Wahrscheinlichkeitsrechnung 2

Dynamische stochastische Systeme mit diskretem Zustandsraum

Friedrich Graef
Institut für Angewandte Mathematik
der Universität Erlangen-Nürnberg

Fassung vom 25. Februar 1998

Inhaltsverzeichnis

Einleitung **5**

1 Wahrscheinlichkeitsräume **9**
1.1 Diskrete Wahrscheinlichkeitsräume 10
1.1.1 Wahrscheinlichkeitsfunktionen 10
1.1.2 Die Momente 11
1.1.3 Erzeugende Funktionen 11
1.1.4 Faltung 15
1.2 Zufallsvariable 16
1.2.1 Stochastische Unabhängigkeit. 16
1.2.2 Erwartungswert und Varianz 16
1.2.3 Faltungen. 17
1.2.4 Bedingte Verteilungen 19
1.3 Konvergenzbegriffe 22
1.3.1 Stochastische Konvergenz von Zufallsvariablen 22
1.3.2 Verteilungskonvergenz 23
1.4 Verzweigungsprozesse 28
1.4.1 Der Ansatz von Watson 28
1.4.2 Die Aussterbewahrscheinlichkeiten. 30
1.4.3 Die totale Familiengröße 33
1.4.4 Zeitdiskrete Bedienungssysteme 40
1.4.5 Literatur 43

2 Markoffketten **45**
2.1 Zeitdiskrete stochastische Prozesse 46
2.1.1 Wahrscheinlichkeiten auf Folgenräumen 46
2.1.2 Die Markoff-Eigenschaft. 50
2.1.3 Spezielle Markoffketten 54
2.2 Homogene Markoffketten 60
2.2.1 Die Mehrschritt-Übergangswahrscheinlichkeiten 60
2.2.2 Die stationäre Verteilung 62

2.2.3 Erreichbarkeit von Zuständen . . . 65
2.2.4 Periodische Zustände . . . 67
2.2.5 Rekurrente und transiente Zustände . . . 68
2.2.6 Positive Rekurrenz und Nullrekurrenz . . . 73
2.2.7 Die Struktur des Zustandsraums . . . 75
2.3 Absorbierende Markoffketten . . . 78
2.3.1 Absorbierende Zustände . . . 78
2.3.2 Die Absorptionswahrscheinlichkeiten . . . 79
2.3.3 Beispiele . . . 80
2.3.4 Die mittlere Schrittzahl bis zur Absorption . . . 83
2.3.5 Graphenreduktion . . . 86
2.4 Ergodische Markoffketten . . . 87
2.4.1 Simulated Annealing . . . 91

3 Markoffprozesse **97**
3.1 Markoff-Prozesse mit diskretem Zustandsraum . . . 98
3.2 Der Poissonprozeß . . . 101
3.3 Homogene Markoffprozesse . . . 106
3.3.1 Markoff-Bedienungssysteme . . . 109
3.4 Die stationäre Verteilung eines Markoffprozesses . . . 112
3.4.1 Definition der stationären Verteilung . . . 112
3.4.2 Berechnung der stationären Verteilung . . . 113
3.4.3 Das M/M/1-System im stationären Zustand . . . 115
3.4.4 Existenz einer stationären Verteilung . . . 118
3.5 Bevölkerungsentwicklungen . . . 123
3.5.1 Die $o(t)$-Notation . . . 123
3.5.2 Der allgemeine Geburts- und Todesprozeß . . . 124
3.5.3 Der zeitkontinuierliche Verzweigungsprozeß . . . 126

4 Anhang **131**

Literaturverzeichnis **133**

Einleitung

Dieser zweite Teil der Vorlesungsreihe Wahrscheinlichkeitsrechnung für Informatiker und Ingenieure befaßt sich mit Zufallsexperimenten, bei denen die Zeit eine wesentliche Rolle spielt. Es werden die folgenden beiden Grundmuster behandelt:

- Folgen von Zufallsexperimenten, wobei das Wahrscheinlichkeitsgesetz für das n-te Experiment von den Ergebnissen der vorher durchgeführten Experimente abhängt.
- Zufallsexperimente, bei denen ein bestimmtes Ereignis zu zufälligen Zeitpunkten eintritt.

Beispiele.

1. Ein typisches Beispiel für die erste Situation sind Glücksspiele der folgenden Art:

> Ein Spieler mit einem Startkapital von $s = 5$ DM möchte in einer Spielbank durch einige Runden Roulette auf ein Endkapital von $z = 100$ DM kommen.

Aus der Sicht der Wahrscheinlichkeitsrechnung hat man es dabei mit einer Folge $X_0, X_1, X_2, \ldots$ von Zufallsvariablen zu tun, wobei X_n den Kapitalbestand des Spielers nach der n-ten Runde (mit $X_0 = s$) darstellt. Die Verteilung der Variable X_n hängt dabei vom Kapitalbestand x_{n-1} vor Beginn der n-ten Runde und von der Strategie des Spielers ab, der vermutlich die früher realisierten Bestände $x_1, x_2, \ldots, x_{n-1}$ in sein Kalkül mit einbezieht.
Die Fragestellungen im Zusammenhang mit diesem Zufallsexperiment sind naheliegend. Es geht um die Wahrscheinlichkeiten der drei Ereignisse

V: Der Spieler geht in irgendeiner Runde n bankrott ($X_n = 0$), ohne sein Ziel erreicht zu haben.

G: Der Spieler erreicht irgendwann einmal sein Ziel ($X_n = z$), ohne vorher bankrott zu gehen.

U: Das Spiel hört niemals auf: Es ist $0 < X_n < z$ für alle Runden $n = 1, 2, , 3 \ldots$.

und um die Frage, wie lange es im Mittel dauert, bis eines der Ereignisse G oder V eintritt.

2. Repräsentative Vertreter für die zweite Art von Zufallsexperimenten sind Bedienungssysteme:

> Zu zufälligen Zeitpunkten $0 < T_1 < T_2 < \cdots$ kommt jeweils ein Kunde an einem Schalter (z.B. der Kasse eines Supermarkts) an. Falls gerade kein anderer Kunde am Schalter steht, wird er sofort bedient. Andernfalls stellt er sich an das Ende einer Warteschlange von Kunden, die der Reihe nach mit Bedienungszeiten von evtl. zufälliger Länge abgefertigt werden.

Die in diesem Fall bedeutsamen Größen sind die Anzahlen W_t der Kunden, die sich zu den Zeitpunkten t im Bedienungssystem aufhalten, also eine durch reelle Zahlen $t \geq 0$ indizierte Menge von Zufallsvariablen. Die mathematische Problemstellung ist klar: Es geht um die Berechnung der Wahrscheinlichkeitsverteilungen $p_k(t) = \mathcal{P}(W_t = k)$ der Zufallsvariablen W_t und die Frage, ob sich das Bedienungssystem irgendwann einmal in einem *stationären Zustand* einpendelt, d.h. um das Verhalten von $p_k(t)$ für $t \longmapsto \infty$.

Stochastische Prozesse

Die beschriebenen Situationen sind Spezialfälle des folgenden allgemeinen wahrscheinlichkeitstheoretischen Modells: Gegeben ist eine Indexmenge $\mathcal{T}$, die als die *Zeitachse* interpretiert wird, und eine Menge $X = (X_t \,;\, t \in \mathcal{T})$ von Zufallsvariablen $X_t : \Omega \longrightarrow \mathcal{S}$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$, die durch die Elemente aus $\mathcal{T}$ indiziert sind und die alle den gleichen Wertebereich $\mathcal{S}$ besitzen. Ein solches Gebilde nennt man einen *stochastischen Prozeß*.
Im Rückblick auf die obigen Beispiele, bei denen der Zustand eines Systems (Roulettetisch, Bedienungsschalter) im Lauf der Zeit verfolgt wird, sind die folgenden Bezeichnungen naheliegend und anschaulich:

- Die Elemente $x \in \mathcal{S}$ des Wertebereichs des stochastischen Prozesses heißen *Zustände*.
- Der Wertebereich $\mathcal{S}$ selbst heißt der *Zustandsraum* des stochastischen Prozesses.

In dieser Vorlesung werden wie schon in den obigen Beispielen nur abzählbare Zustandsräume vorkommen, die keinen Häufungspunkt enthalten. In den meisten Fällen wird es die Menge $\mathcal{S} = \mathbb{N} = \{1, 2, 3, \ldots\}$ der natürlichen oder die Menge $\mathcal{S} = \mathbb{N}_0 = \{0, 1, 2, 3, \ldots\}$ der nichtnegativen ganzen Zahlen sein. Als σ-Algebra auf $\mathcal{S}$, die für die Wahrscheinlichkeitsrechnung ja auch noch

erforderlich ist, benutzt man in diesem Fall die Menge aller Teilmengen von $\mathcal{S}$.
Einen derartigen Zustandsraum nennen wir im folgenden einen *diskreten Zustandsraum*.

Dynamische stochastische Systeme

In dieser Vorlesung werden wir uns mit solchen stochastischen Prozessen beschäftigen, die die sogenannte *Markoff-Eigenschaft* besitzen. Ein derartiger Prozeß ist anschaulich durch eine Art *Gedächtnislosigkeit* charakterisiert: Das zukünftige Verhalten des Prozesses bezogen auf einen beliebigen Gegenwartszeitpunkt t hängt stochastisch nur vom gegenwärtigen Zustand $X_t = x_t$ ab und ist unabhängig von den in der Vergangenheit realisierten Zuständen x_s mit $s < t$. Dies entspricht deterministischen mathematischen Modellen, die in der Analysis als dynamische Systeme bezeichnet werden. Wir verwenden daher für stochastische Prozesse mit der Markoff-Eigenschaft die Bezeichnung *dynamische stochastische Systeme*.

Überblick

Die Vorlesung ist in drei Kapitel eingeteilt. Im ersten Kapitel werden einige Hilfsmittel erarbeitet, die über den Stoff der Vorlesung Wahrscheinlichkeitsrechnung 1 hinausgehen. Als unmittelbare Anwendung betrachten wir die *Verzweigungsprozesse*, die historisch gesehen den Ausgangspunkt der Analyse stochastischer Prozesse darstellen.
Das zweite Kapitel über Markoffketten behandelt zeitdiskrete und das dritte mit dem Titel „Markoffprozesse“ zeitkontinuierliche dynamische stochastische Systeme, jeweils mit diskretem Zustandsraum.

Kapitel 1

Wahrscheinlichkeitsräume

In diesem Kapitel werden einige Themen aus dem Inhalt der Vorlesung Wahrscheinlichkeitsrechnung 1 vertieft, die im folgenden eine wesentliche Rolle spielen. Insbesondere sind das die erzeugenden Funktionen, bedingte Wahrscheinlichkeiten und die Konvergenz von Zufallsvariablen und Verteilungen. Außerdem wird das Konzept des stochastischen Prozesses eingeführt.

1.1 Diskrete Wahrscheinlichkeitsräume

Ein *diskreter Wahrscheinlichkeitsraum* besteht aus einer endlichen oder abzählbar unendlichen Ergebnismenge $\mathcal{X}$ und einer Wahrscheinlichkeitsverteilung $\mathcal{P}$ auf der Menge $2^{\mathcal{X}}$ aller Teilmengen von $\mathcal{X}$. Da die σ-Algebra festliegt, schreiben wir für einen solchen Wahrscheinlichkeitsraum kurz $(\mathcal{X}, \mathcal{P})$.
$\mathcal{P}$ heißt eine *diskrete Wahrscheinlichkeitsverteilung*. Sie ist eindeutig bestimmt durch ihre *Wahrscheinlichkeitsfunktion* $f(x) = \mathcal{P}\{x\}$. Jede Funktion $f : \mathcal{X} \longrightarrow \mathbb{R}$ mit $f(x) \geq 0$ für alle $x \in \mathcal{X}$ und $\sum_{x\in\mathcal{X}} f(x) = 1$ legt durch

$$\mathcal{P}(A) = \sum_{x\in A} f(x)$$

eine diskrete Wahrscheinlichkeitsverteilung $\mathcal{P}$ auf $\mathcal{X}$ fest.

1.1.1 Wahrscheinlichkeitsfunktionen

Im Rahmen dieser Vorlesung wird $\mathcal{X}$ fast immer eine Teilmenge von $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$ sein. Wenn es sich in dieser Situation für die Darstellung als zweckmäßig erweist, werden wir als Ergebnismenge die gesamte Menge $\mathbb{N}_0$ verwenden und solchen x, die nicht in der ursprünglichen Ergebnismenge $\mathcal{X}$ liegen, dafür die Wahrscheinlichkeit $f(x) = 0$ zuordnen.

Bernoulli-Verteilung. Sie besitzt die Wahrscheinlichkeitsfunktion

$$f(x) = \begin{cases} q & \text{falls } x = 0 \\ p & \text{falls } x = 1 \\ 0 & \text{sonst} \end{cases}$$

Dabei ist $0 < p < 1$ und $q = 1 - p$. Interessiert man sich beim Werfen eines regulären Würfels nur dafür, ob eine „Sechs“ geworfen wird oder nicht, so wird dieses Experiment durch eine Bernoulli-Verteilung mit $p = 1/6$ beschrieben, wenn man das Ereignis „Werfen einer Sechs“ mit dem Ergebnis „1“ identifiziert.

Binomialverteilung.

$$f(k) = \begin{cases} \binom{n}{k} p^k q^{n-k} & \text{für } k = 0, 1, \ldots n \\ 0 & \text{sonst} \end{cases}$$

p und q wie oben. Bezogen auf das obige Würfelexperiment ist $f(k)$ die Wahrscheinlichkeit, daß bei n Würfen genau k-mal die „Sechs“ geworfen wird.

Geometrische Verteilung.

$$f(n) = \begin{cases} 0 & \text{für } n = 0 \\ pq^{n-1} & \text{für } n \geq 1 \end{cases}$$

p und q wie oben. Sie stellt das Wahrscheinlichkeitsgesetz für die Wartezeit bis zum ersten Auftreten einer „Sechs“ dar.

Poissonverteilung.

$$f(n) = e^{-\mu}\frac{\mu^n}{n!}$$

mit einem Parameter $\mu > 0$. Die Poissonverteilung findet bei Zählungen Anwendung. $f(n)$ ist dann die Wahrscheinlichkeit, daß ein bestimmtes Ereignis, z.B. die Ankunft eines Kunden in einem Bedienungssystem, innerhalb eines vorgegebenen Zeitintervalls genau n-mal eintritt.

1.1.2 Die Momente

Für eine diskrete Wahrscheinlichkeitsverteilung $\mathcal{P}$ auf der Menge $\mathbb{N}_0$ mit der Wahrscheinlichkeitsfunktion f heißen

$$m_k(\mathcal{P}) = \sum_{n=0}^{\infty} n^k f(n) \quad \text{und} \quad \hat{m}_k(\mathcal{P}) = \sum_{n=0}^{\infty}(n - m_1(\mathcal{P}))^k f(n)$$

das *k-te* bzw. das *k-te zentrale Moment* der Verteilung $\mathcal{P}$. Beim k-ten Moment kann natürlich der Summand mit dem Index $n = 0$ weggelassen werden. Außerdem wird im allgemeinen das Argument $\mathcal{P}$ weggelassen, wenn aus dem Zusammenhang klar ersichtlich ist, um welche Wahrscheinlichkeitsverteilung es sich handelt. Von besonderer Bedeutung sind m_1, der *Mittelwert*, und $\hat{m}_2$, die *Varianz* der Verteilung. Zwischen ihnen besteht die Beziehung

$$\hat{m}_2 = m_2 - (m_1)^2$$

wie sich leicht anhand der Definitionen nachrechnen läßt.

1.1.3 Erzeugende Funktionen

Erzeugende Funktionen sind für diskrete Wahrscheinlichkeitsverteilungen auf $\mathbb{N}_0$ eine zu den Wahrscheinlichkeitsfunktionen äquivalente Beschreibungsmethode, die besonders vorteilhaft ist, wenn man mit Summen von Zufallsvariablen operiert. Es handelt sich dabei um nichts anderes als die aus der Elektrotechnik bekannte z-Transformation.

Für eine Folge $a = (a_0, a_1, a_2, \ldots)$ heißt die Potenzreihe

$$\hat{a}(z) = a_0 + a_1 z + a_2 z^2 + \cdots = \sum_{k=0}^{\infty} a_k z^k \tag{1.1}$$

die *z-Transformierte der Folge a*. Wir betrachten $\hat{a}(z)$ als Funktion einer reellen oder komplexen Variablen z und bezeichnen mit R_a den *Konvergenzradius*, d.h. die größte reelle Zahl mit der Eigenschaft, daß für $|z| < R_a$ die Reihe (1.1) absolut konvergiert.
Für $|z| < R_a$ ist die Funktion $\hat{a}(z)$ beliebig oft differenzierbar und man erhält die Ableitungen durch Differenzieren der einzelnen Summanden:

$$\begin{aligned} \frac{d\hat{a}}{dz}(z) &= a_1 + a_2 \cdot 2 \cdot z + a_3 \cdot 3 \cdot z^2 + \cdots \\ \frac{d^2\hat{a}}{dz^2}(z) &= a_2 \cdot 2 + a_3 \cdot 3 \cdot 2 \cdot z + \cdots \end{aligned}$$

usw.
Umgekehrt ergibt sich aus einem Koeffizientenvergleich mit der Taylorreihe um $z = 0$

$$\hat{a}(z) = \sum_{k=0}^{\infty} \frac{1}{k!} \frac{d^k \hat{a}}{dz^k}(0) z^k$$

daß sich die Folgenelemente a_k aus der z-Transformierten $\hat{a}(z)$ berechnen lassen:

$$a_k = \frac{1}{k!} \frac{d^k \hat{a}}{dz^k}(0)$$

Ein Beispiel für die Anwendungsmöglichkeiten der z-Transformation ist die Berechnung einer expliziten Formel für die Gleider der Fibonacci-Folge ([12], Seite 90).
Fibonacci (*Leonardo von Pisa, 1170–1230, Sohn des Bonacci*) stellte die Aufgabe, die Vermehrung von Kaninchen zahlenmäßig zu erfassen. Er machte dabei die folgenden Annahmen:

1. Zu Beginn des 1. Monats wird die Zählung mit einem Paar neu geborener Kaninchen gestartet.

2. Ein Kaninchen ist erwachsen, wenn es mindestens zwei Monate alt ist.

3. Ein Paar erwachsener Kaninchen erzeugt jeden Monat ein Paar Kaninchen.

4. Kaninchen sterben nie.

Für die Anzahl a_n der im n-ten Monat vorhandenen Kaninchenpaare ergibt sich daraus:

- $a_1 = 1$ und $a_2 = 1$, denn das Kaninchenpaar ist im zweiten Monat noch nicht erwachsen.
- Für $n > 2$ ist a_n die Anzahl der im Vormonat vorhandenen Paare (a_{n-1}) plus der in diesem Monat neu geborenen. Letztere ist identisch mit der im Monat n erwachsenen Paare, was a_{n-2} entspricht. Daraus folgt die Rekursion $a_n = a_{n-1} + a_{n-2}$.

