Numerische Dynamik von Kontrollsystemen

Lars Grüne
Mathematisches Institut
Fakultät für Mathematik und Physik
Universität Bayreuth
95440 Bayreuth
lars.gruene@uni-bayreuth.de
www.uni-bayreuth.de/departments/math/∼lgruene/

Vorlesungsskript

Sommersemester 2004

Vorwort

Dieses Skript ist im Rahmen einer gleichnamigen Vorlesung entstanden, die ich im Sommersemester 2004 an der Universität Bayreuth gehalten habe. Für viele Verbessrungsvorschläge und Korrekturen möchte ich mich bei den Teilnehmerinnen und Teilnehmern der Vorlesung herzlich bedanken.

Eine elektronische Version dieses Skripts sowie die zu dieser Vorlesung gehörigen Übungsaufgaben finden sich im WWW unter dem Link "Lehrveranstaltungen" auf der Seite `http://www.uni-bayreuth.de/departments/math/~lgruene/`.

Bayreuth, August 2004 Lars Grüne

Inhaltsverzeichnis

Kapitel 1

Kontrollsysteme

1.1 Einleitung

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 als Störung, die auf das System wirkt (z.B. Straßenunebenheiten bei einem Auto, Kursschwankungen bei Wechselkursen). Die erste Interpretation entspricht dem Konzept des *Kontrollsystems*, die zweite einem *gestörten System*. Da diese zwei Konzepte mathematisch mit der gleichen Struktur formalisiert werden, werden wir hier durchgehend den Ausdruck „Kontrollsystem“ verwenden. Dieser Begriff hat sich im deutschen Sprachgebrauch hierfür inzwischen etabliert, wenngleich er eine eher schlechte, oder zumindest missverständliche Übersetzung des englischen Ausdrucks „control system“ darstellt. Eine korrektere Übersetzung wäre „gesteuertes System“ oder „Steuersystem“, da es hier um Kontrolle im Sinne von Einflussnahme und nicht im Sinne von Überwachung geht. Wir wollen hier aber bei der geläufigen Bezeichnung bleiben.

Wie bei den dynamischen Systemen, interessiert man sich auch bei Kontrollsystemen für das Langzeitverhalten der Lösungen, also z.B. für die Existenz von asymptotisch stabilen Mengen (z.B. Gleichgewichten) oder für ihre Einzugsbereiche. Durch den zusätzlichen „Freiheitsgrad“, den der Parameter u liefert, ist die Sache hier allerdings etwas komplizierter. Kommt man von den dynamischen Systemen, so ist die konzeptionell ähnlichste Situation dann gegeben, wenn für jeden Anfangswert x_0 eine Kontrollfunktion $u(t)$ fest vorgegeben ist. In diesem Fall kann man viele der üblichen Konzepte dynamischer Systeme recht einfach übertragen. Eine wichtige Methode zur Auswahl eines solchen u ist die *optimale Steuerung*, bei der ein Optimalitätskriterium definiert wird, für das dann die optimalen Lösungen betrachtet werden. Dieses Problem werden wir in der ersten Hälfte der Vorlesung betrachten, wobei der Schwerpunkt auf der Entwicklung eines numerischen Verfahrens zur *globalen* Lösung optimaler Steuerungsprobleme liegt. Diese globale Lösung wird es uns dann ermöglichen, das dynamische Verhalten des optimal gesteuerten Systems direkt aus der Numerik abzulesen.

Im zweiten Teil der Vorlesung werden wir den Parameter u dann „frei“ lassen und zwei weitere Fälle betrachten: Zum einen werden wir das dynamische Verhalten für *alle möglichen* (variierenden) u untersuchen. Zum anderen werden wir (numerische) Verfahren kennen lernen, mit denen man Funktionen u bestimmen kann, die ein *gewünschtes* Langzeitverhalten (z.B. asymptotische Stabilität einer vorgegebenen Menge) erzielen. Auch hier wird die optimale Steuerung eine Rolle spielen, darüberhinaus werden wir aber auch mengenwertige numerische Verfahren einsetzten.

1.2 Definition

In diesem Abschnitt wollen wir die grundlegenden Systeme definieren, mit denen wir uns in dieser Vorlesung beschäftigen wollen. Wir betrachten hierbei zunächst zeitabhängige Kontrollfunktionen $u(t)$; zustandsabhängige Parameter werden wir zu gegebener Zeit einführen.

Definition 1.1 (i) Ein *Kontrollsystem* in *kontinuierlicher Zeit* $\mathbb{T} = \mathbb{R}$ im $\mathbb{R}^d$, $d \in \mathbb{N}$, ist gegeben durch die gewöhnliche Differentialgleichung

$$\frac{d}{dt}x(t) = f(x(t), u(t)), \tag{1.1}$$

wobei $f : \mathbb{R}^d \times U \to \mathbb{R}^d$ ein *parameterabhängiges stetiges Vektorfeld* ist.

(ii) Ein *Kontrollsystem* in *diskreter Zeit* $\mathbb{T} = h\mathbb{Z} = \{hk \,|\, k \in \mathbb{Z}\}$ für ein $h > 0$ im $\mathbb{R}^d$, $d \in \mathbb{N}$, ist gegeben durch die Differenzengleichung

$$x(t+h) = f_h(x(t), u(t)), \tag{1.2}$$

wobei $f_h : \mathbb{R}^d \times U \to \mathbb{R}^d$ eine *stetige Abbildung* ist.

(iii) Die Menge $U \subseteq \mathbb{R}^m$ heißt *Kontrollwertebereich*, und definiert die Werte, die $u(t)$ für $t \in \mathbb{R}$ annehmen darf. U wird in dieser Vorlesung üblicherweise kompakt sein.

(iv) Mit $\mathcal{U}$ bzw. $\mathcal{U}_h$ bezeichnen wir den *Raum der zulässigen Kontrollfunktionen*, also

$$\mathcal{U} := \{u : \mathbb{R} \to U \,|\, u \text{ zulässig}\} \quad \text{bzw.} \quad \mathcal{U}_h := \{u_h : h\mathbb{Z} \to U \,|\, u \text{ zulässig}\}$$

Welche Klassen von Funktionen wir als „zulässig“ definieren, werden wir im folgenden Abschnitt festlegen. □

Bemerkung 1.2 Statt „$\frac{d}{dt}x(t)$“ werden wir meist kurz „$\dot{x}(t)$“ schreiben. □

1.3 Beispiele

Wir wollen einige Beispiele von Kontrollsystemen aus verschiedenen Anwendungsbereichen betrachten.

Beispiel 1.3 Wir betrachten zunächst ein einfaches mechanisches Beispiel. Ein Wagen, der entlang einer festen Führungsschiene fahren kann, lässt sich mit Position $x_1(t)$, Geschwindigkeit $x_2(t)$ und Beschleunigung $a(t)$ beschreiben durch

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= a(t) \end{aligned}$$

Setzen wir $a(t) = u_1(t) - (r + u_2(t))x_2(t)$ mit $u = (u_1, u_2)^T \in U = \mathbb{R} \times \mathbb{R}_0^+$, so modelliert die Kontrolle u_1 hier die externe Beschleunigung, die z.B. durch einen Motor erzeugt werden kann, $r > 0$ ist ein Reibungsfaktor und u_2 modelliert eine Bremse. Modelle dieser Art werden z.B. in der energieoptimalen Steuerung von Schienenfahrzeugen verwendet, wir kommen darauf im nächsten Kapitel zurück. □

Beispiel 1.4 Die einfachste Modellierung von Kapitalströmen in einem Betrieb ist gegeben durch das eindimensionale lineare System

$$\dot{x}(t) = u(t) - \sigma x(t),$$

für einen konstanten Parameter $\sigma > 0$, bei dem x_1 das in einem Betrieb investierte Kapital und u die getätigten Investitionen modelliert[1] Ein etwas allgemeineres Modell ist gegeben durch

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) - \sigma x_1(t) \\ \dot{x}_2(t) &= u \end{aligned}$$

bei dem die Kontrolle nun nicht mehr die Investitionen sondern die Änderung der Investitionen beschreibt. Dieses erweiterte Modell erlaubt im Rahmen der optimalen Steuerung die Modellierung von Umstrukturierungskosten, indem betragsmäßig große Werte von u „bestraft“ werden. Trotz der sehr einfachen Dynamik zeigt dieses Modell im Rahmen der optimalen Steuerung ein interessantes dynamisches Verhalten. □

Beispiel 1.5 Ein weiteres Beispiel, das wir betrachten wollen, ist ein einfaches Modell für ein Ökosystem, ein sogenanntes *Räuber–Beute–Modell*.[2] Es ist gegeben durch die Gleichungen

$$\begin{aligned} \dot{x}_1(t) &= (a_0 - a_2 x_2(t) - a_1 x_1(t) - u(t))x_1(t) \\ \dot{x}_2(t) &= (b_1 x_1(t) - b_0 - b_2 x_2(t) - u(t))x_2(t) \end{aligned}$$

mit konstanten Parametern a_i, $b_i > 0$. Hierbei bezeichnen x_1 und x_2 die Größe der Populationen zweier Spezies in einem begrenzten Lebensraum (wir können uns z.B. zwei Fischarten in einem großen See vorstellen), wobei die zweite Spezies („Räuber“) die erste („Beute“) jagt. Beide werden wiederum von Fischern aus dem See gefangen, wobei die Kontrolle u die Fangrate der Fischer beschreibt. □

[1] In den Wirtschaftswissenschaften wird Kapital meist mit k und Investitionen mit i bezeichnet. Wir wollen hier allerdings bei unseren Standardbezeichnungen für Kontrollsysteme bleiben.

[2] Dieses spezielle Modell stammt aus der Arbeit [19].

Bemerkung 1.6 Bei all diesen Beispielen kann man o.B.d.A annehmen, dass die Menge U kompakt ist, da z.B. weder eine beliebig schnelle Beschleunigung noch beliebig hohe Investitionen noch eine beliebig hohe Fangrate praktisch realisierbar sind. In der numerischen Lösung optimaler Steuerungsprobleme werden wir U immer als kompakt voraussetzen. □

1.4 Ein Existenz– und Eindeutigkeitssatz

Wir werden uns nun damit beschäftigen, welche Wahl des Kontrollfunktionsraumes $\mathcal{U}$ bzw. $\mathcal{U}_h$ sinnvoll ist. Bei der Auswahl von $\mathcal{U}$ spielen zwei Kriterien eine Rolle: Zum einen wollen wir eine hinreichend große Menge an Funktionen zulassen, zum anderen wollen wir eine Existenz– und Eindeutigkeitsaussage für die Lösungen von (1.1) bzw. (1.2) erhalten.

Im zeitdiskreten Fall ist dies einfach, wir lassen für $\mathcal{U}_h$ *alle* möglichen Funktionen u_h von $h\mathbb{Z}$ nach U zu, also

$$\mathcal{U}_h := \{u_h : h\mathbb{Z} \to U\}.$$

Per Induktion sieht man leicht, dass dann für jeden Anfangswert $x_0 \in \mathbb{R}^d$ und jede Funktion $u_h \in \mathcal{U}_h$ eine eindeutige Lösung $\Phi_h(t, x_0, u_h)$ von (1.2) in positiver Zeitrichtung existiert, also eine Funktion $\Phi : h\mathbb{N}_0 \times \mathbb{R}^d \times \mathcal{U}_h$ mit

$$\Phi_h(0, x_0, u_h) = x_0 \quad \text{und} \quad \Phi_h(t+h, x_0, u_h) = f_h(\Phi_h(t, x_0, u_h), u_h(t)).$$

Im kontinuierlichen Fall ist das etwas komplizierter: Aus der Theorie der gewöhnlichen Differentialgleichungen wissen wir, dass z.B. die Wahl $\mathcal{U} = C(\mathbb{R}, U)$ (also die Menge aller stetigen Funktionen mit Werten in U), zusammen mit der Lipschitz–Stetigkeit von f in x einen Existenz– und Eindeutigkeitssatz erlaubt. Stetige Kontrollfunktionen sind allerdings für viele Anwendungen zu einschränkend, z.B. in der optimalen Steuerung, wo man bereits für sehr einfache Probleme nachweisen kann, dass optimale Steuerstrategien unstetig in t sind. Zudem ist es sowohl für die theoretische als auch für die numerische Behandlung von Kontrollsystemen sehr nützlich, wenn zu je zwei Kontrollfunktionen $u_1, u_2 \in \mathcal{U}$ auch die durch die *Konkatenation zur Zeit* $\tau \in \mathbb{R}$

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

gegebene Funktion u wieder in $\mathcal{U}$ liegt, was für den Raum der stetigen Funktionen ebenfalls nicht zutrifft.

Wir werden deshalb eine größere Klasse von Kontrollfunktionen zulassen. Wir erinnern an die folgende Definition.

Definition 1.7 Sei $I = [a, b] \subset \mathbb{R}$ ein abgeschlossenes Intervall.

(i) Eine Funktion $g : I \to \mathbb{R}^m$ heißt *stückweise konstant*, falls eine Zerlegung von I in endlich viele Teilintervalle I_j, $j = 1, \ldots, n$ existiert, so dass g auf I_j konstant ist für alle $j = 1, \ldots, n$.

(ii) Eine Funktion $g : I \to \mathbb{R}^n$ heißt *(Lebesgue–) messbar*, falls eine Folge von stückweise konstanten Funktionen $g_i : I \to \mathbb{R}^m$, $i \in \mathbb{N}$, existiert mit $\lim_{i\to\infty} g_i(x) = g(x)$ für fast alle[3] $x \in I$.

(iii) Eine Funktion $g : \mathbb{R} \to \mathbb{R}^m$ heißt *(Lebesgue–) messbar*, falls für jedes abgeschlossene Teilintervall $I = [a, b] \subset \mathbb{R}$ die Einschränkung $g|_I$ messbar im Sinne von (ii) ist. □

Der folgende Satz zeigt, dass die Wahl messbarer Kontrollfunktionen einen sinnvollen Lösungsbegriff für (1.1) liefert.

Satz 1.8 (Satz von Carathéodory) Betrachte ein Kontrollsystem mit folgenden Eigenschaften:

i) Der Raum der Kontrollfunktionen ist gegeben durch

$$\mathcal{U} = L_\infty(\mathbb{R}, U) := \{u : \mathbb{R} \to U \mid u \text{ ist messbar und essentiell beschränkt}^4\}.$$

ii) Das Vektorfeld $f : \mathbb{R}^d \times U \to \mathbb{R}^d$ ist stetig.

iii) Für jedes $R > 0$ existiert eine Konstante $L_R > 0$, so dass die Abschätzung

$$\|f(x_1, u) - f(x_2, u)\| \le L_R \|x_1 - x_2\|$$

für alle $x_1, x_2 \in \mathbb{R}^d$ und alle $u \in U$ mit $\|x_1\|, \|x_2\|, \|u\| \le R$ erfüllt ist.

Dann gibt es für jeden Punkt $x_0 \in \mathbb{R}^d$ und jede Kontrollfunktion $u \in \mathcal{U}$ ein (maximales) offenes Intervall I mit $0 \in I$ und genau eine absolut stetige Funktion $x(t)$, die die Integralgleichung

$$x(t) = x_0 + \int_0^t f(x(\tau), u(\tau))\, d\tau$$

für alle $t \in I$ erfüllt.

Definition 1.9 Wie bezeichnen die eindeutige Funktion $x(t)$ aus Satz 1.8 mit $\Phi(t, x_0, u)$ und nennen sie die *Lösung* von (1.1) zum *Anfangswert* $x_0 \in \mathbb{R}^d$ und zur *Kontrollfunktion* $u \in \mathcal{U}$. □

Die folgende Beobachtung rechtfertigt diese Definition: Da $\Phi(t, x_0, u)$ absolut stetig ist, ist diese Funktion für fast alle $t \in I$ nach t differenzierbar. Insbesondere folgt also aus dem Satz 1.8, dass $\Phi(t, x_0, u)$ die Differentialgleichung (1.1) für fast alle $t \in I$ erfüllt, d.h. es gilt

$$\dot{\Phi}(t, x_0, u) = f(\Phi(t, x_0, u), u(t))$$

für fast alle $t \in I$.

[3] d.h. für alle x aus einer Menge $J \subseteq I$ mit der Eigenschaft, dass $I \setminus J$ eine Lebesgue–Nullmenge ist

[4] d.h. beschränkt außerhalb einer Lebesgue–Nullmenge

Bemerkung 1.10 Im Weiteren nehmen wir stets an, dass die Voraussetzungen (i)–(iii) von Satz 1.8 erfüllt sind, werden dies aber nur in wichtigen Sätzen explizit formulieren. □

Der Beweis von Satz 1.8 (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 *Mathematical Control Theory* von E.D. Sontag [20, Anhang C].

1.5 Ein einfaches numerisches Verfahren

Wir wollen uns nun überlegen, wie man ein kontinuierliches Kontrollsystem (1.1) durch ein diskretes System (1.2) approximieren kann. Eine übliche Klasse numerischer Verfahren zur Lösung von gewöhnlichen Differentialgleichungen sind die sogenannten *Einschrittverfahren*. Hierbei wird, zu einer Schrittweite $h > 0$, die kontinuierliche Lösungsfunktion $\Phi(t, x_0, u)$ durch eine diskrete *Gitterfunktion* approximiert, im einfachsten Fall auf einem äquidistanten Gitter $h\mathbb{N}_0$ definiert. Diese Gitterfunktion wird durch ein diskretes dynamisches System der Form (1.2) erzeugt, deren Iterationsvorschrift f_h dabei eine mittels des Computers auswertbaren Abbildung ist.

Zur numerischen Simulation von Trajektorien von (1.1) wollen wir nun solch ein Verfahren definieren. Die naheliegende Idee dazu ist sicherlich, für festes $u \in \mathcal{U}$ ein Vektorfeld $g(t, x) = f(x, u(t))$ zu definieren und ein Einschrittverfahren zur Lösung zeitvarianter gewöhnlicher Differentialgleichungen (z.B. ein Runge–Kutta oder ein Taylor Verfahren) auf die Gleichung

$$\dot{x}(t) = g(t, x(t))$$

anzuwenden. Die Tatsache, dass wir messbare Kontrollfunktionen verwenden, führt aber zu Schwierigkeiten, da alle diese Verfahren nämlich zumindest Lipschitz–Stetigkeit von $g(t, x)$ in t verlangen, eine Eigenschaft, die für messbares u im Allgemeinen nicht gegeben ist.

Tatsächlich muss man hier einen Trick anwenden, um Einschrittverfahren definieren zu können, die später einen wichtigen „Baustein" für unseren Algorithmus zur Lösung optimaler Steuerungsprobleme bilden werden. Wir werden hier nur den Fall des Euler–Verfahrens betrachten.

Wir werden einen Konvergenzsatz für allgemeine Systeme der Form (1.1) formulieren, uns im Beweis aber auf Systeme mit der folgenden Konvexitätseigenschaft beschränken.

Definition 1.11 Wir nennen ein Kontrollsystem (1.1) *konvex*, falls die Menge

$$f(x, U_R) := \{f(x, u),\, u \in U_R\} \subset \mathbb{R}^d \quad \text{mit} \quad U_R := \{u \in U \mid \|u\| \leq R\}$$

für jedes $x \in \mathbb{R}^d$ und jedes hinreichend große $R > 0$ konvex ist. □

Wir definieren nun ein numerisches Einschrittverfahren.

Definition 1.12 (Euler–Verfahren für Kontrollsysteme) Für einen Zeitschritt $h > 0$ und einen Kontrollwert $u \in U$ definieren wir das *Euler Verfahren* als das durch die Abbildung

$$f_h(x,u) := x + hf(x,u)$$

definierte zeitdiskrete Kontrollsystem (1.2). Die Lösungen bezeichnen wir mit $\widetilde{\Phi}_h(t, x_0, u_h)$. □

Der folgende Satz fasst die Eigenschaften dieses Verfahrens zusammen. Hierin bezeichnen wir mit $\overline{B}_R(0)$ den abgeschlossenen Ball mit Radius R um 0 im $\mathbb{R}^d$ bzw. $\mathbb{R}^m$.

Satz 1.13 Betrachte ein Kontrollsystem, für das die Voraussetzungen (i)–(iii) von Satz 1.8 gelten. Dann gilt für das Verfahren aus Definition 1.12 und jede Konstante $R > 0$ die folgende Aussage:

(i) Es existiert eine (von R abhängige) Konstante $K > 0$, so dass für jede Kontrollfunktion $u \in \mathcal{U}$ mit $\|u\|_\infty \le R$ und jeden Anfangswert $x_0 \in \overline{B}_R(0)$ eine diskrete Kontrollfunktion $u_h \in \mathcal{U}_h$ existiert, mit der die Abschätzung

$$\|\widetilde{\Phi}_h(t, x_0, u_h) - \Phi(t, x_0, u)\| \le K\sqrt{h}e^{Lt}$$

gilt für alle $t \in h\mathbb{N}_0$ für die die Lösungen in $\overline{B}_R(0)$ liegen.

Ist das Kontrollsystem konvex, so gilt die schärfere Abschätzung

$$\|\widetilde{\Phi}_h(t, x_0, u_h) - \Phi(t, x_0, u)\| \le Kh(e^{Lt} - 1).$$

(ii) Umgekehrt existiert eine von R abhängige Konstante $K > 0$ so dass für jedes $x_0 \in \overline{B}_R(0)$, jede diskrete Kontrollfunktion $u_h \in \mathcal{U}_h$ mit $\|u_h\|_\infty \le R$ und die durch

$$u(\tau) := u_h(t), \;\; \tau \in [t, t+h), \quad t \in h\mathbb{N}_0$$

definierte stückweise konstante (also messbare) Kontrollfunktion die Abschätzung

$$\|\widetilde{\Phi}(t, x_0, u_h) - \Phi(t, x_0, u)\| \le Kh(e^{Lt} - 1)$$

gilt für alle $t \in h\mathbb{N}_0$ für die die Lösungen in $\overline{B}_R(0)$ liegen.

Beweis: Mit $L = L_R$ bezeichnen wir die Lipschitz–Konstante aus Satz 1.8(iii). Zudem sei $M > 0$ eine Konstante, für die die Abschätzung

$$\|f(x,u)\| \le M \text{ für alle } \|x\| \le R,\ \|u\| \le R$$

gilt. Dieses M existiert wegen der Stetigkeit von f.

Wir beginnen mit (i). Wie bereits erwähnt, betrachten wir hier nur den konvexen Fall, da der allgemeine Fall deutlich komplizierter ist (ein Beweis findet sich z.B. in der Arbeit [8] von R. L. V. González und M. M. Tidball). Wir nehmen dabei o.B.d.A. an, dass R so groß ist, dass die Konvexität gemäß Definition 1.5 gilt.

Wir betrachten zunächst zu jeder Kontrollfunktion $u \in \mathcal{U}_R := \{u : \mathbb{R} \to \mathbb{R}^m \mid \|u\| \leq R\}$ und jedem Punkt $x \in \mathbb{R}^d$ den Wert

$$\bar{f}(x,u) = \frac{1}{h}\int_0^h f(x,u(t))\,dt.$$

Aus der Konvexität von $f(x,U_R)$ folgt, dass $\bar{f}(x,u)$ in $f(x,U_R)$ liegt (tatsächlich ist die Konvexität von $f(x,U_R)$ hierfür auch notwendig). Daher gibt es Kontrollwerte

$$\bar{u}(x,u) \in U_R \text{ mit } f(x,\bar{u}(x,u)) = \bar{f}(x,u). \tag{1.3}$$

Für diese Werte zeigen wir zunächst die Abschätzung

$$\|\Phi(h,x,u) - f_h(x,\bar{u}(x,u))\| \leq \frac{M}{2}Lh^2 \tag{1.4}$$

für f_h aus Definition 1.12, unter der Annahme, dass $\Phi(\tau,x,u) \in \overline{B}_R(0)$ gilt für $\tau \in [0,h]$. Es gilt nämlich

$$\begin{aligned}
\Phi(h,x,u) &= x + \int_0^h f(\Phi(t,x,u),u(t))\,dt \\
&= x + \int_0^h f(x,u(t))\,dt + R(h) \;=\; x + h\,\bar{f}(x,u) + R(h) \\
&= x + hf(x,\bar{u}(x,u)) + R(h) \;=\; f_h(x,\bar{u}(x,u)) + R(h)
\end{aligned}$$

mit dem Restterm

$$R(h) = \int_0^h f(\Phi(t,x,u),u(t)) - f(x,u(t))\,dt.$$

Also folgt

$$\|\Phi(h,x,u) - f_h(x,\bar{u}(x,u))\| \leq \|R(h)\|.$$

Der Restterm $R(h)$ lässt sich abschätzen durch

$$\begin{aligned}
\|R(h)\| &\leq \int_0^h L\|\Phi(t,x,u) - x\|\,dt \\
&\leq \int_0^h L\int_0^t \|f(\Phi(\tau,x,u),u(\tau))\|\,d\tau\,dt \\
&\leq \int_0^h L\int_0^t M\,d\tau\,dt \;=\; \frac{M}{2}Lh^2
\end{aligned}$$

womit (1.4) gezeigt ist.

Aus der Reihendarstellung $e^{Lh} = 1 + Lh + L^2h^2/2 + \ldots$ folgt $e^{Lh} \geq 1 + Lh$, damit $Lh^2 \leq h(e^{Lh} - 1)$ und folglich aus (1.4)

$$\|\Phi(h,x,u) - f_h(x,\bar{u}(x,u))\| \leq \frac{M}{2}h(e^{Lh} - 1). \tag{1.5}$$

Betrachte nun eine Kontrollfunktion $u \in \mathcal{U}$ und einen Anfangswert $x_0 \in \mathbb{R}^d$. Für jedes $t \in h\mathbb{N}_0$ betrachte die Funktion $w_t \in \mathcal{U}$ gegeben durch $w_t(\tau) = u(t+\tau)$. Mit dem Existenz– und Eindeutigkeitssatz 1.8 ist leicht zu überprüfen, dass für diese w_i die Identität

$$\Phi(t+h, x_0, u) = \Phi(h, \Phi(t, x_0, u), w_t) \tag{1.6}$$

für alle $i \in \mathbb{N}_0$ gilt. Wir definieren nun die diskrete Kontrollfunktion $u_h : h\mathbb{N}_0 \to U_R$ für $t \in h\mathbb{N}_0$ als

$$u_h(t) = \bar{u}(\Phi(t, x_0, u), w_t)$$

mit $\bar{u}$ aus (1.3).

Aus (1.5) und (1.6) folgt damit für alle $t \in h\mathbb{N}_0$ die Abschätzung

$$\|\Phi(t+h, x_0, u) - f_h(\Phi(t, x_0, u), u(t))\| \leq \frac{M}{2} h(e^{Lh} - 1). \tag{1.7}$$

Wir zeigen die behauptete Ungleichung aus (i) nun durch Induktion über $t \in h\mathbb{N}_0$. Für $t = 0$ ist die Behauptung offensichtlich. Nehmen wir also an, die gewünschte Abschätzung sei für ein $t \in h\mathbb{N}_0$ erfüllt, d.h. es gelte

$$\|\widetilde{\Phi}_h(t, x_0, u_h) - \Phi(t, x_0, u)\| \leq h(e^{Lt} - 1) \tag{1.8}$$

Aus der Annahme (iv) von Satz 1.8 und der obigen Reihendarstellung von e^{Lh} folgt die Lipschitz–Abschätzung

$$\|f_h(x_1, u) - f_h(x_2, u)\| \leq (1 + Lh)\|x_1 - x_2\| \leq e^{Lh}\|x_1 - x_2\| \tag{1.9}$$

für alle $x_1, x_2 \in \mathbb{R}^d$ und alle $u \in U$. Damit erhalten wir

$$\begin{aligned}
&\|\widetilde{\Phi}_h(t+h, x_0, u_h) - \Phi(t+h, x_0, u)\| \\
&= \|f_h(\widetilde{\Phi}_h(t, x_0, u_h), u_h(t)) - \Phi(t+h, x_0, u)\| \\
&\leq \|f_h(\widetilde{\Phi}_h(t, x_0, u_h), u_h(t)) - f_h(\Phi(t, x_0, u), u_h(t))\| \\
&\quad + \|f_h(\Phi(t, x_0, u), u_h(t)) - \Phi(t+h, x_0, u)\| \\
&\overset{(1.9)}{\leq} e^{Lh}\|\widetilde{\Phi}_h(t, x_0, u_h) - \Phi(t, x_0, u)\| + \|f_h(\Phi(t, x_0, u), u_h(t)) - \Phi(t+h, x_0, u)\| \\
&\overset{(1.7)}{\leq} e^{Lh}\|\widetilde{\Phi}_h(t, x_0, u_h) - \Phi(t, x_0, u)\| + \frac{M}{2} h(e^{Lh} - 1) \\
&\overset{(1.8)}{\leq} e^{Lh}\frac{M}{2} h(e^{Lhi} - 1) + \frac{M}{2} h(e^{Lh} - 1) \\
&= \frac{M}{2} h(e^{Lh(i+1)} - e^{Lh} + e^{Lh} - 1) \;=\; \frac{M}{2} h(e^{Lh(i+1)} - 1)
\end{aligned}$$

und damit die gewünschte Abschätzung aus (i).

Zum Beweis von (ii) betrachte eine diskrete Kontrollfunktion $u_h : h\mathbb{Z} \to U$ und die in (i) konstruierte stückweise konstante Kontrollfunktion $u \in \mathcal{U}$. Wenn wir nun eine neue diskrete Kontrollfunktion $\tilde{u}_h$ wie im Beweis von (i) aus diesem u konstruieren, so gilt offenbar $\tilde{u}_h = u_h$, d.h. wir erhalten gerade wieder die Ausgangsfunktion u_h (beachte, dass für die stückweise konstante Funktion u die Konstruktion aus (i) auch ohne die Konvexitätsbedingung funktioniert). Also folgt (ii) indem wir (i) auf dieses u anwenden. □

Bemerkung 1.14 (i) Beachte, dass die diskrete Kontrollfunktion u_h aus Satz 1.13(i) implizit definiert ist und daher im Allgemeinen keine einfache Formel zu ihrer Berechnung angegeben werden kann. Tatsächlich ist die explizite Kenntnis von u_h zur Lösung von optimalen Steuerungsproblemen aber gar nicht notwendig, wie wir in den nächsten Kapiteln sehen werden.

(ii) Ein Sonderfall ergibt sich, falls das Vektorfeld f in (1.1) die *kontroll–affine* Struktur

$$f(x,u) = f_0(x) + \sum_{i=1}^{m} f_i(x)u_i$$

besitzt. In diesem Fall sieht man leicht, dass sich die im Beweis konstruierte diskrete Funktion zu

$$u_h(t) = \frac{1}{h}\int_t^{t+h} u(\tau)d\tau$$

ergibt. Hier lässt sich u_h also explizit angeben. □

Bemerkung 1.15 Das Euler–Verfahren für Kontrollsysteme besitzt also die Konvergenzordnung $O(\sqrt{h})$ bzw. $O(h)$. Natürlich kann die Iterationsvorschrift $f_h(x,u) = x + hf(x,u)$ in Definition 1.12 für konstante Kontrollwerte $u \in U$ durch ein beliebiges numerisches Einschrittverfahren, z.B. ein Runge–Kutta Verfahren höherer Ordnung, ersetzt werden. Man wird hierbei allerdings in Teil (i) des Satzes i.A. trotzdem nur die Konvergenzordnung $O(\sqrt{h})$ bzw. $O(h)$ erzielen, da der numerische Fehler durch den beim Übergang von der messbaren Funktion u zu der „stückweise konstanten“ Funktion u_h entstehenden Fehler dominiert wird. Um höhere Konvergenzordnung zu erzielen, muss man neben $\bar{f}$ weitere (iterierte) Integrale im numerischen Schema berücksichtigen, was die Konstruktion entsprechender Einschrittverfahren sehr aufwändig macht. Insbesondere muss man in der diskreten Approximation (1.2) zu i.A. komplizierten höherdimensionalen Kontrollparametermengen $\overline{U}$ an Stelle von U übergehen. Details finden sich — unter verschiedenen strukturellen Annahmen an f — in den Arbeiten [6] und [11].

Anders ist dies bei Teil (ii) des Satzes. Hier kann man relativ leicht beweisen, dass ein Verfahren höherer Ordnung auch eine entsprechend bessere Konvergenzabschätzung liefert. □

Bemerkung 1.16 Wir werden später sehen, dass es praktisch sein kann, die Werte $u_h(t)$ der diskreten Kontrollfunktion u_h aus einer endlichen Menge $\widetilde{U} \subset U$ zu wählen. Wenn $\widetilde{U}$ hinreichend „dicht“ in U liegt, lässt sich ein ähnliches Resultat wie in Satz 1.13(i) für Funktionen u_h mit Werten $u_h(t) \in \widetilde{U}$ beweisen. □

Kapitel 2

Optimale Steuerung

In diesem Kapitel wollen wir ein Modellproblem der optimalen Steuerung einführen und einige wesentliche Eigenschaften des Problems betrachten.

2.1 Diskontierte Optimale Steuerung

Zunächst müssen wir uns überlegen, was für eine Größe wir eigentlich „optimieren“ wollen. Hier betrachten wir zunächst eine *Kosten– oder Ertragsfunktion* g, die jedem Punkt $(x, u) \in \mathbb{R}^d \times U$ im kombinierten Zustands–Kontrollwerteraum einen Wert zuweist. Integriert bzw. summiert man diese Funktion entlang einer Trajektorie $\Phi(t, x_0, u)$ bzw. $\Phi_h(t, x_0, u)$ und der dazugehörigen Kontrollfunktion u, so erhält man einen Wert, der von dem Anfangswert x_0 und von der Kontrollfunktion u abhängt. Ziel der optimalen Steuerung ist es nun, die Kontrollfunktion $u \in \mathcal{U}$ (in Abhängigkeit von x_0) so zu wählen, dass dieser Wert maximiert oder minimiert wird. Formal können wir das so definieren.

Definition 2.1 Betrachte ein Kontrollsystem (1.1) bzw. (1.2). Für eine Funktion $g : \mathbb{R}^d \times U \to \mathbb{R}$ und einen Parameter $\delta > 0$ definieren wir das *diskontierte Funktional auf unendlichem Zeithorizont* in kontinuierlicher Zeit als

$$J(x, u) := \int_0^\infty e^{-\delta t} g(\Phi(t, x, u), u(t))\, dt. \tag{2.1}$$

und in diskreter Zeit als

$$J_h(x, u_h) := h \sum_{j=0}^{\infty} (1 - \delta h)^j g(\Phi_h(jh, x, u_h), u_h(jh)) \tag{2.2}$$

Das optimale Steuerungsproblem lautet nun: Bestimme die *optimale Wertefunktion*

$$v(x) := \sup_{u \in \mathcal{U}} J(x, u) \;\text{ bzw. }\; v_h(x) := \sup_{u_h \in \mathcal{U}_h} J_h(x, u_h).$$

Hierbei machen wir die folgende Annahmen:

(A1) Der Kontrollwertebereich U sei kompakt.

(A2) In kontinuierlicher Zeit erfülle das Kontrollsystem (1.1) die Voraussetzungen (i)–(iii) von Satz 1.8, wobei die Lipschitz–Konstante $L_R = L$ unabhängig von R sei.

In diskreter Zeit existiere eine Konstante $L > 0$, so dass die Lipschitz–Abschätzung

$$\|f_h(x_1, u) - f_h(x_2, u)\| \leq (1 + Lh)\|x_1 - x_2\|$$

gilt für alle $x_1, x_2 \in \mathbb{R}^d$ und alle $u \in U$.

(A3) Die Funktion g sei stetig und erfülle

$$|g(x, u)| \leq M_g \quad \text{und} \quad |g(x_1, u) - g(x_2, u)| \leq L_g\|x_1 - x_2\|$$

für alle x, x_1, $x_2 \in \mathbb{R}^d$, alle $u \in \mathcal{U}$ und geeignete Konstanten M_g, $L_g > 0$.

□

Die Annahme der globalen Lipschitz–Stetigkeit (d.h., dass die Konstante L_R unabhängig von R ist), dient der Vereinfachung einiger nachfolgender Aussagen und Beweise, sie ist aber nicht wesentlich für die Funktion des Algorithmus.

Bemerkung 2.2 (i) Statt zu maximieren kann—völlig analog—das entsprechende Minimierungsproblem $v(x) := \inf_{u\in\mathcal{U}} J(x, u)$ bzw. $v_h(x) := \inf_{u_h\in\mathcal{U}_h} J_h(x, u_h)$ betrachtet werden.

(ii) Wir setzen hier nicht voraus, dass optimale Kontrollfunktionen $u \in \mathcal{U}$ bzw. $u_h \in \mathcal{U}_h$ existieren[1], deshalb verwenden wir „sup“ statt „max“.

(iii) Über die Kenntnis von v und v_h hinaus ist es natürlich interessant, auch (zumindest näherungsweise) optimale Kontrollfunktionen u und u_h zu berechnen. Tatsächlich kann man diese aus v bzw. v_h berechnen, wir werden später sehen, wie.

