§4 Die Politik-Iteration für die Lösung von sequentiellen Entscheidungsprozessen

Die Politik-Iteration, die wir beschreiben werden, wird die optimale Politik mit einer kleinen Anzahl von Iterationen feststellen. Sie besteht aus zwei Teilen, der Wertbestimmung und der Verbesserung der Politik.

4.1 Optimale Strategien

Betrachten wir nun eine ergodische Markov-Kette mit einer ergodischen Klasse und dem Zustandsraum E mit N Zuständen, die beschrieben wird durch eine Übergangs-Matrix **P** und eine Bewerungs-Matrix **C**. Der zu erwartende gesamte Erlös hängt von der totalen Anzahl der Übergänge ab, die im System vorkommen, so dass diese Grösse über alle Grenzen wächst, wenn die Zahl der Übergänge zunimmt. Eine dienlicherwe Grösse stellt der Durchschnittserlös des Prozesses pro Zeiteinheit dar. Im Kapitel 2 wurde gezeigt, dass diese Grösse sinvoll ist, wenn der Prozess viele Übergänge enthält; wir nannten sie den Gewinn des Prozesses.

Da der Prozess eine ergodische Klasse hat, sind die Grenzwahrscheinlichkeiten π_i, $i \in E$, unabhängig vom Anfangszustand, und als Gewinn g des Systems erhalten wir

$$g = \sum_{i=1}^{N} \pi_i \, q_i,$$

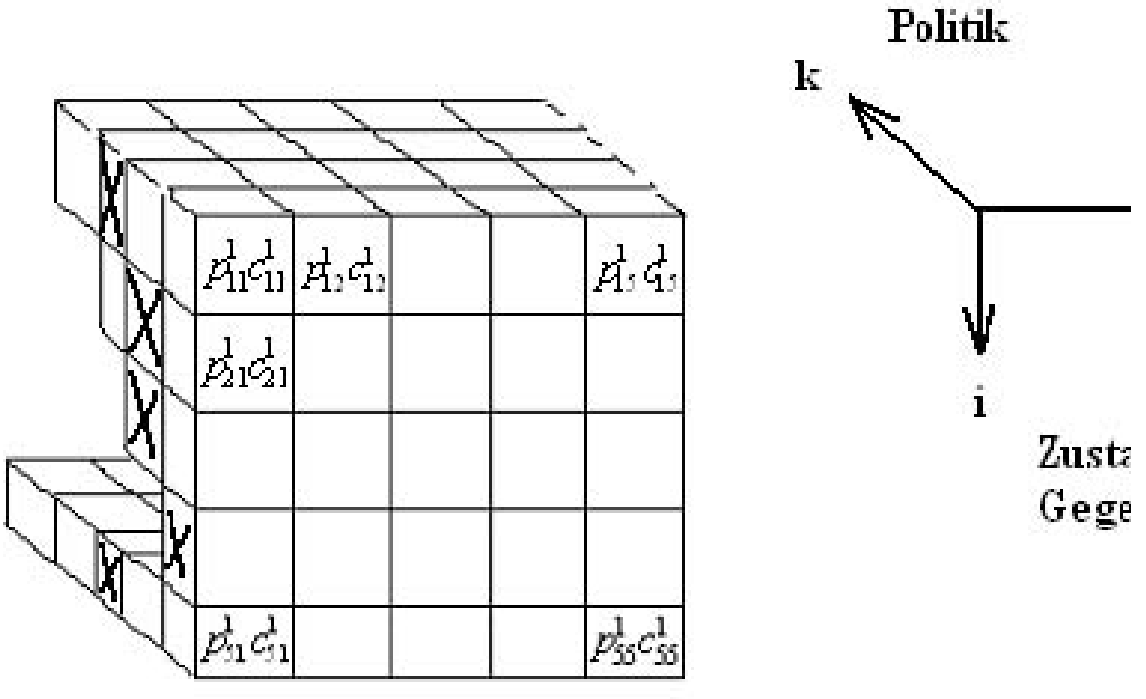
wobei q_i den zu erwartenden unmittelbaren Erlös im Zustand $i \in E$ darstellt und durch die Gleichung (i)

$$q_i = \sum_{j=1}^{N} p_{i\,j} \, c_{i\,j}, \quad i \in E$$

definiert ist.

Jede bewertete Markov-Kette mit einer ergodischen Klasse weist einen durch die Gleichung $g = \sum_{i=1}^{N} \pi_i \, q_i$ bestimmten Gewinn auf. Wenn wir mehrere solche Prozesse haben und wissen möchten, welcher davon auf lange Sicht der vorteilhafteste ist, bestimmen wir die Zugehörigen Gewinne und wählen den Prozess mit dem höchsten Gewinn.

Die sequentielle Entscheidungsprozess im Kapitel 3 zieht viele denkbare Prozesse in Betracht, weil die Strategien in jedem Zustand frei gewählt werden können. Um diese zu illustrieren, bertachten wir die dre-dimensionale Anordnung in nächter Figur, die graphisch die Zustände und Strategien darstellt.



Diese Anordnung veranschaulicht ein Problem mit fünf Zuständen, das vier Strategien im ersten Zustand zulässt, drei im zweiten, zwei im dritten, eine im vierten und fünf im fünften. Auf der Vorderseite die Parameter für die esrte Strategie in jedem Zustand eingezeichnet, die nächste Reihe der Anordnung enthält die Parameter für die Zweite Strategie in jedem Zuständ, usw.

Ein "**X**" gibt an, dass wir eine Spezielle Strategie in einem Zustand mit einer Wahrscheinlichkeits- und Erlösverteilung gewählt haben, welche das Verhalten des Systems zu jeder Zeit bei Eintritt in den betreffenden Zustand beherrscht.

> **4.1.1 Definition:** Die so gewählte Strategie für diesen Zustand nennt man Entscheidung in diesem Zustand. Sie ist nun keine Funktion von *n* mehr.
>
> Alle "**X**" zusammen oder die Entscheidungen aller Zustände nennt man eine Politik.

