Numerik optimaler Steuerung

Lars Grüne
Fachbereich Mathematik
Johann Wolfgang Goethe-Universität
Postfach 111932
60054 Frankfurt am Main, Germany
gruene@math.uni-frankfurt.de
www.math.uni-frankfurt.de/∼gruene

Vorlesungsskript

Sommersemester 2001

Vorwort

Dieses Skript ist im Rahmen einer Vorlesung entstanden, die ich im Sommersemester 2001 am Fachbereich Mathematik der J.W. Goethe–Universität Frankfurt gehalten habe. Ich möchte mich an dieser Stelle bei allen Teilnehmerinnen und Teilnehmern dieser Vorlesung für ihre Anregungen bedanken, die zur Verbesserung dieses Textes beigetragen haben.

Eine elektronische Version dieses Skripts sowie die zu dieser Vorlesung gehörigen Übungsaufgaben und ihre Lösungen als MAPLE–Worksheets finden sich im WWW unter der Adresse `http://www.math.uni-frankfurt.de/~gruene/teaching/optctrl01/`.

Frankfurt am Main, August 2001 LARS GRÜNE

Inhaltsverzeichnis

Kapitel 1

Kontrollsysteme

In diesem Kapitel wollen wir die grundlegenden Systeme definieren, mit denen wir uns in dieser Vorlesung beschäftigen wollen. Der Ausdruck „Kontrollsystem“ 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.

1.1 Definition

Definition 1.1 Ein *Kontrollsystem* 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.

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

Mit dem Symbol $\mathcal{U}$ bezeichnen wir den *Raum der zulässigen Kontrollfunktionen*, also

$$\mathcal{U} := \{u : \mathbb{R} \to U \,|\, u \text{ zulässig}\}.$$

Was „zulässig“ im mathematischen Sinne bedeutet, werden wir im folgenden Abschnitt genauer festlegen. □

Einige Beispiele für Kontrollsysteme werden wir in Kapitel 2 kennen lernen.

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

Wir werden uns nun damit beschäftigen, welche Wahl des Kontrollfunktionsraumes $\mathcal{U}$ sinnvoll ist. Zwei Kriterien spielen dabei 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) erhalten.

1.2 Ein Existenz– und Eindeutigkeitssatz

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 für Anwendungen in der optimalen Steuerung allerdings eine zu einschränkende Klasse, da man bereits für sehr einfache Probleme nachweisen kann, dass optimale Steuerstrategien unstetig in t sind. Wir werden deshalb eine größere Klasse von Kontrollfunktionen zulassen. Wir erinnern an die folgende Definition.

Definition 1.3 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[1] $x \in I$.

(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 1.4 (Satz von Caratheodory) Betrachte ein Kontrollsystem mit folgenden Eigenschaften:

i) Die Menge $U \subset \mathbb{R}^m$ ist kompakt.

ii) Der Raum der Kontrollfunktionen ist gegeben durch

$$\mathcal{U} := \{u : \mathbb{R} \to U \mid u \text{ ist messbar } \}.$$

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

iv) Es existiert eine Konstante $L > 0$, so dass die Abschätzung

$$\|f(x_1, u) - f(x_2, u)\| \leq L\|x_1 - x_2\|$$

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

Dann gibt es für jeden Punkt $x_0 \in \mathbb{R}^d$ und jede Kontrollfunktion $u \in \mathcal{U}$ 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 $t \geq 0$ erfüllt.

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

Definition 1.5 Wie bezeichnen die eindeutige Funktion $x(t)$ aus Satz 1.4 mit $\varphi(t, x_0, u)$ 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 $\varphi(t, x_0, u)$ absolut stetig ist, ist diese Funktion für fast alle $t \geq 0$ nach t differenzierbar. Insbesondere folgt also aus dem Satz 1.4, dass $\varphi(t, x_0, u)$ die Differentialgleichung (1.1) für fast alle $t \geq 0$ erfüllt, d.h. es gilt

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

für fast alle $t \geq 0$.

Bemerkung 1.6 Im Weiteren nehmen wir stets an, dass die Voraussetzungen (i)–(iv) von Satz 1.4 erfüllt sind, werden dies aber nur in wichtigen Sätzen explizit formulieren. □

Der Beweis von Satz 1.4 (auf den wir aus Zeitgründen nicht näher eingehen können) 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 (unter schwächeren Voraussetzungen als hier) zusammen mit einer Einführung in die zugrundeliegende Lebesgue–Maßtheorie z.B. im Buch von E.D. Sontag [16, Anhang C].

1.3 Ein einfaches numerisches Verfahren

Eine übliche Klasse numerischer Verfahren zur Lösung von gewöhnlichen Differentialgleichungen sind die sogenannten *Einschrittverfahren*. Hierbei wird, zu einer Schrittweite $h > 0$, die kontinuierliche Lösungsfunktion $\varphi(t, x_0, u)$ durch eine Folge von Punkten x_i zu den diskreten Zeiten hi, $i \in \mathbb{N}_0$ angenähert. Das Einschrittverfahren besteht dabei im Wesentlichen aus einer (mittels eines Computers) auswertbaren Abbildung, die beschreibt, wie x_{i+1} aus x_i berechnet werden kann.

Die naheliegende Idee ist nun, 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, für weitere Verfahren siehe z.B. die Arbeit [10] von L. Grüne und P. Kloeden.

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

Definition 1.7 Wir nennen ein Kontrollsystem (1.1) *konvex*, falls die Menge $f(x, U) := \{f(x, u),\, u \in U\} \subset \mathbb{R}^d$ für jedes $x \in \mathbb{R}^d$ konvex ist. □

Wir definieren nun ein numerisches Einschrittverfahren.

Definition 1.8 (Euler–Verfahren für Kontrollsysteme) Für einen Zeitschritt $h > 0$ und einen Kontrollwert $u \in U$ definiere die Abbildung

$$\Phi_h(x, u) := x + hf(x, u)$$

und zu einem Anfangswert $x_0 \in \mathbb{R}^d$ und einer Folge von Kontrollwerten $\mathbf{u} = (u_i)_{i\in\mathbb{N}_0}$ betrachte die Folge von Punkten $x_i(x_0, \mathbf{u})$, die rekursiv definiert ist durch $x_0(x_0, \mathbf{u}) = x_0$ und

$$x_{i+1}(x_0, \mathbf{u}) = \Phi_h(x_i(x_0, \mathbf{u}), u_i), \;\; i \geq 0.$$

□

Der folgende Satz fasst die Eigenschaften dieses Verfahrens zusammen.

Satz 1.9 Betrachte ein Kontrollsystem, für das die Voraussetzungen (i)–(iv) von Satz 1.4 gelten. Darüberhinaus erfülle das Vektorfeld f die Abschätzung

$$\|f(x, u)\| \leq M$$

für alle $x \in \mathbb{R}^d$, alle $u \in U$ und eine Konstante $M > 0$. Dann gilt für das Verfahren aus Definition 1.8:

(i) Es existiert eine Konstante $K > 0$, so dass für jede Kontrollfunktion $u \in \mathcal{U}$ und jeden Anfangswert $x_0 \in \mathbb{R}^d$ eine Folge von Kontrollwerten $\mathbf{u} = (u_i)_{i\in\mathbb{N}_0}$ existiert, mit der die Abschätzung

$$\|x_i(x_0, \mathbf{u}) - \varphi(t, x_0, u)\| \leq K\sqrt{h}e^{Lt}$$

für alle $t = ih$, $i \in \mathbb{N}_0$ erfüllt ist.

Ist das Kontrollsystem konvex, so gilt die schärfere Abschätzung

$$\|x_i(x_0, \mathbf{u}) - \varphi(t, x_0, u)\| \leq \frac{M}{2}h(e^{Lt} - 1) \quad \text{für alle } t = ih,\, i \in \mathbb{N}_0.$$

(ii) Umgekehrt gilt für jedes $x_0 \in \mathbb{R}^d$, jede Folge von Kontrollwerten $\mathbf{u} = (u_i)_{i\in\mathbb{N}}$ und die durch

$$u(t) := u_i, \;\; t \in [hi, h(i + 1)), \quad i \in \mathbb{N}_0$$

definierte stückweise konstante (also messbare) Kontrollfunktion die Abschätzung

$$\|x_i(x_0, \mathbf{u}) - \varphi(t, x_0, u)\| \leq \frac{M}{2}h(e^{Lt} - 1) \quad \text{für alle } t = ih,\, i \in \mathbb{N}_0.$$

Beweis: 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 in der Arbeit [7] von R. L. V. González and M. M. Tidball).

Wir betrachten zunächst zu jeder Kontrollfunktion $u \in \mathcal{U}$ 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 $f(x,U)$ folgt, dass $\bar{f}(x,u)$ in $f(x,U)$ liegt (tatsächlich ist die Konvexität von $f(x,U)$ hierfür auch notwendig). Daher gibt es Kontrollwerte

$$\bar{u}(x,u) \in U \text{ mit } f(x,\bar{u}(x,u)) = \bar{f}(x,u). \tag{1.2}$$

Für diese Werte zeigen wir zunächst die Abschätzung

$$\|\varphi(h,x,u) - \Phi_h(x,\bar{u}(x,u))\| \le \frac{M}{2}Lh^2 \tag{1.3}$$

Es gilt nämlich

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

mit dem Restterm

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

Also folgt

$$\|\varphi(h,x,u) - \Phi_h(x,\bar{u}(x,u))\| \le \|R(h)\|.$$

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

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

womit (1.3) gezeigt ist.

Aus der Reihendarstellung $e^{Lh} = 1 + Lh + L^2h^2/2 + \ldots$ folgt $e^{Lh} \ge 1 + Lh$, damit $Lh^2 \le h(e^{Lh} - 1)$ und folglich aus (1.3)

$$\|\varphi(h,x,u) - \Phi_h(x,\bar{u}(x,u))\| \le \frac{M}{2}h(e^{Lh} - 1). \tag{1.4}$$

Betrachte nun eine Kontrollfunktion $u \in \mathcal{U}$ und einen Anfangswert $x_0 \in \mathbb{R}^d$. Für jedes $i \in \mathbb{N}_0$ betrachte die Funktion $w_i \in \mathcal{U}$ gegeben durch $w_i(t) = u(hi + t)$. Es ist leicht zu überprüfen, dass für diese w_i die Identität

$$\varphi(h(i+1), x_0, u) = \varphi(h, \varphi(hi, x_0, u), w_i) \tag{1.5}$$

für alle $i \in \mathbb{N}_0$ gilt. Wir definieren die Folge $\mathbf{u} = (u_i)_{i\in\mathbb{N}_0}$ als

$$u_i = \bar{u}(\varphi(hi, x_0, u), w_i)$$

mit $\bar{u}$ aus (1.2).

