag{3.134}$$

Für nichtlineare Modelle hängen die Werte für $\mathbf{S}$, $\tilde{\mathbf{S}}$ und ρ im Allgemeinen von $\mathbf{p}$ ab. Für festgelegte Werte $\mathbf{p}$ kann eine Prüfung auf Kollinearität vorab, d. h. vor Aufnahme von Messwerten $\mathbf{y}$ und vor einer Berechnung eines Schätzwertes $\hat{\mathbf{p}}$, durchgeführt werden.

Es soll nun die Gültigkeit von (3.133) gezeigt werden. Wegen der Symmetrie der positiv semidefiniten Matrix $\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}$ existiert eine orthogonale $n \times n$ Matrix $\mathbf{V}$ so, dass

$$\mathbf{V}^{\mathrm{T}}\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}\mathbf{V} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\} \tag{3.135}$$

mit den Eigenwerten $\lambda_j \geq 0$ von $\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}$ gilt[2]. Die Spalten von $\mathbf{V}$ sind also normierte

[2] Wenn

$$\tilde{\mathbf{S}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$$

der Singulärwertzerlegung von $\tilde{\mathbf{S}}$ mit $m \geq n$, der $m \times n$ Matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathrm{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_n\} \\ \mathbf{0} \end{bmatrix}$$

und orthogonalen Matrizen $\mathbf{U} \in \mathbb{R}^{m\times m}$ und $\mathbf{V} \in \mathbb{R}^{n\times n}$ entspricht, so gilt

$$\mathbf{V}^{\mathrm{T}}\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}\mathbf{V} = \boldsymbol{\Sigma}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{\Sigma} = \mathrm{diag}\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}\ .$$

Eigenvektoren von $\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}$. Es folgt nun mit $\mathbf{r} = \mathbf{V}\mathbf{z}$

$$\|\tilde{\mathbf{S}}\mathbf{r}\|_2 = \sqrt{\mathbf{r}^{\mathrm{T}}\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}\mathbf{r}} = \sqrt{\mathbf{z}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}\mathbf{V}\mathbf{z}} = \sqrt{\sum_{j=1}^{n} \lambda_j z_j^2}\ . \tag{3.136}$$

Wegen $\|\mathbf{r}\|_2 = \|\mathbf{V}\mathbf{z}\|_2 = \|\mathbf{z}\|_2$ liefert die Minimierung von (3.136) schließlich den zu beweisenden Zusammenhang

$$\min_{\substack{\mathbf{r}\in\mathbb{R}^n,\\ \|\mathbf{r}\|_2=1}} \|\tilde{\mathbf{S}}\mathbf{r}\|_2 = \min_{\substack{\mathbf{z}\in\mathbb{R}^n,\\ \|\mathbf{z}\|_2=1}} \sqrt{\sum_{j=1}^{n} \lambda_j z_j^2} = \sqrt{\lambda_{\min}(\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}})}\ . \tag{3.137}$$

Aufgabe 3.7 (Bestmögliche Unabhängigkeit). Zeigen Sie, dass im Fall $\rho = 1$ für die Eigenwerte der Matrix $\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}$

$$\lambda_j = 1 \qquad \forall\, j = 1, \ldots, n \tag{3.138}$$

gilt und die Spalten der Matrix $\mathbf{S}$ orthogonal sind. Besonders einfach ist dieser Beweis zu führen, wenn Sie zunächst zeigen, dass alle Einträge in der Hauptdiagonale von $\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}$ stets den Wert 1 haben und somit $\mathrm{spur}(\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}) = n$ gilt.

Abhilfe bei Kollinearität

Eine Möglichkeit Kollinearität zu vermeiden oder zu beseitigen ist eine *geänderte Wahl des Modells* (3.95), so dass der Kollinearitätsindex

$$\frac{1}{\rho} \leq 20 \tag{3.139}$$

erfüllt. Eine zweite Möglichkeit mit dem Problem umzugehen besteht darin, nicht die Werte von $\mathbf{p}$ selbst, sondern die Werte eines *transformierten (reduzierten) Parametervektors* $\bar{\mathbf{p}}$ zu schätzen. Konkret soll

$$\mathbf{p} = \mathbf{t}(\bar{\mathbf{p}}) \tag{3.140}$$

mit einer injektiven Abbildung $\mathbf{t} : \mathbb{R}^o \to \mathbb{R}^n$ gelten. Eine Voraussetzung für die eindeutige Schätzbarkeit von $\bar{\mathbf{p}}$ ist natürlich die Zeilenregularität von $(\nabla\mathbf{t})(\bar{\mathbf{p}})$. Die Abbildung $\mathbf{t}(\,\cdot\,)$ wird nun so gewählt, dass die Kollinearität der neuen Sensitivitätsmatrix

$$\bar{\mathbf{S}}(\bar{\mathbf{p}}) = \mathbf{S}(\mathbf{p})|_{\mathbf{p}=\mathbf{t}(\bar{\mathbf{p}})}(\nabla\mathbf{t})^{\mathrm{T}}(\bar{\mathbf{p}}) \tag{3.141}$$

z. B. im Sinne von (3.133) geringer ausfällt als jene der ursprünglichen Sensitivitätsmatrix $\mathbf{S}(\mathbf{p})$ gemäß (3.130). Klarerweise sollte daher

$$\mathrm{Bild}((\nabla\mathbf{t})^{\mathrm{T}}(\bar{\mathbf{p}})) \cap \mathrm{Kern}(\mathbf{S}(\mathbf{p})|_{\mathbf{p}=\mathbf{t}(\bar{\mathbf{p}})}) = \{\mathbf{0}\} \tag{3.142}$$

gelten. Ist $\mathbf{S}(\mathbf{p})$ zwar spaltenregulär aber kollinear, so sollte analog gelten, dass sich die Spalten von $\operatorname{diag}\{\|\mathbf{S}_1\|_2, \|\mathbf{S}_2\|_2, \dots, \|\mathbf{S}_n\|_2\}(\nabla \mathbf{t})^{\mathrm{T}}(\bar{\mathbf{p}})$ möglichst *nicht* entlang von jenen Richtungen erstrecken, welche durch die zu kleinen Eigenwerten λ_j von $\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}$ gehörenden Eigenvektoren aus $\mathbf{V}$ aufgespannt werden. Als *kleine* Eigenwerte sind gemäß (3.134) jene zu verstehen, für die $1/\sqrt{\lambda_j} > 20$ gilt.

Beispiel 3.4 (Vermeidung einer singulären Sensitivitätsmatrix durch Transformation). Die Abbildung

$$\mathbf{h}(\mathbf{p}) = \mathbf{a}p_1 + \mathbf{b}\frac{p_2}{p_3} \tag{3.143}$$

mit festen, orthogonalen Vektoren $\mathbf{a}$ und $\mathbf{b}$ besitzt die Sensitivitätsmatrix

$$\mathbf{S}(\mathbf{p}) = (\nabla \mathbf{h})^{\mathrm{T}}(\mathbf{p}) = \begin{bmatrix} \mathbf{a} & \mathbf{b}\frac{1}{p_3} & -\mathbf{b}\frac{p_2}{p_3^2} \end{bmatrix}, \tag{3.144}$$

welche jedenfalls singulär ist. Dies zeigt sich auch im Wert $\rho = 0$ gemäß (3.133). Augenscheinlich können hier die Parameter p_2 und p_3 nicht unabhängig voneinander geschätzt werden.

Es gilt nun

$$\operatorname{diag}\{\|\mathbf{S}_1\|_2, \|\mathbf{S}_2\|_2, \|\mathbf{S}_3\|_2\} = \operatorname{diag}\left\{\|\mathbf{a}\|_2, \frac{\|\mathbf{b}\|_2}{|p_3|}, \frac{\|\mathbf{b}\|_2|p_2|}{p_3^2}\right\} \tag{3.145a}$$

$$\tilde{\mathbf{S}} = \begin{bmatrix} \frac{\mathbf{a}}{\|\mathbf{a}\|_2} & \frac{\mathbf{b}|p_3|}{\|\mathbf{b}\|_2 p_3} & -\frac{\mathbf{b}p_2}{\|\mathbf{b}\|_2|p_2|} \end{bmatrix} \tag{3.145b}$$

$$\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -\frac{|p_3|p_2}{p_3|p_2|} \\ 0 & -\frac{|p_3|p_2}{p_3|p_2|} & 1 \end{bmatrix} \tag{3.145c}$$

$$\operatorname{diag}\{\lambda_1, \lambda_2, \lambda_3\} = \operatorname{diag}\{1, 2, 0\} \tag{3.145d}$$

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{|p_3|p_2}{\sqrt{2}p_3|p_2|} & \frac{|p_3|p_2}{\sqrt{2}p_3|p_2|} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}. \tag{3.145e}$$

Folglich sollte $o \leq 2$ gelten und die Abbildung $\mathbf{t}(\bar{\mathbf{p}})$ sollte so gewählt werden, dass die Spalten von $\operatorname{diag}\{\|\mathbf{S}_1\|_2, \|\mathbf{S}_2\|_2, \|\mathbf{S}_3\|_2\}(\nabla \mathbf{t})^{\mathrm{T}}(\bar{\mathbf{p}})$ nicht parallel zu $\mathbf{v}_3$ verlaufen. Eine mögliche Wahl, die dies erfüllt, lautet

$$\mathbf{p} = \mathbf{t}(\bar{\mathbf{p}}) = \begin{bmatrix} \bar{p}_1 \\ \bar{p}_2 \\ 1 \end{bmatrix}. \tag{3.146}$$

Mit dieser Wahl gilt

$$\mathbf{h}(\mathbf{t}(\bar{\mathbf{p}})) = \mathbf{a}\bar{p}_1 + \mathbf{b}\bar{p}_2 \tag{3.147a}$$

$$\bar{\mathbf{S}}(\bar{\mathbf{p}}) = \begin{bmatrix} \mathbf{a} & \mathbf{b} \end{bmatrix} \tag{3.147b}$$

$$\tilde{\mathbf{S}} = \begin{bmatrix} \frac{\mathbf{a}}{\|\mathbf{a}\|_2} & \frac{\mathbf{b}}{\|\mathbf{b}\|_2} \end{bmatrix} \tag{3.147c}$$

$$\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{3.147d}$$

und folglich $\rho = 1$. Die Spalten von $\bar{\mathbf{S}}$ sind also bestmöglich unabhängig und eine Schätzung von $\bar{\mathbf{p}} = \begin{bmatrix} \bar{p}_1 & \bar{p}_2 \end{bmatrix}^{\mathrm{T}}$ kann zuverlässig erfolgen.

Beispiel 3.5 (Vermeidung einer kollinearen Sensitivitätsmatrix durch Transformation). Für ein lineares Modell (3.1) mit $n = 3$ ist die Sensitivitätsmatrix

$$\mathbf{S} = \begin{bmatrix} 3.77 & 1.12 & 2.96 \\ -0.55 & -6.13 & -11.09 \\ -2.68 & 1.64 & 3.30 \\ -1.36 & -6.45 & -13.36 \\ 6.13 & 9.67 & 19.95 \end{bmatrix} \tag{3.148}$$

bekannt. Im Falle einer Kollinearität soll eine Abbildung (3.140) gesucht werden, welche die Kollinearität beseitigt.

