

Dynamische Allokationsprobleme und der  
Gittins-Index in diskreter und stetiger Zeit

- Diplomarbeit -

Humboldt-Universität zu Berlin  
Mathematisch-Naturwissenschaftliche Fakultät II  
Institut für Mathematik

eingereicht von: Christian Küchler

geboren am 02.01.1979 in Dresden

Betreuer: Prof. Dr. Peter Bank

Berlin, den 27. Juli 2005

# Zusammenfassung

Gegenstand der vorliegenden Arbeit ist die Untersuchung dynamischer Allokationsprobleme.

Eine Betrachtung des zeitdiskreten Falls und des von Gittins stammenden Konzepts der dynamischen Allokationsindizes bildet den ersten Teil der Arbeit, der sich an den Lösungen von El Karoui und Karatzas [10] und Whittle [29] orientiert. Technisch weniger aufwändig als die Situation stetiger Zeit wird hier die Struktur der Probleme und der Lösung ausführlich beschrieben. Numerische Ergebnisse für das Beispiel Bernoulli-verteilter Auszahlungsprozesse vervollständigen diesen Abschnitt.

Im zweiten Teil der Arbeit werden Allokationsprobleme in stetiger Zeit behandelt. Anhand einer neuen Darstellung des erwarteten Gewinns werden die Gittins-Indexprozesse als Lösungen einer Familie von Darstellungsproblemen optionaler Prozesse charakterisiert. Dieser Ansatz ermöglicht die Herleitung einer Regularität der Pfade der Indexprozesse, einen weiteren Beweis des bekannten Gittins-Indextheorems sowie eine vollständige Beschreibung der Menge der optimalen Strategien. Die Konstruktion einer optimalen Strategie bildet den Abschluss der Arbeit.

## Danksagung

Für die kontinuierliche Unterstützung und zahlreichen wegweisenden Ratschläge während der Entstehung dieser Arbeit danke ich Prof. Peter Bank. Für hilfreiche Gespräche danke ich Dierk Peithmann und für ihren stetigen Beistand meiner Familie.



# Inhaltsverzeichnis

<b>Einleitung</b>	<b>7</b>
<b>1 Mehrarmige Banditen in diskreter Zeit</b>	<b>13</b>
1.1 Übersicht . . . . .	13
1.2 Formulierung des Allokationsproblems . . . . .	14
1.3 Fallende Auszahlungsprozesse . . . . .	16
1.4 Optimales Stoppen . . . . .	19
1.5 Der erweiterte einarmige Bandit . . . . .	22
1.6 Der Gittins-Index . . . . .	25
1.7 Dynamische Programmierung . . . . .	30
1.8 Der Whittle-Ansatz und das Gittins-Indextheorem . . . . .	32
1.9 Eine Zerlegung des Gittins-Index im Bernoulli-Banditen . . . . .	34
1.10 Anhang . . . . .	35
<b>2 Mehrarmige Banditen in stetiger Zeit</b>	<b>43</b>
2.1 Übersicht . . . . .	43
2.2 Formulierung des Allokationsproblems . . . . .	44
2.3 Optionale Projektionen . . . . .	45
2.4 Eine Darstellung der Rendite . . . . .	47
2.5 Das Darstellungsproblem . . . . .	49
2.6 Regularität der Indexprozesse . . . . .	51
2.7 Zwei Schranken für die Rendite . . . . .	53
2.8 Fallende Auszahlungsprozesse . . . . .	54
2.9 Indexstrategien . . . . .	56
2.10 Existenz von Indexstrategien . . . . .	61
2.11 Positive Antizipation der Gittins-Indexprozesse . . . . .	65
2.12 Das Gittins-Indextheorem . . . . .	67
<b>Literaturverzeichnis</b>	<b>69</b>



# Einleitung

Dynamische Allokationsprobleme beschreiben die Aufteilung einer knappen Ressource zwischen mehreren unabhängigen Projekten im Laufe der Zeit. Die verschiedenen Projekte entwickeln sich nur dann weiter, wenn in sie investiert wurde und liefern abhängig von ihrem Zustand zufällige Erträge. Entscheidungen über die Zuteilung der Ressource berücksichtigen vorher erfolgte Auszahlungen und die beobachteten Zustände der Projekte. Gesucht ist eine Strategie, eine Folge von Entscheidungen, die den maximalen Ertrag liefert.

Dieses Konzept spiegelt einen zentralen Konflikt des menschlichen Handelns wider, die Entscheidung zwischen Tätigkeiten, die unmittelbaren Nutzen bringen, und jenen, deren Nutzen erst in der Zukunft sichtbar wird.

Probleme dieser Art können in unterschiedlichen Zusammenhängen auftreten. Beispiele sind das Ausführen verschiedener Aufträge durch eine einzelne Maschine, wiederkehrende Entscheidungen zwischen verschiedenen Forschungsprojekten und medizinische Versuchsreihen. Eine weitere Anwendung bildet die Modellierung von Lernprozessen.

Dynamische Allokationsprobleme werden ihrer Struktur wegen auch als *mehrrarmige Banditen* bezeichnet. Dieser Begriff beschreibt eine Situation, in der sich ein Spieler einer Anzahl von Glücksspielautomaten gegenüber sieht, welche häufig einarmige Banditen genannt werden. Der Spieler muss sich jeweils entscheiden, welchen er durch Einwurf einer Münze betätigt. Durch die Benutzung eines Gerätes erhält er eine zufällige Auszahlung und damit auch Informationen über die den Auszahlungen dieses Gerätes zugrunde liegende Wahrscheinlichkeitsverteilung. Diese kann er bei der Entscheidung über weitere Betätigungen nutzen. Weiter kann man von einem *Zustand* sprechen, in dem sich der Bandit befindet und der sich durch eine Betätigung ändern kann.

Das klassische und sicher meistbehandelte Problem ist das folgende. Es stehen zur Behandlung einer Folge von Patienten verschiedene Mittel zur Verfügung. Das Ergebnis einer Behandlung ist zufällig und die Entscheidung über die Behandlung eines Patienten kann auf der Grundlage der Behandlungsergebnisse der zuvor behandelten Patienten getroffen werden. Gesucht ist eine Strategie, die eine erfolgreiche Behandlung möglichst vieler Patienten ermöglicht. Die Resultate der verschiedenen Behandlungsmöglichkeiten werden

durch unabhängige Folgen von Bernoulli-verteilten Zufallsvariablen modelliert. Diese sind jeweils bedingt auf den gemeinsamen unbekanntem Bernoulli-Parameter der Folge unabhängig.

Allokationsprobleme mit Bernoulli-verteilten Armen wurden seit etwa 1930 untersucht. Die erste Veröffentlichung stammt von Thompson [24]. Er betrachtete zwei unabhängige Bernoulli-Prozesse mit dem Ziel, für einen endlichen Zeithorizont die erwartete Anzahl von Erfolgen zu maximieren. Dazu schlug er eine randomisierte Strategie vor, deren Entscheidungswahrscheinlichkeiten von der Anzahl der beobachteten Erfolge und Misserfolge abhängen.

Robbins [22] formulierte eine Strategie, deren Entscheidung nur vom Ergebnis der letzten Betätigung abhängt: *Betätige denselben Arm bei einem Erfolg, betätige den anderen Arm bei einem Misserfolg*. Sein Ziel war die Maximierung des Erfolgsanteils auf lange Sicht. Diese von ihm betrachtete 'stay on a winner, switch on a loser' Strategie war die Grundlage für die später betrachteten Strategien mit *endlichem Gedächtnis*, die nur die letzten  $r$  Ergebnisse berücksichtigten.

Der Gewinn einer Strategie hängt von den unbekanntem Bernoulli-Parametern ab. Das Konzept, dem Robbins und später andere folgten, besteht aus der Betrachtung einer gewissen Klasse von Strategien. Wenn der Nutzen einer Strategie als Funktion der unbekanntem Parameter den der anderen Strategien dieser Klasse gleichmäßig dominiert, so ist die betrachtete Strategie innerhalb dieser Klasse optimal. Im Allgemeinen kann man jedoch nicht erwarten, eine in diesem Sinne in der Menge aller Strategien optimale Strategie zu finden.

Das Bayessche Konzept, dem neuere Arbeiten folgen, nimmt dagegen an, dass die unbekanntem Parameter einer a priori bekannten Verteilung genügen und versucht, den erwarteten Gewinn einer Strategie unter dieser Verteilung zu maximieren.

Die Dynamik einer Strategie im Bernoulli-Banditen hängt von den sukzessive erfolgten Beobachtungen ab. Für die Kenntnis der bedingten Verteilung der unbekanntem Parameter unter diesen Beobachtungen ist nur eine suffiziente Statistik notwendig. In der betrachteten Situation Bernoulli verteilter Renditen wählt man als solche die Anzahlen der beobachteten Erfolge und Misserfolge. Paare natürlicher Zahlen können damit als *Zustände des Armes* betrachtet werden. Die Entwicklung des Zustandes jedes Armes wird so unter der Bayesschen Betrachtungsweise durch eine Markovkette auf  $\mathbb{N} \times \mathbb{N}$  beschrieben. Insbesondere wird das Allokationsproblem so mit Hilfe der dynamischen Programmierung numerisch lösbar. Die Komplexität der Rechnung ist allerdings sehr hoch, da alle zukünftigen möglichen *gemeinsamen* Entwicklungen der verschiedenen Projekte berücksichtigt werden müssen.

Die wegweisende Arbeit von Gittins und Jones [13] löst schließlich den Fall allgemeiner Renditeprozesse mit Markovschen Zuständen in diskreter Zeit und unter geometrischer Diskontierung. Sie liefert erstmals grundlegende Einsichten in die Struktur mehrarmiger Banditen als spezielle Optimierungsprobleme.

Da die einzelnen Projekte und damit die Entwicklungen der verschiedenen Arme unabhängig voneinander sind, ist die *Reihenfolge* der Betätigungen der unterschiedlichen Arme nur deshalb wichtig, weil die Diskontierung frühere Betätigungen belohnt. Eine optimale Strategie sollte also stets einen Arm wählen, der *in der näheren Zukunft* maximale Renditen liefert. Die Höhe dieser Renditen wird durch den von Gittins eingeführten *dynamischen Allokationsindex* gemessen, welcher seitdem als *Gittins-Index* bekannt ist.

Gittins und Jones zeigen, dass im Markovschen Fall eine Strategie, die den erwarteten Gewinn maximiert, stets einen Arm mit maximalem Index wählt. Dieser hängt nur vom aktuellen Zustand des Armes und der Dynamik seines Renditeprozesses ab. Mit Hilfe dieses Ansatzes kann das mehrdimensionale Allokationsproblem als Familie eindimensionaler Stopp-Probleme formuliert werden. Dadurch erhält man neben einer vereinfachten Struktur auch eine erhebliche Verringerung der numerischen Komplexität des Problems. Das Ergebnis von Gittins und Jones wird meist als Gittins-Indextheorem bezeichnet.

Whittle [29] liefert einen Beweis des Indextheorems, der weitere Einsicht in die Struktur des Problems ermöglicht. Er betrachtet die in der dynamischen Programmierung auftretende Wertfunktion des Problems und nutzt die Unabhängigkeit der Arme, um eine multiplikative Zerlegung dieser Funktion zu erhalten. Varaiya et al. [25] und Mandelbaum [19] erweitern die Lösung über den Markovschen Fall hinaus, dabei formuliert Mandelbaum das Allokationsproblem erstmals als Kontrollproblem mit mehrdimensionalem Zeitparameter. Ein weiterer, dem Ansatz von Whittle folgender Beweis stammt von El Karoui und Karatzas [10]. Erwähnt seien auch noch die umfangreichen Monographien von Gittins [14] und Berry und Fristedt [2]. Letztere beinhaltet neben einer ausführlichen Untersuchung Bernoulli-verteilter Arme auch die Betrachtung anderer Diskontierungsarten.

Eine erste Analyse des zeitstetigen Falls jenseits des Markovschen Rahmens unternimmt Mandelbaum [20]. Er beweist die Optimalität von Strategien, die ausschließlich Arme mit maximalem Gittins-Index betätigen. Dies geschieht unter der Voraussetzung stetiger Renditeprozesse, unter der Annahme stetiger Gittins-Indexprozesse und mit Hilfe einer Approximation der betrachteten Strategien durch solche, die sich wie in diskreter Zeit verhalten. El Karoui und Karatzas [11] zeigen unter der Voraussetzung quasi linksstetiger Filtrationen ebenfalls die Optimalität von Strategien, die nur Arme mit maximalem Gittins-Index betätigen.

Die konkrete Berechnung der Indizes kann kompliziert sein. Kaspi und Mandelbaum [16]

entwickeln für Lévy-Auszahlungsprozesse verschiedene Darstellungen des Index. Weiter zeigen sie für adaptierte Renditeprozesse [17] die Optimalität von Strategien, die ausschließlich Arme mit maximalem Index wählen und zusätzlich die sogenannte *Exkursions-eigenschaft* besitzen. Dieses ist unseres Wissens nach das bisher allgemeinste Resultat in stetiger Zeit.

El Karoui und Karatzas [12] konstruieren eine optimale Indexstrategie, die ebenfalls die Exkursionseigenschaft besitzt. Die von ihnen verwendeten Methoden des Optimalen Stoppens ermöglichen die Verallgemeinerung der Unabhängigkeit der Arme auf die Gültigkeit der *Cairoli-Walsh-Eigenschaft*, also bedingte Unabhängigkeit unter der gemeinsamen Vergangenheit.

Die vorliegende Arbeit besteht aus zwei Teilen. Im ersten Kapitel werden wir uns mit dem zeitdiskreten Fall beschäftigen. Die hier betrachtete Lösung des Allokationsproblems ist bereits bekannt und technisch wenig aufwändig, wir nutzen diesen Rahmen zur ausführlicheren Einführung und Untersuchung der Eigenschaften und Struktur des Problems. Der dazu betrachtete Beweis des Indextheorems im zeitdiskreten Fall folgt der Lösung von El Karoui und Karatzas [10]. Nach der mathematischen Formulierung des Problems untersuchen wir die spezielle Situation fallender Auszahlungsprozesse. Diese ist von besonderer Bedeutung, da verschiedene Ansätze den allgemeinen Fall hierauf zurückführen.

Nach einer Erinnerung an die Grundlagen der Theorie des Optimalen Stoppens wird jeder der Auszahlungsprozesse um eine Familie von Stoppproblemen erweitert. Mit Hilfe dieser kann der Gittins-Index des Armes als Indifferenzwert definiert und die oben erwähnte erwartete Höhe der Auszahlungen in der näheren Zukunft gemessen werden.

Mit dem Ansatz von Whittle [29] wird anschließend die in der Dynamischen Programmierung auftretende Wertfunktion zerlegt. Dies liefert eine Verbindung der Gittins-Indizes mit der Bellman-Gleichung, aus der folgt, dass eine Strategie genau dann optimal ist, wenn sie stets einen Arm mit maximalem Gittins-Index wählt. Weiter werden wir verschiedene im Rahmen dieser Lösung auftretende Größen am Beispiel Bernoulli-verteilter Auszahlungsprozesse numerisch berechnen.

Im zweiten Kapitel wenden wir uns der stetigen Zeit zu und entwickeln eine neue Beweisstrategie des Gittins-Indextheorems, die sich insbesondere grundlegend von der im zeitdiskreten Fall betrachteten unterscheidet. Da keine besonderen Annahmen an die Regularität der Auszahlungsprozesse oder deren Filtrationen getroffen werden, sind der technische Aufwand hier höher und die Formulierung subtiler als in diskreter Zeit.

Nach der Einführung der grundlegenden Begriffe werden wir an einige wesentliche Eigenschaften optionaler Prozesse und optionaler Projektionen erinnern und eine neue Darstellung für den Gewinn einer Strategie herleiten. Diese liefert die Verbindung der von Bank

und El Karoui [1] gelösten Darstellungsprobleme für optionale Prozesse mit unserem Allokationsproblem. Mit Hilfe der Konstruktion von Bank und El Karoui werden wir zeigen, dass die Lösungen der Darstellungsprobleme in unserem Fall unterhalbstetig von rechts sind. Dies ermöglicht eine Charakterisierung der Gittins-Indexprozesse als Lösungen solcher Darstellungsprobleme und die Reduzierung des allgemeinen Allokationsproblems auf eines mit fallenden rechtsstetigen Auszahlungsprozessen. Mit der von El Karoui und Karatzas [12] erstmals in diesem Zusammenhang eingeführten *Synchronization Identity* lässt sich dieses sehr leicht lösen.

Anhand eines einfachen Beispiels werden wir zeigen, dass es im Allgemeinen in stetiger Zeit im Gegensatz zum zeitdiskreten Fall für die Optimalität einer Strategie nicht hinreichend ist, ausschließlich Arme mit maximalem Gittins-Index zu betätigen.

Mit Hilfe der Regularität der Pfade der Indexprozesse werden wir das Gittins-Indextheorem beweisen und zeigen, dass die von Kaspi und Mandelbaum [17] definierten Indexstrategien, die die Exkursionseigenschaft besitzen, genau die optimalen Strategien ergeben.

Durch die Konstruktion einer solchen Indexstrategie werden wir schließlich die Existenz einer optimalen Strategie beweisen. Diese stammt im Wesentlichen von El Karoui und Karatzas [12] und setzt in ihrer ursprünglichen Formulierung die pfadweise Unterhalbstetigkeit von links der Indexprozesse voraus. Wir werden den in unserem allgemeineren Rahmen noch offenen Fall mit Hilfe einer Konstruktion von Kaspi und Mandelbaum [17] lösen, nach der die problematischen Aufwärtssprünge der Indexprozesse vorhersehbar sind.



# Kapitel 1

## Mehrmarmige Banditen in diskreter Zeit

### 1.1 Übersicht

In diesem Kapitel unternehmen wir eine ausführliche Einführung und Untersuchung des Problems des mehrarmigen Banditen und seiner Eigenschaften in diskreter Zeit.

Nach der mathematischen Formulierung des Allokationsproblems und einer ersten Betrachtung des Bernoulli-Banditen in Abschnitt 1.2 beweisen wir in Abschnitt 1.3 die Optimalität *kurzsichtiger* Strategien in der Situation fallender Auszahlungsprozesse. Dies geschieht mit Hilfe eines einfachen Tauscharguments, ähnlich dem von Gittins und Jones [13].

In Abschnitt 1.4 erinnern wir an einige Grundlagen von Snells Theorie des Optimalen Stoppens. Dabei orientieren wir uns an Neveu [21]. Anschließend folgen wir dem Beweis des Indextheorems von El Karoui und Karatzas [10]. Dazu assoziieren wir zunächst in Abschnitt 1.5 mit jedem Arm eine Familie von Stoppproblemen, indem man die Betätigung des Armes mit dem Erhalt einer zusätzlichen Zahlung vergleicht. Mit Hilfe einer Indifferenz- oder Schwellenwertcharakterisierung kann man so die erwartete Höhe der Auszahlungen des Armes in der näheren Zukunft quantifizieren. Der entsprechende Wert wird als Gittins-Index definiert. Dies erfolgt in Abschnitt 1.6, hier findet man neben verschiedenen Darstellungen des Gittins-Index auch numerische Ergebnisse für das Beispiel Bernoulli-verteilter Auszahlungsprozesse. Weiter werden für dieses Beispiel von Berry und Fristedt [2] angegebene untere Schranken für den Index betrachtet.

In Abschnitt 1.7 wird der mehrarmige Bandit ebenfalls um die Möglichkeit zusätzlicher Zahlungen erweitert und die bei der Dynamischen Programmierung auftretenden Größen definiert. Die Unabhängigkeit der verschiedenen Projekte macht mehrarmige Banditen zu besonderen Optimierungsproblemen. Mit Hilfe dieser Eigenschaft wird in Abschnitt 1.8, einem Ansatz von Whittle [29] folgend, eine multiplikative Zerlegung für die in der

Dynamischen Programmierung genutzte Wertfunktion hergeleitet. Jeder der auftretenden Faktoren hängt nur noch vom Auszahlungsprozess jeweils eines Armes ab. Dadurch wird die Struktur der Wertfunktion vereinfacht und eine Verbindung zwischen der Bellman-Gleichung des erweiterten Allokationsproblems und den Gittins-Indizes der Arme aufgezeigt. Es folgt, dass die Wertfunktion unter einer Strategie stets ein Supermartingal ist und ein Martingal genau dann, wenn die Strategie stets einen Arm mit maximalem Gittins-Index aktiviert. Dies beweist schließlich das Indextheorem im zeitdiskreten Fall. In Abschnitt 1.9 kommen wir erneut auf das Beispiel Bernoulli-verteilter Auszahlungsprozesse zurück und betrachten Resultate von Wang [27] und Gittins und Wang [15]. Der Gittins-Index und damit die optimalen Strategien berücksichtigen neben der erwarteten Auszahlung eines Armes im nächsten Zeitpunkt die Möglichkeit einer zukünftigen Verbesserung. Der Index kann damit in zwei Komponenten zerlegt werden.

Mit zunehmender Anzahl von Beobachtungen stehen mehr Informationen über die unbekannt Parameter zur Verfügung und eine bedeutende Veränderung der geschätzten Erfolgswahrscheinlichkeit ist nicht mehr zu erwarten. Es wird gezeigt, dass die mögliche zukünftige Veränderungen bewertende Komponente gegen Null konvergiert und numerische Beobachtungen dieser Asymptotik dargestellt.

## 1.2 Formulierung des Allokationsproblems

Es stehen  $d \in \mathbb{N}$  verschiedene Verwendungsmöglichkeiten für eine knappe Ressource, die wir im Folgenden als Zeit interpretieren, zur Verfügung. Diese Alternativen nennen wir *Arme* des Banditen. Die zufälligen Ergebnisse der Arme werden durch unabhängige stochastische Prozesse

$$(Z_k(t))_{t \geq 1}, \quad k = 1, \dots, d$$

auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{F}, \mathbb{P})$  modelliert und als Auszahlungs- oder Renditeprozesse bezeichnet.  $Z_k(t)$  nehme für gegebene Konstanten  $K \in (0, \infty)$  und  $\alpha \in (0, 1)$  Werte in  $[0, (1 - \alpha)K]$  an. In jedem Zeitpunkt  $t \in \mathbb{N}$  muss genau eine dieser  $d$  Möglichkeiten genutzt werden. Wird zur Zeit  $t$  der Arm  $k$  gewählt, so liefert dieser im Zeitpunkt  $t + 1$  eine zufällige Auszahlung.  $Z_k(t)$  gibt die Höhe der Auszahlung an, die Bandit  $k$  liefert, nachdem er zum  $t$ . Mal betätigt wurde.

Die Bewertung zukünftiger Zahlungen erfolgt mittels geometrischer Diskontierung, der Wert zum Zeitpunkt 0 einer im zukünftigen Zeitpunkt  $t$  erfolgenden Zahlung in Höhe von  $X$  sei  $\alpha^t X$ . Die Auszahlungsprozesse genügen der Integrierbarkeitsbedingung

$$(1.1) \quad \mathbb{E} \left[ \sum_{t=1}^{\infty} \alpha^t Z_k(t) \right] < \infty.$$

Für  $k = 1, \dots, d$  werden die durch sukzessive Beobachtung der Auszahlungen des Armes  $k$  akkumulierten Informationen durch die unabhängigen Filtrationen  $\mathcal{F}^k = (\mathcal{F}^k(t))_{t \geq 0}$  dargestellt. Der Prozess  $Z_k$  sei adaptiert an  $\mathcal{F}^k$ . Hat man die Arme  $1, \dots, d$  jeweils  $s_1, \dots, s_d \in \mathbb{N}_0$  mal betätigt, so werden die dann verfügbaren Informationen durch die  $\sigma$ -Algebra

$$\mathcal{F}(\tilde{s}) \triangleq \bigvee_{i=1}^d \mathcal{F}^k(s_i), \tilde{s} = (s_1, \dots, s_d) \in \mathbb{N}_0^d$$

modelliert.

Eine *Strategie*  $T$  ist ein  $\mathbb{N}_0^d$ -wertiger stochastischer Prozess

$$T(t) = (T_1(t), \dots, T_d(t)).$$

Dabei gibt  $T_k(t)$  die Zeit an, die man bis zum Zeitpunkt  $t$  in Arm  $k$  investiert hat. Da man in jedem Zeitpunkt genau einen Arm aktiviert und bei der Entscheidung für einen Arm nur die bis zu diesem Zeitpunkt verfügbaren Informationen berücksichtigen kann, soll eine Strategie die folgenden Bedingungen erfüllen:

$$(1.2) \quad \begin{aligned} T(0) &= 0 \text{ und } T(t) \text{ ist wachsend in } t, \\ T_1(t) + \dots + T_d(t) &= t \quad \forall t \in [0, \infty), \\ \{T(t+1) = T(t) + \tilde{e}_i, T(t) = \tilde{s}\} &\in \mathcal{F}(\tilde{s}) \quad \forall t \in [0, \infty), s \in \mathbb{N}_0^d, \end{aligned}$$

$\tilde{e}_i$  bezeichnet den  $i$ . Einheitsvektor in  $\mathbb{N}_0^d$ . Strategien können als mehrdimensionale Zeitwechsel interpretiert werden, die Menge der Strategien nennen wir  $\mathcal{S}(0)$ .

Setzt man  $Z_k(0) = 0$ , so liefert die Strategie  $T$  den zufälligen Gewinn

$$(1.3) \quad \mathcal{R}(T) \triangleq \sum_{t=1}^{\infty} \alpha^t \sum_{k=1}^d Z_k(T_k(t)) \cdot (T_k(t) - T_k(t-1)).$$

Das Problem des mehrarmigen Banditen besteht darin, eine Strategie  $\hat{T}$  zu finden, die den erwarteten Gewinn maximiert:

$$(1.4) \quad \mathbb{E}[\mathcal{R}(\hat{T})] = \max_{T \in \mathcal{S}(0)} \mathbb{E}[\mathcal{R}(T)].$$

Das folgende Beispiel beschreibt das klassische Problem des Bernoulli-Banditen.