Die Auswahl einer Politik bestimmt somit die bewertete Markov-Kette, die die Vorgänge im System beschreibt. Die im Diagramm markierte Politik verlangt, dass die Wahsrcheinlichkeits- und Bewertungsmatrizen für das System aus der ersten Strategie im Zustand 4, der zweiten Strategie in den Zuständen 2 und 3 und der dritten Strategie in den Zuständen 1 und 5 bestimmt werden. Man kann die Politik durch einen Entscheidungsvektor ***d***, dessen Elemente die Zahl der in jedem Zustand gewählte Strategie bestimmen, darstellen. In diesem Fall ist also

$$\mathbf{d} = \begin{pmatrix} 3 \\ 2 \\ 2 \\ 1 \\ 3 \end{pmatrix}, \mathbf{P} = \begin{pmatrix} p_{11}^3 & p_{12}^3 & p_{13}^3 & p_{14}^3 & p_{15}^3 \\ p_{21}^2 & p_{22}^2 & p_{23}^2 & p_{24}^2 & p_{25}^2 \\ p_{31}^2 & p_{32}^2 & p_{33}^2 & p_{34}^2 & p_{35}^2 \\ p_{41}^1 & p_{42}^1 & p_{43}^1 & p_{44}^1 & p_{45}^1 \\ p_{51}^3 & p_{52}^3 & p_{53}^3 & p_{54}^3 & p_{55}^3 \end{pmatrix}, \mathbf{C} = \begin{pmatrix} c_{11}^3 & c_{12}^3 & c_{13}^3 & c_{14}^3 & c_{15}^3 \\ c_{21}^2 & c_{22}^2 & c_{23}^2 & c_{24}^2 & c_{25}^2 \\ c_{31}^2 & c_{32}^2 & c_{33}^2 & c_{34}^2 & c_{35}^2 \\ c_{41}^1 & c_{42}^1 & c_{43}^1 & c_{44}^1 & c_{45}^1 \\ c_{51}^3 & c_{52}^3 & c_{53}^3 & c_{54}^3 & c_{55}^3 \end{pmatrix}.$$

Eine optimale Politik wird als eine Politik definiert, die den Gewinn oder den durschnittlichen Ertrag pro Zeiteinheit maximiert. In dem Problem mit fünf Zuständen (in vorieger Figur graphisch dargestellt) gibt es

$$4 \cdot 3 \cdot 2 \cdot 1 \cdot 5 = 120$$

verschiedene Politikrn. Es ist denkbar, dass wir den Gewinn für jede dieser Politiken feststellen können, um so die Politik mit dem höchsten Gewinn zu ermitteln. Wenn dies nurn auch für 120 Politiken durchführbar sein mag, wird das nicht mehr möglich sein für sehr grosse Probleme. Zum Beispiel würde ein Problem mit 50 Zuständen und 50 Strategien in jedem Zustand

$$50^{50} \approx 10^{85}$$

Politiken enthalten.

Die Politik-Iteration, die wir beschreiben werden, wird die optimale Politik mit einer kleinen Anzahl von Iterationen feststellen. Sie besteht aus zwei Teilen, der Wertbestimmung und der Verbesserung der Politik. Zuerst werden wir die Wertbestimmung diskutieren.

4.2 Die Wertbestimmung

Wie nehmen an, dass wir ein System vor uns haben mit einer gegebenen Politik, so dass wir eine gegebene bewertete Markov-Kette spezifizieren haben. Wenn diesem Prozess gestattet würde, während n Stufen oder Übergängen zu arbeiten, könnten wir $v_i(n)$, $i \in E$, bestimmen als den total zu erwartenden Erlös, den das System in $n \in \mathbb{N}$ Schritten einbringt, wenn es vom Zustand $i \in E$ aus unter der gegebenen Politik startet.

Die Grösse $v_i(n)$, $i \in E$, muss der rekursiven Relation (ii) oder (iii) (Abschnitt 2.2), die im Kapitel 2 abgeleitet wurde, genügen,

$$v_i(n) = q_i + \sum_{j=1}^{N} p_{i\,j}\, v_j(n-1), \;\; i \in E, \;\; n \in \mathbb{N}.$$

In dieser Gleichung benötigen wir den oberen Index k nicht, weil durch die Feststellung einer Politik die Wahrscheinlichkeits- und Erlösmatrizen, die das System beschreiben, bestimmt sind.

In Kapitel 2 wurde gezeigt, dass für die Markov-Kette mit einer ergodischen Klasse $v_i(n)$, $i \in E$, die asymptotische Form hat

$$v_i(n) = n\,g + v_i, \;\; i \in E, \;\; n \to \infty.$$

In diesem Kapitel interessieren wir uns nur für Systeme, welche eine sehr grosse Anzahl von Übergängen aufweisen. Folglich ist es gerechtfertigt

4.2.1 Satz: Für die Werte v_i, $i \in E$, gilt die folgende Gleichung

(xi) $$v_i = q_i + \sum_{j=1}^{N} p_{i\,j}\, v_j - g, \;\; i \in E.$$

▼

Beweis: Setzt man die Gleichung $v_i(n) = n\,g + v_i$ in die Gleichung $v_i(n) = q_i + \sum_{j=1}^{N} p_{i\,j}\, v_j(n-1)$, erhält man also

$$n\,g + v_i = q_i + \sum_{j=1}^{N} p_{i\,j}((n-1)\,g + v_j), \;\; i \in E$$

$$n\,g + v_i = q_i + (n-1) \sum_{j=1}^{N} p_{i\,j}\, g + \sum_{j=1}^{N} p_{i\,j}\, v_j.$$

Da

$$\sum_{j=1}^{N} p_{i\,j} = 1,$$

ergeben diese Gleichungen

$$n\,g + \mathrm{v}_i = (n-1)\,g + \sum_{j=1}^{N} p_{i\,j}\,\mathrm{v}_j \text{ oder}$$

$$\mathrm{v}_i = q_i + \sum_{j=1}^{N} p_{i\,j}\,\mathrm{v}_j - g,\ \ i \in E.$$

Wir haben nun ein System von N linearen Gleichungen erhalten, welches die Grössen v_i, $i \in E$, und g in Beziehung mit der Wahrscheinlichkeits- und Bewertungsstruktur des Prozesses bringt. Jedoch hat man N Unbekannte v_i, $i \in E$, und eine Unbekannte g, also insgesamt $N+1$ Unbekannte. Man wird das Wesen dieser Schwierigkeit erkennen, wenn eine Konstante a zu sämtlichen v_i in der Gleichung (xi) addiert wird und wir das Resultat prüfen.

