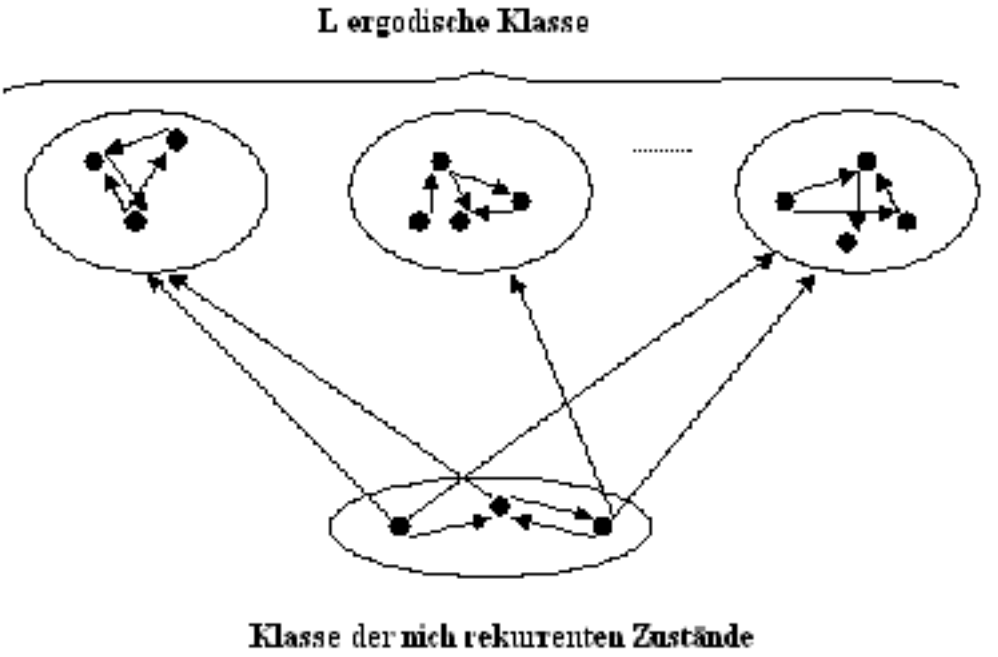
§6 Die Politik-Iterationsmethode für Prozesse mit mehreren ergodischen Klassen

Die Entwicklung im Kapitel 4 setzen voraus, dass jede Politik eine Markov-Kette mit nur einer ergodischen Klasse definiert und demgemäss mit nur einem eindeutig bestimmten Gewinn. Unser Problem war, einfach festzustellen, welche Politik den höchsten Gewinn mit sich bringt. Mit der Methode im Kapitel 4 wurde dieses Ziel erreicht. Diese Iterationsmethode genügt für die meisten Probleme, da wir im allgemeinen ein Problem so definieren können, dass nur ergodische Politiken mit einer ergodischen Klasse vorkommen. Das war der Fall für die Beispiele im Kapitel 5.

6.1 Einführung

Dennoch ist es nicht schwer, sich Prozesse mit mehreren ergodischen Klassen vorzustellen.



6.1.1 Definition: Wenn zuerst die Zustände jeder ergodischen Klasse und dann transiente (nicht rekurrente) Zustände der Markov-Kette durchnumeriert werden, dann heißt die Übergangsmatrix **P** ***kanonisch***, nähmlich

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_1 & 0 & 0 & \dots & 0 \\ 0 & \mathbf{P}_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \mathbf{P}_L & 0 \\ \mathbf{R}_1 & \mathbf{R}_2 & \dots & \mathbf{R}_L & \mathbf{Q} \end{pmatrix},$$

L – die Anzahl der ergodischen Klassen

Q –bezeichnet die Übegangsmatrix für die nicht rekurrenten Zuständen;

$\mathbf{R}_i$ –die Übegangsmatrix für die Übergänge von der Klasse der nicht rekurrenten Zustände in i – te ergodische Klasse;

$\mathbf{P}_i$ –die Übergangsmatrix innerhalb der i – ten ergodischen Klasse

In die Literatur kann man die kanonische Matrix in Form

$$\mathbf{P} = \begin{pmatrix} \mathbf{Q} & \mathbf{R}_1 & \mathbf{R}_2 & \dots & \mathbf{R}_L \\ 0 & \mathbf{P}_1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & \mathbf{P}_L \end{pmatrix}$$

treffen.

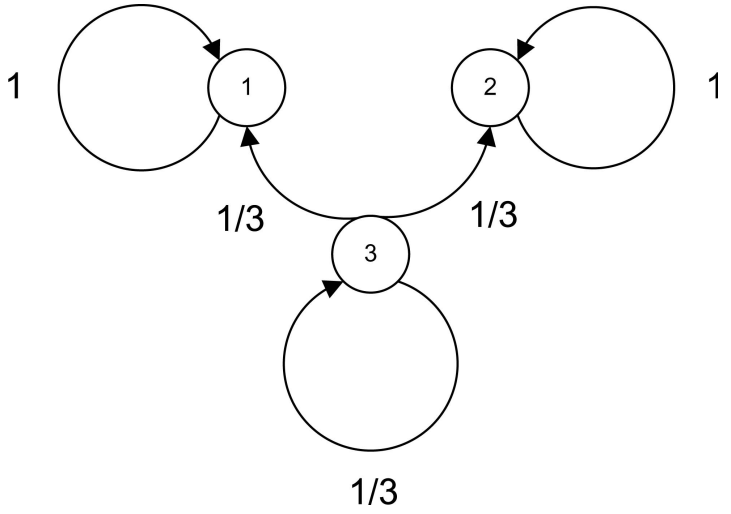
Im Kapitel 1 wurde ein Prozess mit drei Zuständen und einer Übergangsmatrix

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}$$

besprochen,

$\mathbf{Q}$=1/3, $\mathbf{R}_1 = \mathbf{R}_2 = 1/3$, $\mathbf{P}_1 = \mathbf{P}_2 = 1$,

der zwei ergodische Klasse enthält, d.h. $L = 2$,



Nehmen wir an, der Prozess habe einen Vektor von zu erwartenden unmittelbaren Erlösen

$$\mathbf{q} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix},$$

ausgedrückt in Dollars. Als Matrix der Grenzwahrscheinlichkeiten ergab sich im Kapitel 1 (siehe Beispiel 1.5.7)

$$\mathbf{S} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}.$$

Der Gewinnvektor ist damit (Satz 2.5.1)

$$\mathbf{g} = \mathbf{S}\,\mathbf{q} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1.5 \end{pmatrix},$$

und wir interpretieren **g** folgendermassen:

Angenommen man starte den Prozess im Zustand l, dann würde je Übergang \$1.00 verdient. Mit einem Start im Zustand 2 würde man einen Gewinn von \$2.00 je Übergang erzielen. Schliesslich, da das System sich nach vielen Übergangen mit einer ebenso grossen Wahrscheinlichkeit in den Zustand 1 wie in den Zustand 2 begibt, wenn es im Zustand 3 startet, erwartet man von einer solchen Anfangsposition einen durchschnittlichen Gewinn von \$1.50 je Übergang. Diesen Durchschnittswert erhält man durch mehrere von einander unabhängige Versuche mit Anfangszustand 3, da in jedem dieser Versuche letzten Endes entweder \$ 1.00 oder \$ 2.00 je Übergang verdient wird.