Für die z-Transformierte dieser Folge (mit $a_0 := 0$) erhält man aus dieser Rekursionsformel die Gleichung

$$\begin{aligned}\hat{a}(z) &= \sum_{n=1}^{\infty} a_n z^n = a_1 z + a_2 z^2 + \sum_{n=3}^{\infty} (a_{n-1} + a_{n-2}) z^n \\ &= z + z^2 + \sum_{n=2}^{\infty} a_n z^{n+1} + \sum_{n=1}^{\infty} a_n z^{n+2} = z + \sum_{n=1}^{\infty} a_n z^{n+1} + \sum_{n=1}^{\infty} a_n z^{n+2} \\ &= z + z\hat{a}(z) + z^2 \hat{a}(z)\end{aligned}$$

oder

$$\hat{a}(z) = -\frac{z}{z^2 + z - 1}$$

Die Nullstellen des Nennerpolynoms sind $z_1 = \frac{1}{2}(\sqrt{5} - 1)$ und $z_2 = -\frac{1}{2}(\sqrt{5} + 1)$. Daher besitzt diese z-Transformierte die Partialbruchzerlegung

$$\begin{aligned}\hat{a}(z) &= -\left(\frac{A}{z - z_1} + \frac{B}{z - z_2}\right) = \frac{1}{\sqrt{5}}\left(\frac{z_2}{z - z_2} - \frac{z_1}{z - z_1}\right) \\ &= \frac{1}{\sqrt{5}}\left(\frac{1}{1 - \frac{z}{z_1}} - \frac{1}{1 - \frac{z}{z_2}}\right)\end{aligned}$$

Für genügend kleines z kann man die beiden Terme in den runden Klammern in geometrische Reihen entwickeln und erhält die Potenzreihe

$$\hat{a}(z) = \sum_{n=0}^{\infty} \frac{1}{\sqrt{5}} \left[\left(\frac{1}{z_1}\right)^n - \left(\frac{1}{z_2}\right)^n\right] z^n$$

mit den Fibonacci-Zahlen a_n als Koeffizienten:

$$a_n = \frac{1}{\sqrt{5}} \left[\left(\frac{1}{z_1}\right)^n - \left(\frac{1}{z_2}\right)^n\right] = \frac{1}{2^n \sqrt{5}} \left((1 + \sqrt{5})^n - (1 - \sqrt{5})^n\right)$$

Definition 1.1.1 *Die z-Transformierte*

$$\hat{f}(z) = \sum_{k=0}^{\infty} f(k) z^k \tag{1.2}$$

der Wahrscheinlichkeitsfunktion f einer diskreten Wahrscheinlichkeitsverteilung $\mathcal{P}$ auf $\mathbb{N}_0$ heißt die erzeugende Funktion *dieser Verteilung.*

Charakteristische Eigenschaften einer erzeugenden Funktion sind

$$\hat{f}(1) = 1 \tag{1.3}$$

und

$$\hat{f}'(1) = \lim_{z \to 1} \frac{d\hat{f}}{dz}(z) = m_1(\mathcal{P}) \tag{1.4}$$

Die Gleichung (1.3) folgt aus $\sum_k f(k) = 1$. Falls der Mittelwert $m_1(\mathcal{P})$ existiert, d.h. falls $\sum_k |kf(k)| = \sum_k kf(k) < \infty$, konvergiert die Potenzreihe $\frac{d\hat{f}}{dz}(z)$ im Bereich $|z| \leq 1$ absolut. Da man bei Potenzreihen unter dieser Voraussetzung Limesbildung und Summation vertauschen darf, erhält man (1.4).

Beispiele

Bernoulli-Verteilung

$$\hat{f}(z) = q + pz \quad \text{mit} \quad \hat{f}'(1) = p$$

Binomialverteilung

$$\hat{f}(z) = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) (pz)^k q^{n-k} = (q + pz)^n \quad \text{mit}$$
$$\hat{f}'(1) = n(q+p)^{n-1}p = np$$

geometrische Verteilung

$$\hat{f}(z) = \sum_{k=1}^{\infty} pq^{k-1}z^k = pz \sum_{m=0}^{\infty} (qz)^m = \frac{pz}{1-qz} \quad \text{mit}$$
$$\hat{f}'(1) = \frac{p}{(1-q)^2} = \frac{1}{p}$$

Poissonverteilung

$$\hat{f}(z) = \sum_{k=0}^{\infty} e^{-\mu} \frac{(\mu z)^k}{k!} = e^{-\mu} e^{\mu z} = e^{\mu(z-1)} \quad \text{mit}$$
$$\hat{f}'(1) = e^{\mu 0}\mu = \mu$$

1.1.4 Faltung

Für die Wahrscheinlichkeitsrechnung besonders wichtig ist die folgende Rechenregel

Satz 1.1.1 *Sind $a = (a_k)$ und $b = (b_k)$ reelle Zahlenfolgen und $c = a * b$ ihre Faltung, d.h.*

$$c_n = \sum_{k=0}^{n} a_k \cdot b_{n-k}$$

so gilt im Bereich $|z| < \min(R_a, R_b)$ für die entsprechenden z-Transformierten die Beziehung

$$\hat{c}(z) = \hat{a}(z) \cdot \hat{b}(z) \tag{1.5}$$

oder — in Kurzform —

$$\widehat{a * b} = \widehat{a}\,\widehat{b} \tag{1.6}$$

Beweis: Die Formel (1.5) erhält man durch eine einfache Änderung der Summationsreihenfolge, die wegen der absoluten Konvergenz innerhalb des Konvergenzradius erlaubt ist:

$$\begin{array}{rcl@{\qquad}rcl}
\hat{c}(z) & = & \sum_{n=0}^{\infty}\sum_{k=0}^{n} a_k b_{n-k} z^n & = & \sum_{n=0}^{\infty}\sum_{k=0}^{n} (a_k z^k)(b_{n-k} z^{n-k}) \\
& =: & \sum_{n=0}^{\infty}\sum_{k=0}^{n} x_k y_{n-k} & & \\
& = & \left\{ \begin{array}{l} x_0 y_0 + \\ x_0 y_1 + x_1 y_0 + \\ x_0 y_2 + x_1 y_1 + x_2 y_0 + \\ \vdots \end{array} \right\} & = & \left\{ \begin{array}{l} x_0(y_0 + y_1 + \cdots) + \\ x_1(y_0 + y_1 + \cdots) + \\ \vdots \end{array} \right\} \\
& = & (x_0 + x_1 + \cdots)(y_0 + y_1 + \cdots) & = & \hat{a}(z)\hat{b}(z)
\end{array}$$

1.2 Zufallsvariable

Allgemeine Zufallsgrößen und Zufallsvariable wurden bereits in der Vorlesung Wahrscheinlichkeitsrechnung 1 behandelt. In diesem Abschnitt sind einige spezielle Eigenschaften von Zufallsvariablen mit der Menge $\mathbb{N}_0$ als Wertebereich zusammengestellt, die in den folgenden Kapiteln häufig benötigt werden.
Sei $X : \Omega \longrightarrow \mathbb{N}_0$ eine Abbildung auf der Ergebnismenge eines Wahrscheinlichkeitsraums $(\Omega, \mathcal{A}, \mathcal{P})$. Da jede Teilmenge $B \subset \mathbb{N}_0$ abzählbar ist, ist

$$(X \in B) = \sum_{n \in B} (X = n) \tag{1.7}$$

für alle B eine abzählbare Summe von Mengen der Form $(X = n) = \{\omega \in \Omega \ ; \ X(\omega) = n\}$. Wegen der σ-Additivität von $\mathcal{A}$ lassen sich daher alle Eigenschaften mittels dieser Basismengen formulieren.
X ist eine Zufallsvariable genau dann, wenn $(X = n) \in \mathcal{A}$ für alle n, denn wegen (1.7) ist unter dieser Voraussetzung $(X \in B) \in \mathcal{A}$ für alle Teilmengen B von $\mathbb{N}_0$. Die Verteilung $\mathcal{P}^X$ der Zufallsvariablen X ist wegen

$$\mathcal{P}^X(B) = \mathcal{P}(X \in B) = \sum_{n \in B} \mathcal{P}(X = n) = \sum_{n \in B} \mathcal{P}^X\{n\}$$

diskret mit der Wahrscheinlichkeitsfunktion $f^X(n) = \mathcal{P}(X = n)$.

1.2.1 Stochastische Unabhängigkeit.

Die stochastische Unabhängigkeit läßt sich wegen (1.7) ebenfalls mit Hilfe der Basismengen formulieren:
Eine endliche Anzahl $X_1, X_2, \ldots, X_m$ von $\mathbb{N}_0$-wertigen Zufallsvariablen auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ ist *stochastisch unabhängig*, wenn für beliebige nichtnegative ganze Zahlen $n_1, n_2, \ldots, n_m$ gilt

$$\begin{aligned} & \mathcal{P}\left((X_1 = n_1) \cap (X_2 = n_2) \cap \ldots \cap (X_m = n_m)\right) \\ = \quad & \mathcal{P}(X_1 = n_1)\mathcal{P}(X_2 = n_2) \cdots \mathcal{P}(X_m = n_m) \end{aligned}$$

Eine Folge $X_1, X_2, \ldots$ von $\mathbb{N}_0$-wertigen Zufallsvariablen auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ heißt *stochastisch unabhängig*, wenn jede beliebige endliche Teilmenge von X_k's stochastisch unabhängig ist.

1.2.2 Erwartungswert und Varianz

Eine $\mathbb{N}_0$-wertige Zufallsvariable ist eine Treppenfunktion

$$X(\omega) = \sum_{n \in \mathbb{N}_0} n \mathbf{1}_{(X=n)}(\omega)$$

Für den Erwartungswert erhält man daher

$$\mathcal{E}X = \int X(\omega)\mathcal{P}(d\omega) = \sum_n n\mathcal{P}(X=n) = \sum_n n f^X(n) = m_1(\mathcal{P}^X) \qquad (1.8)$$

und für Funktionen $Y = G(X)$ wegen

$$Y(\omega) = G(X(\omega)) = \sum_n G(n)\mathbf{1}_{(X=n)}(\omega)$$

den Erwartungswert

$$\mathcal{E}Y = \sum_n G(n)\mathcal{P}(X=n) = \sum_n G(n) f^X(n) \qquad (1.9)$$

Daraus folgt speziell für $Y = X^2$

$$\mathcal{E}(X^2) = \sum_n n^2 f^X(n) = m_2(\mathcal{P}^X)$$

und für $Y = (X - \mathcal{E}X)^2$

$$\mathrm{var}X = \mathcal{E}(X - \mathcal{E}X)^2 = \sum_n (n - \mathcal{E}X)^2 f^X(n) = \hat{m}_2(\mathcal{P}^X)$$

Der Erwartungswert der Zufallsvariablen $Y = z^X$ ist

$$\mathcal{E}(z^X) = \sum_n z^n f^X(n) = \widehat{f^X}(z),$$

also gleich der erzeugenden Funktion der Zufallsvariablen X an der Stelle z.

1.2.3 Faltungen.

Satz 1.2.1 *X_1 und X_2 seien zwei stochastisch unabhängige Zufallsvariable mit Werten in $\mathbb{N}_0$ und ihre Verteilungen mögen die Wahrscheinlichkeitsfunktionen f_1 bzw. f_2 besitzen. Dann besitzt die Zufallsvariable $Y = X_1 + X_2$ die Verteilung mit der Wahrscheinlichkeitsfunktion*

$$f_1 * f_2(n) \;=\; \sum_{k=0}^{n} f_1(k) f_2(n-k) \qquad (1.10)$$

Beweis: Dies ergibt sich wegen der stochastischen Unabhängigkeit aus

$$\begin{aligned} \mathcal{P}(Y=n) &= \mathcal{P}\left(\bigcup_{k=0}^n (X_1=k) \cap (X_2 = n-k)\right) \\ &= \sum_{k=0}^n \mathcal{P}\left((X_1=k) \cap (X_2=n-k)\right) \\ &= \sum_{k=0}^n \mathcal{P}(X_1=k)\mathcal{P}(X_2=n-k) \end{aligned}$$

$\diamond$

Die Wahrscheinlichkeitsfunktion $f_1 * f_2$ heißt die *Faltung* der Wahrscheinlichkeitsfunktionen f_1 und f_2. Aus den Regeln $X_1 + X_2 = X_2 + X_1$ und $(X_1 + X_2) + X_3 = X_1 + (X_2 + X_3) =: X_1 + X_2 + X_3$ ergibt sich

$$f_1 * f_2 = f_2 * f_1$$
$$(f_1 * f_2) * f_3 = f_1 * (f_2 * f_3) =: f_1 * f_2 * f_3$$

Sind $X_1, X_2, \ldots, X_m$ stochastisch unabhängige Zufallsvariable, die alle die gleiche Wahrscheinlichkeitsverteilung mit der Wahrscheinlichkeitsfunktion f besitzen, so schreibt man für die Wahrscheinlichkeitsfunktion g der Summe $X_1 + X_2 + \cdots + X_m$ anstelle von

$$g = \underbrace{f * f * \cdots * f}_{m\text{-mal}}$$

auch kürzer $g = f^{m*}$.

Das wichtigste Hilfsmittel bei der Berechnung von Faltungen stellen die erzeugenden Funktionen dar. Sind nämlich X und Y stochastisch unabhängige $\mathbb{N}_0$-wertige Zufallsvariable, deren Verteilungen die Wahrscheinlichkeitsfunktionen f^X bzw. f^Y besitzen, so gilt für die Wahrscheinlichkeitsfunktion $f^{X+Y} = f^X * f^Y$ von $X + Y$ nach (1.6) die Gleichung

$$\widehat{f^{X+Y}} = \widehat{f^X}\,\widehat{f^Y} \tag{1.11}$$

Man kann (1.11) auch anders herleiten. Sind X und Y stochastisch unabhängig, dann auch z^X und z^Y und es gilt (s. Beispiel im letzten Abschnitt):

$$\widehat{f^{X+Y}}(z) = \mathcal{E}(z^{X+Y}) = \mathcal{E}(z^X z^Y) = \mathcal{E}(z^X)\mathcal{E}(z^Y) = \widehat{f^X}(z)\widehat{f^Y}(z)$$

Durch mehrfache Anwendung dieser Regel erhält man weiter

$$\widehat{f^{X_1+X_2+\cdots X_n}} = \widehat{f^{X_1}}\,\widehat{f^{X_2}} \ldots \widehat{f^{X_n}}$$

für stochastisch unabhängige Zufallsvariable X_1, X_2, $\ldots X_n$.

Als Anwendungsbeispiel betrachten wir die Herleitung der Binomialverteilung. Diese ist bekanntlich die Verteilung der Summe

$$S = X_1 + X_2 + \cdots + X_n$$

von n stochastisch unabhängigen Bernoulli-verteilten Zufallsvariablen X_k, deren Verteilungen die erzeugende Funktion $\hat{f}(z) = q + pz$ besitzen. Die erzeugende Funktion zu S ist daher

$$\widehat{f^S}(z) = \widehat{f^{n*}}(z) = (\hat{f}(z))^n = (q + pz)^n$$

Die Werte der Wahrscheinlichkeitsfunktion f^S erhält man als die Koeffizienten der Potenzreihenentwicklung der erzeugenden Funktion. In unserem Fall kann man die Binomialformel anwenden:

$$(q+pz)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) (pz)^k q^{n-k} = \sum_{k=0}^{n} \left(\left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k} \right) z^k$$

Ein anderes Beispiel: Sind die X_k Poisson-verteilt mit Parametern μ_k, so ist

$$\begin{aligned} \widehat{f^S}(z) &= \hat{f}_1(z) \cdot \hat{f}_2(z) \cdots \hat{f}_n(z) \\ &= e^{\mu_1(z-1)} e^{\mu_2(z-1)} \cdots e^{\mu_n(z-1)} \\ &= e^{(\mu_1+\mu_2+\cdots+\mu_n)(z-1)} \\ &= e^{\mu(z-1)} \end{aligned}$$

und man entnimmt der Form dieser Funktion, daß die Verteilung von S die Poissonverteilung mit Parameter $\mu = \mu_1 + \mu_2 + \cdots + \mu_n$ ist.

1.2.4 Bedingte Verteilungen

Seien wieder X und Y $\mathbb{N}_0$-wertige Zufallsvariable und $Z = X + Y$. Wenn man die bedingte Wahrscheinlichkeit $\mathcal{P}(Z = n|Y = k)$ berechnen will, so erscheint es naheliegend, die folgende Ersetzung vorzunehmen:

$$\begin{aligned} \mathcal{P}(Z = n|Y = k) &= \mathcal{P}(X + Y = n|Y = k) \\ &= \mathcal{P}(X + k = n) \\ &= \mathcal{P}(X = n - k) \end{aligned}$$

An dem Beispiel $\mathcal{P}((X = 0) \cap (Y = 0)) = \mathcal{P}((X = 1) \cap (Y = 1)) = 1/3$ und $\mathcal{P}((X = 0) \cap (Y = 1)) = \mathcal{P}((X = 1) \cap (Y = 0)) = 1/6$ erhält man aber

$$\mathcal{P}(X = 1) = 1/2$$

und

$$\mathcal{P}(X + Y = 1|Y = 0) = \frac{\mathcal{P}\left((X = 1) \cap (Y = 0)\right)}{\mathcal{P}(Y = 0)} = 1/3$$

Der Unterschied kommt offensichtlich dadurch zustande, daß X und Y stochastisch abhängig sind. Man benötigt als Voraussetzung also auf jeden Fall die stochastische Unabhängigkeit der beteiligten Zufallsgrößen.