4.2.2 Satz: Eines der Werte v_i, $i \in E$, kann eine Konstante gesetzt werden.

▼

Beweis: Addieren wir die Konstante a zu sämtlichen v_i in der Gleichung (xi)

$$\mathrm{v}_i + a = q_i + \sum_{j=1}^{N} p_{i\,j}(\mathrm{v}_j + a) - g \text{ oder}$$

$$\mathrm{v}_i = q_i + \sum_{j=1}^{N} p_{i\,j}\,\mathrm{v}_j - g.$$

Wiederum erhalten wir die ursprünglichen Gleichungen, so dass der absolute Wert von v_i, $i \in E$, durch diese Gleichungen nicht bestimmt werden kann.

Wie dem auch sei, wenn wir eines der v_i gleich Null setzen, zum Beispiel v_N, dann sind nur N Unbekannte vorhanden, und die Gleichungen (xi) können für g und die verbleibenden v_i, $i \in E\backslash\{N\}$, gelöst werden.

4.2.3 Bemerkung: Man beachte dass die so erhaltenen v_i, $i \in E$, nicht diejenigen sind, die durch die Gleichung $\mathrm{v}_i(n) = n\,g + \mathrm{v}_i$ bestimmt wurden, sondern sich von jenen um einen konstanten Betrag unterscheiden. Da die exakten Werte v_i gemäss Gleichung (vi) eine Konstante Grösse

$$\sum_{i=1}^{N} \pi_i\,\mathrm{v}_i(0)$$

enthalten, haben sie aber in Prozessen mit sehr vielen Übergängen keine wesentliche Bedeutung.

4.2.4 Definition: Die v_i, $i \in E$, welche wir durch Lösung der Gleichungen (xi) mit $\mathrm{v}_N = 0$ erhalten, genügen für unseren Zweck. Wir nennen sie die relative Werte der Politik.

Den relativen Werten kann eine physikalische Interpretation gegeben werden. Man beachte die ersten beiden Zustände, 1 und 2. Für irgendein grosses n ergibt die Gleichung (vii) für die Markov-Kette mit einer ergodischen Klasse

$$\mathrm{v}_1(n) = n\,g + \mathrm{v}_1,\ \ \mathrm{v}_2(n) = n\,g + \mathrm{v}_2.$$

Also ist

$$\mathrm{v}_1(n) - \mathrm{v}_2(n) = \mathrm{v}_1 - \mathrm{v}_2$$

für grosse Werte von n. Diese Differenz ist gleich dem Zuwachs des vom System zu erwartenden Erlöses, wenn

man im Zustand 1 statt im Zustand 2 startet. Da Diese Differenz

$$v_1 - v_2$$

unabhängig von jeglichem absoluten Niveau ist, können die relativen Werte verwendet werden, um die Differenz festzustellen. Mit anderen Worten, die Differenz der relativen Werte der beiden Zustände

$$v_1 - v_2$$

ist gleich dem Betrag, den ein vernünftiger Mann gerade bereit wäre, zu zahlen, um den Prozess im Zustand 1 statt 2 zu starten, wenn er die Absicht hat, in dem System über viele Übergänge hinweg zu operieren.

Wir werden diese Interpretation der relativen Werte im Kapitel 5 an Beispielen auswerten.

4.2.5 Satz: Wenn die Gleichungen (xi) mit π_i, $i \in E$, der Grenzwahrscheinlichkeit des i-ten Zustandes, multipliziert und uber i summiert werden, erhält man einen ausdruck, der zur Gleichung

$$g = \sum_{i=1}^{N} \pi_i q_i$$

äquivalent ist.

▼

Beweis: Multiplizieren wir die Gleichungen (xi) mit π_i, $i \in E$, und dann summieren über i, erhalten wir

$$\sum_{i=1}^{N} \pi_i v_i = \sum_{i=1}^{N} \pi_i q_i + \sum_{i=1}^{N} \sum_{j=1}^{N} \pi_i p_{i\,j} v_j - \sum_{i=1}^{N} \pi_i g.$$

Da

$$\pi_j = \sum_{i=1}^{N} \pi_i p_{i\,j} \text{ und } \sum_{i=1}^{N} \pi_i = 1,$$

erhält man also

$$\sum_{i=1}^{N} \pi_i v_i = \sum_{i=1}^{N} \pi_i q_i + \sum_{j=1}^{N} \pi_j v_j - g \text{ oder}$$

$$g = \sum_{i=1}^{N} \pi_i q_i$$

Eine wichtige Frage zu diesem Punkt ist nun: Wenn wir nur den Gewinn der gegebenen Politik suchen, warum verwenden wir dann nicht die Gleichung $g = \sum_{i=1}^{N} \pi_i q_i$ statt der Gleichung (xi)? In der Tat, warum plagen wir uns überhaupt ab, Dinge, wie relative Werte, zu bestimmen? Die Antwort lautet zunächst: Obschon die Gleichung $g = \sum_{i=1}^{N} \pi_i q_i$ den Gewinn des Prozesses feststellt, zeigt sie uns nicht, wie man eine bessere Politik finden kännte. Wir werden noch sehen, dass die relativen Werte den Schlüssel enthalten, bessere und noch bessere Politiken festzustellen, um zuletzt die beste Politik anzugeben.

