Mathematische Kontrolltheorie

Lars Grüne
Mathematisches Institut
Fakultät für Mathematik und Physik
Universität Bayreuth
95440 Bayreuth
lars.gruene@uni-bayreuth.de
http://num.math.uni-bayreuth.de

Vorlesungsskript

Sommersemester 2023

Vorwort

Dieses Skript ist im Rahmen einer gleichnamigen Vorlesung entstanden, die ich im Sommersemester 2023 an der Universität Bayreuth gehalten habe. Kapitel 1–7 behandeln dabei Themen aus der linearen Kontrolltheorie, während Kapitel 8–14 eine Einführung in die Modellprädiktive Regelung für nichtlineare Systeme geben. Gegenüber der vorherigen Auflage aus dem Wintersemester 2020/2021 wurden nur kleinere Korrekturen und Ergänzungen gemacht.

Teile des ersten Teils des Skriptes wurden auf Basis des Skripts [2], der Lehrbücher [19] und [14] sowie der Monographie [11] erstellt, die auch ohne explizite Erwähnung intensiv genutzt wurden. Die Kapitel über die Modellprädiktive Regelung sind überarbeitete Auszüge aus der Monographie [8]. Herzlich bedanken möchte ich mich bei Thomas Lorenz, Lisa Krügel und Jan Zetzmann sowie wie immer bei allen aufmerksamen Studentinnen und Studenten, die mich auf Fehler und Ungenauigkeiten hingewiesen haben.

Die jeweils aktuelle Version dieses Skripts erhalten Sie im Internet über meine Homepage (Google: Lars Grüne).

Bayreuth, Oktober 2023 Lars Grüne

Inhaltsverzeichnis

Kapitel 1

Grundbegriffe

Kontrollsysteme sind dynamische Systeme in kontinuierlicher oder diskreter Zeit, die von einem Parameter $u \in \mathbb{R}^m$ abhängen, der sich — abhängig von der Zeit und/oder vom Zustand des Systems — verändern kann. Dieser Parameter kann verschieden interpretiert werden. Er kann entweder als Steuergröße verstanden werden, also als Größe, die von außen aktiv beeinflusst werden kann (z.B. die Beschleunigung bei einem Fahrzeug, die Investitionen in einem Unternehmen) oder auch als Störung, die auf das System wirkt (z.B. Straßenunebenheiten bei einem Auto, Kursschwankungen bei Wechselkursen). Für das mathematische Fachgebiet, das sich mit der Analyse dieser Systeme beschäftigt, hat sich im deutschen Sprachgebrauch der Begriff „Kontrolltheorie“ etabliert, wenngleich er eine etwas missverständliche Übersetzung des englischen Ausdrucks „control theory“ darstellt, da es hier nicht um Kontrolle im Sinne von Überwachung sondern im Sinne von Einflussnahme von außen geht. Statt von Kontrolle spricht man auch von *Steuerung*, wenn die Parameter u lediglich von der Zeit abhängen und von *Regelung*, wenn die Parameter u vom aktuellen Zustand abhängen. Neben *Mathematischer Kontrolltheorie* ist auch der Ausdruck *Mathematische Systemtheorie* gebräuchlich.

1.1 Lineare Kontrollsysteme

Wir werden uns in dieser Vorlesung mit Kontrollsystemen beschäftigen, die in kontinuierlicher oder in diskreter Zeit definiert sind. In kontinuierlicher Zeit sind Kontrollsysteme durch gewöhnliche oder partielle Differentialgleichungen beschrieben. Wir beschränken uns in dieser Vorlesung in den meisten Fällen auf gewöhnliche Differentialgeichungen. Dann ist das Kontrollsystem durch die Gleichung

$$\dot{x}(t) = f(t, x(t), u(t)) \tag{1.1}$$

beschrieben. Die Variable $t \in \mathbb{R}$ werden wir hierbei stets als *Zeit* interpretieren und die Notation $\dot{x}(t)$ steht kurz für die zeitliche Ableitung $d/dt\, x(t)$. Die Größe $x(t) \in \mathbb{R}^n$ heißt der *Zustand* und $u(t) \in \mathbb{R}^m$ heißt der *Kontrollwert* oder die *Eingangsgröße*, jeweils zur Zeit t. Die Abbildung $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ heißt *Vektorfeld*. Sowohl f als auch die Funktion $u : \mathbb{R} \to \mathbb{R}^m$ müssen gewisse Regularitätseigenschaften erfüllen, damit die Lösungen von (1.1) existieren und eindeutig sind. Wir wollen uns mit diesem allgemeinen Problem

aber zunächst nicht weiter beschäftigen, da wir uns im ersten Teil der Vorlesung nur mit Spezialfall von Kontrollsystemen befassen werden.

In diskreter Zeit ist das allgemeine Modell gegeben durch die Abbildung

$$x(k+1) = f(k, x(k), u(k)). \tag{1.2}$$

Hierbei ist $k \in \mathbb{N}$ ein abstrakter Zeitindex und $f : \mathbb{N} \times \mathbb{R}^n \times U \to \mathbb{R}^n$ die *Übergangsabbildung*. Der abstrakte Zeitindex k steht dabei üblicherweise für eine reale Zeit $t_k \in \mathbb{R}$, oft von der Form $t_n = nT$ für ein festes $T > 0$. Ein zeitdiskretes Kontrollsystem kann das Verhalten eines kontinuierlichen Modells zu den diskreten Zeitpunkten t_k wiedergeben — dieses Vorgehen nennt man Abtastung oder Sampling und das entstehende zeitdiskrete System heißt *Abtastsystem*[1]. In diesem Fall gibt es unterschiedliche Möglichkeiten zur Wahl von U. Z.B. könnte $u(k)$ ein konstanter Kontrollwert aus dem $\mathbb{R}^m$ sein, der im Intervall $[t_k, t_{k+1})$ verwendet wird. In diesem Fall wäre $U = \mathbb{R}^m$. Die Größe $u(k)$ könnte aber auch eine zeitveränderliche Kontrollfunktion sein, die im kontinuierlichen System auf dem Intervall $[t_k, t_{k+1})$ verwendet wird. In diesem Fall wäre U eine Menge von Funktionen.

Fast alle Ergebnisse in dieser Vorlesung gelten sowohl für zeitkontinuierliche als auch für zeitdiskrete Kontrollsysteme, allerdings werden wir meistens nur einen der beiden Fälle beweisen. Im ersten Teil der Vorlesung werden wir die Beweise i.d.R. für zeitkontinuierliche Systeme angeben und im zweiten Teil i.d.R. für zeitdiskrete Systeme.

Im ersten Teil der Vorlesung werden wir uns mit den folgenden speziellen Kontrollsystemen befassen.

Definition 1.1 Ein *lineares zeitinvariantes* Kontrollsystem ist in kontinuierlicher Zeit gegeben durch die Differentialgleichung

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{1.3}$$

mit $A \in \mathbb{R}^{n\times n}$ und $B \in \mathbb{R}^{n\times m}$. In diskreter Zeit ist es gegeben durch die Gleichung

$$x(k+1) = Ax(k) + Bu(k) \tag{1.4}$$

mit $A \in \mathbb{R}^{n\times n}$ und einer linearen Abbildung $B : U \to \mathbb{R}^n$. □

Diese Klasse von Kontrollsystemen ist besonders einfach, da die rechte Seite linear in x und u ist und zudem nicht explizit von der Zeit t abhängt. Trotzdem ist sie bereits so reichhaltig, dass man mit ihr eine große Anzahl realer Prozesse z.B. für technische Anwendungen brauchbar beschreiben kann. Tatsächlich werden in der technischen Praxis auch heute noch viele lineare Modelle eingesetzt, wenn auch nicht immer in der einfachen Form (1.3) (wir werden später in der Vorlesung noch eine wichtige Erweiterung kennen lernen).

Um zu veranschaulichen, warum die Klasse (1.3) oft eine brauchbare Modellierung ermöglicht, betrachten wir ein Modell aus der Mechanik, und zwar ein auf einem Wagen befestigtes umgedrehtes starres Pendel, vgl. Abb. 1.1.

[1] Eine formale Definition des Abtastsystems findet sich in Abschnitt 8.2.

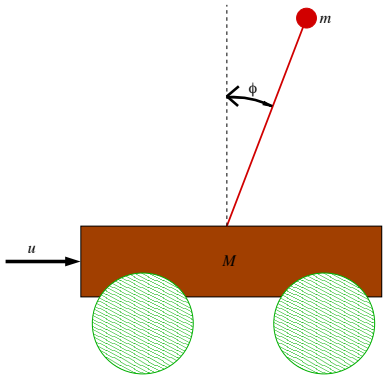


Abbildung 1.1: Schematische Darstellung des Pendels auf einem Wagen

Die Kontrolle u ist hierbei die Beschleunigung des Wagens. Mittels physikalischer Gesetze kann ein "exaktes"[2] Differentialgleichungsmodell hergeleitet werden.

$$\left.\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= -kx_2(t) + g\sin x_1(t) + u(t)\cos x_1(t) \\ \dot{x}_3(t) &= x_4(t) \\ \dot{x}_4(t) &= u \end{aligned}\right\} =: f(x(t), u(t)) \tag{1.5}$$

Hierbei besteht der Zustandsvektor $x \in \mathbb{R}^4$ aus 4 Komponenten: x_1 entspricht dem Winkel ϕ des Pendels (vgl. Abb. 1.1), der entgegen dem Uhrzeigersinn zunimmt, wobei $x_1 = 0$ dem aufgerichteten Pendel entspricht. x_2 ist die Winkelgeschwindigkeit, x_3 die Position des Wagens und x_4 dessen Geschwindigkeit. Die Konstante k beschreibt die Reibung des Pendels (je größer k desto mehr Reibung) und die Konstante $g \approx 9.81 m/s^2$ ist die Erdbeschleunigung.

Sicherlich ist (1.5) von der Form (1.1). Es ist aber nicht von der Form (1.3), da sich die nichtlinearen Funktionen sin und cos nicht mittels der Matrizen A und B darstellen lassen (beachte, dass in A und B nur konstante Koeffizienten stehen dürfen, die Matrizen dürfen also nicht von x abhängen).

Trotzdem kann ein lineares Modell der Form (1.3) verwendet werden, um (1.5) in der Nähe gewisser Punkte zu approximieren. Diese Prozedur, die man *Linearisierung* nennt, ist möglich in der Nähe von Punkten $(x^*, u^*) \in \mathbb{R}^n \times \mathbb{R}^m$, in denen $f(x^*, u^*) = 0$ gilt. In solchen Punkten erhalten wir ein System der Form (1.3), indem wir A und B definieren als

$$A := \frac{\partial f}{\partial x}(x^*, u^*) \quad \text{und} \quad B := \frac{\partial f}{\partial u}(x^*, u^*).$$

Wenn f stetig differenzierbar ist gilt

$$f(x + x^*, u + u^*) = Ax + Bu + o(\|x\| + \|u\|),$$

[2] Das Modell (1.5) ist nicht ganz exakt, da es bereits etwas vereinfacht ist: es wurde angenommen, dass das Pendel so leicht ist, dass es keinen Einfluss auf die Bewegung des Wagens hat. Zudem wurde eine Reihe von Konstanten so gewählt, dass sie sich gegeneinander aufheben.

d.h., für $x \approx 0$ und $u \approx 0$ stimmen $f(x + x^*, u + u^*)$ und $Ax + Bu$ gut überein. Man kann nun beweisen, dass sich diese Näherung auf die Lösungen der Differentialgleichungen (1.1) und (1.3) überträgt.[3]

Für unser Beispiel wenden wir die Linearisierung in $(x^*, u^*) = (0, 0)$ an. Dieses Gleichgewicht entspricht dem aufgerichteten oder invertierten Pendel. Aus der obigen Rechnung ergibt sich ein System der Form (1.3) mit

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ g & -k & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{und} \quad B = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \tag{1.6}$$

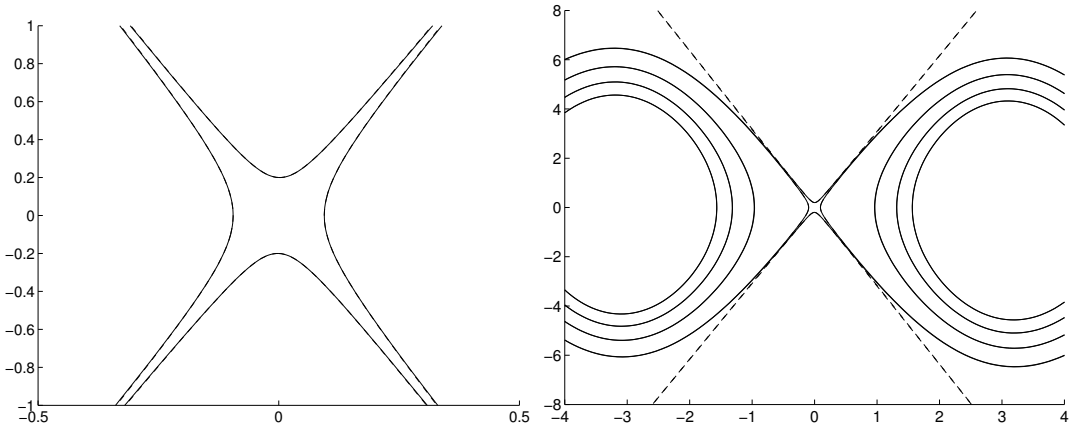


Abbildung 1.2: Vergleich der Lösungen von (1.5) (durchgezogen) mit (1.3, 1.6) (gestrichelt)

Abbildung 1.1 zeigt einen Vergleich der Lösungen von (1.5) (durchgezogen) mit den Lösungen von (1.3, 1.6) (gestrichelt), jeweils für $u \equiv 0$ und mit $k = 0.1$, $g = 9.81$, in zwei verschiedenen Umgebungen um die 0. Dargestellt sind hier für jede der zwei Gleichungen jeweils 4 Lösungskurven der Form

$$\left\{ \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} \,\middle|\, t \in [-10, 10] \right\} \subset \mathbb{R}^2.$$

Während in der kleinen Umgebung im linken Bildausschnitt mit bloßem Auge kein Unterschied zu erkennen ist, weichen die Lösungen in der größeren Umgebung im rechten Ausschnitt deutlich voneinander ab.

1.2 Existenz und Eindeutigkeit

Wann immer man sich mit Differentialgleichungen beschäftigt, muss man zunächst die Existenz und die Eindeutigkeit der Lösungen klären. Wir wollen dies zunächst für das lineare Kontrollsystem (1.3) mit $u \equiv 0$ machen.

[3] Eine mathematisch exakte Formulierung dieser Eigenschaft für unkontrollierte Differentialgleichungen findet sich z.B. als Satz 4.5 in [7].

Hierzu benötigen wir zunächst etwas Notation.

Für eine Matrix $A \in \mathbb{R}^{n\times n}$ bezeichnen wir im Folgenden mit $[A]_{ij} \in \mathbb{R}$ den Eintrag in der i-ten Zeile und j-ten Spalte. Für $A \in \mathbb{R}^{n\times n}$ und $t \in \mathbb{R}$ bezeichnen wir mit At die komponentenweise Multiplikation, also $[At]_{i,j} = [A]_{ij}t$. Für $k \in \mathbb{N}_0$ ist die Matrix-Potenz A^k induktiv mittels $A^0 = \mathrm{Id}$ und $A^{k+1} = AA^k$ definiert.

Zudem benötigen wir die folgende Definition.

Definition 1.2 Für eine Matrix $A \in \mathbb{R}^{n\times n}$ und eine reelle Zahl $t \in \mathbb{R}$ ist die Matrix-Exponentialfunktion gegeben durch

$$e^{At} := \sum_{k=0}^{\infty} A^k \frac{t^k}{k!}.$$

□

Die Konvergenz der unendlichen Reihe in dieser Definition ist dabei als komponentenweise Konvergenz, also als

$$[e^{At}]_{ij} = \sum_{k=0}^{\infty} [A^k \frac{t^k}{k!}]_{ij}, \quad n \in \mathbb{N}_0$$

zu verstehen. Dass die Komponenten dieser Reihe tatsächlich konvergieren, und zwar sogar absolut (also im Betrag), folgt aus dem Majorantenkriterium, denn mit der Zeilensummennorm

$$\alpha = \|A\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^{n} |[A]_{ij}|$$

gilt $|[A^k]_{ij}| \le \|A^k\|_\infty \le \|A\|_\infty^k = \alpha^k$, also

$$\left|[A^k \frac{t^k}{k!}]_{ij}\right| = |[A^k]_{ij}| \left|\frac{t^k}{k!}\right| \le \alpha^k \left|\frac{t^k}{k!}\right| = \frac{(\alpha|t|)^k}{k!}$$

und damit

$$|[e^{At}]_{ij}| \le e^{\alpha|t|},$$

wobei hier auf die rechten Seite die (übliche) skalare Exponentialfunktion steht.

Beachte, dass im Allgemeinen

$$[e^{At}]_{ij} \neq e^{[At]_{ij}}$$

gilt, wobei $e^{[At]_{ij}}$ die (komponentenweise angewandte) skalare Exponentialfunktion ist.

Aus der Definition der Matrix-Exponentialfunktion folgen sofort die Eigenschaften

$$\text{(i)}\ \ e^{A0} = \mathrm{Id} \quad \text{und} \quad \text{(ii)}\ \ Ae^{At} = e^{At}A \tag{1.7}$$

Das folgende Lemma liefert eine weitere wichtige Eigenschaft der Matrix-Exponentialfunktion.

Lemma 1.3 Für beliebiges $A \in \mathbb{R}^{n\times n}$ ist die Funktion $t \mapsto e^{At}$ differenzierbar und es gilt

$$\frac{d}{dt}e^{At} = Ae^{At}$$

für jedes $t \in \mathbb{R}$.

Beweis: Übungsaufgabe.

Satz 1.4 Betrachte die lineare Differentialgleichung

$$\dot{x}(t) = Ax(t) \tag{1.8}$$

mit $x : \mathbb{R} \to \mathbb{R}^n$ und einer gegebenen Matrix $A \in \mathbb{R}^{n\times n}$.

Dann gilt: Für jede *Anfangsbedingung* der Form

$$x(t_0) = x_0 \tag{1.9}$$

mit $t_0 \in \mathbb{R}$ und $x_0 \in \mathbb{R}^n$ existiert genau eine Lösung $x : \mathbb{R} \to \mathbb{R}^n$ von (1.8), die (1.9) erfüllt und die wir mit $x(t; t_0, x_0)$ bezeichnen. Für diese Lösung gilt

$$x(t; t_0, x_0) = e^{A(t-t_0)}x_0. \tag{1.10}$$

Beweis: Wir zeigen zunächst, dass die in (1.10) angegebene Funktion $x(t) = e^{A(t-t_0)}x_0$ sowohl die Differentialgleichung (1.8) als auch die Anfangsbedingung (1.9) erfüllt. Aus Lemma 1.3 folgt

$$\frac{d}{dt}x(t) = \frac{d}{dt}e^{A(t-t_0)}x_0 = Ae^{A(t-t_0)}x_0 = Ax(t),$$

also (1.8). Wegen (1.7)(i) gilt zudem

$$x(t_0) = e^{A(t_0-t_0)}x_0 = e^{A0}x_0 = \mathrm{Id}x_0 = x_0,$$

also (1.9).

Da wir damit (1.10) als Lösung verifiziert haben, folgt insbesondere die Existenz.

Es bleibt die Eindeutigkeit zu zeigen. Hierzu zeigen wir zunächst, dass die Matrix e^{At} invertierbar ist mit

$$(e^{At})^{-1} = e^{-At}. \tag{1.11}$$

Für jedes $y_0 \in \mathbb{R}^n$ löst $y(t) = e^{-At}y_0$ die Differentialgleichung $\dot{y}(t) = -Ay(t)$. Nach Produktregel gilt dann

$$\frac{d}{dt}(e^{-At}e^{At}x_0) = \frac{d}{dt}e^{-At}(e^{At}x_0) + e^{-At}\frac{d}{dt}e^{At}x_0 = -Ae^{-At}e^{At}x_0 + e^{-At}Ae^{At}x_0 = 0,$$

wobei wir im letzten Schritt (1.7)(ii) ausgenutzt haben. Also ist $e^{-At}e^{At}x_0$ konstant in t. Damit gilt für alle $t \in \mathbb{R}$ und alle $x_0 \in \mathbb{R}^n$

$$e^{-At}e^{At}x_0 = e^{-A0}e^{A0}x_0 = \mathrm{Id}\,\mathrm{Id}\,x_0 = x_0,$$

und folglich

$$e^{-At}e^{At} = \mathrm{Id} \;\Rightarrow\; e^{-At} = (e^{At})^{-1}.$$

Mit (1.11) können wir nun die Eindeutigkeit zeigen. Es sei $x(t)$ eine beliebige Lösung von (1.8), (1.9). Dann gilt

$$\begin{aligned}\frac{d}{dt}(e^{-A(t-t_0)}x(t)) &= \frac{d}{dt}e^{-A(t-t_0)}(x(t)) + e^{-A(t-t_0)}\dot{x}(t)\\ &= -Ae^{-A(t-t_0)}x(t) + e^{-A(t-t_0)}Ax(t) \;=\; 0,\end{aligned}$$

wobei wir wiederum (1.7)(ii) ausgenutzt haben. Also ist $e^{-A(t-t_0)}x(t)$ konstant in t, woraus für alle $t \in \mathbb{R}$

$$e^{-A(t-t_0)}x(t) = e^{-A(t_0-t_0)}x(t_0) = \mathrm{Id}x(t_0) = x_0$$

folgt. Multiplizieren wir nun beide Seiten dieser Gleichung mit $e^{A(t-t_0)}$ und verwenden (1.11), so ergibt sich

$$x(t) = e^{A(t-t_0)}x_0.$$

Da $x(t)$ eine beliebige Lösung war, folgt daraus die Eindeutigkeit. □

Eine nützliche Folgerung aus diesem Satz ist das folgende Korollar.

Korollar 1.5 Die Matrix-Exponentialfunktion e^{At} ist die eindeutige Lösung der Matrix-Differentialgleichung

$$\dot{X}(t) = AX(t) \tag{1.12}$$

mit $X : \mathbb{R} \to \mathbb{R}^{n\times n}$ und Anfangsbedingung

$$X(0) = \mathrm{Id}. \tag{1.13}$$

Beweis: Es bezeichne e_j den j-ten Einheitsvektor im $\mathbb{R}^n$. Eine einfache Rechnung zeigt, dass eine matrixwertige Funktion $X(t)$ genau dann eine Lösung von (1.12), (1.13) ist, wenn $X(t)e_j$ eine Lösung von (1.8), (1.9) mit $t_0 = 0$ und $x_0 = e_j$ ist. Mit dieser Beobachtung folgt die Behauptung sofort aus Satz 1.4. □

Das folgende Lemma fasst weitere Eigenschaften der Matrix-Exponentialfunktion zusammen.

Lemma 1.6 Für $A, A_1, A_2 \in \mathbb{R}^{n\times n}$ und $s, t \in \mathbb{R}$ gilt:

(i) $(e^{At})^{-1} = e^{-At}$

(ii) $e^{At}e^{As} = e^{A(t+s)}$

(iii) $e^{A_1t}e^{A_2t} = e^{(A_1+A_2)t}$ falls $A_1A_2 = A_2A_1$

(iv) Für eine invertierbare Matrix $T \in \mathbb{R}^{n\times n}$ gilt

$$e^{T^{-1}ATt} = T^{-1}e^{At}T.$$

Beweis: (i) Wurde im Beweis von Satz 1.4 gezeigt.

(ii) Mit Hilfe von (i) ergibt sich, dass sowohl $e^{At}e^{As}e^{-As}$ als auch $e^{A(t+s)}e^{-As}$ das Matrix-Anfangswertproblem (1.12), (1.13) erfüllen. Da dessen Lösung nach Korollar 1.5 eindeutig ist und e^{-As} invertierbar ist, folgt die behauptete Gleichheit.

(iii) Unter der angegebenen Bedingung $A_1A_2 = A_2A_1$ rechnet man nach, dass beide Ausdrücke das Matrix-Anfangswertproblem (1.12), (1.13) mit $A = A_1 + A_2$ erfüllen. Also müssen die Ausdrücke wegen der Eindeutigkeit nach Korollar 1.5 übereinstimmen.

(iv) Man rechnet nach, dass beide Ausdrücke das Matrix-Anfangswertproblem (1.12), (1.13) mit $T^{-1}AT$ an Stelle von A erfüllen. Wiederum folgt daraus die Gleichheit wegen der Eindeutigkeit der Lösungen nach Korollar 1.5. □

Nach diesen Vorbereitungen kehren wir nun zum linearen Kontrollsystem (1.3) zurück. Zur Formulierung eines Existenz- und Eindeutigkeitssatzes müssen wir einen geeigneten Funktionenraum $\mathcal{U}$ für die Kontrollfunktion $u(\cdot)$ definieren. Sicherlich wären stetige Funktionen geeignet, diese Wahl ist aber zu einschränkend, da wir im Verlauf dieser Vorlesung öfter einmal Kokatenationen von Kontrollfunktionen gemäß der folgenden Definition benötigen werden.

Definition 1.7 Für zwei Funktionen $u_1, u_2 : \mathbb{R} \to \mathbb{R}^m$ und $s \in \mathbb{R}$ definieren wir die *Konkatenation zur Zeit s* als

$$u_1 \&_s u_2(t) := \begin{cases} u_1(t), & t < s \\ u_2(t), & t \geq s \end{cases}$$

□

Selbst wenn u_1 und u_2 stetig sind, wird $u_1\&_s u_2$ im Allgemeinen nicht stetig sein. Wir benötigen also einen Funktionenraum, der abgeschlossen bezüglich der Konkatenation ist. Hier gibt es verschiedene Möglichkeiten, die einfachste ist die folgende.

Definition 1.8 Eine Funktion $u : \mathbb{R} \to \mathbb{R}^m$ heißt *stückweise stetig*, falls für jedes kompakte Intervall $[t_1, t_2]$ eine endliche Folge von Zeiten $t_1 = \tau_1 < \tau_2 < \ldots < \tau_k = t_2$ existiert, so dass $u|_{(\tau_i,\tau_{i+1})}$ beschränkt und stetig ist für alle $i = 1, \ldots, k-1$. Wir definieren $\mathcal{U}$ als den Raum der stückweise stetigen Funktionen von $\mathbb{R}$ nach $\mathbb{R}^m$. □

Sicherlich ist $\mathcal{U}$ abgeschlossen unter Konkatenation, aber auch unter Addition und Multiplikation (wobei wir $(u_1 + u_2)(t) := u_1(t) + u_2(t)$ und $(u_1 \cdot u_2)(t) := u_1(t) \cdot u_2(t)$ definieren). Zudem — und dies ist für unsere Zwecke wichtig — existiert das Riemann-Integral

$$\int_{t_1}^{t_2} u(t)dt$$

über Funktionen $u \in \mathcal{U}$, da es in jedem kompakten Integrationsintervall nur endlich viele Unstetigkeitsstellen gibt.[4]

Mit diesem Funktionenraum können wir nun das entsprechende Resultat formulieren.

[4] Eine Alternative zu den stückweise stetigen Funktionen bietet der Raum der Lebesgue-messbaren Funktionen, wobei das Integral dann als das Lebesgue-Integral gewählt wird. Diesen Raum werden wir bei den nichtlinearen Systemen verwenden, vgl. Kapitel 8. Für lineare Kontrollsysteme bringt die Verwendung Lebesgue-messbarer Kontrollfunktionen keinen Vorteil.

Satz 1.9 Betrachte das lineare Kontrollsystem (1.3)

$$\dot{x}(t) = Ax(t) + Bu(t)$$

mit $x : \mathbb{R} \to \mathbb{R}^n$ und gegebenen Matrizen $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$.

Dann gilt: Für jede *Anfangsbedingung* der Form (1.9)

$$x(t_0) = x_0$$

mit $t_0 \in \mathbb{R}$, $x_0 \in \mathbb{R}^n$ und jede stückweise stetige Kontrollfunktion $u \in \mathcal{U}$ existiert genau eine stetige Funktion $x : \mathbb{R} \to \mathbb{R}^n$, die (1.9) erfüllt und deren Ableitung für jedes t, in dem u stetig ist, existiert und (1.3) erfüllt. Diese eindeutige Funktion nennen wir die Lösung von (1.3), (1.9) und bezeichnen sie mit $x(t; t_0, x_0, u)$. Für diese Lösung gilt

$$x(t; t_0, x_0, u) = e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-s)}Bu(s)ds. \tag{1.14}$$

Beweis: Wir rechnen zunächst nach, dass (1.14) tatsächlich eine Lösung im angegeben Sinne ist. Die Abbildung $t \mapsto \int_{t_0}^{t} g(s)ds$ ist stetig für jede Riemann-integrierbare Funktion, also ist $x(t; t_0, x_0, u)$ stetig in t. In den Stetigkeitsetellen von u gilt

$$\begin{aligned}
&\frac{d}{dt}\left[e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-s)}Bu(s)ds\right] \\
&= \frac{d}{dt}e^{A(t-t_0)}x_0 + \frac{d}{dt}\int_{t_0}^{t} e^{A(t-s)}Bu(s)ds \\
&= Ae^{A(t-t_0)}x_0 + \underbrace{e^{A(t-s)}Bu(s)|_{s=t}}_{=Bu(t)} + \int_{t_0}^{t} Ae^{A(t-s)}Bu(s)ds \\
&= A\left(e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-s)}Bu(s)ds\right) + Bu(t),
\end{aligned}$$

also (1.3). Zudem gilt

$$\underbrace{e^{A(t_0-t_0)}}_{=\mathrm{Id}}x_0 + \underbrace{\int_{t_0}^{t_0} e^{A(t_0-s)}Bu(s)ds}_{=0} = x_0,$$

also (1.9).

Es bleibt die Eindeutigkeit zu zeigen. Dazu betrachten wir zwei beliebige Lösungen $x(t)$, $y(t)$ von (1.3), (1.9) im Sinne des Satzes. Dann gilt zunächst

$$\dot{z}(t) = \dot{x}(t) - \dot{y}(t) = Ax(t) + Bu(t) - Ay(t) - Bu(t) = A(x(t) - y(t)) = Az(t)$$

für alle Punkte in denen u stetig ist. Da z selbst stetig ist, kann $\dot{z}$ in den Unstetigkeitsstellen τ_i von u durch $\dot{z}(\tau_i) = \lim_{t\to\tau_i} Az(t)$ wohldefiniert stetig fortgesetzt werden. Wir erhalten damit eine Funktion, die die Differentialgleichung $\dot{z}(t) = Az(t)$ für alle $t \in \mathbb{R}$ löst. Da zudem

$$z(t_0) = x(t_0) - y(t_0) = x_0 - x_0 = 0$$

gilt, erfüllt z ein Anfangswertproblem der Form (1.8), (1.9), dessen nach Satz 1.4 eindeutige Lösung durch $z(t) = e^{At}0 = 0$ gegeben ist. Also ist $x(t) = y(t)$ für alle $t \in \mathbb{R}$, womit die Eindeutigkeit folgt. □

Eine Folgerung aus diesem Satz ist das folgende Korollar.

Korollar 1.10 Für die Lösungen von (1.3), (1.9) gelten für alle $t, s \in \mathbb{R}$ die Gleichungen

$$x(t; t_0, x_0, u) = x(t; s, x(s; t_0, x_0, u), u)$$

und

$$x(t; t_0, x_0, u) = x(t - s; t_0 - s, x_0, u(s + \cdot)),$$

wobei die Funktion $u(s + \cdot) \in \mathcal{U}$ mittels $u(s + \cdot)(t) = u(s + t)$ definiert ist. Aus der Kombination der beiden Formeln folgt für $t_0 = 0$ dann auch

$$x(t; x_0, u) = x(t - s; x(s; x_0, u), u(s + \cdot)).$$

Beweis: Folgt sofort aus der Darstellung (1.14). □

Bemerkung 1.11 Eine weitere unmittelbare Folgerung aus der Lösungsformel (1.14) ist die Identität

$$x(t; t_0, x_0, u) = x(t; t_0, x_0, 0) + x(t; t_0, 0, u). \tag{1.15}$$

Diese Identität besagt, dass jede Lösung als Überlagerung (oder Superposition) einer unkontrollierten Lösung (also mit Kontrolle 0) und einer Lösung ohne Eigendynamik (also mit Anfangswert 0) zusammengesetzt ist. Sie ist daher als *Superpositionsprinzip* bekannt. □

Bemerkung 1.12 Da wir uns in den folgenden Kapiteln in vielen Fällen auf die Betrachtung von Lösungen mit der speziellen Anfangszeit $t_0 = 0$ beschränken, schreiben wir für $t_0 = 0$ oft kurz $x(t; x_0, u) = x(t; 0, x_0, u)$. □

Bemerkung 1.13 Wenn man die Zeiten $t_n = nT$ betrachtet und ein kontinuierliches Kontrollsystem mit Kontrollfunktionen, die auf den Intervallen $[t_k, t_{k+1})$ konstant gleich $u_T(k)$ sind, so kann man aus der Lösungsformel (1.14) explizite Formeln für die Matrizen A_T und B_T in dem zugehörigen Abtastsystem

$$x_T(k + 1) = A_T x_T(k) + B_T u_T(k)$$

herleiten. Details werden in einer Übungsaufgabe ausgearbeitet. □

Kapitel 2

Kontrollierbarkeit

2.1 Definitionen

Ein wichtiger Aspekt in der Analyse lineare Kontrollsysteme der Form (1.3) ist die Frage der Kontrollierbarkeit. In der allgemeinsten Formulierung ist dies die Frage, für welche Punkte x_0, $x_1 \in \mathbb{R}^n$ und Zeiten t_1 eine Kontrollfunktion $u \in \mathcal{U}$ gefunden werden kann, so dass $x(t_1; x_0, u) = x_1$ gilt, d.h., so dass die zwei Punkte durch eine Lösungstrajektorie verbunden werden. Formal definieren wir dies wie folgt.

Definition 2.1 Betrachte ein lineares Kontrollsystem (1.3).

Ein Zustand $x_0 \in \mathbb{R}^n$ heißt *kontrollierbar* (oder auch *steuerbar*) zu einem Zustand $x_1 \in \mathbb{R}^n$ zur Zeit $t_1 > 0$, falls ein $u \in \mathcal{U}$ existiert mit

$$x_1 = x(t_1; x_0, u).$$

Der Punkt x_1 heißt dann *erreichbar* von x_0 zur Zeit t_1. □

Das folgende Lemma zeigt, dass man den Fall beliebiger $x_0 \in \mathbb{R}^n$ auf $x_0 = 0$ zurückführen kann.

Lemma 2.2 Ein Zustand $x_0 \in \mathbb{R}^n$ ist genau dann kontrollierbar zu einem Zustand $x_1 \in \mathbb{R}^n$ zur Zeit $t_1 > 0$, falls der Zustand $\tilde{x}_0 = 0$ kontrollierbar zu dem Zustand $\tilde{x}_1 = x_1 - x(t_1; x_0, 0)$ zur Zeit t_1 ist.

Beweis: Übungsaufgabe.

Diese Tatsache motiviert, im Weiteren die Kontrollierbarkeit bzw. Erreichbarkeit der 0 speziell zu betrachten.

Definition 2.3 Betrachte ein lineares Kontrollsystem (1.3).

(i) Die *Erreichbarkeitsmenge* (*reachable set*) von $x_0 = 0$ zur Zeit $t \geq 0$ ist gegeben durch

$$\mathcal{R}(t) = \{x(t; 0, u) \,|\, u \in \mathcal{U}\}.$$

(ii) Die *Kontrollierbarkeitsmenge* (*controllable set*) nach $x_1 = 0$ zur Zeit $t \geq 0$ ist gegeben durch

$$\mathcal{C}(t) = \{x_0 \in \mathbb{R}^n \,|\, \text{es existiert } u \in \mathcal{U} \text{ mit } x(t; x_0, u) = 0\}.$$

□

Die Beziehung zwischen diesen beiden Mengen klärt das folgende Lemma.

Lemma 2.4 Die Erreichbarkeitsmenge $\mathcal{R}(t)$ für (1.3) ist gerade gleich der Kontrollierbarkeitsmenge $\mathcal{C}(t)$ für das zeitumgekehrte System

$$\dot{z}(t) = -Az(t) - Bu(t). \tag{2.1}$$

Beweis: Durch Überprüfen des Anfangswertproblems sieht man, dass zwischen den Lösungen von (1.3) und (2.1) für alle $t, s \in \mathbb{R}$ die Beziehung

$$x(s, 0, u) = z(t - s, x(t, 0, u), u(t - \cdot)).$$

Wenn also $x_1 \in \mathcal{R}(t)$ für (1.3) ist und $x(s, 0, u)$ die zugehörige Lösung, so folgt

$$z(0, x(t, 0, u), u(t - \cdot)) = x(t, 0, u) = x_1 \text{ und } z(t, x(t, 0, u), u(t - \cdot)) = x(0, 0, u) = 0,$$

womit $x_1 \in \mathcal{C}(t)$ folgt. Umgekehrt argumentiert man genauso. □

2.2 Analyse von Kontrollierbarkeitseigenschaften

Wir wollen nun die Struktur dieser Mengen klären. Wir leiten die technischen Zwischenresultate dabei für $\mathcal{R}(t)$ her und formulieren nur die Hauptresultate auch für $\mathcal{C}(t)$.

Lemma 2.5 (i) $\mathcal{R}(t)$ ist für alle $t \geq 0$ ein Untervektorraum des $\mathbb{R}^n$.
(ii) $\mathcal{R}(t) = \mathcal{R}(s)$ für alle $s, t > 0$.

Beweis: (i) Zu zeigen ist, dass für x_1, $x_2 \in \mathcal{R}(t)$ und $\alpha \in \mathbb{R}$ auch $\alpha(x_1 + x_2) \in \mathcal{R}(t)$ ist. Für x_1, x_2 in $\mathcal{R}(t)$ existieren Kontrollfunktionen u_1, $u_2 \in \mathcal{U}$ mit

$$x_i = x(t; 0, u_i) = \int_0^t e^{A(t-s)} B u_i(s) ds.$$

Also gilt für $u = \alpha(u_1 + u_2)$ die Gleichung

$$\begin{aligned} x(t; 0, u) &= \int_0^t e^{A(t-s)} B u(s) ds = \int_0^t e^{A(t-s)} B \alpha(u_1(s) + u_2(s)) ds \\ &= \alpha \left(\int_0^t e^{A(t-s)} B u_1(s) ds + \int_0^t e^{A(t-s)} B u_2(s) ds \right) = \alpha(x_1 + x_2), \end{aligned}$$

woraus $\alpha(x_1 + x_2) \in \mathcal{R}(t)$ folgt. Dies beweist (i).

(ii) Wir geben hier einen direkten Beweis, die Aussage folgt aber unabhängig davon auch aus Satz 2.12.

Wir zeigen zuerst die Hilfsaussage

$$\mathcal{R}(t_1) \subseteq \mathcal{R}(t_2) \tag{2.2}$$

für $0 < t_1 < t_2$: Falls $y \in \mathcal{R}(t_1)$ existiert ein $u \in \mathcal{U}$ mit

$$x(t_1; 0, u) = y.$$

Mit der neuen Kontrolle $\tilde{u} = 0\&_{t_2-t_1} u(t_1 - t_2 + \cdot)$ und Korollar 1.10 ergibt sich so

$$x(t_2; 0, \tilde{u}) = x(t_2; t_2 - t_1, \underbrace{x(t_2 - t_1; 0, 0)}_{=0}, \tilde{u}) = x(t_2; t_2 - t_1, 0, \tilde{u}) = x(t_1; 0, u) = y,$$

weswegen $y \in \mathcal{R}(t_2)$ gilt.

Als nächstes zeigen wir, dass für beliebige $0 < t_1 < t_2$ aus der Gleichheit $\mathcal{R}(t_1) = \mathcal{R}(t_2)$ bereits die Gleichheit $\mathcal{R}(t_1) = \mathcal{R}(t)$ für alle $t \geq t_1$ folgt. Um dies zu zeigen sei $x \in \mathcal{R}(2t_2 - t_1)$, es existiere also ein $u \in \mathcal{U}$ mit $x = x(2t_2 - t_1, 0, u)$.

Da $x(t_2, 0, u) \in \mathcal{R}(t_2)$ und $\mathcal{R}(t_2) = \mathcal{R}(t_1)$, existiert ein $v \in \mathcal{U}$ mit $x(t_1, 0, v) = x(t_2, 0, u)$. Definieren wir nun eine Kontrollfunktion $w = v\&_{t_1} u(t_2 - t_1 + \cdot)$, so gilt mit Korollar 1.10

$$\begin{aligned} x(t_2, 0, w) &= x(t_2, t_1, \underbrace{x(t_1, 0, v)}_{=x(t_2,0,u)}, w) \\ &= x(t_2 + t_2 - t_1, t_1 + t_2 - t_1, x(t_2, 0, u), \underbrace{w(t_1 - t_2 + \cdot)}_{=u(\cdot)}) \\ &= x(2t_2 - t_1, 0, u) = x. \end{aligned}$$

Damit gilt also $x \in \mathcal{R}(t_2)$ und folglich $\mathcal{R}(t_1) = \mathcal{R}(t_2) = \mathcal{R}(2t_2 - t_1) = \mathcal{R}(2(t_2 - t_1) + t_1)$. Induktive Wiederholung dieser Konstruktion liefert $\mathcal{R}(t_1) = \mathcal{R}(2^k(t_2 - t_1) + t_1)$ für alle $k \in \mathbb{N}$ und damit wegen (2.2) die Behauptung $\mathcal{R}(t_1) = \mathcal{R}(t)$ für alle $t \geq t_1$.

Nun zeigen wir die Behauptung (ii): Sei dazu $s > 0$ beliebig und sei $0 < s_0 < \ldots < s_{n+1} = s$ eine aufsteigende Folge von Zeiten. Dann ist $\mathcal{R}(s_0), \ldots, \mathcal{R}(s_{n+1})$ nach (2.2) eine aufsteigende Folge von $n+2$ Unterräumen des $\mathbb{R}^n$. Aus $\mathcal{R}(s_{k+1}) \neq \mathcal{R}(s_k)$ folgt daher $\dim \mathcal{R}(s_{k+1}) \geq \dim \mathcal{R}(s_k) + 1$. Wären also die $\mathcal{R}(s_k)$ paarweise verschieden, so müsste $\dim \mathcal{R}(s_{n+1}) \geq n+1$ gelten, was ein Widerspruch zu $\mathcal{R}(s_{n+1}) \subseteq \mathbb{R}^n$ ist, weswegen mindestens zwei der $\mathcal{R}(s_k)$ übereinstimmen müssen. Nach der vorhergehenden Überlegung folgt daraus $\mathcal{R}(t) = \mathcal{R}(s)$ für alle $t \geq s$ und da $s > 0$ beliebig war, folgt die Behauptung. □

Bemerkung 2.6 Da die Menge $\mathcal{R}(t)$ also nicht von t abhängt, schreiben wir im Folgenden oft einfach $\mathcal{R}$. □

Bemerkung 2.7 Die Verbindung von Lemma 2.2 und Lemma 2.5 zeigt also, dass die Menge der von einem Punkt $x_0 \in \mathbb{R}^n$ in einer Zeit $t > 0$ erreichbaren Zustände der affine Unterraum

$$x(t; x_0, 0) + \mathcal{R}$$

ist, dessen Dimension gerade gleich der von $\mathcal{R}$ ist. Beachte, dass diese Menge i.A. nicht unabhängig von t ist. Eine Ausnahme ist der Fall $\mathcal{R} = \mathbb{R}^n$, da dann auch $x(t; x_0, 0) + \mathcal{R} = \mathbb{R}^n$ gilt. In diesem Fall ist jeder Zustand x_0 zu jedem anderen Zustand x_1 kontrollierbar, weswegen wir das System für $\mathcal{R} = \mathbb{R}^n$ *vollständig kontrollierbar* oder kurz *kontrollierbar* nennen. □

Wie in den Übungen zu sehen war, kann es bereits für recht einfache Kontrollsysteme ziemlich schwierig sein, die Mengen $\mathcal{R}$ und $\mathcal{C}$ direkt auszurechnen. Wir wollen daher jetzt eine einfache Charakterisierung dieser Mengen herleiten. Hierzu benötigen wir etwas lineare Algebra.

Definition 2.8 (i) Ein Unterraum $U \subseteq \mathbb{R}^n$ heißt A-invariant für eine Matrix $A \in \mathbb{R}^{n\times n}$, falls $Av \in U$ für alle $v \in U$ (oder kurz $AU \subseteq U$) gilt.

(ii) Für einen Unterraum $V \subseteq \mathbb{R}^n$ und $A \in \mathbb{R}^{n\times n}$ bezeichne

$$\langle A \,|\, V \rangle$$

den kleinsten (bezüglich der Dimension) A-invarianten Unterraum von $\mathbb{R}^n$, der V enthält. □

Beachte, dass ein kleinster solcher Raum existiert und eindeutig ist: Einerseits existiert mit dem $\mathbb{R}^n$ selbst ein A-invarianter Unterraum, der V enthält. Da die Dimension endlich ist, existiert also auch ein solcher Raum kleinster Dimension. Zudem ist der Schnitt zweier A-invarianter Unterräume, die V enthalten, wieder ein A-invarianter Unterraum, der V enthält. Existieren also mehrere A-invariante Unterräume kleinster Dimension, die alle V enthalten, so müssen diese alle übereinstimmen, da ihr Schnitt ansonsten einen solchen Raum kleinerer Dimension bilden würde.

Lemma 2.9 Für einen Unterraum $V \subseteq \mathbb{R}^n$ und $A \in \mathbb{R}^{n\times n}$ gilt

$$\langle A \,|\, V \rangle = V + AV + \ldots + A^{n-1}V.$$

Beweis: "$\supseteq$": Wegen der A-Invarianz von $\langle A \,|\, V \rangle$ und $V \subseteq \langle A \,|\, V \rangle$ gilt

$$A^k V \subseteq \langle A \,|\, V \rangle$$

für alle $k \in \mathbb{N}_0$ und damit $\langle A \,|\, V \rangle \supseteq V + AV + \ldots + A^{n-1}V$.

"$\subseteq$": Es genügt zu zeigen, dass $V + AV + \ldots + A^{n-1}V$ A-invariant ist, da dann wegen $V \subseteq V + AV + \ldots + A^{n-1}V$ sofort $\langle A \,|\, V \rangle \subseteq V + AV + \ldots + A^{n-1}V$ folgt.

Zum Beweis der A-Invarianz betrachte das charakteristische Polynom von A

$$\chi_A(z) = \det(z\mathrm{Id} - A) = z^n + a_{n-1}z^{n-1} + \ldots + a_1 z + a_0.$$

Für dieses gilt nach dem Satz von Cayley-Hamilton

$$\chi_A(A) = A^n + a_{n-1}A^{n-1} + \ldots + a_1 A + a_0 \mathrm{Id} = 0,$$

woraus

$$A^n = -a_{n-1}A^{n-1} - \ldots - a_1 A - a_0 \mathrm{Id}$$

folgt. Sei also $v \in V + AV + \ldots + A^{n-1}V$. Dann lässt sich v darstellen als $v = v_0 + Av_1 + \ldots + A^{n-1}v_{n-1}$ für $v_0, \ldots, v_{n-1} \in V$. Damit folgt

$$\begin{aligned} Av &= Av_0 + A^2 v_1 + \ldots + A^n v_{n-1} \\ &= Av_0 + A^2 v_1 - a_{n-1}A^{n-1}v_{n-1} - \ldots - a_1 A v_{n-1} - a_0 v_{n-1} \\ &= \tilde{v}_0 + A\tilde{v}_1 + \ldots + A^{n-1}\tilde{v}_{n-1} \end{aligned}$$

für geeignete $\tilde{v}_0, \ldots, \tilde{v}_{n-1} \in V$. Damit folgt $Av \in V + AV + \ldots + A^{n-1}V$, also die A-Invarianz. □

Wir werden nun den Spezialfall betrachten, dass $V = \operatorname{im} B$ das Bild der Matrix B ist. In diesem Fall sagt Lemma 2.9, dass

$$\langle A \,|\, \operatorname{im} B\rangle = \{Bx_0 + ABx_1 + \ldots + A^{n-1}Bx_{n-1} \,|\, x_0, \ldots, x_{n-1} \in \mathbb{R}^m\} = \operatorname{im}(B\, AB\, \ldots\, A^{n-1}B),$$

wobei $(B\, AB\, \ldots\, A^{n-1}B) \in \mathbb{R}^{n\times(m\cdot n)}$ ist.

Definition 2.10 Die Matrix $(B\, AB\, \ldots\, A^{n-1}B) \in \mathbb{R}^{n\times(m\cdot n)}$ heißt *Kontrollierbarkeitssmatrix* des Systems (1.3). □

Im Folgenden verwenden wir für $t \in \mathbb{R}$ die Notation

$$W_t := \int_0^t e^{A\tau} B B^T (e^{A\tau})^T d\tau.$$

Beachte, dass $W_t \in \mathbb{R}^{n\times n}$ gilt und W_t damit ein linearer Operator auf dem $\mathbb{R}^n$ ist. Die Matrix W_t wird *Kontrollierbarkeitsgramsche* genannt und ist symmetrisch und positiv semidefinit, denn es gilt

$$x^T W_t x = \int_0^t \underbrace{x^T e^{A\tau} B B^T (e^{A\tau})^T x}_{=\|B^T(e^{A\tau})^T x\|^2 \geq 0} d\tau \geq 0.$$

Für das Bild $\operatorname{im} W_t$ dieses Operators gilt das folgende Lemma.

Lemma 2.11 Für alle $t > 0$ gilt $\langle A \,|\, \operatorname{im} B\rangle = \operatorname{im} W_t$.

Beweis: Wir zeigen $\langle A \,|\, \operatorname{im} B\rangle^\perp = (\operatorname{im} W_t)^\perp$.
"$\subseteq$": Sei $x \in \langle A \,|\, \operatorname{im} B\rangle^\perp$, also $x^T A^k B = 0$ für alle $k \in \mathbb{N}_0$. Dann gilt

$$x^T e^{At} B = \sum_{k=0}^{\infty} \frac{t^k x^T A^k B}{k!} = 0$$

und damit $x^T W_t = 0$, also $x \in (\operatorname{im} W_t)^\perp$.

"$\supseteq$": Sei $x \in (\operatorname{im} W_t)^\perp$ für ein $t > 0$. Dann gilt

$$0 = x^T W_t x = \int_0^t \|B^T (e^{A\tau})^T x\|^2 d\tau,$$

woraus wegen der Stetigkeit des Integranden $x^T e^{A\tau} B = (B^T (e^{A\tau})^T x)^T = 0$ folgt.

Sukzessives Differenzieren von $x^T B e^{A\tau}$ nach τ liefert

$$x^T A^k e^{A\tau} B = 0$$

für alle $k \in \mathbb{N}_0$. Für $\tau = 0$ folgt $x^T A^k B = 0$, also $x \in (\operatorname{im} A^k B)^\perp$ für alle $k \in \mathbb{N}_0$ und damit auch $x \in [\operatorname{im} (B\ AB\ \dots\ A^{n-1}B)]^\perp = \langle A \,|\, \operatorname{im} B\rangle^\perp$. □

Der folgende Satz ist das Hauptresultat über die Struktur der Erreichbarkeits- und Kontrollierbarkeitsmengen.

Satz 2.12 Für das System (1.3) gilt für alle $t > 0$

$$\mathcal{R}(t) = \mathcal{C}(t) = \langle A \,|\, \operatorname{im} B\rangle = \operatorname{im} (B\ AB\ \dots\ A^{n-1}B).$$

Beweis: Die Gleichheit $\langle A \,|\, \operatorname{im} B\rangle = \operatorname{im} (B\ AB\ \dots\ A^{n-1}B)$ wurde bereits in der Rechnung vor Definition 2.10 gezeigt. Wir zeigen $\mathcal{R}(t) = \langle A \,|\, \operatorname{im} B\rangle$ (woraus insbesondere wiederum die Unabhängigkeit von $\mathcal{R}(t)$ von t folgt). Die Aussage für $\mathcal{C}(t)$ folgt dann mit Lemma 2.4, denn es gilt $\langle A \,|\, \operatorname{im} B\rangle = \langle -A \,|\, \operatorname{im} -B\rangle$.

"$\subseteq$": Sei $x = x(t; 0, u) \in \mathcal{R}(t)$. Nach der allgemeinen Lösungsformel ist

$$x = \int_0^t e^{A(t-\tau)} B u(\tau) d\tau.$$

Nun gilt für all $\tau \in [0, t]$ nach Definition von $\langle A \,|\, \operatorname{im} B\rangle$

$$e^{A(t-\tau)} B u(\tau) = \sum_{k=0}^{\infty} \frac{(t-\tau)^k}{k!} A^k B u(\tau) \in \langle A \,|\, \operatorname{im} B\rangle$$

und damit auch $x \in \langle A \,|\, \operatorname{im} B\rangle$, da die Integration über Elemente eines Unterraums wieder ein Element dieses Unterraums ergibt.

"$\supseteq$": Sei $x \in \langle A \,|\, \operatorname{im} B\rangle$ und $t > 0$ beliebig. Dann existiert nach Lemma 2.11 ein ein $z \in \mathbb{R}^n$ mit $x = W_t z$. Definieren wir nun $u \in \mathcal{U}$ durch $u(\tau) := B^T (e^{A(t-\tau)})^T z$ für $\tau \in [0, t]$, so gilt

$$x(t; 0, u) = \int_0^t e^{A(t-\tau)} B B^T (e^{A(t-\tau)})^T z d\tau = W_t z = x,$$

und damit $x \in \mathcal{R}(t)$. □

Beachte, dass der Beweis konstruktiv ist: er liefert eine Formel für die Kontrollfunktion u, mit der man von 0 nach x steuern kann.

Korollar 2.13 (Kalman-Kriterium)
Das System (1.3) ist genau dann vollständig kontrollierbar, wenn

$$\operatorname{rg}(B\,AB\,\ldots\,A^{n-1}B) = n$$

ist. In diesem Fall nennen wir das Matrizenpaar (A, B) *kontrollierbar*.

Wenn (A, B) nicht kontrollierbar ist, kann man den Zustandsraum $\mathbb{R}^n$ wie folgt aufteilen, um das Paar (A, B) in seinen kontrollierbaren und unkontrollierbaren Anteil zu zerlegen.

Lemma 2.14 Sei (A, B) nicht kontrollierbar, d.h., $r := \dim\langle A \,|\, \operatorname{im} B\rangle < n$. Dann existiert ein invertierbares $T \in \mathbb{R}^{n\times n}$, so dass $\widetilde{A} = T^{-1}AT$ und $\widetilde{B} = T^{-1}B$ die Form

$$\widetilde{A} = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}, \quad \widetilde{B} = \begin{pmatrix} B_1 \\ 0 \end{pmatrix}$$

mit $A_1 \in \mathbb{R}^{r\times r}$, $A_2 \in \mathbb{R}^{r\times(n-r)}$, $A_3 \in \mathbb{R}^{(n-r)\times(n-r)}$, $B_1 \in \mathbb{R}^{r\times m}$ besitzen, wobei das Paar (A_1, B_1) kontrollierbar ist. Insbesondere hat das System nach Koordinatentransformation mit T also die Form

$$\begin{aligned} \dot{z}_1(t) &= A_1 z_1(t) + A_2 z_2(t) + B_1 u(t) \\ \dot{z}_2(t) &= A_3 z_2(t) \end{aligned}$$

mit $z_1(t) \in \mathbb{R}^r$ und $z_2(t) \in \mathbb{R}^{n-r}$.

Beweis: Übungsaufgabe.

Beachte, dass sich das charakteristische Polynom einer Matrix bei Koordinatentransformationen nicht verändert. Es gilt also

$$\chi_A(z) = \det(z\mathrm{Id} - A) = \det(z\mathrm{Id} - \widetilde{A}) = \det(z\mathrm{Id} - A_1)\cdot\det(z\mathrm{Id} - A_3) = \chi_{A_1}(z)\cdot\chi_{A_3}(z).$$

Dies motiviert die folgende Definition.

Definition 2.15 Wir nennen χ_{A_1} den *kontrollierbaren* und χ_{A_3} den *unkontrollierbaren Anteil* des charakteristischen Polynoms χ_A. □

Der folgende Satz liefert alternative Charakterisierungen der Kontrollierbarkeit, die ohne die Berechnung der Kontrollierbarkeitsmatrix auskommen. Hierbei bezeichnet $(\lambda\mathrm{Id} - A \,|\, B) \in \mathbb{R}^{n\times(n+m)}$ die Matrix, die durch Nebeneinanderschreiben der Matrizen $\lambda\mathrm{Id} - A$ und B entsteht.

Satz 2.16 (Hautus-Kriterium)
Die folgenden Bedingungen sind äquivalent:

(i) (A, B) ist kontrollierbar

(ii) $\operatorname{rg}(\lambda\mathrm{Id} - A \,|\, B) = n$ für alle $\lambda \in \mathbb{C}$

(iii) $\operatorname{rg}(\lambda\mathrm{Id} - A \,|\, B) = n$ für alle Eigenwerte $\lambda \in \mathbb{C}$ von A

Beweis: Wir beweisen zuerst "(ii) $\Leftrightarrow$ (iii)" und dann "(i) $\Leftrightarrow$ (ii)".

"(ii) $\Rightarrow$ (iii)": klar

"(ii) $\Leftarrow$ (iii)": Es sei $\lambda \in \mathbb{C}$ kein Eigenwert von A. Dann gilt $\det(\lambda \mathrm{Id} - A) \neq 0$, woraus $\mathrm{rg}(\lambda \mathrm{Id} - A) = n$ folgt. Hieraus folgt (ii) wegen $\mathrm{rg}(\lambda \mathrm{Id} - A \,|\, B) \geq \mathrm{rg}(\lambda \mathrm{Id} - A)$.

"(i) $\Leftrightarrow$ (ii)": Wir beweisen dies mit Kontraposition, zeigen also "nicht (i) $\Leftrightarrow$ nicht (ii)".

"nicht (i) $\Leftarrow$ nicht (ii)": Wenn (ii) nicht gilt, existiert ein $\lambda \in \mathbb{C}$ mit $\mathrm{rg}(\lambda \mathrm{Id} - A \,|\, B) < n$. Also existiert ein $p \in \mathbb{R}^n$, $p \neq 0$ mit $p^T(\lambda \mathrm{Id} - A \,|\, B) = 0$, also

$$p^T A = \lambda p^T \text{ und } p^T B = 0.$$

Aus der ersten Gleichheit folgt $p^T A^k = \lambda^k p^T$ und damit insgesamt

$$p^T A^k B = \lambda^k p^T B = 0$$

für $k = 0, \ldots, n-1$. Also gilt $p^T(B \, AB \, \ldots \, A^{n-1}B) = 0$, womit $\mathrm{rg}(B \, AB \, \ldots \, A^{n-1}B) < n$ ist. Also ist (A, B) nicht kontrollierbar.

"nicht (i) $\Rightarrow$ nicht (ii)": Wenn (A, B) nicht kontrollierbar ist, existiert die Zerlegung

$$\widetilde{A} = T^{-1}AT = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}, \quad \widetilde{B} = T^{-1}B = \begin{pmatrix} B_1 \\ 0 \end{pmatrix}$$

gemäß Lemma 2.14 mit Koordinatentransformationsmatrix T.

Sei nun $\lambda \in \mathbb{C}$ ein Eigenwert von A_3^T zum Eigenvektor v. Dann gilt $v^T(\lambda \mathrm{Id} - A_3) = 0$. Damit gilt für $w^T = (0, v^T)$

$$w^T(\lambda \mathrm{Id} - \widetilde{A}) = \big(0^T(\lambda \mathrm{Id} - A_1) + v^T 0, \; 0^T(-A_2) + v^T(\lambda \mathrm{Id} - A_3)\big) = 0$$

und

$$w^T \widetilde{B} = \begin{pmatrix} 0^T B_1 \\ v^T 0 \end{pmatrix} = 0.$$

Mit $p^T = w^T T^{-1} \neq 0$ folgt dann

$$p^T(\lambda \mathrm{Id} - A \,|\, B) = w^T T^{-1}(\lambda \mathrm{Id} - A \,|\, B) = (w^T(\lambda \mathrm{Id} - \widetilde{A})T^{-1} \,|\, w^T \widetilde{B}) = 0,$$

weswegen (ii) nicht gilt. □

Bemerkung 2.17 Für zeitdiskrete Systems (1.4) mit $U = \mathbb{R}^m$ sind die Bedingungen für vollständige Kontrollierbarkeit vollkommen identisch. Es gibt aber einen entscheidenden Unterschied: Während Kontrollierbarkeit im Kontinuierlichen immer Kontrollierbarkeit in beliebig kurzer Zeit bedeutet, braucht man im Zeitdiskreten im schlechtesten Fall bis zu n Zeitschritte. Ein Beispiel hierfür ist das System

$$x(k+1) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x(k) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u(k)$$

mit $x \in \mathbb{R}^2$ und $u \in \mathbb{R}$. Hier gilt $(B \, AB) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, weswegen vollständige Kontrollierbarkeit gilt. Um das System von $(0,0)^T$ nach $(1,1)^T$ zu steuern, sind aber mindestens zwei Zeitschritte notwendig. Lemma 2.5 gilt im Zeitdiskreten tatsächlich nur für $s, t \geq n$. □

Kapitel 3

Stabilität und Stabilisierung

In diesem Kapitel werden wir uns mit dem Problem der Stabilisierung linearer Kontrollsysteme beschäftigen. Bevor wir dieses Problem angehen, müssen wir zunächst klären, was wir unter Stabilität verstehen.

3.1 Definitionen

In diesem und den folgenden zwei Abschnitten werden wir wichtige Resultate der Stabilitätstheorie linearer zeitinvarianter Differentialgleichungen (1.8)

$$\dot{x}(t) = Ax(t)$$

einführen. Die Darstellung wird dabei relativ knapp gehalten; eine ausführlichere Behandlung dieses Themas findet sich z.B. in dem Skript [5] sowie in vielen Lehrbüchern über gewöhnliche Differentialgleichungen. Wir beschränken uns hier auf die Stabilität von Gleichgewichten.

Definition 3.1 Ein Punkt $x^* \in \mathbb{R}^n$ heißt *Gleichgewicht* (auch *Ruhelage* oder *Equilibrium*) einer gewöhnlichen Differentialgleichung, falls für die zugehörige Lösung

$$x(t; x^*) = x^* \text{ für alle } t \in \mathbb{R}$$

gilt. □

Gleichgewichte haben wir bereits ohne formale Definition im einführenden Kapitel betrachtet. Man rechnet leicht nach, dass ein Punkt x^* genau dann ein Gleichgewicht einer allgemeinen zeitinvarianten Differentialgleichung $\dot{x}(t) = f(x(t))$ ist, wenn $f(x^*) = 0$ ist. Für die lineare Differentialgleichung (1.8) ist daher der Punkt $x^* = 0$ immer ein Gleichgewicht. Dieses Gleichgewicht $x^* = 0$ wollen wir in der folgenden Analyse näher betrachten.

Definition 3.2 Sei $x^* = 0$ das Gleichgewicht der linearen Differentialgleichung (1.8).

(i) Das Gleichgewicht $x^* = 0$ heißt *stabil*, falls für jedes $\varepsilon > 0$ ein $\delta > 0$ existiert, so dass die Ungleichung

$$\|x(t; x_0)\| \leq \varepsilon \text{ für alle } t \geq 0$$

für alle Anfangswerte $x_0 \in \mathbb{R}^n$ mit $\|x_0\| \leq \delta$ erfüllt ist.

(ii) Das Gleichgweicht $x^* = 0$ heißt *lokal asymptotisch stabil*, falls es stabil ist und darüberhinaus

$$\lim_{t\to\infty} x(t; x_0) = 0$$

gilt für alle Anfangswerte x_0 aus einer offenen Umgebung N von $x^* = 0$.

(iii) Das Gleichgewicht $x^* = 0$ heißt *global asymptotisch stabil*, falls (ii) mit $U = \mathbb{R}^n$ erfüllt ist.

(iv) Das Gleichgewicht $x^* = 0$ heißt *lokal* bzw. *global exponentiell stabil*, falls Konstanten $c, \sigma > 0$ existieren, so dass die Ungleichung

$$\|x(t; x_0)\| \leq ce^{-\sigma t}\|x_0\| \text{ für alle } t \geq 0$$

für alle x_0 aus einer Umgebung U von $x^* = 0$ (mit $U = \mathbb{R}^n$ im globalen Fall) erfüllt ist. □

Bemerkung 3.3 Die Stabilität aus (i) wird auch „Stabilität im Sinne von Ljapunov“ genannt, da dieses Konzept Ende des 19. Jahrhunderts vom russischen Mathematiker Alexander M. Ljapunov eingeführt wurde. Beachte, dass aus den Definitionen die Implikationen

$$\text{(lokal/global) exponentiell stabil } \Rightarrow \text{ (lokal/global) asymptotisch stabil } \Rightarrow \text{ stabil}$$

folgen. Die zweite Implikation ergibt sich direkt aus der Definition. Dass aus exponentieller Stabilität die asymptotische Stabilität folgt, sieht man folgendermaßen:
Für ein gegebenes ε folgt (i) mit $\delta = \varepsilon/c$, denn damit gilt für $\|x_0\| \leq \delta$ die Ungleichung $\|x(t; x_0)\| \leq ce^{-\sigma t}\|x_0\| \leq c\|x_0\| \leq \varepsilon$. Die in (ii) geforderte Konvergenz ist offensichtlich. □

3.2 Eigenwertkriterien

Der folgende Satz gibt Kriterien an die Matrix A, mit denen man Stabilität leicht überprüfen kann.

Satz 3.4 Betrachte die lineare zeintinvariante Differentialgleichung (1.8) für eine Matrix $A \in \mathbb{R}^{n\times n}$. Seien $\lambda_1, \ldots, \lambda_d \in \mathbb{C}$, $\lambda_l = a_l + ib_l$, die Eigenwerte der Matrix A, die hier so angeordnet seien, dass jedem Eigenwert λ_l ein Jordan-Block J_l in der Jordan'schen Normalform entspricht. Dann gilt:

(i) Das Gleichgewicht $x^* = 0$ ist stabil genau dann, wenn alle Eigenwerte λ_l nichtpositiven Realteil $a_l \leq 0$ besitzen und für alle Eigenwerte mit Realteil $a_l = 0$ der entsprechende Jordan-Block J_l eindimensional ist.

(ii) Das Gleichgewicht $x^* = 0$ ist lokal asymptotisch stabil genau dann, wenn alle Eigenwerte λ_l negativen Realteil $a_l < 0$ besitzen. In diesem Fall nennt man die Matrix A eine *Hurwitz-Matrix* oder kurz *Hurwitz*.

Beweisskizze: Zunächst überlegt man sich, dass alle Stabilitätseigenschaften unter linearen Koordinatentransformationen mit invertierbarer Transformationsmatrix $T \in \mathbb{R}^{n\times n}$ erhalten bleiben, da die Lösungen $y(t; y_0)$ des transformierten Systems mittels

$$y(t; y_0) = T^{-1}x(t; Ty_0)$$

ineinander umgerechnet werden können.

Es reicht also, die Stabilitätseigenschaften für die Jordan'sche Normalform

$$J = \begin{pmatrix} J_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \dots & 0 & J_d \end{pmatrix}$$

mit den Jordan-Blöcken der Form

$$J_l = \begin{pmatrix} \lambda_l & 1 & 0 & \cdots & 0 \\ 0 & \lambda_l & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \lambda_l & 1 \\ 0 & \cdots & \cdots & 0 & \lambda_l \end{pmatrix}, \tag{3.1}$$

$j = 1, \dots, d$, der Matrix A zu beweisen. Wir bezeichnen die zu $\dot{x}(t) = Jx(t)$ gehörigen Lösungen wiederum mit $x(t; x_0)$.

Aus den Eigenschaften der Matrix-Exponentialfunktion folgt nun, dass die allgemeine Lösung

$$x(t; x_0) = e^{Jt}x_0$$

für J die Form

$$x(t; x_0) = \begin{pmatrix} e^{J_1 t} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \dots & 0 & e^{J_d t} \end{pmatrix} x_0$$

besitzt. Weiter rechnet man nach, dass

$$e^{J_l t} = e^{\lambda_l t} \begin{pmatrix} 1 & t & \frac{t^2}{2!} & \cdots & \frac{t^{m-1}}{(m-1)!} \\ 0 & 1 & t & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \frac{t^2}{2!} \\ \vdots & \ddots & \ddots & 1 & t \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

ist, wobei $e^{\lambda_l t}$ die (übliche) skalare Exponentialfunktion ist, für die

$$|e^{\lambda_l t}| = e^{a_l t} \begin{cases} \to 0, & a_l < 0 \\ \equiv 1, & a_l = 0 \\ \to \infty, & a_l > 0 \end{cases}$$

für $t \to \infty$ gilt.