**Beispiel 1.1.** *Es seien  $(Z_1(t))_{t \geq 1}$  und  $(Z_2(t))_{t \geq 1}$  zwei unabhängige Folgen von Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{F}, \mathbb{P})$ . Diese seien Bernoulli-verteilt zu den unbekanntem Parametern  $\Theta_1$  und  $\Theta_2$ . Die Parameter seien ebenfalls unabhängig und gleichverteilt auf  $[0, 1]$ . Bedingt auf  $\Theta_i$  sollen die Zufallsvariablen  $Z_i(t), t \geq 1$ , unabhängig sein, es soll also gelten*

$$(1.5) \quad \mathbb{P}[Z_i(1) = z_1, \dots, Z_i(n) = z_n \mid \Theta_i = \theta] = \prod_{k=1}^n (\theta \cdot 1_{\{z_k=1\}} + (1-\theta) \cdot 1_{\{z_k=0\}}).$$

Die sukzessiven Beobachtungen der Zufallsvariablen liefern Informationen über  $\Theta_i$ , die von einer optimalen Strategie berücksichtigt werden sollten. Wurden die Zufallsvariablen  $Z_i(1), \dots, Z_i(n)$  beobachtet und sind  $a_i(n) = \sum_{j=1}^n 1_{\{Z_i(j)=1\}}$  die Zahl der Erfolge,  $b_i(n) = n - a_i(n)$  die der Misserfolge, so ist  $\Theta_i$  unter diesen Beobachtungen betaverteilt zu den Parametern  $a_n^i + 1, b_n^i + 1$ . Insbesondere ist

$$\mathbb{E}[\Theta_i \mid Z_i(1), \dots, Z_i(n)] = \frac{a_n^i + 1}{a_n^i + b_n^i + 2}.$$

Die bedingte Verteilung von  $\Theta_i$  unter den ersten  $n$  Beobachtungen des Armes  $i$  hängt nur von  $(a_n^i, b_n^i)$  ab. Der Zustand eines Bernoulli-Armes kann durch ein solches Paar charakterisiert werden und bildet eine Markovkette auf  $\mathbb{N} \times \mathbb{N}$ .

Eine mögliche Strategie ist es, stets einen Arm  $i$  mit dem höchsten bedingten Erwartungswert von  $\Theta_i$  zu wählen. Diese einfache Strategie liefert tatsächlich gute Ergebnisse, ist aber nicht optimal. Sind beispielsweise zum Zeitpunkt  $n = n_1 + n_2$  die bedingten Erwartungswerte der  $\Theta_i$  gleich, haben wir den  $i$ . Arm jeweils  $n_i$  mal aktiviert und gilt  $n_1 > n_2$ , so ist die bedingte Varianz von  $\Theta_2$  größer als die von  $\Theta_1$ . Damit ist eine zukünftige Änderung des bedingten Erwartungswertes von  $\Theta_2$  wahrscheinlicher als von  $\Theta_1$ . Von einer positiven Änderung profitiert man in der gesamten Zukunft, im Falle einer negativen Änderung wählt man statt dessen den anderen Arm. Es ist also anzunehmen, dass eine optimale Strategie nicht nur den bedingten Erwartungswert, sondern auch die Möglichkeit zukünftiger Veränderungen dieses Wertes berücksichtigt.

Dieses Beispiel macht deutlich, dass es unter Umständen vorteilhaft ist, kurzfristig geringere Auszahlungen in Kauf zu nehmen um zusätzliche Informationen zu erhalten. Ebenso kann man Beispiele betrachten, in denen die Betätigung eines Armes mit geringeren Auszahlungen in naher Zukunft geboten ist, um dadurch erst den Zugang zu später folgenden Auszahlungen zu erhalten. Im Allgemeinen ist die Bestimmung einer optimalen Strategie also ein komplexes Problem.

### 1.3 Fallende Auszahlungsprozesse

Sind die Renditeprozesse  $(Z_k(t))_{t \geq 0}$   $\mathbb{P}$ -fast sicher monoton fallend in  $t$ , so kann es nicht von Vorteil sein, einen Arm mit geringer kurzfristiger Auszahlung zu aktivieren. Weder wird dieser dadurch zukünftig höhere Auszahlungen liefern, noch können zusätzliche Informationen höhere zukünftige Auszahlungen offenbaren. Die Diskontierung bestraft dagegen eine Verzögerung der Betätigung eines Armes, wenn dieser die höchste Auszahlung im nächsten Zeitpunkt verspricht.

Optimale Strategien in der Situation fallender Auszahlungsprozesse sollten also darin bestehen, kurzfristig zu handeln. Dabei wird jeweils der Arm aktiviert, der die höchste

bedingte erwartete Auszahlung im nächsten Zeitpunkt aufweist.

Einige Beweisstrategien ([28],[17]) führen das allgemeine Problem auf ein Problem mit fallenden Renditeprozessen zurück. Wir werden das Resultat dieses Abschnitts für unsere zeitdiskrete Untersuchung nicht benötigen. Der Beweis beruht auf einem Tauschargument, wie es in ähnlicher Form auch im ursprünglichen Beweis des Indextheorems von Gittins und Jones [13] verwendet wurde.

Unserem Beweis der Optimalität kurzfristiger Strategien liegt die folgende Überlegung zugrunde. Eine Strategie, die zu einem Zeitpunkt nicht einen Arm mit maximaler erwarteter Rendite im nächsten Schritt wählt, kann man verbessern. Dies geschieht, indem man solch einen Arm maximaler Rendite aktiviert und anschließend mit der betrachteten Strategie fortfährt. So erhält man eine Folge von Strategien, die gegen die kurzfristige Strategie konvergiert und deren erwartete Gewinne zunehmen.

**Theorem 1.2.** *Es sei  $\hat{T}$  eine Strategie, die für alle Zeitpunkte  $t$  die folgende Bedingung erfüllt:*

$$\begin{aligned} \text{Es gilt } \hat{T}_k(t+1) &= \hat{T}_k(t) + 1 \text{ nur dann, wenn} \\ \mathbb{E} \left[ Z_k(\hat{T}_k(t+1)) \mid \mathcal{F}(\hat{T}(t)) \right] &= \bigvee_{i=1}^d \mathbb{E} \left[ Z_i(\hat{T}_i(t+1)) \mid \mathcal{F}(\hat{T}(t)) \right]. \end{aligned}$$

*Durch eine solche Strategie  $\hat{T}$  wird der erwartete Gewinn maximiert, das heißt es gilt*

$$\mathbb{E} \left[ \mathcal{R}(\hat{T}) \right] = \max_{T \in \mathcal{S}(0)} \mathbb{E} \left[ \mathcal{R}(T) \right].$$

*Beweis.* Die Strategie  $\hat{T}$  liefert den erwarteten Gewinn

$$\mathbb{E} \left[ \mathcal{R}(\hat{T}) \right] = \mathbb{E} \left[ \sum_{t=1}^{\infty} \alpha^t \bigvee_{k=1}^d \mathbb{E} \left( Z_k(\hat{T}_k(t+1)) \mid \mathcal{F}(\hat{T}(t)) \right) \right].$$

Auf der Menge der Strategien  $\mathcal{S}(0)$  definieren wir den Abstand  $\tilde{d}$  zweier Strategien  $T^1$  und  $T^2$  durch

$$\tilde{d}(T^1, T^2) = \left\| \sum_{t=0}^{\infty} \alpha^t \mathbf{1}_{T^1(t) \neq T^2(t)} \right\|_{\mathbb{L}^{\infty}}.$$

Eine Folge  $(T^n)$  von Strategien konvergiert damit genau dann gegen eine Strategie  $T$ , wenn zu jedem  $t \geq 0$  ein  $n \in \mathbb{N}$  existiert, so dass alle  $T^m$  mit  $m \geq n$   $\mathbb{P}$ -fast-sicher bis zum Zeitpunkt  $t$  mit der Strategie  $T$  übereinstimmen.

Das Funktional  $\mathbb{E}[\mathcal{R}(\cdot)]$  ist stetig auf  $(\mathcal{S}(0), \tilde{d})$ , denn stimmen zwei Strategien  $T^1$  und  $T^2$

bis zum Zeitpunkt  $t$  überein, so gilt wegen der Monotonie der Renditeprozesse

$$\begin{aligned} |\mathbb{E}[\mathcal{R}(T^1)] - \mathbb{E}[\mathcal{R}(T^2)]| &\leq \mathbb{E}\left[\sum_{s=t+1}^{\infty} \alpha^s \sum_{k=1}^d (Z_k(T_k^1(s)) + Z_k(T_k^2(s)))\right] \\ &\leq \alpha^t \mathbb{E}\left[\sum_{s=1}^{\infty} \alpha^s \sum_{k=1}^d 2Z_k(1)\right] \rightarrow 0 \text{ für } t \uparrow \infty. \end{aligned}$$

Es sei nun  $T$  eine beliebige Strategie. Wir definieren rekursiv eine Folge  $(T^n)$ . Dazu sei  $T^0 \triangleq T$ . Zu  $T^n$  sei  $\tau_n$  die Stoppzeit

$$\tau_n \triangleq \inf\{t \geq 0 : T^n(t+1) \neq \hat{T}(t+1)\},$$

also der erste Zeitpunkt, in dem die Strategie  $T^n$  einen anderen Arm wählt als  $\hat{T}$ . Es sei  $k_n$  der Arm, den  $\hat{T}$  in  $\tau_n$  wählt, das heißt  $\hat{T}_{k_n}(\tau_n+1) - \hat{T}_{k_n}(\tau_n) = 1$ . Die Stoppzeit  $\tilde{\tau}_n$  sei der erste Zeitpunkt nach  $\tau_n$ , in dem die Strategie  $T^n$  den Arm  $k_n$  wählt:

$$\tilde{\tau}_n \triangleq \inf\{t \geq \tau_n : T_{k_n}^n(t+1) - T_{k_n}^n(\tau_n) = 1\}.$$

Es ist möglich, dass die Strategie  $T^n$  ab dem Zeitpunkt  $\tau_n$  nicht mehr den Arm  $k_n$  aktiviert, also  $\tilde{\tau}_n = \infty$  gilt.  $T^{n+1}$  wird nun wie folgt definiert:

$$T^{n+1}(t) \triangleq \begin{cases} T^n(t) & t \leq \tau_n \\ T^n(t-1) + e_{k_n} & \tau_n < t < \tilde{\tau}_n + 1 \\ T^n(t) & t \geq \tilde{\tau}_n + 1. \end{cases}$$

Es gilt  $\tau_n \geq n$  und  $T^n(t) = \hat{T}(t)$  für  $t \leq \tau_n + 1$ , also konvergiert die Folge der Strategien  $(T^n)$  gegen  $\hat{T}$  bezüglich  $\tilde{d}$ .

Wir zeigen nun, dass das Funktional  $\mathbb{E}[\mathcal{R}(\cdot)]$  entlang der Folge  $(T^n)$  wächst, woraus die

Behauptung  $\mathbb{E}[\mathcal{R}(T)] \leq \mathbb{E}[\mathcal{R}(\hat{T})]$  folgt. Es gilt

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(T^{n+1})] - \mathbb{E}[\mathcal{R}(T^n)] &= \mathbb{E}\left[\alpha^{\tau_n+1} Z_{k_n}(\hat{T}_{k_n}(\tau_n + 1))\right] \\
&\quad + \mathbb{E}\left[\sum_{t=\tau_n+2}^{\tilde{\tau}_n+1} \alpha^t \sum_{k=1}^d Z_k(T_k^n(t-1))(T_k^n(t-1) - T_k^n(t-2))\right] \\
&\quad - \mathbb{E}\left[\sum_{t=\tau_n+1}^{\tilde{\tau}_n} \alpha^t \sum_{k=1}^d Z_k(T_k^n(t))(T_k^n(t) - T_k^n(t-1))\right] \\
&\quad - \mathbb{E}\left[\alpha^{\tilde{\tau}_n+1} Z_{k_n}(\hat{T}_{k_n}(\tau_n + 1))\right] \\
&= \mathbb{E}\left[(\alpha^{\tau_n+1} - \alpha^{\tilde{\tau}_n+1}) Z_{k_n}(\hat{T}_{k_n}(\tau_n + 1))\right] \\
&\quad + \mathbb{E}\left[(\alpha - 1) \sum_{t=\tau_n+1}^{\tilde{\tau}_n} \alpha^t \sum_{k=1}^d Z_k(T_k^n(t))(T_k^n(t) - T_k^n(t-1))\right] \\
&\geq \mathbb{E}\left[(\alpha^{\tau_n+1} - \alpha^{\tilde{\tau}_n+1}) Z_{k_n}(\hat{T}_{k_n}(\tau_n + 1))\right] \\
&\quad + \mathbb{E}\left[(\alpha - 1) \sum_{t=\tau_n+1}^{\tilde{\tau}_n} \alpha^t Z_{k_n}(\hat{T}_{k_n}(\tau_n + 1))\right] \\
&= 0.
\end{aligned}$$

Die Ungleichung gilt wegen

$$\mathbb{E}\left[Z_{k_n}[\hat{T}_{k_n}(\tau_n) + 1]\right] = \bigvee_{i=1}^d \mathbb{E}\left[Z_i[\hat{T}_i(\tau_n) + 1]\right]$$

und der Monotonie der Renditeprozesse. □

## 1.4 Optimales Stoppen

Wir erinnern in diesem Abschnitt an einige Grundlagen der Theorie des Optimalen Stoppens. Eine Einführung in diese Theorie in diskreter Zeit findet man beispielsweise in Neveu [21].

Es sei  $(X_n)_{n \in \mathbb{N}}$  eine adaptierte Folge integrierbarer Zufallsvariablen auf einem filtrierten Wahrscheinlichkeitsraum  $(\Omega, \mathcal{F}, (\mathcal{F}_n), \mathbb{P})$  mit

$$\mathbb{E}[\sup_{n \in \mathbb{N}} X_n^+] < \infty.$$

Weiter seien die  $(X_n)$  nach unten durch eine integrierbare Zufallsvariable  $\tilde{X}$  beschränkt:

$$(1.6) \quad \inf_{n \in \mathbb{N}} X_n > \tilde{X} \in \mathbb{L}^1.$$

Betrachtet man  $X_n$  als Gewinn eines Spielers bis zum Zeitpunkt  $n$ , so stellt sich die Frage, wie lange er spielen soll. Den Zeitpunkt des Spielendes kann er dynamisch, das

heißt abhängig vom Spielverlauf wählen. Stoppt er im Zeitpunkt  $n$  das Spiel, so soll er dies allerdings nur aufgrund der bis  $n$  verfügbaren Information tun, der zufällige Endzeitpunkt  $\nu$  für das Spiel soll also eine  $(\mathcal{F}_n)$ –Stoppzeit sein. Die Menge der  $\mathbb{P}$ -fast-sicher endlichen Stoppzeiten bezeichnen wir mit  $\Lambda$ . Weiter sei  $\Lambda(n) \triangleq \{\nu \in \Lambda : \nu \geq n\}$ .

Man sucht nun nach einer optimalen Stoppzeit  $\nu_0$ , die den erwarteten Gewinn maximiert:

$$\mathbb{E}[X_{\nu_0}] = \sup_{\nu \in \Lambda} \mathbb{E}[X_\nu].$$

Die Lösung von Snell [23] dieses Problems wird klar, wenn man zunächst annimmt, dass die sukzessiven Gewinne  $(X_n) = (x_n)$  deterministisch sind. Man sucht nach einem  $p_0 \in \mathbb{N}$  mit

$$x_{p_0} = \sup_n x_n.$$

Angenommen, solch ein  $p_0$  existiert. Man kann die fallende Folge

$$y_n \triangleq \sup_{p \geq n} x_p.$$

definieren und macht die folgende Beobachtung. Die Folge  $(y_n)$  dominiert  $(x_n)$  und der kleinste Index  $n$  mit  $y_n = x_n$  ist gerade das optimale  $p_0$ . Weiter gilt

$$y_0 = y_{p_0} = x_{p_0} = \sup_{n \geq 0} x_n.$$

Insbesondere ist  $(y_n)$  die kleinste fallende Folge, die  $(x_n)$  dominiert.

Im stochastischen Fall ist das Supremum über die  $X_n$  a priori nicht bekannt. Die Rolle der Folge  $(y_n)$  übernehmen nun die Zufallsvariablen

$$(1.7) \quad Y_n \triangleq \operatorname{ess\,sup}_{\nu \in \Lambda(n)} \mathbb{E}[X_\nu \mid \mathcal{F}_n].$$

Die Folge  $(Y_n)$  ist nicht mehr fallend, sondern ein Supermartingal.  $(Y_n)$  wird als Snellsche Enveloppe von  $(X_n)$  bezeichnet. Die folgenden Propositionen zeigen, dass im stochastischen Fall analoge Aussagen gelten.

**Proposition 1.3** ([21], Proposition VI-1-2).  *$(Y_n)_{n \in \mathbb{N}}$  ist eine adaptierte Folge integrierbarer Zufallsvariablen. Diese erfüllen die Gleichungen*

$$(1.8) \quad Y_n = \max(X_n, \mathbb{E}[Y_{n+1} \mid \mathcal{F}_n]).$$

*Sind die Zufallsvariablen  $X_n$  positiv, so ist  $(Y_n)$  das kleinste  $(\mathcal{F}_n)$ –Supermartingal, das die Folge  $(X_n)$  dominiert. Weiter gilt*

$$(1.9) \quad \mathbb{E}[Y_n] = \sup_{\nu \in \Lambda(n)} \mathbb{E}[X_\nu].$$

Auch die optimale Stoppzeit ist von ähnlicher Gestalt wie im deterministischen Fall.

**Proposition 1.4** ([21], Propositions VI-1-3,4). *Das Supremum  $\sup_{\nu \in \Lambda} \mathbb{E}[X_\nu]$  wird angenommen, genau dann wenn die von dem Supermartingal  $(Y_n)$  definierte Stoppzeit*

$$(1.10) \quad \nu_0 \triangleq \inf(n \in \mathbb{N} : Y_n = X_n)$$

$\mathbb{P}$ -fast-sicher endlich ist. In diesem Fall ist  $\mathbb{E}[X_{\nu_0}] = \sup_{\nu \in \Lambda} \mathbb{E}[X_\nu]$  und  $\nu_0$  ist die kleinste endliche Stoppzeit, die dies erfüllt.

Weiter ist der Prozess  $(Y_{\nu_0 \wedge n})$  ein integrierbares Martingal.

Die konkrete Berechnung der  $(Y_n)$  erweist sich jedoch meist als nicht ganz einfach. Allerdings ist es unter Umständen möglich, das Problem approximativ zu lösen, indem man von einem endlichen Zeithorizont  $\Xi$  ausgeht. Für diesen setzt man

$$\begin{aligned} Y_\Xi^\Xi &\triangleq X_\Xi, \\ Y_n^\Xi &\triangleq \max(X_n, \mathbb{E}[Y_{n+1}^\Xi | \mathcal{F}_n]) \text{ für } n < \Xi. \end{aligned}$$

Man sucht nun nach einer Stoppzeit  $\nu_0 \leq \Xi$  mit

$$\mathbb{E}[X_{\nu_0^\Xi}] = Y_0^\Xi = \sup_{\nu \leq \Xi} \mathbb{E}[X_\nu].$$

Eine solche ist wieder  $\nu_0^\Xi \triangleq \min(n \leq \Xi : Y_n^\Xi = X_n)$ . Für eine sinnvolle Approximation sollte für  $\Xi \rightarrow \infty$  gelten  $Y^\Xi \rightarrow Y$ .

Wegen  $Y_n^\Xi \leq Y_n^{\Xi+1}$  können wir definieren

$$Y_n^\infty \triangleq \lim_{\Xi \rightarrow \infty} \uparrow Y_n^\Xi.$$

Es sei  $\Lambda^b(n) \triangleq \{\nu \in \Lambda : \nu \geq n \text{ und beschränkt}\}$ . Man zeigt leicht, dass gilt

$$Y_n^\infty = \sup_{\nu \in \Lambda_n^b} \mathbb{E}[X_\nu | \mathcal{F}(n)]$$

und  $Y^\infty$  das kleinste integrierbare Supermartingal ist, das  $X$  dominiert. Insbesondere gilt  $Y^\infty \leq Y$ . Existiert eine integrierbare untere Schranke für den Prozess  $X$ , wie in (1.6) vorausgesetzt, so gilt mit dem Lemma von Fatou für Stoppzeiten  $\nu \in \Lambda_n$ :

$$\mathbb{E}[X_\nu | \mathcal{F}(n)] \leq \liminf_k \mathbb{E}[X_{\nu \wedge k} | \mathcal{F}(n)] \leq Y_n^\infty.$$

Damit ist  $Y = Y^\infty$ . Für beliebige Renditeprozesse muss dies jedoch nicht gelten, ein Beispiel findet man in Neveu [21].

## 1.5 Der erweiterte einarmige Bandit

Mehrmarmige Banditen sind besondere Optimierungsprobleme. Da die Renditeprozesse unabhängig sind und sich nur nach einer Betätigung weiterentwickeln, liefert die Aktivierung eines Armes weder Informationen über andere Arme, noch verändern sich dadurch deren Auszahlungen oder Zustände. Die Reihenfolge, in der die verschiedenen Arme betätigt werden, spielt also nur deshalb eine Rolle, weil die spätere Auszahlung hoher Renditen aufgrund der Diskontierung nachteilig ist. Arme mit hohen Auszahlungen in der näheren Zukunft sollten also eher betätigt werden.

Wie kann man die Höhe der Auszahlungen eines Armes *in der näheren Zukunft* messen? Dies geschieht durch den Vergleich des Armes mit einer deterministischen Auszahlung. Dabei betrachtet man das folgende Stoppproblem. Bis zu einer Stoppzeit  $\tau$  betätigt man ausschließlich diesen Arm und erhält die dabei anfallenden Auszahlungen. In  $\tau$  erfolgt die deterministische Zahlung  $m$ .

Für  $m = 0$  wird man möglicherweise nie stoppen, je größer  $m$  ist, desto eher wird eine weitere Betätigung des Armes nachteilig gegenüber der Zahlung  $m$ . Man betrachtet den Schwellenwert  $M$ , für den man indifferent zwischen einer vorläufigen Betätigung des Armes und späterem Erhalt von  $m$  sowie dem sofortigen Erhalt von  $m$  ist. Dieser Wert kann als Maß für die Höhe der erwarteten Auszahlungen des Armes in der näheren Zukunft interpretiert werden. Man sollte stets einen Arm wählen, der einen hohen Wert  $M$  besitzt. Der erste Schritt zur Lösung unseres Allokationsproblems ist also eine Erweiterung um eine Familie von Stoppproblemen.

Wir betrachten Arm  $k$  mit Auszahlungsprozess  $(Z_k(t))$  für  $k \in \{1, \dots, d\}$ . Beenden wir die Betätigung des Armes im Zeitpunkt  $t$ , so erfolgt in diesem Zeitpunkt eine deterministische Zahlung  $m \in [0, \infty)$ . Das Spielen bis zum Zeitpunkt  $t$  liefert dann gerade

$$(1.11) \quad X_k(t, m) \triangleq \sum_{u=1}^t \alpha^u Z_k(u) + \alpha^t m.$$

Beendet man das Spiel zeitig, so verzichtet man auf Auszahlungen des Armes. Betätigt man dagegen den Arm eine längere Zeit, so verringert sich wegen der Diskontierung der Wert der abschließenden Zahlung. Wir erhalten für  $m \in [0, \infty)$  und  $t \in \mathbb{N}$  das Stoppproblem

$$(1.12) \quad V_k(t, m) \triangleq \operatorname{ess\,sup}_{\tau \in \Lambda(t)} \mathbb{E} \left[ \sum_{u=t+1}^{\tau} \alpha^{u-t} Z_k(u) + \alpha^{\tau-t} m \mid \mathcal{F}^k(t) \right]$$

$V_k$  wird als Wertfunktion bezeichnet und gibt den maximalen erwarteten Gewinn an. Die Snellsche Enveloppe zu  $(X_k(t, m))_{t \in \mathbb{N}}$  ist der Prozess

$$(1.13) \quad Y_k(t, m) = \alpha^t V_k(t, m) + \sum_{u=1}^t \alpha^u Z_k(u).$$

Die Theorie des Optimalen Stoppens liefert die nun folgenden Ergebnisse. Die Stoppzeit

$$(1.14) \quad \sigma_k(t, m) = \inf(u \geq t : V_k(u, m) = m)$$

löst das Problem (1.12). Die Folge

$$(1.15) \quad (Y_k(u \wedge \sigma_k(t, m)))_{u=t}^\infty$$

ist ein Martingal und es gilt die Gleichung der Dynamischen Programmierung

$$(1.16) \quad V_k(t, m) = \max\{m, \alpha \mathbb{E}[Z_k(t+1) + V_k(t+1, m) \mid \mathcal{F}^k(t)]\}.$$

Die Abhängigkeit der Wertfunktion  $V_k(t, m)$  und der optimalen Stoppzeit  $\sigma_k(t, m)$  von  $m$  wollen wir nun untersuchen.

**Beispiel 1.5.** Wir wollen die Wertfunktion  $V_k(t, m)$  eines Armes mit Bernoulli-Auszahlungsprozess berechnen. Dazu betrachten wir das Problem bis zum Zeitpunkt  $\Xi > t$ . Wir setzen  $V_k^\Xi(\Xi, m) \triangleq m$ . Nach den Überlegungen des letzten Abschnitts konvergiert  $V^\Xi(t, m)$  für  $\Xi \rightarrow \infty$  gegen  $V(t, m)$ .