Eine zweite Antwort besagt, dass der erforderliche Rechenaufwand für die Lösung der Gleichungen (xi) für Gewinn und relative Werte ungefäher der gleiche ist wir der, der nötwendig ist, um die Grenzwahrscheinlichkeiten unter Verwendung der Gleichungen

$$\boldsymbol{\pi} = \boldsymbol{\pi}\,\mathbf{P} \text{ und } \quad \sum_{i \in E} \pi_i = 1$$

festzustellen, da beide Berechnungen die Lösung von N linearen Gleichungen erfordern. Vom Gesichtspunkt der Feststellung des Gewinns aus sind die Gleichungen

$$g = \sum_{i=1}^{N} \pi_i q_i$$

und

$$v_i = q_i + \sum_{j=1}^{N} p_{i\,j} v_j - g, \; i \in E$$

gleichwertig. Dennoch ziehen wir die Gleichungen (xi) vor, weil sie die relativen Werte liefern, die, wie wir

zeigen werden, notwendig für die Verbesserung der Politik sind.

Vom Standpunkt der Berechnungen aus gesehen ist es interessant, zu bemerken, dass wir wegen der Linearität der Gleichungen (xi) beachtliche Freiheit haben, unsere Bewertungen festzusetzen. Wenn die Erlöse $c_{i\,j}$

4.2.6 Satz: Wenn die Erlöse $c_{i\,j}$ eines Prozesses mit Gewinn g und relativen Werten v_i, $i \in E$, durch eine lineare Transformation

$$c'_{i\,j} = a\, c_{i\,j} + b$$

dann die Gleichungen (xi) ergeben

$$\mathrm{v}'_i = q'_i + \sum_{j=1}^{N} p_{i\,j}\, \mathrm{v}'_j - g', \; i \in E.$$

Der Gewinn

$$g' = a\, g + b,$$

die relative Werte

$$\mathrm{v}'_i = a\, \mathrm{v}_i, \; i \in E.$$

▼

Beweis: Wegen

$$q_i = \sum_{j=1}^{N} p_{i\,j}\, c_{i\,j},$$

die neuen erwarteten unmittelbaren Erlöse betragen

$$q'_i = a\, q_i + b.$$

Die Gleichungen (xi) ergeben

$$\mathrm{v}_i = \frac{q'_i - b}{a} + \sum_{j=1}^{N} p_{i\,j}\, v_j - g, \; i \in E \text{ oder}$$

$$a\, \mathrm{v}_i = q'_i + \sum_{j=1}^{N} p_{i\,j}\, a\, v_j - a\, g - b \text{ und}$$

$$\mathrm{v}'_i = q'_i + \sum_{j=1}^{N} p_{i\,j}\, \mathrm{v}'_j - g', \; i \in E.$$

Der Gewinn g' des Prozesses mit transformierten Erlösen ist somit $a\, g + b$, wobei die Werte v'_i in diesem Prozess $a\, \mathrm{v}_i$ gleichkommen.

Die Auswirkung der Änderungen der Masseinheiten und des anbsoluten Niveaus des Bewertungssystems bei dem Gewinn und den relativen Werten ist leicht zu berechnen. So können wir alle Erlöse so normieren, dass sie zwischen 0 und 1 liegen, den ganzen sequentiellen Entscheidungsprozess lösen, und dann das Inverse unserer ursprünglichen Transformation verwenden, um den Gewinn und die relativen Werte auf ihre ursprüngliche Höhe zurückzubringen.

Wir haben nun gezeigt, dass wir durch die Lösung von N linearen Gleichungen (xi) mit $\mathrm{v}_N = 0$ für eine gegebene Politik den Gewinn und die relativen Werte dieser Politik feststellen können. Nun werden wir zeigen, wie die relativen Werte benützt werden, um eine Politik mit höherem Gewinn als dem der ursprünglichen Politik festzustellen.

4.3 Die Verbesserung der Politik

Im Kapitel 3 haben wir gesehen, dass, wenn wir bis zur n-ten Stufe eine optimale Politik haben, wir die beste aller Strategien im i-ten Zustand auf $(n+1)$-ten Stufe durch Maximieren von

$$q_i^k + \sum_{j=1}^N p_{i\,j}^k \, \mathrm{v}_j(n)$$

bestimmen können. Für grosse n können wir Gleichung

$$\mathrm{v}_i(n) = n\,g + v_i, \;\; i \in E$$

einsetzen und erhalten

$$q_i^k + \sum_{j=1}^N p_{i\,j}^k \left(n\,g + \mathrm{v}_j\right)$$

als Testgrösse, die in jedem Zustand zu maximieren ist. Da

$$\sum_{j=1}^N p_{i\,j}^k = 1 \text{ für jeden Zustand } i \in E,$$

wird der Beitrag $n\,g$ und irgendeiner additiven Konstanten in v_j ein Teil der Testgrösse, welcher unabhängig von k ist. Also können wir, wenn wir unsere Entscheidung im Zustand $i \in E$ treffen,

$$q_i^k + \sum_{j=1}^N p_{i\,j}^k \, \mathrm{v}_j$$

maximieren in bezug auf die Strategien im i-ten Zustand. Ferner können wir relativen Werte benützlich (gegeben durch die Gleichung (xi)) für die Politik, die bis zur n-ten Stufe angewandt wurde.

4.3.1 Bemerkung: Das Verfahren der Politik-Verbesserung könnte man wie folgt zusammenfassen: Für jeden Zustand $i \in E$ stelle man die Strategie $k \in A(i)$ fest, welche die Testgrösse

$$q_i^k + \sum_{j=1}^N p_{i\,j}^k \, \mathrm{v}_j$$

maximiert, wobei die unter der alten Politik bestimmten relativen Werte verwendet werden. Aus dieser Strategie k wird nun d_i, $i \in E$, die Entscheidung im i-ten Zustand. Wenn dieses Verfahren für jeden Zustand durchgeführt worden ist, haben wir eine Politik bestimmt.

Wir haben nun durch etwas heuristische Mittel eine Methode zur Feststellung einer Politik, welche gegenüber unserer ursprünglichen Politik eine Verbesserung darstellt, beschrieben. Bald werden wir beweisen, dass die neue Politik einen höheren Gewinn als die alte Politik aufweisen wird. Doch vorerst werden wir zeigen, wie die Wertbestimmung und das Verfahren zur Verbesserung der Politik in einem Iterationszyklus verbunden werden, dessen Ziel das Auffinden der Politik mit dem höchsten Gewinn unter sämtlichen möglichen Politiken ist.