Der Gewinn des Systems hängt somit vom Anfangszustand ab. Ein Start im Zustand $i \in E$ ergibt einen Gewinn g_i, sodass wir uns den Gewinn als eine Funktion sowohl des Zustandes als auch des Prozesses denken können. Unsere neue Aufgabe besteht nun darin, die Politik für das System zu finden, die den Gewinn für alle Zustände des Systems maximiert. Glücklicherweise kann die Politik-Iterationsmethode des Kapitels 4 für Prozesse mit verschiedenen Gewinnen erweitert werden. Wir kommen nun zu dieser Erweiterung.

6.2 Die Wertbestimmung

Die Gleichungen (vii) aus dem Satz 2.5.5 geben die asymptotische Form, die der total zu erwartende Erlös annimmt, wenn das System im Zustand $i \in E$ gestartet wird und eine grosse Anzahl von Übergängen stattfindet :

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

Jeder Zustand hat sein eigenes g_i, aber wie im Kapitel 2 erläutert wurde, haben alle Zustände, die derselben ergodischen Klasse angehören, die gleichen Gewinne.

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

(xii) $$g_i = \sum_{j=1}^{N} p_{i\,j}\, g_j\ ,\ i \in E,$$

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

▼

Beweis: Wenn wir den unendlichen Prozess studieren wollen, können wir die Gleichungen (vii) zusammen mit der grundlegenden rekursiven Relation für die total zu erwartenden Erlöse

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

verwenden, woraus folgt:

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

oder

$$n\, g_i + g_i + v_i = q_i + n \sum_{j=1}^{N} p_{i\,j}\, g_j + \sum_{j=1}^{N} p_{i\,j}\, v_j,\ \ i \in E.$$

Soll diese Gleichung für jedes grosse n gelten, dann muss

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

und

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

Wir haben nun zwei Systeme von je N linearen Gleichungen (xii und xiii), die wir benützen können, um die N Unbekannten g_i und die N Unbekannten v_i zu bestimmen. Jedoch sind die Gleichungen (xii) nicht eindeutig lösbar. Die Matrix

I – P

ist singulär, sodass die Lösungen g_i der Gleichungen (xii) arbiträre Konstanten enthalten werden. Die Anzahl der freien Konstanten ist gleich der Anzahl der ergodischen Klassen im Prozess. Die Gleichungen (xii) setzen im wesentlichen die Gewinne jedes Zustandes in Beziehung zu den Gewinnen jeder ergodischen Klasse. Zum

Beispiel gibt es in einem Prozess mit L ergodischen Klassen L unabhängige Gewinne. Die Gewinne der transienten Zustände werden durch die Gleichungen (xii) mit den L unabhängigen Gewinnen in Beziehung gebracht, und sie sind bestimmt, wenn die unabhängigen Gewinne bestimmt sind.

Die N Gleichungen (xiii) müssen nun verwendet werden, um die L unabhängigen Gewinne und auch die N Werte v_i zu bestimmen. Wir haben somit L Unbekannte zu viel. Nehmen wir jedoch an, wir erweitern unser früheres Verfahren so, dass wir für einen Zustand $i \in E$ in jeder ergodischen Klasse v_i gleich Null setzen, sodass insgesamt L Werte v_i, gleich Null sind. Im allgemeinen werden wir in jeder Klasse v_i für den Zustand mit dem höchsten Index gleich Null setzen. Wir sehen, dass die Gleichungen (xiii) nun für die L unabhängigen Gewinne und für die verbleibenden $N - L$ Werte v_i gelöst werden können.

Die durch die Lösung der Gleichungen (xiii) bestimmten v_i, $i \in E$, können immer noch als relative Werte bezeichnet werden, wenn wir daran denken, dass sie relativ sind innerhalb einer Klasse. Bei der Lösung der Gleichungen (xii) und (xiii) begegnen wir ungefähr der gleichen Schwierigkeit wie bei der Berechnung der Grenzwahrscheinlichkeiten-Matrix **S** für einen Prozess mit mehreren ergodischen Klassen. Wir werden sehen, dass die relativen Werte v_i, $i \in E$, ebenso nützlich sind, wie die wahren durch die Gleichungen (vii) bestimmten Werte v_i, soweit es um die Bestimmung der optimalen Politik geht. Um diese Bemerkungen zu illustrieren, wollen wir den Gewinn und die relativen Werte des Zwei-Klassen-Prozesses, welchen wir am Anfang dieses Kapitels besprochen haben, bestimmen. Die Gleichungen (xii) ergeben

$$g_1 = g_1,\ g_2 = g_2,\ g_3 = \frac{1}{3}\, g_1 + \frac{1}{3}\, g_2 + \frac{1}{3}\, g_3.$$

Somit haben wir zwei unabhängige Gewinne g_1 und g_2. Der Gewinn des Zustandes 3 wird durch g_1 und g_2 gemäss

$$g_3 = \frac{1}{2}\, g_1 + \frac{1}{2}\, g_2$$

ausgedrückt. Wenn wir g_1 und g_2 finden können, kennen wir den Gewinn jedes Zustandes.

6.2.2 Definition: Allgemein werden wir mit ${}^1g$ den Gewinn der Klasse 1, mit ${}^2g$ den Gewinn von 2 usw. bezeichnen und dann den Gewinn eines jeden Zustandes durch

$${}^1g, {}^2g, \ldots$$

ausdrücken.