Satz 1.2.2 *Sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein beliebiger Wahrscheinlichkeitsraum, $X : \Omega \longrightarrow \tilde{\Omega}$ eine beliebige, $Y : \Omega \longrightarrow \mathbb{N}_0$ eine $\mathbb{N}_0$-wertige Zufallsgröße und $G : \tilde{\Omega} \times \mathbb{N}_0 \longrightarrow \mathbb{N}_0$ eine meßbare Abbildung. Sind X und Y stochastisch unabhängig, so gilt für beliebige nichtnegative ganze Zahlen k und n, daß*

$$\mathcal{P}(G(X,Y) = n|Y = k) = \mathcal{P}(G(X,k) = n)$$

Beweis:

$$\begin{aligned}(G(X,Y)=n)\cap(Y=k) &= \{\omega\ ;\ G(X(\omega),Y(\omega))=n \text{ und } Y(\omega)=k)\}\\ &= \{\omega\ ;\ G(X(\omega),k)=n \text{ und } Y(\omega)=k\}\\ &= \{\omega\ ;\ G(X(\omega),k)=n\}\cap\{\omega\ ;\ Y(\omega=k\}\\ &= (G(X,k)=n)\cap(Y=k)\end{aligned}$$

Da $Z_k = G(X,k)$ nur von der Variablen X abhängt, sind mit X und Y auch Z_k und Y stochastisch unabhängig, so daß

$$\begin{aligned}\mathcal{P}(G(X,Y)=n|Y=k) &= \frac{\mathcal{P}((G(X,Y)=n)\cap(Y=k))}{\mathcal{P}(Y=k)}\\ &= \frac{\mathcal{P}((Z_k=n)\cap(Y=k))}{\mathcal{P}(Y=k)}\\ &= \frac{\mathcal{P}(Z_k=n)\mathcal{P}(Y=k)}{\mathcal{P}(Y=k)}\\ &= \mathcal{P}(Z_k=n)=\mathcal{P}(G(X,k)=n)\end{aligned}$$

◇

Damit ergibt sich eine Formel für eine Summe von Zufallsvariablen, bei der auch noch die Anzahl der Summanden eine Zufallsvariable ist.

Satz 1.2.3 *$X_1, X_2, \ldots$ sei eine Folge von stochastisch unabhängigen identisch verteilten $\mathbb{N}_0$-wertigen Zufallsvariablen mit erzeugender Funktion $\widehat{f^X}$ und N eine $\mathbb{N}_0$-wertige Zufallsvariable mit erzeugender Funktion $\widehat{f^N}$, die von den X_k stochastisch unabhängig ist. Dann ist die erzeugende Funktion $\widehat{f^S}$ der Zufallsvariable*

$$S = X_1 + X_2 + \cdots + X_N \tag{1.12}$$

gegeben durch

$$\widehat{f^S}(z) = \widehat{f^N}\left(\widehat{f^X}(z)\right)$$

Beweis: Die Ereignisse $(N=n)$, $n=0,1,2,\ldots$, bilden eine Partition der Ergebnismenge. Für die Wahrscheinlichkeitsfunktion f^S der Verteilung von S an der Stelle k erhält man nach dem Satz von der totalen Wahrscheinlichkeit

$$f^S(k) = \mathcal{P}(S=k) = \sum_{n=0}^{\infty} \mathcal{P}(S=k|N=n)\,\mathcal{P}(N=n)$$

Auf die bedingten Wahrscheinlichkeiten in dieser Summe wenden wir den Satz 1.2.2 mit der Funktion

$$G(x,n) = x_1 + x_2 + \cdots + x_n$$

für Folgen $x = (x_1, x_2, \ldots)$ an, wobei $G(x, 0) := 0$. Es gilt danach

$$\begin{aligned}\mathcal{P}(S = k|N = n) &= \mathcal{P}(G(X, N) = k|N = n)\\ &= \mathcal{P}(G(X, n) = k)\\ &= \mathcal{P}(X_1 + X_2 + \cdots + X_n \ = \ k)\end{aligned}$$

Um die folgende Darstellung nicht durch Indizes zu überladen, schreiben wir f für die Wahrscheinlichkeitsfunktion f^X der Verteilung der Variablen X_i. Die Summe von n solchen Variablen besitzt die Wahrscheinlichkeitsfunktion f^{n*} und wir erhalten daher

$$\mathcal{P}(S = k|N = n) = f^{n*}(k)$$

Mit der Wahrscheinlichkeitsfunktion f^N von N ergibt sich damit

$$f^S(k) = \sum_{n=0}^{\infty} f^{n*}(k)\, f^N(n)$$

und für die erzeugende Funktion der Variablen S

$$\begin{array}{rcll}\widehat{f^S}(z) &=& \sum_{k=0}^{\infty} f^S(k)\, z^k & = \sum_{k=0}^{\infty}\sum_{n=0}^{\infty} f^{n*}(k)\, f^N(n) z^k\\ &=& \sum_{n=0}^{\infty}\left(\sum_{k=0}^{\infty} f^{n*}(k)\, z^k\right) f^N(n) & = \sum_{n=0}^{\infty}\left(\widehat{f^{n*}}(z)\right) f^N(n)\\ &=& \sum_{n=0}^{\infty}\left(\hat{f}(z)\right)^n f^N(n) & = \sum_{n=0}^{\infty} f^N(n)(\hat{f}(z))^n\\ &=& \widehat{f^N}(\hat{f}(z)) = \widehat{f^N}(\widehat{f^X}(z)) & \end{array}$$

◇

Nach der Kettenregel für die Differentiation erhält man

$$\widehat{f^S}'(1) = \widehat{f^N}'(1)\widehat{f^X}'(1)$$

Der Erwartungswert $\mathcal{E}S = \widehat{f^S}'(1)$ der Variablen S ist also gleich der mittleren Anzahl der Summanden $\mathcal{E}N = \widehat{f^N}'(1)$ multipliziert mit der mittleren Größe $\mathcal{E}X_k = \widehat{f^X}'(1)$ eines Summanden.

Anwendungsbeispiel: Ist N die Anzahl der Fahrzeuge, die bei einer Verkehrszählung innerhalb einer gewissen Zeitspanne die Meßstelle passieren, und ist $X_i = 1$ oder 0 je nachdem, ob das i-te Fahrzeug ein LKW ist oder nicht, so ist S die Anzahl der LKW's.
Ist N Poisson-verteilt mit Parameter μ und ist $\mathcal{P}(X_i = 1) = p$ sowie $\mathcal{P}(X_i = 0) = q = 1 - p$, so gilt

$$\hat{h}(z) = e^{\mu((q+pz)-1)} = e^{\mu(pz-(1-q))} = e^{\mu(pz-p)} = e^{(p\mu)(z-1)}$$

woraus man ersieht, daß S Poisson-verteilt mit Parameter $p\mu$ ist.

1.3 Konvergenzbegriffe

Da es in dieser Vorlesung darum geht, stochastische Vorgänge im zeitlichen Ablauf zu analysieren, spielt die Frage der Konvergenz eine wesentliche Rolle. In diesem Abschnitt werden zwei in der Wahrscheinlichkeitrechnung wichtige Konvergenzbegriffe vorgestellt. Als erster den der stochastischen Konvergenz von Zufallsvariablen, den wir anhand der Herleitung des Gesetzes der großen Zahlen einführen. Danach befassen wir uns mit der Konvergenz von Wahrscheinlichkeitsverteilungen, insbesondere für diskrete Verteilungen.

1.3.1 Stochastische Konvergenz von Zufallsvariablen

Das Gesetz der großen Zahlen. Ein Zufallsexperiment werde „unendlich oft“ wiederholt, wobei die einzelnen Durchführungen „unabhängig“ voneinander stattfinden und jeweils nur beobachtet wird, ob ein bestimmtes Ereignis eintritt oder nicht. Die Wahrscheinlichkeit für das Eintreten des Ereignisses bei einer Durchführung sei p mit $0 < p < 1$.
Dieser Vorgang läßt sich als Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ mit einer Folge $X_1, X_2, X_3, \ldots$ von stochastisch unabhängigen Zufallsvariablen modellieren, wobei X_k den Ausgang der k-ten Durchführung des Zufallsexperiments beschreibt ([7], Abschnitt V.4). Es ist $X_k = 1$, wenn bei der k-ten Durchführung das bewußte Ereignis eintritt, und $X_k = 0$ andernfalls. Die Zufallsvariablen X_k besitzen eine diskrete Verteilung mit der für alle k gleichen Wahrscheinlichkeitsfunktion $f(1) = \mathcal{P}(X_k = 1) = p$ und $f(0) = \mathcal{P}(X_k = 0) = 1 - p =: q$. Daraus ergibt sich der Erwartungswert zu $\mathcal{E}X_k = 0f(0) + 1f(1) = p$ und die Varianz zu $\mathrm{var}X_k = (0-p)^2 f(0) + (1-p)^2 f(1) = p(1-p) = pq$.
Mit dieser Definition der X_k ist die *relative Häufigkeit* des Eintretens des Ereignisses bei den ersten n Durchführungen offensichtlich durch die Zufallsvariable

$$H_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

gegeben, die nach den entsprechenden Rechenregeln den Erwartungswert

$$\mathcal{E}H_n = \frac{1}{n}\sum_{k=1}^{n} \mathcal{E}X_k = p$$

und die Varianz

$$\mathrm{var}H_n = \left(\frac{1}{n}\right)^2 \sum_{k=1}^{n} \mathrm{var}X_k = \frac{1}{n}pq$$

besitzt.

Die Ungleichung von Tschebischeff, angewandt auf diese Zufallsvariable, lautet dann

$$\mathcal{P}(|H_n - \mathcal{E}H_n| > \varepsilon) \leq \frac{\mathrm{var}H_n}{\varepsilon^2}$$

oder

$$\mathcal{P}(|H_n - p| > \varepsilon) \leq \frac{pq}{n\varepsilon^2}$$

für beliebige $\varepsilon > 0$.
Daraus folgt

$$\lim_{n\to\infty} \mathcal{P}(|H_n - p| > \varepsilon) = 0 \tag{1.13}$$

oder in Worten: Wie klein man auch ein ε wählt, wird es mit wachsender Anzahl n von Durchführungen praktisch unwahrscheinlich, daß sich die relative Häufigkeit von der Wahrscheinlichkeit des Ereignisses um mehr als ε unterscheidet.
Dies ist nicht die Art von Konvergenz, die man in der Analysis verwendet, sondern eine abgeschwächte Form, weshalb man (1.13) auch *das schwache Gesetz der großen Zahlen* nennt.
Unter den genannten Voraussetzungen läßt sich auch ein *starkes Gestz* der großen Zahlen

$$\mathcal{P}(\{\omega \ ; \ \lim_{n\to\infty} H_n(\omega) = p\}) = 1$$

herleiten, das die *fast sichere* Konvergenz der relativen Häufigkeiten gegen die Wahrscheinlichkeit beinhaltet. Wir werden in dieser Vorlesung aber nicht weiter auf diesen Konvergenzbegriff eingehen.

Ausgehend von (1.13) führen wir den folgenden Begriff ein:

Definition 1.3.1 *Eine Folge $X_1, X_2, \ldots$ von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ heißt* stochastisch konvergent *gegen eine Zufallsvariable X, wenn für alle $\varepsilon > 0$*

$$\lim_{n\to\infty} \mathcal{P}(|X_n - X| > \varepsilon) = 0$$

Als Bezeichnungen für diese Eigenschaft verwendet man meist $X = \text{p-lim}\, X_n$ oder

$$X_n \xrightarrow{p} X$$

1.3.2 Verteilungskonvergenz

Eine noch schwächere Form der Konvergenz ist die Verteilungskonvergenz, bei der nicht das Verhalten der Zufallsvariablen selbst, sonder das ihrer Verteilungen untersucht wird. Das wichtigste Beispiel hierfür ist der im ersten

Teil der Vorlesung vorgestellte *Zentrale Grenzwertsatz (ZGS)*. In seiner klassischen Form werden die Partialsummen $S_n = X_1 + X_2 + \ldots + X_n$ einer Folge $X_1, X_2, \ldots$ von stochastisch unabhängigen und identisch verteilten Zufallsvariablen mit Erwartungswerten $\mathcal{E}X_k = \mu$ und Varianzen $\mathrm{var}X_k = \sigma^2$ betrachtet. Unter diesen Voraussetzungen ist $\mathcal{E}S_n = n\mu$ und $\mathrm{var}S_n = n\sigma^2$, so daß die *normierten* Partialsummen

$$S_n^* = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$$

alle die Erwartungswerte 0 und Varianzen 1 besitzen.
Der ZGS besagt, daß für beliebige reelle Zahlen $-\infty \le a < b < \infty$ gilt

$$\lim_{n\to\infty} \mathcal{P}(a < S_n^* \le b) = \int_a^b \varphi(t)\,dt$$

oder

$$\lim_{n\to\infty} \mathcal{P}^{S_n^*}(a,b] = \lim_{n\to\infty} (F_n(b) - F_n(a)) = \Phi(b) - \Phi(a)$$

Dabei sind $\varphi(t)$ bzw. $\Phi(x)$ die Dichte bzw. Verteilungsfunktion der $\mathcal{N}(0,1)$-Verteilung und $F_n(x)$ die Verteilungsfunktion der Verteilung der Zufallsvariablen S_n^*.
Die *Verteilungen* der Zufallsvariablen S_n^* konvergieren somit gegen die $\mathcal{N}(0,1)$-Verteilung, über die Konvergenz der Zufallsvariablen selbst wird nichts ausgesagt. Man hat aber eine Aussage über das Wahrscheinlichkeitsgesetz, nach dem sich die Zufallsvariablen für große n in etwa verhalten, und dies ist für praktische Anwendungen vollkommen ausreichend.

Definition 1.3.2 *Sind $\mathcal{P}_1, \mathcal{P}_2, \ldots$ und $\mathcal{P}$ eindimensionale Verteilungen mit Verteilungsfunktionen $F_1, F_2, \ldots$ und F, so heißt die Folge der $\mathcal{P}_n$* verteilungskonvergent *gegen $\mathcal{P}$, wenn an jeder Stetigkeitsstelle t der Funktion F gilt $\lim_{n\to\infty} F_n(t) = F(t)$.*
Eine Folge von Zufallsvariablen X_n heißt verteilungskonvergent (gegen eine Zufallsvariable X), wenn die Verteilungen dieser Zufallsvariablen verteilungskonvergent (gegen die Verteilung von X) sind.

Als Abkürzung benutzt man die Bezeichnungen v-lim oder $\xrightarrow{v}$.

Verteilungskonvergenz bei diskreten Verteilungen.
Sind $\mathcal{P}_n$ für $n = 1, 2, \ldots$ und $\mathcal{P}$ Verteilungen auf $\mathbb{N}_0$ mit Wahrscheinlichkeitsfunktionen f_n bzw. f, so ist die Verteilungskonvergenz der $\mathcal{P}_n$ gegen $\mathcal{P}$ äquivalent zur punktweisen Konvergenz der entsprechenden Wahrscheinlichkeitsfunktionen:

$$\mathcal{P}_n \xrightarrow{v} \mathcal{P} \;\Leftrightarrow\; \lim_{n\to\infty} f_n(k) = f(k) \;\text{ für alle } k = 0, 1, 2, \ldots$$

Denn in diesem Fall sind die Verteilungsfunktionen F_n bzw. F Treppenfunktionen, die in den Intervallen $(K, K+1)$ mit $K = 0, 1, 2, \ldots$ konstant und damit stetig sind. Für $K < t < K+1$ gilt daher

$$\lim_{n\to\infty} F_n(t) = \lim_{n\to\infty} \sum_{k=0}^{K} f_n(k) = \sum_{k=0}^{K} \lim_{n\to\infty} f_n(k) = \sum_{k=0}^{K} f(k) = F(t)$$

Für unsere Zwecke wichtig ist das Verhalten der erzeugenden Funktionen bei Verteilungskonvergenz.

Satz 1.3.1 *Konvergiert die Folge der Wahrscheinlichkeitsfunktionen f_n punktweise gegen die Wahrscheinlichkeitsfunktion f, so konvergieren die erzeugenden Funktionen $\hat{f}_n(z)$ auf der Menge der komplexen Zahlen z mit $|z| \leq 1$ gleichmäßig gegen die erzeugende Funktion $\hat{f}(z)$.*

Beweis
a) Zu vorgegebenem $\varepsilon > 0$ gibt es ein K, so daß

$$\sum_{k=K}^{\infty} f(k) < \frac{\varepsilon}{4}$$

denn die Summe der $f(k)$ konvergiert. Weiterhin gibt es wegen der Konvergenz der Wahrscheinlichkeitsfunktionen ein n_0, so daß für $n \geq n_0$ und $k = 0, 1, \ldots K-1$

$$|f_n(k) - f(k)| < \frac{\varepsilon}{4K}$$

b) Für $n \geq n_0$ und $|z| \leq 1$ erhält man daraus die Abschätzung

$$\begin{aligned} |\hat{f}_n(z) - \hat{f}(z)| &= |\sum_{k=0}^{\infty} (f_n(k) - f(k)) z^k| \leq \sum_{k=0}^{\infty} |f_n(k) - f(k)| \\ &= \sum_{k=0}^{K-1} |f_n(k) - f(k)| + \sum_{k=K}^{\infty} |f_n(k) - f(k)| \\ &\leq \sum_{k=0}^{K-1} |f_n(k) - f(k)| + \sum_{k=K}^{\infty} f(k) + \sum_{k=K}^{\infty} f_n(k) \end{aligned}$$

Die beiden ersten Summen der letzten Zeile sind gemäß Teil a) kleiner als $\varepsilon/4$.
c) Zur Abschätzung der verbleibenden Summe verwenden wir, daß aus $\sum_{k=0}^{K-1} |f_n(k) - f(k)| < \varepsilon/4$ die Ungleichung

$$\sum_{k=0}^{K-1} f_n(k) \geq \sum_{k=0}^{K-1} f(k) - \frac{\varepsilon}{4}$$

folgt. Damit wird

$$\sum_{k=K}^{\infty} f_n(k) = 1 - \sum_{k=0}^{K-1} f_n(k) \leq 1 - \sum_{k=0}^{K-1} f(k) + \frac{\varepsilon}{4} = \sum_{k=K}^{\infty} f(k) + \frac{\varepsilon}{4} < \frac{\varepsilon}{2}$$

d) Die Teile b) und c) zusammen ergeben die Abschätzung

$$|\hat{f}_n(z) - \hat{f}(z)| < \varepsilon$$

für $n \geq n_0$ und $|z| \leq 1$, d.h. die gleichmäßige Konvergenz. ◇
Die Umkehrung dieser Aussage ist ebenfalls richtig.