Aus (1.4) und (1.5) folgt damit für alle $i \in \mathbb{N}_0$ die Abschätzung

$$\|\varphi(h(i+1), x_0, u) - \Phi_h(\varphi(hi, x_0, u), u_i)\| \leq \frac{M}{2} h(e^{Lh} - 1). \tag{1.6}$$

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

$$\|x_i(x_0, \mathbf{u}) - \varphi(hi, x_0, u)\| \leq h(e^{Lhi} - 1) \tag{1.7}$$

Aus der Annahme (iv) von Satz 1.4 und der obigen Reihendarstellung von e^{Lh} folgt die Lipschitz–Abschätzung

$$\|\Phi_h(x_1, u) - \Phi_h(x_2, u)\| \leq (1 + Lh)\|x_1 - x_2\| \leq e^{Lh}\|x_1 - x_2\| \tag{1.8}$$

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

$$\begin{aligned}
&\|x_{i+1}(x_0, \mathbf{u}) - \varphi(h(i+1), x_0, u)\| \\
&= \|\Phi_h(x_i(x_0, \mathbf{u}), u_i) - \varphi(h(i+1), x_0, u)\| \\
&\leq \|\Phi_h(x_i(x_0, \mathbf{u}), u_i) - \Phi_h(\varphi(hi, x_0, u), u_i)\| \\
&\quad + \|\Phi_h(\varphi(hi, x_0, u), u_i) - \varphi(h(i+1), x_0, u)\| \\
&\overset{(1.8)}{\leq} e^{Lh}\|x_i(x_0, \mathbf{u}) - \varphi(hi, x_0, u)\| + \|\Phi_h(\varphi(hi, x_0, u), u_i) - \varphi(h(i+1), x_0, u)\| \\
&\overset{(1.6)}{\leq} e^{Lh}\|x_i(x_0, \mathbf{u}) - \varphi(hi, x_0, u)\| + \frac{M}{2} h(e^{Lh} - 1) \\
&\overset{(1.7)}{\leq} e^{Lh}\frac{M}{2} h(e^{Lhi} - 1) + \frac{M}{2} h(e^{Lh} - 1) \\
&= \frac{M}{2} h(e^{Lh(i+1)} - e^{Lh} + e^{Lh} - 1) \;=\; \frac{M}{2} h(e^{Lh(i+1)} - 1)
\end{aligned}$$

und damit die gewünschte Abschätzung aus (i).

Zum Beweis von (ii) betrachte eine Folge $\mathbf{u} = (u_i)_{i\in\mathbb{N}_0}$ und die in (ii) konstruierte stückweise konstante Kontrollfunktion $u \in \mathcal{U}$. Wenn wir nun eine neue Folge $\tilde{\mathbf{u}} = (\tilde{u}_i)_{i\in\mathbb{N}_0}$ wie im

Beweis von (i) aus diesem u konstruieren, so gilt $\tilde{u}_i = u_i$ für alle $i \in \mathbb{N}_0$, d.h. wir erhalten gerade wieder die Folge $\mathbf{u}$, von der wir ausgegangen sind (beachte, dass für diese stückweise konstante Funktion u die Konstruktion aus (i) auch ohne die Konvexitätsbedingung funktioniert). Also folgt (ii) indem wir (i) auf dieses u anwenden. □

Bemerkung 1.10 Beachte, dass die Folge $\mathbf{u}$ aus Satz 1.9(i) implizit definiert ist und daher im Allgemeinen keine einfache Formel zu ihrer Berechnung angegeben werden kann.

Tatsächlich ist die explizite Kenntnis von $\mathbf{u}$ zur Lösung von optimalen Steuerungsproblemen aber gar nicht notwendig, wie wir in den nächsten Kapiteln sehen werden. □

Bemerkung 1.11 Wir werden später sehen, dass es praktisch ist, die Werte u_i der Kontrollwertfolge $\mathbf{u}$ aus einer endlichen Menge $\widetilde{U} \subset U$ zu wählen. Wenn $\widetilde{U}$ hinreichend „dicht“ in U liegt ist, lässt sich ein ähnliches Resultat wie in Satz 1.9(i) für Folgen $\mathbf{u}$ mit Werten $u_i \in \widetilde{U}$ beweisen, zumindest solange die zugehörigen Trajektorien eine kompakte Menge nicht verlassen. □

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 man diese Funktion entlang einer Trajektorie $\varphi(t, x_0, u)$ 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). 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* als

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

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

$$v(x) := \sup_{u \in \mathcal{U}} J(x, u).$$

Hierbei machen wir die folgenden Annahmen:

(A1) Das Kontrollsystem (1.1) erfülle die Voraussetzungen (i)–(iv) von Satz 1.4.

(A2) Die Funktion g sei stetig und erfülle

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

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

□

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

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

(iii) Über die Kenntnis von v hinaus ist es natürlich interessant, auch (zumindest näherungsweise) optimale Kontrollfunktionen u zu berechnen. Wir werden uns hier zunächst mit der Berechnung von v beschäftigen und dieses Problem danach behandeln.

(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. □

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 entfernte Werte von $g(\varphi(t,x,u),u(t))$ schwächer gewichtet werden). Wichtiger als diese ökonomische Interpretation sind für uns die mathematischen Auswirkungen des Diskontfaktors. Die offensichtlichsten sind in dem folgenden Lemma zusammengefasst.

Lemma 2.3 (i) Das diskontierte Funktional ist endlich. Genauer gilt

$$|J(x,u)| \le \frac{M_g}{\delta},$$

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

(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

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

Beweis: (i) Es gilt

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

[1] unter gewissen Voraussetzungen lässt sich die Existenz optimaler $u \in \mathcal{U}$ beweisen; dies näher auszuführen würde den Rahmen dieser Vorlesung aber sprengen

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

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

□

2.2 Beispiele

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

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

Wir setzen nun $a(t) = u_1(t) - (r + u_2(t))x_2(t)$ mit $u = (u_1,u_2)^T \in [-1,1] \times [0,1]$. Die Kontrolle u_1 modelliert hier die externe Beschleunigung, die z.B. durch einen Motor erzeugt werden kann, $r > 0$ ist ein Reibungsfaktor und u_2 modelliert eine Bremse.

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 schwächer) so weit beschleunigt, bis er, ungefähr zum Zeitpunkt $t = 2.5$, gerade so schnell ist, dass er durch die Reibung genau im Punkt $x_1 = 0$ stehen bleibt.

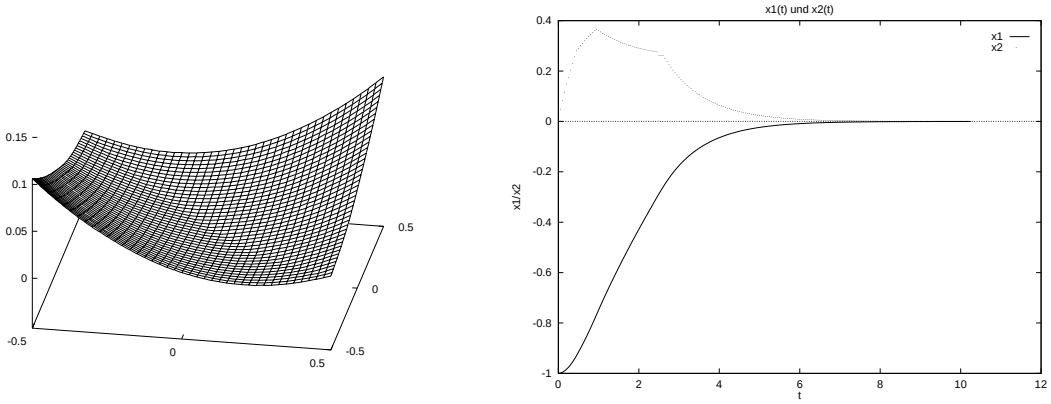


Abbildung 2.1: Wertefunktion und optimale Trajektorie für den gesteuerten Wagen

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

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

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. Die folgende Liste gibt die Parameterwerte sowie eine kurze Erklärung ihrer Bedeutung:

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 Beute	(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

$$g(x,u) = \frac{1}{1+x_1 u} x_1 u + \frac{1}{1+x_2 u} x_2 u - \frac{u}{2}$$

festgelegt. Sie berücksichtigt die Kosten des Fangs durch den Term $u/2$, sowie die Tatsache, dass bei erhöhter Stückzahl $x_1 u$ bzw. $x_2 u$ der erzielte Marktpreis pro Einheit sinkt.

Abbildung 2.2 zeigt die optimale Wertefunktion dieses Problems für Diskontrate $\delta = 5$. Abbildung 2.3 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. In Vergleich dazu zeigt Abbildung 2.4 die Trajektorie zu konstanter Fangrate $u \equiv 0.75$; hier liegt der Ertrag bei 0.29.

[2] dieses spezielle Modell stammt aus der Arbeit [15] von W. Semmler und M. Sieveking

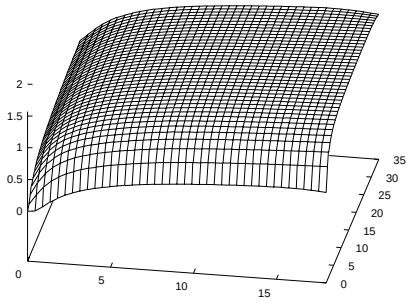


Abbildung 2.2: Wertefunktion für das Räuber–Beute Modell

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 2.4 dargestellte Lösung.