Zunächst werden

$$\tilde{\mathbf{S}} = \begin{bmatrix} 0.4822 & 0.0843 & 0.1104 \\ -0.0703 & -0.4612 & -0.4135 \\ -0.3428 & 0.1234 & 0.1231 \\ -0.1740 & -0.4853 & -0.4982 \\ 0.7841 & 0.7276 & 0.7439 \end{bmatrix} \tag{3.149a}$$

$$\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}} = \begin{bmatrix} 1 & 0.6857 & 0.7107 \\ 0.6857 & 1 & 0.9983 \\ 0.7101 & 0.9983 & 1 \end{bmatrix} \tag{3.149b}$$

$$\mathrm{diag}\{\lambda_1, \lambda_2, \lambda_3\} = \mathrm{diag}\{0.0011, 0.3937, 2.6052\} \tag{3.149c}$$

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{bmatrix} = \begin{bmatrix} 0.003\,31 & 0.8512 & 0.5238 \\ 0.6948 & -0.3967 & 0.5999 \\ -0.7184 & -0.3437 & 0.6048 \end{bmatrix} \tag{3.149d}$$

$$\rho = 0.0333\ , \qquad \frac{1}{\rho} = 29.99 > 20\ . \tag{3.149e}$$

berechnet. Es liegt also eine deutliche Kollinearität vor. Aus dem zu $\lambda_{\min}(\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}) = \lambda_1$ gehörigen Eigenvektor $\mathbf{v}_1$ ist zu erkennen, dass diese Kollinearität vor allem zwischen der zweiten und dritten Spalte der Matrix $\mathbf{S}$ vorliegt.

Da nur der Eigenwert λ_1 die Ungleichung $1/\lambda_1 > 20$ erfüllt, sollte $o \leq 2$ gewählt werden. Da es sich außerdem um ein lineares Modell handelt, reicht hier die Verwendung einer linearen Abbildung

$$\mathbf{p} = \mathbf{t}(\bar{\mathbf{p}}) = \bar{\mathbf{T}}\bar{\mathbf{p}} \,. \tag{3.150}$$

Es muss also eine $n \times o$ Matrix $\bar{\mathbf{T}}$ gewählt werden. Die bestmögliche Wahl für $\bar{\mathbf{T}}$ im Sinne einer geringen verbleibenden Kollinearität ergibt sich aus der Gleichung

$$\begin{bmatrix} \mathbf{v}_2 & \mathbf{v}_3 \end{bmatrix} = \operatorname{diag}\{\|\mathbf{S}_1\|_2, \|\mathbf{S}_2\|_2, \|\mathbf{S}_3\|_2\}\bar{\mathbf{T}} \tag{3.151}$$

in der Form

$$\bar{\mathbf{T}} = \begin{bmatrix} 0.1089 & 0.0670 \\ -0.0298 & 0.0451 \\ -0.0128 & 0.0226 \end{bmatrix} . \tag{3.152}$$

Daraus folgen

$$\bar{\mathbf{S}} = \mathbf{S}\bar{\mathbf{T}} = \begin{bmatrix} 0.3391 & 0.3699 \\ 0.2652 & -0.5636 \\ -0.3830 & -0.0311 \\ 0.2157 & -0.6836 \\ 0.1230 & 1.2971 \end{bmatrix} \tag{3.153a}$$

$$\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{3.153b}$$

und schließlich $\rho = 1$. Die Spalten von $\bar{\mathbf{S}}$ sind also bestmöglich unabhängig und eine Schätzung von $\bar{\mathbf{p}} = \begin{bmatrix} \bar{p}_1 & \bar{p}_2 \end{bmatrix}^{\mathrm{T}}$ kann zuverlässig erfolgen. Wird der Schätzwert für $\bar{\mathbf{p}}$ gemäß Abschnitt 3.1.1 berechnet so stimmt er mit dem Schätzwert $\widehat{\mathbf{Tp}}$ aus Abschnitt 3.1.2 überein, wenn $\mathbf{T} = \bar{\mathbf{T}}^{\dagger}$ gilt.

3.2.4 Schranken für die Kovarianzmatrix des Parameterschätzfehlers

In diesem Abschnitt werden mit Hilfe der *Wahrscheinlichkeitstheorie* weitere Überlegungen zur Erwartungstreue von Schätzern und zur Kovarianzmatrix des Schätzfehlers angestellt. Insbesondere werden Schranken für diese Kovarianzmatrix berechnet.

Es sei $P_{\mathbf{z}}(\mathbf{z})$ die Wahrscheinlichkeitsdichtefunktion (Verteilungsdichtefunktion) einer stetigen Zufallsvariable $\mathbf{z}$. Dementsprechend gilt $P_{\mathbf{v}}(\mathbf{v})$ für die zufällige Störung $\mathbf{v}$ im Modell (3.95). Natürlich sind auch der Ausgangwert $\mathbf{y}$ und der Schätzwert $\hat{\mathbf{p}}$ gemäß (3.109) Zufallsgrößen und es gilt

$$P_{\mathbf{y}}(\mathbf{y};\mathbf{p}) = P_{\mathbf{v}}(\mathbf{y} - \mathbf{h}(\mathbf{p})) \ . \tag{3.154}$$

Die Schreibweise $P_{\mathbf{y}}(\mathbf{y};\mathbf{p})$ signalisiert, dass die Verteilung von $\mathbf{y}$ von der deterministischen Größe $\mathbf{p}$ abhängt. Die Bedingung für Erwartungstreue kann damit in der Form

$$\mathrm{E}(\hat{\mathbf{p}}) = \int \mathbf{k}(\mathbf{y}) P_{\mathbf{y}}(\mathbf{y};\mathbf{p}) \,\mathrm{d}\mathbf{y} = \mathrm{E}(\mathbf{k}(\mathbf{h}(\mathbf{p}) + \mathbf{v})) = \int \mathbf{k}(\mathbf{h}(\mathbf{p}) + \mathbf{v}) P_{\mathbf{v}}(\mathbf{v}) \,\mathrm{d}\mathbf{v} = \mathbf{p} \tag{3.155}$$

angeschrieben werden[3].

Satz 3.1 (Näherungsweise Schranke für die Kovarianzmatrix des Schätzfehlers)**.** *Für die Kovarianzmatrix eines erwartungstreuen Schätzers $\hat{\mathbf{p}} = \mathbf{k}(\mathbf{y})$ für die Parameter $\mathbf{p}$ im nichtlinearen Modell (3.95) gilt näherungsweise (Fehler $\int \mathcal{O}(\|\mathbf{v}\|_2^3) P_{\mathbf{v}}(\mathbf{v}) \,\mathrm{d}\mathbf{v}$ bzw. $\int \mathcal{O}(\|\mathbf{v}\|_2^2) P_{\mathbf{v}}(\mathbf{v}) \,\mathrm{d}\mathbf{v}$)*

$$\mathrm{E}((\hat{\mathbf{p}} - \mathbf{p})(\hat{\mathbf{p}} - \mathbf{p})^{\mathrm{T}}) \approx (\nabla \mathbf{k})^{\mathrm{T}}(\mathbf{h}(\mathbf{p}))\mathbf{Q}(\nabla \mathbf{k})(\mathbf{h}(\mathbf{p})) \gtrsim \left(\mathbf{S}^{\mathrm{T}}(\mathbf{p})\mathbf{Q}^{-1}\mathbf{S}(\mathbf{p})\right)^{-1} \ , \tag{3.156}$$

wobei das (näherungsweise) Größer-Gleich-Zeichen so zu verstehen ist, dass die Matrix $(\nabla \mathbf{k})^{\mathrm{T}}(\mathbf{h}(\mathbf{p}))\mathbf{Q}\ (\nabla \mathbf{k})(\mathbf{h}(\mathbf{p})) - (\mathbf{S}^{\mathrm{T}}(\mathbf{p})\mathbf{Q}^{-1}\mathbf{S}(\mathbf{p}))^{-1}$ (näherungsweise) positiv semidefinit ist.

Der nachfolgende Beweis ist an [3.6] angelehnt.

Beweis. Die Taylorreihenentwicklung von $\mathbf{k}(\,\cdot\,)$ am Punkt $\mathbf{h}(\mathbf{p})$ lautet

$$\hat{\mathbf{p}} = \mathbf{k}(\mathbf{h}(\mathbf{p}) + \mathbf{v}) = \mathbf{k}(\mathbf{h}(\mathbf{p})) + (\nabla \mathbf{k})^{\mathrm{T}}(\mathbf{h}(\mathbf{p}))\mathbf{v} + \mathcal{O}(\|\mathbf{v}\|_2^2) \ . \tag{3.157}$$

Einsetzen in (3.155) liefert unter Berücksichtigung von $\int P_{\mathbf{v}}(\mathbf{v}) \,\mathrm{d}\mathbf{v} = 1$ und $\mathrm{E}(\mathbf{v}) = \mathbf{0}$

$$\begin{aligned} &\int \mathbf{k}(\mathbf{h}(\mathbf{p}) + \mathbf{v}) P_{\mathbf{v}}(\mathbf{v}) \,\mathrm{d}\mathbf{v} \\ &\qquad = \int \left(\mathbf{k}(\mathbf{h}(\mathbf{p})) + (\nabla \mathbf{k})^{\mathrm{T}}(\mathbf{h}(\mathbf{p}))\mathbf{v} + \mathcal{O}(\|\mathbf{v}\|_2^2)\right) P_{\mathbf{v}}(\mathbf{v}) \,\mathrm{d}\mathbf{v} \\ &\qquad = \mathbf{k}(\mathbf{h}(\mathbf{p})) + \int \mathcal{O}(\|\mathbf{v}\|_2^2) P_{\mathbf{v}}(\mathbf{v}) \,\mathrm{d}\mathbf{v} = \mathbf{p} \ . \end{aligned} \tag{3.158}$$

In guter Näherung (Fehler $\int \mathcal{O}(\|\mathbf{v}\|_2^2) P_{\mathbf{v}}(\mathbf{v}) \,\mathrm{d}\mathbf{v}$) kann daher für Erwartungstreue

$$\mathbf{k}(\mathbf{h}(\mathbf{p})) \approx \mathbf{p} \tag{3.159}$$

gefordert werden, so dass sich aus (3.157) für den Schätzfehler

$$\hat{\mathbf{p}} - \mathbf{p} = (\nabla \mathbf{k})^{\mathrm{T}}(\mathbf{h}(\mathbf{p}))\mathbf{v} + \mathcal{O}(\|\mathbf{v}\|_2^2) \tag{3.160}$$

[3] Das Symbol $\int$ in (3.155) ist als Mehrfachintegral über den Ereignisraum der jeweiligen Zufallsvariable zu verstehen.

und in weiterer Folge für dessen Kovarianzmatrix

$$\begin{aligned}&\mathrm{E}((\hat{\mathbf{p}}-\mathbf{p})(\hat{\mathbf{p}}-\mathbf{p})^{\mathrm{T}})\\&\qquad=(\nabla\mathbf{k})^{\mathrm{T}}(\mathbf{h}(\mathbf{p}))\underbrace{\mathrm{E}(\mathbf{v}\mathbf{v}^{\mathrm{T}})}_{\mathbf{Q}}(\nabla\mathbf{k})(\mathbf{h}(\mathbf{p}))+\int\mathcal{O}(\|\mathbf{v}\|_2^3)P_{\mathbf{v}}(\mathbf{v})\,\mathrm{d}\mathbf{v}\end{aligned}\tag{3.161}$$

ergibt. Damit ist der erste Teil von (3.156) gezeigt.

