Kapitel 10

Dynamische Programmierung I (deterministische Probleme)

10.1 Einleitung; Bellman-Prinzip und Bellman-Gleichung

Neben Variationsrechnung und Kontrolltheorie (nach Pontryagin) steht mit der *dynamischen Programmierung* (nach Bellman) ein auf den ersten Blick vollkommen andersartiges Konzept zur Behandlung dynamischer Optimierungsprobleme zur Verfügung. Obwohl die dynamische Programmierung gerade für *stochastische* Probleme (wo also Unsicherheit über die zukünftige Entwicklung der Zustandsvariable und damit auch die zukünftigen Erträge besteht) besser geeignet ist als die Kontrolltheorie, beschränken wir uns in diesem Kapitel auf deterministische Probleme. Da verschiedene Aspekte der dynam. Programmierung, wie z.B. ihr rekursiver Charakter, *closed-loop controls* usw., am deutlichsten bei Problemen mit einer diskreten Einteilung des Zeitintervalls $[0,T]$ in Perioden $t = 0, 1, \dots T-1$ hervortreten, beginnen wir mit *zeitdiskreten* Problemen der Form:

$$\max_{u_0,\dots,u_T \in \mathcal{U}} \sum_{t=0}^{T} f_t(x_t, u_t), \qquad x_0 \text{ gegeben}, \quad x_{t+1} = g_t(x_t, u_t) \ \ (t = 0, \dots, T-1)$$

statt der bisherigen zeitstetigen Probleme (die später im Kapitel behandelt werden):

$$\max_{u(t)\in\mathcal{U} \text{ für } t\in[0,T]} \int_0^T f\big(t, x(t), u(t)\big)\, dt + S\big(T, x(T)\big), \quad x(0) = x_0, \quad \dot{x}(t) = g\big(t, x(t), u(t)\big)$$

(Durchgängig bezeichnet x die Zustandsvariable, u die Kontrollvariable; bei der dynamischen Programmierung wird standardmäßig ein freies x_T bzw. $x(T)$ unterstellt).

Grundlegend für die dynamische Programmierung ist das **Bellman-Prinzip** (Bellman (1957)):

> An optimal policy has the property that, whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

Das heißt: Eine Lösung des „Gesamtproblems“ stellt ab jedem Zeitpunkt $t \in \{1, \dots, T\}$ eine Lösung des „Restproblems“ dar, das ausgeht vom zum Zeitpunkt t erreichten Zustand $x = x_t$. Gegeben ein Zeitpunkt t und ein zu diesem Zeitpunkt erreichter Zustand x, spielt es für die zukünftigen optimalen Entscheidungen $u_s, s \geq t$ also keine Rolle, wie man in den Zustand x gelangt ist. Dieses Prinzip gilt gleichermaßen für zeit-diskrete wie zeit-stetige Probleme. (Es gilt

auch für stochastische Probleme, sofern der stochastische Prozess, dem x_t folgt, die *Markov-Eigenschaft* hat). Das legt es nahe (bzw. erlaubt es bei stochastischen Problemen), das „vom Zeitpunkt t aus bestehende Restproblem" – und insbesondere dessen Optimalwert V_t – *als Funktion des Startzustands* x zu betrachten. Betrachtet man nun die Restprobleme in zwei aufeinanderfolgenden Zeitpunkten t und $t+1$ (bzw. t und $t+dt$ bei zeitstetigen Problemen), so kann man den Wert der Zielfunktion des Restproblems zur Zeit t zerlegen in den unmittelbaren Ertrag $f_t(x,u)$ (bzw. $f(t,x,u)\,dt$) und den Zielfunktionswert des Restproblems zum Zeitpunkt $t+1$ (bzw. $t+dt$). Die vom Zeitpunkt t aus zu treffenden Entscheidungen kann man analog unterteilen in die aktuelle Entscheidung $u=u_t$ und die zukünftigen Entscheidungen $u_s, s>t$. Nach dem Bellman-Prinzip spielt für die zukünftigen Entscheidungen aber nur der Ausgangszustand $\tilde{x}$ eine Rolle, der sich nach der aktuellen Entscheidung ergibt. Ist u die unmittelbare Entscheidung, so ist $\tilde{x}=g_t(x,u)$ und der Optimalwert des zukünftigen Problems lautet $V_{t+1}(g_t(x,u))$. Wenn die Funktion $V_{t+1}(\tilde{x})$ bekannt ist, bestimmt sich die unmittelbare Entscheidung u also durch Maximierung von $f_t(x,u)+V_{t+1}(g_t(x,u))$, und das Ergebnis dieser Maximierung ist gerade $V_t(x)$. Es besteht also folgende Beziehung zwischen den Funktionen V_t und V_{t+1}, die wir als **Bellman-Gleichung** (hier: eines zeitdiskreten deterministischen Problems) bezeichnen:[1]