Man erhält mit (1.13) für  $t < \Xi$  rekursiv

$$V_k^\Xi(t, m) \triangleq \max[m; \alpha \mathbb{E}[Z_k(t+1) \mid \mathcal{F}^k(t)] + \alpha \mathbb{E}[V_k^\Xi(t+1, m) \mid \mathcal{F}^k(t)]]$$

Für  $t = 0$ ,  $\alpha = 0.8$  und  $m = 2.5$  ergeben sich die folgenden Approximationen für  $V_k(t, m)$ :

$$\begin{aligned} V_k^{10}(t, m) &= 2.50910348531, \\ V_k^{100}(t, m) &= 2.53035310302, \\ V_k^{1000}(t, m) &= 2.53035310306. \end{aligned}$$

**Beispiel 1.6.** Eine bessere Approximation liefert die folgende Überlegung. Statt in  $\Xi$  stets die deterministische Zahlung  $m$  zu erhalten, entscheidet man sich zwischen  $m$  oder der Möglichkeit unbegrenzt weiterzuspielen. Das letztere liefert die erwartete Auszahlung  $\frac{\alpha}{1-\alpha} \mathbb{E}[Z_k(T+1) \mid \mathcal{F}^k(\Xi)]$ . Man erhält dann

$$V_k^\Xi(\Xi, m) \triangleq \max[m; \frac{\alpha}{1-\alpha} \mathbb{E}[Z_k(\Xi+1) \mid \mathcal{F}^k(\Xi)]],$$

$V_k^\Xi(t, m)$  wird für  $t < \Xi$  ebenso wie oben rekursiv berechnet und konvergiert für  $\Xi \rightarrow \infty$  monoton. Diese Sichtweise entspricht der Betrachtung von Bhulai und Koole [3], welche im Falle eines mehrarmigen Banditen mit Bernoulli-verteilten Auszahlungsprozessen die folgenden beiden Situationen vergleichen. In der ersten nutzt der Spieler für Entscheidungen nur die ersten  $\Xi$  Beobachtungen jedes Arms. In der anderen erhält er nach jeweils  $\Xi$  Beobachtungen die vollständige Information über den Parameter des Armes. Bhulai und

Koole zeigen, dass die Differenz der Wertfunktionen in diesen unterschiedlichen Situationen mit  $\Xi \rightarrow \infty$  gegen 0 geht. Da die Wertfunktion im Standardfall zwischen diesen beiden Wertfunktionen liegt, erhält man  $V^\Xi(t, m) \rightarrow V(t, m)$ .

Unsere Situation ist vergleichbar mit einem zweiarmligen Banditen, dessen zweiter Arm deterministisch und konstant ist. Es ist klar, dass das Resultat von Bhulai und Koole in diesem Fall ebenfalls gilt und damit konvergiert  $V_k^\Xi(t, m)$  mit  $\Xi \rightarrow \infty$  tatsächlich gegen  $V_k(t, m)$ .

Für gleichen Werte der Parameter wie oben,  $t = 0$ ,  $\alpha = 0.8$  und  $m = 2.5$  ergibt sich

$$\begin{aligned} V_k^{10}(t, m) &= 2.53022260033, \\ V_k^{100}(t, m) &= 2.53035310306, \\ V_k^{1000}(t, m) &= 2.53035310306. \end{aligned}$$

Es ist zu erwarten, dass man früher stoppt wenn die zum Stoppzeitpunkt erfolgende Zahlung  $m$  steigt. Ist  $m$  hinreichend groß, so wird man sofort stoppen. Dies sagt die folgende

**Proposition 1.7.** *Die Abbildung  $m \mapsto \sigma_k(t, m)$  ist  $\mathbb{P}$ -fast sicher fallend und rechtsstetig, für  $m \rightarrow \infty$  gilt  $\sigma_k(t, m) \rightarrow t$ .*

*Beweis in Abschnitt 1.10.*

Wie reagiert die Wertfunktion  $V_k(t, m)$  auf Veränderungen von  $m$ ? Wegen der Rechtsstetigkeit von  $\sigma_k(t, m)$  ändert sich die optimale Stoppzeit nicht, wenn sich  $m$  nur minimal erhöht. Der Wert der Auszahlung erhöht sich also um denselben Betrag, allerdings diskontiert vom erwarteten Zeitpunkt dieser Auszahlung, der in der Zukunft liegen kann. Dieser erwartete Diskontfaktor ist damit die rechtsseitige Ableitung der Wertfunktion.

**Proposition 1.8.** *Die Abbildung  $m \mapsto V_k(t, m)$  ist konvex, wachsend und mit rechtsseitiger Ableitung*

$$(1.17) \quad \frac{\partial^+}{\partial m} V_k(t, m) = \mathbb{E} [\alpha^{\sigma_k(t, m) - t} | \mathcal{F}^k(t)].$$

*Inbesondere ist  $\lim_{m \rightarrow \infty} \frac{\partial^+}{\partial m} V_k(t, m) = 1$  und  $\lim_{m \rightarrow \infty} V_k(t, m) - m = 0$ .*

*Beweis in Abschnitt 1.10.*

Die Abhängigkeit der Wertfunktion  $V_k(0, m)$  eines Bernoulli-Armes von der Endzahlung  $m$  zeigt Abbildung 1.1 auf Seite 25.

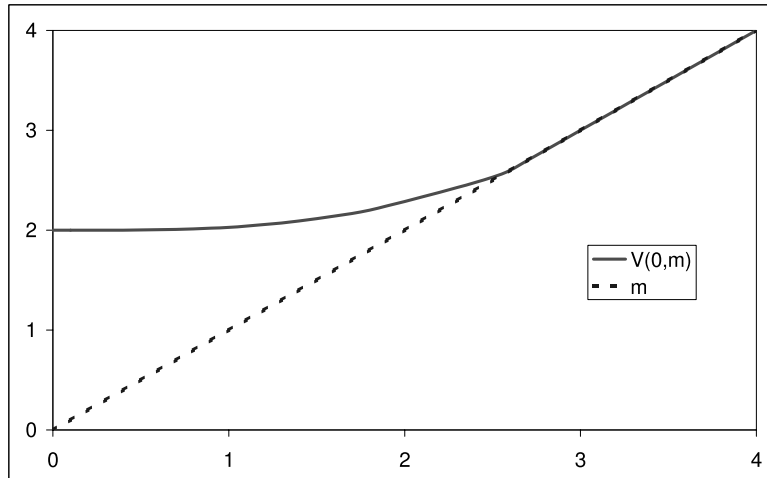


Abbildung 1.1: Wertfunktion  $m \mapsto V(0, m)$  für einen Bernoulli-Arm und  $\alpha = 0.8$

## 1.6 Der Gittins-Index

Wir werden nun den im letzten Abschnitt erwähnten Schwellenwert definieren und dessen Eigenschaften untersuchen.

**Definition 1.9.** Die  $\mathcal{F}^k(t)$ -messbare nichtnegative Zufallsvariable

$$(1.18) \quad M_k(t) \triangleq \text{ess inf}\{X \text{ } \mathcal{F}^k(t) \text{- messbar} : V_k(t, X) = X \text{ } \mathbb{P} \text{- f.s.}\}$$

heißt Gittins-Index zum Zeitpunkt  $t$ .

Diese Definition des Gittins-Index als Schwellenwert stammt von Whittle [29]. Jedoch bezeichnen Whittle ebenso wie Mandelbaum [20] und Weber [28] den Wert  $\frac{1-\alpha}{\alpha} M_k(t)$  als Gittins-Index, eine naheliegende Darstellung, wenn man den Index mittels der Gleichung (1.25) definiert. Wir folgen hier der Notation von El Karoui und Karatzas [10].

Da in unserer Beschreibung des mehrarmigen Banditen die auf eine Betätigung eines Armes folgende Auszahlung *erst im nächsten Zeitpunkt* erfolgt, unterscheidet sich der in (1.18) definierte Gittins-Index von der von El Karoui und Karatzas definierten Größe um den Faktor  $\alpha$ .

**Beispiel 1.10.** Im Fall eines Bernoulli-verteilten Renditeprozesses erhält man mit den Parameterwerten für  $\alpha = 0.8$  den folgenden Gittins-Index im Zeitpunkt 0:

$$M_k(0) = \inf\{m : V_k(0, m) = m\} = 2.565261.$$

$M_k(t)$  ist zufällig, hängt allerdings in diesem Beispiel nur von der Anzahl der Erfolge  $a_t$  und Misserfolge  $b_t$  bis  $t$  ab, da die Beobachtungen bedingt auf den Parameter unabhängig und identisch verteilt sind.

$b_t \setminus a_t$	0	1	2	3	4	5	6
0	2,5653	3,0385	3,2628	3,3968	3,4871	3,5526	3,6024
1	1,7719	2,3591	2,6859	2,8992	3,0492	3,1618	3,2493
2	1,3280	1,9046	2,2636	2,5122	2,6942	2,8362	2,9486
3	1,0516	1,5908	1,9466	2,2083	2,4082	2,5657	2,6932
4	0,8653	1,3591	1,7068	1,9656	2,1723	2,3394	2,4769
5	0,7320	1,1852	1,5160	1,7719	1,9759	2,1469	2,2903
6	0,6325	1,0482	1,3610	1,6105	1,8136	1,9821	2,1280

Tabelle 1.1: Gittins-Indizes eines Bernoulli-Armes nach verschiedenen Anzahlen von Erfolgen  $a_t$  und Misserfolgen  $b_t$  und  $\alpha = 0.8$ .

*Tabelle 1.1 gibt den Wert des Gittins-Index für verschiedene Werte von  $a_t$  und  $b_t$  an. Entsprechende Tafeln - mit der oben erwähnten abweichenden Skalierung - findet man auch im Anhang von Gittins [14].*

Wie bisher gilt für  $\mathcal{F}^k(t)$ -messbare Zufallsvariablen  $X$  die Gleichung

$$(1.19) \quad V_k(t, X) = \operatorname{ess\,sup}_{\tau \in \Lambda_t} \mathbb{E} \left[ \sum_{u=t+1}^{\tau} \alpha^{u-t} Z_k(u) + \alpha^{\tau-t} X \mid \mathcal{F}^k(t) \right]$$

mit optimaler Stopzeit

$$(1.20) \quad \sigma_k(t, X) \triangleq \inf \{ u \geq t : V_k(u, X) = X \}.$$

Weiter gelten ebenfalls für  $\mathcal{F}^k(t)$ -messbare Zufallsvariablen  $X_1, X_2$  mit  $X_1 \geq X_2$  die Ungleichungen

$$(1.21) \quad V_k(t, X_1) \geq V_k(t, X_2) \text{ und}$$

$$(1.22) \quad V_k(t, X_1) - X_1 \leq V_k(t, X_2) - X_2.$$

Im Stoppproblem muss man sich zwischen sofortigem Erhalt von  $X$  und Weiterspielen mit späterem Erhalt von  $X$  entscheiden. Die Wertfunktion ist in einer solchen Situation beschränkt durch die erwartete Rendite bei unbegrenztem Weiterspiel *und* sofortiger Auszahlung  $X$ :

$$X \leq V_k(t, X) \leq X + \mathbb{E} \left[ \sum_{u=t+1}^{\infty} \alpha^{u-t} Z_k(u) \mid \mathcal{F}^k(t) \right].$$

Für  $\tilde{X}$  hinreichend groß ist  $V(t, \tilde{X}) = \tilde{X}$ . Ein solches  $\tilde{X}$  findet man beispielsweise durch die Überlegung im Beweis von Proposition 1.7:

$$\tilde{X} \triangleq \frac{1}{1-\alpha} \mathbb{E} \left[ \sum_{u=t+1}^{\infty} \alpha^{u-t} Z_k(u) \mid \mathcal{F}^k(t) \right].$$

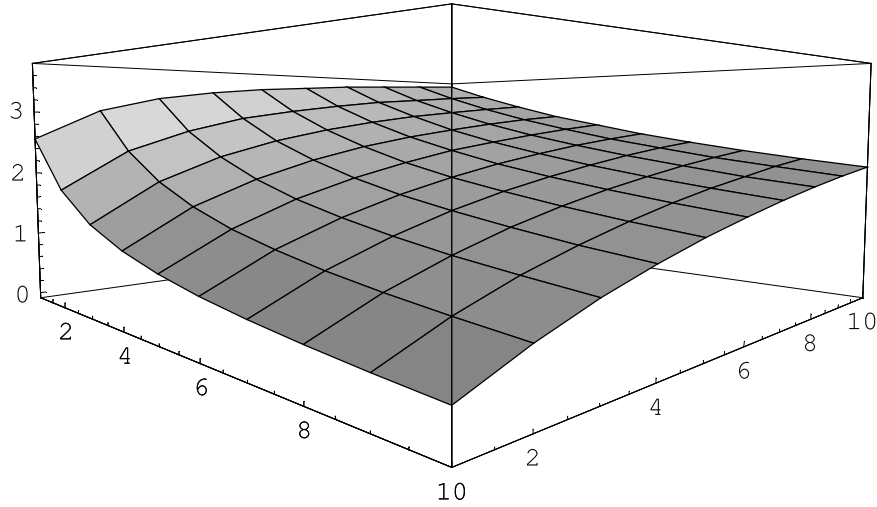


Abbildung 1.2: Darstellung der Gittins-Indizes aus Tabelle 1.1

Damit ist insbesondere die den Gittins-Index  $M$  definierende Menge

$$A_k^+ \triangleq \{X \mathcal{F}^k(t) - \text{messbar} : V_k(t, X) = X \mathbb{P} - f.s.\}$$

nicht leer. Schließlich gilt wegen der Stetigkeit von  $m \mapsto V_k(t, m)$

$$(1.23) \quad V_k(t, M(t)) = M(t).$$

Das folgende Lemma liefert eine zweite Darstellung von  $M_k(t)$  als Schwellenwert.

**Lemma 1.11.** *Es gilt die Darstellung*

$$(1.24) \quad M_k(t) = \text{ess sup} \{X \mathcal{F}^k(t) - \text{messbar} : V_k(t, X) > X \mathbb{P} - f.s.\}.$$

*Beweis in Abschnitt 1.10.*

Die folgende, als *Forward Induction* bezeichnete Charakterisierung (1.25) des Gittins-Index entspricht der ursprünglichen Definition von Gittins und Jones [13]. Diese Darstellung ergibt sich, wenn man statt der Einführung einer deterministischen Zahlung  $m$  das Allokationsproblem um einen zusätzlichen deterministischen Arm erweitert, der in jedem Zeitpunkt eine konstante Auszahlung  $m$  liefert.

**Proposition 1.12.** *Der Gittins-Index  $M_k(t)$  hat die Darstellung*

$$(1.25) \quad \frac{1 - \alpha}{\alpha} M_k(t) = \text{ess sup}_{\tau \geq t+1} \frac{\mathbb{E}[\sum_{u=t+1}^{\tau} \alpha^u Z_k(u) \mid \mathcal{F}^k(t)]}{\mathbb{E}[\sum_{u=t+1}^{\tau} \alpha^u \mid \mathcal{F}^k(t)]}.$$

*Er lässt sich damit als maximal erzielbarer erwarteter Ertrag pro erwarteter Zeiteinheit bei vorläufiger Fortsetzung des Spiels interpretieren.*

Beweis in Abschnitt 1.10.

**Bemerkung 1.13.** Die folgende Stoppzeit realisiert das Supremum in (1.25):

$$\begin{aligned}\sigma_k(t+1, M_k(t)) &\triangleq \min\{s \geq t+1 : V(s, M_k(t)) = M_k(t)\} \\ &= \min\{s \geq t+1 : M_k(s) \leq M_k(t)\}\end{aligned}$$

Die zweite Gleichung liefert eine Verbindung zu Exkursionen des Prozesses  $M_k$  von seinem laufenden Minimum. Diese spielen eine wichtige Rolle in unserer Untersuchung des zeitstetigen Falls.

Die Stoppzeit  $\sigma_k(t+1, M_k(t))$  ist im Allgemeinen nicht leicht zu berechnen. Durch das Einsetzen geeigneter Stoppzeiten kann man allerdings untere Schranken für den Gittins-Index erhalten.

Berry und Fristedt [2] (Theorem 5.4.1) tun dies in der Situation des Bernoulli-verteilter Auszahlungsprozesse. Die Stoppzeit  $\xi \triangleq \inf\{t \geq 1 : Z(t) = 0\}$  sei der Zeitpunkt des ersten Misserfolgs. Zu  $r \in \mathbb{N} \cup \{\infty\}$  setzt man

$$(1.26) \quad \tau_r \triangleq \begin{cases} \xi & a_\xi < r \\ \infty & a_\xi \geq r. \end{cases}$$

Im Zeitpunkt des ersten Misserfolgs stoppt man also, falls davor weniger als  $r$  Erfolge beobachtet wurden, sonst stoppt man nicht. Damit erhält man für den Gittins-Index die folgenden Abschätzung

$$(1.27) \quad M(t) \geq \frac{\alpha}{1-\alpha} \frac{\sum_{u=1}^{\infty} \alpha^u \mathbb{E}[\Theta^{u \wedge (r+1)} \mid \mathcal{F}(t)]}{\sum_{u=1}^{\infty} \alpha^u \mathbb{E}[\Theta^{(u-1) \wedge r} \mid \mathcal{F}(t)]} =: K_r.$$

Für  $t = 0$  gilt  $\Theta \sim \mathcal{U}[0, 1]$  und man erhält

$$(1.28) \quad \begin{aligned}K_r &= \frac{\alpha}{1-\alpha} \frac{\sum_{u=1}^r \frac{\alpha^{u-1}}{u+1} + \frac{\alpha^r}{(r+2)(1-\alpha)}}{\sum_{u=1}^r \frac{\alpha^{u-1}}{u} + \frac{\alpha^r}{(r+1)(1-\alpha)}}, \\ K_\infty &= \frac{\alpha}{1-\alpha} \left( \frac{1}{\alpha} + \frac{1}{\log(1-\alpha)} \right).\end{aligned}$$

Numerische Auswertungen dieser Schranken sind in Tabelle 1.2 dargestellt.

Wir kommen nun zurück zum Stoppproblem (1.12). Zu der Folge der Gittins-Indizes  $(M_k(t))_{t \geq 0}$  definieren wir deren laufende Minima durch

$$\begin{aligned}\underline{M}_k(t, s) &\triangleq \min_{t \leq u \leq s} M_k(u), \\ \underline{M}_k(s) &\triangleq \underline{M}_k(0, s).\end{aligned}$$

$\alpha$	$K_\infty$	$M(0)$	$\alpha$	$K_\infty$	$M(0)$
0,1	0,0565	0,0565	0,6	0,8630	0,8681
0,2	0,1296	0,1297	0,7	1,3953	1,4107
0,3	0,2270	0,2271	0,8	2,5147	2,5653
0,4	0,3616	0,3621	0,9	6,0913	6,3260
0,5	0,5573	0,5590	0,95	13,6576	14,4672

Tabelle 1.2: Der Gittins-Index  $M(0)$  und die untere Schranke  $K_\infty$  aus (1.28) für verschiedene Diskontfaktoren

Die optimale Stoppzeit  $\sigma_k(t, m)$  hängt eng mit dem Prozess  $\underline{M}_k$  zusammen. Es sei  $s \geq t$ . Falls gilt  $\sigma_k(t, m) > s$ , so ist offenbar in jedem Zeitpunkt zwischen  $t$  und  $s$  Weiterspielen der Zahlung von  $m$  vorzuziehen, insbesondere also  $V_k(u, m) > m \forall t \leq u \leq s$ . Wenn wiederum diese Bedingung erfüllt ist, so ist  $m < \underline{M}_k(u) \forall t \leq u \leq s$ . Die Umkehrung gilt ebenfalls, man erhält die folgende Äquivalenz:

$$(1.29) \quad \begin{aligned} \sigma_k(t, m) > s &\Leftrightarrow V_k(u, m) > m \quad \forall t \leq u \leq s \\ &\Leftrightarrow m < \underline{M}_k(t, s). \end{aligned}$$

Daraus folgt die folgende Darstellung der optimalen Stoppzeit  $\sigma_k(t, m)$ :

$$(1.30) \quad \begin{aligned} \sigma_k(t, m) &= \inf\{s \geq t : \underline{M}_k(t, s) \leq m\} \\ &= \inf\{s \geq t : M_k(s) \leq m\}, \end{aligned}$$

sowie die Beziehung

$$\begin{aligned} \underline{M}_k(t, s) &= \inf\{m \geq 0 : \sigma_k(t, m) \leq s\} \\ &= \sup\{m \geq 0 : \sigma_k(t, m) > s\}. \end{aligned}$$

Diese Zusammenhänge von  $\sigma_k$  und  $\underline{M}_k$  liefern die folgenden Darstellungen der erwarteten Gewinne und der Wertfunktion in Abhängigkeit der Gittins-Indizes.

**Proposition 1.14.** *Es gelten  $\mathbb{P}$ -f.s. die Beziehungen*

$$(1.31) \quad \mathbb{E}\left[\sum_{u=\sigma(t,m)+1}^{\infty} \alpha^u Z(u) \mid \mathcal{G}(t)\right] = (1-\alpha)\mathbb{E}\left[\sum_{u=\sigma(t,m)}^{\infty} \alpha^u \underline{M}(t, u) \mid \mathcal{G}(t)\right],$$

$$(1.32) \quad V(t, 0) = \mathbb{E}\left[\sum_{u=t+1}^{\infty} \alpha^{u-t} Z(u) \mid \mathcal{G}(t)\right] = (1-\alpha)\mathbb{E}\left[\sum_{u=t}^{\infty} \alpha^{u-t} \underline{M}(t, u) \mid \mathcal{G}(t)\right],$$

$$(1.33) \quad V(t, m) = (1-\alpha)\mathbb{E}\left[\sum_{u=t}^{\infty} \alpha^{u-t} (m \vee \underline{M}(t, u)) \mid \mathcal{G}(t)\right].$$

*Beweis in Abschnitt 1.10.*

**Bemerkung 1.15.** Die Gleichung (1.32) stellt die erwartete Auszahlung eines Armes mit Renditeprozess  $Z$  in Abhängigkeit der unteren Einhüllenden  $\underline{M}$  des Gittins-Indexprozess des Armes dar. Diese Identität spielt beispielsweise in ([17]) eine wichtige Rolle. Eine hilfreiche Interpretation dieser Gleichung liefert Weber [28].

Der Auszahlungsprozess wird als einarmiger Bandit betrachtet, dessen Aktivierung jeweils einen gewissen Betrag kostet. Im Zeitpunkt 0 wird dieser als fairer Preis festgelegt und entspricht gerade dem Gittins-Index. Dieser Preis bleibt unverändert so lange ein rationaler Spieler unter diesen Bedingungen zu spielen bereit ist. Genau dann ist dies nicht mehr der Fall, wenn der Gittins-Indexprozess ein neues Minimum erreicht. Da der Preis fair war, ist der erwartete Gewinn des Spiels bis zu diesem zufälligen Zeitpunkt 0. Um ein Weiterspielen zu ermöglichen wird nun ein neuer Preis festgesetzt, man wählt wieder den aktuellen Gittins-Index. So wird ein faires Spiel konstruiert, die erwarteten Einnahmen und Ausgaben sind gleich und das bedeutet (1.32).

Für unsere Betrachtung der zeitstetigen Situation ist eine Variante dieser Gleichung ebenfalls von großer Bedeutung.

## 1.7 Dynamische Programmierung

Nach der Betrachtung der einzelnen Arme und ihrer Gittins-Indizes kommen wir nun zurück zum mehrdimensionalen Allokationsproblem. Dieses wollen wir, weiter El Karoui und Karatzas [10] folgend, mit Hilfe der untersuchten Stoppprobleme und Dynamischer Programmierung lösen. In diesem Abschnitt führen wir zunächst einige dazu notwendige Begriffe und Notationen ein. Näheres zum Problem des Optimalen Stoppens für mehrparametrische Prozesse findet man beispielsweise bei Mandelbaum und Vanderbei [18].

Zu  $\tilde{s} \in \mathbb{N}^d$  bezeichnen wir die Menge der Strategien  $T$ , für die gilt  $T(0) = \tilde{s}$ , mit  $\mathcal{S}(\tilde{s})$ . Der Gewinn unter einer Strategie  $T$  war von der Form

$$\mathcal{R}(T) \triangleq \sum_{t=1}^{\infty} \alpha^t \sum_{k=1}^d Z_k(T_k(t)) \cdot (T_k(t) - T_k(t-1)).$$

Die Abbildung

$$(1.34) \quad \Phi(\tilde{s}) \triangleq \operatorname{ess\,sup}_{T \in \mathcal{S}(\tilde{s})} \mathbb{E}[\mathcal{R}(T) \mid \mathcal{F}(\tilde{s})]$$

heißt *Wertfunktion* im Punkt  $\tilde{s}$ . In  $\tilde{s} = (s_1, \dots, s_d)$  wurde für  $k = 1, \dots, d$  der Arm  $k$  genau  $s_k$  mal betätigt.  $\Phi$  gibt den maximal erreichbaren erwarteten Gewinn nach  $\tilde{s}$  an und erfüllt die Bellman-Gleichung der Dynamischen Programmierung

$$(1.35) \quad \Phi(\tilde{s}) = \max_{1 \leq j \leq d} [\alpha \mathbb{E}[Z_j(s_j + 1) \mid \mathcal{F}(\tilde{s})] + \alpha \mathbb{E}[\Phi(\tilde{s} + \tilde{e}_j) \mid \mathcal{F}(\tilde{s})]],$$

$\tilde{e}_j$  sei wieder der  $j$ . Einheitsvektor in  $\mathbb{N}_0^d$ .

Wir erweitern das obige Problem um die Möglichkeit, das Spiel zu beenden und dafür eine Zahlung  $m \geq 0$  zu erhalten. Dazu benötigen wir den Begriff einer Stoppzeit bezüglich der mehrparametrischen Filtration  $\mathcal{F}$  mit Werten in  $\mathbb{N}_0^d$ .