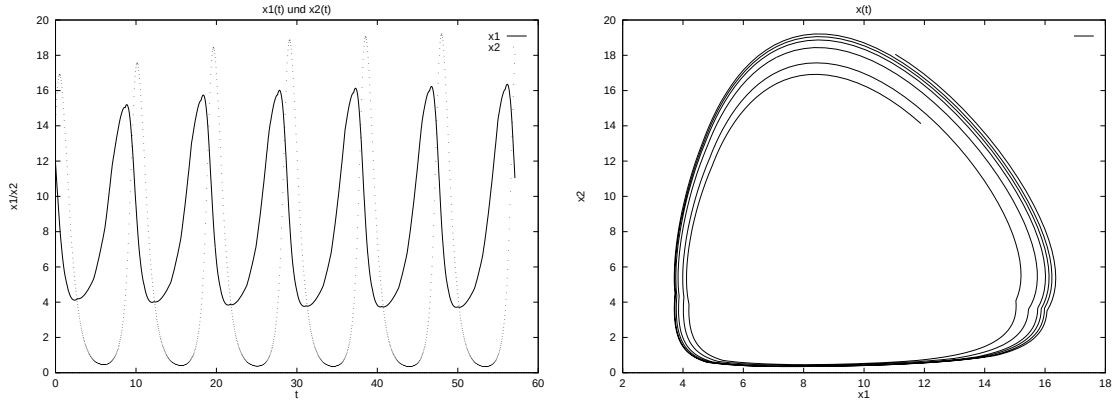


Abbildung 2.3: Optimale Trajektorie für das Räuber–Beute Modell

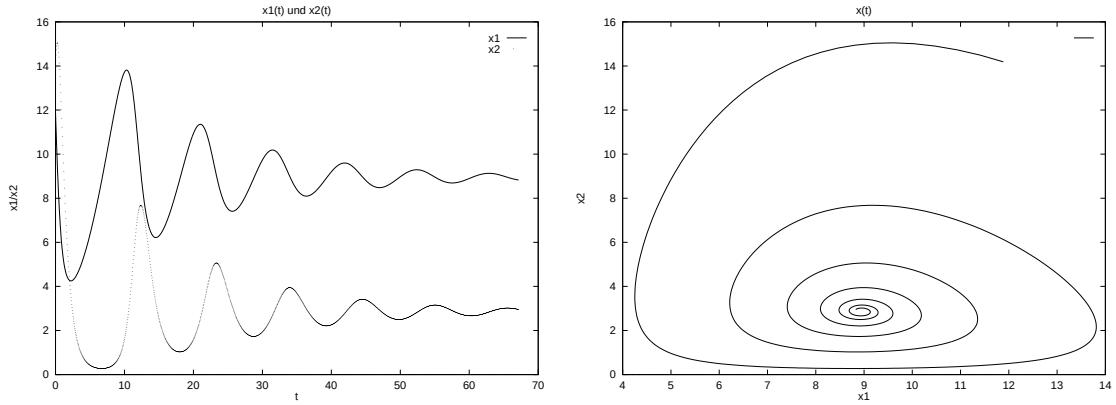


Abbildung 2.4: Trajektorie zu $u \equiv 0.75$ für das Räuber–Beute Modell

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 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: Wir zeigen die erste Ungleichung, die zweite folgt analog. Ohne Beschränkung der Allgemeinheit gelte

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

Wähle nun $\varepsilon > 0$ und $b_\varepsilon \in B$ so, dass

$$a_1(b_\varepsilon) > \sup_{b_1 \in B} a_1(b_1) - \varepsilon.$$

Klarerweise gilt dann $\sup_{b_2 \in B} a_2(b_2) \geq a_2(b_\varepsilon)$ und damit

$$\begin{aligned} \sup_{b_1 \in B} a_1(b_1) - \sup_{b_2 \in B} a_2(b_2) &< a_1(b_\varepsilon) - a_2(b_\varepsilon) + \varepsilon \\ &\leq |a_1(b_\varepsilon) - a_2(b_\varepsilon)| + \varepsilon \\ &\leq \sup_{b \in \mathcal{U}} |a_1(b) - a_2(b)| + \varepsilon \end{aligned}$$

Da $\varepsilon > 0$ beliebig war, folgt die Behauptung. □

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

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

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

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

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

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

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

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

für $L \neq \delta$, bzw.

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

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

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

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

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

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

Beweis: Aus der Lipschitz Stetigkeit des Vektorfeldes f folgt die Abschätzung

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

Mit Gronwalls Lemma[3] folgt daraus

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

[3]Dieses Lemma sollte aus der Einführung in die gewöhlichen Differentialgleichungen bekannt sein, siehe z.B. das Buch von B. Aulbach [1]

Damit ergibt sich

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

falls $\delta \neq L$ oder

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

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

$$|J(x,u) - J(y,u)| \leq K\|x-y\|^\gamma \tag{2.2}$$

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 (2.2) 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 sogennante *Bellman'sche Optimalitätsprinzip*[4], auch *Prinzip der Dynamischen Programmierung* genannt. Es besagt, 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. Eine andere Sichtweise dieses Prinzips ist, dass Endstücke optimaler Trajektorien wieder optimale Trajektorien sind. 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

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

[4]Benannt nach dem amerikanischen Mathematiker Richard E. Bellman, dem die „Entdeckung" dieses Prinzips zugeschrieben wird

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

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

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

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

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

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

Damit ergibt sich

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

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 $w : \mathbb{R}^d \to \mathbb{R}$ eine beschränkte Funktion, die für ein $T > 0$ das Optimalitätsprinzip

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

für alle $x \in \mathbb{R}^d$ erfüllt. Dann ist $w = v$.

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

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

und

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

die Ungleichung

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

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 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

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

gilt. Durch Umstellen der Terme erhält man

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

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

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

Mit einigen technischen Überlegungen, die wir hier nicht ausführen wollen, sieht man, dass das Infimum tatsächlich über konstante Kontrollfunktionen $u \equiv u_0 \in U$ genommen werden kann (was insbesondere die Verwendung der Ableitungen von φ und $\int e^{-\delta t} g(\cdots)\, dt$ rechtfertigt). Für konstante Kontrollen gilt aber

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

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. Crandal, L.C. Evans und P.L. Lions um 1980 eingeführt (siehe [3, 4, 12]), 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 in der Zeit

Das numerische Verfahren, mit dessen Herleitung und Analyse wir nun beginnen wollen, besteht aus zwei voneinander weitgehend unabhängigen Schritten. In diesem Kapitel werden wir den ersten Schritt betrachten, die Diskretisierung in der Zeit.

Aus Definition 1.8 und Satz 1.9 wissen wir, dass wir zu einem Kontrollsystem eine diskrete Approximation konstruieren können, die jedem Anfangswert $x_0 \in \mathbb{R}^d$ und jeder Kontrollfolge $\mathbf{u} = (u_i)_{i\in\mathbb{N}_0}$ eine Punktfolge $x_i(x_0, \mathbf{u})$ liefert. Wir definieren nun, analog zu Definition 2.1, das folgende zeitdiskrete optimale Steuerungsproblem. Hierzu bezeichnen wir den Raum der Kontrollfolgen mit $U^{\mathbb{N}_0}$.

Definition 3.1 Betrachte das optimale Steuerungsproblem aus Definition 2.1 mit Kontrollsystem (1.1) und der zugehörigen Euler–Diskretisierung aus Definition 1.8 mit Schrittweite $h > 0$. Wir definieren das *zeitdiskrete diskontierte Funktional auf unendlichem Zeithorizont* als

$$J_h(x, \mathbf{u}) := h \sum_{i=0}^{\infty} e^{-\delta hi} g(x_i(x, \mathbf{u}), u_i) \tag{3.1}$$

mit $\mathbf{u} = (u_i)_{i\in\mathbb{N}_0}$. Das zeitdiskrete optimale Steuerungsproblem lautet nun: Bestimme die *zeitdiskrete optimale Wertefunktion*

$$v_h(x) := \sup_{\mathbf{u}\in U^{\mathbb{N}_0}} J_h(x, \mathbf{u}).$$

□

3.1 Diskretisierungsfehler

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 3.2 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 v_h.

Satz 3.3 Betrachte ein optimales Steuerungsproblem aus Definition 2.1 sowie das zugehörige zeitdiskrete optimale Steuerungsproblem aus Definition 3.1. Wir nehmen an, dass das zugrundeliegende Kontrollsystem die Voraussetzungen von Satz 1.4 und Satz 1.9 erfüllt. Dann gelten für die optimalen Wertefunktionen v und 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) $v(x) \leq v_h(x) + K(h^{\frac{\gamma}{2}} + h)$

Ist das optimale Steuerungsproblem konvex, so gilt die schärfere Abschätzung

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

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

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

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

im allgemeinen Fall, bzw.

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

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 [7] von R. L. V. González and M. M. Tidball. Ein Beweis, der die Hamilton–Jacobi–Bellman Gleichung verwendet, findet sich im Buch von M. Bardi und I. Capuzzo Dolcetta [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 $\mathbf{u} \in U^{\mathbb{N}_0}$ so, dass die Identitäten

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

und

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

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

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

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 das größte ganze Zahl $i \in \mathbb{N}_0$, mit $hi \leq t$. Damit gilt

$$J_h(x, \mathbf{u}) = \int_0^\infty e^{-\delta h[t]_h} g(x_{[t]_h}(x, \mathbf{u}), u_{[t]_h})) \, dt.$$

Mit der Dreiecksungleichung erhalten wir

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

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 $\|\varphi(t,x,u) - \varphi(h[t]_h,x,u)\| \leq Mh$ gilt, und damit

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

Zum Abschätzen von (3.6) verwenden wir die Ungleichung

$$|e^{-\delta h[t]_h} - e^{-\delta t}| \leq e^{-\delta t} |e^{\delta h} - 1| \leq e^{-\delta t} e^1 \delta h$$

(die letzte Ungleichung folgt aus dem Mittelwertsatz und $\delta h \leq 1$), womit gilt

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

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

$$\int_0^\infty e^{-\delta h[t]_h} g(\varphi(h[t]_h,x,u),u(t)) \, dt = \int_0^\infty e^{-\delta h[t]_h} g(\varphi(h[t]_h,x,u),u_{[t]_h}) \, dt$$

folgt. Desweiteren folgt aus (3.2), dass $\mathbf{u}$ eine Folge ist, für die Satz 1.9(i) im konvexen Fall gilt, also insbesondere

$$\|\varphi(h[t]_h,x,u) - x_{[t]_h}(x,\mathbf{u})\| \leq Che^{Lt}$$

gilt für ein $C > 0$. Daraus können wir schließen, dass

$$\begin{aligned}
&\left| \int_0^T e^{-\delta h[t]_h} g(\varphi(h[t]_h,x,u),u(t)) \, dt - \int_0^T e^{-\delta h[t]_h} g(x_{[t]_h}(x,\mathbf{u}),u_{[t]_h}) \, dt \right| \\
&\leq e^1 \int_0^T e^{-\delta t} L_g Che^{Lt} \, dt \leq e^1 L_g Ch \frac{e^{(L-\delta)T} - 1}{L - \delta}
\end{aligned}$$