4.4 Der Iterations-Zyklus

Der grundliegende Iterationszyklus ist in nächstem Satz beschrieben.

4.4.1 Satz: Der Iterationszyklus für die Ermittlung der optimalen Politik,

Schritt 0 (Initialisierung): Fixieren wir eine Anfangspolitik **d**

Schritt 1 (Wertbestimmung): Man verwende $p_{i\,j}$ und q_i für die gegebene Politik und löse

$$\mathrm{v}_i = q_i + \sum_{j=1}^N p_{i\,j} \, \mathrm{v}_j - g, \;\; i \in E$$

für alle relativen Werte v_i, $i \in E$, und g, unter der Annahme, dass z.B. $\mathrm{v}_N = 0$ ist.

Schritt 2 (**Verbesserung der Politik**): Für jeden Zustand $i \in E$ stelle man die Strategie k' der neuen Politik **d**' fest, welche

$$q_i^k + \sum_{j=1}^{N} p_{i\,j}^k \, \mathrm{v}_j$$

maximiert unter Verwendung der relative Werte v_i, $i \in E$, der vorherigen Politik **d**.

k' ergibt dann die neue Entscheidung im i-ten Zustand, $q_i^{k'}$ wird q_i , und $p_{i\,j}^{k'}$ wird $p_{i\,j}$.

Schritt 3 (**Prüfung von Konvergenz**): Der Iterationszyklus endet, wenn die Politiken **d** und **d**' bei zwei aufeinanderfolgenden Iterationen stimmen überein. Sonst muss den Schritt 1 mit den neuen Werten wiederholt werden.

▼

Beweis: Siehe Abschnitt 4.6.

Der Schritt 1, die Wertbestimmung, liefert g und v_i, $i \in E$, entsprechend der gegebenen Wahl von q_i und $p_{i\,j}$. Der Schritt 2 ergibt $p_{i\,j}$ und q_i, welche den Gewinn für eine gegebene Menge der v_i, $i \in E$, erhöhen. Mit anderen Worten, die Wertbestimmung ergibt Werte in Abhängigkeit der Politik, währemd die Verbesserung der Politik die Politik als eine Funktion der Werte liefert.

4.4.2 Bemerkung: Wir können den Iterationszyklus in beiden Schritten 1 oder 2 beginnen lassen. Wenn wir die Wertbestimmung als ersten Schritt gewählt haben, muss eine Anfangspolitik ausgewählt werden. Soll der Zyklus mit dem Schritt 2 gestartet werden, so ist eine Menge von Anfangswerten nötig. Wenn a priori kein Grund besteht, eine besondere Anfangspolitik oder eine gewisse Menge von Anfangswerten zu wählen, ist es ser oft zweckdienstlich, den Prozess mit der Verbesserung der Politik mit allen $v_i = 0$, $i \in E$, zu starten.

In diesem Fall wird die Verbesserung der Politik wie folgt eine Politik auserwählen:

Für jedes $i \in E$ wird die Alternative k' bestimmt, welche q_i^k maximiert, und dann $d_i = k'$ gesetzt.

Folglich wird dieses Anfangsverfahren bewirken, dass die Verbesserung der Politik als Anfangspolitik jene auswählt, die in jedem Zustand den zu erwartenden unmittelbaren Erlös maximiert. Mit dieser Politik führt die Iteration dann zur Wertbestimmung, und der Iterationszyklus kann beginnen.

Die Wahl einer Anfangspolitik, die den zu erwartenden unmittelbaren Erlös maximiert, ist in den meisen Fällen durchaus befriedigend.

An dieser Stelle ist es angebracht, einige Worte zu sagen über den Schritt 3: Wie beendet man den Iterationszyklus, wenn er seine Aufgabe erfüllt hat? Die Regel ist ganz einfach: Die optimale Politik ist erreicht (g wird maximiert), wenn die Politiken bei zwei aufeinander folgenden Iterationen identisch sind. Um zu vermeiden, dass die Verbesserung der Politik zwischen gleich guten Strategien in einem bestimmten Zustand hin- und herspringt, muss einfach das alte d_i unverändert bleiben, wenn die Testgrösse für dieses d_i mindestens ebenso gross ist wie für jede andere Strategie bei der Bestimmung der neuen Politik.

Kurz zusammengefasst besitzt die soeben beschriebene Politik-Iteration folgende Eigenschaften:

4.4.3 Satz:

1. Die Lösung des sequentiellen Entscheidungsprozesses beschränkt sich auf die Lösung von Systemen linearer Gleichungen und auf nachfolgende Vergleiche.

2. Jede im Iterationszyklus nachfolgend bestimmte Politik weist einen höheren Gewinn auf als die vorhergehende.

3. Der Iterationszyklus endet bei der Politik mit dem höchsten Gewinn, welcher im Bereich des Problems erlangt werden kann. Diese Politik wird im allgemeinen nach einer kleinen Anzahl von Iterationsschritten gefunden.

▼

Beweis: Siehe Abschnitt 4.6.

Bevor wir die Punkte 2 und 3 beweisen, wollen wir die Politik-Iteration in der Praxis kennen lernen, indem wir sie beim Problem des Spielzeugfabrikanten anwenden.

4.5 Das Problem des Spielzeugfabrikanten

Die Daten für das Problem des Spielzeugfabrikanten finden wir in nächste Tabelle:

Zustand	Strategie k	p_{i1}^k	p_{i2}^k	c_{i1}^k	c_{i2}^k	q_i^k
1 (erfolgreich)	1 keine Reklame	0.5	0.5	9	3	6
1 (erfolgreich)	2 Reklame	0.8	0.2	4	4	4
2 (erfolglos)	1 keine Forschung	0.4	0.6	3	−7	−3
2 (erfolglos)	2 Forschung	0.7	0.3	1	−19	−5

Wir haben zwei Zustände und zwei Strategien in jedem Zustand, so dass vier mögliche Politiken für den Spielzeugfabrikanten bestehen, jede mit den zugehörigen Wahrscheinlichkeiten und Bewertungen. Er möchte nur wissen, welche der vier Politiken

$$\mathbf{d} = \begin{pmatrix} 1 \\ 1 \end{pmatrix},\ \mathbf{d} = \begin{pmatrix} 1 \\ 2 \end{pmatrix},\ \mathbf{d} = \begin{pmatrix} 2 \\ 1 \end{pmatrix},\ \mathbf{d} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

er über eine unabsehbare zukünftige Zeitspanne hinweg befolgen soll, um seinen durchschnittlichen wöchentlichen Verdienst so gross wie möglich zu halten.