$$V_t(x) \;=\; \max_{u\in\mathcal{U}} \big\{ f_t(x,u) + V_{t+1}\big(g_t(x,u)\big)\big\} \quad \text{[Bellman-Gl. im zeitdiskreten determin. Fall]} \quad (*)$$

Ausgehend von $V_T(x) = \max_{u\in\mathcal{U}} f_T(x,u)$ kann man mit der Beziehung $(*)$ *rekursiv* (in zeitlich absteigender Reihenfolge) $V_{T-1}(x), V_{T-2}(x), \ldots, V_0(x)$ bestimmen und erhält damit schließlich den Optimalwert $V(x)=V_0(x)$ des Ausgangsproblems als Funktion des Startzustands $x=x_0$.

Die Optimalwertfunktionen der Restprobleme, $V_{T-n}(x)$, an sich sind oft gar nicht so interessant. Was man mit der Bellman-Rekursion aber auch bekommt, ist eine Darstellung der optimalen Kontrolle $u=u_t$ zur Zeit t als Funktion des Zustands $x=x_t$ zu diesem Zeitpnkt: $u_t=u_t^*(x)$. (Dies enthält die eigentliche Bedeutung des Begriffs „*optimal policy*": Man sagt, welche Entscheidung u optimalerweise zu treffen ist, wenn man den Zustand x des Systems kennt). Man spricht dann auch von **closed-loop controls** oder **feedback controls**. Will man die **open-loop controls** $u=u_t$ bzw. $u=u(t)$, d.h. die optimale Kontrolle als Funktion der Zeit t (nicht als Funktion des Zustands x), wie sie in der Kontrolltheorie ermittelt werden, aus den closed loop controls bestimmen, so ist eine erneute Rekursion, diesmal in zeitlich aufsteigender Reihenfolge, erforderlich: Ausgehend von x_0 bestimmt man $u_0=u_0^*(x_0)$, dann $x_1=g_0(x_0,u_0)$, dann $u_1=u_1^*(x_1)$ usw.

Die Bellman-Gleichung erfasst sehr kompakt das Entscheidungsproblem, vor dem beispielsweise ein Unternehmen zum Zeitpunkt t steht, dessen Erträge (pro Periode) in den folgenden Perioden $s=t,t+1,\ldots,T$ als Funktion $f_s(x,u)$ des Zustands x des Unternehmens (beispielsweise Kapitalstock) und der Entscheidungsvariable u (z.B. Höhe der Investition pro Periode) durch $f_s(x,u)$ gegeben sind: Ist x der Zustand zum Zeitpunkt t, so muss die unmittelbare Entscheidung u die Summe aus unmittelbarem Ertrag, das ist $f_t(x,u)$, und dem Wert des Unternehmens, der sich nach der Entscheidung ergeben würde, das ist $V_{t+1}\big(x_{t+1}(x,u)\big)$, maximieren. Beim „Wert des Unternehmens, der sich nach der Entscheidung ergeben würde" muss das UN nämlich davon ausgehen, dass es auch ab der Periode $t+1$ optimal entscheidet – und dies ist in V_{t+1} erfasst. Dass V_{t+1} im Zustand $x_{t+1}(x,u)$ ausgewertet werden muss, berücksichtigt, dass die unmittelbare Entscheidung $u=u_t$ einen Effekt auf den Zustand x_{t+1} zu Beginn der nächsten Periode hat.

[1] Analoge Beziehungen gelten auch für zeitstetige und stochastische Probleme; bei zeitstetigen Problemen erhält man i.w. eine partielle Differentialgleichung für die Funktion $V(t,x)$, die dann als Hamilton-Jacobi-Bellman-Gleichung bezeichnet wird. Die Bellman-Gleichung des Problems stellt jeweils Dreh- und Angelpunkt der dynamischen Programmierung dar und wird auch als die **fundamentale Gleichung für Optimalität** bezeichnet.