Analog zum eindimensionalen Fall heißt eine messbare Abbildung

$$\tilde{\nu} : \Omega \rightarrow \mathbb{N}_0^d$$

*Stoppzeitpunkt* bezüglich der Filtration  $\mathcal{F}$ , wenn  $\{\tilde{\nu} = \tilde{s}\} \in \mathcal{F}(\tilde{s})$  für alle  $\tilde{s} \in \mathbb{N}_0^d$ . Die  $\sigma$ -Algebra der  $\tilde{\nu}$ -Vergangenheit definiert man wie im eindimensionalen Fall:

$$\mathcal{F}(\tilde{\nu}) \triangleq \{A \in \mathcal{F} : A \cap \{\tilde{\nu} = \tilde{s}\} \in \mathcal{F}(\tilde{s}) \forall \tilde{s} \in \mathbb{N}_0^d\}.$$

Gemäß der Definition einer Strategie (1.2) gilt für eine Strategie  $T$  und  $t \in \mathbb{N}_0, \tilde{s} \in \mathbb{N}_0^d$

$$\{T(t) = \tilde{s}\} = \bigcup_{i=1}^d \{T(t+1) - T(t) = \tilde{e}_i, T(t) = \tilde{s}\} \in \mathcal{F}(\tilde{s}).$$

Zu einer Strategie  $T$  ist insbesondere die Filtration  $(\mathcal{F}(T(t)))_{t \geq 0}$  wohldefiniert. Diese beschreibt die unter der Strategie  $T$  sukzessive verfügbaren Informationen.

Ein Paar  $\Pi = (T, \tau)$ , bestehend aus einer Strategie  $T \in \mathcal{S}(\tilde{s})$  und einer  $\mathbb{N}_0$ -wertigen  $\mathcal{F}(T(\cdot))$ -Stoppzeit  $\tau$  nennen wir *erweiterte Strategie*, die Menge solcher Paare bezeichnen wir mit  $\mathcal{P}(\tilde{s})$ .

Analog zu (1.3) und (1.12) definieren wir für  $m \geq 0$  den Gewinn unter der erweiterten Strategie  $\Pi$

$$\mathcal{R}(\Pi, m) \triangleq \sum_{t=1}^{\tau} \alpha^t \sum_{k=1}^d Z_k(T_k(t)) \cdot (T_k(t) - T_k(t-1)) + \alpha^{\tau} m$$

sowie die Wertfunktion im Zeitpunkt  $\tilde{s} \in \mathbb{N}_0^d$

$$(1.36) \quad \Phi(\tilde{s}, m) \triangleq \text{ess sup}_{\Pi \in \mathcal{P}(\tilde{s})} \mathbb{E}[\mathcal{R}(\Pi, m) \mid \mathcal{F}(\tilde{s})].$$

Wieder gilt die Bellman-Gleichung, die nun die folgende Form annimmt

$$(1.37) \quad \Phi(\tilde{s}, m) = \max \left[ m, \max_{1 \leq j \leq d} [\alpha \mathbb{E}[Z_j(s_j + 1) \mid \mathcal{F}(\tilde{s})] + \alpha \mathbb{E}[\Phi(\tilde{s} + \tilde{e}_j, m) \mid \mathcal{F}(\tilde{s})]] \right]$$

Neben einer Strategie muss man nun zusätzlich eine optimale Stoppzeit wählen. Die Erweiterung ermöglicht jedoch die Nutzung der betrachteten eindimensionalen Stoppprobleme zur Lösung des mehrdimensionalen Allokationsproblems. Dies ist Thema des nächsten Abschnitts.

## 1.8 Der Whittle-Ansatz und das Gittins-Indextheorem

In diesem Abschnitt beweisen wir das zentrale Gittins-Indextheorem, welches die Klasse der optimalen Strategien charakterisiert. Dies geschieht mit Hilfe der von Whittle [29] stammenden multiplikativen Zerlegung der Wertfunktion. Wir folgen dabei weiter El Karoui und Karatzas [10].

In dem um die Zahlung  $m$  erweiterten Allokationsproblem ist ein Weiterspielen so lange optimal, wie es mindestens einen Arm gibt, dessen Betätigung vorteilhafter als die Zahlung  $m$  erscheint. Dies ist genau dann der Fall, wenn der Gittins-Index eines Armes größer oder gleich  $m$  ist. Man betätigt also ausschließlich Arme, deren Gittins-Indizes hinreichend groß sind und spielt, bis die Indizes aller Arme  $m$  erreicht oder unterschritten haben. Vom Zeitpunkt  $\tilde{s} \in \mathbb{N}_0^d$  betätigt man dazu Arm  $k$  jeweils  $\sigma_k(s_k, m)$  mal bevor man stoppt. Die Stoppzeit  $\tau$  einer optimalen erweiterten Strategie  $\Pi = (T, \tau)$  sollte also die Gestalt  $\tau = \sum_{k=1}^d \sigma_k(s_k, m)$  besitzen.

Im eindimensionalen Stoppproblem mit Endauszahlung  $m$  ist mit (1.17) die rechtsseitige Ableitung der Wertfunktion  $m \mapsto V_k(t, m)$  der erwartete Diskontfaktor des optimalen Stoppzeitpunkts:

$$\frac{\partial^+}{\partial m} V_k(t, m) = \mathbb{E}[\alpha^{\sigma_k(t, m) - t} \mid \mathcal{F}^k(t)].$$

Nehmen wir nun an, dass die Wertfunktion des erweiterten Allokationsproblems  $m \mapsto \Phi(\tilde{s}, m)$  ebenfalls rechtsseitig differenzierbar und die Ableitung wieder der erwartete Diskontfaktor des optimalen Stoppzeitpunkts ist. Wegen der Unabhängigkeit der Filtrationen könnten wir dann annehmen, dass gilt

$$\begin{aligned} \frac{\partial^+}{\partial m} \Phi(\tilde{s}, m) &= \mathbb{E}[\alpha^\tau \mid \mathcal{F}(\tilde{s})] \\ &= \mathbb{E}[\alpha^{\sum_{k=1}^d \sigma_k(s_k, m)} \mid \mathcal{F}(\tilde{s})] \\ &= \prod_{k=1}^d \mathbb{E}[\alpha^{\sigma_k(s_k, m)} \mid \mathcal{F}(\tilde{s})] \\ &= \prod_{k=1}^d \frac{\partial^+}{\partial m} V_k(s_k, m). \end{aligned}$$

Damit würde folgen, dass

$$(1.38) \quad \Phi(\tilde{s}, m) = \begin{cases} \Phi(\tilde{s}) & m = 0 \\ K - \int_m^K \prod_{i=1}^d \frac{\partial^+}{\partial x} V_i(s_i, x) dx & 0 < m < K \\ m & m \geq K. \end{cases}$$

Diesen Ansatz werden wir indirekt verifizieren, weiter ([10]) folgend. Wir setzen

$$(1.39) \quad F(\tilde{s}, m) \triangleq \begin{cases} K - \int_m^K \prod_{i=1}^d \frac{\partial^+}{\partial x} V_i(s_i, x) dx & 0 \leq m < K \\ m & m \geq K. \end{cases}$$

Die Form von  $F$  ermöglicht den Beweis des folgenden Theorems. Die Gleichung (1.40) liefert eine Verbindung der Bellman-Gleichung mit dem Gittins-Index, die es erlaubt, die Optimalität von Indexstrategien zu zeigen.

**Theorem 1.16.** *Die Funktion  $F(\tilde{s}, m)$  erfüllt die Gleichung der Dynamischen Programmierung (1.37) des erweiterten  $d$ -armigen Banditen. Insbesondere gilt*

$$(1.40) \quad F(\tilde{s}, m) = \alpha \mathbb{E}[Z_i(s_i + 1) \mid \mathcal{F}(\tilde{s})] + \alpha \mathbb{E}[F(\tilde{s} + \tilde{e}_i, m) \mid \mathcal{F}(\tilde{s})]$$

auf der Menge  $\{m < M_i(s_i) = \max_{1 \leq j \leq d} M_j(s_j)\}$ .

*Beweis in Abschnitt 1.10.*

Um für die Abbildung  $F : \mathbb{N}_0^d \rightarrow \mathbb{R}$  mit

$$F(\tilde{s}) \triangleq F(\tilde{s}, 0) = K - \int_0^K \prod_{i=1}^d \frac{\partial^+}{\partial m} V_i(s_i, m) dm$$

die gewünschte Identität  $F = \Phi$  zu erhalten, betrachtet man den für  $\tilde{s} \in \mathbb{N}^d$  und  $T \in \mathcal{S}(\tilde{s})$  definierten Prozess  $Q_T$  mit

$$Q_T(t) \triangleq \alpha^t F(T(t)) + \sum_{u=1}^t \alpha^u \sum_{k=1}^d Z_k(T_k(u))(T_k(u) - T_k(u-1)).$$

Mit Hilfe der Darstellung (1.40) erhält man nun das folgende Resultat.

**Korollar 1.17.**  *$Q_T$  ist ein beschränktes Supermartingal bezüglich der Filtration  $(\mathcal{F}(T(t)))$ .  $Q_{\hat{T}}$  ist ein Martingal genau dann, wenn  $\hat{T}$  eine Indexstrategie ist.*

*Beweis in Abschnitt 1.10.*

$Q_T$  konvergiert für jede Strategie  $T$  gegen den von  $T$  erzielten Gewinn  $\mathcal{R}[T]$ . Aus der Supermartingaleigenschaft folgt für beliebige Strategien  $T \in \mathcal{S}(\tilde{s})$

$$F(\tilde{s}) = Q_T(0) \geq \mathbb{E}[\mathcal{R}(T) \mid \mathcal{F}(\tilde{s})].$$

Die Gleichheit gilt genau dann, wenn  $T$  eine Indexstrategie ist. Man erhält für Indexstrategien  $\hat{T}$  die folgenden Identitäten

$$\mathbb{E}[\mathcal{R}(\hat{T}) \mid \mathcal{F}(\tilde{s})] = F(\hat{T}(0)) = F(\tilde{s}) = \operatorname{ess\,sup}_{T \in \mathcal{S}(\tilde{s})} \mathbb{E}[\mathcal{R}(T) \mid \mathcal{F}(\tilde{s})] = \Phi(\tilde{s}).$$

$Q_T(t)$  ist damit tatsächlich der erwartete Wert des Gewinns unter der Strategie  $T$  zum Zeitpunkt  $t$  und es folgt das bekannte Indextheorem.

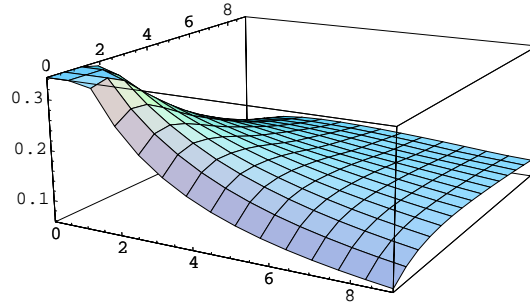


Abbildung 1.3: Entwicklung des Informationsteils  $L(a, b)$  des Gittins-Index für  $\alpha = 0.8$

**Theorem 1.18.** *Es existiert eine optimale Strategie für das Problem (1.4). Die Mengen der optimalen Strategien und der Indexstrategien sind gleich.*

## 1.9 Eine Zerlegung des Gittins-Index im Bernoulli-Banditen

Das Indextheorem rechtfertigt die einleitenden Überlegungen, nach denen es unter Umständen optimal ist, einen Arm mit geringerer erwarteter Auszahlung im nächsten Schritt zu aktivieren. Dies ist genau dann der Fall, wenn dessen Gittins-Index größer als die der anderen Arme ist.

Hat man bis zu einem Zeitpunkt  $a$  Erfolge und  $b$  Misserfolge eines Armes beobachtet, so ist der unbekannte Parameter  $\Theta$  des Armes unter dieser Beobachtung betaverteilt mit den Koeffizienten  $a + 1$  und  $b + 1$ . Die erwartete Auszahlung des Armes in der nächsten Runde unter dieser Beobachtung ist  $\mathbb{E}[\Theta \mid a, b] = \frac{a+1}{a+b+2}$ .

Unter vollständiger Kenntnis der Parameter ist es optimal, stets den Arm mit der größten erwarteten Auszahlung der nächsten Runde zu betätigen. Die Differenz zwischen dem entsprechend skalierten Gittins-Index  $\frac{1-\alpha}{\alpha}M(a, b)$  eines Armes und diesem Erwartungswert kann deshalb als *Maß für den Nutzen zusätzlicher Information* über den unbekannt Parameter  $\Theta$  angesehen werden. Den Wert

$$(1.41) \quad L(a, b) \triangleq \frac{1-\alpha}{\alpha}M(a, b) - \frac{a+1}{a+b+2}$$

nennen wir in Anlehnung an Gittins und Wang [15] *Informationsanteil des Gittins-Index*  $M(a, b)$ .

Hat man bis zum Zeitpunkt  $t$  genau  $a_i$  Erfolge und  $b_i$  Misserfolge zweier Arme  $i = 1, 2$ , beobachtet und gilt  $\frac{a_1+1}{a_1+b_1+2} = \frac{a_2+1}{a_2+b_2+2}$ , so sollte der Informationsanteil des Armes, der bis  $t$  seltener betätigt wurde, größer sein. Dies gilt tatsächlich nach dem folgenden Theorem von Gittins und Wang [15].

**Theorem 1.19.** *Es sei  $\frac{a_1+1}{a_1+b_1+2} \leq \frac{a_2+1}{a_2+b_2+2}$  und  $a_1 + b_1 \leq a_2 + b_2$ . Dann gilt*

$$(1.42) \quad M(a_1, b_1) \leq M(a_2, b_2).$$

*Die Ungleichung ist strikt, wenn dies für mindestens eine der beiden vorausgesetzten Ungleichungen gilt.*

Damit folgt, dass der Informationsanteil der Gittins-Indizes entlang von Pfaden  $(a, b)$ , auf denen der bedingte Erwartungswert  $\frac{a+1}{a+b+2}$  konstant ist, mit zunehmender Anzahl von Beobachtungen fällt. Tatsächlich gilt noch mehr, der Informationsanteil konvergiert mit steigender Zahl von Beobachtungen gleichmäßig gegen 0. Dies zeigt das folgende Lemma von Wang [27], das wir hier mit Hilfe eines Arguments von Berry und Fristedt [2] beweisen. Abbildung 1.3 illustriert diese Konvergenz.

**Lemma 1.20.** *Für den Informationsanteil  $L$  des Gittins-Index gilt*

$$(1.43) \quad L(a, b) \rightarrow 0 \text{ für } a + b \rightarrow \infty.$$

*Beweis.* Gilt die Aussage nicht, dann existieren  $\epsilon > 0$  und eine Folge  $(a_n, b_n)$  mit  $a_n + b_n \rightarrow \infty$ , so dass  $\frac{1-\alpha}{\alpha}M(a_n, b_n) - \frac{a_n+1}{a_n+b_n+2} > \epsilon$ . Wegen  $\frac{a_n+1}{a_n+b_n+2} \in [0, 1]$  existiert eine Teilfolge  $(a_k, b_k)$  mit  $\frac{a_k+1}{a_k+b_k+2} \rightarrow \lambda$  für ein  $\lambda \in [0, 1]$ .

Die Folge der Betaverteilungen mit den Parametern  $(a_k + 1, b_k + 1)$  konvergiert gegen das Diracmaß  $\delta_\lambda$ . Der Gittins-Index ist nach Korollar 5.3.3 von Berry und Fristedt [2] stetig bezüglich der zugrundeliegenden Verteilung und der Index eines Bernoulli-Armes mit bekanntem Parameter  $\lambda$  hat wegen der Darstellung (1.25) den Wert  $\frac{\alpha}{1-\alpha}\lambda$ . Damit gilt  $L(a_k, b_k) = \frac{1-\alpha}{\alpha}M(a_k, b_k) - \frac{a_k+1}{a_k+b_k+2} \rightarrow 0$  und die Behauptung ist bewiesen.  $\square$

Dieses Resultat stimmt mit der in Abschnitt 1.5 erwähnten Betrachtung von Bhulai und Koole [3] überein, ohne dass diese den Gittins-Index verwenden.

Nach Gittins und Wang [15] konvergiert  $L(a, b)$  entlang von Kurven  $(a, b)$  mit konstantem Index  $M(a, b)$  von der Ordnung  $O(\frac{1}{n})$ . Die in Abbildung dargestellten 1.4 numerischen Ergebnisse lassen vermuten, dass dies ebenfalls gleichmäßig in  $(a, b)$  gilt, das heißt  $L(a, b) = O(\frac{1}{a+b})$ .

## 1.10 Anhang

*Beweis von Proposition 1.7.* Die Abbildung

$$m \mapsto V_k(t, m) - m$$

ist  $\mathbb{P}$ -fast sicher fallend, also gilt für  $m_1 \leq m_2$

$$V_k(t, m_1) = m_1 \Rightarrow V_k(t, m_2) = m_2.$$

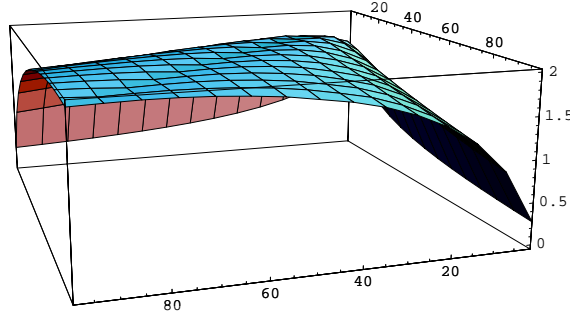


Abbildung 1.4: Asymptotik des Informationsanteils:  $(a + b) \cdot L(a, b)$ ,  $\alpha = 0.8$

und damit  $\sigma_k(t, m_2) \leq \sigma_k(t, m_1)$ .

Nehmen wir an,  $\sigma_k(t, m)$  sei nicht  $\mathbb{P}$ -fast sicher rechtsstetig in  $m$ . Da die Stoppzeit mit fallendem  $m$  monoton steigt, existieren  $n \in \mathbb{N}$  und  $\delta_m > 0$  mit  $\mathbb{P}[A] > 0$  für

$$A \triangleq \{\sigma(t, m + \delta) = n < \sigma(t, m) \forall \delta \leq \delta_m\}.$$

Auf  $A$  gilt  $V_k(\sigma(t, m + \delta), m) = V_k(n, m) > m \forall \delta \leq \delta^*$ . Also existiert ein  $\epsilon > 0$  mit

$$\mathbb{P}[A \cap \{V_k(\sigma_k(t, m + \delta), m) > m + \epsilon \forall \delta \leq \delta^*\}] > 0.$$

Damit ist auch

$$\mathbb{P}[A \cap \{\liminf_{\delta \downarrow 0} V_k(\sigma_k(t, m + \delta), m) \geq m + \epsilon\}] > 0.$$

Ohne Beschränkung der Allgemeinheit gelte  $\sigma_k(t, m + \delta) < \infty$  für  $\delta > 0$ . Dann gilt

$$V_k(\sigma_k(t, m + \delta), m) \leq V_k(\sigma_k(t, m + \delta), m + \delta) = m + \delta.$$

und damit  $\liminf_{\delta \downarrow 0} V_k(\sigma_k(t, m + \delta), m) \leq m$ . Dies liefert den Widerspruch und damit die Rechtsstetigkeit.

Ist nun  $m$  so groß, dass die erwartete Gesamtauszahlung des Armes kleiner als der bei einer Verzögerung der Auszahlung  $m$  durch die Diskontierung entstehende Verlust ist, so wird sofortiges Stoppen optimal. Es sei

$$\mathbb{E}\left[\sum_{u=t+1}^{\infty} \alpha^{u-t} Z_k(u) \mid \mathcal{F}^k(t)\right] < (1 - \alpha)m.$$

Nun ist nach (1.16)  $\sigma_k(t, m) = t$  äquivalent dazu, dass

$$m \geq \alpha \mathbb{E}[Z_k(t+1) + V_k(t+1, m) \mid \mathcal{F}^k(t)],$$

und dies gilt wegen

$$\alpha \mathbb{E}[Z_k(t+1) + V_k(t+1, m) \mid \mathcal{F}^k(t)] \leq \alpha \mathbb{E}\left[\sum_{u=t+1}^{\infty} \alpha^{u-t} Z_k(u) \mid \mathcal{F}^k(t)\right] + \alpha m < m.$$

Also konvergiert  $\sigma_k(t, m)$  monoton fallend gegen  $t$  für  $m \rightarrow \infty$ .  $\square$

*Beweis von Proposition 1.8.* Es ist klar, dass  $m \mapsto V_k(t, m)$  wachsend ist. Weiter existiert eine Folge von Stoppzeiten  $(\tau_n)$ , so dass

$$V_k(t, m) = \sup_{n \geq 0} \mathbb{E} \left[ \sum_{u=t+1}^{\tau_n} \alpha^{u-t} Z_k(u) + \alpha^{\tau_n-t} m \mid \mathcal{F}^k(t) \right].$$

Als Supremum über affin-lineare Abbildungen ist  $V(t, m)$  konvex.

Es sei  $\delta > 0$ . Wegen (1.15) und  $\sigma_k(t, m) \geq \sigma_k(t, m + \delta) \geq t$  folgt aus dem Stoppsatz

$$(1.44) \quad Y_k(t, m) = \mathbb{E}[Y_k(\sigma_k(t, m + \delta), m) \mid \mathcal{F}^k(t)].$$

Zusammen mit (1.13) erhält man

$$\begin{aligned} V_k(t, m) &= \alpha^{-t} Y_k(t, m) - \sum_{u=1}^t \alpha^{u-t} Z_k(u) \\ &= \alpha^{-t} \mathbb{E}[Y_k(\sigma_k(t, m + \delta), m) \mid \mathcal{F}^k(t)] - \sum_{u=1}^t \alpha^{u-t} Z_k(u) \\ &= \mathbb{E}[\alpha^{\sigma_k(t, m + \delta) - t} V_k(\sigma_k(t, m + \delta), m) + \sum_{u=t+1}^{\sigma_k(t, m + \delta)} \alpha^{u-t} Z_k(u) \mid \mathcal{F}^k(t)] \\ &= \mathbb{E}[\alpha^{\sigma_k(t, m + \delta) - t} V_k(\sigma_k(t, m + \delta), m) \\ &\quad + V_k(t, m + \delta) - \alpha^{\sigma_k(t, m + \delta) - t} (m + \delta) \mid \mathcal{F}^k(t)] \\ &\geq \mathbb{E}[\alpha^{\sigma_k(t, m + \delta) - t} m + V_k(t, m + \delta) - \alpha^{\sigma_k(t, m + \delta) - t} (m + \delta) \mid \mathcal{F}^k(t)] \\ &= \mathbb{E}[V_k(t, m + \delta) - \alpha^{\sigma_k(t, m + \delta) - t} \delta \mid \mathcal{F}^k(t)]. \end{aligned}$$

Wegen der Rechtsstetigkeit von  $\sigma_k(t, m)$  ergibt sich

$$(1.45) \quad \limsup_{\delta \downarrow 0} \frac{V_k(t, m + \delta) - V_k(t, m)}{\delta} \leq \limsup_{\delta \downarrow 0} \mathbb{E}[\alpha^{\sigma_k(t, m + \delta) - t} \mid \mathcal{F}^k(t)] = \mathbb{E}[\alpha^{\sigma(t, m) - t} \mid \mathcal{F}^k(t)]$$

Andererseits ist  $Y_k(t, m + \delta)$  ein Supermartingal, also gilt

$$Y_k(t, m + \delta) \geq \mathbb{E}[Y_k(\sigma_k(t, m), m + \delta) \mid \mathcal{F}^k(t)],$$

und damit

$$\begin{aligned} V_k(t, m + \delta) &= \alpha^{-t} Y_k(t, m + \delta) - \sum_{u=1}^t \alpha^{u-t} Z_k(u) \\ &\geq \alpha^{-t} \mathbb{E}[Y_k(\sigma_k(t, m), m + \delta) - \sum_{u=1}^t \alpha^u Z_k(u) \mid \mathcal{F}^k(t)] \\ &= \mathbb{E}[\alpha^{\sigma_k(t, m) - t} V_k(\sigma_k(t, m), m + \delta) + \sum_{u=t+1}^{\sigma_k(t, m)} \alpha^u Z_k(u) \mid \mathcal{F}^k(t)] \\ &\geq \mathbb{E}[\alpha^{\sigma_k(t, m) - t} (m + \delta) + V_k(t, m) - \alpha^{\sigma_k(t, m) - t} m \mid \mathcal{F}^k(t)]. \end{aligned}$$

Daraus folgt

$$(1.46) \quad \liminf_{\delta \downarrow 0} \frac{V_k(t, m + \delta) - V_k(t, m)}{\delta} \geq \mathbb{E}[\alpha^{\sigma_k(t, m) - t} \mid \mathcal{F}^k(t)]$$

$$(1.45) \text{ und } (1.46) \text{ liefern } \frac{\partial^+}{\partial m} V_k(t, m) = \mathbb{E}[\alpha^{\sigma_k(t, m) - t} \mid \mathcal{F}^k(t)].$$

Wegen  $\sigma_k(t, m) \downarrow t$  für  $m \rightarrow \infty$  gilt  $\lim_{m \rightarrow \infty} \frac{\partial^+}{\partial m} V_k(t, m) = 1$  und aus der Form von  $\sigma_k(t, m)$  folgt schließlich  $\lim_{m \rightarrow \infty} V_k(t, m) = m$ .  $\square$

*Beweis von Lemma 1.11.* Wir definieren die Menge

$$A^- \triangleq \{X \text{ } \mathcal{F}^k(t)\text{-messbar} : V_k(t, X) > X \text{ } \mathbb{P}\text{-f.s.}\},$$

die insbesondere alle negativen  $\mathcal{F}^k(t)$ -messbaren Zufallsvariablen enthält.

Zu  $X_1 \in A^+$ ,  $X_2 \in A^-$  ist  $V_k(t, X_1 \vee X_2) - (X_1 \vee X_2)$  echt positiv auf der Menge  $\{X_2 > X_1\}$ .

Wegen  $(X_1 \vee X_2) \geq X_1$  und (1.22) gilt  $\mathbb{P}$ -f.s.

$$V_k(t, X_1 \vee X_2) - (X_1 \vee X_2) = 0,$$

also ist  $X_1 \geq X_2$  und damit  $A^+ \geq A^-$  sowie

$$Z^+ \triangleq \text{ess inf } A^+ \geq \text{ess sup } A^- \triangleq Z^-.$$

Wir nehmen an, die Menge  $B \triangleq \{Z^+ > Z^-\}$  besitze positive Wahrscheinlichkeit. Wir wählen ein  $X^- \in A^-$  und setzen  $Y \triangleq X^- 1_{B^c} + \frac{Z^+ + Z^-}{2} 1_B$ . Wegen  $Y \not\leq Z^-$  gilt  $Y \notin A^-$  und damit  $V_k(t, Y) = Y$  mit positiver Wahrscheinlichkeit.