Satz 1.3.2 *Seien $\hat{f}_n(z)$ erzeugende Funktionen zu Wahrscheinlichkeitsfunktionen $f_n(k)$, die für alle komplexen Zahlen z mit $|z| \leq 1$ konvergieren. Dann ist der Grenzwert $\hat{g}(z) = \lim_{n\to\infty} \hat{f}_n(z)$ erzeugende Funktion zu einer Wahrscheinlichkeitsfunktion g mit $g(k) = \lim_{n\to\infty} f_n(k)$ für alle $k \in \mathbb{N}_0$.*

Der Beweis benutzt einige Hilfsmittel aus der Funktionentheorie, die für diese Vorlesung nicht vorausgesetzt wird. Es werden daher nur die wesentlichen Beweisschritte skizziert.
Zunächst folgt aus $|\hat{f}_n(z)| \leq \hat{f}_n(1) = 1$ und dem Satz von Vitali, daß $\hat{g}(z)$ eine holomorphe Funktion mit einer Potenzreihenentwicklung

$$\hat{g}(z) = \sum_{k=0}^{\infty} g(k) z^k$$

für $|z| < 1$ ist. Mit Hilfe der Cauchyschen Integralformel zeigt man dann daß für alle k

$$g(k) = \lim_{n\to\infty} f_n(k)$$

Die Funktion g ist damit aber noch nicht notwendig eine Wahrscheinlichkeitsfunktion. Denn aus

$$\sum_{k=0}^{K} g(k) = \sum_{k=0}^{K} \lim_{n\to\infty} f_n(k) = \lim_{n\to\infty} \sum_{k=0}^{K} f_n(k) \leq \lim_{n\to\infty} \sum_{k=0}^{\infty} f_n(k) = 1$$

für alle K folgt lediglich die Ungleichung $\sum_{k=0}^{\infty} g(k) \leq 1$. Die fehlende Eigenschaft folgt aus der Voraussetzung, daß die Funktionenfolge auch für $|z| = 1$ konvergieren muß. Speziell für $z = 1$ gilt

$$\sum_{k=0}^{\infty} g(k) = \hat{g}(1) = \lim_{n\to\infty} \hat{f}_n(1) = 1$$

◇

Die Voraussetzung der Konvergenz auch für *alle* z mit $|z| = 1$ ist erforderlich. Dies zeigt das Beispiel $f_n(n) = 1$ und $f_n(k) = 0$ für $k \neq n$ mit $\hat{f}_n(z) = z^n$. Die $\hat{f}_n(z)$ konvergieren für $|z| < 1$ und $z = 1$, aber z.B. *nicht* für $z = -1$. Entsprechend ist $g(k) = \lim f_n(k) = 0$ für alle k auch keine Wahrscheinlichkeitsfunktion.

Beispiel: $\mathcal{P}_1, \mathcal{P}_2, \ldots$ sei eine Folge von Binomialverteilungen $\mathcal{B}(n, p_n)$ so, daß die Mittelwerte konstant bleiben: $np_n = \mu > 0$. Die erzeugenden Funktionen sind dann gleich

$$\hat{f}_n(z) = (p_n z + q_n)^n = \left(\frac{\mu}{n} z + 1 - \frac{\mu}{n}\right)^n = \left(1 + \frac{\mu(z-1)}{n}\right)^n$$

mit dem Limes

$$\hat{f}(z) = \lim_{n \to \infty} \hat{f}_n(z) = e^{\mu(z-1)}$$

Diese Binomialverteilungen sind also verteilungskonvergent gegen die Poissonverteilung mit dem Parameter μ.
Wie man leicht nachrechnet, ist die Aussage auch dann noch richtig, wenn $np_n = \mu_n$ mit $\lim_{n \to \infty} \mu_n = \mu$.

1.4 Verzweigungsprozesse

Mitte des 19. Jahrhunderts war einigen Leuten beim Studium der Bevölkerungsentwicklung in England aufgefallen, daß die meisten „berühmten“ Familiennamen irgendwann einmal ausgestorben waren und sie fragten sich, ob es irgendeine Gesetzmäßigkeit für die Anzahl der Generationen einer Familie gebe. *Francis Galton* stellte 1873 in der *Educational Times* dazu das folgende Problem:

> Problem 4001: A large nation, of whom we will only concern ourselves with the adult males, N in number, and who each bear separate surnames, colonise a district. Their law of population is such that, in each generation, p_0 per cent of the adult males have no male children who reach adult life; p_1 have only one such male child; p_2 have two; and so on up to p_5 who have five.
>
> Find (1) what proportion of the surnames will have become extinct after r generations; and (2) how many instances there will be of the same surname being held by m persons.

Es ging nur eine Lösung ein, die Galton aber nicht zufriedenstellte, genauer, die er als “frightful mull” bezeichnete. Er drängte daher einen Bekannten, Reverend Henry William Watson, seines Zeichens Pfarrer, Alpinist und Mathematiker, sich mit dem Problem zu beschäftigen. Dieser fand schließlich als Lösung, daß jede Familie mit Wahrscheinlichkeit Eins irgendwann einmal ausstirbt.

1.4.1 Der Ansatz von Watson

Um das Problem im Rahmen der Wahrscheinlichkeitsrechnung zu formulieren, denken wir uns einen der N Ahnherrn fest ausgewählt und betrachten die Anzahlen Z_r seiner männlichen Nachkommen in den Generationen $r = 0, 1, 2, \ldots$ als Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$. Den Prozentsätzen der Frage (1) entspricht dann die Wahrscheinlichkeit $a_r = \mathcal{P}(Z_r = 0)$, daß es keine r-te Generation mehr gibt, und Frage (2) ist in dieser Formulierung die nach dem Aussehen der Wahrscheinlichkeitsfunktionen $f_r(k) = \mathcal{P}(Z_r = k)$.
Die 0-te Generation besteht nur aus dem Ahnherrn selbst. Somit ist $Z_0 = 1$, $f_0(1) = 1$ und $f_0(k) = 0$ für $k \neq 1$ mit der erzeugenden Funktion $\hat{f}_0(z) = z$.
Für die männlichen Mitglieder $i = 1, 2, \ldots, Z_r$ der r-ten Generation seien $X_i^{(r)}$ die Anzahlen der männlichen Nachkommen. Dann ist $Z_1 = X_1^{(0)}$ und, für $r \geq 1$,

$$Z_{r+1} = X_1^{(r)} + X_2^{(r)} + \cdots + X_{Z_r}^{(r)}$$

eine Zufallsvariable von der in Formel (1.12) betrachteten Art.
Wir setzen voraus, daß die $X_i^{(s)}$ stochastisch unabhängig sind, woraus insbesondere folgt, daß für alle r die Variable Z_r von den $X_i^{(r)}$ unabhängig ist. Das Fortpflanzungsgesetz sei über alle Generationen gleich und durch die Wahrscheinlichkeitsfunktion $f(k) = \mathcal{P}(X_i^{(s)} = k) = p_k$ mit der erzeugenden Funktion

$$\hat{f}(z) = \sum_{k=0}^{\infty} p_k \, z^k$$

gegeben. Die Beschränkung auf $k \leq 5$ aus der Galtonschen Problemstellung benötigen wir nicht.
Ist $\hat{f}_r(z)$ die erzeugende Funktion der Variable Z_r, so ist wegen $Z_1 = X_1^{(0)}$ erst einmal

$$\hat{f}_1(z) = \hat{f}(z)$$

und nach Satz 1.2.3 bekommt man für $r > 1$ die Rekursionsformel

$$\hat{f}_r(z) = \hat{f}_{r-1}\left(\hat{f}(z)\right) , \tag{1.14}$$

im Einzelnen also

$$\begin{array}{rcl} \hat{f}_2(z) & = & \hat{f}\left(\hat{f}(z)\right) \\ \hat{f}_3(z) & = & \hat{f}\left(\hat{f}\left(\hat{f}(z)\right)\right) \qquad = \quad \hat{f}\left(\hat{f}_2(z)\right) \\ \hat{f}_4(z) & = & \hat{f}\left(\hat{f}\left(\hat{f}\left(\hat{f}(z)\right)\right)\right) \quad = \quad \hat{f}\left(\hat{f}_3(z)\right) , \end{array}$$

woraus man die alternative Rekursionsformel

$$\hat{f}_r(z) = \hat{f}\left(\hat{f}_{r-1}(z)\right) \tag{1.15}$$

ersieht.
Letztere kann man sich auch direkt überlegen. Wenn $i = 1, 2, \ldots, Z_1$ die Söhne des Ahnherrn und $Z_i^{(r-1)}$ die Anzahl der männlichen Nachkommen des i-ten Sohns in der $(r-1)$-ten Generation sind, so gilt

$$Z_r = Z_1^{(r-1)} + Z_2^{(r-1)} + \cdots + Z_{Z_1}^{(r-1)}$$

wobei — wenn sich das Fortpflanzungsgesetz nicht mit der Zeit ändert — die Zufallsvariablen $Z_i^{(r-1)}$ die gleiche Verteilung wie Z_{r-1} mit der erzeugenden Funktion $\hat{f}_{r-1}(z)$ besitzen. Aus dem Satz 1.2.3 erhält man danach die Formel (1.15).

Wie man sich am Beispiel $\hat{f}(z) = e^{\mu(z-1)}$ leicht überzeugt, gibt es wohl kaum eine Chance, für die Funktionen $\hat{f}_r$ eine vernünftige geschlossene Form zu finden und daraus die Verteilung der Variablen Z_r zu berechnen. Etwas günstiger sieht es bei den *Aussterbewahrscheinlichkeiten*

$$a_r = \mathcal{P}(Z_r = 0) = \hat{f}_r(0)$$

aus. Es ist $a_1 = \hat{f}(0) = p_0$ und nach (1.15)

$$a_r = \hat{f}(a_{r-1})\,. \tag{1.16}$$

Die Formel ist auch für $r = 1$ richtig, wenn man $a_0 = 0$ als Startwert benutzt. Die Ereignisse $(Z_r = 0)$ bilden offensichtlich eine aufsteigende Mengenfolge (*wenn* $Z_r = 0$*, dann ist zwangsläufig auch* $Z_{r+1} = 0$). Nach einem Satz aus der Vorlesung Wahrscheinlichkeitsrechnung 1 ist aus diesem Grunde die Folge der Zahlen $(a_r) = \mathcal{P}(Z_r = 0)$ monoton nichtfallend und konvergent mit dem Limes

$$a := \lim_{r\to\infty} a_r = \mathcal{P}(\bigcup_{r=0}^{\infty}(Z_r = 0))$$

also der Wahrscheinlichkeit des Ereignisses $\bigcup_{r=0}^{\infty}(Z_r = 0)$, daß der Familienname *irgendwann einmal* ausstirbt.
Die erzeugende Funktion $\hat{f}$ ist stetig, weshalb $\lim_{r\to\infty} \hat{f}(a_r) = \hat{f}(a)$, und man erhält aus (1.16) die Gleichung

$$a = \hat{f}(a) \tag{1.17}$$

Wie uns war auch Watson bekannt, daß für erzeugende Funktionen stets $\hat{f}(1) = 1$ gilt, und er schloß daraus, daß mit Wahrscheinlichkeit 1 jeder Familienname irgendwann einmal ausstirbt.
Galton war mit dieser Lösung nicht zufrieden, nicht weil er an dem Resultat $a = 1$ zweifelte, sondern deshalb, weil er gern eine Formel für die Funktionen $\hat{f}_r$ gehabt hätte, und er drängte Watson, doch in dieser Richtung weiter zu forschen. Der hatte aber anscheinend keine Lust mehr. Auch bei anderen Mathematikern hatte Galton keinen Erfolg und so blieb die ganze Sache liegen bis etwa in das Jahr 1929, in dem — unabhängig und offenbar in Unkenntnis der Ergebnisse von Watson — der dänische Ingenieur *Agner Krarup Erlang* sich mit dem Problem beschäftigte und zwar deshalb, weil die historisch bedeutende dänische Familie der Krarups vom Aussterben bedroht war.

1.4.2 Die Aussterbewahrscheinlichkeiten.

Erlang kam auf einem anderen Weg zu der Gleichung (1.17). In der Form

$$a \;=\; p_0 + p_1\, a + p_2\, a^2 + \cdots \tag{1.18}$$

geschrieben läßt sie sich wie folgt interpretieren:
Die Wahrscheinlichkeit a, daß ein Familienname irgendwann einmal ausstirbt, ist die Summe aus

0) der Wahrscheinlichkeit p_0, daß der Ahnherr keinen männlichen Nachkommen hat,

1) der Wahrscheinlichkeit p_1, daß er genau einen Sohn hat, multipliziert mit der Wahrscheinlichkeit a, daß die von diesem Sohn ausgehende Linie irgendwann einmal ausstirbt,

2) der Wahrscheinlichkeit p_2, daß er genau zwei Söhne hat, multipliziert mit der Wahrscheinlichkeit a, daß die Linie des ersten Sohns ausstirbt, mal der Wahrscheinlichkeit a für das Aussterben der Linie des zweiten Sohns,

und entsprechendes für die übrigen Summanden.
Im Gegensatz zu Galton und Watson untersuchte er aber die Funktion $\hat{f}$ genauer:
Im *Spezialfall* $p_0 = 1$, d.h. eines Ahnherrn, der mit Wahrscheinlichkeit 1 keine Söhne hat, sind alle anderen p_k gleich Null, so daß sich die Gleichung (1.18) auf $a = p_0 = 1$ reduziert.
Im folgenden betrachten wir nur den *Normalfall* $0 \leq p_0 < 1$. Hier gibt es mindestens ein $k > 0$ mit $p_k > 0$ und die Funktion $\hat{f}(z)$ ist im Bereich $0 \leq z \leq 1$ streng monoton steigend und konvex mit $\hat{f}(0) = p_0$ und $\hat{f}(1) = 1$. Hinsichtlich der Lösung der Gleichung (1.17) kann man die Menge der möglichen Funktionsverläufe in zwei Klassen einteilen, wie sie in der Abbildung 1.1 dargestellt sind. Ein Schnittpunkt des Graphen von $\hat{f}(z)$ mit der Diagonale $y = z$ entspricht einer Lösung a von (1.17) und man sieht, daß diese Gleichung auch eine Lösung $0 \leq a < 1$ besitzen kann.

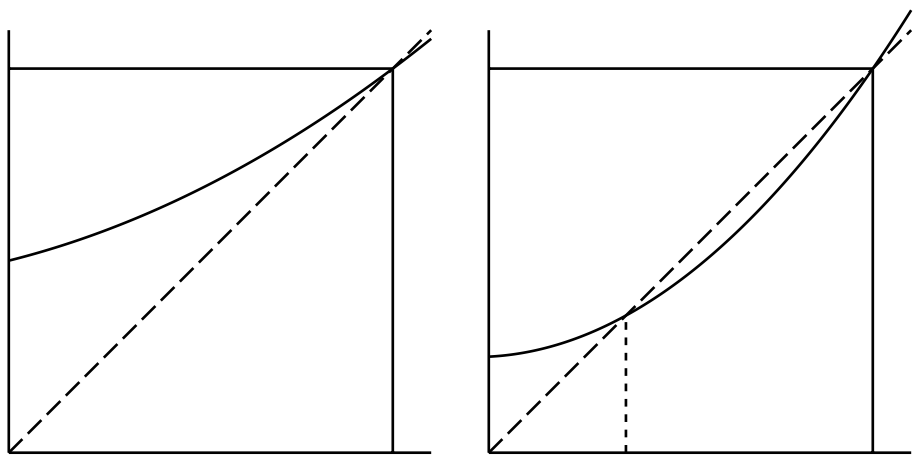


Abbildung 1.1: Lösungen der Gleichung $a = \hat{f}(a)$.

Erlang wollte wie Galton die Frage der exakten Lösung des Problems, insbesondere der Konvergenz der Folge (a_r), als Aufgabe in einer Zeitschrift—der

Matematisk Tidskrift—zur Diskussion stellen. Er starb, bevor der Zeitschriftenband herauskam.
Die Lösung wurde schließlich 1930 von Steffensen gefunden.

Satz 1.4.1 *Die Folge (a_r) ist monoton steigend.*

Beweis durch vollständige Induktion: Es ist $0 = a_0 \leq a_1 = p_0$. Wenn für irgendein $r \geq 1$ gilt $a_{r-1} \leq a_r$, so ist, weil die Funktion $\hat{f}$ monoton steigt, auch $a_r = \hat{f}(a_{r-1}) \leq \hat{f}(a_r) = a_{r+1}$. ◇

Satz 1.4.2 *Ist a irgendeine Lösung von (1.17), so gilt $a_r \leq a$ für alle r.*

Beweis durch vollständige Induktion: Es ist $a_0 = 0 \leq a$ und wenn $a_r \leq a$, so gilt $a_{r+1} = \hat{f}(a_r) \leq \hat{f}(a) = a$. ◇
Da es wenigstens eine Lösung von (1.17) gibt, nämlich 1, ist die Folge (a_r) nach oben beschränkt und besitzt daher einen Limes $\tilde{a}$. Dieser Limes ist—wie bereits diskutiert—eine Lösung von (1.17) und außerdem nach dem letzten Satz höchstens so groß wie irgendeine Lösung von (1.17), woraus folgt:

Satz 1.4.3 *Die Folge der a_r konvergiert gegen die kleinste der Lösungen der Gleichung (1.17) im Intervall* $[0, 1]$.