Diese Bezeichnung kann nicht verwendet werden, bevor die Zustände in bezug auf ihre Zugehörigkeit zu einer Klasse identifiziert sind. In diesem Problem ist

$$g_1 = {}^1g\ g_2 = {}^2g \text{ und } g_3 = \frac{1}{2}\, {}^1g + \frac{1}{2}\, {}^2g.$$

Die Gleichungen (xiii) ergeben

$$v_1 = 1 + v_1 - g_1$$

$$v_2 = 2 + v_2 - g_2$$

$$v_3 = 3 + \frac{1}{3}\, v_1 + \frac{1}{3}\, v_2 + \frac{1}{3}\, v_3 - g_3.$$

Wenn wir nun g_3 durch g_1 und g_2 ausdrücken und dann den relativen Wert eines Zustandes in jeder ergodischen Klasse gleich Null setzen, so dass

$$v_1 = v_2 = 0,$$

erhalten wir

$$g_1 = 1,\ g_2 = 2,\ \ v_3 = 3 + \frac{1}{3}\, v_3 - \frac{1}{2}\, g_1 - \frac{1}{2}\, g_2.$$

Die Lösung dieses Gleichungssystems ist $g_1 = 1$, $g_2 = 2$, $v_3 = 2.25$, so dass

$g_1 = 1, g_2 = 2, g_3 = 1.5$

$v_1 = 0, v_2 = 0, v_3 = 2.25$

die Gewinne und relativen Werte für jeden Zustand des Prozesses darstellen.

Die Gewinne sind natürlich dieselben wie diejenigen, die wir früher erhalten haben.

6.3 Die Verbesserung der Politik

Wir werden nun zeigen, wie die Gewinne und relativen Werte einer Politik verwendet werden können, um die optimale Politik für ein System zu bestimmen, im Anschluss an die Argumentation im Kapitel 4 können wir, wenn wir jetzt eine bis zur Stufe n angewandte Politik haben, die beste Entscheidung für den i-ten Zustand auf der Stufe $n + 1$ treffen, indem wir

$$q_i^k + \sum_{j=1}^N p_{ij}^k v_j(n)$$

maximieren in bezug auf alle möglichen Strategien im Zustand $i \in E$. Für grosse n können wir in diesem Ausdruck die Gleichungen (vii) einsetzen und erhalten

$$q_i^k + \sum_{j=1}^N p_{ij}^k (n g_j + v_j)$$

oder

$$n \sum_{j=1}^N p_{ij}^k g_j + q_i^k + \sum_{j=1}^N p_{ij}^k v_j$$

als Testgrösse, die zu maximieren ist. Ist n gross, wird dieser Ausdruck natürlich maximiert durch die Strategie, die die Gewinntestgrösse

$$\sum_{j=1}^N p_{ij}^k g_j$$

maximiert, wobei die g_j die Gewinne der alten Politik sind. Wenn jedoch alle Strategien den selben Wert

$$\sum_{j=1}^N p_{ij}^k g_j$$

haben oder wenn eine Anzahl von Strategien den gleichen Maximalwert der Gewinngrösse haben, wird die Unklarheit beseitigt durch die Wahl der Strategie, welche die Werttestgrösse

$$q_i^k + \sum_{j=1}^N p_{ij}^k v_j$$

maximiert, wobei die relativen Werte der alten Politik benützt werden. Die relativen Werte können für den Testwert gebraucht werden, weil, wie wir sehen werden, der Test durch eine Konstante, die zu den v_i, $i \in E$, aller Zustände einer ergodischen Klasse addiert wird, nicht beeinflusst wird. Der allgemeine Iterationszyklus ist im Satz 6.3.1 dargestellt. Man beachte, dass er in unseren Iterationszyklus vom Satz 4.4.1 übergeht für ergodische Prozesse mit einer ergodischen Klasse. Wir werden nun ein Beispiel mit mehr als nur einer ergodischen Klasse behandeln und anschliessend die wesentlichen Optimalitälsbeweise erbringen.

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

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

Schritt 1 (Wertbestimmung): Man verwende p_{ij} und q_i für die gegebene Politik und löse

$$g_i = \sum_{j=1}^N p_{ij} g_j, \; i \in E,$$

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

für alle relativen Werte v_i, $i \in E$, und g_i, indem man den Wert eines v_i in jeder ergodischen Klasse gleich Null setzt.

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

$$\sum_{j=1}^{N} p_{ij}^{k} g_j$$

maximiert, indem man die Gewinne der vorhergehenden Politik verwendet. Diese Strategie ist die neue Entscheidung im i-ten Zustand.

Wenn

$$\sum_{j=1}^{N} p_{ij}^{k} g_j$$

für alle Strategien gleich ist, oder wenn mehrere Strategien gleich gut sind gemäss diesem Test, muss die Entscheidung auf Grund der relativen Werte anstatt des Gewinnes getroffen werden. Demzufolge bestimme man, wenn der Gewinntest misslingen sollte, die Strategie k', die

$$q_i^k + \sum_{j=1}^{N} p_{ij}^{k} v_j$$

maximiert, wobei die relativen Werte der vorangehenden Politik benutzt werden.

Damit hat man die neue Entscheidung im i-ten Zustand bestimmt.

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 6.5

6.3.2 Bemerkung: Ungeachtet dessen, ob der Politik-Verbesserungstest auf den Gewinnen oder Werten basiert, belasse man die alte Entscheidung unverändert, wenn sie im i-ten Zustand einen abenso grossen Wert der Testgrösse liefert wie jede andere Strategie. Diese Regel sichert die Konvergenz im Falle gleichwertigen Politiken.

6.4 Ein Beispiel

Wir wollen nun die optimale Politik für das System mit drei Zuständen bestimmen, dessen Übergangswahrscheinlichkeiten und Erlöse in der nächten Tabelle aufgeführt sind. Alle Übergangswahrscheinlichkeiten sind entweder gleich 1 oder gleich 0, erstens wegen der Rechenvereinfachung und zweitens, um zu zeigen, dass sich bei einer solchen Struktur keine Schwierigkeiten ergeben. Dieses System erlaubt Politiken, zu denen Prozesse mit mehreren ergodischen Klassen gehören.

Zustand i	Strategie k	p_{i1}^k	p_{i2}^k	p_{i3}^k	q_i^k
1	1	1	0	0	1
1	2	0	1	0	2
1	3	0	0	1	3
2	1	1	0	0	6
2	2	0	1	0	4
2	3	0	0	1	5
3	1	1	0	0	8
3	2	0	1	0	9
3	3	0	0	1	7

Schritt 0.