Auf  $B^c$  ist  $V_k(t, Y) > Y$ , also existiert  $\mathcal{F}^k(t) \ni C \subset B$  mit  $\mathbb{P}(C) > 0$  und  $V_k(t, Y) = Y$  auf  $C$ . Schließlich definiert man zu einem beliebigen  $Y^+ \in A^+$  die Zufallsvariable  $Y^* \triangleq 1_C \frac{Z^+ + Z^-}{2} + 1_{C^c} Y^+$ . Es gilt  $V_k(t, Y^*) = Y^*$ , aber  $Y^* \not\leq Z^+$  und damit  $Y^* \notin A^+$ . Dies liefert den Widerspruch.  $\square$

*Beweis von Proposition 1.12.* Sei  $X \in A^-$ . Dann gilt

$$0 < V_k(t, X) - X = \mathbb{E}\left[\sum_{u=t+1}^{\sigma(t, X)} \alpha^{u-t} Z_k(u) + \alpha^{\sigma_k(t, X) - t} X - X \mid \mathcal{F}^k(t)\right].$$

Wegen  $X \in A^-$  ist  $\sigma_k(t, X) \geq t + 1$  und es gilt

$$\frac{1 - \alpha}{\alpha} X < \frac{\mathbb{E}[\sum_{u=t+1}^{\sigma_k(t, X)} \alpha^u Z_k(u) \mid \mathcal{F}^k(t)]}{\mathbb{E}[\sum_{u=t+1}^{\sigma_k(t, X)} \alpha^u \mid \mathcal{F}^k(t)]} \leq \text{ess sup}_{\tau \geq t+1} \frac{\mathbb{E}[\sum_{u=t+1}^{\tau} \alpha^u Z_k(u) \mid \mathcal{F}^k(t)]}{\mathbb{E}[\sum_{u=t+1}^{\tau} \alpha^u \mid \mathcal{F}^k(t)]}.$$

Damit ist  $M_k(t)$  kleiner gleich der rechten Seite. Andererseits gilt für  $X \in A^+$  und jede Stoppzeit  $\tau \geq t + 1$ :

$$X = V_k(t, X) \geq \mathbb{E}\left[\sum_{u=t+1}^{\tau} \alpha^{u-t} Z_k(u) + \alpha^{\tau-t} X \mid \mathcal{F}^k(t)\right],$$

also gilt

$$\frac{1-\alpha}{\alpha}X \geq \frac{\mathbb{E}[\sum_{u=t+1}^{\tau} \alpha^u Z_k(u) \mid \mathcal{F}^k(t)]}{\mathbb{E}[\sum_{u=t+1}^{\tau} \alpha^u \mid \mathcal{F}^k(t)]}$$

und damit ist

$$\frac{1-\alpha}{\alpha}X \geq \operatorname{ess\,sup}_{\tau \geq t+1} \frac{\mathbb{E}[\sum_{u=t+1}^{\tau} \alpha^u Z_k(u) \mid \mathcal{F}^k(t)]}{\mathbb{E}[\sum_{u=t+1}^{\tau} \alpha^u \mid \mathcal{F}^k(t)]}$$

und  $M_k(t)$  größer gleich der rechten Seite. Dies liefert die Behauptung.  $\square$

*Beweis von Proposition 1.14.* Wegen (1.29) gilt

$$\alpha^{\sigma_k(t,m)-t} = (1-\alpha) \sum_{u=0}^{\infty} \alpha^u 1_{\{\sigma_k(t,m) \leq u+t\}} = (1-\alpha) \sum_{u=t}^{\infty} \alpha^{u-t} 1_{\{\underline{M}_k(t,u) \leq m\}}.$$

Damit erhalten wir zusammen mit (1.17),(1.29) und dem Satz von Fubini:

$$\begin{aligned} V_k(t, m) - V_k(t, 0) &= \int_0^m \frac{\partial^+}{\partial x} V_k(t, x) dx \\ &= \int_0^m \mathbb{E}[\alpha^{\sigma_k(t,x)-t} \mid \mathcal{F}^k(t)] dx \\ &= (1-\alpha) \mathbb{E} \left[ \int_0^m \sum_{u=t}^{\infty} \alpha^{u-t} 1_{\{\underline{M}_k(t,u) \leq x\}} dx \mid \mathcal{F}^k(t) \right] \\ &= (1-\alpha) \mathbb{E} \left[ \sum_{u=t}^{\infty} \alpha^{u-t} (m - \underline{M}_k(t, u))^+ \mid \mathcal{F}^k(t) \right] \\ (1.47) \quad &= (1-\alpha) \mathbb{E} \left[ \sum_{u=\sigma_k(t,m)}^{\infty} \alpha^{u-t} (m - \underline{M}_k(t, u)) \mid \mathcal{F}^k(t) \right]. \end{aligned}$$

Andererseits gilt ohne Einschränkung  $\sigma_k(t, 0) = \infty$  und wegen

$$V_k(t, m) = \mathbb{E}[\alpha^{\sigma_k(t,m)-t} m + \sum_{u=t+1}^{\sigma_k(t,m)} \alpha^{u-t} Z_k(u) \mid \mathcal{F}^k(t)]$$

die Gleichung

$$V_k(t, m) - V_k(t, 0) = \mathbb{E}[\alpha^{\sigma_k(t,m)-t} m - \sum_{u=\sigma_k(t,m)+1}^{\infty} \alpha^{u-t} Z_k(u) \mid \mathcal{F}^k(t)].$$

Daraus folgt (1.31), mit  $m \rightarrow \infty$  erhält man (1.32). Weiter gilt

$$\begin{aligned} V_k(t, 0) &= \mathbb{E} \left[ \sum_{u=t+1}^{\infty} \alpha^{u-t} Z_k(u) \mid \mathcal{F}^k(t) \right] \\ &= (1-\alpha) \mathbb{E} \left[ \sum_{u=t}^{\infty} \alpha^{u-t} \underline{M}_k(t, u) \mid \mathcal{F}^k(t) \right] \end{aligned}$$

Zusammen mit (1.47) und (1.29) folgt dann (1.33).  $\square$

*Beweis von Theorem 1.16.* Wegen  $0 \leq \prod_{i=1}^d \frac{\partial^+}{\partial m} V_i(s_i, m) \leq 1$  ist zunächst  $F(\tilde{s}, m) \geq m$ . Auf der Menge  $\{m \geq \max_{1 \leq j \leq d} M_j(s_j)\}$  gilt für  $x \geq m$  die Gleichung  $\frac{\partial^+}{\partial x} V_i(s_i, x) = 1$  und damit  $F(\tilde{s}, m) = m$ .

Hingegen gilt auf der komplementären Menge  $\{m < \max_{1 \leq j \leq d} M_j(s_j)\}$  für ein  $i$  und  $m \leq x < M_i$  die Ungleichung  $\frac{\partial^+}{\partial x} V_i(s_i, x) < 1$ , also  $F(\tilde{s}, m) > m$ .

Zu  $i \in \{1, \dots, d\}$  setzen wir  $s^{(i)} \triangleq (s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_d) \in \mathbb{N}_0^{d-1}$  und  $P_i(s^{(i)}, m) \triangleq \prod_{j \neq i} \frac{\partial^+}{\partial m} V_j(s_j, m)$ . Die Abbildung  $m \mapsto P_i(s^{(i)}, m)$  ist wachsend, das dadurch induzierte Maß auf  $[0, \infty)$  nennen wir  $\mathbb{P}_i(s^{(i)})$ . Mittels partieller Integration erhält man für  $m \leq K$

$$\begin{aligned} F(\tilde{s}, m) &= K - \int_m^K \prod_{i=1}^d \frac{\partial^+}{\partial x} V_i(s_i, x) dx \\ &= P_i(s^{(i)}, m) V_i(s_i, m) + \int_m^K V_i(s_i, x) \mathbb{P}_i(s^{(i)})(dx). \end{aligned}$$

Damit gilt

$$\begin{aligned} F(\tilde{s} + \tilde{e}_i, m) &= P_i(s^{(i)}, m) V_i(s_i + 1, m) \\ &\quad + \int_m^K V_i(s_i + 1, x) \mathbb{P}_i(s^{(i)})(dx), \\ \mathbb{E}[F(\tilde{s} + \tilde{e}_i, m) \mid \mathcal{F}(\tilde{s})] &= P_i(s^{(i)}, m) \mathbb{E}[V_i(s_i + 1, m) \mid \mathcal{F}(\tilde{s})] \\ &\quad + \int_m^K \mathbb{E}[V_i(s_i + 1, x) \mid \mathcal{F}(\tilde{s})] \mathbb{P}_i(s^{(i)})(dx). \end{aligned}$$

Setzt man nun

$$\varphi_i(t, m) \triangleq V_i(t, m) - \alpha \mathbb{E}[Z_i(t+1) \mid \mathcal{F}(\tilde{s})] - \alpha E[V_i(t+1, m) \mid \mathcal{F}(\tilde{s})],$$

so folgt

$$\begin{aligned} &F(\tilde{s}, m) - \alpha \mathbb{E}[Z_i(s_i + 1) \mid \mathcal{F}(\tilde{s})] - \alpha E[F(\tilde{s} + \tilde{e}_i, m) \mid \mathcal{F}(\tilde{s})] \\ &= P_i(s^{(i)}, m) V_i(s_i, m) + \int_m^K V_i(s_i, x) \mathbb{P}_i(s^{(i)})(dx) \\ &\quad - \alpha \mathbb{E}[Z_i(s_i + 1) \mid \mathcal{F}(\tilde{s})] - \alpha P_i(s^{(i)}, m) \mathbb{E}[V_i(s_i + 1, m) \mid \mathcal{F}(\tilde{s})] \\ &\quad - \alpha \int_m^K \mathbb{E}[V_i(s_i + 1, x) \mid \mathcal{F}(\tilde{s})] \mathbb{P}_i(s^{(i)})(dx) \\ &= P_i(s^{(i)}, m) \varphi_i(s_i, m) + \int_m^K \varphi_i(s_i, x) \mathbb{P}_i(s^{(i)})(dx) \\ &\quad - \alpha(1 - P_i(s^{(i)}, m)) \mathbb{E}[Z_i(s_i + 1) \mid \mathcal{F}(\tilde{s})] \\ &\quad + \int_m^K \alpha \mathbb{E}[Z_i(s_i + 1) \mid \mathcal{F}(\tilde{s})] \mathbb{P}_i(s^{(i)})(dx) \\ (1.48) \quad &= P_i(s^{(i)}, m) \varphi_i(s_i, m) + \int_m^K \varphi_i(s_i, x) \mathbb{P}_i(s^{(i)})(dx), \end{aligned}$$

die letzte Gleichung gilt wegen  $P_i(s^{(i)}, K) = 1$ . Wegen (1.16) ist  $\varphi_i \geq 0$  und damit auch die rechte Seite von (1.48). Damit ist also

$$(1.49) \quad F(\tilde{s}, m) \geq \alpha \mathbb{E}[Z_i(s_i + 1) \mid \mathcal{F}(\tilde{s})] - \alpha E[F(\tilde{s} + \tilde{e}_i, m) \mid \mathcal{F}(\tilde{s})].$$

Betrachten wir nun die Menge  $\{m < M_i = \max_{1 \leq j \leq d} M_j\}$ . Für  $m < M_i(s_i)$  gilt  $\varphi_i(s_i, m) = 0$ . Dagegen ist für  $m \geq M_i(s_i)$  und alle  $j \neq i$   $\frac{\partial^+}{\partial x} V_j(s_j, x) = 1$  und damit  $P_i(s^{(i)}, m) = 1$ . Mit diesen Eigenschaften liefert (1.48)

$$(1.50) \quad \begin{aligned} & F(\tilde{s}, m) - \alpha \mathbb{E}[Z_i(t+1) \mid \mathcal{F}(\tilde{s})] - \alpha E[F(\tilde{s} + \tilde{e}_i, m) \mid \mathcal{F}(\tilde{s})] \\ &= \int_{M_i}^K \varphi_i(s_i, x) \mathbb{P}_i(s^{(i)})(dx) = 0. \end{aligned}$$

Es bleibt zu zeigen, dass

$$\begin{aligned} & \{m < M_i = \max_{1 \leq j \leq d} M_j\} \\ &= \{m < \max_{1 \leq j \leq d} [\alpha \mathbb{E}[Z_j(s_j + 1) \mid \mathcal{F}(\tilde{s})] + \alpha E[F(\tilde{s} + \tilde{e}_j, m) \mid \mathcal{F}(\tilde{s})]]\} \end{aligned}$$

$\mathbb{P}$ -fast sicher erfüllt ist. Dies gilt, denn ist  $m \geq M_i \triangleq \max_{1 \leq j \leq d} M_j$ , so gilt

$$m = F(\tilde{s}, m) \geq \alpha \mathbb{E}[Z_j(t+1) \mid \mathcal{F}(\tilde{s})] + \alpha E[F(\tilde{s} + \tilde{e}_j, m) \mid \mathcal{F}(\tilde{s})]$$

für alle  $j$  und damit auch für das Maximum.

Gilt dagegen  $m < M_i$ , so erhält man

$$m < F(\tilde{s}, m) = \alpha \mathbb{E}[Z_i(t+1) \mid \mathcal{F}(\tilde{s})] + \alpha E[F(\tilde{s} + \tilde{e}_i, m) \mid \mathcal{F}(\tilde{s})],$$

die rechte Seite kann man durch das Maximum über  $j \in \{1, \dots, d\}$  abschätzen.

Damit erfüllt  $F$  die Gleichung der Dynamischen Programmierung (1.37) und das Theorem ist bewiesen.  $\square$

*Beweis von Korollar 1.17.* Ohne Einschränkung nehmen wir an, die Strategie  $T$  wählt im Zeitpunkt  $t$  den Arm  $k$ . Dann gilt wegen (1.49)

$$\begin{aligned} & \alpha^{-t} \mathbb{E}[Q_T(t+1) - Q_T(t) \mid \mathcal{F}(T(t))] \\ &= \mathbb{E}[\alpha F(T(t) + e_k) + \alpha Z^k(T(t) + e_k) \mid \mathcal{F}(T(t))] - F(T(t)) \\ &\leq 0, \end{aligned}$$

und  $Q_T$  ist für beliebige Strategien  $T$  ein Supermartingal. Ein Gleichheitszeichen und damit ein Martingal erhält man, wenn die Strategie stets einen Arm mit maximalem Gittins-Index wählt:  $M_k = \max_{1 \leq j \leq d} M_j$  wegen (1.50). Ist  $T$  keine Indexstrategie, so ist der Term  $\int_{M_i}^K \varphi_i(s_i, m) \mathbb{P}_i(s^{(i)})(dm)$  in (1.50) für ein  $\tilde{s}$  ungleich 0 und damit  $Q_T$  kein Martingal.  $\square$



# Kapitel 2

## Mehrrarmige Banditen in stetiger Zeit

### 2.1 Übersicht

Die Untersuchung dynamischer Allokationsprobleme und ihrer Lösungen in stetiger Zeit sowie ein weiterer Beweis des Gittins-Indextheorems sind die Themen dieses Kapitels.

Die Formulierung des Problems erfolgt in Abschnitt 2.2, Abschnitt 2.3 erinnert an einige Eigenschaften optionaler Prozesse und optionaler Projektionen. Eine neue Darstellung für den erwarteten Ertrag einer Strategie wird in Abschnitt 2.4 hergeleitet. Der erwartete Beitrag der einzelnen Arme zum Ertrag einer Strategie führt auf die von Bank und El Karoui [1] gelösten Darstellungsprobleme für optionale Prozesse. Genauer bleibt die Darstellung des erwarteten Ertrags gültig, wenn man die Auszahlungsprozesse der Arme durch die laufenden Infima gewisser optionaler Prozesse ersetzt. Dies wird in Abschnitt 2.5 gezeigt. Falls diese Prozesse pfadweise unterhalbstetig von rechts sind, so gilt nach Bank und El Karoui eine Darstellung, die sie als die bekannten Gittins-Indexprozesse identifiziert.

Dass die Lösungen der Darstellungsprobleme in dem hier betrachteten Fall tatsächlich die gewünschte Regularität besitzen, wird in Abschnitt 2.6 demonstriert.

Die Darstellung des Ertrags in Abhängigkeit der Indexprozesse wird in Abschnitt 2.7 genutzt, um für diesen mit Hilfe zweier Abschätzungen eine obere Schranke zu erhalten. Gilt für eine Strategie zweimal die Gleichheit in den betrachteten Abschätzungen, so ist diese optimal.

Eine der Ungleichungen führt das Allokationsproblem auf eines mit fallenden rechtsstetigen Auszahlungsprozessen zurück. Dieses wird, El Karoui und Karatzas [12] folgend, in Abschnitt 2.8 gelöst.

Ein einfaches Beispiel zeigt in Abschnitt 2.9, dass im Gegensatz zum zeitdiskreten Fall nun Strategien, die ausschließlich Arme mit maximalem Gittins-Index wählen, nicht optimal sein müssen. Anschließend werden hier die von Kaspi und Mandelbaum [17] eingeführten

*Indexstrategien* und ihre Verbindung zu den *Strategien vom Index-Typ* von El Karoui und Karatzas [11] studiert.

In Abschnitt 2.12 wird gezeigt, dass der erwartete Ertrag einer Strategie genau dann die erwähnte obere Schranke realisiert, wenn dies eine Indexstrategie im Sinne von Kaspi und Mandelbaum ist. Zusammen mit der in den Abschnitten 2.10 und 2.11 konstruierten Indexstrategie beweist dies das Gittins-Indextheorem und charakterisiert die Menge der optimalen Strategien vollständig.

## 2.2 Formulierung des Allokationsproblems

Es seien  $d$  nichtnegative stochastische Prozesse  $(h_p(t))_{t \geq 0}$ , die die Auszahlungsraten der Arme  $p = 1, \dots, d$  modellieren. Der Prozess  $h_p$  sei adaptiert an eine Filtration  $\mathcal{F}^p$ , die die üblichen Bedingungen erfüllt. Die Filtrationen  $\mathcal{F}^p, p = 1, \dots, d$ , seien unabhängig.

Weiter seien wie im zeitdiskreten Fall

$$(2.1) \quad \mathcal{F}(\tilde{s}) \triangleq \bigvee_{p=1}^d \mathcal{F}^p(s_p),$$

$$(2.2) \quad \begin{aligned} \overline{\mathcal{F}}^p(t) &\triangleq \mathcal{F}(\infty, \dots, t, \dots, \infty) \\ &= \mathcal{F}^p(t) \vee \bigvee_{k \neq p} \mathcal{F}^k(\infty). \end{aligned}$$

$h_p$  erfülle für den vorgegebenen Diskontierungsparameter  $\alpha > 0$  die Integrierbarkeitsbedingung

$$(2.3) \quad \mathbb{E} \left[ \int_0^\infty e^{-\alpha t} h_p(t) dt \right] < \infty.$$

Eine *Strategie*  $T$  ist ein  $[0, \infty)^d$ -wertiger stochastischer Prozess mit

$$(2.4) \quad T(0) = 0 \text{ und } T(t) \text{ ist wachsend in } t,$$

$$(2.5) \quad T_1(t) + \dots + T_d(t) = t \quad \forall t \in [0, \infty),$$

$$(2.6) \quad \{T(t) \leq \tilde{s}\} \in \mathcal{F}(\tilde{s}) \quad \forall t \in [0, \infty), \tilde{s} \in [0, \infty)^d.$$

Mit diesen Eigenschaften ist  $\mathbb{P}$ -f.s. jeder Pfad  $t \mapsto T_p(t)$  einer Strategie  $T$  absolutstetig bezüglich des Lebesguemaßes und es gilt

$$T_p(t) = \int_0^\infty \chi_p(s) ds$$

für geeignete Prozesse  $\chi_p$  mit  $\chi_p(t) \in [0, 1]$  und  $\sum_{p=1}^d \chi_p(t) = 1$ .  $\chi_p(t)$  kann als Rate interpretiert werden, mit der die Strategie  $T$  den Arm  $p$  im Zeitpunkt  $t$  betätigt. Der Unterschied zum zeitdiskreten Fall besteht darin, dass nun mehrere Arme gleichzeitig

betätigt werden können. Die Menge der Strategien bezeichnen wir mit  $\mathcal{S}$ .

Welche Informationen zum Zeitpunkt  $t \in [0, \infty)$  zur Verfügung stehen, hängt von der Strategie  $T$  ab, mit der bis  $t$  gespielt wurde. Diese sind definiert durch die  $\sigma$ -Algebra

$$\mathcal{F}(T(t)) = \{A \in \mathcal{F} : A \cap \{T(t) \leq \tilde{s}\} \in \mathcal{F}(\tilde{s})\}.$$

Zu jeder Strategie  $T$  erhält man so eine Filtration  $(\mathcal{F}(T(t)))_{t \geq 0}$ , die die unter dieser Strategie sukzessive verfügbaren Informationen darstellt und den üblichen Bedingungen genügt. Die Strategie  $T$  liefert den zufälligen Wert

$$(2.7) \quad \mathcal{R}(T) = \sum_{p=1}^d \int_0^\infty e^{-\alpha t} h_p(T_p(t)) dT_p(t).$$

Gesucht ist eine Strategie  $\hat{T}$ , die den maximalen erwarteten Wert liefert:

$$(2.8) \quad \mathbb{E}[\mathcal{R}(\hat{T})] = \max_{T \in \mathcal{S}} \mathbb{E}[\mathcal{R}(T)].$$

Der erste Schritt unserer Untersuchung ist die Herleitung einer weiteren Darstellung der Auszahlung  $\mathcal{R}(T)$  und ihres Erwartungswertes. Dazu benötigen wir einige Eigenschaften optionaler Prozesse, die im nächsten Abschnitt zusammengefasst sind.

## 2.3 Optionale Projektionen

Gegeben sei ein filtrierter Wahrscheinlichkeitsraum  $(\Omega, \mathcal{F}, (\mathcal{G}(t)), \mathbb{P})$ , der die üblichen Bedingungen erfüllt.

**Definition 2.1.** Die  $\sigma$ -Algebra  $\Theta$  auf  $\mathbb{R}_+ \times \Omega$ , die von den adaptierten reellwertigen Prozessen mit càdlàg Pfaden erzeugt wird, heißt optionale  $\sigma$ -Algebra.

Ein stochastischer Prozess  $X$  heißt optional, wenn die Abbildung

$$\begin{aligned} \mathbb{R}_+ \times \Omega &\rightarrow \mathbb{R} \\ (t, \omega) &\mapsto X(t, \omega) \end{aligned}$$

messbar bezüglich  $\Theta$  ist.

Eine wichtige Folgerung aus Meyers Theorem über optionale Schnitte ([6]-IV-84) ist, dass ein optionaler Prozess durch seine Werte zu allen Stoppzeiten festgelegt ist.

**Theorem 2.2** ([6]-IV-86). Es seien  $X$  und  $Y$  zwei optionale Prozesse und für jede Stoppzeit  $\mu$  gelte  $\mathbb{P}$ -f.s.  $X(\mu) = Y(\mu)$  auf  $\{\mu < \infty\}$ . Dann sind  $X$  und  $Y$  ununterscheidbar.

Im Folgenden benötigen wir den Begriff der optionalen Projektion eines stochastischen Prozesses:

**Theorem 2.3** ([7]-VI-43). *Es sei  $H$  ein produktmessbarer stochastischer Prozess*

$$H : \Omega \times [0, \infty) \rightarrow \mathbb{R}_+.$$

*Dann existiert ein optionaler Prozess  $\pi_{\mathcal{G}}[H]$ , der für alle  $\mathcal{G}$ -Stoppzeiten  $\mu$   $\mathbb{P}$ -f.s. die folgende Bedingung erfüllt:*

$$\mathbb{E}[H(\mu)1_{\{\mu < \infty\}} \mid \mathcal{G}(\mu)] = \pi_{\mathcal{G}}[H](\mu)1_{\{\mu < \infty\}}.$$

$\pi_{\mathcal{G}}[H]$  heißt optionale Projektion von  $H$  bezüglich der Filtration  $\mathcal{G}$ . Ist  $H$  selbst bereits optional, so ist  $\pi_{\mathcal{G}}[H] = H$ .

Sind  $X, X^n, Y$  produktmessbare nichtnegative Prozesse mit  $X^n \uparrow X \leq Y$ , so folgt  $\lim \pi_{\mathcal{G}}[X_n] = \pi_{\mathcal{G}}[X] \leq \pi_{\mathcal{G}}[Y]$ . Weiter gilt das folgende

**Theorem 2.4** ([5]-V-24). *Es seien  $H_1$  und  $H_2$  produktmessbare nichtnegative Prozesse mit  $\pi_{\mathcal{G}}[H_1] = \pi_{\mathcal{G}}[H_2]$ . Dann gilt für alle wachsenden rechtsstetigen  $\mathcal{G}$ -adaptierten Prozesse  $A$*

$$\mathbb{E} \left[ \int_0^\infty H_1(t) dA(t) \right] = \mathbb{E} \left[ \int_0^\infty H_2(t) dA(t) \right].$$

Gewisse Pfadeneigenschaften bleiben unter der optionalen Projektion erhalten:

**Theorem 2.5** ([7]-VI-47). *Es sei  $H$  produktmessbar und von der Klasse  $(D)$ . Besitzt  $H$  rechtsstetige (bzw. càdlàg) Pfade, so gilt dies auch für  $\pi_{\mathcal{G}}[H]$ .*

Das folgende Lemma spielt eine wichtige Rolle bei der Betrachtung des Darstellungsproblems (2.13). Es sagt, dass die optionalen Projektionen bezüglich zweier verschiedener Filtrationen unter gewissen Bedingungen gleich sind.