Wir wollen annehmen, dass wir apriori nicht wissen, welche Politik die beste ist. Wenn wir nun

$$v_1 = v_2 = 0$$

setzen und damit Politik-Verbesserung gehen, wird natürlich als Anfangspolitik diejenige gewählt, welche den zu erwartenden unmittelbaren Erlös in jedem Zustand maximiert

$$\max_k q_i^k = \max_k \sum_{j=1}^N p_{ij}^k c_{ij}^k.$$

```
Pm1 = {{0.5, 0.5}, {0.4, 0.6}};
Pm2 = {{0.8, 0.2}, {0.7, 0.3}};
Cm1 = {{9, 3}, {3, -7}};
Cm2 = {{4, 4}, {1, -19}};
q1 = {Sum[Pm1[[1]][[j]] Cm1[[1]][[j]], {j, 1, 2}],
  Sum[Pm1[[2]][[j]] Cm1[[2]][[j]], {j, 1, 2}]}
q2 = {Sum[Pm2[[1]][[j]] Cm2[[1]][[j]], {j, 1, 2}],
  Sum[Pm2[[2]][[j]] Cm2[[2]][[j]], {j, 1, 2}]}
```

```
{6., -3.}
```

```
{4., -5.}
```

Für den Spielzeugfabrikant besteht diese Politik aus der Wahl der Strategie $k = 1$ in beiden Zuständen 1 und 2.

Schritt 0. Für diese Politik ist

$$\mathbf{d} = \begin{pmatrix} 1 \\ 1 \end{pmatrix},\ \mathbf{P} = \begin{pmatrix} p_{11}^1 & p_{12}^1 \\ p_{21}^1 & p_{22}^1 \end{pmatrix} = \begin{pmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \end{pmatrix},\ \mathbf{C} = \begin{pmatrix} c_{11}^1 & c_{12}^1 \\ c_{21}^1 & c_{22}^1 \end{pmatrix} = \begin{pmatrix} 9 & 3 \\ 3 & -7 \end{pmatrix}$$

$$\mathbf{q} = \begin{pmatrix} q_1^1 \\ q_2^1 \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^N p_{1\,j}^1\, c_{1\,j}^1 \\ \sum_{j=1}^N p_{2\,j}^1\, c_{2\,j}^1 \end{pmatrix} = \begin{pmatrix} p_{11}^1\, c_{11}^1 + p_{12}^1\, c_{12}^1 \\ p_{21}^1\, c_{21}^1 + p_{22}^1\, c_{22}^1 \end{pmatrix} = \begin{pmatrix} 6 \\ -3 \end{pmatrix}.$$

Schritt 1. Nun können wir mit der *Wertbestimmung* beginnen, die unsere Anfangspolitik auswertet.

Aus den Gleichungen (xi) folgt

$$\mathrm{v}_1 = q_1 + p_{11}\,\mathrm{v}_1 + p_{12}\,\mathrm{v}_2 - g = 6 + 0.5\,\mathrm{v}_1 + 0.5\,\mathrm{v}_2 - g,$$

$$\mathrm{v}_2 = q_2 + p_{21}\,\mathrm{v}_1 + p_{22}\,\mathrm{v}_2 - g = -3 + 0.4\,\mathrm{v}_1 + 0.6\,\mathrm{v}_2 - g,$$

wobei $p_{i\,j} = p_{i\,j}^1,\ q_i = q_i^1$.

Wir setzen

$$\mathrm{v}_2 = 0,$$

dann ergibt die Lösung dieser Gleichungen

$$\mathrm{v}_1 = 6 + 0.5\,\mathrm{v}_1 - g,$$

$$0 = -3 + 0.4\,\mathrm{v}_1 - g \quad \Longrightarrow \quad g = 1,\ \mathrm{v}_1 = 10,\ \mathrm{v}_2 = 0.$$

```
Solve[{v1 == 6 + 0.5 v1 - g, 0 == -3 + 0.4 v1 - g}, {v1, g}]
```

```
{{v1 → 10., g → 1.}}
```

Wir erinnern uns, dass früher für diese Politik bei Anwendung einer anderen Methode auch der Gewinn von 1 herauskam.

Schritt 2. Wir sind nun in der Lage, mit der *Politik-Verbesserung* zu beginnen (siehe nächste Tabelle):

Zustand i	Strategie d (i) = k	$c_i^k + \sum_{j=1}^M p_{ij}^k\, V_i$
1	1	6 + 0.5 · 10 + 0.5 · 0 = 11
1	2	4 + 0.8 · 10 + 0.2 · 0 = 12 ←
2	1	−3 + 0.4 · 10 + 0.6 · 0 = 1
2	2	−5 + 0.7 · 10 + 0.3 · 0 = 2 ←

Die Politik-Verbesserung ergibt:

$$\max_k \left[q_i^k + \sum_{j=1}^N p_{i\,j}^k\, \mathrm{v}_j\right] \text{ für } i \in E.$$