Die Ableitung von (3.159) nach $\mathbf{p}$ lautet

$$(\nabla\mathbf{k})^{\mathrm{T}}(\mathbf{h}(\mathbf{p}))(\nabla\mathbf{h})^{\mathrm{T}}(\mathbf{p})=(\nabla\mathbf{k})^{\mathrm{T}}(\mathbf{h}(\mathbf{p}))\mathbf{S}=\mathbf{E}+\int\mathcal{O}(\|\mathbf{v}\|_2^2)P_{\mathbf{v}}(\mathbf{v})\,\mathrm{d}\mathbf{v}\;.\tag{3.162}$$

Hierbei wird für eine kompaktere Notation $\mathbf{p}$ nicht explizit als Argument von $\mathbf{S}$ angegeben, d. h. $\mathbf{S}=\mathbf{S}(\mathbf{p})$. Durch Ausmultiplizieren kann leicht die foglende Identität bewiesen werden.

$$\begin{aligned}&\mathbf{Q}-\mathbf{S}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\\&\qquad=\Big(\mathbf{Q}-\mathbf{S}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\Big)\mathbf{Q}^{-1}\Big(\mathbf{Q}-\mathbf{S}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\Big)\end{aligned}\tag{3.163}$$

Aus ihr folgt

$$\mathbf{Q}-\mathbf{S}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\geq 0\tag{3.164}$$

und nach links- und rechtsseitiger Multiplikation mit $(\nabla\mathbf{k})^{\mathrm{T}}(\mathbf{h}(\mathbf{p}))$ bzw. $(\nabla\mathbf{k})(\mathbf{h}(\mathbf{p}))$ unter Berücksichtigung von (3.162) schließlich

$$(\nabla\mathbf{k})^{\mathrm{T}}(\mathbf{h}(\mathbf{p}))\mathbf{Q}(\nabla\mathbf{k})(\mathbf{h}(\mathbf{p}))-(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1}+\int\mathcal{O}(\|\mathbf{v}\|_2^2)P_{\mathbf{v}}(\mathbf{v})\,\mathrm{d}\mathbf{v}\geq 0\;.\tag{3.165}$$

Damit ist auch der zweite Teil von (3.156) gezeigt. □

Aufgabe 3.8. Zeigen Sie, dass Satz 3.1 im Falle eines linearen erwartungstreuen Schätzers $\hat{\mathbf{p}}=\mathbf{K}\mathbf{y}$ nicht nur näherungsweise sondern exakt gilt.

Vorbereitend auf den nächsten Satz, der eine rigorosere Schranke für die Kovarianzmatrix eines erwartungstreuen Schätzers formuliert, werden zwei Maße für den *Informationsgehalt* der Zufallsgröße $\mathbf{y}$ bezüglich $\mathbf{p}$ definiert. Weiterführende Informationen dazu sind auch in [3.6, 3.15] zu finden.

Definition 3.2 (Score-Funktion)**.** Wird eine Zufallsvariable $\mathbf{y}$ durch Parameter $\mathbf{p}$ beeinflusst und ist $P_{\mathbf{y}}(\mathbf{y};\mathbf{p})$ deren Wahrscheinlichkeitsdichtefunktion, so gilt für die Score-Funktion

$$\mathbf{s}=\left(\frac{\partial\ln(P_{\mathbf{y}}(\mathbf{y};\mathbf{p}))}{\partial\mathbf{p}}\right)^{\mathrm{T}}=\frac{1}{P_{\mathbf{y}}(\mathbf{y};\mathbf{p})}\left(\frac{\partial P_{\mathbf{y}}(\mathbf{y};\mathbf{p})}{\partial\mathbf{p}}\right)^{\mathrm{T}}.\tag{3.166}$$

Die Score-Funktion hängt von $\mathbf{y}$ und $\mathbf{p}$ ab. Sie entspricht der Ableitung der sogenannten *Log-Likelihood-Funktion* $\ln(P_{\mathbf{y}}(\mathbf{y};\mathbf{p}))$ nach $\mathbf{p}$, d. h. sie beschreibt die (lokale) Sensitivität von $\ln(P_{\mathbf{y}}(\mathbf{y};\mathbf{p}))$ bezüglich $\mathbf{p}$. Klarerweise beeinflusst diese Sensitivität die Genauigkeit

mit der $\mathbf{p}$ basierend auf einer Realisierung der Zufallsvariable $\mathbf{y}$ geschätzt werden kann. Für den Erwartungswert der Score-Funktion gilt

$$\begin{aligned} \mathrm{E}(\mathbf{s}) &= \int \left(\frac{\partial \ln(P_{\mathbf{y}}(\mathbf{y};\mathbf{p}))}{\partial \mathbf{p}} \right)^{\mathrm{T}} P_{\mathbf{y}}(\mathbf{y};\mathbf{p}) \, \mathrm{d}\mathbf{y} \\ &= \int \frac{1}{P_{\mathbf{y}}(\mathbf{y};\mathbf{p})} \left(\frac{\partial P_{\mathbf{y}}(\mathbf{y};\mathbf{p})}{\partial \mathbf{p}} \right)^{\mathrm{T}} P_{\mathbf{y}}(\mathbf{y};\mathbf{p}) \, \mathrm{d}\mathbf{y} \\ &= \left(\frac{\partial}{\partial \mathbf{p}} \int P_{\mathbf{y}}(\mathbf{y};\mathbf{p}) \, \mathrm{d}\mathbf{y} \right)^{\mathrm{T}} = \left(\frac{\partial 1}{\partial \mathbf{p}} \right)^{\mathrm{T}} = \mathbf{0} \, . \end{aligned} \tag{3.167}$$

Hier wurde davon ausgegangen, dass die Integration über $\mathbf{y}$ und die Ableitung nach $\mathbf{p}$ in ihrer Reihenfolge vertauscht werden dürfen. Dies ist erfüllt, wenn der Ereignisraum von $\mathbf{y}$ (Integrationsgebiet) unbeschränkt ist oder sein Rand nicht von $\mathbf{p}$ abhängt.

Definition 3.3 (Fisher-Informationsmatrix)**.** Die Kovarianzmatrix der Score-Funktion

$$\begin{aligned} \mathbf{I}(\mathbf{p}) = \mathrm{E}(\mathbf{s}\mathbf{s}^{\mathrm{T}}) &= \mathrm{E}\left(\left(\frac{\partial \ln(P_{\mathbf{y}}(\mathbf{y};\mathbf{p}))}{\partial \mathbf{p}} \right)^{\mathrm{T}} \frac{\partial \ln(P_{\mathbf{y}}(\mathbf{y};\mathbf{p}))}{\partial \mathbf{p}} \right) \\ &= -\mathrm{E}\left(\frac{\partial}{\partial \mathbf{p}} \left(\frac{\partial \ln(P_{\mathbf{y}}(\mathbf{y};\mathbf{p}))}{\partial \mathbf{p}} \right)^{\mathrm{T}} \right) \end{aligned} \tag{3.168}$$

wird als Fisher-Informationsmatrix bezeichnet. Sie existiert, wenn die Score-Funktion für alle Werte $\mathbf{y}$ mit $P_{\mathbf{y}}(\mathbf{y};\mathbf{p}) > 0$ existiert und finit ist.

Dass auch die Fisher-Informationsmatrix ein Maß für den Informationsgehalt der Zufallsgröße $\mathbf{y}$ bezüglich $\mathbf{p}$ ist, wird anhand der zweiten Zeile von (3.168) klar, welche den Erwartungswert der negativen Krümmungen der Log-Likelihood-Funktion $\ln(P_{\mathbf{y}}(\mathbf{y};\mathbf{p}))$ bezüglich $\mathbf{p}$ enthält. Je größer die Werte der Fisher-Informationsmatrix sind, desto spitzer ist die Log-Likelihood-Funktion $\ln(P_{\mathbf{y}}(\mathbf{y};\mathbf{p}))$ bezüglich $\mathbf{p}$. Um die Gültigkeit der zweiten Zeile von (3.168) zu zeigen, wird (3.167) nach $\mathbf{p}$ abgeleitet

$$\begin{aligned} \mathbf{0} = \frac{\mathrm{d}\,\mathrm{E}(\mathbf{s})}{\mathrm{d}\mathbf{p}} &= \frac{\partial}{\partial \mathbf{p}} \left(\int \left(\frac{\partial \ln(P_{\mathbf{y}}(\mathbf{y};\mathbf{p}))}{\partial \mathbf{p}} \right)^{\mathrm{T}} P_{\mathbf{y}}(\mathbf{y};\mathbf{p}) \, \mathrm{d}\mathbf{y} \right) \\ &= \int \frac{\partial}{\partial \mathbf{p}} \left(\frac{\partial \ln(P_{\mathbf{y}}(\mathbf{y};\mathbf{p}))}{\partial \mathbf{p}} \right)^{\mathrm{T}} P_{\mathbf{y}}(\mathbf{y};\mathbf{p}) \, \mathrm{d}\mathbf{y} + \int \left(\frac{\partial \ln(P_{\mathbf{y}}(\mathbf{y};\mathbf{p}))}{\partial \mathbf{p}} \right)^{\mathrm{T}} \frac{\partial P_{\mathbf{y}}(\mathbf{y};\mathbf{p})}{\partial \mathbf{p}} \, \mathrm{d}\mathbf{y} \\ &= \mathrm{E}\left(\frac{\partial}{\partial \mathbf{p}} \left(\frac{\partial \ln(P_{\mathbf{y}}(\mathbf{y};\mathbf{p}))}{\partial \mathbf{p}} \right)^{\mathrm{T}} \right) \\ &\quad + \int \left(\frac{\partial \ln(P_{\mathbf{y}}(\mathbf{y};\mathbf{p}))}{\partial \mathbf{p}} \right)^{\mathrm{T}} \frac{\partial \ln(P_{\mathbf{y}}(\mathbf{y};\mathbf{p}))}{\partial \mathbf{p}} P_{\mathbf{y}}(\mathbf{y};\mathbf{p}) \, \mathrm{d}\mathbf{y} \, . \end{aligned} \tag{3.169}$$

Hier wurde wieder davon ausgegangen, dass die Integration über $\mathbf{y}$ und die Ableitung nach $\mathbf{p}$ in ihrer Reihenfolge vertauscht werden dürfen.