(iv) Es gibt eine ganze Reihe anderer Funktionale, die man im Rahmen der optimalen Steuerung maximieren oder minimieren kann. Viele davon können numerisch mit ähnlichen Methoden gelöst werden, wie wir sie in dieser Vorlesung kennen lernen werden, einige werden wir später noch kennen lernen. □

Unser hier betrachtetes Modellproblem des diskontierten Funktionals J mit dem exponentiellen *Diskontfaktor* $e^{-\delta t}$ und positiver *Diskontrate* $\delta > 0$ stammt ursprünglich aus der Ökonomie und modelliert die Tatsache, dass der Ertrag in naher Zukunft wichtiger ist als derjenige in ferner Zukunft (beachte, dass $e^{-\delta t}$ für $t \to \infty$ monoton gegen 0 strebt, und damit zeitlich weit in der Zukunft liegende Werte von $g(\Phi(t, x, u), u(t))$ schwächer gewichtet werden). Wichtiger als diese ökonomische Interpretation sind für uns die mathematischen Auswirkungen des Diskontfaktors. Die offensichtlichsten sind in dem folgenden Lemma zusammengefasst.

[1] unter gewissen Voraussetzungen lässt sich die Existenz optimaler Kontrollfunktionen beweisen; dies näher auszuführen würde den Rahmen dieser Vorlesung aber sprengen

Lemma 2.3 Es sei $\delta h < 1$ (was im kontinuierlichen Fall „$h = 0$“ keine Einschränkung bedeutet). Dann gilt

(i) Das diskontierte Funktional ist endlich. Genauer gilt dann

$$|J(x,u)| \le \frac{M_g}{\delta} \text{ und } |J_h(x,u_h)| \le \frac{M_g}{\delta},$$

insbesondere also auch $|v(x)|, |v_h(x)| \le M_g/\delta$.

(ii) „Für ε–optimale Steuerung reicht die Betrachtung endlicher Zeitintervalle“, oder formal: Seien $x \in \mathbb{R}^d$ und $\varepsilon > 0$ gegeben, dann gibt es ein $T_\varepsilon > 0$ und eine Kontrollfunktion $u_\varepsilon \in \mathcal{U}$, so dass für jede Kontrollfunktion $u \in \mathcal{U}$ mit $u(t) = u_\varepsilon(t)$ für $t \in [0, T_\varepsilon]$ gilt

$$J(x,u) \ge v(x) - \varepsilon.$$

Dies gilt analog in diskreter Zeit.

Beweis: Wir beweisen die Aussagen in kontinuierlicher Zeit, die analogen Aussagen in diskreter Zeit sind Übungsaufgaben.

(i) Es gilt

$$\begin{aligned}
|J(x,u)| &= \left| \int_0^\infty e^{-\delta t} g(\Phi(t,x,u),u(t))\,dt \right| \\
&\le \int_0^\infty e^{-\delta t} \,|g(\Phi(t,x,u),u(t))|\,dt \\
&\le \int_0^\infty e^{-\delta t}\, M_g\,dt \\
&\le M_g \int_0^\infty e^{-\delta t}\,dt \\
&= M_g \left[-\frac{1}{\delta}\, e^{-\delta t} \right]_0^\infty \;=\; \frac{M_g}{\delta}
\end{aligned}$$

(ii) Sei $T_\varepsilon = -\log[\varepsilon\delta/(4M_g)]/\delta$ und $u_\varepsilon \in \mathcal{U}$ so gewählt, dass $J(x,u_\varepsilon) \ge v(x) - \varepsilon/2$. Dann gilt für u aus der Behauptung

$$\begin{aligned}
J(x,u) &= \int_0^\infty e^{-\delta t} g(\Phi(t,x,u),u(t))\,dt \\
&= \int_0^{T_\varepsilon} e^{-\delta t} g(\Phi(t,x,u),u(t))\,dt + \int_{T_\varepsilon}^\infty e^{-\delta t} g(\Phi(t,x,u),u(t))\,dt \\
&= \int_0^{T_\varepsilon} e^{-\delta t} g(\Phi(t,x,u_\varepsilon),u_\varepsilon(t))\,dt + \int_{T_\varepsilon}^\infty e^{-\delta t} g(\Phi(t,x,u),u(t))\,dt \\
&= \int_0^\infty e^{-\delta t} g(\Phi(t,x,u_\varepsilon),u_\varepsilon(t))\,dt \\
&\qquad + \int_{T_\varepsilon}^\infty e^{-\delta t} g(\Phi(t,x,u),u(t))\,dt - \int_{T_\varepsilon}^\infty e^{-\delta t} g(\Phi(t,x,u_\varepsilon),u_\varepsilon(t))\,dt \\
&\ge v(x) - \frac{\varepsilon}{2} - 2M_g \left[-\frac{1}{\delta}\, e^{-\delta t} \right]_{T_\varepsilon}^\infty \\
&= v(x) - \frac{\varepsilon}{2} - 2M_g \left(\frac{\varepsilon\delta}{4M_g} \right) \;=\; v(x) - \varepsilon
\end{aligned}$$

□

2.2 Beispiele

2.2.1 Der gesteuerte Wagen

Betrachte das einfache Modell eines Wagens aus Beispiel 1.3 gegeben durch

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= u_1(t) - (r + u_2(t))x_2(t) \end{aligned}$$

Wir setzen nun $u = (u_1, u_2)^T \in [-1, 1] \times [0, 1]$. Die Kontrolle u_1 modelliert hier also die externe Beschleunigung, die z.B. durch einen Motor erzeugt werden kann, $r > 0$ ist ein Reibungsfaktor und u_2 modelliert eine Bremse.

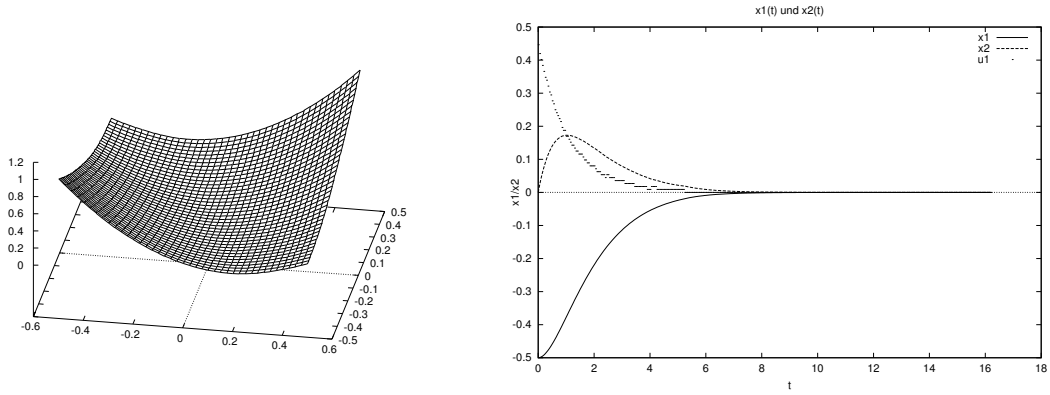

Abbildung 2.1: Wertefunktion und optimale Trajektorie für den gesteuerten Wagen

Wählen wir als Zielfunktion $g(x, u) = \|x\|^2 + u_1^2$, so ist es zur Minimierung des Funktionals $J(x, u)$ offensichtlich eine gute Strategie, den Wagen in die Position $x_1 = 0$ mit Geschwindigkeit $x_2 = 0$ zu bringen. Andererseits erhöht jeder Einsatz des über u_1 gesteuerten Motors das Funktional, so dass das Manöver mit möglichst wenig Motorkraft durchgeführt werden sollte. Abbildung 2.1 zeigt einen Ausschnitt der numerisch berechneten optimale Wertefunktion dieses Problems für Diskontrate $\delta = 0.1$ und Reibungsfaktor $r = 1$, sowie die zugehörige optimale Trajektorie für den Anfangswert $x = (-1, 0)^T$, dargestellt als x_1 und x_2 in Abhängigkeit von t. Der Wagen wird (zunächst stark, dann immer schwächer) gerade so stark beschleunigt, dass er ohne Verwendung der Bremse durch die Reibung genau im Punkt $x_1 = 0$ stehen bleibt.

Stellt man mehrere optimale Trajektorien gemeinsam dar, so stellt man fest, dass alle diese Kurven gegen das Gleichgewicht $(0, 0)$ streben, tatsächlich ist dieser Punkt „asymptotisch stabil“ für das optimal gesteuerte System. Wir werden im zweiten Teil der Vorlesung optimale Steuerungsprobleme betrachten, mit denen man dieses Verhalten systematisch erzielen kann.

2.2.2 Investitionsmodell

Wir betrachten das einfache lineare Investitionsmodell

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) - \sigma x_1(t) \\ \dot{x}_2(t) &= u \end{aligned}$$

aus Beispiel 1.4. Ziel ist es hier, den diskontierten cash flow zu maximieren, der mittels der Ertragsfunktion

$$g(x_1, x_2, u) = k_1\sqrt{x_1} - \frac{x_1}{1 + k_2 x_1^4} - c_1 x_2 - \frac{c_2}{2}x_2^2 - \frac{\alpha}{2}u^2$$

gemessen wird; diese Funktion wurde in dem Artikel [15] eingeführt. Mit den Parametern $\sigma = 0.25$, $k_1 = 2$, $k_2 = 0.0117$, $c_1 = 0.75$, $c_2 = 2.5$ und $\alpha = 12$ sowie der Diskontrate $\delta = 0.04$ ergibt sich das in Abbildung 2.2 dargestellte Ergebnis.

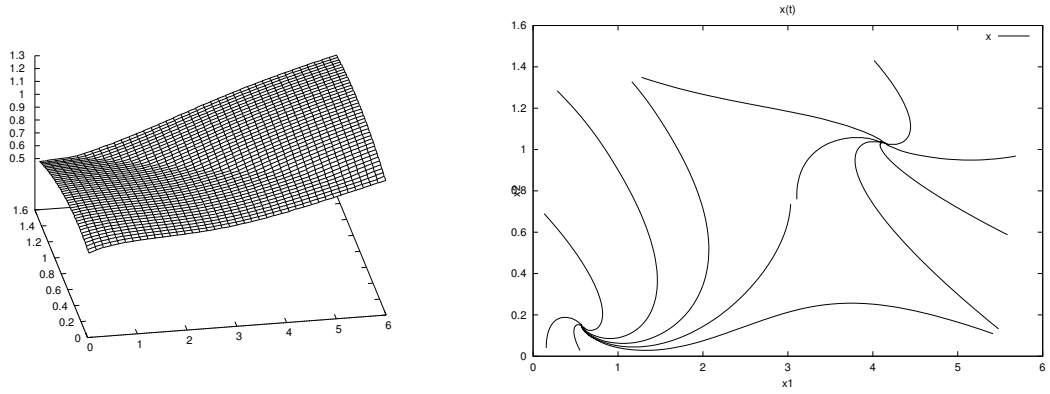

Abbildung 2.2: Wertefunktion und optimale Trajektorien für das Investitionsmodell

Interessant ist hier, dass für das optimal gesteuerte System zwei asymptotisch stabile „optimale Gleichgewichte“ existieren, deren Einzugsbereiche durch eine eindimensionale Kurve getrennt sind. Betrachtet man die optimale Wertefunktion genauer, so sieht man, dass die optimale Wertefunktion entlang genau dieser Kurve einen Knick besitzt, also nicht differenzierbar ist. Dieses Verhalten ist typisch und kann in der Numerik ausgenutzt werden, um solche „Trennkurven“ von Einzugsbereichen effizient zu berechnen.

2.2.3 Das Räuber–Beute Modell

Zum Abschluss unserer Beispiele betrachten wir das *Räuber–Beute–Modell* aus Beispiel 1.5, gegeben durch

$$\begin{aligned} \dot{x}_1(t) &= (a_0 - a_2 x_2(t) - a_1 x_1(t) - u(t))x_1(t) \\ \dot{x}_2(t) &= (b_1 x_1(t) - b_0 - b_2 x_2(t) - u(t))x_2(t). \end{aligned}$$

Die folgende Liste stellt die verwendeten Parameterwerte sowie eine kurze Erklärung ihrer Bedeutung dar:

a_0	: Geburtenrate der Beute	(1.04)
a_1	: Stressfaktor der Beute	(0.01)
a_2	: Sterberate der Beute, abhängig von vorhandenen Räubern	(0.07)
b_0	: Sterberate der Räuber	(1.01)
b_1	: Geburtenrate der Räuber, abhängig von vorhandener Beute	(0.2)
b_2	: Stressfaktor der Räuber	(0.01)
$u(t)$	: Fangrate der Fischer	(0–3)

Ziel des optimalen Steuerungproblems ist es nun, den Ertrag der Fischer zu maximieren. Der Ertrag des Fangs wird hierbei durch die Ertragsfunktion

$$g(x,u) = \frac{1}{1+x_1u}x_1u + \frac{1}{1+x_2u}x_2u - \frac{u}{2}$$

festgelegt. Sie berücksichtigt die Kosten des Fangs durch den Term $u/2$, sowie die Tatsache, dass bei erhöhter Stückzahl x_1u bzw. x_2u der erzielte Marktpreis pro Einheit sinkt.

Abbildung 2.3 zeigt die optimale Wertefunktion dieses Problems für Diskontrate $\delta = 5$. Abbildung 2.4 zeigt eine optimale Trajektorie sowohl in zeitabhängiger Darstellung als auch in der (x_1, x_2)–Ebene. Die Fangrate der Fischer schwankt hierbei zwischen 0.5 und 0.75, der Ertrag liegt bei etwa 0.3. Im Vergleich dazu zeigt Abbildung 2.5 die Trajektorie zu konstanter Fangrate $u \equiv 0.75$; hier liegt der Ertrag bei 0.29.

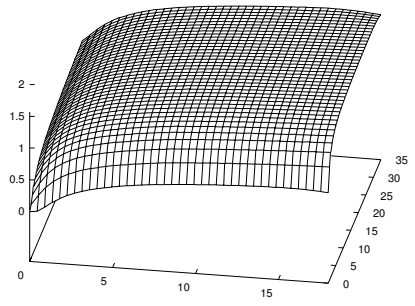

Abbildung 2.3: Wertefunktion für das Räuber–Beute Modell

Bemerkenswert an diesem Beispiel ist die Tatsache, dass die optimale Fangstrategie periodisch ist, d.h., es wird abhängig vom vorhandenen Fischbestand entschieden, ob mehr oder weniger gefischt wird. Zu jeder konstanten Fangrate $u \in [0,3]$ hingegen läuft das System in ein Gleichgewicht, wie z.B. die in Abbildung 2.5 dargestellte Lösung.

2.3 Stetigkeit der optimalen Wertefunktion

Um die optimale Wertefunktion $v(x)$ numerisch approximieren zu können, müssen wir zunächst betrachten, welche Regularitätseigenschaften sie besitzt. Wir können im Allgemeinen nicht davon ausgehen, dass $v(x)$ differenzierbar ist; tatsächlich können wir noch

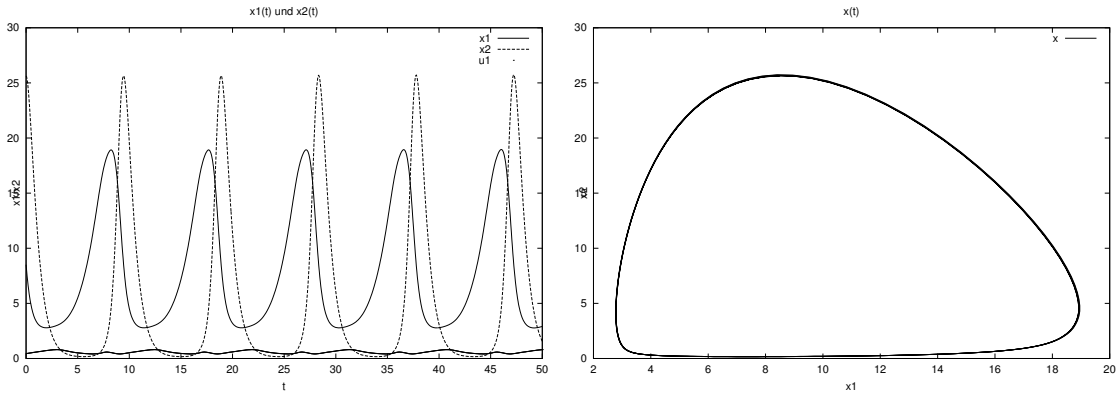

Abbildung 2.4: Optimale Trajektorie für das Räuber–Beute Modell

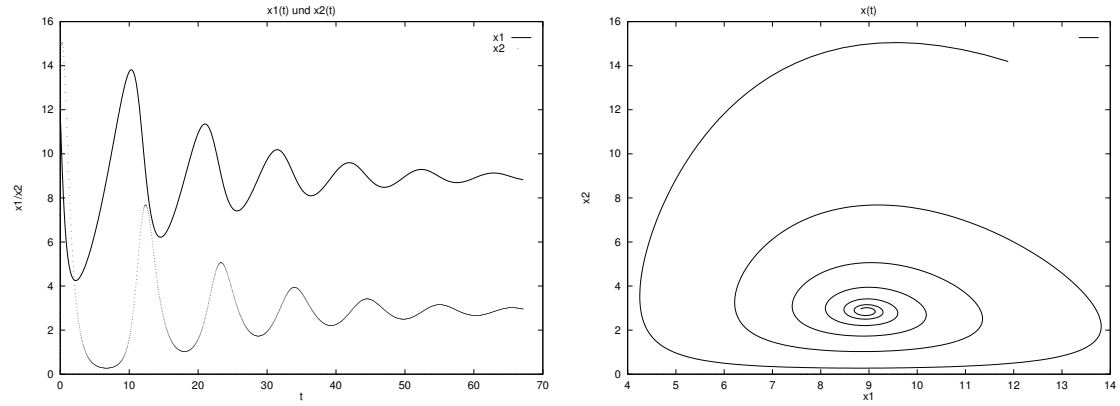

Abbildung 2.5: Trajektorie zu $u \equiv 0.75$ für das Räuber–Beute Modell

nicht einmal Lipschitz Stetigkeit erwarten. Wir können aber eine Abschwächung der Lipschitz Stetigkeit, die sogenannte *Hölder Stetigkeit* beweisen. Hierzu benötigen wir zunächst zwei vorbereitende Lemmata.

Lemma 2.4 Sei B eine beliebige Menge und betrachte zwei Abbildungen $a_1, a_2 : B \to \mathbb{R}$. Dann gelten die Abschätzungen

$$\left| \sup_{b_1 \in B} a_1(b_1) - \sup_{b_2 \in B} a_2(b_2) \right| \leq \sup_{b \in B} |a_1(b) - a_2(b)|$$

und

$$\left| \inf_{b_1 \in B} a_1(b_1) - \inf_{b_2 \in B} a_2(b_2) \right| \leq \sup_{b \in B} |a_1(b) - a_2(b)|.$$

Beweis: Übungsaufgabe □

Lemma 2.5 Sei $\delta > 0$ und $\phi : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ eine Funktion mit $\phi(t) \leq M$ für ein $M > 0$ und alle $t \geq 0$. Desweiteren nehmen wir an, dass Konstanten $C, D, L > 0$, $C \leq M$, existieren,

so dass die Ungleichung

$$\int_0^T e^{-\delta t}\phi(t)\,dt \le C\frac{e^{(L-\delta)T}-1}{L-\delta} + D$$

im Fall $L \neq \delta$ oder die Ungleichung

$$\int_0^T e^{-\delta t}\phi(t)\,dt \le CT + D$$

im Fall $L = \delta$ für alle $T > 0$ gilt. Dann gilt

$$\int_0^\infty e^{-\delta t}\phi(t)\,dt \le KC^\gamma + D$$

mit $\gamma = 1$, falls $\delta > L$, $\gamma \in (0,1)$ beliebig, falls $\delta = L$ und $\gamma = \delta/L$, falls $\delta < L$ und einer Konstanten $K > 0$.

Beweis: Für jedes $T > 0$ gilt

$$\begin{aligned}\int_0^\infty e^{-\delta t}\phi(t)\,dt &\le \int_0^T e^{-\delta t}\phi(t)\,dt + \int_T^\infty e^{-\delta t}\phi(t)\,dt \\ &\le C\frac{e^{(L-\delta)T}-1}{L-\delta} + D + M\frac{e^{-\delta T}}{\delta}\end{aligned}$$

für $L \neq \delta$, bzw.

$$\int_0^\infty e^{-\delta t}\phi(t)\,dt \le CT + D + M\frac{e^{-\delta T}}{\delta}$$

für $L = \delta$. Wählen wir nun $T = \infty$ für $\delta > L$, $T = \frac{1}{\delta}\log\frac{M}{C}$ für $\delta = L$ und $T = \frac{1}{L}\log\frac{M}{C}$ für $\delta < L$, so ergibt sich

$$\int_0^\infty e^{-\delta t}\phi(t)\,dt \le D + \begin{cases} \frac{C}{\delta - L}, & \text{falls } \delta > L \\ C\left(\frac{1}{\delta} + \frac{1}{\delta}\log\frac{M}{C}\right), & \text{falls } \delta = L \\ C^{\frac{\delta}{L}}\left(\frac{M^{L-\frac{\delta}{L}}}{L-\delta} + \frac{M^{1-\frac{\delta}{L}}}{\delta}\right), & \text{falls } \delta < L \end{cases}$$

und damit die Behauptung, wobei wir im Fall $\delta = L$ ausnutzen, dass für alle $\gamma \in (0,1)$ ein $B > 0$ existiert mit $C\left(\log\frac{M}{C}\right) \le BC^\gamma$ für alle $C \in [0, M]$. □

Mit diesen Teilresultaten können wir nun die Stetigkeitseigenschaften von v beweisen.

Satz 2.6 Betrachte das optimale Steuerungsproblem aus Definition 2.1. Ist $\delta > L$, so ist die optimale Wertefunktion Lipschitz stetig mit Konstante $L_g/(\delta - L)$. Ist $\delta \le L$, so ist die optimale Wertefunktion *Hölder stetig*, d.h., es existieren Konstanten $K, \gamma > 0$ so dass für alle $x, y \in \mathbb{R}^d$ die Abschätzung

$$|v(x) - v(y)| \le K\|x - y\|^\gamma$$

gilt. Hierbei ist $\gamma = \delta/L$, falls $\delta < L$ und $\gamma \in (0,1)$ beliebig, falls $\delta = L$.

Beweis: Wir beweisen das Resultat für das kontinuierliche Problem, die Aussage für das zeitdiskrete Problem folgt analog mit Abschätzung der entsprechenden Summen an Stelle der Integrale.

Aus der Lipschitz Stetigkeit des Vektorfeldes f folgt die Abschätzung

$$\|\Phi(t,x,u) - \Phi(t,y,u)\| \leq \|x-y\| + \int_0^t L\|\Phi(\tau,x,u) - \Phi(\tau,y,u)\|d\tau.$$

Mit Gronwalls Lemma[2] folgt daraus

$$\|\Phi(t,x,u) - \Phi(t,y,u)\| \leq \|x-y\|e^{Lt}.$$

Damit ergibt sich

$$\begin{aligned}
&\int_0^T e^{-\delta T}\|g(\Phi(t,x,u),u(t)) - g(\Phi(t,y,u),u(t))\|\,dt \\
&\quad\leq \int_0^T e^{-\delta T}L_g\|x-y\|e^{Lt}\,dt \\
&\quad\leq L_g\|x-y\|\frac{e^{(L-\delta)T}-1}{L-\delta}
\end{aligned}$$

falls $\delta \neq L$ oder

$$\int_0^T e^{-\delta T}\|g(\Phi(t,x,u),u(t)) - g(\Phi(t,y,u),u(t))\|\,dt \leq L_g\|x-y\|T$$

falls $\delta = L$. Mit Lemma 2.5 (für $C = L_g\|x-y\|$ und $D = 0$) erhalten wir also für jedes $u \in \mathcal{U}$ und alle x, $y \in \mathbb{R}^d$ die Abschätzung

$$|J(x,u) - J(y,u)| \leq K\|x-y\|^\gamma \tag{2.3}$$

mit $\gamma = 1$, falls $\delta > L$, und γ wie in der Formulierung des Satzes, falls $\delta \leq L$. Aus Lemma 2.4 mit $a_1(u) = J(x,u)$ und $a_2(u) = J(y,u)$ folgt

$$|v(x) - v(y)| \leq \sup_{u\in\mathcal{U}}|J(x,u) - J(y,u)|,$$

was zusammen mit (2.3) die Behauptung liefert. □

2.4 Das Bellman'sche Optimalitätsprinzip

Wir werden nun die Eigenschaft der optimalen Wertefunktion beweisen, die die Basis für die numerische Approximation darstellt. Es handelt sich dabei um das sogenannte *Bellman'sche Optimalitätsprinzip*[3], auch *Prinzip der Dynamischen Programmierung* genannt. Es besagt,

[2]Dieses Lemma sollte aus der Einführung in die gewöhnlichen Differentialgleichungen bekannt sein, siehe z.B. das Buch *Gewöhnliche Differentialgleichungen* von B. Aulbach [1] oder mein Skript zur „Numerik dynamischer Systeme"

[3]Benannt nach dem amerikanischen Mathematiker Richard E. Bellman (1920–1984), dem die Erfindung dieses Prinzips zugeschrieben wird

dass Endstücke optimaler Trajektorien wieder optimale Trajektorien sind. Eine andere Sichtweise dieses Prinzips ist, dass wir den optimalen Wert $v(x)$ in einem Punkt erhalten, wenn wir für eine (beliebig kurze oder lange) endliche Zeit optimal steuern und dabei den Wert von v in dem erreichten Punkt berücksichtigen. Formal lässt sich dies wie folgt fassen.

Satz 2.7 Betrachte das optimale Steuerungsproblem aus Definition 2.1. Dann erfüllt die optimale Wertefunktion $v(x)$ für jedes $x \in \mathbb{R}^d$ und jedes $T > 0$ die Gleichung

$$v(x) = \sup_{u \in \mathcal{U}} \left\{ \int_0^T e^{-\delta t} g(\Phi(t,x,u), u(t))\, dt + e^{-\delta T} v(\Phi(T,x,u)) \right\}$$

bzw. für jedes $k \in \mathbb{N}$ die Gleichung

$$v_h(x) = \sup_{u_h \in \mathcal{U}_h} \left\{ h \sum_{j=0}^{k} (1-\delta h)^j g(\Phi_h(jh,x,u_h), u_h(jh)) + (1-\delta h)^{k+1} v_h(\Phi_h((k+1)h,x,u_h)) \right\}$$

Beweis: Wiederum beweisen wir das Resultat in kontinuierlicher Zeit, der zeitdiskrete Fall folgt analog.

„$\leq$“: Seien $x \in \mathbb{R}^d$, $T > 0$ und $u \in \mathcal{U}$ beliebig. Dann gilt

$$\begin{aligned} J(x,u) &= \int_0^\infty e^{-\delta t} g(\Phi(t,x,u), u(t)) dt \\ &= \int_0^T e^{-\delta t} g(\Phi(t,x,u), u(t)) dt \;+\; \int_T^\infty e^{-\delta t} g(\Phi(t,x,u), u(t)) dt \\ &\leq \int_0^T e^{-\delta t} g(\Phi(t,x,u), u(t)) dt \;+\; e^{-\delta T} v(\Phi(t,x,u)) \end{aligned}$$

und da dies für jedes beliebige $u \in \mathcal{U}$ gilt, gilt die Ungleichung auch für das Supremum und damit für $v(x)$.

„$\geq$“: Seien $x \in \mathbb{R}^d$, $T > 0$ und $\varepsilon > 0$ beliebig. Wähle $u_1 \in \mathcal{U}$ so, dass

$$\begin{aligned} &\sup_{u \in \mathcal{U}} \left\{ \int_0^T e^{-\delta t} g(\Phi(t,x,u), u(t)) dt + e^{-\delta T} v(\Phi(T,x,u)) \right\} \\ &\leq \int_0^T e^{-\delta t} g(\Phi(t,x,u_1), u_1(t)) dt + e^{-\delta T} v(\Phi(T,x,u_1)) \;+\; \varepsilon \end{aligned}$$

Dadurch ist u_1 auf $[0,T]$ festgelegt. Wähle nun $u_1|_{(T,\infty)}$ so, dass

$$J(\Phi(T,x,u_1), u_1(T+\cdot)) \geq v(\Phi(T,x,u_1)) \;-\; \varepsilon$$

Damit ergibt sich

$$\begin{aligned}
&\sup_{u\in\mathcal{U}}\left\{\int_0^T e^{-\delta t}g(\Phi(t,x,u),u(t))dt + e^{-\delta T}v(\Phi(T,x,u))\right\}\\
&\leq \int_0^T e^{-\delta t}g(\Phi(t,x,u_1),u_1(t))dt + e^{-\delta T}J(\Phi(T,x,u_1),u_1(T+\cdot)) \;+\; (1+e^{-\delta T})\varepsilon\\
&\leq \int_0^T e^{-\delta t}g(\Phi(t,x,u_1),u_1(t))dt + \int_T^\infty e^{-\delta t}g(\Phi(t,x,u_1),u_1(t))dt \;+\; (1+e^{-\delta T})\varepsilon\\
&= J(x,u_1) \;+\; (1+e^{-\delta T})\varepsilon \quad\leq\quad v(x) \;+\; (1+e^{-\delta T})\varepsilon
\end{aligned}$$

und da $\varepsilon > 0$ beliebig war somit die Behauptung. □

Der folgende Satz zeigt, dass $v(x)$ durch das Optimalitätsprinzip tatsächlich sogar eindeutig bestimmt ist.

Satz 2.8 Betrachte das optimale Steuerungsproblem aus Definition 2.1 mit optimaler Wertefunktion v. Sei ein $T > 0$ gegeben und sei Sei $w : \mathbb{R}^d \to \mathbb{R}$ eine beschränkte Funktion, die das Optimalitätsprinzip

$$w(x) = \sup_{u\in\mathcal{U}}\left\{\int_0^T e^{-\delta t}g(\Phi(t,x,u),u(t))\,dt + e^{-\delta T}w(\Phi(T,x,u))\right\}$$

(bzw. das Gegenstück in diskreter Zeit) für alle $x \in \mathbb{R}^d$ erfüllt. Dann ist $w = v$ (bzw. $w = v_h$).

Beweis: Für alle $x \in \mathbb{R}^d$ erhalten wir mit Lemma 2.4 angewendet auf

$$a_1(u) = \int_0^T e^{-\delta t}g(\Phi(t,x,u),u(t))\,dt + e^{-\delta T}w(\Phi(T,x,u))$$

und

$$a_2(u) = \int_0^T e^{-\delta t}g(\Phi(t,x,u),u(t))\,dt + e^{-\delta T}v(\Phi(T,x,u))$$

die Ungleichung

$$\begin{aligned}
|w(x)-v(x)| \;&\leq\; \sup_{u\in\mathcal{U}}\Bigg|\int_0^T e^{-\delta t}g(\Phi(t,x,u),u(t))\,dt + e^{-\delta T}w(\Phi(T,x,u))\\
&\qquad\qquad - \int_0^T e^{-\delta t}g(\Phi(t,x,u),u(t))\,dt + e^{-\delta T}v(\Phi(T,x,u))\Bigg|\\
&= \sup_{u\in\mathcal{U}} e^{-\delta T}|w(\Phi(T,x,u)) - v(\Phi(T,x,u))|\\
&\leq e^{-\delta T}\sup_{y\in\mathbb{R}^d}|w(y)-v(y)|.
\end{aligned}$$

Da dies für alle $x \in \mathbb{R}^d$ gilt, folgt

$$\sup_{y\in\mathbb{R}^d} |w(y) - v(y)| \leq e^{-\delta T} \sup_{y\in\mathbb{R}^d} |w(y) - v(y)|$$

und damit

$$(1 - e^{-\delta T}) \sup_{y\in\mathbb{R}^d} |w(y) - v(y)| \leq 0.$$

Da $1 - e^{-\delta T} > 0$, folgt daraus $\sup_{y\in\mathbb{R}^d} |w(y) - v(y)| = 0$, also $w = v$. □

2.5 Die Hamilton–Jacobi–Bellman Gleichung

In diesem Abschnitt werden wir eine partielle Differentialgleichung erster Ordnung kennen lernen, die von der kontinuierlichen Wertefunktion v erfüllt wird. Zwar werden wir diese Gleichung im Weiteren nicht verwenden, wegen Ihrer Bedeutung für die Theorie der optimalen Steuerung wollen wir sie aber zumindest kurz vorstellen.

Satz 2.9 Gegeben sei ein optimales Steuerungproblem aus Definition 2.1 mit optimaler Wertefunktion v. Betrachte die partielle Differentialgleichung

$$\delta v(x) + \inf_{u\in U}\{-Dv(x) \cdot f(x,u) - g(x,u)\} = 0,$$

die sogenannte *Hamilton–Jacobi–Bellman Gleichung*.

Dann gilt: Ist die optimale Wertefunktion v differenzierbar in $x \in \mathbb{R}^d$, so erfüllt v die Hamilton–Jacobi–Bellman Gleichung in diesem Punkt.

Beweis: Das Optimalitätsprinzip besagt, dass für alle $T > 0$ die Gleichung

$$v(x) = \sup_{u\in\mathcal{U}} \left\{ \int_0^T e^{-\delta t} g(\Phi(t,x,u), u(t))\, dt + e^{-\delta T} v(\Phi(T,x,u)) \right\}$$

gilt. Durch Umstellen der Terme erhält man

$$\inf_{u\in\mathcal{U}} \left\{ \frac{v(x) - e^{-\delta T} v(\Phi(T,x,u))}{T} - \frac{1}{T} \int_0^T e^{-\delta t} g(\Phi(t,x,u), u(t))\, dt \right\} = 0.$$

Da v nach Annahme in x differenzierbar ist, folgt damit für $T \to 0$

$$\inf_{u\in\mathcal{U}} \left\{ \delta v(x) - Dv(x) \cdot \frac{d}{d\tau}\bigg|_{\tau=0} \Phi(\tau,x,u) - \frac{d}{d\tau}\bigg|_{\tau=0} \int_0^\tau e^{-\delta t} g(\Phi(t,x,u), u(t))\, dt \right\} = 0.$$

Mit einigen technischen Überlegungen, die wir hier nicht ausführen wollen, sieht man, dass das Infimum (im Limes für $T \to 0$, nicht für festes $T > 0$!) tatsächlich über konstante Kontrollfunktionen $u \equiv u_0 \in U$ genommen werden kann. Für konstante Kontrollen gilt aber

$$\frac{d}{d\tau}\bigg|_{\tau=0} \Phi(\tau,x,u) = f(x,u_0) \text{ und } \frac{d}{d\tau}\bigg|_{\tau=0} \int_0^\tau e^{-\delta t} g(\Phi(t,x,u), u(t))\, dt = g(x,u_0)$$

und damit die Behauptung. □

Wir haben bereits erwähnt, dass man nicht erwarten kann, dass die optimale Wertefunktion v differenzierbar ist. Es gibt aber einen Lösungsbegriff für partielle Differentialgleichungen, der ohne Differenzierbarkeit auskommt, da er mit sogenannten Sub– und Superdifferentialen arbeitet, d.h. statt

$$Dv(x)(y-x) = v(y) - v(x) \pm o(\|y-x\|)$$

wird nur „$\leq$“ oder „$\geq$“ verlangt. Dieses Konzept der sogenannten *Viskositätslösungen* wurde von M.G. Crandall, L.C. Evans und P.L. Lions um 1980 eingeführt, (siehe z.B. [3, 4, 16]), und erlaubt insbesondere einen Existenz– und Eindeutigkeitssatz für nichtdifferenzierbare Lösungen von Hamilton–Jacobi–Bellman Gleichungen. Für Interessierte empfiehlt sich das Buch [2] von M. Bardi und I. Capuzzo Dolcetta.

Wir wollen hier noch erwähnen, dass jedes numerische Schema zur Berechnung von optimalen Wertefunktionen dadurch auch als Schema zur Lösung von Hamilton–Jacobi–Bellman Gleichungen interpretiert werden kann, und umgekehrt. Insbesondere kann man numerische Schemata mit Hilfe dieser Gleichung analysieren, ohne das zugrundeliegende optimale Steuerungsproblem explizit zu betrachten.

Kapitel 3

Diskretisierung des optimalen Steuerungsproblems

3.1 Diskretisierung in der Zeit

Das numerische Verfahren, mit dessen Herleitung und Analyse wir nun beginnen wollen, besteht aus zwei voneinander weitgehend unabhängigen Schritten. In diesem Abschnitt werden wir den ersten Schritt betrachten, die Diskretisierung in der Zeit, in der wir ein zeitkontinuierliches Problem (1.1) durch ein zeitdiskretes Problem (1.2) approximieren. Wenn das Problem von vornherein in diskreter Zeit vorliegt, ist dieser Schritt nicht nötig.

Aus Definition 1.12 und Satz 1.13 wissen wir, dass wir zu einem Kontrollsystem eine diskrete Euler–Approximation konstruieren können, die jedem Anfangswert $x \in \mathbb{R}^d$ und jeder diskreten Kontrollunktion $u_h : h\mathbb{Z} \to U$ eine zeitdiskrete approximative Lösung $\widetilde{\Phi}_h(t, x, u_h)$ liefert.

Für ein gegebenes kontinuierliches optimales Steuerungsproblem betrachten wir nun das zu dieser Euler–Approximation gehörige diskrete optimale Steuerungsproblem gegeben durch

$$\tilde{v}_h(x) := \sup_{u_h \in \mathcal{U}_h} \widetilde{J}_h(x, u_h) \quad \text{mit} \quad \widetilde{J}_h(x, u_h) := h \sum_{j=0}^{\infty} (1 - \delta h)^j g(\widetilde{\Phi}_h(jh, x, u_h), u_h(jh)). \quad (3.1)$$

3.1.1 Diskretisierungsfehler

Wir wollen nun untersuchen in welcher Form $\tilde{v}_h$ eine Approximation von v darstellt. Wir werden beweisen, dass $\tilde{v}_h \to v$ konvergiert und den Diskretisierungsfehler abschätzen. Wie schon bei der Diskretisierung der Trajektorien werden wir uns beim Beweis des Konvergenzsatzes für $v_h \to v$ hier auf eine einfachere Klasse von optimalen Steuerungsproblemen beschränken.