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

$$\left| \int_0^\infty e^{-\delta h[t]_h} g(\varphi(h[t]_h, x, u), u(t))\, dt - \int_0^\infty e^{-\delta h[t]_h} g(x_{[t]_h}(x, \mathbf{u}), u_{[t]_h})\, dt \right| \leq K_3 h^\gamma$$

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

Wir zeigen nun (i). Analog zur Konstruktion im Beweis von Satz 1.9 folgt aus der Konvexität, dass zu beliebigem $u \in \mathcal{U}$ eine Folge $\mathbf{u} \in U^{\mathbb{N}_0}$ existiert, so dass (3.2) und (3.3) erfüllt sind. Also folgt aus (3.4) für alle $u \in \mathcal{U}$ die Existenz von $\mathbf{u} \in U^{\mathbb{N}_0}$ mit

$$v_h(x) \geq J_h(x, \mathbf{u}) \geq J(x, u) - K h^\gamma$$

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

Zum Beweis von (ii) sei $\mathbf{u} \in U^{\mathbb{N}_0}$ beliebig. Dann erfüllt die stückweise konstante Kontrollfunktion $u(t) = u_{h[t]_h}$ offenbar (3.2) und (3.3). Mit (3.4) folgt

$$v(x) \geq J(x, u) \geq J_h(x, \mathbf{u}) - K h^\gamma,$$

also (ii), da wir auch hier auf der rechten Seite zum Supremum (jetzt über $\mathbf{u}$) übergehen können. □

Bemerkung 3.4 Analog zum Satz 1.9 können wir die Menge U durch eine hinreichend große endliche Menge $\widetilde{U} \subset U$ ersetzen, so dass die Aussage von Satz 3.3 gültig bleibt, vgl. Bemerkung 1.11. Wir können also o.B.d.A. annehmen, dass U eine endliche Menge ist. □

Der folgende Satz formuliert ein Optimalitätsprinzip für v_h.

Satz 3.5 Betrachte das optimale Steuerungsproblem aus Definition 2.1. Dann erfüllt die optimale Wertefunktion $v_h(x)$ für jedes $x \in \mathbb{R}^d$ und jedes $k \geq 0$ die Gleichung

$$v_h(x) = \sup_{\mathbf{u} \in U^{\mathbb{N}_0}} \left\{ h \sum_{i=0}^{k} e^{-\delta h i} g(x_i(x, \mathbf{u}), u_i) + e^{-\delta h (k+1)} v_h(x_{k+1}(x, \mathbf{u})) \right\}. \tag{3.8}$$

Beweis: Völlig analog zu Satz 2.7. □

Bemerkung 3.6 Beachte, dass das Supremum hier für festes k nur über Folgen der Form $(u_0, \ldots, u_k)$ genommen wird. Da außerdem x_i stetig in $(u_0, \ldots, u_i)$ ist (bzgl. einer passenden Topologie für diesen Folgenraum), können wir das Supremum hier auch als Maximum schreiben, falls v_h stetig ist. Tatsächlich lässt sich die Stetigkeit von v_h ähnlich wie die Stetigkeit von v beweisen. Insbesodere 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) angenommen wird. □

3.2 Ein Iterationsverfahren

Aus dem Optimalitätsprinzip in Satz 3.5 lässt sich eine Iterationsformel zur Berechnung von v_h herleiten, die wir als Basis für unsere numerische Approximation verwenden werden. Beachte dafür, dass die Folge $x_i(x, \mathbf{u})$ im Euler–Verfahren aus Definition 1.8 durch die Abbildung $\Phi_h(x, u)$ definiert ist.

Definition 3.7 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

$$T_h(w)(x) := \max_{u \in U} \Big\{ hg(x, u) + e^{-\delta h} w(\Phi_h(x, u)) \Big\}.$$

Hierbei bezeichnet $C(\mathbb{R}^d, \mathbb{R})$ die Menge der stetigen Funktionen von $\mathbb{R}^d$ nach $\mathbb{R}$. □

Wir werden zeigen, dass diese Iteration tatsächlich gegen v_h konvergiert.

Satz 3.8 Betrachte ein zeitdiskretes optimales Steuerungsproblem aus Definition 3.1 mit optimaler Wertefunktion v_h. Dann gilt für die in Definition 3.7 definierten Funktionen die Abschätzung

$$\|v_h^i - v_h\|_\infty \le e^{-\delta h i} \frac{hM_g}{1 - e^{-\delta h}} \le e^{-\delta h i} \frac{e^1 M_g}{\delta}.$$

Hierbei ist die *L_∞–Norm* $\|\cdot\|_\infty$ definiert durch $\|w\|_\infty = \sup_{x \in \mathbb{R}^d} |w(x)|$.

Insbesondere folgt also 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 (das analog für diskrete Kontrollfolgen $\mathbf{u}$ gilt)

$$|T_h(w_1)(x) - T_h(w_2)(x)| \le e^{-\delta h} \sup_{u \in U} |w_1(\Phi_h(x, u)) - w_2(\Phi_h(x, u))| \le e^{-\delta h} \|w_1 - w_2\|_\infty$$

und damit

$$\|T_h(w_1) - T_h(w_2)\|_\infty \le e^{-\delta h} \|w_1 - w_2\|_\infty.$$

Mit dem Optimalitätsprinzip aus Satz 3.5 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

$$\|v_h - v_h^{i+1}\|_\infty = \|T_h(v_h) - T_h(v_h^i)\|_\infty \le e^{-\delta h} \|v_h - v_h^i\|_\infty.$$

Wie im Beweis von Lemma 2.3(i) sieht man $\|v_h\|_\infty \le hM_g/(1 - e^{-\delta h})$, woraus

$$\|v_h - v_h^0\|_\infty = \|v_h\|_\infty \le \frac{hM_g}{1 - e^{-\delta h}} = e^{-\delta h 0} \frac{hM_g}{1 - e^{-\delta h}}$$

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

Die Ungleichung $\frac{hM_g}{1-e^{-\delta h}} \le \frac{e^1 M_g}{\delta}$ folgt dann aus dem Mittelwertsatz, der auf den Ausdruck $1 - e^{-r}$ angewendet wird. □

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

Lemma 3.9 Die Funktionen v_h^i aus Definition 3.7 sind Lipschitz–stetig mit Konstanten $L_0 = 0$ und

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

für $i \ge 1$.

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

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

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

$$\begin{aligned} |v_h^{i+1}(x) - v_h^{i+1}(y)| &= |T_h(v_h^i)(x) - T_h(v_h^i)(y)| \\ &\le \sup_{u\in U} \left\{ |hg(x,u) - hg(y,u) + e^{-\delta h} v_h^i(\Phi_h(x,u)) - e^{-\delta h} v_h^i(\Phi_h(y,u))| \right\}. \end{aligned}$$

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

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

Im Fall $i = 0$ folgt $hL_g + L_i(1+hL)e^{-\delta h} = hL_g = L_1$, also die Behauptung. Im Fall $i \ge 1$ folgt

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

also ebenfalls die Behauptung. □

3.3 Zustandsraumbeschränkung

Bevor wir im nächsten Kapitel einen eingeschränkten endlichdimensionalen Funktionenraum einführen, um die Iterationsvorschrift aus Definition 3.7 in eine implementierbare Form zu bringen, wollen wir uns noch kurz Gedanken zur Einschränkung von v_h auf eine kompakte Menge $\Omega \subset \mathbb{R}^d$ machen. Dies ist nötig, da wir v_h numerisch nicht im ganzen $\mathbb{R}^d$ berechnen können. Oftmals ergibt sich eine „interessante“ Menge Ω aus dem Modell, indem man sich z.B. auf einen physikalisch interessanten Bereich einschränkt. Formal ist das ganze 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 $\Phi_h(x,u) \in \Omega$. 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 $\Phi_h(x,u) \in \Omega$. 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 $\Phi_h(x,u_x^*) \in A$ für alle $x \in A$ und ein u_x^* aus Bemerkung 3.6, 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 $\Phi_h(x,u) \notin \Omega$. Dann können wir entweder die Punkte $\Phi_h(x,u)$ zurückprojizieren (d.h. wir ersetzen $\Phi_h(x,u)$ 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 Aussagen über optimal invariante Mengen aus (2). Wir werden das in den Übungen an Beispielen genauer diskutieren.

Kapitel 4

Diskretisierung im Ort

Obwohl alle Größen, die in der Iterationsvorschrift aus Definition 3.7 vorkommen, im Rechner—bis auf Rundungsfehler—auswertbar sind (zumindest wenn wir annehmen, dass U eine endliche Menge ist, vgl. Bemerkung 3.4), 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).

4.1 Funktionen auf Gittern

Wir müssen uns also zur Berechnung der v_h auf einen endlichdimensionalen Funktionenraum einschränken. Wir werden uns hier 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 4.1 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, $i = 1, \dots, P = P_1 P_2$, mit Kantenlängen $k_1 = (b_1 - a_1)/P_1$ und $k_2 = (b_2 - a_2)/P_2$, so dass

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

Mit E_i, $i = 1, \dots, N = (P_1 + 1)(P_2 + 1)$ 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 4.1 zeigt ein solches Gitter.