Die Einträge von $e^{J_l t}$ sind also genau dann beschränkt, wenn die Bedingung aus (i) erfüllt ist. Weil zudem für jedes $k \in \mathbb{N}$ und jedes $\varepsilon > 0$ ein $c > 0$ existiert mit

$$e^{a_l t} t^k \leq c e^{(a_l + \varepsilon)t}, \tag{3.2}$$

konvergieren die Einträge von $e^{J_l t}$ genau dann gegen 0, wenn (ii) erfüllt ist.

Dieses Verhalten der Matrix-Einträge überträgt sich bei der Matrix-Vektor-Multiplikation $e^{Jt}x_0$ auf die Lösungen, weswegen es äquivalent zur Stabilität bzw. asymptotischen Stabilität ist. □

Der Beweis von (iii) zeigt tatsächlich globale exponentielle Stabilität, da die Einträge in (3.2) exponentiell gegen 0 konvergieren. Die Konsequenz dieser Tatsache formulieren wir explizit in dem folgenden Satz.

Satz 3.5 Betrachte die lineare zeintinvariante Differentialgleichung (1.8) für eine Matrix $A \in \mathbb{R}^{n \times n}$. Seien $\lambda_1, \ldots, \lambda_d \in \mathbb{C}$, $\lambda_l = a_l + i b_l$, die Eigenwerte der Matrix A. Dann sind die folgenden vier Eigenschaften äquivalent.

(i) Alle Eigenwerte λ_l besitzen negativen Realteil $a_l < 0$, d.h. die Matrix ist Hurwitz.

(ii) Das Gleichgewicht $x^* = 0$ ist lokal asymptotisch stabil.

(iii) Das Gleichgewicht $x^* = 0$ ist global exponentiell stabil, wobei die Konstante $\sigma > 0$ aus Definition 3.2(iv) beliebig aus dem Intervall $(0, -\max_{l=1,\ldots,d} a_l)$ gewählt werden kann.

(iv) Die Norm der Matrix-Exponentialfunktion erfüllt $\|e^{At}\| \leq c e^{-\sigma t}$ für σ aus (iii) und eine von σ abhängige Konstante $c > 0$.

Beweis: (iii) $\Rightarrow$ (ii) folgt mit Bemerkung 3.3, (ii) $\Rightarrow$ (i) folgt aus Satz 3.4(iii) und (i) $\Rightarrow$ (iii) wurde im Beweis von Satz 3.4(iii) gezeigt. Schließlich folgt (iii) $\Leftrightarrow$ (iv) sofort aus der Definition der induzierten Matrix-Norm (und gilt dann für alle Normen auf $\mathbb{R}^{n \times n}$, weil diese äquivalent sind). □

Beispiel 3.6 Wir betrachten das Pendelmodell aus Kapitel 1 für $u \equiv 0$ und ohne Berücksichtigung der Bewegung des Wagens. Die Linearisierung im unteren (= herunterhängenden) Gleichgewicht $x^* = \pi$ liefert

$$A = \begin{pmatrix} 0 & 1 \\ -g & -k \end{pmatrix}$$

mit Eigenwerten

$$\lambda_{1/2} = -\frac{1}{2}k \pm \frac{1}{2}\sqrt{k^2 - 4g}.$$

Hierbei ist $\sqrt{k^2 - 4g}$ entweder komplex oder $< k$, weswegen man in jedem Fall $\mathrm{Re}\lambda_{1/2} < 0$ und damit exponentielle Stabilität erhält.

Die Linearisierung im oberen (= aufgerichteten) Gleichgewicht $x^* = 0$ lautet

$$A = \begin{pmatrix} 0 & 1 \\ g & -k \end{pmatrix}$$

liefert. Hier erhält man die Eigenwerte

$$\lambda_{1/2} = -\frac{1}{2}k \pm \frac{1}{2}\sqrt{k^2 + 4g},$$

deren größerer wegen $\sqrt{k^2 + 4g} > k$ immer positiv ist. Man erhält also keine Stabilität. □

Bemerkung 3.7 Für zeitdiskrete Systeme bleibt Satz 3.5 im Prinzip gleich, allerdings ändert sich in (i) die Bedingung "Realteil $a_l < 0$" zu "Betrag $|\lambda_l| < 1$" und in (iv) wird aus $\|e^{At}\| \leq ce^{-\sigma t}$ die Ungleichung $\|A^k\| \leq ce^{-\sigma k}$. Eine Matrix, bei der alle Eigenwerte die Ungleichung $|\lambda_l| < 1$ erfüllen, heißt *Schur-stabil*. □

3.3 Ljapunov Funktionen

In diesem Kapitel werden wir ein wichtiges Hilfsmittel zur Untersuchung asymptotisch stabiler Differentialgleichungen behandeln, nämlich die sogenannten Ljapunov Funktionen. Asymptotische (und auch exponentielle Stabilität) verlangen nur, dass die Norm $\|x(t)\|$ einer Lösung für $t \to \infty$ abnimmt. Für viele Anwendungen wäre es aber viel einfacher, wenn die Norm streng monoton in t fallen würde. Dies ist natürlich im Allgemeinen nicht zu erwarten. Wir können die strenge Monotonie aber erhalten, wenn wir die euklidische Norm $\|x(t)\|$ durch eine allgemeinere Funktion, nämlich gerade die Ljapunov Funktion, ersetzen.

Für lineare Systeme können wir uns auf sogenannte quadratische Ljapunov Funktionen beschränken, wie sie durch die folgende Definition gegeben sind.

Definition 3.8 Sei $A \in \mathbb{R}^{n\times n}$. Eine stetig differenzierbare Funktion $V : \mathbb{R}^n \to \mathbb{R}_0^+$ heißt *(quadratische) Ljapunov Funktion* für A, falls positive reelle Konstanten $c_1, c_2, c_3 > 0$ existieren, so dass die Ungleichungen

$$c_1\|x\|^2 \leq V(x) \leq c_2\|x\|^2$$

und

$$DV(x)Ax \leq -c_3\|x\|^2$$

für alle $x \in \mathbb{R}^n$ gelten. □

Der folgende Satz zeigt, dass die Existenz einer Ljapunov Funktion exponentielle Stabilität der zugehörigen Differentialgleichung impliziert.

Satz 3.9 Seien $A \in \mathbb{R}^{n\times n}$ eine Matrix und $x(t; x_0)$ die Lösungen des zugehörigen linearen Anfangswertproblems (1.8), (1.9). Dann gilt: Falls eine quadratische Ljapunov Funktion mit Konstanten $c_1, c_2, c_3 > 0$ existiert, so erfüllen alle Lösungen die Abschätzung

$$\|x(t; x_0)\| \leq ce^{-\sigma t}\|x_0\|$$

für $\sigma = c_3/2c_2$ und $c = \sqrt{c_2/c_1}$, d.h. das Gleichgewicht $x^* = 0$ ist exponentiell stabil und die Matrix A ist Hurwitz.

Beweis: Aus der Ableitungsbedingung für $x = x(\tau, x_0)$ folgt

$$\left.\frac{d}{dt}\right|_{t=\tau} V(x(t;x_0)) = DV(x(\tau;x_0))\dot{x}(\tau;x_0) = DV(x(\tau;x_0))Ax(\tau;x_0) \leq -c_3\|x(\tau;x_0)\|^2$$

Wegen $-\|x\|^2 \leq -V(x)/c_2$ folgt damit für $\lambda = c_3/c_2$ die Ungleichung

$$\frac{d}{dt}V(x(t;x_0)) \leq -\lambda V(x(t;x_0)).$$

Aus dieser Differentialungleichung folgt die Ungleichung

$$V(x(t;x_0)) \leq e^{-\lambda t}V(x_0),$$

(siehe z.B. den Beweis von [7, Satz 8.2]). Mit den Abschätzungen für $V(x)$ erhalten wir daraus

$$\|x(t;x_0)\|^2 \leq \frac{1}{c_1}e^{-\lambda t}V(x_0) \leq \frac{c_2}{c_1}e^{-\lambda t}\|x_0\|^2$$

und damit durch Ziehen der Quadratwurzel auf beiden Seiten die Ungleichung

$$\|x(t;x_0)\| \leq ce^{-\sigma t}\|x_0\|$$

für $c = \sqrt{c_2/c_1}$ und $\sigma = \lambda/2$. □

Wir wollen uns nun mit einer speziellen Klasse von Ljapunov Funktionen beschäftigen, bei denen V durch eine Bilinearform der Form x^TPx dargestellt wird, wobei $P \in \mathbb{R}^{n\times n}$.

Wir erinnern daran, dass eine Matrix $P \in \mathbb{R}^{n\times n}$ *positiv definit* heißt, falls $x^TPx > 0$ ist für alle $x \in \mathbb{R}^n$ mit $x \neq 0$. Das folgende Lemma fasst zwei Eigenschaften bilinearer Abbildungen zusammen.

Lemma 3.10 Sei $P \in \mathbb{R}^{n\times n}$. Dann gilt: (i) Es existiert eine Konstante $c_2 > 0$, so dass

$$-c_2\|x\|^2 \leq x^TPx \leq c_2\|x\|^2 \text{ für alle } x \in \mathbb{R}^n.$$

(ii) P ist positiv definit genau dann, wenn eine Konstante $c_1 > 0$ existiert mit

$$c_1\|x\|^2 \leq x^TPx \text{ für alle } x \in \mathbb{R}^n.$$

Beweis: Aus der Bilinearität folgt für alle $x \in \mathbb{R}^n$ mit $x \neq 0$ und $y = x/\|x\|$ die Gleichung

$$x^TPx = \|x\|^2y^TPy. \tag{3.3}$$

Da y^TPy eine stetige Abbildung in $y \in \mathbb{R}^n$ ist, nimmt sie auf der kompakten Menge $\{y \in \mathbb{R}^n \mid \|y\| = 1\}$ ein Maximum $c_{\max}$ und ein Minimum $c_{\min}$ an.

(i) Die Ungleichung (i) folgt nun aus (3.3) mit $c_2 = \max\{c_{\max}, -c_{\min}\}$.

(ii) Falls P positiv definit ist, ist $c_{\min} > 0$, und (ii) folgt mit $c_1 = c_{\min}$. Andererseits folgt die positive Definitheit von P sofort aus (ii), also erhalten wir die behauptete Äquivalenz. □

Hiermit erhalten wir die folgende Aussage.

Lemma 3.11 Seien A, $P \in \mathbb{R}^{n\times n}$ und $c_3 > 0$ so, dass die Funktion $V(x) = x^T Px$ die Ungleichung

$$DV(x)Ax \leq -c_3\|x\|^2$$

für alle $x \in \mathbb{R}^n$ erfüllt. Dann gilt: P ist genau dann positiv definit ist, wenn A Hurwitz ist. In diesem Fall ist V eine quadratische Ljapunov Funktion.

Beweis: Falls P positiv definit ist, folgt aus Lemma 3.10(ii) sofort, dass V eine quadratische Ljapunov Funktion ist, womit $x^* = 0$ exponentiell stabil und A folglich Hurwitz ist.

Falls P nicht positiv definit ist, gibt es ein $x_0 \in \mathbb{R}^n$ mit $x_0 \neq 0$ und $V(x_0) \leq 0$. Weil sich verschiedene Lösungen der Differentialgleichung nicht schneiden können, kann die Lösung $x(t; x_0)$ mit $x_0 \neq 0$ niemals 0 werden. Daher folgt aus der Ableitungsbedingung, dass $V(x(t; x_0))$ für alle $t \geq 0$ streng monoton fällt. Insbesondere gibt es also ein $c > 0$, so dass $V(x(t; x_0)) \leq -c$ für alle $t \geq 1$. Mit der ersten Abschätzung aus Lemma 3.10(i) folgt dann

$$\|x(t; x_0)\|^2 \geq c/c_2 > 0 \text{ für alle } t \geq 1.$$

Also konvergiert $x(t; x_0)$ nicht gegen den Nullpunkt, weswegen $x^* = 0$ nicht exponentiell stabil und A folglich nicht Hurwitz ist. □

Wir können das Ableitungskriterium vereinfachen, indem wir die bilineare Form der Ljapunov Funktion ausnutzen.

Lemma 3.12 Für eine bilineare Funktion $V(x) = x^T Px$ sind äquivalent:

(i) $DV(x)Ax \leq -c_3\|x\|^2$ für alle $x \in \mathbb{R}^n$ und eine Konstante $c_3 > 0$

(ii) Die Matrix $C = -A^T P - PA$ ist positiv definit.

Beweis: Wegen $x^T Py = y^T P^T x$ gilt $\frac{d}{dx}(x^T Py)Ax = \frac{d}{dx}(y^T P^T x)Ax = y^T P^T Ax = x^T A^T Py$. Daraus folgt nach Produktregel

$$DV(x)Ax = x^T A^T Px + x^T PAx = x^T(A^T P + PA)x = -x^T Cx.$$

Bedingung (i) ist also äquivalent zu

$$x^T Cx \geq c_3\|x\|^2 \text{ für alle } x \in \mathbb{R}^n.$$

Wegen Lemma 3.10 (ii) ist diese Bedingung genau dann für ein $c_3 > 0$ erfüllt, wenn C positiv definit ist. □

Die Gleichung in Lemma 3.12 (iii) wird auch *Ljapunov Gleichung* genannt. Es liegt nun nahe, diese Gleichung zur Konstruktion von Ljapunov Funktionen zu verwenden. Die Frage ist, wann kann man zu einer gegebenen Matrix A und einer gegebenen positiv definiten Matrix C eine Matrix P finden, so dass $A^T P + PA = -C$ gilt? Das folgende Lemma beantwortet diese Frage.

Lemma 3.13 Für eine Matrix $A \in \mathbb{R}^{n\times n}$ und eine positiv definite Matrix $C \in \mathbb{R}^{n\times n}$ hat die Ljapunov Gleichung

$$A^T P + PA = -C \tag{3.4}$$

genau dann eine (sogar eindeutige) positiv definite Lösung $P \in \mathbb{R}^{n\times n}$, wenn A Hurwitz ist.

Beweis: Falls eine positiv definite Lösung P von (3.4) existiert, ist die Funktion $V(x) = x^T Px$ wegen den Lemmas 3.12 und 3.11 eine quadratische Lyapunov Funktion, womit A Hurwitz ist.

Sei umgekehrt A Hurwitz und C positiv definit. Wir zeigen zunächst, dass (3.4) lösbar ist. O.B.d.A. können wir annehmen, dass A in Jordan'scher Normalform vorliegt, denn für $\widetilde{A} = TAT^{-1}$ sieht man leicht, dass P (3.4) genau dann löst, wenn $\widetilde{P} = (T^{-1})^T PT^{-1}$ die Gleichung

$$\widetilde{A}^T\widetilde{P} + \widetilde{P}\widetilde{A} = -(T^{-1})^T CT^{-1}$$

löst. Wir können also annehmen, dass A von der Form

$$A = \begin{pmatrix} \alpha_1 & \beta_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \beta_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \alpha_{n-1} & \beta_{n-1} \\ 0 & \cdots & \cdots & 0 & \alpha_n \end{pmatrix} \tag{3.5}$$

ist, wobei die α_i gerade Eigenwerte von A sind und die β_i entweder 0 oder 1 sind. Schreibt man die Spalten von P untereinander in einen Spaltenvektor $p \in \mathbb{R}^{n^2}$, und macht das gleiche für die Matrix C und einen Vektor c, so ist (3.4) äquivalent zu einem Gleichungssystem

$$\overline{A}p = -c,$$

mit einer geeigneten Matrix $\overline{A} \in \mathbb{C}^{n^2 \times n^2}$. Falls A in der Form (3.5) ist, sieht man durch Nachrechnen der Koeffizienten, dass $\overline{A}$ eine untere Dreiecksmatrix ist, d.h.

$$\overline{A} = \begin{pmatrix} \bar{\alpha}_1 & 0 & 0 & \cdots & 0 \\ * & \bar{\alpha}_2 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \bar{\alpha}_{n^2-1} & 0 \\ * & \cdots & \cdots & * & \bar{\alpha}_{n^2} \end{pmatrix},$$

wobei $*$ beliebige Werte bezeichnet, und die $\bar{\alpha}_i$ von der Form $\bar{\alpha}_i = \lambda_j + \lambda_k$ für Eigenwerte der Matrix A sind. Aus der linearen Algebra ist bekannt, dass

(i) bei einer Dreiecksmatrix die Elemente auf der Diagonalen gerade die Eigenwerte sind

(ii) eine Matrix genau dann invertierbar ist, wenn alle Eigenwerte ungleich Null sind.

Da A Hurwitz ist und folglich alle λ_i negativen Realteil haben, sind die $\bar{a}_i$ alle ungleich Null, also ist die Matrix $\overline{A}$ wegen (i) und (ii) invertierbar. Demnach gibt es genau eine Lösung des Gleichungssystems $\overline{A}p = c$ und damit genau eine Lösung P der Ljapunov Gleichung (3.4).

Es bleibt zu zeigen, dass diese Lösung P positiv definit ist. Wegen Lemma 3.12 erfüllt P alle Voraussetzungen von Lemma 3.11. Da A Hurwitz ist, muss P also nach Lemma 3.11 positiv definit sein. □

Der folgende Satz fasst das Hauptresultat dieses Abschnitts zusammen.

Satz 3.14 Für $A \in \mathbb{R}^{n\times n}$ gilt: Eine quadratische Ljapunov Funktion für die lineare Differentialgleichung (1.8) existiert genau dann, wenn $x^* = 0$ exponentiell stabil ist, d.h. wenn die Matrix A Hurwitz ist.

Beweis: Sei eine quadratische Ljapunov Funktion V gegeben. Dann ist A nach Satz 3.9 Hurwitz.

Sei A umgekehrt Hurwitz. Dann existiert nach Lemma 3.13 eine positiv definite Matrix P, die die Ljapunov Gleichung (3.4) für eine positiv definite Matrix C löst. Wegen Lemma 3.12 und Lemma 3.11 ist $V(x) = x^T Px$ dann eine quadratische Ljapunov Funktion. □

Die Existenz einer quadratischen Ljapunov Funktion ist also eine notwendige und hinreichende Bedingung für die exponentielle Stabilität des Gleichgewichts $x^* = 0$ und liefert damit eine Charakterisierung, die äquivalent zu der Eigenwertbedingung aus Satz 3.5 ist.

Beispiel 3.15 Für das im unteren Gleichgewicht linearisierte Pendelmodell mit

$$A = \begin{pmatrix} 0 & 1 \\ -g & -k \end{pmatrix}$$

ist die bilineare Ljapunov Funktion zu $C = \mathrm{Id}$ gegeben durch die Matrix

$$P = \begin{pmatrix} \frac{k^2+g^2+g}{2gk} & \frac{1}{2g} \\ \frac{1}{2g} & \frac{g+1}{2gk} \end{pmatrix}.$$

□

Bemerkung 3.16 Für zeitdiskrete Systeme ändert sich die untere Ungleichung in Definition 3.8 zu

$$V(Ax) - V(x) \leq -c_3\|x\|^2.$$

Die Lyapunov-Gleichung (3.4) ändert sich dadurch zu

$$A^T PA - P = -C. \tag{3.6}$$

Mit diesen Änderungen bleiben alle Sätze in diesem Abschnitt gültig. □

3.4 Das Stabilisierungsproblem für lineare Kontrollsysteme

Wir haben nun das technische Werkzeug, um uns wieder den linearen Kontrollsystemen zu widmen. In den Übungen haben wir gesehen, dass die Vorausberechnung einer Kontrollfunktion $u(t)$ auf großen Zeithorizonten i.A. nicht funktioniert, um ein System in einen gegebenen Punkt (o.B.d.A. 0) zu steuern und dort zu halten – selbst die geringen Fehler einer genauen numerischen Simulation reichten dort aus, um die Lösung weit von dem gewünschten Punkt zu entfernen.

Wir machen daher nun einen anderen Ansatz. Statt die Kontrolle als Steuerung – abhängig von t – anzusetzen, wählen wir nun eine Regelung, in der wir die Kontrollfunktion in jedem Zeitpunkt zustandsabhängig als $u(t) = F(x(t))$ für eine zu bestimmende Funktion

$F : \mathbb{R}^n \to \mathbb{R}^m$ ansetzen. Eine solche Funktion, die jedem Zustand einen Kontrollwert zuordnet, nennt man *Feedback* (auch *Zustandsfeedback*, *(Zustands-)Rückführung* oder kurz *Regler*). Da unser System linear ist, liegt es nahe, die Feedback-Funktion F linear zu wählen, also $u = Fx$ für ein $F \in \mathbb{R}^{m\times n}$. Dies hat den Vorteil, dass das entstehende System

$$\dot{x}(t) = Ax(t) + BFx(t) = (A + BF)x(t)$$

eine lineare zeitinvariante Differentialgleichung wird, auf die wir die Theorie der vorhergehenden Abschnitte anwenden können.

Um nun einen Zustand nach 0 zu steuern und ihn dort zu halten, können wir das folgende Stabilisierungsproblem lösen.

Definition 3.17 Gegeben sei ein lineares Kontrollsystem (1.3)

$$\dot{x}(t) = Ax(t) + Bu(t)$$

mit Matrizen $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$. Das *(Feedback-) Stabilisierungsproblem* für (1.3) besteht darin, eine lineare Abbildung $F : \mathbb{R}^m \to \mathbb{R}^n$ (bzw. die dazugehörige Matrix $F \in \mathbb{R}^{m\times n}$) zu finden, so dass die lineare gewöhnliche Differentialgleichung

$$\dot{x}(t) = (A + BF)x(t)$$

asymptotisch stabil ist. Diese Gleichung wird als *geschlossener Regelkreis* oder *closed loop System* bezeichnet. □

Aus unseren Kriterien für asymptotische Stabilität kann man leicht das folgende Lemma ableiten.

Lemma 3.18 Gegeben seien zwei Matrizen $A \in \mathbb{R}^{n\times n}$ und $B \in \mathbb{R}^{n\times m}$. Dann löst die Matrix $F \in \mathbb{R}^{m\times n}$ das Stabilisierungsproblem, falls alle Eigenwerte der Matrix $A + BF \in \mathbb{R}^{n\times n}$ negativen Realteil haben.

Wir werden uns im weiteren Verlauf mit der Frage beschäftigen, wann – zu gegebenen Matrizen A und B – eine solche Matrix F existiert und wie man sie berechnen kann.

Beispiel 3.19 Als einfaches und intuitiv lösbares Beispiel für ein Stabilisierungsproblem betrachten wir ein (sehr einfaches) Modell für eine Heizungsregelung. Nehmen wir an, dass wir wir die Temperatur x_1 in einem Raum an einem festgelegten Messpunkt regeln wollen. Der Einfachheit halber sei die gewünschte Temperatur durch Verschiebung der Skala auf $x_1^* = 0$ festgesetzt[1]. In dem Raum befindet sich ein Heizkörper mit Temperatur x_2, auf die wir mit der Kontrolle u Einfluss nehmen können. Die Veränderung von x_2 sei durch die Differentialgleichung $\dot{x}_2(t) = u(t)$ beschrieben, d.h. die Kontrolle u regelt die Zunahme (falls $u > 0$) bzw. Abnahme (falls $u < 0$) der Temperatur. Für die Temperatur x_1 im Messpunkt

[1] Die Größe x_1 sollte also als Abweichung von der gewünschten Temperatur und nicht als absoluter Wert interpretiert werden.

nehmen wir an, dass sie der Differentialgleichung $\dot{x}_1(t) = -x_1(t) + x_2(t)$ genügt, d.h. für konstante Heiztemperatur x_2 ergibt sich

$$x_1(t) = e^{-t}x_1(0) + (1 - e^{-t})x_2(0).$$

Mit anderen Worten nehmen wir an, dass die Raumtemperatur x_1 im Messpunkt exponentiell gegen die Temperatur des Heizkörpers konvergiert.

Aus diesem Modell erhalten wir das Kontrollsystem

$$\dot{x}(t) = \begin{pmatrix} -1 & 1 \\ 0 & 0 \end{pmatrix} x(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u(t).$$

Eine naheliegende Regelstrategie ergibt sich nun wie folgt: Falls $x_1 > x_1^* = 0$ ist, so vermindern wir die Temperatur in x_2, d.h., wir wählen $u < 0$. Im umgekehrten Fall, d.h. falls $x_1 < x_1^* = 0$ ist, erhöhen wir die Temperatur und setzen $u > 0$. Da unser Feedback linear sein soll, lässt sich dies durch die Wahl $F(x) = -\lambda x_1$ für ein $\lambda > 0$ erreichen, oder, in Matrix-Schreibweise $F = (-\lambda, 0)$ (beachte, dass hier $n = 2$ und $m = 1$ ist, F also eine 1×2-Matrix bzw. ein 2-dimensionaler Zeilenvektor ist). Damit erhalten wir das rückgekoppelte System

$$\dot{x}(t) = \begin{pmatrix} -1 & 1 \\ -\lambda & 0 \end{pmatrix} x(t).$$

Berechnet man die Eigenwerte für $\lambda > 0$, so sieht man, dass alle Realteile negativ sind. Wir haben also (ohne es zu wollen) das Stabilisierungsproblem gelöst und folglich konvergieren $x_1(t)$ und $x_2(t)$ für alle beliebige Anfangswerte exponentiell schnell gegen 0, insbesondere konvergiert x_1 exponentiell schnell gegen die gewünschte Temperatur $x_1^* = 0$. Damit haben wir bewiesen, dass unser von Hand konstruierter Regler tatsächlich das gewünschte Ergebnis erzielt.

Falls wir die Temperatur x_2 am Heizkörper messen können, so können wir auch $F(x) = -\lambda x_2$, bzw. in Matrix-Schreibweise $F = (0, -\lambda)$ setzen. Wiederum sieht man durch Betrachtung der Eigenwerte, dass das rückgekoppelte System für alle $\lambda > 0$ exponentiell stabil ist und damit das gewünschte Verhalten erzielt wird. Das Verhalten dieses Systems mit den zwei verschiedenen Feedbacks ist allerdings recht unterschiedlich. Wir werden dies in den Übungen genauer untersuchen. □

Bemerkung 3.20 In der Praxis ist der Zustand $x(t)$ eines Systems oft nicht vollständig messbar, stattdessen hat man nur Zugriff auf einen *Ausgangsvektor* $y = Cx$ für eine Matrix $C \in \mathbb{R}^{d \times n}$. In diesem Fall kann ein Feedback F nur vom Ausgangsvektor y abhängen, man spricht von einem *Ausgangsfeedback*.

Tatsächlich haben wir im obigen Beispiel so etwas Ähnliches gemacht, indem wir zur Konstruktion von F nur die „Information“ aus der Variablen x_1 bzw. x_2 verwendet haben. Wir werden im Folgenden zunächst annehmen, dass alle Zustände messbar sind und den allgemeinen Fall in Kapitel 4 behandeln. □

3.5 Lösung des Stabilisierungsproblems mit eindimensionaler Kontrolle

In diesem Abschnitt werden wir Bedingungen untersuchen, unter denen wir eine Lösung für das Stabilisierungsproblems aus Definition 3.17 mit eindimensionaler Kontrolle finden können. Insbesondere werden wir eine hinreichende und notwendige Bedingung an die Matrizen A und B in (1.3) angeben, unter der das Problem lösbar ist. Die einzelnen Schritte der Herleitung liefern dabei ein konstruktives Verfahren zur Berechnung eines stabilisierenden Feedbacks.

Bei der Herleitung werden wieder einmal Koordinatentransformationen eine wichtige Rolle spielen. Für eine Transformationsmatrix $T \in \mathbb{R}^{n\times n}$ ist das zu

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{3.7}$$

gehörige transformierte System

$$\dot{x}(t) = \widetilde{A}x(t) + \widetilde{B}u(t) \tag{3.8}$$

durch $\widetilde{A} = T^{-1}AT$ und $\widetilde{B} = T^{-1}B$ gegeben. Ein Feedback F für (3.7) wird mittels $\widetilde{F} = FT$ in eines für (3.8) transformiert; dies folgt sofort aus der Bedingung $T^{-1}(A + BF)T = \widetilde{A} + \widetilde{B}\widetilde{F}$.

Wir haben in Lemma 2.14 bereits gesehen, dass man Paare (A, B) mittels einer geeigneten Koordinatentransformation in die Form

$$\widetilde{A} = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}, \quad \widetilde{B} = \begin{pmatrix} B_1 \\ 0 \end{pmatrix},$$

d.h. in ein kontrollierbares Paar (A_1, B_1) und einen unkontrollierbaren Rest zerlegen kann.

Wir benötigen hier noch eine zweite Koordinatentransformation, die für kontrollierbare Systeme gilt, bei denen u eindimensional ist. In diesem Fall haben wir $m = 1$, also $B \in \mathbb{R}^{n\times 1}$, d.h. die Matrix B ist ein n-dimensionaler Spaltenvektor.

Lemma 3.21 Sei $A \in \mathbb{R}^{n\times n}$ und $B \in \mathbb{R}^{n\times 1}$. Dann gilt: Das Paar (A, B) ist kontrollierbar genau dann, wenn es eine Koordinatentransformation S gibt, so dass

$$\widetilde{A} = S^{-1}AS = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{pmatrix} \quad \text{und } \widetilde{B} = S^{-1}B = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

ist, wobei die Werte $\alpha_i \in \mathbb{R}$ gerade die Koeffizienten des charakteristischen Polynoms von A sind, d.h. $\chi_A(z) = z^n - \alpha_n z^{n-1} - \cdots - \alpha_2 z - \alpha_1$.

Beweis: Wir zeigen zunächst, dass für Matrizen $\widetilde{A}$ der angegebenen Form die α_i gerade die Koeffizienten des charakteristischen Polynoms sind. Dies folgt durch Induktion über n:

Für $n = 1$ ist die Behauptung sofort klar. Für den Induktionsschritt sei $A_n \in \mathbb{R}^{n\times n}$ von der Form des Satzes und $A_{n+1} \in \mathbb{R}^{n\times n}$ gegeben durch

$$A_{n+1} = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ 0 & & & \\ \vdots & & A_n & \\ \alpha_0 & & & \end{pmatrix}.$$

Entwickeln wir nun $\det(z\mathrm{Id}_{\mathbb{R}^{n+1}} - A_{n+1})$ nach der ersten Spalte, so ergibt sich

$$\chi_{A_{n+1}} = z\chi_{A_n}(z) - \alpha_0 = z^{n+1} - \alpha_n z^n - \cdots - \alpha_1 z - \alpha_0,$$

also nach Umnummerierung der α_i gerade der gewünschte Ausdruck.

Nehmen wir nun an, dass S existiert. Durch Nachrechnen sieht man leicht, dass

$$\widetilde{R} = (\widetilde{B}\,\widetilde{A}\widetilde{B}\,\ldots\,\widetilde{A}^{n-1}\widetilde{B}) = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & \iddots & * \\ 0 & 1 & * & * \\ 1 & * & \cdots & * \end{pmatrix} \tag{3.9}$$

gilt, wobei $*$ beliebige Werte bezeichnet. Diese Matrix hat vollen Rang, denn durch Umordnung der Zeilen (dies ändert den Rang nicht) erhalten wir eine obere Dreiecksmatrix mit lauter Einsen auf der Diagonalen, welche offenbar invertierbar ist, also vollen Rang besitzt. Daher ist $(\widetilde{A}, \widetilde{B})$ kontrollierbar und da Kontrollierbarkeit unter Koordinatentransformationen erhalten bleibt, ist auch das Paar (A, B) kontrollierbar.

Sei umgekehrt (A, B) kontrollierbar. Dann ist die Matrix $R = (B\,AB\,\ldots\,A^{n-1}B)$ invertierbar, folglich existiert R^{-1}. Wir zeigen nun zunächst, dass $R^{-1}AR = \widetilde{A}^T$ ist. Dazu reicht es zu zeigen, dass $AR = R\widetilde{A}^T$ ist. Dies folgt (unter Verwendung des Satzes von Cayley-Hamilton) aus der Rechnung

$$\begin{aligned} AR &= A(B\,AB\,\ldots\,A^{n-1}B) = (AB\,A^2B\,\ldots\,A^{n-1}B\;\; A^nB) \\ &= (AB\,A^2B\,\ldots\,A^{n-1}B\;\; \alpha_nA^{n-1}B + \cdots + \alpha_1B) \\ &= (B\,AB\,\ldots\,A^{n-1}B) \begin{pmatrix} 0 & \cdots & 0 & \alpha_1 \\ 1 & \cdots & 0 & \alpha_2 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & \alpha_n \end{pmatrix} = R\widetilde{A}^T \end{aligned}$$

Mit $\widetilde{R}$ aus (3.9) folgt mit analoger Rechnung die Gleichung $\widetilde{R}^{-1}\widetilde{A}\widetilde{R} = \widetilde{A}^T$ und damit

$$\widetilde{A} = \widetilde{R}\widetilde{A}^T\widetilde{R}^{-1} = \widetilde{R}R^{-1}AR\widetilde{R}^{-1}.$$

Aus den Definitionen von R und $\widetilde{R}$ folgt $R(1, 0, \ldots, 0)^T = B$ und $\widetilde{R}(1, 0, \ldots, 0)^T = \widetilde{B}$, also $R\widetilde{R}^{-1}\widetilde{B} = B$. Damit ergibt sich $S = R\widetilde{R}^{-1}$ als die gesuchte Transformation. □

Die durch Lemma 3.21 gegebene Form der Matrizen A und B wird auch *Regelungsnormalform* genannt. Beachte, dass sich die Koordinatentransformation S allein durch Kenntnis von A, B und den Koeffizienten des charakteristischen Polynoms von A berechnen lässt.

Mit Hilfe der Regelungsnormalform können wir nun die Lösung des Stabilisierungsproblems für $u \in \mathbb{R}$ angehen.

Zunächst drücken wir das Stabilisierungsproblem mit Hilfe des charakteristischen Polynoms aus. Dies können wir für beliebige Kontrolldimensionen machen.

Definition 3.22 Betrachte ein Kontrollsystem (1.3) mit Matrizen $A \in \mathbb{R}^{n\times n}$ und $B \in \mathbb{R}^{n\times m}$. Ein Polynom χ heißt *vorgebbar* für das Kontrollsystem, falls ein lineares Feedback $F \in \mathbb{R}^{m\times n}$ existiert, so dass $\chi = \chi_{A+BF}$ ist für das charakteristische Polynom χ_{A+BF} der Matrix $A + BF$. □

Da wir wissen, dass die Nullstellen des charakteristischen Polynoms gerade die Eigenwerte der zugehörigen Matrix sind, erhalten wir aus Lemma 3.18 sofort die folgende Charakterisierung.

Lemma 3.23 Betrachte ein Kontrollsystem (1.3) mit Matrizen $A \in \mathbb{R}^{n\times n}$ und $B \in \mathbb{R}^{n\times m}$. Dann gilt: Das Stabilisierungsproblem ist genau dann lösbar, falls ein vorgebbares Polynom existiert, dessen Nullstellen über $\mathbb{C}$ alle negativen Realteil haben.

Der folgende Satz zeigt die Beziehung zwischen der Kontrollierbarkeit von (A, B) und der Vorgebbarkeit von Polynomen.

Satz 3.24 Betrachte ein Kontrollsystem (1.3) mit Matrizen $A \in \mathbb{R}^{n\times n}$ und $B \in \mathbb{R}^{n\times 1}$, d.h. mit eindimensionaler Kontrolle. Dann sind die folgenden zwei Eigenschaften äquivalent.

(i) Das Paar (A, B) ist kontrollierbar.

(ii) Jedes Polynom der Form $\chi(z) = z^n - \beta_n z^{n-1} - \cdots - \beta_2 z - \beta_1$ mit $\beta_1, \ldots, \beta_n \in \mathbb{R}$ ist vorgebbar.

Beweis: (i) $\Rightarrow$ (ii): Sei (A, B) kontrollierbar und sei S die Koordinatentransformation aus Lemma 3.21. Wir setzen

$$\widetilde{F} = (\beta_1 - \alpha_1 \;\; \beta_2 - \alpha_2 \;\; \ldots \;\; \beta_n - \alpha_n) \in \mathbb{R}^{1\times n}.$$

Dann gilt

$$\begin{aligned}
\widetilde{A} + \widetilde{B}\widetilde{F} &= \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} (\beta_1 - \alpha_1 \;\; \beta_2 - \alpha_2 \;\; \ldots \;\; \beta_n - \alpha_n) \\
&= \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{pmatrix} + \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 \\ \beta_1 - \alpha_1 & \beta_2 - \alpha_2 & \cdots & \beta_n - \alpha_n \end{pmatrix} \\
&= \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \beta_1 & \beta_2 & \cdots & \beta_n \end{pmatrix}.
\end{aligned}$$

Aus der zweiten Aussage von Lemma 3.21 folgt, dass $\chi_{\widetilde{A}+\widetilde{B}\widetilde{F}} = \chi$ ist. Also ist, nach Rücktransformation, $F = \widetilde{F}S^{-1}$ die gesuchte Feedback Matrix, da das charakteristische Polynom einer Matrix invariant unter Koordinatentransformationen ist.

(ii) $\Rightarrow$ (i): Wir zeigen die Implikation „nicht (i) $\Rightarrow$ nicht (ii)“:

Sei (A, B) nicht kontrollierbar. Sei T die Koordinatentransformation aus Lemma 2.14. Dann ergibt sich für jedes beliebige Feedback $\widetilde{F} = (F_1 \;\; F_2)$

$$\widetilde{A} + \widetilde{B}\widetilde{F} = \begin{pmatrix} A_1 + B_1F_1 & A_2 + B_1F_2 \\ 0 & A_3 \end{pmatrix} =: \widetilde{D}.$$

Für das charakteristische Polynom dieser Matrix gilt

$$\chi_{\widetilde{D}} = \chi_{A_1+B_1F_1}\chi_{A_3},$$

daher sind (beachte, dass (A_1, B_1) kontrollierbar ist) die vorgebbaren Polynome gerade von der Form $\chi = \chi_k\chi_u$, wobei χ_k ein beliebiges normiertes Polynom vom Grad d ist und $\chi_u = \chi_{A_3}$ ist. Dies sind sicherlich weniger als die in (ii) angegebenen Polynome, weshalb (ii) nicht gelten kann. □

Natürlich ist es zur Stabilisierung nicht notwendig, dass jedes Polynom vorgebbar ist, wir brauchen lediglich eines zu finden, dessen Nullstellen nur negative Realteile haben. Der Beweis von Satz 3.24 lässt bereits erahnen, wann dies möglich ist.

Satz 3.25 Betrachte ein Kontrollsystem (1.3) mit Matrizen $A \in \mathbb{R}^{n\times n}$ und $B \in \mathbb{R}^{n\times 1}$, d.h. mit eindimensionaler Kontrolle. Seien $A_1 \in \mathbb{R}^{d\times d}$, $A_2 \in \mathbb{R}^{d\times(n-d)}$, $A_3 \in \mathbb{R}^{(n-d)\times(n-d)}$ und $B_1 \in \mathbb{R}^{d\times 1}$ die Matrizen aus Lemma 2.14 mit der Konvention, dass $A_1 = A$ und $B_1 = B$ ist, falls (A, B) kontrollierbar ist.

Dann sind die vorgebbaren Polynome von (1.3) gerade die Polynome der Form $\chi = \chi_k\chi_{A_3}$, wobei χ_k ein beliebiges normiertes Polynom vom Grad d und χ_{A_3} das charakteristische Polynom der Matrix A_3, also gerade der unkontrollierbare Anteil des charakteristischen Polynoms χ_A ist, vgl. Definition 2.15. Hierbei machen wir die Konvention $\chi_{A_3} = 1$ falls $d = n$.

Insbesondere gilt: Das Stabilisierungsproblem ist genau dann lösbar, wenn alle Eigenwerte von A_3 negativen Realteil haben (die Eigenwerte von A_3 werden auch "unkontrollierbare Eigenwerte" genannt). In diesem Fall nennen wir das Paar (A, B) *stabilisierbar*.

Beweis: Die erste Behauptung folgt sofort aus dem zweiten Teil des Beweises von Satz 3.24. Die Aussage über das Stabilisierungsproblem folgt dann sofort aus Lemma 3.23. □

Bemerkung 3.26 Alle Aussagen dieses Abschnitts gelten auch für zeitdiskrete Systeme, wenn man die Bedingung "Realteil des Eigenwerts kleiner als 0" durch "Betrag des Eigenwerts kleiner als 1" ersetzt. □

3.6 Lösung des Stabilisierungsproblems mit mehrdimensionaler Kontrolle

Die Resultate für mehrdimensionale Kontrolle $m > 1$ sind völlig analog zu denen für eindimensionale Kontrolle. Bei einer direkten Herangehensweise sind allerdings die Beweise etwas aufwändiger, da wir nicht direkt auf Lemma 3.21 zurückgreifen können. Wir werden den mehrdimensionalen Fall deswegen auf den Fall $m = 1$ zurückführen, indem wir das folgende Lemma verwenden, das als *Heymanns Lemma* bezeichnet wird.

Lemma 3.27 Betrachte ein Kontrollsystem (1.3) mit Matrizen $A \in \mathbb{R}^{n\times n}$ und $B \in \mathbb{R}^{n\times m}$. Das Paar (A, B) sei kontrollierbar. Sei $v \in \mathbb{R}^m$ ein Vektor mit $\overline{B} = Bv \neq 0$. Dann gibt es eine Matrix $\overline{F} \in \mathbb{R}^{m\times n}$, so dass das Kontrollsystem

$$\dot{x}(t) = (A + B\overline{F})x(t) + \overline{B}\bar{u}(t)$$

mit eindimensionaler Kontrolle $\bar{u}(t)$ kontrollierbar ist.

Beweis: Mittels der rekursiven Vorschrift $x_{i+1} := Ax_i + Bu_i$ mit geeigneten u_i konstruieren wir uns zunächst linear unabhängige Vektoren $x_1, \ldots, x_n \in \mathbb{R}^n$ mit der folgenden Eigenschaft: Für alle $l \in \{1, \ldots, n\}$ gilt

$$Ax_i \in V_l \text{ für } i = 1, \ldots, l-1 \text{ mit } V_l = \langle x_1, \ldots, x_l\rangle. \tag{3.10}$$

Setze dazu $x_1 = \overline{B}$ (wir können die $n \times 1$ Matrix $\overline{B}$ als Spaltenvektor auffassen) und beachte, dass die Eigenschaft (3.10) für $l = 1$ und jedes $x_1 \neq 0$ trivialerweise erfüllt ist.

Für $k \in 1, \ldots, n-1$ und gegebene linear unabhängige Vektoren $x_1, \ldots, x_k$, die (3.10) für $l \in \{1, \ldots, k\}$ erfüllen, konstruieren wir nun wie folgt einen Vektor x_{k+1}, so dass $x_1, \ldots, x_k, x_{k+1}$ linear unabhängig sind und (3.10) für $l \in \{1, \ldots, k+1\}$ erfüllen:

1. Fall: $Ax_k \notin V_k$: Setze $u_k := 0 \in \mathbb{R}^m$ und $x_{k+1} = Ax_k$.

2. Fall: $Ax_k \in V_k$: Wegen (3.10) folgt dann, dass V_k A-invariant ist. Aus Kapitel 2 wissen wir, dass $\langle A \,|\, \mathrm{im}\, B\rangle = \mathrm{im}\, R$ für die Erreichbarkeitsmatrix $R = (B\, AB \ldots A^{n-1}B)$ der kleinste A-invariante Unterraum ist, der das Bild von B enthält. Da (A, B) kontrollierbar ist, ist $\langle A \,|\, \mathrm{im}\, B\rangle = \mathbb{R}^n$. Weil V_k nun ein A-invarianter Unterraum mit $\dim V_k = k < n$ ist, kann dieser das Bild von B also nicht enthalten. Folglich gibt es ein $u_k \in \mathbb{R}^m$ mit $Ax_k + Bu_k \notin V_k$ und wir setzen $x_{k+1} = Ax_k + Bu_k$.

Wir konstruieren nun die gesuchte Abbildung $\overline{F}$ aus den Vektoren $x_1, \ldots, x_n$. Da die x_i linear unabhängig sind, ist die Matrix $X = (x_1 \ \ldots \ x_n)$ invertierbar, und wir können $\overline{F} := UX^{-1}$ für $U = (u_1, \ldots, u_n) \in \mathbb{R}^{m\times n}$ definieren, wobei die u_i für $i = 1, \ldots, n-1$ die in der obigen Rekursion verwendeten Kontrollvektoren sind und wir $u_n := 0 \in \mathbb{R}^m$ setzen. Damit gilt $\overline{F}x_i = u_i$ und deswegen $(A + B\overline{F})x_i = x_{i+1}$ für $i = 1, \ldots, n-1$. Wegen $\overline{B} = x_1$ folgt somit

$$(\overline{B} \ \ (A + B\overline{F})\overline{B} \ \ \ldots \ \ (A + B\overline{F})^{n-1}\overline{B}) = X,$$

also hat $(\overline{B} \ \ (A+B\overline{F})\overline{B} \ \ \ldots \ \ (A+B\overline{F})^{n-1}\overline{B})$ den Rang n, weswegen das Paar $(A+B\overline{F}, \overline{B})$ kontrollierbar ist. □

Mit diesem Resultat lassen sich nun die Sätze 3.24 und 3.25 leicht auf beliebige Kontrolldimensionen verallgemeinern.

Satz 3.28 Betrachte ein Kontrollsystem (1.3) mit Matrizen $A \in \mathbb{R}^{n\times n}$ und $B \in \mathbb{R}^{n\times m}$. Dann sind die folgenden zwei Eigenschaften äquivalent.

(i) Das Paar (A, B) ist kontrollierbar.

(ii) Jedes Polynom der Form $\chi(z) = z^n - \beta_n z^{n-1} - \cdots - \beta_2 z - \beta_1$ mit $\beta_1, \ldots, \beta_n \in \mathbb{R}$ ist vorgebbar.

Beweis: (i) $\Rightarrow$ (ii): Sei (A, B) kontrollierbar und χ gegeben. Seien $\overline{F} \in \mathbb{R}^{n\times m}$ und $\overline{B} \in \mathbb{R}^{n\times 1}$ die Matrizen aus Lemma 3.27 für ein $v \in \mathbb{R}^m$ mit $Bv \neq 0$ (beachte, dass solch ein $v \in \mathbb{R}^n$ existiert, da (A, B) kontrollierbar ist, also $B \neq 0$ ist). Dann ist das Paar $(A+B\overline{F}, \overline{B})$ kontrollierbar und aus Satz 3.24 folgt die Existenz eines Feedbacks $F_1 \in \mathbb{R}^{1\times n}$, so dass

$$\chi_{A+B\overline{F}+\overline{B}F_1} = \chi$$

ist. Wegen

$$A + B\overline{F} + \overline{B}F_1 = A + B\overline{F} + BvF_1 = A + B(\overline{F} + vF_1)$$

ist also $F = \overline{F} + vF_1$ das gesuchte Feedback.

(ii) $\Rightarrow$ (i): Völlig analog zum Beweis von Satz 3.24. □

Satz 3.29 Betrachte ein Kontrollsystem (1.3) mit Matrizen $A \in \mathbb{R}^{n\times n}$ und $B \in \mathbb{R}^{n\times m}$. Seien $A_1 \in \mathbb{R}^{d\times d}$, $A_2 \in \mathbb{R}^{d\times(n-d)}$, $A_3 \in \mathbb{R}^{(n-d)\times(n-d)}$ und $B_1 \in \mathbb{R}^{d\times m}$ die Matrizen aus Lemma 2.14 mit der Konvention, dass $A_1 = A$ und $B_1 = B$ ist, falls (A, B) kontrollierbar ist.

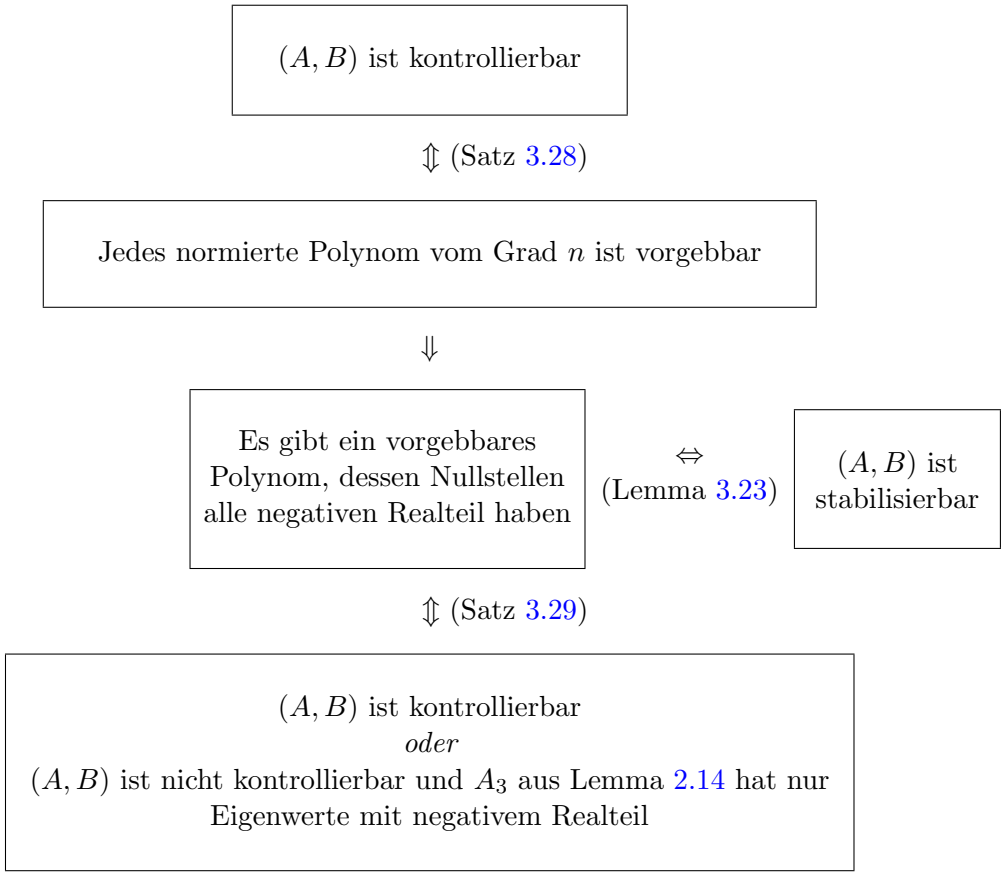
Dann sind die vorgebbaren Polynome von (1.3) gerade die Polynome der Form $\chi = \chi_k\chi_u$, wobei χ_k ein beliebiges normiertes Polynom vom Grad d und χ_u das charakteristische Polynom der Matrix A_3 ist, mit der Konvention $\chi_u = 1$ falls $d = n$.

Insbesondere gilt: Das Stabilisierungsproblem ist genau dann lösbar, wenn alle Eigenwerte von A_3 negativen Realteil haben. In diesem Fall nennen wir das Paar (A, B) *stabilisierbar*.

Beweis: Völlig analog zum Beweis von Satz 3.25. □

Bemerkung 3.30 Satz 3.29 wird oft als *Polverschiebungssatz* bezeichnet, da die Nullstellen des charakteristischen Polynoms in der Regelungstechnik als "Pole" bezeichnet werden (den Grund erklärt Bemerkung 5.15) und dieser Satz gerade angibt wie man diese Nullstellen durch geeignete Wahl des Feedbacks „verschieben" kann. □

Wir können die wesentlichen Ergebnisse über das Stabilisierungsproblem wie folgt schematisch darstellen:



Ersetzt man überall "negativer Realteil" durch "Betrag kleiner 1", so gelten diese Aussagen analog für zeitdiskrete Systeme.

3.7 Lokale Stabilisierung nichtlinearer Systeme

In diesem Abschnitt zeigen wir, dass ein lineares stabilisierendes Feedback zur lokalen Stabilisierung eines nichtlinearen Kontrollsystems verwendet werden kann. Grundlage dafür ist der folgende Satz aus der Theorie gewöhnlicher Differentialgleichungen.

Satz 3.31 Betrachte eine nichtlineare Differentialgleichung

$$\dot{x} = g(x) \tag{3.11}$$

mit Gleichgewicht $x^* \in \mathbb{R}^n$ und stetig differenzierbarem Vektorfeld $g : \mathbb{R}^n \to \mathbb{R}^n$. Betrachte zudem die Linearisierung

$$\dot{y} = \widehat{A}y \qquad \text{mit } \widehat{A} = \frac{d}{dx}g(x^*). \tag{3.12}$$

Dann ist das Gleichgewicht x^* lokal exponentiell stabil für Gleichung (3.11) genau dann, wenn das Gleichgewicht $y^* = 0$ global exponentiell stabil für Gleichung (3.12) ist.

Ein Beweis findet sich z.B. in [7, Satz 8.8].

Betrachten wir nun das nichtlineare Kontrollsystem

$$\dot{x} = f(x,u)$$

und seine Linearisierung

$$\dot{y} = Ay + Bv \qquad \text{mit } A = \frac{\partial}{\partial x} f(x^*,u^*),\ B = \frac{\partial}{\partial u} f(x^*,u^*).$$

Aus den Überlegungen in Kapitel 1 folgt, dass f, A und B die Beziehung $f(x,u) \approx A(x - x^*) + B(u - u^*)$ verbunden sind. Folglich müssen y und v als $y = x - x^*$ und $v = u - u^*$ gewählt werden.

Sei nun F ein stabilisierendes lineares Feedback für das lineare Kontrollsystem. Für das lineare System errechnet sich die Kontrolle dann als $v = Fu$, was für u und x die Beziehung $u = u^* + F(x - x^*)$ ergibt. Setzen wir diese in f ein, so erhalten wir die Differentialgleichung

$$\dot{x} = f(x, u^* + F(x - x^*)) =: g(x). \tag{3.13}$$

Die Linearisierung dieser Gleichung ist gegeben durch

$$\dot{y} = \widehat{A} y$$

mit

$$\widehat{A} = \frac{d}{dx} g(x^*) = \frac{d}{dx}\Big|_{x=x^*} f(x, u^* + F(x - x^*)) = \frac{\partial}{\partial x} f(x^*,u^*) + \frac{\partial}{\partial u} f(x^*,u^*) F = A + BF.$$

Da F das lineare System exponentiell stabilisiert, ist $y^* = 0$ folglich exponentiall stabil für (3.12) und aus Satz 3.31 folgt, dass x^* ein lokal exponentiell stabiles Gleichgewicht für das nichtlineare System mit linearem Feedback (3.13) ist. Das stabilisierende lineare Feedback stabilisiert das nichtlineare System also lokal in x^*.

Beispiel 3.32 Betrachte das nichtlineare invertierte Pendel (1.5)

$$\left.\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= -kx_2(t) + g\sin x_1(t) + u(t)\cos x_1(t) \\ \dot{x}_3(t) &= x_4(t) \\ \dot{x}_4(t) &= u(t) \end{aligned}\right\} =: f(x(t),u(t)).$$

Die Linearisierung in $(x^*,u^*) = (0,0)$ ergibt hier

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ g & -k & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{und} \quad B = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

vgl. (1.6). In den Übungen wurde ein stabilisierendes lineares Feedback $F : \mathbb{R}^4 \to \mathbb{R}$ für dieses lineare System berechnet. Die zugehörige Matrix $F \in \mathbb{R}^{1\times 4}$ lautet

$$F = \Big(-\frac{g+k^2}{g^2} - \frac{4k}{g} - 6 - g,\ -\frac{k}{g^2} - \frac{4}{g} - 4 + k,\ \frac{1}{g},\ \frac{k}{g^2} + \frac{4}{g} \Big)$$

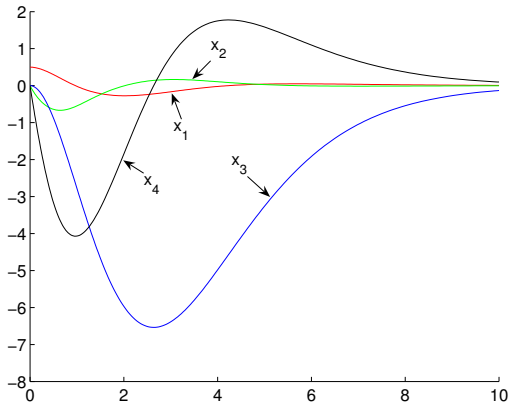


Abbildung 3.1: Lösungstrajektorie des nichtlinearen Pendels mit stabilisierendem linearem Feedback

Abbildung (3.1) zeigt, dass dieses Feedback auch das nichtlineare Pendel stabilisiert. Die Abbildung zeigt die Komponenten der Trajektorie $x(t, x_0, F)$ für $x_0 = (1/2,\, 0,\, 0,\, 0)^T$.

□

Kapitel 4

Beobachtbarkeit und Beobachter

Die im letzten Kapitel vorgestellte Lösung des Stabilisierungsproblems geht davon aus, dass der gesamte Vektor $x(t)$ zur Verfügung steht, um den Kontrollwert $u(t) = Fx(t)$ zu berechnen. Dies ist in der Praxis im Allgemeinen nicht der Fall. Dort kann man nur davon ausgehen, gewisse von $x(t)$ abhängige Werte $y(t) = C(x(t)) \in \mathbb{R}^k$ zu kennen, aus denen $u(t)$ dann berechnet werden muss. Da wir uns in diesem Teil der Vorlesung mit linearen Systemen beschäftigen, nehmen wir wieder an, dass die Funktion $C : \mathbb{R}^n \to \mathbb{R}^k$ linear ist, also durch eine Matrix $C \in \mathbb{R}^{k\times n}$ gegeben ist.

Definition 4.1 Ein *lineares Kontrollsystem mit Ausgang* ist gegeben durch[1] die Gleichungen

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t) \tag{4.1}$$

mit $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ und $C \in \mathbb{R}^{k\times n}$. □

In diesem Kapitel werden wir Bedingungen herleiten, unter denen das Stabilisierungsproblem für (4.1) lösbar ist und zeigen, wie man den Feedback-Regler in diesem Fall konstruieren muss.

4.1 Beobachtbarkeit und Dualität

Die wichtigste Frage bei der Analyse von (4.1) ist, wie viel "Information" in dem Ausgang $y(t) = Cx(t)$ enthalten ist. Dies wird durch die folgenden Definitionen formalisiert.

Definition 4.2 (i) Zwei Zustände $x_1, x_2 \in \mathbb{R}^n$ heißen *unterscheidbar*, falls ein $u \in \mathcal{U}$ und ein $t \geq 0$ existiert mit

$$Cx(t, x_1, u) \neq Cx(t, x_2, u).$$

(ii) Das System (4.1) heißt *beobachtbar*, falls alle Zustände $x_1, x_2 \in \mathbb{R}^n$ mit $x_1 \neq x_2$ unterscheidbar sind. □

[1] Manchmal wird auch die Variante $y(t) = Cx(t) + Du(t)$ mit $D \in \mathbb{R}^{k\times m}$ betrachtet. Die hier betrachtete Form erhält man dann durch die Wahl $D = 0$.

Das folgende Lemma zeigt, dass die Unterscheidbarkeit wegen der Linearität des Systems einfacher ausgedrückt werden kann.

Lemma 4.3 Zwei Zustände $x_1, x_2 \in \mathbb{R}^n$ sind genau dann unterscheidbar, wenn ein $t \geq 0$ existiert mit

$$Cx(t, x_1 - x_2, 0) \neq 0.$$

Beweis: Aus dem Superpositionsprinzip (1.15) folgt die Gleichung

$$x(t, x_1, u) - x(t, x_2, u) = x(t, x_1 - x_2, 0),$$

woraus wegen der Linearität von C sofort die Behauptung folgt. □

Aus diesem Lemma folgt, dass die Beobachtbarkeit von (4.1) nicht von u und damit nicht von B abhängt. Falls das System (4.1) beobachtbar ist, nennen wir daher das Paar (A, C) *beobachtbar*.

Zudem motiviert das Lemma die folgende Definition.

Definition 4.4 (i) Wir nennen $x_0 \in \mathbb{R}^n$ *beobachtbar*, falls ein $t \geq 0$ existiert mit

$$Cx(t, x_0, 0) \neq 0$$

und *unbeobachtbar auf* $[0, t]$, falls

$$Cx(s, x_0, 0) = 0$$

für alle $s \in [0, t]$.

(ii) Wir definieren die Mengen der *unbeobachtbaren Zustände auf* $[0, t]$ für $t > 0$ durch

$$\mathcal{N}(t) := \{x_0 \in \mathbb{R}^n \,|\, Cx(s, x_0, 0) = 0 \text{ für alle } s \in [0, t]\}$$

und die Menge der *unbeobachtbaren Zustände* durch

$$\mathcal{N} := \bigcap_{t>0} \mathcal{N}(t).$$

□

Das folgende Lemma zeigt die Struktur dieser Mengen auf.

Lemma 4.5 Für alle $t > 0$ gilt

$$\mathcal{N} = \mathcal{N}(t) = \bigcap_{i=0}^{n-1} \ker(CA^i).$$

Insbesondere ist $\mathcal{N}$ also ein linearer Unterraum, der zudem A-invariant ist, also $A\mathcal{N} \subseteq \mathcal{N}$ erfüllt.

Beweis: Ein Zustand $x_0 \in \mathbb{R}^n$ liegt genau dann in $\mathcal{N}(t)$, wenn gilt

$$0 = Cx(s, x_0, 0) = Ce^{As}x_0 \text{ für alle } s \in [0, t]. \tag{4.2}$$

Sei nun $x_0 \in \bigcap_{i=0}^{n-1} \ker(CA^i)$. Dann gilt mit dem Satz von Cayley-Hamilton $CA^i x_0 = 0$ für alle $i \in \mathbb{N}_0$. Aus der Reihendarstellung von e^{As} folgt damit $Ce^{As}x_0 = 0$ für alle $s \geq 0$ und daher (4.2), also $x_0 \in \mathcal{N}(t)$.

Sei umgekehrt $x_0 \in \mathcal{N}(t)$. Dann gilt nach (4.2) $Ce^{As}x_0 = 0$. Durch i-maliges Ableiten dieses Ausdrucks in $s = 0$ folgt

$$CA^i x_0 = 0, \quad i \in \mathbb{N}_0$$

und damit insbesondere $x_0 \in \ker CA^i$, $i = 0, \ldots, n-1$. Also folgt $x_0 \in \mathcal{N}(t)$.

Die A-Invarianz folgt mit dem Satz von Cayley-Hamilton aus der Darstellung von $\mathcal{N}$. □

Offenbar gibt es hier eine gewisse Ähnlichkeit mit der Kontrollierbarkeit, speziell mit den Mengen $\mathcal{R}(t)$ und $\mathcal{R}$. Wir zeigen nun, dass dies mehr als eine oberflächliche Ähnlichkeit ist, wenn wir ein geeignetet definiertes duales System einführen.

Definition 4.6 Zu einem durch (A, B, C) gegebenen Kontrollsystem (4.1) definieren wir das *duale System* durch die Matrizen (A^T, C^T, B^T). Ausgeschrieben lautet das duale System zu

$$\dot{x}(t) = Ax(t) + Bu(t), \quad y(t) = Cx(t), \qquad x(t) \in \mathbb{R}^n,\ u(t) \in \mathbb{R}^m,\ y(t) \in \mathbb{R}^k$$

also

$$\dot{x}(t) = A^T x(t) + C^T u(t), \quad y(t) = B^T x(t), \qquad x(t) \in \mathbb{R}^n,\ u(t) \in \mathbb{R}^k,\ y(t) \in \mathbb{R}^m.$$

□

In Worten ausgedrückt erhält man das duale System also durch Transponieren und Vertauschen von B und C, also von Eingangs- und Ausgangsmatrix.

Satz 4.7 Für ein durch (A, B, C) gegebenes Kontrollsystem (4.1) und das zugehörige durch (A^T, C^T, B^T) gegebene duale System definiere

$$\begin{array}{rclrcl} \mathcal{R} & = & \langle A \,|\, \operatorname{im} B \rangle & \mathcal{N} & = & \bigcap_{i=0}^{n-1} \ker(CA^i) \\ \mathcal{R}^T & = & \langle A^T \,|\, \operatorname{im} C^T \rangle & \mathcal{N}^T & = & \bigcap_{i=0}^{n-1} \ker(B^T(A^T)^i). \end{array}$$

Dann gilt

$$\mathcal{R}^T = \mathcal{N}^\perp \quad \text{und} \quad \mathcal{N}^T = \mathcal{R}^\perp.$$

Insbesondere gilt

$$(A, B, C) \text{ kontrollierbar} \iff (A^T, C^T, B^T) \text{ beobachtbar}$$

$$(A, B, C) \text{ beobachtbar} \iff (A^T, C^T, B^T) \text{ kontrollierbar}.$$

Beweis: Betrachte die Matrix

$$M = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} \in \mathbb{R}^{(n\cdot k)\times n}.$$

Für diese Matrix gilt mit Lemma 4.5 offenbar

$$\mathcal{N} = \ker M.$$

Andererseits ist

$$M^T = (C^T\, A^T C^T\, \ldots\, (A^T)^{n-1} C^T) \in \mathbb{R}^{n\times(n\cdot k)}$$

gerade die Erreichbarkeitsmatrix des dualen Systems, vgl. Definition 2.10, weswegen $\mathcal{R}^T = \operatorname{im} M^T$ gilt. Aus der linearen Algebra ist bekannt:

$$\operatorname{im} M^T = (\ker M)^\perp.$$

Hieraus folgt die erste Behauptung wegen

$$\mathcal{R}^T = \operatorname{im} M^T = (\ker M)^\perp = \mathcal{N}^\perp.$$

Durch Vertauschen der beiden Systeme folgt analog $\mathcal{R} = (\mathcal{N}^T)^\perp$, woraus die zweite Aussage wegen

$$\mathcal{R}^\perp = \left((\mathcal{N}^T)^\perp\right)^\perp = \mathcal{N}^T$$

folgt □

Wir können damit alle Aussagen zur Kontrollierbarkeit auf die Beobachtbarkeit übertragen und formulieren dies explizit für Korollar 2.13 und Lemma 2.14.

Definition 4.8 Die Matrix $(C^T, A^T C^T\, \ldots\, (A^T)^{n-1} C^T) \in \mathbb{R}^{n\times(k\cdot n)}$ heißt *Beobachtbarkeitsmatrix* des Systems (1.3). □

Korollar 4.9 Das System (4.1) ist genau dann beobachtbar, wenn

$$\operatorname{rg}(C^T, A^T C^T\, \ldots\, (A^T)^{n-1} C^T) = n$$

ist.

Beweis: Folgt aus Korollar 2.13 angewendet auf das duale System. □

Wir formulieren nun noch das Analogon zu der Zerlegung

$$\widetilde{A} = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}, \quad \widetilde{B} = \begin{pmatrix} B_1 \\ 0 \end{pmatrix}$$

aus Lemma 2.14.

Lemma 4.10 Sei (A, C) nicht beobachtbar, d.h., $\dim \mathcal{N} = l > 0$. Dann existiert ein invertierbares $T \in \mathbb{R}^{n\times n}$, so dass $\widetilde{A} = T^{-1}AT$, $\widetilde{B} = T^{-1}B$ und $\widetilde{C} = CT$ die Form

$$\widetilde{A} = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}, \quad \widetilde{B} = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \quad \widetilde{C} = (0 \;\; C_2)$$

mit $A_1 \in \mathbb{R}^{l\times l}$, $A_2 \in \mathbb{R}^{l\times(n-l)}$, $A_3 \in \mathbb{R}^{(n-l)\times(n-l)}$, $B_1 \in \mathbb{R}^{l\times m}$, $B_2 \in \mathbb{R}^{(n-l)\times m}$ und $C_2 \in \mathbb{R}^{k\times(n-l)}$ besitzen, wobei das Paar (A_3, C_2) beobachtbar ist.