Definition 3.1 Wir nennen das optimale Steuerungsproblem aus Definition 2.1 *konvex*, falls die Menge

$$\left\{ \begin{pmatrix} f(x,u) \\ g(x,u) \end{pmatrix}, u \in U \right\} \subset \mathbb{R}^{d+1}$$

für jedes $x \in \mathbb{R}^d$ konvex ist. □

Der folgende Satz zeigt die Beziehung zwischen v und $\tilde{v}_h$.

Satz 3.2 Betrachte ein optimales Steuerungsproblem aus Definition 2.1 sowie das zugehörige Euler–diskretisierte optimale Steuerungsproblem (3.1). Wir nehmen an, dass das zugrundeliegende Kontrollsystem die Voraussetzungen von Satz 1.8 und Satz 1.13 erfüllt. Dann gelten für die optimalen Wertefunktionen v und $\tilde{v}_h$ und alle $h \in [0, 1/\delta]$ die folgenden Abschätzungen für alle $x \in \mathbb{R}^d$, $\gamma \in (0, 1]$ aus Satz 2.6 und eine passende Konstante $K > 0$.

(i) $v(x) \leq \tilde{v}_h(x) + K(h^{\frac{\gamma}{2}} + h)$

Ist das optimale Steuerungsproblem konvex, so gilt die schärfere Abschätzung

$v(x) \leq \tilde{v}_h(x) + K(h^{\gamma} + h)$

(ii) $\tilde{v}_h(x) \leq v(x) + K(h^{\gamma} + h)$

Insbesondere gilt also für eine passend Konstante $\widetilde{K} > 0$ und alle $x \in \mathbb{R}^d$ die Abschätzung

$$|v(x) - \tilde{v}_h(x)| \leq \widetilde{K} h^{\frac{\gamma}{2}}$$

im allgemeinen Fall, bzw.

$$|v(x) - \tilde{v}_h(x)| \leq \widetilde{K} h^{\gamma}$$

im konvexen Fall.

Beweis: Wie bereits erwähnt, beschränken wir uns im Teil (i) wieder auf den konvexen Fall. Ein Beweis für den nicht–konvexen Fall finden sich—mit ähnlichen Techniken, wie wir sie hier verwenden—in der Arbeit [8] von R. L. V. González and M. M. Tidball. Ein Beweis, der die Hamilton–Jacobi–Bellman Gleichung verwendet, findet sich in [2, Theorem 1.5 in Kapitel VI].

Wir zeigen nun zunächst die folgende Eigenschaft: Seien $x \in \mathbb{R}^d$, $u \in \mathcal{U}$ und $u_h \in \mathcal{U}_h$ so, dass die Identitäten

$$f(\Phi(hi, x, u), u_h(hi)) = \frac{1}{h} \int_{hi}^{h(i+1)} f(\Phi(hi, x, u), u(t))\, dt \tag{3.2}$$

und

$$g(\Phi(hi, x, u), u_h(hi)) = \frac{1}{h} \int_{hi}^{h(i+1)} g(\Phi(hi, x, u), u(t))\, dt \tag{3.3}$$

für alle $i \in \mathbb{N}_0$ gelten. Dann gilt die Abschätzung

$$|J(x, u) - \widetilde{J}_h(x, u_h)| \leq K(h^{\gamma} + h) \tag{3.4}$$

für γ aus Satz 2.6 und eine passende Konstante $K > 0$, wobei γ und K unabhängig von x und u sind.

Zum Beweis von (3.4) definieren wir zunächst $[t]_h$ als die größte ganze Zahl $i \in \mathbb{N}_0$, mit $hi \leq t$. Wir schreiben zudem kurz $\beta = (1 - \delta h)$. Damit gilt

$$\widetilde{J}_h(x, u_h) = \int_0^\infty \beta^{[t]_h} g(\widetilde{\Phi}_h(h[t]_h, x, u_h), u(h[t]_h))\, dt.$$

Mit der Dreiecksungleichung erhalten wir

$$\begin{aligned}
&|J(x,u) - \widetilde{J}_h(x,u_h)| \\
&\leq \left| \int_0^\infty e^{-\delta t} g(\Phi(t,x,u), u(t))\, dt - \int_0^\infty e^{-\delta t} g(\Phi(h[t]_h, x, u), u(t))\, dt \right| && (3.5) \\
&\quad + \left| \int_0^\infty e^{-\delta t} g(\Phi(h[t]_h, x, u), u(t))\, dt - \int_0^\infty \beta^{[t]_h} g(\Phi(h[t]_h, x, u), u(t))\, dt \right| && (3.6) \\
&\quad + \left| \int_0^\infty \beta^{[t]_h} g(\Phi(h[t]_h, x, u), u(t))\, dt - \int_0^\infty \beta^{[t]_h} g(\widetilde{\Phi}_h(h[t]_h, x, u_h), u_h(h[t]_h))\, dt \right| && (3.7)
\end{aligned}$$

Wir schätzen nun die Terme (3.5)–(3.7) einzeln ab. Für (3.5) nutzen wir aus, dass aus der Beschränktheit $|f(x,u)| \leq M$ die Ungleichung $\|\Phi(t,x,u) - \Phi(h[t]_h, x, u)\| \leq Mh$ gilt, und damit

$$\begin{aligned}
\left| \int_0^\infty e^{-\delta t} g(\Phi(t,x,u), u(t))\, dt - \int_0^\infty e^{-\delta t} g(\Phi(h[t]_h, x, u), u(t))\, dt \right| &\leq \int_0^\infty e^{-\delta t} L_g M h\, dt \\
&= K_1 h.
\end{aligned}$$

Zum Abschätzen von (3.6) schreiben wir

$$\beta = 1 - \delta h = e^{-\theta \delta h},$$

was für $\theta = -\ln(1 - \delta h)/(\delta h)$ gilt. Aus der Taylor Entwicklung $-\ln(1-r) = r + r^2/2 + r^3/3 + \ldots$ folgt $\theta > 1$ und $\theta - 1 \leq C_1 h$ für alle hinreichend kleinen h. Für $i \in \mathbb{N}_0$ folgt damit

$$|\beta^i - e^{-\delta i h}| = |e^{-\theta \delta i h} - e^{-\delta i h}| \leq \max\{e^{-\delta i h},\, e^{-\theta \delta i h}\} |\theta \delta i h - \delta i h| \leq e^{-\delta i h} C_1 h \delta i h,$$

wobei wir in der ersten Ungleichung den Mittelwertsatz der Differentialgleichung für e^r und in der zweiten Ungleichung $\theta > 1$ und $\theta - 1 \leq C_1 h$ verwendet haben.

Für allgemeine $t \geq 0$ gilt damit

$$\begin{aligned}
|\beta^{[t]_h} - e^{-\delta t}| &\leq |\beta^{[t]_h} - e^{-\delta h [t]_h}| + |e^{-\delta h [t]_h} - e^{-\delta t}| \\
&\leq e^{-\delta h [t]_h} C_1 h \delta h [t]_h + e^{-\delta t} C_2 h \;\leq\; C_3 h e^{-\delta t} (\delta t + 1)
\end{aligned}$$

wobei der zweite Summand abgeschätzt wurde mittels

$$|e^{-\delta h [t]_h} - e^{-\delta t}| \leq |e^{-\delta (t-h)} - e^{-\delta t}| = e^{-\delta t} |e^{\delta h} - 1| \leq e^{-\delta t} C_2 h.$$

Hierbei folgt die letzte Ungleichung aus der Reihendarstellung von $e^{\delta h}$.

Damit gilt

$$\begin{aligned}
&\left| \int_0^\infty e^{-\delta t} g(\Phi(h[t]_h, x, u), u(t))\, dt - \int_0^\infty \beta^{[t]_h} g(\Phi(h[t]_h, x, u), u(t))\, dt \right| \\
&\quad \leq \int_0^\infty e^{-\delta t} (\delta t + 1) C_3 h\, dt \;=\; K_2 h,
\end{aligned}$$

wobei wir

$$\int_0^\infty e^{-\delta t}dt = \frac{1}{\delta} \quad \text{und} \quad \int_0^\infty e^{-\delta t}\delta t dt = \frac{1}{\delta}$$

verwendet haben.

Für (3.7) beachte, dass aus (3.3) die Identität

$$\int_0^\infty \beta^{[t]_h} g(\Phi(h[t]_h, x, u), u(t))\, dt = \int_0^\infty \beta^{[t]_h} g(\Phi(h[t]_h, x, u), u_h(h[t]_h))\, dt$$

folgt. Desweiteren folgt aus (3.2), dass u_h eine diskrete Kontrollfunktion ist, für die Satz 1.13(i) im konvexen Fall gilt, also insbesondere

$$\|\Phi(h[t]_h, x, u) - \widetilde{\Phi}_h(h[t]_h, x, u_h)\| \le Che^{Lt}$$

gilt für ein $C > 0$. Wegen $\beta^{[t]_h} \le e^1 e^{-\delta t}$ folgt

$$\begin{aligned}&\left|\int_0^T \beta^{[t]_h} g(\Phi(h[t]_h, x, u), u(t))\, dt - \int_0^T \beta^{[t]_h} g(\widetilde{\Phi}_h(h[t]_h, x, u_h), u_h(h[t]_h))\, dt\right| \\ &\le e^1 \int_0^T e^{-\delta t} L_g Che^{Lt}\, dt \le e^1 L_g Ch \frac{e^{(L-\delta)T} - 1}{L - \delta}\end{aligned}$$

für $\delta \neq L$, bzw. $\le e^1 L_g ChT$, falls $\delta = L$. Aus Lemma 2.5 (mit $C = e^1 L_g Ch$ und $D = 0$) folgt also

$$\left|\int_0^\infty \beta^{[t]_h} g(\Phi(h[t]_h, x, u), u(t))\, dt - \int_0^\infty \beta^{[t]_h} g(\widetilde{\Phi}_h(h[t]_h, x, u_h), u_h(h[t]_h))\, dt\right| \le K_3 h^\gamma$$

und damit (3.4) mit $K = \max\{K_1 + K_2,\, K_3\}$.

Wir zeigen nun (i). Analog zur Konstruktion im Beweis von Satz 1.13 folgt aus der Konvexität, dass zu beliebigem $u \in \mathcal{U}$ eine diskrete Kontrollfunktion $u_h \in \mathcal{U}_h$ existiert, so dass (3.2) und (3.3) erfüllt sind. Also folgt aus (3.4) für alle $u \in \mathcal{U}$ die Existenz von $u_h \in \mathcal{U}_h$ mit

$$\tilde{v}_h(x) \ge \widetilde{J}_h(x, u_h) \ge J(x, u) - K(h^\gamma + h)$$

und damit (i), da wir auf der rechten Seite zum Supremum über u übergehen können.

Zum Beweis von (ii) sei $u_h \in \mathcal{U}_h$ beliebig. Dann erfüllt die stückweise konstante Kontrollfunktion $u(t) = u_h(h[t]_h)$ offenbar (3.2) und (3.3). Mit (3.4) folgt

$$v(x) \ge J(x, u) \ge \widetilde{J}_h(x, u_h) - K(h^\gamma + h),$$

also (ii), da wir auch hier auf der rechten Seite zum Supremum (jetzt über u_h) übergehen können. □

Bemerkung 3.3 Analog zum Satz 1.13 können wir die Menge U durch eine hinreichend große endliche Menge $\widetilde{U} \subset U$ ersetzen, so dass die Aussage von Satz 3.2 gültig bleibt, vgl. Bemerkung 1.16. Wir können also bei Bedarf annehmen, dass U eine endliche Menge ist.

□

Bemerkung 3.4 Nach Satz 2.7 erfüllt $\tilde{v}_h$ das Optimalitätsprinzip

$$\tilde{v}_h(x) = \sup_{u_h \in \mathcal{U}_h} \left\{ h \sum_{i=0}^{k} \beta^i g(\widetilde{\Phi}_h(ih, x, u_h), u_h(ih)) + \beta^{k+1} \tilde{v}_h(\widetilde{\Phi}_h((k+1)hx, u_h)) \right\}. \quad (3.8)$$

Da $\widetilde{\Phi}_h$ und g stetig in $(u_h(0), \ldots, u_h(k))$ sind und $\tilde{v}_h$ stetig ist, können wir das Supremum hier tatsächlich als Maximum schreiben. Insbesondere folgt daraus, dass wir für $k = 0$ zu jedem $x \in \mathbb{R}^d$ mindestens ein $u_x^* \in U$ finden, so dass das Supremum in (3.8) für ein $u_h \in \mathcal{U}_h$ mit $u(0) = u_x^*$ angenommen wird. □

3.1.2 Ein Iterationsverfahren

Aus dem Optimalitätsprinzip (3.8) lässt sich eine Iterationsformel zur Berechnung zeitdiskreter optimaler Wertefunktionen v_h herleiten, die wir als Basis für unsere numerische Approximation verwenden werden. Beachte dafür, dass die diskrete Lösungsfunktion $\Phi_h(t, x, u_h)$ eines diskreten Kontrollsystems durch die Iteration einer Abbildung $f_h(x, u)$ definiert ist.

Wir werden nun ein Iterationsverfahren definieren und analysieren, das für die Wertefunktionen v_h allgemeiner zeitdiskreter optimaler Steuerungsprobleme funktioniert, insbesondere aber auch für die oben eingeführte zeitliche Diskretisierung $\tilde{v}_h$.

Definition 3.5 Wir definieren iterativ Funktionen $v_h^i : \mathbb{R}^d \to \mathbb{R}$, $i = 0, 1, \ldots$ mittels $v_h^0(x) = 0$ und $v_h^{i+1}(x) = T_h(v_h^i)(x)$ für alle $x \in \mathbb{R}^d$, wobei der Operator $T_h : C(\mathbb{R}^d, \mathbb{R}) \to C(\mathbb{R}^d, \mathbb{R})$ gegeben ist durch

$$T_h(w)(x) := \max_{u \in U} \{hg(x, u) + \beta w(f_h(x, u))\}$$

mit $\beta = 1 - \delta h$. Hierbei bezeichnet $C(\mathbb{R}^d, \mathbb{R})$ die Menge der stetigen Funktionen von $\mathbb{R}^d$ nach $\mathbb{R}$. □

Wir werden zeigen, dass die durch diese Iteration erzeugte Funktionenfolge tatsächlich gegen v_h konvergiert.

Satz 3.6 Betrachte das zeitdiskretisierte optimale Steuerungsproblem (3.1) mit optimaler Wertefunktion v_h. Es sei $\delta h < 1$. Dann gilt für die in Definition 3.5 definierten Funktionen die Abschätzung

$$\|v_h^i - v_h\|_\infty \leq \beta^i \frac{M_g}{\delta}.$$

Wegen $\beta < 1$ folgt also insbesondere folgt die Konvergenz $v_h^i(x) \to v_h(x)$ gleichmäßig für alle $x \in \mathbb{R}^d$.

Beweis: Betrachte zwei beliebige Funktionen w_1, $w_2 : \mathbb{R}^d \to \mathbb{R}$. Dann folgt aus Lemma 2.4

$$|T_h(w_1)(x) - T_h(w_2)(x)| \leq \beta \sup_{u \in U} |w_1(f_h(x, u)) - w_2(f_h(x, u))| \leq \beta \|w_1 - w_2\|_\infty$$

und damit

$$\|T_h(w_1) - T_h(w_2)\|_\infty \leq \beta \|w_1 - w_2\|_\infty,$$

vgl. Übungsaufgabe 4/Blatt 2. Mit dem Optimalitätsprinzip (3.8) für $k = 0$ ergibt sich nun die Gleichung $v_h = T_h(v_h)$. Damit und mit der Definition der v_h^i erhalten wir

$$\|v_h - v_h^{i+1}\|_\infty = \|T_h(v_h) - T_h(v_h^i)\|_\infty \leq \beta \|v_h - v_h^i\|_\infty.$$

Wie im Beweis von Lemma 2.3(i) sieht man $\|v_h\|_\infty \leq hM_g/(\delta h)$, woraus

$$\|v_h - v_h^0\|_\infty = \|v_h\|_\infty \leq \frac{hM_g}{\delta h} = \beta^0 \frac{M_g}{\delta}$$

folgt. Also ergibt sich die Behauptung leicht durch Induktion über i. □

Das folgende Lemma zeigt, dass jedes der v_h^i Lipschitz stetig ist, wobei wir sogar die Lipschitz–Konstante explizit angeben können.

Lemma 3.7 Es sei $\delta h < 1$. Dann sind die Funktionen v_h^i aus Definition 3.5 Lipschitz–stetig mit Konstanten $L_0 = 0$ und

$$L_i \leq hL_g \sum_{k=0}^{i-1} e^{(L-\delta)hk} \leq \begin{cases} e^{h(\delta-L)} \frac{L_g}{\delta - L}, & \delta > L \\ hiL_g, & \delta = L \\ \frac{L_g}{L-\delta} e^{(L-\delta)hi}, & \delta < L \end{cases} \tag{3.9}$$

für $i \geq 1$.

Beweis: Die zweite Ungleichung in (3.9) ist klar für $\delta = L$; für $\delta \neq L$ folgt sie aus

$$hL_g \sum_{k=0}^{i-1} e^{(L-\delta)hk} \leq CL_g \int_0^{hi} e^{(L-\delta)t} dt = C \frac{L_g}{L-\delta} (e^{(L-\delta)hi} - 1)$$

mit $C = e^{\max\{h(\delta-L),0\}}$. Wir zeigen nun mittels Induktion die erste Ungleichung in (3.9). Die Funktion v_h^0 ist konstant, also Lipschitz–stetig mit Konstante $L_0 = 0$. Nehmen wir nun an, dass v_h^i Lipschitz–stetig mit Konstante L_i ist.

Mit Lemma 2.4 ergibt sich

$$\begin{aligned} |v_h^{i+1}(x) - v_h^{i+1}(y)| &= |T_h(v_h^i)(x) - T_h(v_h^i)(y)| \\ &\leq \sup_{u \in U} \left\{ |hg(x,u) - hg(y,u) + \beta v_h^i(f_h(x,u)) - \beta v_h^i(f_h(y,u))| \right\}. \end{aligned}$$

Für jedes $u \in U$ lässt sich dieser Term abschätzen durch

$$\begin{aligned} &|hg(x,u) - hg(y,u) + \beta v_h^i(\Phi_h(x,u)) - \beta v_h^i(\Phi_h(y,u))| \\ &\leq |hg(x,u) - hg(y,u)| + \beta |v_h^i(\Phi_h(x,u)) - v_h^i(\Phi_h(y,u))| \\ &\leq hL_g \|x-y\| + L_i(1+hL)\beta \|x-y\| \leq (hL_g + L_i(1+hL)e^{-\delta h}) \|x-y\| \end{aligned}$$

wobei wir im letzten Schritt die Ungleichung $\beta \leq e^{-\delta h}$ verwendet haben. Im Fall $i = 0$ folgt nun $hL_g + L_i(1 + hL)e^{-\delta h} = hL_g = L_1$, also die Behauptung. Im Fall $i \geq 1$ folgt

$$\begin{aligned} hL_g + L_i(1 + hL)e^{-\delta h} &\leq hL_g + L_i e^{hL} e^{-\delta h} \leq hL_g + e^{h(L-\delta)} hL_g \sum_{k=0}^{i-1} e^{(L-\delta)hk} \\ &= hL_g + hL_g \sum_{k=0}^{i-1} e^{(L-\delta)h(k+1)} = L_{i+1}, \end{aligned}$$

also ebenfalls die Behauptung. □

3.1.3 Zustandsraumbeschränkung

Bevor wir im nächsten Abschnitt einen approximierenden endlichdimensionalen Funktionenraum einführen, um die Iterationsvorschrift aus Definition 3.5 in eine implementierbare Form zu bringen, wollen wir uns noch kurz Gedanken zur Einschränkung des Definitionsbereiches von v_h bzw. $\tilde{v}_h$ auf eine kompakte Menge $\Omega \subset \mathbb{R}^d$ machen. Wir betrachten dabei in der Optimierung nur solche Lösungen Φ_h bzw. $\widetilde{\Phi}_h$, die für alle positiven Zeiten in Ω bleiben. Dies ist nötig, da wir v_h numerisch nicht im ganzen $\mathbb{R}^d$ berechnen können. Oftmals ergibt sich eine geeignete Menge Ω aus dem Modell, indem man sich z.B. auf einen physikalisch interessanten Bereich einschränkt. Formal ist die Sache etwas komplizierter, im Wesentlichen gibt es die folgenden drei Möglichkeiten:

(1) Die Menge Ω ist *stark invariant*, d.h. für alle $x \in \Omega$ und alle $u \in U$ gilt $f_h(x, u) \in \Omega$. In diesem Fall gibt es kein Problem.

(2) Die Menge Ω ist *schwach invariant*, d.h. für alle $x \in \Omega$ gibt es (mindestens) ein $u \in U$ mit $f_h(x, u) \in \Omega$. In diesem Fall berücksichtigen wir nur diese $u \in U$ bei der Optimierung; wir optimieren damit nur über die Trajektorien, die für alle Zeiten in Ω bleiben. Falls es eine Teilmenge von Ω gibt, die *optimal invariant* ist, d.h. eine Menge $A \subseteq \Omega$ mit der Eigenschaft, dass $f_h(x, u_x^*) \in A$ für alle $x \in A$ und ein u_x^* aus Bemerkung 3.4, so folgt, dass sich v_h auf A dadurch nicht ändert.

(3) Die Menge Ω ist *nicht invariant*, d.h. es gibt ein $x \in \Omega$, so dass für alle $u \in U$ gilt $f_h(x, u) \notin \Omega$. Dann können wir entweder die Punkte $f_h(x, u)$ zurückprojizieren (d.h. wir ersetzen $f_h(x, u)$ durch den nächstgelegenen Punkt in Ω), oder wir definieren eine Funktion $\tilde{v}_h : \Omega^c \to \mathbb{R}$ und benutzen den entsprechenden Wert in der Iteration für Punkte außerhalb Ω. In diesem Fall ist es nicht a priori klar, dass die so erhaltene Lösung noch etwas mit v_h zu tun hat. Unter gewissen Voraussetzungen gilt aber die Aussage über optimal invariante Mengen aus (2). Wir werden das in den Übungen an Beispielen genauer diskutieren.

3.2 Diskretisierung im Raum

Obwohl alle Größen, die in der Iterationsvorschrift aus Definition 3.5 vorkommen, im Rechner — bis auf Rundungsfehler — auswertbar sind (zumindest wenn wir annehmen, dass

wir das „max“ berechnen können, z.B. wenn U eine endliche Menge ist, vgl. Bemerkung 3.3), können wir diese Vorschrift nicht direkt implementieren. Der Grund dafür ist, dass die Funktionen v_h^i für unendlich viele Punkte berechnet werden müssen. Selbst wenn wir uns auf eine kompakte Menge $\Omega \subset \mathbb{R}^d$ einschränken, was wir im Folgenden machen werden, löst dies das Problem noch nicht, denn auch in einer beliebig kleinen kompakten Menge gibt es im Allgemeinen unendlich viele Punkte (klarerweise ist es nicht zweckmäßig, sich auf eine endliche Menge zu beschränken).

3.2.1 Funktionen auf Gittern

Wir müssen uns also zur Berechnung von v_h auf einen endlichdimensionalen Funktionenraum einschränken. Wir werden uns hier in der Darstellung der Methoden auf den Fall $d = 2$ einschränken; die verwendeten Techniken lassen sich aber leicht auf beliebige Dimensionen verallgemeinern. Um technische Komplikationen zu vermeiden, werden wir annehmen, dass die kompakte Menge $\Omega \subset \mathbb{R}^2$, auf der wir v_h berechnen wollen, ein Rechteck ist.

Definition 3.8 Sei $\Omega \subset \mathbb{R}^2$ gegeben durch $\Omega = [a_1, b_1] \times [a_2, b_2]$ mit Werten $a_1 < b_1$ und $a_2 < b_2$. Ein *(regelmäßiges) Rechteckgitter* Γ auf Ω ist eine Menge von Rechtecken R_i, $i = 0, \dots, P-1$, $P = P_1 P_2$, mit Kantenlängen $k_1 = (b_1 - a_1)/P_1$ und $k_2 = (b_2 - a_2)/P_2$, so dass

$$\bigcup_{i=0}^{P-1} R_i = \Omega \quad \text{und} \quad \operatorname{int} R_i \cap \operatorname{int} R_j = \emptyset \text{ für alle } i, j = 0, \dots, P-1,\ i \neq j.$$

Mit E_i, $i = 0, \dots, N-1$, $N = (P_1 + 1)(P_2 + 1)$ bezeichnen wir die Eckpunkte (oder Knotenpunkte) des Gitters. Der Wert $k = \sqrt{k_1^2 + k_2^2}$ bezeichnet den maximalen Durchmesser eines Rechtecks. □

Abbildung 3.1 zeigt ein solches Gitter.

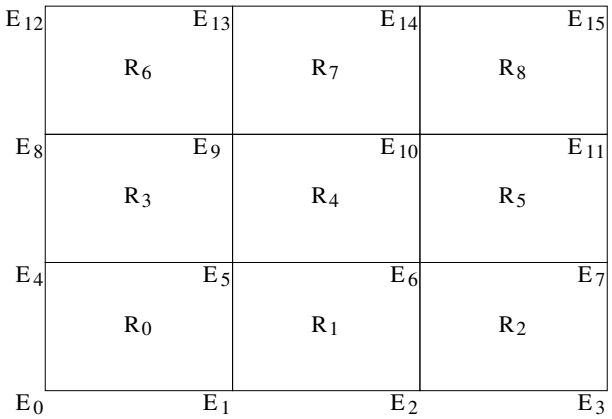

Abbildung 3.1: Beispielgitter

Wir definieren nun den Funktionenraum, den wir zur Approximation von v_h verwenden wollen.

Definition 3.9 (i) Sei $A \subset \mathbb{R}^2$. Eine Funktion $w : A \to \mathbb{R}$ heißt *affin bilinear*, falls es Konstanten $\alpha_0, \ldots, \alpha_3$ gibt, so dass für alle $x = (x_1, x_2)^T \in A$ die Identität $w(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1 x_2$ gilt.

(ii) Betrachte eine rechteckförmige Menge $\Omega \subset \mathbb{R}^2$ mit Rechteckgitter Γ. Wir definieren den Raum der stetigen und stückweise affin bilinearen Funktionen auf Ω bezüglich Γ als

$$\mathcal{W} := \{w : \Omega \to \mathbb{R} \,|\, w \text{ ist stetig und } w|_{R_i} \text{ ist affin bilinear für jedes } i = 0, \ldots, P-1\}.$$

□

Das folgende Lemma fasst die für uns wichtigen Eigenschaften von $\mathcal{W}$ zusammen.

Lemma 3.10 (i) Jede Funktion $w \in \mathcal{W}$ ist eindeutig durch ihre Werte $w(E_i)$ in den Eckpunkten des Gitters bestimmt.

(ii) Für jedes Rechteck $R_i = [c_1, d_1] \times [c_2, d_2]$ mit den Eckpunkten

$$E_{i_0} = (c_1, c_2)^T, \;\; E_{i_1} = (d_1, c_2)^T, \;\; E_{i_2} = (c_1, d_2)^T, \;\; E_{i_3} = (d_1, d_2)^T$$

lässt sich $w|_{R_i}$ für $x = (x_1, x_2)^T \in R_i$ schreiben als

$$w(x) = \sum_{j=0}^{3} \mu_j(x) w(E_{i_j})$$

mit

$$\begin{aligned} \mu_0(x) &= (1 - y_1(x))(1 - y_2(x)), & \mu_1(x) &= y_1(x)(1 - y_2(x)), \\ \mu_2(x) &= (1 - y_1(x)) y_2(x), & \mu_3(x) &= y_1(x) y_2(x) \end{aligned}$$

und

$$y_l(x) = \frac{x_l - c_l}{d_l - c_l} \text{ für } l = 1, 2.$$

Insbesondere gilt hierbei $\mu_j(x) \geq 0$ für $j = 0, \ldots, 3$ und $\sum_{j=0}^{3} \mu_j(x) = 1$.

Beweis: (i) Übungsaufgabe.

(ii) Man rechnet leicht nach, dass die angebene Funktion tatsächlich affin bilinear auf R_i ist (es tauchen nur Terme der Form α_0, $\alpha_1 x_1$, $\alpha_2 x_2$ oder $\alpha_3 x_1 x_2$ auf) und in den Eckpunkten des Rechtecks mit w übereinstimmt. Also folgt die Aussage aus Teil (i). Die behaupteten Eigenschaften für die $\mu_j(x)$ sind ebenfalls leicht zu sehen, wenn man ausnutzt, dass $y_l(x) \in [0, 1]$ gilt. □

Wir können also jede Funktion $w \in \mathcal{W}$ mit ihren Werten $w(E_i)$ in den Eckpunkten des Gitters identifizieren. Insbesondere ist der Funktionenraum $\mathcal{W}$ somit ein N–dimensionaler Vektorraum über $\mathbb{R}$.

3.2.2 Die vollständige Diskretisierung

Wir werden nun den iterativen Algorithmus aus Definition 3.5 für Funktionen aus $\mathcal{W}$ formulieren. Wie wir im letzten Abschnitt gesehen haben, reicht es aus, die Werte in den Eckpunkten des Gitters zu berechnen. Wir müssen also den Operator T_h in jedem Schritt nur an den Punkten E_i, $i = 0, \dots, N-1$ auswerten, d.h. wir berechnen eine Folge von Funktionen $\hat{v}_h^j \in \mathcal{W}$ mittels

$$\hat{v}_h^{j+1}(E_i) = \max_{u \in U} \left\{ hg(E_i, u) + \beta \hat{v}_h^j(f_h(E_i, u)) \right\},$$

mit $\beta = 1 - \delta h$. Schreiben wir $V^j = (V_1^j, \dots, V_N^j)^T \in \mathbb{R}^N$ mit $V_i^j = \hat{v}_h^j(E_i)$, so können wir diese Iteration auf $\mathcal{W}$ nun als eine Iteration auf N–dimensionalen Vektoren formulieren. Zu einem gegebenen Gitter berechnen wir also sukzessive Vektoren $V^j \in \mathbb{R}^N$ gemäß der folgenden Vorschrift.

Definition 3.11 Betrachte ein zeitdiskretes optimales Steuerungsproblem und ein Rechteckgitter Γ mit P Rechtecken und N Eckpunkten. Zu jedem $u \in U$ und jedem $i = 0, \dots, N-1$ sei $B(i,u)$ der N–dimensionale Zeilenvektor, für den für jedes $w \in \mathcal{W}$ und $W = (w(E_0), \dots, w(E_{N-1}))^T \in \mathbb{R}^N$ mit der üblichen Matrixmultiplikation gilt

$$w(\Phi_h(E_i, u)) = B(i,u)W$$

(beachte, dass $B(i,u)$ unabhängig von $w \in \mathcal{W}$ ist und höchstens 4 Einträge $\neq 0$ besitzt, nämlich gerade die μ_l aus Lemma 3.10(ii) in globaler Nummerierung, welche ≥ 0 sind und sich zu 1 aufsummieren). Desweiteren sei $G(i,u) = hg(E_i, u)$. Dann berechnen wir Vektoren V^j iterativ durch $V^0 := (0, \dots, 0)^T$ und dem *Gesamtschrittverfahren*

$$V_i^{j+1} := \max_{u \in U} \{ G(i,u) + \beta B(i,u) V^j \} \quad \text{für } i = 0, \dots, N-1$$

oder dem *Einzelschrittverfahren*

$$V^{j+1} := V^j, \quad V_i^{j+1} := \max_{u \in U} \{ G(i,u) + \beta B(i,u) V^{j+1} \} \quad \text{für } i = 0, \dots, N-1.$$

□

Bemerkung 3.12 (i) Im Allgemeinen ist das Einzelschrittverfahren vorteilhafter, da wir für jedes $i > 1$ bereits die aktuellen Werte V_k^{j+1} für $0 \leq k < i$ berücksichtigen, und somit eine (leicht) schnellere Konvergenz erwarten können. Außerdem müssen im Gesamtschrittverfahren jeweils zwei Vektoren V^j und V^{j+1} gespeichert werden, während das Einzelschrittverfahren auf einem einzigen Vektor durchgeführt werden kann.

(ii) Wenn wir U als endliche Menge annehmen (i.A. als endliche Approximation einer gegebenen kontinuierlichen Menge), also $U = \{u_1, \dots, u_q\}$ für ein $q \in \mathbb{N}$ gilt, kann das Maximum in dieser Iteration für jedes $i = 1, \dots, N$ durch Vergleich der Werte

$$G(x, u_k) + \beta B(i, u_k) V^j, \quad k = 1, \dots, q$$

bzw.

$$G(x, u_k) + \beta B(i, u_k) V^{j+1}, \quad k = 1, \dots, q$$

bestimmt werden.

Dieses Vorgehen ist die einfachste Art, das Maximum zu bestimmen und liefert für relativ grobe Genauigkeit brauchbare Ergebnisse. Für hohe Genauigkeit ist es sinnvoller, hier ein besseres kontinuierliches Optimierungsverfahren zu verwenden.

(iii) Falls U endlich ist und genügend Speicherplatz zur Verfügung steht, empfiehlt es sich in der praktischen Implementierung, die Werte $G(i,u)$ und die Vektoren $B(i,u_k)$ im Voraus zu berechnen und zu speichern, da dies der aufwändigste Teil des Algorithmus' ist. Natürlich sollte man dabei die Vektoren nicht komplett speichern sondern nur diejenigen Einträge, die ungleich Null sind. □

Das folgende Lemma gibt ein Abbruchkriterium für diese Iteration.

Lemma 3.13 Betrachte die Iterationsvorschrift aus Definition 3.11 und sei $\delta h < 1$. Dann konvergieren die Vektoren V^j für $j \to \infty$ komponentenweise gegen den Vektor V, der eindeutig bestimmt ist durch

$$V_i = \max_{u\in U}\{G(i,u) + \beta B(i,u)V\} \quad \text{für } i = 0,\dots,N-1.$$

Für die mit $\hat{v}_h^j$, $j = 1,\dots,\infty$ und $\hat{v}_h$ bezeichneten zugehörigen Funktionen aus $\mathcal{W}$ gilt darüberhinaus: Falls $|V_i^j - V_i^{j+1}| \le \varepsilon$ für alle $i = 0,\dots,N-1$, so folgt

$$\|\hat{v}_h^j - \hat{v}_h\|_\infty \le \frac{\varepsilon}{h\delta}.$$

Beweis: Beachte zunächst, dass mit Lemma 2.4 für beliebige Vektoren $V, W \in \mathbb{R}^N$ und alle $i = 0,\dots,N-1$ die Ungleichung

$$\left|\max_{u\in U}\{G(i,u) + \beta B(i,u)V\} - \max_{u\in U}\{G(i,u) + \beta B(i,u)W\}\right| \le \beta\|V - W\|_\infty \tag{3.10}$$

mit $\beta = 1 - \delta h$ folgt, wobei wir $\sum_{k=0}^{N-1} B(i,u)_k = 1$ ausgenutzt haben und $\|\cdot\|_\infty$ die L_∞–Norm im $\mathbb{R}^N$ bezeichnet. Also definiert (3.10) eine Kontraktion auf dem $\mathbb{R}^N$ bzgl. der L_∞–Norm, weswegen der Vektor V existiert. Dieser ist zudem eindeutig, denn für jeden weiteren solchen Vektor W gilt mit (3.10)

$$\|V - W\|_\infty \le \beta\|V - W\|_\infty < \|V - W\|_\infty,$$

woraus $W = V$ folgt.

Ebenfalls mit (3.10) sieht man leicht, dass die Vektoren V^j die Abschätzung

$$\|V^{j+1} - V\|_\infty \le \beta\|V^j - V\|_\infty$$

erfüllen. Daraus folgt die behauptete Konvergenz.

Außerdem folgt

$$\begin{aligned}\|V^j - V\|_\infty &\leq \|V^j - V^{j+1}\|_\infty + \|V^{j+1} - V\|_\infty \\ &\leq \varepsilon + \beta\|V^j - V^\infty\|_\infty\end{aligned}$$

und daraus

$$\|V^j - V\|_\infty \leq \frac{\varepsilon}{1-\beta} = \frac{\varepsilon}{h\delta}.$$

Die entsprechende Aussage für die Funktionen $\hat{v}_h^j$ folgt nun leicht mit der Darstellung aus Lemma 3.10(ii). □

In der Praxis zeigt sich, dass die Iterationsvorschrift aus Definition 3.11 recht langsam gegen V konvergiert, insbesondere für kleine h und δ. Wir später alternative Iterationen besprechen, die deutlich schneller konvergieren.

3.2.3 Diskretisierungsfehler

Wir wollen nun den Fehler abschätzen, der durch die Diskretisierung im Ort entsteht, d.h. wir wollen eine Abschätzung für die Differenz

$$\|v_h - \hat{v}_h\|_\infty$$

herleiten. Dazu betrachten wir zunächst die Projektion einer beliebigen Funktion nach $\mathcal{W}$.

Definition 3.14 Für eine Funktion $q : \Omega \to \mathbb{R}$ und ein Gitter Γ bezeichnen wir mit $\pi_{\mathcal{W}} q$ die (eindeutige) Funktion $w \in \mathcal{W}$ mit

$$w(E_i) = q(E_i) \text{ für alle } i = 0, \ldots, N-1$$

□

Das folgende Lemma gibt Auskunft über den dabei entstehenden Projektionsfehler, der auch als Interpolationsfehler bezeichnet wird.