Wir wollen mit der Politik beginnen, die die zu erwartenden unmittelbaren Erlöse maximiert. Diese Politik setzt sich zusammen aus der dritten Strategie im ersten Zustand, der ersten Strategie im zweiten Zustand und der zweiten Strategie im dritten Zustand,

$$d(1) = \text{argmax}_k \{q_1^k\} = \{1, 2, 3\} = 3$$

$$d(2) = \text{argmax}_k \{q_2^k\} = \{6, 4, 5\} = 1$$

$$d(3) = \text{argmax}_k \{q_3^k\} = \{8, 9, 7\} = 2$$

Für diese Politik gilt also

$$\mathbf{d} = \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}, \mathbf{P} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \mathbf{q} = \begin{pmatrix} 3 \\ 6 \\ 9 \end{pmatrix}$$

Schritt 1. Wertbestimmung.

Nun können wir die Iteration mit der Auswertung der Politik beginnen. Die Gleichungen (xii)

$$g_i = \sum_{j=1}^{N} p_{i\,j}\, g_j \,,\; i \in E$$

ergeben

$$g_1 = g_3,\, g_2 = g_1,\, g_3 = g_2.$$

Diese Resultate zeigen, dass es nur eine ergodische Klasse gibt und dass alle drei Zustände Elemente dieser Klasse sind. Wenn wir ihren Gewinn mit g bezeichnen, dann ist

$$g_1 = g_2 = g_3 = g.$$

Der relative Wert v_3 wird willkürlich gleich Null gesetzt. Wenn wir diese Resultate in die Gleichungen (xiii) einsetzen, erhalten wir folgende Gleichungen:

$$v_1 = 3 - g$$

$$v_2 = 6 + v_1 - g$$

$$v_3 = 0 = 9 + v_2 - g.$$

Dieses System hat die Lösung

$g = 6$, $v_1 = v_2 = -3$, so dass

$$g_1 = g_2 = g_3 = 6.$$

Schritt 2. Wir sind nun in der Lage, eine Politik-Verbesserung durchzuführen, wie aus nächster Tabelle

ersichtlich ist.

Zustand i	Strategie k	Gewinntestgrösse $\sum_{j=1}^{N} p_{ij}^{k} g_i$	Werttestgrösse $q_i^k + \sum_{j=1}^{N} p_{ij}^{k} v_j$
1	1	6	$1 + (-3) = -2$
1	2	6	$2 + (-3) = -1$
1	3	6	$3 + 0 = 3 \leftarrow$
2	1	6	$6 + (-3) = 3$
2	2	6	$4 + (-3) = 1$
2	3	6	$5 + 0 = 5 \leftarrow$
3	1	6	$8 + (-3) = 5$
3	2	6	$9 + (-3) = 6$
3	3	6	$7 + 0 = 7 \leftarrow$

Da der Gewinntest in allen Fällen nicht schlüssig ist, ist ein Werttest nötig.

Die neue Politik lautet

$$\mathbf{d} = \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix}.$$

Schritt 0.

$$\mathbf{d} = \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix}, \mathbf{P} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \mathbf{q} = \begin{pmatrix} 3 \\ 5 \\ 7 \end{pmatrix}.$$

Schritt 1.

$$g_1 = g_3,\ g_2 = g_3,\ g_3 = g_3.$$

Wir schreiben

$$g_1 = g_2 = g_3 = g,$$

setzen für $v_3 = 0$, benutzen die Gleichungen (xiii) und erhalten

$$v_1 = 3 - g$$

$$v_2 = 5 - g$$

$$v_3 = 0 = 7 - g.$$

Daraus folgt:

$$g = 7,\ v_1 = -4,\ v_2 = -2,$$

und somit

$$g_1 = g_2 = g_3 = 7,$$

$$v_1 = -4,\ v_2 = -2,\ v_3 = 0.$$

Schritt 2. Die Politik-Verbesserung wird in nächter Tabelle gezeigt.

Zustand i	Strategie k	Gewinntestgrösse $\sum_{j=1}^{N} p_{ij}^{k} g_i$	Werttestgrösse $q_i^k + \sum_{j=1}^{N} p_{ij}^{k} v_j$
1	1	7	−3
1	2	7	0
1	3	7	3 ←
2	1	7	2
2	2	7	2
2	3	7	5 ←
3	1	7	4
3	2	7	7
3	3	7	7 ←

Da der Gewinntest wiederum unbestimmt war, musste man sich auf den Vergleich der relativen Werte stützen. Im Zustand 3 ist die Werttestgrösse der Strategien 2 und 3 gleich gross. Da jedoch die Strategie 3 unsere alte Entscheidung war, bleibt sie unverändert. Somit haben wir zweimal hintereinander die gleiche Politik erhalten, die demzufolge die optimale Politik sein muss. Die optimale Politik weist in allen Zuständen einen Gewinn von 7 auf.

Die Politik

$$\mathbf{d} = \begin{pmatrix} 3 \\ 3 \\ 2 \end{pmatrix},$$

die wegen der Gleichheit der Werttestgrösse im Zustand 3 möglich wäre, ist ebenfalls optimal.

Obschon dieses System die Möglichkeit von Prozessen mit mehreren ergodischen Klassen zulässt, begegnen wir diesem Verhalten nicht, wenn wir die Politik, welche die zu erwartenden unmittelbaren Erlöse maximiert, als erste wählen. Dennoch wird dieses Verhalten durch die Wahl einer anderen Anfangspolitik hervorgerufen.

Man nehme folgende Anfangspolitik an

Schritt 0.

$$\mathbf{d} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}, \mathbf{P} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \mathbf{q} = \begin{pmatrix} 3 \\ 4 \\ 8 \end{pmatrix}.$$

Schritt 1. Für die Auswertung dieser Politik wenden wir vorerst die Gleichungen (xii) an und erhalten

$$g_1 = g_3,\ g_2 = g_2,\ g_3 = g_1.$$

Wir haben $L = 2$ ergodische Klassen. Die Klasse 1 setz sich aus den Zuständen 1 und 3 zusammen, die Klasse 2 besteht nur aus dem Zustand 2. Daher ist

$$g_1 = g_3 = {}^1g$$

$$g_2 = {}^2g,$$

und wir können

$$v_2 = v_3 = 0$$

setzen. Die Gleichungen (xiii) ergeben dann

$$v_1 = 3 - {}^1g$$

$$v_2 = 0 = 4 - {}^2g$$

$$v_3 = 0 = 8 + v_1 - {}^1g.$$

Die Lösung dieser Gleichungen sind ${}^1g = \frac{11}{2}$, ${}^2g = 4$, $v_1 = -\frac{5}{2}$ und somit

$$g_1 = \frac{11}{2},\ g_2 = 4,\ g_3 = \frac{11}{2}$$

und

$$v_1 = -\frac{5}{2},\ v_2 = 0,\ v_3 = 0.$$

Schritt 2. Die nächste Tabelle veranschaulicht die Politik-Verbesserung.