**Lemma 2.6.** *Es sei  $\mathcal{A} \subset \mathcal{F}$  eine von  $\mathcal{G}(\infty)$  unabhängige  $\sigma$ -Algebra und die Filtration  $(\bar{\mathcal{G}}(t))_{t \geq 0}$  definiert durch  $\bar{\mathcal{G}}(t) \triangleq \mathcal{G}(t) \vee \mathcal{A}$ .*

*Dann sind für jede  $\mathcal{G}(\infty) \otimes B[0, \infty)$ -messbare Abbildung  $H : \Omega \times [0, \infty) \rightarrow \mathbb{R}_+$  die optionalen Projektionen  $\pi_{\mathcal{G}}[H]$  und  $\pi_{\bar{\mathcal{G}}}[H]$  ununterscheidbar.*

*Beweis.* Wir setzen

$$\begin{aligned} \mathcal{P} &\triangleq \{H : \Omega \times [0, \infty) \rightarrow \mathbb{R}_+ : H \text{ ist } \mathcal{G}(\infty) \otimes B[0, \infty)\text{-messbar und beschränkt}\}, \\ \mathcal{H} &\triangleq \{H \in \mathcal{P} : \pi_{\mathcal{G}}[H] \text{ ist ununterscheidbar von } \pi_{\bar{\mathcal{G}}}[H]\}. \end{aligned}$$

$\mathcal{H}$  ist ein Vektorraum, der alle konstanten Abbildungen enthält. Weiter ist  $\mathcal{H}$  unter monotoner Konvergenz abgeschlossen und damit eine monotone Klasse.

Wir betrachten die Menge

$$\mathcal{C} \triangleq \{H \in \mathcal{P} : H(\omega, t) = 1_A(\omega)1_{[a,b)}(t) \text{ für } A \in \mathcal{G}(\infty) \text{ und } a, b \in \mathbb{R}_+\}.$$

Für  $H \in \mathcal{P}$  und  $t \in \mathbb{R}_+$  gilt wegen der vorausgesetzten Unabhängigkeit  $\mathbb{P}$ -f.s.

$$\begin{aligned}
\pi_{\bar{\mathcal{G}}}[H](t) &= \mathbb{E}[H(t) \mid \bar{\mathcal{G}}(t)] \\
&= \mathbb{E}[H(t) \mid \mathcal{G}(t) \vee \mathcal{A}] \\
&= \mathbb{E}[H(t) \mid \mathcal{G}(t)] \\
(2.9) \qquad &= \pi_{\mathcal{G}}[H](t).
\end{aligned}$$

Ist  $H \in \mathcal{C}$ , so sind die Pfade von  $H$  rechtsstetig. Nach Theorem 2.5 besitzen die optionalen Projektionen ebenfalls rechtsstetige Pfade und sind damit wegen (2.9) ununterscheidbar. Also gilt  $\mathcal{C} \subset \mathcal{H}$ . Weiter ist  $\mathcal{C}$  abgeschlossen bezüglich punktweiser Multiplikation.

Aus dem Satz über monotone Klassen folgt, dass  $\mathcal{H} = \mathcal{P}$ , der Satz über monotone Konvergenz liefert schließlich die Behauptung für beliebige nichtnegative produktmessbare Abbildungen.  $\square$

**Bemerkung 2.7.** *Wir werden Lemma 2.6 auf die Filtrationen  $\mathcal{F}^p$  und  $\bar{\mathcal{F}}^p$  anwenden. Tatsächlich behält es seine Gültigkeit auch unter den allgemeineren Voraussetzungen von El Karoui und Karatzas [12] an die mehrparametrische Filtration  $\mathcal{F}$ .*

*Statt der Bedingung (2.1) setzt man dort neben den üblichen Bedingungen die schwächere Cairoli-Walsh-Eigenschaft voraus.*

*Eine mehrparametrische Filtration  $(\mathcal{F}(\tilde{s}))_{s \in [0, \infty)^d}$  besitzt die Cairoli-Walsh-Eigenschaft, wenn folgendes gilt:*

$$(2.10) \qquad \mathcal{F}(\tilde{s}), \mathcal{F}(\tilde{r}) \text{ sind bedingt unabhängig, gegeben } \mathcal{F}(\tilde{s} \wedge \tilde{r}).$$

*Daraus folgt nach Cairoli und Walsh [4]: Für alle beschränkten oder nichtnegativen Zufallsvariablen  $X$  und  $\tilde{r}, \tilde{s} \in [0, \infty)^d$  gilt*

$$\mathbb{E}[X \mid \mathcal{F}(\tilde{r} \wedge \tilde{s})] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{F}(\tilde{r})] \mid \mathcal{F}(\tilde{s})].$$

*Die Struktur der Filtrationen spielt im Beweis des Lemmas nur in den Gleichungen (2.9) eine Rolle. Besitzt  $\mathcal{F}$  die Cairoli-Walsh-Eigenschaft, so gelten diese Identitäten ebenfalls. Dazu setzt man  $\tilde{s} = (0, \dots, 0, t, 0, \dots, 0)$  und  $\tilde{r} = (\infty, \dots, \infty, t, \infty, \dots, \infty)$  und erhält*

$$\begin{aligned}
\pi_{\bar{\mathcal{F}}^k}[H](t) &= \mathbb{E}[H(t) \mid \bar{\mathcal{F}}^k(t)] \\
&= \mathbb{E}[\mathbb{E}[H(t) \mid \bar{\mathcal{F}}^k(t)] \mid \mathcal{F}^k(t)] \\
&= \mathbb{E}[H(t) \mid \mathcal{F}^k(t)] \\
&= \pi_{\mathcal{F}^k}[H](t).
\end{aligned}$$

## 2.4 Eine Darstellung der Rendite

Die Interaktion der verschiedenen Zeitskalen macht die Maximierung von (2.7) kompliziert. Eine andere Darstellung kann man mit der folgenden Überlegung herleiten. Der

Beitrag des Arms  $p$  zum Gewinn einer Strategie  $T$  ist der aus einer ununterbrochenen Betätigung dieses Armes resultierende Gewinn, verringert um die wegen der möglichen Betätigung anderer Arme durch die Strategie  $T$  entstehenden Diskontierungsverluste. Erhöht sich  $T^p$  während des Zeitraumes  $\Delta t$  um den Betrag  $\Delta T^p$ , so verringert sich der Wert der darauffolgenden Zahlungen des Armes  $p$  durch die zusätzliche Diskontierung mit dem Faktor  $e^{-\alpha(\Delta t - \Delta T^p(t))}$ . Dies motiviert die folgende Darstellung.

**Lemma 2.8.** *Der Wert einer Strategie  $T$  hat die Darstellung*

$$(2.11) \quad \mathcal{R}(T) = \sum_{p=1}^d \int_0^\infty e^{-\alpha s} h_p(s) ds + \sum_{p=1}^d \int_0^\infty \int_{T_p(t)}^\infty e^{-\alpha s} h_p(s) ds de^{-\alpha(t-T_p(t))},$$

für seinen Erwartungswert gilt

$$(2.12) \quad \mathbb{E}[\mathcal{R}(T)] = \sum_{p=1}^d \mathbb{E} \left[ \int_0^\infty e^{-\alpha s} h_p(s) ds \right] + \sum_{p=1}^d \mathbb{E} \left[ \int_0^\infty \mathbb{E} \left[ \int_{T_p(t)}^\infty e^{-\alpha s} h_p(s) ds \mid \overline{\mathcal{F}}^p(T_p(t)) \right] de^{-\alpha(t-T_p(t))} \right].$$

*Beweis.* Die Pfade des nichtnegativen Prozesses

$$K_p(t) \triangleq \int_{T_p(t)}^\infty e^{-\alpha s} h_p(s) ds$$

sind  $\mathbb{P}$ -f.s. fallend und stetig. Spielt man bis zum Zeitpunkt  $t$  mit Strategie  $T$ , so ist  $K_p(t)$  die zukünftige Rendite von Arm  $p$ , wenn nur noch dieser Arm betätigt wird.

Wegen  $dK_p(t) = -e^{-\alpha T_p(t)} h_p(T_p(t)) dT_p(t)$  erhält der Beitrag des Armes  $p$  die folgende Form

$$\int_0^\infty e^{-\alpha t} h_p(T_p(t)) dT_p(t) = - \int_0^\infty e^{-\alpha(t-T_p(t))} dK_p(t).$$

Partielle Integration liefert

$$\int_0^\infty e^{-\alpha t} h_p(T_p(t)) dT_p(t) = \int_0^\infty e^{-\alpha s} h_p(s) ds + \int_0^\infty K_p(t) de^{-\alpha(t-T_p(t))},$$

und damit folgt (2.11).

Der Prozess  $K_p$  besitzt stetige Pfade und ist von der Klasse (D), damit sind nach Theorem 2.5 die Pfade von  $\pi[K_p]$  càdlàg. Mit Theorem 2.4 folgt (2.12).  $\square$

Die Darstellung (2.12) liefert die Verbindung zu den von Bank und El Karoui [1] betrachteten Darstellungsproblemen. Genauer suchen wir Prozesse  $M_p$  mit

$$(2.13) \quad \begin{aligned} & \mathbb{E} \left[ \int_{T_p(t)}^{\infty} e^{-\alpha t} h_p(t) ds \mid \overline{\mathcal{F}}^p(T_p(t)) \right] \\ &= \mathbb{E} \left[ \int_{T_p(t)}^{\infty} \alpha e^{-\alpha s} \inf_{v \in [T_p(t), s]} M_p(v) ds \mid \overline{\mathcal{F}}^p(T_p(t)) \right] \end{aligned}$$

für  $p = 1, \dots, d$  und  $t \geq 0$ . Dies ist eine Verallgemeinerung der Darstellung (3.7) von El Karoui und Karatzas [11], in der  $M_p$  der Gittins-Indexprozess des Armes  $p$  ist. Eine Interpretation dieser Gleichung findet man in Bemerkung 1.15.

## 2.5 Das Darstellungsproblem

In diesem Abschnitt werden wir zunächst das Darstellungsproblem (2.13) mit der kleineren Filtration  $\mathcal{F}^p$  betrachten, zu dem nach Bank und El Karoui [1] eine optionale Lösung  $M_p$  existiert.

Falls  $M_p$  pfadweise unterhalbstetig von rechts ist erhält man mit *Theorem 1* von [1] eine Darstellung, die  $M_p$  als den Gittins-Indexprozess des Armes  $p$  identifiziert. Diese Regularität der Pfade von  $M_p$  werden wir im nächsten Abschnitt überprüfen.

Die Gültigkeit der Darstellung (2.13) für die größere Filtration  $\overline{\mathcal{F}}^p$  folgt schließlich aus Lemma 2.6.

Bank und El Karoui [1] lösen eine Verallgemeinerung des folgenden Darstellungsproblems. Auf einem filtrierten Wahrscheinlichkeitsraum  $(\Omega, \mathcal{G}, (\mathcal{G}(t)), \mathbb{P})$ , der die üblichen Bedingungen erfüllt, betrachtet man einen reellwertigen optionalen Prozess  $(X(t))_{t \in [0, \infty]}$  der Klasse  $D$ , der unterhalbstetig in Erwartung ist und  $X(\infty) = 0$  erfüllt. Weiter sei

$$f : [0, \infty] \times \mathbb{R} \rightarrow \mathbb{R}$$

eine Abbildung, die die folgenden Bedingungen der *Assumption 1* in [1] erfüllt:

(i) Für jedes  $t \in [0, \infty]$  ist die Funktion  $f(t, \cdot) : \mathbb{R} \rightarrow \mathbb{R}$  stetig und streng monoton wachsend von  $-\infty$  nach  $+\infty$ .

(ii) Für jedes  $l \in \mathbb{R}$  gelte  $\int_0^{\infty} |f(t, l)| dt < +\infty$ .

Bank und El Karoui zeigen, dass ein optionaler Prozess  $M_p$  mit Werten in  $\mathbb{R} \cup \{-\infty\}$  existiert, so dass  $X$  für jede  $\mathcal{G}$ -Stoppzeit  $\eta$  mit Werten in  $[0, \infty]$  die Darstellung

$$(2.14) \quad X(\eta) = \mathbb{E} \left[ \int_{\eta}^{\infty} f(t, \inf_{\eta \leq v \leq t} M_p(v)) dt \mid \mathcal{G}(\eta) \right]$$

besitzt.  $M_p$  erfüllt für jede Stoppzeit  $\eta$  die Integrabilitätsbedingung

$$f(t, \inf_{\eta \leq v \leq t} M_p(v)) 1_{[\eta, \infty)}(t) \in L^1(\mathbb{P} \otimes dt)$$

und wird mit Hilfe besonderer Snellscher Enveloppes konstruiert.

Um dieses Resultat für unser Darstellungsproblem (2.13) zu nutzen, sei  $X$  die optionale Projektion des Prozesses  $H$

$$(2.15) \quad H(t) \triangleq \int_t^\infty e^{-\alpha s} h_p(s) ds$$

bezüglich der Filtration  $\mathcal{F}^p$ . Weiter seien

$$\begin{aligned} f(t, l) &\triangleq \alpha e^{-\alpha t} l, \\ \mathcal{G}(t) &\triangleq \mathcal{F}^p(t). \end{aligned}$$

$\Lambda$  bezeichne die Menge der  $\mathcal{F}^p$ -Stoppzeiten mit Werten in  $[0, \infty]$  und für  $\eta \in \Lambda$  seien

$$\Lambda(\eta) \triangleq \{\mu \in \Lambda, \mu \geq \eta\}, \quad \Lambda^>(\eta) \triangleq \{\mu \in \Lambda(\eta), \mu > \eta \text{ auf } \{\eta < \infty\}\}.$$

Die Gleichung (2.14) erhält damit in unserer Situation für  $\eta \in \Lambda$  die Gestalt

$$\mathbb{E} \left[ \int_\eta^\infty e^{-\alpha s} h_p(s) ds \mid \mathcal{F}^p(\eta) \right] = \mathbb{E} \left[ \int_\eta^\infty \alpha e^{-\alpha s} \inf_{\eta \leq v \leq t} M_p(v) dt \mid \mathcal{F}^p(\eta) \right].$$

Falls  $M_p$  pfadweise unterhalbstetig von rechts ist, gilt nach *Theorem 1* in [1] für  $\eta \in \Lambda$  die als *Forward Induction* bekannte Darstellung

$$(2.16) \quad M_p(\eta) = \operatorname{ess\,sup}_{\mu \in \Lambda^>(\eta)} \frac{\mathbb{E} \left[ \int_\eta^\mu e^{-\alpha t} h_p(t) dt \mid \mathcal{F}^p(\eta) \right]}{\mathbb{E} \left[ \int_\eta^\mu e^{-\alpha t} dt \mid \mathcal{F}^p(\eta) \right]}.$$

Damit ist  $M_p$  der bekannte Gittins-Indexprozess des Armes  $p$ . Setzt man

$$I(t) \triangleq \int_t^\infty e^{-\alpha s} \inf_{v \in [t, s]} M_p(v) ds,$$

so stimmen insbesondere die optionalen Projektionen von  $H$  und  $I$  bezüglich  $\mathcal{F}^p$  in allen  $\mathcal{F}^p$ -Stoppzeiten überein und sind damit nach Theorem 2.2 ununterscheidbar:

$$(2.17) \quad \pi_{\mathcal{F}^p}(H) = \pi_{\mathcal{F}^p}(I).$$

Für die Gültigkeit der gewünschten Darstellung (2.13) ist notwendig, dass dies auch für die optionalen Projektionen bezüglich der größeren Filtration  $\overline{\mathcal{F}}^p$  gilt.

Wendet man zweimal Lemma 2.6 auf die Filtrationen  $\mathcal{F}^p$  und  $\overline{\mathcal{F}}^p$  und die Prozesse  $H$  und  $I$  an, so erhält man mit (2.17) die folgenden Identitäten

$$\pi_{\overline{\mathcal{F}}^p}(H) = \pi_{\mathcal{F}^p}(H) = \pi_{\mathcal{F}^p}(I) = \pi_{\overline{\mathcal{F}}^p}(I).$$

Damit gilt (2.13). Aus der pfadweisen Unterhalbstetigkeit von rechts der Prozesse  $M_p$  folgt weiter, dass die fallenden Prozesse

$$\left( \inf_{v \in [T_p(t), s]} M_p(v) \right)_{s \geq T_p(t)}$$

càdlàg und damit ebenfalls optional sind.

Im nächsten Abschnitt werden wir zeigen, dass die von Bank und El Karoui konstruierte Lösung  $M_p$  des Darstellungsproblems in unserem Fall tatsächlich die gewünschte Regularität aufweist.

## 2.6 Regularität der Indexprozesse

Wir benötigen die folgenden Begriffe.

**Definition 2.9** ([8]). *Es sei  $X$  ein optionaler Prozess mit Werten in  $\overline{\mathbb{R}}$ .*

(i)  *$X$  heißt oberhalbstetig von rechts, wenn für jede fallende Folge von Stoppzeiten  $(\mu_n)$  auf der Menge  $\{\mu = \lim_n \mu_n\}$   $\mathbb{P}$ -f.s. gilt  $X(\mu) \geq \limsup_n X(\mu_n)$ .*

(ii)  *$X$  heißt oberhalbstetig von rechts in Erwartung, wenn  $X(\mu)$  integrierbar ist für jede Stoppzeit  $\mu$  und für jede fallende Folge von Stoppzeiten  $(\mu_n) \rightarrow \mu$  gilt  $\mathbb{E}[X(\mu)] \geq \liminf_n \mathbb{E}[X(\mu_n)]$ .*

(iii)  *$X$  heißt unterhalbstetig von rechts (in Erwartung), wenn  $-X$  oberhalbstetig (in Erwartung) ist.*

In unserer Situation ist der Prozess  $X$  stetig in Erwartung und rechtsstetig. Wir werden nun mit Hilfe der Konstruktion von Bank und El Karoui zeigen, dass daraus die gewünschte Regularität für den Prozess  $M_p$  folgt.

*Lemma 3.11* in [1] zeigt die Existenz einer mit  $l \in \mathbb{R}$  indizierten Familie Snellscher Enveloppes  $Y^l$ , die unter anderem die folgenden Bedingungen erfüllt:

(i) *Die Abbildung*

$$\begin{aligned} Y : \Omega \times [0, \infty] \times \mathbb{R} &\rightarrow \mathbb{R} \\ (\omega, t, l) &\mapsto Y^l(\omega, t) \end{aligned}$$

*ist produktmessbar.*

(ii) Für  $l \in \mathbb{R}$  ist der Prozess  $Y^l : \Omega \times [0, \infty] \rightarrow \mathbb{R}$  optional und für  $\eta \in \Lambda$  gilt

$$Y^l(\eta) = \operatorname{ess\,inf}_{\mu \in \Lambda(\eta)} \mathbb{E} \left[ X(\mu) - \int_{\eta}^{\mu} f(t, l) dt \mid \mathcal{G}(\eta) \right] \quad \mathbb{P} - f.s. .$$

(iii) Für  $(\omega, t) \in \Omega \times [0, \infty]$  ist die Abbildung  $l \mapsto Y^l(\omega, t)$  stetig und monoton fallend mit  $\lim_{l \downarrow -\infty} Y^l(\omega, t) = X(\omega, t)$ .

In unserem Fall nimmt  $Y^l$  die folgende Form an:

$$(2.18) \quad Y^l(\eta) = \operatorname{ess\,inf}_{\mu \in \Lambda(\eta)} \mathbb{E} \left[ \int_{\mu}^{\infty} e^{-\alpha s} h_p(s) ds - \int_{\eta}^{\mu} \alpha e^{-\alpha t} l dt \mid \mathcal{G}(\eta) \right].$$

**Lemma 2.10.** Die Pfade des optionalen Prozesses  $Y^l$  sind  $\mathbb{P}$ -f.s. rechtsstetig.

*Beweis.* Wir zeigen zunächst, dass  $Y^l$  oberhalbstetig von rechts in Erwartung ist. Dazu sei  $(\eta_n)$  eine fallende Folge von Stoppzeiten mit  $\eta_n \rightarrow \eta$ . Wir müssen zeigen, dass

$$\begin{aligned} & \mathbb{E} \left[ \operatorname{ess\,sup}_{\mu \geq \eta} \mathbb{E} \left[ - \int_{\mu}^{\infty} e^{-\alpha s} h_p(s) ds + \int_{\eta}^{\mu} \alpha e^{-\alpha t} l dt \mid \mathcal{G}(\eta) \right] \right] \\ & \leq \limsup_n \mathbb{E} \left[ \operatorname{ess\,sup}_{\mu \geq \eta_n} \mathbb{E} \left[ - \int_{\mu}^{\infty} e^{-\alpha s} h_p(s) ds + \int_{\eta_n}^{\mu} \alpha e^{-\alpha t} l dt \mid \mathcal{G}(\eta_n) \right] \right]. \end{aligned}$$

Die rechte Seite ist gleich

$$\begin{aligned} & \limsup_n \left( \operatorname{ess\,sup}_{\mu \geq \eta_n} \mathbb{E} \left[ - \int_{\mu}^{\infty} e^{-\alpha s} h_p(s) ds + \int_{\eta_n}^{\mu} \alpha e^{-\alpha t} l dt \right] \right) \\ & = \operatorname{ess\,sup}_{\mu > \eta} \mathbb{E} \left[ - \int_{\mu}^{\infty} e^{-\alpha s} h_p(s) ds + \int_{\eta}^{\mu} \alpha e^{-\alpha t} l dt \right]. \end{aligned}$$

Da  $X$  stetig in Erwartung ist, erhält man die gewünschte Ungleichung.  $Y^l$  ist also oberhalbstetig von rechts in Erwartung und gleichgradig integrierbar, damit gilt nach *Théorème 12* von [8], dass  $Y^l$  oberhalbstetig von rechts ist.

Analog zeigt man, dass  $Y^l$  ebenfalls unterhalbstetig von rechts ist und damit nach [7]-VI-50 rechtsstetige Pfade besitzt.  $\square$

**Bemerkung 2.11.** Ist  $X$  wie in [1] nur unterhalbstetig in Erwartung, so muss  $Y^l$  im Allgemeinen nicht oberhalbstetig in Erwartung sein.

Analog zur Schwellenwertcharakterisierung (1.18) im zeitdiskreten Fall kann man nun den Prozess  $M_p$  wie folgt definieren:

$$(2.19) \quad M_p(\omega, t) \triangleq \inf \{ -l \in \mathbb{R} : Y^l(\omega, t) = X(\omega, t) \} \quad \text{für } (\omega, t) \in \Omega \times [0, \infty].$$

$M_p$  ist nach Lemma 3.12 in [1] ebenfalls optional mit Werten in  $(-\infty, +\infty]$  und nimmt in unserem Fall nur positive Werte an. Das folgende Lemma liefert die gewünschte Regularität von  $M_p$ .

**Lemma 2.12.** *Die Pfade des optionalen Prozesses  $M_p$  sind  $\mathbb{P}$ -f.s. unterhalbstetig von rechts.*

*Beweis.* Nach Proposition 2 in [8] genügt es zu zeigen, dass  $M_p(\mu) \leq \lim_n M_p(\mu_n)$   $\mathbb{P}$ -f.s. für jede Folge von Stoppzeiten  $\mu_n \downarrow \mu$  mit  $\mu_n > \mu$  auf  $\{\mu < \infty\}$  und für die  $\lim_{n \rightarrow \infty} M_p(\mu_n)$   $\mathbb{P}$ -f.s. existiert.

Es sei  $(\mu_n)$  eine solche Folge und  $Z \triangleq \lim_n M_p(\mu_n)$ . Wir müssen zeigen, dass  $Y^{-Z}(\mu) = X(\mu)$ . Dazu sei  $\epsilon > 0$  gegeben. Zu  $\omega \in \Omega$  existiert dann  $n_\epsilon^\omega$ , so dass für alle  $n \geq n_\epsilon^\omega$  gilt

$$\begin{aligned} M_p(\omega, \mu_n(\omega)) &\geq Z(\omega) - \epsilon \text{ und damit} \\ Y^{-(Z(\omega) - \epsilon)}(\omega, \mu_n(\omega)) &= X(\mu_n(\omega)). \end{aligned}$$

$Y^l$  ist stetig in  $l$ , damit erhält man wegen der Rechtsstetigkeit der Prozesse  $X$  und  $Y^l$

$$Y^{-(Z(\omega) - \epsilon)}(\omega, \mu(\omega)) \geq \limsup_{n \rightarrow \infty} Y^{-(Z(\omega) - \epsilon)}(\omega, \mu_n(\omega)) = X(\mu(\omega))$$

Dies gilt für jedes  $\epsilon > 0$ , es folgt

$$Y^{-Z(\omega)}(\omega, \mu(\omega)) = X(\mu(\omega)).$$

Der optionale Prozess  $M_p$  ist damit unterhalbstetig von rechts, nach (2.42) in [9] besitzen seine Pfade dann ebenfalls diese Eigenschaft.  $\square$

## 2.7 Zwei Schranken für die Rendite

Die Darstellung (2.12) setzt die erwartete Rendite einer Strategie in Verbindung mit den unteren Einhüllenden der Gittins-Indexprozesse  $M_p$ . Diese bezeichnen wir von nun an mit

$$\begin{aligned} \underline{M}_p(u, s) &\triangleq \inf_{v \in [u, s]} M_p(v), \\ \underline{M}_p(s) &\triangleq \underline{M}_p(0, s). \end{aligned}$$

Da  $\underline{M}_p(0, s) \leq \underline{M}_p(T_p(t), s)$  gilt für  $s \geq T_p(t)$  und die Prozesse  $e^{-\alpha(t - T_p(t))}$  monoton fallen, folgt aus (2.12)

$$\begin{aligned} \mathbb{E} [\mathcal{R}(T)] &\leq \sum_{p=1}^d \mathbb{E} \left[ \int_0^\infty e^{-\alpha s} \underline{M}_p(0, s) ds \right] + \\ &+ \sum_{p=1}^d \mathbb{E} \left[ \int_0^\infty \mathbb{E} \left[ \int_{T_p(t)}^\infty e^{-\alpha s} \underline{M}_p(0, s) ds \mid \overline{\mathcal{F}}^p(T_p(t)) \right] de^{-\alpha(t - T_p(t))} \right] \\ (2.20) \quad &= \mathbb{E} [\tilde{\mathcal{R}}(T)]. \end{aligned}$$

Dabei ist  $\tilde{\mathcal{R}}(T)$  der Gewinn von  $T$  an einem  $d$ -armigen Banditen mit den fallenden pfadweise rechtsstetigen Renditeprozessen  $\tilde{h}_p(s) = \underline{M}_p(0, s)$ .