Lemma 3.15 Sei $q : \Omega \to \mathbb{R}$ eine Lipschitz–stetige Funktion mit Lipschitz–Konstante L_q. Dann gilt

$$\|q - \pi_{\mathcal{W}} q\|_\infty \leq L_q k$$

mit dem Wert k aus Definition 3.8.

Beweis: Sei $x \in \Omega$ ein beliebiger Punkt und R_i ein Gitterrechteck, in dem dieser Punkt liegt. Seien $E_{i_0}, \ldots, E_{i_3}$ die Eckpunkte dieses Rechtecks. Dann gilt $\|x - E_{i_j}\| \leq k$ für $j = 0, \ldots, 3$ und somit $|q(x) - q(E_{i_j})| \leq L_q k$. Mit Lemma 3.10 folgt

$$|q(x) - \pi_{\mathcal{W}} q(x)| = \left| q(x) - \sum_{j=0}^{3} \mu_j(x) q(E_{i_j}) \right|$$

$$
\begin{aligned}
&= \left| \sum_{j=0}^{3} \mu_j(x) q(x) - \sum_{j=0}^{3} \mu_j(x) q(E_{i_j}) \right| \\
&\leq \sum_{j=0}^{3} \mu_j(x) \, |q(x) - q(E_{i_j})| \;=\; \sum_{j=0}^{3} \mu_j(x) \, L_q k \;=\; L_q k
\end{aligned}
$$

wobei wir im zweiten und im letzten Schritt ausgenutzt haben, dass $\sum_{j=0}^{3} \mu_j(x) = 1$ gilt. □

Mit Hilfe dieses Lemmas können wir nun zunächst den Fehler zwischen v_h^j und $\hat{v}_h^j$ abschätzen. Hierbei bezeichnet L, wie üblich, die Lipschitz–Konstante des Vektorfeldes f.

Lemma 3.16 Betrachte die Funktionen v_h^j und $\hat{v}_h^j$ aus den Definitionen 3.5 und 3.11. Dann gelten die Abschätzungen

$$\|v_h^j - \hat{v}_h^j\|_\infty \leq 2M_g e^{\delta h} \int_0^{jh} e^{-\delta t} \, dt \tag{3.11}$$

und, falls $\delta > L$,

$$\|v_h^j - \hat{v}_h^j\|_\infty \leq C \frac{k}{h} \tag{3.12}$$

bzw., falls $\delta < L$

$$\|v_h^j - \hat{v}_h^j\|_\infty \leq C \frac{k}{h} \int_0^{jh} e^{(L-\delta)t} \, dt \tag{3.13}$$

mit einer geeigneten Konstante $C > 0$.

Beweis: Beachte zunächst, dass man aus der Definition von $\pi_{\mathcal{W}}$ leicht die Gleichungen

$$\hat{v}_h^{j+1} = \pi_{\mathcal{W}} \hat{v}_h^{j+1} = \pi_{\mathcal{W}} T_h(\hat{v}_h^j)$$

erhält. Wir zeigen nun zunächst (3.11). Aus den Definitionen folgt, dass für die betrachteten Funktionen die Ungleichungen

$$\|v_h^{j+1}\|_\infty \leq hM_g + \beta \|v_h^j\|_\infty \quad \text{und} \quad \|\hat{v}_h^{j+1}\|_\infty \leq hM_g + \beta \|\hat{v}_h^j\|_\infty$$

gelten (beachte, dass $\|\pi_{\mathcal{W}} q\|_\infty \leq \|q\|_\infty$ gilt). Durch Induktion erhalten wir

$$\|v_h^j\|_\infty \leq \sum_{i=0}^{j-1} \beta^i hM_g \quad \text{und} \quad \|\hat{v}_h^j\|_\infty \leq \sum_{i=0}^{j-1} \beta^i hM_g.$$

Aus $\|v_h^j - \hat{v}_h^j\|_\infty \leq \|v_h^j\|_\infty + \|\hat{v}_h^j\|_\infty$ und

$$\sum_{i=0}^{j-1} \beta^i hM_g \leq e^{\delta h} \int_0^{jh} e^{-\delta t} M_g \, dt$$

folgt damit (3.11).

Zum Beweis von (3.12) und (3.13) verwenden wir Lemma 3.7, welches besagt, dass v_h^j Lipschitz–stetig mit der dort angegebenen Konstante L_j ist. Damit ergibt sich

$$\begin{aligned}\|v_h^{j+1} - \hat{v}_h^{j+1}\|_\infty &\leq \|v_h^{j+1} - \pi_{\mathcal{W}} v_h^{j+1}\|_\infty + \|\pi_{\mathcal{W}} v_h^{j+1} - \pi_{\mathcal{W}} \hat{v}_h^{j+1}\|_\infty \\ &\leq kL_{j+1} + \beta \|v_h^j - \hat{v}_h^j\|_\infty,\end{aligned}$$

wobei wir zur Abschätzung des zweiten Terms in der letzten Ungleichung die Abschätzung $\|\pi_{\mathcal{W}} q_1 - \pi_{\mathcal{W}} q_2\|_\infty \leq \|q_1 - q_2\|_\infty$ und Lemma 2.4 verwendet haben.
Aus $v_h^0 = \hat{v}_h^0$ erhalten wir nun mittels Induktion die Abschätzung

$$\|v_h^j - \hat{v}_h^j\|_\infty \leq \sum_{i=1}^{j} \beta^{j-i} k L_i \leq \sum_{i=1}^{j} e^{-\delta h(j-i)} k L_i.$$

Für $\delta > L$ gilt $L_j \leq e^{h(\delta - L)} L_g/(\delta - L) =: \widetilde{C}$ und damit können wir die rechte Seite abschätzen durch

$$\sum_{i=1}^{j} e^{-\delta h(j-i)} k \widetilde{C} \leq e^{\delta h} \frac{k}{h} \widetilde{C} \int_0^{jh} e^{-\delta t}\, dt \leq e^{\delta h} \frac{k}{h} \widetilde{C} \int_0^{\infty} e^{-\delta t}\, dt = C \frac{k}{h}.$$

Für $\delta < L$ haben wir $L_j \leq L_g e^{(L-\delta)jh}/(L-\delta)$ und damit

$$\sum_{i=1}^{j} e^{-\delta h(j-i)} k \frac{L_g}{L-\delta} e^{(L-\delta)ih} \leq e^{(L-\delta)h} \frac{k}{h} \frac{L_g}{L-\delta} \int_0^{jh} e^{(L-\delta)t}\, dt = C \frac{k}{h} \int_0^{jh} e^{(L-\delta)t}\, dt.$$

□

Mit diesem Lemma können wir nun den folgenden Satz beweisen.

Satz 3.17 Betrachte ein zeitdiskretes optimales Steuerungsproblem aus Definition 2.1 auf einer kompakten Rechteckmenge Ω mit Zeitschritt h. Sei v_h die zugehörige optimale Wertefunktion. Betrachte weiterhin ein Gitter Γ auf Ω mit Durchmesser k. Dann gilt für die Funktion $\hat{v}_h$ aus Lemma 3.13 die Abschätzung

$$\|v_h - \hat{v}_h\|_\infty \leq K \left(\frac{k}{h}\right)^\gamma,$$

für $\gamma = 1$, falls $\delta > L$, $\gamma \in (0,1)$ beliebig, falls $\delta = L$ und $\gamma = \delta/L$ falls $\delta < L$, und für eine geeignete Konstante $K > 0$.

Beweis: Im Fall $\delta > L$ folgt die Behauptung direkt aus der Abschätzung 3.12 im Lemma 3.16, die für alle $j \geq 0$ und damit auch für $j \to \infty$ gilt.

Im Fall $\delta < L$ erhalten wir aus Lemma 3.16 die Abschätzung

$$\|v_h - \hat{v}_h\|_\infty \leq \int_0^{\infty} e^{-\delta t} \phi(t)\, dt$$

mit

$$\phi(t) = \min\left\{ e^{\delta h} 2M_g,\, C \frac{k}{h} e^{Lt} \right\}.$$

Diese Funktion erfüllt die Voraussetzung von Lemma 2.5 mit $M = e^{\delta h} 2M_g$, $D = 0$ und $C = C\frac{k}{h}$, also folgt

$$\int_0^\infty e^{-\delta t}\Phi(t)\,dt \le K\left(\frac{k}{h}\right)^\gamma$$

für γ aus der Behauptung und ein geeignetes $K > 0$.

Im Falle $\delta = L$ können wir zu gegebenem $\gamma \in (0,1)$ o.B.d.A. $L = \delta/\gamma > \delta$ annehmen, und erhalten die gewünschte Aussage somit aus dem Fall $\delta < L$. □

Kombinieren wir nun Satz 3.17 mit Satz 3.2 so erhalten wir die folgende Aussage.

Satz 3.18 Betrachte ein optimales Steuerungsproblem aus Definition 2.1 auf einer kompakten Rechteckmenge Ω mit optimaler Wertefunktion v, das zugehörige zeitdiskrete optimale Steuerungsproblem aus Definition 3.1 zu einem $h > 0$ sowie ein Gitter Γ auf Ω mit Durchmesser k. Dann gilt für die Funktion $\hat{v}_h$ aus Lemma 3.13 die Abschätzung

$$\|v - \hat{v}_h\|_\infty \le Kh^{\gamma/2} + K\left(\frac{k}{h}\right)^\gamma,$$

für $\gamma = 1$, falls $\delta > L$, $\gamma \in (0,1)$ beliebig, falls $\delta = L$ und $\gamma = \delta/L$ falls $\delta < L$, und für eine geeignete Konstante $K > 0$. Falls das optimale Steuerungsproblem konvex ist, erhalten wir sogar

$$\|v - \hat{v}_h\|_\infty \le Kh^{\gamma} + K\left(\frac{k}{h}\right)^\gamma.$$

Aus dem Satz ergibt sich die Forderung, dass wir, um Konvergenz von $\hat{v}_h$ gegen v zu erhalten, h und k so gegen Null streben lassen müssen, dass die Bedingung $k/h \to 0$ ebenfalls erfüllt ist. Praktische Tests zeigen aber, dass man auch im Fall $k \approx h$ gute Ergebnisse erhält. Der Grund dafür ist, dass tatsächlich eine stärkere Abschätzung gilt, die von M. Falcone und T. Giorgi in dem Artikel [5] bewiesen wurde. Diese lautet

$$\|v - \hat{v}_h\|_\infty \le Kh^{\gamma/2} + K\left(\frac{k}{\sqrt{h}}\right)^\gamma$$

für alle hinreichend kleinen k, $h > 0$ mit der Eigenschaft, dass $k \le C_1 h$ für eine Konstante $C_1 > 0$ gilt. Für $k = C_1 h$ folgt also insbesondere die Abschätzung $\|v - \hat{v}_h\|_\infty \le \widetilde{K} h^{\gamma/2}$ für ein geeignetes $\widetilde{K} > 0$.

Der Beweis ist ziemlich kompliziert und verwendet explizit, dass die Funktion v die Viskositätslösung der Hamilton–Jacobi–Bellman Gleichung aus Satz 2.9 ist. Wir werden deshalb nicht näher auf ihn eingehen.

Kapitel 4

Numerik des optimalen Steuerungsproblems

Ausgehend von der im letzten Kapitel betrachteten Diskretisierung wollen wir uns in diesem Kapitel mit weiter gehenden Aspekten der numerischen Lösung beschäftigen. Wir betrachten dabei insbesondere effiziente Strategien zur iterativen Berechnung der optimalen Wertefunktion, Methoden der adaptiven Diskretisierung und ein Verfahren zur Berechnung approximativ optimaler Trajektorien, mit dem wir beginnen wollen.

4.1 Berechnung approximativ optimaler Trajektorien

In diesem Abschnitt wollen wir zeigen, wie wir aus der approximativen Wertefunktion approximativ optimale Trajektorien berechnen können. Wir werden zunächst das zeitdiskrete Problem betrachten (unter der Annahme, dass wir v_h kennen), dann den durch die Approximation $\hat{v}_h$ entstehenden Fehler analysieren, und schließlich — im Falle einer vorhergehenden zeitlichen Diskretisierung — zum kontinuierlichen Problem übergehen.

4.1.1 Zeitdiskrete optimale Trajektorien

Wie erinnern uns an das Optimalitätsprinzip für v_h, das für $k = 0$ als

$$v_h(x) = \max_{u \in U} \{hg(x,u) + \beta v_h(f_h(x,u))\} \tag{4.1}$$

geschrieben werden kann. Die folgende Definition basiert auf diesem Prinzip

Definition 4.1 Wir definieren eine Abbildung $u^* : \mathbb{R}^d \to U$, indem wir zu jedem $x \in \mathbb{R}^d$ ein $u_x \in U$ wählen, so dass das Maximum in (4.1) angenommen wird (dieses u_x existiert, wird aber im Allgemeinen nicht eindeutig sein), und $u^*(x) = u_x$ setzen. □

Bemerkung 4.2 Dies Vorschrift definiert eine Kontrollstrategie, die vom aktuellen Zustand x und nicht von der Zeit t abhängt, also keine diskrete Kontrollfunktion im bisher

bekannten Sinne ist. Eine solche Kontrollstrategie nennt man *Zustandsrückführung* oder *Zustandsfeedback*. □

Mittels u^* definieren wir nun zu jedem Anfangswert x eine diskrete Kontrollfunktion $u_h^x \in \mathcal{U}_h$ gemäß der folgenden iterativen Vorschrift:

$$u_h^x(0) := u^*(x), \qquad u_h^x(ih) = u^*(\Phi_h(ih, x, u_h^x)) \quad \text{für} \quad i = 1, 2, \ldots. \tag{4.2}$$

Beachte, dass $u_h^x(ih)$ wohldefiniert ist, da die Lösung $\Phi_h(ih, x, u_h^x)$ nur von den Werten $(u_h^x(0), u_h^x(h), \ldots, u_h^x((i-1)h))$ abhängt, die für ein gegebenes $i \in \mathbb{N}$ bereits iterativ definiert sind.

Der folgende Satz zeigt, dass die so definierten Kontrollen tatsächlich optimale Kontrollen für das zeitdiskrete Problem sind.

Satz 4.3 Die in (4.2) definierte diskrete Kontrollfunktion ist optimal für das zeitdiskrete optimale Steuerungsproblem, d.h. für alle $x \in \mathbb{R}^d$ gilt

$$J_h(x, u_h^x) = v_h(x).$$

Beweis: Wir wählen ein $x \in \mathbb{R}^d$, schreiben kurz $x_j = \Phi_h(jh, x, u_h^x)$ und $u_j = u_h^x(jh)$, und zeigen per Induktion für alle $k \geq 0$ die Gleichung

$$v_h(x) = h \sum_{i=0}^{k} \beta^i g(x_i, u_i) + \beta^{k+1} v_h(x_{k+1}). \tag{4.3}$$

Hieraus folgt die Behauptung für $k \to \infty$, denn da v_h beschränkt ist gilt

$$\lim_{k\to\infty} \left(h \sum_{i=0}^{k} \beta^i g(x_i, u_i) + \beta^{k+1} v_h(x_{k+1}) \right) = h \sum_{i=0}^{\infty} \beta^i g(x_i, u_i) = J_h(x, u_h^x).$$

Zum Beweis von (4.3) verwenden wir, dass aus der Definition von $u^*(x)$ und u_h^x für alle $j \geq 0$ folgt

$$\begin{aligned} v_h(x_j) &= hg(x_j, u^*(x_j)) + \beta v_h(f_h(x_j, u^*(x_j)) \\ &= hg(x_j, u_j) + \beta v_h(f_h(x_j, u_j)) \quad = \quad hg(x_j, u_j) + \beta v_h(x_{j+1}). \end{aligned} \tag{4.4}$$

Für $k = 0$ folgt (4.3) nun direkt aus (4.4) mit $j = 0$. Gelte also (4.3) für ein $k \geq 0$. Dann folgt mit (4.4) für $j = k + 1$

$$\begin{aligned} v_h(x) &= h \sum_{i=0}^{k} \beta^i g(x_i, u_i) + \beta^{k+1} v_h(x_{k+1}) \\ &= h \sum_{i=0}^{k} \beta^i g(x_i, u_i) + \beta^{k+1} \Big(hg(x_{k+1}, u_{k+1}) + \beta v_h(x_{k+2}) \Big) \\ &= h \sum_{i=0}^{k+1} \beta^i g(x_i, u_i) + \beta^{k+2} v_h(x_{k+2}), \end{aligned}$$

womit (4.3) für $k + 1$ gezeigt ist und damit für alle $k \geq 0$ gilt. □

4.1.2 Numerische Berechnung approximativ optimaler Kontrollen

Analog zu Definition 4.1 definieren wir nun eine numerische Kontrollstrategie $\hat{u}^*$. Betrachte dazu den Ausdruck

$$hg(x,u) + \beta \hat{v}_h(f_h(x,u)) \tag{4.5}$$

für die Funktion $\hat{v}_h$ aus Lemma 3.13.

Definition 4.4 Wir definieren eine Abbildung $\hat{u}^* : \Omega \to U$, indem wir zu jedem $x \in \Omega$ ein $u_x \in U$ wählen, so dass das Maximum in (4.5) angenommen wird (dieses u_x existiert, wird aber im Allgemeinen wiederum nicht eindeutig sein), und $\hat{u}^*(x) = u_x$ setzen. □

Analog zu (4.2) definieren wir nun zu jedem Anfangswert x eine diskrete Kontrollfunktion $\hat{u}_h^x \in U^{\mathbb{N}_0}$ gemäß der folgenden iterativen Vorschrift:

$$\hat{u}_h^x(0) := \hat{u}^*(x), \qquad \hat{u}_h^x(ih) = \tilde{u}^*(\Phi_h(hi, x, \hat{u}_h^x)) \quad \text{für} \quad i = 1, 2, \ldots. \tag{4.6}$$

Satz 4.5 Die in (4.6) definierte diskrete Kontrollfunktion u_h^x ist approximativ optimal für das zeitdiskrete optimale Steuerungsproblem. Genauer gilt für alle $x \in \mathbb{R}^d$ die Abschätzung

$$|J_h(x, \hat{u}_h^x) - v_h(x)| \le C \frac{k^\gamma}{h^{\gamma+1}}$$

für $\gamma \in [0,1]$ aus Satz 3.17 und eine geeignete Konstante $C > 0$.

Beweis: Der Beweis verläuft analog zu dem von Satz 4.3. Wir schreiben kurz $x_k = \Phi_h(ih, x, \hat{u}_h^x)$, $u_i = \hat{u}_h^x(ih)$, und beweisen

$$v_h(x) = h \sum_{i=0}^{l} \beta^i g(x_i, u_i) + \beta^{l+1} v_h(x_{l+1}) + 2K_l \sum_{i=0}^{k} \beta^{i+1} \frac{k^\gamma}{h^\gamma}. \tag{4.7}$$

für Konstanten $K_l > 0$ mit $K_l \le K$ für ein von l unabhängiges > 0. Hieraus folgt die Behauptung analog zum Beweis von Satz 4.3, denn es gilt

$$2K \sum_{i=0}^{k} \beta^{i+1} \le C \frac{1}{h}$$

für ein geeignetes $C > 0$.

Zum Beweis von (4.7) beachte, dass aus Satz 3.17, Gleichung (4.1) und der Definition von $\hat{u}^*$ und u_i für jedes $i \ge 0$ folgt

$$\begin{aligned} v_h(x_i) &= \max_{u \in U} \{hg(x_i,u) + \beta v_h(f_h(x_i,u))\} \\ &= \max_{u \in U} \{hg(x_i,u) + \beta \hat{v}_h(f_h(x_i,u))\} + R_1(x_i) \\ &= hg(x_i,u_i) + \beta \hat{v}_h(f_h(x_i,u_i)) + R_1(x_i) \\ &= hg(x_i,u_i) + \beta v_h(f_h(y, \hat{u}^*(y))) + R_1(x_i) + R_2(x_i) \end{aligned}$$

mit

$$|R_m(x_i)| \le \beta \|v_h - \hat{v}_h\|_\infty \le \beta K \frac{k^\gamma}{h^\gamma}$$

für $m = 1, 2$ und alle $i \ge 0$.

Nun folgt (4.7) ganz analog zum Beweis von (4.3) im Beweis von Satz 4.3, wobei sich die R_m–Terme zu dem angegebenen Fehlerterm aufsummieren. □

4.1.3 Das zeitkontinuierliche Problem

Für den Fall, dass das betrachtete zeitdiskrete Problem eine Diskretisierung eines kontinuierlichen Problems darstellt, wollen wir nun abschließend zeigen, wie wir aus der approximativ optimalen diskreten Kontrollfunktion $\hat{u}_h^x$ eine messbare Kontrollfunktion $\hat{u}^x$ konstruieren können, die auch für das ursprüngliche zeitkontinuierliche Problem approximativ optimal ist. Hierzu setzen wir

$$\hat{u}^x(t) := \hat{u}_h^x(hi), \quad t \in [ih, (i+1)h). \tag{4.8}$$

Satz 4.6 Die in (4.8) definierte stückweise konstante kontinuierliche Kontrollfunktion $\hat{u}^x \in \mathcal{U}$ ist approximativ optimal für das optimale Steuerungsproblem aus Definition 2.1. Genauer gilt für alle $x \in \mathbb{R}^d$ die Abschätzung

$$|J(x, \hat{u}^x) - v(x)| \leq C \left(h^{\gamma/2} + \frac{k^\gamma}{h^{\gamma+1}} \right)$$

für $\gamma \in [0, 1]$ aus Satz 3.17 und eine geeignete Konstante $C > 0$. Ist das optimale Steuerungsproblem konvex, so gilt

$$|J(x, \hat{u}^x) - v(x)| \leq C \left(h^\gamma + \frac{k^\gamma}{h^{\gamma+1}} \right).$$

Beweis: Beachte, dass die stückweise konstante Kontrollfunktion $\hat{u}^x$ aus (4.8) die Bedingungen (3.2) und (3.3) aus dem Beweis von Satz 3.2 erfüllt. Also folgt aus (3.4) die Abschätzung

$$|\widetilde{J}_h(x, \hat{u}_h^x) - J(x, \hat{u}^x)| \leq K_1 h^\gamma \tag{4.9}$$

für eine passende Konstante $K_1 > 0$. Die Behauptung folgt nun mit Dreiecksungleichung aus (4.9) und den Sätzen 4.5 und 3.2. □

4.2 Alternative Iterationsverfahren

In den Übungen haben wir festgestellt, dass die Iterationsvorschrift aus Definition 3.11 zur Berechnung des Vektors V insbesondere für kleine $\delta > 0$ und $h > 0$ recht langsam konvergiert. In diesem Abschnitt wollen wir zwei Methoden besprechen, mit denen die Berechnung von V schneller durchgeführt werden kann.

4.2.1 Das kontrollierte Gauß–Seidel–Verfahren

Wir erinnern zunächst an die bekannte Iteration zur Berechnung von V, die in etwas anderer Notation lautet:

$$V^0 := (0, \ldots, 0)^T; \quad V^{j+1} := V^j, \quad V_i^{j+1} := S(V^{j+1})_i, \;\; i = 0, \ldots, N-1, \;\; j = 0, 1, \ldots \tag{4.10}$$

mit

$$S(W)_i = \max_{u\in U}\left\{G(i,u) + \beta\sum_{k=0}^{N-1} B(i,u)_k W_k\right\}$$

für $W \in \mathbb{R}^N$. Dieses Verfahren wird in der Literatur oft als *sukzessive Approximation* bezeichnet. Beachte, dass die „:=“ in (4.10) Zuweisungen sind; insbesondere geht in der letzten Zuweisung in (4.10) der Wert V_i^{j+1} auch auf der rechten Seite ein.

Eine Idee für eine alternative Iteration liegt nun darin, die letzte Zuweisung in (4.10) als Gleichung aufzufassen, d.h. einen Wert V_i^{j+1} zu bestimmen, so dass

$$V_i^{j+1} = S(V^{j+1})_i \tag{4.11}$$

erfüllt ist. Die Gleichung (4.11) ist explizit lösbar, denn es gilt

$$\begin{aligned}
V_i^{j+1} = S(V^{j+1})_i = \max_{u\in U}\left\{G(i,u) + \beta\sum_{k=0}^{N-1} B(i,u)_k V_k^{j+1}\right\}
\end{aligned}$$

$$\iff \quad \begin{cases} \forall u\in U: \;\; V_i^{j+1} \geq G(i,u) + \beta\sum_{k=0}^{N-1} B(i,u)_k V_k^{j+1} \\ \exists u\in U: \;\; V_i^{j+1} = G(i,u) + \beta\sum_{k=0}^{N-1} B(i,u)_k V_k^{j+1} \end{cases}$$

$$\iff \quad \begin{cases} \forall u\in U: \;\; V_i^{j+1} \geq G(i,u) + \beta\sum_{\substack{k=0\\k\neq i}}^{N-1} B(i,u)_k V_k^{j+1} + \beta B(i,u)_i V_i^{j+1} \\ \exists u\in U: \;\; V_i^{j+1} = G(i,u) + \beta\sum_{\substack{k=0\\k\neq i}}^{N-1} B(i,u)_k V_k^{j+1} + \beta B(i,u)_i V_i^{j+1} \end{cases}$$

$$\iff \quad \begin{cases} \forall u\in U: \;\; V_i^{j+1} \geq \frac{G(i,u) + \beta\sum_{\substack{k=0\\k\neq i}}^{N-1} B(i,u)_k V_k^{j+1}}{1-\beta B(i,u)_i} \\ \exists u\in U: \;\; V_i^{j+1} = \frac{G(i,u) + \beta\sum_{\substack{k=0\\k\neq i}}^{N-1} B(i,u)_k V_k^{j+1}}{1-\beta B(i,u)_i} \end{cases}$$

$$\iff \quad V_i^{j+1} = \max_{u\in U}\left\{\frac{G(i,u) + \beta\sum_{\substack{k=0\\k\neq i}}^{N-1} B(i,u)_k V_k^{j+1}}{1-\beta B(i,u)_i}\right\}$$

Wir definieren also das sogenannte *kontrollierte Gauß–Seidel–Verfahren* analog zur obigen Iteration durch

$$V^0 := (0,\ldots,0)^T; \quad V^{j+1} := V^j, \quad V_i^{j+1} := \widetilde{S}(V^{j+1})_i, \;\; i = 0,\ldots,N-1, \;\; j = 0,1,\ldots \tag{4.12}$$

mit

$$\widetilde{S}(W)_i = \max_{u\in U}\left\{\frac{G(i,u) + \beta\sum_{\substack{k=0\\k\neq i}}^{N-1} B(i,u)_k W_k}{1-\beta B(i,u)_i}\right\}$$

für $W \in \mathbb{R}^N$.

Um die Konvergenz des Verfahrens zu analysieren, betrachten wir zu $W \in \mathbb{R}^N$ und $i \in \{0, \dots, N-1\}$ den Vektor $\widetilde{W}^i = (W_0, \dots, W_{i-1}, \widetilde{S}(W)_i, W_{i+1}, \dots, W_{N-1})^T$. Damit gilt

$$S(\widetilde{W}^i)_i = [\widetilde{W}^i]_i = \widetilde{S}(W)_i$$

und daher für je zwei Vektoren W_1, $W_2 \in \mathbb{R}^N$

$$\begin{aligned} |\widetilde{S}(W_1)_i - \widetilde{S}(W_2)_i| &= |S(\widetilde{W}_1^i)_i - S(\widetilde{W}_2^i)_i| \\ &\leq \beta \|\widetilde{W}_1^i - \widetilde{W}_2^i\|_\infty = \beta \max_{j=0,\dots,N-1} |[\widetilde{W}_1^i - \widetilde{W}_2^i]_j| \end{aligned}$$

Wird das Maximum nun in $j = i$ angenommen, so folgt

$$|\widetilde{S}(W_1)_i - \widetilde{S}(W_2)_i| \leq \beta |[\widetilde{W}_1^i - \widetilde{W}_2^i]_i| = \beta |\widetilde{S}(W_1)_i - \widetilde{S}(W_2)_i|$$

und damit wegen $\beta < 1$

$$|\widetilde{S}(W_1)_i - \widetilde{S}(W_2)_i| = 0.$$

Andernfalls erhalten wir

$$|\widetilde{S}(W_1)_i - \widetilde{S}(W_2)_i| \leq \beta |[\widetilde{W}_1^i - \widetilde{W}_2^i]_j| \leq \beta \|W_1 - W_2\|_\infty.$$

Tatsächlich gilt also in beiden Fällen

$$|\widetilde{S}(W_1)_i - \widetilde{S}(W_2)_i| \leq \beta \|W_1 - W_2\|_\infty,$$

weswegen $\widetilde{S}$ eine Kontraktion mit Rate $\beta < 1$ ist und die Iteration wegen des Banach'schen Fixpunktsatzes (linear) gegen den Fixpunkt $V \in \mathbb{R}^n$ konvergiert. Insbesondere gilt also das Abbruchkriterium aus Lemma 3.13 auch für die Iteration (4.12).

Der Name "Gauß–Seidel" ergibt sich aus der folgenden Beobachtung: Falls U eine einpunktige Menge $U = \{u_0\}$ ist, so fällt die Maximierung weg und der Operator S lässt sich als $S(V) = \widetilde{A}V + b$ schreiben. Die Lösung der Fixpunktgleichung $V = S(V)$ ist dann gegeben durch $S(V) - V = 0$, also durch $(\widetilde{A} - \mathrm{Id})V = -b$, was gerade ein lineares Gleichungssystem im $\mathbb{R}^N$ ist. In diesem Fall liefert die Iteration (4.12) gerade das klassische Gauß–Seidel Verfahren zur Lösung linearer Gleichungssysteme.

Dieses Verfahren hat aber auch eine interessante geometrische Interpretation: Wenn wir (zur leichteren Interpretation) annehmen, dass $G(i, u) \geq 0$ gilt für alle i und u, so sieht man leicht, dass sowohl die Iteration mit S als auch diejenige mit $\widetilde{S}$ monoton wachsend sind, d.h. es gilt $V_i^{j+1} \geq V_i^j$. In diesem Fall liegen die Vektoren V^j in einem Polyeder im $\mathbb{R}^N$, dessen Spitze gerade durch V gegeben ist. Abbildung 4.1 zeigt eine schematische Darstellung der beiden Iterationen.

Die Tatsache, dass man in jeder Koordinate jeweils bis zum Rand des Polyeders „aufsteigt" ist der Grund, dass das Verfahren auch *Koordinatenaufstiegsverfahren* genannt wird und unter diesem Namen in [9] eingeführt wurde.

Offenbar ist der Beschleunigungseffekt am größten, wenn der Nenner in der Definition von $\widetilde{S}_i$ klein wird.[1] Dies wiederum ist gerade dann der Fall, wenn $B(i, u)_i$ für das maximierende u^* groß ist. Dieser Fall tritt ein, falls $\Phi_h(E_i, u^*) \approx E_i$ gilt, d.h., falls es ein optimales

[1] im ungünstigsten Fall ist er gleich 1; dann stimmen S_i und $\widetilde{S}_i$ überein

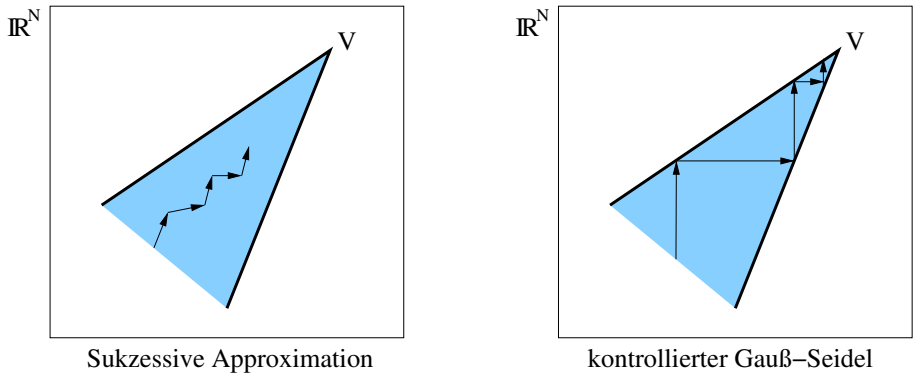

Abbildung 4.1: Iterationsverfahren

Gleichgewicht in der Nähe des Eckpunktes E_i gibt.[2] Tatsächlich lässt sich numerisch beobachten, dass das Koordinatenaufstiegsverfahren besonders effizient ist, wenn das System ein optimales Gleichgewicht besitzt. Sind die optimalen Trajektorien hingegen periodisch, ist der Zeitgewinn weniger groß, im Allgemeinen aber immer noch vorhanden. Beispiele, in denen (4.12) langsamer konvergiert als (4.10) sind nicht bekannt.

4.2.2 Strategie–Iteration

Beide bisher betrachteten Iterationen beruhen auf der Tatsache, dass die Iterationsoperatoren eine Kontraktion mit Kontraktionsrate β sind. Man kann also i.A. hier eine lineare Konvergenzordnung gegen den Vektor V erwarten.

Eine weitere Iterationsvariante, die schneller konvergiert, ist bereits seit den 60er Jahren bekannt. Die sogenannte *Strategie–Iteration* (engl.: policy iteration) nutzt die folgende Tatsache aus:

Wenn wir einen Vektor $\tilde{u} = (\tilde{u}_0, \ldots, \tilde{u}_{N-1})^T \in U^N$ wählen und statt des maximierenden $u \in U$ in $S(V_i^{j+1})$ den Kontrollwert $\tilde{u}_i$ einsetzen, so konvergiert die Iteration gegen einen Vektor $V^{\tilde{u}}$, der durch das Gleichungssystem

$$V_i^{\tilde{u}} = G(i, \tilde{u}_i) + \beta B(i, \tilde{u}_i) V^{\tilde{u}}, \quad i = 0, \ldots, N-1 \tag{4.13}$$

eindeutig bestimmt ist. Man sieht leicht, dass $V_i^{\tilde{u}} \leq V_i$ für alle $i = 0, \ldots, N-1$ gilt. Da die Maximierung hier wegfällt, lässt sich der Vektor V^{u^*} viel schneller als der Vektor V berechnen. Zur Lösung dieses linearen Gleichungssystems kann man z.B. das (klassische) Gauß–Seidel–Verfahren nehmen, das gerade durch die Iteration (4.12) ohne Maximierung gegeben ist, (4.13) kann also mittels der Iteration

$$V^{\tilde{u},j+1} := V^{\tilde{u},j}, \quad V_i^{\tilde{u},j+1} := \widetilde{S}(V^{j+1}, \tilde{u})_i, \quad i = 0, \ldots, N-1, \quad j = 0, 1, \ldots \tag{4.14}$$

mit geeignetem Startvektor $V^{\tilde{u},0}$ und

$$\widetilde{S}(W, \tilde{u})_i = \frac{G(i, \tilde{u}_i) + \beta \sum_{\substack{k=0 \\ k \neq i}}^{N-1} B(i, \tilde{u}_i)_k W_k}{1 - \beta B(i, \tilde{u}_i)_i}$$

[2] dies ist natürlich keine präzise mathematische Aussage sondern eine heuristische Beobachtung

für $W \in \mathbb{R}^N$, $\tilde{u} \in \mathbb{R}^N$ gelöst werden.

Alternativ kann man (4.13) lösen, indem man ausnutzt, dass sich aus (4.13) das schwach besetzte lineare Gleichungssystem

$$AV^{\tilde{u}} = b, \qquad A = \mathrm{Id}_{\mathbb{R}^N} - \beta \begin{pmatrix} B(0,\tilde{u}_1) \\ \vdots \\ B(N-1,\tilde{u}_1) \end{pmatrix}, \quad b = \begin{pmatrix} G(0,\tilde{u}_1) \\ \vdots \\ G(N-1,\tilde{u}_1) \end{pmatrix}$$

herleiten lässt, für dessen Lösung es weitere iterative numerische Verfahren gibt, z.B. das präkonditionierte CGS– oder BiCGStab–Verfahren.

Die Idee der Strategie–Iteration liegt nun darin, zu gegebenem V^j einen Kontrollvektor $\tilde{u}^j$ so zu wählen, dass dessen Komponenten gerade die maximierenden Kontrollen für die Iterationsvorschrift $S(V^j)_i$ sind, und damit $V^{j+1} = V^{\tilde{u}^j}$ zu berechnen. Formal lässt sich dieses Verfahren wie folgt beschreiben.

(1) Setze $V^0 := (0,\dots,0)^T \in \mathbb{R}^N$, $j = 0$

(2) Wähle $\tilde{u}^j \in U^N$ mit $\widetilde{S}(V^j)_i = G(i,\tilde{u}_i^j) + \beta\ B(i,\tilde{u}_i^j)V^j$ für $i = 0,\dots,N-1$

(3) Berechne $V^{j+1} = V^{\tilde{u}^j}$

(4) Falls $\|V^j - V^{j+1}\|_\infty > \varepsilon$ setze $j := j+1$ und gehe zu (2)

Schritt (3) kann hierbei z.B. mittels (4.14) gelöst werden, wobei der Anfangsvektor als $V^{\tilde{u},0} = V^j$ gewählt wird.

In der Arbeit [17] von M. L. Puterman and S. Brumelle wurde gezeigt, dass die Vektoren V^j lokal quadratisch gegen V konvergieren; allerdings muss dabei die Zeit zur Berechnung von $V^{\tilde{u}^j}$ berücksichtigt werden, die i.A. deutlich länger ist als die Durchführung eines Iterationsschritts in (4.10).