Zustand i	Strategie k	Gewinntestgrösse $\sum_{j=1}^{M} p_{ij}^{k}\, g_i$	Werttestgrösse $q_i^k + \sum_{j=1}^{M} p_{ij}^{k}\, v_j$
1	1	11 / 2	−3 / 2
1	2	4	2
1	3	11 / 2	3 ←
2	1	11 / 2	7 / 2
2	2	4	4
2	3	11 / 2	5 ←
3	1	11 / 2	11 / 2
3	2	4	9
3	3	11 / 2	7 ←

In diesem Fall wurde die Politik-Verbesserung mit Hilfe sowohl der Gewinne als auch der Werte durchgeführt. Der Gewinntest bestimmte zwei Strategien in jedem Zustand, und dann traf der Werttest unter diesen die Auswahl. Die erhaltene Politik ist die schon früher gefundene optimale Politik, und somit ist es nicht nötig, das Verfahren fortzusetzen, weil es sich ja nur um eine Wiederholung unserer früheren Arbeit handelt würde.

Im obigen Beispiel wurde mit einer Politik mit zwei ergodischen Klassen begonnen, und man gelangte zu einer optimalen Politik mit einer ergodischen Klasse. Der Leser sollte mit solchen Politiken wie

$$\mathbf{d} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \text{ und } \mathbf{d} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

beginnen, um zu sehen, wie die optimale Politik mit einem Gewinn von 7 für alle Zustände über verschiedene Wege erreicht werden kann. Mail beachte, dass es in keinem Fall notwendig ist, die wirklichen Grenzwerte v_i zu benützen ; die relativen Werte sind ausreichend für Politik-Verbesserungszwecke.

6.5 Eigenschaften des Iterationszyklus

Im Folgenden zeigen wir, dass der Iterationszyklus im Satz 6.3.1 zu der Politik führen wird, die in jedem. Zustand einen höheren Gewinn aufweist als jede andere Politik.

Angenommen eine Politik A wurde ausgewertet, so dass ihre Gewinne und Werte bekannt sind. Diese Gewinne und Werte werden in der Politik-Verbesserung verwendet, um eine neue Politik B zu bestimmen. Wir müssen die Beziehungen zwischen A und B bestimmen. Wenn die Entscheidung im Zustand $i \in E$ auf Grund der Gewinne getroffen wurde, wissen wir, dass

$$\sum_{j=1}^{N} p_{ij}^{B}\, g_j^{A} > \sum_{j=1}^{N} p_{ij}^{A}\, g_j^{A},$$

wobei wir die oberen Indizes A und B verwenden wollen, um die zu jeder Politik gehörenden Grössen zu bezeichnen. Insbesondere wollen wir ψ_i definieren gemäss

(6.5.1) $$\psi_i = \sum_{j=1}^{N} p_{ij}^{B}\, g_j^{A} - \sum_{j=1}^{N} p_{ij}^{A}\, g_j^{A}.$$

Die Zahl ψ_i ist grösser als Null, wenn die Entscheidung im i-ten Zustand auf dem Gewinn beruht, und sie ist gleich Null, wenn sie auf den Werten basiert. Wenn ψ_i gleich null ist, so dass eine Entscheidung auf Grund der Werte getroffen wurde, wissen wir, dass

$$q_i^B + \sum_{j=1}^N p_{i\,j}^B \, \mathrm{v}_j^A \geq q_i^A + \sum_{j=1}^N p_{i\,j}^A \, \mathrm{v}_j^A.$$

Wenn wir

(6.5.2) $$\gamma_i = q_i^B + \sum_{j=1}^N p_{i\,j}^B \, \mathrm{v}_j^A - q_i^A - \sum_{j=1}^N p_{i\,j}^A \, \mathrm{v}_j^A$$

setzen, dann ist

$$\gamma_i \geq 0.$$

Wenn ψ_i und γ_i gleich Null ist, dann sind die Politiken A und B gleichwertig, soweit es sich um die Testgrössen im Zustand $i \in E$ handelt. In einem solchen Fall würden wir willkürlich die Entscheidung im Zustand i benutzen, die zur Politik A gehört.

Nun können wir für beide Politiken A und B die Gleichungen für die Auswertung der Politik gemäss Gleichungen (xii) und (xiii) schreiben.

Für die Politik A ergibt sich

(6.5.3) $$g_i^A = \sum_{j=1}^N p_{i\,j}^A \, g_j^A, \; i \in E$$

(6.5.4) $$\mathrm{v}_i^A = q_i^A + \sum_{j=1}^N p_{i\,j}^A \, \mathrm{v}_j^A - g_i^A, \; i \in E.$$

Die entsprechenden Relationen für die Politik B sind

(6.5.5) $$g_i^B = \sum_{j=1}^N p_{i\,j}^B \, g_j^B, \; i \in E$$

(6.5.6) $$\mathrm{v}_i^B = q_i^B + \sum_{j=1}^N p_{i\,j}^B \, \mathrm{v}_j^B - g_i^B, \; i \in E.$$

Substrahiert man Gleichung (6.5.3) von Gleichung (6.5.5) so folgt

$$g_i^B - g_i^A = \sum_{j=1}^N p_{i\,j}^B \, g_j^B - \sum_{j=1}^N p_{i\,j}^A \, g_j^A.$$

Benutzt man die Gleichung (6.5.1) um

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

zu eleminieren, und setzt man

$$g_i^\Delta = g_i^B - g_i^A,$$

dann ist

(6.5.7) $$g_i^\Delta = \psi_i + \sum_{j=1}^N p_{i\,j}^B \, g_j^\Delta, \; i \in E.$$

Analog folgt, wenn die Gleichung (6.5.4) von der Gleichung (6.5.6) subtrahiert wird,

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

Gleichung (6.5.2) kann zur Elimination von

$$q_i^B - q_i^A$$

benutzt werden. Dann folgt, wenn wir

$$\mathrm{v}_i^\Delta = \mathrm{v}_i^B - \mathrm{v}_i^A$$

setzen,

(6.5.8) $$\mathrm{v}_i^\Delta = \gamma_i + \sum_{j=1}^N p_{i\,j}^B \, \mathrm{v}_j^\Delta - g_i^\Delta, \; i \in E.$$

Wir haben nun festgestellt, dass die Veränderungen der Gewinne und Werte die beiden Gleichungssysteme (6.5.7) und (6.5.8) erfüllen müssen.