Satz 3.2 (Cramér-Rao-Schranke). *Es sei* $\hat{\mathbf{p}} = \mathbf{k}(\mathbf{y})$ *ein erwartungstreuer Schätzer für die Parameter* $\mathbf{p}$ *im nichtlinearen Modell* (3.95) *und es seien die folgenden Regularitätsbedingungen erfüllt.*

a) Die Fisher-Informationsmatrix existiert und ist positiv definit.

b) Die Integration über $\mathbf{y}$ *und die Ableitung nach* $\mathbf{p}$ *dürfen in ihrer Reihenfolge vertauscht werden*

Dann ist die Kovarianzmatrix des Schätzfehlers in der Form

$$\mathrm{E}((\hat{\mathbf{p}} - \mathbf{p})(\hat{\mathbf{p}} - \mathbf{p})^{\mathrm{T}}) \geq \mathbf{I}^{-1}(\mathbf{p}) \tag{3.170}$$

nach unten beschränkt, wobei das Größer-Gleich-Zeichen so zu verstehen ist, dass $\mathrm{E}((\hat{\mathbf{p}} - \mathbf{p})(\hat{\mathbf{p}} - \mathbf{p})^{\mathrm{T}}) - \mathbf{I}^{-1}(\mathbf{p})$ *eine positiv semidefinite Matrix ist.*

Der nachfolgende Beweis ist an [3.6] angelehnt. Dort wird zusätzlich eine Erweiterung der Cramér-Rao-Schranke für nicht erwartungstreue Schätzer vorgestellt.

Beweis. Die Ableitung der Bedingung (3.155) für Erwartungstreue des Schätzers nach $\mathbf{p}$ liefert

$$\begin{aligned}\int \mathbf{k}(\mathbf{y}) \frac{\partial P_{\mathbf{y}}(\mathbf{y};\mathbf{p})}{\partial \mathbf{p}} \,\mathrm{d}\mathbf{y} &= \int \mathbf{k}(\mathbf{y}) \frac{1}{P_{\mathbf{y}}(\mathbf{y};\mathbf{p})} \frac{\partial P_{\mathbf{y}}(\mathbf{y};\mathbf{p})}{\partial \mathbf{p}} P_{\mathbf{y}}(\mathbf{y};\mathbf{p}) \,\mathrm{d}\mathbf{y} \\ &= \mathrm{E}(\hat{\mathbf{p}}\mathbf{s}^{\mathrm{T}}) = \mathbf{E} \,.\end{aligned} \tag{3.171}$$

Wird davon der aus (3.167) folgende Ausdruck

$$\mathbf{p}\,\mathrm{E}(\mathbf{s}^{\mathrm{T}}) = \mathrm{E}(\mathbf{p}\mathbf{s}^{\mathrm{T}}) = \mathbf{0} \tag{3.172}$$

abgezogen, so ergibt sich

$$\mathrm{E}((\hat{\mathbf{p}} - \mathbf{p})\mathbf{s}^{\mathrm{T}}) = \mathbf{E} \,. \tag{3.173}$$

Die Matrix

$$\mathrm{E}\left(\begin{bmatrix}\hat{\mathbf{p}} - \mathbf{p} \\ \mathbf{s}\end{bmatrix}\begin{bmatrix}(\hat{\mathbf{p}} - \mathbf{p})^{\mathrm{T}} & \mathbf{s}^{\mathrm{T}}\end{bmatrix}\right) = \begin{bmatrix}\mathrm{E}((\hat{\mathbf{p}} - \mathbf{p})(\hat{\mathbf{p}} - \mathbf{p})^{\mathrm{T}}) & \mathbf{E} \\ \mathbf{E} & \mathbf{I}(\mathbf{p})\end{bmatrix} \geq 0 \tag{3.174}$$

ist natürlich positiv semidefinit, wobei hier (3.173) verwendet wurde. Die positive Semidefinitheit von (3.174) änder sich nicht, wenn der Ausdruck links- und rechtsseitig mit einer beliebigen Matrix und ihrer Transponierten multipliziert wird. Im Speziellen gilt

$$\begin{aligned}\begin{bmatrix}\mathbf{E} & -\mathbf{I}^{-1}(\mathbf{p})\end{bmatrix}\begin{bmatrix}\mathrm{E}((\hat{\mathbf{p}} - \mathbf{p})(\hat{\mathbf{p}} - \mathbf{p})^{\mathrm{T}}) & \mathbf{E} \\ \mathbf{E} & \mathbf{I}(\mathbf{p})\end{bmatrix}\begin{bmatrix}\mathbf{E} \\ -\mathbf{I}^{-1}(\mathbf{p})\end{bmatrix} \\ = \mathrm{E}((\hat{\mathbf{p}} - \mathbf{p})(\hat{\mathbf{p}} - \mathbf{p})^{\mathrm{T}}) - \mathbf{I}^{-1}(\mathbf{p}) \geq 0 \,.\end{aligned} \tag{3.175}$$

□

Ist die Fisher-Informationsmatrix zwar regulär aber numerisch schlecht konditioniert (großes Verhältnis aus maximalem zu minimalem Eigenwert von $\mathbf{I}(\mathbf{p})$, d. h. $\lambda_{\max}/\lambda_{\min} \gg 1$), so kann die Cramér-Rao-Schranke gemäß (3.170) sehr große Werte annehmen. Die Parameterwerte $\mathbf{p}$ sind dann nicht zuverlässig schätzbar [3.14]. Es ist in diesem Zusammenhang zu beachten, dass die Eigenwerte der Fisher-Informationsmatrix auch von den gewählten Einheiten der Parameter $\mathbf{p}$ abhängen.

Ein erwartungstreuer Schätzer wird als *effizient* bezeichnet, wenn seine Fehlerkovarianzmatrix die Cramér-Rao-Schranke gemäß Satz 3.2 erreicht, wenn also

$$\mathrm{E}((\hat{\mathbf{p}} - \mathbf{p})(\hat{\mathbf{p}} - \mathbf{p})^{\mathrm{T}}) = \mathbf{I}^{-1}(\mathbf{p}) \tag{3.176}$$

gilt [3.16]. Bedingungen für die Existenz eines effizienten Schätzers werden z. B. in [3.15] angegeben. Jeder effiziente Schätzer ist erwartungstreu und minimiert die aufsummierten Einzelvarianzen der Schätzfehler. Umgekehrt muss nicht jeder Schätzer, der erwartungstreu ist und die aufsummierten Einzelvarianzen der Schätzfehler minimiert, effizient sein.

3.2.5 Normalverteilte Störung

In diesem Abschnitt wird kurz der Fall betrachtet, dass die Störung $\mathbf{v} \in \mathbb{R}^m$ im Modell (3.95) normalverteilt ist mit dem Erwartungswert $\mathbf{0}$ und der Kovarianzmatrix $\mathbf{Q} \geq 0$, d. h. $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$.

Definition 3.4 (Normalverteilung)**.** Zufallsvariablen $\mathbf{z} \in \mathbb{R}^m$ genügen einer *Normalverteilung* mit dem Erwartungswert $\boldsymbol{\mu}$ und der Kovarianzmatrix $\boldsymbol{\Sigma} \geq 0$, wenn für ihre Wahrscheinlichkeitsdichtefunktion

$$P_{\mathbf{z}}(\mathbf{z}) = \frac{1}{\sqrt{(2\pi)^m \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu})\right) \tag{3.177}$$

gilt. Dafür wird die abgekürzte Notation $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ verwendet.

Satz 3.3 (Affine Transformation einer normalverteilten Zufallsvariable)**.** *Wird eine normalverteilte Zufallsvariable $\mathbf{z} \in \mathbb{R}^m$ mit $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ einer affinen Transformation*

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \mathbf{b} \tag{3.178}$$

mit zeilenregulärer Matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ und $n \leq m$ unterzogen, so ist die resultierende Zufallsvariable $\mathbf{x} \in \mathbb{R}^n$ normalverteilt mit dem Erwartungswert $\mathbf{A}\boldsymbol{\mu} + \mathbf{b}$ und der Kovarianzmatrix $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}$, d. h.

$$\mathbf{x} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}) \ . \tag{3.179}$$

Ein Beweis dieses Satzes findet sich z. B. in [3.17, 3.18].

Aufgabe 3.9. Zeigen Sie, dass im Satz 3.3 die Beziehungen $\mathrm{E}(\mathbf{x}) = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$ und $\mathrm{E}((\mathbf{x} - \mathrm{E}(\mathbf{x}))(\mathbf{x} - \mathrm{E}(\mathbf{x}))^{\mathrm{T}}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}$ gelten.

Beispiel 3.6 (Cramér-Rao-Schranke bei normalverteilter Störung). Für die Störung $\mathbf{v}$ im Modell (3.95) gilt $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ und es liegt der reguläre Fall vor (positiv definite Kovarianzmatrix $\mathbf{Q}$, spaltenreguläre Sensitivitätsmatrix $\mathbf{S}(\mathbf{p}) = (\nabla \mathbf{h})^{\mathrm{T}}(\mathbf{p})$). Es soll die Cramér-Rao-Schranke für erwartungstreue Schätzer $\hat{\mathbf{p}}$ von $\mathbf{p}$ berechnet werden.

Aus $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ folgt mit dem Modell (3.95) und dem Satz 3.3

$$\mathbf{y} \sim \mathcal{N}(\mathbf{h}(\mathbf{p}), \mathbf{Q}) \tag{3.180}$$

und somit für die zugehörige Wahrscheinlichkeitsdichtefunktion

$$P_{\mathbf{y}}(\mathbf{y}; \mathbf{p}) = \frac{1}{\sqrt{(2\pi)^m \det(\mathbf{Q})}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{h}(\mathbf{p}))^{\mathrm{T}} \mathbf{Q}^{-1} (\mathbf{y} - \mathbf{h}(\mathbf{p}))\right) \tag{3.181}$$

sowie für die Log-Likelihoodfunktion

$$\ln(P_{\mathbf{y}}(\mathbf{y}; \mathbf{p})) = -\frac{1}{2}\ln((2\pi)^m \det(\mathbf{Q})) - \frac{1}{2}(\mathbf{y} - \mathbf{h}(\mathbf{p}))^{\mathrm{T}} \mathbf{Q}^{-1} (\mathbf{y} - \mathbf{h}(\mathbf{p})) \ . \tag{3.182}$$

Weiters ergibt sich für die Score-Funktion

$$\mathbf{s} = \left(\frac{\partial \ln(P_{\mathbf{y}}(\mathbf{y}; \mathbf{p}))}{\partial \mathbf{p}}\right)^{\mathrm{T}} = (\nabla \mathbf{h})(\mathbf{p}) \mathbf{Q}^{-1} (\mathbf{y} - \mathbf{h}(\mathbf{p})) = \mathbf{S}^{\mathrm{T}}(\mathbf{p}) \mathbf{Q}^{-1} (\mathbf{y} - \mathbf{h}(\mathbf{p})) \tag{3.183}$$

und für die Fisher-Informationsmatrix

$$\mathbf{I}(\mathbf{p}) = \mathbf{S}^{\mathrm{T}}(\mathbf{p}) \mathbf{Q}^{-1} \underbrace{\mathrm{E}\Big((\mathbf{y} - \mathbf{h}(\mathbf{p}))(\mathbf{y} - \mathbf{h}(\mathbf{p}))^{\mathrm{T}}\Big)}_{\mathbf{Q}} \mathbf{Q}^{-1} \mathbf{S}(\mathbf{p}) = \mathbf{S}^{\mathrm{T}}(\mathbf{p}) \mathbf{Q}^{-1} \mathbf{S}(\mathbf{p}) \ , \tag{3.184}$$

woraus die Cramér-Rao-Schranke

$$\mathrm{E}((\hat{\mathbf{p}} - \mathbf{p})(\hat{\mathbf{p}} - \mathbf{p})^{\mathrm{T}}) \geq (\mathbf{S}^{\mathrm{T}}(\mathbf{p}) \mathbf{Q}^{-1} \mathbf{S}(\mathbf{p}))^{-1} \tag{3.185}$$

folgt. Sie stimmt also in diesem Fall mit der näherungsweisen Schranke gemäß Satz 3.1 überein. Ein Vergleich mit (3.17) zeigt, dass der BLUE Schätzer (3.16) im Falle eines linearen Modells die Cramér-Rao-Schranke erreicht, d. h. er ist effizient.