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

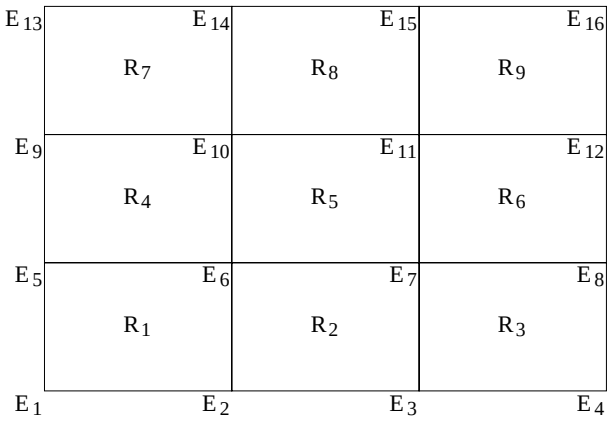


Abbildung 4.1: Beispielgitter

Definition 4.2 (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 Recheckgitter Γ. Wir definieren den Raum der stetigen und stückweise affin bilinearen Funktionen auf Ω bezüglich Γ als

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

□

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

Lemma 4.3 (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

$$E_{i_1} = (c_1, c_2)^T, \;\; E_{i_2} = (d_1, c_2)^T, \;\; E_{i_3} = (c_1, d_2)^T, \;\; E_{i_4} = (d_1, d_2)^T$$

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

$$w(x) = \sum_{j=1}^{4} \mu_j(x) w(E_{i_j})$$

mit

$$\begin{aligned} \mu_1(x) &= (1 - y_1(x))(1 - y_2(x)), & \mu_2(x) &= y_1(x)(1 - y_2(x)), \\ \mu_3(x) &= (1 - y_1(x))y_2(x), & \mu_4(x) &= y_1(x)y_2(x) \end{aligned}$$

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 $j = 1, \ldots, 4$ und $\sum_{j=1}^{4} \mu_j(x) = 1$.

Beweis: (i) Seien w und $\tilde{w}$ zwei affin bilineare Funktionen auf R_i die in den Eckpunkten von R_i übereinstimmen. Mit den Bezeichnungen aus (ii) folgt aus der Definition der affin bilinearen Funktionen, dass die Koeffizienten α_i beider Funktionen die Gleichungen

$$\begin{aligned}
\alpha_0 + \alpha_1 c_1 + \alpha_2 c_2 + \alpha_3 c_1 c_2 &= w(E_{i_1}) \\
\alpha_0 + \alpha_1 d_1 + \alpha_2 c_2 + \alpha_3 d_1 c_2 &= w(E_{i_2}) \\
\alpha_0 + \alpha_1 c_1 + \alpha_2 d_2 + \alpha_3 c_1 d_2 &= w(E_{i_3}) \\
\alpha_0 + \alpha_1 d_1 + \alpha_2 d_2 + \alpha_3 d_1 d_2 &= w(E_{i_4})
\end{aligned}$$

erfüllen müssen. Berechnet man die Determinante der zugehörigen Matrix (leicht mit MAPLE), so erhält man $(d_1 - c_1)^2(d_2 - c_2)^2$. Da die Eckpunkte paarweise verschieden sind, ist diese verschieden von Null und damit besitzt dieses Gleichungssystem also genau eine Lösung. Folglich stimmen die Koeffizienten α_i von $\tilde{w}$ und w überein und damit gilt $\tilde{w} = w$.

(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 $\mathbb{R}$.

4.2 Die vollständige Diskretisierung

Wir werden nun den iterativen Algorithmus aus Definition 3.7 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, $i = 1, \ldots, N$ auswerten, d.h. wir berechnen eine Folge von Funktionen $v^j_{h,\Gamma} \in \mathcal{W}$ mittels

$$v^{j+1}_{h,\Gamma}(E_i) = \max_{u \in U} \left\{ h g(E_i, u) + e^{-\delta h} v^j_{h,\Gamma}(\Phi_h(E_i, u)) \right\}.$$

Schreiben wir $V^j = (V^j_1, \ldots, V^j_N)^T \in \mathbb{R}^N$ mit $V^j_i = v^j_{h,\Gamma}(E_i)$, 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 4.4 Betrachte ein zeitdiskretes optimales Steuerungsproblem mit einer endlichen Menge von Kontrollwerten U. Betrachte weiterhin ein Rechteckgitter Γ mit P Rechtecken und N Eckpunkten. Zu jedem $u \in U$ und jedem $i = 1, \ldots, N$ sei $B(i, u)$ der N–dimensionale Zeilenvektor, für den für jedes $w \in \mathcal{W}$ und $W = (w(E_1), \ldots, w(E_N))^T \in \mathbb{R}^N$ 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, welche zudem ≥ 0 sind und sich zu 1 aufsummieren). Desweiteren sei $G(i,u) = hg(E_i, u)$. Dann berechnen wir Vektoren V^j iterativ durch $V^0 := (0, \ldots, 0)^T$ und dem *Gesamtschrittverfahren*

$$V_i^{j+1} := \max_{u \in U} \{G(i,u) + e^{-\delta h} B(i,u) V^j\} \quad \text{für } i = 1, \ldots, N$$

oder dem *Einzelschrittverfahren*

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

□

Bemerkung 4.5 (i) Im Allgemeinen ist das Einzelschrittverfahren vorteilhafter, da wir für jedes $i > 1$ bereits die aktuellen Werte V_k^{j+1} für $0 < k < i$ 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) Da wir U als endliche Menge angenommen haben, also $U = \{u_1, \ldots, u_q\}$ für ein $q \in \mathbb{N}$ gilt, kann das Maximum in dieser Iteration für jedes $i = 1, \ldots, N$ durch Vergleich der Werte

$$G(x, u_k) + e^{-\delta h} B(i, u_k) V^j, \quad k = 1, \ldots, q$$

bzw.

$$G(x, u_k) + e^{-\delta h} B(i, u_k) V^{j+1}, \quad k = 1, \ldots, q$$

bestimmt werden.

(iii) Falls 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 aufwendigste Teil des Algorithmus' ist. Natürlich sollte man dabei die Vektoren nicht komplett speichern sondern nur diejenigen Einträge, die ungleich Null sind (Details siehe Übung). □

Das folgende Lemma gibt ein Abbruchkriterium für diese Iteration.

Lemma 4.6 Betrachte die Iterationsvorschrift aus Definition 4.4. Dann konvergieren die Vektoren V^j für $j \to \infty$ komponentenweise gegen den Vektor V^∞, der eindeutig bestimmt ist durch

$$V_i^\infty = \max_{u \in U} \{G(i,u) + e^{-\delta h} B(i,u) V^\infty\} \quad \text{für } i = 1, \ldots, N.$$

Für die mit $v_{h,\Gamma}^j$, $j = 1, \ldots, \infty$ bezeichneten zugehörigen Funktionen aus $\mathcal{W}$ gilt darüberhinaus: Falls $|V_i^j - V_i^{j+1}| \leq \varepsilon$ für alle $i = 1, \ldots, N$, so folgt

$$\|v_{h,\Gamma}^j - v_{h,\Gamma}^\infty\|_\infty \leq e^1 \frac{\varepsilon}{h\delta}.$$

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

$$\left| \max_{u \in U} \{G(i,u) + e^{-\delta h} B(i,u)V\} - \max_{u \in U} \{G(i,u) + e^{-\delta h} B(i,u)W\} \right| \leq e^{-\delta h} \|V - W\|_\infty \quad (4.1)$$

folgt, wobei wir $\sum_{k=1}^N B(i,u)_k = 1$ ausgenutzt haben und $\|\cdot\|_\infty$ die L_∞–Norm im $\mathbb{R}^N$ bezeichnet. Insbesondere folgt daraus, dass es genau einen Vektor V^∞ gibt, der die angegebene Gleichung erfüllt, denn für jeden weiteren solchen Vektor W^∞ gilt mit (4.1)

$$\|V^\infty - W^\infty\|_\infty \leq e^{-\delta h} \|V^\infty - W^\infty\|_\infty,$$

woraus $W^\infty = V^\infty$ folgt.

Ebenfalls mit (4.1) sieht man leicht, dass die Vektoren V^j die Abschätzung

$$\|V^{j+1} - V^\infty\|_\infty \leq e^{-\delta h} \|V^j - V^\infty\|_\infty$$

erfüllen. Daraus folgt die behauptete Konvergenz.

Außerdem folgt

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

und daraus

$$\|V^j - V^\infty\|_\infty \leq \frac{\varepsilon}{1 - e^{-\delta h}} \leq e^1 \frac{\varepsilon}{h\delta},$$

wobei die letzte Ungleichung aus dem Mittelwertsatz für die Funktion e^{-r} folgt. Die entsprechende Aussage für die Funktionen $v^j_{h,\Gamma}$ folgt nun leicht mit der Darstellung aus Lemma 4.3(ii). □

In der Praxis zeigt sich, dass die Iterationsvorschrift aus Definition 4.4 recht langsam gegen V^∞ konvergiert, insbesondere für kleine h und δ. Wir werden in einem späteren Kapitel alternative Iterationen besprechen, die deutlich schneller konvergieren.

4.3 Diskretisierungsfehler

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

$$\|v_h - v^\infty_{h,\Gamma}\|_\infty$$

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

Definition 4.7 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

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

□

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

Lemma 4.8 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 4.1.

Beweis: Sei $x \in \Omega$ ein beliebiger Punkt und R_i ein Gitterrechteck, in dem dieser Punkt liegt. Seien $E_{i_1}, \ldots, E_{i_4}$ die Eckpunkte dieses Rechtecks. Dann gilt $\|x - E_{i_j}\| \leq k$ für alle $j = 1, \ldots, 4$ und somit $|q(x) - q(E_{i_j})| \leq L_q k$. Mit Lemma 4.3 folgt

$$\begin{aligned} |q(x) - \pi_{\mathcal{W}} q(x)| &= \left| q(x) - \sum_{j=1}^{4} \mu_j(x) q(E_{i_j}) \right| \\ &= \left| \sum_{j=1}^{4} \mu_j(x) q(x) - \sum_{j=1}^{4} \mu_j(x) q(E_{i_j}) \right| \\ &\leq \sum_{j=1}^{4} \mu_j(x) \, |q(x) - q(E_{i_j})| \;=\; \sum_{j=1}^{4} \mu_j(x) \, k \;=\; k \end{aligned}$$

wobei wir im zweiten und im letzten Schritt ausgenutzt haben, dass $\sum_{j=1}^{4} \mu_j(x) = 1$ gilt. □

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

Lemma 4.9 Betrachte die Funktionen v_h^j und $v_{h,\Gamma}^j$ aus den Definitionen 3.7 und 4.4. Dann gelten die Abschätzungen

$$\|v_h^j - v_{h,\Gamma}^j\|_\infty \leq 2 M_g e^{\delta h} \int_0^{jh} e^{-\delta t} \, dt \tag{4.2}$$

und, falls $\delta > L$,

$$\|v_h^j - v_{h,\Gamma}^j\|_\infty \leq C \frac{k}{h} \tag{4.3}$$

bzw., falls $\delta < L$

$$\|v_h^j - v_{h,\Gamma}^j\|_\infty \leq C \frac{k}{h} \int_0^{jh} e^{(L-\delta)t} \, dt \tag{4.4}$$

mit einer geeigneten Konstante $C > 0$.