Für den weiteren Spezialfall, $p_0 = 0$, bei dem (1.17) die Lösungen 0 und 1 besitzt, wissen wir damit, daß $a = 0$ die *richtige* Lösung ist. Wenn es mit Wahrscheinlichkeit 0 keinen bzw. Wahrscheinlichkeit 1 mindestens einen Nachkommen gibt, muß die Wahrscheinlichkeit für das Aussterben der Familie null sein. Im folgenden wird daher vorausgesetzt, daß $0 < p_0 < 1$. Dann sind alle $p_k < 1$ und es gibt wenigstens ein $k \geq 1$ mit $p_k > 0$.
Für die Existenz einer Lösung $a < 1$ ist, wie man den Abbildungen 1.1 entnimmt, offensichtlich die Steigung der Kurve $z \longmapsto \hat{f}(z)$ im Punkt $z = 1$ entscheidend, d.h. die erste Ableitung

$$\mu = \hat{f}'(1)$$

Wie aus dem ersten Abschnitt bekannt, ist $\mu = \sum_k k\, p_k$ der Mittelwert der Verteilung mit der Wahrscheinlichkeitsfunktion $f(k) = p_k$, also die *mittlere Anzahl der männlichen Nachkommen* eines Mitglieds der Familie. Dabei sind drei Fälle zu unterscheiden.

Subkritischer Fall ($\mu < 1$): (In Abbildung 1.1 links). Hier gibt es nur die Lösung $a = 1$, d.h. mit Wahrscheinlichkeit Eins stirbt eine solche Familie irgendwann einmal aus.

Kritischer Fall ($\mu = 1$): Es gilt zwar das Gleiche wie im subkritischen Fall. Da man jedoch bei durchschnittlich einem Nachfolger intuitiv zu der Annahme neigt, daß eine solche Familie nicht aussterben kann, wird dieser Fall auch mathematisch einige Besonderheiten aufweisen, auf die wir noch zu sprechen kommen werden.

Superkritischer Fall ($\mu > 1$): (In Abbildung 1.1 rechts). In diesem Fall liegt die kleinste Lösung im Bereich $0 < a < 1$, d.h. mit positiver Wahrscheinlichkeit $1 - a$ wird eine solche Familie niemals aussterben.

Diskussion des superkritischen Falls.
Wie man aus der Skizze 1.2 ersieht, konvergiert für jeden Startwert $0 < z < 1$ die durch

$$\begin{aligned} s_0 &= z \\ s_r &= \hat{f}(s_{r-1}) \end{aligned}$$

rekursiv definierte Folge gegen a. Da $s_r = \hat{f}_r(z)$, konvergieren die Werte

$$\hat{f}_r(z) = p_0^{(r)} + p_1^{(r)}\, z + p_2^{(r)}\, z^2 + \cdots$$

der erzeugenden Funktionen also für jedes $0 < z < 1$ und $r \to \infty$ gegen die Konstante a. Dies bedeutet, daß für $k \geq 1$ die Koeffizienten

$$p_k^{(r)} = \mathcal{P}(Z_r = k)$$

gegen Null konvergieren und $\lim_{r\to\infty} p_0^{(r)} = a$. Für jede beliebige nichtnegative ganze Zahl K gilt daher

$$\begin{aligned} \mathcal{P}(Z_r > K) &= \textstyle\sum_{k>K} p_k^{(r)} \\ &= 1 - \textstyle\sum_{k=0}^{K} p_k^{(r)} \\ &\longrightarrow 1 - a \end{aligned}$$

Für $r \to \infty$ gibt es im superkritischen Fall also nur zwei Möglichkeiten: Mit Wahrscheinlichkeit a wird die Generationengröße Z_r irgendeinmal 0 und mit Wahrscheinlichkeit $1 - a$ wird sie beliebig groß.

1.4.3 Die totale Familiengröße

Das Verhalten des Verzweigungsprozesses im subkritischen und kritischen Fall läßt sich am besten an den Zufallsvariablen

$$\begin{aligned} Y_r &= Z_0 + Z_1 + \cdots + Z_r \\ &= 1 + Z_1 + \cdots + Z_r \end{aligned}$$

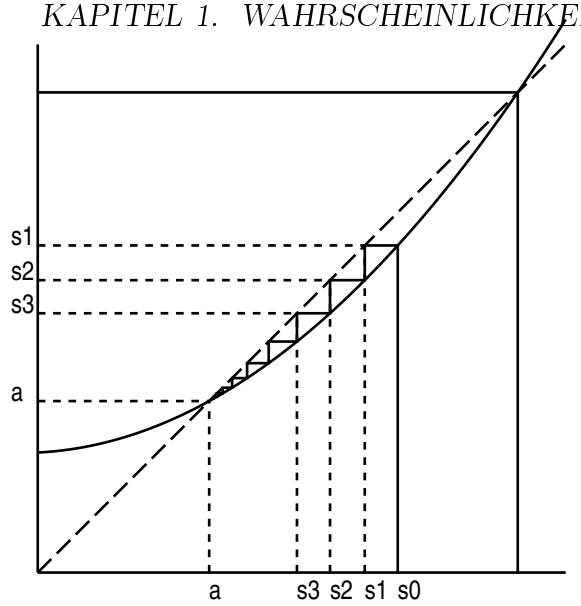


Abbildung 1.2: Konvergenz der $\hat{f}_r(z)$

ablesen, die *die Gesamtzahlen* der männlichen Mitglieder der Familie bis zur r-ten Generation einschließlich darstellen. Zur Berechnung der erzeugenden Funktionen $\hat{g}_r(z)$ der Verteilungen dieser Zufallsvariablen benötigt man die folgende Formel:

Satz 1.4.4 *Sei X eine $\mathbb{N}_0$-wertige Zufallsvariable und $Y = 1+X$. Dann gilt für die entsprechenden erzeugenden Funktionen die Beziehung*

$$\widehat{f^Y}(z) = z\,\widehat{f^X}(z) \tag{1.19}$$

Beweis: Dies läßt sich auf zwei verschiedene Arten ausrechnen:
Version 1, über Wahrscheinlichkeitsfunktionen: Es ist

$$\begin{array}{rcccl} f^Y(0) & = & \mathcal{P}(Y=0) & = & 0 \\ f^Y(k) & = & \mathcal{P}(Y=k) & = & \mathcal{P}(X=k-1) \;=\; f^X(k-1) \end{array}$$

Daher gilt für die erzeugende Funktion

$$\begin{array}{rcl} \widehat{f^Y}(z) & = & f^Y(0) + f^Y(1)\,z + f^Y(2)\,z^2 + f^Y(3)\,z^3 + \cdots \\ & = & 0 + f^X(0)\,z + f^X(1)\,z^2 + f^X(2)\,z^3 + \cdots \\ & = & z\,(f^X(0) + f^X(1)\,z + f^X(2)\,z^2 + \cdots) \\ & = & z\,\widehat{f^X}(z) \end{array}$$

Version 2, über Zufallsvariable: Die konstante Zufallsvariable $I(\omega) = 1$ ist stochastisch unabhängig von jeder anderen Zufallsvariablen und besitzt, da

sie mit Wahrscheinlichkeit 1 nur den Wert 1 annimmt, die erzeugende Funktion $\widehat{f^I}(z) = z$. Nach den Rechenregeln für erzeugende Funktionen ergibt sich daraus

$$\widehat{f^Y}(z) = \widehat{f^{I+X}}(z) = \widehat{f^I}(z)\,\widehat{f^X}(z) = z\,\widehat{f^X}(z)$$

◇

Für die Variable Y_1 erhält man damit

$$\hat{g}_1(z) = z\,\hat{f}(z) \tag{1.20}$$

Zur Bestimmung einer Rekursionsformel gehen wir wieder nach der Methode vor, die zur Rekursionsformel (1.17) führte: Für die Söhne $i = 1, 2, \ldots, Z_1$ und $r \geq 1$ sei $Y_i^{(r-1)}$ die Gesamtzahl der männlichen Mitglieder der Linie, die von diesem Sohn ausgeht, bis einschließlich der $(r-1)$-ten Generation — von Sohn i aus gerechnet. Dann gilt offensichtlich

$$Y_r \;=\; 1 + Y_1^{(r-1)} + Y_2^{(r-1)} + \cdots + Y_{Z_1}^{(r-1)}$$

Wenn sich das Fortpflanzungsgesetz im Lauf der Zeit nicht ändert, dann besitzen die $Y_i^{(r-1)}$ die gleiche Verteilung wie Y_{r-1} mit der erzeugenden Funktion $\hat{g}_{r-1}$, so daß wie im Fall (1.12) die Zufallsvariable

$$\tilde{Y}_r \;=\; Y_1^{(r-1)} + Y_2^{(r-1)} + \cdots + Y_{Z_1}^{(r-1)}$$

die erzeugende Funktion $\hat{f}\,(\hat{g}_{r-1}(z))$ und nach (1.19) $Y_r = 1 + \tilde{Y}_r$ die erzeugende Funktion

$$\hat{g}_r(z) \;=\; z\,\hat{f}\,(\hat{g}_{r-1}(z)) \tag{1.21}$$

besitzt. Mit dem Startwert $\hat{g}_0(z) = z$ ist (1.20) als Fall $r = 1$ mit in (1.21) enthalten.

Man sieht sofort, daß $\hat{g}_r(0) = 0$ für alle r gilt. Außerdem ist wie bei allen erzeugenden Funktionen $\hat{g}_r(1) = 1$ und $\hat{g}_r(z) \geq 0$ für $0 \leq z \leq 1$.

Satz 1.4.5 *Für alle* $0 < z < 1$ *ist die Folge* $(\hat{g}_r(z))$ *monoton fallend.*

Beweis durch vollständige Induktion: Da $\hat{f}(z)$ als Funktion von z in diesem Bereich monoton steigt, ist $\hat{f}(z) \leq \hat{f}(1) = 1$ und damit

$$\hat{g}_1(z) = z\,\hat{f}(z) \leq z = \hat{g}_0(z)$$

Gilt für ein $r \geq 1$, daß $\hat{g}_r(z) \leq \hat{g}_{r-1}(z)$, so folgt wieder aus der Monotonie von $\hat{f}$, daß

$$\hat{g}_{r+1}(z) = z\,\hat{f}(\hat{g}_r(z)) \leq z\,\hat{f}(\hat{g}_{r-1}(z)) = \hat{g}_r(z)$$

$\diamond$

Für alle $0 \leq z < 1$ konvergiert daher die Folge der $\hat{g}_r(z)$ gegen eine Funktion $\hat{g}(z)$, die — wie sich aus der Rekursionsformel (1.21) ergibt — der Gleichung

$$\hat{g}(z) \;=\; z\,\hat{f}\,(\hat{g}(z)) \tag{1.22}$$

genügt.

Diskussion der Funktion $\hat{\mathbf{g}}(\mathbf{z})$
Bei fest vorgegebenem z mit $0 < z < 1$ besagt die Gleichung (1.22), daß $\hat{s} = \hat{g}(z)$ eine Lösung der Gleichung

$$s \;=\; z\,\hat{f}(s)$$

im Bereich $0 \leq s \leq 1$ ist. Wie man aus der Abbildung 1.3 ersieht, gilt $\hat{s} < a$, wo a die Aussterbewahrscheinlichkeit des Verzweigungsprozesses ist. Je näher z bei 1, desto näher liegt der Graph der Funktion $s \longmapsto z\hat{f}(s)$ beim Graphen der Funktion $s \longmapsto \hat{f}(s)$ und dementsprechend die Schnittstelle $\hat{s}$ des Graphen von $z\hat{f}(s)$ mit der Diagonale bei a.

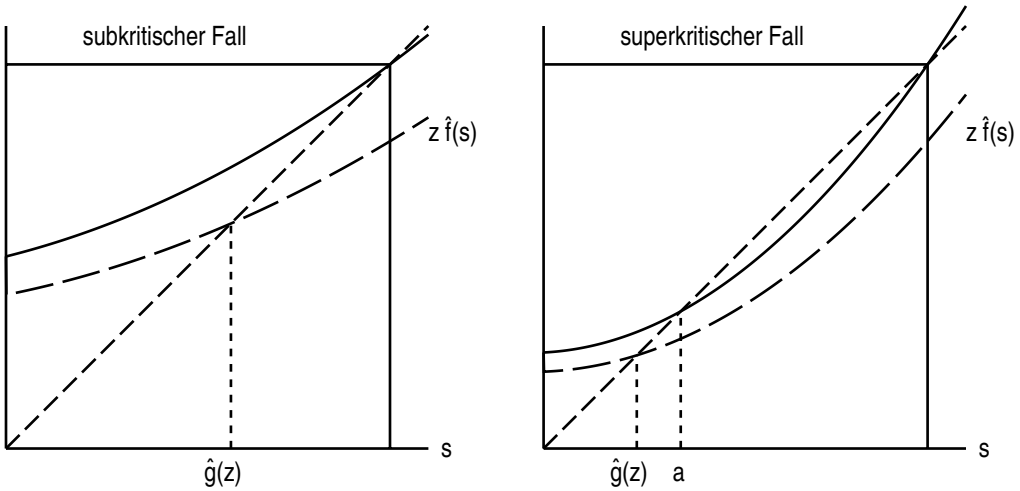


Abbildung 1.3: Die Funktion $s \longmapsto z\,\hat{f}(s)$

Diese Beobachtungen fassen wir zusammen im

Satz 1.4.6 *Ist a die Aussterbewahrscheinlichkeit des Verzweigungsprozesses, so gilt*

$$\lim_{z \to 1} \hat{g}(z) = a$$

Beweisskizze:
1) Im subkritischen und superkritischen Fall ist $\alpha := \hat{f}'(a) < 1$. Zur Abschätzung von $a - \hat{s}$ benutzen wir die Darstellung

$$\begin{aligned} a - \hat{s} &= \hat{f}(a) - z\hat{f}(\hat{s}) \\ &= (1-z)\hat{f}(a) + z(\hat{f}(a) - \hat{f}(\hat{s})) \end{aligned}$$

Nach dem Mittelwertsatz der Differentialrechnung und wegen der Monotonie der Funktion $\hat{f}'(s)$ (die zweite Ableitung ist ebenfalls nichtnegativ) gilt für eine Zwischenstelle $\hat{s} \leq s \leq a$

$$\hat{f}(a) - \hat{f}(\hat{s}) = \hat{f}'(s)(a - \hat{s}) \leq \hat{f}'(a)(a - \hat{s}) = \alpha(a - \hat{s})$$

Daraus ergibt sich die Ungleichung

$$a - \hat{s} \leq (1-z)a + z\alpha(a - \hat{s})$$

oder

$$0 \leq a - \hat{s} \leq \frac{(1-z)a}{1 - \alpha z}$$

Wegen $1 - \alpha > 0$ konvergiert für $z \to 1$ die rechte Seite gegen Null und damit $\hat{s}$ gegen a.
2) Im kritischen Fall ist $\alpha = 1$, so daß die obige Abschätzung nicht möglich ist. Es gilt aber offensichtlich $\hat{f}(s) - s \geq 0$ für alle s und $s = 1$ ist die einzige Nullstelle dieser Funktion. Aus der Abbildung 1.3 ersieht man, daß $\hat{s} \geq \hat{f}(\hat{s}) - (1 - z)$ bzw.

$$1 - z \geq \hat{f}(\hat{s}) - \hat{s} \geq 0$$

Für $z \to 1$ konvergiert daher $\hat{f}(\hat{s}) - \hat{s}$ gegen Null und infolgedessen $\hat{s}$ gegen 1. ◇

Folgerung: Es ist $\lim_{z \to 1} \hat{g}(z) = 1$ genau dann, wenn $\mu = \hat{f}'(1) \leq 1$, d.h. im subkritischen und kritischen Fall.
Zur Interpretation dieses Verhaltens der Funktion $\hat{g}$ mittels Zufallsvariablen verwenden wir den folgende Satz, den wir aber hier nicht beweisen wollen:

Satz 1.4.7 *Die Funktion $\hat{g}(z)$ besitzt im Bereich $0 \leq z < 1$ eine Potenzreihendarstellung*

$$\hat{g}(z) = \sum_{k=0}^{\infty} \rho_k z^k$$

mit Koeffizienten $\rho_k \geq 0$ und $\sum_{k=0}^{\infty} \rho_k \leq 1$.

Es ist daher gerade $\sum_{k=0}^{\infty} \rho_k = \lim_{z \to 1} \hat{g}(z)$.
Ist

$$\hat{g}_r(z) = \sum_{k=0}^{\infty} q_k^{(r)} \, z^k$$

die Potenzreihendarstellung der erzeugenden Funktion der Zufallsvariablen Y_r, so folgt aus der Konvergenz dieser erzeugenden Funktionen diejenige der Koeffizienten der entsprechenden Potenzreihen. Es gilt also

$$q_k^{(r)} \;=\; \mathcal{P}(Y_r = k) \;\stackrel{r \to \infty}{\longrightarrow}\; \rho_k$$

Falls $\sum_{k=0}^{\infty} \rho_k = 1$, ist die Funktion $k \longmapsto \rho_k$ eine Wahrscheinlichkeitsfunktion, d.h. es gibt eine Zufallsvariable $Y : \Omega \longrightarrow \mathbb{N}_0$ mit $\mathcal{P}(Y = k) = \rho_k$ und der Eigenschaft

$$\lim_{r \to \infty} \mathcal{P}(Y_r = k) = \mathcal{P}(Y = k)$$

Man sagt in einem solchen Fall, daß die Folge (Y_r) *verteilungskonvergent* gegen Y ist und symboliert das mit

$$Y = \text{v-lim}_{r \to \infty} Y_r$$

Die Zufallsvariable Y läßt sich als die *totale Familiengröße* interpretieren und aus der Tatsache, daß Y nur Werte aus dem Bereich $\mathbb{N}_0$ annimmt, folgert man, daß im hier vorliegenden Fall die totale Familiengröße mit Wahrscheinlichkeit Eins endlich ist.