10.2 Bellman-Gleichung für zeitdiskrete determinist. Probleme

10.2.1 Formulierung im Gegenwartswert (present value)

Wir betrachten in diesem Abschnitt folgendes dynamische Optimierungsproblem in der Zustandsvariable $\boldsymbol{x} \in \mathbb{R}^n$ und der Entscheidungsvariable $\boldsymbol{u} \in \mathbb{R}^m$ (da es prinzipiell und auch formal überhaupt keinen Unterschied macht, lassen wir gleich mehrdimensionale Zustands- und Kontrollvariablen zu, obwohl in den einfachsten Beispielen immer $n = m = 1$ sein wird):

$$\max_{\boldsymbol{u}_t \in \mathcal{U}_t,\ t=0,\dots,T} \sum_{t=0}^{T} f_t(\boldsymbol{x}_t, \boldsymbol{u}_t) \quad \text{unt. d. NB} \quad \begin{cases} \boldsymbol{x}_0 \text{ gegeben,} \\ \boldsymbol{x}_{t+1} = \boldsymbol{g}_t(\boldsymbol{x}_t, \boldsymbol{u}_t) \ \text{ für } \ t = 0, \dots, T-1 \end{cases} \tag{1}$$

Hierbei sind, für $t = 0, \dots, T-1$, $f_t(\boldsymbol{x}, \boldsymbol{u}) \in \mathbb{R}$ und $\boldsymbol{g}_t(\boldsymbol{x}, \boldsymbol{u})$ gegebene Funktionen, die den „Momentanertrag in der Periode von t zu $t+1$“ (f_t) bzw. die „Bewegungsgleichung der Zustandsvariable“ ($\boldsymbol{g}_t$) beschreiben. Man beachte, dass $\boldsymbol{g}_t$ hier nicht den Zuwachs $\Delta\boldsymbol{x}_t = \boldsymbol{x}_{t+1} - \boldsymbol{x}_t$, sondern direkt $\boldsymbol{x}_{t+1}$ wiedergibt. Die Funktion $f_T(\boldsymbol{x}, \boldsymbol{u})$ entspricht dem früheren „Terminalwert“ $S(T, \boldsymbol{x})$, der hier auch von der Kontrollgröße $\boldsymbol{u}$ zum Zeitpunkt T abhängen kann. Anders als in der Kontrolltheorie gilt das Folgende auch dann ohne Modifikationen, wenn die „Kontrollbereiche“ $\mathcal{U}_t \subset \mathbb{R}^m$ vom Zustand $\boldsymbol{x}$ zur Zeit t abhängen (d.h. die Kontrollvariablenrestriktion könnte auch $\boldsymbol{u}_t \in \mathcal{U}_t(\boldsymbol{x}_t)$ lauten). Man beachte, dass wir das Problem als Funktion des Anfangswertes $\boldsymbol{x}$ der Zustandsvariable betrachten.

Lösungskonzept: Gesucht sind Funktionen $\boldsymbol{u}_t^*(\boldsymbol{x})$, so dass man mit der Rekursion $\boldsymbol{u}_t = \boldsymbol{u}_t^*(\boldsymbol{x}_t)$, $\boldsymbol{x}_{t+1} = \boldsymbol{g}_t(\boldsymbol{x}_t, \boldsymbol{u}_t)$ das Problem (1) für jeden Anfangswert $\boldsymbol{x}_0$ löst. Wir nennen dies eine **Markov-Politik**-Lösung des Problems (obwohl der Begriff bei determinist. Problemen unangebracht erscheint; der wichtige Punkt bei einer ‘Markov-policy’ ist, dass die aktuelle Entscheidung $\boldsymbol{u}$ auf Basis des aktuellen Systemzustands $\boldsymbol{x}$ getroffen werden kann – die Vergangenheit von $\boldsymbol{x}$ spielt dafür keine Rolle).

Zur Ermittlung der optimalen Politiken, d.h. der Funktionen $\boldsymbol{u}_t^*(\boldsymbol{x})$, betrachtet die dynamische Programmierung den Optimalwert des Restproblems, das sich vom Zeitpunkt t an noch stellt, als Funktion von dessen Initialzustand $\boldsymbol{x} = \boldsymbol{x}_t$:

$$V_t(\boldsymbol{x}) := \max_{\boldsymbol{u}_s \in \mathcal{U}_s,\ s=t,\dots,T} \sum_{s=t}^{T} f_s(\boldsymbol{x}_s, \boldsymbol{u}_s) \quad \text{unt. d. NB} \quad \begin{cases} \boldsymbol{x}_t = \boldsymbol{x} \text{ und für } s = t, \dots, T-1: \\ \boldsymbol{x}_{s+1} = \boldsymbol{g}_s(\boldsymbol{x}_s, \boldsymbol{u}_s) \end{cases} \tag{2}$$