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

$$v_{h,\Gamma}^{j+1} = \pi_{\mathcal{W}} v_{h,\Gamma}^{j+1} = \pi_{\mathcal{W}} T_h(v_{h,\Gamma}^j)$$

erhält. Wir zeigen nun zunächst (4.2). Aus den Definitionen folgt, dass für die betrachteten Funktionen die Ungleichungen

$$\|v_h^{j+1}\|_\infty \le hM_g + e^{-\delta h}\|v_h^j\|_\infty \quad \text{und} \quad \|v_{h,\Gamma}^{j+1}\|_\infty \le hM_g + e^{-\delta h}\|v_{h,\Gamma}^j\|_\infty$$

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

$$\|v_h^j\|_\infty \le \sum_{i=0}^{j-1} e^{-\delta h i} hM_g \quad \text{und} \quad \|v_{h,\Gamma}^j\|_\infty \le \sum_{i=0}^{j-1} e^{-\delta h i} hM_g.$$

Aus $\|v_h^j - v_{h,\Gamma}^j\|_\infty \le \|v_h^j\|_\infty + \|v_{h,\Gamma}^j\|_\infty$ und

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

folgt damit (4.2).

Zum Beweis von (4.3) und (4.4) verwenden wir Lemma 3.9, welches besagt, dass v_h^j Lipschitz–stetig mit der dort angegebenen Konstante L_j ist. Damit ergibt sich

$$\begin{aligned} \|v_h^{j+1} - v_{h,\Gamma}^{j+1}\|_\infty &\le \|v_h^{j+1} - \pi_{\mathcal{W}} v_h^{j+1}\|_\infty + \|\pi_{\mathcal{W}} v_h^{j+1} - \pi_{\mathcal{W}} v_{h,\Gamma}^{j+1}\|_\infty \\ &\le kL_{j+1} + e^{-\delta h}\|v_h^j - v_{h,\Gamma}^j\|_\infty, \end{aligned}$$

wobei wir zur Abschätzung des zweiten Terms in der letzten Ungleichung die Abschätzung $\|\pi_{\mathcal{W}} q_1 - \pi_{\mathcal{W}} q_2\|_\infty \le \|q_1 - q_2\|_\infty$ und Lemma 2.4 verwendet haben.

Aus $v_h^0 = v_{h,\Gamma}^0$ erhalten wir nun mittels Induktion die Abschätzung

$$\|v_h^j - v_{h,\Gamma}^j\|_\infty \le \sum_{i=1}^{j} e^{-\delta h(j-i)} kL_i.$$

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

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

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

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

□

Mit diesem Lemma können wir nun den folgenden Satz beweisen.

Satz 4.10 Betrachte ein zeitdiskretes optimales Steuerungsproblem aus Definition 3.1 auf einer kompakten Rechteckmenge Ω mit Zeitschritt h und einer endlichen Menge von Kontrollwerten U. Sei v_h die zugehörige optimale Wertefunktion. Betrachte weiterhin ein Gitter Γ auf Ω mit Durchmesser k. Dann gilt für die Funktion $v^\infty_{h,\Gamma}$ aus Lemma 4.6 die Abschätzung

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

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 4.3 im Lemma 4.9, die für alle $j \geq 0$ und damit auch für $j \to \infty$ gilt.

Im Fall $\delta < L$ erhalten wir aus Lemma 4.9 die Abschätzung

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

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

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

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 4.10 mit Satz 3.3 so erhalten wir die folgende Aussage.

Satz 4.11 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 $v^\infty_{h,\Gamma}$ aus Lemma 4.6 die Abschätzung

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

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

$$\|v - v^\infty_{h,\Gamma}\|_\infty \leq Kh^{\gamma} + K\left(\frac{k}{h}\right)^\gamma.$$

Aus dem Satz ergibt sich die Forderung, dass wir, um Konvergenz von $v_{h,\Gamma}^{\infty}$ 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 [5] bewiesen wurde. Diese lautet

$$\|v - v_{h,\Gamma}^{\infty}\|_{\infty} \leq Kh^{\gamma/2} + K\left(\frac{k}{\sqrt{h}}\right)^{\gamma}$$

für alle hinreichend kleinen k, $h > 0$ mit der Eigenschaft, dass $k \leq C_1 h$ für eine Konstante $C_1 > 0$ gilt. Für $k = C_1 h$ folgt also insbesondere die Abschätzung $\|v - v_{h,\Gamma}^{\infty}\|_{\infty} \leq \widetilde{K} h^{\gamma/2}$ 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 5

Berechnung approximativ optimaler Trajektorien

In diesem Kapitel 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 $v_{h,\Gamma}^{\infty}$ entstehenden Fehler analysieren, und schließlich zum kontinuierlichen Problem übergehen.

5.1 Zeitdiskrete optimale Trajektorien

Wie erinnern uns an das Optimalitätsprinzip für v_h aus Satz 3.5, das für $k = 0$ als

$$v_h(x) = \max_{u \in U} \{hg(x,u) + e^{-\delta h} v_h(\Phi_h(x,u))\} \tag{5.1}$$

geschrieben werden kann. Die folgende Definition basiert auf diesem Prinzip

Definition 5.1 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 (5.1) angenommen wird (dieses u_x wird im Allgemeinen nicht eindeutig sein), und $u^*(x) = u_x$ setzen. □

Bemerkung 5.2 Dies Vorschrift definiert eine Kontrollstrategie, die von x und nicht von t abhängt, also keine Kontrollfunktion bzw. Kontrollfolge im bisher bekannten Sinne ist. Eine solche Kontrollstrategie nennt man *Zustandsrückführung* oder *Zustandsfeedback*. □

Mittels u^* definieren wir nun zu jedem Anfangswert x eine Kontrollfolge $\mathbf{u}^x \in U^{\mathbb{N}_0}$, $\mathbf{u}^x = (u_0^x, u_1^x, \ldots)$, gemäß der folgenden iterativen Vorschrift:

$$u_0^x := u^*(x), \qquad u_i^x = u^*(x_i(x, \mathbf{u}^x)) \quad \text{für} \quad i = 1, 2, \ldots. \tag{5.2}$$

Beachte, dass u_i^x wohldefiniert ist, da $x_i(x, \mathbf{u}^x)$ nur von den Werten $(u_0^x, u_1^x, \ldots, u_{i-1}^x)$ abhängt, die für ein gegebenes $i \geq 1$ bereits iterativ definiert sind.

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

Satz 5.3 Die in (5.2) definierte Kontrollfolge ist optimal für das zeitdiskrete optimale Steuerungsproblem, d.h. für alle $x \in \mathbb{R}^d$ gilt

$$J_h(x, \mathbf{u}^x) = v_h(x).$$

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

$$v_h(x) = h \sum_{i=0}^{k} e^{-\delta ih} g(x_i, u_i^x) + e^{-\delta(k+1)h} v_h(x_{k+1}). \tag{5.3}$$

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

$$\lim_{k\to\infty} \left(h \sum_{i=0}^{k} e^{-\delta ih} g(x_i, u_i^x) + e^{-\delta(k+1)h} v_h(x_{k+1}) \right) = h \sum_{i=0}^{\infty} e^{-\delta ih} g(x_i, u_i^x) = J_h(x, \mathbf{u}^x).$$

Zum Beweis von (5.3) verwenden wir, dass aus der Definition von $u^*(x)$ und $\mathbf{u}^x$ für alle $j \geq 0$ folgt

$$\begin{aligned} v_h(x_j) &= hg(x_j, u^*(x_j)) + e^{-\delta h} v_h(\Phi_h(x_j, u^*(x_j)) \\ &= hg(x_j, u_j^x) + e^{-\delta h} v_h(\Phi_h(x_j, u_j^x)) \;=\; hg(x_j, u_j^x) + e^{-\delta h} v_h(x_{j+1}). \end{aligned} \tag{5.4}$$

Für $k = 0$ folgt (5.3) nun direkt aus (5.4) mit $j = 0$. Gelte also (5.3) für ein $k \geq 0$. Dann folgt mit (5.4) für $j = k + 1$

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

womit (5.3) für $k + 1$ gezeigt ist und damit für alle $k \geq 0$ gilt. □

5.2 Numerische Approximation

Analog zu Definition 5.1 definieren wir nun eine numerische Kontrollstrategie $\tilde{u}^*$. Betrachte dazu den Ausdruck

$$hg(x, u) + e^{-\delta h} v_{h,\Gamma}^{\infty}(\Phi_h(x, u)) \tag{5.5}$$

für die Funktion $v_{h,\Gamma}^{\infty}$ aus Lemma 4.6.

Definition 5.4 Wir definieren eine Abbildung $\tilde{u}^* : \mathbb{R}^d \to U$, indem wir zu jedem $x \in \mathbb{R}^d$ ein $u_x \in U$ wählen, so dass das Maximum in (5.5) angenommen wird (dieses u_x wird im Allgemeinen wiederum nicht eindeutig sein), und $\tilde{u}^*(x) = u_x$ setzen. □

Analog zu (5.2) definieren wir nun zu jedem Anfangswert x eine Kontrollfolge $\tilde{\mathbf{u}}^x \in U^{\mathbb{N}_0}$, $\tilde{\mathbf{u}}^x = (\tilde{u}_0^x, \tilde{u}_1^x, \ldots)$, gemäß der folgenden iterativen Vorschrift:

$$\tilde{u}_0^x := \tilde{u}^*(x), \qquad \tilde{u}_i^x = \tilde{u}^*(x_i(x, \tilde{\mathbf{u}}^x)) \quad \text{für} \quad i = 1, 2, \ldots. \tag{5.6}$$

Satz 5.5 Die in (5.6) definierte Kontrollfolge ist approximativ optimal für das zeitdiskrete optimale Steuerungsproblem. Genauer gilt für alle $x \in \mathbb{R}^d$ die Abschätzung

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

für $\gamma \in [0, 1]$ aus Satz 4.10 und eine geeignete Konstante $C > 0$.