Beweis: Lemma 2.14 angewendet auf das duale System (A^T, C^T) liefert $\widehat{T}$ mit

$$\widehat{T}^{-1}A^T\widehat{T} = \begin{pmatrix} \widehat{A}_1 & \widehat{A}_2 \\ 0 & \widehat{A}_3 \end{pmatrix}, \quad \widehat{T}^{-1}C^T = \begin{pmatrix} \widehat{C}_1 \\ 0 \end{pmatrix}.$$

Also folgt mit $S = (\widehat{T}^T)^{-1}$

$$S^{-1}AS = \begin{pmatrix} \widehat{A}_1^T & 0 \\ \widehat{A}_2^T & \widehat{A}_3^T \end{pmatrix}, \quad CS = \begin{pmatrix} \widehat{C}_1^T & 0 \end{pmatrix}.$$

Durch eine weitere Koordinatentransformation

$$Q = \begin{pmatrix} 0 & \mathrm{Id}_{\mathbb{R}^{n-l}} \\ \mathrm{Id}_{\mathbb{R}^l} & 0 \end{pmatrix}$$

folgt die behauptete Zerlegung mit $T = SQ$ und

$$A_1 = \widehat{A}_3^T,\ A_2 = \widehat{A}_2^T,\ A_3 = \widehat{A}_1^T,\ C_2 = \widehat{C}_1^T.$$

Als Alternative hier noch ein **direkter Beweis**, der ohne Lemma 2.14 auskommt:

Es sei $v_1, \ldots, v_l$ eine Basis von $\mathcal{N}$, also $\mathcal{N} = \langle v_1, \ldots, v_l\rangle$, die wir durch $w_1, \ldots, w_{n-l}$ zu einer Basis des $\mathbb{R}^n$ ergänzen. Definiere nun $T := (v_1, \ldots, v_l, w_1, \ldots, w_{n-l})$. Bezeichnen wir mit e_i wie üblich den i-ten Einheitsvektor im $\mathbb{R}^n$, so gilt $Te_i = v_i$, $i = 1, \ldots, l$, $Te_i = w_{i-l}$, $i = l+1, \ldots, n$, $T^{-1}v_i = e_i$, $i = 1, \ldots, l$ und $T^{-1}w_i = e_{i+l}$, $i = 1, \ldots, n-l$.

Wir zeigen zunächst die Struktur von $\widetilde{A}$. Angenommen, ein Eintrag im 0-Block von $\widetilde{A}$ ist ungleich Null. Dann gilt

$$\widetilde{A}e_i \notin \langle e_1, \ldots, e_l\rangle = T^{-1}\mathcal{N}$$

für ein $i \in \{1, \ldots, l\}$. Andererseits folgt aus der A-Invarianz von $\mathcal{N}$

$$\widetilde{A}e_i = T^{-1}ATe_i = T^{-1}Av_i \in T^{-1}\mathcal{N},$$

was ein Widerspruch ist.

Die Struktur von $\widetilde{C}$ folgt aus

$$\mathcal{N} = \bigcap_{i=0}^{n-1} \ker(CA^i) \subseteq \ker C.$$

Es muss also $v_i \in \ker C$ gelten und damit $\widetilde{C}e_i = CTe_i = Cv_i = 0$. Also müssen die ersten l Spalten von $\widetilde{C}$ gleich 0 sein.

Es bleibt, die Beobachtbarkeit von (A_3, C_2) zu zeigen. Für jedes $\tilde{x} \in \mathbb{R}^{n-l}$, $\tilde{x} \neq 0$ gilt

$$C_2 A_3^i \tilde{x} = \widetilde{C}\widetilde{A}^i \begin{pmatrix} 0 \\ \tilde{x} \end{pmatrix} = CA^i T \begin{pmatrix} 0 \\ \tilde{x} \end{pmatrix},$$

wobei wir in der ersten Gleichung die Struktur von $\widetilde{A}$ und $\widetilde{C}$ ausgenutzt haben. Wegen

$$w := T \begin{pmatrix} 0 \\ \tilde{x} \end{pmatrix} \notin \mathcal{N}$$

existiert nun ein $i \in \{0, \ldots, n-1\}$ mit $CA^i w \neq 0$ und damit $C_2 A_3^i \tilde{x} \neq 0$. Da $\tilde{x} \neq 0$ beliebig war, folgt

$$\bigcap_{i=0}^{n-1} \ker(C_2 A_3^i) = \{0\},$$

also die Beobachtbarkeit von (A_3, C_2). □

Bemerkung 4.11 Alle Aussagen in diesem Abschnitt gelten auch für zeitdiskrete Systeme. Die einzige Änderung ergibt sich in Lemma 4.5, das — analog zur Kontrollierbarkeit, vgl. Bemerkung 2.17 — im Zeitdiskreten nur für $t \geq n$ gilt. □

4.2 Entdeckbarkeit

Wir haben gesehen, dass (vollständige) Kontrollierbarkeit zwar hinreichend, nicht jedoch notwendig zur Lösung des Stabilisierungsproblems ist. Notwendig ist nur, dass das Paar (A, B) stabilisierbar ist, was nach Satz 3.29 genau dann der Fall ist, wenn alle Eigenwerte des unkontrollierbaren Anteils A_3 der Matrix A negative Realteile besitzen.

Ähnlich verhält es sich mit der Beobachtbarkeit. Um das Stabilisierungsproblem für das System (4.1) zu lösen, braucht man die Beobachtbarkeit nicht. Es reicht eine schwächere Bedingung, die durch die folgende Definition gegeben ist.

Definition 4.12 Das System (4.1) heißt *entdeckbar* (oder auch *asymptotisch beobachtbar*), falls

$$\lim_{t \to \infty} x(t, x_0, 0) = 0 \quad \text{für alle } x_0 \in \mathcal{N}.$$

□

Dies bedeutet, dass die Lösungen für unbeobachtbare Anfangswerte und $u \equiv 0$ bereits gegen 0 konvergieren. Anschaulich gesprochen wird die Information über diese Anfangswerte für das Stabilisierungsproblem nicht benötigt, da die zugehörigen Lösungen ja bereits gegen 0 konvergieren, also asymptotisch (und damit auch exponentiell) stabil sind.

Das folgende Lemma charakterisiert die Entdeckbarkeit für die Zerlegung aus Lemma 4.10.

Lemma 4.13 System (4.1) ist genau dann entdeckbar, wenn die Matrix A_1 aus Lemma 4.10 Hurwitz ist, also nur Eigenwerte mit negativem Realteil besitzt.

Beweis: Beachte zunächst, dass die Entdeckbarkeit unter Koordinatenwechseln erhalten bleibt, wir können also alle Rechnungen in der Basis von Lemma 4.10 durchführen.

In der Basis von Lemma 4.10 ist $\mathcal{N}$ gerade durch

$$\mathcal{N} = \left\{ x_0 \in \mathbb{R}^n \;\middle|\; x_0 = \begin{pmatrix} x_0^1 \\ 0 \end{pmatrix},\ x_0^1 \in \mathbb{R}^l \right\}$$

gegeben. Aus der Form der Matrix $\widetilde{A}$ folgt damit, dass alle Lösungen zu Anfangswerten $x_0 \in \mathcal{N}$ als

$$x(t, x_0, 0) = e^{\widetilde{A}t} x_0 = \begin{pmatrix} e^{A_1 t} x_0^1 \\ 0 \end{pmatrix}$$

geschrieben werden können.

Aus der Entdeckbarkeit folgt nun $x(t, x_0, 0) \to 0$ für alle $x \in \mathcal{N}$, also $e^{A_1 t} x_0^1 \to 0$ für alle $x_0^1 \in \mathbb{R}^l$. Dies ist nur möglich, wenn A_1 Hurwitz ist.

Umgekehrt folgt aus der Hurwitz-Eigenschaft von A_1 die Konvergenz $e^{A_1 t} x_0^1 \to 0$ für alle $x_0^1 \in \mathbb{R}^l$, also $x(t, x_0, 0) \to 0$ für alle $x \in \mathcal{N}$ und damit die Entdeckbarkeit. □

Der folgende Satz zeigt, dass die Entdeckbarkeit gerade die duale Eigenschaft zur Stabilisierbarkeit ist.

Satz 4.14 (A, C) ist entdeckbar genau dann, wenn (A^T, C^T) stabilisierbar ist.

Beweis: Wir bezeichnen die Komponenten der Zerlegung aus Lemma 4.10 angewendet auf (A, C) mit A_1, A_2, A_3, C_2 und die Komponenten der Zerlegung aus Lemma 2.14 angewendet auf (A^T, C^T) mit $\widehat{A}_1, \widehat{A}_2, \widehat{A}_3, \widehat{C}_1$. Aus dem Beweis von Lemma 4.10 folgt mit dieser Notation gerade $A_1 = \widehat{A}_3^T$.

Nach Lemma 4.13 folgt, dass Entdeckbarkeit von (A, C) gerade äquivalent zur Hurwitz-Eigenschaft von A_1 ist. Andererseits folgt aus Satz 3.29, dass (A^T, C^T) genau dann stabilisierbar ist, wenn $\widehat{A}_3$ Hurwitz ist. Da die Eigenwerte von $\widehat{A}_3$ und $\widehat{A}_3^T = A_1$ übereinstimmen, folgt die behauptete Äquivalenz. □

Bemerkung 4.15 Um die Aussagen dieses Abschnitts ins Zeitdiskrete zu übertragen, müssen lediglich die Eigenwertbedingungen von "negativ" auf "Betrag kleiner als 1" geändert werden. □

4.3 Dynamische Beobachter

Ein naheliegender Ansatz zur Lösung des Stabilisierungsproblems für (4.1) ist die Wahl $u(t) = Fy(t)$. Dies kann funktionieren (vgl. Beispiel 3.19, wo wir $C = (0\ \ 1)$ und $C = (1\ \ 0)$ betrachtet haben), muss aber nicht, wie das kontrollierbare und beobachtbare System (4.1) mit

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \text{und} \quad C = (1\ \ 0)$$

zeigt, vgl. die Übungen. Tatsächlich ist dieses System nicht einmal dann stabilisierbar, wenn $F(y(t))$ eine beliebige stetige Funktion $F : \mathbb{R} \to \mathbb{R}$ sein darf.

Wir wollen daher nun eine Methode zur Stabilisierung entwickeln, die immer funktioniert, wenn (4.1) stabilisierbar und entdeckbar ist. Die Methode funktioniert für ein durch die Matrizen (A, B, C) gegebenes System (4.1) wie folgt:

(1) Entwerfe ein stabilisierendes lineares Feedback F für (A, B)

(2) Entwerfe einen Algorithmus, der aus den gemessenen Ausgängen $y(s)$, $s \in [0, t]$, einen Schätzwert $z(t) \approx x(t)$ ermittelt

(3) Regle das System (4.1) mittels $u(t) = Fz(t)$.

Schritt (1) können wir mit den Methoden aus Kapitel 3 bereits lösen. In diesem Abschnitt werden wir Schritt (2) betrachten und im folgenden Abschnitt dann beweisen, dass die Methode mit den Schritten (1)–(3) tatsächlich funktioniert.

Der Grund dafür, dass das obige Beispiel nicht stabilisiert werden kann, liegt darin, dass Beobachtbarkeit nicht verlangt, dass $Cx_0 \neq 0$ ist für $x_0 \neq 0$. Es wird lediglich verlangt, dass $Cx(t; t_0, x_0, 0) \neq 0$ für $t > 0$. Um zu erkennen, dass der Schätzwert $z(t) \neq 0$ sein sollte (und das Feedback reagieren muss), muss der Algorithmus in Schritt (2) also die Ausgangswerte über einen längeren Zeitraum verwenden, nicht nur den aktuellen Wert. Dies erreichen wir, indem wir den Schätzwert $z(t)$ als Lösung eines geeignet formulierten Kontrollsystem definieren, in dem neben der Kontrollfunktion $u(t)$ der Ausgang $y(t)$ von (4.1) eine weitere Eingangsfunktion bildet. Die folgende Definition formalisiert diese Idee.

Definition 4.16 Ein *dynamischer Beobachter* (oder auch *Luenberger-Beobachter*) für (4.1) ist ein lineares Kontrollsystem der Form

$$\dot{z}(t) = Jz(t) + Ly(t) + Ku(t) \tag{4.3}$$

mit $J \in \mathbb{R}^{n\times n}$, $L \in \mathbb{R}^{n\times k}$, $K \in \mathbb{R}^{n\times m}$, so dass für alle Anfangswerte x_0, $z_0 \in \mathbb{R}^n$ und alle Kontrollfunktionen $u \in \mathcal{U}$ für die Lösungen $x(t, x_0, u)$ und $z(t, z_0, u, y)$ von (4.1), (4.3) mit $y(t) = Cx(t, x_0, u)$ die Abschätzung

$$\|x(t, x_0, u) - z(t, z_0, u, y)\| \leq ce^{-\sigma t}\|x_0 - z_0\|$$

für geeignete Konstanten $c, \sigma > 0$ gilt. □

In der Praxis kann System (4.3) z.B. numerisch gelöst werden, um die Werte $z(t)$ zu bestimmen.

Der folgende Satz zeigt, wann ein dynamischer Beobachter existiert; im Beweis wird dieser explizit konstruiert.

Satz 4.17 Ein dynamischer Beobachter für (4.1) existiert genau dann, wenn das System entdeckbar ist.

Beweis: "⇐" Da (4.1) entdeckbar ist, ist (A^T, C^T) stabilisierbar. Wir können also ein lineares Feedback $\widehat{F} \in \mathbb{R}^{k\times n}$ finden, so dass $A^T + C^T\widehat{F}$ Hurwitz ist. Mit $G = \widehat{F}^T$ ist dann auch $A + GC = (A^T + C^T\widehat{F})^T$ Hurwitz.

Wir wählen nun in (4.3) $J = A + GC$, $L = -G$ und $K = B$, also

$$\dot{z}(t) = (A + GC)z(t) - Gy(t) + Bu(t).$$

Schreiben wir kurz $x(t) = x(t, x_0, u)$, $z(t) = z(t, z_0, u, y)$ und $e(t) = z(t) - x(t)$, so gilt für $e(t)$ die Differentialgleichung

$$\begin{aligned}\dot{e}(t) &= \dot{z}(t) - \dot{x}(t)\\ &= (A + GC)z(t) - Gy(t) + Bu(t) - Ax(t) - Bu(t)\\ &= (A + GC)z(t) - GCx(t) - Ax(t)\\ &= (A + GC)(z(t) - x(t)) \;=\; (A + GC)e(t)\end{aligned}$$

Aus der Hurwitz-Eigenschaft von $A + GC$ folgt damit

$$\|e(t)\| \le ce^{-\sigma t}\|e(0)\|$$

für geeignetes $c, \sigma > 0$, was wegen $e(t) = z(t) - x(t)$ und $e(0) = z_0 - x_0$ gerade die gewünschte Abschätzung liefert.

"⇒" Sei $x_0 \in \mathcal{N}$, also $y(t) = Cx(t, x_0, 0) = 0$ für alle $t \ge 0$. Für $z_0 = 0$ gilt damit $z(t, z_0, 0, y) = z(t, 0, 0, 0) = 0$. Damit folgt aus der Eigenschaft des dynamischen Beobachters

$$\|x(t, x_0, 0)\| = \|x(t, x_0, 0) - z(t, z_0, 0, y)\| \le ce^{-\sigma t}\|x_0 - z_0\| = ce^{-\sigma t}\|x_0\| \to 0$$

für $t \to \infty$. Also gilt $x(t, x_0, 0) \to 0$ und damit die Entdeckbarkeit. □

4.4 Lösung des Stabilisierungsproblems mit Ausgang

Wir wollen nun die im vorherigen Abschnitt angegebene Methode zur Stabilisierung analysieren und zeigen, dass diese zum Erfolg führt, wenn man in Schritt (2) den dynamischen Beobachter (4.3) verwendet.

Aus den Schritten (1)–(3) unter Verwendung von (4.3) in Schritt (2) ergibt sich die Feedback-Gleichung

$$u(t) = Fz(t), \quad \dot{z}(t) = Jz(t) + Ly(t) + KFz(t). \tag{4.4}$$

Diese Form von Feedback nennt man *dynamisches Ausgangsfeedback*, da $u(t)$ aus dem *Ausgang* $y(t) = Cx(t)$ berechnet wird und das Feedback eine "interne" *Dynamik* besitzt, die gerade durch die Differentialgleichung für z gegeben ist[2].

[2] Im Gegensatz dazu nennt man das in Kapitel 3 behandelte Feedback $u(t) = Fx(t)$ *statisches Zustandsfeedback*.

Definition 4.18 Ein dynamisches Ausgangsfeedback (4.4) löst das *Stabilisierungsproblem mit Ausgang*, wenn das durch Einsetzen von (4.4) entstehende System von Differentialgleichungen

$$\begin{aligned} \dot{x}(t) &= Ax(t) + BFz(t) \\ \dot{z}(t) &= Jz(t) + LCx(t) + KFz(t) \end{aligned}$$

mit Lösungen $\binom{x(t)}{z(t)} \in \mathbb{R}^{2n}$ exponentiell stabil ist. □

Satz 4.19 Gegeben sei ein Kontrollsystem (4.1) mit Matrizen (A, B, C). Dann ist das Stabilisierungsproblem mit Ausgang genau dann im Sinne von Definition 4.18 lösbar, wenn (A, B) stabilisierbar und (A, C) entdeckbar ist.

In diesem Fall ist (4.4) mit dem im Beweis von Satz 4.17 konstruierten dynamischen Beobachter (4.3) und einem stabilisierendes Feedback $F \in \mathbb{R}^{m\times n}$ für (A, B) ein stabilisierendes dynamisches Feedback.

Beweis: "⇐": Es sei (A, B) stabilisierbar und (A, C) entdeckbar. Es sei $F \in \mathbb{R}^{m\times n}$ ein stabilisierendes Feedback für (A, B) und (4.3) der im Beweis von Satz 4.17 konstruierte dynamischen Beobachter. Dann ergibt sich das mittels (4.4) geregelte System zu

$$\begin{aligned} \begin{pmatrix} \dot{x}(t) \\ \dot{z}(t) \end{pmatrix} &= \begin{pmatrix} A & BF \\ LC & J + KF \end{pmatrix} \begin{pmatrix} x(t) \\ z(t) \end{pmatrix} \\ &= \begin{pmatrix} A & BF \\ -GC & A + GC + BF \end{pmatrix} \begin{pmatrix} x(t) \\ z(t) \end{pmatrix} \\ &= T^{-1} \begin{pmatrix} A + BF & BF \\ 0 & A + GC \end{pmatrix} T \begin{pmatrix} x(t) \\ z(t) \end{pmatrix}. \end{aligned}$$

mit

$$T = \begin{pmatrix} \mathrm{Id}_{\mathbb{R}^n} & 0 \\ -\mathrm{Id}_{\mathbb{R}^n} & \mathrm{Id}_{\mathbb{R}^n} \end{pmatrix}, \qquad T^{-1} = \begin{pmatrix} \mathrm{Id}_{\mathbb{R}^n} & 0 \\ \mathrm{Id}_{\mathbb{R}^n} & \mathrm{Id}_{\mathbb{R}^n} \end{pmatrix}$$

Da die exponentielle Stabilität unter Koordinatentransformationen erhalten bleibt, reicht es nun nachzuweisen, dass die Matrix in der letzten Zeile der Rechnung Hurwitz ist. Für Blockdreiecksmatrizen sind die Eigenwerte nun aber gerade gleich den Eigenwerten der Diagonalblöcke $A + BF$ und $A + GC$. Da $A + BF$ nach Wahl von F Hurwitz ist und $A + GC$ nach Wahl von G im Beweis von Satz 4.17 ebenfalls Hurwitz ist, hat obige Matrix also nur Eigenwerte mit negativem Realteil und ist damit Hurwitz.

"⇒": Mit der Koordinatentransformation T aus Lemma 2.14 erhält man für das transformierte System die Gleichungen

$$\begin{aligned} \dot{x}^1(t) &= A_1 x^1(t) + A_2 x^2(t) + B_1 F z(t) \\ \dot{x}^2(t) &= A_3 x^2(t) \\ \dot{z}(t) &= Jz(t) + LCx(t) + KFz(t) \end{aligned}$$

mit $x(t) = T \begin{pmatrix} x^1(t) \\ x^2(t) \end{pmatrix}$. Nehmen wir nun an, dass (A, B) nicht stabilisierbar ist. Dann besitzt A_3 Eigenwerte mit positivem Realteil, der Ursprung ist also nicht asymptotisch stabil für die Gleichung $\dot{x}^2(t) = A_3 x^2(t)$ und es gibt daher einen Anfangswert x_0^2 mit $x^2(t, x_0^2) \not\to 0$. Wählen wir also

$$x_0 = T \begin{pmatrix} x_0^1 \\ x_0^2 \\ z_0 \end{pmatrix} \in \mathbb{R}^{2n}$$

mit x_0^1, z_0 beliebig, so gilt $x(t, x_0, Fz) \not\to 0$ für jede Wahl des dynamischen Feedbacks. Dies widerspricht der Tatsache, dass das Stabilisierungsproblem lösbar ist, das Paar (A, B) ist also stabilisierbar.

Die Entdeckbarkeit von (A, C) folgt analog zum Beweis von "⇒" in Satz 4.17. □

Bemerkung 4.20 Die Konstruktionen und Aussagen in diesem und dem vorhergehenden Abschnitt gelten mit den offensichtlichen Änderungen analog für zeitdiskrete Systeme. □

Kapitel 5

Analyse im Frequenzbereich

Ein nicht unerheblicher Teil der modernen Kontroll- und Systemtheorie ist aus der Elektrotechnik heraus entstanden, in der das Verhalten von Schaltungen mit Eingangs- und Ausgangssignalen betrachtet wird. Als Beispiel kann hierbei z.B. ein Verstärker dienen, der ein Eingangssignal (von einem Mikrophon, einem Handy etc.) in ein Ausgangssignal umwandelt, das dann an die Lautsprecher geschickt wird. Ein anderes Beispiel ist ein (analoges) Radio, in dem das Eingangssignal (die elektromagnetischen Wellen) in ein hörbares Ausgangssignal umgewandelt wird. Stellen wir uns den Verstärker bzw. das Radio als Kontrollsystem vor, so können wir das Eingangssignal gemäß mit u und das Ausgangssignal mit y bezeichnen. Dies ändert die Interpretation dieser Funktionen: $u(t)$ ist nun ein von außen kommendes Signal (statt einer von uns wählbaren Kontrollfunktion) und $y(t)$ ist ein Ausgangssignal, das bestimmten Kriterien genügen soll (statt einer Messgröße). Es ändert aber zunächst nichts an der mathematischen Darstellung des Zusammenhangs zwischen u und y über das System (4.1). Der Anfangswert wird bei dieser Betrachtung üblicherweise als $x_0 = 0$ gewählt. Man geht also davon aus, dass sich das System bis zur Zeit $t = 0$ in der Ruhelage 0 befindet und ab dann durch das Eingangssignal $u(t)$, $t \geq 0$, beeinflusst wird.

Die beiden genannten Anwendungsbeispiele zeigen, dass Frequenzen eine wichtige Rolle bei dieser Betrachtungsweise spielen. Aus diesem Grunde werden u und y bei dieser Art der Betrachtung nicht als Funktionen der Zeit sondern der Frequenz dargestellt. Zu diesem Zweck betrachten wir zunächst die sogenannte Laplace-Transformation.

5.1 Laplace-Transformation

Es sei $\mathbb{K} = \mathbb{R}$ oder $\mathbb{C}$ und $\mathbb{R}_0^+ = [0, \infty)$. Wir bezeichnen mit $L^1_{loc}(\mathbb{R}_0^+, \mathbb{K}^m)$ die Menge aller Funktionen $u : \mathbb{R}_0^+ \to \mathbb{K}^m$, die auf jedem kompakten Intervall in $\mathbb{R}_0^+$ Lebesgue-integrierbar sind und mit $L^1(\mathbb{R}_0^+, \mathbb{K}^m)$ die Menge der Funktionen $u : \mathbb{R}_0^+ \to \mathbb{K}^m$, die auf ganz $\mathbb{R}_0^+$ Lebesgue-integrierbar sind. Für ein $u \in L^1_{loc}(\mathbb{R}_0^+, \mathbb{K}^m)$ und ein $\alpha \in \mathbb{R}$ definiere $u_\alpha : \mathbb{R}_0^+ \to \mathbb{K}^m$ mittels $u_\alpha(t) := u(t)e^{-\alpha t}$. Dann definieren wir den Raum der α-*exponentiell integrierbaren Funktionen* als

$$\mathcal{E}_\alpha(\mathbb{K}^m) := \{u \in L^1_{loc}(\mathbb{R}_0^+, \mathbb{K}^m) \,|\, u_\alpha \in L^1(\mathbb{R}_0^+, \mathbb{K}^m)\}.$$

Beispiel 5.1 Die Funktion $u(t) = e^t$ liegt als stetige Funktion offenbar in $L^1_{loc}(\mathbb{R}^+_0, \mathbb{R})$, wegen

$$\int_0^t e^\tau d\tau = e^t - 1 \to \infty$$

für $t \to \infty$ liegt sie aber nicht in $L^1(\mathbb{R}^+_0, \mathbb{R})$. Für $\alpha > 1$ gilt

$$\int_0^t u_\alpha(\tau)d\tau = \int_0^t e^\tau e^{-\alpha\tau}d\tau = \frac{1}{1-\alpha}(e^{(1-\alpha)t} - 1) \to \frac{1}{\alpha - 1}$$

für $t \to \infty$. Damit existiert zunächst das unendliche Riemann-Integral und wegen $u_\alpha(t) \geq 0$ auch das unendliche Lebesgue-Integral. Folglich liegt $u(t) = e^t$ in $\mathcal{E}_\alpha(\mathbb{R})$ für alle $\alpha > 1$. □

Definition 5.2 Die Funktionen in $\mathcal{E}_\alpha(\mathbb{K}^m)$ heißen *Laplace-transformierbar*. Die (einseitige) *Laplace-Transformation* für ein $u \in \mathcal{E}_\alpha(\mathbb{K}^m)$ ist für alle $s \in \mathbb{C}_\alpha := \{s \in \mathbb{C} \,|\, \mathrm{Re}(s) > \alpha\}$ definiert als

$$\hat{u}(s) := (\mathcal{L}u)(s) := \int_0^\infty u(t)e^{-st}dt.$$

Die Laplace-Transformierte $\hat{u} = \mathcal{L}u$ ist damit eine Funktion von $\mathbb{C}_\alpha$ nach $\mathbb{C}^m$. □

Beispiel 5.3 Laplace-Transformationen einiger Funktionen von $\mathbb{R}^+_0$ nach $\mathbb{R}$ mit $a \in \mathbb{C}$, $\omega \in \mathbb{R}$, $m \in \mathbb{N}_0$:

(a)	$u(t) = 1$	$\Rightarrow$	$\hat{u}(s) = \frac{1}{s}$	für $\mathrm{Re}(s) > 0$
(b)	$u(t) = \sin(\omega t)$	$\Rightarrow$	$\hat{u}(s) = \frac{\omega}{\omega^2 + s^2}$	für $\mathrm{Re}(s) > 0$
(c)	$u(t) = \cos(\omega t)$	$\Rightarrow$	$\hat{u}(s) = \frac{s}{\omega^2 + s^2}$	für $\mathrm{Re}(s) > 0$
(d)	$u(t) = e^{at}$	$\Rightarrow$	$\hat{u}(s) = \frac{1}{s - a}$	für $\mathrm{Re}(s) > \mathrm{Re}(a)$
(e)	$u(t) = e^{at}\sin(\omega t)$	$\Rightarrow$	$\hat{u}(s) = \frac{\omega}{\omega^2 + (s-a)^2}$	für $\mathrm{Re}(s) > \mathrm{Re}(a)$
(f)	$u(t) = e^{at}\cos(\omega t)$	$\Rightarrow$	$\hat{u}(s) = \frac{s-a}{\omega^2 + (s-a)^2}$	für $\mathrm{Re}(s) > \mathrm{Re}(a)$
(g)	$u(t) = \frac{t^m}{m!}e^{at}$	$\Rightarrow$	$\hat{u}(s) = \frac{1}{(s-a)^{m+1}}$	für $\mathrm{Re}(s) > \mathrm{Re}(a)$

□

Bemerkung 5.4 Wenngleich das Integral in der Laplace-Transformation nur für die hier angegebenen Werte von $\mathrm{Re}(s)$ definiert ist, ist der berechnete Ausdruck für einen größeren Bereich von Werten von s definiert. In (d) beispielsweise ist $\hat{u}(s)$ für alle $s \neq a$ definiert. Im Folgenden werden wir für $\hat{u}$ stets alle Argumente $s \in \mathbb{C}$ zulassen, für die der berechnete Ausdruck definiert ist. □

Die Umkehrung der Laplace-Transformation ist gegeben durch

$$(\mathcal{L}^{-1}\hat{u})(t) := \frac{1}{2\pi i}\int_{\beta-i\infty}^{\beta+i\infty} e^{st}\hat{u}(s)ds = \frac{e^{\beta t}}{2\pi i}\int_{-\infty}^{\infty} e^{i\omega t}\hat{u}(\beta+i\omega)d\omega.$$

Genauer gilt für alle $u \in \mathcal{E}_\alpha(\mathbb{K}^m)$ und beliebiges $\beta > \alpha$ die Gleichung $\mathcal{L}^{-1}\mathcal{L}u(t) = u(t)$ für fast alle $t \in \mathbb{R}_0^+$; falls u stetig ist gilt dies sogar für alle $t \in \mathbb{R}_0^+$, vgl. [11, Theorem A.3.19].

Im Folgenden sind einige wichtige Rechenregeln für die Laplace-Transformation aufgeführt. Dabei sind $a, a_1, a_2 \in \mathbb{R}$ und $u, u_1, u_2 \in \mathcal{E}_\alpha(\mathbb{K}^m)$. Weitere Annahmen sind unten zusammengefasst.

$$\begin{aligned}
(i) \quad & \mathcal{L}(a_1u_1 + a_2u_2)(s) &&= a_1\hat{u}_1(s) + a_2\hat{u}_2(s)\\
(ii) \quad & \mathcal{L}(u(a\,\cdot))(s) &&= \frac{1}{a}\hat{u}\left(\frac{s}{a}\right), \quad \text{für } a > 0\\
(iii) \quad & \mathcal{L}(u(\cdot - a))(s) &&= e^{-sa}\hat{u}(s), \quad \text{für } a > 0\\
(iv) \quad & \mathcal{L}(e^{a\,\cdot}u)(s) &&= \hat{u}(s-a)\\
(v) \quad & \mathcal{L}(\dot{u})(s) &&= s\hat{u}(s) - u(0)\\
(vi) \quad & \mathcal{L}\left(\int_0^{\cdot} u(\tau)d\tau\right)(s) &&= \frac{1}{s}\hat{u}(s)\\
(vii) \quad & \mathcal{L}(\cdot^k u)(s) &&= (-1)^k\frac{d^k\hat{u}}{ds^k}(s)\\
(viii) \quad & \mathcal{L}(u_1 \star u_2)(s) &&= \hat{u}_1(s)\hat{u}_2(s)\\
(ix) \quad & \lim_{t\to 0, t>0} u(t) &&= \lim_{s\to\infty} s\hat{u}(s)
\end{aligned}$$

In (iii) setzen wir dabei voraus, dass u auf $[-a, \infty)$ definiert ist mit $u(t) = 0$ für alle $t \in [-a, 0]$. In (v) nehmen wir an, dass u auf $(-\varepsilon, \infty)$ für ein $\varepsilon > 0$ definiert und in s differenzierbar ist. Falls u in 0 unstetig ist, muss $u(0)$ in (v) durch $\lim_{t\to 0, t<0} u(t)$ ersetzt werden. In $(viii)$ ist $u_1 \star u_2(t) = \int_0^t u_1(t-\tau)u_2(\tau)d\tau$ die *Faltung*.

5.2 Die Übertragungsfunktion

Die Übertragungsfunktion dient dazu, das Eingangs-Ausgangsverhalten eines Systems mit Hilfe der Laplace-Transformation auszudrücken. Mit dem Eingangs-Ausgangsverhalten bezeichnet man die Abbildung $u \mapsto y$ mit $y(t) = Cx(t, 0, u)$, also die Abbildung, die der Eingangsfunktion u die Ausgangsfunktion der zugehörigen Lösung mit Anfangswert $x_0 = 0$ zuordnet.

Wir betrachten nun, wie diese Abbildung für die Laplace-transformierten Signale aussieht. Dazu betrachten wir wieder das System (4.1), also

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t)$$

mit $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ und $C \in \mathbb{R}^{k\times n}$.

Satz 5.5 Betrachte das Kontrollsystem (4.1). Sei $u \in \mathcal{U}$, $u \in \mathcal{E}_\alpha(\mathbb{R}^m)$ und $y(t) = Cx(t,0,u)$. Dann ist y Laplace-transformierbar und es gilt

$$\hat{y}(s) = G(s)\hat{u}(s)$$

mit $G(s) = C(s\mathrm{Id} - A)^{-1}B$.

Beweis: Gemäß (1.14) gilt

$$y(t) = C\int_0^t e^{A(t-\tau)}Bu(\tau)d\tau.$$

Da $u \in \mathcal{E}_\alpha(\mathbb{R}^m)$ gilt, ist u exponentiell beschränkt, ebenso ist $\|e^{At}\|$ durch $e^{\|A\|t}$ exponentiell beschränkt. Folglich ist der Integrand exponentiell beschränkt, damit auch das Integral und weil x und y als Ergebnisse einer Integration zudem stetig sind, gilt $x \in \mathcal{E}_\alpha(\mathbb{R}^n)$, $y \in \mathcal{E}_\alpha(\mathbb{R}^k)$ für geeignetes (hinreichend großes) $\alpha > 0$.

Wenden wir nun die Laplace-Transformation auf (4.1) an, so erhalten wir mit Rechenregeln (i), (v) und $x_0 = 0$

$$s\hat{x}(s) = A\hat{x}(s) + B\hat{u}(s), \qquad \hat{y}(s) = C\hat{x}(s)$$

für alle $s \in \mathbb{C}$ mit $\mathrm{Re}(s) > \alpha$. Die erste Gleichung ist äquivalent zu

$$s\hat{x}(s) - A\hat{x}(s) = B\hat{u}(s) \quad \Leftrightarrow \quad (s\mathrm{Id} - A)\hat{x}(s) = B\hat{u}(s).$$

Für alle $s \in \mathbb{C}$, die keine Eigenwerte von A sind (also insbesondere für s mit hinreichend großem Realteil) ist die Matrix auf der linken Seite invertierbar und es folgt

$$\hat{x}(s) = (s\mathrm{Id} - A)^{-1}B\hat{u}(s) \quad \Rightarrow \quad \hat{y}(s) = C\hat{x}(s) = C(s\mathrm{Id} - A)^{-1}B\hat{u}(s) = G(s)\hat{u}(x).$$

□

Definition 5.6 Die Funktion $G : \mathbb{C} \to \mathbb{C}^{k\times m}$ aus Satz 5.5 heißt *Übertragungsfunktion* (auf englisch *transfer function*). □

Bemerkung 5.7 (i) Aus der Darstellung

$$(s\mathrm{Id} - A)^{-1} = \frac{1}{\det(s\mathrm{Id} - A)}\mathrm{adj}(s\mathrm{Id} - A)$$

mit der adjunkten Matrix $\mathrm{adj}(s\mathrm{Id} - A)$ folgt, dass $G : \mathbb{C} \to \mathbb{C}^{k\times m}$ eine matrixwertige Funktion mit rationalen Einträgen ist, d.h. mit Einträgen der Form

$$g_{ij}(s) = \frac{p_{ij}(s)}{q_{ij}(s)} \tag{5.1}$$

mit Polynomen p_{ij}, q_{ij}, für deren Grad gilt[1] $\deg p_{ij} < \deg q_{ij} \leq n$.

(ii) Die sogenannte *Realisierungstheorie* befasst sich mit der Frage, ob es zu einer gegebenen Funktion $G : \mathbb{C} \to \mathbb{C}^{k\times m}$ ein Kontrollsystem (4.1) gibt, so dass G die Übertragungsfunktion

[1] Für Ausgänge der Form $y(t) = Cx(t) + Du(t)$ gilt $G(s) = D + C(s\mathrm{Id} - A)^{-1}B$ und $\deg p_{ij} \leq \deg q_{ij} \leq n$.

dieses Kontrollsystems ist. Man kann zeigen, dass das für jede propere[2] rationale Matrixfunktion tatsächlich der Fall ist, allerdings sind A, B, C dabei in der Regel nicht eindeutig.

(iii) Definieren wir $g(t) := Ce^{At}B$, so folgt aus der Lösungsdarstellung

$$y(t) = \int_0^t Ce^{A(t-\tau)}Bu(\tau)d\tau = \int_0^t g(t-\tau)u(\tau)d\tau = g \star u(t).$$

Mit der Rechenregel (*viii*) der Laplace-Transformation ergibt sich

$$\hat{y}(s) = \mathcal{L}(g \star u)(s) = \hat{g}(s)\hat{u}(s).$$

Also gilt für die Übertragungsfunktion $G = \hat{g}$ (wenn wir die Definition der Laplace-Transformation in der natürlichen Weise auf matrixwertige Funktionen verallgemeinern).

□

Beispiel 5.8 Wir betrachten das herunterhängende und das invertierte linearisierte Pendel, jeweils ohne Berücksichtigung der Wagenkoordinaten, also

$$A = \begin{pmatrix} 0 & 1 \\ -g & -k \end{pmatrix}, \qquad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

und

$$A = \begin{pmatrix} 0 & 1 \\ g & -k \end{pmatrix}, \qquad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

In beiden Fällen sei $C = (1\ 0)$, d.h. der Ausgang misst die Position des Pendels.

Für das herunterhängende Pendel ergibt sich dann

$$(s\mathrm{Id} - A)^{-1} = \begin{pmatrix} s & -1 \\ g & s+k \end{pmatrix}^{-1} = \begin{pmatrix} \frac{s+k}{ks+s^2+g} & \frac{1}{ks+s^2+g} \\ \frac{-g}{ks+s^2+g} & \frac{s}{ks+s^2+g} \end{pmatrix}$$

und damit

$$G(s) = C(s\mathrm{Id} - A)^{-1}B = \frac{1}{ks+s^2+g}.$$

Analog ergibt sich für das invertierte Pendel

$$G(s) = C(s\mathrm{Id} - A)^{-1}B = \frac{1}{ks+s^2-g}.$$

□

[2] Proper heißt, dass $\deg p_{ij} \leq \deg q_{ij}$ für alle i, j.

5.3 Eingangs-Ausgangs Stabilität

Wir führen nun einen Stabilitätsbegriff ein, der zu der Eingangs-Ausgangs-Sichtweise der Übertragungsfunktion G passt.

Definition 5.9 Ein Kontrollsystem heißt *Eingangs-Ausgangs-stabil* (kurz E/A-stabil), falls eine Konstante $K > 0$ existiert, so dass für jede auf $\mathbb{R}_0^+$ beschränkte Funktion $u \in \mathcal{U}$ und den zugehörigen Ausgang

$$y(t) = C \int_0^t e^{A(t-\tau)} B u(s) d\tau$$

zum Anfangswert $x_0 = 0$ die Ungleichung $\|y\|_\infty \leq K\|u\|_\infty$ gilt. □

Bemerkung 5.10 (i) Man kann zeigen, dass E/A-Stabilität äquivalent zu der Implikation "$\|u\|_\infty < \infty \Rightarrow \|y\|_\infty < \infty$" ist. In dieser Form findet sich die Definition der E/A-Stabilität in vielen Büchern. Der Beweis dieser Äquivalenz verlangt aber einige technische Abschätzungen, die wir hier aus Zeitgründen vermeiden. Für unsere Zwecke ist die obige Definition im Folgenden günstiger.

(ii) Um den bisherigen Stabilitätsbegriff (A bzw. das geregelte System mit Feedback ist exponentiell stabil, d.h. alle Eigenwerte von A bzw. des geregelten Systems haben negativen Realteil) von dem Begriff der E/A-Stabilität zu unterscheiden, nennen wir die Stabilität von A auch *Zustandsstabilität*. □

Eine erste hinreichende und notwendige Bedingung gibt das folgende Lemma.

Lemma 5.11 Ein System (4.1) ist genau dann E/A-stabil, falls für $g(t) = Ce^{At}B$ gilt

$$g_{\max} := \int_0^\infty \|g(t)\| dt < \infty. \tag{5.2}$$

Beweis: "⇒": Das System sei E/A-stabil. Wir zeigen

$$\int_0^\infty |\gamma_{ij}(t)| dt \leq K \tag{5.3}$$

für alle Komponentenfunktionen γ_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, m$ von $g = (\gamma_{ij})_{i=1,\ldots,k, j=1,\ldots,m}$, woraus (5.2) folgt.

Zu gegebenem $t > 0$ sei dazu u gegeben durch $u(\tau) := \operatorname{sgn}(\gamma_{ij}(t-\tau))e_j$ für $\tau \in [0, t]$. Damit gilt $[g(t-\tau)u(\tau)]_i = |\gamma_{ij}(t-\tau)|$. Setzen wir $u(\tau) = 0$ für $\tau > t$, so gilt $\|u\|_\infty = 1$ und damit für den zugehörigen Ausgang $\|y\|_\infty \leq K$, folglich auch $|y_i(t)| \leq K$ für alle $t \geq 0$. Damit folgt

$$K \geq |y_i(t)| = \left| \int_0^t [g(t-\tau)u(\tau)]_i d\tau \right| = \left| \int_0^t |\gamma_{ij}(t-\tau)| d\tau \right| = \int_0^t |\gamma_{ij}(t-\tau)| d\tau = \int_0^t |\gamma_{ij}(\tau)| d\tau.$$

Weil dies für alle $t \geq 0$ gilt, folgt (5.3).

"⇐": Es sei $g_{\max} < \infty$ und es sei u ein Eingangssignal mit $\|u\|_\infty < \infty$. Dann gilt für alle $t \geq 0$

$$\|y(t)\| = \left\| \int_0^t g(t-\tau)u(\tau)d\tau \right\| \leq \int_0^t \|g(t-\tau)\|\|u(\tau)\|d\tau \leq \int_0^t \|g(t-\tau)\|d\tau \|u\|_\infty = g_{\max}\|u\|_\infty.$$

Folglich ist das System E/A-stabil mit $K = g_{\max}$. □

Korollar 5.12 Falls (4.1) zustandsstabil ist, also A Hurwitz ist, so ist (4.1) auch E/A-stabil.

Beweis: Falls (4.1) zustandsstabil ist, ist A Hurwitz. Also gilt nach Satz 3.5 die Ungleichung $\|e^{At}\| \leq ce^{-\sigma t}$ für Konstanten $c, \sigma > 0$ und alle $t \geq 0$. Damit folgt $\|g(t)\| \leq \|C\|ce^{-\sigma t}\|B\|$ und damit

$$\int_0^\infty \|g(t)\|dt \leq \int_0^\infty \|C\|ce^{-\sigma t}\|B\|dt = \frac{c\|C\|\|B\|}{\sigma} < \infty.$$

□

Die Umkehrung dieses Korollars gilt offensichtlich nicht; ein einfaches Gegenbeispiel erhalten wir, wenn wir $C = 0$ setzen, da das System dann wegen $y(t) \equiv 0$ für alle $u \in \mathcal{U}$ trivialerweise E/A-stabil mit $K = 0$ ist, egal ob die Matrix A stabil ist oder nicht.

Die Überprüfung des Kriteriums (5.2) ist im Allgemeinen mühsam, weil hier ein uneigentliches Integral abgeschätzt werden muss. Falls aber die Übertragungsfunktion G bekannt ist, so lässt sich dies Kriterium leicht anhand dieser Funktion überprüfen. Dabei heißt $s^* \in \mathbb{C}$ Polstelle einer rationalen (Matrix-)Funktion G, wenn s^* Polstelle für mindestens eine ihrer Komponentenfunktionen ist, was wiederum bedeutet, dass $j, k \in \mathbb{N}_0$ existieren mit $j < k$, so dass s^* k-fache Nullstelle des Nennerpolynoms und j-fache Nullstelle des Zählerpolynoms ist (wobei $j = 0$ bedeutet, dass s^* keine Nullstelle ist). Beachte, dass s^* genau dann eine Polstelle ist, wenn $\|G(s)\|$ in jeder Umgebung von s^* unbeschränkt ist.

Satz 5.13 Gegeben sei ein Kontrollsystem (4.1) mit Übertragungsfunktion G. Dann ist das System genau dann E/A-stabil, wenn alle Polstellen s^* von G in der offenen linken komplexen Halbebene $\mathbb{C}^- = \{z \in \mathbb{C} \,|\, \mathrm{Re}(z) < 0\}$ liegen, also $\mathrm{Re}(s^*) < 0$ erfüllen.

Beweis: "⇒": Wenn das System E/A-stabil ist, gilt nach Lemma 5.11 die Ungleichung $g_{\max} = \int_0^\infty \|g(t)\|dt < \infty$. Damit folgt für alle $s \in \mathbb{C}$ mit $\mathrm{Re}(s) \geq 0$ die Ungleichung

$$\|G(s)\| = \left\| \int_0^\infty g(t)e^{-st}dt \right\| \leq \int_0^\infty \|g(t)\| \underbrace{|e^{-st}|}_{\leq 1} dt \leq \int_0^\infty \|g(t)\|dt = g_{\max},$$

weswegen G keine Polstellen außerhalb von $\mathbb{C}^-$ haben kann.

"⇐": Es seien $\gamma_{ij}(t)$ die Komponenten der Funktion $g(t) = Ce^{At}B$. Aus Bemerkung 5.7 folgt, dass die Einträge von G durch $g_{ij} = \hat{\gamma}_{ij}$ gegeben sind. Aus der Form der Matrix-Exponentialfunktion folgt, dass die $\gamma_{ij}(t)$ von der Form

$$\gamma_{ij}(t) = \sum_{p=1}^{q} \mu_p e^{\lambda_p t} \frac{t^{k_p}}{k_p!}$$

sind, wobei die λ_j Eigenwerte von A sind. Aus Beispiel 5.3(g) folgt daher

$$g_{ij}(s) = \hat{\gamma}_{ij}(s) = \sum_{p=1}^{q} \mu_p \frac{1}{(s-\lambda_p)^{k_p+1}}.$$

Hieraus folgt, dass die Polstellen von G gerade durch die λ_p gegeben sind. Aus der Annahme an die Polstellen von G folgt daher, dass alle λ_p in $\mathbb{C}^-$ liegen. Daraus folgt wiederum, dass das Integral $\int_0^\infty \gamma_{ij}(t)dt$ für alle i, j endlich ist, womit auch $\int_0^\infty \|g(t)\|dt < \infty$ ist. Gemäß Lemma 5.11 ist das System damit E/A-stabil. □

Beispiel 5.14 Für das Pendel sieht man mit diesem Kriterium leicht, dass das herunterhängende Pendel E/A-stabil ist, weil die Polstellen (also die Nullstellen des Nenners) gegeben sind durch $-k/2 \pm \sqrt{k^2 - 4g}/2$ und damit stets negativen Realteil besitzen. Analog sieht man beim invertierten Pendel an den Polstellen $-k/2 \pm \sqrt{k^2 + 4g}/2$, von denen einer positiven Realteil besitzt, dass das invertierte Pendel nicht E/A-stabil ist. □

Bemerkung 5.15 (i) Der Beweis zeigt, dass alle Polstellen von G Eigenwerte von A sind. Dies erklärt den Namen Polverschiebungssatz für Satz 3.29.

(ii) Im Allgemeinen sind nicht alle Eigenwerte von A Polstellen von G. Zum einen fehlen diejenigen Eigenwerte, für die der zugehörige Eigenraum in $\mathcal{N}$ liegt, für die man also die darin liegenden Lösungen nicht beobachten kann. Zum anderen fehlen die Eigenwerte, deren Eigenräume man von $x_0 = 0$ aus nicht erreichen kann, weil sie nicht in der Erreichbarkeitsmenge $\mathcal{R}$ liegen.

Falls das System kontrollierbar und beobachtbar ist, sind alle Eigenwerte von A Pole von G, was man auch beim Vergleich von Beispiel 5.14 mit Beispiel 3.6 sieht. Falls das System stabilisierbar und entdeckbar ist, sind alle instabilen Eigenwerte (also diejenigen mit positivem Realteil) Pole von G. In diesen Fällen ist Zustandsstabilität äquivalent zur E/A-Stabilität. □

5.4 Feedbacks im Frequenzbereich

Um ein Feedback bzw. eine Rückführung im Frequenzbereich formulieren zu können, müssen wir das Konzept zuerst etwas erweitern. Dazu beobachten wir zuerst, dass wir sowohl das statische Feedback-Konzept mit $u(t) = Fx(t)$ als auch das dynamische Konzept mit der $u(t) = Fz(t)$ und der Differentialgleichung $\dot{z}(t) = (J + KF)z(t) + Ly(t)$ leicht Laplacetransformieren können. Es ergeben sich die Übertragungsfunktionen

$$K(s) = F \quad \text{bzw.} \quad K(s) = F(s\mathrm{Id} - M)^{-1}L,$$

wobei wir im ersten Fall $C = \mathrm{Id}$ annehmen und im zweiten Fall kurz $M = J + KF$ geschrieben haben. Ein geschlossener Regelkreis kann also immer als eine Verkopplung zweier Übertragungsfunktionen G und K dargestellt werden. Konsistent mit dem E/A-Konzept wäre es nun, wenn solch eine Verkopplung selbst wieder eine Übertragungsfunktion wäre. Dazu brauchen wir aber einen Eingang für unser geregeltes System, den wir bisher

nicht hatten, da der ursprüngliche Eingang ja mit $u = Fx$ bzw. $u = Fz$ "belegt" ist. Zur Abhilfe führen wir einen neuen Eingang $w(t)$ ein, indem wir $Fx(t)$ bzw. $Ly(t)$ durch $F(x(t) + w(t))$ bzw. $L(y(t) + w(t))$ ersetzen.

Satz 5.16 Gegeben seien zwei Übertragungsfunktionen G und K passender Dimension, die mittels $\hat{y}(s) = G(s)\hat{u}(s)$ und $\hat{u}(s) = K(\hat{y}(s) + \hat{w}(s))$ verkoppelt sind. Dann gilt

$$\hat{y}(s) = (\mathrm{Id} - G(s)K(s))^{-1}G(s)K(s)\hat{w}(s)$$

für alle $s \in \mathbb{C}$ für die $\mathrm{Id} - G(s)K(s)$ invertierbar ist.

Beweis: Aus den beiden angegebenen Gleichungen folgt

$$\hat{y}(s) = G(s)\hat{u}(s) = G(s)K(\hat{y}(s) + \hat{w}(s)).$$

Umstellen liefert, dass diese Gleichung äquivalent ist zu

$$(\mathrm{Id} - G(s)K(s))\hat{y}(s) = G(s)K(s)\hat{w}(s),$$

woraus die behauptete Gleichung sofort folgt. □

Das Feedback-Stabilisierungsproblem besteht im Frequenzraum nun darin, eine Übertragungsfunktion K zu finden, so dass $(\mathrm{Id}-G(s)K(s))^{-1}G(s)K(s)$ stabil ist, also nur Polstellen in $\mathbb{C}^-$ besitzt. Dafür gibt es insbesondere im Fall, dass u und y eindimensional sind, eine ganze Reihe von Techniken, die wir hier aus Zeitgründen aber nicht besprechen wollen.

Wir wollen stattdessen noch kurz darauf eingehen, was die Rolle des neuen Eingangsignals im stabilisierten System ist. Dazu betrachten wir der Einfachheit halber den Fall eines statischen stabilisierenden Feedbacks $u = Fx$ und $C = \mathrm{Id}$. Dann ergeben sich die Lösungen des geregelten Systems mit dem neuen Eingang zu

$$x(t) = e^{(A+BF)t}x_0 + \underbrace{\int_0^t e^{(A+BF)(t-\tau)}BFw(\tau)d\tau}_{=:v(t)}.$$

Exponentielle Stabilität ist nun äquivalent dazu, dass $e^{(A+BF)t}$ gegen 0 konvergiert für $t \to \infty$. Damit gilt

$$\|x(t) - v(t)\| \leq ce^{-\sigma t}\|x_0\|,$$

d.h. die Lösung konvergiert gegen $v(t)$. Stabilität stellt also sicher, dass die Lösung unabhängig vom Anfangswert gegen eine wohldefinierte Grenzfunktion konvergiert, die nur vom Eingang $w(t)$ abhängt. Dies ist eine neue Interpretation der Stabilität, die äquivalent zur E/A-Stabilität ist und daher wie diese aus der Stabilität des Systems im Sinne von Kapitel 3 und 4 folgt. Im Fall $w \equiv 0$ gilt für diese Grenzfunktion $v \equiv 0$ und wir befinden uns gerade wieder in der Situation dieser Kapitel.

5.5 Grafische Analyse

Wir betrachten in diesem Abschnitt zwei in der Regelungstechnik übliche grafische Darstellungsweisen. Diese sind auf Systeme mit eindimensionalem Eingang und Ausgang, also $m = k = 1$ anwendbar. Beachte, dass die Übertragungsfunktion G in diesem Fall eine skalare Funktion ist. Systeme dieser Art werden als SISO-Systeme (Single Input Single Output) bezeichnet.

Das Bodediagramm

Das Bodediagramm[3] dient dazu, den Zusammenhang zwischen u und y grafisch zu veranschaulichen. Insbesondere wird durch diese Interpretation klar, warum die Betrachtung der Laplace-Transformierten "Analyse im Frequenzbereich" genannt wird. Zur Vorbereitung benötigen wir zunächst den folgenden Satz.

Satz 5.17 Betrachte die Übertragungsfunktion $G : \mathbb{C} \to \mathbb{C}$ für ein E/A-stabiles SISO-System der Form (4.1). Dann konvergiert das Ausgangssignal $y(t)$ zum Eingangssignal $u(t) = \sin(\omega t)$ für $t \to \infty$ gegen die Funktion

$$y_\infty(t) = |G(i\omega)| \sin(\omega t + \varphi(\omega)),$$

wobei φ eine Argumentfunktion[4] von $\omega \mapsto G(\omega i)$ ist.

Beweis: Siehe [11, Proposition 2.3.22].

Die Werte der Übertragungsfunktion G entlang der imaginären Achse $i\mathbb{R}$ — der sogenannte *Frequenzgang* von G — haben also eine ganz konkrete Bedeutung für das Verhalten des Ausgangs $y(t)$ bei sinusförmigen Eingängen $u(t)$: Das Ausgangssignal wird gerade dadurch erzeugt, dass das Eingangssignal um $|G(i\omega)|$ verstärkt wird und die Phase um $\varphi(\omega)$ verschoben wird.

Abbildung 5.5 illustriert dies an Hand des (herunterhängenden) Pendelmodells mit $k = 0.1$ und $g = 9.81$.

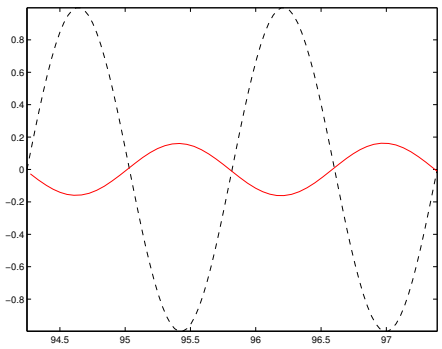


Abbildung 5.1: Eingang (schwarz gestrichelt) mit Frequenz $\omega = 4$ und zugehöriger Ausgang (rot) für das herunterhängende Pendel

Hier ist der numerisch simulierte Ausgang für den Eingang $u(t) = \sin(\omega t)$ für $\omega = 4$ zu sehen. Man erkennt, dass das Ausgangssignal eine Amplitude von etwa 0.16 besitzt und die Phase um ca. π gegenüber dem Eingangssignal verschoben ist; das Pendel pendelt

[3] Hendrik Wade Bode (1905–1982), US-amerikanischer Elektrotechniker

[4] Sei I ein Intervall. Eine stetige Funktion $\varphi : I \to \mathbb{R}$ heißt *Argumentfunktion* einer Funktion $\gamma : I \to \mathbb{C} \setminus \{0\}$, wenn $\gamma(t) = |\gamma(t)|e^{i\varphi(t)}$ gilt für alle $t \in I$. Wir schreiben dann kurz $\varphi = \arg \gamma$.

also gegenläufig zur periodischen Wagenbewegung und mit kleineren Ausschlägen. Für die zugehörige Übertragungsfunktion gilt $|G(i4)| = 0.1612$ und $\arg(G(i4)) = -3.077$, was diese Beobachtung genau bestätigt.

Diese direkte Beziehung zwischen Übertragungsfunktion und Ausgangssignal bedeutet umgekehrt, dass durch das Messen der Amplitude und der Phase des Ausgangs bei sinusförmigem Eingang die Werte $G(i\omega) = |G(i\omega)|e^{\varphi(\omega)}$ leicht errechnet werden können. Die Übertragungsfunktion kann auf der imaginären Achse also durch experimentelle Messungen bestimmt werden.

Diese Tatsache gewinnt durch einen Satz aus der Funktionentheorie besondere Bedeutung: Man kann nämlich beweisen, dass die Funktion $G(i\omega)$ durch ihre Werte auf $i\mathbb{R}$ eindeutig bestimmt ist. Genauer folgt aus der Integralformel von Cauchy für E/A-stabile Systeme (4.1) die Darstellung

$$G(s) = \frac{1}{2\pi i}\int_{-\infty}^{\infty} \frac{G(i\omega)}{i\omega - s} d\omega$$

für alle $s \in \mathbb{C}$ mit $\mathrm{Re}(s) > 0$ (beachte, dass hier wichtig ist, dass kein "$Du(t)$" in der Formel für $y(t)$ in (4.1) auftaucht; ansonsten muss die Formel modifiziert werden). Da zudem $G(i\omega) \to 0$ gilt für $\omega \to \pm\infty$, kann das obige Integral durch ein Integral mit kompaktem Integrationsintervall approximiert werden. Folglich kann die komplette Übertragungsfunktion eines E/A-stabilen Systems aus Messdaten für sinusförmige Eingangssignale rekonstruiert werden, vgl. [14, Abschnitt 6.5.3].

Grafisch werden diese Messdaten nun in dem sogenannten Bodediagramm dargestellt, wobei für die Frequenz und für den Betrag $|G(i\omega)|$ logarithmische Skalen verwendet wird. In Abbildung 5.2 ist dieses Diagramm für das herunterhängende Pendel, wiederum mit $k = 0.1$ und $g = 9.81$ dargestellt.

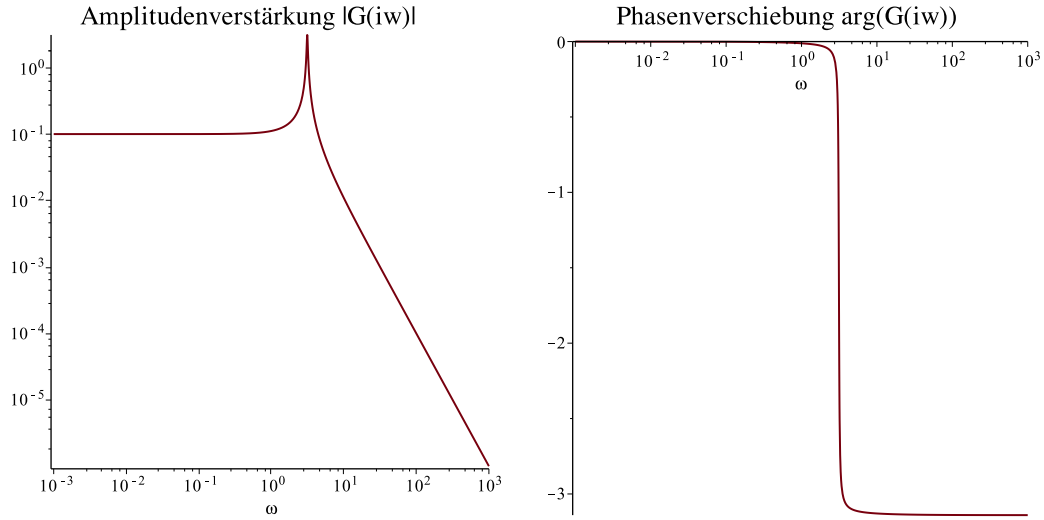


Abbildung 5.2: Bodediagramm für das herunterhängende Pendel

Das linke Diagramm besagt, dass das Eingangssignal zunächst schwach, mit steigender Frequenz bis zu etwa $\omega = 3$ dann aber immer stärker verstärkt wird, während die Verstärkung

für größere ω dann wieder abnimmt. Die Phase bleibt dabei für kleine ω fast unverändert, um dann ab etwa $\omega = 3$ abrupt um ca. $-\pi$ verschoben zu werden. Genau dies Verhalten zeigt sich in den numerischen Simulationen in Abbildung 5.3.

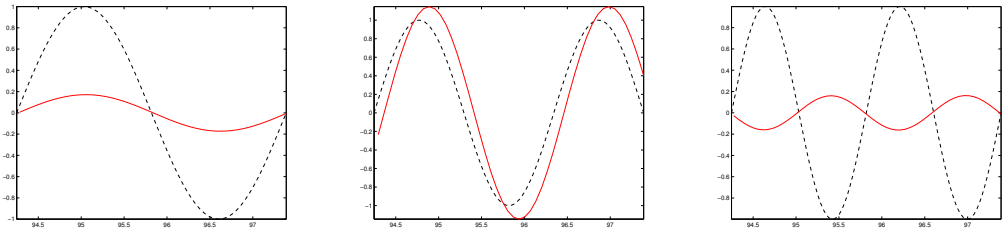


Abbildung 5.3: Eingang (schwarz gestrichelt) und Ausgang (rot) für das herunterhängende Pendel mit $\omega = 2, 3, 4$ von links nach rechts

Das Nyquistdiagramm

Das Nyquistdiagramm[5] dient dazu, um zu prüfen, ob ein Feedbacksystem E/A-stabil ist. Wie beim Bodediagramm kann die Grafik dabei allein aus Messwerten erstellt werden und die Stabilität damit experimentell verifiziert werden.

Die Übertragungsfunktion eines Feedbacksystems ist nach Satz 5.16 im SISO-Fall gegeben durch

$$G_{cl} := \frac{G(s)K(s)}{1 - G(s)K(s)}.$$

Diese ist nach Satz 5.13 genau dann E/A-stabil, wenn keine Polstellen in der abgeschlossenen rechten Halbebene liegen. Hinreichend dafür ist, dass $F(s) := 1 - G(s)K(s)$ keine Nullstellen in der abgeschlossenen rechten Halbebene besitzt, was genau dann der Fall ist, wenn $G_0(s) := -G(s)K(s)$ in der rechten Halbebene nie den Wert -1 annimmt.

Das Nyquistdiagramm[6] stellt nun die Werte von $G_0(\omega i)$ für $\omega \in (-\infty, \infty)$, grafisch dar. Praktisch wird dies dadurch näherungsweise realisiert, dass Werte von $-R$ bis R für ein großes $R \in \mathbb{R}$ an Stelle von $\pm\infty$ verwendet werden. Da $G(s)K(s)$ die Übertragungsfunktion der Hintereinanderschaltung von Feedback und System ist, können diese Werte dieses Produkts wiederum experimentell ermittelt werden.

In Abbildung 5.4 sind diese Kurven für das invertierte Pendel mit $G(s) = 1/(ks + s^2 - g)$ mit $k = 0.1$ und $g = 9.81$ und das statische Feedback $K = -1$ (links) und $K = -10$ (rechts) dargestellt.

Betrachtung der Zähler- und Nennerpolynome in G_0 liefert nun das folgende Stabilitätskriterium.

Nyquistkriterium: Es sei $n^+ \in \mathbb{N}$ die Anzahl der Polstellen von G_0 mit positivem Realteil, zudem habe G_0 keine Polstellen mit Realteil gleich 0. Dann ist das Feedbacksystem mit

[5] Harry Nyquist (1889–1976), US-Amerikanischer Elektrotechniker

[6] Wir stellen hier nur die Version für $D = 0$ vor, siehe z.B. [14, Abschnitt 8.5] für den allgemeinen Fall.

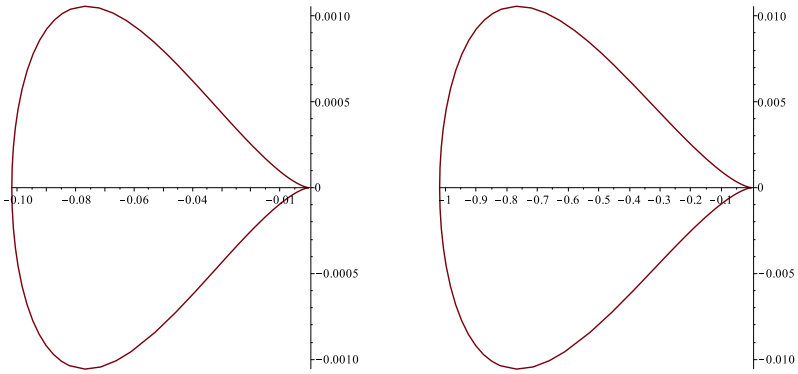


Abbildung 5.4: Nyquistdiagramm für das invertierte Pendel mit $K = -1$ (links) und $K = -10$ (rechts)

Übertragungsfunktion G_{cl} genau dann E/A-stabil, wenn die Ortskurve $G(\omega i)$ für $\omega = -\infty \ldots, \infty$ den Punkt $-1 = -1 + 0i \in \mathbb{C}$ genau n^+-mal entgegen dem Uhrzeigersinn umläuft. Im Fall $n^+ = 0$ gilt Stabilität genau dann, wenn die Ortskurve den Punkt -1 keinmal im Uhrzeigersinn umläuft.

In unserem Beispiel aus Abbildung 5.4 hat G_0 wegen $K = const$ gerade die gleichen Polstellen wie G; also existiert eine Polstelle mit positivem Realteil und keine mit Realteil 0. Folglich muss die Ortskurve einmal entgegen dem Uhrzeigersinn um den Punkt $-1 + 0i$ laufen. Dies ist in der linken Kurve für $K = -1$ offenbar nicht der Fall. Es trifft aber in der rechten Kurve für $K = -10$ zu (die Umlaufrichtung ist in dieser Grafik natürlich nicht zu sehen, verläuft aber tatsächlich entgegen dem Uhrzeigersinn). Eine Analyse im Zeitbereich zeigt, dass die zugehörige closed-loop Matrix für $K = -1$ bzw. $K = -10$ gegeben ist durch

$$A = \begin{pmatrix} 0 & 1 \\ g - K & -k \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 8.81 & -0.1 \end{pmatrix} \quad \text{bzw.} \quad A = \begin{pmatrix} 0 & 1 \\ -0.19 & -0.1 \end{pmatrix}.$$

Eine Analyse der Eigenwerte dieser Matrix bestätigt die Instabilität für $K = -1$ und die Stabilität für $K = -10$. Tatsächlich liegt die Grenze zwischen Instabilität und Stabilität gerade bei $K = -9.81$.

Bemerkung 5.18 Auch für zeitdiskrete Systeme ist eine Betrachtung im Frequenzbereich möglich. Statt der Laplace-Transformation verwendet man dort die sogenannte z-Transformation, auf die wir hier aus Zeitgründen nicht weiter eingehen wollen. □

Kapitel 6

Optimale Stabilisierung

Die in Kapitel 3 vorgestellte Methode zur Berechnung stabilisierender Feedbacks hat den Nachteil, dass man zwar die Eigenwerte bestimmen kann, ansonsten aber relativ wenig Einflussmöglichkeiten auf die Dynamik des geregelten Systems hat. So ist es z.B. oft so, dass große Werte der Kontrollvariablen u nur mit großem Energieaufwand zu realisieren sind (wie im Pendelmodell, wo u gerade die Beschleunigung des Wagens ist), weswegen man große Werte vermeiden möchte. Im Heizungsmodell andererseits möchte man z.B. Überschwingen (d.h. starke Schwankungen bis zum Erreichen der gewünschten Temperatur) vermeiden.

Wir werden deshalb in diesem Kapitel einen Ansatz verfolgen, der größeren Einfluss auf das Verhalten des geregelten Systems ermöglicht, indem wir Methoden der Optimierung zur Berechnung der Feedback-Matrix F verwenden. Dabei können die gewünschten Eigenschaften durch die verwendete Kostenfunktion bestimmt werden. Wir nehmen dabei aus Vereinfachungsgründen wieder an, dass wie in Kapitel 3 der gesamte Zustandsvektor x für die Regelung zur Verfügung steht. Falls das nicht der Fall ist, kann ein dyamischer Beobachter gemäß Kapitel 4 verwendet werden. Wir beschränken uns hier suf Optimierungsprobleme, die direkt mit dem Stabilisierungsproblem in Zusammenhang stehen. Allgemeinere Probleme werden wir später in der Vorlesung im Rahmen der Modellprädiktiven Regelung betrachten.