Die Gleichungen (6.5.8) sind identisch mit den Gleichungen (xiii), mit der Ausnahme, dass sie die Differenzen der Gewinne und Werte enthalten anstatt der absoluten Grössen und dass γ_i vorkommt statt q_i. Hingegen unterscheiden sich die Gleichungen (6.5.7) von den Gleichungen (xii) um die Grösse ψ_i. Andernfalls, wenn ψ_i Null wäre, würden die Gleichungen (6.5.7) im gleichen Verhältnis zu den Gleichungen (xii) stehen wie die Gleichungen (xiii) zu (6.5.8).

Nun wollen wir die Gleichungen (6.5.7) näher untersuchen. Die durch die Parameter p_{ij}^B und q_i^B bestimmte Politik B kann natürlich viele ergodische Klassen haben. Wenn es L ergodische Klassen im Prozess gibt, sind wir imstande, L Gruppen von Zuständen zu bestimmen, die die Eigenschaft haben, dass, wenn das System in irgendeinem Zustand innerhalb einer Gruppe gestartet wurde, es immer Übergänge innerhalb dieser Gruppe machen wird. Zusätzlich werden wir eine $L + 1$-ste Gruppe von transienten Zuständen haben mit der Eigenschaft, dass das System, wenn es in irgendeinem Zustand dieser Gruppe gestartet wird, letzten Endes in eine der L ergodischen Klassen übergeht.

Durch eine Umordnung der Zustände ist es möglich, die Matrix $\mathbf{P}^B$ in folgender kanonischer Form zu schreiben:

$$\mathbf{P}^B = \begin{pmatrix} {}^{11}\mathbf{P} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & {}^{22}\mathbf{P} & \dots & \mathbf{0} & \mathbf{0} \\ \dots & \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & {}^{L\,L}\mathbf{P} & \mathbf{0} \\ {}^{L+1,1}\mathbf{P} & {}^{L+1,\,2}\mathbf{P} & \dots & {}^{L+1,\,L}\mathbf{P} & {}^{L+1,L+1}\mathbf{P} \end{pmatrix}.$$

Die quadratischen Untermatrizen

$${}^{11}\mathbf{P}, {}^{22}\mathbf{P}, \dots, {}^{L\,L}\mathbf{P}$$

sind die Übergangsmatrizen für die Klassen 1, 2, ..., L nach der Umordnung. Jede ist selbst eine stochastische Matrix. Untermatrizen der Form ${}^{rs}\mathbf{P}$ setzen sich aus Nullelementen zusammen, sofern

$$r \neq s \text{ und } r \neq L + 1.$$

Die Untermatrix

$${}^{L+1,L+1}\mathbf{P}$$

ist die Matrix der Übergangswahrscheinlichkeiten innerhalb der transienten Zustände. Einige der Elemente der Untermatrizen

$${}^{L+1,s}\mathbf{P}, \; s = 1, 2, \dots, L$$

müssen positive sein.

Wird dieselbe Umordnung auf die Vektoren

$$\mathbf{g}^\Delta, \mathbf{v}^\Delta, \boldsymbol{\psi}, \boldsymbol{\gamma}, \text{ und } \boldsymbol{\pi}$$

angewendet, erhalten wir eine Menge von Vektoren, die aus $L + 1$ Teilvektoren zusammengesetzt sind:

$$\mathbf{g}^\Delta = \begin{pmatrix} {}^1\mathbf{g}^\Delta \\ {}^2\mathbf{g}^\Delta \\ \vdots \\ {}^L\mathbf{g}^\Delta \\ {}^{L+1}\mathbf{g}^\Delta \end{pmatrix}, \; \mathbf{v}^\Delta = \begin{pmatrix} {}^1\mathbf{v}^\Delta \\ {}^2\mathbf{v}^\Delta \\ \vdots \\ {}^L\mathbf{v}^\Delta \\ {}^{L+1}\mathbf{v}^\Delta \end{pmatrix}, \; \boldsymbol{\psi} = \begin{pmatrix} {}^1\boldsymbol{\psi} \\ {}^2\boldsymbol{\psi} \\ \vdots \\ {}^L\boldsymbol{\psi} \\ {}^{L+1}\boldsymbol{\psi} \end{pmatrix}, \; \boldsymbol{\gamma} = \begin{pmatrix} {}^1\boldsymbol{\gamma} \\ {}^2\boldsymbol{\gamma} \\ \vdots \\ {}^L\boldsymbol{\gamma} \\ {}^{L+1}\boldsymbol{\gamma} \end{pmatrix}$$

$$\boldsymbol{\pi} = \left({}^1\boldsymbol{\pi}, {}^2\boldsymbol{\pi}, \dots, {}^L\boldsymbol{\pi}, {}^{L+1}\boldsymbol{\pi}\right).$$

Der Vektor $\boldsymbol{\pi}$ charakterisiert die Grenzverteilung für den L ergodische Klassen enhaltenden Prozess. Jeder Teilvektor ${}^r\boldsymbol{\pi}$ ist der Vektor der Grenzwahrscheinlichkeiten, wenn das System in einem Zustand der r-ten Klasse startet. Es gilt

$${}^r\boldsymbol{\pi} = {}^r\boldsymbol{\pi} \cdot {}^{r\,r}\mathbf{P}$$

und die Summe der Komponenten von jedem ${}^r\boldsymbol{\pi}$ ist gleich Eins für $r = 1, 2, \dots, L$. Die Komponenten des Teilvek-

tors ${}^{L+1}\boldsymbol{\pi}$ sind alle gleich Null, weil alle Zustände in der Gruppe $L + 1$ transient sind.