Beweis: Der Beweis verläuft analog zu dem von Satz 5.3. Wir schreiben kurz $x_k = x_k(x, \tilde{\mathbf{u}}^x)$ und beweisen

$$v_h(x) = h \sum_{i=0}^{k} e^{-\delta i h} g(x_i, \tilde{u}_i^x) + e^{-\delta(k+1)h} v_h(x_{k+1}) + 2K \sum_{i=0}^{k} e^{-\delta(i+1)h} \frac{k^\gamma}{h^\gamma}. \tag{5.7}$$

für ein $K > 0$. Hieraus folgt die Behauptung analog zum Beweis von Satz 5.3, denn es gibt ein $C > 0$ so dass für alle $k > 0$ die Abschätzung

$$2K \sum_{i=0}^{k} e^{-\delta(i+1)h} \le C \frac{1}{h}$$

gilt.

Zum Beweis von (5.7) beachte, dass aus Satz 4.10, Gleichung (5.1) und der Definition von $\tilde{u}^*$ für jedes $y \in \mathbb{R}^d$ folgt

$$\begin{aligned} v_h(y) &= \max_{u \in U} \{h g(y, u) + e^{-\delta h} v_h(\Phi_h(y, u))\} \\ &= \max_{u \in U} \{h g(y, u) + e^{-\delta h} v_{h,\Gamma}^\infty(\Phi_h(y, u))\} + R_1(y) \\ &= h g(y, \tilde{u}^*(y)) + e^{-\delta h} v_{h,\Gamma}^\infty(\Phi_h(y, \tilde{u}^*(y))) + R_1(y) \\ &= h g(y, \tilde{u}^*(y)) + e^{-\delta h} v_h(\Phi_h(y, \tilde{u}^*(y))) + R_2(y) \end{aligned}$$

mit $|R_2(y)| \le e^{-\delta h} 2K k^\gamma / h^\gamma$.

Nun folgt (5.7) ganz analog zum Beweis von (5.3) im Beweis von Satz 5.3. □

5.3 Das zeitkontinuierliche Problem

Wir wollen nun abschließend zeigen, dass wir aus der approximativ optimalen Kontrollfolge $\tilde{\mathbf{u}}^x$ eine messbare Kontrollfunktion konstruieren können, die auch für das ursprüngliche zeitkontinuierliche Problem approximativ optimal ist. Hierzu setzen wir

$$u^x(t) := \tilde{u}_i^x, \quad t \in [ih, (i+1)h). \tag{5.8}$$

Satz 5.6 Die in (5.8) definierte Kontrollfunktion ist approximativ optimal für das optimale Steuerungsproblem aus Definition 2.1. Genauer gilt für alle $x \in \mathbb{R}^d$ die Abschätzung

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

für $\gamma \in [0, 1]$ aus Satz 4.10 und eine geeignete Konstante $C > 0$. Ist das optimale Steuerungsproblem konvex, so gilt sogar

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

Beweis: Beachte, dass die stückweise konstante Kontrollfunktion aud (5.8) die Bedingungen (3.2) und (3.3) aus dem Beweis von Satz 3.3 erfüllt. Also folgt aus (3.4) die Abschätzung

$$|J_h(x, \tilde{\mathbf{u}}^x) - J(x, u^x)| \leq K_1 h^\gamma \tag{5.9}$$

für eine passende Konstante $K_1 > 0$. Die Behauptung folgt nun mit Dreiecksungleichung aus (5.9) und den Sätzen 5.5 und 3.3. □

Kapitel 6

Beschleunigung der Iteration

In den Übungen haben wir festgestellt, dass die Iterationsvorschrift aus Definition 4.4 zur Berechnung des Vektors V^∞ insbesondere für kleine $\delta > 0$ und $h > 0$ recht langsam konvergiert. In diesem Kapitel wollen wir zwei Methoden besprechen, mit denen die Berechnung von V^∞ schneller durchgeführt werden kann.

6.1 Das Koordinatenaufstiegsverfahren

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

$$V^0 := (0, \ldots, 0)^T; \quad V^{j+1} := V^j, \quad V_i^{j+1} := S(V^{j+1})_i, \;\; i = 1, \ldots, N, \;\; j = 0, 1, \ldots \tag{6.1}$$

mit

$$S(V)_i = \max_{u \in U} \{G(i,u) + e^{-\delta h} \sum_{k=1}^{N} B(i,u)_k V_k\}$$

für $V \in \mathbb{R}^N$. Dieses Verfahren wird in der Literatur oft als *sukzessive Approximation* bezeichnet. Beachte, dass die „:=“ in (6.1) Zuweisungen sind; insbesondere geht in der letzten Zuweisung in (6.1) 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 (6.1) als Gleichung aufzufassen, d.h. einen Wert V_i^{j+1} zu bestimmen, so dass

$$V_i^{j+1} = S(V^{j+1})_i \tag{6.2}$$

erfüllt ist. Die Gleichung (6.2) ist explizit lösbar, denn es gilt

$$V_i^{j+1} = S(V^{j+1})_i = \max_{u \in U} \{G(i,u) + e^{-\delta h} \sum_{k=1}^{N} B(i,u)_k V_k^{j+1}\}$$

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

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

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

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

Wir definieren also das sogenannte *Koordinatenaufstiegsverfahren* analog zur obigen Iteration durch

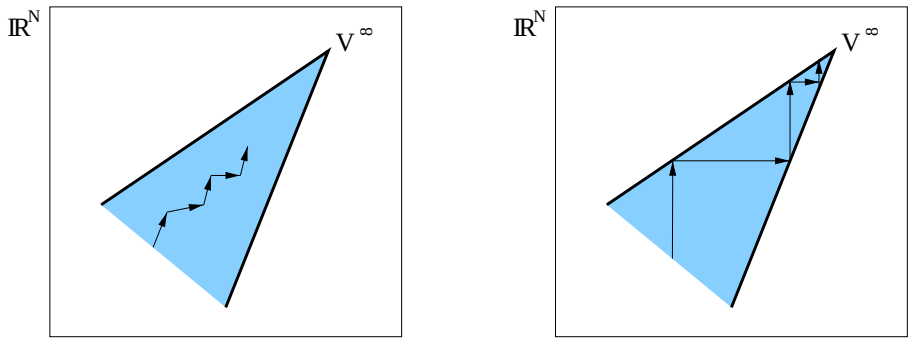
$$V^0 := (0, \ldots, 0)^T; \quad V^{j+1} := V^j, \quad V_i^{j+1} := \widetilde{S}(V^{j+1})_i, \quad i = 1, \ldots, N, \quad j = 0, 1, \ldots \tag{6.3}$$

mit

$$\widetilde{S}(V)_i = \max_{u \in U} \left\{ \frac{G(i,u) + e^{-\delta h} \sum_{\substack{k=1 \\ k \neq i}}^{N} B(i,u)_k V_k}{1 - e^{-\delta h} B(i,u)_i} \right\}$$

für $V \in \mathbb{R}^N$.

Dieses Verfahren aus [8] hat 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 6.1 zeigt eine schematische Darstellung der beiden Iterationen.



Sukzessive Approximation Koordinatenaufstiegsverfahren

Abbildung 6.1: Iterationsverfahren

Die Tatsache, dass man in jeder Koordinate jeweils bis zum Rand des Polyeders „aufsteigt“ ist der Grund für den Namen Koordinatenaufstiegsverfahren.

Es ist leicht nachzuprüfen, dass Lemma 4.6 auch für die Iteration (6.3) gilt, und dass die Iteration tatsächlich gegen den Vektor V^∞ konvergiert.

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 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 (6.3) langsamer konvergiert als (6.1) sind nicht bekannt.

6.2 Strategie–Iteration

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

Wenn wir einen Vektor $\tilde{u} = (\tilde{u}_1,\ldots,\tilde{u}_N)^T \in U^N$ 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 $V^{\tilde{u},\infty}$, der durch das Gleichungssystem

$$V_i^{\tilde{u},\infty} = G(i,\tilde{u}_i) + e^{-\delta h} B(i,\tilde{u}_i) V^{\tilde{u},\infty}, \quad i = 1,\ldots,N \tag{6.4}$$

eindeutig bestimmt ist. Man sieht leicht, dass $V_i^{u^*,\infty} \le V_i^\infty$ für alle $i = 1,\ldots,N$ gilt. Da die Maximierung hier wegfällt, lässt sich—vor allem wenn U eine große Menge ist—der Vektor $V^{u^*,\infty}$ viel schneller als der Vektor V^∞ berechnen. Dies gilt insbesondere, wenn man zur Berechnung von $V^{u^*,\infty}$ nicht die Iterationen (6.1) oder (6.3) mit fixiertem $\tilde{u}$ verwendet, sondern ausnutzt, dass sich aus (6.4) das schwach besetzte lineare Gleichungssystem

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

herleiten lässt, für dessen Lösung es effiziente numerische Verfahren gibt.[3]

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 $V^{j+1} = V^{\tilde{u}^j,\infty}$ zu berechnen. Formal lässt sich dieses Verfahren wie folgt beschreiben.

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

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

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

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

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

[3] In der Praxis haben sich hier z.B. präkonditionierte CGS– oder BiCGStab–Verfahren bewährt

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

In der Arbeit [13] von M.L. Puterman and S. Brumelle wurde gezeigt, dass die Vektoren $\widetilde{V}^j$ lokal quadratisch gegen V^∞ konvergieren; allerdings muss dabei die Zeit zur Berechnung von $V^{\tilde{u}^j,\infty}$ berücksichtigt werden, die unter Umständen recht lang sein kann.

In der Praxis zeigt sich, dass dieses Verfahren am Anfang recht langsam konvergiert, oft langsamer als die Iteration (6.1). Es liegt daher nahe, die beiden Verfahren zu kombinieren. Dies erreicht man, indem man Schritt (2) wie folgt modifiziert.

(2) (a) Setze $V^0 := \widetilde{V}^j$ und $k := 0$

(b) Setze $V^{k+1} := V^k$ und berechne $V_i^{k+1} := S(V^k)_i$, $i = 1, \ldots, N$;
sei dabei $\tilde{u}^{j,k} \in U^N$ der Vektor, der die maximierenden Kontrollwerte enthält

(c) Falls $\|V^{k+1} - V^k\|_\infty \leq \varepsilon$ beende die Berechnung;
Falls ein Abbruchkriterium (s.u.) erfüllt ist, setze $\tilde{u}^j = \tilde{u}^{j,k}$ und gehe zu (3);
ansonsten setze $k := k + 1$ und gehe zu (2b)

Natürlich kann man in diesem Verfahren auch die Iteration $\widetilde{S}$ an Stelle von S verwenden, wodurch eine weitere Beschleunigung erzielt werden kann.