6.1 Grundlagen der optimalen Steuerung

In diesem Abschnitt werden wir einige Grundlagen der optimalen Steuerung herleiten, die zur Lösung unseres Problems nötig sind. Da es für die abstrakten Resultate keinen Unterschied macht, ob die Dynamik linear oder nichtlinear ist, betrachten wir hier allgemeine Kontrollsysteme der Form

$$\dot{x}(t) = f(x(t), u(t)), \tag{6.1}$$

unter der Annahme, dass $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ stetig ist und dass für alle $R > 0$ ein $L_R > 0$ existiert, so dass die Lipschitz-Bedingung

$$\|f(x_1, u) - f(x_2, u)\| \leq L_R \|x_1 - x_2\| \tag{6.2}$$

für alle $x_1, x_2 \in \mathbb{R}^n$ und alle $u \in \mathbb{R}^m$ mit $\|x_1\|, \|x_2\|, \|u\| \leq R$ erfüllt ist (vgl. Satz 8.1). Unter dieser Bedingung kann man den aus der Theorie der gewöhnlichen Differentialgleichungen bekannten Existenz- und Eindeutigkeitssatz so modifizieren, dass er für jede stückweise stetige Kontrollfunktion $u \in \mathcal{U}$ und jeden Anfangswert x_0 die Existenz einer eindeutige Lösung $x(t, x_0, u)$ mit $x(0, x_0, u) = x_0$ liefert.

Wir definieren nun das optimale Steuerungsproblem, mit dem wir uns im Folgenden beschäftigen wollen.

Definition 6.1 Für eine stetige nichtnegative *Kostenfunktion* $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_0^+$ definieren wir das *Kostenfunktional*

$$J(x_0, u) := \int_0^\infty g(x(t, x_0, u), u(t))dt.$$

Das optimale Steuerungsproblem ist damit gegeben durch das Optimierungsproblem

$$\text{Minimiere } J(x_0, u) \text{ über } u \in \mathcal{U} \text{ für jedes } x_0 \in \mathbb{R}^n.$$

Die Funktion

$$V(x_0) := \inf_{u \in \mathcal{U}} J(x_0, u)$$

wird als *optimale Wertefunktion* dieses optimalen Steuerungsproblems bezeichnet. Ein Paar $(x^*, u^*) \in \mathbb{R}^n \times \mathcal{U}$ mit $J(x^*, u^*) = V(x^*)$ wird als *optimales Paar* bezeichnet. □

Als Funktionenraum $\mathcal{U}$ wählen wir hierbei wie bisher den Raum der stückweise stetigen Funktionen, und nehmen dabei zusätzlich an, dass jede Funktion u auf jedem kompakten Intervall beschränkt ist und dass die Funktionen u rechtsseitig stetig sind, d.h, dass für alle $t_0 \in \mathbb{R}$ die Bedingung $\lim_{t \searrow t_0} u(t) = u(t_0)$ gilt. Beachte dass wir die zweite Annahme o.B.d.A. machen können, da die Lösung nicht vom dem Wert von u in der Sprungstelle abhängt.

Bemerkung 6.2 Im Zeitdiskreten mit der Dynamik

$$x(k+1) = f(x(k), u(k))$$

und Anfangswert $x(0) = x_0$ lautet das Kostenfunktional

$$J(x_0, u) := \sum_{k=0}^{\infty} g(x(k, x_0, u), u(k)).$$

□

Beachte, dass das Funktional $J(x_0, u)$ nicht endlich sein muss. Ebenso muss das Infimum in der Definition von V kein Minimum sein.

Der erste Satz dieses Kapitels liefert eine Charakterisierung der Funktion V.

Satz 6.3 (Prinzip der dynamischen Programmierung oder **Bellman'sches Optimalitätsprinzip)**
(i) Für die optimale Wertefunktion gilt für jedes $\tau > 0$

$$V(x_0) = \inf_{u \in \mathcal{U}} \left\{ \int_0^\tau g(x(t, x_0, u), u(t))dt + V(x(\tau, x_0, u)) \right\}.$$

(ii) Für ein optimales Paar (x^*, u^*) gilt für jedes $\tau > 0$

$$V(x^*) = \int_0^\tau g(x(t, x^*, u), u^*(t))dt + V(x(\tau, x^*, u^*)).$$

Beweis: (i) Wir zeigen zunächst

$$V(x_0) \leq \int_0^\tau g(x(t, x_0, u), u(t))dt + V(x(\tau, x_0, u))$$

für alle $u \in \mathcal{U}$ und alle $\tau > 0$. Sei dazu $x_\tau = x(\tau, x_0, u)$, $\varepsilon > 0$ beliebig und $u_\tau \in \mathcal{U}$ so gewählt, dass

$$J(x_\tau, u_\tau) \leq V(x_\tau) + \varepsilon$$

gilt. Sei $\tilde{u} = u \&_\tau u_\tau(\cdot - \tau)$ (vgl. Definition 1.7). Dann gilt

$$\begin{aligned}
V(x_0) &\leq \int_0^\infty g(x(t, x_0, \tilde{u}), \tilde{u}(t))dt \\
&= \int_0^\tau g(x(t, x_0, \tilde{u}), \tilde{u}(t))dt + \int_\tau^\infty g(x(t, x_0, \tilde{u}), \tilde{u}(t))dt \\
&= \int_0^\tau g(x(t, x_0, u), u(t))dt + \int_\tau^\infty g(\underbrace{x(t, x_0, \tilde{u})}_{=x(t-\tau, x_\tau, u_\tau)}, u_\tau(t-\tau))dt \\
&= \int_0^\tau g(x(t, x_0, u), u(t))dt + \int_0^\infty g(x(t, x_\tau, u_\tau), u_\tau(t))dt \\
&= \int_0^\tau g(x(t, x_0, u), u(t))dt + J(x_\tau, u_\tau) \leq \int_0^\tau g(x(t, x_0, u), u(t))dt + V(x_\tau) + \varepsilon.
\end{aligned}$$

Da $\varepsilon > 0$ beliebig war, folgt die behauptete Ungleichung.

Als zweiten Schritt zeigen wir

$$V(x_0) \geq \inf_{u \in \mathcal{U}} \left\{ \int_0^\tau g(x(t, x_0, u), u(t))dt + V(x(\tau, x_0, u)) \right\}.$$

Sei dazu wiederum $\varepsilon > 0$ beliebig. Wir wählen u_0 so, dass $V(x_0) \geq J(x_0, u_0) - \varepsilon$ gilt und

schreiben $x_\tau = x(\tau, x_0, u_0)$. Damit folgt

$$\begin{aligned}
V(x_0) &\geq \int_0^\infty g(x(t,x_0,u_0),u_0(t))dt - \varepsilon \\
&= \int_0^\tau g(x(t,x_0,u_0),u_0(t))dt + \int_\tau^\infty g(x(t,x_0,u_0),u_0(t))dt - \varepsilon \\
&= \int_0^\tau g(x(t,x_0,u_0),u_0(t))dt + \int_0^\infty g(x(t,x(\tau,x_0,u_0),u_0(\cdot+\tau)),u_0(t+\tau))dt - \varepsilon \\
&= \int_0^\tau g(x(t,x_0,u_0),u_0(t))dt + J(x(\tau,x_0,u_0),u_0(\cdot+\tau)) - \varepsilon \\
&\geq \int_0^\tau g(x(t,x_0,u_0),u_0(t))dt + V(x(\tau,x_0,u_0)) - \varepsilon \\
&\geq \inf_{u\in\mathcal{U}} \left\{ \int_0^\tau g(x(t,x_0,u),u(t))dt + V(x(\tau,x_0,u)) \right\} - \varepsilon
\end{aligned}$$

woraus die Behauptung folgt, da $\varepsilon > 0$ beliebig war.

(ii) Aus (i) folgt sofort die Ungleichung

$$V(x^*) \leq \int_0^\tau g(x(t,x^*,u^*),u^*(t))dt + V(x(\tau,x^*,u^*)).$$

Die umgekehrte Ungleichung folgt aus

$$\begin{aligned}
V(x^*) &= \int_0^\infty g(x(t,x^*,u^*),u^*(t))dt \\
&= \int_0^\tau g(x(t,x^*,u^*),u^*(t))dt + \int_\tau^\infty g(x(t,x^*,u^*),u^*(t))dt \\
&= \int_0^\tau g(x(t,x^*,u^*),u^*(t))dt + \int_0^\infty g(x(t,x(\tau,x^*,u^*),u^*(\cdot+\tau)),u^*(t+\tau))dt \\
&= \int_0^\tau g(x(t,x^*,u^*),u^*(t))dt + J(x(\tau,x^*,u^*),u^*(\cdot+\tau)) \\
&\geq \int_0^\tau g(x(t,x^*,u^*),u^*(t))dt + V(x(\tau,x^*,u^*)).
\end{aligned}$$

□

Eine Folgerung dieses Prinzips liefert das folgende Korollar.

Korollar 6.4 Sei (x^*, u^*) ein optimales Paar. Dann ist $(x(\tau, x^*, u^*), u^*(\cdot + \tau))$ für jedes $\tau > 0$ ein optimales Paar.

Beweis: Übungsaufgabe.

Anschaulich besagt Korollar 6.4, dass Endstücke optimaler Trajektorien selbst wieder optimale Trajektorien sind.

Die bisherigen Aussagen gelten analog (und mit analogen Beweisen) auch im Zeitdiskreten. Dort gilt für alle $K \in \mathbb{N}$

$$V(x_0) = \inf_{u\in\mathcal{U}} \left\{ \sum_{k=0}^{K-1} g(x(k,x_0,u),u(k)) + V(x(K,x_0,u)) \right\} \tag{6.3}$$

sowie für alle optimalen Paare (x^*, u^*)

$$V(x^*) = \sum_{k=0}^{K-1} g(x(k, x^*, u), u^*(k)) + V(x(K, x^*, u^*)).$$

Für die folgende Aussage, mit der wir durch einen geschickten Grenzübergang für $\tau \to 0$ die Gleichung aus Satz 6.3 als (partielle) Differentialgleichung ausdrücken können, gibt es kein zeitdiskretes Gegenstück.

Satz 6.5 (Hamilton-Jacobi-Bellman Differentialgleichung)
Es sei g stetig in x und u. Zudem sei $O \subseteq \mathbb{R}^n$ offen und $V|_O$ endlich.

(i) Wenn V in $x_0 \in O$ stetig differenzierbar ist, so folgt

$$DV(x_0) \cdot f(x_0, u_0) + g(x_0, u_0) \geq 0$$

für alle $u_0 \in \mathbb{R}^m$.

(ii) Wenn (x^*, u^*) ein optimales Paar ist und V stetig differenzierbar in $x^* \in O$ ist, so folgt

$$\min_{u \in \mathbb{R}^m} \{DV(x^*) \cdot f(x^*, u) + g(x^*, u)\} = 0, \tag{6.4}$$

wobei das Minimum in $u^*(0)$ angenommen wird. Gleichung (6.4) wird *Hamilton-Jacobi-Bellman Gleichung* genannt.

Beweis: Wir zeigen zunächst für alle $u \in \mathcal{U}$ die Hilfsbehauptung

$$\lim_{\tau \searrow 0} \frac{1}{\tau} \int_0^\tau g(x(t, x_0, u), u(t))dt = g(x_0, u(0)).$$

Wegen der (rechtssitigen) Stetigkeit von x und u in t und der Stetigkeit von g existiert zu $\varepsilon > 0$ ein $t_1 > 0$ mit

$$|g(x(t, x_0, u), u(t)) - g(x_0, u(0))| \leq \varepsilon$$

für alle $t \in [0, t_1)$. Damit folgt für $\tau \in (0, t_1]$

$$\begin{aligned} \left| \frac{1}{\tau} \int_0^\tau g(x(t, x_0, u), u(t))dt - g(x_0, u(0)) \right| &\leq \frac{1}{\tau} \int_0^\tau |g(x(t, x_0, u), u(t)) - g(x_0, u(0))|dt \\ &\leq \frac{1}{\tau} \int_0^\tau \varepsilon \;=\; \varepsilon \end{aligned}$$

und damit die Aussage für den Limes, da $\varepsilon > 0$ beliebig war.

Hiermit folgen nun beide Behauptungen:

(i) Aus Satz 6.3(i) folgt für $u(t) \equiv u_0 \in \mathbb{R}^m$

$$V(x_0) \leq \int_0^\tau g(x(t, x_0, u), u(t))dt + V(x(\tau, x_0, u))$$

und damit

$$\begin{aligned} DV(x_0)f(x_0,u(0)) &= \lim_{\tau\searrow 0}\frac{V(x(\tau,x_0,u))-V(x_0)}{\tau} \\ &\geq \lim_{\tau\searrow 0} -\frac{1}{\tau}\int_0^\tau g(x(t,x_0,u),u(t))dt \;=\; -g(x_0,u(0)), \end{aligned}$$

also die Behauptung.

(ii) Aus (i) folgt

$$\inf_{u\in\mathbb{R}^m}\{DV(x^*)\cdot f(x^*,u)+g(x^*,u)\}\geq 0.$$

Aus Satz 6.3(ii) folgt zudem

$$V(x^*)=\int_0^\tau g(x(t,x^*,u^*),u^*(t))dt+V(x(\tau,x^*,u^*)).$$

Damit gilt

$$\begin{aligned} DV(x^*)f(x^*,u^*(0)) &= \lim_{\tau\searrow 0}\frac{V(x(\tau,x^*,u^*))-V(x^*)}{\tau} \\ &= \lim_{\tau\searrow 0} -\frac{1}{\tau}\int_0^\tau g(x(t,x^*,u^*),u^*(t))dt \;=\; -g(x^*,u^*(0)), \end{aligned}$$

woraus die Existenz des Minimums in $u=u^*(0)$ und die behauptete Gleichheit folgt. □

Satz 6.5 gibt *notwendige* Optimalitätsbedingungen, d.h. Bedingungen die die optimale Wertefunktion bzw. ein optimales Paar erfüllen *muss* — vorausgesetzt die optimale Wertefunktion ist stetig differenzierbar. Im Allgemeinen folgt aus der Erfüllung der angegebenen notwendigen Bedingungen aber noch nicht, dass eine Funktion tatsächlich eine optimale Wertefunktion ist oder ein Paar ein optimales Paar. Hierzu braucht man *hinreichende* Optimalitätsbedingungen, die wir im Folgenden untersuchen.

Zur Herleitung der hinreichenden Bedingungen brauchen wir zusätzliche Annahmen, für deren genaue Ausgestaltung es verschiedene Möglichkeiten gibt. Da wir die optimale Steuerung auf das Stabilisierungsproblem anwenden wollen, verwenden wir dazu die folgende Definition.

Definition 6.6 Für das Kontrollsystem gelte $f(0,0)=0$, d.h. der Nullpunkt ist ein Gleichgewicht für $u=0$. Dann nennen wir das optimale Steuerungsproblem *nullkontrollierend*, falls die Implikation

$$J(x_0,u)<\infty \quad\Rightarrow\quad x(t,x_0,u)\to 0 \text{ für } t\to\infty$$

gilt. □

Nun können wir die hinreichende Bedingung formulieren.

Satz 6.7 (Hinreichende Optimalitätsbedingung)
Betrachte ein nullkontrollierendes optimales Steuerungsproblem. Es sei $W:\mathbb{R}^n\to\mathbb{R}_0^+$ eine stetig differenzierbare Funktion, die die Hamilton-Jacobi-Bellman Gleichung

$$\min_{u\in\mathbb{R}^m}\{DW(x)f(x,u)+g(x,u)\}=0$$

erfüllt und für die $W(0) = 0$ gilt.

Zu gegebenem $x^* \in \mathbb{R}^n$ sei $u^* \in \mathcal{U}$ eine Kontrollfunktion, so dass für die zugehörige Lösung $x(t, x^*, u^*)$ und alle $t \geq 0$ das Minimum in der obigen Gleichung für $x = x(t, x^*, u^*)$ in $u = u^*(t)$ angenommen wird.

Dann ist (x^*, u^*) ein optimales Paar und es gilt

$$V(x(t, x^*, u^*)) = W(x(t, x^*, u^*))$$

für alle $t \geq 0$.

Beweis: Wir zeigen die Aussage für $t = 0$. Für $t > 0$ folgt sie durch Anwendung des Beweises auf $(x(t, x^*, u^*), u^*(t + \cdot))$. Es sei $u \in \mathcal{U}$ und $x(t) = x(t, x^*, u)$ die zugehörige Lösungsfunktion. Wir zeigen zunächst die Ungleichung

$$J(x^*, u) \geq W(x^*).$$

Im Falle $J(x^*, u) = \infty$ ist nichts zu zeigen, es reicht also den Fall $J(x^*, u) < \infty$ zu betrachten. Aus der Hamilton-Jacobi-Bellman Gleichung folgt

$$\frac{d}{dt} W(x(t)) = DW(x(t)) f(x(t), u(t)) \geq -g(x(t), u(t)),$$

und damit mit dem Hauptsatz der Differential- und Integralrechnung

$$W(x(T)) - W(x^*) = \int_0^T \frac{d}{dt} W(x(t)) dt \geq -\int_0^T g(x(t), u(t)) dt.$$

Daraus folgt

$$J(x^*, u) = \lim_{T \to \infty} \int_0^T g(x(t), u(t)) dt \geq \lim_{T \to \infty} \big(W(x^*) - W(x(T)) \big) = W(x^*).$$

für alle $T > 0$. Die letzte Gleichung folgt dabei, weil das Problem nullkontrollierend ist und $J(x^*, u) < \infty$ gilt, weswegen $x(T) \to 0$ für $T \to \infty$ und damit wegen der Stetigkeit von W und $W(0) = 0$ auch $W(x(T)) \to 0$ gilt.

Beachte, dass aus dieser Ungleichung insbesondere $V(x^*) = \inf_{u \in \mathbb{U}} J(x^*, u) \geq W(x^*)$ folgt. Zum Abschluss des Beweises reicht es daher,

$$J(x^*, u^*) = W(x^*)$$

zu zeigen. Für die Kontrolle u^* und die zugehörige Lösung $x^* = x(t, x^*, u^*)$ folgt aus der Hamilton-Jacobi-Bellman Gleichung

$$\frac{d}{dt} W(x^*(t)) = DW(x^*(t)) f(x^*(t), u^*(t)) = -g(x^*(t), u^*(t)),$$

und analog zu oben

$$J(x^*, u^*) = \lim_{T \to \infty} \int_0^T g(x^*(t), u^*(t)) dt = \lim_{T \to \infty} \big(W(x^*) - W(x(T)) \big) = W(x^*).$$

□

Beachte, dass beide Sätze dieses Abschnitts nur anwendbar sind, wenn V bzw. W differenzierbar sind. Diese Annahme ist im allgemeinen nichtlinearen Fall sehr einschränkend[1]. Zudem ist es im Allgemeinen sehr schwierig, die Funktion V mittels dieser Gleichung zu bestimmen, selbst wenn sie differenzierbar ist.

Im linearen Fall hingegen vereinfacht sich das Problem und die Hamilton-Jacobi-Bellman Gleichung so weit, dass eine explizite Lösung möglich ist, wie wir im folgenden Abschnitt sehen werden.

6.2 Das linear-quadratische Problem

Wir kommen nun zurück zu unserem linearen Kontrollsystem (1.3)

$$\dot{x}(t) = Ax(t) + Bu(t) =: f(x(t), u(t)).$$

Um eine schöne Lösungstheorie zu erhalten, müssen wir auch für die Kostenfunktion $g(x, u)$ eine geeignete Struktur annehmen.

Definition 6.8 Eine quadratische Kostenfunktion $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_0^+$ ist gegeben durch

$$g(x, u) = (x^T\, u^T) \begin{pmatrix} Q & N \\ N^T & R \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix}$$

mit $Q \in \mathbb{R}^{n\times n}$, $N \in \mathbb{R}^{n\times m}$ und $R \in \mathbb{R}^{m\times m}$, so dass $G := \begin{pmatrix} Q & N \\ N^T & R \end{pmatrix}$ symmetrisch und positiv definit ist. □

Hieraus ergibt sich der Name "linear-quadratisches" optimales Steuerungsproblem: die Dynamik ist linear und die Kostenfunktion ist quadratisch.

Wir zeigen zunächst, dass dieses Problem nullkontrollierend ist.

Lemma 6.9 Das linear-quadratische Problem ist nullkontrollierend im Sinne von Definition 6.6.

Beweis: Wir zeigen zunächst die Ungleichungen

$$g(x, u) \geq c_1\|x\|^2 \;\text{ und }\; g(x, u) \geq c_2\|f(x, u)\|^2 \tag{6.5}$$

für geeignete Konstanten $c_1, c_2 > 0$.

Da die Matrix G positiv definit ist, folgt aus Lemma 3.10 die Ungleichung

$$g(x, u) \geq c_1 \left\| \begin{pmatrix} x \\ u \end{pmatrix} \right\|^2 \geq c_1\|x\|^2, \tag{6.6}$$

[1] Die nichtlineare Theorie dieser Gleichungen verwendet den verallgemeinerten Lösungsbegriff der "Viskositätslösungen", der auch für nichtdifferenzierbare Funktionen V sinnvoll ist.

also die erste Abschätzung in (6.5). Wegen

$$\|f(x,u)\|^2 = (x^T, u^T) \begin{pmatrix} A & A^TB \\ B^TA & B \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix}$$

folgt ebenfalls aus Lemma 3.10

$$\|f(x,u)\|^2 \le c_3 \left\| \begin{pmatrix} x \\ u \end{pmatrix} \right\|^2,$$

woraus wir mit (6.6) und $c_2 = c_1/c_3$ die zweite Abschätzung in (6.5) erhalten.

Es sei nun $u \in \mathcal{U}$ und $x(t) = x(t, x_0, u)$ die zugehörige Lösungsfunktion. Es gelte

$$J(x_0, u) < \infty.$$

Zu zeigen ist also, dass

$$\lim_{t\to\infty} x(t) = 0$$

gilt. Dazu nehmen wir an, dass $x(t) \not\to 0$. Es existiert also ein $\varepsilon > 0$ und eine Folge $t_k \to \infty$, so dass $\|x(t_k)\| \ge \varepsilon$ gilt. O.B.d.A. gelte $t_{k+1} - t_k \ge \varepsilon/2$. Nun wählen wir $\delta = \varepsilon/4$ und unterscheiden für jedes $k \in \mathbb{N}$ zwei Fälle:

1. Fall: $\|x(t)\| \ge \varepsilon/2$ für alle $t \in [t_k, t_k + \delta]$. In diesem Fall erhalten wir aus (6.5) für diese t die Ungleichung $g(x(t), u(t)) \ge c_1\varepsilon^2/4$ und es folgt

$$\int_{t_k}^{t_k+\delta} g(x(t), u(t))dt \ge c_1\delta\varepsilon^2/4 = c_1\varepsilon^3/16.$$

2. Fall: $\|x(t)\| < \varepsilon/2$ für ein $t \in [t_k, t_k + \delta]$. In diesem Fall folgt

$$\left\| \int_{t_k}^{t} f(x(\tau), u(\tau))d\tau \right\| = \|x(t_k) - x(t)\| \ge \|x(t_k)\| - \|x(t)\| \ge \varepsilon/2.$$

Aus der zweiten Abschätzung in (6.5) erhalten wir

$$g(x,u) \ge c_2\|f(x,u)\|^2 \ge \left\{ \begin{array}{ll} 0, & \|f(x,u)\| \le 1 \\ c_2\|f(x,u)\|, & \|f(x,u)\| > 1 \end{array} \right\} \ge c_2(\|f(x,u)\| - 1)$$

und damit

$$\int_{t_k}^{t_k+\delta} g(x(\tau), u(\tau))d\tau \ge c_2 \int_{t_k}^{t_k+\delta} \|f(x(\tau), u(\tau))\| - 1 d\tau \ge c_2(\varepsilon/2 - \delta) \ge c_2\varepsilon/4.$$

Mit $\gamma = \min\{c_1\varepsilon^3/16,\, c_2\varepsilon/4\} > 0$ ergibt sich

$$J(x_0,u) = \int_0^\infty g(x(t), u(t))dt \ge \sum_{k=1}^{\infty} \int_{t_k}^{t_k+\delta} g(x(t), u(t))dt \ge \sum_{k=1}^{\infty} \gamma = \infty,$$

ein Widerspruch. □

Wir können also Satz 6.7 verwenden, um die Optimalität einer Lösung des linear-quadratischen Problems nachzuweisen.

Um eine Kandidatin für die optimale Wertefunktion zu finden, machen wir den Ansatz

$$W(x) = x^T P x \tag{6.7}$$

für eine symmetrische und positiv definite Matrix $P \in \mathbb{R}^{n\times n}$.

A priori wissen wir nicht, ob dieser Ansatz gerechtfertigt ist – wir nehmen dies zunächst einfach an und untersuchen die Folgerungen dieser Annahme.

Lemma 6.10 Falls das linear-quadratische optimale Steuerungsproblem eine optimale Wertefunktion der Form (6.7) besitzt, so sind alle optimalen Paare (x^*, u^*) von der Form

$$u^*(t) = Fx(t, x^*, F)$$

mit $F \in \mathbb{R}^{m\times n}$ gegeben durch

$$F = -R^{-1}(B^T P + N^T),$$

wobei $x(t, x^*, F)$ die Lösung des mittels F geregelten Systems

$$\dot{x}(t) = (A + BF)x(t) = Ax(t) + Bu^*(t)$$

mit Anfangsbedingung $x(0, x^*, F) = x^*$ bezeichnet.

Darüberhinaus ist das mittels F geregelte System exponentiell stabil.

Beweis: Die optimale Wertefunktion der Form (6.7) ist stetig differenzierbar und erfüllt $W(0) = 0$, weswegen sowohl Satz 6.5 als auch Satz 6.7 anwendbar ist.

Wenn W die optimale Wertefunktion ist, so folgt aus Satz 6.5(ii), dass die optimale Kontrolle $u = u^*(t)$ für $x = x(t, x^*, u^*)$ den Ausdruck

$$DW(x) \cdot f(x, u) + g(x, u) \tag{6.8}$$

minimiert. Umgekehrt folgt aus Satz 6.7, dass jede Kontrollfunktion, die (6.8) entlang der zugehörigen Trajektorie minimiert, ein optimales Paar erzeugt. Wir müssen also zeigen, dass das angegebene Feedback gerade solche Lösungen und Kontrollfunktionen erzeugt und dass das angegebene u^* die einzige Kontrollfunktion ist, die (6.8) minimiert.

Der zu minimierende Ausdruck ist unter den gemachten Annahmen gerade gleich

$$\begin{aligned} & DW(x) \cdot f(x, u) + g(x, u) \\ = \;& x^T P(Ax + Bu) + (Ax + Bu)^T Px + x^T Qx + x^T Nu + u^T N^T x + u^T Ru \\ = \;& 2x^T P(Ax + Bu) + x^T Qx + 2x^T Nu + u^T Ru \;\; =: \;\; h(u), \end{aligned}$$

da P symmetrisch ist. Da R wegen der positiven Definitheit von G ebenfalls positiv definit sein muss, ist die zweite Ableitung von h nach u positiv definit, die Funktion h ist also

strikt konvex in u. Folglich ist jede Nullstelle der Ableitung von h nach u ein globales Minimum. Diese Nullstellen sind gerade gegeben durch

$$\begin{aligned} & 0 &=& Dh(u) = 2x^T PB + 2x^T N + 2u^T R \\ \Leftrightarrow\ & -2u^T R &=& 2x^T PB + 2x^T N \\ \Leftrightarrow\ & -Ru &=& B^T Px + N^T x \\ \Leftrightarrow\ & u &=& -R^{-1}(B^T Px + N^T x) \ = \ Fx, \end{aligned}$$

was die Behauptung zeigt.

Die exponentielle Stabilität des geregelten Systems folgt aus der Hamilton-Jacobi-Bellman Gleichung. Diese impliziert wegen der positiven Definitheit von g nach Lemma 3.10

$$DW(x) \cdot f(x, Fx) = -g(x, Fx) \leq -c\|(x^T, (Fx)^T)^T\|^2 \leq -c\|x\|^2$$

für ein geeignetes $c > 0$. Da P zudem positiv definit ist, ist das System nach Lemma 3.11 exponentiell stabil mit Lyapunov Funktion $W(x)$. □

Wenn die optimale Wertefunktion also von der Form (6.7) ist, so erhalten wir eine besonders schöne Lösung: Nicht nur lassen sich die optimalen Kontrollen u^* explizit berechnen, sie liegen darüberhinaus auch in linearer Feedback-Form vor und liefern als (natürlich gewünschtes) Nebenprodukt ein stabilisierendes Feedback.

Wie müssen also untersuchen, wann V die Form (6.7) annehmen kann. Das nächste Lemma gibt eine hinreichende Bedingung dafür an, dass die optimale Wertefunktion diese Form besitzt. Zudem liefert es eine Möglichkeit, P zu berechnen.

Lemma 6.11 Wenn die Matrix $P \in \mathbb{R}^{n\times n}$ eine symmetrische und positiv definite Lösung der algebraischen Riccati-Gleichung[2]

$$PA + A^T P + Q - (PB + N)R^{-1}(B^T P + N^T) = 0 \tag{6.9}$$

ist, so ist die optimale Wertefunktion des Problems gegeben durch $V(x) = x^T Px$.

Insbesondere existiert höchstens eine symmetrische und positiv definite Lösung P von (6.9).

Beweis: Wir zeigen zunächst, dass die Funktion $W(x) = x^T Px$ die Hamilton-Jacobi-Bellman Gleichung (6.4) löst.

Im Beweis von Lemma 6.10 wurde bereits die Identität

$$\min_{u\in U}\{DW(x) \cdot f(x, u) + g(x, u)\} = DW(x) \cdot f(x, Fx) + g(x, Fx)$$

für die Matrix $F = -R^{-1}(B^T P + N^T)$ gezeigt. Mit

$$\begin{aligned} & F^T B^T P + F^T RF + F^T N^T \\ = \ & -(N + PB)R^{-1}B^T P + (N + PB)R^{-1}RR^{-1}(B^T P + N^T) - (N + PB)R^{-1}N^T = 0 \end{aligned}$$

[2]benannt nach Jacopo Francesco Riccati, italienischer Mathematiker, 1676–1754

ergibt sich

$$\begin{aligned} & DW(x) \cdot f(x, Fx) + g(x, Fx) \\ = \; & x^T(P(A+BF) + (A+BF)^T P + Q + NF + F^T N^T + F^T RF)x \\ = \; & x^T(PA + A^T P + Q + (PB+N)F + \underbrace{F^T B^T P + F^T RF + F^T N^T}_{=0})x \\ = \; & x^T(PA + A^T P + Q + (PB+N)F)x \\ = \; & x^T(PA + A^T P + Q - (PB+N)R^{-1}(B^T P + N^T))x. \end{aligned}$$

Wenn die algebraische Riccati-Gleichung (6.9) erfüllt ist, so ist dieser Ausdruck gleich Null, womit die Hamilton-Jacobi-Bellman Gleichung erfüllt ist.

Um $V(x) = W(x)$ zu zeigen weisen wir nun nach, dass die Voraussetzungen von Satz 6.7 erfüllt sind. Aus der positiven Definitheit von P folgt $W(x) \geq 0$ und $W(0) = 0$. Wie oben gezeigt erfüllt $W(x) = x^T Px$ die Hamilton-Jacobi-Bellman Gleichung, zudem wurde die in Lemma 6.10 mittels des Feedbacks F angegebene optimale Kontrolle u^* im Beweis gerade so konstruiert, dass sie die in Satz 6.7 and u^* geforderten Bedingungen erfüllt. Also folgt die Behauptung $V(x) = W(x)$ aus Satz 6.7.

Die Eindeutigkeit der symmetrischen und positiv definiten Lösung P folgt aus der Tatsache, dass jede solche Lösung die Gleichung $V(x) = x^T Px$ für alle $x \in \mathbb{R}^n$ erfüllt, wodurch P eindeutig bestimmt ist. □

Bemerkung 6.12 Beachte, dass die Eindeutigkeitsaussage dieses Lemmas nur für die symmetrischen und positiv definiten Lösungen gilt. Die algebraische Riccati-Gleichung (6.9) kann durchaus mehrere Lösungen P haben, von denen dann aber höchstens eine positiv definit sein kann. □

Die Lemmata 6.10 und 6.11 legen die folgende Strategie zur Lösung des linear-quadratischen Problems nahe:

> Finde eine positiv definite Lösung P der algebraischen Riccati-Gleichung (6.9) und berechne daraus das optimale lineare Feedback F gemäß Lemma 6.10.

Dies liefert ein optimales lineares Feedback, das nach Lemma 6.10 zugleich das Stabilisierungsproblem löst.

Die wichtige Frage ist nun, unter welchen Voraussetzungen man die Existenz einer positiv definiten Lösung der algebraischen Riccati-Gleichung erwarten kann. Der folgende Satz zeigt, dass dieses Vorgehen unter der schwächsten denkbaren Bedingung an A und B funktioniert.

Satz 6.13 Für das linear-quadratische optimale Steuerungsproblem sind die folgenden Aussagen äquivalent:

(i) Das Paar (A, B) ist stabilisierbar.

(ii) Die algebraische Riccati-Gleichung (6.9) besitzt genau eine symmetrische und positiv definite Lösung P.

(iii) Die optimale Wertefunktion ist von der Form (6.7).

(iv) Es existiert ein optimales lineares Feedback, welches das Kontrollsystem stabilisiert.

Beweis: "(i) $\Rightarrow$ (ii)": Betrachte die Riccati-Differentialgleichung

$$\dot{P}(t) = P(t)A + A^T P(t) + Q - (P(t)B + N)R^{-1}(B^T P(t) + N^T)$$

mit Matrix-wertiger Lösung $P(t)$, die die Anfangsbedingung $P(0) = 0$ erfüllt. Aus der Theorie der gewöhnlichen Differentialgleichungen folgt, dass die Lösung $P(t)$ zumindest für t aus einem Intervall der Form $[0, t^*)$ existiert, wobei t^* maximal gewählt sei. Durch Nachrechnen sieht man, dass auch $P(t)^T$ eine Lösung ist, die ebenfalls $P(0)^T = 0$ erfüllt. Wegen der Eindeutigkeit muss also $P(t) = P(t)^T$ sein, d.h. die Lösung ist symmetrisch.

Wir wollen zunächst zeigen, dass diese Lösung für alle $t \geq 0$ existiert, dass also $t^* = \infty$ gilt. Wir nehmen dazu an, dass $t^* < \infty$ ist.

Mit analogen Rechnungen wie im Beweis von Lemma 6.10 rechnet man nach, dass die Funktion $W(t, t_1, x) := x^T P(t_1 - t)x$ für alle $t_1 - t \in [0, t^*)$ und alle $u \in U$ die Ungleichung

$$\frac{\partial}{\partial t} W(t, t_1, x) + \frac{\partial}{\partial x} W(t, t_1, x) \cdot f(x, u) + g(x, u) \geq 0 \tag{6.10}$$

erfüllt. Für jede Lösung $x(t, x_0, u)$ des Kontrollsystems mit beliebigem $u \in \mathcal{U}$ folgt daraus

$$\frac{d}{dt} W(t, t_1, x(t, x_0, u)) = \frac{\partial}{\partial t} W(t, t_1, x) + \frac{\partial}{\partial x} W(t, t_1, x) \cdot f(x, u) \geq -g(x, u).$$

Der Hauptsatz der Differential- und Integralrechnung unter Ausnutzung von $W(t_1, t_1, x) = 0$ liefert nun

$$W(0, t_1, x_0) = -\int_0^{t_1} \frac{d}{dt} W(t, t_1, x) dt \leq \int_0^{t_1} g(x(t, x_0, u), u(t)) dt \tag{6.11}$$

für $t_1 \in [0, t^*)$. Ebenfalls analog zu Lemma 6.10 rechnet man nach, dass für $u = u^* = -R^{-1}(B^T P(t) + N^T)x$ definierte Kontrollfunktion Gleichheit in (6.10) gilt, woraus mit analoger Rechnung für die durch $u^*(t) = -R^{-1}(B^T P(t) + N^T)x(t, x_0, u^*)$ definierte Kontrollfunktion die Gleichung

$$W(0, t_1, x_0) = \int_0^{t_1} g(x(t, x_0, u^*), u^*(t)) dt \tag{6.12}$$

gilt. Da G positiv definit und die Lösungen $x(t, x_0, u^*)$ stetig sind, ist $W(0, t_1, x_0) > 0$ für $x_0 \neq 0$, weswegen $P(t_1)$ positiv definit ist. Mit der speziellen Wahl $u \equiv 0$ folgt aus (6.11), dass $W(0, t_1, x_0) = x^T P(t_1)x$ gleichmäßig beschränkt ist für alle $t_1 \in [0, t^*)$. Wegen der Symmetrie gilt für die Einträge von $P(t)$ die Gleichung

$$[P(t)]_{ij} = e_i^T P(t) e_j = \frac{1}{2}((e_i + e_j)^T P(t)(e_i + e_j) - e_i^T P(t) e_i - e_j^T P(t) e_j), \tag{6.13}$$

weswegen also auch diese für $t \in [0, t^*)$ gleichmäßig beschränkt sind. Aus der Theorie der gewöhnlichen Differentialgleichungen ist bekannt, dass, wenn die rechte Seite der Differentialgleichung global definiert ist (was bei uns der Fall ist, weil sie für alle $P \in \mathbb{R}^{n\times n}$ definiert ist) und $t^* < \infty$ gilt, die Norm der Lösung gegen unendlich strebt für $t \nearrow t^*$. Dies wiederum ist nur möglich, wenn mindestens ein Eintrag von $P(t)$ unbeschränkt wächst. Da hier allerdings alle Einträge beschränkt sind, ist $t^* < \infty$ nicht möglich.

Die Lösung $P(t)$ ist also eine für alle $t \geq 0$ definierte symmetrische und positiv definite matrixwertige Funktion. Zudem folgt aus (6.12) für alle $s \geq t$ und alle $x \in \mathbb{R}^n$ die Ungleichung

$$x^T P(s)x \geq x^T P(t)x.$$

Wir zeigen nun, dass $P_\infty := \lim_{t\to\infty} P(t)$ existiert. Dazu wählen wir ein stabilisierendes Feedback F für das Paar (A, B) und setzen $u_F(t) = Fx(t, x_0, F)$. Damit erhalten wir aus (6.11) und der Abschätzung

$$g(x, Fx) \leq K\|x\|^2$$

die Ungleichung

$$\begin{aligned} W(0, t_1, x_0) &\leq \int_0^{t_1} g(x(\tau, x_0, F), u_F(\tau))d\tau \\ &\leq \int_0^{t_1} K(Ce^{-\sigma t}\|x_0\|)^2 dt \\ &\leq \underbrace{\int_0^{\infty} KC^2 e^{-2\sigma t} dt}_{=\frac{KC^2}{2\sigma}=:D<\infty} \|x_0\|^2 \quad \leq \quad D\|x_0\|^2. \end{aligned}$$

Daraus folgt $x^T P(t)x \leq D\|x\|^2$ für alle $t \geq 0$, womit $x^T P(t)x$ für jedes feste $x \in \mathbb{R}^n$ beschränkt und monoton ist und damit für $t \to \infty$ konvergiert. Mit e_j bezeichnen wir den j-ten Basisvektor. Definieren wir

$$l_{ij} = \lim_{t\to\infty} (e_i + e_j)^T P(t)(e_i + e_j) \quad \text{und} \quad l_j = \lim_{t\to\infty} e_j^T P(t) e_j,$$

so folgt aus (6.13)

$$\lim_{t\to\infty} [P(t)]_{ij} = \frac{1}{2}(l_{ij} - l_i - l_j).$$

Dies zeigt, dass der Limes $P_\infty := \lim_{t\to\infty} P(t)$ existiert. Diese Matrix ist symmetrisch und wegen

$$x^T P_\infty x \geq x^T P(t)x > 0 \quad \text{für alle } x \neq 0 \text{ und beliebiges } t > 0$$

positiv definit.

Wir zeigen schließlich, dass P_∞ die algebraische Riccati-Gleichung löst. Aus der qualitativen Theorie der gewöhnlichen Differentialgleichungen ist bekannt, dass aus $P(t) \to P_\infty$ folgt, dass P_∞ ein Gleichgewicht der Riccati-DGL sein muss.[3] Daraus folgt sofort, dass P_∞ die algebraische Riccati-Gleichung erfüllt, was die Existenz einer symmetrischen und positiv definiten Lösung zeigt. Die Eindeutigkeit folgt aus Lemma 6.11.

[3] siehe z.B.Lemma 7.2 in [7]

"(ii) ⇒ (iii)": Folgt aus Lemma 6.11

"(iii) ⇒ (iv)": Folgt aus Lemma 6.10.

"(iv) ⇒ (i)": Da ein stabilisierendes Feedback existiert, ist das Paar (A, B) stabilisierbar. □

Bemerkung 6.14 Die im Beweis von "(i)⇒(ii)" verwendete Hilfsfunktion $W(t_0, t_1)$ ist tatsächlich die optimale Wertefunktion des optimalen Steuerungsproblems

$$\text{Minimiere } J(t_0, t_1, x_0, u) := \int_{t_0}^{t_1} g(x(t, t_0, x_0, u), u(t))dt$$

auf endlichem Zeithorizont $[t_0, t_1]$, wobei $x(t, t_0, x_0, u)$ die Lösung des Kontrollsystems mit Anfangszeit t_0 und Anfangswert x_0, also $x(t_0, t_0, x_0, u) = x_0$, bezeichnet. □

Diese Beobachtung lässt sich sogar noch verallgemeinern, was wir (ohne Beweise) kurz skizzieren:

Für das linear quadratische Problem auf endlichem Zeithorizont mit Endkosten $l(x) = x^T Lx$ für eine positiv definite Matrix $L \in \mathbb{R}^n \times n$, also

$$\text{Minimiere } J(t_0, t_1, x_0, u) := \int_{t_0}^{t_1} g(x(t, t_0, x_0, u), u(t))dt + l(x(t_1, t_0, x_0, u))$$

ergibt sich die optimale Wertefunktion als

$$W(t_0, t_1) = x^T P(t_1 - t_0)x,$$

wobei $P(\cdot)$ wie im obigen Beweis die Lösung der Riccati-Differentialgleichung ist, nun aber mit Anfangsbedingung $P(0) = L$.

Das optimale Feedback ist dann analog zum unendlichen Horizont gegeben durch

$$F(t) = -R^{-1}(B^T P(t_1 - t) + N^T),$$

hängt aber nun von der Zeit t ab. Das auf $[t_0, t_1]$ optimal geregelte System lautet also

$$\dot{x}(t) = (A + BF(t))x(t).$$

Beachte, dass $F(t)$ für $t_1 \to \infty$ gegen F aus Lemma 6.10 konvergiert.

Bemerkung 6.15 Für zeitdiskrete Systeme lassen sich analoge Resultate herleiten. Hier baut man nicht auf der Hamilton-Jacobi-Bellman Gleichung sondern direkt auf dem Optimalitätsprinzip (6.3) für $K = 1$ auf. Damit kommt man auf die zeitdiskrete algebraische Riccati-Gleichung

$$A^T PA - P - (A^T PB + N)(B^T PB + R)^{-1}(B^T PA + N^T) + Q = 0.$$

Die Formel für das optimale Feedback lautet $F = (B^T PB + R)^{-1}(B^T PA + N^T)$. □

6.3 Linear-quadratische Ausgangsregelung

Wir haben im vorhergehenden Abschnitt stets vorausgesetzt, dass die Matrix G in der Definition von $g(x,u)$ positiv definit ist. In den Übungsaufgaben haben wir gesehen, dass das LQ-Problem i.A. nicht nullkontrollierend ist und dass auch das Lösungsverfahren i.A. nicht funktioniert, wenn diese Bedingung verletzt ist.

Es gibt aber trotzdem Gründe, diese Bedingung abzuschwächen. Betrachten wir wie in Kapitel 4 ein Kontrollsystem mit Ausgang (4.1), also

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t),$$

so ist es sinnvoll, das Optimierungskriterium nur von y und nicht von x abhängig zu machen, d.h. eine Kostenfunktion der Form $\tilde{g}(y,u)$ zu betrachten. Formal wählt man dazu die Teilmatrizen Q und N von G von der Form

$$Q = C^T \widetilde{Q} C,\ N = C^T \widetilde{N}$$

für Matrizen $\widetilde{Q}$ und $\widetilde{N}$ passender Dimension. Dann gilt

$$\begin{aligned} g(x,u) &= (x^T\, u^T) \underbrace{\begin{pmatrix} Q & N \\ N^T & R \end{pmatrix}}_{=:G} \begin{pmatrix} x \\ u \end{pmatrix} = (x^T\, u^T) \begin{pmatrix} C^T\widetilde{Q}C & C^T\widetilde{N} \\ \widetilde{N}^T C & R \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \\ &= (y^T\, u^T) \underbrace{\begin{pmatrix} \widetilde{Q} & \widetilde{N} \\ \widetilde{N}^T & R \end{pmatrix}}_{=:\widetilde{G}} \begin{pmatrix} y \\ u \end{pmatrix} =: \tilde{g}(y,u). \end{aligned} \tag{6.14}$$

Wir wählen dabei $\widetilde{Q}$ und $\widetilde{N}$ so, dass $\widetilde{G}$ symmetrisch und positiv definit ist. Die Matrix G ist nun allerdings nicht mehr positiv definit. Trotzdem lassen sich die Resultate aus dem vorhergehenden Abschnitt auf dieses neue G übertragen. Dazu muss man betrachten, wo und wie die positive Definitheit in den Beweisen eingeht:

(i) In Lemma 6.9 wird die positive Definitheit von G ausgenutzt, um zu zeigen, dass das Problem nullkontrollierend ist.

(ii) In Lemma 6.10 wird die positive Definitheit der Teilmatrix R implizit ausgenutzt, da die Inverse R^{-1} verwendet wird.

(iii) Im Beweis von Teil "(i)⇒(ii)" von Satz 6.13 wird die positive Definitheit von G verwendet um zu zeigen, dass $P(t)$ positiv definit ist.

Punkt (ii) ist hierbei unproblematisch, denn R ist weiterhin positiv definit. Punkt (i) und (iii) klären wir im Folgenden. Wesentlich dafür ist die Aussage des folgenden Lemmas.

Lemma 6.16 Das Paar (A,C) sei beobachtbar. Dann existiert für jedes $t_1 > 0$ ein $c > 0$, so dass für g aus (6.14) die Abschätzung

$$J(0,t_1,x_0,u) = \int_0^{t_1} g(x(t;x_0,u),u(t))dt \geq c\|x_0\|^2$$

für alle $x_0 \in \mathbb{R}^n$ und alle $u \in \mathcal{U}$ gilt.

Beweis: Aus der allgemeinen Lösungsformel

$$x(t;x_0,u) = e^{At}x_0 + \int_0^t e^{A(t-s)}Bu(s)ds = x(t;x_0,0) + x(t;0,u)$$

folgt für alle $\alpha > 0$ die Gleichung

$$x(t;\alpha x_0,\alpha u) = \alpha x(t;x_0,u).$$

Daraus folgt für $x_0 \neq 0$ und $\alpha = \|x_0\|$

$$J(0,t_1,x_0,u) = \alpha^2 J(0,t_1,x_0/\alpha,u/\alpha) = \|x_0\|^2 J(0,t_1,x_0/\|x_0\|,u/\|x_0\|).$$

Um die Behauptung zu zeigen reicht es also aus, die Existenz von $c > 0$ mit

$$J(0,t_1,x_0,u) \geq c \quad \text{für alle } x_0 \in \mathbb{R}^n \text{ mit } \|x_0\| = 1 \text{ und alle } u \in \mathcal{U} \tag{6.15}$$

zu zeigen.

Um (6.15) zu zeigen, betrachten wir zunächst

$$J(0,t_1,x_0,0) = \int_0^{t_1} x(t;x_0,0)^T Q x(t;x_0,0)dt = \int_0^{t_1} y(t)^T \widetilde{Q} y(t)dt.$$

Da (A,C) beobachtbar ist, gilt für $x_0 \neq 0$ nach Lemma 4.5 $y(\tau) \neq 0$ für ein $\tau \in [0,t_1]$. Da $y(t)$ stetig ist, folgt $y(t) \neq 0$ auf einem Intervall um τ, woraus wegen der positiven Definitheit von $\widetilde{Q}$ die Ungleichung $J(0,t_1,x_0,0) > 0$ folgt. Da $J(0,t_1,x_0,0)$ stetig in x_0 ist, existiert auf der kompakten Menge $\{x_0 \in \mathbb{R}^n \mid \|x_0\| = 1\}$ das Minimum $c_0 > 0$, weswegen

$$J(0,t_1,x_0,0) \geq c_0 \tag{6.16}$$

für alle $x_0 \in \mathbb{R}^n$ mit $\|x_0\| = 1$ gilt.

Zur Abschätzung von $J(0,t_1,x_0,u)$ wählen wir nun ein beliebiges $x_0 \in \mathbb{R}^n$ mit $\|x_0\| = 1$ sowie ein $\varepsilon > 0$. Für Kontrollen u mit

$$\int_0^{t_1} u(t)^T R u(t)dt > \varepsilon \tag{6.17}$$

gilt $\int_0^{t_1} u(t)^T u(t)dt \geq k_1\varepsilon$, wobei $k_1 = 1/\|R\|$ und folglich wegen der positiven Definitheit von $\widetilde{G}$ mit $k_2 = 1/\|\widetilde{G}^{-1}\|$

$$J(0,t_1,x_0,u) = \int_0^{t_1} \underbrace{(y(t)^T u(t)^T)\widetilde{G}\begin{pmatrix} y(t) \\ u(t) \end{pmatrix}}_{\geq k_2 \left\| \binom{y(t)}{u(t)} \right\|^2 \geq k_2\|u(t)\|^2} dt \geq k_1 k_2 \varepsilon > 0. \tag{6.18}$$

Es bleibt also die Ungleichung zu zeigen für die Kontrollen $u \in \mathcal{U}$ mit

$$\int_0^{t_1} u(t)^T R u(t)dt \leq \varepsilon. \tag{6.19}$$

Da R positiv definit ist, folgt

$$\|u(t)\|^2 \leq c_1 u(t)^T R u(t)$$

für ein $c_1 > 0$ und damit

$$\int_0^{t_1} \|u(t)\|^2 dt \leq c_1 \varepsilon.$$

Zudem gilt

$$\|u(t)\| \leq \begin{cases} \sqrt{\varepsilon}, & \|u(t)\|^2 \leq \varepsilon \\ \|u(t)\|^2/\sqrt{\varepsilon}, & \|u(t)\|^2 > \varepsilon. \end{cases}$$

Damit folgt

$$\int_0^{t_1} \|u(t)\| dt \leq \int_0^{t_1} \max\{\sqrt{\varepsilon}, \|u(t)\|^2/\sqrt{\varepsilon}\} dt \leq \int_0^{t_1} \sqrt{\varepsilon} + \|u(t)\|^2/\sqrt{\varepsilon} dt = (c_1 + t_1)\sqrt{\varepsilon}.$$

Aus der allgemeinen Lösungsformel folgt damit die Existenz einer Konstanten $c_2 > 0$, so dass

$$\|x(t; 0, u)\| \leq c_2 \sqrt{\varepsilon} \tag{6.20}$$

für alle $t \in [0, t_1]$ gilt. Ebenso folgt aus der Lösungsformel

$$\|x(t; x_0, 0)\| \leq c_3 \|x_0\| = c_3 \tag{6.21}$$

für eine geeignete Konstante $c_3 > 0$ und alle $t \in [0, t_1]$. Insbesondere folgt damit

$$\|x(t; x_0, u)\| \leq c_4 \tag{6.22}$$

für $c_4 = c_2\sqrt{\varepsilon} + c_3$.

Für das Funktional gilt nun

$$J(0, t_1, x_0, u) \geq \int_0^{t_1} x(t; x_0, u)^T Q x(t; x_0, u) dt + 2 \int_0^{t_1} x(t; x_0, u)^T N u(t) dt.$$

Für den zweiten Summanden gilt dabei wegen (6.22) die Abschätzung

$$2 \int_0^{t_1} x(t; x_0, u)^T N u(t) dt \geq -2c_4 \|N\| \int_0^{t_1} \|u(t)\| dt \geq -2c_4 \|N\| (c_1 + t_1)\sqrt{\varepsilon} =: -c_5\sqrt{\varepsilon}.$$

Aus der Abschätzung

$$(x_1 + x_2)^T Q (x_1 + x_2) = x_1^T Q x_1 + x_2^T Q x_2 + 2x_1^T Q x_2 \geq x_1^T Q x_1 + 2x_1^T Q x_2$$

folgt für den ersten Summanden mit $x_1(t) = x(t; x_0, 0)$, $x_2(t) = x(t; 0, u)$ und der Cauchy-Schwarz-Ungleichung

$$\begin{aligned} \int_0^{t_1} x(t; x_0, u)^T Q x(t; x_0, u) dt &\geq \int_0^{t_1} x_1(t)^T Q x_1(t) + \int_0^{t_1} 2x_1(t)^T Q x_2(t) dt \\ &\geq c_0 - 2\|N\| \sqrt{\int_0^{t_1} \|x_1(t)\|^2 dt} \sqrt{\int_0^{t_1} \|x_2(t)\|^2 dt} \\ &\geq c_0 - 2\|N\| c_3 \sqrt{t_1 c_2^2 \varepsilon} \;=:\; c_0 - c_6\sqrt{\varepsilon}. \end{aligned}$$

Insgesamt ergibt sich damit mit $c_7 := c_5 + c_6$

$$J(0, t_1, x_0, u) \geq c_0 - c_7\sqrt{\varepsilon}.$$

Wählen wir nun $\varepsilon = c_0^2/(2c_7)^2$ (womit $c_7\sqrt{\varepsilon} = c_0/2$ gilt), so folgt letztendlich im Fall (6.19)

$$J(0, t_1, x_0, u) \geq c_0/2.$$

Zusammen der Abschätzung (6.18) für den Fall (6.17) erhalten wir also

$$J(0, t_1, x_0, u) \geq \max\{c_0/2,\ k_1k_2c_0^2/(4c_7)^2\} =: c$$

und folglich (6.15). □

Nun können wir die Punkte (i) und (iii) in der obigen Aufstellung klären. Als erstes betrachten wir Punkt (i), d.h. wir verallgemeinern wir Lemma 6.9 auf die neue Kostenfunktion (6.14).

Lemma 6.17 Das Paar (A, C) sei beobachtbar. Dann ist das linear quadratische Problem mit g aus (6.14) nullkontrollierend.

Beweis: Wir beweisen

$$x(t; x_0, u) \not\to 0 \quad \Rightarrow \quad J(x_0, u) = \infty.$$

Gelte also $x(t; x_0, u) \not\to 0$. Dann existiert eine Folge von Zeiten $t_k \to \infty$ und ein $\varepsilon > 0$, so dass $\|x(t_k; x_0, u)\| \geq \varepsilon$. O.B.d.A. gelte $t_{k+1} - t_k \geq 1$. Mit Lemma 6.16, $x_k = x(t_k; x_0, u)$ und $u_k(\cdot) = u(t_k + \cdot)$ folgt dann

$$\int_{t_k}^{t_k+1} g(x(t; x_0, u), u(t))dt = \int_0^1 g(x(t; x_k, u_k), u_k(t))dt = J(0, 1, x_k, u_k) \geq c\varepsilon^2.$$

Damit folgt

$$\begin{aligned} J(x_0, u) &= \int_0^\infty g(x(t; x_0, u), u(t))dt \\ &\geq \sum_{k=1}^\infty \int_{t_k}^{t_k+1} g(x(t; x_0, u), u(t))dt \ \geq \ \sum_{k=1}^\infty \varepsilon^2 \ = \ \infty. \end{aligned}$$

□

Es bleibt Punkt (iii) nachzuweisen, also dass der Beweis "(i)⇒(ii)" von Satz 6.13 auch für g aus (6.14) gilt. Dies zeigt der folgende Satz.

Satz 6.18 Das Paar (A, C) sei beobachtbar. Dann gilt Satz 6.13 auch für das linear quadratische Problem mit g aus (6.14).

Beweis: Mit Lemma 6.17 an Stelle von Lemma 6.9 folgen alle Beweisteile bis auf "(i)⇒(ii)" ganz analog zu Satz 6.13.

Im Beweis von "(i)⇒(ii)" wird die positive Definitheit von G nur an einer Stelle benutzt, nämlich um zu zeigen dass

$$W(0,t_1,x_0) = \int_0^{t_1} g(x(t,x_0,u^*),u^*(t))dt$$

in Gleichung (6.12) positiv ist für alle $x_0 \neq 0$. Dies folgt aber mit Lemma 6.16 und der Beobachtbarkeitsannahme ebenfalls für g aus (6.14). Damit lässt sich der Beweis unverändert übernehmen und die Aussage folgt. □

Bemerkung 6.19 Die zugehörige Riccati-Gleichung lautet ausgeschrieben

$$PA + A^TP + C^T\widetilde{Q}C - (PB + C^T\widetilde{N})R^{-1}(B^TP + \widetilde{N}^TC)$$

und das optimale Feedback

$$F = -R^{-1}(B^TP + \widetilde{N}^TC).$$

Beachte, dass sowohl $V(x) = x^TPx$ als auch Fx i.A. *nicht* von der Form $y^T\widetilde{P}y$ oder $\widetilde{F}y$ sind. Um F für ein Kontrollsystem der Form (4.1) in Abhängigkeit von y zu implementieren, benötigen wir also nach wie vor einen Beobachter. ▫

Kapitel 7

Der Kalman Filter

Wir haben bereits in Kapitel 4 eine Möglichkeit gesehen, wie man aus dem gemessenen Ausgang $y(t) = Cx(t)$ den Zustand $x(t)$ eines Kontrollsystems mittels eines dynamischen Beobachters $z(t)$ rekonstruieren kann. Allerdings stand bei den dortigen Überlegungen in erster Linie die asymptotische Stabilität des geregelten Systems im Vordergrund und nicht so sehr die Güte der Approximation $z(t) \approx x(t)$.

Mit Hilfe der im letzten Kapitel entwickelten linear quadratischen optimalen Steuerung wollen wir nun eine Methode entwickeln, mit der eine – in einem gewissen Sinne – optimale Zustandsschätzung $z(t) \approx x(t)$ erzielt werden kann.

Die Lösung dieses linear quadratischen Zustandsschätzproblems wird durch den sogenannten Kalman Filter (oder auch LQ-Schätzer) geliefert. Dieser Filter findet sich heutzutage – in der ein oder anderen Variante – in unzähligen technischen Anwendungen, von Radargeräten über Satelliten bis zu Smartphones. Hier betrachten wir eine deterministische, zeitkontinuierliche Variante auf unendlichem Zeithorizont, weil wir für diese Version direkt auf den Ergebnissen des letzten Kapitels aufbauen können.

7.1 Zustandsschätzung auf unendlichem Zeithorizont

Wir betrachten zunächst das folgende, etwas anders formulierte Problem: Gegeben sei ein Kontrollsystem mit Ausgang (4.1) mit der etwas geänderten Notation $B = D$ und $u = v$, also

$$\dot{x}(t) = Ax(t) + Dv(t), \qquad y(t) = Cx(t), \tag{7.1}$$

wobei (A, C) beobachtbar sei.

Gegeben sei weiterhin eine Funktion $y_m : \mathbb{R} \to \mathbb{R}^l$. Ziel ist es nun, mit Hilfe der Lösungen von (7.1) eine konstruktiv berechenbare Funktion $x^*(t)$ zu finden, so dass $y(t) = Cx^*(t)$ die Funktion $y_m(t)$ gut approximiert. Die Interpretation ist, dass $y_m(t) = Cx_m(t)$ gemessene Ausgangswerte einer Lösung x_m der Differentialgleichung $\dot{x}_m = Ax_m$ mit der gleichen Matrix A wie in (7.1) sind, aus denen der Zustand $x_m(t)$ möglichst gut geschätzt werden soll. Die Erweiterung dieser Problemstellung auf Lösungen x_m von Kontrollsystemen mit zusätzlicher Kontrolle u betrachten wir im nachfolgenden Abschnitt.

Der Kalman-Filter, den wir in den folgenden Schritten herleiten werden, löst dieses Problem optimal im Sinne einer "indirekten" kleinsten Quadrate-Approximation, die in zwei Schritten vorgeht:

Im *ersten Schritt* wählen wir symmetrische und positiv definite Matrizen $\widetilde{M}$ und N passender Dimension und berechnen für jedes $\tau \geq 0$ und jeden Anfangswert x_0 zur Anfangszeit $t_0 = \tau$ eine Kontrollfunktion $v : (-\infty, \tau] \to \mathbb{R}^n$, so dass die zugehörige Lösung $x_\tau(t) = x(t; \tau, x_0, v)$ das Funktional

$$J_\tau(x_0, v) := \int_{-\infty}^{\tau} (Cx_\tau(t) - y_m(t))^T \widetilde{M}(Cx_\tau(t) - y_m(t)) + v(t)^T N v(t) dt \tag{7.2}$$

minimiert. Wir nehmen dabei an, dass die optimale Wertefunktion

$$P_\tau(x_0) := \inf_{v \in \mathcal{U}} J_\tau(x_0, v)$$

endlich ist.

Im *zweiten Schritt* wählen wir dann $x^*(\tau)$ so, dass $P_\tau(x^*(\tau))$ minimal wird, d.h. dass

$$P_\tau(x^*(\tau)) = \min_{x_0 \in \mathbb{R}^n} P_\tau(x_0)$$

gilt.

Der Ansatz mag auf den ersten Blick etwas umständlich erscheinen. Er führt aber auf eine sehr einfach zu implementierende Lösung, die wir nun herleiten wollen.

Zunächst einmal transformieren wir die Zeit so, dass das Integral in (7.2) von 0 bis ∞ läuft, wie dies in unserem üblichen linear-quadratischen Problem der Fall ist.

Dazu setzen wir $x^\tau(t; x_0, v) := x(\tau - t; x_0, v)$ und $y_m^\tau(t) = y_m(\tau - t)$. Dann gilt mit der Abkürzung $x^\tau(t) = x^\tau(t; x_0, v)$ für

$$J_\tau^-(x_0, v) := \int_0^{\infty} (Cx^\tau(t) - y_m^\tau(t))^T \widetilde{M}(Cx^\tau(t) - y_m^\tau(t)) + v(t)^T N v(t) dt \tag{7.3}$$

die Gleichheit $J_\tau^-(x_0, v) = J_\tau(x_0, v(\tau - \cdot))$ und damit insbesondere

$$P_\tau^-(x_0) := \inf_{v \in \mathcal{U}} J_\tau^-(x_0, v) = P_\tau(x_0).$$

Beachte, dass $x^\tau(t; x_0, v)$ Lösung des Kontrollsystems

$$\dot{x}^\tau(t) = -Ax^\tau(t) - Dv(\tau - t)$$

ist. Mit einer weiteren Transformation können wir (7.3) nun (fast) auf die Form unseres linear quadratischen Ausgangsregelungsproblems gemäß Definition 6.1 mit g aus (6.14) bringen:

Dazu erweitern wir den Zustand $x \in \mathbb{R}^n$ des Systems um eine Komponente $x_{n+1}(t) \equiv const$, also $\dot{x}_{n+1}(t) \equiv 0$. Dies erreichen wir durch die Wahl

$$\bar{x} := \begin{pmatrix} x \\ x_{n+1} \end{pmatrix}, \quad \overline{A} := \begin{pmatrix} -A & 0 \\ 0 & 0 \end{pmatrix} \text{ und } \overline{D} := \begin{pmatrix} -D \\ 0 \end{pmatrix}.$$

Definieren wir nun

$$\overline{M}_\tau(t) := \begin{pmatrix} C^T\widetilde{M}C & -C^T\widetilde{M}y_m^\tau(t) \\ -y_m^\tau(t)^T\widetilde{M}C & y_m^\tau(t)^T\widetilde{M}y_m^\tau(t) \end{pmatrix}$$

und $g(t,\bar{x},v) := \bar{x}^T\overline{M}_\tau(t)\bar{x} + v^TNv$ so folgt für $\bar{x} = \begin{pmatrix} x \\ 1 \end{pmatrix}$

$$g(t,\bar{x},v) = (Cx - y_m^\tau(t))^T\widetilde{M}(Cx - y_m^\tau(t)) + v(t)^TNv(t)dt.$$

Folglich gilt für $\bar{x}_0 = \begin{pmatrix} x_0 \\ 1 \end{pmatrix}$ und $\bar{x}^\tau(t,\bar{x}_0,v) = \begin{pmatrix} x^\tau(t,x_0,v) \\ 1 \end{pmatrix}$

$$J_\tau^-(x_0,v) = \int_0^\infty g(t,\bar{x}^\tau(t;\bar{x}_0,v),v(t))dt =: \overline{J}_\tau(\bar{x}_0,v).$$

Mit $\overline{P}_\tau$ bezeichnen wir wie üblich die optimale Wertefunktion. Dieses Problem ist von der üblichen LQ-Form mit Ausnahme der Tatsache, dass g nun explizit von der Zeit abhängt. Tatsächlich sind aber die im Beweis von Satz 6.13 verwendeten Gleichungen weiterhin gültig, wenn wir die Zeit in $\overline{M}(t)$ passend berücksichtigen. Genauer gilt (was wir hier aus Zeitgründen nicht beweisen):

Betrachte für $t \in [0,\sigma]$ die Lösung der Riccati-Differentialgleichung

$$\dot{\overline{Q}}_{\tau,\sigma}(t) = \overline{Q}_{\tau,\sigma}(t)\overline{A} + \overline{A}^T\overline{Q}_\tau(t) + \overline{M}_\tau(\sigma - t) - \overline{Q}_{\tau,\sigma}(t)\overline{D}N^{-1}\overline{D}^T\overline{Q}_{\tau,\sigma}(t) \tag{7.4}$$

mit Anfangsbedingung $\overline{Q}_{\tau,\sigma}(0) = 0$. Dann gilt die Konvergenz

$$\overline{P}_\tau(\bar{x}) := \lim_{\sigma\to\infty} \bar{x}^T\overline{Q}_{\tau,\sigma}(\sigma)\bar{x}.$$

Nun zerlegen wir $\overline{Q}_{\tau,\sigma}(t)$ passend zur Definition von $\overline{A}$: Schreiben wir

$$\overline{Q}_{\tau,\sigma}(t) = \begin{pmatrix} Q_{\tau,\sigma}(t) & q_{\tau,\sigma}(t) \\ q_{\tau,\sigma}(t)^T & \alpha_{\tau,\sigma}(t) \end{pmatrix},$$

so folgt aus der Form der Matrizen $\overline{A}$ und $\overline{D}$, dass $Q_{\tau,\sigma}(t)$ die Gleichung

$$\dot{Q}_{\tau,\sigma}(t) = -Q_{\tau,\sigma}(t)A - A^TQ_{\tau,\sigma}(t) + C^T\widetilde{M}C - Q_{\tau,\sigma}(t)DN^{-1}D^TQ_{\tau,\sigma}(t)$$

erfüllt. Dies ist aber genau die Riccati-Differentialgleichung aus dem Beweis von Satz 6.13. Zudem sind alle Daten und damit auch $Q_{\tau,\sigma}(t) = Q(t)$ unabhängig von τ und σ. Es folgt also

$$\lim_{\sigma\to\infty} Q(\sigma) = Q,$$

wobei Q die algebraische Riccati-Gleichung

$$-QA - A^TQ + C^T\widetilde{M}C - QDN^{-1}D^TQ = 0 \tag{7.5}$$

löst.

Damit erhalten wir mit $\bar{x}_0^T = (x_0^T, 1)$ und $q_\tau = \lim_{\sigma\to\infty} q_{\tau,\sigma}(\sigma)$, $\alpha_\tau = \lim_{\sigma\to\infty} \alpha_{\tau,\sigma}(\sigma)$

$$P_\tau(x_0) = \overline{P}_\tau(\bar{x}_0) = \lim_{\sigma\to\infty} \bar{x}_0^T \overline{Q}_{\tau,\sigma}(\sigma)\bar{x}_0 = x_0^T Q x_0 + 2x_0^T q_\tau + \alpha_\tau.$$

Der im zweiten Schritt des Ansatzes gesuchte Wert $x^*(\tau)$ ergibt sich damit (durch Ableiten des Ausdrucks und Umstellen nach x_0) zu

$$x^*(\tau) = -Q^{-1}q_\tau = -Sq_\tau$$

für $S := Q^{-1}$. Durch Multiplikation von (7.5) mit S von links und rechts sowie mit -1 folgt, dass S die sogenannte *duale Riccati-Gleichung*

$$AS + SA^T - SC^T\widetilde{M}CS + DN^{-1}D^T = 0 \tag{7.6}$$

löst.