Im Fall $\sum_{k=0}^{\infty} \rho_k < 1$ hat man es nicht mit einer Wahrscheinlichkeitsfunktion, sondern mit einer sog. *degenerierten Wahrscheinlichkeitsfunktion* zu tun. Hier befassen wir uns mit dem Ereignis

$$A \;=\; \bigcap_{k=0}^{\infty} \bigcup_{r=0}^{\infty} (Y_r \geq k)$$

daß es zu jeder beliebigen Unterschranke k wenigstens eine Variable Y_r gibt, deren Wert mindestens gleich k ist, oder mit anderen Worten, dem Ereignis, daß die Folge der Y_r unbeschränkt wächst.

Satz 1.4.8

$$\mathcal{P}(A) = 1 - \sum_{k=0}^{\infty} \rho_k$$

Beweis: Wegen $(Y_r \geq k) \supset (Y_r \geq k{+}1)$ bilden die Mengen $A_k = \bigcup_{r=0}^{\infty} (Y_r \geq k)$ eine absteigende Folge, so daß

$$\mathcal{P}(A) = \mathcal{P}(\bigcap_{k=0}^{\infty} A_k) = \lim_{k \to \infty} \mathcal{P}(A_k)$$

Wegen $Y_1 \leq Y_2 \leq \ldots$ gilt $(Y_r \geq k) \subset (Y_{r+1} \geq k)$, so daß

$$\begin{aligned} \mathcal{P}(A_k) &= \lim_{r\to\infty} \mathcal{P}(Y_r \geq k) \\ &= \lim_{r\to\infty} (1 - \mathcal{P}(Y_r < k)) \\ &= 1 - \lim_{r\to\infty} \sum_{i<k} q_i^{(r)} \\ &= 1 - \sum_{i<k} \rho_i \end{aligned}$$

und daraus wiederum

$$\begin{aligned} \mathcal{P}(A) &= \lim_{k\to\infty} \mathcal{P}(A_k) \\ &= 1 - \sum_{i=0}^{\infty} \rho_i \\ &> 0 \end{aligned}$$

d.h. die Folge (Y_r) und damit die totale Familiengröße ist mit positiver Wahrscheinlichkeit nicht beschränkt.

Im Fall $\sum_k \rho_k = 1$ kann man schließlich den subkritischen und kritischen Fall durch den Erwartungswert der Variablen Y, also die mittlere totale Familiengröße unterscheiden. Für $0 \leq z < 1$ erhält man mit Produkt- und Kettenregel für die Differentiation aus der Gleichung (1.22) die erste Ableitung

$$\hat{g}'(z) = \hat{f}\,(\hat{g}(z)) \;+\; z\,\hat{f}'\,(\hat{g}(z))\;\hat{g}'(z)$$

und daraus

$$\hat{g}'(z)\left[1 - z\,\hat{f}'\,(\hat{g}(z))\right] \;=\; \hat{f}\,(\hat{g}(z))$$

Wegen $z < 1$, $\hat{g}(z) < 1$ und $\hat{f}'(s) \leq \hat{f}'(1) \leq 1$ für $s \leq 1$ kann man die Gleichung durch den Term in eckigen Klammern dividieren und erhält

$$\hat{g}'(z) \;=\; \frac{\hat{f}\,(\hat{g}(z))}{1 - z\,\hat{f}'\,(\hat{g}(z))} \tag{1.23}$$

Für $z \to 1$ konvergiert $\hat{g}(z)$ im hier betrachteten Fall gegen 1 und damit $\hat{f}\,(\hat{g}(z))$ gegen 1 und $z\,\hat{f}'\,(\hat{g}(z))$ gegen $\hat{f}'(1) = \mu$. Für die mittlere totale Familiengröße $\mathcal{E}\,Y = \lim_{z\to 1} \hat{g}'(z)$ ergibt sich daraus

subkritischer Fall Wegen $\mu < 1$ ist der Erwartungswert ist endlich:

$$\mathcal{E}\,Y \;=\; \frac{1}{1-\mu}$$

kritischer Fall Mit $\mu = 1$ stirbt die Familie zwar mit Wahrscheinlichkeit Eins aus, aber die mittlere Familiengröße ist $\mathcal{E}\,Y = \infty$, d.h. die Lebensdauer der Familie ist im Mittel unendlich lange.

1.4.4 Zeitdiskrete Bedienungssysteme

Mit den bisher erarbeiteten Hilfsmitteln ist eine Analyse von Bedienungssystemen möglich, bei denen die Zeit nicht kontinuierlich sondern in Zeitintervallen fester Länge gemessen wird.

Die Zeitachse sei ab einem Punkt in Intervalle $i = 1, 2, \ldots$ gleicher Länge eingeteilt, wobei die Länge so gewählt ist, daß innerhalb eines solchen Intervalls höchstens ein Kunde ankommt. X_i mit $i = 1, 2, \ldots$ seien die Variablen, die den Ankunftsstrom beschreiben, d.h. es ist $X_i = 1$, wenn im Intervall i ein Kunde ankommt, und $X_i = 0$, wenn dies nicht der Fall ist. Wir nehmen an, daß diese Variablen stochastisch unabhängig sind. Es sei $\mathcal{P}(X_i = 1) := p$ und $\mathcal{P}(X_i = 0) := q = 1 - p$. Die erzeugende Funktion für eine solche Variable ist dann $q + pz$.

Die Kunden werden in der Reihenfolge ihres Eintreffens an einem Schalter bedient, wobei die Bedienungszeiten der einzelnen Kunden stochastisch unabhängige und identisch verteilte Zufallsvariable sind. Wir nehmen an, daß die Bedienung immer am Anfang eines Zeitintervalls beginnt und ein ganzzahliges Vielfaches der Länge eines Zeitintervalls dauert.

Es sei β_k die Wahrscheinlichkeit, daß die Bedienung genau k Zeitintervalle dauert, $\hat{b}(z) = \sum_{k=0}^{\infty} \beta_k \, z^k$ die erzeugende Funktion der Bedienungszeit und $\bar{\beta} = \hat{b}'(1)$ die mittlere Bedienungszeit.

Für unsere Analyse nehmen wir an, daß unmittelbar vor Beginn des ersten Intervalls ein Kunde angekommen ist, dessen Bedienung mit dem ersten Intervall beginnt.

Der Zusammenhang mit Verzweigungsprozessen ergibt sich durch den folgenden Ansatz:

Jeder Kunde hat als Nachkommen die Kunden, die während seiner Bedienungszeit in das System eintreten.

Das Bedienungssystem wird damit durch einen Verzweigungsprozeß mit den Variablen $Z_0, Z_1, Z_2, \ldots$ beschrieben, wo $Z_0 = 1$ für den ersten Kunden steht, der in das System eintritt, Z_1 die Anzahl der Kunden ist, die während der Bedienung des ersten Kunden eintreffen, Z_2 die Anzahl der Ankünfte in der Zeit, in der die Kunden $i = 1, 2, \ldots, Z_1$ aus der ersten Ankunftswelle bedient werden usw.

Die entscheidende Größe für die Analyse ist die erzeugende Funktion $\hat{f}(z)$ der Verteilung der Nachkommen, die gleichzeitig die erzeugende Funktion der Variablen Z_1 ist.

Wenn wir mit B die Bedienungsdauer des ersten Kunden bezeichnen, so ist

$$Z_1 = X_1 + X_2 + \cdots + X_B$$

woraus sich nach Satz 1.2.3 die Formel

$$\hat{f}(z) = \hat{b}(q + pz)$$

ergibt. Die mittlere Anzahl von Ankünften während der Bedienung eines Kunden ist dann nach der Kettenregel für die Differentiation gleich

$$\mu = \hat{f}'(1) = \hat{b}'(1)\, p = \bar{\beta}\, p \tag{1.24}$$

Die Größe $\mu = \bar{\beta}\, p$ läßt sich wie folgt anschaulich interpretieren: Ist i_0 ein Intervall, in dem ein Kunde angekommen ist, so stellt die Folge $X_{i_0+1}, X_{i_0+2}, \ldots$ eine Bernoulli-Versuchsreihe dar und die Wartezeit bis zum nächsten Eintreffen eines Kunden ist geometrisch verteilt mit Parameter p und dem Mittelwert $1/p$. Dieser Mittelwert $1/p$ ist also die mittlere Zeit, die zwischen der Ankunft zweier Kunden verstreicht. Sie wird auch als *mittlere Zwischenankunftszeit* bezeichnet.
Wegen $\bar{\beta}\, p = \frac{\bar{\beta}}{1/p}$ ist μ das Verhältnis von mittlerer Bedienungszeit und mittlerer Zwischenankunftszeit, womit die aus den obigen Formeln gezogenen Konsequenzen auch anschaulich ganz natürlich sind:

Im superkritischen Fall $\mu > 1$ ist die Anzahl der Ankünfte größer als die Anzahl der in der selben Zeit bedienten Personen. Das System ist also *instabil*. Bei $\mu < 1$ kommen im Mittel weniger Kunden an als bedient werden könnten. Das Wartesystem dürfte also ein stabiles Verhalten aufweisen. Bei $\mu = 1$ halten sich *im Mittel* Ankünfte und Abgänge die Waage. Was das für Konsequenzen für das System hat, muß noch genauer untersucht werden.

Die Beschäftigungsperiode. Zur Analyse des kritischen Falls betrachten wir als neue Größe die *Beschäftigungsperiode*, d.h. das Zeitintervall, in dem der Schalterbeamte von der Ankunft des ersten Kunden an ununterbrochen arbeiten muß, bis er mangels weiterer wartender Kunden eine Pause einlegen darf.
Falls die mittlere Dauer einer solchen Beschäftigungsperiode nicht endlich ist, bedeutet das, daß i.a. die Warteschlange niemals abgebaut wird und jede beliebige Länge erreichen kann. Für die Praxis muß man diesen Zustand wohl zu den instabilen Fällen rechnen.
Um die Beschäftigungsperiode mit den Methoden der Verzweigungsprozesse zu analysieren, gehen wir von dem folgenden Ansatz aus:
Jedes Zeitintervall, in dem der Schalter beschäftigt ist, erzeugt eine Anzahl von neuen Zeitintervallen dieser Art, die zur gleichen Beschäftigungsperiode gehören, nämlich die Bedienungszeitintervalle eines in diesem Intervall evtl. ankommenden Kunden.
Wir wollen ein Intervall, das zur Beschäftigungsperiode gehört, kurz ein *Beschäftigungsintervall* nennen.

Ist i ein solches Beschäftigungsintervall und U_i die Anzahl der Beschäftigungsintervalle, die es erzeugt, so gilt

$$\begin{aligned} p_k &:= \mathcal{P}(U_i = k) \\ &= \mathcal{P}(U_i = k|X_i = 0)\,\mathcal{P}(X_i = 0) + \mathcal{P}(U_i = k|X_i = 1)\,\mathcal{P}(X_i = 1) \end{aligned}$$

wobei unter der Bedingung $(X_i = 0)$, daß im Intervall i kein Kunde ankommt, natürlich $\mathcal{P}(U_i = 0|X_i = 0) = 1$ und $\mathcal{P}(U_i = k|X_i = 0) = 0$ für $k > 0$.
Falls ein Kunde ankommt, ist die Anzahl der erzeugten Beschäftigungsintervalle gleich der Länge der Bedienungszeit dieses Kunden, so daß

$$\mathcal{P}(U_i = k|X_i = 1) = \beta_k$$

Daraus erhält man

$$\begin{aligned} p_0 &= q + p\,\beta_0 \\ p_k &= p\,\beta_k \qquad \text{für } k > 0 \end{aligned}$$

mit der erzeugenden Funktion $\hat{f}(z) = \sum_{k=0}^{\infty} p_k\, z^k = q + p\,\hat{b}(z)$ und dem gleichen Mittelwert wie in (1.24), nämlich $\mu = \hat{f}'(1) = p\,\hat{b}'(1) = \bar{\beta}\,p$.
Im Fall $\mu \leq 1$ existiert die totale Anzahl der Beschäftigungsintervalle, die durch i entstehen, als v-Limes $Y^{(i)}$ der Variablen

$$Y_r^{(i)} = 1 + Z_1^{(i)} + \cdots + Z_r^{(i)}$$

des zugehörigen Verzweigungsprozesses mit der erzeugenden Funktion $\hat{g}(z)$, die sich aus der Gleichung (1.22) ergibt.

Zur Berechnung der Länge der Beschäftigungsperiode müssen wir beim ersten Kunden anfangen, der das System betritt. Ist B seine Bedienungszeit, so erzeugt er Beschäftigungsintervalle $i = 1, 2, \ldots, B$, von denen jedes einzelne mit einer Totalzahl $Y^{(i)}$ von Intervallen zur Beschäftigungsperiode beiträgt. Die Länge der Beschäftigungsperiode ist somit $Y = Y^{(1)} + Y^{(2)} + \cdots Y^{(B)}$ mit der erzeugenden Funktion

$$\hat{h}(z) = \hat{b}\,(\hat{g}(z))$$

Zur Bestimmung der mittleren Länge der Beschäftigungsperiode berechnen wir die Ableitung

$$\hat{h}'(z) = \hat{b}'\,(\hat{g}(z))\;\hat{g}'(z)$$

mit der Ableitung $\hat{g}'(z)$ wie in (1.23) und $\hat{f}'(z) = p\,\hat{b}'(z)$ mit

$$\hat{f}'(1) = p\,\hat{b}'(1) = \bar{\beta}\,p = \mu$$

Ist $\mu \leq 1$, so konvergiert $\hat{g}(z)$ für $z \to 1$ gegen Eins.

Im Fall $\mu = 1$, also dem kritischen Fall, gilt $\hat{g}'(z) \to \infty$ und damit auch $\hat{h}'(z) \to \infty$, wenn $z \to 1$, d.h. die mittlere Länge der Beschäftigungsperiode ist unendlich und das System instabil.

Im Fall $\mu < 1$ ist diese Länge gleich

$$\hat{h}'(1) = \lim_{z \to 1} \hat{h}'(z) = \frac{\bar{\beta}}{1 - p\,\bar{\beta}}$$

1.4.5 Literatur

Die historischen Anmerkungen wurden dem Artikel [9] von David G. Kendall entnommen, der mathematische Inhalt den Büchern [7] und [6] von Feller und Engel. Ausführliche Darstellungen der Theorie der Verzweigungsprozesse sind [2] und [14].

Kapitel 2

Markoffketten

2.1 Zeitdiskrete stochastische Prozesse

Zur mathematischen Beschreibung eines Zufallsprozesses, der zu diskreten Zeitpunkten $t_1 < t_2 < t_3 < \ldots$ beobachtet wird, gibt es zwei mögliche Ansätze. Einmal kann man ihn als eine Folge $Y_1, Y_2, Y_3 \ldots$ von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ mit Werten im Zustandsraum $\mathcal{S}$ ansehen, wobei Y_i die Beobachtung zum Zeitpunkt t_i repräsentiert. Zum anderen kann man den Prozeß als Ganzes betrachten, d.h. als Zufallsexperiment, bei dem unendliche Folgen $(s_1, s_2, s_3, \ldots)$ von Zuständen $s_i \in \mathcal{S}$ als Ergebnisse anfallen.

2.1.1 Wahrscheinlichkeiten auf Folgenräumen

In diesem Abschnitt werden Wahrscheinlichkeiten auf der Ergebnismenge

$$\mathcal{S}^\infty = \{s = (s_1, s_2, \ldots)\ ;\ s_t \in \mathcal{S}\} \tag{2.1}$$

der Folgen von Elementen s_n aus einem diskreten Zustandsraum $\mathcal{S}$ eingeführt (*Wenn es für die Darstellung zweckmäßiger ist, werden wir mit der Numerierung der Folgenelemente auch zuweilen bei* 0 *beginnen*).
Da diese Menge überabzählbar viele Elemente enthält, muß präzisiert werden, auf welcher σ-Algebra von Teilmengen dieses Ergebnisraums die Wahrscheinlichkeiten betrachtet werden. Sie wird mittels der *Zustandsvariablen*

$$X_n : \mathcal{S}^\infty \longrightarrow \mathcal{S} \tag{2.2}$$

beschrieben, die durch $X_n(s_1, s_2, \ldots) := s_n$ definiert sind.

Definition 2.1.1 *$\mathcal{Z}_\mathcal{S}$ ist die kleinste σ-Algebra auf $\mathcal{S}^\infty$, die alle Mengen der Form*

$$(X_n = k) = \{s = (s_1, s_2, \ldots)\ ;\ s_n = k\}$$

enthält. Sie wird zuweilen als die σ-Algebra der Zylindermengen *bezeichnet.*

Die am häufigsten verwendeten Ereignisse aus $\mathcal{Z}_\mathcal{S}$ sind die Mengen

$$(X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_m} = x_m) = \bigcap_{j=1}^{m} (X_{t_j} = x_j) \tag{2.3}$$

aller Folgen mit festen Werten x_j an den Komponenten $t_1 < t_2 < \ldots < t_m$.
Zur Charakterisierung von Wahrscheinlichkeiten auf $\mathcal{Z}_\mathcal{S}$ kann man sich auf noch einfachere Mengen zurückziehen.

Definition 2.1.2 *Eine Menge der Gestalt*

$$[x_1, x_2, \ldots, x_m] = (X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m) \tag{2.4}$$

heißt ein Pfad (der Länge m).