Die Gleichungen (6.5.7) und (6.5.8) lauten in Vektorform

(6.5.9) $$\mathbf{g}^{\Delta} = \boldsymbol{\psi} + \mathbf{P}^{B}\,\mathbf{g}^{\Delta}$$

(6.5.10) $$\mathbf{v}^{\Delta} = \boldsymbol{\gamma} + \mathbf{P}^{B}\,\mathbf{v}^{\Delta} - \mathbf{g}^{\Delta}.$$

Wenn die Teilformen in der Gleichung (6.5.9) verwendet werden, erhalten wir

(6.5.11) $${}^{r}\mathbf{g}^{\Delta} = {}^{r}\boldsymbol{\psi} + {}^{r\,r}\mathbf{P}\cdot\,{}^{r}\mathbf{g}^{\Delta},\ r = 1,\ 2,\ \ldots,\ L$$

(6.5.12) $${}^{L+1}\mathbf{g}^{\Delta} = {}^{L+1}\boldsymbol{\psi} + \sum_{s=1}^{L+1} {}^{L+1,s}\mathbf{P}\cdot\,{}^{s}\mathbf{g}^{\Delta}.$$

Die Unterteilung transformiert die Gleichung (6.5.10) in

(6.5.13) $${}^{r}\mathbf{v}^{\Delta} = {}^{r}\boldsymbol{\gamma} + {}^{r\,r}\mathbf{P}\cdot\,{}^{r}\mathbf{v}^{\Delta} - {}^{r}\mathbf{g}^{\Delta},\ r = 1,\ 2,\ \ldots,\ L$$

und

(6.5.14) $${}^{L+1}\mathbf{v}^{\Delta} = {}^{L+1}\boldsymbol{\gamma} + \sum_{s=1}^{L+1} {}^{L+1,s}\mathbf{P}\cdot\,{}^{s}\mathbf{v}^{\Delta} - {}^{L+1}\mathbf{g}^{\Delta}.$$

Nehmen wir an, die Gleichungen (6.5.11) werde mit ${}^{r}\boldsymbol{\pi}$ von links multipliziert, so dass

$${}^{r}\boldsymbol{\pi}\cdot{}^{r}\mathbf{g}^{\Delta} = {}^{r}\boldsymbol{\pi}\cdot{}^{r}\boldsymbol{\psi} + {}^{r}\boldsymbol{\pi}\cdot{}^{r\,r}\mathbf{P}\cdot\,{}^{r}\mathbf{g}^{\Delta}.$$

Da

$${}^{r}\boldsymbol{\pi} = {}^{r}\boldsymbol{\pi}\cdot{}^{r\,r}\boldsymbol{P}$$

gilt, folgt

(6.5.15) $${}^{r}\boldsymbol{\pi}\cdot{}^{r}\boldsymbol{\psi} = 0.$$

Weil sämtliche Zustände in der r-ten Klasse nicht-transient sind, enthält ${}^{r}\boldsymbol{\pi}$ nur positive Elemente. Von unseren früheren Überlegungen her wissen wir, dass alle ψ_i grösser oder gleich Null sind. Aus der Gleichung (6.5.15) geht hervor, dass in jeder Gruppe r für $r = 1,\ 2,\ \ldots,\ L$ ψ_i gleich Null sein muss. Daraus folgt, dass in jeder ergodischen Klasse der Politik B die Entscheidung in jedem Zustand auf dem Werttest anstatt auf dem Gewinntest basieren muss.

Somit wird aus den Gleichungen (6.5.11)

(6.5.16) $${}^{r}\mathbf{g}^{\Delta} = {}^{r\,r}\mathbf{P}\cdot\,{}^{r}\mathbf{g}^{\Delta}.$$

Wir wissen, dass die Lösung dieser Gleichungen ergibt, dass alle

$${}^{r}g_i^{\Delta} = {}^{r}g^{\Delta}$$

sind, so dass alle Zustände in der r-ten Gruppe die gleiche Gewinnerhöhung erfahren, wenn die Politik A nach B wechselt. Wenn dieses Ergebnis in Gleichung (6.5.13) ausgenutzt wird, finden wir die Beziehung

(6.5.17) $${}^{r}g^{\Delta} = {}^{r}\boldsymbol{\pi}\cdot{}^{r}\boldsymbol{\gamma}$$

Somit ist die Gewinnerhöhung für jeden Zustand in der r-ten Gruppe gleich dem Skalarprodukt aus dem Vektor der Grenzwahrscheinlichkeiten für die r-te Gruppe und dem Vektor der Zuwächse der Wcrttestgrösse für diese Gruppe. Da für jede Gruppe r mit $r \le L$,

${}^{r}\psi_i = 0$ gilt,

ist ${}^{r}\gamma_i \ge 0$. Gleichung (6.5.17) zeigt, dass eine Gewinnerhöhung für jeden nichttransienten Zustand der Politik B erfolgt, falls nicht die Politiken A und B äquivalent sind.

Nun müssen wir noch untersuchen, ob der Gewinn der transienten Zustände der Politik B grösser wird oder nicht. Gleichung (6.5.12) zeigt, dass

(6.5.18) $$\left({}^{L+1}I - {}^{L+1,L+1}\mathbf{P}\right){}^{L+1}\mathbf{g}^{\Delta} = {}^{L+1}\boldsymbol{\psi} + \sum_{s=1}^{L} {}^{L+1,s}\mathbf{P}\cdot\,{}^{s}\mathbf{g}^{\Delta},$$

wobei ${}^{L+1}I$ eine Einheitsmatrix von der Dimension der Zahl der Zustände in der transienten Gruppe $L + 1$ ist. Folglich kann die Gewinnänderung der transienten Zustände so dargestellt werden

(6.5.19) $${}^{L+1}\mathbf{g}^{\Delta} = \left({}^{L+1}I - {}^{L+1,L+1}\mathbf{P}\right)^{-1}\left({}^{L+1}\boldsymbol{\psi} + \sum_{s=1}^{L} {}^{L+1,s}\mathbf{P} \cdot {}^{s}\mathbf{g}^{\Delta}\right).$$

Es kann gezeigt werden, dass $\left({}^{L+1}I - {}^{L+1,L+1}\mathbf{P}\right)^{-1}$ existiert und keine negativen Elemente enthält. Wir wissen, dass alle ψ_i grösser oder gleich null sind, dass einige der Elemente der Matrizen ${}^{L+1,s}\mathbf{P}$ für $s = 1, 2, \ldots, L$

positiv und dass keine negativ sind, dass die Gewinnänderungen für die L nichttransienten Gruppen nicht negativ sein können. Folglich kann die Gewinnänderung für alle transienten Zustände der Gruppe $L+1$ nicht negativ sein, und sie wird nur positiv sein, wenn eine oder beide von zwei Bedingungen erfüllt sind. Erstens wird der Gewinn eines transienten Zustandes grösser, wenn sein stochastisches Verhalten so geändert wird, dass es wahrscheinlicher ist, in Klassen mit einem höheren Gewinn zu gelangen. Zweitens wird der Gewinn des transienten Zustandes zunehmen, wenn die Gewinne der Klassen, in welche der transiente Zustand übergeht, erhöht werden.