Das Bellman-Prinzip besagt gerade, dass man mit der Markov-Politik-Lösung $\boldsymbol{u}_t^*(\boldsymbol{x}), \dots, \boldsymbol{u}_T^*(\boldsymbol{x})$ des Problems (2) die optimalen Politiken des gesamten, vom Zeitpunkt 0 an betrachteten Problems (1), im restlichen Zeitraum $t, \dots, T$ bekommt. Damit lässt sich zeigen:

Satz 10.1 (Bellman-Gleichung als notwend. und hinreich. Bed. für Optimalität)
a) *Sofern die Optimalwertfunktionen $V_t(\boldsymbol{x})$ aus (2) existieren (d.h. $V_t(\boldsymbol{x}) < \infty\ \forall\, t, \boldsymbol{x}$), erfüllt jede Markov-Politik-Lösng $\boldsymbol{u}_t^*(\boldsymbol{x})$ ($t = 0, \dots, T$) des Optimierungsproblems (1) die Bedingungen*

$$\begin{aligned} \boldsymbol{u}_T^*(\boldsymbol{x}) &= \arg \max_{\boldsymbol{u} \in \mathcal{U}_T(\boldsymbol{x})} f_T(\boldsymbol{x}, \boldsymbol{u}) \\ \boldsymbol{u}_t^*(\boldsymbol{x}) &= \arg \max_{\boldsymbol{u} \in \mathcal{U}_t(\boldsymbol{x})} \big\{ f_t(\boldsymbol{x}, \boldsymbol{u}) + V_{t+1}\big(\boldsymbol{g}_t(\boldsymbol{x}, \boldsymbol{u})\big) \big] \big\}, \ t = T\text{-}1,\ T\text{-}2, \dots, 0 \end{aligned} \tag{ARG}$$

und die Optimalwertfunktionen $V_t(\boldsymbol{x})$ erfüllen die Bellman-Gleichung

$$\begin{aligned} V_T(\boldsymbol{x}) &= \max_{\boldsymbol{u} \in \mathcal{U}_T(\boldsymbol{x})} f_T(\boldsymbol{u}, \boldsymbol{x}) \\ V_t(\boldsymbol{x}) &= \max_{\boldsymbol{u} \in \mathcal{U}_t(\boldsymbol{x})} \big\{ f_t(\boldsymbol{x}, \boldsymbol{u}) + V_{t+1}\big(\boldsymbol{g}_t(\boldsymbol{x}, \boldsymbol{u})\big) \big\}, \quad t = T\text{-}1,\ T\text{-}2, \dots, 0 \end{aligned} \tag{BGL}$$

b) *Wenn Funktionen $V_t(\boldsymbol{x})$ ($< \infty\ \forall\, t, \boldsymbol{x}$) existieren, die die Bellman-Gleichungen (BGL) erfüllen, und durch (ARG) eine Markov-Politik definiert wird, dann ist diese Politik eine Lösung des Optimierungsproblems (1).*

Die Bellman-Gl. (BGL) reduziert das dynamische Optimierungsproblem auf eine **Sequenz statischer Optimierungsprobleme**, die in zeitlich negativer Richung zu durchlaufen ist.

Beispiel 1:

$$\max_{u_t\in\mathbb{R}} \sum_{t=0}^{T} \left(x_t - u_t^2\right) \quad \text{unt. d. NB} \quad \begin{cases} x_0 \text{ gegeben,} \\ x_{t+1} \;=\; x_t + u_t \;\; \text{für} \;\; t = 0,\ldots,T-1 \end{cases}$$

Hier ist $f_t(x,u) = x - u^2$ und $g_t(x,u) = x + u$ für $t = 0,\ldots,T-1$ sowie $f_T(x,u) = x^2 - u^2$. Wir bestimmen zunächst $V_T(x)$, $V_{T-1}(x)$, $V_{T-2}(x)$ und $V_{T-3}(x)$. Dabei wird sich das Bildungsgesetz für $V_{T-n}(x)$ abzeichnen, das sich dann durch vollständige Induktion über n nachweisen lässt:
$V_T(x)$: Da $f_T(x,u) = x - u^2$, bestimmt sich $V_T(x)$ als

$$V_T(x) \;=\; \max_{u\in\mathbb{R}} \left\{ f_T(x,u) \right\} \overset{\text{(hier)}}{=} \max_{u\in\mathbb{R}} \left\{ x - u^2 \right\}$$