Ein Pfad ist also die Menge aller unendlichen Folgen mit einem festen Anfangsteil $(x_1, x_2, \ldots, x_m)$ aus der Menge

$$\mathcal{S}^m = \{(y_1, y_2, \ldots, y_m)\ ;\ y_j \in \mathcal{S}\} \tag{2.5}$$

Die Mengen (2.3) lassen sich als endliche Summe von Pfaden darstellen:

$$(X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_m} = x_m) = \sum_{I(x,t)} [y_1, y_2, \ldots, y_{t_m}] \tag{2.6}$$

wobei über die Menge $I(x,t)$ aller Anfangsteile $(y_1, y_2, \ldots, y_{t_m})$ mit $y_{t_j} = x_j$ für $j = 1, \ldots, m$ summiert wird. Aus diesem Grund kann man auch sagen, daß $\mathcal{Z}_\mathcal{S}$ die kleinste σ-Algebra ist, die alle Pfade enthält.
Für das folgende sind zwei Eigenschaften von Pfaden von Bedeutung:

$$\sum_{(x_1,x_2,\ldots,x_n)\in\mathcal{S}^n} [x_1, x_2, \ldots, x_n] = \mathcal{S}^\infty \tag{2.7}$$

und

$$\sum_{x\in\mathcal{S}} [x_1, x_2, \ldots, x_n, x] = [x_1, x_2, \ldots, x_n] \tag{2.8}$$

Wahrscheinlichkeiten auf $\mathcal{Z}_\mathcal{S}$.
Ist $\mathcal{P}$ eine Wahrscheinlichkeit auf $\mathcal{Z}_\mathcal{S}$ und $f_n(x_1, \ldots, x_n) := \mathcal{P}[x_1, \ldots, x_n]$, so folgt aus (2.7) und (2.8) offensichtlich, daß

$$f_n(x_1, \ldots, x_n) \geq 0 \tag{2.9}$$

$$\sum_{(x_1,\ldots,x_n)\in\mathcal{S}^n} f_n(x_1, \ldots, x_n) = 1 \tag{2.10}$$

und

$$f_n(x_1, \ldots, x_n) = \sum_{x\in\mathcal{S}} f_{n+1}(x_1, \ldots, x_n, x) \tag{2.11}$$

Die Formeln (2.9) und (2.10) besagen, daß die Funktionen f_n Wahrscheinlichkeitsfunktionen auf den Mengen $\mathcal{S}^n$ sind und (2.11) stellt den Zusammenhang zwischen den verschiedenen Dimensionen her.
Diese drei Eigenschaften charakterisieren bereits vollständig die Wahrscheinlichkeit $\mathcal{P}$, wie der folgende Satz zeigt:

Satz 2.1.1 *Für $n = 1, 2, \ldots$ seien Funktionen $f_n(x_1, \ldots, x_n)$ mit den Eigenschaften (2.9), (2.10) und (2.11) gegeben. Dann gibt es genau eine Wahrscheinlichkeit $\mathcal{P}$ auf $\mathcal{Z}_\mathcal{S}$ mit*

$$\mathcal{P}[x_1, \ldots, x_n] = f_n(x_1, \ldots, x_n)$$

Definition 2.1.3 *Eine Wahrscheinlichkeit $\mathcal{P}$ auf $\mathcal{Z}_\mathcal{S}$ heißt ein zeitdiskreter stochastischer Prozeß mit Zustandsraum $\mathcal{S}$*

Wegen (2.11) läßt sich die Forderung (2.10) vereinfachen. Man kann sie durch

$$\sum_{x \in \mathcal{S}} f_1(x) = 1 \tag{2.12}$$

ersetzen.
Wegen (2.6) gilt außerdem

$$\mathcal{P}(X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_m} = x_m) = \sum_{I(x,t)} f_{t_m}(x_1, \ldots, x_{t_m})$$

Beispiel: Unter einer *Bernoulli-Versuchsreihe* versteht man eine unendliche unabhängige Wiederholung eines Zufallsexperiments, bei dem nur die Ergebnisse 1 (*für Erfolg*) und -1 (*für Mißerfolg*) eintreten können. Eine entsprechende Wahrscheinlichkeit $\mathcal{P}$ auf $\mathcal{Z}_\mathcal{S}$ mit $\mathcal{S} = \{1, -1\}$ muß daher die Bedingung

$$\begin{array}{rcl} \mathcal{P}(X_{t_1} = \delta_1, X_{t_2} = \delta_2, \ldots, X_{t_m} = \delta_m) & = & \mathcal{P}(\bigcap_{j=1}^m (X_{t_j} = \delta_j)) \\ & = & \prod_{j=1}^m \mathcal{P}(X_{t_j} = \delta_j) \\ & = & f(\delta_1) f(\delta_2) \cdots f(\delta_m) \end{array} \tag{2.13}$$

erfüllen, wobei δ_1, δ_2, ..., δ_m aus $\{1, -1\}$ und $f(1) = \mathcal{P}(X_n = 1) =: p$ bzw. $f(-1) = \mathcal{P}(X_n = -1) = 1 - p =: q$ die für alle Einzelexperimente gleichen Wahrscheinlichkeiten für Erfolg bzw. Mißerfolg sind.
Die Funktionen $f_m(\delta_1, \delta_2, \ldots, \delta_m) = f(\delta_1) f(\delta_2) \cdots f(\delta_m)$ besitzen offensichtlich die Eigenschaften (2.9), (2.10) und (2.11), so daß es eine eindeutig bestimmte Wahrscheinlichkeit $\mathcal{P}$ mit

$$\mathcal{P}[\delta_1, \delta_2, \ldots, \delta_m] = f(\delta_1) f(\delta_2) \cdots f(\delta_m)$$

und folglich (2.13) gibt. Sie heißt eine Bernoulli-Versuchreihe mit Erfolgswahrscheinlichkeit p.
Wenn man für einen Vektor $\delta = (\delta_1, \delta_2, \ldots, \delta_m)$ mit $\#\delta$ die Anzahl der Komponenten bezeichnet, die den Wert 1 haben, so lautet (2.13)

$$\mathcal{P}(X_{t_1} = \delta_1, X_{t_2} = \delta_2, \ldots, X_{t_m} = \delta_m) = p^{\#\delta} q^{m - \#\delta}$$

Zeitdiskrete stochastische Prozesse.
Ist $Y = (Y_1, Y_2, \ldots)$ eine Folge von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ mit Werten im Zustandsraum $\mathcal{S}$, so besitzen die Funktionen

$$f_n(x_1, \ldots, x_n) = \mathcal{P}((Y_1 = x_1) \cap \ldots \cap (Y_n = x_n)) \tag{2.14}$$

die Eigenschaften (2.9), (2.11) und (2.12). Sie legen daher eine Wahrscheinlichkeit $\mathcal{P}^Y$ auf $\mathcal{Z}_\mathcal{S}$ fest, die—wie man leicht nachrechnet—die Gleichung

$$\begin{aligned} & \mathcal{P}^Y(X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_m} = x_m) \\ = \; & \mathcal{P}(Y_{t_1} = x_1, Y_{t_2} = x_2, \ldots, Y_{t_m} = x_m) \end{aligned} \tag{2.15}$$

erfüllt. $\mathcal{P}^Y$ ist eine Wahrscheinlichkeit, bezüglich der die Folge der Zustandsvariablen X_t die gleichen Eigenschaften besitzt wie die der Folge (Y_t) auf $(\Omega, \mathcal{A}, \mathcal{P})$. In Analogie zu endlichdimensionalen Verteilungen könnte man $\mathcal{P}^Y$ die Verteilung der Folge (Y_t) nennen.
Unter Inkaufnahme sprachlicher Mehrdeutigkeit nennen wir sowohl die Wahrscheinlichkeit $\mathcal{P}^Y$ auf $\mathcal{Z}_\mathcal{S}$ als auch die Folge $Y = (Y_1, Y_2, \ldots)$ einen zeitdiskreten stochastischen Prozeß mit Zustandsraum $\mathcal{S}$, wobei es sich bei den Y_t dann auch um die Zustandsvariablen X_t auf $\mathcal{S}^\infty$ handeln kann.

Beispiel: Die *eindimensionale Irrfahrt* ist ein sehr vereinfachtes Modell der Brownschen Bewegung. Ein Teilchen wandert auf den Punkten der reellen Zahlenachse mit ganzzahligen Koordinaten, indem es bei jedem Zeittakt aus der momentanen Position zufällig zum linken oder rechten Nachbarn springt. Die Folge der Sprünge ist dabei eine Bernoulli-Versuchsreihe mit der Erfolgswahrscheinlichkeit p, wobei unter *Erfolg* ein Sprung nach rechts und unter Mißerfolg ein Sprung nach links zu verstehen ist.
Ist s_0 die Position des Teilchens zum Startzeitpunkt, so ist die Position nach dem n-ten Sprung durch die Zufallsvariable

$$S_n(\delta_1, \delta_2, \ldots) = s_0 + \delta_1 + \delta_2 + \cdots + \delta_n$$

auf dieser Bernoulli-Versuchsreihe gegeben. Die Folge $S = (S_1, S_2, \ldots)$ dieser Zufallsvariablen bildet selbst wieder einen stochastischen Prozeß mit dem Zustandsraum $\mathbb{Z} = \{\ldots -2, -1, 0, 1, 2, \ldots\}$, der Menge der ganzen Zahlen.
Dieser Prozeß kann nach Konstruktion die Zustände $s_0, s_1, \ldots, s_n$ nur dann nacheinander durchlaufen, wenn die Differenzen $\delta_i = s_i - s_{i-1}$ die Werte $+1$ oder -1 besitzen und in diesem Fall ist $(S_1 = s_1, \ldots, S_n = s_n) = [\delta_1, \ldots, \delta_n]$ mit der Wahrscheinlichkeit

$$\mathcal{P}(S_1 = s_1, \ldots, S_n = s_n) = \mathcal{P}[\delta_1, \ldots, \delta_n] = f(\delta_1) f(\delta_2) \cdots f(\delta_n) \tag{2.16}$$

Für alle anderen Folgen von Zuständen ist die Wahrscheinlichkeit dieses Ereignisses Null. Mit den Größen

$$p_{ik} = \begin{cases} p & \text{falls} \quad k = i+1 \\ q & \text{falls} \quad k = i-1 \\ 0 & \text{sonst} \end{cases} \tag{2.17}$$

kann man diese Fälle zu der Formel

$$\mathcal{P}(S_1 = s_1, \ldots, S_n = s_n) = p_{s_0 s_1} p_{s_1 s_2} \ldots p_{s_{n-1} s_n}$$

zusammenfassen und gemäß (2.14) ist durch

$$\mathcal{P}^S[s_1, \ldots, s_n] = p_{s_0 s_1} p_{s_1 s_2} \ldots p_{s_{n-1} s_n} \tag{2.18}$$

ein neuer stochastischer Prozeß mit dem Zustandsraum $\mathbb{Z}$ festgelegt, den man die *eindimensionale Irrfahrt* nennt.

2.1.2 Die Markoff-Eigenschaft.

Sei $X = (X_0, X_1, \ldots)$ ein zeitdiskreter stochastischer Prozeß mit diskretem Zustandsraum $\mathcal{S}$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$.
Wie in der Einleitung erwähnt, werden in dieser Vorlesung stochastische Prozesse betrachtet, die durch eine Art Gedächtnislosigkeit ausgezeichnet sind. Das Wahrscheinlichkeitsgesetz für das Verhalten des Prozesses ab einem Zeitpunkt t hängt nur von dem zum Zeitpunkt t erreichten Zustand $X_t = i$ ab, aber nicht davon, wie der Prozeß in diesen Zustand gelangt ist. Um diese Eigenschaft mathematisch formulieren zu können, definieren wir die *Zukunft* des Prozesses bezogen auf einen Gegenwartszeitpunkt t als die Teil-σ-Algebra $\mathcal{Z}_t$, die von den Ereignissen der Form $(X_s = k)$ mit $s > t$ erzeugt wird, und die *Vergangenheit* als die Teil-σ-Algebra $\mathcal{V}_t$, die von den $(X_s = k)$ mit $s < t$ aufgespannt wird.

Definition 2.1.4 *Ein zeitdiskreter stochastischer Prozeß $\mathcal{P}$ besitzt die* Markoff-Eigenschaft*, wenn für alle Zeitpunkte $t > 0$, für alle Ereignisse $A \in \mathcal{Z}_t$, alle $B \in \mathcal{V}_t$ und alle Zustände $i \in \mathcal{S}$ mit $\mathcal{P}((X_t = i) \cap B) > 0$ gilt*

$$\mathcal{P}\left(A | (X_t = i) \cap B\right) = \mathcal{P}\left(A | (X_t = i)\right) \tag{2.19}$$

Eine andere Formulierung der Markoff-Eigenschaft besagt, daß bei bekannter Gegenwart die Vergangenheit und die Zukunft stochastisch unabhängig

sind. Eine entsprechende mathematische Formel erhält man aus (2.19) durch Umformen der linken Seite. Mit der Abkürzung $G := (X_t = i)$ erhält man

$$\mathcal{P}(A|G \cap B) = \frac{\mathcal{P}(A \cap B \cap G)}{\mathcal{P}(B \cap G)} = \frac{\mathcal{P}(A \cap B \cap G)}{\mathcal{P}(G)} \frac{\mathcal{P}(G)}{\mathcal{P}(B \cap G)} = \frac{\mathcal{P}(A \cap B|G)}{\mathcal{P}(B|G)}$$

und (2.19) wird zu

$$\frac{\mathcal{P}(A \cap B|G)}{\mathcal{P}(B|G)} = \mathcal{P}(A|G)$$

Multipliziert beide Seiten dieser Gleichung mit $\mathcal{P}(B|G)$, so erhält man die entsprechende Unabhängigkeitsaussage:

$$\mathcal{P}(A \cap B|G) = \mathcal{P}(A|G)\mathcal{P}(B|G) \tag{2.20}$$

Zum Nachweis der Markoff-Eigenschaft muß man wissen, für welche Mengen A und B die Gültigkeit von (2.19) mindestens nachgeprüft werden muß. Diese Angabe liefert der folgende Satz, der hier ohne Beweis nur zitiert wird:

Satz 2.1.2 *Der stochastische Prozeß X besitzt die Markoff-Eigenschaft genau dann, wenn für alle Zeitpunkte $t > 0$ und alle Zustände $x_0, x_1, \ldots, x_{t-1}, i, k$ mit $\mathcal{P}(X_t = i, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) > 0$ gilt*

$$\begin{array}{rl} & \mathcal{P}(X_{t+1} = k|X_t = i, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) \\ = & \mathcal{P}(X_{t+1} = k|X_t = i) \end{array} \tag{2.21}$$

Mit dieser Formel lassen sich die Pfadwahrscheinlichkeiten des Prozesses X berechnen. Seien

$$\begin{array}{rcll} p_i & = & \mathcal{P}(X_0 = i) & \text{und} \\ p_{ik}(t) & = & \mathcal{P}(X_t = k|X_{t-1} = i) & \end{array} \tag{2.22}$$

Dann gilt der

Satz 2.1.3 *Ist $X = (X_0, X_1, \ldots)$ ein zeitdiskreter stochastischer Prozeß mit dem diskreten Zustandsraum $\mathcal{S}$, der die Eigenschaft (2.21) besitzt, so gilt*

$$0 \leq p_k \leq 1 \text{ für alle } k \in \mathcal{S} \text{ und } \sum_{k \in \mathcal{S}} p_k = 1 \tag{2.23}$$

$$0 \leq p_{ik}(t) \leq 1 \text{ und } \sum_{k \in \mathcal{S}} p_{ik}(t) = 1 \tag{2.24}$$

für alle $i, k \in \mathcal{S}$ und alle $t \geq 1$, und

$$\mathcal{P}(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = p_{x_0} p_{x_0 x_1}(1) \cdots p_{x_{n-1} x_n}(n) \tag{2.25}$$

für alle Zustände $x_0, x_1, \ldots, x_n$.

Beweis: Die Ereignisse $(X_t = k)$ bilden für $k \in \mathcal{S}$ eine Partition des Ergebnisraums $\mathcal{S}^\infty$. Für die Wahrscheinlichkeiten p_k und $p_{ik}(t)$ derartiger Ereignisse gilt daher (2.23) und (2.24). Falls $\mathcal{P}(X_{t-1} = i) = 0$, kann man sich die Zahlen $p_{ik}(t)$ so wählen, daß sie (2.24) erfüllen.
Die Gleichung (2.25) beweist man durch vollständige Induktion. Es ist

$$\begin{aligned}\mathcal{P}(X_0 = x_0, X_1 = x_1) &= \mathcal{P}\left((X_1 = x_1) \cap (X_0 = x_0)\right) \\ &= \mathcal{P}\left(X_1 = x_1 | X_0 = x_0\right) \mathcal{P}(X_0 = x_0) \\ &= p_{x_0} p_{x_0 x_1}(1)\end{aligned}$$

Ist die Behauptung für die Zustände $x_0, x_1, \ldots, x_n$ richtig, so gilt

$$\begin{aligned}& \mathcal{P}(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n, X_{n+1} = x_{n+1}) \\ = \;& \mathcal{P}\left((X_{n+1} = x_{n+1}) \cap (X_n = x_n, \ldots, X_0 = x_0)\right) \\ = \;& \mathcal{P}\left(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_0 = x_0\right) \\ & \mathcal{P}\left(X_n = x_n, \ldots, X_0 = x_0\right) \\ = \;& \mathcal{P}\left(X_{n+1} = x_{n+1} | X_n = x_n\right) \cdot \mathcal{P}(X_n = x_n, \ldots, X_0 = x_0) \\ = \;& p_{x_0} p_{x_0 x_1}(1) \cdots p_{x_{n-1} x_n}(n) \cdot p_{x_n x_{n+1}}(n+1)\end{aligned}$$

◇

Sind die X_t die Zustandsvariablen auf $\mathcal{S}^\infty$, so ist

$$(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = [x_0, x_1, \ldots, x_n]$$

ein Pfad und der obige Satz kann wie folgt formuliert werden:

Satz 2.1.4 *Besitzt der stochastische Prozeß $\mathcal{P}$ auf $\mathcal{S}^\infty$ die Markoff-Eigenschaft, so sind die Pfadwahrscheinlichkeiten durch*

$$\mathcal{P}[x_0, x_1, \ldots, x_n] = p_{x_0} p_{x_0 x_1}(1) \cdots p_{x_{n-1} x_n}(n) \tag{2.26}$$

gegeben.