Es bleibt q_τ zu berechnen. Aus der Riccati-Differentialgleichung (7.4) folgt für $q_{\tau,\sigma}(t)$ die Differentialgleichung

$$\dot{q}_{\tau,\sigma}(t) = -A^T q_{\tau,\sigma}(t) - Q(t)DN^{-1}D^T q_{\tau,\sigma}(t) - C^T\widetilde{M}y_m(\tau - \sigma + t)$$

mit Anfangsbedingung $q_{\tau,\sigma}(0) = 0$. Hieraus folgt

$$\dot{q}_{\tau+s,\sigma+s}(t) = \dot{q}_{\tau,\sigma}(t)$$

und da diese beiden Lösungen für $t = 0$ übereinstimmen, folgt

$$q_{\tau+s,\sigma+s}(t) = q_{\tau,\sigma}(t).$$

Damit folgt

$$\begin{aligned}\left.\frac{d}{ds}\right|_{s=0} q_{\tau+s,\sigma+s}(\sigma+s) &= \dot{q}_{\tau,\sigma}(\sigma)\\ &= -A^T q_{\tau,\sigma}(\sigma) - Q(\sigma)DN^{-1}D^T q_{\tau,\sigma}(\sigma) - C^T\widetilde{M}y_m(\tau)\end{aligned}$$

und folglich mit $\sigma \to \infty$

$$\frac{d}{d\tau}q_\tau = -A^T q_\tau - QDN^{-1}D^T q_\tau - C^T\widetilde{M}y_m(\tau).$$

Damit erhalten wir schließlich mit (7.6)

$$\begin{aligned}\dot{x}^*(\tau) &= -S\frac{d}{d\tau}q_\tau\\ &= SA^T q_\tau + DN^{-1}D^T q_\tau + SC^T\widetilde{M}y_m(\tau)\\ &= -SA^T S^{-1}x^*(\tau) - DN^{-1}D^T S^{-1}x^*(\tau) + SC^T\widetilde{M}y_m(\tau)\\ &= (-SA^T - DN^{-1}D^T)S^{-1}x^*(\tau) + SC^T\widetilde{M}y_m(\tau)\\ &= (AS - SC^T\widetilde{M}CS)S^{-1}x^*(\tau) + SC^T\widetilde{M}y_m(\tau)\\ &= Ax^*(\tau) - SC^T\widetilde{M}(Cx^*(\tau) - y_m(\tau))\\ &= Ax^*(\tau) + L(Cx^*(\tau) - y_m(\tau))\end{aligned}$$

mit $L = -SC^T\widetilde{M}$.

Diese Differentialgleichung ist der sogenannte *Kalman-Filter*. Seine Anwendung ist wie folgt: Ist $x^*(t)$ bekannt, so kann $x^*(s)$, $s > t$, durch Lösen der Differentialgleichung auf dem Intervall $[t, s]$ (analytisch oder numerisch) aus den Daten $y_m|_{[t,s]}$ berechnet werden. Der Kalman-Filter eignet sich also zur rekursiven Online-Implementierung.

Zwei Eigenschaften des Kalman-Filters wollen wir hier noch explizit festhalten:

(i) Die Matrix L hängt nicht von y_m ab. Um L zu berechnen, muss lediglich eine der beiden Riccati-Gleichungen (7.5) oder (7.6) gelöst werden.

(ii) Die Matrix $A+LC$ ist Hurwitz. Die Matrix L^T ist nämlich das LQ-optimale Feedback des zur dualen Riccati-Gleichung (7.6) gehörigen dualen optimalen Steuerungsproblems ist. Daher ist $A^T + C^T L^T$ asymptotisch stabil und folglich auch $A + LC = (A^T + C^T L^T)^T$, weil diese beiden Matrizen die gleichen Eigenwerte besitzen.

7.2 Der Kalman-Filter als Beobachter

Wir wollen den Kalman-Filter nun für das in der Einführung dieses Kapitels skizzierte Beobachterproblem anwenden.

Gegeben sei dazu ein Kontrollsystem mit Ausgang (4.1), also

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t),$$

mit beobachtbarem Paar (A, C). Gegeben seien weiterhin ein unbekannter Anfangswert x_0 sowie eine bekannte Kontrollfunktion $u(t)$, $t \geq 0$, die zugehörigen Ausgangswerte $y(t) = Cx(t; x_0, u)$, $t \geq 0$, sowie eine Schätzung z_0 des Anfangswerts x_0. Gesucht ist nun eine Kurve $z(t)$, $t \geq 0$, mit $z(0) = z_0$ im $\mathbb{R}^n$, so dass der Schätzfehler $Cz(t) - y(t)$ in einem geeigneten Sinne möglichst klein wird und so, dass $z(t)$ nur von $y|_{[0,t]}$ abhängt (also aus den zur Zeit t bekannten Daten berechenbar ist). Der Ausgang $y(t)$ spielt hier also die Rolle der Messgröße $y_m(t)$ im Kalman-Filter.

Zur Lösung des Problems machen wir den Ansatz

$$\dot{z}(t) = Az(t) + Bu(t) + v(t), \tag{7.7}$$

wobei $v : \mathbb{R} \to \mathbb{R}^n$ so bestimmt werden soll, dass $z(t)$ eine möglichst gute Schätzung ist. Um den Term $Bu(t)$ aus der Gleichung zu eleminieren, definieren wir den *Schätzfehler* $e(t) := z(t) - x(t)$. Dieser erfüllt die Gleichung

$$\dot{e}(t) = Ae(t) + v(t), \tag{7.8}$$

d.h. wir haben hier ein Kontrollsystem (7.1) mit $D = \mathrm{Id}$ und $x = e$. Der Fehler e spielt hier also die Rolle des x in (7.1).

Wir wollen uns nun überlegen, wie das Gegenstück der Messgröße y_m für das e-System lautet. Wir bezeichnen diese mit e_m. In Abschnitt 7.1 haben wir (in der Notation dieses

Abschnitts) die Größe $Ce(t) - e_m(t)$ minimiert, hier wollen wir die Größe $Cz(t) - y(t)$ minimieren. Also muss gelten

$$Ce(t) - e_m(t) = Cz(t) - y(t) \tag{7.9}$$

und damit

$$e_m(t) = y(t) + Ce(t) - Cz(t) = y(t) + Cz(t) - \underbrace{Cx(t)}_{=y(t)} - Cz(t) = 0.$$

Die Messwerte für die e-Gleichung sind also konstant gleich Null. Dies liegt daran, dass wir die gemessene Größe $y_m(t) = y(t) = Cx(t)$ durch die Definition von e bereits in die Gleichung für e einbezogen haben.

Berechnen wir nun gemäß dem vorhergehenden Abschnitt das Feedback L für den Kalman-Filter für (7.8), so ergibt sich die Filtergleichung wegen $e_m \equiv 0$ zu

$$\dot{e}^*(t) = (A + LC)e^*(t).$$

Dies ist äquivalent zu

$$\dot{z}(t) = Az(t) + Bu(t) + L(Cz(t) - y(t)) \tag{7.10}$$

und liefert damit eine online implementierbare Beobachtergleichung (beachte die strukturelle Ähnlichkeit zum dynamischen Beobachter in Kapitel 4) zur Berechnung von $z(t)$, die nur noch (analytisch oder numerisch) gelöst werden muss. Beachte, dass die optimalen Schätzungen für die e und die z-Variable mittels $e^*(t) = z^*(t) - x(t)$ zusammenhängen. Während wir den Kalman Filter formal auf die e-Gleichung (7.8) anwenden, verwenden wir zur Berechnung des Schätzers $z^*(t)$ die z-Gleichung (7.10), denn ansonsten bräuchten wir den unbekannten Zustand $x(t)$, um $z^*(t)$ aus $e^*(t)$ zu berechnen.

Nachdem wir hier keine Messwerte $y(t)$ für $t < 0$ gegeben haben, können wir den optimalen Startwert $e^*(0)$ hier nicht wie im vorhergehenden Abschnitt berechnen. Aber selbst wenn wir es könnten, würde uns dies nichts nützen, denn für (7.10) müssten wir dann ja $z(0) = e^*(0) + x_0$ verwenden — der Wert x_0 ist aber unbekannt. Es liegt also nahe, in (7.10) den Schätzwert $z_0 \approx x_0$ als Anfangswert zu verwenden. Weil $A - LC$ is Hurwitz ist, konvergiert der Schätzfehler $e^*(t)$ für $t \to \infty$ gegen 0, d.h. die Approximation $z(t) \approx x(t)$ wird mit wachsendem t immer besser. Da unserem Ansatz aber ein LQ-optimales Steuerungsproblem zu Grunde liegt, kann man erwarten, dass die Schätzung $z(t)$ ausgehend von $z(0) = z_0$ in einem gewissen Sinne optimal ist.

Um zu sehen, welcher Art diese Optimalität ist, setzen wir $y(t)$ für $t < 0$ so fort, dass sich $e^*(0) = z_0 - x_0$ und damit $z(0) = z_0$ als Lösung des Kalman-Filters ergibt. Wir erzeugen also gewissermaßen "künstliche" Messwerte, für die der Kalman Filter zur Zeit $t = 0$ gerade den Schätzwert z_0 liefert. Dies ist gerade dann der Fall, wenn wir $y(t)$ mittels

$$y(t) = \begin{cases} Cx(t; z_0, 0), & t < 0 \\ Cx(t; x_0, u), & t \geq 0 \end{cases} \tag{7.11}$$

aus der Vorwärtslösung von (4.1) für x_0 und u und der Rückwärtslösung für z_0 und $u \equiv 0$ zusammensetzen: Für $v \equiv 0$ und $e(0) = 0$ gilt dann wegen $e_m \equiv 0$

$$Ce(t) - e_m = 0$$

für alle $t < 0$ und es folgt $J_0(0,0) = 0$ für das Optimalitätskriterium (7.2), folglich auch $P_0(0) = 0$ und somit $e^*(0) = 0$. Damit folgt $z^*(0) = z_0 - e^*(0) = z_0$.

Der aus dem Anfangswert z_0 berechnete Approximationswert $z(t)$ ist also gerade der Endwert derjenigen Lösung von (7.7), welche die zusammengesetzte Kurve (7.11) im Sinne von (7.2) am Besten approximiert.

Der große Vorteil des Kalman-Filters ist es, dass er auch bei ungenauen Daten $\tilde{y}(t) \approx y(t)$ gute Approximationen liefert. Dies kann mit stochastischen Methoden mathematisch rigoros formuliert und bewiesen werden.

Auch für den Kalman-Filter existiert eine zeitdiskrete Version. In diesem Fall wird die Differentialgleichung (7.10) zu einer Differenzengleichung

$$z(k+1) = Az(k) + Bu(k) + L(Cz(k) - y(k)).$$

Da diese leichter zu implementieren ist als die Differentialgleichung (7.10) (die man ja zuerst noch numerisch lösen muss) und zudem mit diskreten Messwerten $y(k)$ auskommt (welche technisch leichter zu messen sind als kontinuierliche Messwerte $y(t)$), wird in der Praxis der zeitdiskrete Kalman-Filter oft bevorzugt.

Kapitel 8

Nichtlineare Kontrollsysteme

In diesem und den folgenden Kapiteln werden wir uns mit nichtlinearen Kontrollsystemen der allgemeinen zeitkontinuierlichen Form

$$\dot{x}(t) = f(x(t), u(t)) \tag{8.1}$$

bzw. der zeitdiskreten Form

$$x(k+1) = f(x(k), u(k)), \tag{8.2}$$

kurz auch geschrieben als $x^+ = f(x, u)$, befassen. Ein Beispiel für ein nichtlineares Kontrollsystem in kontinuierlicher Zeit ist das bereits bekannte nichtlineare Pendel auf dem Wagen (1.5). Während wir den Zustands- und Kontrollwerteraum für zeitkontinuierliche Systeme als $\mathbb{R}^n$ bzw. $\mathbb{R}^m$ gewählt haben, können wir bei zeitdiskreten Systemen beliebige metrische Räume X und U als Zustands- und Kontrollraum verwenden.

In den folgenden beiden Abschnitten fassen wir einige wichtige Grundlagen zusammen.

8.1 Zeitkontinuierliche Systeme

Im kontinuierlichen betrachten wir Kontrollfunktionen mit Werten in $U \subset \mathbb{R}^m$. Die Funktion $f : \mathbb{R}^n \times U \to \mathbb{R}^n$ ist ein parameterabhängiges stetiges Vektorfeld. Den Raum der Kontrollfunktionen bezeichnen wir weiterhin mit $\mathcal{U}$, werden diesen aber im Vergleich zu den vorhergehenden Kapiteln im folgenden Abschnitt in Zusammenhang mit einem Existenz- und Eindeutigkeitsresultat erweitern. Genauer verwenden wir Kontrollfunktionen aus $L^\infty(\mathbb{R}, U)$ und den folgenden Satz von Carathéodory.

Satz 8.1 (Satz von Carathéodory) Betrachte ein Kontrollsystem mit folgenden Eigenschaften:

i) Der Raum der Kontrollfunktionen ist gegeben durch

$$\mathcal{U} = L^\infty(\mathbb{R}, U) := \{u : \mathbb{R} \to U \,|\, u \text{ ist messbar und essentiell beschränkt}^1\}.$$

[1] d.h. beschränkt außerhalb einer Lebesgue–Nullmenge

ii) Das Vektorfeld $f : \mathbb{R}^n \times U \to \mathbb{R}^n$ ist stetig.

iii) Für jedes $R > 0$ existiert eine Konstante $L_R > 0$, so dass die Abschätzung

$$\|f(x_1, u) - f(x_2, u)\| \leq L_R \|x_1 - x_2\|$$

für alle $x_1, x_2 \in \mathbb{R}^n$ und alle $u \in U$ mit $\|x_1\|, \|x_2\|, \|u\| \leq R$ erfüllt ist.

Dann gibt es für jeden Punkt $x_0 \in \mathbb{R}^n$, jede Zeit $t_0 \in \mathbb{R}$ und jede Kontrollfunktion $u \in \mathcal{U}$ ein (maximales) offenes Intervall I mit $t_0 \in I$ und genau eine absolut stetige[2] Funktion $x(t)$, die die Integralgleichung

$$x(t) = x_0 + \int_{t_0}^{t} f(x(\tau), u(\tau))\, d\tau$$

für alle $t \in I$ erfüllt.

Definition 8.2 Wie bezeichnen die eindeutige Funktion $x(t)$ aus Satz 8.1 mit $x_u(t; t_0, x_0)$ und nennen sie die *Lösung* von (8.1) zum *Anfangswert* $x_0 \in \mathbb{R}^n$ und zur *Kontrollfunktion* $u \in \mathcal{U}$. Im Fall $t_0 = 0$ schreiben wir kurz $x_u(t, x_0, u) = x(t; 0, x_0)$. □

Die folgende Beobachtung rechtfertigt diese Definition: Da $x_u(t; t_0, x_0)$ absolut stetig ist, ist diese Funktion für fast alle $t \in I$ nach t differenzierbar. Insbesondere folgt also aus dem Satz 8.1, dass $x_u(t; t_0, x_0)$ die Differentialgleichung (8.1) für fast alle $t \in I$ erfüllt, d.h. es gilt

$$\dot{x}_u(t; t_0, x_0) = f(x_u(t; t_0, x_0), u(t))$$

für fast alle $t \in I$.

Bemerkung 8.3 Im Weiteren nehmen wir stets an, dass die Voraussetzungen (i)–(iii) von Satz 8.1 erfüllt sind, werden dies aber nur in wichtigen Sätzen explizit formulieren. □

Der Beweis von Satz 8.1 (auf den wir aus Zeitgründen nicht näher eingehen) verläuft ähnlich wie der Beweis des entsprechenden Satzes für stetige gewöhnliche Differentialgleichungen, d.h. mit dem Banach'schen Fixpunktsatz angewendet auf einen passenden Funktionenraum. Er findet sich zusammen mit einer Einführung in die zugrundeliegende Lebesgue–Maßtheorie z.B. in dem Buch *Mathematical Control Theory* von E.D. Sontag [19, Anhang C].

Aus dem Eindeutigkeitssatz folgen wie bei stetigen gewöhnlichen Differentialgleichungen für alle $t, s \in \mathbb{R}$ die Beziehungen

$$x_u(t; t_0, x_0) = x_u(t; s, x_u(s; t_0, x_0)) \tag{8.3}$$

(die sogenannte Kozykluseigenschaft) und

$$x_u(t; t_0, x_0) = x_{u(s+\cdot)}(t - s; t_0 - s, x_0),$$

die wir in Korollar 1.10 bereits für lineare Systeme formuliert haben. Aus der zweiten Gleichung folgt mit $s = t_0$ insbesondere

$$x_u(t; t_0, x_0) = x_{u(t_0+\cdot)}(t - t_0, x_0). \tag{8.4}$$

[2] Eine Funktion heißt absolut stetig, wenn sie als Integral über eine L^∞-Funktion geschrieben werden kann.

8.2 Abtastsysteme

Wie im ersten Kapitel schon erwähnt, kann jedem zeitkontinuierlichem Kontrollsystem, das die Voraussetzungen des Satzes von Carathéodory erfüllt, durch Abtastung ein zeitdiskretes System zugeordnet werden. Dieses entsteht einfach daraus, dass wir den Zustand des kontinuierlichen Systems nur zu den Zeitpunkten kT für $k \in \mathbb{N}$ und eine feste Abtastzeit[3] $T > 0$ betrachten. Bezeichnet $\hat{x}(t, x_0, \hat{u})$ die zeitkontinuierliche Lösung, so sind die Zustände $x(k)$ des Abtastsystems gegeben durch

$$x(k) = \hat{x}_{\hat{u}}(kT, x_0).$$

Mit Hilfe von (8.3) und (8.4) folgt

$$x(k+1) = \hat{x}_{\hat{u}}((k+1)T; kT, \hat{x}_{\hat{u}}(kT, x_0)) = \hat{x}_{\hat{u}}((k+1)T; kT, x(k)) = \hat{x}_{\hat{u}(kT+\cdot)}(T, x(k)).$$

Definieren wir für die Kontrollfunktion $\hat{u}(\cdot)$ die Funktionen $u(k) : [0, T] \to \mathbb{R}$ mittels

$$u(k)(t) := \hat{u}(kT + t), \quad t \in [0, T]$$

so ergibt sich

$$x(k+1) = \hat{x}_{u(k)}(T, x(k)) =: f(x(k), u(k)), \tag{8.5}$$

wodurch das zeitdiskrete *Abtastsystem* definiert ist. Im Allgemeinen ist dabei $u(k) \in L^\infty([0, T], U)$. Wie in Kapitel 1 bereits erläutert, ist es ist aber möglich (und in der technischen Praxis üblich), $u(k)$ aus einer eingeschränkteren Menge zu wählen. Sehr verbreitet ist die Wahl, $u(k)$ einfach als konstante Funktion zu wählen. Die zugehörige zeitkontinuierliche Kontrollfunktion $\hat{u}$ ist dann stückweise konstant. Manchmal werden die $u(k)$ auch als Polynome gewählt, dann ist $\hat{u}$ eine stückweise polynomiale (aber i.d.R. in den "Nahtstellen" kT nicht stetige) Funktion.

Wir werden im weiteren Verlauf der Vorlesung zumeist zeitdiskrete Systeme verwenden, weil für diese die Methode der Modellprädiktiven Regelung, die im Folgenden im Vordergund stehen soll, einfacher handzuhaben ist. Wir werden aber an einigen Stellen auf Eigenheiten der Abtastsysteme eingehen.

[3] englisch: Abtastung = sampling, Abtastzeit = sampling time, Abtastsystem = sampled-data system

Kapitel 9

Introduction to Model Predictive Control

In this introduction, we present the basics of Model Predictive Control (henceforth abbreviated as MPC) in an informal way. In particular, we introduce the central idea of iterative optimal control on a moving finite horizon.

MPC is a method for obtaining an approximately optimal feedback control for an optimal control problem on an infinite or indefinite time horizon. Feedback here means that the control at time k is of the form $u(k) = \mu(x(k))$ for a map $\mu : X \to U$. We have already seen how linear quadratic optimal control leads to an optimal feedback control. The decisive property that makes the approach via the Riccati equation computationally feasible is that the optimal value function V is of quadratic form $V(x) = x^T P x$. This means that we only have to determine the coefficients of the matrix P, whose number is of the order $O(n^2)$. However, as soon as the cost is nonquadratic, the dynamics is nonlinear or state and/or control constraints are introduced into the problem, the function V is no longer quadratic. This means that an exact representation by finitely many coefficients is in general no longer possible. The same holds for the optimal feedback law, which is in general a rather complicated function in x for which already the storage poses challenging problems, known as the "curse of dimensionality". This implies that the direct computation and storage of an approximately optimal feedback law is computationally intractable even for problems in moderate space dimensions, say 5–10.

In contrast to this, nowadays there exist powerful optimization algorithms which can compute single optimal trajectories in very short time, even for high dimensional systems like accurately discretized PDEs. The key idea of MPC is now to use this computational approach for obtaining a feedback law which is near optimal for infinite horizon problems.

In order to describe the idea of MPC, consider the discrete time model

$$x^+ = f(x, u) \tag{9.1}$$

where $f : X \times U \to X$ is a known and in general nonlinear map which assigns to a state x and a control value u the successor state x^+ at the next time instant and X and U are metric spaces. Starting from the current state $x(j)$, for any given control sequence

$u(0), \dots, u(N-1)$ with *horizon length* $N \geq 2$, we can now iterate (9.1) in order to construct a prediction trajectory x_u defined by

$$x_u(0) = x(j), \quad x_u(k+1) = f(x_u(k), u(k)), \quad k = 0, \dots, N-1. \tag{9.2}$$

Proceeding this way, we obtain predictions $x_u(k)$ for the state of the system $x(j+k)$ for k time steps into the future, depending on the chosen control sequence $u(0), \dots, u(N-1)$.

Now we use optimal control in order to determine $u(0), \dots, u(N-1)$. To this end, we fix a cost function $\ell(x, u)$. This function may be very general. In the simplest case, X and U are vector spaces with norms and ℓ penalizes the distance of x to some "reference state" x_*; for simplicity we assume $x_* = 0$. Typically, one does not penalize the deviation of the state from the reference but also—if desired—the distance of the control values $u(k)$ to a reference control u_*, which here we also choose as $u_* = 0$. A common and popular choice for such a function is the quadratic function

$$\ell(x_u(k), u(k)) = \|x_u(k)\|^2 + \lambda \|u(k)\|^2,$$

where $\|\cdot\|$ denotes the norms[1] of the spaces X and U and $\lambda \geq 0$ is a weighting parameter for the control, which could also be chosen as 0 if no control penalization is desired. The purpose of MPC with a stage cost penalizing the distance to an equilibrium is that the optimal control should drive the system towards the reference state $x_* = 0$, in order to stabilize the system at this state, just as in the linear quadratic case. MPC with such stage costs is thus called *stabilizing MPC*. In contrast to this, MPC with more general cost function is often called *economic MPC*.

Regardless which cost function is used, the optimal control problem now reads

$$\text{minimize} \quad J_N(x(j), u(\cdot)) := \sum_{k=0}^{N-1} \ell(x_u(k), u(k))$$

with respect to all admissible[2] control sequences $u(0), \dots, u(N-1)$ with x_u generated by (9.2).

Let us assume that this optimal control problem has a solution which is given by the minimizing control sequence $u^\star(0), \dots, u^\star(N-1)$, i.e.,

$$\min_{u(0),\dots,u(N-1)} J_N(x(j), u(\cdot)) = \sum_{k=0}^{N-1} \ell(x_{u^\star}(k), u^\star(k)).$$

In order to get the desired feedback value $\mu(x(j))$, we now set $\mu(x(j)) := u^\star(0)$, i.e., we apply the first element of the optimal control sequence. This procedure is sketched in Fig. 9.1.

We now apply this feedback law, i.e., the first element of $u^\star$, on the time interval from j to $j+1$. Thus we obtain

$$x(j+1) = f(x(j), \mu(x(j))) \tag{9.3}$$

[1] For simplicity of notation we use the same symbol for the in gereral different norms on X and U.
[2] The meaning of "admissible" will be defined in Sect. 11.2.

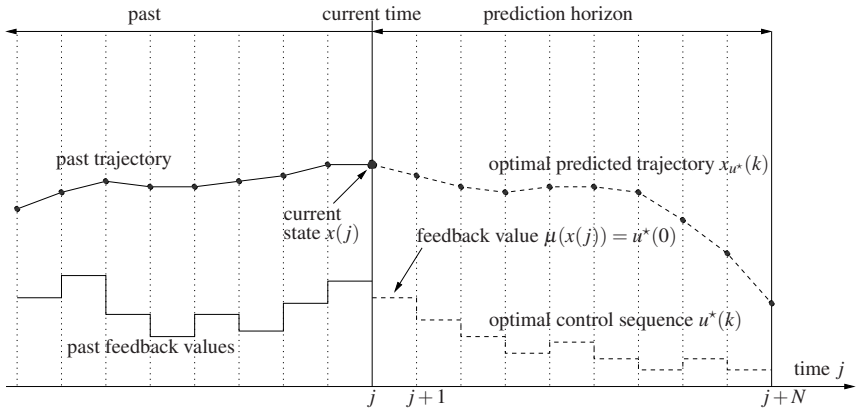


Abbildung 9.1: Illustration of the MPC step at time j

System (9.3) is called the *MPC closed-loop system*.

At the following time instants $j+1, j+2, \ldots$ we repeat the procedure with the new measurements $x(j+1), x(j+2), \ldots$ in order to derive the feedback values $\mu(x(j+1)), \mu(x(j+2)), \ldots$. In other words, we obtain the feedback law μ by an *iterative online optimization* over the predictions generated by our model (9.1). This is the first key feature of model predictive control.

From the prediction horizon point of view, proceeding this iterative way the trajectories $x_u(k)$, $k = 0, \ldots, N$ provide a prediction on the discrete interval $j, \ldots, j+N$ at time j, on the interval $j+1, \ldots, j+N+1$ at time $j+1$, on the interval $j+2, \ldots, j+N+2$ at time $j+2$, and so on. Hence, the prediction horizon is moving and this *moving horizon* is the second key feature of model predictive control.

Regarding terminology, another term which is often used alternatively to *model predictive control* is *receding horizon control*. While the former expression stresses the use of model based predictions, the latter emphasizes the moving horizon idea. Despite these slightly different literal meanings, we prefer and follow the common practice to use these names synonymously. In addition, one often uses the term Nonlinear Model Predictive Control (NMPC) if one wants to indicate that our model (9.1) need not be a linear map.

9.1 Motivating examples

In this section we present three motivating examples (the corresponding numerical simulations and experiments will only be presented in the lectures), which show different phenomema which can be observed when using MPC.

The first example is the classical inverted pendulum, which is available as a real experiment at the Chair of Applied Mathematics. The cost function ℓ here penalizes the distance to the upright equilibrium. The ordinary differential equation system (which is similar to (1.5) but a little more complex in order to take into acount the motor dynamics) is sampled with sampling time $T = 50$ms. The video shows that this time is enough to solve the optimal

control problem numerically in each sampling interval[3].

The second example is a very simple economic problem of optimal investment. Let $x \geq 0$ be the amount of capital invested in a company. The invested capital x yields a return of $Ax^\alpha - x$ in one time unit (e.g., a year), i.e., after one time step the amount of capital is Ax^α. The control u describes the amount of capital which is invested again in the next time step. Hence, the amount of money to be consumed is $Ax^\alpha - u$. The utility of consumption is measured by a classical logarithmic utility function $\ln(Ax^\alpha - u)$. We want to maximize this utility over several time steps, hence we want to minimize the cost function $\ell(x,u) = -\ln(Ax^\alpha - u)$. We note that this cost function is not of the form of a function which penalizes the distance from a reference point x_*. Numerical simultions for $A = 5$ and $\alpha = 0.34$ and state constraint set $\mathbb{X} = [0, 10]$ show that the finite horizon optimal solutions always end up at $x = 0$, i.e., at the end of the optimization horizon all money is spent (which is natural). However, for longer horizons the solutions spend quite some time in the vicinity of the point $x^e \approx 2.2344$ and the MPC closed-loop (9.3) converges to an equilibrium near this point. Further tests reveal that the limit point of the MPC closed-loop itself converges as $N \to \infty$.

There are many questions which arise from this behaviour: Why does the MPC closed-loop converge to a point far away from the endpoint of the finite horizon optimal trajectories? How do we characterize this point and its limit for $N \to \infty$? Is the MPC closed-loop trajectory approximately optimal in some sense? And how can we check whether an optimal control problem has such a behavior?

The third example is a simple partial differential equation control system governed by the 1d heat equation on $\Omega = (0, L)$. We consider the equation either with distributed control

$$\begin{aligned} y_t(x,t) &= y_{xx}(x,t) + \mu y(x,t) + \hat{u}(x,t) && \text{on } \Omega \times (0,\infty) \\ y(0,t) &= y(L,t) \;=\; 0 && \text{on } (0,\infty) \\ y(x,0) &= y_0(x) && \text{on } \Omega \end{aligned}$$

or with boundary control.

$$\begin{aligned} y_t(x,t) &= y_{xx}(x,t) + \mu y(x,t) && \text{on } \Omega \times (0,\infty) \\ y(0,t) &= 0, \;\; y(L,t) \;=\; \hat{u}(t) && \text{on } (0,\infty) \\ y(x,0) &= y_0(x) && \text{on } \Omega \end{aligned}$$

We set $\mu = 15$, which implies that $y \equiv 0$ is an unstable equilibrium for $u \equiv 0$. In order to stabilize this equilibrium, we consider the cost functions $\ell(y,u) = \|y\|_{L^2}^2 + \lambda\|u\|^2$ ("L^2-cost") and $\ell(y,u) = \|y_x\|_{L^2}^2 + \lambda\|u\|^2$ ("∇-cost"). As usual in MPC, it depends on the length of the horizon N whether the equilibrium $y \equiv 0$ is indeed stable. The simulations — all with sampling time $T = 0.01$ — show that depending on the parameters L and λ as well as on the type of the cost the minimal horizon length needed for stabilization differs significantly. This immediately leads to the question how we can estimate this minimal horizon length and whether we can tune, e.g., the stage cost ℓ such that this horizon becomes small.

[3] In practice, the state $x(j)$ must be computed from sensor data using a suitable observed, as, e.g., the Kalman filter or variants thereof. Also, in practice the MPC problem is initialized with the state $x(j-1)$ such that the time span until time j can be fully used in order to solve the optimal control problem. Both aspects will be neglected in the analysis of MPC schemes we will present in this lecture.

As we will see later, in all these examples we can prove that MPC yields approximately optimal infinite horizon trajectories. Hence, the problem on (rather short) finite horizons already contains enough information to compute near optimal solutions on an infinite horizon, a property that can be seen as a complexity reduction technique in time. In the subsequent analysis, we will in particular investigate the mechanisms behind this complexity reduction.

Kapitel 10

Stability of discrete time nonlinear systems

10.1 Stability definitions

In the introduction, we already specified one of the goals of model predictive control, namely to control the state $x(n)$ of the system toward a reference point x_* and then keep it close to this point. In this section we formalize what we mean by "toward" and "close to" using concepts from stability theory of nonlinear systems. These concepts will also turn out to be useful for the analysis of MPC schemes in which ℓ does not penalize the distance to an equilibrium x_*.

We assume that the states $x(k)$ are generated by a difference equation of the form

$$x^+ = g(x) \tag{10.1}$$

for a not necessarily continuous map $g : X \to X$ via the usual iteration $x(k+1) = g(x(k))$. Similar to before, we write $x(k, x_0)$ for the trajectory satisfying the initial condition $x(0, x_0) = x_0 \in X$. Allowing g to be discontinuous is important for our MPC application, because g will later represent the MPC closed-loop system (9.3), i.e., $g(x) = f(x, \mu(x))$. Since μ is obtained as an outcome of an optimization algorithm, in general we cannot expect μ to be continuous and thus g will in general be discontinuous, too.

Nonlinear stability properties can be expressed conveniently via so-called comparison functions which were first introduced by Hahn in 1967 [10] and popularized in nonlinear control theory during the 1990s by Sontag, particularly in the context of input-to-state stability [17]. Although we mainly deal with discrete time systems, we stick to the usual continuous time definition of these functions using the notation $\mathbb{R}_0^+ = [0, \infty)$.

Definition 10.1 [Comparison functions] We define the following classes of comparison functions.

$$\mathcal{K} := \{\alpha : \mathbb{R}_0^+ \to \mathbb{R}_0^+ \,|\, \alpha \text{ is continuous \& strictly increasing with } \alpha(0) = 0\}$$

$$\mathcal{K}_\infty := \{\alpha : \mathbb{R}_0^+ \to \mathbb{R}_0^+ \,|\, \alpha \in \mathcal{K},\ \alpha \text{ is unbounded}\}$$

$$\mathcal{L} := \{\delta : \mathbb{R}_0^+ \to \mathbb{R}_0^+ \,|\, \delta \text{ is continuous \& strictly decreasing with } \lim_{t\to\infty} \delta(t) = 0\}$$

$$\mathcal{KL} := \{\beta : \mathbb{R}_0^+ \times \mathbb{R}_0^+ \to \mathbb{R}_0^+ \,|\, \beta \text{ is continuous}, \beta(\cdot, t) \in \mathcal{K}\, \forall t \geq 0,\ \beta(r, \cdot) \in \mathcal{L}\, \forall r > 0\}.$$

□

Using this function, we can now introduce the concept of asymptotic stability. Here, for arbitrary $x_1, x_2 \in X$ we denote the distance from x_1 to x_2 by

$$|x_1|_{x_2} := d_X(x_1, x_2).$$

Furthermore, we use the ball

$$\mathcal{B}_\eta(x_*) := \{x \in X \,|\, |x|_{x_*} < \eta\}$$

and we say that a set $Y \subseteq X$ is *forward invariant* for (10.1) if $g(x) \in Y$ holds for all $x \in Y$.

Definition 10.2 [Asymptotic stability] Let $x_* \in X$ be an equilibrium for (10.1), i.e., $g(x_*) = x_*$. Then we say that x_* is *locally asymptotically stable* if there exist $\eta > 0$ and a function $\beta \in \mathcal{KL}$ such that the inequality

$$|x(n, x_0)|_{x_*} \leq \beta(|x_0|_{x_*}, n) \tag{10.2}$$

holds for all $x_0 \in \mathcal{B}_\eta(x_*)$ and all $n \in \mathbb{N}_0$.

We say that x_* is *asymptotically stable on a forward invariant set* Y with $x_* \in Y$ if there exists $\beta \in \mathcal{KL}$ such that (10.2) holds for all $x_0 \in Y$ and all $n \in \mathbb{N}_0$ and we say that x_* is *globally asymptotically stable* if x_* is asymptotically stable on $Y = X$.

If one of these properties holds then β is called *attraction rate*. □

Note that asymptotic stability on a forward invariant set Y implies local asymptotic stability if Y contains a ball $\mathcal{B}_\eta(x_*)$. However, we do not necessarily require this property.

Asymptotic stability thus defined consists of two main ingredients:

(i) The smaller the initial distance from x_0 to x_* is, the smaller the distance from $x(n)$ to x_* becomes for all future n, or formally: for each $\varepsilon > 0$ there exists $\delta > 0$ such that $|x(n, x_0)|_{x_*} \leq \varepsilon$ holds for all $n \in \mathbb{N}_0$ and all $x_0 \in Y$ (or $x_0 \in \mathcal{B}_\eta(x_*)$) with $|x_0|_{x_*} \leq \delta$.

This fact is easily seen by choosing δ so small that $\beta(\delta, 0) \leq \varepsilon$ holds, which is possible since $\beta(\cdot, 0) \in \mathcal{K}$. Since β is decreasing in its second argument, for $|x_0|_{x_*} \leq \delta$ from (10.2) we obtain

$$|x(n, x_0)|_{x_*} \leq \beta(|x_0|_{x_*}, n) \leq \beta(|x_0|_{x_*}, 0) \leq \beta(\delta, 0) \leq \varepsilon.$$

(ii) As the system evolves, the distance from $x(n, x_0)$ to x_* becomes arbitrarily small, or formally: for each $\varepsilon > 0$ and each $R > 0$ there exists $N > 0$ such that $|x(n, x_0)|_{x_*} \leq \varepsilon$ holds for all $n \geq N$ and all $x_0 \in Y$ (or $x_0 \in \mathcal{B}_\eta(x_*)$) with $|x_0|_{x_*} \leq R$. This property easily follows from (10.2) by choosing $N > 0$ with $\beta(R, N) \leq \varepsilon$ and exploiting the monotonicity properties of β.

These two properties are known as (i) stability (in the sense of Lyapunov) and (ii) attraction. In the literature, asymptotic stability is often defined via these two properties. In fact, for continuous time (and continuous) systems (i) and (ii) are known to be equivalent to the continuous time counterpart of Definition 10.2, cf. [13, Sect. 3]. We conjecture that the arguments in this reference can be modified in order to prove that equivalence also holds for our discontinuous discrete time setting.

Asymptotic stability includes the desired properties of the MPC closed loop described earlier: whenever we are already close to the reference equilibrium we want to stay close; otherwise we want to move toward the equilibrium.

Asymptotic stability also includes that eventually the distance of the closed-loop solution to the equilibrium x_* becomes arbitrarily small. Occasionally, this may be too demanding. For instance, we will see that in general we cannot expect this behavior for stage costs ℓ which do not penalize the distance to x_*. In this case, one can relax the asymptotic stability definition to practical asymptotic stability as follows. Here we only consider the case of asymptotic stability on a forward invariant set Y.

Definition 10.3 [P-practically asymptotic stability] Let Y be a forward invariant set and let $P \subset Y$ be a subset of Y. Then we say that a point $x_* \in Y$ is *P-practically asymptotically stable on Y* if there exists $\beta \in \mathcal{KL}$ such that (10.2) holds for all $x_0 \in Y$ and all $n \in \mathbb{N}_0$ with $x(n, x_0) \notin P$. □

Fig. 10.1 illustrates practical asymptotic stability (on the right) as opposed to "usual" asymptotic stability (on the left).

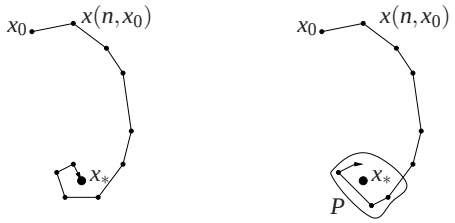


Abbildung 10.1: Sketch of asymptotic stability (left) as opposed to practical asymptotic stability (right)

This definition is typically used with P contained in a small ball around the equilibrium, i.e., $P \subseteq \mathcal{B}_\delta(x_*)$ for some small $\delta > 0$. In this case one obtains the estimate

$$|x(n, x_0)|_{x_*} \leq \max\{\beta(|x_0|_{x_*}, n),\, \delta\} \tag{10.3}$$

for all $x_0 \in Y$ and all $n \in \mathbb{N}_0$, i.e., the system behaves like an asymptotically stable system until it reaches the ball $\mathcal{B}_\delta(x_*)$. Note that x_* does not need to be an equilibrium in Definition 10.3.

10.2 Lyapunov functions

In order to verify that our MPC controller achieves asymptotic stability we will utilize the concept of Lyapunov functions.

Definition 10.4 [Lyapunov function] Consider a system (10.1), a point $x_* \in X$ and let $S \subseteq X$ be a subset of the state space. A function $V : S \to \mathbb{R}_0^+$ is called a *Lyapunov function* on S if the following conditions are satisfied:

(i) There exist functions $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ such that

$$\alpha_1(|x|_{x_*}) \leq V(x) \leq \alpha_2(|x|_{x_*}) \tag{10.4}$$

holds for all $x \in S$.

(ii) There exists a function $\alpha_V \in \mathcal{K}$ such that

$$V(g(x)) \leq V(x) - \alpha_V(|x|_{x_*}) \tag{10.5}$$

holds for all $x \in S$ with $g(x) \in S$.

□

The following theorem shows that the existence of a Lyapunov function ensures asymptotic stability.

Theorem 10.5 [Asymptotic stability using Lyapunov functions] Let x_* be an equilibrium of (10.1) and assume there exists a Lyapunov function V on S. If S contains a ball $\mathcal{B}_\nu(x_*)$ with $g(x) \in S$ for all $x \in \mathcal{B}_\nu(x_*)$ then x_* is locally asymptotically stable with $\eta = \alpha_2^{-1} \circ \alpha_1(\nu)$. If $S = Y$ holds for some forward invariant set $Y \subseteq X$ containing x_* then x_* is asymptotically stable on Y. If $S = X$ holds then x_* is globally asymptotically stable.

Proof: The idea of the proof lies in showing that by (10.5) the function $V(x(n, x_0))$ is strictly decreasing in n and converges to 0. Then by (10.4) we can conclude that $x(n, x_0)$ converges to x_*. The function β from Definition 10.2 will be constructed from α_1, α_2 and α_V. In order to simplify the notation, throughout the proof we write $|x|$ instead of $|x|_{x_*}$.

First, if S is not forward invariant, define the value $\gamma := \alpha_1(\nu)$ and the set $\widetilde{S} := \{x \in S \,|\, V(x) < \gamma\}$. Then from (10.4) we get

$$x \in \widetilde{S} \Rightarrow \alpha_1(|x|) \leq V(x) < \gamma \Rightarrow |x| < \alpha_1^{-1}(\gamma) = \nu \Rightarrow x \in \mathcal{B}_\nu(x_*),$$

observing that each $\alpha \in \mathcal{K}_\infty$ is invertible with $\alpha^{-1} \in \mathcal{K}_\infty$.

Hence, for each $x \in \widetilde{S}$ inequality (10.5) applies and consequently $V(g(x)) \leq V(x) < \gamma$ implying $g(x) \in \widetilde{S}$. If $S = Y$ for some forward invariant set $Y \subseteq X$ we define $\widetilde{S} := S$. With these definitions, in both cases the set $\widetilde{S}$ becomes forward invariant.

Now we define $\alpha'_V := \alpha_V \circ \alpha_2^{-1}$. Note that concatenations of $\mathcal{K}$-functions are again in $\mathcal{K}$, hence $\alpha'_V \in \mathcal{K}$. Since $|x| \geq \alpha_2^{-1}(V(x))$, using monotonicity of α_V this definition implies

$$\alpha_V(|x|) \geq \alpha_V \circ \alpha_2^{-1}(V(x)) = \alpha'_V(V(x)).$$

Hence, along a trajectory $x(n, x_0)$ with $x_0 \in \widetilde{S}$, from (10.5) we get the inequality

$$V(x(n+1, x_0)) \leq V(x(n, x_0)) - \alpha_V(|x(n, x_0)|) \leq V(x(n, x_0)) - \alpha'_V(V(x(n, x_0))). \quad (10.6)$$

For the construction of β we need the last expression in (10.6) to be strictly increasing in $V(x(n, x_0))$. To this end we define

$$\tilde{\alpha}_V(r) := \min_{s \in [0,r]} \{\alpha'_V(s) + (r-s)/2\}.$$

Straightforward computations show that this function satisfies $r_2 - \tilde{\alpha}_V(r_2) > r_1 - \tilde{\alpha}_V(r_1) \geq 0$ for all $r_2 > r_1 \geq 0$ and $\min\{\alpha'_V(r/2), r/4\} \leq \tilde{\alpha}_V(r) \leq \alpha'_V(r)$ for all $r \geq 0$. In particular, (10.6) remains valid and we get the desired monotonicity when α'_V is replaced by $\tilde{\alpha}_V$.

We inductively define a function $\beta_1 : \mathbb{R}_0^+ \times \mathbb{N}_0 \to \mathbb{R}_0^+$ via

$$\beta_1(r, 0) := r, \quad \beta_1(r, n+1) = \beta_1(r, n) - \tilde{\alpha}_V(\beta_1(r, n)). \quad (10.7)$$

By induction over n using the properties of $\tilde{\alpha}_V(r)$ and Inequality (10.6) one easily verifies the following inequalities:

$$\beta_1(r_2, n) > \beta_1(r_1, n) \geq 0 \text{ for all } r_2 > r_1 \geq 0 \text{ and all } n \in \mathbb{N}_0 \quad (10.8)$$
$$\beta_1(r, n_1) > \beta_1(r, n_2) > 0 \text{ for all } n_2 > n_1 \geq 0 \text{ and all } r > 0 \quad (10.9)$$
$$V(x(n, x_0)) \leq \beta_1(V(x_0), n) \text{ for all } n \in \mathbb{N}_0 \text{ and all } x_0 \in \widetilde{S} \quad (10.10)$$

From (10.9) it follows that $\beta_1(r, n)$ is monotone decreasing in n and by (10.8) it is bounded from below by 0. Hence, for each $r \geq 0$ the limit $\beta_1^\infty(r) = \lim_{n\to\infty} \beta_1(r, n)$ exists. We claim that $\beta_1^\infty(r) = 0$ holds for all r. Indeed, convergence implies $\beta_1(r, n) - \beta_1(r, n+1) \to 0$ as $n \to \infty$ which together with (10.7) yields $\tilde{\alpha}_V(\beta_1(r, n)) \to 0$. On the other hand, since $\tilde{\alpha}_V$ is continuous, we get $\tilde{\alpha}_V(\beta_1(r, n)) \to \tilde{\alpha}_V(\beta_1^\infty(r))$. This implies

$$\tilde{\alpha}_V(\beta_1^\infty(r)) = 0$$

which because of $\tilde{\alpha}_V(r) \geq \min\{\alpha_V(r/2), r/4\}$ and $\alpha_V \in \mathcal{K}$ is only possible if $\beta_1^\infty(r) = 0$.

Consequently, $\beta_1(r, n)$ has all properties of a $\mathcal{KL}$ function except that it is only defined for $n \in \mathbb{N}_0$. Defining the linear interpolation

$$\beta_2(r, t) := (n + 1 - t)\beta_1(r, n) + (t - n)\beta_1(r, n+1)$$

for $t \in [n, n+1)$ and $n \in \mathbb{N}_0$, we obtain a function $\beta_2 \in \mathcal{KL}$ which coincides with β_1 for $t = n \in \mathbb{N}_0$. Finally, setting

$$\beta(r, t) := \alpha_1^{-1} \circ \beta_2(\alpha_2(r), t)$$

we can use (10.10) in order to obtain

$$\begin{aligned} |x(n,x_0)| &\leq \alpha_1^{-1}(V(x(n,x_0))) \leq \alpha_1^{-1} \circ \beta_1(V(x_0),n) \\ &= \alpha_1^{-1} \circ \beta_2(V(x_0),n) \leq \alpha_1^{-1} \circ \beta_2(\alpha_2(|x_0|,n) = \beta(|x_0|,n), \end{aligned}$$

for all $x_0 \in \widetilde{S}$ and all $n \in \mathbb{N}_0$. This is the desired inequality (10.2). If $\widetilde{S} = S = Y$ this shows the claimed asymptotic stability on Y and global asymptotic stability if $Y = X$. If $\widetilde{S} \neq S$, then in order to satisfy the local version of Definition 10.2 it remains to show that $x \in \mathcal{B}_\eta(x_*)$ implies $x \in \widetilde{S}$. Since by definition of η and γ we have $\eta = \alpha_2^{-1}(\gamma)$, we get

$$x \in \mathcal{B}_\eta(x_*) \Rightarrow |x| < \eta = \alpha_2^{-1}(\gamma) \Rightarrow V(x) \leq \alpha_2(|x|) < \gamma \Rightarrow x \in \widetilde{S}.$$

This finishes the proof. □

Likewise, P-practical asymptotic stability can be ensured by a suitable Lyapunov function condition provided the set P is forward invariant.

Theorem 10.6 [P-practical asymptotic stability]
Consider forward invariant sets Y and $P \subset Y$ and a point $x_* \in P$. If there exists a Lyapunov function V on $S = Y \setminus P$ then x_* is P-practically asymptotically stable on Y.

Proof: The same construction of β as in the proof of Theorem 10.5 yields

$$|x(n,x_0)|_{x_*} \leq \beta(|x|_{x_*},n) \tag{10.2}$$

for all $n = 0,\ldots,n^*-1$, where $n^* \in \mathbb{N}_0$ is minimal with $x(n^*,x_0) \in P$. This follows with the same arguments as in the proof of Theorem 10.5 by restricting the times considered in (10.6) and (10.10) to $n = 0,\ldots,n^*-2$ and $n = 0,\ldots,n^*-1$, respectively.

Since forward invariance of P ensures $x(n,x_0) \in P$ for all $n \geq n^*$, the times n for which $x(n,x_0) \notin P$ holds are exactly $n = 0,\ldots,n^*-1$. Since these are exactly the times at which (10.2) is required, this yields the desired P-practical asymptotic stability. □

For continuous time systems $\dot{x} = g(x)$ all the concepts introduced in this section can be carried over directly. Particularly, the definitions of asymptotic and P-practical asymptotic stability are identical. In the definition of Lyapunov functions, condition (10.4) stays the same while condition (10.5) becomes

$$V(x(t,x_0)) \leq V(x_0) - \int_0^t \alpha_V(|x(t,x_0)|_{x_*}).$$

This is equivalent to

$$\frac{V(x(t,x_0)) - V(x_0)}{t} \leq -\frac{1}{t}\int_0^t \alpha_V(|x(t,x_0)|_{x_*})$$

and if V is continuously differentiable, then by letting $t \to 0$ one obtains the equivalent characterization

$$DV(x_0)g(x_0) \leq -\alpha_V(|x_0|_{x_*}). \tag{10.11}$$

Now it is obvious that this concept generalizes Definition 3.8, which we used in the linear case. With this definition of a Lyapunov function, all results in this section remain valid in the continuous time case.

Kapitel 11

Model predictive control schemes

11.1 The MPC algorithm without terminal conditions

We start this chapter by formulating the basic MPC algorithm already sketched in Chapter 9 in a more rigorous way. Here, the stage cost $\ell : X \times U \to \mathbb{R}$ is a general function. In the case of sampled data systems we can take the continuous time nature of the underlying model into account by defining the stage cost ℓ as an integral over a continuous time running cost function $L : X \times U \to \mathbb{R}_0^+$ on a sampling interval. Using the continuous time solution $\hat{x}$ from (8.5), we can define

$$\ell(x, u) := \int_0^T L(\hat{x}(t, x, u), u(t))dt. \tag{11.1}$$

Defining ℓ this way, we can incorporate the intersampling behavior of the sampled data system, i.e., the behavior of the continuous time solution between two sampling times t_k and t_{k+1}, explicitly into our optimal control problem.

Given such a cost function ℓ and a prediction horizon length $N \geq 2$, we can now formulate the basic MPC scheme as an algorithm. In the optimal control problem ($\mathrm{OCP_N}$) within this algorithm we introduce a set of control sequences $\mathbb{U}^N(x_0) \subseteq U^N$ over which we optimize. This set may include constraints depending on the initial value x_0. Details about how this set should be chosen will be discussed in Sect. 11.2. For the moment we simply set $\mathbb{U}^N(x_0) := U^N$ for all $x_0 \in X$.

Algorithm 11.1 (Basic MPC algorithm)

At each time instant $j = 0, 1, 2 \ldots$:

(1) Measure the state $x(j) \in X$ of the system

(2) Set $x_0 := x(j)$, solve the optimal control problem

$$\begin{aligned} &\text{minimize} \quad J_N(x_0, u(\cdot)) := \sum_{k=0}^{N-1} \ell(x_u(k, x_0), u(k)) \\ &\text{with respect to} \quad u(\cdot) \in \mathbb{U}^N(x_0), \quad \text{subject to} \\ &x_u(0, x_0) = x_0, \quad x_u(k+1, x_0) = f(x_u(k, x_0), u(k)) \end{aligned} \tag{$\mathrm{OCP_N}$}$$

and denote the obtained optimal control sequence by $u^\star(\cdot) \in \mathbb{U}^N(x_0)$.

(3) Define the MPC-feedback value $\mu_N(x(j)) := u^\star(0) \in U$ and use this control value in the next sampling period.

□

Observe that in this algorithm we have assumed that an optimal control sequence $u^\star(\cdot)$ exists. Sufficient conditions for this existence are briefly discussed after Definition 12.1, below.

The MPC closed loop system resulting from Algorithm 11.1 is given by (9.3) with state feedback law $\mu = \mu_N$, i.e.,

$$x^+ = f(x, \mu_N(x)). \tag{11.2}$$

The trajectories of this system will be denoted by $x_{\mu_N}(n)$ or, if we want to emphasize the initial value $x_0 = x_{\mu_N}(0)$, by $x_{\mu_N}(n, x_0)$.

During our theoretical investigations we will neglect the fact that computing the solution of ($\mathrm{OCP_N}$) in Step (2) of the algorithm usually needs some computation time τ_c which — in the case when τ_c is relatively large compared to the sampling period T — may not be negligible in a real time implementation.

In our abstract formulations of the MPC Algorithm 11.1 only the first element $u^\star(0)$ of the respective minimizing control sequence is used in each step, the remaining entries $u^\star(1), \ldots, u^\star(N-1)$ are discarded. In the practical implementation, however, these entries play an important role because numerical optimization algorithms for solving ($\mathrm{OCP_N}$) (or its variants) usually work iteratively: starting from an initial guess $u^0(\cdot)$ an optimization algorithm computes iterates $u^i(\cdot)$, $i = 1, 2, \ldots$ converging to the minimizer $u^\star(\cdot)$ and a good choice of $u^0(\cdot)$ is crucial in order to obtain fast convergence of this iteration, or even to ensure convergence, at all. Here, the minimizing sequence from the previous time step can be efficiently used in order to construct such a good initial guess. Ways to implement this idea will be discussed in the excercises.

11.2 Constraints

One of the main reasons for the success of MPC (and MPC in general) is its ability to explicitly take constraints into account. Here, we consider constraints both on the control as well as on the state. To this end, we introduce a nonempty *state constraint set* $\mathbb{X} \subseteq X$ and for each $x \in \mathbb{X}$ we introduce a nonempty *control constraint set* $\mathbb{U}(x) \subseteq U$. Of course,

$\mathbb{U}$ may also be chosen independent of x. The idea behind introducing these sets is that we want the trajectories to lie in $\mathbb{X}$ and the corresponding control values to lie in $\mathbb{U}(x)$. This is made precise in the following definition.

Definition 11.2 [Admissibility] Consider a control system (8.2) and the state and control constraint sets $\mathbb{X} \subseteq X$ and $\mathbb{U}(x) \subseteq U$.

(i) The states $x \in \mathbb{X}$ are called *admissible states* and the control values $u \in \mathbb{U}(x)$ are called *admissible control values for* x. The elements of the set $\mathbb{Y} := \{(x,u) \in X \times U \,|\, x \in \mathbb{X}, u \in \mathbb{U}(x)\}$ are called *admissible pairs*.

(ii) For $N \in \mathbb{N}$ and an initial value $x_0 \in \mathbb{X}$ we call a control sequence $u \in U^N$ and the corresponding trajectory $x_u(k, x_0)$ *admissible for* x_0 *up to time* N, if

$$(x_u(k,x_0), u(k)) \in \mathbb{Y} \text{ for all } k = 0, \ldots, N-1 \quad \text{and} \quad x_u(N, x_0) \in \mathbb{X}$$

holds. We denote the set of admissible control sequences for x_0 up to time N by $\mathbb{U}^N(x_0)$.

(iii) A control sequence $u \in U^\infty$ and the corresponding trajectory $x_u(k, x_0)$ are called *admissible for* x_0 if they are admissible for x_0 up to every time $N \in \mathbb{N}$. We denote the set of admissible control sequences for x_0 by $\mathbb{U}^\infty(x_0)$.

(iv) A feedback law $\mu : X \to U$ is called *admissible* if $\mu(x) \in \mathbb{U}^1(x)$ holds for all $x \in \mathbb{X}$.

Whenever the reference to x or x_0 is clear from the context we will omit the additional "for x" or "for x_0". □

Since we can (and will) identify control sequences with only one element with the respective control value, we can consider $\mathbb{U}^1(x_0)$ as a subset of U, which we already implicitly did in the definition of admissibility for the feedback law μ, above. However, in general $\mathbb{U}^1(x_0)$ does not coincide with $\mathbb{U}(x_0) \subseteq U$ because using $x_u(1,x) = f(x,u)$ and the definition of $\mathbb{U}^N(x_0)$ we get $\mathbb{U}^1(x) := \{u \in \mathbb{U}(x) \mid f(x,u) \in \mathbb{X}\}$. With this subtle difference in mind, one sees that our admissibility condition (iv) on μ ensures both $\mu(n,x) \in \mathbb{U}(x)$ and $f(x, \mu(n,x)) \in \mathbb{X}$ whenever $x \in \mathbb{X}$.

Furthermore, our definition of $\mathbb{U}^N(x)$ implies that even if $\mathbb{U}(x) = \mathbb{U}$ is independent of x the set $\mathbb{U}^N(x)$ may depend on x for some or all $N \in \mathbb{N}_\infty$.

Often, in order to be suitable for optimization purposes these sets are assumed to be compact and convex. For our theoretical investigations, however, we do not need any regularity requirements of this type except that these sets are nonempty.

MPC is well suited to handle constraints because these can directly be inserted into Algorithm 11.1. In fact, since we already formulated the corresponding optimization problem ($\mathrm{OCP_N}$) with state dependent control value sets, the constraints are readily included if we use $\mathbb{U}^N(x_0)$ from Definition 11.2(ii) in ($\mathrm{OCP_N}$). However, when doing so we have to make sure that the constraints in ($\mathrm{OCP_N}$) can be satisfied for all j, i.e., that we do not optimize over an empty set because $\mathbb{U}^N(x_0) = \emptyset$. This is formalized in the following definition.

Definition 11.3 (i) An initial condition $x_0 \in \mathbb{X}$ is called *feasible* for ($\mathrm{OCP_N}$) if the constraints imposed in ($\mathrm{OCP_N}$) can be satisfied, i.e, if $\mathbb{U}^N(x_0) \neq \emptyset$.

(ii) A MPC algorithm 11.1 is called *recursively feasible* on a set $A \subseteq \mathbb{X}$ if each $x \in A$ is feasible for ($\mathrm{OCP_N}$) and $x \in A$ implies $f(x, \mu_N(x)) \in A$ (implying that $f(x, \mu_N(x))$ is again feasible). □

One easily sees that recursive feasibility implies that $x_{\mu_N}(j)$ is feasible for all $j \in \mathbb{N}$ if $x_{\mu_N}(0) \in A$. In order to ensure recursive feasibility of $A = \mathbb{X}$ for Algorithm 11.1, we need the following assumption.

Assumption 11.4 [Viability] For each $x \in \mathbb{X}$ there exists $u \in \mathbb{U}(x)$ such that $f(x, u) \in \mathbb{X}$ holds. □

The property defined in this assumption is called *viability* or *weak (or controlled) forward invariance* of $\mathbb{X}$. It excludes the situation that there are states $x \in \mathbb{X}$ from which the trajectory leaves the set $\mathbb{X}$ for all admissible control values. Hence, it ensures $\mathbb{U}^N(x_0) \neq \emptyset$ for all $x_0 \in \mathbb{X}$ and all $N \in \mathbb{N}_\infty$. Thus, it ensures that any $x_0 \in \mathbb{X}$ is feasible for ($\mathrm{OCP_N}$) and hence ensures that $\mu_N(x)$ is well defined for each $x \in \mathbb{X}$. We will see after the next example that viability of $\mathbb{X}$ also implies recursive feasibility and admissibility of the closed loop. Furthermore, a straightforward induction shows that under Assumption 11.4 any finite admissible control sequence $u(\cdot) \in \mathbb{U}^N(x_0)$ can be extended to an infinite admissible control sequence $\tilde{u}(\cdot) \in \mathbb{U}^\infty(x_0)$ with $u(k) = \tilde{u}(k)$ for all $k = 0, \ldots, N-1$.

In order to see that the construction of a constraint set $\mathbb{X}$ meeting Assumption 11.4 is usually a nontrivial task, we consider the following Example.

Example 11.5 Consider

$$x^+ = f(x, u) = \begin{pmatrix} x_1 + x_2 + u/2 \\ x_2 + u \end{pmatrix},$$

which can be seen as a sampled-data model for a car on a one-dimensional road with position x_1, speed x_2 and piecewise constant acceleration u. Assume we want to constrain all variables, i.e., the position x_1, the velocity x_2 and the acceleration u to the interval $[-1, 1]$. For this purpose one could define $\mathbb{X} = [-1, 1]^2$ and $\mathbb{U}(x) = \mathbb{U} = [-1, 1]$. Then, however, for $x = (1, 1)^\top$, one immediately obtains

$$x_1^+ = x_1 + x_2 + u/2 = 2 + u/2 \geq 3/2$$

for all u, hence $x^+ \notin \mathbb{X}$ for all $u \in \mathbb{U}$. Thus, in order to find a viable set $\mathbb{X}$ we need to either tighten or relax some of the constraints. For instance, relaxing the constraint on u to $\mathbb{U} = [-2, 2]$ the viability of $\mathbb{X} = [-1, 1]^2$ is guaranteed, because then by elementary computations one sees that for each $x \in \mathbb{X}$ the control value

$$u = \begin{cases} 0, & x_1 + x_2 \in [-1, 1] \\ 2 - 2x_1 - 2x_2, & x_1 + x_2 > 1 \\ -2 - 2x_1 - 2x_2, & x_1 + x_2 < -1 \end{cases}$$

is in $\mathbb{U}$ and satisfies $f(x, u) \in \mathbb{X}$. A way to achieve viability without changing $\mathbb{U}$ is by tightening the constraint on x_2 by defining

$$\mathbb{X} = \{(x_1, x_2)^T \in \mathbb{R}^2 \mid x_1 \in [-1, 1], x_2 \in [-1, 1] \cap [-3/2 - x_1, 3/2 - x_1]\}, \tag{11.3}$$

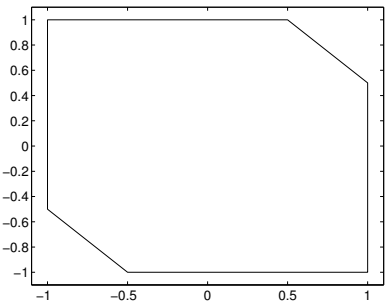


Abbildung 11.1: Illustration of the set $\mathbb{X}$ from (11.3)

see Fig. 11.5. Again, elementary computations show that for each $x \in \mathbb{X}$ and

$$u = \begin{cases} 1, & x_2 < -1/2 \\ -2x_2, & x_2 \in [-1/2, 1/2] \\ -1, & x_2 > 1/2 \end{cases}$$

the desired properties $u \in \mathbb{U}$ and $f(x,u) \in \mathbb{X}$ hold. □

This example shows that finding viable constraint sets $\mathbb{X}$ (and the corresponding $\mathbb{U}$ or $\mathbb{U}(x)$) is a tricky task already for very simple systems. Still, Assumption 11.4 significantly simplifies the subsequent analysis, cf. Theorem 11.6, below. For this reason we will impose this condition in our theoretical investigations for schemes without stabilizing terminal conditions. The assumption can be avoided if suitable terminal constraints are employed. We will discuss this extension of the scheme in Section 11.3.

The following theorem shows that the viability assumption ensures recursive feasibility of Algorithm 11.1 and that the resulting MPC closed loop satisfies the desired constraints.

Theorem 11.6 [Recursive Feasibility and Admissibility] Consider Algorithm 11.1 using $\mathbb{U}^N(x_0)$ from Def. 11.2(ii) in the optimal control problem ($\mathrm{OCP_N}$) for constraint sets $\mathbb{X} \subset X$, $\mathbb{U}(x) \subset U$, $x \in \mathbb{X}$, satisfying Assumption 11.4. Consider the MPC closed loop system (11.2). Then the MPC algorithm is recursively feasible on $A = \mathbb{X}$ and for any $x_{\mu_N}(0) \in \mathbb{X}$ the constraints are satisfied along the solution of (11.2), i.e.,

$$(x_{\mu_N}(n), \mu_N(x_{\mu_N}(n))) \in \mathbb{Y} \tag{11.4}$$

for all $n \in \mathbb{N}$. Thus, the MPC-feedback μ_N is admissible in the sense of Definition 11.2(iv).

Proof: First, recall from the discussion after Assumption 11.4 that under this assumption the optimal control problem ($\mathrm{OCP_N}$) is feasible for each $x \in \mathbb{X}$, hence $\mu_N(x)$ is well defined for each $x \in \mathbb{X}$.

We now show that $x_{\mu_N}(n) \in \mathbb{X}$ implies $\mu_N(x_{\mu_N}(n)) \in \mathbb{U}(x_{\mu_N}(n))$ and $x_{\mu_N}(n+1) \in \mathbb{X}$. This implies recursive feasibility of $A = \mathbb{X}$, and admissibility follows by induction from $x_{\mu_n}(0) \in \mathbb{X}$.

The viability of $\mathbb{X}$ from Assumption 11.4 ensures that whenever $x_{\mu_N}(n) \in \mathbb{X}$ holds in Algorithm 11.1 then $x_0 \in \mathbb{X}$ is feasible for the respective optimal control problem ($\mathrm{OCP_N}$). Since the optimization is performed with respect to admissible control sequences only, also the optimal control sequence $u^\star(\cdot)$ is admissible for $x_0 = x_{\mu_N}(n)$. This implies $\mu_N(x_{\mu_N}(n)) = u^\star(0) \in \mathbb{U}^1(x_{\mu_N}(n)) \subseteq \mathbb{U}(x_{\mu_N}(n))$ and thus also

$$x_{\mu_N}(n+1) = f(x_{\mu_N}(n), \mu_N(x_{\mu_N}(n))) = f(x(n), u^\star(0)) \in \mathbb{X},$$

i.e., $x_{\mu_N}(n+1) \in \mathbb{X}$. □

In the underlying optimization algorithms for solving ($\mathrm{OCP_N}$), usually the constraints cannot be specified via sets $\mathbb{X}$ and $\mathbb{U}(x)$. Rather, one uses so-called *equality* and *inequality constraints* in order to specify $\mathbb{X}$ and $\mathbb{U}(x)$ according to the following definition.

Definition 11.7 Given functions $G_i^S : X \times U \to \mathbb{R}$, $i \in \mathcal{E}^S = \{1, \dots, p_g\}$ and $H_i^S : X \times U \to \mathbb{R}$, $i \in \mathcal{I}^S = \{p_g + 1, \dots, p_g + p_h\}$ with $p_g, p_h \in \mathbb{N}_0$, we define the constraint sets $\mathbb{X}$ and $\mathbb{U}(x)$ via

$$\mathbb{X} := \left\{ x \in X \;\middle|\; \begin{array}{l} \text{there exists } u \in U \text{ with } G_i^S(x,u) = 0 \text{ for all } i \in \mathcal{E}^S \\ \text{and } H_i^S(x,u) \geq 0 \text{ for all } i \in \mathcal{I}^S \end{array} \right\}$$

and, for $x \in \mathbb{X}$

$$\mathbb{U}(x) := \left\{ u \in U \;\middle|\; \begin{array}{l} G_i^S(x,u) = 0 \text{ for all } i \in \mathcal{E}^S \text{ and} \\ H_i^S(x,u) \geq 0 \text{ for all } i \in \mathcal{I}^S \end{array} \right\}$$

Here, the functions G_i^S and H_i^S do not need to depend on both arguments. The functions G_i^S, H_i^S not depending on u are called *pure state constraints*, the functions G_i^S, H_i^S not depending on x are called *pure control constraints* and the functions G_i^S, H_i^S depending on both x and u are called *mixed constraints*. □

Observe that if we do not have mixed constraints then $\mathbb{U}(x)$ is independent of x.

The reason for defining $\mathbb{X}$ and $\mathbb{U}(x)$ via these (in)equality constraints is purely algorithmic: the plain information "$x_u(k, x_0) \notin \mathbb{X}$" does not yield any information for the optimization algorithm in order to figure out how to find an admissible $u(\cdot)$, i.e., a $u(\cdot)$ for which "$x_u(k, x_0) \in \mathbb{X}$" holds. In contrast to that, an information of the form "$H_i^S(x_u(k, x_0), u(k)) < 0$" together with additional knowledge about H_i^S (provided, e.g., by the derivative of H_i^S) enables the algorithm to compute a "direction" in which $u(\cdot)$ needs to be modified in order to reach an admissible $u(\cdot)$.