In der Praxis zeigt sich, dass dieses Verfahren am Anfang recht langsam konvergiert, oft deutlich langsamer als die Iteration (4.12). Es liegt daher nahe, die beiden Verfahren zu kombinieren. Die führt zu dem folgenden Verfahren.

(1) Setze $V^0 := (0,\dots,0)^T \in \mathbb{R}^N$, $k = 0$

(2) (a) Setze $V^{k+1} := V^k$ und berechne $V_i^{k+1} := \widetilde{S}(V^{k+1})_i$, $i = 0,\dots,N-1$;
speichere dabei in $\tilde{u}^k \in U^N$ die maximierenden Kontrollwerte

(b) Falls $\|V^{k+1} - V^k\|_\infty \le \varepsilon$ beende den Algorithmus;
Falls ein Abbruchkriterium (s.u.) erfüllt ist gehe zu (3);
ansonsten setze $k := k+1$ und gehe zu (2a)

(3) Berechne $V^{k+1} = V^{\tilde{u}^k}$, setze $k := k+1$ und gehe zu (2)

Die Frage nach einem guten Abbruchkriterium in (2b) kann nur experimentell beantwortet werden und hängt stark vom zugrundeliegenden optimalen Steuerungsproblem ab. In jedem

Fall zeigen Experimente aber, dass es günstiger ist, das Verhalten der Strategien $\tilde{u}^k$ als das der Vektoren V^k zu betrachten.

In der Diplomarbeit [18] von A. Seeck wird vorgeschlagen zu prüfen, wie viele Einträge von $\tilde{u}^k$ und $\tilde{u}^{k-1}$ übereinstimmen. Falls die Anzahl der übereinstimmenden Einträge größer als eine bestimmte Prozentzahl ist, wird zu (3) übergegangen. Hier haben sich Werte zwischen 80 und 100 Prozent als geeignet erwiesen. In der Arbeit [7] von R. L. V. González und C. A. Sagastizábal hingegen wird ein Wert $q \in \mathbb{N}$ festgelegt und erst dann zu (3) übergegangen, wenn die $q+1$ aufeinanderfolgenden Kontrollvektoren $\tilde{u}^{k-q}, \ldots, \tilde{u}^k$ exakt übereinstimmen. Leider werden in dieser Arbeit keine Vorschläge für eine gute Wahl von q gemacht. Praktische Erfahrungen zeigen, dass Werte zwischen $q = 1$ und $q = 5$ recht gute Ergebnisse zeigen, der Wert $q = 1$ entspricht dabei dem 100 Prozent Kriterium aus der Arbeit von Seeck.

Beide Kriterien funktionieren nur, wenn U eine endliche Menge ist und das Maximum durch Vergleich bestimmt wird. Ein weiteres Kriterium, das auch für unendliche Mengen funktioniert (d.h., wenn das Maximum mit einem kontinuierlichen Optimierungsverfahren bestimmt wurde), liegt darin, einen Wert $\alpha > 0$ vorzugeben, und zu Schritt (3) überzugehen, wenn

$$\|\tilde{u}^k - \tilde{u}^{k-1}\|_1 \leq N\alpha$$

gilt (hier könnte man natürlich auch die 2–Norm nehmen, die 1–Norm ist aber schneller zu berechnen). Hier entspricht der Wert $\alpha = 0$ dem 100%–Kriterium von Seeck.

4.3 Kontinuierliche Optimierung

Die bisherige Vorgehensweise in der Maximierung über U, nämlich die Diskretisierung der Menge und die Bestimmung des Maximums durch Vergleiche, stößt rasch an ihre Grenzen, wenn eine höhere numerische Genauigkeit erwünscht ist, da die Berechnung der Werte für viele diskrete Werte $u \in U$ dann sehr lange dauert.

In diesem kurzen Abschnitt wollen wir ein einfaches heuristisches Verfahren zur Berechnung des Maximums über U in den Iterationen kennen lernen, das ohne Diskretisierung der Menge U auskommt. Sicherlich gibt es eine ganze Reihe von besseren und mathematisch fundierteren Verfahren, deren Behandlung allerdings den Rahmen dieser Vorlesung sprengen würde.

Ziel ist es, für jedes $i = 0, \ldots, N-1$ den Ausdruck

$$G(i,u) + \beta \sum_{k=0}^{N-1} B(i,u)_k V_k =: F(u)$$

bzw. im Falle der Gauß–Seidel Iteration den Ausdruck

$$\frac{G(i,u) + \beta \sum_{\substack{k=0 \\ k \neq i}}^{N-1} B(i,u)_k V_k}{1 - \beta B(i,u)_i} =: F(u)$$

über $u \in U$ zu maximieren.

Wir betrachten das folgende Verfahren, das man als *rekursive Suche* bezeichnen kann. Wir nehmen dabei an, dass $U = [u^-, u^+]$ ein kompaktes Intervall ist.

(0) Wähle ganze Zahlen $n_i \geq 3$ für $i = 0, 1, \ldots$; setze $i := 0$, $u_0^- = u^-$, $u_0^+ = u^+$.

(1) Setze $\Delta_i = (u_i^+ - u_i^-)/n_i$, $U_i = \{u_i^-, u_i^- + \Delta_i, u_i^- + 2\Delta_i, \ldots, u_i^- + n_i \Delta i\}$ und bestimme $\max_{u \in U_i} F(u)$ sowie ein $u_i^* \in U_i$ mit $F(u_i^*) = \max_{u \in U_i} F(u)$.

(2) Falls eine vorgegebene Anzahl von Iterationen erreicht ist, beende den Algorithmus und gebe u_i^* als approximative Maximalstelle aus

(3) Falls $u_i^* = u_i^-$, setze, $u_{i+1}^- := u_i^*$, $u_{i+1}^+ := u_i^* + 2\Delta_i$
Falls $u_i^* = u_i^+$, setze, $u_{i+1}^- := u_i^* - 2\Delta_i$, $u_{i+1}^+ := u_i^*$
sonst, setze $u_{i+1}^- := u_i^* - \Delta_i$, $u_{i+1}^+ := u_i^* + \Delta_i$.
Setze $i := i + 1$ und gehe zu (1)

Obwohl diese rekursive Suche im Allgemeinen kein rigoroses Konvergenzresultat zulässt, lässt sich in einem Spezialfall doch ein entsprechender Satz beweisen.

Definition 4.7 Sei $U \subset \mathbb{R}$ ein kompaktes Intervall. Eine Funktion $F : U \to \mathbb{R}$ heißt *unimodal*, falls ein $u^* \in U$ existiert, so dass $F(u^*) = \sup_{u \in U} F(u)$ ist und für alle $u_1, u_2 \in U$ die Implikationen

$$u_1 < u_2 < u^* \quad \Rightarrow \quad F(u_1) < F(u_2) < F(u_2)$$

und

$$u_1 > u_2 > u^* \quad \Rightarrow \quad F(u_1) < F(u_2) < F(u_2)$$

gelten. □

Satz 4.8 Falls die Funktion F *unimodal* ist, so gilt

$$u^* \in [u_i^-, u_i^+]$$

für alle $i = 0, 1, \ldots$, d.h. die Maximalstelle wird durch den Algorithmus eingeschachtelt. Darüberhinaus gilt die Abschätzung

$$|u^* - u_i^*| \leq 2^i \frac{u_0^+ - u_0^-}{n_0 \cdots n_i}$$

für alle $i \geq 0$.

Beweis: Übungsaufgabe □

Im tatsächlichen Algorithmus wird unsere Funktion F diese Bedingung nur in Ausnahmefällen erfüllen. Wählt man aber die Anfangszerlegung n_0 fein genug, so ist es nicht unwahrscheinlich, dass die Einschränkung der Funktion auf die Menge $[u_1^-, u_1^+]$ unimodal ist, was die im Allgemeinen recht guten Eigenschaften des Algorithmus erklärt, falls n_0 hinreichend groß ist.

Bemerkung 4.9 (i) Wenn die Funktion F Lipschitz stetig un u ist, so folgt aus $|u_i^* - u^*| \leq \varepsilon$ die Abschätzung $|F(u_i^*) - F(u^*)| \leq L\varepsilon$, also wird auch das Maximum selbst gut approximiert.

(ii) Wir wollen den Aufwand dieses Verfahrens im Vergleich mit der äquidistanten Diskretisierung von U abschätzen. Nehmen wir an, es sei $U = [0, 1]$ und wir wollen die Maximalstelle mit einer Genauigkeit von $\varepsilon = 10^{-3}$ ermitteln. Mit einer äquidistanten Diskretisierung von U benötigen wir dazu $10^3/2 = 500$ Teilintervalle, also 501 Punkte und damit ebenso viele Auswertungen von F. Mit der rekursiven Suche (unter der Annahme, dass diese konvergiert) mit $n_0 = 10$ und $n_i = 4$ für $i \geq 1$ gilt für den Fehler die Abschätzung

$$2^i \frac{u_0^+ - u_0^-}{n_0 \cdots n_i} = \frac{1}{10 \cdot 2^i},$$

so dass der gewünschte Fehler wegen $1/(10 \cdot 2^7) \approx 0.78 \cdot 10^{-3}$ für $i = 7$ erreicht wird. Die dafür nötige Anzahl von Auswertungen von F beträgt

$$11 + 5i = 46,$$

wir kommen also bereits bei dieser relativ bescheidenen Genauigkeit mit weniger als einem Zehntel der F–Auswertungen aus. □

4.4 Fehlerschätzung

In diesem abschließenden Abschnitt wollen wir der folgenden Frage nachgehen: Lässt sich der vollständig diskretisierten Lösung $\hat{v}_h$ „ansehen“, wie groß die Differenz $\|v_h - \hat{v}_h\|_\infty$ ist, ohne dass wir v_h kennen? Bisher haben wir in Satz 3.17 eine Abschätzung *a–priori*, also ohne Kenntnis von $\hat{v}_h$ allein aus den Daten des Problems gewonnen. Nun wollen wir den Fehler *a–posteriori*, d.h. unter Einbeziehung von $\hat{v}_h$ abschätzen.

4.4.1 Fehlerschätzer

Natürlich ist es nicht möglich, die Differenz $\|v_h - \hat{v}_h\|_\infty$ ohne Kenntnis von v_h exakt anzugeben; es gibt aber die Möglichkeit, den Fehler über eine geeignete Größe abzuschätzen. Formal macht man dies gemäß der folgenden Definition.

Definition 4.10 Betrachte das vollständig diskrete optimale Steuerungsproblem auf einem Rechteckgitter Γ mit $P \in \mathbb{N}$ Rechtecken $R_0, \ldots, R_{P-1}$. Ein *lokaler a–posteriori Fehlerschätzer* (in der $\|\cdot\|_\infty$–Norm) ist eine Menge von Werten $\eta_0, \ldots, \eta_{P-1}$ mit den folgenden Eigenschaften.

(i) Der Wert η_i lässt sich aus den Daten des optimalen Steuerungsproblems und aus der Funktion $\hat{v}_h$ in einer Umgebung $\mathcal{N}(R_i)$ berechnen.

(ii) Es gibt Konstanten C_1, $C_2 > 0$ (unabhängig vom Gitter Γ), so dass für den Wert $\eta := \max_{i=0,\ldots,P-1} \eta_i$ die Abschätzungen

$$C_1 \eta \leq \|v_h - \hat{v}_h\|_\infty \leq C_2 \eta$$

gelten. Man sagt auch, dass der Fehlerschätzer *effizient* und *zuverlässig*[3] ist.

[3] Effizient: großer Fehlerschätzer ⇒ großer Fehler (kein Überschätzen)
Zuverlässig: kleiner Fehlerschätzer ⇒ kleiner Fehler (kein Unterschätzen)

Gilt darüberhinaus die Abschätzung

$$C_1\eta_i \leq \sup_{x\in\mathcal{N}(R_i)} |v_h(x) - v^\infty_{h,\Gamma}(x)|$$

so heißt der Fehlerschätzer *lokal effizient*. □

Analog zur lokalen Effizienz lässt sich die lokale Zuverlässigkeit definieren; diese Eigenschaft ist allerdings im Allgemeinen schwer zu erhalten. Die lokale Effizienz ist insbesondere wichtig, wenn wir auf Basis der Fehlerschätzer ein neues Gitter Γ_1 konstruieren wollen, auf dem wir eine genauere Approximation berechnen wollen: Da wir wissen, dass große η_i einen großen Fehler in der Nähe implizieren, liegt es nahe, die Regionen mit großen η_i genauer zu diskretisieren, während die Diskretisierung in Regionen mit kleinen η_i gleich bleibt. Dies führt zur sogenannten *adaptiven Gittererzeugung*, die tatsächlich die Hauptmotivation für die Konstruktion von Fehlerschätzern darstellt, und heute ein wichtiges numerisches Hilfsmittel zur Lösung partieller Differentialgleichungen aller Art darstellt. Dies motiviert auch den Begriff „effizient“: Mit dieser Strategie wird nur dort verfeinert, wo sich tatsächlich große Fehler in der Lösung befinden.

Hier können wir auf diese adaptive Gitterkonstruktion aus Zeitgründen nicht näher eingehen; für Interessierte empfehlen sich die Arbeiten [10] und L. Grüne und W. Semmler, *Using Dynamic Programming with Adaptive Grid Scheme for Optimal Control Problems in Economics*, erscheint im Journal of Economic Dynamics and Control, Vorabversion erhältlich auf `http://www.elsevier.com/locate/jedc` unter den Links “Access full text articles” und dann “Articles in Press”.

In dieser Vorlesung werden wir uns darauf beschränken zu zeigen, wie sich Fehlerschätzer gemäß Definition 4.10 konstruieren lassen (diese Konstruktion stammt ebenfalls aus den zitierten Arbeiten).

4.4.2 Konstruktion der Fehlerschätzer

Die Fehlerschätzer, die wir nun betrachten wollen, gehören zur Klasse der sogenannten *residualen* Fehlerschätzer. Die Grundidee dabei ist die folgende: Die Gleichung, die wir lösen wollen, lautet

$$v_h = T_h(v_h)$$

mit dem Operator T_h aus Definition 3.5. Wie wir im Beweis von Lemma 3.16 beobachtet haben, lösen wir aber tatsächlich (zumindest approximativ) die Gleichung

$$\hat{v}_h = \pi_{\mathcal{W}} T_h(\hat{v}_h),$$

wobei $\pi_{\mathcal{W}}$ die Projektion auf den Raum der stückweise affin bilinearen Funktionen auf dem Gitter Γ bezeichnet. Die Idee besteht nun darin, das Residuum des Operators T_h bzgl. $\hat{v}_h$ auszurechnen, d.h., den Wert

$$\|\hat{v}_h - T_h(\hat{v}_h)\|$$

zu berechnen.

Definition 4.11 Wir definieren eine Funktion $\eta : \Omega \to \mathbb{R}_0^+$ mittels

$$\begin{aligned} \eta(x) &:= |\hat{v}_h(x) - T_h(\hat{v}_h)(x)| \\ &= \left| \hat{v}_h(x) - \left(\max_{u \in U} \{ hg(x,u) + \beta \hat{v}_h(f_h(x,u)) \} \right) \right| \end{aligned}$$

Basierend auf $\eta(x)$ definieren wir einen lokalen Fehlerschätzer mittels

$$\eta_i := \max_{x \in R_i} \eta(x)$$

für $i = 0, \ldots, P-1$. □

Offenbar hängen die Werte η_i in dieser Definition tatsächlich nur von den Daten des zeitdiskreten (oder zeitdiskretisierten) optimalen Steuerungsproblems (genauer von f_h, g, δ und h) ab, sowie von den Werten der Funktion $\hat{v}_h$ in der Umgebung

$$\mathcal{N}(R_i) := \{ y \in \Omega \,|\, \text{es gibt ein } x \in R_i \text{ mit } \|y - x\| \leq M_h \}, \tag{4.15}$$

wobei M_h eine obere Schranke für $\|x - f_h(x,u)\|$ für alle $x \in \Omega$ und $u \in U$ ist (im Falle der Euler–Diskretisierung gilt $M_h = hM$, wobei M eine Schranke für $\|f(x,u)\|$ ist). Der folgende Satz zeigt, dass auch die anderen Eigenschaften der lokalen Fehlerschätzer erfüllt sind.

Satz 4.12 Es sei $\delta h < 1$. Für den Fehlerschätzer aus Definition 4.11 gelten dann die Ungleichungen

$$\frac{1}{2}\,\eta \leq \|v_h - \hat{v}_h\|_\infty \leq \frac{1}{\delta h}\,\eta$$

mit $\eta = \max_{i=0,\ldots,P-1} \eta_i$. Darüberhinaus gilt

$$\frac{1}{2}\,\eta_i \leq \sup_{x \in \mathcal{N}(R_i)} |v_h(x) - \hat{v}_h(x)|$$

für die Umgebung $\mathcal{N}(R_i)$ aus (4.15).

Beweis: Mit Lemma 2.4 folgt, dass für je zwei stetige Funktionen $v_1, v_2 : \Omega \to \mathbb{R}$ die Ungleichung

$$|T_h(v_1)(x) - T_h(v_2)(x)| \leq \beta \sup_{y \in B_M(x)} |v_1(y) - v_2(y)| \tag{4.16}$$

gilt.

Wir zeigen nun zunächst die Abschätzung für η_i. Aus der Gleichung $T_h(v_h) = v_h$ ergibt sich

$$\begin{aligned} |\hat{v}_h(x) - T_h(\hat{v}_h)| &= |\hat{v}_h(x) - v_h(x) + T_h(v_h)(x) - T_h(\hat{v}_h)(x)| \\ &\leq |\hat{v}_h(x) - v_h(x)| + |T_h(\hat{v}_h)(x) - T_h(v_h)(x)| \\ &\leq 2 \sup_{y \in B_{hM}(x)} |\hat{v}_h(y) - v_h(y)|, \end{aligned}$$

wobei die letzte Ungleichung aus (4.16) und $|1-\delta h| < 1$ folgt. Die Ungleichung für η_i folgt nun leicht durch Bilden des Maximums über $x \in R_i$.

Die erste Ungleichung für η folgt sofort aus dieser Abschätzung durch Maximumsbildung über $i = 0, \ldots, P-1$.

Die zweite Abschätzung für η erhalten wir wiederum mit $T_h(v_h) = v_h$ und (4.16) aus

$$\begin{aligned} |v_h(x) - \hat{v}_h(x)| &= |T_h(v_h)(x) - \hat{v}_h(x)| \\ &= |T_h(v_h)(x) - \hat{v}_h(x) + T_h(\hat{v}_h)(x) - T_h(\hat{v}_h)(x)| \\ &\leq |T_h(\hat{v}_h)(x) - \hat{v}_h(x)| + |T_h(v_h)(x) - T_h(\hat{v}_h)(x)| \\ &\leq \eta(x) + \beta \max_{y\in\Omega} |v_h(y) - \hat{v}_h(y)| \end{aligned}$$

Durch Bilden des Maximums über $x \in \Omega$ ergibt sich

$$\|v_h - \hat{v}_h\|_\infty \leq \eta + \beta \|v_h - \hat{v}_h\|_\infty$$

und daraus

$$(1-\beta)\|v_h - \hat{v}_h\|_\infty \leq \eta,$$

also wegen $1 - \beta = \delta h$ die gewünschte Ungleichung. □

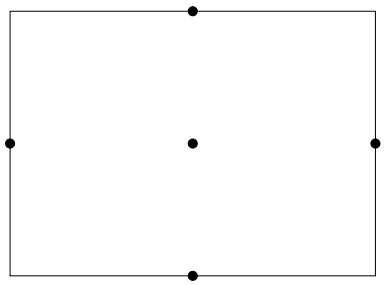

Abbildung 4.2: Testpunkte für die Auswertung von $\eta(x)$

Zwar beruht der Fehlerschätzer η_i tatsächlich nur auf Werten, die wir numerisch auswerten können. Allerdings ist es praktisch leider nicht möglich, das Maximum $\max_{x\in R_i} \eta(x)$ exakt auszurechnen, da wir die Funktion $\eta(x)$ an unendlich vielen Punkten auswerten müssten. Glücklicherweise lässt sich beweisen, dass auch die Funktion $\hat{v}_h$ Hölder stetig ist[4], womit es gerechtfertigt ist, das Maximum über R_i durch Auswertung von $\eta(x)$ in einer Menge von Testpunkten approximativ zu bestimmen. In der numerischen Praxis haben sich hierbei für zweidimensionale Rechteckgitter die in Abbildung 4.2 angegebenen 5 Testpunkte als geeignet erwiesen. Dieses Muster lässt sich leicht auf höhere Dimensionen verallgemeinern.

[4]Tatsächlich ist $\hat{v}_h$ als affin bilineare Funktion sogar Lipschitz stetig, die Lipschitz–Konstante hängt aber von der Wahl des Gitters Γ ab, während die Hölder–Stetigkeit mit von Γ unabhängigen Konstanten bewiesen werden kann

Kapitel 5

Stabilitätsanalyse optimaler Steuerungsprobleme

Bei der Analyse optimaler Steuerungsprobleme spielt das Langzeitverhalten der Lösungen eine wichtige Rolle, sowohl in technischen Anwendungen, die wir im nächsten Kapitel genauer betrachten werden, als auch in der Ökonomie. In der klassischen ökonomischen Theorie sind vor allem die Gleichgewichte interessant, die hier die Punkte darstellen, gegen die die optimalen Lösungen nach einer gewissen Zeit konvergieren. Traditionell werden diese Gleichgewichte dabei mit statischen Methoden bestimmt, d.h. durch Lösen linearer oder nichtlinearer Gleichungen, z.B. beim Bestimmen des Gleichgewichts von Angebot und Nachfrage.

In den letzten Jahren werden auch in der Ökonomie zunehmend kompliziertere dynamische Modelle verwendet, bei denen nicht nur die Gleichgewichte analysiert werden, sondern auch das Verhalten der optimalen Lösungen außerhalb der Gleichgewichte. Hierbei stellte es sich schnell heraus, dass die klassische Gleichgewichtstheorie unzureichend ist, denn in komplexeren Modellen existieren nicht unbedingt optimale Gleichgewichte, da die optimalen Lösungen nicht notwendigerweise gegen Gleichgewichte konvergieren. Aber selbst, wenn optimale Gleichgewichte existieren — wie in den Modellen, die wir in diesem Kapitel und den zugehörigen Übungen betrachten werden — müssen diese nicht eindeutig sein, so dass man nicht mehr von „dem“ optimalen Gleichgewicht sprechen kann, sondern mehrere Gleichgewichte auftreten und es vom Anfangswert abhängt, gegen welches die optimale Lösung konvergiert.

5.1 Begriffe aus der Stabilitätstheorie

5.1.1 Unkontrollierte Systeme

Bevor wir uns den Kontrollsystemen widmen, wiederholen wir zunächst einige Begriffe aus der Stabilitätstheorie für unkontrollierte dynamische Systeme. Hierunter verstehen wir hier vereinfachend die Lösungsabbildung $\Phi(t, x)$ einer Differential– oder Differenzengleichungen der Form (1.1) oder (1.2), die nicht von einer Kontrollfunktion u abhängt und die für alle

Anfangswerte $x \in \mathbb{R}^d$ und alle Zeiten $t \geq 0$ bzw. $t \in h\mathbb{N}_0$ existiert.[1]

Stabilitätskonzepte kann man für sehr allgemeine Mengen definieren, vgl. die Vorlesung zur numerischen Dynamik. In dieser Vorlesung werden wir uns auf Gleichgewichte beschränken.

Definition 5.1 Ein Punkt x^* heißt *Gleichgewicht* (oder *Ruhelage* oder *Equilibrium*) eines dynamischen Systems, falls

$$\Phi(t, x^*) = x^*$$

für alle $t \geq 0$ bzw. $t \in h\mathbb{N}_0$ gilt. □

Wir verwenden für unsere Stabilitätsdefinitionen die folgenden Klassen von Vergleichsfunktionen, die eine sehr kompakte Notation der entsprechenden Definitionen erlauben.

Definition 5.2 Wir definieren die folgende Klassen von Funktionen

$$\mathcal{K}_\infty := \left\{ \alpha : \mathbb{R}_0^+ \to \mathbb{R}_0^+ \;\middle|\; \begin{array}{l} \alpha \text{ ist stetig, streng monoton wachsend,} \\ \text{unbeschränkt und erfüllt } \alpha(0) = 0 \end{array} \right\}$$

und

$$\mathcal{KL} := \left\{ \beta : \mathbb{R}_0^+ \times \mathbb{R}_0^+ \to \mathbb{R}_0^+ \;\middle|\; \begin{array}{l} \beta \text{ ist stetig, } \beta(\cdot, t) \in \mathcal{K}_\infty \text{ für jedes } t \geq 0 \\ \text{und } \beta(r, t) \text{ ist streng monoton fallend in } t \\ \text{mit } \lim_{t \to \infty} \beta(r, t) = 0 \text{ für jedes } r > 0 \end{array} \right\}.$$

□

Abbildung 5.1 veranschaulicht eine typische $\mathcal{K}_\infty$–Funktion (links) bzw. $\mathcal{KL}$–Funktion (links und rechts).

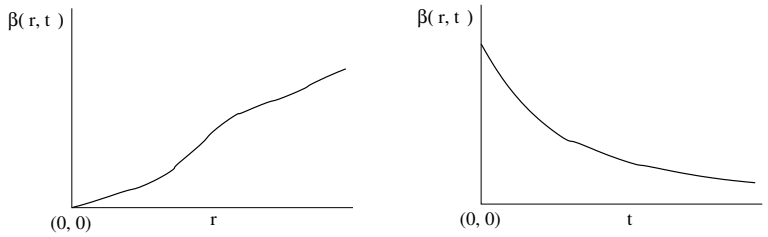

Abbildung 5.1: Typische $\mathcal{KL}$–Funktion β

Mit diesen Funktionen definieren wir die folgenden Eigenschaften.

Definition 5.3 (i) Ein Gleichgewicht x^* heißt *stabil*, falls eine Umgebung $\mathcal{N}(x^*)$ und eine $\mathcal{K}_\infty$ Funktion α existiert, so dass für alle $x \in \mathcal{N}(x^*)$ die Ungleichung

$$\|\Phi(t, x) - x^*\| \leq \alpha(\|x - x^*\|)$$

[1] Strenggenommen handelt es sich hier um ein semi–dynamisches System, da für dynamische Systeme die Lösungen für alle $t \in \mathbb{R}$ bzw. $t \in h\mathbb{Z}$ existieren müssen, was wir hier aber nicht benötigen.

für alle $t \geq 0$ bzw. $t \in h\mathbb{N}_0$ gilt.

(ii) Ein Gleichgewicht x^* heißt *asymptotisch stabil*, falls eine Umgebung $\mathcal{N}(x^*)$ und eine $\mathcal{KL}$ Funktion β existiert, so dass für alle $x \in \mathcal{N}(x^*)$ die Ungleichung

$$\|\Phi(t,x) - x^*\| \leq \beta(\|x - x^*\|, t)$$

für alle $t \geq 0$ bzw. $t \in h\mathbb{N}_0$ gilt. Falls dabei $\mathcal{N}(x^*) = \mathbb{R}^d$ ist, so heißt x^* *global asymptotisch stabil*.

(iii) Ein Gleichgewicht x^* heißt *instabil*, falls (i) nicht gilt, also falls für jede Umgebung $\mathcal{N}(x^*)$ und jede $\mathcal{K}_\infty$ Funktion α ein Punkt $x \in \mathcal{N}(x^*)$ und ein $t \geq 0$ existieren mit

$$\|\Phi(t,x) - x^*\| > \alpha(\|x - x^*\|).$$

□

Die Stabilitätsdefinitionen mittels $\mathcal{K}_\infty$ und $\mathcal{KL}$–Funktionen wurden in den späten 1950er Jahren von W. Hahn [13, 14] eingeführt und gerieten danach etwas in Vergessenheit, bis sie in den letzten Jahren in vielen Anwendungsbereichen, speziell in der Kontrolltheorie und der numerischen Dynamik, wieder entdeckt wurden. Äquivalent kann man diese Eigenschaften — ähnlich der Stetigkeit — mittels eines ε–δ Formalismus definieren, der sich in vielen Lehrbüchern findet. Der Vorteil der hier verwendeten Definition ist zum einen die elegante kurze Formulierung und zum anderen die direkte quantitative Information, da z.B. die Funktion β angibt, wie schnell die Lösungen aus der Umgebung gegen x^* konvergieren.

Der Nachteil dieser Definition ist die implizite Charakterisierung der Instabilität. Der folgende Satz gibt eine äquivalente aber anschaulichere Bedingung.

Satz 5.4 Ein Gleichgewicht x^* ist genau dann instabil im Sinne von Definition 5.3 (iii), wenn die folgende Eigenschaft gilt:
Es existiert eine Umgebung $\mathcal{N}_1(x^*)$ so dass für jede kleinere Umgebung $\mathcal{N}_2(x^*) \subset \mathcal{N}_1(x^*)$ ein $x \in \mathcal{N}_2(x^*)$ und ein $t \geq 0$ existieren mit $\Phi(t,x) \notin \mathcal{N}_1(x^*)$.

Beweis: Übungsaufgabe

Stabilität und Instabilität von Gleichgewichten für nicht–kontrollierte dynamische Systeme können durch hinreichende Bedingungen an die Linearisierung überprüft werden. Z.B. für Differentialgleichungen der Form

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

ist ein Punkt $x^* \in \mathbb{R}^d$ genau dann ein Gleichgewicht, falls $f(x^*) = 0$ gilt. In diesem Fall betrachtet man die Ableitung (oder *Linearisierung*)

$$A = \frac{d}{dx} f(x^*) \in \mathbb{R}^{d \times d},$$

berechnet die Eigenwerte λ_i dieser Matrix und betrachtet $\sigma := \max_i \Re(\lambda_i)$, also den maximalen Realteil der Eigenwerte. Ist $\sigma < 0$, so ist das Gleichgewicht x^* asymptotisch stabil, ist $\sigma > 0$, so ist es instabil. Für $\sigma = 0$ lassen anhand der Linearisierung keine Aussagen machen.

Beispiel 5.5 Betrachte die eindimensionale DGL

$$\dot{x}(t) = x(t)(1 - x(t))(1 + x(t)).$$

Hier existieren drei Gleichgewichte $x_1^* = -1$, $x_2^* = 0$ und $x_3^* = 1$. Die Ableitung von $f(x) = x(1-x)(1+x) = x - x^3$ lautet

$$\frac{d}{dx} f(x) = 1 - 3x^2,$$

also

$$\frac{d}{dx} f(x_1^*) = \frac{d}{dx} f(x_3^*) = -2 \text{ und } \frac{d}{dx} f(x_2^*) = 1.$$

Im $\mathbb{R}^1$ sind dies natürlich gerade die Eigenwerte, weswegen x_1^* und x_3^* asymptotisch stabil sind während x_2^* instabil ist. □

Für eindimensionale DGL lassen sich die Stabilitätseigenschaften auch direkt aus f ablesen. Wenn nämlich für ein Gleichgewicht x^* eine Umgebung $\mathcal{N}(x^*)$ existiert, so dass

$$f(x) < 0 \text{ für } x \in \mathcal{N}(x^*) \text{ mit } x > x^*$$

und

$$f(x) > 0 \text{ für } x \in \mathcal{N}(x^*) \text{ mit } x < x^*$$

gilt, so ist x^* asymptotisch stabil. Schwächen wir die Ungleichungen zu "$\leq$" bzw. "$\geq$" ab, so folgt Stabilität. Falls eine Umgebung $\mathcal{N}(x^*)$ existiert mit

$$f(x) > 0 \text{ für } x \in \mathcal{N}(x^*) \text{ mit } x > x^*$$

oder

$$f(x) < 0 \text{ für } x \in \mathcal{N}(x^*) \text{ mit } x < x^*$$

so folgt Instabilität.

Beispiel 5.6 Betrachte die DGL aus Beispiel 5.5. Wählen wir die Umgebungen $\mathcal{N}(x_1^*) = (-2, -1/2)$, $\mathcal{N}(x_2^*) = (-1/2, 1/2)$ und $\mathcal{N}(x_3^*) = (1/2, 2)$, so prüft man wiederum leicht nach, dass x_1^* und x_3^* asymptotisch stabil sind während x_2^2 instabil ist. □

Dieses Szenarium lässt sich schön grafisch veranschaulichen, indem man das Vorzeichen der Funktion f auf einer Geraden durch entsprechende Pfeile veranschaulicht, siehe Abb. 5.2.

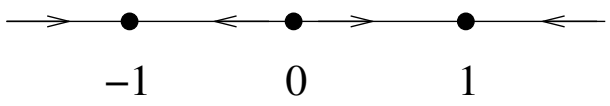

Abbildung 5.2: Schematische Stabilitätsdarstellung einer 1d DGL

Falls in einer DGL mehrere asymptotisch stabile Gleichgewichte existieren, so ist zudem interessant, für welche Anfangswerte die Lösungen gegen welche Gleichgewichte konvergieren. Formal definiert man dazu den Einzugsbereich.

Definition 5.7 Für ein Gleichgewicht x^* eines dynamischen Systems definieren wir den *Einzugsbereich* (oder auch *Stabilitätsumgebung*) als

$$\mathcal{D}(x^*) := \{x \in \mathbb{R}^d \mid \Phi(t,x) \to x^* \text{ für } t \to \infty\}.$$

□

Im Falle einer eindimensionalen DGL ist diese Menge einfach zu bestimmen, da zwei Einzugsbereiche immer durch ein instabiles Gleichgewicht getrennt sind. So erhält man im Beispiel 5.5 die Einzugsbereiche

$$\mathcal{D}(x_1^*) = (-\infty, 0), \;\; \mathcal{D}(x_2^*) = \{0\}, \;\; \mathcal{D}(x_3^*) = (0, \infty).$$

Für asymptotisch stabile Gleichgewichte ist der Einzugsbereich immer eine offene Menge.

5.1.2 Optimale Steuerungsprobleme

Wir wollen die bisher eingeführten Elemente der Stabilitätstheorie nun auf unsere optimalen Steuerungsprobleme übertragen. Die Grundidee liegt dabei darin, die optimalen Steuerungsprobleme als dynamische Systeme aufzufassen, indem wir annehmen, dass zu jedem Anfangswert eine optimale Kontrollfunktion $u^x \in \mathcal{U}$ existiert, für das also

$$v(x) = J(x, u^x)$$

gilt und die zugehörige Lösung mit $\Phi^*(t,x) = \Phi(t,x,u^x)$ bezeichnen. Die Annahme, dass eine optimale Kontrollfunktion existiert ist in vielen praktischen Problemen tatsächlich erfüllt und kann auch rigoros bewiesen werden. Wir wollen aber nicht voraussetzen, dass diese eindeutig ist, da wir damit — wie wir bald sehen werden — gerade die interessanten Phänomene ausschließen würden. Folglich können wir nicht annehmen, dass Φ^* eindeutig definiert ist, wir müssen also berücksichtigen, dass es zu einem Anfangswert u.U. mehrere Lösungen gibt. Wir werden daher stets von *einer optimalen Lösung* bzw. *den Lösungen* $\Phi^*(t,x)$ mit Anfangswert x sprechen.

Mit dieser Konvention können wir die Definitionen aus dem vorhergehenden Abschnitt nun auf optimal gesteuerte Kontrollsysteme erweitern:

Definition 5.8 (i) Wir nennen einen Punkt $x^* \in \mathbb{R}^d$ ein *optimales Gleichgewicht*, falls eine optimale Lösung $\Phi^*(t,x^*)$ mit $\Phi^*(t,x^*) = x^*$ für alle $t \geq 0$ existiert.

(ii) Ein optimales Gleichgewicht x^* heißt *stabil* bzw. *asymptotisch stabil*, falls Definition 5.3 (i) bzw. (ii) für alle in $\mathcal{N}(x^*)$ startenden Lösungen $\Phi^*(t,x)$ erfüllt ist.

(iii) Ein optimales Gleichgewicht x^* heißt *instabil*, falls in jeder Umgebung $\mathcal{N}(x^*)$ mindestens eine Lösung $\Phi^*(t,x)$ startet, die Definition 5.3 (iii) erfüllt. □

In optimalen Gleichgewichten kann man leicht den Wert der optimale Wertefunktion berechnen.

Lemma 5.9 Es sei x^* ein optimales Gleichgewicht und es seien $U_0 \subseteq U$ diejenigen Kontrollwerte, für die $f(x^*, u_0) = 0$ gilt. Dann ist die optimale Wertefunktion v in x^* gegeben durch

$$v(x^*) = \frac{1}{\delta} \max_{u_0 \in U_0} g(x^*, u_0) \tag{5.1}$$

und eine zugehörige optimale Kontrollfunktion ist gegeben durch

$$u(t) \equiv u_0^*,$$

wobei $u_0^* \in U_0$ ein maximierender Kontrollwert aus (5.1) ist.

Die Aussage gilt analog in diskreter Zeit.