Die Frage, was ein gutes Abbruchkriterium in (2c) ist, kann nur experimentell beantwortet werden und hängt stark vom zugrundeliegenden optimalen Steuerungsproblem ab. In der Arbeit [14] von A. Seeck wird vorgeschlagen zu prüfen, wie viele Einträge von $\tilde{u}^{j,k}$ und $\tilde{u}^{j,k-1}$ 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 [6] von R. L. V. González and C. A. Sagastizábal hingegen wird ein Wert $q \in \mathbb{N}$ festgelegt und erst dann zu (3) übergegangen, wenn die $q + 1$ aufeinanderfolgenden Kontrollvektoren $\tilde{u}^{j,k-q}, \ldots, \tilde{u}^{j,k}$ 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.[4]

[4]Der Wert $q = 1$ entspricht dem 100 Prozent Kriterium aus der Arbeit von Seeck

Kapitel 7

Fehlerschätzung

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

7.1 Definition der Fehlerschätzer

Natürlich ist es nicht möglich, die Differenz $\|v_h - v_{h,\Gamma}^{\infty}\|_{\infty}$ 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 7.1 Betrachte das vollständig diskrete optimale Steuerungsproblem auf einem Rechteckgitter Γ mit $P \in \mathbb{N}$ Rechtecken $R_1, \ldots, R_P$. Ein *lokaler a–posteriori Fehlerschätzer* (in der $\|\cdot\|_{\infty}$–Norm) ist eine Menge von Werten $\eta_1, \ldots, \eta_P$ mit den folgenden Eigenschaften.

(i) Der Wert η_i lässt sich aus den Daten des optimalen Steuerungsproblems und aus der Funktion $v_{h,\Gamma}^{\infty}$ 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 $\eta := \max_{i=1,\ldots,P} \eta_i$ die Abschätzungen

$$C_1 \eta \leq \|v_h - v_{h,\Gamma}^{\infty}\|_{\infty} \leq C_2 \eta$$

gelten. Man sagt auch, dass der Fehlerschätzer *effizient* und *zuverlässig*[1] ist.

Gilt darüberhinaus die Abschätzung

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

so heißt der Fehlerschätzer *lokal effizient*. □

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

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 [9] für die mathematischem Grundlagen und [11] für Details der Implementierung.

In dieser Vorlesung werden wir uns darauf beschränken zu zeigen, wie sich Fehlerschätzer gemäß Definition 7.1 konstruieren lassen (diese Konstruktion stammt ebenfalls aus den zitierten Arbeiten).

7.2 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 3.7. Wie wir im Beweis von Lemma 4.9 beobachtet haben, lösen wir aber tatsächlich (zumindest approximativ) die Gleichung

$$v_{h,\Gamma}^{\infty} = \pi_{\mathcal{W}} T_h(v_{h,\Gamma}^{\infty}).$$

Die Idee besteht nun darin, das Residuum des Operators T_h bzgl. $v_{h,\Gamma}^{\infty}$ auszurechnen, d.h., den Wert

$$\|v_{h,\Gamma}^{\infty} - T_h(v_{h,\Gamma}^{\infty})\|$$

zu berechnen.

Definition 7.2 Wir definieren eine Funktion $\eta : \Omega \to \mathbb{R}_0^+$ mittels

$$\begin{aligned} \eta(x) &:= \left| v_{h,\Gamma}^{\infty}(x) - T_h(v_{h,\Gamma}^{\infty})(x) \right| \\ &= \left| v_{h,\Gamma}^{\infty}(x) - \left(\max_{u \in U} \left\{ hg(x,u) + e^{-\delta h} v_{h,\Gamma}^{\infty}(x + hf(x,u)) \right\} \right) \right| \end{aligned}$$

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

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

□

Offenbar hängen die Werte η_i in dieser Definition tatsächlich nur von den Daten des optimalen Steuerungsproblems (genauer von f, g, δ und h) ab, sowie von den Werten der Funktion $v_{h,\Gamma}^{\infty}$ in der Umgebung

$$\mathcal{N}(R_i) := \{y \in \Omega \,|\, \text{es gibt ein } x \in R_i \text{ mit } \|y - x\| \leq hM\}, \tag{7.1}$$

wobei M eine obere Schranke für $\|f(x,u)\|$ für alle $x \in \Omega$ und $u \in U$ ist. Der folgende Satz zeigt, dass auch die anderen Eigenschaften der lokalen Fehlerschätzer erfüllt sind.

Satz 7.3 Für den Fehlerschätzer aus Definition 7.2 gelten die Ungleichungen

$$\frac{1}{2}\,\eta \leq \|v_h - v_{h,\Gamma}^{\infty}\|_{\infty} \leq \frac{1}{1 - e^{-\delta h}}\,\eta$$

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

$$\frac{1}{2}\,\eta_i \leq \sup_{x \in \mathcal{N}(R_i)} |v_h(x) - v_{h,\Gamma}^{\infty}(x)|$$

für die Umgebung $\mathcal{N}(R_i)$ aus (7.1).

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

$$|T_h(v_1)(x) - T_h(v_2)(x)| \leq e^{-\delta h} \sup_{y \in B_{hM}(x)} |v_1(y) - v_2(y)| \tag{7.2}$$

gilt.

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

$$\begin{aligned} |v_{h,\Gamma}^{\infty}(x) - T_h(v_{h,\Gamma}^{\infty})| &= |v_{h,\Gamma}^{\infty}(x) - v_h(x) + T_h(v_h)(x) - T_h(v_{h,\Gamma}^{\infty})(x)| \\ &\leq |v_{h,\Gamma}^{\infty}(x) - v_h(x)| + |T_h(v_{h,\Gamma}^{\infty})(x) - T_h(v_h)(x)| \\ &\leq 2 \sup_{y \in B_{hM}(x)} |v_{h,\Gamma}^{\infty}(y) - v_h(y)|, \end{aligned}$$

wobei die letzte Ungleichung aus (7.2) und $e^{-\delta h} < 1$ 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 $i = 1, \ldots, P$.

Die zweite Abschätzung für η erhalten wir wiederum mit $T_h(v_h) = v_h$ und (7.2) aus

$$\begin{aligned} |v_h(x) - v_{h,\Gamma}^{\infty}(x)| &= |T_h(v_h)(x) - v_{h,\Gamma}^{\infty}(x)| \\ &= |T_h(v_h)(x) - v_{h,\Gamma}^{\infty}(x) + T_h(v_{h,\Gamma}^{\infty})(x) - T_h(v_{h,\Gamma}^{\infty})(x)| \\ &\leq |T_h(v_{h,\Gamma}^{\infty})(x) - v_{h,\Gamma}^{\infty}(x)| + |T_h(v_h)(x) - T_h(v_{h,\Gamma}^{\infty})(x)| \\ &\leq \eta(x) + e^{-\delta h} \max_{y \in \Omega} |v_h(y) - v_{h,\Gamma}^{\infty}(y)| \end{aligned}$$

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

$$\|v_h - v_{h,\Gamma}^{\infty}\|_{\infty} \leq \eta + e^{-\delta h} \|v_h - v_{h,\Gamma}^{\infty}\|_{\infty}$$

und daraus

$$(1 - e^{-\delta h})\|v_h - v_{h,\Gamma}^{\infty}\|_{\infty} \leq \eta,$$

also die gewünschte Ungleichung. □

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 $v_{h,\Gamma}^{\infty}$ Hölder stetig ist, 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 7.1 angegebenen 5 Testpunkte als geeignet erwiesen. Dieses Muster lässt sich leicht auf höhere Dimensionen verallgemeinern.

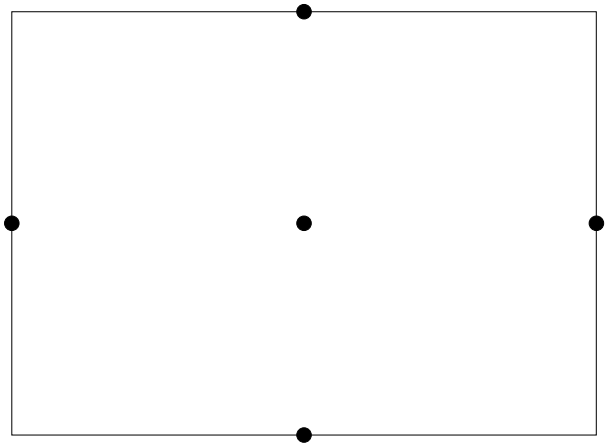


Abbildung 7.1: Testpunkte für die Auswertung von $\eta(x)$

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 viscocity 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. McEneaney et al., ed., Birkhäuser, Boston, 1999, pp. 288–303.

[6] 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.

[7] 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.

[8] L. Grüne, *Numerische optimale Steuerung und Stabilisierung.* Diplomarbeit, Institut für Mathematik, Universität Augsburg, 1994.

[9] ——, *An adaptive grid scheme for the discrete Hamilton-Jacobi-Bellman equation*, Numer. Math., 75 (1997), pp. 319–337.

[10] L. Grüne and P. E. Kloeden, *Higher order numerical schemes for affinely controlled nonlinear systems*, Numer. Math., (2001). Online version appeared April, 5. Printed version to appear.

[11] L. Grüne, M. Metscher, and M. Ohlberger, *On numerical algorithm and interactive visualization for optimal control problems*, Comput. Vis. Sci., 1 (1999), pp. 221–229.

[12] P. L. Lions, *Generalized solutions of Hamilton-Jacobi equations*, Pitman, London, 1982.

[13] M. L. Puterman and S. Brumelle, *On the convergence of policy iteration in stationary dynamic programming*, Math. of Operations Research, 4 (1979).

[14] A. Seeck, *Iterative Lösungen der Hamilton-Jacobi-Bellman-Gleichung bei unendlichem Zeithorizont*. Diplomarbeit, Universität Kiel, 1997.

[15] W. Semmler and M. Sieveking, *On optimal exploitation of interacting resources*, Journal of Economics, 59 (1994), pp. 23–49.

[16] E. D. Sontag, *Mathematical Control Theory*, Springer Verlag, New York, 2nd ed., 1998.

Index