**Bemerkung 2.13.** *Im Allgemeinen muss eine Strategie  $T$  für das ursprünglich betrachtete Optimierungsproblem keine Strategie für den Banditen mit den betrachteten fallenden Renditeprozessen sein, wenn man diesen mit seiner kanonischen Informationsstruktur betrachtet. Die Filtration spielt in der Situation fallender rechtsstetiger Renditeprozesse jedoch nur eine untergeordnete Rolle, da die optimale Strategie hier sogar pfadweise optimal in der größeren Menge der antizipierenden Strategien ist.*

**Definition 2.14.** *Ein stochastischer Prozess  $\tilde{T}$  mit Werten in  $[0, \infty)^d$  heißt antizipierende Strategie, wenn er die Wachstumsbedingungen (2.4) und (2.5) erfüllt. Die Menge der antizipierenden Strategien bezeichnen wir mit  $\mathcal{Q}$ .*

Es gilt also für den erwarteten Wert einer Strategie im Optimierungsproblem (2.8)

$$(2.21) \quad \mathbb{E}[\mathcal{R}(T)] \leq \mathbb{E}[\tilde{\mathcal{R}}(T)] \leq \sup_{\tilde{T} \in \mathcal{Q}} \mathbb{E}[\tilde{\mathcal{R}}(\tilde{T})].$$

Eine Strategie  $T$ , für die zweimal die Gleichheit gilt, ist offensichtlich optimal für das Problem (2.8). Wir werden zeigen, dass dies genau für die in Abschnitt 2.9 definierten Indexstrategien erfüllt ist. Existiert eine solche, so muss auch jede andere optimale Strategie Indexstrategie sein.

## 2.8 Fallende Auszahlungsprozesse

Um die zweite Ungleichung in (2.21) zu untersuchen, betrachten wir nun einen Banditen mit fallenden, pfadweise rechtsstetigen Auszahlungsprozessen. Wie im zeitdiskreten Fall bedeutet hier das Verzögern hoher Renditen eines Armes zu Gunsten niedrigerer Renditen eines anderen Armes einen Verlust durch die Diskontierung. Da man also stets den Arm mit der höchsten gegenwärtigen Rendite wählen sollte, liefern Informationen über zukünftige Renditen keinen Vorteil. Die Filtration kann beliebig erweitert werden, ohne die Lösung des Problems zu verändern.

Wir folgen dem Beweis von *Theorem 3.7* in [12]. Die dort genutzte *Synchronization Identity* ermöglicht eine sehr kurze Lösung. In Lemma 2.22 des nächsten Abschnittes werden wir zeigen, dass eine antizipierende Strategie in der Situation fallender rechtsstetiger Renditen genau dann die Synchronization Identity erfüllt, wenn sie stets einen Arm  $p$  mit der größten gegenwärtigen Rendite wählt. Optimale Strategien sind also ganz wie im zeitdiskreten Fall kurzsichtig. Die Formulierung dieser Eigenschaft durch die Synchronization Identity ermöglicht jedoch den kurzen Beweis von El Karoui und Karatzas unseres Lemmas 2.17. Wir fixieren  $\omega \in \Omega$  und erhalten für  $p = 1, \dots, d$  die rechtsstetigen fallenden Abbildungen

$$t \mapsto \underline{M}_p(t), t \in [0, \infty)$$

und deren rechtsstetige Inversen

$$\sigma_p(m) \triangleq \inf\{t \geq 0 : \underline{M}_p(t) \leq m\}, m \in [0, \infty).$$

Die Summe der Inversen bezeichnen wir mit

$$\underline{\tau}(m) \triangleq \sum_{p=1}^d \underline{\sigma}_p(m), m \in [0, \infty).$$

Bis zur Kalenderzeit  $\underline{\tau}(m)$  ist es möglich, den Banditen so zu spielen, dass die Auszahlungsraten nicht unterhalb von  $m$  liegen. Eine optimale Strategie sollte dies auch tun, da anderenfalls Arme mit geringerer Auszahlung früher gespielt werden müssen, was den Gewinn aufgrund der Diskontierung verringert.

Gilt für ein  $p$  nun  $T_p(\underline{\tau}(m)) > \underline{\sigma}_p(m)$ , so wurde Arm  $p$  betätigt, obwohl seine Rendite zwischenzeitlich  $m$  erreicht oder unterschritten hatte und Arme mit höherer Rendite zur Verfügung standen. Falls umgekehrt  $T_p(\underline{\tau}(m)) < \underline{\sigma}_p(m)$  gilt, so wurde ein von  $p$  verschiedener Arm mit geringerer Rendite als  $m$  betätigt, obwohl Arm  $p$  eine höhere Rendite versprach. Es sollte also  $T_p(\underline{\tau}(m)) = \underline{\sigma}_p(m)$  gelten.

**Definition 2.15.** ([12]) Eine antizipierende Strategie  $T$  für den betrachteten fallenden Banditen erfüllt die Synchronization Identity, wenn sie die folgenden Gleichungen erfüllt:

$$(2.22) \quad T_p(\underline{\tau}(m)) = \underline{\sigma}_p(m) \quad \forall m \in [0, \infty), p = 1, \dots, d.$$

**Bemerkung 2.16.** Man zeigt leicht, dass (2.22) äquivalent zu

$$(2.23) \quad \sum_{p=1}^d T_p(t) \wedge \underline{\sigma}_p(m) = t \wedge \underline{\tau}(m) \quad \forall m, t \in [0, \infty),$$

ist. Siehe dazu auch Proposition 3.8 in [12]. Wir werden beide Darstellungen nutzen, (2.23) zum Beweis von Lemma 2.17 zur Optimalität in der Situation fallender Renditeprozesse, (2.22) um die Verbindung der Synchronization Identity zu den im nächsten Abschnitt definierten Indexstrategien aufzuzeigen.

**Lemma 2.17.** Eine antizipierende Strategie  $\hat{T}$  maximiert die Rendite eines mehrarmigen Banditen mit fallenden rechtsstetigen Renditeprozessen  $\underline{M}_p$  genau dann, wenn  $\hat{T}$  (2.23) erfüllt.

*Beweis.* Wir setzen für  $\lambda, t \in [0, \infty)$

$$\begin{aligned} A_p(t, \lambda, T) &\triangleq T_p(t) \wedge \underline{\sigma}_p(\lambda), \\ A(t, \lambda, T) &\triangleq \sum_{p=1}^d A_p(t, \lambda, T) \leq t \wedge \underline{\tau}(\lambda). \end{aligned}$$

Damit gilt für jede antizipierende Strategie  $T$

$$\begin{aligned}
\mathcal{R}(T) &= \sum_{p=1}^d \int_0^\infty e^{-\alpha t} \underline{M}_p(T_p(t)) dT_p(t) \\
&= \sum_{p=1}^d \int_0^\infty e^{-\alpha t} \left( \int_0^\infty 1_{(\lambda < \underline{M}^p(T_p(t)))} d\lambda \right) dT_p(t) \\
&= \sum_{p=1}^d \int_0^\infty \left( \int_0^\infty e^{-\alpha t} 1_{(\lambda < \underline{M}^p(T_p(t)))} dT_p(t) \right) d\lambda \\
&= \sum_{p=1}^d \int_0^\infty \left( \int_0^\infty e^{-\alpha t} dA_p(t, \lambda, T) \right) d\lambda \\
&= \sum_{p=1}^d \int_0^\infty \left( \int_0^\infty \alpha e^{-\alpha t} A_p(t, \lambda, T) dt \right) d\lambda \\
&= \int_0^\infty \left( \int_0^\infty \alpha e^{-\alpha t} A(t, \lambda, T) dt \right) d\lambda \\
&\leq \int_0^\infty \left( \int_0^\infty \alpha e^{-\alpha t} (t \wedge \underline{\tau}(\lambda)) dt \right) d\lambda \\
&= \frac{1}{\alpha} \int_0^\infty 1 - e^{-\alpha \underline{\tau}(\lambda)} d\lambda.
\end{aligned}$$

Der letzte Term ist unabhängig von der gewählten antizipierenden Strategie und liefert eine obere Schranke für den erzielbaren Gewinn. Eine antizipierende Strategie  $T$  erzielt genau dann diesen maximalen Gewinn, wenn Gleichheit gilt, also genau dann wenn

$$(2.24) \quad A(t, \lambda, T) = t \wedge \underline{\tau}(\lambda) \quad \forall \lambda, t \in [0, \infty).$$

Dies ist die Synchronization Identity (2.23) für  $T$ . □

**Bemerkung 2.18.** *In der Situation allgemeiner Renditeprozesse ist die Synchronization Identity (2.23) notwendig, aber nicht hinreichend für die Optimalität. Man zeigt leicht, dass die nicht optimale Strategie  $T$  aus dem unten folgenden Beispiel 2.19 die Eigenschaft (2.23) besitzt.*

## 2.9 Indexstrategien

In diesem Abschnitt definieren wir Indexstrategien. Wir folgen dabei Kaspi und Mandelbaum [16], die zum ersten Mal die wichtige *Exkursionseigenschaft* (2.27) formulieren. Die von El Karoui und Karatzas [12] konstruierte optimale Strategie, die wir im nächsten Abschnitt betrachten, besitzt ebenfalls diese Eigenschaft.

Im zeitdiskreten Fall sind genau jene Strategien optimal, welche einen Arm nur dann betätigen, wenn sein Gittins-Index maximal unter denen aller Arme ist. Steigt der Index

eines solchen Armes nach einer Betätigung, so ist er größer als der jedes anderen Armes. Eine Indexstrategie wird dann ausschließlich diesen Arm betätigen, bis sein Index den anfänglichen Wert wieder erreicht. Insbesondere ist dieses Vorgehen notwendig für die Optimalität einer Strategie.

In der zeitstetigen Situation besitzen Strategien, die nur Arme mit maximalem Gittins-Index betätigen, nicht immer diese Eigenschaft. Werden zwei Arme gleichzeitig betätigt, so können die Gittins-Indizes beider Arme simultan steigen. Diese Situation kann in diskreter Zeit nicht eintreten, muss aber im Hinblick auf das folgende Beispiel berücksichtigt werden. Dieses zeigt, dass es im zeitstetigen Fall nicht mehr hinreichend für die Optimalität einer Strategie ist, ausschließlich Arme mit maximalem Gittins-Index zu aktivieren.

**Beispiel 2.19.** *Wir betrachten einen zweiarmigen Banditen mit den Filtrationen  $\mathcal{F}^i$ ,  $i = 1, 2$ . Es sei  $\mathcal{F}^i(s) = (\Omega, \emptyset)$  für  $s < 1$  und  $\mathcal{F}^i(s) = \sigma(A^i)$  für  $s > 1$  und unabhängige  $A^i \subset \Omega$  mit  $\mathbb{P}[A^i] = \frac{1}{2}$ .*

*Gegeben seien weiter die Renditeprozesse*

$$(2.25) \quad h_i(s) = \begin{cases} 2 & s < 1 \\ 3 \cdot 1_{A^i} + 1_{\Omega \setminus A^i} & s \geq 1. \end{cases}$$

*Mit Hilfe der Forward Induction (2.16) kann man die Gittins-Indizes  $M_i(s)$  für  $s < 1$  berechnen. Man zeigt leicht, dass die Stoppzeiten, die die wesentlichen Suprema realisieren, die Form  $\tau_i = \infty \cdot 1_{A^i} + 1_{\Omega \setminus A^i}$  besitzen. Damit ergibt sich für die Gittins-Indizes*

$$M_i(s) = \frac{4e^{-\alpha s} - e^{-\alpha}}{2e^{-\alpha s} - e^{-\alpha}} \text{ für } 0 \leq s < 1.$$

*$M_i(s)$  ist stetig und wachsend für  $s \in [0, 1)$  mit Werten im Intervall  $(2, 3)$  und  $M_i(1-) = 3$ . Für  $s \geq 1$  ist  $M_i(s) = h_i(s)$ .*

*Wir definieren die Strategie  $T$  wie folgt.  $T$  soll bis zum Zeitpunkt  $t = 2$  beide Arme mit Rate  $\frac{1}{2}$  betätigen, anschließend Arm 1 wählen, falls dieser dann die Rendite 3 besitzt, sonst Arm 2. Der erwartete Gewinn unter dieser Strategie beträgt*

$$\begin{aligned} \mathbb{E}[\mathcal{R}(T)] &= \int_0^2 2e^{-\alpha t} dt + \frac{3}{4} \int_2^\infty 3e^{-\alpha t} dt + \frac{1}{4} \int_2^\infty e^{-\alpha t} dt \\ &= \frac{1}{\alpha} \left( 2 + \frac{1}{2} e^{-2\alpha} \right). \end{aligned}$$

*Offensichtlich wählt  $T$  stets einen Arm mit maximalem Gittins-Index.*

*Wir betrachten die Strategie  $S$ , die bis  $s = 1$  ausschließlich Arm 1 betätigt. Falls dessen Rendite sich dann auf 3 erhöht, wählt  $S$  nur noch diesen Arm. Anderenfalls betätigt  $S$*

nur noch Arm 2. Der erwartete Wert ist

$$\begin{aligned}\mathbb{E}[\mathcal{R}(S)] &= \int_0^1 2e^{-\alpha t} dt + \frac{1}{2} \int_1^\infty 3e^{-\alpha t} dt + \frac{1}{2} \int_1^2 2e^{-\alpha t} dt \\ &\quad + \frac{1}{4} \int_2^\infty e^{-\alpha t} dt + \frac{1}{4} \int_2^\infty 3e^{-\alpha t} dt \\ &= \frac{1}{\alpha} \left( 2 + \frac{1}{2} e^{-\alpha} \right)\end{aligned}$$

Damit ist  $T$  offenbar nicht optimal.

**Bemerkung 2.20.** Betrachtet man dasselbe Problem mit einer diskreten Zeitstruktur, so wird eine - in dieser Situation stets optimale - Indexstrategie ebenso wie  $S$  zunächst ausschließlich einen Arm betätigen. Die Indexstrategie  $T$  ist - im Gegensatz zu  $S$  - nicht der Grenzwert einer Folge von optimalen Strategien auf einer Folge feiner werdender diskreter Zeitstrukturen.

Überlegungen zur Approximation von zeitstetigen durch diskrete Strategien findet man in [20], Section 6.

Unter Umständen ist es offenbar vorteilhaft, ausschließlich einen Arm zu betätigen. Diese Beobachtung motiviert die folgende

**Definition 2.21.** Eine Strategie  $T$  heißt Indexstrategie, wenn sie die folgenden Bedingungen erfüllt.

(i)  $T$  folgt dem größten  $M_p$ :

$$(2.26) \quad \sum_{p=1}^d \int_0^\infty 1_{\{M_p(T_p(t)) < \max_{1 \leq k \leq d} M_k(T_k(t))\}} dT_p(t) = 0.$$

(ii)  $T$  besitzt die Exkursionseigenschaft:

$$(2.27) \quad \sum_{p=1}^d \int_0^\infty 1_{\{M_p(T_p(t)) > \underline{M}_p(T_p(t))\}} d(t - T_p(t)) = 0.$$

Eine Indexstrategie wählt also stets einen Arm mit maximalem Gittins-Index und betätigt während einer Exkursion des Indexprozesses  $M_p$  eines Armes von seinem laufenden Minimum  $\underline{M}_p$  diesen Arm mit voller Intensität. Beginnen die Indizes zweier Arme gleichzeitig eine Exkursion, so muss sich eine Indexstrategie für einen der beiden Arme entscheiden und für die Dauer der Exkursion ausschließlich diesen Arm betätigen.

Die Strategie  $S$  in Beispiel 2.19 ist eine Indexstrategie. Die Strategie  $T$  folgt dem größten  $M_p$ , ist allerdings keine Indexstrategie. Denn für  $p = 1, 2$  und  $t \in [0, 2)$  ist  $d(t - T_p(t)) = \frac{1}{2} dt$  und es gilt  $M_p(T_p(t)) > \underline{M}_p(0, T_p(t))$ , die Gittins-Indizes beider Arme sind also auf einer

Exkursion.

Aus  $S$  kann man durch Vertauschung der Reihenfolge der Betätigungen der beiden Arme eine weitere Indexstrategie konstruieren. Im Allgemeinen kann es also mehrere Indexstrategien geben.

Wir haben im letzten Abschnitt die von El Karoui und Karatzas [12] eingeführte Synchronization Identity genutzt. Weiter werden dort Strategien betrachtet, die *dem größten  $\underline{M}_p$  folgen*, also nur Arme betätigen, deren untere Einhüllende  $\underline{M}_p$  der Gittins-Indexprozesse  $M_p$  maximal ist unter denen der verschiedenen Arme. Das folgende Lemma verdeutlicht den Zusammenhang dieser Begriffe untereinander und zu den von uns betrachteten Indexstrategien nach Mandelbaum.

**Lemma 2.22.** *Für jede Strategie  $T$  sind die beiden folgenden Eigenschaften äquivalent:*

(i)  $T$  folgt dem größten  $\underline{M}_p$ :

$$(2.28) \quad \sum_{p=1}^d \int_0^\infty 1_{\{\underline{M}_p(T_p(t)) < \max_{1 \leq k \leq d} \underline{M}_k(T_k(t))\}} dT_p(t) = 0.$$

(i)'  $T$  erfüllt die Synchronization Identity:

$$(2.29) \quad T_p(\tau(m)) = \underline{\sigma}_p(m) \quad \forall m \in [0, \infty), p = 1, \dots, d.$$

Erfüllt eine Strategie  $T$  die Eigenschaft (2.26), so auch (2.28) und (2.29).

Die Umkehrung gilt, falls  $T$  zusätzlich (2.27) erfüllt.

*Beweis.* Es sei  $T$  eine Strategie.

(2.26)  $\Rightarrow$  (2.28) : Es sei  $t \geq 0$  mit  $\underline{M}_p(T_p(t)) < \max_{1 \leq k \leq d} \underline{M}_k(T_k(t))$ . Wir nehmen an, für  $t$  gelte  $M_p(T_p(t)) = \max_{1 \leq k \leq d} M_k(T_k(t))$ . Dann ist  $M_p(T_p(t)) > \underline{M}_p(T_p(t))$ , also existiert ein  $s < t$  mit

$$M_p(T_p(s)) < \max_{1 \leq k \leq d} \underline{M}_k(T_k(t)) \leq \max_{1 \leq k \leq d} M_k(T_k(r))$$

für  $r \in [s, t]$ . Insbesondere ist  $T_p(t) > T_p(s)$ . Wegen (i) gilt

$$0 = \int_s^t 1_{\{M_p(T_p(u)) < \max_{1 \leq k \leq d} M_k(T_k(u))\}} dT_p(u) = \int_s^t 1 dT_p(u) = T_p(t) - T_p(s).$$

Dies ist ein Widerspruch, es folgt  $M_p(T_p(t)) < \max_{1 \leq k \leq d} M_k(T_k(t))$  und damit die Behauptung.

(2.28)  $\Rightarrow$  (2.29) : Wir nehmen an,  $T$  erfülle nicht (2.29). Dann existieren  $i, j \in \{1, \dots, d\}$  und  $m \in [0, \infty)$  mit  $T_i(\tau(m)) < \underline{\sigma}_i(m)$  und  $T_j(\tau(m)) > \underline{\sigma}_j(m)$ .

Weiter gibt es  $s < \tau(m)$  mit  $T_j(s) = \underline{\sigma}_j(m)$  und  $T_j(\tau(m)) - T_j(s) > 0$ . Allerdings gilt wegen  $\underline{M}_j(T_j(r)) \leq m < \underline{M}_i(T_i(\tau(m)))$  für  $r \in [s, t]$  und (i)'  $T_j(\tau(m)) - T_j(s) = 0$ .

(2.29)  $\Rightarrow$  (2.28) : Dies entspricht *Proposition 3.12* in [12].

Wir nehmen an, (2.28) sei nicht erfüllt. Wegen der pfadweisen Unterhalbstetigkeit von rechts der Prozesse  $M_p$  existieren  $p, k \in \{1, \dots, d\}$ ,  $m \geq 0$  und  $0 \leq s < t < \infty$  mit  $T_p(t) - T_p(s) > 0$  und  $\underline{M}_p(T_p(r)) < m < \underline{M}_k(T_k(r))$  für  $r \in [s, t]$ .

Gilt nun  $t < \tau(m)$ , so ist  $T_p(\tau(m)) > \underline{\sigma}_p(m)$ , im Widerspruch zu (i)". Falls  $t \geq \tau(m)$ , so gilt  $T_k(\tau(m)) < m$ , ebenfalls ein Widerspruch.

(2.29) + (2.27)  $\Rightarrow$  (2.26) : Angenommen, (2.26) gelte nicht. Wie oben erhält man aus der Regularität der Pfade von  $M_p$  die Existenz von  $p, k \in \{1, \dots, d\}$ ,  $m \geq 0$  und  $0 \leq s < t < \infty$  mit  $T_p(t) - T_p(s) > 0$  und  $M_p(T_p(r)) < m < M_k(T_k(r))$  für  $r \in [s, t]$ .

Im Fall  $t < \tau(m)$  liefert wieder  $T_p(\tau(m)) > \underline{\sigma}_p(m)$  einen Widerspruch zu (i)". Ebenfalls einen Widerspruch liefert die Annahme  $t \geq \tau(m)$ ,  $\underline{M}_k(T_k(t)) > m$ . Einzig der Fall  $t \geq \tau(m^*)$ ,  $\underline{M}_k(T_k(t)) \leq m < M_k(T_k(t))$  bleibt möglich.

Damit existiert für  $t \geq 0$  mit  $M_p(T_p(t)) < \max_{1 \leq k \leq d} M_k(T_k(t))$  ein Arm  $k \neq p$ , dessen Gittins-Indexprozess  $M_k$  sich im Zeitpunkt  $T_k(t)$  auf einer Exkursion von seinem laufenden Minimum  $\underline{M}_k$  befindet.

Da  $T$  nach Voraussetzung die Exkursionseigenschaft besitzt, wächst  $T_k$  zur Zeit  $t$  mit Rate 1 und  $T_p$  ändert sich nicht. Formal bedeutet das

$$\begin{aligned} & \int_0^\infty 1_{\{M_p(T_p(t)) < \max_{1 \leq k \leq d} M_k(T_k(t))\}} dT_p(t) \\ & \leq \sum_{k \neq p} \int_0^\infty 1_{\{M_k(T_k(t)) > \underline{M}_k(T_k(t))\}} dT_p(t) \\ & \leq \sum_{k \neq p} \int_0^\infty 1_{\{M_k(T_k(t)) > \underline{M}_k(T_k(t))\}} d(t - T_k(t)) = 0. \end{aligned}$$

Das Gleichheitszeichen folgt aus (2.27), damit ist die Behauptung bewiesen.  $\square$

**Bemerkung 2.23.** Für die Folgerung (2.29)  $\Rightarrow$  (2.26) ist die Exkursionseigenschaft (2.27) hinreichend, aber nicht notwendig. Dies illustriert Strategie  $T$  aus Beispiel 2.19.

**Bemerkung 2.24.** Kaspı und Mandelbaum [17] weisen darauf hin, dass durch die Definition des Begriffs der Indexstrategie nicht klar ist, für welchen der Arme sich eine solche entscheidet, wenn die Indizes mehrerer Arme gleichzeitig eine Exkursion beginnen. Sie bemerken, dass die Synchronization Identity von El Karoui und Karatzas [12] dieses Problem nicht löst. Die Bedeutung der Synchronization Identity ist mit Lemma 2.22 klar und hat tatsächlich nichts mit diesen Entscheidungen zu tun.

Konstruiert man eine Indexstrategie, so muss sich diese in den beschriebenen Situationen

natürlich für einen Arm entscheiden. Die von El Karoui und Karatzas [12] konstruierte Strategie, die wir im nächsten Abschnitt genauer betrachten werden, löst das erwähnte Problem durch die Verwendung eines der von Mandelbaum [20] definierten Prioritätschemata.

**Bemerkung 2.25.** El Karoui und Karatzas [11] definieren (Def 7.1) Strategien vom Index-Typ als jene, die dem größten  $\underline{M}_p$  folgen. Sie beweisen die Optimalität solcher Strategien (Thm 8.1) unter der Voraussetzung der Quasi-Linksstetigkeit der Filtrationen. Diese ist beispielsweise erfüllt, wenn die Dynamik der Arme durch Diffusionsprozesse bestimmt wird.

Gilt nicht zusätzliche die pfadweise Unterhalbstetigkeit der Indexprozesse von links, so scheint es uns problematisch, nur die laufenden Minima der Indexprozesse zu berücksichtigen. Betrachtet man eine Variante des Beispiels 2.19, in welchem die Sprungzeiten der Auszahlungsprozesse nicht vorhersehbar sind, so wäre mit dem obigen Resultat die folgende Strategie optimal. Man betätigt beide Arme mit Rate  $\frac{1}{2}$  bis einer der Auszahlungsprozesse nach unten springt, anschließend den anderen Arm. Dies ist jedoch keine Indexstrategie und somit nach dem unten folgenden Gittins-Indextheorem nicht optimal.