Für den Zustand 1:

$$\max_k \left[c_1^k + \sum_{j=1}^M p_{1\,j}^k\, \mathrm{v}_j\right] = \max\left\{c_1^1 + p_{11}^1\,\mathrm{v}_1 + p_{12}^1\,\mathrm{v}_2,\ c_1^2 + p_{11}^2\,\mathrm{v}_1 + p_{12}^2\,\mathrm{v}_2\right\} =$$

$$= \max\{6 + 0.5\cdot 10 + 0.5\cdot 0,\ 4 + 0.8\cdot 10 + 0.2\cdot 0\} =$$

$$= \max\{11,\ 12\} \quad \Rightarrow d(1) = 2$$

Für den Zustand 2:

$$\max_k \left[c_2^k + \sum_{j=1}^M p_{2\,j}^k\, \mathrm{v}_j\right] = \max\left\{c_2^1 + p_{21}^1\,\mathrm{v}_1 + p_{22}^1\,\mathrm{v}_2,\ c_2^2 + p_{21}^2\,\mathrm{v}_1 + p_{22}^2\,\mathrm{v}_2\right\} =$$

$$= \max\{-3 + 0.4\cdot 10 + 0.6\cdot 0,\ -5 + 0.7\cdot 10 + 0.3\cdot 0\} =$$

$$= \max\{1,\ 2\} \quad \Rightarrow \ d(2) = 2$$

d.h. in jedem Zustand liefert die zweite Strategie einen höheren Wert der Testgrösse $q_i^k + \sum_{j=1}^N p_{i\,j}^k \, v_j$ als die erste Strategie. Somit wird die durch die zweite Strategie in jedem Zustand gebildete Politik einen höheren Gewinn aufweisen als unsere ursprüngliche Politik.

Schritt 3. Dennoch müssen wir unser Verfahren fortführen, weil wir ja noch nicht wissen, ob die neue Politik die beste ist, die wir finden können.

Schritt 0. Für diese Politik ist

$$\mathbf{d} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \mathbf{P} = \begin{pmatrix} p_{11}^2 & p_{12}^2 \\ p_{21}^2 & p_{22}^2 \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 \\ 0.7 & 0.3 \end{pmatrix}, \mathbf{C} = \begin{pmatrix} c_{11}^2 & c_{12}^2 \\ c_{21}^2 & c_{22}^2 \end{pmatrix} = \begin{pmatrix} 4 & 4 \\ 1 & -19 \end{pmatrix}$$

$$\mathbf{q} = \begin{pmatrix} q_1^2 \\ q_2^2 \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^N p_{1\,j}^2 \, c_{1\,j}^2 \\ \sum_{j=1}^N p_{2\,j}^2 \, c_{2\,j}^2 \end{pmatrix} = \begin{pmatrix} p_{11}^2 \, c_{11}^2 + p_{12}^2 \, c_{12}^2 \\ p_{21}^2 \, c_{21}^2 + p_{22}^2 \, c_{22}^2 \end{pmatrix} = \begin{pmatrix} 4 \\ -5 \end{pmatrix}$$

Schritt 1. In diesem Fall werden die Gleichungen (xi) zu

$$v_1 = q_1 + p_{11} \, v_1 + p_{12} \, v_2 - g = 4 + 0.8 \, v_1 + 0.2 \, v_2 - g,$$

$$v_2 = q_2 + p_{21} \, v_1 + p_{22} \, v_2 - g = -5 + 0.7 \, v_1 + 0.3 \, v_2 - g,$$

wobei $p_{i\,j} = p_{i\,j}^2$, $q_i = q_i^2$.

Mit $v_2 = 0$ erhalten wir in der Wertbestimmung folgende Resultate:

$$g = 2, \, v_1 = 10, \, v_2 = 0.$$

```
Solve[{v1 == 4 + 0.8 v1 - g, 0 == -5 + 0.7 v1 - g}, {v1, g}]
```

```
{{v1 → 10., g → 2.}}
```

Der Gewinn der Strategie

$$\mathbf{d} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

ist somit doppelt so gross wie der der ursprpnglichen Politik.

Schritt 2. Wiederun müssen wir die Politik-Verbesserung durchführen. Aber da die relative Werte zufällig die gleichen sind wie diejenigen im vorhergehenden Iterationsschritt, wiederholen sich die Berechnungen, siehe Tabelle:

Zustand i	Strategie d (i) = k	$c_i^k + \sum_{j=1}^M p_{ij}^k \, v_i$
1	1	6 + 0.5 · 10 + 0.5 · 0 = 11
1	2	4 + 0.8 · 10 + 0.2 · 0 = 12 ←
2	1	−3 + 0.4 · 10 + 0.6 · 0 = 1
2	2	−5 + 0.7 · 10 + 0.3 · 0 = 2 ←

Wiederum wird die Politik

$$\mathbf{d} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

festgestellt.

Schritt 3. Da wir hintereinander zweimal die gleiche Politik gefunden haben, ist die optimale Politik bestimmt. Der Iterationszyklus endet.

Der Spielzeugfabrikant sollte die zweite Strategie in jedem Zustand verforgen. Wenn er das tut, wird er durchschnittlich zwei Einheiten in der Woche verdienen, und das bedeutet einen höheren mittleren Erlöss als den, den jede andere Politik zu bieten hat. Der Leser kann zum Beispiel nachprüfen, dass die beiden Politiken

$$\mathbf{d} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \text{ und } \mathbf{d} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

kleinere Gewinne aufweisen.

Für die optimale Politik ist

$$v_1 = 10 \text{ und } v_2 = 0,$$

so dass

$$v_1 - v_2 = 10$$

folgt. Das heisst, dass, wenn auch der Spielzeugfabrikant die optimale Politik befolgt, indem er Reklame macht und Forschung treibt, er zu jeder Zeit bereit ist, einen Erfinder für ein erfolgreiches Spielzeug, wenn er selbst kein solches besitzt, bis zu 10 Einheiten zu zahlen. Die relativen Werte der optimalen Politik können also verwendet werden, um dem Spielzeugfabrikanten bei einmaligen Entscheidungen darüber, ob er bei schlechtem Geschäftsgang die Lizenz für ein erfolgreiches Spielzeug kaufen soll, zu helfen.

Die optimale Politik für den Spielzeugfabrikanten wurde mittels der Wertiteration im Kapitel 3 bestimmt. Die Ähnlichkeiten und Verschiedenheiten der beiden Methoden sollten nun klar sein.