> *Bemerkung* 3.4. Basierend auf den Ergebnissen dieses Beispiels wird klar, dass die in den Beispielen 3.1 und 3.2 berechneten Gramschen Matrizen $\mathbf{G}$ im Falle von normalverteilten Messstörungen $\mathbf{v}_k$ bzw. $\mathbf{v}(t)$ gleich der Fisher-Informationsmatrix $\mathbf{I}(\mathbf{x}_0)$ sind und dass die in diesen Beispielen berechneten BLUE Schätzer für den Anfangszustand $\mathbf{x}_0$ die Cramér-Rao-Schranke erreichen, d. h. effizient sind.

Verteilung eines linearen Schätzers

Es soll nun geklärt werden, welche statistische Verteilung ein linearer Schätzer $\hat{\mathbf{p}} = \mathbf{K}\mathbf{y}$ für $\mathbf{p}$ bei normalverteilter Störung $\mathbf{v}$ besitzt. Aus $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ folgt mit dem nichtlinearen

Modell (3.95) wieder

$$\mathbf{y} \sim \mathcal{N}(\mathbf{h}(\mathbf{p}), \mathbf{Q}) \ . \tag{3.186}$$

Die Anwendung des Satzes 3.3 auf den linearen Schätzer $\hat{\mathbf{p}} = \mathbf{K}\mathbf{y}$ zeigt, dass $\hat{\mathbf{p}}$ normalverteilt mit dem Erwartungswert $\mathbf{K}\mathbf{h}(\mathbf{p})$ und der Kovarianzmatrix $\mathbf{K}\mathbf{Q}\mathbf{K}^{\mathrm{T}}$ ist, d. h.

$$\hat{\mathbf{p}} \sim \mathcal{N}(\mathbf{K}\mathbf{h}(\mathbf{p}), \mathbf{K}\mathbf{Q}\mathbf{K}^{\mathrm{T}}) \ . \tag{3.187}$$

Speziell ergibt sich im Falle des linearen Modells (3.1), also für $\mathbf{h}(\mathbf{p}) = \mathbf{S}\mathbf{p}$, und bei Verwendung der zeilenregulären Matrix $\mathbf{K} = (\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}$ (BLUE Schätzer gemäß (3.16))

$$\hat{\mathbf{p}} \sim \mathcal{N}(\mathbf{p}, (\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1}) \tag{3.188}$$

(siehe dazu auch (3.17)).

Vertrauensbereich für die zu schätzenden Parameter

Bislang wurden verschiedene Möglichkeiten der Konstruktion eines Schätzers $\hat{\mathbf{p}}$ für die unbekannten Systemparameter $\mathbf{p} \in \mathbb{R}^n$ vorgestellt. In der Regel weicht $\hat{\mathbf{p}}$ vom wahren Wert $\mathbf{p}$ ab. Es kann daher nützlich sein, alternativ (oder in Ergänzung) zu einem Schätzer $\hat{\mathbf{p}}$ ein (möglichst eingeschränktes) Gebiet $K(\mathbf{y})$ im Parameterraum $\mathbb{R}^n$ anzugeben, von dem behauptet werden kann, dass es die unbekannten Systemparameter $\mathbf{p}$ mit vorgebener Wahrscheinlichkeit enthält. Ein solches Gebiet wird mit der nachfolgenden Definition angegeben.

Definition 3.5. Ein Gebiet $K(\mathbf{y}) \subseteq \mathbb{R}^n$ in Abhängigkeit der zufälligen Ausgangswerte $\mathbf{y}$ wird als *Vertrauensbereich* (Konfidenzbereich) für die zu schätzenden Parameter $\mathbf{p}$ mit dem *Niveau* $1-\alpha$ bezeichnet, wenn die Wahrscheinlichkeit, dass $K(\mathbf{y})$ den Punkt $\mathbf{p}$ enthält, (unabhängig von den konkreten Werten $\mathbf{p}$) die Bedingung

$$P(\mathbf{p} \in K(\mathbf{y})) \geq 1 - \alpha \tag{3.189}$$

erfüllt. Der Wert $\alpha \in [0, 1]$ wird als *Irrtumswahrscheinlichkeit* bezeichnet.

Für eine detailliertere Definition wird auf [3.18, 3.19] verwiesen. Praktisch wird die Irrtumswahrscheinlichkeit α häufig im einstelligen Prozentbereich gewählt. Für ein gegebenens Niveau $1-\alpha$ ist das Gebiet $K(\mathbf{y})$ nicht eindeutig, da verschiedene Berechnungsvorschriften $K(\mathbf{y})$ die Bedingung (3.189) erfüllen können. Während $\hat{\mathbf{p}}$ eine *Punktschätzung* für die unbekannten Parameter $\mathbf{p}$ realisiert, kann $K(\mathbf{y})$ als *Bereichsschätzung* für $\mathbf{p}$ mit Irrtumswahrscheinlichkeit α verstanden werden. Es sei betont, dass $\mathbf{p}$ eine *deterministische* Größe und $K(\mathbf{y})$ ein *zufälliges* Gebiet ist.

Für den Fall einer normalverteilten Störung $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ wird nun angelehnt an [3.20] eine konkrete Berechnungsvorschrift für einen Vertrauensbereich $K(\mathbf{y})$ angegeben.

Definition 3.6 (Mahalanobis-Distanz)**.** Gilt für Zufallsvariablen $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ mit $\mathbf{z} \in \mathbb{R}^n$, so wird die Größe

$$d(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sqrt{(\mathbf{z} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu})} \tag{3.190}$$

(vgl. den Exponenten in (3.177)) als *Mahalanobis-Distanz* bezeichnet und ihr Quadrat $d^2(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ genügt der sogenannten χ^2_n-Verteilung mit n Freiheitsgraden[4].

Es sei $\chi^2_n(\alpha)$ das α-Quantil der χ^2_n-Verteilung, d. h. für χ^2_n-verteilte Zufallsvariable x gilt $P(x \leq \chi^2_n(\alpha)) = \alpha$. Definitionen der χ^2_n-Verteilung, ihrer Wahrscheinlichkeitsdichtefunktion und ihrer Verteilungsfunktion sind z. B. in [3.18, 3.19] zu finden.

Gemäß (3.187) genügt ein linearer erwartungstreuer Schätzer $\hat{\mathbf{p}} = \mathbf{K}\mathbf{y}$ der Normalverteilung

$$\hat{\mathbf{p}} \sim \mathcal{N}(\mathbf{p}, \mathbf{K}\mathbf{Q}\mathbf{K}^{\mathrm{T}}) \ . \tag{3.191}$$

Folglich ist die Mahalanobis-Distanz

$$d(\hat{\mathbf{p}}, \mathbf{p}, \mathbf{K}\mathbf{Q}\mathbf{K}^{\mathrm{T}}) \ , \tag{3.192}$$

welche natürlich ebenfalls eine Zufallszahl ist, ein skalares Maß für den Schätzfehler $\hat{\mathbf{p}} - \mathbf{p}$. Im Parameterraum $\mathbb{R}^n$ bilden alle Punkte $\tilde{\mathbf{p}}$ mit einer Mahalanobis-Distanz $d(\tilde{\mathbf{p}}, \mathbf{p}, \mathbf{K}\mathbf{Q}\mathbf{K}^{\mathrm{T}}) \leq \bar{d}$ ein Hyperellipsoid mit dem Mittelpunkt $\mathbf{p}$. Für eine bestimmte Distanz $\bar{d} \in [0, \infty)$ lässt sich diese Punktmenge sofort in der Form

$$\left\{ \tilde{\mathbf{p}} \in \mathbb{R}^n \,\middle|\, \sqrt{(\tilde{\mathbf{p}} - \mathbf{p})^{\mathrm{T}} (\mathbf{K}\mathbf{Q}\mathbf{K}^{\mathrm{T}})^{-1} (\tilde{\mathbf{p}} - \mathbf{p})} \leq \bar{d} \right\} \tag{3.193}$$

angeben. Die Wahrscheinlichkeit, dass $\hat{\mathbf{p}}$ in dieser Menge liegt, entspricht der Wahrscheinlichkeit, dass $d(\hat{\mathbf{p}}, \mathbf{p}, \mathbf{K}\mathbf{Q}\mathbf{K}^{\mathrm{T}}) \leq \bar{d}$ gilt[5]. Da nun $d^2(\hat{\mathbf{p}}, \mathbf{p}, \mathbf{K}\mathbf{Q}\mathbf{K}^{\mathrm{T}}) = d^2(\mathbf{p}, \hat{\mathbf{p}}, \mathbf{K}\mathbf{Q}\mathbf{K}^{\mathrm{T}}) \sim \chi^2_n$ gilt, stellt die Menge

$$K(\mathbf{y}) = \left\{ \tilde{\mathbf{p}} \in \mathbb{R}^n \,\middle|\, (\mathbf{K}\mathbf{y} - \tilde{\mathbf{p}})^{\mathrm{T}} (\mathbf{K}\mathbf{Q}\mathbf{K}^{\mathrm{T}})^{-1} (\mathbf{K}\mathbf{y} - \tilde{\mathbf{p}}) \leq \chi^2_n(1 - \alpha) \right\} \tag{3.194}$$

einen Vertrauensbereich für $\mathbf{p}$ mit dem Niveau $1 - \alpha$ dar. Diese Menge definiert im Parameterraum ein Hyperellipsoid mit dem Mittelpunkt $\mathbf{K}\mathbf{y}$. Für den Fall $n = 2$ veranschaulicht Abbildung 3.4 den Vertrauensbereich für $\mathbf{p}$ anhand von zwei Realisierungen der Zufallsvariablen $\mathbf{y}$.