Beweis: Es sei u^{x^*} die optimale Steuerung, für die $\Phi^*(x^*, t) = \Phi(t, x^*, u^{x^*})$ ein optimales Gleichgewicht ist und $u_0^* \in U_0$ ein maximierender Kontrollwert aus (5.1). Da $\Phi(t, x^*, u^{x^*}) = x^*$ ist für alle $t \geq 0$, muss $f(x^*, u^{x^*}(t)) = 0$ sein für fast alle $t \geq 0$, weswegen $u^{x^*}(t) \in U_0$ für fast alle $t \geq 0$ gilt. Damit folgt

$$\begin{aligned} v(x^*) &= \int_0^\infty e^{-\delta t} g(\Phi(t, x^*, u^{x^*}), u^{x^*}(t)) dt \\ &\leq \int_0^\infty e^{-\delta t} g(x^*, u_0^*) dt \\ &= \frac{1}{\delta} \max_{u_0 \in U_0} g(x^*, u_0). \end{aligned}$$

Andererseits gilt für $u \equiv u_0^*$

$$\begin{aligned} v(x^*) &\geq \int_0^\infty e^{-\delta t} g(\Phi(t, x^*, u), u(t)) dt \\ &= \int_0^\infty e^{-\delta t} g(x^*, u_0^*) dt \\ &= \frac{1}{\delta} \max_{u_0 \in U_0} g(x^*, u_0). \end{aligned}$$

Dies zeigt sowohl (5.1) als auch die Tatsache, dass $u(t) \equiv u_0^*$ optimal ist. □

Numerisch können wir optimale Gleichgewichte durch die approximative Berechnung optimaler Trajektorien bestimmen. Durch gezielte Simulationen optimaler Trajektorien kann man damit die Gleichgewichte sowie ihre Einzugsbereiche bestimmen.

Effizienter ist es allerdings, die approximativ optimale Feedbackabbildung $\hat{u}^*(x)$ aus Definition 4.4 zu verwenden, um die Dynamik des optimal gesteuerten System zu analysieren. Diese Abbildung kann zum einen direkt als Funktion (analog zur Wertefunktion v) grafisch dargestellt werden. Zudem kann die Abbildung $f(x, u^*(x))$ z.B. in Vektorfeldform mit Pfeilen grafisch dargestellt werden, womit man einen globalen Überblick über das Verhalten der Trajektorien des optimal gesteuerten Systems erhält. Abbildung 5.3 zeigt beide Darstellungen für das Investitionsmodell aus Abschnitt 2.2.2.

Dieses Beispiel zeigt ein typisches und sehr interessantes Phänomen optimal gesteuerter Systeme. Sowohl aus den Trajektoriensimulationen als auch aus dem Vektorfeld kann man

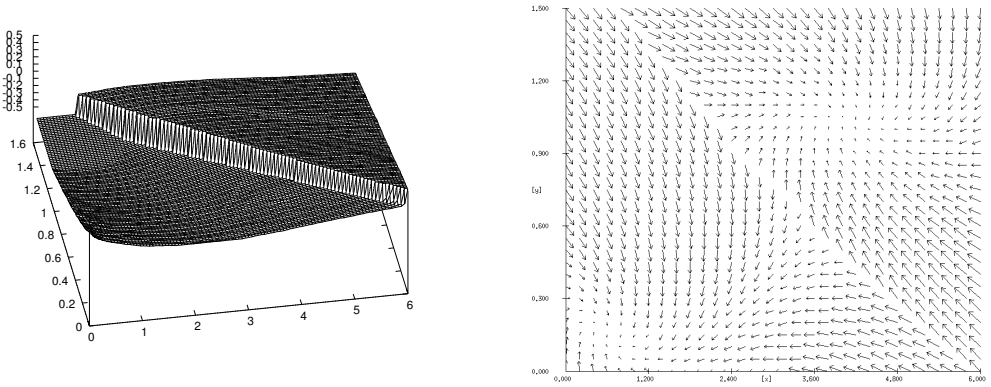

Abbildung 5.3: Optimale Kontrolle und optimales Vektorfeld für das Investitionsmodell aus Abschnitt 2.2.2

sehen, dass zwei asymptotisch stabile optimale Gleichgewichte existieren, deren Einzugsbereiche durch eine in etwa diagonal von links oben nach rechts unten verlaufende Kurve getrennt sind. An der Trennlinie der Einzugsbereiche sind das optimale Vektorfeld und das optimale Feedback $u^*(x)$ unstetig. Bei einem ungesteuerten dynamischen System würde man hier die Existenz eines instabilen Gleichgewichtes erwarten, bei unserem optimal gesteuerten System hingegen tritt dies nicht auf. Tatsächlich besteht die ganze Trennlinie aus Punkten, für die die optimale Steuerung nicht eindeutig ist: Es ist sowohl optimal, nach links unten zu steuern als auch optimal, nach rechts oben zu steuern. Es ist aber für keinen dieser Punkte optimal, in dem Punkt zu bleiben; die Kurve, an der ein solches Verhalten auftritt wird in der ökonomischen oft *Skiba–Kurve* genannt.

Dieses dynamische Verhalten erklärt auch den „Knick“ in der optimalen Wertefunktion, der in Abbildung 2.2 zu erkennen ist. Würde man das optimale Steuerungproblem mit der Nebenbedingung “konvergiere gegen das rechte obere Gleichgewicht“ lösen, so würde man eine Wertefunktion v_1 erhalten, die im rechten oberen Bereich mit v übereinstimmen würde. Analog würde man unter der Nebenbedingung “konvergiere gegen das linke untere Gleichgewicht“ eine Wertefunktion v_2 erhalten, die im unteren linken Bereich mit v übereinstimmen würde. Bei beiden Funktionen könnte man erwarten, dass sie differenzierbar sind (dies ist oft der Fall, wenn es nur ein optimales Gleichgewicht gibt). Die Funktion v ergibt sich nun gerade als Maximum der zwei Teilfunktionen v_1 und v_2[2], also

$$v(x) = \max\{v_1(x),\, v_2(x)\}$$

und ist als solche i.A. nichtdifferenzierbar an der Übergangsstelle. Dies erklärt, warum man Trennlinien zwischen Einzugsbereichen optimaler Gleichgewichte fast immer an Nichtdifferenzierbarkeitsstellen der optimalen Wertefunktion erkennen kann, da der Übergang zwischen v_1 und v_2 nur in sehr seltenen Ausnahmefällen differenzierbar ist.

Aus ökonomischer Sicht zeigt dieses Modell, dass es kein eindeutiges ökonomisches Gleichgewicht geben muss, sondern dass mehrere Gleichgewichte nebeneinander existieren können,

[2]Tatsächlich gibt es Lösungsverfahren, die diese Tatsache ausnutzen, indem sie zunächst die optimalen Gleichgewichte ermitteln und dann von diesen aus „rückwärts“ entlang der optimalen Trajektorien rechnen.

wobei es auf den Anfangszustand ankommt, gegen welches die optimalen Lösungen konvergieren. Schaut man sich die Struktur der hier verwendeten Ertragsfunktion g an, so erscheint das Verhalten auf den ersten Blick durchaus natürlich, da jede Veränderung der Investitionen durch den Term $-\alpha u^2/2$ „bestraft“ wird: Zwar wäre es aus ökonomischer Sicht durchaus sinnvoll, zu dem größeren Gleichgewicht zu steuern, da dies langfristig einen größeren Gewinn liefert, unter dem gegebenen Optimalitätskriterium ist dies aber schlicht zu “teuer”.

Diese intuitive Argumentation ist zwar einleuchtend, kann aber leicht zu falschen Schlüssen führen, denn tatsächlich ist die Situation komplexer, als es auf den ersten Blick erscheint. Dies erkennt man, wenn man das Modell mit verschiedenen Diskontraten δ löst. Bereits mit relativ kleinen Veränderungen von δ (und damit implizit des Zeithorizonts, über den optimiert wird) kann man das dynamische Verhalten erheblich ändern, und zwar in verschiedene Richtungen, vgl. Übungsaufgabe 15.

5.2 Bifurkationen

Oftmals ist man nicht an den Gleichgewichten und deren Stabilitätseigenschaften für ein einzelnes System interessiert, sondern daran, wie sich diese Eigenschaften in Abhängigkeit von einem Parameter verändern. Wenn sich die Anzahl der Gleichgewichte hierbei verändert, so spricht man von einer *Bifurkation* oder *Verzweigung*. Wir wollen dies zunächst an zwei nichtkontrolliertenen Beispielen illustrieren, bevor wir uns überlegen, wie wir unseren optimalen Steuerungsalgorithmus zur systematischen Bifurkationsanalyse verwenden können.

Beispiel 5.10 Betrachte die 1d Differentialgleichung

$$\dot{x}(t) = f(x(t), \lambda) \quad \text{mit} \quad f(x, \lambda) = -x^3 + \lambda x = -x(x^2 - \lambda),$$

wobei $\lambda \in \mathbb{R}$ ein fester (d.h. nicht zeitvarianter) Parameter ist.

Durch Berechnung der Nullstellen von f sieht man, dass die Gleichung für $\lambda \leq 0$ das eindeutige Gleichgewicht $x^*_{\lambda,1} = 0$ besitzt, während für $\lambda > 0$ die drei Gleichgewichte

$$x^*_{\lambda,1} = 0, \;\; x^*_{\lambda,2} = \sqrt{\lambda}, \;\; x^*_{\lambda,2} = -\sqrt{\lambda}$$

existieren. Die Analyse der Vorzeichen von f ergibt, dass $x^*_{\lambda,1}$ für $\lambda \leq 0$ asymptotisch stabil ist, während es für $\lambda > 0$ instabil ist. Die Gleichgewichte $x^*_{\lambda,2}$ und $x^*_{\lambda,3}$ sind asymptotisch stabil für $\lambda > 0$.

Das Gleichgewicht $x^*_{\lambda,1}$ ändert also seine Stabilitätseigenschaften, wenn λ sein Vorzeichen wechselt, dabei entstehen (bzw. verschwinden, je nach dem aus welcher λ–Richtung man schaut) die zwei Gleichgewichte $x^*_{\lambda,2}$ und $x^*_{\lambda,3}$.

Schematisch stellt man dieses Verhalten in einem *Bifurkationsdiagramm* dar, indem man die Gleichgewichte in Abhängigkeit von λ in der (x, λ)–Ebene darstellt. Asymptotisch stabile Gleichgewichte werden dabei mit einer durchgezogenen, instabile mit einer gepunkteten Linie gezeichnet. Zusätzlich kann man die Vorzeichen von f durch Pfeile symbolisieren. Abbildung 5.4 zeigt diese Darstellung für dieses Beispiel.

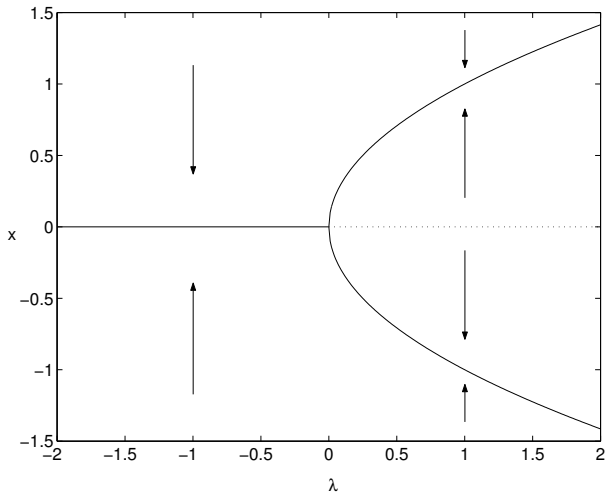

Abbildung 5.4: Bifurkationsdiagramm für Beispiel 5.10

Die hier vorliegende Bifurkation wird als *Pitchfork–Bifurkation* (Pitchfork = Heugabel) bezeichnet. □

Beispiel 5.11 Betrachte die 1d Differentialgleichung

$$\dot{x}(t) = f(x(t), \lambda) \quad \text{mit} \quad f(x, \lambda) = -x^3 + 3x^2 - \lambda,$$

mit $\lambda \in \mathbb{R}$.

Hier sind die Ausdrücke für die Nullstellen etwas komplizierter, weswegen wir direkt das Bifurkations–Diagramm in Abbildung 5.5 betrachten.

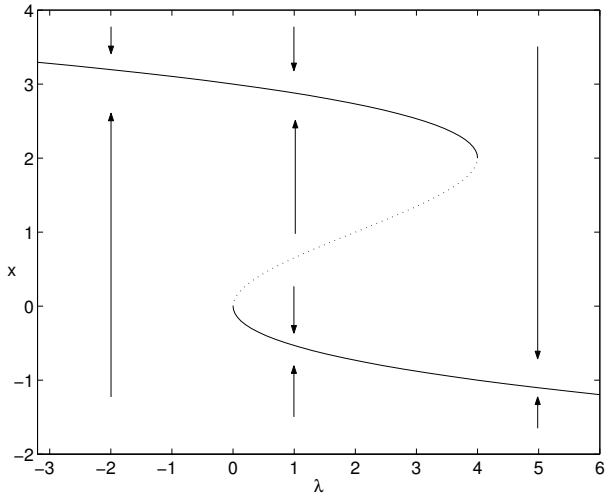

Abbildung 5.5: Bifurkationsdiagramm für Beispiel 5.11

Hier gilt: Für $\lambda < 4$ existiert ein asymptotisch stabiles Gleichgewicht $x^*_{\lambda,1}$, für $\lambda \in [0,4]$ existiert ein instabiles Gleichgewicht $x^*_{\lambda,2}$ und für $\lambda > 0$ ein asymptotisch stabiles Gleichgewicht $x^*_{\lambda,3}$. Im Bereich $\lambda \in (0,4)$, in dem alle Gleichgewichte gemeinsam existieren, gilt dabei $x^*_{\lambda,1} > x^*_{\lambda,2} > x^*_{\lambda,2}$.

Dieses Bifurkations–Szenario heißt Umkehrpunkt– oder Faltungs–Bifurkation. □

Um ein Bifurkationsszenario übersichtlich darzustellen, genügt es, das Vorzeichen des Vektorfeldes f zu plotten. Plottet man z.B. ein positives Vorzeichen in weiß und ein negatives in schwarz, so erhält man für das zweite Beispiel die alternative Darstellung aus Abbildung 5.6. Diese Art der Darstellung lässt sich mit sehr einfachen Mitteln (z.B. dem `fill` Befehl in MATLAB) leicht erzeugen.

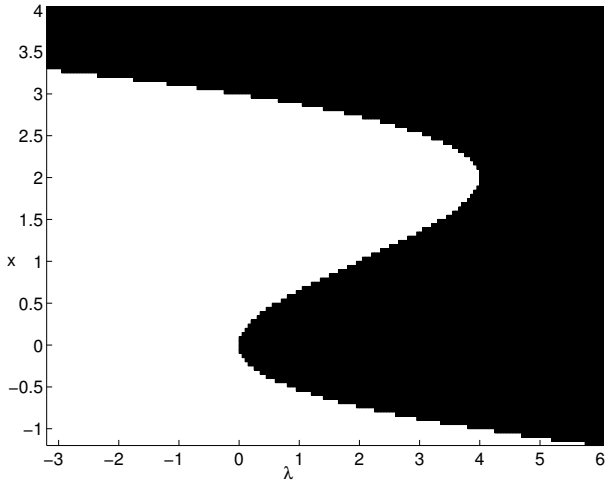

Abbildung 5.6: Bifurkationsdiagramm für Beispiel 5.11, alternative Darstellung

Wir kommen nun zurück zur optimalen Steuerung. Wir wollen eine numerische Bifurkationsanalyse für das folgende eindimensionale optimale Steuerungsproblem durchführen.

Beispiel 5.12 Wir betrachten das zeitkontinuierliche optimale Steuerungsproblem mit

$$f(x,u) = u - 0.55\,x + \frac{x^2}{1+x^2}$$

und Ertragsfunktion

$$g(x,u) = 2\sqrt{u} - \frac{\lambda}{2}x^2$$

mit Parameter $\lambda \geq 0$.

Die Interpretation dieses makroökonomischen Modells, das auf W. Brock und D. Starret zurückgeht, ist die folgende: Wir nehmen an, dass ein Unternehmen ein Werk in der Nähe eines Sees betreibt und durch die Produktion phosphathaltiges Abwasser entsteht, dass in den See eingeleitet ist. Die Variable x beschreibt den Phostphatgehalt des Sees und die Gleichung

$$\dot{x} = -0.55\,x + \frac{x^2}{1+x^2}$$

beschreibt die Selbstreinigungsfähigkeit des Sees, d.h. sie gibt an, wie schnell das Phosphat abgebaut werden kann: recht schnell für niedrige Belastung x, relativ langsam für Belastungen $x \approx 1$ und wieder etwas schneller für sehr große x (die Gleichung ergibt sich aus Erkenntnissen über das Algenwachstum bei starker Phosphatbelastung, ist aber natürlich vereinfacht). Die Kontrollvariable u beschreibt nun die Höhe der Produktion des Werks, der Einfachheit halber ist sie so normiert, dass die Menge der Phosphateinleitung gerade gleich u ist.

Das Ziel des optimalen Steuerungsproblems ist es nun, die optimale Lösung für einen geeigneten Kompromiss zwischen der Höhe der Produktion und der Verschmutzung des Sees zu finden. Die Ertragsfunktion besitzt daher zwei Komponenten: Zum einen geht die Höhe der Produktion positiv ein, d.h. je größer die Produktion desto höher der Ertrag, allerdings nicht linear sondern über die Wurzelfunktion, d.h. ein doppelte so hohe Produktion wird nur mit dem Faktor $\sqrt{2}$ stärker positiv bewertet. Auf der anderen Seite geht die Verschmutzung des Sees negativ ein, wobei der Faktor $\lambda \geq 0$ bestimmt, wie stark diese Komponente gewichtet werden soll: für $\lambda = 0$ spielt die Phosphatbelastung keine Rolle während diese für wachsendes λ immer stärker negativ in den Ertrag eingeht.

Die zu untersuchende Frage bei diesem Problem ist nun, wie das Langzeitverhalten und damit insbesondere die asymptotisch stabilen optimalen Gleichgewichte von dem Parameter λ abhängen. Sicherlich würde man erwarten, dass die Belastung des Sees für wachsendes λ abnimmt, dass sich das optimale Gleichgewicht also nach links bewegt. □

Um ein solches Bifurkationsverhalten global zu analysieren, kann man wir folgt vorgehen.

Algorithmus 5.13 (Bifurkationsdiagramm)

(1) Füge den Parameter λ als weiteren Zustand zu der Gleichung hinzu, indem $x_1 = x$ und $x_2 = \lambda$ gesetzt wird. Da der Parameter zeitlich invariant ist, führt dies auf die 2d DGL

$$\begin{aligned} \dot{x}_1(t) &= f(x_1(t), u(t)) \\ \dot{x}_2(t) &= 0 \end{aligned}$$

(2) Löse das optimale Steuerungsproblem auf $\Omega = [x^-, x^+] \times [\lambda^-, \lambda^+]$. Damit erhält man mit einer Rechnung die Lösung für das gesamte Parameterintervall $[\lambda^-, \lambda^+]$

(3) Stelle das Vorzeichen des Vektorfeldes $f(x_1, \hat{u}^*(x_1, x_2))$ gemäß Abbildung 5.6 dar. Damit erhält man das schematische Bifurkationsdiagramm des optimalen Steuerungsproblems. □

Die numerische Durchführung dieser Analyse ist eine Übungsaufgabe.

5.3 Auswirkung numerischer Fehler

Wir haben in diesem Kapitel gesehen, dass die globale numerische Lösung des optimalen Steuerungsproblems durch geeignete grafische Darstellung der Daten detaillierte Einsicht in das dynamische Verhalten des optimal gesteuerten Systems gibt. Einen Aspekt haben

wir hierbei allerdings bisher vernachlässigt, nämlich die Frage in wie weit wir uns auf die numerisch ermittelten Ergebnisse verlassen können. In der Vorlesung zur numerischen Dynamik im vergangenen Semester haben wir gesehen, dass Vorsicht geboten ist, wenn man vom numerischen Langzeitverhalten auf das echte Langzeitverhalten schließen will, dass es aber auch eine Reihe von Kriterien gibt, unter denen dies erlaubt ist. Für die hier vorliegenden optimalen Steuerungsprobleme ist eine entsprechende Theorie noch nicht entwickelt. Man kann zwar durchaus erwarten, dass sich Resultate aus der Numerik der dynamischen Systeme auf die optimalen Steuerungsprobleme übertragen lassen, wie aber die Voraussetzungen dafür genau lauten müssen und wie die Beweise im Detail zu führen sind ist bisher ungeklärt.

Kapitel 6

Stabilität von Kontrollsystemen

Während wir im letzten Kapitel die Stabilitätseigenschaften optimal gesteuerter Systeme betrachtet haben, wollen wir in diesem Kapitel das Stabilitätsverhalten der Kontrollsysteme betrachten, ohne dass wir vorher eine "ausgewählte" Kontrollfunktion festlegen. Dazu werden wir geeignete optimale Steuerungsprobleme formulieren, mit denen man dieses analysieren und beeinflussen kann. Wie bisher nehmen wir in diesem Kapitel an, dass der Kontrollwertebereich U kompakt ist und dass f die Lipschitz–Bedingung (A2) aus Abschnitt 2.1 erfüllt.

6.1 Starke und schwache asymptotische Stabilität

Wir wollen uns bei unseren Stabilitätsuntersuchungen auf asymptotische Stabilität konzentrieren. Für Kontrollsysteme (1.1) bzw. (1.2) gibt es dabei zwei naheliegende Erweiterungen von Definition 5.3(ii).

Definition 6.1 (i) Ein Punkt $x^* \in \mathbb{R}^d$ heißt *starkes* (oder *robustes*) Gleichgewicht, falls $\Phi(t, x^*, u) = x^*$ ist für alle $u \in \mathcal{U}$ und alle $t \geq 0$ (bzw. für alle $u \in \mathcal{U}_h$ und alle $t \in h\mathbb{N}$).

(ii) Ein starkes Gleichgewicht x^* heißt *stark asymptotisch stabil* (oder *robust asymptotisch stabil*), falls eine offene Umgebung $\mathcal{N}(x^*)$ und eine $\mathcal{KL}$–Funktion β existieren, so dass für alle $x \in \mathcal{N}(x^*)$ die Ungleichung

$$\|\Phi(t, x, u) - x^*\| \leq \beta(\|x - x^*\|, t)$$

für alle $u \in \mathcal{U}$ und alle $t \geq 0$ (bzw. für alle $u \in \mathcal{U}_h$ und alle $t \in h\mathbb{N}$) gilt.

(iii) Ein Punkt $x^* \in \mathbb{R}^d$ heißt *schwaches* (oder *kontrolliertes*) Gleichgewicht, falls ein $u \in \mathcal{U}$ (bzw. $u \in \mathcal{U}_h$) existiert, so dass $\Phi(t, x^*, u) = x^*$ ist für alle $t \geq 0$ (bzw. für alle $t \in h\mathbb{N}$).

(iv) Ein schwaches Gleichgewicht x^* heißt *schwach asymptotisch stabil* (oder *asymptotisch kontrollierbar*), falls eine offene Umgebung $\mathcal{N}(x^*)$ und eine $\mathcal{KL}$–Funktion β existieren, so dass für jedes $x \in \mathcal{N}(x^*)$ ein $u^x \in \mathcal{U}$ (bzw. $u^x \in \mathcal{U}_h$) existiert, so dass die Ungleichung

$$\|\Phi(t, x, u) - x^*\| \leq \beta(\|x - x^*\|, t)$$

für alle $t \geq 0$ (bzw. für alle $t \in h\mathbb{N}$) gilt. □

Diesen Definitionen liegen implizit zwei verschiedene Interpretationen der Funktionen $u \in \mathcal{U}$ zu Grunde — wenn man davon ausgeht, dass asymptotische Stabilität eine wünschenswerte Eigenschaft ist. Im Falle der starken Stabilität würde man $\mathcal{U}$ bzw. $\mathcal{U}_h$ als eine Menge von Störungen verstehen, die z.B. von außen auf das System einwirken oder durch Modellierungs– oder Diskretisierungsfehler hervorgerufen werden. Die Eigenschaft besagt dann, dass das Gleichgewicht asymptotisch stabil ist, egal welche Störung (aus der erlaubten Menge $\mathcal{U}$ bzw. $\mathcal{U}_h$) gerade auf das System wirkt.

Im Falle der schwachen Stabilität liegt — wie in den vorangegangenen Kapiteln — die Interpretation von $\mathcal{U}$ als Menge von Kontrollfunktionen zu Grunde. Hier kann man durch Auswahl eines geeigneten $u^x \in \mathcal{U}$ die asymptotische Stabilität „erzwingen“, selbst wenn Lösungen zu anderen u existieren, die möglicherweise ein ganz anderes Verhalten aufweisen.

In beiden Fällen spielt der Einzugsbereich der jeweiligen Gleichgewichte eine wichtige Rolle. Für die obigen Stabilitätskonzepte ist dieser wie folgt definiert.

Definition 6.2 (i) Der (gleichmäßige) *starke* (oder *robuste*) *Einzugsbereich* eines stark asymptotisch stabilen Gleichgewichtes $x^* \in \mathbb{R}^d$ ist definiert als

$$\mathcal{D}^+(x^*) := \left\{ x \in \mathbb{R}^d \;\middle|\; \begin{array}{l} \text{es existiert eine (möglicherweise von } x \text{ abhängige) Funktion} \\ \gamma \in \mathcal{L} \text{ mit } \|\Phi(t,x,u) - x^*\| \le \gamma(t) \text{ für alle } t \ge 0,\ u \in \mathcal{U} \end{array} \right\},$$

wobei

$$\mathcal{L} := \{\gamma : \mathbb{R}_0^+ \to \mathbb{R}_0^+ \,|\, \gamma \text{ ist stetig und streng monoton fallend mit } \lim_{t\to\infty} \gamma(t) = 0\}$$

(ii) Der *schwache* (oder *kontrollierte*) *Einzugsbereich* eines schwach asymptotisch stabilen Gleichgewichtes $x^* \in \mathbb{R}^d$ ist definiert als

$$\mathcal{D}^-(x^*) := \{x \in \mathbb{R}^d \,|\, \text{es existiert ein } u \in \mathcal{U} \text{ mit } \Phi(t,x,u) \to x^* \text{ für } t \to \infty\}$$

□

Bemerkung 6.3 Die Funktion $\gamma \in \mathcal{L}$ in der Definition des starken Einzugsbereiches bewirkt, dass die Konvergenzgeschwindigkeit gleichmäßig in $u \in \mathcal{U}$ ist. Man kann zeigen, dass bei dem entsprechenden nicht gleichmäßige Einzugsbereich höchstens Randpunkte von $\mathcal{D}(x^*)$ hinzukommen, der Beweis ist allerdings kompliziert und nicht trivial. □

Zur Analyse der Einzugsbereiche führen wir die folgenden Zeiten ein.

Definition 6.4 Es sei $\mathcal{N}(x^*)$ die Umgebung aus Definition 6.1 (ii) bzw. (iv). Dann definieren wir die Zeiten

$$t(x,u) := \inf\{t \ge 0 \,|\, \Phi(t,x,u) \in \mathcal{N}(x^*)\}$$

(mit der Konvention $\inf \emptyset = \infty$),

$$t^+(x) := \sup_{u\in\mathcal{U}} t(x,u)$$

und

$$t^-(x) := \inf_{u\in\mathcal{U}} t(x,u).$$

□

Für diese Zeiten gilt das folgende Lemma. Beachte dabei, dass die geforderte Beschränktheitsbedingung an die $\mathcal{N}(x^*)$ immer erfüllt werden kann, wenn man diese Umgebungen nur geeignet einschränkt.

Lemma 6.5 Falls die Umgebungen $\mathcal{N}(x^*)$ aus Definition 6.1 (ii) bzw. (iv) beschränkt sind, gelten die Gleichungen

(i) $\mathcal{D}^+(x^*) = \{x \in \mathbb{R}^d \,|\, t^+(x) < \infty\}$

(ii) $\mathcal{D}^-(x^*) = \{x \in \mathbb{R}^d \,|\, t^-(x) < \infty\}$

Beweis: Wir zeigen (i), (ii) folgt mit ähnlichen (sogar etwas einfacheren) Argumenten.

„$\subseteq$“: Es sei $\varepsilon > 0$ so gewählt, dass $B_\varepsilon(x^*) \subseteq \mathcal{N}(x^*)$ gilt. Es sei $x \in \mathcal{D}^+(x^*)$ und es sei $t_0 > 0$ so gewählt, dass $\gamma(t_0) < \varepsilon$ gilt für das zu x gehörige $\gamma \in \mathcal{L}$. Dann gilt für jedes $u \in \mathcal{U}$ $\|\Phi(t_0, x, u) - x^*\| \leq \gamma(t_0) < \varepsilon$, also $\Phi(t_0, x, u) \in \mathcal{N}(x^*)$ und folglich $t(x, u) \leq t_0$ für alle $u \in \mathcal{U}$. Damit gilt auch $t^+(x) \leq t_0 < \infty$.

„$\supseteq$“: Es sei $\delta > 0$ so gewählt, dass $\mathcal{N}(x^*) \subset B_\delta(x^*)$ gilt. Es sei $x \in \mathbb{R}^d$ gegeben mit $t^+(x) < \infty$. Dann existiert für jedes $u \in \mathcal{U}$ eine Zeit $t_u \leq t^+(x)$ mit $\Phi(t_u, x, u) \in \mathcal{N}(x^*)$. Für $t \geq 0$ gilt damit

$$\|\Phi(t_u + t, x, u) - x^*\| \leq \beta(\|\Phi(t, x, u) - x^*\|, t) \leq \beta(\delta, t).$$

Aus der Lipschitz–Stetigkeit von f folgt mit Hilfe des Gronwall Lemmas die Abschätzung $\sup_{t\in[0,t^+(x)]} \|\Phi(t, x, u) - x^*\| \leq C$ für ein $C > 0$ und alle $u \in \mathcal{U}$. Definieren wir also

$$\gamma(t) = e^{t^+(x)} C e^{-t} + \beta(\delta, t - t^+(x)),$$

wobei wir β mittels $\beta(\delta, t) = e^{-t}\beta(\delta, 0)$ für $t < 0$ stetig fortsetzen, so ist dies eine Funktion aus $\mathcal{L}$ mit $\gamma(t + t^+(x)) \geq \beta(\delta, t)$ für $t \geq 0$ und $\gamma(t) \geq C$ für $t \in [0, t^+(x)]$. Damit folgt $\|\Phi(t, x, u) - x^*\| \leq \gamma(t)$ und damit $x \in \mathcal{D}^+(x^*)$, was zu zeigen war. □

Satz 6.6 Die Mengen $\mathcal{D}^+(x^*)$ und $\mathcal{D}^-(x^*)$ sind offene Mengen.

Beweis: Es sei $x \in \mathcal{D}^+(x^*)$. Wir zeigen, dass eine offene Umgebung $\mathcal{N}_2(x)$ existiert mit $\mathcal{N}_2(x) \subset \mathcal{D}^+(x^*)$. Sei dazu $\varepsilon > 0$ so gewählt, dass $B_\varepsilon(x^*) \subset \mathcal{N}(x^*)$ gilt. Wir wählen $t_0 > 0$, so dass $\gamma(t_0) \leq \varepsilon/2$ gilt für die Funktion $\gamma \in \mathcal{L}$ aus der Definition von $\mathcal{D}^+(x^*)$. Da die Lipschitz–Konstante des Kontrollsystems unabhängig von u ist und die Lösungen $\Phi(t, x, u)$ für alle u durch γ gleichmäßig beschränkt sind, liefert Gronwalls Lemma eine Konstante K, so dass die Abschätzung

$$\|\Phi(t_0, x, u) - \Phi(t_0, y, u)\| \leq K\|x - y\|$$

für alle y aus einer hinreichend kleinen Umgebung $\mathcal{N}_1(x)$ gilt. Für die Umgebung $\mathcal{N}_2(x) = \mathcal{N}_1(x) \cap B_{\varepsilon/(4K)}(x)$ gilt also

$$\|\Phi(t_0, x, u) - \Phi(t_0, y, u)\| \leq \varepsilon/4$$

für alle $y \in \mathcal{N}_2(x)$ und damit mit Dreiecksungleichung

$$\|\Phi(t_0, y, u) - x^*\| \leq 3\varepsilon/4 < \varepsilon$$

für alle $u \in \mathcal{U}$. Damit folgt $t(y, u) \leq t_0$ für alle u, folglich $t^+(y) < \infty$ und damit nach Lemma 6.5 $y \in \mathcal{D}^+(x)$ für alle $y \in \mathcal{N}_2(x)$, was zu zeigen war.

Der Beweis für $\mathcal{D}^-(x^*)$ folgt ähnlich. □

6.2 Ein optimales Steuerungsproblem

Wir wollen nun ein geeignetes optimales Steuerungsproblem formulieren, mit dem wir $\mathcal{D}^+$ und $\mathcal{D}^-$ charakterisieren können. Hierzu machen wir die vereinfachende Annahme, dass die Funktion $\beta \in \mathcal{KL}$ aus Definition 6.1 (ii) bzw. (iv) von der Form

$$\beta(r,t) \leq Ce^{-\sigma t}r \tag{6.1}$$

für geeignete Konstanten $C, \sigma > 0$ ist. Diese Annahme ist einschränkend, aber in vielen Beispielen erfüllt. Tatsächlich kann alles, was wir im Folgenden machen werden, auf allgemeine $\beta \in \mathcal{KL}$ verallgemeinert werden, allerdings werden eine Reihe von Rechnungen und Abschätzungen dabei komplizierter, weswegen wir uns hier auf den einfacheren Fall (6.1) beschränken. Desweiteren werden wir im Folgenden die Annahme $x^* = 0$ für das betrachtete Gleichgewicht machen, was sich immer durch eine geeignete Verschiebung des Koordinatensystems erzielen lässt. Abkürzend schreiben wir dann $\mathcal{D}^+ = \mathcal{D}^+(x^*)$ und $\mathcal{D}^- = \mathcal{D}^-(x^*)$.

Wir betrachten nun die folgenden optimalen Steuerungsprobleme.

Definition 6.7 Wir betrachten ein (stark oder schwach) asymptotisch stabiles Gleichgewicht $x^* = 0$ und eine Funktion $g : \mathbb{R}^d \times U \to \mathbb{R}^d$ mit den folgenden Eigenschaften.

(1) g ist stetig, beschränkt und global Lipschitz–stetig in x mit Konstante L unabhängig von u.

(2) $g(0,u) = 0$ für alle $u \in U$.

(3) $\inf_{u\in U,\, \|x\|\geq\varepsilon} g(x,u) =: c_\varepsilon > 0$ für alle $\varepsilon > 0$.

Mittels dieses g definieren wir das *undiskontierte* Funktional

$$J(x,u) = \int_0^\infty g(\Phi(t,x,u),u(t))dt$$

bzw.

$$J_h(x,u) = h\sum_{k=0}^{\infty} g(\Phi_h(hk,x,u),u(hk))$$

mit der Konvention $J(x,u) = \infty$ bzw. $J_h(x,u) = \infty$ falls das Integral bzw. die Summe nicht konvergieren.

Zu diesen Funktionalen definieren wir die optimalen Wertefunktionen

$$V^+(x) = \sup_{u\in\mathcal{U}} J(x,u),\;\; V^-(x) = \inf_{u\in\mathcal{U}} J(x,u)$$

bzw.

$$V_h^+(x) = \sup_{u\in\mathcal{U}_h} J_h(x,u),\;\; V_h^-(x) = \inf_{u\in\mathcal{U}_h} J_h(x,u).$$

□

Bemerkung 6.8 Aus Annahme (3) folgt sofort $V^+(x) \geq 0$ und $V^-(x) \geq 0$ mit $V^+(x) = 0$ bzw. $V^-(x) = 0$ genau dann, wenn $x = 0$ gilt. Beachte, dass auch die optimalen Wertefunktionen den Wert ∞ annehmen können. □

Der folgende Satz klärt den Zusammenhang zwischen diesen optimalen Wertefunktionen und den Einzugsbereichen $\mathcal{D}^+$ und $\mathcal{D}^-$.

Satz 6.9 Für die optimalen Wertefunktionen V^+ und V^- gilt

(i) $\mathcal{D}^+ = \{x \in \mathbb{R}^d \,|\, V^+(x) < \infty\}$

(ii) $\mathcal{D}^- = \{x \in \mathbb{R}^d \,|\, V^-(x) < \infty\}$

Beweis: Wir zeigen den Beweis für (i), der Beweis für (ii) verläuft ähnlich. O.B.d.A. nehmen wir die Menge $\mathcal{N}(0)$ als beschränkt an.

„$\subseteq$“: Es sei $x \in \mathcal{D}^+$. Nach Lemma 6.5 gilt dann $t^+(x) < \infty$, d.h., für jedes $u \in \mathcal{U}$ existiert eine Zeit t_u mit

$$\Phi(t_u, x, u) \in \mathcal{N}(0),$$

wobei $t_u \leq t^+(x)$ gilt. Aus der Definition der starken asymptotischen Stabilität folgt

$$\|\Phi(t_u + t, x, u)\| \leq \beta(\|\Phi(t_u, x, u)\|, t) \leq Ce^{-\sigma t}\|\Phi(t_u, x, u)\|.$$

Da $\mathcal{N}(0)$ beschränkt ist, existiert ein $\delta > 0$ mit $\mathcal{N}(0) \subset B_\delta(0)$. Damit folgt

$$\|\Phi(t_u + t, x, u)\| \leq Ce^{-\sigma t}\delta.$$

Da $g(x, u)$ nach Annahme beschränkt ist, existiert eine Konstante M_g mit $|g(x, u)| \leq M_g$ für alle x, u. Aus der globalen Lipschitz Stetigkeit von g und $g(0, u) = 0$ folgt

$$|g(x, u)| = |g(x, u) - g(0, u)| \leq L\|x - 0\| = L\|x\|,$$

und damit

$$\begin{aligned} J(x,u) &= \int_0^\infty g(\Phi(t,x,u),u(t))dt \\ &= \int_0^{t_u} g(\Phi(t,x,u),u(t))dt + \int_{t_u}^\infty g(\Phi(t,x,u),u(t))dt \\ &\leq \int_0^{t_u} M_g dt + \int_{t_u}^\infty L\|\Phi(t,x,u)\|dt \\ &\leq M_g t_u + \int_0^\infty LCe^{-\sigma t}\delta dt \;\leq\; M_g t^+(x) + \frac{LC\delta}{\sigma} \end{aligned}$$

Da dies für alle $u \in \mathcal{U}$ gilt, folgt auch

$$V^+(x) \leq M_g t^+(x) + \frac{LC\delta}{\sigma} < \infty,$$

was zu zeigen war.