4.5.1 Bemerkung: Man beachte, wie die Politik-Iteration bei Erreichen der Politikkonvergenz von selbst endet. Es gibt kein vergleichbares Verhalten in der Wertiteration. Die Politik-Iteration ist einfach in der Form und in der Interpretation, was sie vom rechnerischen Standpunkt uas empfehlenswert erscheinen lässt. Jedoch müssen wir immer daran denken, dass sie nur für unbegrenzte Prozesse oder für solche, deren Ende noch weit entfernt ist, angewendet werden kann.

4.6 Ein Beweis für die Eigenschaften der Politik-Iteration

Angenommen wir hätten eine Politik A für den Verlauf des Systems ausgewertet, und die Politik-Verbesserung habe eine Politik B, die verschieden von A ist, ergeben. Wenn wir die obere Indizes A und B verwenden, um die zugehörigen Grössen für die Politiken A und B zu bezeichnen, möchten wir beweisen, dass

$$g^B > g^A$$

gilt.

Aus der Definition der Politik-Verbesserung geht hervor, da B gegenüber A vorgezogen wurde:

(4.6.1) $$q_i^B + \sum_{j=1}^N p_{ij}^B v_j^A \geq q_i^A + \sum_{j=1}^N p_{ij}^A v_j^A, \ i \in E.$$

Sei

(4.6.2) $$\gamma_i = q_i^B + \sum_{j=1}^N p_{ij}^B v_j^A - q_i^A - \sum_{j=1}^N p_{ij}^A v_j^A,$$

so dass

$$\gamma_i \geq 0$$

ist. Die Grösse γ_i ist die Verbesserung der Testgrösse, welche die Politik-Verbesserung im i-ten Zustand erreichen konnte.

Für die einzelnen Politiken A und B ergeben die Gleichungen (xi)

(4.6.3) $$v_i^B = q_i^B + \sum_{j=1}^N p_{ij}^B v_j^B - g^B, \ i \in E,$$

(4.6.4) $$v_i^A = q_i^A + \sum_{j=1}^N p_{ij}^A v_j^A - g^A, \ i \in E.$$

Wird die Gleichung (4.6.4) von der Gleichung (4.6.3) subtrahiert, erhalten wir folgendes Resultat:

(4.6.5) $$v_i^B - v_i^A = q_i^B - q_i^A + \sum_{j=1}^N p_{ij}^B v_j^B - \sum_{j=1}^N p_{ij}^A v_j^A + g^A - g^B.$$

Wenn die Gleichung (4.6.2) nach

$$q_i^B - q_i^A$$

aufgelöst, d.h.

$$q_i^B - q_i^A = \gamma_i - \sum_{j=1}^N p_{ij}^B v_j^A - \sum_{j=1}^N p_{ij}^A v_j^A$$

und das Resultat in Gleichung (4.6.5) eingesetzt wird, haben wir

$$v_i^B - v_i^A = \gamma_i - \sum_{j=1}^N p_{ij}^B v_j^A - \sum_{j=1}^N p_{ij}^A v_j^A + \sum_{j=1}^N p_{ij}^B v_j^B - \sum_{j=1}^N p_{ij}^A v_j^A + g^A - g^B$$

oder

(4.6.6) $$v_i^B - v_i^A = \gamma_i + \sum_{j=1}^N p_{ij}^B \left(v_j^B - v_j^A\right) + g^A - g^B.$$

Sei

$$g^\Delta = g^B - g^A \text{ und } v_i^\Delta = v_i^B - v_i^A.$$

Dann wird aus der Gleichung (4.6.6)

(4.6.7) $$v_i^\Delta = \gamma_i + \sum_{j=1}^N p_{ij}^B v_j^\Delta - g^\Delta, \ i \in E.$$

Die Gleichungen (4.6.7) sind in der Form identisch mit den Gleichungen (xi), ausser dass sie in Differenzenform statt in absoluten Grössen geschrieben werden. Genau wie die aus den Gleichungen (xi) erhaltene Lösung für g

$$g = \sum_{i=1}^N \pi_i q_i$$

ist, so ist die Lösung für g^Δ in den Gleichungen (4.6.7)

(4.6.8) $$g^\Delta = \sum_{i=1}^N \pi_i^B \gamma_i,$$

wobei π_i^B die Grenzwahrscheinlichkeit des Zustandes $i \in E$ für die Politik B ist.

Da alle $\pi_i^B \geq 0$ und alle $\gamma_i \geq 0$ sind, gilt

$$g^\Delta \geq 0.$$

Insbesondere wird g^B grösser sein als g^A, wenn eine Vergrösserung der Testgrösse in irgendeinem Zustand, der nicht-transient bezüglich der Politik B ist, erzielt werden kann. Aus der Gleichung (4.6.8) ersehen wir, dass die durch die Verbesserungen in jedem nicht-transienten Zustand der neuen Politik verursachten Gewinnerhöhungen additiv sind. Auch dann, wenn wir unsere Politik-Verbesserung nur bei einem Zustand durchführen und die anderen Entscheidungen unverändert lassen, wird der Gewinn des Systems erhöht, wenn dieser Zustand unter der neuen Politik nicht-transient ist.

Nun zeigen wir noch, dass es nicht vorkommen kann, dass eine bessere Politik existiert und nicht in irgendeinem Iterationsschritt mit Hilfe der Politik-Verbesserung entdeckt wird.

Wir nehmen an, dass für zwei Strategien A und B

$$g^B > g^A$$

gilt, aber die Politik-Verbesserung gegen die Politik A konvergiert. Dann gilt für alle Zustände

$$\gamma_i \leq 0,$$

wobei γ_i durch die Gleichung (4.6.2) definiert ist. Da $\pi_i^B \geq 0$ für alle $i \in E$, liefert Gleichung (4.6.8)

$$g^B - g^A = \sum_{i=1}^{N} \pi_i^B \gamma_i \leq 0.$$

Aber nach Annahme ist $g^B > g^A$, so dass wir zu einem Wiederspruch gelangen. Es ist somit unmöglich, dass eine bessere Politik unentdeckt bleibt.

Das nächste Kapitel wird weitere Beispiele für die Politik-Iterationsmethode bringen und zeigen, wie sie auf verschiedene Probleme angewendet werden kann.