Wird im Falle des linearen Modells (3.1), d. h. $\mathbf{h}(\mathbf{p}) = \mathbf{S}\mathbf{p}$, die zeilenreguläre Matrix $\mathbf{K} = (\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}$ verwendet (BLUE Schätzer gemäß (3.16)), so vereinfacht sich (3.194) zu

$$K(\mathbf{y}) = \left\{ \tilde{\mathbf{p}} \in \mathbb{R}^n \,\middle|\, (\mathbf{K}\mathbf{y} - \tilde{\mathbf{p}})^{\mathrm{T}} (\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}) (\mathbf{K}\mathbf{y} - \tilde{\mathbf{p}}) \leq \chi^2_n(1 - \alpha) \right\} . \tag{3.195}$$

[4] Um dies zu sehen, kann $\mathbf{z}$ unter Berücksichtigung von Satz 3.3 auf die standardnormalverteilten Zufallsvariablen

$$\mathbf{Q}^{-1/2}(\mathbf{z} - \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}, \mathbf{E})$$

transformiert werden. Die Zufallsvariable $d^2(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ entspricht folglich den aufsummierten Quadraten von n unabhängigen standardnormalverteilten Zufallsvariablen und ist auf dem Intervall $[0, \infty)$ definiert. Eine solche Summe genügt per Definition der χ^2_n-Verteilung mit n Freiheitsgraden und ihr Erwartungswert ist n.

[5] Klarerweise gilt auf der durch $d(\tilde{\mathbf{p}}, \mathbf{p}, \mathbf{K}\mathbf{Q}\mathbf{K}^{\mathrm{T}}) = \bar{d}$ definierten Oberfläche, dass die Wahrscheinlichkeitsdichtefunktion $P_{\hat{\mathbf{p}}}(\hat{\mathbf{p}})$ konstant ist.

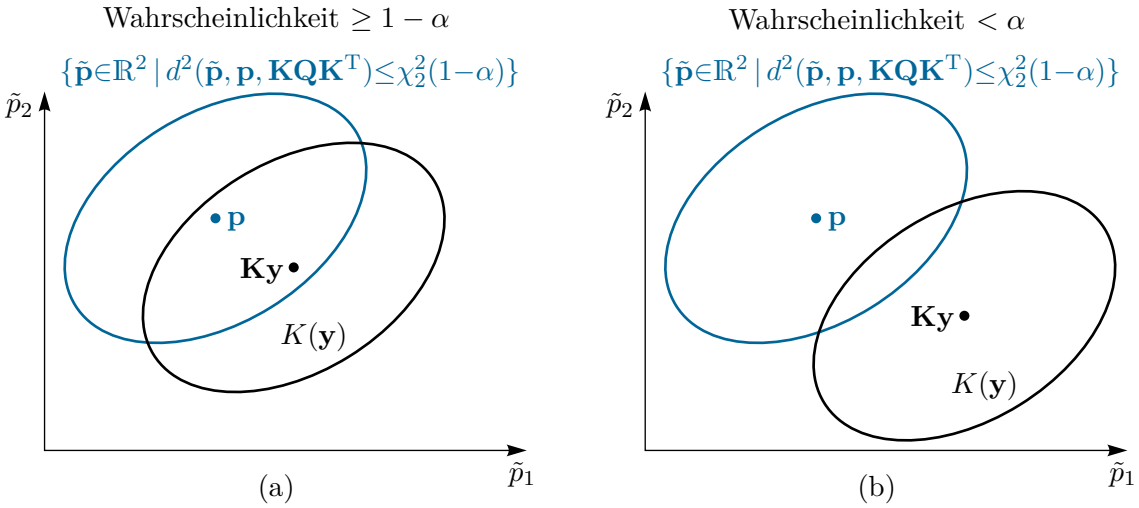


Abbildung 3.4: Vertrauensbereich für $\mathbf{p} \in \mathbb{R}^2$ mit dem Niveau $1-\alpha$, (a) korrekte Bereichsschätzung: $\mathbf{p} \in K(\mathbf{y})$, (b) irrtümliche Bereichsschätzung: $\mathbf{p} \notin K(\mathbf{y})$.

3.3 Optimale Versuchsplanung

Die beim Entwurf eines Schätzers für Systemparameter $\mathbf{p} \in \mathbb{R}^n$ erreichbare Schätzgüte hängt erheblich von den verfügbaren Ausgangswerten $\mathbf{y} \in \mathbb{R}^m$ und deren Messgenauigkeit ab. Die *optimale Versuchsplanung* (englisch: *optimal design of experiments*) [3.21–3.23] verfolgt nun das Ziel, die Beschaffung von Ausgangswerten (im Versuch oder im laufenden Betrieb eines Systems) so zu planen, dass die erreichbare Schätzgüte maximiert wird. Diese Planungsaufgabe kann z. B. folgende Entwurfsfreiheitsgrade umfassen:

- Auswahl von physikalischen Messgrößen
- Zeiträume und Zeitpunkte von Messungen
- Type, Präzision, Anzahl und örtliche Positionierung von Sensoren
- Aufbau, Funktion, Konfiguration und Größe des Systems bzw. Versuchs
- Anregung und Betriebszustand des Systems (Anfangsbedingungen, Randbedingungen, Eingangstrajektorie)

Natürlich wird die erreichbare Schätzgüte auch von der Wahl des Modells (siehe (3.1) oder (3.95)) und der Auswahl an zu schätzenden Systemparametern $\mathbf{p}$ beeinflusst. Alle weiteren nicht in $\mathbf{p}$ enthaltenen Systemparameter müssen anderweitig festgelegt werden, z. B. basierend auf Vorwissen, speziellen Messungen, Datenblättern, Literatur, etc.

Als Beurteilungskriterium für die erreichbare Schätzgüte wird meist eine untere Schranke $\mathbf{R}^{-1}$ für die Kovarianzmatrix des Parameterschätzfehlers $\mathrm{E}((\hat{\mathbf{p}}-\mathbf{p})(\hat{\mathbf{p}}-\mathbf{p})^{\mathrm{T}})$ in der Form

$$\mathrm{E}((\hat{\mathbf{p}}-\mathbf{p})(\hat{\mathbf{p}}-\mathbf{p})^{\mathrm{T}}) \geq \mathbf{R}^{-1} \tag{3.196}$$

oder eine daraus abgeleitete skalare Funktion $r(\mathbf{R})$ herangezogen. Konkret wird bei der Planung die (nichtlineare) Optimierungsaufgabe

$$\min \; r(\mathbf{R}) \tag{3.197}$$

formuliert und gelöst. Das Lösen von (3.197) kann numerisch anspruchsvoll sein.

Welche Matrix?

Für $\mathbf{R}$ kann gemäß den Sätzen 3.1 und 3.2

$$\mathbf{R} = \mathbf{S}^{\mathrm{T}}(\mathbf{p})\mathbf{Q}^{-1}\mathbf{S}(\mathbf{p}) \tag{3.198a}$$

oder

$$\mathbf{R} = \mathbf{I}(\mathbf{p}) \tag{3.198b}$$

verwendet werden.

Für die Wahl (3.198b) spricht, dass es sich bei $\mathbf{I}^{-1}(\mathbf{p})$ um eine rigorose Schranke handelt. Die Berechnung der Fisher-Informationsmatrix setzt jedoch die Kenntnis der Wahrscheinlichkeitsdichtefunktion $P_{\mathbf{y}}(\mathbf{y};\mathbf{p})$ voraus.

Die Wahl (3.198a) kann wie folgt motiviert werden:

- $\mathbf{S}^{\mathrm{T}}(\mathbf{p})\mathbf{Q}^{-1}\mathbf{S}(\mathbf{p})$ kann einfach und ohne Kenntnis von Wahrscheinlichkeitsdichtefunktionen berechnet werden.
- Für ein lineares Modell (3.1) und den BLUE Schätzer (3.16) entspricht $(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1}$ exakt der Kovarianzmatrix des Parameterschätzfehlers (siehe (3.17)). Im Optimierungsproblem der dazu korrespondierenden gewichteten linearen Least-Squares Methode (siehe Aufgabe 3.1) entspricht $\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}$ der Hessematrix der Kostenfunktion.
- In der gewichteten nichtlinearen Least-Squares Optimierungsaufgabe (3.112) mit der Gewichtungsmatrix $\mathbf{W} = \mathbf{Q}^{-1}$ entspricht $\mathbf{S}^{\mathrm{T}}(\mathbf{p})\mathbf{Q}^{-1}\mathbf{S}(\mathbf{p})$ der Gauss-Newton Approximation der Hessematrix der Kostenfunktion.
- Wie in Beispiel 3.6 gezeigt, gilt bei normalverteilter Störung für die Fisher-Informationsmatrix $\mathbf{I}(\mathbf{p}) = \mathbf{S}^{\mathrm{T}}(\mathbf{p})\mathbf{Q}^{-1}\mathbf{S}(\mathbf{p})$, d. h. (3.198a) und (3.198b) sind identisch.
- Wie in Abschnitt 3.2.5 gezeigt, gilt bei einem linearen Modell (3.1) mit normalverteilter Störung und Verwendung des BLUE Schätzers (3.16) für den Schätzwert $\hat{\mathbf{p}} \sim \mathcal{N}(\mathbf{p}, (\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1})$ und es kann für $\mathbf{p}$ der Vertrauensbereich

  $$K(\mathbf{y}) = \left\{ \tilde{\mathbf{p}} \in \mathbb{R}^n \;\middle|\; (\mathbf{K}\mathbf{y} - \tilde{\mathbf{p}})^{\mathrm{T}}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})(\mathbf{K}\mathbf{y} - \tilde{\mathbf{p}}) \leq \chi^2_n(1-\alpha) \right\} . \tag{3.199}$$

  mit dem Niveau $1-\alpha$ angegeben werden. $K(\mathbf{y})$ definiert im Parameterraum $\mathbb{R}^n$ ein Hyperellipsoid, dessen Größe, Form und Orientierung ausschließlich durch $\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}$ festgelegt sind. Natürlich soll dieses Hyperellipsoid möglichst klein sein.

Alternativ zu (3.198) kann $\mathbf{R}$ im Sinne einer Minimierung der in Abschnitt 3.2.3 besprochenen Kollinearität (allerdings ohne direkten Bezug zur Kovarianzmatrix des Schätzfehlers) in der Form

$$\mathbf{R} = \tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}} \tag{3.200}$$

mit (dem zumeist von $\mathbf{p}$ abhängigen) $\tilde{\mathbf{S}}$ gemäß (3.132) gewählt werden. Diese Wahl hat gegenüber (3.198) den Vorteil, dass $\mathbf{R}$ hier nicht mehr von den Einheiten bzw. der Skalierung der Parameter $\mathbf{p}$ abhängt.