Somit haben wir gezeigt, dass in dem im Satz 6.3.1 dargestellten Iterationszyklus der Gewinn keines Zustandes vermindert werden kann, und dass der Gewinn irgendeines Zustandes zunehmen muss, wenn nicht äquivalente Politiken vorliegen.

Nun müssen wir zeigen, dass mit dem Iterationszyklus die Politik mit dem höchsten Gewinn in allen Zuständen gefunden wird.

Wir nehmen an, dass die Politik B in irgendeinem Zustand einen höheren Gewinn aufweist als die Politik A, dass aber der Iterationszyklus gegen die Politik A konvergiert. Folglich sind alle

$$\psi_i \le 0,$$

und wenn

$$\psi_i = 0$$

ist, dann folgt

$$\gamma_i \le 0.$$

Gleichung (6.5.17) zeigt, dass alle ${}^{r}g^{\Delta}$ nichtpositiv sind, so dass kein nichttransienter Zustand der Politik B einen höheren Gewinn haben kann als der gleiche Zustand unter der Politik A. Da Gleichung (6.5.19) zeigt, dass alle ${}^{L+1}\mathbf{g}^{\Delta}$ nichtpositiv sind, kann kein transienter Zustand der Politik B einen höheren Gewinn haben als der gleiche Zustand unter der Politik A. Infolgedessen kann kein Zustand unter der Politik B einen höheren Gewinn haben als unter der Politik A und trotzdem die Konvergenz des Iterationszyklus gegen die Politik A aufweisen.

Somit haben wir gezeigt, dass der Iterationszyklus den Gewinn aller Zustände erhöht, bis er in der Politik mit dem höchsten Gewinn in allen Zuständen, d. b. in der optimalen Politik angelangt ist.

■ Beispiel

Diese Diskussion wollen wir nun anhand unseres Beispiels in Tabelle des Abschnitts 6.4 illustrieren. Erinnern wir uns an den Fall, in dem die Politik

$$\mathbf{d} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}, \mathbf{P} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \mathbf{q} = \begin{pmatrix} 3 \\ 4 \\ 8 \end{pmatrix}$$

abgeändert wurde in die Politik

$$\mathbf{d} = \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix}, \mathbf{P} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \mathbf{q} = \begin{pmatrix} 3 \\ 5 \\ 7 \end{pmatrix}$$

mittels der Politik-Verbesserung (letzte Tabelle im Abschnitt 6.4).

Zustand i	Strategie k	Gewinntestgrösse $\sum_{j=1}^{M} p_{ij}^{k} g_i$	Werttestgrösse $q_i^k + \sum_{j=1}^{M} p_{ij}^{k} v_j$
1	1	11 / 2	− 3 / 2
1	2	4	2
1	3	11 / 2	3 ←
2	1	11 / 2	7 / 2
2	2	4	4
2	3	11 / 2	5 ←
3	1	11 / 2	11 / 2
3	2	4	9
3	3	11 / 2	7 ←

Die erste Politik nennen wir Politik *A*, die zweite die Politik *B*. Aus diese Tabelle geht hevor, dass

$$\boldsymbol{\psi} = \begin{pmatrix} \frac{11}{2} - \frac{11}{2} \\ \frac{11}{2} - 4 \\ \frac{11}{2} - \frac{11}{2} \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{3}{2} \\ 0 \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} 3-3 \\ 5-4 \\ 7-\frac{11}{2} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ \frac{3}{2} \end{pmatrix}.$$

Wenn die Zustände 3 und 1 vertauscht werden, haben wir

$$\mathbf{P}^{B} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \boldsymbol{\psi} = \begin{pmatrix} 0 \\ \frac{3}{2} \\ 0 \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} \frac{3}{2} \\ 1 \\ 0 \end{pmatrix}, \mathbf{g}^{\Delta} = \begin{pmatrix} {}^{1}\mathbf{g}^{\Delta} \\ {}^{2}\mathbf{g}^{\Delta} \end{pmatrix}.$$

Demgemäß ist $L = 1$, es gibt eine ergodische Klasse, uns es ist

$${}^{11}\mathbf{P} = 1.$$

Wir sehen, dass die Entscheidung im neuen Zustand 1 (alter Zustand 3) auf den Werten anstatt auf den Gewinn beruht. die Grenzverteilung für $s = 1$

$${}^{1}\boldsymbol{\pi} = 1.$$

Also folgt aus Gleichung (6.5.17)

$${}^{1}g^{\Delta} = \frac{3}{2}.$$

Da

$${}^{22}\mathbf{P} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

sehen wir aus Gleichung (6.5.19), dass

$${}^{2}\boldsymbol{g}^{\Delta} = {}^{2}\boldsymbol{\psi} + {}^{21}\mathbf{P} \cdot {}^{1}\mathbf{g}^{\Delta} = \begin{pmatrix} \frac{3}{2} \\ 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} \cdot \frac{3}{2} = \begin{pmatrix} 3 \\ \frac{3}{2} \end{pmatrix},$$

so dass

$$\mathbf{g}^{\Delta} = \begin{pmatrix} \frac{3}{2} \\ 3 \\ \frac{3}{2} \end{pmatrix}.$$

Wenn nun die Vertauschung der (neuen) Zustände 1 und 3 wiederholt wird, bleibt der Vektor $\mathbf{g}^{\Delta}$ unverändert. Folglich linden wir, dass, indem wir von der Politik *A* zur Politik *B* übergehen, der Gewinn in den Zuständen 1 und 3 um $\frac{3}{2}$ hätte zunehmen sollen, während im Zustand 2 die Gewinnerhöhung 3 Einheiten hätte betragen müssen. Wenn wir nun die früher gelösten Gleichungen der Politik-Auswertung für die Politiken *A* und *B* nachschlagen, so sehen wir, dass diese Werte tatsächlich angenommen wurden.

Wir haben gesehen, dass der sequentielle Entscheidungsprozess mit mehreren ergodischen Klassen mittels einer Methode, die sehr ähnlich derjenigen für ergodische Prozesse mit einer ergodischen Klasse ist, gelöst werden

kann. Dennoch ermöglicht in den meisten praktischen Problemen die Kenntnis des Prozesses die Anwendung der einfacheren Methode (für Prozesse mit nur einer ergodischen Klasse).