„$\supseteq$“: Es sei $V^+(x) < \infty$. Aus Übungsaufgabe 16 wissen wir bereits, dass dann $\Phi(t, x, u) \to 0$ konvergiert für alle $u \in \mathcal{U}$, wir müssen aber noch die Gleichmäßigkeit zeigen, was wir mit Hilfe von Lemma 6.5 machen, indem wir $t^+(x) < \infty$ zeigen.

Sei dazu $\varepsilon > 0$ so, dass $B_\varepsilon(0) \subset \mathcal{N}(0)$ gilt und sei c_ε die Konstante aus der Definition von g. Wir nehmen an, dass $t^+(x) = \infty$ ist. Dann existiert eine Folge von Kontrollfunktionen $u_k \in \mathcal{U}$, $k \in \mathbb{N}$, so dass $t(x, u_k) \to \infty$ für $k \to \infty$ gilt. Damit folgt

$$\begin{aligned} V^+(x) &\geq \int_0^\infty g(\Phi(t, x, u_k), u_k(t))dt \\ &\geq \int_0^{t(x,u_k)} c_\varepsilon dt \;=\; t(x, u_k)c_\varepsilon \;\to\; \infty \end{aligned}$$

für $k \to \infty$, ein Widerspruch. □

Ganz analog zum Beweis von Satz 2.7 beweist man das Optimalitätsprinzip für V^+ und V^-. Hier gilt für jedes $T > 0$

$$V^+(x) = \sup_{u \in \mathcal{U}} \left\{ \int_0^T g(\Phi(t, x, u), u(t))dt + V^+(\Phi(T, x, u)) \right\}$$

und

$$V^-(x) = \inf_{u \in \mathcal{U}} \left\{ \int_0^T g(\Phi(t, x, u), u(t))dt + V^-(\Phi(T, x, u)) \right\}.$$

Mit Hilfe des sogenannten Maximal– bzw. Minimalzeitproblems lassen sich auch die Zeiten t^+ und t^- als Lösung (genauer als optimale Wertefunktion) eines optimalen Steuerungsproblems definieren, weswegen es zunächst etwas unklar ist, warum wir zu dem hier definierten weiteren Problem übergehen. Der Grund dafür ist, dass die Zeiten t^+ und t^- im Allgemeinen unstetige Funktionen in x sind und sich für eine numerische Approximation daher schlecht eignen. Die Wertefunktionen V^+ und V^- hingegen sind stetig, wie der folgende Satz zeigt.

Satz 6.10 Die Funktionen V^+ und V^- sind stetig auf $\mathcal{D}^+$ bzw. $\mathcal{D}^-$.

Beweis: Wir zeigen die Aussage für V^+; die Aussage für V^- beweist man mit der gleichen Idee aber einigen technischen Änderungen, siehe Proposition 3.1(iii) in [12]

Es sei zunächst $x \in \mathcal{N}(0)$. Wie im Beweis von Satz 6.9, Teil „$\subseteq$“ ,nun mit $t^+(x) = 0$ und $\delta = \|x\|$, erhalten wir

$$V^+(x) \leq \frac{LC}{\sigma}\|x\|.$$

Für zwei beliebige Punkte $x, y \in \mathbb{R}^d$ gilt wegen der globalen Lipschitz Stetigkeit von f nach Gronwall's Lemma für jedes $u \in \mathcal{U}$ die Abschätzung

$$\|\Phi(t, x, u) - \Phi(t, y, u)\| \leq e^{Lt}\|x - y\|. \tag{6.2}$$

Wähle nun ein $x \in \mathcal{D}^+$ und ein $\varepsilon > 0$. Zum Beweis der Stetigkeit zeigen wir, dass ein $\delta > 0$ existiert, so dass $|V^+(x) - V^+(y)| \leq \delta$ gilt für alle $y \in B_\delta(x)$.

Da die Lösungen $\Phi(t, x, u)$ gleichmäßig gegen 0 konvergieren, existiert ein $T > 0$, so dass $\|\Phi(T, x, u)\| \leq \varepsilon\frac{\sigma}{8LC}$ ist. Wegen (6.2) existiert ein $\delta_1 > 0$, so dass

$$\|\Phi(T, y, u)\| \leq \varepsilon\frac{\sigma}{4LC}$$

für alle $y \in B_{\delta_1}(0)$, also

$$V^+(\Phi(T,y,u)) \leq \frac{\varepsilon}{4}.$$

Ebenfalls auf Grund von (6.2) finden wir ein δ_2, so dass

$$\max_{t\in[0,T]} \|\Phi(t,x,u) - \Phi(T,y,u)\| \leq \frac{\varepsilon}{2L_gT}$$

gilt für alle $y \in B_{\delta_2}(0)$ und alle $u \in \mathcal{U}$.

Mit Lemma 2.4 angewendet auf das Optimalitätsprinzip gilt dann für $y \in B_\delta(x)$ und $\delta = \min\{\delta_1, \delta_2\}$

$$\begin{aligned}
&|V^+(x) - V^+(y)| \\
&\leq \sup_{u\in\mathcal{U}} \left| \int_0^T g(\Phi(t,x,u),u(t))dt - \int_0^T g(\Phi(t,y,u),u(t))dt \right. \\
&\qquad \left. +\ V^+(\Phi(T,x,u)) - V^+(\Phi(T,y,u)) \right| \\
&\leq \sup_{u\in\mathcal{U}} \int_0^T |g(\Phi(t,x,u),u(t)) - g(\Phi(t,y,u),u(t))|dt \\
&\qquad +\ |V^+(\Phi(T,x,u))| + |V^+(\Phi(T,y,u))| \\
&\leq \sup_{u\in\mathcal{U}} TL \max_{t\in[0,T]} \|\Phi(t,x,u) - \Phi(T,y,u)\| + |V^+(\Phi(T,x,u))| + |V^+(\Phi(T,y,u))| \\
&\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} \ = \ \varepsilon
\end{aligned}$$

was zu zeigen war. □

Bemerkung 6.11 Für V^+ lässt sich darüberhinaus (lokale) Lipschitz Stetigkeit zeigen, falls g die Abschätzung

$$\|g(x,u) - g(y,u)\| \leq K \max\{\|x\|, \|y\|\}^{\frac{\sigma}{L}} \|x - y\|$$

erfüllt. □

6.3 Zubovs Methode

Die Charakterisierung der Einzugsbereiche mittels der optimalen Wertefunktionen V^+ und V^- ist für unsere Zwecke noch nicht direkt anwendbar und zwar aus zwei Gründen:

(a) Die Wertefunktionen sind unbeschränkt

(b) Das diskrete Optimalitätsprinzip führt nicht auf eine Kontraktion

Beide Gründe machen uns bei der numerischen Approximation Probleme, weswegen wir nun eine Transformation des Problems betrachten, mit der sich diese Probleme umgehen lassen.

Hierzu definieren wir die transformierten optimalen Wertefunktionen

$$v^+(x) = 1 - e^{-V^+(x)} \quad \text{und} \quad v^-(x) = 1 - e^{-V^-(x)}$$

mit der Konvention $e^{-\infty} = 0$.

Für gewöhnliche Differentialgleichungen ohne Kontrolle wurde diese Methode in den 1960er Jahren durch den russischen Mathematiker V. Zubov eingeführt.

Der folgende Satz fasst die wichtigsten Eigenschaften der Funktionen v^+ und v^- zusammen.

Satz 6.12 (i) $v^+(x) \in [0,1]$, $v^-(x) \in [0,1]$ für alle $x \in \mathbb{R}^d$

(ii) $\mathcal{D}^+ = \{x \in \mathbb{R}^d \,|\, v^+(x) < 1\}$, $\mathcal{D}^- = \{x \in \mathbb{R}^d \,|\, v^-(x) < 1\}$

(iii) v^+ und v^- sind stetig auf ganz $\mathbb{R}^d$

(iv) Es gilt

$$v^+(x) = \sup_{u \in \mathcal{U}} \int_0^\infty e^{-\int_0^t g(\Phi(s,x,u),u(s))ds} g(\Phi(t,x,u),u(t))dt,$$

(analog für v^-) d.h. wir haben ein diskontiertes optimales Steuerungsproblem mit Diskontfaktor

$$G(t,x,u) := e^{-\int_0^t g(\Phi(s,x,u),u(s))ds}. \tag{6.3}$$

(v) Es gilt das Optimalitätsprinzip in den zwei äquivalenten Formulierungen

$$\begin{aligned} v^+(x) &= \sup_{u \in \mathcal{U}} \left\{ \int_0^T G(t,x,u) g(\Phi(t,x,u),u(t))dt + G(T,x,u) v^+(\Phi(T,x,u)) \right\} \\ &= \sup_{u \in \mathcal{U}} \left\{ 1 - G(T,x,u) + G(T,x,u) v^+(\Phi(T,x,u)) \right\} \end{aligned}$$

(analog für v^-).

Beweis: (i) Dies folgt sofort aus Bemerkung 6.8.

(ii) Folgt sofort aus Satz 6.9.

(iii) Wir zeigen die Eigenschaft für v^+, für v^- folgt sie völlig analog. Wegen der Stetigkeit von V^+ folgt sofort, dass v^+ stetig auf $\mathcal{D}^+$ ist. Außerhalb von $\mathcal{D}^+$ ist $v^+ \equiv 1$, also ebenfalls stetig. Es bleibt zu zeigen, dass v^+ stetig auf dem Rand $\partial\mathcal{D}^+$ ist, also $v^+(x_k) \to 1$ für eine Folge $x_k \to x \in \partial\mathcal{D}^+$. Für jede solche Folge muss $t^+(x) \to \infty$ gelten, da ansonsten (wegen der Stetigkeit der Lösungen) $x \in \mathcal{D}^+$ gälte, was der Offenheit von $\mathcal{D}^+$ widerspricht. Analog zum Beweis von Satz 6.9 folgt daher $V^+(x_k) \to \infty$ und damit $v^+(x_k) \to 1$.

(iv) Nach Definition von v^+ (analog für v^-) gilt

$$\begin{aligned} v^+(x) &= 1 - e^{-V^+(x)} \\ &= 1 - e^{-\sup_{u \in \mathcal{U}} \int_0^\infty g(\Phi(t,x,u),u(t))dt} \\ &= \sup_{u \in \mathcal{U}} \left\{ 1 - e^{-\int_0^\infty g(\Phi(t,x,u),u(t))dt} \right\} \end{aligned}$$

Nach Übungsaufgabe 19 gilt für jedes $T > 0$ und jedes $u \in \mathcal{U}$

$$1 - e^{-\int_0^T g(\Phi(s,x,u),u(s))ds} = \int_0^T e^{-\int_0^t g(\Phi(s,x,u),u(s))ds} g(\Phi(t,x,u),u(t))dt \tag{6.4}$$

womit die Behauptung folgt.

(v) Beide Formulierungen folgen aus dem Optimalitätsprinzip für V^+ bzw. V^- unter Verwendung von (6.4). □

Die entsprechende zeitdiskrete Transformation

$$v_h^+(x) = 1 - e^{-V_h^+(x)}$$

führt für jedes $k \in \mathbb{N}_0$ auf das zeitdiskrete Optimalitätsprinzip

$$\begin{aligned} v_h^+(x) &= \sup_{u \in \mathcal{U}_h} \left\{ \sum_{j=0}^{k} G_h(j,x,u)(1 - e^{-hg(\Phi_h(jh,x,u),u(jh))}) \right. \\ &\qquad \left. + G_h(k+1,x,u) v_h^+(\Phi_h((k+1)h,x,u)) \right\}, \end{aligned}$$

wobei

$$G_h(k,x,u) := e^{-h\sum_{j=0}^{k-1} g(\Phi(jh,x,u),u(jh))}$$

mit der Konvention $\sum_{j=0}^{-1} = 0$ (analog für v_h^-).

Wie in Definition 3.5 wollen wir dieses Prinzip für $k = 0$ als Basis einer Iteration verwenden. Für $k = 0$ erhalten wir gerade

$$v_h^+(x) = \sup_{u \in U} \left\{ 1 - e^{-hg(x,u)} + e^{-hg(x,u)} v_h^+(f_h(x,u)) \right\}.$$

Im einfachen diskontierten Fall haben wir dabei $e^{-\delta h}$ gleich in der Definition des diskreten Problems durch die lineare Approximation $1 - \delta h$ ersetzt, was wir hier mittels $e^{-hg(x,u)} \approx 1 - hg(x,u)$ genau so machen. Somit erhalten wir

$$v_h^+(x) = \sup_{u \in U} \left\{ hg(x,u) + (1 - hg(x,u)) v_h^+(f_h(x,u)) \right\}$$

und den zugehörigen Operator

$$T_h(w)(x) = \sup_{u \in U} \left\{ hg(x,u) + (1 - hg(x,u)) w(f_h(x,u)) \right\}.$$

Ganz analog zum einfachen diskontierten Fall sieht man, dass T_h die Ungleichung

$$|T_h(w_1)(x) - T_h(w_2)(x)| \leq \sup_{u \in U} (1 - hg(x,u)) \|w_1 - w_2\|_\infty$$

erfüllt. Das Problem ist nun, dass der Faktor $\beta(x,u) := (1 - hg(x,u))$ für $x \neq 0$ zwar < 1 ist; allerdings gilt diese Abschätzung nicht gleichmäßig und ist für $x = 0$ sogar verletzt. Der Operator T_h ist also keine Kontraktion.

Um dies Problem zu beheben, werden wir unser optimales Steuerungsproblem geeignet modifizieren. Dazu wählen wir einen *Regularisierungsparameter* $\rho > 0$, setzen

$$g_\rho(x,u) = \max\{\rho,\, g(x,u)\} \tag{6.5}$$

und betrachten das durch

$$v_\rho^+(x) = \sup_{u\in\mathcal{U}} \int_0^\infty e^{-\int_0^t g_\rho(\Phi(s,x,u),u(s))ds} g(\Phi(t,x,u),u(t))dt,$$

(analog für v^-) definierte Problem. Hier lautet das Optimalitätsprinzip

$$v_\rho^+(x) = \sup_{u\in\mathcal{U}} \left\{ \int_0^T G_\rho(t,x,u) g(\Phi(t,x,u),u(t))dt + G_\rho(T,x,u) v_\rho^+(\Phi(T,x,u)) \right\}$$

mit

$$G_\rho(t,x,u) := e^{-\int_0^t g_\rho(\Phi(s,x,u),u(s))ds}.$$

Mit der gleichen Herleitung wie oben erhalten wir damit den Operator

$$T_{\rho,h}(w)(x) = \sup_{u\in U} \left\{ hg(x,u) + (1 - hg_\rho(x,u)) w(\Phi_h(T,x,u)) \right\},$$

der nun tatsächlich eine Kontraktion mit Konstante $1-h\rho < 1$ ist. Modifiziert man unseren Algorithmus derart, dass man die Konstante $\beta = 1-\delta h$ überall durch $1-hg_\rho(x,u)$ ersetzt, so löst dieser nun das hier gegebene Problem.

Bemerkung 6.13 Analog zu den Aussagen in den Kapiteln 2 und 3 kann man zeigen, dass v_ρ^+ und v_ρ^- Hölder stetig sind und dass der Diskretisierungsfehler von der gleichen Form ist, wobei δ durch ρ ersetzt wird. □

Durch die Einführung des Parameters ρ haben wir nun unser ursprüngliches optimales Steuerungsproblem so modifiziert, dass es mit unseren Methoden numerisch gelöst werden kann. Es stellt sich aber die Frage, in wie weit die Funktionen v_ρ^+ und v_ρ^- noch eine brauchbare Lösung für unser Problem darstellen. Der folgende Satz zeigt zum einem, dass die ρ–Funktionen gleichmäßig gegen die Originalfunktionen konvergieren und zum anderen, dass die Charakterisierung von DD^+ und $\mathcal{D}^-$ für alle hinreichend kleinen $\rho > 0$ *exakt* erhalten bleibt.

Satz 6.14 (i) Es gelten die Ungleichungen

$$v_\rho^+(x) \le v^+(x) \quad \text{und} \quad v_\rho^-(x) \le v^-(x)$$

für alle $\rho > 0$ und alle $x \in \mathbb{R}^d$.

(ii) Es gilt die Konvergenz

$$\|v_\rho^+ - v^+\|_\infty \to 0 \quad \text{und} \quad \|v_\rho^- - v^-\|_\infty \to 0$$

für $\rho \to 0$.

(iii) Falls $\rho > 0$ in (6.5) so klein ist, dass die Bedingung

$$g_\rho(x,u) = g(x,u)$$

für alle $x \notin \mathcal{N}(0)$ und alle $u \in U$ gilt, so gelten die Charakterisierungen

$$\mathcal{D}^+ = \{x \in \mathbb{R}^d \,|\, v_\rho^+(x) < 1\} \quad \text{und} \quad \mathcal{D}^- = \{x \in \mathbb{R}^d \,|\, v_\rho^-(x) < 1\}.$$

Beweis: Wir zeigen die Behauptungen für v_ρ^+, für v_ρ^- laufen die Beweise völlig analog.

(i) und (ii): siehe Übungsaufgabe 20.

(ii) Aus der Stetigkeit von v^+ und g und aus $v^+(x) > 0$ für $x \neq 0$ und $g(0,u) = 0$ folgt, dass für jedes $\delta > 0$ ein $\rho > 0$ mit

$$\{x \in \mathbb{R}^d \,|\, \inf_{u \in U} g(x,u) \leq \rho\} \subseteq \{x \in \mathbb{R}^d \,|\, v^+(x) \leq \delta\}$$

existiert, es gilt also die Implikation

$$v^+(x) > \delta \;\; \Rightarrow \;\; g_\rho(x,u) = g(x,u) \text{ für alle } u \in U.$$

Wir wählen ein beliebiges $\delta > 0$, das zugehörige $\rho > 0$ sowie ein beliebiges $\gamma > 0$ und ein $x \in \mathbb{R}^d$. Wir wählen ein $u_\gamma \in \mathcal{U}$ mit

$$v^+(x) \leq \int_0^\infty G(t,x,u_\gamma) g(\Phi(t,x,u_\gamma), u_\gamma(t)) dt + \gamma.$$

Wenn $v^+(x) = 1$ ist, folgt nach (iii) $v^+(x) = v_\rho^+(x)$ für hinreichend kleines ρ. Für $v^+(x) < 1$ ist auch das Integral < 1, weswegen $\Phi(t,x,u_\gamma) \to 0$ und damit $v^+(\Phi(t,x,u_\gamma)) \to 0$ gilt. Es sei $T \geq 0$ die minimale Zeit mit $v^+(\Phi(T,x,u_\gamma)) \leq \delta$. Dann gilt

$$\begin{aligned}
& v^+(x) - v_\rho^+(x) - \gamma \\
\leq \;& \int_0^\infty G(t,x,u_\gamma) g(\Phi(t,x,u_\gamma), u_\gamma(t)) dt - \int_0^\infty G_\rho(t,x,u_\gamma) g(\Phi(t,x,u_\gamma), u_\gamma(t)) dt \\
\leq \;& \int_0^T \underbrace{(G(t,x,u_\gamma) - G_\rho(t,x,u_\gamma)) g(\Phi(t,x,u_\gamma), u_\gamma(t))}_{=0} dt + G(T,x,u_\gamma) v^+(\Phi(T,x,u_\gamma)) \\
\leq \;& \delta.
\end{aligned}$$

Da das gewählte ρ nicht von x abhängt, erhalten wir mit (i) also

$$\|v^+ - v_\rho^+\|_\infty \leq \delta,$$

womit die Behauptung folgt. □

6.4 Das Feedback–Stabilisierungsproblem

In diesem Abschnitt betrachten wir das sogenannte Feedback–Stabilisierungsproblem. Wir werden zeigen, dass die zur Zubov–Wertefunktion gehörigen optimalen Steuerungen das Problem lösen und das die numerischen optimalen Steuerungen, die wir aus der numerischen Lösung der Zubov–Wertefunktion erhalten, dieses Problem in einem geeigneten Sinne "approximativ" lösen. Wir definieren zunächst, was genau wir unter dem Feedback–Stabilisierungsproblem verstehen. Um die Beschreibung technisch einfach zu halten, beschränken wir uns dabei auf Probleme in diskreter Zeit.

Definition 6.15 Gegeben sei ein zeitdiskretes Kontrollsystem

$$x(t+h) = f_h(x(t), u(t))$$

mit einem schwach asymptotisch stabilen Gleichgewicht $x^* = 0$. Das Feedback–Stabilisierungsproblem besteht darin, eine Feedback–Abbildung $F : \mathbb{R}^n \to U$ zu finden, so dass das Gleichgewicht $x^* = 0$ für die Differenzengleichung

$$x(t+h) = f_h(x(t), F(x(t))) \tag{6.6}$$

asymptotisch stabil ist. □

Im Folgenden bezeichnen wir die Lösung der Differenzengleichung (6.6) mit $\Phi_h(t, x, F)$.

Zur Lösung des Stabilisierungsproblems betrachten wir die Lösung v_h^- des diskreten optimalen Steuerungsproblems aus dem vorangegangenen Abschnitt, und zwar in der linearisierten Form, also mit dem Optimalitätsprinzip

$$v_h^-(x) = \min_{u \in U} \left\{ hg(x,u) + (1 - hg(x,u)) v_h^-(f_h(x,u)) \right\}. \tag{6.7}$$

Durch die Linearisierung $e^{-hg(x,u)} \approx 1 - hg(x,u)$ ist dies nicht mehr exakt die Transformation $1 - e^{-V_h^-(x)}$, sondern eine Approximation dieser Funktion, die aber ebenso wie die Originalfunktion den Einzugsbereich mittels

$$\mathcal{D}^- = \{x \in \mathbb{R}^d \,|\, v_h^-(x) < 1\}$$

charakterisiert.

Wir definieren die Feedback–Abbildung F nun ganz analog zu dem optimalen Feedback aus Definition 4.1, wählen also $F(x)$ so, dass das Minimum in (6.7) in $u = F(x)$ angenommen wird, d.h. so, dass

$$v_h^-(x) = hg(x,u) + (1 - hg(x, F(x))) v_h^-(f_h(x, F(x))) \tag{6.8}$$

gilt. Damit gilt der folgende Satz.

Satz 6.16 Das Feedback F löst das Stabilisierungsproblem. Darüberhinaus ist der Einzugsbereich

$$\mathcal{D}_F := \{x \in \mathbb{R}^d \,|\, \Phi_h(t, x, F) \to 0 \text{ für } t \to \infty\}$$

des Feedback–Systems (6.6) gerade gleich $\mathcal{D}^-$, also gleich dem schwachen Einzugsbereich von x^*.

Beweis: Wir beweisen zunächst die asymptotische Stabilität. Dazu müssen wir zeigen, dass eine $\mathcal{KL}$–Funktion β und eine Umgebung $\mathcal{N}(0)$ existieren, so dass für alle $x \in \mathcal{N}(0)$ die Ungleichung

$$\|\Phi_h(t,x,F)\| \leq \beta(\|x\|,t)$$

gilt. Dazu wählen wir die Umgebung

$$\mathcal{N}(0) = \{x \in \mathbb{R}^d \,|\, v_h^-(x) < 1/2\}.$$

Für alle $x \in \mathbb{R}^d$ gilt die Gleichung

$$\begin{aligned} v^-(x) &= hg(x,F(x)) + (1 - hg(x,F(x)))v_h^-(f_h(x,F(x))) \\ &= v_h^-(f_h(x,F(x))) + hg(x,F(x))(1 - v_h^-(f_h(x,F(x)))), \end{aligned}$$

und damit

$$v_h^-(f_h(x,F(x))) - v_h^-(x) \leq hg(x,F(x))(v_h^-(f_h(x,F(x))) - 1).$$

Da die rechte Seite dieser Ungleichung ≤ 0 ist (beachte, dass $v_h^- \leq 1$ ist), folgt, dass v_h^- entlang der Lösungen $\Phi_h(t,x,F)$ nicht wachsen kann, woraus für alle $t \in h\mathbb{N}$ die Implikation

$$v_h^-(x) \leq 1/2 \;\Rightarrow\; v_h^-(\Phi_h(t,x,F)) \leq 1/2$$

und damit

$$x \in \mathcal{N}(0) \;\Rightarrow\; \Phi_h(t,x,F) \in \mathcal{N}(0)$$

folgt. Auf $\mathcal{N}(0)$ gilt wegen $v_h^-(x) < 1/2$ sogar

$$v_h^-(f_h(x,F(x))) - v_h^-(x) \leq -hg(x,F(x))/2 \leq -hc_{\|x\|}/2, \tag{6.9}$$

wobei

$$c_\varepsilon = \inf_{\|x\| \geq \varepsilon,\, u \in U} g(x,u)$$

die Konstanten aus der Annahme an g sind. Indem wir diese Konstanten falls nötig verkleinern, können wir o.B.d.A. annehmen, dass sie die Ungleichung $c_\varepsilon \leq \varepsilon/h$ erfüllen. Da $c_\varepsilon > 0$ ist für $\varepsilon > 0$, fällt die Funktion v_h^- streng monoton entlang von Lösungen so lange $\Phi_h(t,x,F) \in \mathcal{N}(0) \setminus \{0\}$ ist; Funktionen mit dieser Eigenschaft werden *Lyapunov–Funktionen* genannt. Wir definieren nun rekursiv eine Funktion $\mu : \mathbb{R}_0^+ \times \mathbb{R}_0^+ \to \mathbb{R}_0^+$ mittels

$$\mu(r,0) = r, \quad \mu(r,t+h) = \mu(r,t) - hc_{\mu(r,t)}/2$$

für $t \in h\mathbb{N}_0$, die wir (z.B. mittels linearer Interpolation) stetig für $t \in \mathbb{R}_0^+$ fortsetzen. Dies ist eine $\mathcal{KL}$–Funktion, denn da c_ε monoton in ε ist, ist μ streng monoton in r für alle t. Zudem gilt $\mu(r,t) \to 0$ für $t \to \infty$: Zunächst ist $\mu(r,t)$ streng monoton fallend in t, und wegen $c_\varepsilon \leq \varepsilon/h$ nach unten durch 0 beschränkt. Also konvergiert $\mu(r,t) \searrow r^*$ für $t \to \infty$. Nehmen wir nun an, dass $r^* > 0$ ist und wählen ein $t \in h\mathbb{N}$, so dass $\mu(r,t) \leq r^* + hc_{r^*}/4$ gilt, so folgt

$$\mu(r,t+h) = \mu(r,t) - hc_{\mu(r,t)}/2 \leq \mu(r,t) - hc_{r^*}/2 \leq r^* + hc_{r^*}/4 - hc_{r^*}/2 < r^*,$$

ein Widerspruch. Also ist $r^* = 0$ und folglich ist $\mu \in \mathcal{KL}$.

Aus der Konstruktion von μ folgt für $x \in \mathcal{N}(0)$ die Ungleichung

$$v_h^-(f_h(x, F(x))) \leq \mu(v_h^-(x), h)$$

und damit per Induktion unter Ausnutzung der Monotonie von μ die Ungleichung

$$v_h^-(\Phi_h(t, x, F)) \leq \mu(v_h^-(x), t).$$

Aus der $\mathcal{KL}$–Funktion μ konstruieren wir nun die gesuchte $\mathcal{KL}$–Funktion β. Dazu definieren wir

$$\alpha_1(r) = \inf_{x \in \mathcal{N}(0), \|x\| \geq r} v_h^-(x) \quad \text{und} \quad \alpha_2(r) = \sup_{x \in \mathcal{N}(0), \|x\| \leq r} v_h^-(x)$$

für alle r für die die Menge $\{x \in \mathcal{N}(0) \mid \|x\| \geq r\}$ nicht leer ist. Beide Funktionen sind stetig und streng monoton wachsend für hinreichend kleine r und wir können sie für große r so fortsetzen, dass sie in $\mathcal{K}_\infty$ liegen. Für diese Funktionen gilt für $x \in \mathcal{N}(0)$ die Ungleichung

$$\alpha_1(\|x\|) \leq v_h^-(x) \leq \alpha_2(\|x\|).$$

Mit Hilfe dieser Funktionen definieren wir nun

$$\beta(r, t) = \alpha_1^{-1}(\mu(\alpha_2(r), t)).$$

Aus den Stetigkeits– und Monotonieeigenschaften von μ, α_1 und α_2 folgt, dass $\beta \in \mathcal{KL}$ liegt.

Für $x \in \mathcal{N}(0)$ gilt damit

$$\begin{aligned} \|\Phi_h(t, x, F)\| &\leq \alpha_1^{-1}(v_h^-(\Phi_h(t, x, F))) \leq \alpha_1^{-1}(\mu(v_h^-(x), t)) \\ &\leq \alpha_1^{-1}(\mu(\alpha_2(\|x\|), t)) = \beta(\|x\|, t) \end{aligned}$$

und damit die gewünschte Ungleichung für die asymptotische Stabilität.

Es bleibt zu zeigen, dass der Einzugsbereich $\mathcal{D}_F$ von x^* für (6.6) gerade gleich $D^- = \{x \in \mathbb{R}^d \mid v_h^-(x) < 1\}$ ist. Wir wählen dazu einen Punkt $x \in \mathcal{D}^- \setminus \mathcal{N}(0)$, setzen $\gamma = 1 - v_h^-(x) > 0$ und $\sigma = \inf_{x \notin \mathcal{N}(0),\, u \in U} g(x, u) > 0$. Analog zu (6.9) erhalten wir die Ungleichung

$$v_h^-(f_h(x, F(x))) - v_h^-(x) \leq -h\gamma\sigma/2$$

und damit per Induktion

$$v_h^-(\Phi_h(jh, x, F)) - v_h^-(x) \leq -jh\gamma\sigma/2, \tag{6.10}$$

solange $\Phi_h((j-1)h, x, F) \notin \mathcal{N}(0)$ ist. Wählen wir nun das minimale $j \in \mathbb{N}$ mit $v_h^-(x) - jh\gamma\sigma/2 < 1/2$, so gilt entweder $\Phi_h(kh, x, F) \in \mathcal{N}(0)$ für ein $k < j$ oder es gilt (6.10) und damit $v_h^-(\Phi_h(jh, x, F)) < 1/2$ und damit $\Phi_h(jh, x, F) \in \mathcal{N}(0)$. In beiden Fällen existiert ein $\tau \geq 0$ mit $\Phi_h(\tau, x, F) \in \mathcal{N}(0)$, woraus $\Phi_h(t, x, F) \to 0$ für $\tau \to \infty$ und damit $x \in \mathcal{D}_F$ folgt. □

Bemerkung 6.17 Da man aus jeder Feedback–Abbildung mittels

$$u_h(t) = F(\Phi_h(t, x, F))$$

eine zeitabhängige Kontrollfunktion $u_h \in \mathcal{U}_h$ gewinnen kann, folgt sofort, dass der Einzugsbereich $\mathcal{D}_F$ nicht größer als $\mathcal{D}^-$ sein kann. Für jede Feedback–Abbildung gilt also $\mathcal{D}_F \subseteq \mathcal{D}^-$. Wir haben mit dieser Konstruktion also nicht nur das Stabilisierungsproblem gelöst, sondern dazu noch den maximal möglichen Einzugsbereich realisiert. □

Numerisch können wir das stabilisierende F in dieser Form nicht realisieren, da wir v_h^- nicht exakt sondern nur approximativ berechnen können. Unsere Approximation besteht dabei aus zwei Schritten: Zuerst wird v_h^- durch $v_{\rho,h}^-$ approximiert und dann wird diese Funktion auf dem Gitter Γ durch eine Funktion $\hat{v}_{\rho,h}^-$ approximiert. Auf Basis dieser Funktion definieren wir eine Feedback–Abbildung $\widehat{F}$ mittels

$$\begin{aligned} & hg(x, \widehat{F}(x)) + (1 - hg_\rho(x, \widehat{F}(x)))\hat{v}_{\rho,h}^-(f_h(x, \widehat{F}(x))) \\ &= \inf_{u \in U}\{hg(x,u) + (1 - hg_\rho(x,u))\hat{v}_{\rho,h}^-(f_h(x,u))\}. \end{aligned} \tag{6.11}$$

Leider kann man nun nicht mehr erwarten, dass dieses $\widehat{F}$ das Stabilisierungsproblem exakt löst. Man kann aber immer noch eine geeignete approximative Form der asymptotischen Stabilität beweisen, wie sie in dem folgenden Satz formuliert ist.

Satz 6.18 Gegeben sei eine kompakte Teilmenge $K \subset \mathcal{D}^-$, und ein $\delta > 0$. Dann existiert eine $\mathcal{KL}$–Funktion β so dass für hinreichend genaues $\rho > 0$ und hinreichend genaue Approximation $\hat{v}_{\rho,h}^-$ von $v_{\rho,h}^-$ sowie die Feedback–Abbildung $\widehat{F}$ die Ungleichung

$$\|\Phi_h(t, x, \widehat{F})\| \leq \max\{\beta(\|x\|, t), \delta\}$$

für alle $x \in \mathcal{N}(0) = \{x \in \mathbb{R}^d \,|\, v_h^-(x) < 1/4\}$. Darüberhinaus enthält der Einzugsbereich

$$\mathcal{D}_{\widehat{F}} := \{x \in \mathbb{R}^d \,|\, \Phi_h(t, x, \widehat{F}) \leq \delta \text{ für alle hinreichend großen } t \in h\mathbb{N}\}$$

des Feedback–Systems (6.6) die Menge K, es gilt also $K \subset \mathcal{D}_{\widehat{F}}$.

Beweis: Für das gegebene $\delta > 0$ wählen wir ein $\varepsilon \in (0, 1/2]$, so dass die Inklusion

$$B_1 := \{x \in \mathbb{R}^d \,|\, v_h^-(x) < \varepsilon\} \subseteq B_\delta(0)$$

gilt und setzen

$$B_2 := \{x \in \mathbb{R}^d \,|\, v_h^-(x) < \varepsilon/2\}.$$

Wir definieren

$$\sigma := \min\{\inf_{x \notin B_2,\, u \in U} g(x,u), \varepsilon/h\} \quad \text{und} \quad \gamma = \inf_{x \in K}(1 - v_h^-(x)) \in (0,1).$$

Nun wählen wir ρ so klein und das Gitter so fein, dass die Ungleichungen

$$\|\hat{v}_{\rho,h}^- - v_h^-\|_\infty \leq h\gamma\sigma/16$$

und

$$\|g_\rho - g\|_\infty \leq \gamma\sigma/16$$

gelten.

Aus dem Optimalitätsprinzip für v_h^- und den obigen Ungleichungen erhalten wir

$$\begin{aligned} v_h^-(x) &= \inf_{u\in U}\left\{hg(x,u) + (1-hg(x,u))v_h^-(f_h(x,u))\right\} \\ &\geq \inf_{u\in U}\left\{hg(x,u) + (1-hg_\rho(x,u))\hat{v}_{\rho,h}^-(f_h(x,u))\right\} - h\gamma\sigma/8 \\ &= hg(x,\widehat{F}(x)) + (1-hg_\rho(x,\widehat{F}(x)))\hat{v}_{\rho,h}^-(f_h(x,\widehat{F}(x))) - h\gamma\sigma/8 \\ &\geq hg(x,\widehat{F}(x)) + (1-hg(x,\widehat{F}(x)))v_h^-(f_h(x,\widehat{F}(x))) - h\gamma\sigma/4. \end{aligned}$$

Daraus folgt

$$v_h^-(f_h(x,\widehat{F}(x))) \leq v_h^-(x) + hg(x,\widehat{F}(x))(v_h^-(f_h(x,\widehat{F}(x))) - 1) + h\gamma\sigma/4,$$

woraus direkt die Abschätzungen

$$v_h^-(f_h(x,\widehat{F}(x))) \leq v_h^-(x) + \min\{\gamma/2,\, \varepsilon/4\}$$

folgen.

Wir betrachten nun zunächst ein $x \in B_1$ und zeigen, dass $f_h(x,\widehat{F}(x)) \in B_1$ gilt. Für $x \in B_1$ gilt entweder $v_h^-(x) \leq \varepsilon/2$ und wir erhalten

$$v_h^-(f_h(x,\widehat{F}(x))) \leq \varepsilon/2 + h\gamma\sigma/4 < \varepsilon/2 + \varepsilon/2 = \varepsilon$$

oder es gilt $v_h^-(x) \geq \varepsilon/2$, also $x \notin B_2$ und wir erhalten

$$\begin{aligned} v_h^-(f_h(x,\widehat{F}(x))) &\leq v_h^-(x) + h\underbrace{g(x,\widehat{F}(x))}_{\geq\sigma}\underbrace{(v_h^-(f_h(x,\widehat{F}(x))) - 1)}_{\leq\varepsilon+\varepsilon/4-1<-3/8} + h\gamma\sigma/4 \\ &\leq v_h^-(x) - 3h\sigma/8 + h\gamma\sigma/4 \;\leq\; v_h^-(x) \;\leq\; \varepsilon. \end{aligned}$$

In beiden Fällen gilt also $v_h^-(f_h(x,\widehat{F}(x))) < \varepsilon$ und damit $f_h(x,\widehat{F}(x)) \in B_1$. Per Induktion erhalten wir damit

$$x \in B_1 \quad\Rightarrow\quad \Phi_h(t,x,\widehat{F}) \in B_1 \tag{6.12}$$

für alle $t \in h\mathbb{N}$ (man sagt, B_1 ist eine *vorwärts invariante* Menge).