Die Berechnung von $\mathbf{R}$ kann *ohne* Kenntnis der Ausgangswerte $\mathbf{y}$ erfolgen, also bereits vorab. Bei dieser Berechnung werden jedoch (im Allgemeinen) die *unbekannten* (zu schätzenden) Systemparameter $\mathbf{p}$ benötigt. Ein gängiger Ausweg aus diesem Dilemma ist hier die Verwendung von Schätzwerten für $\mathbf{p}$ anstatt der wahren Werten $\mathbf{p}$. Die Optimierungsaufgabe (3.197) und ihre Lösung sind daher nur lokal in der Nähe der verwendeten Parameterwerte gültig [3.21]. Gegebenenfalls sind die optimale Versuchsplanung und die Schätzung von $\mathbf{p}$ iterativ zu wiederholen.

Es soll noch kurz die Bedeutung der *Eigenwerte* und *Eigenvektoren* der Matrix $\mathbf{R}$ diskutiert werden. Die Matrix $\mathbf{R}$ und ihre Inverse $\mathbf{R}^{-1}$ sind symmetrisch und somit diagonalisierbar. Folglich sind ihre Eigenvektoren orthogonal. Es seien λ_i mit $i = 1, \ldots, n$ die Eigenwerte der Matrix $\mathbf{R}$. Somit sind $1/\lambda_i$ die Eigenwerte der Matrix $\mathbf{R}^{-1}$. Die implizite Gleichung

$$\mathbf{p}^{\mathrm{T}}\mathbf{R}\mathbf{p} = c^2 \tag{3.201}$$

definiert im Parameterraum $\mathbb{R}^n$ ein Hyperellipsoid, dessen Halbachsen die Längen $c/\sqrt{\lambda_i}$ aufweisen und entlang der zugehörigen Eigenvektoren von $\mathbf{R}$ ausgerichtet sind. Bei Verwendung von (3.198) folgt die Interpretation von Hyperellipsoiden (für verschiedene Werte c) im Sinne der Kovarianzmatrix des Schätzfehlers bzw. eines Vertrauensbereiches für $\mathbf{p}$ direkt aus den vorangegangenen Überlegungen. Bei Verwendung von (3.200) können diese Hyperellipsoide (für verschiedene Werte c) im Sinne der Kollinearität interpretiert werden. Ohne Beschränkung der Allgemeinheit kann $c = 1$ gesetzt werden, so dass für die weitere geometrische Interpretation das durch

$$\mathbf{p}^{\mathrm{T}}\mathbf{R}\mathbf{p} = 1 \tag{3.202}$$

bestimmte Ellipsoid herausgegriffen wird.

Welches Beurteilungskriterium?

Je nach Wahl der Funktion $r(\mathbf{R})$ in der Optimierungsaufgabe (3.197) können nun folgende Beurteilungskriterien zur optimalen Versuchsplanung unterschieden werden. Weitere Beurteilungskriterien und Details zu den nachfolgend angegebenen Funktionen sind z. B. in [3.14, 3.21, 3.23] zu finden. So nicht anders angegeben, wird in den nachfolgenden Interpretationen von der Verwendung von (3.198) ausgegangen.

- **A-Optimalität:** Es wird die Funktion

$$r(\mathbf{R}) = \mathrm{spur}(\mathbf{R}^{-1}) = \sum_{i=1}^{n} \frac{1}{\lambda_i} \tag{3.203}$$

minimiert, wobei λ_i die Eigenwerte der Matrix $\mathbf{R}$ sind. Damit wird versucht, die Summe (oder gleichwertig das arithmetische Mittel, A steht für average) der Einzelvarianzen der Parameterschätzfehler (siehe z. B. (3.4)) zu minimieren. Geometrisch

kann dies so gedeutet werden, dass die pythagoreische Summe der Halbachsenlängen des in (3.202) definierten Hyperellipsoids minimiert wird. Dies entspricht einer Minimierung der Diagonalenlänge jenes kleinstmöglichen Hyperquaders der das Hyperellipsoid umschreibt und dessen Kanten parallel zu den Achsen des Hyperellipsoids sind.

- **D-Optimalität:** Es wird die Funktion

$$r(\mathbf{R}) = \det(\mathbf{R}^{-1}) = \frac{1}{\det(\mathbf{R})} = \prod_{i=1}^{n} \frac{1}{\lambda_i} \tag{3.204}$$

minimiert. Die Determinante einer Kovarianzmatrix wird als *generalisierte Varianz* bezeichnet. Die D-Optimalität (D steht für Determinante) versucht daher die generalisierte Varianz des Parameterschätzfehlers zu minimieren. Geometrisch kann dies so gedeutet werden, dass der Rauminhalt des in (3.202) definierten Hyperellipsoids minimiert wird. Dies entspricht einer Minimierung des Rauminhalts jenes kleinstmöglichen Hyperquaders der das Hyperellipsoid umschreibt und dessen Kanten parallel zu den Achsen des Hyperellipsoids sind. Die D-Optimialität hat den Nachteil, dass sie gute (kleine) Werte für $r(\mathbf{R})$ liefert sobald die Länge einer Halbachse des Hyperellipsoids klein wird. Die übrigen Halbachsenlängen können weiterhin groß sein und damit eine hohe Varianz von Parameterschätzfehlern erlauben.

- **E-Optimalität:** Es wird die Funktion

$$r(\mathbf{R}) = \max_i \frac{1}{\lambda_i} = \frac{1}{\min_i \lambda_i} \tag{3.205}$$

minimiert. Damit wird versucht, die größtmögliche (E steht für extreme) Varianz des Parameterschätzfehlers (entspricht der Varianz der am unzuverlässigsten schätzbaren Linearkombination $\mathbf{t}^{\mathrm{T}}\mathbf{p}$ mit $\|\mathbf{t}\|_2 = 1$) zu minimieren. Geometrisch kann dies so gedeutet werden, dass die Länge der längsten Halbachse des in (3.202) definierten Hyperellipsoids minimiert wird. Wird $\mathbf{R} = \tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}$ verwendet, so entspricht die E-Optimalität einer Minimierung der Größe $1/\rho$ aus (3.134).

- **Erweiterte E-Optimalität:** Es wird die Funktion

$$r(\mathbf{R}) = \frac{\max_i \lambda_i}{\min_i \lambda_i} \tag{3.206}$$

minimiert[6]. Dies entspricht dem Verhältnis aus maximalem zu minimalem Eigenwert der Matrix $\mathbf{R}$ und wird spektrale Konditionszahl von $\mathbf{R}$ genannt [3.24]. Damit wird versucht, die Varianzen aller Parameterschätzfehler möglichst aneinander anzugleichen. Geometrisch kann dies so gedeutet werden, dass das in (3.202) definierte Hyperellipsoid möglichst zu einer Kugel wird. Es ist zu beachten, dass hierbei keine Rücksicht auf das Ausmaß der Varianzen bzw. die Größe der Kugel genommen wird. Bei der Verwendung von $\mathbf{R} = \tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}$ ist dies unkritisch, da $\tilde{\mathbf{S}}$ die normierten Spalten von $\mathbf{S}$ enthält.

[6] Natürlich gilt dann $r(\mathbf{R}) = r(\mathbf{R}^{-1})$.

Beispiel 3.7 (Optimale Anordnung von Sensoren). Ein Diffusionsprozess führt zu einem stationären Konzentrationsprofil $u(x)$ im Bereich $x \in [0,1]$, welches durch das Zweipunktrandwertproblem

$$0 = \frac{\mathrm{d}^2}{\mathrm{d}x^2} u(x) + 6p_1 x + 2p_2 \qquad x \in [0,1] \tag{3.207a}$$

mit homogenen Randbedingungen

$$u(0) = u(1) = 0 \tag{3.207b}$$

definiert ist. Es sollen die Positionen $x_1, \ldots, x_m$ von m Sensoren für die Größen $u(x_1), \ldots, u(x_m)$ so optimiert werden, dass aus den Messwerten bestmöglich die unbekannten Modellparameter $\mathbf{p} = \begin{bmatrix} p_1 & p_2 \end{bmatrix}^{\mathrm{T}}$ geschätzt werden können. Die Sensoren sind identisch und ihre Messwerte werden von statistisch unabhängigen Störungen $v_1, \ldots, v_m$ mit Erwartungswert 0 verfälscht.

Die Lösung von (3.207) lautet

$$u(x) = \begin{bmatrix} x(1-x)(1+x) & x(1-x) \end{bmatrix} \mathbf{p} \ , \tag{3.208}$$

so dass sich die Ausgangswerte

$$\mathbf{y} = \mathbf{S}\mathbf{p} + \mathbf{v} \tag{3.209a}$$

mit

$$\mathbf{S} = \begin{bmatrix} x_1(1-x_1)(1+x_1) & x_1(1-x_1) \\ \vdots & \vdots \\ x_m(1-x_m)(1+x_m) & x_m(1-x_m) \end{bmatrix} \tag{3.209b}$$

ergeben. Da es sich um identische Sensoren handelt und auf sie statistisch unabhängige Störungen wirken, ist die Annahme $\mathrm{E}(\mathbf{v}\mathbf{v}^{\mathrm{T}}) = \sigma^2 \mathbf{E}$ mit einer beliebigen unbekannten Varianz σ^2 gerechtfertigt. Daher wird

$$\mathbf{R} = \mathbf{S}^{\mathrm{T}}\mathbf{S} \tag{3.210}$$

in der Optimierungsaufgabe

$$\min_{\mathbf{x} \in [0,1]^m} \ r(\mathbf{R}) \tag{3.211}$$

(siehe (3.197)) zur optimalen Anordnung der Sensoren verwendet.

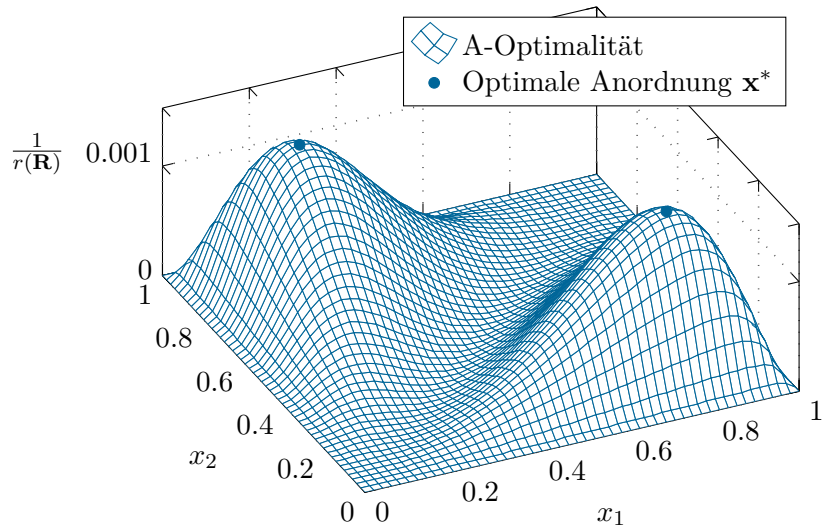


Abbildung 3.5: Kehrwert der Beurteilungsfunktion $r(\mathbf{R})$ für A-Optimalität und $m = 2$ Sensoren.