Das Maximum über $u \in \mathbb{R}$ von $h(u) := x - u^2$ (mit x als Parameter) wird offensichtlich in $u = 0$ angenommen. Also ist $u_T^*(x) = 0$ und

$$V_T(x) \;=\; x$$

$V_{T-1}(x)$: Gemäß Bellman-Gl. ergibt sich $V_{T-1}(x)$ als:

$$\begin{aligned} V_{T-1}(x) \quad &= \quad \max_{u\in\mathbb{R}} \left\{ f_{T-1}(x,u) + V_T\big(g_{T-1}(x,u)\big) \right\} \\ &\overset{\text{(hier)}}{=} \max_{u\in\mathbb{R}} \left\{ \;\; x - u^2 \;\; + \;\; x + u \;\; \right\} \;=\; \max_{u\in\mathbb{R}} \left\{ 2x - u^2 + u \right\} \end{aligned}$$

Das Maximum über $u \in \mathbb{R}$ der Parabel $h(u) := 2x - u^2 + u$ (mit x als Parameter) wird dort angenommen, wo $h'(u) = -2u + 1 = 0$ ist, also $u = \frac{1}{2}$ (und somit $u_{T-1}^*(x) = \frac{1}{2}$). Somit ist

$$V_{T-1}(x) \;=\; 2x - \left(\tfrac{1}{2}\right)^2 + \tfrac{1}{2} \;=\; 2x + \tfrac{1}{4}$$

$V_{T-2}(x)$: Gemäß Bellman-Gl. ergibt sich $V_{T-2}(x)$ als:

$$\begin{aligned} V_{T-2}(x) \quad &= \quad \max_{u\in\mathbb{R}} \left\{ f_{T-2}(x,u) + V_{T-1}\big(g_{T-2}(x,u)\big) \right\} \\ &\overset{\text{(hier)}}{=} \max_{u\in\mathbb{R}} \left\{ x - u^2 + 2(x+u) + \tfrac{1}{4} \right\} \;=\; \max_{u\in\mathbb{R}} \left\{ 3x - u^2 + 2u + \tfrac{1}{4} \right\} \end{aligned}$$

Das Maximum über $u \in \mathbb{R}$ der Parabel $h(u) := 3x - u^2 + 2u + \frac{1}{4}$ (mit x als Parameter) wird dort angenommen, wo $h'(u) = -2u + 2 = 0$ ist, also $u = 1$ (und somit $u_{T-2}^*(x) = 1$). Also ist

$$V_{T-2}(x) \;=\; 3x - 1^2 + 2 \cdot 1 + \tfrac{1}{4} \;=\; 3x + \tfrac{5}{4}$$

$V_{T-3}(x)$: Gemäß Bellman-Gl. ergibt sich $V_{T-3}(x)$ als:

$$\begin{aligned} V_{T-3}(x) \quad &= \quad \max_{u\in\mathbb{R}} \left\{ f_{T-3}(x,u) + V_{T-2}\big(g_{T-2}(x,u)\big) \right\} \\ &\overset{\text{(hier)}}{=} \max_{u\in\mathbb{R}} \left\{ x - u^2 + 3(x+u) + \tfrac{5}{4} \right\} \;=\; \max_{u\in\mathbb{R}} \left\{ 4x - u^2 + 3u + \tfrac{5}{4} \right\} \end{aligned}$$

Das Maximum über $u \in \mathbb{R}$ der Parabel $h(u) := 4x - u^2 + 3u + \frac{5}{4}$ (mit x als Parameter) wird dort angenommen, wo $h'(u) = -2u + 3 = 0$ ist, also $u = \frac{3}{2}$ (und somit $u_{T-3}^*(x) = \frac{3}{2}$). Also ist

$$V_{T-3}(x) \;=\; 4x - \left(\tfrac{3}{2}\right)^2 + 3 \cdot \tfrac{3}{2} + \tfrac{5}{4} \;=\; 4x - \tfrac{9}{4} + \tfrac{9}{2} + \tfrac{5}{4} \;=\; 4x + \tfrac{9}{4} + \tfrac{5}{4} \;=\; 4x + \tfrac{14}{4}$$