In our theoretical investigations we will use the notationally more convenient set characterization of the constraints via $\mathbb{X}$ and $\mathbb{U}(x)$ or $\mathbb{U}^N(x)$. In the practical implementation of our MPC method, however, we will use their characterization via the inequality constraints from Definition 11.7.

11.3 The MPC algorithm with terminal conditions

In this section we discuss an important variant of the basic MPC Algorithm 11.1. This algorithm adds a constraint on the terminal state $x_u(N, x_0)$ of the trajectory over which we

optimize in ($\mathrm{OCP_N}$), as well as a weight on this term. This combination of constraint and weight on the terminal state is called *terminal conditions*. As we will see, under suitable assumptions on the terminal conditions, the behavior of the MPC closed-loop can significantly improve. The main disadvantage of terminal condition is that a rigorous derivation of a constraint and a weight meeting these assumptions can be very dificult for complex control systems.

The terminal constraint is of the form

$$x_u(N, x_0) \in \mathbb{X}_0 \text{ for a } \textit{terminal constraint set } \mathbb{X}_0 \subseteq \mathbb{X}. \tag{11.5}$$

Of course, in the practical implementation the constraint set $\mathbb{X}_0$ is again expressed via (in)equalities of the form given in Definition 11.7.

When using terminal constraints, the MPC-feedback law is only defined for those states x_0 for which the optimization problem within the MPC algorithm is feasible also for these additional constraints, i.e., for which there exists an admissible control sequence with corresponding trajectory starting in x_0 and ending in the terminal constraint set. Such initial values are again called *feasible* and the set of all feasible initial values form the feasible set. This set along with the corresponding admissible control sequences is formally defined as follows.

Definition 11.8 [Feasible set and admissible control sequences]
For $\mathbb{X}_0$ from (11.5) we define the *feasible set* for horizon $N \in \mathbb{N}$ by

$$\mathbb{X}_N := \{x_0 \in \mathbb{X} \mid \text{there exists } u(\cdot) \in \mathbb{U}^N(x_0) \text{ with } x_u(N, x_0) \in \mathbb{X}_0\}$$

and for each $x_0 \in \mathbb{X}_N$ we define the set of *admissible control sequences* by

$$\mathbb{U}^N_{\mathbb{X}_0}(x_0) := \{u(\cdot) \in \mathbb{U}^N(x_0) \mid x_u(N, x_0) \in \mathbb{X}_0\}.$$

□

Note that in $\mathbb{X}_N = \mathbb{X}$ and $\mathbb{U}^N_{\mathbb{X}_0}(x) = \mathbb{U}^N(x)$ holds if $\mathbb{X}_0 = \mathbb{X}$, i.e., if no additional terminal constraints are imposed.

The additional weight on the terminal state $x_u(N)$ is formalized by means of a terminal cost of the form $F(x_u(N, x_0))$ with $F : \mathbb{X}_0 \to \mathbb{R}$ in the optimization objective.

Together this leads to the following MPC algorithms extending the basic Algorithms 11.1. Note that compared to these basic algorithms only the optimal control problems are different, i.e., the part in the boxes in Step (2).

Algorithm 11.9 (MPC algorithm with terminal conditions)

At each time instant $j = 0, 1, 2 \ldots$:

(1) Measure the state $x(j) \in X$ of the system.

(2) Set $x_0 := x(j)$, solve the optimal control problem

$$\boxed{\begin{aligned} &\text{minimize} \quad J_N(x_0, u(\cdot)) := \sum_{k=0}^{N-1} \ell(x_u(k, x_0), u(k)) + F(x_u(N, x_0)) \\ &\text{with respect to} \quad u(\cdot) \in \mathbb{U}_{\mathbb{X}_0}^N(x_0), \quad \text{subject to} \\ &x_u(0, x_0) = x_0, \quad x_u(k+1, x_0) = f(x_u(k, x_0), u(k)) \end{aligned}} \tag{OCP$_{\mathrm{N,e}}$}$$

and denote the obtained optimal control sequence by $u^\star(\cdot) \in \mathbb{U}_{\mathbb{X}_0}^N(x_0)$.

(3) Define the MPC-feedback value $\mu_N(x(j)) := u^\star(0) \in U$ and use this control value in the next sampling period.

□

We end this section with three useful results on the sets of admissible control sequences from Definition 11.8.

Lemma 11.10 Let $x_0 \in \mathbb{X}_N$, $N \in \mathbb{N}$ and $K \in \{0, \dots, N\}$ be given.

(i) For each $u(\cdot) \in \mathbb{U}_{\mathbb{X}_0}^N(x_0)$ we have $x_u(K, x_0) \in \mathbb{X}_{N-K}$.

(ii) For each $u(\cdot) \in \mathbb{U}_{\mathbb{X}_0}^N(x_0)$ the control sequences $u_1 \in U^K$ and $u_2 \in U^{N-K}$ uniquely defined by the relation

$$u(k) = \begin{cases} u_1(k), & k = 0, \dots, K-1 \\ u_2(k-K), & k = K, \dots, N-1 \end{cases} \tag{11.6}$$

satisfy $u_1 \in \mathbb{U}_{\mathbb{X}_{N-K}}^K(x_0)$ and $u_2 \in \mathbb{U}_{\mathbb{X}_0}^{N-K}(x_{u_1}(K, x_0))$.

(iii) For each $u_1(\cdot) \in \mathbb{U}_{\mathbb{X}_{N-K}}^K(x_0)$ there exists $u_2(\cdot) \in \mathbb{U}_{\mathbb{X}_0}^{N-K}(x_{u_1}(K, x_0))$ such that $u(\cdot)$ from (11.6) satisfies $u \in \mathbb{U}_{\mathbb{X}_0}^N(x_0)$.

Proof: (i) Using (8.4) we obtain the identity

$$x_{u(K+\cdot)}(N-K, x_u(K, x_0))) = x_u(N, x_0) \in \mathbb{X}_0,$$

which together with the definition of $\mathbb{X}_{N-K}$ implies the assertion.

(ii) The relation (11.6) together with (8.4) implies

$$x_u(k, x_0) = \begin{cases} x_{u_1}(k, x_0), & k = 0, \dots, K \\ x_{u_2}(k-K, x_{u_1}(K, x_0)), & k = K, \dots, N \end{cases} \tag{11.7}$$

For $k = 0, \dots, K-1$ this identity and (11.6) yield

$$u_1(k) = u(k) \in \mathbb{U}(x_u(k, x_0)) = \mathbb{U}(x_{u_1}(k, x_0))$$

and for $k = 0, \dots, N-K-1$ we obtain

$$u_2(k) = u(k+K) \in \mathbb{U}(x_u(k+K, x_0)) = \mathbb{U}(x_{u_2}(k, x_{u_1}(K, x_0))),$$

implying $u_1 \in \mathbb{U}^K(x_0)$ and $u_2 \in \mathbb{U}^{N-K}(x_{u_1}(K, x_0))$. Furthermore, (11.7) implies the equation $x_{u_2}(N-K, x_{u_1}(K, x_0)) = x_u(N, x_0) \in \mathbb{X}_0$ which proves $u_2 \in \mathbb{U}^{N-K}_{\mathbb{X}_0}(n+K, x_{u_1}(K, x_0))$. This, in turn, implies that $\mathbb{U}^{N-K}_{\mathbb{X}_0}(n+K, x_{u_1}(K, x_0))$ is nonempty, hence $x_{u_1}(K, x_0) \in \mathbb{X}_{N-K}$ and consequently $u_1 \in \mathbb{U}^K_{\mathbb{X}_{N-K}}(n, x_0)$ follows.

(iii) By definition, for each $x \in \mathbb{X}_{N-K}(n+K)$ there exists $u_2 \in \mathbb{U}^{N-K}_{\mathbb{X}_0}(n+K, x)$. Choosing such a u_2 for $x = x_{u_1}(K, x_0) \in \mathbb{X}_{N-K}(n+K)$ and defining u via (11.6), similar arguments as in Part (ii), above, show the claim $u \in \mathbb{U}^N_{\mathbb{X}_0}(n, x_0)$. □

A straightforward corollary of this lemma is the following.

Corollary 11.11 For each $x \in \mathbb{X}_N$ the MPC-feedback law μ_N obtained from Algorithm 11.9 satisfies

$$f(x, \mu_N(x)) \in \mathbb{X}_{N-1}.$$

□

Proof: Since $\mu_N(x)$ is the first element $u^\star(0)$ of the optimal control sequence $u^\star \in \mathbb{U}^N_{\mathbb{X}_0}(x)$ we get $f(x, \mu_N(x)) = x_{u^\star}(1, x)$. Now Lemma 11.10(i) yields the assertion. □

The final result shows that with terminal conditions we can obtain Theorem 11.6 without having to assume viability of $\mathbb{X}$ — if in exchange we assume viability of the terminal constraint set $\mathbb{X}_0$.

Theorem 11.12 [Recursive Feasibility and Admissibility] Consider Algorithm 11.9 for constraint sets $\mathbb{X} \subset X$, $\mathbb{U}(x) \subset U$, $x \in \mathbb{X}$, and a terminal constraint set $\mathbb{X}_0$ which satisfies Assumption 11.4. Consider the MPC closed loop system (11.2). Then the MPC algorithm is recursively feasible on $A = \mathbb{X}_N$ and for $x_{\mu_N}(0) \in \mathbb{X}_N$ the constraints are satisfied along the solution of (11.2), i.e.,

$$(x_{\mu_N}(n), \mu_N(x_{\mu_N}(n))) \in \mathbb{Y} \tag{11.8}$$

for all $n \in \mathbb{N}$. Thus, the MPC-feedback μ_N is admissible in the sense of Definition 11.2(iv).

Proof: We show that under the viability assumption on $\mathbb{X}_0$ the inclusion $\mathbb{X}_{N-1} \subseteq \mathbb{X}_N$ holds. Then recursive feasibility follows from Corollary 11.11 and admissibility follows as in the proof of Theorem 11.6.

In order to show the inclusion $\mathbb{X}_{N-1} \subseteq \mathbb{X}_N$, consider $x \in \mathbb{X}_{N-1}$. Then there is an admissible control $u \in \mathbb{U}^{N-1}_{\mathbb{X}_0}(x)$, implying $x_u(N-1, x) \in \mathbb{X}_0$. Viability of $\mathbb{X}_0$ implies the existence of a control value $\tilde{u} \in \mathbb{U}(x_u(N-1, x))$ with $f(x_u(N-1, x), \tilde{u}) \in \mathbb{X}_0$. This implies that the control sequence

$$\hat{u} = (u(0), \dots, u(N-1), \tilde{u})$$

is admissible and satisfies $x_{\hat{u}}(N, x) = f(x_u(N-1, x), \tilde{u}) \in \mathbb{X}_N$. This implies $x \in \mathbb{X}_N$ and thus the desired inclusion. □

Kapitel 12

Dynamic programming

This chapter repeats and extends some of the results from Section 6.1. As we will see, dynamic programming is not only important for deriving the Riccati equation but also as a basis for analyzing MPC schemes in the next chapters. We first consider finite horizon problems and then discuss infinite horizon problems.

12.1 Finite horizon problems

In this section we provide one of the classical tools in optimal control, the *dynamic programming principle*. We will formulate and prove the results in this section for ($\mathrm{OCP_{N,e}}$), since all other optimal control problems introduced above can be obtained as special cases of this problem. We will first formulate the principle for the open loop control sequences in ($\mathrm{OCP_{N,e}}$) and then derive consequences for the MPC-feedback law μ_N. The dynamic programming principle is often used as a basis for numerical algorithms. In contrast to this, here we will exclusively use the principle for analyzing the behavior of MPC closed loop systems. The reason for this is that the numerical effort of solving ($\mathrm{OCP_{N,e}}$) via dynamic programming usually grows exponentially with the dimension of the state of the system. In contrast to this, the computational effort of solving a single problem of type ($\mathrm{OCP_N}$) or ($\mathrm{OCP_{N,e}}$) scales much more moderately with the space dimension.

We start by defining some objects we need in the sequel.

Definition 12.1 Consider the optimal control problem ($\mathrm{OCP_{N,e}}$) with initial value $x_0 \in \mathbb{X}$ and optimization horizon $N \in \mathbb{N}_0$.

(i) The function

$$V_N(x_0) := \inf_{u(\cdot) \in \mathbb{U}_{\mathbb{X}_0}^N(x_0)} J_N(x_0, u(\cdot))$$

is called *optimal value function*.

(ii) A control sequence $u^\star(\cdot) \in \mathbb{U}_{\mathbb{X}_0}^N(x_0)$ is called *optimal control sequence* for x_0, if

$$V_N(x_0) = J_N(x_0, u^\star(\cdot))$$

holds. The corresponding trajectory $x_{u^\star}(\cdot, x_0)$ is called *optimal trajectory*.

□

In our MPC Algorithms 11.1 and 11.9 we have assumed that an optimal control sequence $u^\star(\cdot)$ exists, cf. the comment after Algorithms 11.1. In general, this is not necessarily the case but under reasonable continuity and compactness conditions the existence of $u^\star(\cdot)$ can be rigorously shown. Examples of such theorems for a general infinite-dimensional state space can be found in Keerthi and Gilbert [12] or Doležal [3]. While for formulating and proving the dynamic programming principle we will not need the existence of $u^\star(\cdot)$, for all subsequent results we will assume that $u^\star(\cdot)$ exists, in particular when we derive properties of the MPC-feedback law μ_N. While we conjecture that most of the subsequent results in this lecture notes can be generalized to the case when μ_N is defined via an approximately minimizing control sequence, we decided to use the existence assumption because it considerably simplifies the presentation of the results in these lecture notes.

The following theorem introduces the *dynamic programming principle*. It gives an equation which relates the optimal value functions for different optimization horizons N and for different points in space.

Theorem 12.2 [Dynamic programming principle] Consider the optimal control problem ($\mathrm{OCP}_{\mathrm{N,e}}$) with $x_0 \in \mathbb{X}_N$ and $N \in \mathbb{N}_0$. Then for all $N \in \mathbb{N}$ and all $K = 1, \ldots, N$ the equation

$$V_N(x_0) = \inf_{u(\cdot) \in \mathbb{U}^K_{\mathbb{X}_{N-K}}(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k, x_0), u(k)) + V_{N-K}(x_u(K, x_0)) \right\} \tag{12.1}$$

holds. If, in addition, an optimal control sequence $u^\star(\cdot) \in \mathbb{U}^N_{\mathbb{X}_0}(x_0)$ exists for x_0, then we get the equation

$$V_N(x_0) = \sum_{k=0}^{K-1} \ell(x_{u^\star}(k, x_0), u^\star(k)) + V_{N-K}(x_{u^\star}(K, x_0)). \tag{12.2}$$

In particular, in this case the "inf" in (12.1) is a "min".

Proof: First observe that from the definition of J_N for $u(\cdot) \in \mathbb{U}^N_{\mathbb{X}_0}(x_0)$ we immediately obtain

$$J_N(x_0, u(\cdot)) = \sum_{k=0}^{K-1} \ell(x_u(k, x_0), u(k)) + J_{N-K}(x_u(K, x_0), u(\cdot + K)). \tag{12.3}$$

Since $u(\cdot + K)$ equals $u_2(\cdot)$ from Lemma 11.10(ii) we obtain $u(\cdot + K) \in \mathbb{U}^{N-K}_{\mathbb{X}_0}(x_u(K, x_0))$.

We now prove (12.1) by proving "$\geq$" and "$\leq$" separately. From (12.3) we obtain

$$\begin{aligned} J_N(x_0, u(\cdot)) &= \sum_{k=0}^{K-1} \ell(x_u(k, x_0), u(k)) \\ &\qquad + \; J_{N-K}(x_u(K, x_0), u(\cdot + K)) \\ &\geq \sum_{k=0}^{K-1} \ell(x_u(k, x_0), u(k)) + V_{N-K}(x_u(K, x_0)). \end{aligned}$$

Since this inequality holds for all $u(\cdot) \in \mathbb{U}_{\mathbb{X}_0}^N(x_0)$, it also holds when taking the infimum on both sides. Hence we get

$$\begin{aligned} V_N(x_0) &= \inf_{u(\cdot) \in \mathbb{U}_{\mathbb{X}_0}^N(x_0)} J_N(x_0, u(\cdot)) \\ &\geq \inf_{u(\cdot) \in \mathbb{U}_{\mathbb{X}_0}^N(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k, x_0), u(k)) \right. \\ &\qquad\qquad \left. + \; V_{N-K}(x_u(K, x_0)) \right\} \\ &= \inf_{u_1(\cdot) \in \mathbb{U}_{\mathbb{X}_{N-K}}^K(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_{u_1}(k, x_0), u(k)) \right. \\ &\qquad\qquad \left. + \; V_{N-K}(x_{u_1}(K, x_0)) \right\}, \end{aligned}$$

i.e., (12.1) with "$\geq$". Here in the last step we used the fact that by Lemma 11.10(ii) the control sequence u_1 consisting of the first K elements of $u(\cdot) \in \mathbb{U}_{\mathbb{X}_0}^N(x_0)$ lies in $\mathbb{U}_{\mathbb{X}_{N-K}}^K(x_0)$ and, conversely, by Lemma 11.10(iii) each control sequence in $u_1(\cdot) \in \mathbb{U}_{\mathbb{X}_{N-K}}^K(x_0)$ can be extended to a sequence in $u(\cdot) \in \mathbb{U}_{\mathbb{X}_0}^N(x_0)$. Thus, since the expression in braces does not depend on $u(K), \ldots, u(N-1)$, the infima coincide.

In order to prove "$\leq$", fix $\varepsilon > 0$ and let $u^\varepsilon(\cdot)$ be an approximately optimal control sequence for the right hand side of (12.3), i.e.,

$$\begin{aligned} &\sum_{k=0}^{K-1} \ell(x_{u^\varepsilon}(k, x_0), u^\varepsilon(k)) + J_{N-K}(x_{u^\varepsilon}(K, x_0), u^\varepsilon(\cdot + K)) \\ &\leq \inf_{u(\cdot) \in \mathbb{U}_{\mathbb{X}_0}^N(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k, x_0), u(k)) \right. \\ &\qquad\qquad \left. + \; J_{N-K}(x_u(K, x_0), u(\cdot + K)) \right\} + \varepsilon. \end{aligned}$$

Now we use the decomposition (11.6) of $u(\cdot)$ into $u_1 \in \mathbb{U}_{\mathbb{X}_{N-K}}^K(x_0)$ and $u_2 \in \mathbb{U}_{\mathbb{X}_0}^{N-K}(x_{u_1}(K, x_0))$

from Lemma 11.10(ii). This way we can proceed

$$\begin{aligned}
&\inf_{u(\cdot)\in\mathbb{U}^N_{\mathbb{X}_0}(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k,x_0),u(k)) \right. \\
&\qquad\qquad \left. + \; J_{N-K}(x_u(K,x_0),u(\cdot+K)) \right\} \\
&= \inf_{\substack{u_1(\cdot)\in\mathbb{U}^K_{\mathbb{X}_{N-K}}(x_0) \\ u_2(\cdot)\in\mathbb{U}^{N-K}_{\mathbb{X}_0}(x_{u_1}(K,x_0))}} \left\{ \sum_{k=0}^{K-1} \ell(x_{u_1}(k,x_0),u_1(k)) \right. \\
&\qquad\qquad \left. + \; J_{N-K}(x_{u_1}(K,x_0),u_2(\cdot)) \right\} \\
&= \inf_{u_1(\cdot)\in\mathbb{U}^K_{\mathbb{X}_{N-K}}(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_{u_1}(k,x_0),u_1(k)) \right. \\
&\qquad\qquad \left. + \; V_{N-K}(x_{u_1}(K,x_0)) \right\}
\end{aligned}$$

Now (12.3) yields

$$\begin{aligned}
V_N(x_0) &\le J(x_0,u^\varepsilon(\cdot)) \\
&= \sum_{k=0}^{K-1} \ell(x_{u^\varepsilon}(k,x_0),u^\varepsilon(k)) + J_{N-K}(x_{u^\varepsilon}(K,x_0),u^\varepsilon(\cdot+K)) \\
&\le \inf_{u(\cdot)\in\mathbb{U}^K_{\mathbb{X}_{N-K}}(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k,x_0),u(k)) \right. \\
&\qquad\qquad \left. + \; V_{N-K}(x_u(K,x_0)) \right\} + \varepsilon.
\end{aligned}$$

Since the first and the last term in this inequality chain are independent of ε and since $\varepsilon > 0$ was arbitrary, this shows (12.1) with "$\le$" and thus (12.1).

In order to prove (12.2) we use (12.3) with $u(\cdot) = u^\star(\cdot)$. This yields

$$\begin{aligned}
V_N(x_0) &= J(x_0,u^\star(\cdot)) \\
&= \sum_{k=0}^{K-1} \ell(x_{u^\star}(k,x_0),u^\star(k)) + J_{N-K}(x_{u^\star}(K,x_0),u^\star(\cdot+K)) \\
&\ge \sum_{k=0}^{K-1} \ell(x_{u^\star}(k,x_0),u^\star(k)) + V_{N-K}(x_{u^\star}(K,x_0)) \\
&\ge \inf_{u(\cdot)\in\mathbb{U}^K_{\mathbb{X}_{N-K}}(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k,x_0),u(k)) + V_{N-K}(x_u(K,x_0)) \right\} \\
&= V_N(x_0),
\end{aligned}$$

where we used the (already proven) equality (12.1) in the last step. Hence, the two "$\geq$" in this chain are actually "=" which implies (12.2). □

The following corollary states an immediate consequence of the dynamic programming principle. It shows that tails of optimal control sequences are again optimal control sequences for suitably adjusted optimization horizon, time instant and initial value.

Corollary 12.3 If $u^\star(\cdot)$ is an optimal control sequence for initial value $x_0 \in \mathbb{X}_N$ and optimization horizon $N \geq 2$, then for each $K = 1, \ldots, N-1$ the sequence $u^\star_K(\cdot) = u^\star(\cdot + K)$, i.e.,

$$u^\star_K(k) = u^\star(K + k), \quad k = 0, \ldots, N - K - 1$$

is an optimal control sequence for initial value $x_{u^\star}(K, x_0)$, time instant K and optimization horizon $N - K$. □

Proof: Inserting $V_N(x_0) = J_N(x_0, u^\star(\cdot))$ and the definition of $u^\star_k(\cdot)$ into (12.3) we obtain

$$V_N(x_0) = \sum_{k=0}^{K-1} \ell(x_{u^\star}(k, x_0), u^\star(k)) + J_{N-K}(x_{u^\star}(K, x_0), u^\star_K(\cdot))$$

Subtracting (12.2) from this equation yields

$$0 = J_{N-K}(x_{u^\star}(K, x_0), u^\star_K(\cdot)) - V_{N-K}(x_{u^\star}(K, x_0))$$

which shows the assertion. □

The next theorem relates the MPC-feedback law μ_N defined in the MPC Algorithms 11.1 and 11.9 to the dynamic programming principle. Here we use the argmin operator in the following sense: for a map $a : U \to \mathbb{R}$, a nonempty subset $\widetilde{U} \subseteq U$ and a value $u^\star \in \widetilde{U}$ we write

$$u^\star = \operatorname*{argmin}_{u \in \widetilde{U}} a(u) \tag{12.4}$$

if and only if $a(u^\star) = \inf_{u \in \widetilde{U}} a(u)$ holds. Whenever (12.4) holds the existence of the minimum $\min_{u \in \widetilde{U}} a(u)$ follows. However, we do not require uniqueness of the minimizer $u^\star$. In case of uniqueness equation (12.4) can be understood as an assignment, otherwise it is just a convenient way of writing "$u^\star$ minimizes $a(u)$".

Theorem 12.4 [Dynamic programming and MPC] Consider the optimal control problem ($\mathrm{OCP}_{\mathrm{N,e}}$) with $x_0 \in \mathbb{X}_N$ and $N \in \mathbb{N}_0$ and an admissible feedback law $\mu : \mathbb{X} \to U$ in the sense of Definition 11.2(iv). Then μ satisfies

$$\mu(x_0) = \operatorname*{argmin}_{u \in \mathbb{U}^1_{\mathbb{X}_{N-1}}(x_0)} \{\ell(x_0, u) + V_{N-1}(f(x_0, u))\} \tag{12.5}$$

if and only if μ satisfies

$$V_N(x_0) = \ell(x_0, \mu(x_0)) + V_{N-1}(f(x_0, \mu(x_0))), \tag{12.6}$$

where in (12.5) we interpret $\mathbb{U}^1_{\mathbb{X}_{N-1}}(x_0)$ as a subset of U, i.e., we identify the one element sequence $u = u(\cdot)$ with its only element $u = u(0)$. Moreover, if an optimal control sequence $u^\star$ exists then the MPC-feedback law $\mu(x_0) = \mu_N(x_0) = u^*(0)$ satisfies both (12.5) and (12.6).

Proof: Equation (12.6) follows from (12.5) by using (12.1) for $K = 1$ and the minimizing property of μ.

Conversely, assume (12.6). Inserting $x_u(1, x_0) = f(x_0, u)$ into the dynamic programming principle (12.1) for $K = 1$ we obtain

$$V_N(x_0) = \inf_{u \in \mathbb{U}^1_{\mathbb{X}_{N-1}}(x_0)} \{\ell(x_0, u) + V_{N-1}(1, f(x_0, u))\}. \tag{12.7}$$

This implies that the right hand sides of (12.6) and (12.7) coincide. Thus, the definition of argmin in (12.4) with $a(u) = \ell(x_0, u) + V_{N-1}(1, f(x_0, u))$ and $\widetilde{U} = \mathbb{U}^1_{\mathbb{X}_{N-1}}(x_0)$ yields (12.5).

Finally, if $u^\star$ exists, then (12.6) (and thus also (12.5)) follows fur $\mu = \mu_n$ from the existence by inserting $u^\star(0) = \mu_N(x_0)$ and $x_{u^\star}(1, x_0) = f(x_0, \mu_N(x_0))$ into (12.2) for $K = 1$. □

Our final corollary in this section shows that we can reconstruct the whole optimal control sequence $u^\star(\cdot)$ using the feedback from (12.5).

Corollary 12.5 Consider the optimal control problem ($\mathrm{OCP}_{\mathrm{N,e}}$) with $x_0 \in \mathbb{X}$ and $N \in \mathbb{N}_0$ and consider admissible feedback laws $\mu_{N-k} : \mathbb{X} \to U$, $k = 0, \ldots, N-1$, in the sense of Definition 11.2(iv). Denote the solution of the closed loop system

$$x(0) = x_0, \quad x(k+1) = f(x(k), \mu_{N-k}(x(k))), \ k = 0, \ldots, N-1 \tag{12.8}$$

by $x_\mu(\cdot)$ and assume that the μ_{N-k} satisfy (12.5) with horizon $N - k$ instead of N and initial value $x_0 = x_\mu(k)$ for $k = 0, \ldots, N-1$. Then

$$u^\star(k) = \mu_{N-k}(x_\mu(k)), \quad k = 0, \ldots, N-1 \tag{12.9}$$

is an optimal control sequence for initial value x_0 and the solution of the closed loop system (12.8) is a corresponding optimal trajectory. □

Proof: Applying the control (12.9) to the dynamics (12.8) we immediately obtain

$$x_{u^\star}(k) = x_\mu(k), \quad k = 0, \ldots, N-1.$$

Hence, we need to show that

$$V_N(x_0) = J_N(x_0, u^\star) = \sum_{k=0}^{N-1} \ell(x_\mu(k), u^\star(k)) + F(x(N)).$$

Using (12.9) and (12.6) for $N - k$ instead of N and $x_0 = x_\mu(k)$ we get

$$V_{N-k}(x_\mu(k)) = \ell(x_\mu(k), u^\star(k)) + V_{N-k-1}(x_\mu(k+1))$$

for $k = 0, \ldots, N-1$. Summing these equalities for $k = 0, \ldots, N-1$ and eliminating the identical terms $V_{N-k}(x_\mu(k))$, $k = 1, \ldots, N-1$ on both sides we obtain

$$V_N(x_0) = \sum_{k=0}^{N-1} \ell(x_\mu(k), u^\star(k)) + V_0(x(N))$$

Since by definition of J_0 we have $V_0(x) = F(x)$, this shows the assertion. □

12.2 Infinite horizon problems

In this section we present the counterparts of the result from the previous section for infinite horizon problems. These are defined by as follows.

$$\boxed{\begin{aligned} &\text{minimize} \quad J_\infty(x_0, u(\cdot)) := \limsup_{K\to\infty} \sum_{k=0}^{K-1} \ell(x_u(k, x_0), u(k)) \\ &\text{with respect to} \;\; u(\cdot) \in \mathbb{U}^\infty(x_0), \quad \text{subject to} \\ &x_u(0, x_0) = x_0, \quad x_u(k+1, x_0) = f(x_u(k, x_0), u(k)) \end{aligned}} \tag{OCP$_\infty$}$$

We assume that for all $x_0 \in \mathbb{X}$

$$-\infty < \inf_{u(\cdot)\in\mathbb{U}^\infty(x_0)} \liminf_{K\to\infty} \sum_{k=0}^{K-1} \ell(x_u(k, x_0), u(k)) = \inf_{u(\cdot)\in\mathbb{U}^\infty(x_0)} \limsup_{K\to\infty} \sum_{k=0}^{K-1} \ell(x_u(k, x_0), u(k)) < \infty,$$

which in particular implies that the optimal value function V_∞, as defined in the following definition, assumes finite values for all $x_0 \in \mathbb{X}$ and and that there is no admissible control sequence $\hat{u}(\cdot)$ for which

$$\liminf_{K\to\infty} \sum_{k=0}^{K-1} \ell(x_{\hat{u}}(k, x_0), \hat{u}(k)) < V_\infty(x_0)$$

holds for some $x_0 \in \mathbb{X}$.

Definition 12.6 Consider the optimal control problem (OCP$_\infty$) with initial value $x_0 \in \mathbb{X}$.

(i) The function
$$V_\infty(x_0) := \inf_{u(\cdot)\in\mathbb{U}^\infty(x_0)} J_\infty(x_0, u(\cdot))$$
is called *optimal value function*.

(ii) A control sequence $u^\star(\cdot) \in \mathbb{U}^\infty(x_0)$ is called *optimal control sequence* for x_0 if
$$V_\infty(x_0) = J_\infty(x_0, u^\star(\cdot))$$
holds. The corresponding trajectory $x_{u^\star}(\cdot, x_0)$ is called *optimal trajectory*.

□

The first result we state is the dynamic programming principle.

Theorem 12.7 [Dynamic programming principle] Consider the optimal control problem (OCP$_\infty$) with $x_0 \in \mathbb{X}$. Then for all $K \in \mathbb{N}$ the equation

$$V_\infty(x_0) = \inf_{u(\cdot)\in\mathbb{U}^K(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k, x_0), u(k)) + V_\infty(x_u(K, x_0)) \right\} \tag{12.10}$$

holds. If, in addition, an optimal control sequence $u^\star(\cdot)$ exists for x_0, then we get the equation

$$V_\infty(x_0) = \sum_{k=0}^{K-1} \ell(x_{u^\star}(k, x_0), u^\star(k)) + V_\infty(x_{u^\star}(K, x_0)). \tag{12.11}$$

In particular, in this case the "inf" in (12.10) is a "min".

Proof: From the definition of J_∞ for $u(\cdot) \in \mathbb{U}^\infty(x_0)$ we immediately obtain

$$J_\infty(x_0, u(\cdot)) = \sum_{k=0}^{K-1} \ell(x_u(k, x_0), u(k)) + J_\infty(x_u(K, x_0), u(\cdot + K)), \tag{12.12}$$

where $u(\cdot + K)$ denotes the shifted control sequence defined by $u(\cdot + K)(k) = u(k + K)$, which is admissible for $x_u(K, x_0)$.

We now prove (12.10) by showing "$\geq$" and "$\leq$" separately: From (12.12) we obtain

$$\begin{aligned} J_\infty(x_0, u(\cdot)) &= \sum_{k=0}^{K-1} \ell(x_u(k, x_0), u(k)) + J_\infty(x_u(K, x_0), u(\cdot + K)) \\ &\geq \sum_{k=0}^{K-1} \ell(x_u(k, x_0), u(k)) + V_\infty(x_u(K, x_0)). \end{aligned}$$

Since this inequality holds for all $u(\cdot) \in \mathbb{U}^\infty$, it also holds when taking the infimum on both sides. Hence we get

$$\begin{aligned} V_\infty(x_0) &= \inf_{u(\cdot) \in \mathbb{U}^\infty(x_0)} J_\infty(x_0, u(\cdot)) \\ &\geq \inf_{u(\cdot) \in \mathbb{U}^K(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k, x_0), u(k)) + V_\infty(x_u(K, x_0)) \right\}, \end{aligned}$$

i.e., (12.10) with "$\geq$".

In order to prove "$\leq$", fix $\varepsilon > 0$ and let $u^\varepsilon(\cdot)$ be an approximately optimal control sequence for the right hand side of (12.12), i.e.,

$$\begin{aligned} &\sum_{k=0}^{K-1} \ell(x_{u^\varepsilon}(k, x_0), u^\varepsilon(k)) + J_\infty(x_{u^\varepsilon}(K, x_0), u^\varepsilon(\cdot + K)) \\ &\leq \inf_{u(\cdot) \in \mathbb{U}^\infty(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k, x_0), u(k)) + J_\infty(x_u(K, x_0), u(\cdot + K)) \right\} + \varepsilon. \end{aligned}$$

Now we decompose $u(\cdot) \in \mathbb{U}^\infty(x_0)$ analogously to Lemma 11.10(ii) and (iii) into $u_1 \in \mathbb{U}^K(x_0)$ and $u_2 \in \mathbb{U}^\infty(x_{u_1}(K, x_0))$ via

$$u(k) = \begin{cases} u_1(k), & k = 0, \ldots, K-1 \\ u_2(k-K), & k \geq K \end{cases}$$

This implies

$$\begin{aligned}
&\inf_{u(\cdot)\in\mathbb{U}^\infty(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k,x_0),u(k)) + J_\infty(x_u(K,x_0),u(\cdot+K)) \right\} \\
&= \inf_{\substack{u_1(\cdot)\in\mathbb{U}^K(x_0) \\ u_2(\cdot)\in\mathbb{U}^\infty(x_{u_1}(K,x_0))}} \left\{ \sum_{k=0}^{K-1} \ell(x_{u_1}(k,x_0),u_1(k)) + J_\infty(x_{u_1}(K,x_0),u_2(\cdot)) \right\} \\
&= \inf_{u_1(\cdot)\in\mathbb{U}^K(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_{u_1}(k,x_0),u_1(k)) + V_\infty(x_{u_1}(K,x_0)) \right\}
\end{aligned}$$

Now (12.12) yields

$$\begin{aligned}
V_\infty(x_0) &\le J_\infty(x_0,u^\varepsilon(\cdot)) \\
&= \sum_{k=0}^{K-1} \ell(x_{u^\varepsilon}(k,x_0),u^\varepsilon(k)) + J_\infty(x_{u^\varepsilon}(K,x_0),u^\varepsilon(\cdot+K)) \\
&\le \inf_{u(\cdot)\in\mathbb{U}^K(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k,x_0),u(k)) + V_\infty(x_u(K,x_0)) \right\} + \varepsilon,
\end{aligned}$$

i.e.,

$$\begin{aligned}
&V_\infty(x_0) \\
&\le \inf_{u(\cdot)\in\mathbb{U}^K(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k,x_0),u(k)) + V_\infty(x_u(K,x_0)) \right\} + \varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary and the expressions in this inequality are independent of ε, this inequality also holds for $\varepsilon = 0$, which shows (12.10) with "$\le$" and thus (12.10).

In order to prove (12.11) we use (12.12) with $u(\cdot) = u^\star(\cdot)$. This yields

$$\begin{aligned}
V_\infty(x_0) &= J_\infty(x_0,u^\star(\cdot)) \\
&= \sum_{k=0}^{K-1} \ell(x_{u^\star}(k,x_0),u^\star(k)) + J_\infty(x_{u^\star}(x_0),u^\star(\cdot+K)) \\
&\ge \sum_{k=0}^{K-1} \ell(x_{u^\star}(k,x_0),u^\star(k)) + V_\infty(x_{u^\star}(K,x_0)) \\
&\ge \inf_{u(\cdot)\in\mathbb{U}^K(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k,x_0),u(k)) + V_\infty(x_u(K,x_0)) \right\} \\
&= V_\infty(x_0),
\end{aligned}$$

where we used the (already proved) equality (12.10) in the last step. Hence, the two "$\ge$" in this chain are actually "=" which implies (12.11). □

The following corollary states an immediate consequence from the dynamic programming principle. It shows that tails of optimal control sequences are again optimal control sequences for suitably adjusted initial value and time.

Corollary 12.8 If $u^\star(\cdot)$ is an optimal control sequence for (OCP_∞) with initial value x_0, then for each $K \in \mathbb{N}$ the sequence $u^\star_K(\cdot) = u^\star(\cdot + K)$, i.e.,

$$u^\star_K(k) = u^\star(K + k), \quad k = 0, 1, \ldots$$

is an optimal control sequence for initial value $x_{u^\star}(K, x_0)$ and initial time K. □

Proof: Inserting $V_\infty(x_0) = J_\infty(x_0, u^\star(\cdot))$ and the definition of $u^\star_K(\cdot)$ into (12.12) we obtain

$$V_\infty(x_0) = \sum_{k=0}^{K-1} \ell(x_{u^\star}(k, x_0), u^\star(k)) + J_\infty(x_{u^\star}(x_0), u^\star_K(\cdot))$$

Subtracting (12.11) from this equation yields

$$0 = J_\infty(x_{u^\star}(x_0), u^\star_K(\cdot)) - V_\infty(x_{u^\star}(K, x_0))$$

which shows the assertion. □

The next two results are the analogues of Theorem 12.4 and Corollary 12.5 in the infinite horizon setting.

Theorem 12.9 Consider the optimal control problem ($\mathrm{OCP}_{\mathrm{N,e}}$) with $x_0 \in \mathbb{X}_N$ and $N \in \mathbb{N}_0$ and an admissible feedback law $\mu : \mathbb{X} \to U$ in the sense of Definition 11.2(iv). Then μ satisfies

$$\mu(x_0) = \operatorname*{argmin}_{u \in \mathbb{U}^1_{\mathbb{X}_{N-1}}(x_0)} \{\ell(x_0, u) + V_\infty(f(x_0, u))\} \tag{12.13}$$

if and only if μ satisfies

$$V_\infty(x_0) = \ell(x_0, \mu(x_0)) + V_\infty(f(x_0, \mu(x_0))), \tag{12.14}$$

where again we interpret $\mathbb{U}^1_{\mathbb{X}_{N-1}}(x_0)$ as a subset of U. Moreover, if an optimal control sequence $u^\star$ exists then the MPC-feedback law $\mu(x_0) = \mu_\infty(x_0) = u^*(0)$ satisfies both (12.13) and (12.14).

Proof: The proof is identical to the finite horizon counterpart Theorem 12.4. □

As in the finite horizon case, the following corollary shows that the feedback law (12.13) can be used in order to construct the optimal control sequence.

Corollary 12.10 Consider the optimal control problem (OCP_∞). Let $x_0 \in \mathbb{X}$ and consider an admissible feedback law $\mu : \mathbb{X} \to U$ in the sense of Definition 11.2(iv). Denote the solution of the closed loop system

$$x(0) = x_0, \quad x(k+1) = f(x(k), \mu_\infty(x(k))), \ k = 0, 1, \ldots \tag{12.15}$$

by x_μ, assume that μ_∞ satisfies (12.13) for initial values $x_0 = x_\mu(k)$ for all $k = 0, 1, \ldots$ and that

$$\lim_{k \to \infty} V_\infty(x_\mu(k)) \geq 0.$$

Then

$$u^\star(k) = \mu_\infty(x_{u^\star}(k, x_0)), \quad k = 0, 1, \ldots \tag{12.16}$$

is an optimal control sequence for initial time n and initial value x_0 and the solution of the closed loop system (12.15) is a corresponding optimal trajectory. □

Proof: From (12.16) for $x(n)$ from (12.15) we immediately obtain

$$x_{u^\star}(k) = x(k), \quad k = 0, 1, \ldots.$$

Hence we need to show that

$$V_\infty(x_0) = J_\infty(x_0, u^\star).$$

Using (12.16) and (12.14) we get

$$V_\infty(x(k)) = \ell(x(k), u^\star(k)) + V_\infty(x(k+1))$$

for $k = 0, 1, \ldots$. Summing these equalities for $k = 0, \ldots, K-1$ for arbitrary $K \in \mathbb{N}$ and eliminating the identical terms $V_\infty(k, x_0)$, $k = 1, \ldots, K-1$ on the left and on the right we obtain

$$V_\infty(x_0) = \sum_{k=0}^{K-1} \ell(x(k), u^\star(k)) + V_\infty(x(K)).$$

Taking the upper limit for $K \to \infty$ and using $\lim_{k\to\infty} V(x_\mu(k)) \geq 0$ as well as the assumption on the lower limit after (OCP_∞) implies that

$$\limsup_{K\to\infty} \sum_{k=0}^{K-1} \ell(x(k), u^\star(k)) \leq V_\infty(x_0) \leq \liminf_{K\to\infty} \sum_{k=0}^{K-1} \ell(x(k), u^\star(k)),$$

which implies that the limit

$$\sum_{k=0}^{\infty} \ell(x(k), u^\star(k)) = \lim_{K\to\infty} \sum_{k=0}^{K-1} \ell(x(k), u^\star(k))$$

exists and equals $V_\infty(x_0)$. □

We note that the condition $\lim_{k\to\infty} V(x_\mu(k)) \geq 0$ is always satisfied when $\ell(x, u) \geq 0$ for all $x \in \mathbb{X}$, $u \in \mathbb{U}(x)$.

Corollary 12.10 implies that infinite horizon optimal control is nothing but MPC with $N = \infty$: Formula (12.16) for $k = 0$ yields that if we replace the optimization problem (OCP_N) in Algorithm 11.1 by (OCP_∞), then the feedback law resulting from this algorithm equals μ_∞. In fact, the infinite horizon problem can be seen as a discrete time nonlinear version of linear quadratic optimal control. Our last theorem (the only one that does not have a finite horizon counterpart in Section 12.1) shows that just like for linear quadratic optimal control, the optimal feedback law stabilizes an equilibrium, provided suitable inequalities are satisfied.

Theorem 12.11 [Asymptotic stability] Consider the optimal control problem (OCP_∞) for the control system (8.2) and an equilibrium $x_* \in \mathbb{X}$. Assume that there exist $\alpha_1, \alpha_2, \alpha_3 \in \mathcal{K}_\infty$ such that the inequalities

$$\alpha_1(|x|_{x_*}) \leq V_\infty(x) \leq \alpha_2(|x|_{x_*}) \quad \text{and} \quad \ell(x, u) \geq \alpha_3(|x|_{x_*}) \tag{12.17}$$

hold for all $x \in \mathbb{X}$ and $u \in U$. Assume furthermore that an optimal feedback μ_∞ exists, i.e., an admissible feedback law $\mu_\infty : \mathbb{X} \to U$ satisfying (12.13) for all $x \in \mathbb{X}$. Then this optimal feedback asymptotically stabilizes the closed loop system

$$x^+ = g(x) = f(x, \mu_\infty(x))$$

on $\mathbb{X}$ in the sense of Definition 10.2.

Proof: For the closed loop system, (12.14) and the last inequality in (12.17) yield

$$\begin{aligned} V_\infty(x) &= \ell(x, \mu_\infty(x)) + V_\infty(f(x, \mu_\infty(x))) \\ &\geq \alpha_3(|x|_{x_*}) + V_\infty(f(x, \mu_\infty(x))). \end{aligned}$$

Together with the first two inequalities in (12.17) this shows that V_∞ is a Lyapunov function on $\mathbb{X}$ in the sense of Definition 10.4 with $\alpha_V = \alpha_3$. Thus, Theorem 10.5 yields asymptotic stability on $\mathbb{X}$. □

Kapitel 13

Analysis of general MPC schemes

13.1 Preliminaries

In this section we analyze the properties of the MPC closed-loop (11.2) for "general" stage costs ℓ. Of course, it is easy to see that ℓ cannot be completely general. Some properties must be met in order to obtain good closed-loop behavior and one of the main tasks in this chapter will be to figure out what these properties are. In the literature, this class of MPC schemes is often called "economic" MPC, because in practice the stage cost often models some economic goal, like maximal yield or minimum energy consumption. The next example is a very simply optimal control problem which falls into the last class.

Example 13.1 An example, which will serve as an illustration for all results in this section, is the 1d discrete-time system with dynamics and stage cost

$$x^+ = 2x + u \quad \text{and} \quad \ell(x,u) = u^2$$

and state and control constraint sets $\mathbb{X} = [-2,2]$ and $\mathbb{U}(x) = \mathbb{U} = [-3,3]$, i.e., $\mathbb{Y} = [-2,2] \times [-3,3]$.

The uncontrolled system is unstable, hence for initial values $x_0 \neq 0$ the solution will leave the admissible set $\mathbb{X}$ if no control is used. Hence, control action is needed in order to keep the system inside $\mathbb{X}$. Interpreting the stage cost $\ell(x,u) = u^2$ as the energy of the current control action, the control objective can be formulated as "keep the state inside $\mathbb{X}$ with minimal control effort". □

In what follows, two aspects will be investigated: the qualitative property of the MPC closed-loop trajectory (as, e.g., stability) and its quantitative properties measured in terms of the stage cost function. For the second purpose, three different quantities can be considered:

The first quantity is the infinite horizon closed-loop performance

$$J_\infty^{cl}(x_0,\mu) := \sum_{k=0}^{\infty} \ell(x_\mu(k), \mu(x_\mu(k))).$$

This would be the "natural" measure if we consider MPC as an approximation to an infinite horizon problem. However, as the infinite sum may not converge, we also look at other measures. We also consider the finite horizon closed-loop performance

$$J_K^{cl}(x_0, \mu) := \sum_{k=0}^{K-1} \ell(x_\mu(k), \mu(x_\mu(k))) \tag{13.1}$$

and the averaged infinite horizon performance

$$\overline{J}_\infty^{cl}(x_0, \mu) := \limsup_{K\to\infty} \frac{1}{K} J_K^{cl}(x_0, \mu).$$

Throughout this chapter by $(x^e, u^e) \in \mathbb{Y}$ we denote an equilibrium of the system, i.e., $f(x^e, u^e) = x^e$. Of particular interest are optimal equilibria according to the following definition.

Definition 13.2 An equilibrium $(x^e, u^e) \in \mathbb{Y}$ is called an *optimal equilibrium* if it yields the lowest value of the cost function among all admissible equilibria, i.e.,

$$\ell(x^e, u^e) \le \ell(x, u) \qquad \text{for all } (x, u) \in \mathbb{Y} \text{ with } f(x, u) = x.$$

□

Example 13.3 In Example 13.1, the equilibria are of the form $(x, -x)$ with cost $\ell(x, -x) = x^2$. Thus, the (unique) optimal equilibrium is given by $(x^e, u^e) = (0, 0)$. □

The following lemma shows that an optimal equilibrium always exists when f and ℓ are continuous and $\mathbb{Y}$ is compact.

Lemma 13.4 If the constraint set $\mathbb{Y} \subset X \times U$ is compact, the maps $\ell : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ and $f : \mathbb{X} \times \mathbb{U} \to \mathbb{X}$ are continuous, and there exists an equilibrium in $\mathbb{Y}$, then there exists an optimal equilibrium, i.e., a pair $(x^e, u^e) \in \mathbb{Y}$ with $f(x^e, u^e) = x^e$ such that

$$\ell(x^e, u^e) = \inf\{\ell(x, u) \,|\, (x, u) \in \mathbb{Y}, f(x, u) = x\}.$$

Proof: Since pre-images of closed sets under continuous mappings are closed, the set $\{(x, u) \in \mathbb{Y} \,|\, f(x, u) = x\}$ is closed, hence compact, and nonempty. Thus, the continuous function ℓ attains a minimum. □

Hence, assuming the existence of an optimal equilibrium is not an overly restrictive assumption.

13.2 Averaged performance with terminal conditions

In this and in the following two sections we consider the MPC Algorithm 11.9 with optimal control problem ($\mathrm{OCP}_{\mathrm{N,e}}$). We note that the terminal condition is only added to the open-loop functional $J_N(x_0, u)$ used in the MPC Algorithm 11.9 but not to the closed-loop performance index $J_K^{cl}(x, \mu)$ from (13.1), which is still defined without terminal cost or constraints according to (13.1). As before, the optimal value function is defined by

$$V_N(x) := \inf_{u(\cdot)\in\mathbb{U}_{\mathbb{X}_0}^N(x)} J_N(x, u(\cdot))$$

and we assume the existence of an optimal control sequence for each feasible initial condition x in order to synthesize the MPC feedback law μ_N according to Algorithm 11.9.

The following assumption links an equilibrium—which will later be chosen as an optimal equilibrium—to the terminal conditions. For its formulation, recall the definition of the feasible sets $\mathbb{X}_N$ from Definition 11.8(i).

Assumption 13.5 [Terminal conditions] (a) The set $\mathbb{X}_0$ is bounded and there is an equilibrium $(x^e, u^e) \in \mathbb{Y}$ with $x^e \in \mathbb{X}_0$ and $F(x^e) = 0$ such that for each $x \in \mathbb{X}_0$ there exists $u \in \mathbb{U}$ with $f(x, u) \in \mathbb{X}_0$ and

$$F(f(x, u)) \le F(x) - \ell(x, u) + \ell(x^e, u^e)$$

(b) There exists $N_0 \in \mathbb{N}$ and $\eta > 0$ such that $\mathbb{X}_{N_0}$ contains the ball $\mathcal{B}_\eta(x^e)$. □

Condition (a) is a compatibility condition between the stage cost ℓ and the terminal cost F. The simplest way to satisfy condition (a) is by setting $\mathbb{X}_0 = \{x^e\}$ and $F \equiv 0$. However, using a terminal constraint set with only one point may cause convergence problems in the numerical optimization routine for solving ($\mathrm{OCP}_{\mathrm{N,e}}$). For ℓ with $\ell(x^e, u^e) = 0$ and $\ell(x, u) > 0$ otherwise, a systematic way to construct F with this property is via a linear quadratic approximation of the problem near x^e.

Observe that the requirement $F(x^e) = 0$ in Assumption 13.5(a) can be made without loss of generality because the inequality is invariant with respect to adding a constant to F. Assumption 13.5(b) is a nondegeneracy condition which prevents that the feasible sets $\mathbb{X}_N$ have empty interior for any $N \in \mathbb{N}$.

Under these assumptions we can formulate the first result.

Theorem 13.6 Consider the MPC Algorithm 11.9. Let Assumption 13.5(a) be satisfied, let $N \ge 2$ and assume V_N is bounded from below on $\mathbb{X}_N$. Then, for any $N \ge 2$ and any $x \in \mathbb{X}_N$ the averaged closed-loop performance satisfies the inequality

$$\overline{J}_\infty^{cl}(x, \mu_N) \le \ell(x^e, u^e). \tag{13.2}$$

Proof: Let $\hat{x} \in \mathbb{X}_{N-1}$ and let $\hat{u}^\star$ be the optimal control sequence for this initial value with horizon $N-1$, i.e., $V_{N-1}(\hat{x}) = J_{N-1}(\hat{x}, \hat{u}^\star)$. Let $\tilde{u}$ be the control value from Assumption 13.5(a) for $\tilde{x} = x_{\hat{u}^\star}(N-1, \hat{x})$. Then, for the control sequence $u = (\hat{u}^\star(0), \ldots, \hat{u}^\star(N-1), \tilde{u})$ we obtain $x_u(N, \hat{x}) = f(\tilde{x}, \tilde{u})$ and thus Assumption 13.5(a) implies

$$\begin{aligned} V_N(\hat{x}) &\leq J_N(\hat{x}, u) \\ &= J_{N-1}(\hat{x}, \hat{u}^\star) - F(\tilde{x}) + \ell(\tilde{x}, \tilde{u}) + F(f(\tilde{x}, \tilde{u})) \\ &\leq V_{N-1}(\hat{x}) + \ell(x^e, u^e) \end{aligned}$$

Using the dynamic programming principle this inequality applied with $\hat{x} = f(x, \mu_N(x))$ implies

$$\ell(x, \mu_N(x)) = V_N(x) - V_{N-1}(f(x, \mu_N(x))) \leq V_N(x) - V_N(f(x, \mu_N(x))) + \ell(x^e, u^e)$$

and we can conclude

$$\begin{aligned} J_K^{cl}(x_0, \mu_N) &= \sum_{k=0}^{K-1} \ell(x_{\mu_N}(k), \mu_N(x_{\mu_N}(k))) \\ &\leq \sum_{k=0}^{K-1} \Big[V_N(x_{\mu_N}(k)) - V_N(x_{\mu_N}(k+1)) + \ell(x^e, u^e) \Big] \\ &= V_N(x_0) - V_N(x_{\mu_N}(K)) + K\ell(x^e, u^e) \\ &\leq V_N(x_0) - M + K\ell(x^e, u^e), \end{aligned}$$

where $M \in \mathbb{R}$ is a lower bound on V_N. This yields

$$\overline{J}_\infty^{cl}(x_0, \mu_N) \leq \limsup_{K \to \infty} \left(\frac{V_N(x_0)}{K} - \frac{M}{K} + \ell(x^e, u^e) \right) = \ell(x^e, u^e).$$

□

We note that the boundedness assumption on V_N is satisfied if ℓ is continuous, $\mathbb{Y}$ is compact and F is bounded from below, because in this case both ℓ and F, and thus also V_N, are bounded from below.

Clearly, the estimate from Theorem 13.6 is particularly powerful if $\ell(x^e, u^e)$ is the best, i.e., the smallest possible value that $\overline{J}_\infty^{cl}(x_0, \mu_N)$ can attain. The next definition provides a property which is sufficient for this fact, as the subsequent Proposition 13.9 shows.

Definition 13.7 [Dissipativity and strict dissipativity] We say that an optimal control problem with stage cost ℓ is *strictly dissipative* at an equilibrium $(x^e, u^e) \in \mathbb{Y}$ if there exists a *storage function* $\lambda : \mathbb{X} \to \mathbb{R}$ bounded from below and satisfying $\lambda(x^e) = 0$, and a function $\rho \in \mathcal{K}_\infty$ such that for all $(x, u) \in \mathbb{Y}$ the inequality

$$\ell(x, u) - \ell(x^e, u^e) + \lambda(x) - \lambda(f(x, u)) \geq \rho(|x|_{x^e}) \tag{13.3}$$

holds.

We say that an optimal control problem with stage cost ℓ is *dissipative* at (x^e, u^e) if the same conditions hold with $\rho \equiv 0$. □

We note that the assumption $\lambda(x^e) = 0$ can be made without loss of generality because adding a constant to λ does not invalidate (13.3).

The classical physical interpretation of the storage function is that $\lambda(x)$ quantifies the amount of energy stored in the system at state x. The function $s(x,u) := \ell(x,u) - \ell(x^e,u^e)$ is called the *supply rate* and measures the (possibly negative) amount of energy supplied to the system via the input u at state x. With this interpretation, strict dissipativity then means that a certain amount of energy, quantified by $\rho(|x|_{x^e})$, is dissipated to the environment in each time step. Of course, in the context of general optimal control problems considered in this chapter the storage function and the supply rate need not have an energy interpretation.

Example 13.8 (i) Any optimal control problem with stage cost satisfying $\ell(x^e,u^e) = 0$ and $\ell(x,u) \geq \rho(|x-x^e|)$ is strictly dissipative with $\lambda \equiv 0$. Hence, MPC problems with stage cost penalizing the distance to a desired equilibrium $x_* = x^e$, as they typically appear in stabilization problems, are always strictly dissipative.

(ii) It is straightforward to check that Example 13.1 is dissipative with $\lambda \equiv 0$ and strictly dissipative with $\lambda(x) = -x^2/2$, both at $(x^e,u^e) = (0,0)$. Note that the storage function $\lambda = -x^2/2$ is bounded from below since $\mathbb{X}$ is bounded. Indeed, for an unbounded state constraint set $\mathbb{X}$ the system would not be strictly dissipative. In this example, the supply rate $s(x,u) = \ell(x,u) = u^2$ does have an energy interpretation and the storage function $\lambda(x)$ shows that the equilibrium (x^e,u^e) is the state in which the stored energy $\lambda(x)$ becomes maximal.

(iii) A somewhat more involved computation shows that the second example from Section 9.1 is strictly dissipative at $x^e = 1/\sqrt[\alpha-1]{\alpha A}$ with storage function $\lambda(x) = \alpha(x-x^e)/x^e$.

(iv) For linear quadratic problems with $Q = C^TC$ and $\mathbb{X} = \mathbb{R}$ it can be shown that strict dissipativity is equivalent to detectability of (A,C), i.e., there are no unobservable eigenvalues $\lambda \in \mathbb{C}$ with $|\lambda| \geq 0$. If $\mathbb{X} \subset \mathbb{R}$ is compact, then strict dissipativity is equivalent to the fact that no unobservable eigenvalues with $|\lambda| = 1$ exist. □

Proposition 13.9 For an optimal control problem ($\mathrm{OCP_N}$) that is dissipative at (x^e,u^e), the point (x^e,u^e) is an optimal equilibrium and the inequality

$$\limsup_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} \ell(x_u(k,x),u(k)) \geq \ell(x^e,u^e) \tag{13.4}$$

holds for all $x \in \mathbb{X}$ and all admissible control sequences $u \in \mathbb{U}^\infty(x)$. □

Proof: Consider an arbitrary equilibrium $(x,u) \in \mathbb{Y}$. Then the identity $x = f(x,u)$ and (13.3) imply

$$\ell(x,u) - \ell(x^e,u^e) = \ell(x,u) - \ell(x^e,u^e) + \lambda(x) - \lambda(f(x,u)) \geq 0$$

which yields $\ell(x^e,u^e) \leq \ell(x,u)$ and thus (x^e,u^e) is an optimal equilibrium.

Moreover, using again (13.3) and denoting by M a lower bound on λ we have

$$\begin{aligned}\sum_{k=0}^{K-1} \ell(x_u(k,x),u(k)) &\geq \sum_{k=0}^{K-1} \ell(x^e,u^e) - \lambda(x_u(k,x)) + \lambda(x_u(k+1,x)) \\ &= K\ell(x^e,u^e) - \lambda(x) + \lambda(x_u(K,x)) \\ &\geq K\ell(x^e,u^e) - \lambda(x) + M\end{aligned}$$

for any $u \in \mathbb{U}^\infty(x)$. This yields

$$\limsup_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} \ell(x_u(k,x),u(k)) \geq \limsup_{K\to\infty} \left(\ell(x^e,u^e) - \frac{\lambda(x)-M}{K} \right) = \ell(x^e,u^e).$$

□

The property expressed by inequality (13.4) is known as *optimal operation at steady state*. It has been shown in [15] that under a controllability condition on the system the converse of Proposition 13.9 is also true, i.e., that optimal operation at a steady state implies dissipativity.

An immediate consequence of Proposition 13.9 is the following corollary.

Corollary 13.10 Consider the MPC Algorithm 11.9 with dissipative optimal control problem ($\mathrm{OCP}_{\mathrm{N,e}}$). Then for all $x \in \mathbb{X}_N$

$$\overline{J}_\infty^{cl}(x,\mu_N) = \inf_{u\in\mathbb{U}^\infty(x)} \limsup_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} \ell(x_u(k,x),u(k)).$$

□

Hence, if dissipativity holds, then Theorem 13.6 ensures infinite horizon averaged optimality of the MPC closed loop.

Example 13.11 Since Example 13.1 is dissipative (see Example 13.8(ii)), the MPC closed loop must be infinite horizon averaged optimal. Indeed, as Fig. 13.1 shows, the closed-loop solution converges to the optimal equilibrium. Since the control (not shown in the figure) does the same, $\ell(x_{\mu_N}(k),\mu_N(x_{\mu_N}(k))) \to 0$ as $k \to \infty$ follows which implies $\overline{J}_\infty^{cl}(x,\mu_N) = 0$, which is clearly optimal since $\ell \geq 0$. □

13.3 Asymptotic stability with terminal conditions

One might conjecture that optimal operation at the steady state (x^e,u^e) implies that closed-loop solutions satisfying (13.2) must also converge to x^e. However, under the assumptions imposed in Theorem 13.6 and Proposition 13.9 this is not necessarily the case. To see this, it suffices to consider an optimal control problem with $\ell \equiv 0$. Such a problem clearly satisfies all assumptions (with terminal cost $F \equiv 0$ and storage function $\ell \equiv 0$), yet every trajectory

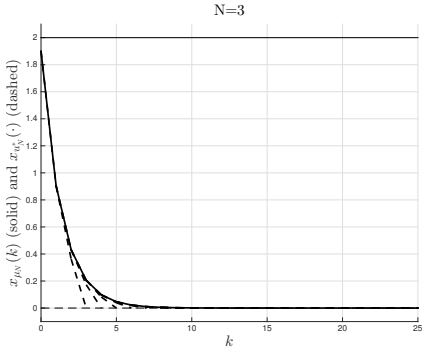


Abbildung 13.1: MPC closed-loop solution (solid) and open-loop predictions (dashed) for Example 13.1 with terminal constraint $\mathbb{X}_0 = \{0\}$ and horizon $N = 3$. The solid line at $x = 2$ indicates the upper bound of the admissible set $\mathbb{X}$

is an optimal trajectory and thus optimal trajectories obviously need not converge to x^e. In order to achieve this — and, in fact, even asymptotic stability of x^e — we need to assume strict dissipativity.

Under this assumption, we establish asymptotic stability by proving the existence of a Lyapunov function. This Lyapunov function will be built from the optimal value function of an auxiliary optimal control problem with stage cost

$$\tilde{\ell}(x,u) := \ell(x,u) - \ell(x^e,u^e) + \lambda(x) - \lambda(f(x,u)) \tag{13.5}$$

and terminal cost

$$\widetilde{F}(x) := F(x) + \lambda(x).$$

These costs are usually called *rotated* or *modified* costs. The name "rotated cost" stems from the fact that for linear f and strictly convex ℓ the graph of $\tilde{\ell}$ is obtained by rotating the graph of ℓ. The corresponding functional is given by

$$\widetilde{J}_N(x_0,u(\cdot)) = \sum_{k=0}^{N-1} \tilde{\ell}(x_u(k,x_0),u(k)) + \widetilde{F}(x_u(N,x_0))$$

and the optimal value function by

$$\widetilde{V}_N(x_0) := \inf_{u(\cdot)\in\mathbb{U}_{\mathbb{X}_0}^N(x_0)} \widetilde{J}_N(x_0,u(\cdot)).$$

It is an easy exercise to check that the equalities $\tilde{\ell}(x^e,u^e) = 0$ and $\widetilde{F}(x^e) = 0$ and — under Assumption 13.5(a) — that the inequality

$$\widetilde{F}(f(x,u)) \le \widetilde{F}(x) - \tilde{\ell}(x,u) \tag{13.6}$$

holds for each $x \in \mathbb{X}_0$ and the control u from Assumption 13.5(a). Moreover, for any $x \in \mathbb{X}_N$ and $u \in \mathbb{U}_{\mathbb{X}_0}^N(x)$ one easily verifies the identity

$$\widetilde{J}_N(x,u) = J_N(x,u) + \lambda(x) - N\ell(x^e,u^e). \tag{13.7}$$

Since the last two terms in (13.7) are independent of u, this implies that the optimal trajectories for J_N and $\widetilde{J}_N$ coincide and that the optimal value functions satisfy

$$\widetilde{V}_N(x) = V_N(x) + \lambda(x) - N\ell(x^e, u^e). \tag{13.8}$$

Since $\tilde{\ell}(x^e, u^e) = 0$ and $\widetilde{F}(x^e) = 0$, using the constant control $u \equiv u^e$ yields

$$\widetilde{V}_N(x^e) \le \widetilde{J}_N(x^e, u) = 0 \quad \text{and thus} \quad V_N(x^e) \le J_N(x^e, u) = N\ell(x^e, u^e) \tag{13.9}$$

using (13.8) and $\lambda(x^e) = 0$.

We now turn to show that $\widetilde{V}_N$ is a Lyapunov function for the MPC closed-loop system. For the rigorous proof of this property, we need the following continuity assumption on F, λ and V_N in x^e.

Assumption 13.12 [Continuity of F, λ and V_N at x^e] There exists γ_F, γ_λ and $\gamma_V \in \mathcal{K}_\infty$ such that the following properties hold.
(a) For all $x \in \mathbb{X}_0$ it holds that

$$|F(x) - F(x^e)| \le \gamma_F(|x|_{x^e}).$$

(b) For all $x \in \mathbb{X}$ it holds that

$$|\lambda(x) - \lambda(x^e)| \le \gamma_\lambda(|x|_{x^e}).$$

(c) For each $N \in \mathbb{N}$ and each $x \in \mathbb{X}_N$ it holds that

$$|V_N(x) - V_N(x^e)| \le \gamma_V(|x|_{x^e}).$$

□

Note that γ_V in (c) is independent of N. We will comment at the end of this section on conditions under which (c) can be ensured.

Theorem 13.13 Consider the MPC Algorithm 11.9 with strictly dissipative optimal control problem ($\mathrm{OCP}_{N,e}$) and bounded $\mathbb{X}_0$. Let Assumptions 13.5(a) and 13.12 be satisfied. Then the optimal equilibrium x^e is asymptotically stable for the MPC closed loop on $\mathbb{X}_N$.

Proof: We show that the modified optimal value function $\widetilde{V}_N$ is a Lyapunov function for the closed-loop system in the sense of Definition 10.4 for $x_* = x^e$. Then the assertion follows from Theorem 10.5. To this end we first check an auxiliary inequality. As in the proof of Theorem 13.6, from Assumption 13.5(a) we obtain $\ell(x, \mu_N(x)) \le V_N(x) - V_N(f(x, \mu_N(x))) + \ell(x^e, u^e)$ which we can rewrite as

$$V_N(x) \ge \ell(x, \mu_N(x)) + V_N(f(x, \mu_N(x))) - \ell(x^e, u^e). \tag{13.10}$$

Using (13.8) this implies

$$\begin{aligned}
\widetilde{V}_N(x) &= V_N(x) + \lambda(x) - N\ell(x^e, u^e) \\
&\ge \ell(x, \mu_N(x)) + V_N(f(x, \mu_N(x))) - \ell(x^e, u^e) + \lambda(x) - N\ell(x^e, u^e) \\
&= \ell(x, \mu_N(x)) + \widetilde{V}_N(f(x, \mu_N(x))) - \lambda(f(x, \mu_N(x))) - \ell(x^e, u^e) + \lambda(x) \\
&= \tilde{\ell}(x, \mu_N(x)) + \widetilde{V}_N(f(x, \mu_N(x))).
\end{aligned}$$

In order to check that $\widetilde{V}_N$ satisfies Definition 10.4, we now have to check the inequalities

$$\alpha_1(|x|_{x^e}) \le \widetilde{V}_N(x) \le \alpha_2(|x|_{x^e}) \quad \text{and} \quad \tilde{\ell}(x,u) \ge \alpha_3(|x|_{x^e}) \tag{13.11}$$

for α_1, α_2, $\alpha_3 \in \mathcal{K}_\infty$. The third inequality follows immediately from the definition of $\tilde{\ell}$ and strict dissipativity for $\alpha_3 = \rho$ from Definition 13.7. For the inequalities involving α_1 and α_2 we first need to establish a lower bound for $\widetilde{F}$.