Als nächstes betrachten wir ein $x \in \mathcal{N}(0)\setminus B_1$. Wie für B_1 beweist man, dass $\mathcal{N}(0)$ vorwärts invariant ist, weswegen auch $f_h(x,\widehat{F}(x)) \in \mathcal{N}(0)$ liegt. Damit gilt

$$\begin{aligned} v_h^-(f_h(x,\widehat{F}(x))) &\leq v_h^-(x) + h\underbrace{g(x,\widehat{F}(x))}_{\geq c_{\|x\|}}\underbrace{(v_h^-(f_h(x,\widehat{F}(x))) - 1)}_{\leq -3/4} + h\gamma\sigma/4 \\ &\leq v_h^-(x) - 3hc_{\|x\|}/4 + \underbrace{h\gamma\sigma/4}_{\leq hc_{\|x\|}/4} \\ &\leq v_h^-(x) - hc_{\|x\|}/2. \end{aligned}$$

Wir erhalten also die gleiche Ungleichung wie im Beweis von Satz 6.16, woraus mit dem gleichen β wie im dortigen Beweis die Ungleichung

$$\|\Phi_h(t,x,\widehat{F})\| \le \beta(\|x\|,t) \tag{6.13}$$

folgt, und zwar für alle $x \in \mathcal{N}(0)$ und alle $t \in h\mathbb{N}_0$ für die $\Phi_h(\tau,x,\widehat{F}) \notin B_1$ gilt für alle $\tau = 0,h,\dots,t-h$.

Falls $\Phi_h(\tau,x,\widehat{F}) \in B_1$ gilt für ein $\tau \le t$, so folgt aus (6.12), dass

$$\Phi_h(t,x,\widehat{F}) = \Phi_h(t-\tau,\underbrace{\Phi_h(\tau,x,\widehat{F})}_{\in B_1},\widehat{F}) \in B_1 \tag{6.14}$$

gilt. Zusammen erhalten wir also, dass für jedes $x \in \mathcal{N}(0)$ und jedes $t \in h\mathbb{N}_0$ entweder (6.13) oder (6.14) gilt, woraus mit der Wahl von B_1 die gewünschte Ungleichung

$$\|\Phi_h(t,x,\widehat{F})\| \le \max\{\beta(\|x\|,t),\delta\}$$

folgt.

Die Tatsache, dass $K \subset \mathcal{D}_{\widehat{F}}$ gilt, folgt ganz analog zum Beweis von Satz 6.16 aus der Ungleichung

$$v_h^-(f_h(x,\widehat{F}(x))) \le v_h^-(x) + h\underbrace{g(x,\widehat{F}(x))}_{\ge\sigma}\underbrace{(v_h^-(f_h(x,\widehat{F}(x))) - 1)}_{\le -\gamma+\gamma/2=-\gamma/2} + h\gamma\sigma/4 \le -h\gamma\sigma/4$$

für $x \in K \setminus \mathcal{N}(0)$, die sich aus der Wahl von γ und σ ergibt und aus der

$$\Phi_h(t,x,\widehat{F}) \in \mathcal{N}(0)$$

für alle $x \in K$ und alle hinreichend großen t folgt. □

6.5 Numerische Berechnung von Einzugsbereichen

Im Prinzip bieten die numerischen Approximationen $\hat{v}_h^+$ und $\hat{v}^-$ der Funktionen v^+ und v^- (bzw. v_h^+ und v_h^-) die Möglichkeit, Einzugsbereiche numerisch zu berechnen, indem man z.B. die Menge $\{x \in \mathbb{R}^d \,|\, v^+(x) < 1\}$ mit Hilfe der numerischen Approximation $\hat{v}_h^+$ annähert. Hierbei taucht aber ein praktisches Problem auf: Da $\hat{v}_h^+$ mit einem numerischen Fehler behaftet ist, macht die Unterscheidung zwischen $\hat{v}_h^+(x) = 1$ $\hat{v}_h^+(x) < 1$ keinen rechten Sinn. Um sicher zu gehen, dass ein Punkt wirklich im Einzugsbereich ist, muss man eine Menge der Form

$$\mathcal{D}_\varepsilon = \{x \in \mathbb{R}^d \,|\, \hat{v}_h^+(x) < 1-\varepsilon\}$$

betrachten, wobei $\varepsilon > 0$ ein Sicherheitsparameter ist, der in der Größenordnung des numerischen Fehlers liegt. Mit diesem Verfahren "verschenkt" man allerdings einen gewissen Teil des Einzugsbereiches, nämlich in etwa die Menge

$$\{x \in \mathbb{R}^d \,|\, v^+(x) < 1-\varepsilon\}. \tag{6.15}$$

Wenn man also an einer guten Approximation des Einzugsbereiches interessiert ist, so wäre es gut, wenn die Menge (6.15) möglichst klein wäre. Wir wollen uns daher zunächst überlegen, wie man diese Menge durch eine geeignete Wahl der zu integrierenden Funktion g möglichst klein machen kann.

Ideal wäre es natürlich, wenn die Funktion v^+ überhaupt keinen Übergangsbereich zwischen 0 und 1 hätte, wenn sie also auf ganz $\mathcal{D}^+$ exakt gleich Null wäre, womit (6.15) die leere Menge wäre. In diesem Fall hätten wir

$$v^+(x) = 1 - \chi_{\mathcal{D}^+}(x),$$

wobei $\chi_{\mathcal{D}^+}$ die *charakteristische Funktion* der Menge $\mathcal{D}^+$ ist, also

$$\chi_{\mathcal{D}^+}(x) = \begin{cases} 1, & x \in \mathcal{D}^+ \\ 0, & x \notin \mathcal{D}^+ \end{cases}$$

Da wir wissen, dass v^+ stetig ist, ist dies so nicht möglich, wir können aber zumindest $v^+(x) \approx 1 - \chi_{\mathcal{D}^+}(x)$ erzielen.

Um dies zu erreichen, erweitern wir unser optimales Steuerungsproblem um einen Parameter $\delta > 0$, indem wir das Funktional

$$J_\delta(x,u) = \int_0^\infty \delta g(\Phi(t,x,u),u(t))dt$$

und die zugehörigen optimalen Wertefunktionen

$$V_\delta^+(x) = \sup_{u\in\mathcal{U}} J_\delta(x,u) \quad \text{und} \quad V_\delta^+(x) = \inf_{u\in\mathcal{U}} J_\delta(x,u)$$

sowie die Transformationen

$$v_\delta^+(x) = 1 - e^{-V_\delta^+(x)} \quad \text{und} \quad v_\delta^-(x) = 1 - e^{-V_\delta^-(x)}$$

(und analog in diskreter Zeit) betrachten. Hier gilt das folgende Lemma.

Lemma 6.19 Für jede kompakte Teilmenge $K \subset \mathcal{D}^+$ gilt

$$\max_{x\in K} v_\delta^+(x) \to 0$$

für $\delta \to 0$. Die analogen Aussagen gelten für v^-, v_h^+ und v_h^-.

Beweis: Aus der Definition von J_δ folgt sofort die Gleichung

$$J_\delta(x,u) = \delta J(x,u).$$

Damit erhalten wir

$$V_\delta^+(x) = \delta V^+(x).$$

Da $K \subset \mathcal{D}^+$ ist, ist $V^+(x) < \infty$ für alle $x \in K$, da K zudem kompakt ist und V^+ stetig auf $\mathcal{D}^+$ ist, folgt sogar

$$\max_{x\in K} V^+(x) =: V_K < \infty.$$

Damit gilt

$$\begin{aligned}\max_{x\in K} v_\delta^+(x) &= \max_{x\in K} 1-e^{-V_\delta^+(x)} = \max_{x\in K} 1-e^{-\delta V^+(x)}\\ &= 1-e^{-\delta \max_{x\in K} V^+(x)} = 1-e^{-\delta V_K} \to 0\end{aligned}$$

für $\delta \to 0$. □

Korollar 6.20 Es gilt

$$v_\delta^+ \to 1-\chi_{\mathcal{D}^+} \quad \text{für } \delta\to 0$$

gleichmäßig auf kompakten Teilmengen $K\subset\mathbb{R}^d$ mit $K\cap\partial\mathcal{D}^+=\emptyset$.

Beweis: Wegen $K\cap\partial\mathcal{D}^+ = \mathit{emptyset}$ ist $K_1 := K\cap\mathcal{D}^+$ kompakt und nach Lemma 6.19 folgt

$$\max_{x\in K_1} v_\delta^+(x) \to 0 = 1-\chi_{\mathcal{D}^+}\Big|_{K_1}$$

für $\delta\to 0$. Wegen $v_\delta^+(x)\geq 0$ folgt also die gleichmäßige Konvergenz auf K_1. Auf $K_2 = K\setminus\mathcal{D}^+$ gilt für alle $\delta>0$

$$v_\delta^+\Big|_{K_2} \equiv 1 \equiv 1-\chi_{\mathcal{D}^+}\Big|_{K_2},$$

weswegen die gleichmäßige Konvergenz auf ganz K folgt. □

Für unsere numerische Approximation müssen wir die Regularisierung $g \rightsquigarrow g_\rho$ durchführen und dabei nun betrachten, wie der Parameter δ hier eingebunden wird. Im Prinzip gibt es hier zwei Möglichkeiten der Regularisierung, nämlich

$$\delta g \rightsquigarrow \max\{\delta g,\rho\} \quad \text{oder} \quad \delta g \rightsquigarrow \delta\max\{g,\rho\}$$

Die erste Möglichkeit scheidet aus, da wir im Konvergenzresultat Satz 6.14(iii) die Bedingung

$$g_\rho(x,u) = g(x,u)$$

für alle $x\notin\mathcal{N}(0)$ benötigen, die im ersten Fall für festes ρ und $\delta\to 0$ verletzt wird. Es bleibt also die zweite Möglichkeit, für die die gleichmäßige Konvergenz aus Korollar 6.20 für hinreichend kleines ρ erhalten bleibt (dies folgt sofort aus Satz 6.14(i) und (iii) und Korollar 6.20).

Trotzdem ist auch diese Bedingung nicht ganz zufriedenstellend, da wir ja in Bemerkung 6.13 festgestellt haben, dass die Größe ρ die Rolle des δ in der Diskretisierungsfehleranalyse in Kapitel 3 ist. Mit der Regularisierung

$$\delta g \rightsquigarrow \delta\max\{g,\rho\} = \delta g_\rho$$

spielt nun aber das Produkt $\delta\rho$ die Rolle von δ, d.h., je kleiner δ wird, desto größer wird unsere Abschätzung für den Diskretisierungsfehler. Der positive Effekt, den wir uns in der numerischen Berechnung von $\mathcal{D}^+$ durch die Wahl eines kleinen δ erhoffen würden, würde damit durch einen größeren Diskretisierungsfehler wieder zunichte gemacht.

In der Praxis liefert die numerische Approximation für feste Diskretisierungsparameter und $\delta\to 0$ trotzdem gute Ergebnisse, wie das folgende Beispiel illustriert.

Beispiel 6.21 Wir betrachten die (unkontrollierte) Differentialgleichung

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

mit $x \in \mathbb{R}^2$. Diese Gleichung hat ein asymptotisch stabiles Gleichgewicht in $x^* = 0$, dessen Einzugsbereich man aus dem Phasenportrait in Abbildung 6.1(links) leicht als $\mathcal{D} = \{x \in \mathbb{R}^d \,|\, \|x\| < 1\}$ identifiziert.

Abbildung 6.1(rechts) zeigt die Lösung der Zubov–Gleichung mit $g(x) = \|x\|$, $\delta = 0.001$, $\rho = 0.1$, $h = 0.1$ und $k = 0.0177$. Obwohl die Abschätzung des Diskretisierungsfehlers für diese Parameter unbrauchbar große Schranken liefert, ist die numerische Lösung brauchbar, da sie "optisch" nicht von der charakteristischen Funktion $1 - \chi_{\mathcal{D}}$ zu unterscheiden ist.

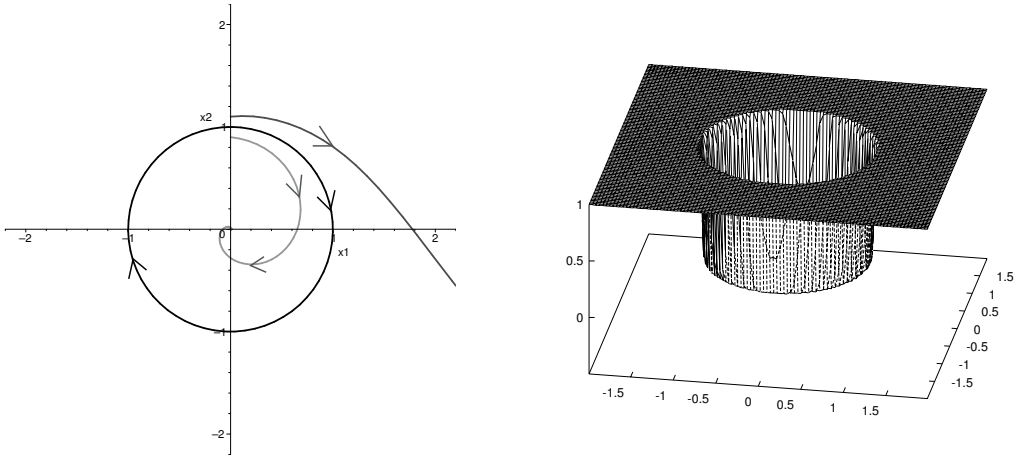

Abbildung 6.1: Phasenportrait und optimale Wertefunktion für Beispiel 6.21

□

Im Rest dieses Abschnitts wollen wir untersuchen, warum die numerische Approximation so gut funktioniert, obwohl die theoretischen Fehlerschranken schlechte Werte liefern. Heuristisch kann man dies wie folgt begründen: In einer kleinen Umgebung des Randes $\partial\mathcal{D}^+$, in dem die exakte Lösung sehr steil ist, wird man tatsächlich punktweise einen großen Fehler erhalten, d.h., hier ist die Fehlerabschätzung, die ja den Fehler in der ∞–Norm misst, durchaus nicht zu grob. Die guten numerischen Ergebnisse erklären sich dadurch, dass die numerische Approximation außerhalb dieses steilen Bereiches exakt ist, in dem Sinne, dass sie dort die gleichen Konvergenzeigenschaften wie die exakte Funktion (vgl. Korollar 6.20) erfüllt.

Wir beschränken uns dabei wie im letzten Abschnitt auf zeitdiskrete Systeme, wodurch die Analyse technisch etwas einfacher wird. Wir formulieren unsere Resultate für das Maximierungsproblem, sie gelten aber analog für das Minimierungsproblem.

Tatsächlich stellt sich bei genauerer Untersuchung heraus, dass wir in dem obigen numerischen Beispiel ein wenig "Glück" gehabt haben, denn im Allgemeinen kann die Approximation der charakteristischen Funktion tatsächlich sehr schlecht werden. Bisher haben

wir in unseren Fehlerabschätzungen immer die Kontraktionseigenschaft verwendet. Da die Kontraktionskonstante bei Verwendung der Funktion δg_ρ nicht mehr gleichmäßig in δ ist, erlaubt dies keine Aussage für $\delta \to 0$.

Um trotzdem zu garantieren, dass die Approximation funktioniert, müssen wir g vor der Regularisierung zunächst noch modifizieren. Wir verwenden dazu die Funktion

$$g_\lambda(x,u) = \max\{g(x,u) - \lambda, 0\}$$

sowie ihre Regularisierung

$$g_{\lambda,\rho}(x,u) = \max\{g_\lambda(x,u), \rho\}.$$

Durch geeignete Modifikation des Beweises von Satz 6.14(iii) man beweisen, dass für hinreichend kleine λ und ρ auch die durch

$$\tilde{v}^+_{\delta,h}(x) = \max_{u\in U}\{h\delta g_\lambda(x,u) + (1 - h\delta g_{\lambda,\rho}(x,u))\tilde{v}^+_{\delta,h}(f_h(x,u))\}$$

gegebene Funktion $\tilde{v}^+_{\delta,h}$ für jedes $\delta > 0$ den Einzugsbereich $\mathcal{D}^+$ mittels $\tilde{v}^+_{\delta,h}(x) < 1$ charakterisiert. Zudem gilt sich auch hier die Konvergenz

$$\tilde{v}_{\delta,h} \to 1 - \chi_{\mathcal{D}^+_h}$$

für $\delta \to 0$. Diese Eigenschaften gelten, falls die Bedingung

$$g_{\lambda,\rho}(x,u) = g_\lambda(x,u) \text{ für alle } x \in \mathcal{N}(0) \text{ und alle } u \in U \tag{6.16}$$

gilt.

Im Gegensatz zu der bisher betrachteten Funktion ist $\tilde{v}^+_{\delta,h}$ nun aber nicht mehr strikt positiv außerhalb von $x^* = 0$, statt dessen gilt $\tilde{v}^+_{\delta,h} = 0$ in einer Umgebung von $x^* = 0$.

Für diese Funktion wollen wir nun den folgenden Sachverhalt untersuchen: Wir betrachten die Funktion $\tilde{v}_{\delta,h}$ mit festem ρ und $\lambda > 0$, die so klein gewählt sind, dass (6.16) gilt. Für ein Gitter Γ und die Approximation von $\tilde{v}_{\delta,h}$ durch eine Gitterfunktion $\hat{v}^+_{\delta,h}$ betrachten wir dann den Grenzübergang

$$\lim_{\delta\to 0} \hat{v}^+_{\delta,h},$$

in der Hoffnung, dass wir beweisen können, dass die Grenzfunktion die charakteristische Funktion $1-\chi_{\mathcal{D}^+}$ in einer geeigneten Weise approximiert, so wie das in dem obigen Beispiel offenbar der Fall ist.

Unser erstes Lemma zeigt, dass $\hat{v}^+_{\delta,h}$ in einer geeigneten Umgebung von $x^* = 0$ für hinreichend feines Gitter Γ konstant gleich Null ist.

Lemma 6.22 Für alle $\rho, \lambda > 0$ existiert eine von $\delta > 0$ unabhängige Umgebung $\mathcal{N}_1(0)$ um $x^* = 0$, so dass für alle hinreichend feinen Gitter Γ, alle $x \in \mathcal{N}_1(0)$ und alle $\delta > 0$ die Gleichung

$$\hat{v}^+_{\delta,h}(x) = 0$$

gilt.

Beweis: Als Hilfsfunktion für den Beweis verwenden wir die optimale Wertefunktion v_h^+ des Zubov Problems für $\rho = \lambda = 0$ und $\delta = 1$. Wir betrachten die Menge

$$B := \{x \in \mathbb{R}^d \,|\, g_\lambda(x,u) = 0 \text{ für alle } u \in U\}$$

und wählen ein $\varepsilon > 0$ mit

$$N := \{x \in \mathbb{R}^d \,|\, v_h^+(x) < \varepsilon\} \subset B.$$

Da die Wertefunktion v_h^+ entlang von Lösungen abnimmt, existiert ein $\eta > 0$, so dass die Ungleichung

$$v_h^+(f_h(x,u)) < \varepsilon - \eta$$

für alle $x \in \mathcal{N}_1(0)$ und alle $u \in U$ gilt. Da v_h^+ auf der kompakten Menge B_1 gleichmäßig stetig ist, existiert ein $\gamma > 0$, so dass die Ungleichung

$$\|x - y\| \le \gamma \;\;\Rightarrow\;\; |v_h^+(x) - v_h^+(y)| \le \eta$$

für alle $x, y \in B_1$ gilt. Wir wählen nun das Gitter Γ so fein, dass der Durchmesser aller Elemente R_i mit $R_i \cap B \neq \emptyset$ die Ungleichung $\operatorname{diam}(R_i) \le \gamma$ erfüllt und behaupten, dass mit dieser Wahl des Gitters und dieser Umgebung $\mathcal{N}_1(0) = \bigcup_{R_i \subset N} R_i$ die Behauptung erfüllt ist (wobei wir annehmen, dass das Gitter so fein ist, dass $\mathcal{N}_1(0)$ tatsächlich eine Umgebung der 0 ist). Dazu genügt es zu zeigen, dass $\hat{v}_{\delta,h}^+(E_j) = 0$ ist für alle Gitterknoten $E_j \in N$.

Aus der Definition der Gitterfunktion $\hat{v}_{\delta,h}^+$ folgt für jeden Knoten E_j des Gitters die Gleichung

$$\hat{v}_{\delta,h}^+(E_j) = \max_{u \in U}\{h\delta g_\lambda(E_j,u) + (1 - h\delta g_{\lambda,\rho}(E_j,u)) \sum_{l=0}^{3} \mu_l(f_h(E_j,u))\hat{v}_{\delta,h}^+(E_{j_l})\},$$

wobei die E_{j_l} gerade die Eckpunkte eines Elements R_i sind, in denen $f_h(E_j,u)$ liegt. Wir betrachten nun die Knoten $E_j \in N$. Für diese gilt $v_h^+(f_h(E_j,u)) < \varepsilon - \eta$, und da die Eckpunkte E_{j_l} den Abstand $\le \operatorname{diam}(R_i) \le \gamma$ von $f_h(E_j,u)$ haben, folgt

$$v_h^+(E_{j_l}) < \varepsilon \;\;\Rightarrow\;\; E_{j_l} \in N.$$

Wenn $E_j \in N$ liegt, liegen also auch alle auftretenden Eckpunkte $E_{j_l} \in N$.

Für $E_j \in N$ gilt zudem

$$g_\lambda(E_j,u) = 0 \quad \text{und} \quad g_{\lambda,\rho}(E_j,u) = \rho.$$

Für jeden Knoten $E_j \in N$ erhalten wir damit

$$\hat{v}_{\delta,h}^+(E_j) = \max_{u \in U}\{(1 - h\delta\rho) \sum_{l=0}^{3} \mu_l(f_h(E_j,u))\hat{v}_{\delta,h}^+(E_{j_l})\} \le \max_{E_l \in N}\{(1 - h\delta\rho)\hat{v}_{\delta,h}^+(E_l)\},$$

wobei wir in der letzten Ungleichung $E_{j_l} \in N$ und $\sum_{l=0}^{3} \mu_l(f_h(E_j,u)) = 1$ ausgenutzt haben. Sei nun E_l^* ein Knoten, in dem das Maximum angenommen wird. Dann folgt

$$\hat{v}_{\delta,h}^+(E_l^*) \le (1 - h\delta\rho)\hat{v}_{\delta,h}^+(E_l^*)$$

und damit $\hat{v}_{\delta,h}^+(E_l^*) = 0$, woraus die Behauptung folgt. □

Das nächste Lemma zeigt, dass $\hat{v}_{\delta,h} \to 0$ konvergiert auf kompakten Teilmengen von $\mathcal{D}_h^+$.

Lemma 6.23 Es sei seien λ, $\rho > 0$ so gegeben, dass die Konvergenz (6.16) gilt. Dann gilt für jede kompakte Menge $K \subset \mathcal{D}_h^+$ und jedes hinreichend feine Gitter (wobei die benötigte Feinheit von der Wahl von K abhängt) die Konvergenz

$$\max_{x \in K} \hat{v}_{\delta,h}(x) \to 0$$

für $\delta \to 0$.

Beweis: Wir verwenden wiederum die Funktion v_h^+ als Hilfsfunktion. Da $K \subset \mathcal{D}_h^+$ ist, folgt $\sup_{x \in K} v_h^+(x) =: \sigma < 1$. Wir betrachten die Menge

$$K_1 = \{x \in \mathbb{R}^d \,|\, v_h^+(x) \leq \sigma\},$$

in der die Menge K enthalten ist, sowie die Menge $\mathcal{N}_1(0)$ aus dem vorhergehenden Lemma. Da v_h^+ entlang von Lösungen abnimmt, gibt es ein $\eta > 0$, so dass die Ungleichungen

$$v_h^+(f_h(x,u)) \leq v_h^+(x) - \eta.$$

Wegen der gleichmäßigen Stetigkeit von v_h^+ existiert ein $\gamma > 0$ mit

$$\|x - y\| \leq \gamma \;\; \Rightarrow \;\; |v_h^+(x) - v_h^+(y)| \leq \eta/2$$

für alle $x, y \in K_1$. Zudem gilt nach der Konstruktion von $\mathcal{N}_1(0)$ im Beweis des vorhergehenden Lemmas für alle hinreichend feinen Gitter die Inklusion

$$f_h(E_j, u) \in R_i \subset \mathcal{N}_1(0)$$

für alle Gitterknoten E_j Wir wählen nun das Gitter so fein, dass diese Inklusion gilt und zudem $\operatorname{diam}(R_i) \leq \gamma$ gilt für alle Elemente R_i, die K_1 schneiden.

Wir betrachten nun für $m \in \mathbb{N}_0$ die durch

$$X_m := \{E_j \in K_1 \,|\, v_h^+(E_j) \leq \sigma - m\eta/2\} \cup \{E_j \in \mathcal{N}_1(0)\}$$

gegebenen Mengen von Gitterpunkten. Nach der obigen Konstruktion und der Wahl der Gitterfeinheit gilt für $E_j \in X_m$ und die Eckpunkte E_{j_l} jedes Rechtecke R_i mit $f_h(E_j, u) \in R_i$ die Inklusion $E_{j_l} \in X_{m+1}$. Zudem existiert ein $m^* > 0$ mit $X_m \subset \mathcal{N}_1(0)$ für alle $m \geq m^*$.

Aus der Definition der Gitterfunktion $\hat{v}_{\delta,h}^+$ folgt für jeden Knoten E_j des Gitters die Gleichung

$$\hat{v}_{\delta,h}^+(E_j) = \max_{u \in U}\{h\delta g_\lambda(E_j, u) + (1 - h\delta g_{\lambda,\rho}(E_j, u)) \sum_{l=0}^{3} \mu_l(f_h(E_j, u))\hat{v}_{\delta,h}^+(E_{j_l})\},$$

aus der wir für $m < m^*$ die Ungleichung

$$\max_{E_j \in X_m} \hat{v}_{\delta,h}^+(E_j) \leq h\delta M_g + \max_{E_j \in X_{m+1}} \hat{v}_{\delta,h}^+(E_j)$$

erhalten. Für $m = m^*$ gilt

$$\max_{E_j \in X_{m^*}} \hat{v}_{\delta,h}^+(E_j) = 0,$$

weswegen wir per "Rückwärtsinduktion" von $m^* - 1$ nach 0 die Abschätzung

$$\max_{E_j \in X_m} \hat{v}^+_{\delta,h}(E_j) \leq (m^* - m)h\delta M_g$$

für alle $m \in \mathbb{N}_0$ erhalten. Damit folgt die Behauptung, da mit den Knotenwerten $\hat{v}^+_{\delta,h}$ auf K_1 auch die interpolierten Werte zwischen den Knoten gleichmäßig nach 0 konvergieren. □

Um die gewünschte "approximative" Konvergenz gegen die charakteristische Funktion $1 - \chi_{\mathcal{D}^+_h}$ zu beweisen, bleibt noch zu zeigen, dass $\hat{v}^+_{\delta,h}$ außerhalb von $\mathcal{D}^+_h$ gegen 1 konvergiert. Dies ist i.A. allerdings nicht der Fall. Ein einfaches Beispiel dafür ist ein System, bei dem alle Lösungen $\Phi_h(t, x)$ mit Ausnahme eines einzigen Anfangswertes $x^u \in \mathbb{R}^d$ gegen 0 konvergieren. Wenn dieser Anfangswert nicht zufällig mit einem Knoten E_i der Diskretisierung zusammenfällt, so wird hier numerisch nie der Wert 1 erreicht.

Im Folgenden führen wir daher eine geeignete Robustheitsbedingung ein, unter der wie die Konvergenz beweisen können. Wir betrachten dazu eine Abbildung

$$s_h : \mathbb{R}^d \times U \to \mathbb{R}^d$$

und das gestörte Kontrollsystem

$$x(t + h) = f_h(x(t), u(t)) + s_h(x(t), u(t)). \tag{6.17}$$

Wir bezeichnen den Einzugsbereich des gestörten Systems mit $\mathcal{D}^+_{h,s_h}$ und machen die folgende Annahme.

$$\begin{aligned} &\text{Für jedes } \varepsilon > 0 \text{ existiert ein } \delta > 0, \text{ so dass für alle Funktionen} \\ &s_h : \mathbb{R}^d \times U \to \mathbb{R}^d \text{ mit } \|s_h\|_\infty \leq \delta \text{ die Inklusion } \mathcal{D}^+_{h,s_h} \subseteq B_\varepsilon(\mathcal{D}^+_h) \text{ gilt.} \end{aligned} \tag{6.18}$$

Man kann leicht Kontrollsysteme angeben, für die die Annahme (6.18) nicht erfüllt ist, in den meisten praktischen Fällen gilt sie jedoch. Sie ist z.B. erfüllt, wenn der Einzugsbereich "abstoßend" ist, in dem Sinne, dass alle Lösungen aus einer Umgebung von $\mathcal{D}^+$ diese nach einer gewissen Zeit verlassen.

Unter dieser Bedingung kann nun der folgende Satz bewiesen werden.

Satz 6.24 Es sei ein Kontrollsystem mit Einzugsbereich $\mathcal{D}^+$ gegeben, der die Annahme 6.18 erfüllt. Weiterhin seien λ, $\rho > 0$ so gegeben, dass die Konvergenz (6.16) gilt. Dann existiert für jede kompakte Menge $K \subset \mathbb{R}^d$ mit $K \cap \partial\mathcal{D}^+ = \emptyset$ ein hinreichend feines Gitter Γ, so dass die Konvergenz

$$\hat{v}^+_{\delta,h}(x) \to 1 - \chi_{\mathcal{D}^+_h}(x)$$

gleichmäßig für $x \in K$ gilt.

Beweis: Da K kompakt ist, nimmt die stetige Abstandsfunktion $d(x, \partial\mathcal{D}^+)$ für $x \in K$ ein Minimum an, das größer als Null sein muss. Also existiert ein $\varepsilon > 0$, so dass sogar

$$K \cap B_\varepsilon(\partial\mathcal{D}^+) = \emptyset$$

gilt. Wir können die Menge K also in zwei Mengen K_1 und K_2 zerlegen, so dass $B_\varepsilon(K_1) \subseteq \mathcal{D}^+$ und $K_2 \cap B_\varepsilon(\mathcal{D}^+) = \emptyset$ gilt. Auf K_1 gilt die behauptete Konvergenz dann auf Grund von Lemma 6.23. Es bleibt zu zeigen, dass $\hat{v}^+_{\delta,h}$ auf K_2 gegen 1 konvergiert, wofür wir zeigen werden, dass $\hat{v}_{\delta,h} \equiv 1$ auf K_2 ist. Dazu wählen wir das Gitter so fein, dass $\operatorname{diam}(R_i) \leq \max\{\delta, \varepsilon/2\}$ gilt für das zu $\varepsilon/2$ gehörige δ aus Annahme 6.18. Damit ist sicher gestellt, dass jedes Rechteck R_i, das K_2 schneidet, die Menge $B_{\varepsilon/2}(\mathcal{D}^+)$ nicht schneidet. Für einen beliebigen Eckpunkt E_j eines solchen Rechtecks gilt dann die bereits mehrfach verwendete Rekursion

$$\hat{v}^+_{\delta,h}(E_j) = \max_{u \in U}\{h\delta g_\lambda(E_j,u) + (1 - h\delta g_{\lambda,\rho}(E_j,u)) \sum_{l=0}^{3} \mu_l(f_h(E_j,u))\hat{v}^+_{\delta,h}(E_{j_l})\}.$$

Für jedes $u \in U$ wählen wir nun denjenigen Eckpunkt $E^*_{j_l}(u)$ aus, für den $\hat{v}^+_{\delta,h}(E^*_{j_l}(u))$ minimal wird. Dann finden wir eine Abbildung s_h, so dass $f_h(E_j,u)+s_h(E_j,u) = E^*_{j_l}(u)$ ist. Da diese Abbildung nach Voraussetzung über die Rechteckgröße die Bedingung $\|s_h\|_\infty \leq \delta$ erfüllt, gilt $E_j \notin \mathcal{D}^+_{h,s_h}$ und es existiert ein $u^* \in U$, so dass $f_h(E_j,u^*) + s_h(E_j,u^*) \notin \mathcal{D}^+_{h,s_h}$ gilt. Unter Verwendung von (6.16) gilt dann

$$\hat{v}^+_{\delta,h}(E_j) \geq h\delta g_\lambda(E_j,u^*) + (1 - h\delta g_\lambda(E_j,u^*))\hat{v}^+_{\delta,h}(E^*_{j_l}(u^*)).$$

Bezeichnen wir mit E^*_j denjenigen Knoten, für den $\hat{v}^+_{\delta,h}(E_j)$ unter allen Knoten $E_j \notin \mathcal{D}^+_{h,s_h}$ minimal wird, so erhalten wir

$$\hat{v}^+_{\delta,h}(E^*_j) \geq h\delta g_\lambda(E^*_j,u^*) + (1 - h\delta g_\lambda(E^*_j,u^*))\hat{v}^+_{\delta,h}(E^*_j)$$

und damit

$$\hat{v}^+_{\delta,h}(E^*_j) - (1 - h\delta g_\lambda(E^*_j,u^*))\hat{v}^+_{\delta,h}(E^*_j) \geq h\delta g_\lambda(E^*_j,u^*) \;\Leftrightarrow\; \hat{v}^+_{\delta,h}(E^*_j) = 1.$$

Da die Ungleichung $\hat{v}^+_{\delta,h} \leq 1$ direkt aus der Iterationsvorschrift folgt, erhalten wir damit

$$\hat{v}^+_{\delta,h}(E_j) = 1$$

für alle Knoten $E_j \notin \mathcal{D}^+_{h,s_h}$, damit $\hat{v}^+_{\delta,h} \equiv 1$ auf allen Rechtecken, die K_2 schneiden, also die Behauptung. □

Literaturverzeichnis

[1] B. Aulbach, *Gewöhnliche Differentialgleichungen*, Spektrum Verlag, Heidelberg, 1997.

[2] M. Bardi and I. Capuzzo Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman equations*, Birkhäuser, Boston, 1997.

[3] M. G. Crandall, L. C. Evans, and P. L. Lions, *Some properties of viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.

[4] M. G. Crandall and P. L. Lions, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[5] M. Falcone and T. Giorgi, *An approximation scheme for evolutive Hamilton-Jacobi equations*, in Stochastic analysis, control, optimization and applications, W. M. McEneaney, G. Yin, and Q. Zhang, eds., Birkhäuser, Boston, 1999, pp. 288–303.

[6] R. Ferretti, *High-order approximations of linear control systems via Runge-Kutta schemes*, Computing, 58 (1997), pp. 351–364.

[7] R. L. V. González and C. A. Sagastizábal, *Un algorithme pour la résolution rapide d'équations discrètes de Hamilton-Jacobi-Bellman*, C. R. Acad. Sci., Paris, Sér. I, 311 (1990), pp. 45–50.

[8] R. L. V. González and M. M. Tidball, *On a discrete time approximation of the Hamilton-Jacobi equation of dynamic programming.* INRIA Rapports de Recherche Nr. 1375, 1991.

[9] L. Grüne, *Numerische optimale Steuerung und Stabilisierung.* Diplomarbeit, Institut für Mathematik, Universität Augsburg, 1994.

[10] ——, *An adaptive grid scheme for the discrete Hamilton-Jacobi-Bellman equation*, Numer. Math., 75 (1997), pp. 319–337.

[11] L. Grüne and P. E. Kloeden, *Higher order numerical schemes for affinely controlled nonlinear systems*, Numer. Math., 89 (2001), pp. 669–690.

[12] L. Grüne and F. Wirth, *Computing control Lyapunov functions via a Zubov type algorithm*, in Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, Australia, 2000, pp. 2129–2134.

[13] W. Hahn, *Theorie und Anwendung der direkten Methode von Ljapunov*, Ergebnisse der Mathematik und ihrer Grenzgebiete 22, Springer–Verlag Berlin, Göttingen, Heidelberg, 1959.

[14] ——, *Stability of Motion*, Springer–Verlag Berlin, Heidelberg, 1967.

[15] J. L. Haunschmied, P. M. Kort, R. F. Hartl, and G. Feichtinger, *A DNS–curve in a two state capital accumulation model: a numerical analysis*, Journal of Economic Dynamics & Control, (2003), pp. 701–716.

[16] P. L. Lions, *Generalized solutions of Hamilton-Jacobi equations*, Pitman, London, 1982.

[17] M. L. Puterman and S. Brumelle, *On the convergence of policy iteration in stationary dynamic programming*, Math. of Operations Research, 4 (1979).

[18] A. Seeck, *Iterative Lösungen der Hamilton-Jacobi-Bellman-Gleichung bei unendlichem Zeithorizont*. Diplomarbeit, Universität Kiel, 1997.

[19] W. Semmler and M. Sieveking, *On optimal exploitation of interacting resources*, Journal of Economics, 59 (1994), pp. 23–49.

[20] E. D. Sontag, *Mathematical Control Theory*, Springer Verlag, New York, 2nd ed., 1998.

Index