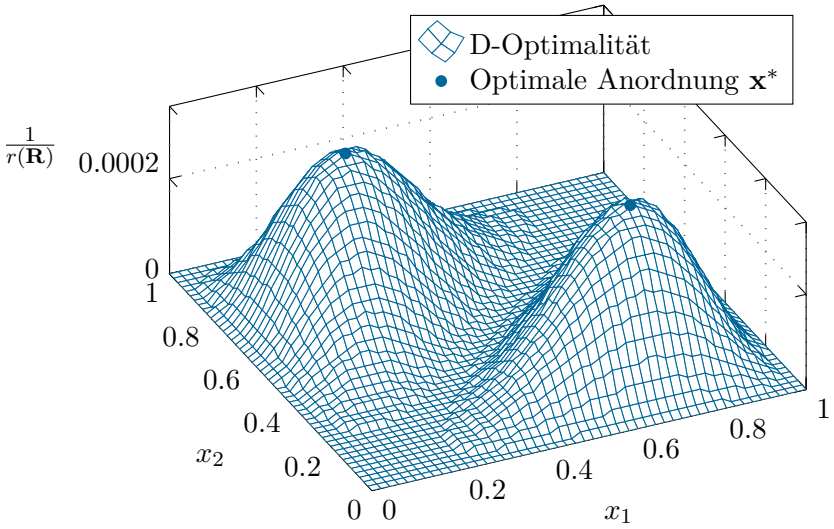


Abbildung 3.6: Kehrwert der Beurteilungsfunktion $r(\mathbf{R})$ für D-Optimalität und $m = 2$ Sensoren.

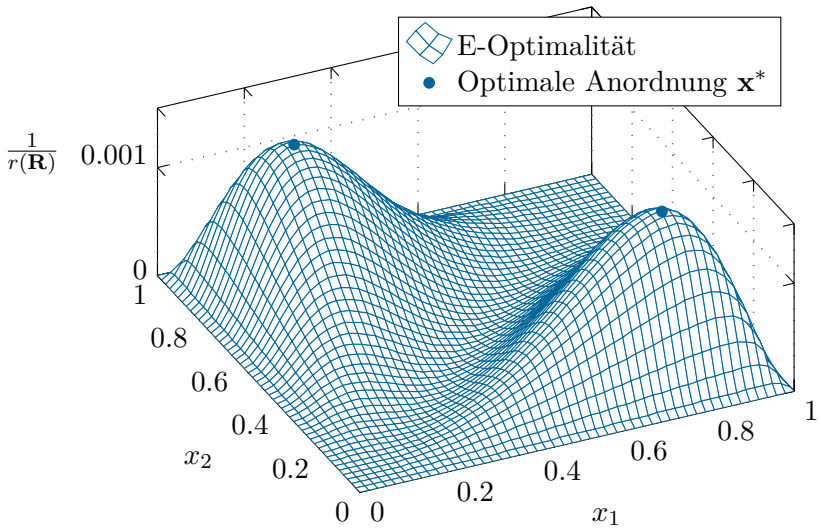


Abbildung 3.7: Kehrwert der Beurteilungsfunktion $r(\mathbf{R})$ für E-Optimalität und $m = 2$ Sensoren.

Für $m = 2$ Sensoren und Verwendung der Kriterien A-, D- und E-Optimalität sind die Werte $r(\mathbf{R})$ in den Abbildungen 3.5 bis 3.7 gezeigt. Konkret sind hier für übersichtlichere Zahlenwerte die (zu maximierenden) Kehrwerte $1/r(\mathbf{R})$ dargestellt. Wenn im Fall $m = 2$ zumindest ein Sensor an einem der Ränder 0 oder 1 positioniert ist, so ist $\mathbf{S}$ nicht spaltenregulär und es gilt $1/r(\mathbf{R}) \to 0$.

Die Optimierungsaufgabe (3.211) kann numerisch, z. B. mit der MATLAB-Funktion fmincon, gelöst werden. Die daraus folgenden (global) optimalen Positionen der Sensoren sind in Tabelle 3.2 angegeben. Für $m > 2$ existieren in der Regel weitere lokale Optima.

Anzahl an Sensoren m	Optimale Sensorpositionen $\mathbf{x}^*$ A-Optimalität	D-Optimalität	E-Optimalität
2	$\begin{bmatrix}0.221\,46\\0.797\,60\end{bmatrix}$	$\begin{bmatrix}0.276\,39\\0.723\,61\end{bmatrix}$	$\begin{bmatrix}0.220\,84\\0.798\,51\end{bmatrix}$
3	$\begin{bmatrix}0.196\,39\\0.196\,39\\0.772\,96\end{bmatrix}$	$\begin{bmatrix}0.276\,39\\0.276\,39\\0.723\,61\end{bmatrix}$	$\begin{bmatrix}0.195\,63\\0.195\,63\\0.773\,79\end{bmatrix}$
4	$\begin{bmatrix}0.221\,46\\0.221\,46\\0.797\,60\\0.797\,60\end{bmatrix}$	$\begin{bmatrix}0.276\,39\\0.276\,39\\0.723\,61\\0.723\,61\end{bmatrix}$	$\begin{bmatrix}0.220\,84\\0.220\,84\\0.798\,51\\0.798\,51\end{bmatrix}$
5	$\begin{bmatrix}0.207\,02\\0.207\,02\\0.207\,02\\0.782\,89\\0.782\,89\end{bmatrix}$	$\begin{bmatrix}0.276\,39\\0.276\,39\\0.276\,39\\0.723\,61\\0.723\,61\end{bmatrix}$	$\begin{bmatrix}0.206\,32\\0.206\,32\\0.206\,32\\0.783\,76\\0.783\,76\end{bmatrix}$

Tabelle 3.2: Optimale Sensorpositionen.

Werden in diesem Beispiel $m > 2$ Sensoren verwendet, so werden stets alle Sensoren auf nur zwei verschiedene optimale Positionen aufgeteilt. Für gerade Werte m sind das stets die selben Positionen. Daraus folgt als bevorzugte Versuchsvariante die Verwendung von zwei Sensoren und deren wiederholtes Auslesen.

Für dieses Beispiel liefern A- und E-Optimalität sehr ähnliche Ergebnisse. Die D-Optimalität führt unabhängig von m immer zu den selben optimalen Sensorpositionen. Für dieses Beispiel sollte weder die erweiterte E-Optimalität noch die Matrix $\mathbf{R} = \tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}$ verwendet werden, da sich damit optimale Sensorpositionen beliebig nahe an den Rändern 0 und 1 ergeben. Je näher Sensoren an diesen Rändern platziert werden, desto kleiner ist die Sensitivität ihrer Messwerte bezüglich $\mathbf{p}$. Folglich ist eine Schätzung von $\mathbf{p}$ basierend auf diesen Messwerten mit größeren Fehlern (Varianzen) verbunden. Die Absolutwerte dieser Sensitivitäten spielen in der erweiterten E-Optimalität aufgrund der Verhältnisbildung in (3.206) keine Rolle. Ebenso spielen sie in $\tilde{\mathbf{S}}$ keine Rolle, da $\tilde{\mathbf{S}}$ die normierten Spalten von $\mathbf{S}$ enthält.

3.4 Literatur

[3.1] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. New Jersey: Prentice-Hall, 1993, Bd. 1.

[3.2] C. Byrne, *Signal Processing: A Mathematical Approach*, 2. Aufl. Boca Raton: CRC Press, 2015.

[3.3] D. Sengupta und S. Jammalamadaka, *Linear Models: An Integrated Approach* (Series on Multivariate Analysis). New Jersey: World Scientific, 2003, Bd. 6.

[3.4] W. Kemmetmüller, *Skriptum zur VO Regelungssysteme (WS 2023/2024)*, Institut für Automatisierungs- und Regelungstechnik, TU Wien, 2023. Adresse: https://www.acin.tuwien.ac.at/master/regelungssysteme/.

[3.5] Y. Eldar und A. Oppenheim, „Covariance shaping least-squares estimation," *IEEE Transactions on Signal Processing*, Jg. 51, Nr. 3, S. 686–697, März 2003.

[3.6] T. Söderström und P. Stoica, *System Identification*. New York: Prentice Hall, 1989.

[3.7] A. Steinböck, *Skriptum zur VU Optimierung (WS 2023/2024)*, Institut für Automatisierungs- und Regelungstechnik, TU Wien, 2023. Adresse: https://www.acin.tuwien.ac.at/master/optimierung/.

[3.8] S. Kay, *Intuitive Probability and Random Processes using MATLAB*, 1. Aufl. New York: Springer, 2005.

[3.9] D. Montgomery und G. Runger, *Applied Statistics and Probability for Engineers*, 7. Aufl. New York: John Wiley & Sons, 2018.

[3.10] K. Ramachandran und C. Tsokos, *Mathematical Statistics with Applications*. Amsterdam: Academic Press, 2009.

[3.11] D. Belsley, E. Kuh und R. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Hoboken: John Wiley & Sons, 2004.

[3.12] R. Brun, P. Reichert und H. Künsch, „Practical identifiability analysis of large environmental simulation models," *Water Resources Research*, Jg. 37, Nr. 4, S. 1015–1030, 2001.

[3.13] P. Tan, M. Steinbach, A. Karpatne und V. Kumar, *Introduction to Data Mining*, 2. Aufl. New York: Pearson, 2019.

[3.14] K. McLean und K. McAuley, „Mathematical modelling of chemical processes—Obtaining the best model predictions and parameter estimates using identifiability and estimability procedures," *The Canadian Journal of Chemical Engineering*, Jg. 90, Nr. 2, S. 351–366, 2012.

[3.15] F. Hlawatsch, *Skriptum zur VO Parameter Estimation Methods (SS 2023)*, Institute of Telecommunications, TU Wien, 2023.

[3.16] J. Norton, *An Introduction to Identification*. London: Academic Press, 1988.

[3.17] Y. Tong, *The Multivariate Normal Distribution* (Springer Series in Statistics). New York: Springer, 1990.

[3.18] H. Georgii, *Stochastik - Einführung in die Wahrscheinlichkeitstheorie und Statistik*, 5. Aufl. Berlin: De Gruyter, 2015.

[3.19] K. Wichmann, *Auswertung von Messdaten - Statistische Methoden für Geo- und Ingenieurwissenschaften*. München: Oldenbourg, 2007.

[3.20] B. Wang, W. Shi und Z. Miao, „Confidence analysis of standard deviational ellipse and its extension into higher dimensional Euclidean space," *PLOS ONE*, Jg. 10, Nr. 3, e0118537, 2015.

[3.21] A. Emery und A. Nenarokomov, „Optimal experiment design," *Measurement Science and Technology*, Jg. 9, Nr. 6, S. 864–876, Juni 1998.

[3.22] F. Pukelsheim, *Optimal Design of Experiments* (Classics in Applied Mathematics). SIAM - Society for Industrial und Applied Mathematics, 2006.

[3.23] A. Atkinson, A. Donev und R. Tobias, *Optimum Experimental Designs with SAS* (Oxford Statistical Science Series). Oxford: Oxford University Press, 2007.

[3.24] G. Golub und C. Van Loan, *Matrix Computations* (Johns Hopkins Studies in the Mathematical Sciences), 4. Aufl. Baltimore: Johns Hopkins University Press, 2013.