To this end, for each $x \in \mathbb{X}_0$ we denote the control u from (13.6) by $\mu_0(x)$. Then (13.6) and strict dissipativity implies

$$\widetilde{F}(f(x,\mu_0(x))) \le \widetilde{F}(x) - \tilde{\ell}(x,\mu_0(x)) \le \widetilde{F}(x) - \rho(|x|_{x^e}).$$

By induction along the closed-loop solution for the feedback law μ_0 we then obtain

$$\widetilde{F}(x_{\mu_0}(K,x)) \le \widetilde{F}(x) - \sum_{k=0}^{K-1} \rho(|x_{\mu_0}(k,x)|_{x^e}).$$

This implies that $x_{\mu_0}(K,x) \to x^e$ as $K \to \infty$, because otherwise the sum on the right hand side of this inequality grows unboundedly which implies $\widetilde{F}(x_{\mu_0}(K,x)) \to -\infty$ and contradicts Assumption 13.12(a) and (b) since $x_{\mu_0}(K,x)$ is contained in the bounded set $\mathbb{X}_0$. Again by Assumption 13.12(a) and (b) this implies $\widetilde{F}(x_{\mu_0}(K,x)) \to \widetilde{F}(x^e) = 0$ as $K \to \infty$ from which we can finally conclude

$$\widetilde{F}(x) \ge \lim_{K\to\infty} \sum_{k=0}^{K-1} \rho(|x_{\mu_0}(k,x)|_{x^e}) \ge \rho(|x|_{x^e}) \ge 0.$$

From this, the definitions of $\widetilde{J}_N$ and $\widetilde{V}_N$ immediately imply $\widetilde{V}_N(x) \ge \tilde{\ell}(x,\mu_N(x)) \ge \rho(|x|_{x^e})$ and thus the inequality for α_1 in (13.11) with $\alpha_1 = \rho$.

Together with (13.9) this implies $\widetilde{V}_N(x^e) = 0$ and the second inequality in (13.11) follows from (13.8) and Assumption 13.12(b) and (c) with $\alpha_2 = \gamma_\lambda + \gamma_V$. □

Observe that in the case of stabilizing stage costs according to Example 13.8(i), we obtain $\lambda \equiv 0$ and $\ell(x^e,u^e) = 0$, and thus $\widetilde{V}_N = V_N$. This implies that the optimal value function itself is a Lyapunov function.

We end this section by discussing sufficient conditions for the bound on V_N required in Assumption 13.12(c). In the case of equilibrium terminal conditions, i.e., $\mathbb{X}_0 = \{x^e\}$ and $F \equiv 0$, this property can be ensured by the condition that x^e is reachable from every $x \in \mathbb{X}_N$ with suitable bounded costs. In case ℓ and f are continuous, it is sufficient to assume that the control sequence steering x to x^e is sufficiently close to the constant control with value u^e. For details we refer to [1], particularly to part 2 of Assumption 2 in [1].

In case $\mathbb{X}_0$ contains a neighborhood of x^e, using Assumption 13.5(a) inductively yields the inequality

$$V_N(x) \le F(x) + N\ell(x^e,u^e)$$

while from (13.8) and $\widetilde{V}_N \ge 0$ we obtain

$$V_N(x) \ge -\lambda(x) + N\ell(x^e,u^e).$$

Since from (13.9) we moreover know $V_N(x^e) = N\ell(x^e, u^e)$, this implies Assumption 13.12(c) for $x \in \mathbb{X}_0$ provided Assumption 13.12(a) and (b) hold. For $x \in \mathbb{X}_N \setminus \mathbb{X}_0$ the inequality follows from boundedness of V_N which in turn follows from boundedness of ℓ along the optimal trajectories.

Example 13.14 According to Example 13.8, the optimal control problem from Example 13.1 is strictly dissipative. Moreover, one easily verifies that x^e is reachable in two steps from each $x \in \mathbb{X}$ with cost $4x^2$, which implies the upper bound on V_N for the terminal constraint set $\mathbb{X}_0 = \{0\}$. Hence, we expect the MPC closed loop to be asymptotically stable, which was already illustrated in Fig. 13.1. □

13.4 Non-averaged performance with terminal conditions

The averaged performance result from Theorem 13.6 provides a useful estimate for large times k. However, it also has two significant weaknesses. First, it does not provide an advantage over a stabilizing MPC algorithm. Indeed, for any combination of a continuous stage cost and a terminal condition for which the MPC closed-loop solution converges to x^e and the corresponding control sequence converges to u^e, the value $\ell(x_{\mu_N}(k), \mu_N(x_{\mu_N}(k)))$ converges to $\ell(x^e, u^e)$ from which $\overline{J}^{cl}_\infty(x, \mu_N) = \ell(x^e, u^e)$ follows. Hence, Theorem 13.6 only states that the economic MPC scheme does not perform worse than a stabilizing one. Second, the averaged estimate does not allow any statement about the finite time behavior of the closed-loop trajectory. Indeed, on any finite time interval of arbitrary length the closed-loop trajectory could behave arbitrarily bad as long as eventually it converges to the equilibrium. Clearly, this is not what we would expect an MPC closed-loop trajectory to do and it is also not consistent with what we see in numerical simulations, e.g., in Fig. 13.1. Hence, in this section we derive estimates for the non-averaged infinite and finite horizon performance $J^{cl}_\infty(x, \mu_N)$ and $J^{cl}_K(x, \mu_N)$, respectively. For the infinite horizon estimate the additional condition $\ell(x^e, u^e) = 0$ will be imposed, in order to make sure that the infinite sum in $J^{cl}_\infty(x, \mu_N)$ converges. For the finite horizon performance, such a condition is not needed. As we already know that — under the conditions of Theorem 13.13 — the equilibrium x^e is asymptotically stable, the finite horizon value $J^{cl}_K(x, \mu_N)$ measures the performance of the solution during the transient phase, i.e., until it reaches a small neighborhood of x^e. This is why we also call this value *transient performance*.

Since $J^{cl}_K(x, \mu_N)$ and $J^{cl}_\infty(x, \mu_N)$ do not involve any terminal constraints or costs, in our analysis we will also need to consider the optimal control problems ($\mathrm{OCP_N}$) and (OCP_∞) without terminal constraints and terminal costs. In order not to confuse these problems with those using terminal conditions, in this section we denote the functionals and the optimal value functions of the unconstrained problems ($\mathrm{OCP_N}$) and (OCP_∞) by J^{uc}_N, V^{uc}_N, J^{uc}_∞ and V^{uc}_∞,respectively. We emphasize that we use the same stage cost ℓ in all problems. This implies that if one of the problems is strictly dissipative then all problems are. If this is the case, we also consider ($\mathrm{OCP_N}$) for the rotated cost $\tilde{\ell}$ and denote the corresponding functional by $\widetilde{J}^{uc}_N$. A straightforward computation reveals that J^{uc}_N and $\widetilde{J}^{uc}_N$ are related by the identity

$$\widetilde{J}^{uc}_N(x, u) = J^{uc}_N(x, u) + \lambda(x) - \lambda(x_u(N, x)) - N\ell(x^e, u^e). \tag{13.12}$$

Observe that compared to (13.7) the additional term $\lambda(x_u(N,x))$ appears here due to the absence of the terminal conditions.

In order to establish our theorems on transient performance, we will need a few preparatory results. The first statement shows that the finite horizon optimal trajectories most of the time stay close to the optimal equilibrium x^e.

Proposition 13.15 Assume that the optimal control problem ($\mathrm{OCP_N}$) is strictly dissipative with bounded storage function λ and $\rho \in \mathcal{K}_\infty$. Then for each $\delta > 0$ there exists $\sigma_\delta \in \mathcal{L}$ such that for all $N, P \in \mathbb{N}$, $x \in \mathbb{X}$ and $u \in \mathbb{U}^N(x)$ with $J_N^{uc}(x,u) \le N\ell(x^e,u^e) + \delta$, the set $\mathcal{Q}(x,u,P,N) := \{k \in \{0,\dots,N-1\} \,|\, |x_u(k,x)|_{x^e} \ge \sigma_\delta(P)\}$ has at most P elements. □

Proof: We fix $\delta > 0$ and claim that the assertion holds with $\sigma_\delta(P) := \rho^{-1}((2M+\delta)/P)$ where M is a bound on $|\lambda|$. To prove this claim, assume that there are N, P, x and u such that $J_N^{uc}(x,u) \le N\ell(x^e,u^e) + \delta$ but $\mathcal{Q}(x,u,P,N)$ contains at least $P+1$ elements. Then from (13.12) we can estimate

$$\widetilde{J}_N^{uc}(x,u) \le J_N^{uc}(x,u) + 2M - N\ell(x^e,u^e) \le 2M + \delta.$$

On the other hand, (13.3), (13.5) and the fact that $\mathcal{Q}(x,u,P,N)$ contains at least $P+1$ elements imply

$$\begin{aligned}\widetilde{J}_N^{uc}(x,u) &\ge \sum_{k=0}^{N-1} \tilde{\ell}(x_u(k,x),u(k)) \ge \sum_{k=0}^{N-1} \rho(|x_u(k,x)|_{x^e}) \ge \sum_{\substack{k\in\{0,\dots,N-1\}\\ |x_u(k,x)|_{x^e} > \sigma_\delta(P)}} \rho(\sigma_\delta(P)) \\ &\ge (P+1)\rho(\sigma_\delta(P)) \ge (P+1)\frac{2M+\delta}{P} > 2M+\delta\end{aligned}$$

which is a contradiction. □

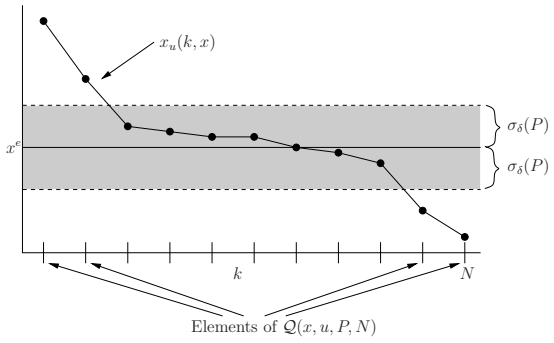


Abbildung 13.2: Illustration of the set $\mathcal{Q}(x,u,P,N)$ defined in Proposition 13.15

We denote the property described by Proposition 13.15 as the *turnpike property*. For an illustration we refer to Fig. 13.2. In fact, there are various variants of the turnpike property known in optimal control, of which the one described by Proposition 13.15 is just a particular version.

We remark that the boundedness assumption on λ can be restrictive in case $\mathbb{X}$ is unbounded. However, for bounded subsets of the state constraint set $\mathbb{X}$ it is not a very strong assumption. Hence, it can be assumed to hold if either $\mathbb{X}$ itself is bounded or if near optimal trajectories are guaranteed to stay in a bounded subset of $\mathbb{X}$.

Example 13.16 Since Example 13.1 is strictly dissipative with bounded storage function (cf. Example 13.8), we expect the system to have the turnpike property. The numerical optimal trajectories depicted in Fig. 13.3 support this claim. □

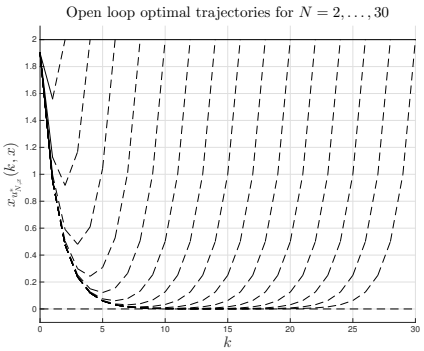


Abbildung 13.3: Open loop optimal trajectories (without terminal conditions) for Example 13.1 with different optimization horizons N. The turnpike property is clearly visible

Next we derive upper and lower bounds for V_∞^{uc}.

Lemma 13.17 Assume that the optimal control problem ($\mathrm{OCP_N}$) is strictly dissipative with bounded storage function λ, that $\ell(x^e, u^e) = 0$ and that Assumptions 13.5(a) and 13.12 hold. Then there is $C > 0$ such that the inequalities

$$-C \leq V_\infty^{uc}(x) \leq \gamma_V(|x|_{x^e}) \tag{13.13}$$

hold for all $x \in \bigcup_{N\in\mathbb{N}} \mathbb{X}_N$ with γ_V from Assumption 13.12(c).

Proof: For $x \in \mathbb{X}_N$, using the control sequence $u(k) = \mu_N(x_{\mu_N}(k,x))$ induced by the closed loop, from (13.10) with $\ell(x^e, u^e) = 0$ for any $K > 0$ we obtain

$$J_K^{uc}(x,u) \quad = \quad \sum_{k=0}^{K-1} \ell(x_u(k,x), u(k)) \leq V_N(x) - V_N(x_u(K,x)).$$

By asymptotic stability of x^e for this solution we obtain $x_u(K,x) \to x^e$ and thus, since $V_N(x^e) = N\ell(x^e,u^e) = 0$ by (13.9), Assumption 13.12(c) yields $V_N(x_u(K,x)) \to 0$ as $K \to \infty$. Using Assumption 13.12(c) and $V_N(x^e) = 0$, this implies

$$V_\infty^{uc}(x) \leq \limsup_{K\to\infty} J_K^{uc}(x,u) \leq V_N(x) \leq \gamma_V(|x|_{x^e}).$$

Moreover, the fact that $\widetilde{J}_N^{uc}(x,u) \geq 0$, (13.12) and the boundedness of λ imply $J_N^{uc}(x,u) \geq -C$ for some $C \geq 0$ and all x, u and N. This implies $V_\infty^{uc}(x) \geq -C$. □

Using the inequality ensured by this lemma we can prove an infinite horizon version of the turnpike property from Proposition 13.15.

Proposition 13.18 Assume that the optimal control problem ($\mathrm{OCP_N}$) is strictly dissipative, that $\mathbb{X}$ is bounded, that $\ell(x^e,u^e) = 0$ and that the inequalities (13.13) hold for all $x \in \bigcup_{N\in\mathbb{N}_0} \mathbb{X}_N$. Then there exists $\sigma_\infty \in \mathcal{L}$ such that for all $P \in \mathbb{N}$, $x \in \mathbb{X}$ and $u \in \mathbb{U}^\infty(x)$ with $J_\infty^{uc}(x,u) \leq V_\infty^{uc}(x) + 1$, the set $\mathcal{Q}(x,u,P,\infty) := \{k \in \mathbb{N}_0 \,|\, |x_u(k,x)|_{x^e} \geq \sigma_\infty(P)\}$ has at most P elements. □

Proof: First note that by Lemma 13.17 and the assumption we get

$$J_\infty^{uc}(x,u) \leq \sup_{x\in\bigcup_{N\in\mathbb{N}}\mathbb{X}_N} V_\infty^{uc}(x) + 1 \leq \sup_{x\in\mathbb{X}} \gamma_V(|x|_{x^e}) + 1 =: \delta.$$

Now we can proceed as in the proof of Proposition 13.15: denoting by M a bound on $|\lambda|$, from (13.12) and $\ell(x^e,u^e) = 0$ we obtain

$$\widetilde{J}_\infty^{uc}(x,u) = \limsup_{K\to\infty} \widetilde{J}_K^{uc}(x,u) \leq \limsup_{K\to\infty} J_K^{uc}(x,u) + 2M \leq \delta + 2M.$$

Setting $\sigma_\infty(K) := \rho^{-1}((2M+\delta)/K)$, the assumption that $\mathcal{Q}(x,u,P,\infty)$ contains more than P elements then again yields a contradiction to this inequality. □

We note that this theorem implies $x_u(k,x) \to x^e$ as $k \to \infty$, because otherwise $\mathcal{Q}(x,u,P,\infty)$ would contain infinitely many elements for sufficiently large $P \in \mathbb{N}$. Using this fact we can improve the lower bound on V_∞^{uc} from Lemma 13.17.

Lemma 13.19 Under the assumptions of Proposition 13.18, the inequality $V_\infty^{uc}(x) \geq -\lambda(x)$ holds for all $x \in \bigcup_{N\in\mathbb{N}_0} \mathbb{X}_N$.

Proof: Let $u \in \mathbb{U}^\infty(x)$ be such that $J_\infty^{uc}(x,u) \leq V_\infty^{uc}(x) + \varepsilon$ for an $\varepsilon \in (0,1)$. As explained above, Proposition 13.18 implies that $x_u(k,x) \to x^e$ as $k \to \infty$. The definition of V_∞^{uc} and (13.12) then imply that

$$\begin{aligned} V_\infty^{uc}(x) + \varepsilon \;&\geq\; \limsup_{K\to\infty} J_K^{uc}(x,u) \\ &=\; \limsup_{K\to\infty} \Big(-\lambda(x) + \underbrace{\widetilde{J}_K^{uc}(x,u)}_{\geq 0} + \underbrace{\lambda(x_u(K,x)}_{\to\lambda(x^e)=0} \Big) \;\geq\; -\lambda(x). \end{aligned}$$

This implies the assertion since $\varepsilon \in (0,1)$ was arbitrary. □

Our final preparatory result is needed for estimating the finite horizon transient performance. It thus concerns the optimal value of the problem with control functions u that steer a given initial value $x \in \mathbb{X}$ to the closed ball $\overline{\mathcal{B}}_\kappa(x^e)$ with radius $\kappa > 0$ around x^e. In order to simplify the notation, we briefly write

$$\mathbb{U}_\kappa^K(x) := \mathbb{U}_{\overline{\mathcal{B}}_\kappa(x^e)}^K(x) \tag{13.14}$$

using the notation from Definition 11.8 with $\overline{\mathcal{B}}_\kappa(x^e)$ in place of $\mathbb{X}_0$. We remark that Theorem 13.13 yields the existence of a $\beta \in \mathcal{KL}$ such that for all $x \in \mathbb{X}_N$ and all K with $\beta(|x|_{x^e}, K) \leq \kappa$ the control u obtained from the MPC feedback law via $u(k) = \mu_N(x_{\mu_N}(k,x))$ is contained in $\mathbb{U}_\kappa^K(x)$. This, in particular, shows that this set is nonempty for sufficiently large K.

The next lemma shows that the infimum of $J_K^{uc}(x,u)$ over $u \in \mathbb{U}_\kappa^K(x)$ and the corresponding approximately optimal trajectories behave similar to those of the infinite horizon problem. More precisely, part (a) of the following lemma is similar to Lemma 13.17, part (b) to Lemma 13.19 and part (c) to Proposition 13.18. Note that since we only consider finite horizon problems here, we do not need to assume $\ell(x^e, u^e) = 0$.

Lemma 13.20 Assume that the optimal control problem ($\mathrm{OCP_N}$) is strictly dissipative with bounded storage function λ and that Assumptions 13.5(a) and 13.12 hold. Fix $\kappa_0 > 0$ and let β be a $\mathcal{KL}$-function characterizing the asymptotic stability of the closed loop, whose existence is guaranteed by Theorem 13.13. Then for any $\kappa \in (0, \kappa_0]$, any $x \in \bigcup_{N\in\mathbb{N}_0} \mathbb{X}_N$ and $K_0 \in \mathbb{N}$ minimal with $\beta(|x|_{x^e}, K_0) \leq \kappa$, the following holds.

(a) For all $K \geq K_0$ the inequality

$$\inf_{u\in\mathbb{U}_\kappa^K(x)} J_K^{uc}(x,u) - K\ell(x^e,u^e) \leq \gamma_V(|x|_{x^e}) + \gamma_V(\kappa)$$

holds with $\gamma_V \in \mathcal{K}_\infty$ from Assumption 13.12(c).

(b) For all $K \in \mathbb{N}$ with $\mathbb{U}_\kappa^K(x) \neq \emptyset$ the inequality

$$-\gamma_\lambda(|x|_{x^e}) - \gamma_\lambda(\kappa) \leq \inf_{u\in\mathbb{U}_\kappa^K(x)} J_K^{uc}(x,u) - K\ell(x^e,u^e)$$

holds with γ_λ from Assumption 13.12(b).

(c) If in addition $\mathbb{X}$ is bounded then there exists $\sigma \in \mathcal{L}$ such that for all $K \geq K_0$, all $P \in \mathbb{N}$ and any $u \in \mathbb{U}_\kappa^K(x)$ with $J_K^{uc}(x,u) \leq \inf_{u\in\mathbb{U}_\kappa^K(x)} J_K^{uc}(x,u) + 1$ there is $k \leq \min\{P, K-1\}$ such that $|x_u(k,x)|_{x^e} < \sigma(\min\{P, K-1\})$.

Proof: (a) The proof of this inequality works similarly to the first part of the proof of Lemma 13.17. For $x \in \mathbb{X}_N$, we choose the control u obtained from the MPC feedback law via $u(k) = \mu_N(x_{\mu_N}(k,x))$. By Theorem 13.13 and the choice of K_0, this control lies in $\mathbb{U}_\kappa^K(x)$. As in the proof of Lemma 13.17, from (13.10) — now with $\ell(x^e,u^e) \neq 0$ — for this u we get

$$J_K^{uc}(x,u) \leq V_N(x) - V_N(x_u(K,x)) + K\ell(x^e,u^e)$$

and from Assumption 13.12(c) and $|x_u(K,x)|_{x^e} < \kappa$ we obtain the assertion.

(b) Let $\varepsilon > 0$ and take a control $u_\varepsilon \in \mathbb{U}_\kappa^K(x)$ with $\inf_{u\in\mathbb{U}_\kappa^K(x)} J_K^{uc}(x,u) + \varepsilon \geq J_K^{uc}(x,u_\varepsilon)$. Then by (13.12), Assumption 13.12(b) and $\lambda(x^e) = 0$, and recalling that strict dissipativity implies $\widetilde{J}_K^{uc}(x,u_\varepsilon) \geq 0$ we get

$$\begin{aligned}
\inf_{u\in\mathbb{U}_\kappa^K(x)} J_K^{uc}(x,u) + \varepsilon &\geq J_K^{uc}(x,u_\varepsilon) \\
&= \underbrace{-\lambda(x)}_{\geq -\gamma_\lambda(|x|_{x^e})} + \underbrace{\widetilde{J}_K^{uc}(x,u_\varepsilon)}_{\geq 0} + \underbrace{\lambda(x_{u_\varepsilon}(K,x))}_{\geq -\gamma_\lambda(\kappa)} + K\ell(x^e,u^e) \\
&\geq -\gamma_\lambda(|x|_{x^e}) - \gamma_\lambda(\kappa) + K\ell(x^e,u^e).
\end{aligned}$$

This implies (b) since $\varepsilon > 0$ was arbitrary.

(c) The assumptions and (a) imply that Proposition 13.15 can be applied with $\delta = \sup_{x\in\mathbb{X}} \gamma(|x|_{x^e}) + \gamma(\kappa_0) + 1$ for all $x \in \mathbb{X}$ and all $\kappa \in (0, \kappa_0]$. We set $\sigma = \sigma_\delta$ from this proposition. Since the set $\mathcal{Q}(x, u, \min\{P, K-1\}, K)$ has at most $\min\{P, K-1\}$ elements, there exists at least one $k \in \{0, \ldots, \min\{P, K-1\}\}$ with $k \notin \mathcal{Q}(x, u, \min\{P, K-1\}, K)$, which thus satisfies $|x_u(k, x)|_{x^e} \le \sigma(\min\{P, K-1\})$. □

We now have all the tools to prove the two main theorems of this section. The first theorem gives an upper bound for the non-averaged infinite horizon performance of the MPC closed-loop trajectory. We recall that when considering the infinite horizon problem we demand $\ell(x^e, u^e) = 0$. Taking into account the inequality $V^{uc}_\infty(x) \le J^{cl}_\infty(x, \mu_N)$ which follows immediately from the definition of these functions, the theorem shows that economic MPC delivers an approximately (non-averaged) infinite horizon optimal closed-loop solution for which the approximation error tends to 0 as the horizon N tends to infinity.

Theorem 13.21 Consider the MPC Algorithm 11.9 with strictly dissipative optimal control problem ($\mathrm{OCP}_{\mathrm{N,e}}$). Assume that $\mathbb{X}$ is bounded, that $\ell(x^e, u^e) = 0$ and that Assumptions 13.5 and 13.12 hold. Then there exists $\delta_1 \in \mathcal{L}$ such that the inequalities

$$J^{cl}_\infty(x, \mu_N) \le V_N(x) \le V^{uc}_\infty(x) + \delta_1(N)$$

hold for all $x \in \mathbb{X}_N$.

Proof: In order to prove the first inequality, from (13.10) we obtain $\ell(x, \mu_N(x)) \le V_N(x) - V_N(f(x, \mu_N(x)))$. This implies for any $K \in \mathbb{N}$

$$J^{cl}_K(x, \mu_N) \;=\; \sum_{k=0}^{K-1} \ell(x_{\mu_N}(k, x), \mu_N(x_{\mu_N}(k, x))) \le V_N(x) - V_N(x_{\mu_N}(K, x)). \quad (13.15)$$

Now from Theorem 13.13 we know that $|x_{\mu_N}(k, x)|_{x^e} \le \beta(|x|_{x^e}, k) \le \beta(M, k) =: \nu(k)$, where $M := \max_{x,y\in\mathbb{X}} d(x, y)$. Note that $\nu \in \mathcal{L}$. Moreover, by (13.9) we have $V_N(x^e) = N\ell(x^e, u^e) = 0$ and from Assumption 13.12(c) we know the existence of $\gamma_V \in \mathcal{K}$ with $|V_N(x)| = |V_N(x) - V_N(x^e)| \le \gamma_V(|x|_{x^e})$ for all $x \in \mathbb{X}$. Together this yields

$$|V_N(x_{\mu_N}(K, x))| \le \gamma_V(\nu(K)).$$

Since $\gamma_V(\nu(K)) \to 0$ for $K \to \infty$, this inequality together with (13.15) yields the first inequality by letting $K \to \infty$.

For the second inequality, we note that it is sufficient to prove the inequality for all sufficiently large N, because by boundedness of V_N and V^{uc}_∞, for small N the inequality can always be satisfied by choosing $\delta_1(N)$ sufficiently large without violating the requirement $\delta_1 \in \mathcal{L}$. Consider σ_∞ from Proposition 13.18, pick N_0 and η from Assumption 13.5(b), choose N_1 such that $\sigma_\infty(N_1) < \eta$, fix $0 < \varepsilon < 1$ and choose an admissible control u_ε satisfying $J^{uc}_\infty(x, u_\varepsilon) \le V^{uc}_\infty(x) + \varepsilon$. Then for $N \ge 2N_1$ we use Proposition 13.18 with $P = \lfloor N/2 \rfloor$. We thus obtain the existence of $k \in \{0, \ldots, P-1\}$ such that $|x_{u_\varepsilon}(k, x)|_{x^e} < \sigma_\infty(P) \le \sigma_\infty(N_1) < \eta$, implying $x_u(k, x) \in \mathbb{X}_{N_1} \subseteq \mathbb{X}_{N_2}$ and thus $u_\varepsilon \in \mathbb{U}^k_{\mathbb{X}_{N_2}}(x)$ for all $N_2 \ge N_1$. Particularly, this holds for $N_2 = N - k$, implying

$u_\varepsilon \in \mathbb{U}^k_{\mathbb{X}_{N-k}}(x)$. Now, from Assumption 13.12(c) applied to V_{N-k} we can conclude (again using $V_N(x^e) = 0$)

$$|V_{N-k}(x_{u_\varepsilon}(k,x))| \le \gamma_V(\sigma_\infty(P)).$$

Moreover, Lemma 13.19 and the bound on λ yield

$$\begin{aligned} V^{uc}_\infty(x) + \varepsilon \ge J^{uc}_\infty(x,u_\varepsilon) &\ge J^{uc}_k(x,u_\varepsilon) + V_\infty(x_{u_\varepsilon}(k,x)) \\ &\ge J^{uc}_k(x,u_\varepsilon) - \lambda(x_{u_\varepsilon}(k,x)) \ge J^{uc}_k(x,u_\varepsilon) - \gamma_\lambda(\sigma_\infty(P)). \end{aligned}$$

Together with the dynamic programming principle (12.1) these inequalities imply

$$\begin{aligned} V_N(x) &= \inf_{u \in \mathbb{U}^k_{\mathbb{X}_{N-k}}(x)} \{J^{uc}_k(x,u) + V_{N-k}(x_u(k,x))\} \le J^{uc}_k(x,u_\varepsilon) + V_{N-k}(x_{u_\varepsilon}(k,x)) \\ &\le V^{uc}_\infty(x) + \gamma_V(\sigma_\infty(P)) + \gamma_\lambda(\sigma_\infty(P)) + \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, this proves the assertion for $\delta_1(N) = \gamma_V(\sigma_\infty(\lfloor N/2 \rfloor)) + \gamma_\lambda(\sigma_\infty(\lfloor N/2 \rfloor))$. □

Since x^e is asymptotically stable for the MPC closed-loop trajectories, the closed-loop solutions converge towards x^e as $k \to \infty$. More precisely, given a time K, by Theorem 13.13 the solutions are guaranteed to satisfy $x_{\mu_N}(k,x) \in \overline{\mathcal{B}}_\kappa(x^e)$ for all $k \ge K$ and $\kappa = \beta(|x|_{x^e}, K)$ for β from Theorem 13.13. We denote the time span $\{0, \dots, K-1\}$ during which the system is (possibly) outside $\overline{\mathcal{B}}_\kappa(x^e)$ as *transient time* and the related finite horizon functional $J^{uc}_K(x,u)$ as *transient performance*. The next theorem then shows that among all possible trajectories from x to $\overline{\mathcal{B}}_\kappa(x^e)$, the MPC closed loop has the best transient performance up to error terms vanishing as $K \to \infty$ and $N \to \infty$. Again, in order to simplify the notation, we use $\mathbb{U}^K_\kappa(x)$ from (13.14). We remark that unlike the previous theorem here we do not need to assume $\ell(x^e, u^e) = 0$.

Theorem 13.22 Consider the MPC Algorithm 11.9 with strictly dissipative optimal control problem ($\mathrm{OCP_{N,e}}$). Assume that $\mathbb{X}$ is bounded and that Assumptions 13.5 and 13.12 hold. Then there exist $\delta_1, \delta_2 \in \mathcal{L}$ such that for all $x \in \mathbb{X}_N$ the inequality

$$J^{cl}_K(x, \mu_N) \le \inf_{u \in \mathbb{U}^K_\kappa(x)} J^{uc}_K(x,u) + \delta_1(N) + \delta_2(K)$$

holds with $\kappa = \beta(|x|_{x^e}, K)$ and $\beta \in \mathcal{KL}$ characterizing the asymptotic stability of the closed loop guaranteed by Theorem 13.13.

Proof: We can without loss of generality assume $\ell(x^e, u^e) = 0$ because the claimed inequality is invariant under adding constants to ℓ. Moreover, similar to the proof of Theorem 13.21 it is sufficient to prove the inequality for all sufficiently large K and N, because by boundedness of all functions involved for small N and K the inequality can always be achieved by choosing $\delta_1(N)$ and $\delta_2(K)$ sufficiently large. As in the first step of the previous proof we obtain $|V_N(x_{\mu_N}(K,x))| \le \gamma_V(\nu(K))$. It is thus sufficient to show the existence of $\delta_1, \tilde{\delta}_2 \in \mathcal{L}$ with

$$V_N(x) \le \inf_{u \in \mathbb{U}^K_\kappa(x)} J^{uc}_K(x) + \delta_1(N) + \tilde{\delta}_2(K) \tag{13.16}$$

for all $x \in \mathbb{X}_N$ because then the assertion follows from (13.15) with $\delta_2 = \gamma_V \circ \nu + \tilde{\delta}_2$.

In order to prove (13.16), consider σ from Lemma 13.20(c), which we apply with $P = \lfloor N/2 \rfloor$ and pick $u_\varepsilon \in \mathbb{U}_\kappa^K(x)$ with $J_K^{uc}(x, u_\varepsilon) \leq \inf_{u \in \mathbb{U}_\kappa^K(x)} J_K^{uc}(x,u) + \varepsilon$ with an arbitrary but fixed $\varepsilon \in (0,1)$. This yields the existence of $k \in \{0, \dots, \lfloor N/2 \rfloor\}$, $k \leq K-1$ with $|x_{u_\varepsilon}(k,x)|_{x^e} \leq \sigma(\min\{P, K-1\})$. Since u_ε steers x to $\overline{\mathcal{B}}_\kappa(x^e)$, the shifted sequence $u_\varepsilon(k+\cdot)$ lies in $\mathbb{U}_\kappa^{K-k}(x_{u_\varepsilon}(k,x))$, implying that this set is nonempty. Hence, we can apply Lemma 13.20(b) in order to conclude $J_{K-k}^{uc}(x_{u_\varepsilon}(k,x), u_\varepsilon(k+\cdot)) \geq -\gamma_\lambda(\sigma(\min\{N, K-1\})) - \gamma_\lambda(\kappa)$. This implies

$$\begin{aligned}\inf_{u \in \mathbb{U}_\kappa^K(x)} J_K^{uc}(x,u) + \varepsilon \;&\geq\; J_K^{uc}(x, u_\varepsilon) \;=\; J_k^{uc}(x, u_\varepsilon) + J_{K-k}^{uc}(x_{u_\varepsilon}(k,x), u_\varepsilon(k+\cdot)) \\ &\geq\; J_k^{uc}(x, u_\varepsilon) - \gamma_\lambda(\sigma(\min\{N, K-1\})) - \gamma_\lambda(\kappa)\end{aligned}$$

Moreover, by choosing N and K sufficiently large we can ensure $\sigma(\min\{P, K-1\}) < \eta$ for η from Assumption 13.5(b), implying $u_\varepsilon \in \mathbb{U}_{\mathbb{X}_Q}^k(x)$ for all $Q \geq N_0$ and N_0 from Assumption 13.5(b). Particularly, choosing $N \geq 2N_0$ implies $N - k \geq N_0$ and thus $u_\varepsilon \in \mathbb{U}_{\mathbb{X}_{N-k}}^k(x)$.

Using this relation, the inequality derived above, the dynamic programming principle (12.1) and Assumption 13.12(c) for V_{N-k} we obtain

$$\begin{aligned}V_N(x) \;&=\; \inf_{u \in \mathbb{U}_{\mathbb{X}_{N-k}}^k(x)} \{J_k^{uc}(x,u) + V_{N-k}(x_u(k,x))\} \;\leq\; J_k^{uc}(x,u_\varepsilon) + V_{N-k}(x_{u_\varepsilon}(k,x)) \\ \leq\; &\inf_{u \in \mathbb{U}_\kappa^K(x)} J_K^{uc}(x,u) + \gamma_\lambda(\sigma(\min\{P, K-1\})) + \gamma_\lambda(\kappa) + \varepsilon \\ &\qquad + \gamma_V(\sigma(\min\{P, K-1\})).\end{aligned}$$

This shows the desired inequality (13.16) for

$$\delta_1(N) = \gamma_V(\sigma(\lfloor N/2 \rfloor)) + \gamma_\lambda(\sigma(\lfloor N/2 \rfloor))$$

and, using the choice of κ,

$$\tilde{\delta}_2(K) \;=\; \gamma_V(\sigma(K-1)) + \gamma_\lambda(\sigma(K-1)) \;+\; \gamma_\lambda(\beta(M,K))$$

with $M = \max_{x,y \in \mathbb{X}} d(x,y)$ and $\beta \in \mathcal{KL}$ characterizing the asymptotic stability of the closed loop. □

Example 13.23 Fig. 13.4 illustrates how $J_K^{cl}(x, \mu_N)$ depends on N for Example 13.1. The value $K = 30$ is so large that the effect of the term $\delta_2(K)$ is negligible and not visible in the figure, hence $J_K^{cl}(x, \mu_N)$ converges to $\inf_{u \in \mathbb{U}_\kappa^K(x)} J_K^{uc}(x,u)$ for increasing N. □

13.5 Averaged optimality without terminal conditions

In this and in the subsequent sections we discuss the case in which we do not impose terminal conditions on the problem, i.e., we consider the MPC Algorithm 11.1 with optimal control problem ($\mathrm{OCP_N}$). The corresponding functionals and optimal value functions will, as usual, be denoted by J_N and V_N and their infinite horizon counterparts by J_∞ and V_∞, i.e., we do not use the superscript notation J_N^{uc} etc. anymore in the sequel. The results are presented in parallel to Sects. 13.2–13.4.

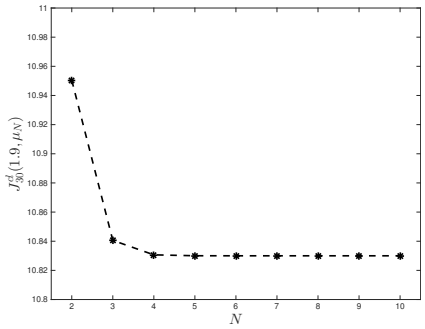


Abbildung 13.4: Value of $J_K^{cl}(x, \mu_N)$ for $K = 30$, $x = 1.9$ and varying N with terminal constraint $\mathbb{X} = \{0\}$

Since we do not impose any terminal conditions, we do not need Assumptions 13.5 and 13.12(a) anymore. However, we still need Part (b) and (a relaxed version of) Part (c) of Assumption 13.12, where the latter now refers to the optimal value function of the unconstrained problem ($\mathrm{OCP_N}$).

Assumption 13.24 [Continuity of λ and V_N at x^e] There exist γ_λ and $\gamma_V \in \mathcal{K}_\infty$ and $\omega \in \mathcal{L}$ such that the following properties hold.
(a) For all $x \in \mathbb{X}$ it holds that

$$|\lambda(x) - \lambda(x^e)| \leq \gamma_\lambda(|x|_{x^e}).$$

(b) For each $N \in \mathbb{N}$ and each $x \in \mathbb{X}$ it holds that

$$|V_N(x) - V_N(x^e)| \leq \gamma_V(|x|_{x^e}) + \omega(N).$$

□

Note that (b) implies viability of $\mathbb{X}$ which we assume for simplicity in this section. If desired, this condition could be relaxed (see [4] for details in a continuous time setting). One method of ensuring the continuity from (b) without requiring explicit knowledge of V_N is by assuming strict dissipativity and local controllability around x^e, see [16] or [6, Sect. 6].

We observe that Propositions 13.15 and 13.18 remain valid, as the assumptions, statements and proofs do not involve any terminal constraints or costs. Based on these two propositions, we can prove the following two auxiliary results, which lead to the main result of this section. In what follows, we denote by $u_\infty^\star$ and $u_N^\star$ the optimal control sequences for (OCP_∞) and ($\mathrm{OCP_N}$), respectively, for initial value $x \in \mathbb{X}$.

Lemma 13.25 If Assumption 13.24 and the assumptions of Proposition 13.15 hold, then the equation

$$V_N(x) = J_M(x, u_N^\star) + V_{N-M}(x^e) + R_1(x, M, N) \tag{13.17}$$

holds with $|R_1(x, M, N)| \leq \gamma_V(\sigma_\delta(P)) + \omega(N - M)$ for all $x \in \mathbb{X}$, all $N \in \mathbb{N}$, all $P \in \mathbb{N}$ and all $M \notin \mathcal{Q}(x, u_N^\star, P, N)$, with σ_δ from Proposition 13.15 with $\delta = \gamma_V(|x|_{x^e}) + \omega(N)$.

Proof: Observe that using the constant control $u \equiv u^e$ we can estimate $V_N(x^e) \leq J_N(x^e, u) = N\ell(x^e, u^e)$. Thus, using Assumption 13.24 we get $J_N(x, u_N^\star) \leq N\ell(x^e, u^e) + \gamma_V(|x|_{x^e}) + \omega(N)$, hence Proposition 13.15 applies to the optimal trajectory with $\delta = \gamma_V(|x|_{x^e}) + \omega(N)$. This in particular ensures $|x_{u_N^\star}(M, x)|_{x^e} \leq \sigma_\delta(P)$ for all $M \notin \mathcal{Q}(x, u_N^\star, P, N)$.

Now the dynamic programming principle (12.2) yields

$$V_N(x) = J_M(x, u_N^\star) + V_{N-M}(x_{u_N^\star}(M, x)).$$

Hence, (13.17) holds with $R_1(x, M, N) = V_{N-M}(x_{u_N^\star}(M, x)) - V_{N-M}(x^e)$. Then for any $P \in \mathbb{N}$ and any $M \notin \mathcal{Q}(x, u_N^\star, P, N)$ this implies $|R_1(x, M, N)| \leq \gamma_V(|x_{u_N^\star}(M, x)|_{x^e}) + \omega(N - M) \leq \gamma_V(\sigma_\delta(P)) + \omega(N - M)$ and thus the assertion. □

Lemma 13.26 If Assumption 13.24 and the assumptions of Proposition 13.15 hold, then the equation

$$V_N(x) \leq V_{N-1}(x) + \ell(x^e, u^e) + R_2(x, N)$$

holds with $|R_2(x, N)| \leq \nu_2(|x|_{x^e}, N) = 2\gamma_V(\sigma_\delta(\lfloor N/2 \rfloor)) + 2\omega(\lfloor N/2 \rfloor - 1)$ for all $x \in \mathbb{X}$, all $N \in \mathbb{N}$ and σ_δ from Proposition 13.15 with $\delta = \gamma_V(|x|_{x^e}) + \omega(N - 1)$.

Proof: Given $x \in \mathbb{X}$, consider the optimal control $u_{N-1}^\star$ for horizon length $N - 1$ and σ_δ from Proposition 13.15 with $\delta = \gamma_V(|x|_{x^e})$. Then Lemma 13.25 applied with $N-1$ in place of N and $P = \lfloor N/2 \rfloor$ implies the existence of $M \in \{0, \ldots, \lfloor N/2 \rfloor - 1\}$ with

$$V_{N-1}(x) = J_M(x, u_{N-1}^\star) + V_{N-M-1}(x^e) + R_1(x, M, N - 1)$$

with $|R_1(x, M, N - 1)| \leq \gamma_V(\sigma_\delta(\lfloor N/2 \rfloor)) + \omega(\lfloor N/2 \rfloor - 1)$. The construction in the proof of Lemma 13.25 moreover yields $|x_{u_{N-1}^\star}(M, x)|_{x^e} \leq \sigma_\delta(\lfloor N/2 \rfloor)$. Using $u(k) = u_{N-1}^\star(k)$ for $k = 0, \ldots, M - 1$ and $u(M + k) = u_{N-M}^\star(k)$ with the optimal control $u_{N-M}^\star$ for initial value $x_{u_N^\star}(M, x)$ and horizon $N - M$ for $k = M, \ldots, N - 1$ yields

$$J_N(x, u) = J_M(x, u_{N-1}^\star) + V_{N-M}(x_{u_N^\star}(M, x)) = J_M(x, u_{N-1}^\star) + V_{N-M}(x^e) + \widehat{R}_1(x, M, N)$$

with $|\widehat{R}_1(x, M, N)| \leq \gamma_V(\sigma_\delta(\lfloor N/2 \rfloor)) + \omega(\lfloor N/2 \rfloor)$. Since for initial value x^e we can always stay at the equilibrium for one step and use the optimal control for initial value x^e for the remaining horizon, we obtain the inequality $V_{N-M}(x^e) \leq \ell(x^e, u^e) + V_{N-M-1}(x^e)$. Together this yields

$$\begin{aligned} V_N(x) \leq J_N(x, u) &= J_M(x, u_{N-1}^\star) + V_{N-M}(x^e) + \widehat{R}_1(x, M, N) \\ &\leq J_M(x, u_{N-1}^\star) + \ell(x^e, u^e) + V_{N-M-1}(x^e) + \widehat{R}_1(x, M, N) \\ &= V_{N-1}(x) + \ell(x^e, u^e) - R_1(x, M, N - 1) + \widehat{R}_1(x, M, N) \end{aligned}$$

and thus the claim with $R_2(x, N) = \widehat{R}_1(x, M, N) - R_1(x, M, N - 1)$. □

Now we can state the theorem on the infinite horizon average performance.

Theorem 13.27 Consider the MPC Algorithm 11.1 with strictly dissipative optimal control problem ($\mathrm{OCP_N}$) with bounded storage function λ. Let Assumption 13.24 hold and

assume V_N is bounded from below on $\mathbb{X}$. Then, for any $N \geq 2$ and any $x \in \mathbb{X}$ the averaged closed-loop performance satisfies the inequality

$$\overline{J}_\infty^{cl}(x, \mu_N) \leq \ell(x^e, u^e) + \delta_1(N) \tag{13.18}$$

with $\delta_1(N) \leq 2\gamma_V(\sigma_\delta(\lfloor N/2 \rfloor)) + 2\omega(\lfloor N/2 \rfloor - 1)$ for σ_δ from Proposition 13.15 with $\delta = \sup_{k\in\mathbb{N}} \gamma_V(|x_{\mu_N}(k)|_{x^e}) + \omega(N-1)$ and γ_V and ω from Assumption 13.24.

Proof: Abbreviate $x_{\mu_N}(k) = x_{\mu_N}(k, x)$. From the dynamic programming principle (12.2) and Lemma 13.26 applied with $x = x_{\mu_N}(k+1)$ we obtain

$$\begin{aligned} \ell(x_{\mu_N}(k), \mu_N(x_{\mu_N}(k))) &= V_N(x_{\mu_N}(k)) - V_{N-1}(x_{\mu_N}(k+1)) \\ &\leq V_N(x_{\mu_N}(k)) - V_N(x_{\mu_N}(k+1)) + \ell(x^e, u^e) + \underbrace{\nu_2(|x_{\mu_N}(k+1)|_{x^e}, N)}_{=:\tilde{\nu}_2(k,N)}. \end{aligned}$$

Thus we obtain

$$\begin{aligned} \overline{J}_\infty^{cl}(x, \mu_N) &= \limsup_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} \ell(x_{\mu_N}(k), \mu_N(x_{\mu_N}(k))) \\ &= \limsup_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} \Big(V_N(x_{\mu_N}(k)) - V_N(x_{\mu_N}(k+1)) + \ell(x^e, u^e) + \tilde{\nu}_2(k, N) \Big) \\ &= \ell(x^e, u^e) + \sup_{k\in\mathbb{N}} \tilde{\nu}_2(k, N) + \limsup_{K\to\infty} \frac{V_N(x_0) - V_N(x_{\mu_N}(K))}{K} \\ &\leq \ell(x^e, u^e) + \sup_{k\in\mathbb{N}} \tilde{\nu}_2(k, N) + \limsup_{K\to\infty} \frac{V_N(x_0) + M}{K} = \ell(x^e, u^e) + \sup_{k\in\mathbb{N}} \tilde{\nu}_2(k, N) \end{aligned}$$

where $-M$ is a lower bound on V_N on $\mathbb{X}$. This shows the claim with $\delta_1(N) = \sup_{k\in\mathbb{N}} \tilde{\nu}_2(k, N)$ which satisfies the stated bounds because σ_δ is increasing in δ. □

The difference between this and the corresponding result with terminal conditions is that we get the error term $\delta_1(N)$ on the right hand side of the estimate, which does, however, tend to 0 as $N \to \infty$ provided $\delta < \infty$. This is always the case for bounded state constraints $\mathbb{X}$. In case of unbounded $\mathbb{X}$, Theorem 13.34 from the next section can be used to obtain a bound for $|x_{\mu_N}(k)|_{x^e}$ which is independent of k.

Example 13.28 Fig. 13.5 shows $\overline{J}_\infty^{cl}(x, \mu_N)$ for Example 13.1 depending on N. The plot in the logarithmic scale shows that the value converges to the optimal value $\ell(0, 0) = 0$ exponentially fast, hence the error $\delta_1(N)$ also vanishes exponentially fast. This is actually not a coincidence. However, an analysis of the rate of convergence is beyond the scope of this lecture. We refer to [9] for details. □

13.6 Asymptotic stability without terminal conditions

Now we turn to analyzing the stability properties of the MPC closed-loop solutions without terminal conditions. As in the case with terminal conditions, our goal is to assume strict

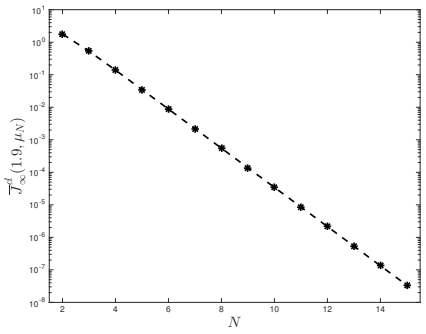


Abbildung 13.5: Value of $\overline{J}^{cl}_{\infty}(x, \mu_N)$ for $x = 1.9$ without terminal conditions depending on N

dissipativity and to use the optimal value function for the modified stage cost $\tilde{\ell}$ from (13.5) as a Lyapunov function, but now without imposing terminal conditions. The crucial difference is that without terminal conditions the optimal trajectories of the original and the modified problem no longer coincide.

In order to see why, we refer to the optimal control problem ($\mathrm{OCP_N}$) with stage cost $\tilde{\ell}$ as ($\widetilde{\mathrm{OCP_N}}$) and, as before, denote the corresponding functional and the optimal value function by $\widetilde{J}_N$ and $\widetilde{V}_N$. Due to the fact that we no longer impose terminal conditions, the relations between V_N and $\widetilde{V}_N$ are not the same as in Sect. 13.3. For J_N and $\widetilde{J}_N$, instead of (13.7) we now have (13.12), which in the notation of this section reads

$$\widetilde{J}_N(x,u) = J_N(x,u) + \lambda(x) - \lambda(x_u(N,x)) - N\ell(x^e,u^e). \tag{13.19}$$

Unfortunately, in contrast to (13.7), this equation does not allow for an easy derivation of a relation between the optimal value functions of the form (13.8), because of the additional u-dependent term $\lambda(x_u(N,x))$ on the right hand side of (13.19). A first consequence of this fact is that the continuity Assumption 13.24(b) does not immediately carry over to $\widetilde{V}_N$. Hence, we need to introduce this as an independent assumption.

Assumption 13.29 [Continuity of $\widetilde{V}_N$ at x^e] There exists $\gamma_{\widetilde{V}} \in \mathcal{K}_\infty$ such that for each $N \in \mathbb{N}$ and each $x \in \mathbb{X}$ it holds that

$$|\widetilde{V}_N(x) - \widetilde{V}_N(x^e)| \le \gamma_{\widetilde{V}}(|x|_{x^e}).$$

□

In case strict dissipativity holds, $\tilde{\ell}$ is positive definite w.r.t. the equilibrium x^e, hence we obtain $\widetilde{V}_N(x^e) = 0$ and $\widetilde{V}_N(x) \ge 0$ for all $x \in \mathbb{X}$. Thus, the inequality in Assumption 13.29 is equivalent to $\widetilde{V}_N(x) \le \gamma_{\widetilde{V}}(|x|_{x^e})$ which can be guaranteed under conditions which guarantee that the system can be controlled to x^e with sufficiently low cost.

Unlike continuity, a straightforward check of Definition 13.7 (with storage function $\lambda \equiv 0$) shows that strict dissipativity carries over from ($\mathrm{OCP_N}$) to ($\widetilde{\mathrm{OCP_N}}$), even with the same ρ. Thus, in particular, all the previous lemmas that apply to ($\mathrm{OCP_N}$) in case of strict

dissipativity also apply to $(\widetilde{\mathrm{OCP}_\mathrm{N}})$. As a general rule, we denote all parameters, sets etc. referring to $(\widetilde{\mathrm{OCP}_\mathrm{N}})$ with a tilde, e.g., the set $\mathcal{Q}(x,u,N,P)$ from Proposition 13.15 will be denoted by $\widetilde{\mathcal{Q}}(x,u,N,P)$ when this proposition is applied to $(\widetilde{\mathrm{OCP}_\mathrm{N}})$.

As already mentioned above, from the definition we cannot directly deduce a simple relation like (13.8) between V_N and $\widetilde{V}_N$. The reason why we can still use $\widetilde{V}_N$ as an — at least practical — Lyapunov function lies in the fact that we can still establish an approximate version of (13.8). To this end, we first need the following preparatory lemma.

Lemma 13.30 If Assumption 13.24 and the assumptions of Proposition 13.15 hold, then the equation

$$V_N(x^e) = M\ell(x^e,u^e) + V_{N-M}(x^e) - R_3(x,P,N)$$

holds with $0 \le R_3(x,P,N) \le \gamma_V(\sigma_\delta(P)) + \omega(N-M) + \gamma_\lambda(\sigma_\delta(P))$ for all $N,P \in \mathbb{N}$ and for all $M \notin \mathcal{Q}(x,u_N^\star,N,P)$, where $u_N^\star \in \mathbb{U}^N(x^e)$ is the optimal control of $(\mathrm{OCP_N})$ for initial value x^e and σ_δ is from Proposition 13.15 with $\delta = \omega(N-M)$.

Proof: The inequality $V_N(x^e) \le M\ell(x^e,u^e) + V_{N-M}(x^e)$ follows from the dynamic programming principle (12.1) using the control $u \equiv u^e$. For the opposite inequality consider the optimal control $u_N^\star \in \mathbb{U}^N(x^e)$ for initial value x^e. As in the proof of Lemma 13.25 we can apply Proposition 13.15 with $\delta = \gamma_V(|x|_{x^e}) + \omega(N)$ in order to conclude that for each $M \notin \mathcal{Q}(x,u_N^\star,N,P)$ we have

$$\begin{aligned}
V_N(x^e) &= \sum_{k=0}^{M-1} \ell(x_{u_N^\star}(k), u_N^\star(k)) + V_{N-M}(x_{u_N^\star}(M)) \\
&= -\lambda(x^e) + \lambda(x_{u_N^\star}(M)) + M\ell(x^e,u^e) + \sum_{k=0}^{M-1} \underbrace{\tilde{\ell}(x_{u_N^\star}(k), u_N^\star(k))}_{\ge 0} + V_{N-M}(x_{u_N^\star}(M)) \\
&\ge M\ell(x^e,u^e) + V_{N-M}(x^e) + \Big[V_{N-M}(x_{u_N^\star}(M)) - V_{N-M}(x^e)\Big] + \Big[\lambda(x_{u_N^\star}(M)) - \lambda(x^e)\Big] \\
&\ge M\ell(x^e,u^e) + V_{N-M}(x^e) - \gamma_V(\sigma_\delta(P)) - \omega(N-M) - \gamma_\lambda(\sigma_\delta(P))
\end{aligned}$$

which shows the claim. □

Now we can prove the approximate relation of the form (13.8) between $\widetilde{V}_N$ and V_N.

Lemma 13.31 If Assumptions 13.24 and 13.29 as well as the assumptions of Proposition 13.15 hold, then the equation

$$\widetilde{V}_N(x) = V_N(x) + \lambda(x) - V_N(x^e) + R_4(x,N)$$

holds with $|R_4(x,N)| \le \nu_4(|x|_{x^e},N)$ with

$$\begin{aligned}
\nu_4(|x|_{x^e},N) = \max\{&\gamma_V(\tilde{\sigma}_{\tilde{\delta}}(\lfloor N/3\rfloor)) + \gamma_V(\sigma_\delta(\lfloor N/3\rfloor)) + \gamma_{\widetilde{V}}(\tilde{\sigma}_{\tilde{\delta}}(\lfloor N/3\rfloor)) \\
&+ \gamma_\lambda(\sigma_\delta(\lfloor N/3\rfloor)) + \gamma_\lambda(\tilde{\sigma}_{\tilde{\delta}}(\lfloor N/3\rfloor)) + 3\omega(\lfloor N/3\rfloor)\,, \\
&\gamma_{\widetilde{V}}(\sigma_\delta(\lfloor N/3\rfloor)) + \gamma_V(\sigma_\delta(\lfloor N/3\rfloor)) + \gamma_\lambda(\sigma_\delta(\lfloor N/3\rfloor)) \\
&+2\omega(\lfloor N/3\rfloor)\}
\end{aligned}$$

with σ_δ and $\tilde{\sigma}_{\tilde{\delta}}$ from Proposition 13.15 applied to $(\mathrm{OCP_N})$ and $(\widetilde{\mathrm{OCP}_\mathrm{N}})$, respectively, with $\delta = \gamma_V(|x|_{x^e}) + \omega(N)$ and $\tilde{\delta} = \gamma_{\widetilde{V}}(|x|_{x^e})$.

Proof: Fix $x \in \mathbb{X}$ and let $u_N^\star$ and $\tilde{u}_N^\star \in \mathbb{U}^N(x)$ denote the optimal control minimizing $J_N(x,u)$ and $\widetilde{J}_N(x,u)$, respectively. We note that if ($\mathrm{OCP_N}$) is strictly dissipative then ($\widetilde{\mathrm{OCP_N}}$) is strictly dissipative, too, with bounded storage function $\lambda \equiv 0$ and same $\rho \in \mathcal{K}_\infty$. Moreover, $V_N(x) \le N\ell(x^e,u^e) + \gamma_V(|x|_{x^e}) + \omega(N)$ and $\widetilde{V}_N(x) \le N\tilde{\ell}(x^e,u^e) + \gamma_{\tilde{V}}(|x|_{x^e})$, since $V_N(x^e) \le N\ell(x^e,u^e)$ and $\widetilde{V}_N(x^e) = 0$. Hence, Proposition 13.15 applies to the optimal trajectories for both problems, yielding $\sigma_\delta \in \mathcal{L}$ and $\mathcal{Q}(x,u_N^\star,P,N)$ for ($\mathrm{OCP_N}$) and $\tilde{\sigma}_{\tilde{\delta}}$ and $\widetilde{\mathcal{Q}}(x,\tilde{u}_N^\star,P,N)$ for ($\widetilde{\mathrm{OCP_N}}$). For all $M \notin \widetilde{\mathcal{Q}}(x,\tilde{u}_N^\star,P,N)$ we can estimate

$$\begin{aligned}
V_N(x) &\le J_M(x,\tilde{u}_N^\star) + V_{N-M}(x_{\tilde{u}_N^\star}(M)) \\
&\le J_M(x,\tilde{u}_N^\star) + V_{N-M}(x^e) + \gamma_V(\tilde{\sigma}_{\tilde{\delta}}(P)) + \omega(N-M) \\
&\le \widetilde{J}_M(x,\tilde{u}_N^\star) - \lambda(x) + \lambda(x^e) + M\ell(x^e,u^e) + V_{N-M}(x^e) + \gamma_V(\tilde{\sigma}_{\tilde{\delta}}(P)) \\
&\quad + \gamma_\lambda(\tilde{\sigma}_{\tilde{\delta}}(P)) + \omega(N-M) \\
&\le \widetilde{V}_N(x) - \widetilde{R}_1(x,P,N) - \lambda(x) \\
&\quad + V_N(x^e) + R_3(x,P,N) + \gamma_V(\tilde{\sigma}_{\tilde{\delta}}(P)) + \gamma_\lambda(\tilde{\sigma}_{\tilde{\delta}}(P)) + \omega(N-M),
\end{aligned}$$

where we have applied the dynamic programming principle (12.1) in the first inequality, Proposition 13.15 for ($\widetilde{\mathrm{OCP_N}}$) and Assumption 13.24(b) respectively Assumption 13.24(a) and (13.19) in the second and third inequality and Lemma 13.25 (applied to ($\widetilde{\mathrm{OCP_N}}$), hence with remainder term denoted by $\widetilde{R}_1$) and Lemma 13.30 (applied to ($\mathrm{OCP_N}$)) in the last step. Moreover, $\lambda(x^e) = 0$ and $\widetilde{V}_N(x^e) = 0$ were used.

By exchanging the two optimal control problems and using the same inequalities as above, we get

$$\begin{aligned}
\widetilde{V}_N(x) &\le V_N(x) - R_1(x,P,N) + \lambda(x) - V_N(x^e) + \gamma_{\tilde{V}}(\sigma_\delta(P)) + \gamma_\lambda(\sigma_\delta(P)) \\
&\quad + \omega(N-M)
\end{aligned}$$

for all $M \notin \mathcal{Q}(x,u_N^\star,P,N)$. Here we can omit the negative $-R_3$-term. Now, choosing $P = \lfloor N/3 \rfloor$, the union $\mathcal{Q}(x,\tilde{u}_N^\star,P,N) \cup \mathcal{Q}(x,u_N^\star,P,N)$ has at most $2N/3$ elements, hence there exists $M \le 2N/3$ for which both inequalities hold. This yields $N - M \ge \lfloor N/3 \rfloor$ and thus

$$\begin{aligned}
|R_1(x,P,N)| &\le \gamma_V(\sigma_\delta(\lfloor N/3 \rfloor)) + \omega(\lfloor N/3 \rfloor), \\
|\widetilde{R}_1(x,P,N)| &\le \gamma_{\tilde{V}}(\tilde{\sigma}_{\tilde{\delta}}(\lfloor N/3 \rfloor)) + \omega(\lfloor N/3 \rfloor) \text{ and} \\
R_3(x,P,N) &\le \gamma_V(\sigma_\delta(\lfloor N/3 \rfloor)) + \omega(\lfloor N/3 \rfloor) + \gamma_\lambda(\sigma_\delta(\lfloor N/3 \rfloor))
\end{aligned}$$

which shows the claim. □

The following proposition shows in which sense $\widetilde{V}_N$ is a Lyapunov function for the system. This will be used in the subsequent theorem in order to prove semiglobal practical asymptotic stability of the closed loop.

Proposition 13.32 Consider the MPC Algorithm 11.1 with strictly dissipative optimal control problem ($\mathrm{OCP_N}$) with bounded storage function λ and $\rho \in \mathcal{K}_\infty$ and let Assumptions 13.24 and 13.29 hold. Then for each $\Theta > 0$ there exists $\delta_1 \in \mathcal{L}$ such that for all $N \ge 2$ with $\delta_1(N) \le \Theta$ the optimal value function $\widetilde{V}_N$ of ($\widetilde{\mathrm{OCP_N}}$) is a Lyapunov function for the closed loop on $S = Y \setminus \mathbb{P}$ for the forward invariant sets $Y = \widetilde{V}_N^{-1}([0,\Theta])$ and $\mathbb{P} = \widetilde{V}_N^{-1}([0,\delta_1(N)])$.

□

Proof: We have to check that Definition 10.4 is satisfied and that Y and P are forward invariant. The lower bound in (10.4) follows with $\alpha_1 = \rho$ because strict dissipativity implies $\tilde{\ell}(x,u) \geq \rho(|x|_{x^e})$ and thus

$$\widetilde{V}_N(x) = \inf_{u \in \mathbb{U}^N(x)} \sum_{k=0}^{N-1} \tilde{\ell}(x_u(k,x),u(k)) \geq \inf_{u \in \mathbb{U}^N(x)} \sum_{k=0}^{N-1} \rho(|x_u(k,x)|_{x^e}) \geq \rho(|x|_{x^e}).$$

The upper bound in (10.4) follows from Assumption 13.29 and $\widetilde{V}_N(x^e) = 0$ with $\alpha_2 = \gamma_{\widetilde{V}}$.

In order to obtain inequality (10.5) we abbreviate $x^+ = f(x,\mu_N(x))$. Now, for all $x \in Y$ we obtain $\widetilde{V}_N(x) \leq \Theta$, which implies $|x|_{x^e} \leq \rho^{-1}(\Theta)$. In order to obtain a similar estimate for $|x^+|_{x^e}$, we observe that $\widetilde{V}_N(x) \leq \Theta$ implies $V_N(x) \leq \Theta - \lambda(x) + M + N\ell(x^e,u^e)$, where $M > 0$ denotes a bound on λ. Thus, Theorem 12.4 and strict dissipativity yield

$$\begin{aligned} V_{N-1}(x^+) &= V_N(x) - \ell(x,\mu_N(x)) \leq V_N(x) + \lambda(x) - \lambda(x^+) - \ell(x^e,u^e) \\ &\leq \Theta - \lambda(x^+) + M + (N-1)\ell(x^e,u^e). \end{aligned}$$

This implies

$$\widetilde{V}_{N-1}(x^+) \leq V_{N-1}(x^+) + \lambda(x^+) + M - (N-1)\ell(x^e,u^e) \leq \Theta + 2M$$

and we can conclude that $|x^+|_{x^e} \leq \rho^{-1}(\Theta + 2M)$. Hence, we can compute

$$\begin{aligned} \widetilde{V}_N(x^+) &= V_N(x^+) + \lambda(x^+) - V_N(x^e) + R_4(x^+,N) \\ &= V_{N-1}(x^+) + \ell(x^e,u^e) + \lambda(x^+) - V_N(x^e) + R_2(x^+,N) + R_4(x^+,N) \\ &= V_N(x) - \ell(x,\mu_N(x)) + \ell(x^e,u^e) + \lambda(x^+) - V_N(x^e) \\ &\qquad + R_2(x^+,N) + R_4(x^+,N) \\ &= \widetilde{V}_N(x) \underbrace{-\ell(x,\mu_N(x)) + \ell(x^e,u^e) + \lambda(x^+) - \lambda(x)}_{=-\tilde{\ell}(x,\mu_N(x))} \\ &\qquad + R_2(x^+,N) + R_4(x^+,N) - R_4(x,N). \end{aligned}$$

where we used Lemma 13.31 for $x = x^+$ for the first equality, Lemma 13.26 for the second, equation (12.6) for the third and Lemma 13.31 in the last step. Defining $\nu(N) = \nu_2(\rho^{-1}(\Theta + 2M),N) + 2\nu_4(\rho^{-1}(\Theta+2M),N)$ with ν_2 and ν_4 from Lemma 13.26 and Lemma 13.31, respectively, we thus obtain

$$\widetilde{V}_N(x^+) \leq \widetilde{V}_N(x) - \rho(|x|_{x^e}) + \nu(N) \leq \widetilde{V}_N(x) - \chi(\widetilde{V}_N(x)) + \nu(N) \tag{13.20}$$

for $\chi := \rho \circ \alpha_2^{-1}(r)$. Now we set $\delta_1(N) = \max\{\chi^{-1}(2\nu(N)),\, \chi^{-1}(\nu(N)) + \nu(N)\}$. Then for all $x \in S = Y \setminus \mathbb{P}$ we obtain $\widetilde{V}_N(x) \geq \delta_1(N)$ and thus $\chi(\widetilde{V}_N(x)) \geq 2\nu(N)$ which implies

$$\widetilde{V}_N(x^+) \leq \widetilde{V}_N(x) - \chi(\widetilde{V}_N(x))/2 \leq \widetilde{V}_N(x) - \chi(\alpha_1(|x|_{x^e}))/2$$

and thus (10.5) with $\alpha_V(r) = \chi(\alpha_1(r))/2$. This inequality also shows that all $x \in Y \setminus \mathbb{P}$ are mapped to Y, since $x \in Y \setminus \mathbb{P} = S$ implies $\widetilde{V}_N(x) \leq \Theta$, hence $\widetilde{V}_N(x^+) < \widetilde{V}_N(x) \leq \Theta$ and thus $x^+ \in Y$.

Finally, to prove forward invariance of $\mathbb{P}$ (which then also implies forward invariance of Y) we recall that $x \in \mathbb{P}$ if and only if $\widetilde{V}_N(x) \le \delta_1(N)$. Now we pick $x \in \mathbb{P}$ and distinguish two cases.

Case 1: $\chi(\widetilde{V}_N(x)) \ge \nu(N)$. In this case from (13.20) we obtain

$$\widetilde{V}_N(x^+) \le \widetilde{V}_N(x) - \chi(\widetilde{V}_N(x)) + \nu(N) \le \widetilde{V}_N(x) \le \delta_1(N).$$

Case 2: $\chi(\widetilde{V}_N(x)) < \nu(N)$. In this case from (13.20) we obtain

$$\begin{aligned} \widetilde{V}_N(x^+) &\le \widetilde{V}_N(x) - \chi(\widetilde{V}_N(x)) + \nu(N) \le \widetilde{V}_N(x) + \nu(N) \\ &< \chi^{-1}(\nu(N)) + \nu(N) \le \delta_1(N). \end{aligned}$$

Hence, in both cases we get $\widetilde{V}_N(x^+) \le \delta_1(N)$ and thus $x^+ \in \mathbb{P}$, which proves the forward invariance of $\mathbb{P}$. □

We note that for small values of N the inequality $\delta_1(N) \ge \Theta$ may hold, in which case the set S on which $\widetilde{V}_N$ is a Lyapunov function is empty.

The final theorem on practical asymptotic stability is now an easy consequence of Proposition 13.32. To this end, we use the following notion of semiglobal practical stability.

Definition 13.33 We call the MPC closed loop system (11.2) *semiglobally practically asymptotically stable with respect to the optimization horizon* N if there exists $\beta \in \mathcal{KL}$ such that the following property holds: for each $\delta > 0$ and $\Delta > \delta$ there exists $N_{\delta,\Delta} \in \mathbb{N}$ such that for all $N \ge N_{\delta,\Delta}$ and all $x \in \mathbb{X}$ with $|x|_{x_*} \le \Delta$ the inequality

$$|x_{\mu_N}(k,x)|_{x_*} \le \max\{\beta(|x|_{x_*},k),\delta\}$$

holds for all $k \in \mathbb{N}_0$. □

Theorem 13.34 Consider the MPC Algorithm 11.1 with strictly dissipative optimal control problem ($\mathrm{OCP_N}$) with bounded storage function λ and $\rho \in \mathcal{K}_\infty$ and let Assumptions 13.24 and 13.29 hold. Then the equilibrium x^e is semiglobally practically asymptotically stable on $\mathbb{X}$ with respect to the optimization horizon N.