**Bemerkung 2.26.** Kaspi und Mandelbaum [16],[17], auf die die oben eingeführten Indexstrategien zurückgehen, formulieren deren Eigenschaften nicht durch Integraldarstellungen wie (2.26) und (2.28), sondern in Abhängigkeit der Pfade der Abbildungen  $t \mapsto T_p(t)$ . Wir geben hier ihre Formulierung wieder:

Eine Strategie  $T$  heißt Indexstrategie, wenn jedes  $T_p = (T_p(t))$  nur dann rechtsseitig wächst, wenn gilt

$$M_p(T_p(t)) = \max_{i=1,\dots,d} M_i(T_i(t))$$

und alle Zeitpunkte  $t$ , in denen  $T_p(t)$  linksseitig und nicht rechtsseitig wächst oder rechtsseitig mit Rate kleiner als 1 wächst, im Abschluss der folgenden Menge liegen

$$(2.30) \quad \{s \geq 0 : M_p(s) = \underline{M}_p(s)\}.$$

Dies ist die zufällige Menge von Zeitpunkten, in denen  $M_p$  gleich seinem laufenden Minimum ist.

Diese Definition entspricht den Darstellungen (2.26) und (2.27), die für unseren Beweis allerdings zweckmäßiger erscheinen.

## 2.10 Existenz von Indexstrategien

Die Konstruktion einer Indexstrategie erscheint zunächst nicht schwierig, man aktiviert stets einen Arm mit maximalem Gittins-Index und wählt während der Exkursion des

Indexprozesses eines Armes von seinem laufenden Minimum ausschließlich diesen Arm. Allerdings ist es denkbar, dass während der Betätigung zweier Arme beide Gittins-Indexprozesse eine Exkursion durch einen Sprung nach oben beginnen. Unter einer Indexstrategie dürfen solche Situationen nicht auftreten, da diese sonst nach dem Sprung beide Arme mit voller Intensität betätigen müsste.

Wir werden auf der Grundlage der Konstruktion von El Karoui und Karatzas [12], die allerdings die pfadweise Unterhalbstetigkeit von links der Indexprozesse voraussetzt, die Existenz einer Indexstrategie beweisen. Um auf diese Annahme verzichten zu können, müssen wir die oben beschriebenen Situationen gleichzeitiger Sprünge untersuchen.

Für  $p \in \{1, \dots, d\}$  setzen wir

$$\begin{aligned}\sigma_p(m) &\triangleq \inf\{u \geq 0 : M_p(u) \leq m\}, m \in [0, \infty), \\ \tau(m) &\triangleq \sum_{p=1}^d \sigma_p(m), m \in [0, \infty), \\ N(u) &\triangleq \inf\{m \geq 0 : \tau(m) \leq u\}, 0 \leq u < \infty, \\ \mathcal{D} &\triangleq \{u \geq 0 : \tau(N(u)-) > u\}.\end{aligned}$$

Mit Hilfe dieser Bezeichnungen lassen sich die Situationen beschreiben, die unter einer Strategie, welche die Synchronization Identity erfüllt, auftreten können. Wir unterscheiden für  $t \geq 0$  wie El Karoui und Karatzas die folgenden Fälle, einzig IIb ist neu.

**Fall I.**  $t \in [0, \infty) \setminus \mathcal{D}$ . Dann gilt  $\tau(N(t)-) = t$  und wir setzen

$$(2.31) \quad T_i(t) \triangleq \sigma_i(N(t)-).$$

Erfüllt  $T$  die Synchronization Identity, so sind dies die Zeitpunkte  $t$ , in denen die Gittins-Indizes aller Arme gleich ihren laufenden Minima sind und weiter fallen. Eine Indexstrategie sollte die Arme so betätigen, dass die Indizes gleichermaßen fallen. Dieses Vorgehen beschreibt (2.31).

Für  $t \in [0, \infty) \setminus \mathcal{D}$  gilt  $\tau(N(t)-) = t$  und damit

$$\sum_{p=1}^d T_p(t) = \sum_{p=1}^d \sigma_p(N(t)-) = \tau(N(t)-) = t.$$

**Fall II.**  $t \in \mathcal{D}$ . Dann ist  $t \in [\tau(N(t)+), \tau(N(t)-))$  und wir unterscheiden weiter:

**Fall IIa.**  $t \in [\tau(N(t)), \tau(N(t)-))$ . Im Zeitpunkt  $\tau(N(t))$  haben die Gittins-Indizes aller Arme das Niveau  $N(t)$  erreicht. Man folgt hier einem der von Mandelbaum [20] definierten *Prioritätsschemata* und spielt zunächst den ersten Arm bis sein Index unter  $N(t)$  fällt,

anschließend den zweiten Arm, und so weiter. Da  $T$  hier stets nur einen Arm aktiviert, können nicht mehrere Exkursionen gleichzeitig beginnen. El Karoui und Karatzas geben eine explizite Darstellung dieser Vorgehensweise an. Man setzt

$$\begin{aligned} y_0 &\triangleq \tau(N(t)), \\ y_i &\triangleq y_{i-1} - \Delta\sigma_i(N(t)) \quad i = 1, \dots, d, \end{aligned}$$

wobei  $\Delta\sigma_i(N(t)) \triangleq \sigma_i(N(t)) - \sigma_i(N(t)-)$  und erhält die disjunkte Zerlegung

$$[\tau(N(t)), \tau(N(t)-)) = \cup_{i=1}^d [y_{i-1}, y_i).$$

Nun wählt man das eindeutige Intervall  $[y_{k-1}, y_k)$ , in dem  $t$  enthalten ist. Zu diesem Zeitpunkt hat man bereits die ersten  $k - 1$  Arme gespielt und wählt jetzt Arm  $k$ :

$$(2.32) \quad T_i(t) \triangleq \begin{cases} \sigma_i(N(t)-) & i = 1, \dots, k-1 \\ \sigma_i(N(t)) + t - y_{i-1} & i = k \\ \sigma_i(N(t)) & i = k+1, \dots, d. \end{cases}$$

**Fall IIb.**  $t \in [\tau(N(t)+), \tau(N(t))]$ . In diesem Fall gab es nach einem Intervall vom Typ I im Zeitpunkt  $\tau(N(t)+)$  offenbar einen Arm  $i^*$ , dessen Gittins-Index  $M_{i^*}$  nach oben gesprungen ist und eine Exkursion begonnen hat. Hier ist das oben beschriebene problematische Ereignis des Sprunges eines zweiten Armes denkbar. Dass im Zeitpunkt  $\tau(N(t)+)$  tatsächlich  $\mathbb{P}$ -f.s. der Gittins-Index höchstens eines Armes nach oben springen kann, sagt die folgende Proposition, die wir im nächsten Abschnitt interpretieren und beweisen werden.

**Proposition 2.27.** *Für  $p = 1, \dots, d$  sei  $J_p \triangleq \{m \geq 0 : \sigma_p(m+) < \sigma_p(m)\}$  die zufällige Menge von Niveaus, von denen der Gittins-Index  $M_p$  unmittelbar nach Erreichen eines neuen Minimums nach oben springt. Dann gilt für  $p \neq k$*

$$(2.33) \quad \mathbb{P}[J_p \cap J_k \neq \emptyset] = 0.$$

Da es keinen Sprung des Index eines weiteren Armes nach oben gibt, ist der des Armes  $i^*$  maximal und auf einer Exkursion. Entsprechend betätigt man ausschließlich diesen Arm, bis sein Index tatsächlich  $N(t)$  erreicht und definiert für  $\mathbb{P}$ -f.a.  $\omega$ :

$$(2.34) \quad T_i(t) \triangleq \begin{cases} \sigma_i(N(t)+) + t - \tau(N(t)+) & i = i^* \\ \sigma_i(N(t)+) & i \neq i^* \end{cases}$$

$T$  erfüllt nach Konstruktion die Synchronization Identity sowie die Messbarkeitsbedingung (2.6). Da  $T$  ebenfalls die Exkursionseigenschaft besitzt, gilt mit Lemma 2.22 das folgende

**Theorem 2.28.** *Die beschriebene Konstruktion liefert eine Indexstrategie  $T$ .*

**Bemerkung 2.29.** *Es gilt die Beziehung  $\sigma_p(m+) = \underline{\sigma}_p(m)$ . Im Allgemeinen sind die Pfade der Abbildungen  $m \mapsto \sigma_p(m)$  weder links- noch rechtsstetig. Sind die Pfade der Indexprozesse nicht unterhalbstetig von links, so kann  $\sigma_p(m) > \sigma_p(m+)$  gelten. Beispiel 2.34 zeigt, dass solche Situationen existieren.*

*Dies entspricht dem oben beschriebenen Fall IIb, welcher nicht in der Konstruktion von El Karoui und Karatzas [12] berücksichtigt wird. Aus demselben Grund muss die in Fall II ihrer Konstruktion betrachtete Stoppzeit  $\tau(N(t))$  durch  $\tau(N(t)+)$  ersetzt werden. Dies entspricht der Formulierung unseres Falls IIa.*

**Bemerkung 2.30.** *Um weitere Indexstrategien zu erhalten, kann man im Fall IIa andere Prioritätsschemata als das angegebene  $\Pi(p) = p$  verwenden.*

Kaspi und Mandelbaum [16] konstruieren ebenfalls eine Indexstrategie, hier spielen die erwähnten Prioritätsschemata eine zentrale Rolle.

**Definition 2.31** ([20]). *Ein Prioritätsschema  $\Pi$  ist eine Permutation von  $\{1, \dots, d\}$ . Eine dem größten  $M_p$  folgende Strategie  $T$  folgt dem statischen Prioritätsschema  $\Pi$ , wenn  $T_p$  mit  $\Pi(p) > \Pi(k)$  nur dann in  $t$  wächst, wenn der Index des priorisierten Armes  $k$  in unmittelbarer Zukunft fällt:*

$$(2.35) \quad \underline{M}_k(T_k(t)) > \underline{M}_k(u) \quad \forall u > T_k(t).$$

Zur Konstruktion einer Indexstrategie ist ein statisches Prioritätsschema nützlich, da es verhindert, dass die Gittins-Indizes zweier Arme gleichzeitig eine Exkursion beginnen. Im Allgemeinen ist eine solche Priorität jedoch nicht notwendig für die Optimalität einer Strategie.

Wir möchten an dieser Stelle auf zwei Punkte der Konstruktion von Kaspi und Mandelbaum hinweisen.

Sind  $M_p$  nicht Index-, sondern nur beliebige càdlàg Prozesse, wie von Kaspi und Mandelbaum vorausgesetzt, so liefert das in [16] zitierte Theorem von Walsh [26] nicht unbedingt das angegebene Prioritätsschema. Ein Gegenbeispiel liefern die Funktionen

$$\begin{aligned} M_1(t) &\triangleq (2-t) \cdot 1_{\{t < 1\}} + 3 \cdot 1_{\{t \geq 1\}}, \\ M_2(t) &\triangleq 2 \cdot 1_{\{t < 1\}} + 1 \cdot 1_{\{t \geq 1\}}. \end{aligned}$$

Wegen Proposition 2.32 ist  $M_1$  jedoch kein Indexprozess. Möglicherweise gilt die Aussage für Indexprozesse, allerdings müssen dabei mögliche Sprünge der Indexprozesse berücksichtigt werden.

Kaspi und Mandelbaum schließen, dass eine dem größten  $\underline{M}_p$  und einem statischen Prioritätsschema folgende Strategie auch dem größten  $M_p$  folgt. Das zitierte Theorem 12 aus [20] setzt allerdings stetige Rendite- und Indexprozesse voraus. Bei einer Verallgemeinerung dieses Theorems auf die Situation von [16] spielt Proposition 2.27 ebenfalls eine Rolle.

## 2.11 Positive Antizipation der Gittins-Indexprozesse

Der Gittins-Index eines Armes bewertet die zukünftigen Gewinne dieses Arms mit Hilfe eines Stoppproblems. Sollten die Auszahlungen geringer ausfallen als erwartet, beendet man die Betätigung des Armes. Unter dieser Betrachtungsweise misst der Index nur die *Chance auf höhere Gewinne*. Dass diese Erklärung tatsächlich sinnvoll ist und der Gittins-Index die Möglichkeit höherer Gewinne in der näheren Zukunft antizipiert, zeigt die folgende Proposition von Kaspi und Mandelbaum [17].

Dazu seien für jeden der Arme die folgenden zufälligen Größen definiert, die Bezeichnung des Armes  $p \in \{1, \dots, d\}$  wird dabei vernachlässigt.

$$\begin{aligned}\mathcal{M} &\triangleq \text{clos}\{t \geq 0 : M(t) = \underline{M}(t)\}, \\ D_t &\triangleq \inf\{u > t : u \in \mathcal{M}\}, \\ G &\triangleq \{t > 0 : D_{t-} = t, D_t > t\}, \\ N &\triangleq \{t \in G : M(t) > \underline{M}(t-)\}.\end{aligned}$$

$\mathcal{M}$  ist der Abschluss der Menge, auf der der Index gleich seiner unteren Einhüllenden ist.  $D_t$  ist der erste Zeitpunkt nach  $t$ , in dem der Index - wegen Unterhalbstetigkeit von rechts - gleich seiner unteren Einhüllenden ist. Die Menge  $G$  besteht aus den Zeitpunkten, in denen der Index eine Exkursion beginnt und  $N$  aus solchen, in denen dies durch einen Sprung nach oben geschieht.

Besteht, wie in Beispiel 2.19, in einem vorhersehbaren Zeitpunkt  $\tau$  die Möglichkeit auf höhere zukünftige Auszahlungen, so bedeutet die erwähnte vollständige Antizipation hoher Gewinne, dass bei Eintreten eines solchen Ereignisses der Indexprozess in  $\tau$  nicht nach oben springen kann. Dies würde bedeuten  $\tau \notin N$ . Tatsächlich gilt die folgende

**Proposition 2.32** ([17] Prop.10). *Für jede vorhersehbare  $(\overline{\mathcal{F}}^p(D_t))_{t \geq 0}$ -Stoppzeit  $\tau$  gilt  $\mathbb{P}[\tau \in N] = 0$ .*

Mit Hilfe dieser Aussage kann man die zur Konstruktion einer Indexstrategie wichtige Proposition 2.27 beweisen. Dazu argumentieren wir wie Kaspi und Mandelbaum im Beweis von *Theorem 1* von [17].

Es sei  $m \geq 0$ , dann ist der zufällige Zeitpunkt  $\sigma_p(m+)$  eine vorhersehbare  $(\overline{\mathcal{F}}^p(D_t))$ -Stoppzeit mit der ankündigenden Folge  $\sigma_p(m + \frac{1}{n})$ . Nach Proposition 2.32 gilt damit  $\mathbb{P}$ -f.s.  $M_p(\sigma(m+)) = \underline{M}_p(\sigma(m+)) = m$ . Insbesondere ist damit die Abbildung  $m \mapsto \sigma_p(m)$  in jedem  $m$   $\mathbb{P}$ -f.s. rechtsstetig.

**Bemerkung 2.33.** *Dass dies nicht pfadweise gelten muss, zeigt Beispiel 2.34, insbesondere ist die Formulierung von Lemma 2.2 in [11] selbst im Fall quasi linksstetiger Filtrationen nicht pfadweise zu verstehen.*

**Beweis von Proposition 2.27.** Für jedes  $\omega \in \Omega$  ist die Menge  $J_p(\omega)$  abzählbar. Es gelte also  $J_p = \cup_{n \in \mathbb{N}} \{Y_p^n\}$ . Damit gilt wegen Unabhängigkeit der Renditeprozesse

$$\begin{aligned}
\mathbb{P}[J_p \cap J_k \neq \emptyset] &= \mathbb{P}[\cup_{n,m \in \mathbb{N}} Y_p^n = Y_k^m] \\
&\leq \sum_{n,m \in \mathbb{N}} \mathbb{P}[Y_p^n = Y_k^m] \\
&= \sum_{n,m \in \mathbb{N}} \int_0^\infty \int_0^\infty 1_{y_p^n = y_k^m} d\mathbb{P}_{(Y_p^n, Y_k^m)}(y_p^n, y_k^m) \\
&= \sum_{n,m \in \mathbb{N}} \int_0^\infty \int_0^\infty 1_{y_p^n = y_k^m} d\mathbb{P}_{Y_p^n}(y_p^n) d\mathbb{P}_{Y_k^m}(y_k^m) \\
&= \sum_{n,m \in \mathbb{N}} \int_0^\infty \mathbb{P}[Y_p^n = y_k^m] d\mathbb{P}_{Y_k^m}(y_k^m) \\
&= 0.
\end{aligned}$$

□

Ereignisse zu nicht vorhersehbaren Stoppzeiten kann der Gittins-Index nicht antizipieren und damit müssen die Pfade der Indexprozesse im Allgemeinen nicht unterhalbstetig von links sein. Dies illustriert das folgende

**Beispiel 2.34.** *Es sei  $\xi$  eine exponentialverteilte Zufallsvariable mit  $P_\xi(dx) = \lambda e^{-\lambda x} dx$  für ein  $\lambda > 0$ . Weiter sei  $A \in \mathcal{F}$  unabhängig von  $\xi$  mit  $\mathbb{P}[A] = \frac{1}{2}$  und  $(\mathcal{F}(t))_{t \geq 0}$  die von dem Prozess*

$$(1_A(\omega) + 2 \cdot 1_{\Omega \setminus A}(\omega)) \cdot 1_{[\xi(\omega), \infty)}(t)_{t \geq 0}$$

erzeugte  $\sigma$ -Algebra. Wir betrachten den Auszahlungsprozess

$$(2.36) \quad h(s) \triangleq \begin{cases} 2e^{-s} & s < \xi \\ 3e^{-s} \cdot 1_A + e^{-s} \cdot 1_{\Omega \setminus A} & s \geq \xi. \end{cases}$$

Mit der Forward Induction lässt sich der dazugehörige Gittins-Indexprozess  $M$  berechnen. Wegen des schnellen Fallens der Renditen wird für hinreichend kleines  $\lambda$  das Supremum stets durch sofortiges Stoppen approximiert, das heißt es gilt  $M(s) = h(s)$ .

Auf  $A$  gilt damit  $M(\xi-) < M(\xi)$ , die Pfade des Indexprozesses sind also nicht  $\mathbb{P}$ -f.s. unterhalbstetig von links. Insbesondere ist die Abbildung  $m \mapsto \sigma(m)$  nicht  $\mathbb{P}$ -f.s. rechtsstetig in  $m = h(\xi-) = 2e^{-\xi}$ .

## 2.12 Das Gittins-Indextheorem

Zur ersten Gleichheit in (2.21) ist äquivalent, dass  $\mathbb{P}$ -f.s. gilt

$$(2.37) \quad \int_0^\infty \int_{T_p(t)}^\infty e^{-\alpha s} (\underline{M}_p(0, s) - \underline{M}_p(T_p(t), s)) ds de^{-\alpha(t-T_p(t))} = 0$$

für  $p = 1, \dots, d$ . Dazu wiederum äquivalent ist

$$\int_0^\infty \int_{T_p(t)}^\infty 1_{\{\underline{M}_p(0, s) < \underline{M}_p(T_p(t), s)\}} ds d(t - T_p(t)) = 0.$$

Wegen der pfadweisen Rechtsstetigkeit von  $s \mapsto \underline{M}_p(0, s) - \underline{M}_p(T_p(t), s)$  ist dafür hinreichend und notwendig, dass

$$(2.38) \quad M_p(T_p(t)) = \underline{M}_p(0, T_p(t)) \text{ für } d(u - T_p(u)) - \text{f.a. } t,$$

und damit

$$\int_0^\infty 1_{\{M_p(T_p(t)) > \underline{M}_p(T_p(t))\}} d(t - T_p(t)) = 0.$$

Dies ist die Exkursionseigenschaft (2.27) der Strategie  $T$ . Nun ergibt sich das zentrale

**Theorem 2.35.** *Die Menge der für das Problem (2.8) optimalen Strategien und die Menge der Indexstrategien sind gleich. Insbesondere existiert eine optimale Strategie.*

*Beweis.* Es sei  $T$  eine Indexstrategie.  $T$  folgt dem größten  $\underline{M}_p$ , also gilt die zweite Gleichheit in (2.21). Weiter besitzt  $T$  die Exkursionseigenschaft, damit gilt auch die erste Gleichheit.

Ist umgekehrt  $T$  eine optimale Strategie, so muss  $T$  ebenfalls zweimal die Gleichheit erfüllen, da eine Indexstrategie existiert. Aus der ersten Gleichheit folgt die Exkursionseigenschaft, wegen der zweiten erfüllt  $T$  die Synchronization Identity. Damit ist  $T$  nach Lemma 2.22 ebenfalls eine Indexstrategie.  $\square$



# Literaturverzeichnis

- [1] P.Bank, N.El Karoui (2004) *A Stochastic Representation Theorem with Applications to Optimization and Obstacle Problems*, The Annals of Probability, 32(1B), 1030-1067
- [2] D.A.Berry, B.Fristedt (1985) *Bandit Problems - Sequential Allocation of Experiments*, Chapman and Hall
- [3] S.Bhulai, G.Koole (2000) *On the Value of Learning for Bernoulli Bandits with Unknown Parameter*, IEEE Transactions on Automatic Control, Vol.45, 2135-2140
- [4] R.Cairoli, J.B.Walsh (1975) *Stochastic integrals in the plane*, Acta Mathematic., 134, 111-183
- [5] C.Dellacherie (1972) *Capacités et Processus Stochastiques*, Springer, Berlin
- [6] C.Dellacherie, P.-A. Meyer (1975) *Probabilités et potentiel, Chapitres I à IV*, Hermann, Paris
- [7] C.Dellacherie, P.-A. Meyer (1980) *Probabilités et potentiel, Chapitres V à VIII*, Hermann, Paris
- [8] C.Dellacherie, E.Lenglart (1981) *Sur des problèmes de régularisation, de recollement et d'interpolation en théorie des processus*, Séminaire de Probabilités XVI, Lecture Notes in Mathematics, 920, 298-313, Springer
- [9] N.El Karoui, (1981) *Les aspects probabilistes du contrôle stochastique* Ecole d'Eté de Probabilités de Saint-Flour IX-1979, Lecture Notes in Mathematics, 876, 74-238, Springer
- [10] N.El Karoui, I.Karatzas (1993) *General Gittins index processes in discrete time*, Proc. Natl. Acad. Sci., 90, 1223-1236
- [11] N.El Karoui, I.Karatzas (1994) *Dynamic Allocation Problems in Continuous Time*, The Annals of Applied Probability, 4(2), 255-286
- [12] N.El Karoui, I.Karatzas (1997) *Synchronization and Optimality for Multi-Armed Bandit Problems in Continuous Time*, Comput.Appl.Math, 16(2), 117-151
- [13] J.C.Gittins, D.M.Jones (1974) *A dynamic allocation index for the sequential design of experiments*, Progress in Statistics (ed. J. Gani), 241-266, North-Holland, Amsterdam
- [14] J.C.Gittins (1989) *Multi-armed Bandit Allocation Indices*, John Wiley and Sons

- [15] J.Gittins, Y.G.Wang (1992) *The learning component of dynamic allocation indices*, Annals of Statistics, 20(3), 1625-1636
- [16] H.Kaspi, A.Mandelbaum (1995) *Lévy Bandits: Multi-Armed Bandits driven by Lévy Processes*, The Annals of Applied Probability, 5(2), 541-565
- [17] H.Kaspi, A.Mandelbaum (1998) *Multi Armed Bandits in Discrete and Continuous Time*, The Annals of Applied Probability, 8(4), 1270-1290
- [18] A.Mandelbaum, R.J.Vanderbei (1981) *Optimal Stopping and Supermartingales over Partially Ordered Sets*, Zur Wahrscheinlichkeitstheorie verwandte Gebiete, 57, 253-264
- [19] A.Mandelbaum (1987) *Discrete Multi-Armed Bandits and Multiparameter Processes*, Probab.Theory Rel.Fields, 71, 129-147
- [20] A.Mandelbaum (1987) *Continuous Multi-Armed Bandits and Multiparameter Processes*, The Annals of Probability, 15(4), 1527-1556
- [21] J.Neveu (1975) *Discrete Parameter Martingales*, North-Holland, Amsterdam
- [22] H.Robbins: *Some aspects of the sequential design of experiments*, Bull.Amer.Math.Soc., 58, 527-535
- [23] J.L.Snell (1952) *Applications of martingale system theorems*, Trans. Amer. Math. Soc., 73, 293-312
- [24] W.R.Thompson (1933) *On the likelihood that one unknown probability exceeds another in view of the evidence of two samples*, Biometrika, 25, 275-294
- [25] P.Varaiya, J.Walrand, C.Buyukkoc(1985): *Extension of the multi-armed bandit problem: the discounted case*, IEEE Trans. Autom.Control, AC-30, 426-439
- [26] J.B.Walsh (1980): *Optional Increasing Paths*, Lecture Notes in Mathematics, 863, ed. H.Korezlioglu,172-201, Springer
- [27] Y.G.Wang (1991): *Gittins Indices and Constrained Allocation in Clinical Trials*, Biometrika, 78(1), 101-111
- [28] R.Weber (1992) *On the Gittins Index for Multiarmed Bandits*, The Annals of Applied Probability, 2(4), 1024-1033
- [29] P.Whittle (1980) *Multi-armed Bandits and the Gittins Index*, J.R. Stat. Soc., 42(2), 143-149

# Thesen

Der Gittins-Indexprozess lässt sich als Lösung eines Darstellungsproblems für optionale Prozesse charakterisieren und das bekannte Gittins-Indextheorem mit Hilfe eines solchen beweisen.

Diese Charakterisierung ermöglicht es, eine gewisse Regularität der Pfade der Indexprozesse herzuleiten. Genauer sind die Pfade der Gittins-Indexprozesse unterhalbstetig von rechts.

Anhand einfacher Beispiele erkennt man, dass es im Gegensatz zum zeitdiskreten Fall für die Optimalität einer Strategie in der Situation stetiger Zeit nicht hinreichend ist, ausschließlich Arme mit maximalem Gittins-Index zu betätigen.

Die optimalen Strategien im zeitstetigen Fall können vollständig charakterisiert werden. Wesentlich dafür sind die Pfadeigenschaften der Indexprozesse.

Die Konstruktion einer Indexstrategie zeigt die Existenz von Lösungen der betrachteten dynamischen Allokationsprobleme in stetiger Zeit.

# Erklärung

Ich versichere, die vorliegende Arbeit nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt zu haben.

Berlin, den 27. Juli 2005