Proof: Fixing $\Delta > \delta > 0$, the assertion follows immediately from Proposition 13.32 and Theorem 10.6 when choosing $\Theta = \alpha_2(\Delta)$ (implying $\overline{\mathcal{B}}_\Delta(x^e) \subset Y$) and $N_{\delta,\Delta} > 0$ so large that $\rho(\delta_1(N_{\delta,\Delta})) \le \delta$ holds for δ from Definition 13.33 and $\delta_1(N)$ from Proposition 13.32 (implying $\mathbb{P} \subset \overline{\mathcal{B}}_\delta(x^e)$). □

We will see in the next chapter that this result can be strengthened to "real" asymptotic stability for stabilizing stage cost under suitable additional assumptions.

Example 13.35 Fig. 13.6 shows the trajectories (open loop dashed, MPC closed loop solid) of Example 13.1 without terminal conditions for $N = 5$ and $N = 10$. One clearly sees the practical asymptotic stability of the closed loop and the turnpike phenomenon for the open-loop trajectories. □

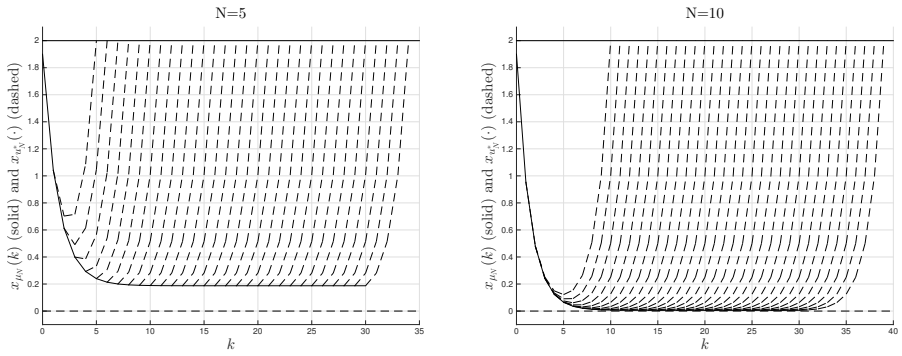


Abbildung 13.6: MPC closed-loop solution (solid) and open-loop predictions (dashed) for Example 13.1 without terminal conditions and horizon $N = 5$ (left) and $N = 10$ (right). The solid line at $x = 2$ indicates the upper bound of the admissible set $\mathbb{X}$

13.7 Non-averaged performance without terminal conditions

Our final results in this chapter concern the adaptation of the results from Sect. 13.4 to the case without terminal conditions. In order to generalize Theorem 13.21, we need the following continuity assumption on the infinite horizon optimal value function and the two subsequent auxiliary results.

Assumption 13.36 [Continuity of V_∞ at x^e] There exists $\gamma_{V_\infty} \in \mathcal{K}_\infty$ such that for each $x \in \mathbb{X}$ it holds that

$$|V_\infty(x) - V_\infty(x^e)| \leq \gamma_{V_\infty}(|x|_{x^e}).$$

□

Lemma 13.37 If Assumption 13.36 and the assumptions of Proposition 13.18 hold, then the equation

$$V_\infty(x) = J_M(x, u_\infty^\star) + V_\infty(x^e) + R_5(x, M) \tag{13.21}$$

holds with $|R_5(x, M)| \leq \gamma_{V_\infty}(\sigma_\infty(P))$ for all $x \in \mathbb{X}$, all $P \in \mathbb{N}$ and all $M \notin \mathcal{Q}(x, u_\infty^\star, P, \infty)$, where $u_\infty^\star \in \mathbb{U}^\infty(x)$ denotes the infinite horizon optimal control for initial value x and σ_∞ is from Proposition 13.18.

Proof: The dynamic programming principle (12.11) yields

$$V_\infty(x) = J_M(x, u_\infty^\star) + V_\infty(x_{u_\infty^\star}(M, x)).$$

Hence, (13.21) holds with $R_5(x, M) = V_\infty(x_{u_\infty^\star}(M, x)) - V_\infty(x^e)$. Then for any $P \in \mathbb{N}$ and $M \notin \mathcal{Q}(x, u_\infty^\star, P, \infty)$ we obtain $|R_5(x, M)| \leq \gamma_{V_\infty}(\|x_{u_\infty^\star}(M, x) - x^e\|) \leq \gamma_{V_\infty}(\sigma_\infty(P))$ and thus the assertion. □

Lemma 13.38 If Assumptions 13.24 and 13.36 and the assumptions of Propositions 13.15 and 13.18 hold, then the equation

$$J_M(x, u^\star_\infty) = J_M(x, u^\star_N) + R_6(x, M, N) \tag{13.22}$$

holds with $|R_6(x, M, N)| \leq \max\{\gamma_V(\sigma_\delta(P)) + \gamma_V(\sigma_\infty(P)) + 2\omega(N-M), \gamma_{V_\infty}(\sigma_\infty(P)) + \gamma_{V_\infty}(\sigma_\delta(P))\}$ for all $P \in \mathbb{N}$, all $x \in \mathbb{X}$ and all $M \in \{0, \dots, N\} \setminus (\mathcal{Q}(x, u^\star_N, P, N) \cup \mathcal{Q}(x, u^\star_\infty, P, \infty))$, with σ_∞ from Proposition 13.18 and σ_δ from Proposition 13.15 with $\delta = |x|_{x^e}$.

Proof: The finite horizon dynamic programming principle (12.1), (12.2) implies that $u = u^\star_N$ minimizes the expression $J_M(x, u) + V_{N-M}(x_u(M, x))$. Together with the error term R_1 from Lemma 13.25 and $\widehat{R}_1(x, M, N) = V_{N-M}(x_{u^\star_\infty}(M, x)) - V_{N-M}(x^e)$ this yields

$$\begin{aligned} J_M(x, u^\star_N) + V_{N-M}(x^e) &= J_M(x, u^\star_N) + V_{N-M}(x_{u^\star_N}(M, x)) - R_1(x, M, N) \\ &\leq J_M(x, u^\star_\infty) + V_{N-M}(x_{u^\star_\infty}(M, x)) - R_1(x, M, N) \\ &= J_M(x, u^\star_\infty) + V_{N-M}(x^e) - R_1(x, M, N) + \widehat{R}_1(x, M, N). \end{aligned}$$

Similar to the proof of Lemma 13.25 one sees that $|\widehat{R}_1(x, M, N)| \leq \gamma_V(\sigma_\infty(P)) + \omega(N-M)$ for all $M \notin \mathcal{Q}(x, u^\star_\infty, P, \infty)$.

Conversely, the infinite horizon dynamic programming principle (12.11) implies that $u^\star_\infty$ minimizes the expression $J_M(x, u^\star_\infty) + V_\infty(x_{u^\star_\infty}(M, x))$. Using the error terms R_5 from Lemma 13.37 and $\widehat{R}_5(x, M, N) = V_\infty(x_{u^\star_N}(M, x)) - V_\infty(x^e)$ we obtain

$$\begin{aligned} J_M(x, u^\star_\infty) + V_\infty(x^e) &= J_M(x, u^\star_\infty) + V_\infty(x_{u^\star_\infty}(M, x)) - R_5(x, M) \\ &\leq J_M(x, u^\star_N) + V_\infty(x_{u^\star_N}(M, x)) - R_5(x, M) \\ &= J_M(x, u^\star_N) + V_\infty(x^e) - R_5(x, M) + \widehat{R}_5(x, M, N). \end{aligned}$$

As in the proof of Lemma 13.25 one sees that Proposition 13.15 applies to $x_{u^\star}(\cdot, x)$ with $\delta = \gamma_V(|x|_{x^e})$. Hence, similar to the proof of Lemma 13.37 one obtains $|\widehat{R}_5(x, M, N)| \leq \gamma_{V_\infty}(\sigma_\delta(P))$ for all $M \notin \mathcal{Q}(x, u^\star_N, P, N)$. Together with the estimates for R_1 and R_5 from Lemma 13.25 and 13.37 this yields

$$\begin{aligned} |R_6(x, M, N)| &= |J_M(x, u^\star_\infty) - J_M(x, u^\star_N)| \\ &\leq \max\{|R_1(x, M, N)| + |\widehat{R}_1(x, M, N)|, |R_5(x, M)| + |\widehat{R}_5(x, M, N)|\} \\ &\leq \max\{\gamma_V(\sigma_\delta(P)) + \gamma_V(\sigma_\infty(P)) + 2\omega(N-M), \gamma_{V_\infty}(\sigma_\infty(P)) + \gamma_{V_\infty}(\sigma_\delta(P))\} \end{aligned}$$

and thus the claim. □

Now we can establish a version of Theorem 13.21 for economic MPC without terminal conditions. We will discuss after the proof how Theorem 13.39 relates to Theorem 13.21.

Theorem 13.39 Consider the MPC Algorithm 11.1 with strictly dissipative optimal control problem ($\mathrm{OCP_N}$) with bounded storage function λ, assume that $\ell(x^e, u^e) = 0$ and $\mathbb{X}$ is bounded and let Assumptions 13.24 and 13.36 hold. Then the inequality

$$J^{cl}_K(x, \mu_N) + V_\infty(x_{\mu_N}(K)) \leq V_\infty(x) + K\delta_1(N) \tag{13.23}$$

holds for all $K \in \mathbb{N}$ and all sufficiently large $N \in \mathbb{N}$ with

$$\begin{aligned}\delta_1(N) &:= 2\gamma_V(\sigma_\delta(\lfloor (N-1)/8 \rfloor)) + 2\gamma_V(\sigma_\infty(\lfloor (N-1)/8 \rfloor)) \\ &\quad + 2\gamma_{V_\infty}(\sigma_\delta(\lfloor (N-1)/8 \rfloor)) + 4\gamma_{V_\infty}(\sigma_\infty(\lfloor (N-1)/8 \rfloor)) + 4\omega(\lfloor N/2 \rfloor)\end{aligned}$$

with σ_∞ from Proposition 13.18 and σ_δ from Proposition 13.15 with $\delta = \sup_{x\in\mathbb{X}} |x|_{x^e}$.

Proof: We pick $x \in \mathbb{X}$ and abbreviate $x^+ := f(x, \mu_N(x))$. For the corresponding optimal control $u_N^\star$ Corollary 12.3 yields that $u_N^\star(\cdot + 1)$ is an optimal control for initial value x^+ and horizon $N-1$. Hence, for each $M \in \{1, \dots, N\}$ we obtain

$$\begin{aligned}\ell(x, \mu_N(x)) = V_N(x) - V_{N-1}(x^+) &= J_N(x, u_N^\star) - J_{N-1}(x^+, u_N^\star(\cdot + 1)) \\ &= J_M(x, u_N^\star) - J_{M-1}(x^+, u_N^\star(\cdot + 1)),\end{aligned}$$

where the last equality follows from the fact that the omitted terms in the sums defining $J_M(x, u_N^\star)$ and $J_{M-1}(x^+, u_N^\star(\cdot + 1))$ coincide. Using Lemma 13.37 for N, x and M and for $N-1$, x^+ and $M-1$, respectively, yields

$$\begin{aligned}V_\infty(x) - V_\infty(x^+) &= J_M(x, u_\infty^\star) + V_\infty(x^e) + R_5(x, M) \\ &\quad - J_{M-1}(x^+, u_\infty^\star) - V_\infty(x^e) - R_5(x^+, M-1) \\ &= J_M(x, u_\infty^\star) - J_{M-1}(x^+, u_\infty^\star) + R_5(x, M) - R_5(x^+, M-1).\end{aligned}$$

Putting the two equations together and using Lemma 13.38 yields

$$\ell(x, \mu_N(x)) = V_\infty(x) - V_\infty(x^+) + R_7(x, M, N). \tag{13.24}$$

with

$$R_7(x, M, N) = -R_6(x, M, N) + R_6(x^+, M-1, N-1) - R_5(x, M) + R_5(x^+, M-1).$$

From Lemma 13.37 and 13.38 we obtain the bound

$$\begin{aligned}|R_7(x, M, N)| &\le 2\gamma_V(\sigma_\delta(P)) + 2\gamma_V(\sigma_\infty(P)) + 2\gamma_{V_\infty}(\sigma_\delta(P)) + 4\gamma_{V_\infty}(\sigma_\infty(P)) \\ &\quad + 4\omega(N - M)\end{aligned}$$

provided we choose $M \in \{1, \dots, N\}$ with $M \notin \mathcal{Q}(x, u_N^\star, P, N) \cup \mathcal{Q}(x, u_\infty^\star, P, \infty)$ and $M-1 \notin \mathcal{Q}(x^+, u_N^\star(\cdot + 1), P, N-1) \cup \mathcal{Q}(x^+, u_\infty^\star(\cdot + 1), P, \infty)$. Since each of the four $\mathcal{Q}$ sets contains at most P elements, their union contains at most $4P$ elements and hence if $N > 8P$ then there is at least one such M with $M \le N/2$.

Thus, choosing $P = \lfloor (N-1)/8 \rfloor$ yields the existence of $M \le N/2$ such that

$$|R_7(x, M, N)| \le \delta_1(N). \tag{13.25}$$

Applying (13.24), (13.25) for $x = x_{\mu_N}(k, x)$, $k = 0, \dots, K-1$, we can conclude

$$\begin{aligned}J_K^{cl}(x, \mu_N) &= \sum_{k=0}^{K-1} \ell(x_{\mu_N}(k, x), \mu_N(x_{\mu_N}(k, x))) \\ &\le \sum_{k=0}^{K-1} \Big(V_\infty(x_{\mu_N}(k, x)) - V_\infty(x_{\mu_N}(k+1, x)) + \delta_1(N) \Big) \\ &\le V_\infty(x) - V_\infty(x_{\mu_N}(K, x)) + K\delta_1(N).\end{aligned}$$

This proves the claim. □

The interpretation of (13.23) is as follows. If we follow the MPC closed-loop trajectory up to some time K and then continue by using the infinite horizon optimal trajectory starting at $x_{\mu_N}(K,x)$, then the value of the overall trajectory exceeds the infinite horizon optimal value by at most $K\delta(N)$. Although seemingly different, it is indeed closely related to Theorem 13.21 because of the following fact: the inequality from Theorem 13.21 holds for all $x \in \mathbb{X}$ if and only if

$$J_K^{cl}(x,\mu_N) + V_\infty(x_{\mu_N}(K,x)) \le V_\infty(x) + \delta_1(N) \tag{13.26}$$

holds for all $x \in \mathbb{X}$ and all $K \ge 1$. This is because $J_K^{cl}(x,\mu_N)+V_\infty(x_{\mu_N}(K,x)) \le J_\infty^{cl}(x,\mu_N)$ for all $K \ge 1$, hence Theorem 13.21 implies (13.26). Conversely, since the assumptions of Theorem 13.21 imply $V_\infty(x_{\mu_N}(K,x)) \to V_\infty(x^e) = 0$ for $K \to \infty$, the validity of (13.26) for all $K \ge 1$ implies the inequality from Theorem 13.21 by letting $K \to \infty$. Comparing (13.23) with (13.26) one immediately sees the difference between the case with and without terminal conditions: without terminal conditions we get the additional factor K in front of the error term, which implies that for large K the error may increase and that for $K \to \infty$ and fixed N the solution may be far from optimal. A numerical illustration of this effect can be found in Example 13.41(iii), below. However, note that the estimate from Theorem 13.27 shows that the averaged value still behaves well for $K \to \infty$, hence the behavior of the trajectories cannot completely deteriorate.

Finally, we formulate and prove the counterpart of Theorem 13.22 for the case without terminal conditions. To this end, recall the definition of $\mathbb{U}_\kappa^K(x)$ from (13.14).

Theorem 13.40 Consider the MPC Algorithm 11.1 with strictly dissipative optimal control problem ($\mathrm{OCP_N}$) with bounded storage function λ and $\rho \in \mathcal{K}_\infty$, let $\mathbb{X}$ be bounded and let Assumptions 13.24 and 13.29 hold. Then there exist $\delta_1, \delta_2, \delta_3 \in \mathcal{L}$ such that for all $x \in \mathbb{X}$ the inequality

$$J_K^{cl}(x,\mu_N) \le \inf_{u \in \mathbb{U}_\kappa^K(x)} J_K(x,u) + \delta_1(N) + K\delta_2(N) + \delta_3(K)$$

holds with $\kappa = \max\{\beta(|x|_{x^e}, K), \rho^{-1}(\delta_1(N))\}$, with δ_1 from Proposition 13.32 and β characterizing the semiglobal practical asymptotic stability in Theorem 13.34.

Proof: First observe that the assumptions of this theorem include those of Theorem 13.34. Hence, from the proof of Theorem 13.34 we obtain the identity

$$\tilde\ell(x,\mu_N(x)) = \widetilde V_N(x) - \widetilde V_N(f(x,\mu_N(x))) + R_2(x,N) + R_4(f(x,\mu_N(x)),N) + R_4(x,N)$$

with $|R_2(x,N) + R_4(f(x,\mu_N(x),N) + R_4(x,N)| \le \nu_2(a,N) + 2\nu_4(a,N) =: \delta_2(N)$, with ν_2 and ν_4 from Lemma 13.26 and 13.31, respectively, and $a = \sup_{x\in\mathbb{X}} |x|_{x^e}$. Summing this cost along the closed-loop trajectory yields

$$\sum_{k=0}^{K-1} \tilde\ell(x_{\mu_N}(k,x),\mu_N(x_{\mu_N}(k,x))) \le \widetilde V_N(x) - \widetilde V_N(x_{\mu_N}(K)) + K\delta_2(N). \tag{13.27}$$

Now the dynamic programming principle (12.1) and Assumption 13.29 yield for all $K \in \{1,\dots,N\}$ and all $u \in \mathbb{U}_\kappa^K(x)$

$$\widetilde{J}_K(x,u) = \underbrace{\widetilde{J}_K(x,u) + \widetilde{V}_{N-K}(x_u(K,x))}_{\geq \widetilde{V}_N(x)} - \underbrace{\widetilde{V}_{N-K}(x_u(K,x))}_{\leq \gamma_{\widetilde{V}}(\kappa)} \geq \widetilde{V}_N(x) - \gamma_{\widetilde{V}}(\kappa). \tag{13.28}$$

Due to the non-negativity of $\tilde{\ell}$, for $K \geq N$ we get $\widetilde{J}_K(x,u) \geq \widetilde{V}_N(x)$ for all $u \in \mathbb{U}^K(x)$. Hence (13.28) holds for all $K \in \mathbb{N}$. Moreover, we have $\widetilde{V}_N(x) \geq 0$. Using (13.27), (13.28) and (13.12) and the definition of δ_2, for all $u \in \mathbb{U}_\kappa^K(x)$ we obtain

$$\begin{aligned}
J_K^{cl}(x,\mu_N(x)) &= \sum_{k=0}^{K-1} \tilde{\ell}(x_{\mu_N}(k,x), \mu_N(x_{\mu_N}(k,x))) - \lambda(x) + \lambda(x_{\mu_N}(K,x)) \\
&\leq \widetilde{V}_N(x) - \widetilde{V}_N(x_{\mu_N}(K,x)) + K\delta_2(N) - \lambda(x) + \lambda(x_{\mu_N}(K,x)) \\
&\leq \widetilde{J}_K(x,u) + \gamma_{\widetilde{V}}(\kappa) - \widetilde{V}_N(x_{\mu_N}(K,x)) + K\delta_2(N) - \lambda(x) + \lambda(x_{\mu_N}(K,x)) \\
&= J_K(x,u) + \gamma_{\widetilde{V}}(\kappa) - \widetilde{V}_N(x_{\mu_N}(K,x)) + K\delta_2(N) - \lambda(x_u(K,x)) + \lambda(x_{\mu_N}(K,x)) \\
&\leq J_K(x,u) + \gamma_{\widetilde{V}}(\kappa) + K\delta_2(N) + 2\gamma_\lambda(\kappa).
\end{aligned}$$

Using the definition of κ we can estimate and define

$$\begin{aligned}
\gamma_{\widetilde{V}}(\kappa) + 2\gamma_\lambda(\kappa) \leq\ & \underbrace{\sup_{x\in\mathbb{X}} \gamma_{\widetilde{V}}(\beta(|x|_{x^e},K)) + 2\gamma_\lambda(\beta(|x|_{x^e},K))}_{=:\delta_3(K)} \\
& + \underbrace{\gamma_{\widetilde{V}}(\rho^{-1}(\delta(N))) + 2\gamma_\lambda(\rho^{-1}(\delta(N)))}_{=:\delta_1(N)}
\end{aligned}$$

which finishes the proof. □

Example 13.41 (i) Fig. 13.7 illustrates how $J_K^{cl}(x,\mu_N)$ depends on N in Example 13.1. As in Fig. 13.4, the value $K = 30$ is so large that the effect of the term $\delta_2(K)$ is negligible and not visible in the figure, hence $J_K^{cl}(x,\mu_N)$ converges to $\inf_{u\in\mathbb{U}_\kappa^K(x)} J_K^{uc}(x,u)$ for increasing N.

(ii) We note that the error estimate depends on the bound on the storage function λ which enters in several of the previous estimates. This dependence is actually visible when computing $J_K^{cl}(x,\mu_N)$ via numerical simulations. In Example 13.1 the bound on λ increases with increasing $\mathbb{X}$ (cf. Example 13.8). Fig. 13.8 shows that increasing the state constraint set from $\mathbb{X} = [-2,2]$ to $\mathbb{X} = [-3,3]$ indeed considerably increases the error, although the optimal trajectories and thus the limiting values for $J_K^{cl}(x,\mu_N)$ for $N \to \infty$ are independent of the choice of $\mathbb{X}$.

(iii) Finally we observe that the main structural difference between Theorem 13.22 and 13.40 lies in the factor K in the error estimate in Theorem 13.40 without terminal conditions. This predicts a deterioration of the value $J_K^{cl}(x,\mu_N)$ for fixed N and growing K in the case without terminal conditions, which should not appear if terminal conditions are used. This effect can again be seen in numerical simulations for Example 13.1, see Fig. 13.9. Here the increase of $J_K^{cl}(x,\mu_N)$ for increasing K is clearly visible in the left figure, i.e., for

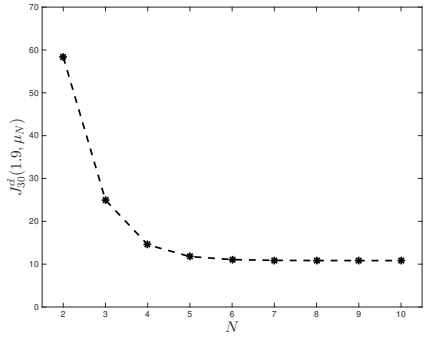


Abbildung 13.7: Value of $J_K^{cl}(x, \mu_N)$ for $K = 30$, $x = 1.9$ and varying N without terminal conditions

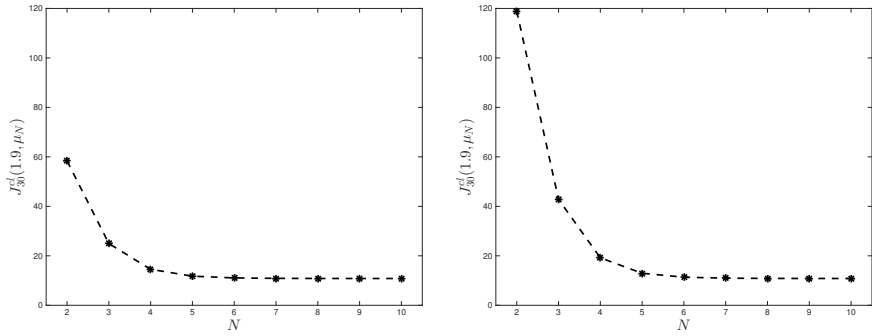


Abbildung 13.8: Value of $J_K^{cl}(x, \mu_N)$ for $K = 30$, $x = 1.9$ and varying N without terminal conditions for $\mathbb{X} = [-2, 2]$ on the left and $\mathbb{X} = [-3, 3]$ on the right

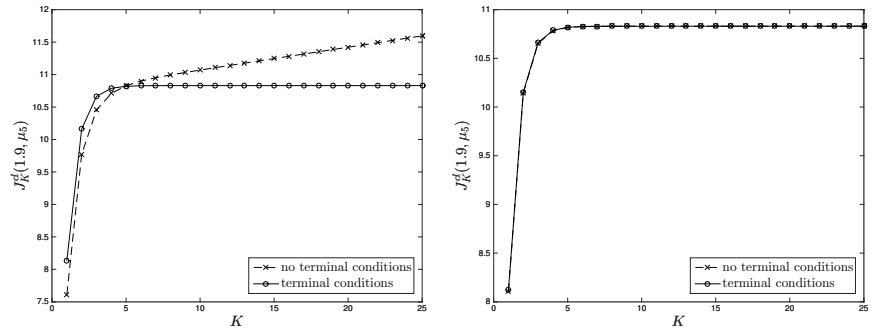


Abbildung 13.9: Value of $J_K^{cl}(x, \mu_N)$ for varying K, $x = 1.9$ and $N = 5$ on the left and $N = 10$ on the right, both with and without terminal conditions $\mathbb{X}_0 = \{0\}$ and $F \equiv 0$

$N = 5$. In the right figure, N has been increased to $N = 10$, due to which the $\delta_2(N)$-term in Theorem 13.40 becomes so small that its effect is not visible anymore for the range of K depicted in the figure.

□

Kapitel 14

Analysis of stabilizing MPC schemes

In this chapter we look at the particular — but practically very relevant — special case in which the stage cost ℓ penalizes the distance from a desired equilibrium. More precisely, we consider stage costs satisfying the conditions

$$\ell(x_*, u_*) = 0 \quad \text{and} \quad \ell(x, u) \geq \alpha_3(|x|_{x_*}) \tag{14.1}$$

for all $x \in \mathbb{X}$ and a $\mathcal{K}_\infty$-function α_3. In normed spaces X and U, the simplest choice for such a function is

$$\ell(x, u) = \|x - x_*\| + \lambda\|u - u_*\|$$

for a control penalization parameter $\lambda \geq 0$.

As we have already observed in Example 13.8(i), problems of this kind are always strictly dissipative (with storage function $\lambda \equiv 0$). Hence, all results of the previous chapter apply and — under the stated conditions — we can conclude asymptotic stability for the scheme with terminal conditions and semiglobal practical asymptotic stability without terminal conditions. In practice, however, one often observes "real" asymptotic stability also in the case without terminal conditions. Also, schemes without terminal conditions are often preferred in practice, because for complex systems the design of terminal conditions satisfying Assumption 13.5 is very difficult if not impossible. Hence, in this chapter we will analyze stabilizing MPC schemes without terminal conditions.

14.1 A relaxed dynamic programming theorem

The basis for the considerations in this chapter is the following fundamental, yet simple to prove theorem.

Theorem 14.1 [Asymptotic stability and suboptimality estimate] Consider a stage cost $\ell : X \times U \to \mathbb{R}_0^+$ and a function $V : X \to \mathbb{R}_0^+$. Let $\mu : \mathbb{X} \to U$ be an admissible feedback law and let $S \subseteq \mathbb{X}$ be a forward invariant set for the closed loop system

$$x^+ = g(x) = f(x, \mu(x)). \tag{14.2}$$

Assume there exists $\alpha \in (0,1]$ such that the *relaxed dynamic programming inequality*

$$V(x) \geq \alpha \ell(x, \mu(n,x)) + V(f(x, \mu(n,x))) \tag{14.3}$$

holds for all $x \in S$. Then the *suboptimality estimate*

$$J_\infty^{cl}(x,\mu) \leq V(x)/\alpha \tag{14.4}$$

holds for all $x \in S$.

If, in addition, ℓ satisfies (14.1) and there exist $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ such that the inequalities

$$\alpha_1(|x|_{x_*}) \leq V(x) \leq \alpha_2(|x|_{x_*}) \tag{14.5}$$

hold for all $x \in \mathbb{X}$, $u \in \mathbb{U}$ and an equilibrium $x_* \in \mathbb{X}$, then the closed loop system (14.2) is asymptotically stable on S in the sense of Definition 10.2.

Proof: In order to prove (14.4) consider $x \in S$ and the trajectory $x_\mu(\cdot)$ of (14.2) with $x_\mu(0) = x$. By forward invariance of the sets S this trajectory satisfies $x_\mu(k) \in S$. Hence from (14.3) for all $k \in \mathbb{N}_0$ we obtain

$$\alpha \ell(x_\mu(k), \mu(x_\mu(k))) \leq V(x_\mu(k)) - V(x_\mu(k+1)).$$

Summing over k yields for all $K \in \mathbb{N}$

$$\alpha \sum_{k=0}^{K-1} \ell(x_\mu(k), \mu(x_\mu(k))) \leq V(x_\mu(0)) - V(x_\mu(K)) \leq V(x)$$

since $V(x_\mu(K)) \geq 0$ and $x_\mu(0) = x$. Since the stage cost ℓ is nonnegative, the term on the left is monotone increasing and bounded, hence for $K \to \infty$ it converges to $\alpha J_\infty^{cl}(x,\mu)$. Since the right hand side is independent of K, this yields (14.4).

The stability assertion now immediately follows by observing that V satisfies all assumptions of Theorem 10.5 with $\alpha_V = \alpha\, \alpha_3$. □

14.2 Bounds on V_N

The central assumption we will use in order to ensure asymptotic stability and performance bounds imposes upper bounds on the optimal value functions V_N. These bounds are formulated relative to the stage cost ℓ. To this end, we define

$$\ell^*(x) := \inf_{u \in U} \ell(x,u). \tag{14.6}$$

With this notation, we can formulate our central assumption.

Assumption 14.2 [Bound on V_N] Consider the optimal control problem ($\mathrm{OCP_N}$). We assume that there exist functions $B_K \in \mathcal{K}_\infty$, $K \in \mathbb{N}$ such that for each $x \in \mathbb{X}$ the inequality

$$V_K(x) \leq B_K(\ell^*(x)) \tag{14.7}$$

holds for all $K \in \mathbb{N}$. □

We observe that $V_K(x) \geq \ell(x, u^\star(0)) \geq \ell^*(x)$ implies $B_K(r) \geq r$.

Before we state consequences from this assumption in the next section, we discuss a sufficient controllability condition which ensures Assumption 14.2. To this end, we first slightly enlarge the class of $\mathcal{KL}$-functions introduced in Definition 10.1.

Definition 14.3 We say that a continuous function $\beta : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is of class $\mathcal{KL}_0$ if for each $r > 0$ we have $\lim_{t\to\infty} \beta(r,t) = 0$ and for each $t \geq 0$ we either have $\beta(\cdot,t) \in \mathcal{K}_\infty$ or $\beta(\cdot,t) \equiv 0$. □

Compared to the class $\mathcal{KL}$, here we do not assume monotonicity in the second argument and we allow for $\beta(\cdot,t)$ being identically zero for some t. This allows for tighter bounds for the actual controllability behavior of the system. It is, however, easy to see that each $\beta \in \mathcal{KL}_0$ can be overbounded by a $\tilde{\beta} \in \mathcal{KL}$, e.g., by setting $\tilde{\beta}(r,t) = \max_{\tau\geq t} \beta(r,\tau) + e^{-t}r$. Using the $\mathcal{KL}_0$ functions we now formulate our controllability assumption.

Assumption 14.4 [Asymptotic controllability wrt. ℓ] Consider the optimal control problem ($\mathrm{OCP_N}$). We assume that the system is *asymptotically controllable with respect to ℓ with rate $\beta \in \mathcal{KL}_0$*, i.e., for each $x \in \mathbb{X}$ and each $N \in \mathbb{N}$ there exists an admissible control sequence $u_x \in \mathbb{U}^N(x)$ satisfying

$$\ell(x_{u_x}(n,x), u_x(n)) \leq \beta(\ell^*(x), n)$$

for all $n \in \{0,\dots,N-1\}$. □

An important special case for $\beta \in \mathcal{KL}_0$ is

$$\beta(r,n) = C\sigma^n r \tag{14.8}$$

for real constants $C \geq 1$ and $\sigma \in (0,1)$, i.e., *exponential controllability*.

The following lemma links Assumptions 14.4 and 14.2.

Lemma 14.5 If Assumption 14.4 holds then Assumption 14.2 holds. More precisely, for each $K \in \mathbb{N}$ and each $x \in \mathbb{X}$ the inequality

$$V_K(x) \leq J_K(x, u_x) \leq B_K(\ell^*(x)) \tag{14.9}$$

holds for u_x from Assumption 14.4 and

$$B_K(r) := \sum_{n=0}^{K-1} \beta(r,n). \tag{14.10}$$

Proof: The inequality follows immediately from

$$V_K(x) \leq J_K(x,u_x) \;=\; \sum_{n=0}^{K-1} \ell(x(n,u_x), u_x(n))$$

$$\leq \sum_{n=0}^{K-1} \beta(\ell^*(x), n) \;=\; B_K(\ell^*(x)).$$

□ □

In the special case (14.8) the values B_K, $K \in \mathbb{N}$, evaluate to

$$B_K(r) = C\frac{1-\sigma^K}{1-\sigma} r.$$

It is easily seen that if the state trajectories itself are exponentially controllable to some equilibrium x_* then exponential controllability, i.e., Assumption 14.4 with β from (14.8), holds if ℓ has polynomial growth. In particular, this covers the usual linear-quadratic setting for stabilizable systems.

However, even if the system itself is not exponentially controllable, exponential controllability in the sense of Assumption 14.4 can be achieved by proper choice of ℓ, as the following example shows.

Example 14.6 Consider the control system

$$x^+ = x + ux^3$$

with $\mathbb{X} = [-1, 1]$ and $U = [-1, 1]$. The system is controllable to $x_* = 0$, which can be seen by choosing $u = -1$. This results in the system $x^+ = x - x^3$ whose solutions approach $x_* = 0$ monotonically for $x_0 \in \mathbb{X}$.

However, the system it is not exponentially controllable to 0: exponential controllability would mean that there exist constants $C > 0$, $\sigma \in (0, 1)$ such that for each $x \in \mathbb{X}$ there is $u_x \in \mathbb{U}^\infty(x)$ with

$$|x_{u_x}(n, x)| \leq C\sigma^n |x|.$$

This implies that by choosing $n^* > 0$ so large such that $C\sigma^{n^*} \leq 1/2$ holds the inequality

$$|x_{u_x}(n^*, x)| \leq |x|/2 \tag{14.11}$$

must hold for each $x \in \mathbb{X}$. However, for each $x \geq 0$ the restriction $u \in [-1, 1]$ implies $x^+ \geq x - x^3 = (1 - x^2)x$ which by induction yields

$$x_u(n^*, x) \geq (1 - x^2)^{n^*} x$$

for all $u \in \mathbb{U}^\infty(x)$ which contradicts (14.11) for $x < 1 - 2^{-1/n^*}$.

On the other hand, since $|x| \leq 1$ we obtain $(1 - x^2)^2(2x^2 + 1) = 1 + 2x^6 - 3x^4 \leq 1$ which implies

$$\frac{1}{(1-x^2)^2} \geq 2x^2 + 1 \quad \Rightarrow \quad -\frac{1}{2x^2(1-x^2)^2} \leq -\frac{2x^2+1}{2x^2} = -1 - \frac{1}{2x^2}.$$

Hence, choosing

$$\ell(x, u) = \ell(x) = e^{-\frac{1}{2x^2}},$$

for $u \equiv -1$ we obtain

$$\ell(x^+) = \ell(x - x^3) = e^{-\frac{1}{2x^2(1-x^2)^2}} = e^{-\frac{1}{2x^2(1-x^2)^2}} \leq e^{-1}e^{-\frac{1}{2x^2}} = e^{-1}\ell(x).$$

By induction this implies Assumption 14.4 with β from (14.8) with $C = 1$ and $\sigma = e^{-1}$. □

For certain results it will be useful that β in Assumption 14.4 has the property

$$\beta(r, n+m) \leq \beta(\beta(r,n), m) \quad \text{for all } r \geq 0, n, m \in \mathbb{N}_0. \tag{14.12}$$

Inequality (14.12), often referred to as *submultiplicativity*, ensures that any sequence of the form $b_n = \beta(r,n)$, $r > 0$, also fulfills $b_{n+m} \leq \beta(b_n, m)$. It is, for instance, always satisfied in the exponential case (14.8). If needed, this property can be assumed without loss of generality, because by Sontag's $\mathcal{KL}$-Lemma [18, Proposition 7] the function β in Assumption 14.4 can be replaced by a β of the form $\beta(r,t) = \alpha_1(\alpha_2(r)e^{-t})$ for $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$. Then, (14.12) is easily verified if $\alpha_2 \circ \alpha_1(r) \geq r$, which is equivalent to $\alpha_1 \circ \alpha_2(r) \geq r$, which in turn is a necessary condition for Assumption 14.4 to hold for $n = 0$ and $\beta(r,t) = \alpha_1(\alpha_2(r)e^{-t})$.

14.3 Implications of the bounds on V_N

In this section we will use the bound on the V_N from Assumption 14.2 in order to establish two lemmas which yield bounds for optimal value functions and functionals along pieces of optimal trajectories. In the subsequent section, these bounds will then be used for the calculation of α in (14.3).

In order to be able to calculate α in (14.3), we will need an upper bound for $V_N(f(x, \mu_N(x)))$. To this end, recall from Step (3) of Algorithm 11.1 that $\mu_N(x_0)$ is the first element of the optimal control sequence $u^\star(\cdot)$ for ($\mathrm{OCP_N}$) with initial value x_0. In particular, this implies $f(x_0, \mu_N(x_0)) = x_{u^\star}(1, x_0)$. Hence, if we want to derive an upper bound for $V_N(f(x_0, \mu_N(x_0)))$ then we can alternatively derive an upper bound for $V_N(x_{u^\star}(1, x_0))$. This will be done in the following lemma.

Lemma 14.7 Suppose Assumption 14.2 holds and consider $x_0 \in \mathbb{X}$ and an optimal control $u^\star \in \mathbb{U}^N(x_0)$ for ($\mathrm{OCP_N}$). Then for each $j = 0, \ldots, N-2$ the inequality

$$V_N(x_{u^\star}(1, x_0)) \leq J_j(x_{u^\star}(1, x_0)), u^\star(1 + \cdot)) + B_{N-j}(\ell^*(x_{u^\star}(1 + j, x_0)))$$

holds for B_K from (14.7).

Proof: We define the control sequence

$$\tilde{u}(n) = \begin{cases} u^\star(1+n), & n \in \{0, \ldots, j-1\} \\ u_x(n-j), & n \in \{j, \ldots, N-1\}, \end{cases}$$

where u_x is an optimal control for initial value $x = x_{u^\star}(1 + j, x_0)$ and $N = N - j$. By construction, this control sequence is admissible for $x_{u^\star}(1, x_0)$ and we obtain

$$V_N(x_{u^\star}(1, x_0)) \leq J(x_{u^\star}(1, x_0), \tilde{u})$$

$$\begin{aligned} &= J_j(x_{u^\star}(1,x_0),u^\star(1+\cdot)) + J_{N-j}(x_{u^\star}(1+j,x_0),u_x) \\ &\le J_j(x_{u^\star}(1,x_0),u^\star(1+\cdot)) + B_{N-j}(\ell^*(x_{u^\star}(1+j,x_0))) \end{aligned}$$

where we used $J_{N-j}(x_{u^\star}(1+j,x_0),u_x) = V_{N-j}(x_{u^\star}(1+j,x_0))$ and Assumption 14.2 in the last step. This is the desired inequality. □

In words, the idea of this proof is as follows. The upper bound for each $j \in \{0,\dots,N-2\}$ is obtained from a specific trajectory. We follow the optimal trajectory for initial value x_0 and horizon N for j steps and for the point x reached this way we use the optimal control sequence for initial value x and horizon $N-j$ for another $N-j$ steps.

In the next lemma we derive upper bounds for the J_k-terms along tails of the optimal trajectory $x_{u^\star}$, which will later be used in order to bound the right hand side of the inequality from Lemma 14.7. To this end we use that these tails are optimal trajectories themselves.

Lemma 14.8 Suppose Assumption 14.2 holds and consider $x_0 \in \mathbb{X}$ and an optimal control $u^\star \in \mathbb{U}^N(x_0)$ for $(\mathrm{OCP_N})$. Then for each $k = 0,\dots,N-1$ the inequality

$$J_{N-k}(x_{u^\star}(k,x_0),u^\star(k+\cdot)) \le B_{N-k}(\ell^*(x_{u^\star}(k,x_0)))$$

holds for B_K from (14.7).

Proof: Corollary 12.3 implies $J_{N-k}(x_{u^\star}(k,x_0),u^\star(k+\cdot)) = V_{N-k}(x_{u^\star}(k,x_0))$. Hence the assertion follows immediately from Assumption 14.2. □

Remark 14.9 Since $u^\star \in \mathbb{U}^N(x_0)$ we obtain $x_{u^\star}(k,x_0) \in \mathbb{X}$ for $k = 0,\dots,N$. For $k = 0,\dots,N-1$ this property is crucial for the proof of Lemma 14.7 because it ensures that an optimal control for initial value $x = x_{u^\star}(1+j,x_0)$ exists. Note, however, that we do not need $x_{u^\star}(N,x_0) \in \mathbb{X}$. In fact, all results in this and the ensuing sections remain true if we remove the state constraint on $x_{u^\star}(N,x_0) \in \mathbb{X}$ from the definition of $\mathbb{U}^N(x_0)$ or replace it by some weaker constraint. □

14.4 Computation of α

We will now use the inequalities derived in the previous section in order to compute α for which (14.3) holds for all $x \in \mathbb{X}$. When trying to put together these inequalities in order to bound $V_N(x_{u^\star}(1,x_0))$ from above, one notices that the functionals in Lemma 14.7 and 14.8 are not exactly the same. Hence, in order to combine these results into a closed form which is suitable for computing α we need to look at the single terms of the stage cost ℓ contained in these functionals.

To this end, let $u^\star$ be an optimal control for $(\mathrm{OCP_N})$ with initial value $x_0 = x$. Then from the definition of V_N and μ_N it follows that (14.3) is equivalent to

$$\sum_{k=0}^{N-1} \ell(x_{u^\star}(k,x),u^\star(k)) \ge \alpha\ell(x,u^\star(0)) + V_N(x_{u^\star}(1,x)). \tag{14.13}$$

Thus, in order to compute α for which (14.3) holds for all $x \in \mathbb{X}$ we can equivalently compute α for which (14.13) holds for all optimal trajectories $x_{u^\star}(\cdot, x)$ with initial values $x \in \mathbb{X}$.

For this purpose we now consider arbitrary real values $\lambda_0, \ldots, \lambda_{N-1}, \nu \geq 0$ and start by deriving necessary conditions which hold if these values coincide with the cost along an optimal trajectory $\ell(x_{u^\star}(k, x), u^\star(k))$ and an optimal value $V_N(x_{u^\star}(1, x))$, respectively.

Proposition 14.10 Suppose Assumption 14.2 holds and consider $N \geq 1$, values $\lambda_n \geq 0$, $n = 0, \ldots, N-1$, and a value $\nu \geq 0$. Consider $x \in \mathbb{X}$ and assume that there exists an optimal control sequence $u^\star \in \mathbb{U}^N(x)$ for ($\mathrm{OCP_N}$) such that

$$\lambda_k = \ell(x_{u^\star}(k, x), u^\star(k)), \quad k = 0, \ldots, N-1$$

holds. Then

$$\sum_{n=k}^{N-1} \lambda_n \leq B_{N-k}(\lambda_k), \quad k = 0, \ldots, N-2 \tag{14.14}$$

holds. If, furthermore,

$$\nu = V_N(x_{u^\star}(1, x))$$

holds then

$$\nu \leq \sum_{n=0}^{j-1} \lambda_{n+1} + B_{N-j}(\lambda_{j+1}), \quad j = 0, \ldots, N-2 \tag{14.15}$$

holds. □

Proof: If the stated conditions hold, then λ_n and ν must meet the inequalities given in Lemmas 14.7 and 14.8, which is exactly (14.15) and (14.14). □

Using this proposition we can give a sufficient condition for (14.13) and thus for (14.3). The idea behind the following proposition is to express the terms in inequality (14.13) using the values $\lambda_0, \ldots, \lambda_{N-1}$ and ν introduced above.

Proposition 14.11 Consider $N \geq 1$ and $B_K \in \mathcal{K}_\infty$, $K = 2, \ldots, N$ and assume that all values $\lambda_n \geq 0$, $n = 0, \ldots, N-1$ and $\nu \geq 0$ fulfilling (14.14) and (14.15) satisfy the inequality

$$\sum_{n=0}^{N-1} \lambda_n - \nu \geq \alpha \lambda_0 \tag{14.16}$$

for some $\alpha \in (0, 1]$. Then for this α and each optimal control problem ($\mathrm{OCP_N}$) satisfying Assumption 14.2 inequality (14.3) holds for μ_N from Algorithm 11.1 and all $x \in \mathbb{X}$. □

Proof: Consider a control system satisfying Assumption 14.2 and an optimal control sequence $u^\star \in \mathbb{U}^N(x)$ for initial value $x \in \mathbb{X}$. Then by Proposition 14.10 the values $\lambda_k = \ell(x_{u^\star}(k, x), u^\star(k))$ and $\nu = V_N(x_{u^\star}(1, x))$ satisfy (14.14) and (14.15), hence by assumption also (14.16). Thus, using $\ell(x, u^\star(0)) = \ell(x_{u^\star}(0, x), u^\star(0)) = \lambda_0$ we obtain

$$V_N(x_{u^\star}(1, x)) + \alpha \ell(x, u^\star(0)) = \nu + \alpha \lambda_0 \quad \leq \quad \sum_{k=0}^{N-1} \lambda_k = \sum_{k=0}^{N-1} \ell(x_{u^\star}(k, x), u^\star(k)).$$

This proves (14.13) and thus also (14.3). □

Proposition 14.11 is the basis for computing α as specified in the following theorem.

Theorem 14.12 [Abstract optimization problem] Consider $N \geq 1$ and $B_K \in \mathcal{K}_\infty$, $K = 2, \ldots, N$ and assume that the optimization problem

$$\begin{aligned} \alpha := \inf_{\lambda_0,\ldots,\lambda_{N-1},\nu} & \frac{\sum_{n=0}^{N-1} \lambda_n - \nu}{\lambda_0} \\ & \text{subject to the constraints (14.14), (14.15), and} \\ & \lambda_0 > 0, \lambda_1, \ldots, \lambda_{N-1}, \nu \geq 0 \end{aligned} \tag{14.17}$$

has an optimal value $\alpha \in (0, 1]$. Then for this α and each optimal control problem ($\mathrm{OCP_N}$) satisfying Assumption 14.2 inequality (14.3) holds for μ_N from Algorithm 11.1 and all $x \in \mathbb{X}$.

Proof: Consider arbitrary values $\lambda_0, \ldots, \lambda_{N-1}, \nu \geq 0$ satisfying (14.14) and (14.15).

If $\lambda_0 > 0$ then the definition of Problem (14.17) immediately implies (14.16).

If $\lambda_0 = 0$, then inequality (14.14) for $k = 0$ together with $B_K(0) = 0$ implies $\lambda_1, \ldots, \lambda_{N-1} = 0$. Thus, (14.15) for $j = 1$ yields $\nu = 0$ and again (14.16) holds.

Hence, (14.16) holds in both cases and Proposition 14.11 yields the assertion. □

Remark 14.13 (i) Theorem 14.12 shows Inequality (14.3) for all $x \in \mathbb{X}$ if Assumption 14.2 or, alternatively, Assumption 14.4 holds for all $x \in \mathbb{X}$ and $K = 2, \ldots, N$.

If we want to establish Inequality (14.3) only for states $x_0 \in Y$ for a subset $Y \subset \mathbb{X}$, then from the proofs of the Lemmas 14.7 and 14.8 it follows that Proposition 14.10 holds for all $x_0 \in Y$ (instead of for all $x_0 \in \mathbb{X}$) under the following condition:

$$\begin{aligned} & \text{(14.7) holds for } x = x_{u^\star}(k, x_0) \text{ for all } k = 0, \ldots, N-1, \text{ all } x_0 \in Y \\ & \text{and all } K = 2, \ldots, N, \text{ where } u^\star \text{ is the optimal control for } J_N(x_0, u). \end{aligned} \tag{14.18}$$

This implies that under condition (14.18) Theorem 14.12 holds for all $x_0 \in Y$ and consequently (14.3) holds for all $x_0 \in Y$.

(ii) A further relaxation of the assumptions of Theorem 14.12 can be obtained by observing that if we are interested in establishing Inequality (14.3) only for states $x_0 \in Y$, then in (14.17) we only need to optimize over those λ_i which correspond to optimal trajectories starting in Y. In particular, if we know that $\inf_{x_0 \in Y} \ell^*(x_0) \geq \zeta$ for some $\zeta > 0$, then the constraint $\lambda_0 > 0$ can be tightened to $\lambda_0 \geq \zeta$. □

The following lemma shows that the optimization problem (14.17) specializes to a linear program if the functions $B_K(r)$ are linear in r.

Lemma 14.14 If the functions $B_K(r)$ from (14.7) in the constraints (14.14), (14.15) are linear in r, then α from Problem (14.17) coincides with

$$\begin{aligned} \alpha := & \min_{\lambda_0,\dots,\lambda_{N-1},\nu} \sum_{n=0}^{N-1} \lambda_n - \nu \\ & \text{subject to the (now linear) constraints (14.14), (14.15), and} \\ & \lambda_0 = 1, \lambda_1, \dots, \lambda_{N-1}, \nu \geq 0. \end{aligned} \tag{14.19}$$

In particular, this holds if Assumption 14.4 holds with functions $\beta(r,t)$ being linear in r.

Proof: Due to the linearity, all sequences $\bar{\lambda}_0, \dots, \bar{\lambda}_{N-1}, \bar{\nu}$ satisfying the constraints in (14.17) can be written as $\gamma\lambda_0, \dots, \gamma\lambda_{N-1}, \gamma\nu$ for some $\lambda_0, \dots, \lambda_{N-1}, \nu$ satisfying the constraints in (14.19), where $\gamma = \bar{\lambda}_0$. Since

$$\frac{\sum_{n=0}^{N-1} \bar{\lambda}_n - \bar{\nu}}{\bar{\lambda}_0} = \frac{\sum_{n=0}^{N-1} \gamma\lambda_n - \gamma\nu}{\gamma\lambda_0} = \frac{\sum_{n=0}^{N-1} \lambda_n - \nu}{\lambda_0} = \sum_{n=0}^{N-1} \lambda_n - \nu,$$

the values α in Problems (14.17) and (14.19) coincide. □

The next result gives an explicit bound for Problem (14.19) and thus also (14.17) if the functions B_K are linear.

Proposition 14.15 If the functions $B_K(r)$ from (14.7) in the constraints (14.14), (14.15) are linear in r, then the solution of Problems (14.17) and (14.19) satisfies the inequality

$$\alpha \geq \tilde{\alpha}_N \tag{14.20}$$

for

$$\tilde{\alpha}_N := 1 - (\gamma_2 - 1)(\gamma_N - 1) \prod_{k=2}^{N-1} \left(\frac{\gamma_k - 1}{\gamma_k} \right) \quad \text{with } \gamma_k = B_k(r)/r. \tag{14.21}$$

□

Proof: We prove the theorem by showing the inequality

$$\lambda_{N-1} \leq (\gamma_N - 1) \prod_{k=2}^{N-1} \left(\frac{\gamma_k - 1}{\gamma_k} \right) \lambda_0 \tag{14.22}$$

for all feasible $\lambda_0, \dots, \lambda_{N-1}$. From this (14.20) follows since (14.15) with $j = N-2$ implies

$$\nu \leq \sum_{n=1}^{N-2} \lambda_n + \gamma_2 \lambda_{N-1}$$

and thus (14.22), $\gamma_2 \geq 1$ and $\lambda_0 = 1$ yield

$$\sum_{n=0}^{N-1} \lambda_n - \nu \geq \lambda_0 + (1 - \gamma_2)\lambda_{N-1} \geq \lambda_0 - (\gamma_2 - 1)(\gamma_N - 1) \prod_{k=2}^{N-1} \left(\frac{\gamma_k - 1}{\gamma_k} \right) \lambda_0 = \tilde{\alpha}_N$$

for all feasible $\lambda_1, \dots, \lambda_{N-1}$ and ν, which yields $\alpha \geq \tilde{\alpha}_N$.

In order to prove (14.22), we start by observing that (14.14) with $j = p$ implies

$$\sum_{k=p+1}^{N-1} \lambda_k \leq (\gamma_{N-p} - 1)\lambda_p \tag{14.23}$$

for $p = 0, \dots, N-2$. From this we can conclude

$$\lambda_p + \sum_{k=p+1}^{N-1} \lambda_k \geq \frac{\sum_{k=p+1}^{N-1} \lambda_k}{\gamma_{N-p} - 1} + \sum_{k=p+1}^{N-1} \lambda_k = \frac{\gamma_{N-p}}{\gamma_{N-p} - 1} \sum_{k=p+1}^{N-1} \lambda_k.$$

Using this inequality inductively for $p = 1, \dots, N-2$ yields

$$\sum_{k=1}^{N-1} \lambda_k \geq \prod_{k=1}^{N-2} \left(\frac{\gamma_{N-k}}{\gamma_{N-k} - 1} \right) \lambda_{N-1} = \prod_{k=2}^{N-1} \left(\frac{\gamma_k}{\gamma_k - 1} \right) \lambda_{N-1}.$$

Using (14.23) for $p = 0$ we then obtain

$$(\gamma_N - 1)\lambda_0 \geq \sum_{k=1}^{N-1} \lambda_k \geq \prod_{k=2}^{N-1} \left(\frac{\gamma_k}{\gamma_k - 1} \right) \lambda_{N-1}$$

which implies (14.22). □

A much more complicated proof (see [8, Proposition 6.18]) shows that the optimal α_N is given by

$$\alpha_N := 1 - \frac{(\gamma_N - 1) \prod_{k=2}^{N} (\gamma_k - 1)}{\prod_{k=2}^{N} \gamma_k - \prod_{k=2}^{N} (\gamma_k - 1)} \quad \text{with } \gamma_k = B_k(r)/r, \tag{14.24}$$

A comparison of the two formulas (14.24) and (14.20) can be found in Remark 14.17, below.

14.5 Main Stability and Performance Results

We are now ready to state our main result on stability and performance of stabilizing MPC without terminal conditions.

Theorem 14.16 [Stability without terminal conditions] Consider the MPC Algorithm 11.1 with optimization horizon $N \in \mathbb{N}$ and stage cost ℓ satisfying $\alpha_3(|x|_{x_*}) \leq \ell^*(x) \leq \alpha_4(|x|_{x_*})$ for suitable $\alpha_3, \alpha_4 \in \mathcal{K}_\infty$. Suppose that Assumption 14.2 holds and that $\alpha = \alpha_N$ from Formula (14.24) or $\alpha = \tilde{\alpha}_N$ from Formula (14.20) satisfies $\alpha \in (0, 1]$. Then the MPC closed loop system (11.2) with MPC-feedback law μ_N is asymptotically stable on $\mathbb{X}$.

In addition, the inequality

$$J_\infty^{cl}(x, \mu_N) \leq V_N(x)/\alpha \leq V_\infty(x)/\alpha$$

holds for each $x \in \mathbb{X}$.

Proof: First note that $V_N \leq V_\infty$ follows immediately from $\ell \geq 0$. Hence, the assertion follows readily from Theorem 14.1 if we prove the inequalities (14.3) and (14.5). Inequality (14.3) follows directly from Theorem 14.12 and Proposition 14.15 or [8, Proposition 6.18].

Regarding (14.5), observe that the inequality for ℓ follows immediately from our assumptions. From the definition of V_N we get

$$V_N(x) = \inf_{u \in \mathbb{U}^N(x)} J_N(x,u) \geq \inf_{u \in \mathbb{U}^N(x)} \ell(x,u(0)) = \ell^*(x) \geq \alpha_3(|x|_{x_*}),$$

thus the lower inequality for V_N follows with $\alpha_1 = \alpha_3$. The upper inequality in (14.5) follows from Assumption 14.2 and the upper bound on ℓ^* via

$$V_N(x) \leq B_N(\ell^*(x)) \leq B_N(\alpha_4(|x|_{x_*})),$$

i.e., for $\alpha_2 = B_N \circ \alpha_4$. □

Remark 14.17 Let us compare the two different bounds on α given by $\tilde{\alpha}_N$ from (14.20) and α_N from (14.24). In order to illustrate that the criterion $\tilde{\alpha}_N > 0$ is more conservative than the criterion $\alpha_N > 0$, we consider the case where $\gamma_k = \gamma$ for all k, i.e., the γ_k are independent of k, and compute the minimal N for which $\tilde{\alpha}_N > 0$ and $\alpha_N > 0$, respectively, hold. For $\gamma_k = \gamma$ the expressions simplify to

$$\tilde{\alpha}_N = 1 - \frac{(\gamma-1)^N}{\gamma^{N-2}} \quad \text{and} \quad \alpha_N = 1 - \frac{(\gamma-1)^N}{\gamma^{N-1} - (\gamma-1)^{N-1}}.$$

Thus, an optimization horizon N for which $\tilde{\alpha}_N > 0$ must satisfy

$$N > 2 + 2\frac{\ln \gamma}{\ln \gamma - \ln(\gamma-1)}$$

while the same condition for $\tilde{\alpha}_N > 0$ is given by

$$N > 2 + \frac{\ln(\gamma-1)}{\ln \gamma - \ln(\gamma-1)}.$$

This means that the estimate for the minimal stabilizing horizon based on $\tilde{\alpha}_N$ is about twice as large as the estimate based on α_N.

In this context, it is interesting to look at the asymptotic behavior of the bounds on N for $\gamma \to \infty$. For large γ the denominator is approximately $1/\gamma$. This implies that asymptotically for $\gamma \to \infty$ the first estimate for N behaves like $2\gamma \ln \gamma$ while the second behaves like $\gamma \ln \gamma$.

□

The class of systems which is covered by Theorem 14.16 is quite large, since, e.g., exponential controllability holds on compact sets $\mathbb{X}$ whenever the linearization of f in x_* is stabilizable and ℓ is quadratic.

The following simple example illustrates the use of Theorem 14.16 for the case of a nonexponentially controllable system.

Example 14.18 We reconsider Example 14.6, i.e.,

$$x^+ = x + ux^3 \quad \text{with} \quad \ell(x,u) = e^{-\frac{1}{2x^2}}.$$

As shown in Example 14.6, Assumption 14.4 holds with $\beta(r,k) = C\sigma^k r$ with $C = 1$ and $\sigma = e^{-1}$. The bounds in Assumption 14.2 resulting from this β according to (14.10) are

$$B_K(r) = C\frac{1-\sigma^K}{1-\sigma}r = C\frac{1-e^{-K}}{1-e^{-1}}r,$$

thus Theorem 14.16 is applicable and we obtain $\alpha \geq \alpha_N$ with α_N from Formula (14.24). The γ_k in Formula (14.24) are given by

$$\gamma_k = C\frac{1-e^{-k}}{1-e^{-1}}.$$

A straightforward computation reveals that for these values Formula (14.24) simplifies to

$$1 - \frac{(\gamma_N - 1)\prod_{k=2}^{N}(\gamma_k - 1)}{\prod_{k=2}^{N}\gamma_k - \prod_{k=2}^{N}(\gamma_k - 1)} = 1 - e^{-N}.$$

Hence, for $N = 2$ we obtain $\alpha = 1 - e^{-2} \approx 0.865$ and for $N = 3$ we get $\alpha \geq 1 - e^{-3} \approx 0.95$. Hence, Theorem 14.16 ensures asymptotic stability for all $N \geq 2$ and — since $1/0.95 \approx 1.053$ — for $N = 3$ the performance of the MPC controller is at most about 5.3% worse than the infinite horizon controller. □

While in this simple example the computation of α via Formula (14.24) is possible, in many practical examples this will not be the case. However, Formula (14.24) can still be used to obtain valuable information for the design of MPC schemes. This aspect will be discussed at the end of this section.

Although the main benefit of the approach developed in this chapter compared to other approaches is that we can get rather precise quantitative estimates, it is nevertheless good to know that our approach also guarantees asymptotic stability for sufficiently large optimization horizons N under suitable assumptions. This is the statement of our final stability result.

Theorem 14.19 [Stability for sufficiently large N] Consider the MPC Algorithm 11.1 with optimization horizon $N \in \mathbb{N}$ and stage cost ℓ satisfying $\alpha_3(|x|_{x_*}) \leq \ell^*(x) \leq \alpha_4(|x|_{x_*})$ for suitable $\alpha_3, \alpha_4 \in \mathcal{K}_\infty$. Suppose that Assumption 14.2 holds for linear $B_K \in \mathcal{K}_\infty$ of the form $B_K(r) = \gamma_K r$ with $\gamma_\infty := \sup_{k\in\mathbb{N}} \gamma_k < \infty$.

Then the MPC closed loop system (11.2) with MPC-feedback law μ_N is asymptotically stable on $\mathbb{X}$ provided N is sufficiently large.

Furthermore, for each $C > 1$ there exists $N_C > 0$ such that

$$J_\infty^{cl}(x, \mu_N) \leq CV_N(x) \leq CV_\infty(x)$$

holds for each $x \in \mathbb{X}$ and each $N \geq N_C$.

Proof: The assertion follows immediately from Theorem 14.16 if we show that $\tilde{\alpha}_N \to 1$ holds in (14.20) as $N \to \infty$. Since all factors in (14.20) are monotone increasing in γ_k and the product has a negative sign, we obtain

$$\tilde{\alpha}_N \geq 1 - (\gamma_\infty - 1)^2 \left(\frac{\gamma_\infty - 1}{\gamma_\infty} \right)^{N-2}.$$

Since $(\gamma_\infty - 1)/\gamma_\infty < 1$ we obtain that

$$\left(\frac{\gamma_\infty - 1}{\gamma_\infty} \right)^{N-2} \to 0$$

as $N \to \infty$ and thus $\tilde{\alpha}_N \to 1$. □

Remark 14.20 For B_K of the form (14.10), a sufficient condition for the γ_k being bounded by γ_∞ is that Assumption 14.4 holds for a $\beta \in \mathcal{KL}_0$ which is linear in its first argument and is summable, i.e.,

$$\sum_{k=0}^{\infty} \beta(r, k) < \infty \quad \text{for all } r > 0.$$

□

Theorem 14.19 justifies what is often done in practice: we set up an MPC scheme using a reasonable stage cost ℓ and increase N until the closed loop system becomes stable.

Of course, Theorem 14.19 immediately leads to the question how large the optimization horizon N needs to be for achieving stability or a certain performance. As the computational cost grows with the length of a horizon, this is also important for the practical implementability of the MPC scheme. We investigate this question for the case that the asymptotic controllability condition from Assumption 14.4 holds with the exponential functions $\beta(r, n) = C\sigma^N r$ from (14.8). To this end, we look at the minimal horizon N for which α_N is larger than a certain threshold depending on the parameters C and σ. This dependence is illustrated in Figure 14.5 for thresholds 0 and 0.5.

As we see, the two parameters C and σ play a very different role. While for fixed $\sigma > 0$ it is always possible to reduce the necessary horizon to $N = 2$, i.e., to the shortest possible horizon, by making C smaller, this is not possible for fixed C by reducing σ. Hence, the constant C plays a more important role for obtaining stability and performance with small optimization horizon N. Particularly, any tuning of the stage cost ℓ which leads to a reduction of C is likely also to reduce the necessary optimization horizon. In the lecture, it will be shown how this observation can explain the parameter dependence of the stability behavior of the third example in Section 9.1.

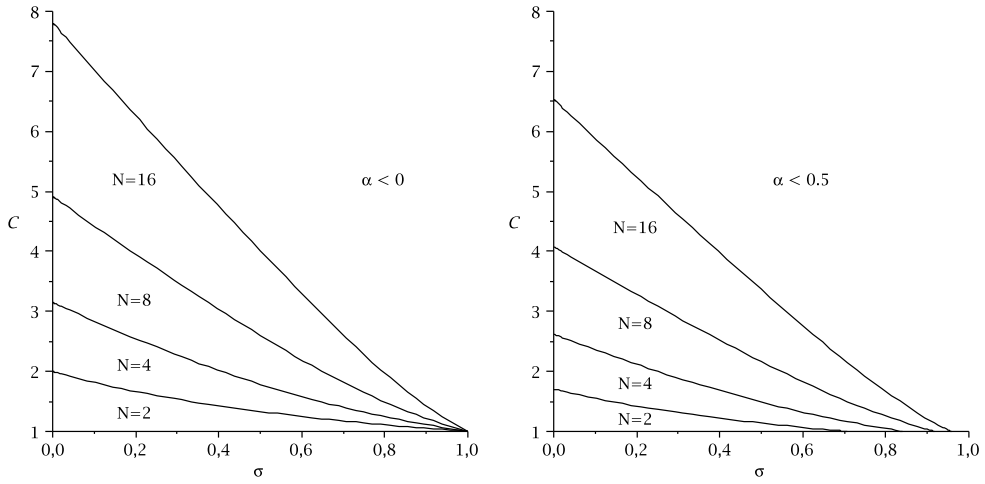


Abbildung 14.1: Suboptimality regions for different optimization horizons N depending on C and σ in (14.8) for $\alpha_N > 0$ (left) and $\alpha_N > 0.5$ (right)

Literaturverzeichnis

[1] D. Angeli, R. Amrit, and J. B. Rawlings, *On average performance and stability of economic model predictive control*, IEEE Trans. Autom. Control, 57 (2012), pp. 1615–1626.

[2] F. Colonius, *Einführung in die Steuerungstheorie*. Vorlesungsskript, Universität Augsburg, 1992, eine aktuelle Version ist erhältlich unter dem Link "Lehre" auf `scicomp.math.uni-augsburg.de/~colonius/`.

[3] J. Doležal, *Existence of optimal solutions in general discrete systems*, Kybernetika, 11 (1975), pp. 301–312.

[4] T. Faulwasser and D. Bonvin, *On the design of economic NMPC based on approximate turnpike properties*, in Proceedings of the 54th IEEE Conference on Decision and Control — CDC 2015, 2015, pp. 4964–4970.

[5] L. Grüne, *Stabilität und Stabilisierung linearer Systeme*. Vorlesungsskript, Universität Bayreuth, 2003, `www.math.uni-bayreuth.de/~lgruene/linstab0203/`.

[6] L. Grüne, *Economic receding horizon control without terminal constraints*, Automatica, 49 (2013), pp. 725–734.

[7] L. Grüne and O. Junge, *Gewöhnliche Differentialgleichungen. Eine Einführung aus der Perspektive der Dynamischen Systeme*, Springer Spektrum, 2. aktualisierte auflage ed., 2016.

[8] L. Grüne and J. Pannek, *Nonlinear Model Predictive Control. Theory and Algorithms*, Springer-Verlag, London, 2nd ed., 2017.

[9] L. Grüne and M. Stieler, *Asymptotic stability and transient optimality of economic MPC without terminal conditions*, J. Proc. Control, 24 (2014), pp. 1187–1196.

[10] W. Hahn, *Stability of Motion*, Springer–Verlag Berlin, Heidelberg, 1967.

[11] D. Hinrichsen and A. J. Pritchard, *Mathematical systems theory I*, vol. 48 of Texts in Applied Mathematics, Springer, Heidelberg, 2010. Modelling, state space analysis, stability and robustness, Corrected reprint [of MR2116013].

[12] S. S. Keerthi and E. G. Gilbert, *An existence theorem for discrete-time infinite horizon optimal control problems*, IEEE Trans. Automat. Contr., 30 (1985), pp. 907–909.

[13] Y. Lin, E. D. Sontag, and Y. Wang, *A smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.

[14] J. Lunze, *Regelungstechnik 1*, Springer, 10 ed., 2010.

[15] M. A. Müller, D. Angeli, and F. Allgöwer, *On necessity and robustness of dissipativity in economic model predictive control*, IEEE Trans. Autom. Control, 60 (2015), pp. 1671–1676.

[16] M. A. Müller and L. Grüne, *Economic model predictive control without terminal constraints for optimal periodic behavior*, Automatica, 70 (2016), pp. 128–139.

[17] E. D. Sontag, *A "universal" construction of Artstein's theorem on nonlinear stabilization*, Systems Control Lett., 13 (1989), pp. 117–123.

[18] E. D. Sontag, *Comments on integral variants of ISS*, Syst. Control Lett., 34 (1998), pp. 93–100.

[19] ——, *Mathematical Control Theory*, Springer Verlag, New York, 2nd ed., 1998.