Umgekehrt gilt

Satz 2.1.5 *Seien p_k und $p_{ik}(t)$ reelle Zahlen mit den Eigenschaften (2.23) und (2.24). Dann wird durch (2.26) in eindeutiger Weise ein zeitdiskreter stochastischer Prozeß $\mathcal{P}$ festgelegt, der die Markoff-Eigenschaft besitzt. Insbesondere gilt (2.22) für die Zustandsvariablen.*

Beweis: **1.** Aus (2.23) und (2.24) folgen die Eigenschaften (2.9), (2.12) und (2.11) für die Funktionen

$$f_{n+1}(x_0, x_1, \ldots, x_n) = p_{x_0} p_{x_0 x_1}(1) \cdots p_{x_{n-1} x_n}(n)$$

so daß durch (2.26) ein stochastischer Prozeß definiert wird.
2. Für die Zustandsvariablen X_t ist

$$(X_{n+1} = k, X_n = i) = \sum_{x_0, x_1, \ldots, x_{n-1}} [x_0, x_1, \ldots, x_{n-1}, i, k]$$

und

$$(X_n = i) = \sum_{x_0, x_1, \ldots, x_{n-1}} [x_0, x_1, \ldots, x_{n-1}, i]$$

so daß nach (2.26)

$$\begin{aligned}
\mathcal{P}(X_{n+1} = k, X_n = i) &= \sum_{x_0, x_1, \ldots, x_{n-1}} \mathcal{P}[x_0, x_1, \ldots, x_{n-1}, i, k] \\
&= \sum_{x_0, x_1, \ldots, x_{n-1}} p_{x_0} p_{x_0 x_1}(1) \cdots p_{x_{n-1} i}(n) p_{ik}(n+1) \\
&= \left(\sum_{x_0, x_1, \ldots, x_{n-1}} p_{x_0} p_{x_0 x_1}(1) \cdots p_{x_{n-1} i}(n) \right) p_{ik}(n+1) \\
&= \mathcal{P}(X_n = i) p_{ik}(n+1)
\end{aligned}$$

Für $\mathcal{P}(X_n = i) > 0$ erhält man damit

$$p_{ik}(n+1) = \frac{\mathcal{P}(X_{n+1} = k, X_n = i)}{\mathcal{P}(X_n = i)} = \mathcal{P}(X_{n+1} = k | X_n = i)$$

3. Weiter ist

$$\begin{aligned}
& \mathcal{P}(X_{n+1} = k | X_n = i, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) \\
= \; & \frac{\mathcal{P}(X_{n+1} = k, X_n = i, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0)}{\mathcal{P}(X_n = i, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0)} \\
= \; & \frac{\mathcal{P}[x_0, \ldots, x_{n-1}, i, k]}{\mathcal{P}[x_0, \ldots, x_{n-1}, i]} \\
= \; & \frac{p_{x_0} p_{x_0 x_1}(1) \cdots p_{x_{n-1} i}(n) p_{ik}(n+1)}{p_{x_0} p_{x_0 x_1}(1) \cdots p_{x_{n-1} i}(n)} \\
= \; & p_{ik}(n+1) = \mathcal{P}(X_{n+1} = k | X_n = i)
\end{aligned}$$

und damit die Markoff-Eigenschaft (2.21) gegeben. ◇

Definition 2.1.5 *Der stochastische Prozeß $\mathcal{P}$ aus Satz 2.1.5 bzw. eine Folge $X = (X_0, X_1, \ldots)$ von Zufallsvariablen mit der Eigenschaft (2.25) heißt eine* Markoff-Kette *mit Zustandsraum $\mathcal{S}$, Übergangswahrscheinlichkeiten $p_{ik}(t)$ und Startwahrscheinlichkeiten p_k.*

2.1.3 Spezielle Markoffketten

Markoffketten mit festem Startzustand.
Ist der Zustand $i \in \mathcal{S}$ fest vorgegeben und $p_i = \mathcal{P}(X_0 = i) > 0$, so ist

$$\begin{aligned} & \mathcal{P}\left((X_1 = x_1) \cap \cdots (X_n = x_n) | (X_0 = i)\right) \\ = \; & \frac{p_i \, p_{ix_1}(1) \cdots p_{x_{n-1}x_n}(n)}{p_i} \\ = \; & p_{ix_1}(1) \cdots p_{x_{n-1}x_n}(n) \\ =: \; & \mathcal{P}_i([x_1, \ldots, x_n]) \end{aligned}$$

Dabei werden hier Pfade $[x_1, \ldots, x_n]$ aus der Menge $\mathcal{S}_1^\infty = \{s = (s_1, s_2, \ldots) \; ; \; s_t \in \mathcal{S}\}$ der Folgen betrachtet, die mit dem Index 1 starten. Es gilt der

Satz 2.1.6 *Für jedes $i \in \mathcal{S}$ ist durch*

$$\mathcal{P}_i[x_1, \ldots, x_n] = p_{in_1}(1) \cdots p_{x_{n-1}x_n}(n) \tag{2.27}$$

eine Wahrscheinlichkeit auf der σ-Algebra der Zylindermengen der Folgenmenge $\mathcal{S}_1^\infty$ gegeben.

$\mathcal{P}_i$ beschreibt das Verhalten einer Markoffkette, die im Zustand i startet. Da die Anfangsverteilung im allgemeinen eine untergeordnete Rolle spielt, wird eine Markoffkette häufig auch als eine Menge von Wahrscheinlichkeiten $\mathcal{P}_i$ der Form (2.27) definiert.
Ein Beispiel dafür stellt die eindimensionale Irrfahrt (2.18) mit der Startposition s_0 dar.

Zeitlich homogene Markoffketten

Definition 2.1.6 *Falls die Übergangswahrscheinlichkeiten $p_{ik}(t)$ einer Markoff-Kette nicht von der Zeit t abhängen, d.h. falls*

$$p_{ik}(t) = p_{ik}$$

für alle i, k und t gilt, heißt sie eine zeitlich homogene Markoffkette.

Ein Beispiel für eine zeitlich homogene Markoffkette ist wiederum die eindimensionale Irrfahrt (2.18).
Wegen der zeitlichen Homogenität ist es für die Berechnung der Pfadwahrscheinlichkeiten unerheblich, ab welchem Zeitpunkt man einen Pfad durchläuft.
Es gilt der

Satz 2.1.7 *Für alle Zeitpunkte $t \geq 0$, $r \in \mathbb{N}$ und alle Zustände $i, n_1, n_2, \ldots, n_r$ ist*

$$\begin{aligned}\mathcal{P}\left(X_{t+1} = n_1, \ldots, X_{t+r} = n_r | X_t = i\right) &= \mathcal{P}\left(X_1 = n_1, \ldots, X_r = n_r | X_0 = i\right) \\ &= \mathcal{P}_i[n_1, \ldots, n_r] \\ &= p_{in_1} p_{n_1 n_2} \cdots p_{n_{r-1} n_r}\end{aligned}$$

Beweis: Das Ereignis $A = (X_{t+1} = n_1, \ldots, X_{t+r} = n_r)$ liegt in der Zukunft $\mathcal{Z}_t$. Für einen beliebigen Pfad $B = (X_0 = x_0, \ldots, X_{t-1} = x_{t-1})$ in der Vergangenheit $\mathcal{V}_t$ folgt aus der Markoff-Eigenschaft (2.19), daß

$$\begin{aligned}\mathcal{P}(A|X_t = i) &= \mathcal{P}\left(A|(X_t = i) \cap B\right) \\ &= \frac{\mathcal{P}(A \cap (X_t = i) \cap B)}{\mathcal{P}((X_t = i) \cap B)} \\ &= \frac{\mathcal{P}[x_0, \ldots x_{t-1}, i, n_1, \ldots, n_r]}{\mathcal{P}[x_0, \ldots x_{t-1}, i]} \\ &= \frac{p_{x_0} p_{x_0 x_1} \cdots p_{x_{t-1} i} p_{in_1} \cdots p_{n_{r-1} r}}{p_{x_0} p_{x_0 x_1} \cdots p_{x_{t-1} i}} \\ &= p_{in_1} \cdots p_{n_{r-1} r} \\ &= \mathcal{P}_i[n_1, \ldots, n_r]\end{aligned}$$

◇

Stochastische Matrizen und Vektoren

Eine zeitlich homogene Markoffkette ist vollständig charakterisiert durch die für alle Zeitpunkte gleichen Übergangswahrscheinlichkeiten p_{ik} und evtl. Startwahrscheinlichkeiten p_k. Ist der Zustandsraum $\mathcal{S}$ — wie im folgenden stets bei allgemeinen Betrachtungen — eine der Mengen $\{1, 2, 3, \ldots\}$ oder $\{1, 2, 3, \ldots, n\}$, so kann man die p_{ik} in natürlicher Weise zu der Matrix

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots \\ p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

zusammenfassen. P heißt die *Matrix der Übergangswahrscheinlichkeiten* der Markoffkette oder kurz die *Übergangsmatrix*. Sie zeichnet sich dadurch aus, daß alle ihre Komponenten nichtnegativ ($p_{ik} \geq 0$) und alle Zeilensummen gleich Eins sind ($\sum_{k \in \mathcal{S}} p_{ik} = 1$).

Eine Matrix mit diesen Eigenschaften nennt man eine *stochastische Matrix* und entsprechend nennt man einen Zeilenvektor $u = (u_1, u_2, \ldots)$ mit nichtnegativen Komponenten und Komponentensumme $\sum_{i \in \mathcal{S}} u_i = 1$ einen *stochastischen Vektor*.

Satz 2.1.8 *Ist u ein stochastischer Vektor und sind P und Q stochastische Matrizen, dann ist $v = uP$ wieder ein stochastischer Vektor und $R = QP$ eine stochastische Matrix.*

Beweis: Die Komponenten von v und R sind offensichtlich nichtnegativ und

$$\sum_{k\in\mathcal{S}} v_k = \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{S}} u_i p_{ik} = \sum_{i\in\mathcal{S}} u_i \left(\sum_{k\in\mathcal{S}} p_{ik}\right) = \sum_{i\in\mathcal{S}} u_i = 1 \tag{2.28}$$

v ist daher ein stochastischer Vektor.
Wählt man als u die j-te Zeile der Matrix Q, so ist $v = uP$ die j-te Zeile der Matrix R und aus (2.28) folgt, daß die Zeilensumme Eins ist. Da dies für alle j gilt, ist R stochastisch. ◇

Beispiele für Markoffketten
Die klassischen Beispiele für Markoffketten sind einfache Modelle für physikalische Phänomene aus der statistischen Mechanik.
Irrfahrten kann man als diskretisierte Brownsche Bewegungen ansehen: Ein Teilchen macht zu festen Zeittakten $t = 1, 2, \ldots$ einen Sprung einer festen Länge in eine zufällige Richtung. Zustand X_t ist die Position des Teilchens nach dem t-ten Sprung.
Im *eindimensionalen Fall* (2.18) hat man bei Sprüngen nach links oder rechts jeweils der Länge 1 und beim Start in einer ganzzahligen Koordinate als Zustandsraum $\mathcal{S}$ die Menge der ganzen Zahlen und die Übergangswahrscheinlichkeiten

$$\mathcal{P}(X_t = k | X_{t-1} = i) = \left\{ \begin{array}{lll} p & \text{für} & k = i+1 \\ q & \text{für} & k = i-1 \\ 0 & \text{sonst} & \end{array} \right\} = p_{ik}$$

Faßt man die p_{ik} zu einer Matrix zusammen, ergibt sich

$$P = \begin{pmatrix} \ddots & \ddots & \ddots & & & & \\ & q & 0 & p & & 0 & \\ & & q & 0 & p & & \\ & 0 & & q & 0 & p & \\ & & & & \ddots & \ddots & \ddots \end{pmatrix}$$

Markoffketten mit einer so einfachen Struktur kann man sich durch einen Graphen veranschaulichen, dessen Knoten die möglichen Zustände der Markoffkette repräsentieren, wobei Knoten i und k durch einen gerichteten Bogen verbunden sind, wenn die Übergangswahrscheinlichkeit p_{ik} positiv ist. Abb. 2.1 zeigt den Graphen für die eindimensionale Irrfahrt.

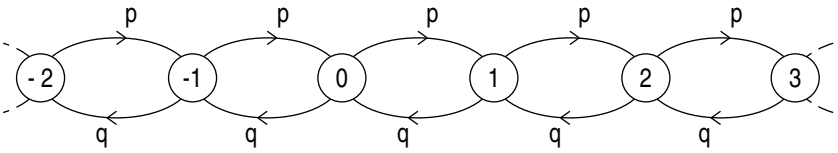


Abbildung 2.1: Die eindimensionale Irrfahrt

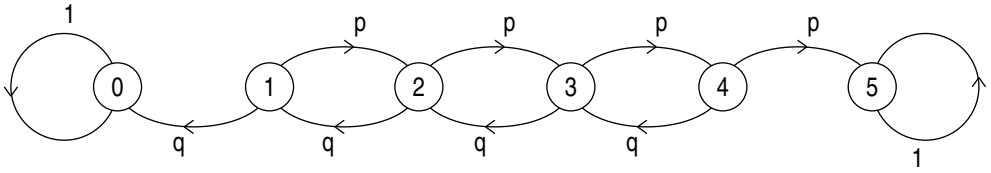


Abbildung 2.2: Irrfahrt mit absorbierenden Zuständen

Im *zweidimensionalen Fall* hat man als Zustandsraum die Menge $\mathcal{S} = \mathbb{Z}^2 = \{s = (m,n)\ ;\ m,n \in \mathbb{Z}\}$ und Übergänge zu den vier benachbarten Zuständen:

$$\mathcal{P}(X_t = s' | X_{t-1} = s) = \begin{cases} q_{10} & \text{falls} \quad s' = s + (1,0) \\ q_{-10} & \text{falls} \quad s' = s + (-1,0) \\ q_{01} & \text{falls} \quad s' = s + (0,1) \\ q_{0-1} & \text{falls} \quad s' = s + (0,-1) \\ 0 & \text{sonst} \end{cases}$$

Irrfahrten mit absorbierenden Zuständen. Falls ein Teilchen irgendwo zwischen den Positionen 0 und n startet und in letzteren bleibt, wenn es sie einmal erreicht hat, hat man (für $n = 5$) den Graphen 2.2

und die Übergangsmatrix

$$P = \begin{pmatrix} 1 & 0 & 0 & & \dots & & 0 \\ q & 0 & p & & & & \\ & q & 0 & p & & & \\ & & q & 0 & p & & \\ & & & & \ddots & & \\ & & & & q & 0 & p \\ 0 & & \dots & & 0 & 0 & 1 \end{pmatrix}$$

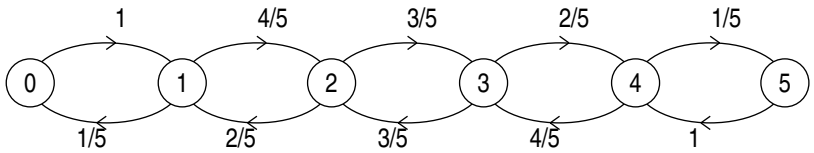


Abbildung 2.3: Ehrenfestsches Diffusionsmodell

Ehrenfestsches Diffusionsmodell. n Moleküle sind auf zwei nebeneinander stehende Behälter A und B verteilt. Die Wand zwischen den beiden Behältern ist durchlässig und innerhalb eines Zeitintervalls fester Länge Δt wechselt genau ein zufällig ausgewähltes Molekül den Behälter. Der Zustand X_t sei die Anzahl der Moleküle im Behälter A nach dem t-ten Zeitintervall. Ist $X_t = i$, so wird mit Wahrscheinlichkeit i/n das Molekül aus dem Behälter A ausgewählt und wechselt nach B. Mit Wahrscheinlichkeit $(n-i)/n$ wird es aus B gewählt und wandert nach A. Die für jedes t gleichen Übergangswahrscheinlichkeiten sind daher

$$\begin{aligned} p_{ii-1} &= \frac{i}{n} \\ p_{ii+1} &= \frac{n-i}{n} \\ p_{ik} &= 0 \qquad \text{für alle anderen } k \end{aligned}$$

mit dem für $n = 5$ in Abb. 2.3 dargestellten Graphen.

Bernoulli-Laplace-Diffusionsmodell. Hier sind auf die beiden Behälter zwei inkompressible Flüssigkeiten verteilt, wobei jede mit n Molekülen vertreten ist und in jedem der Behälter genau n Moleküle Platz finden. Durch die Anzahl i der Moleküle der Flüssigkeit 1, die sich in Behälter A befinden, ist daher die Verteilung der übrigen Moleküle auf die Behälter eindeutig festgelegt: A enthält dann noch $n - i$ von Flüssigkeit 2 und in B ist das Verhältnis genau umgekehrt. Als Zustandsraum wählt man daher die Menge $\mathcal{S} = \{0, 1, \ldots, n\}$ der möglichen Werte von i.
Wird zu den Zeitpunkten $t = 1, 2, 3, \ldots$ zufällig ein Molekül aus A und unabhängig davon eines aus B ausgewählt und in den jeweils anderen Behälter transferiert, so sind — bis auf die Fälle $i = 0$ und $i = n$ — immer die drei Zustandsübergänge $i \mapsto i-1$, $i \mapsto i$ und $i \mapsto i+1$ mit den Wahrscheinlichkeiten

$$\begin{aligned} p_{ii-1} &= \frac{i}{n} \cdot \frac{i}{n} \\ p_{ii} &= \frac{i}{n} \cdot \frac{n-i}{n} + \frac{n-i}{n} \cdot \frac{i}{n} \\ p_{ii+1} &= \frac{n-i}{n} \cdot \frac{n-i}{n} \end{aligned}$$

möglich. In der Graphenstruktur äußert sich das dahingehend, daß zusätzlich zu den Bögen der Abb. 9 noch Schleifen an den Knoten $i = 1, 2, 3, 4$ hinzukommen, die wieder in den Knoten i zurückführen.

Verzweigungsprozesse. Ein Verzweigungsprozeß ist ein zeitdiskreter stochastischer Prozeß mit der Größe der t-ten Generation als Zustand X_t, d.h. dem Zustandsraum $\mathcal{S} = \mathbb{N}_0$. Besteht die t-te Generation aus i Mitgliedern, so hängt die weitere Entwicklung offensichtlich nur noch von dieser Zahl i ab. Insbesondere ist unter dieser Voraussetzung die Größe der nächsten Generation verteilt wie eine Zufallsvariable der Form $S = X_1 + X_2 + \cdots + X_i$, wo X_j die Anzahl der Nachkommen des j-ten Mitglieds repräsentiert. Besitzen die X_j die Wahrscheinlichkeitsfunktion f, dann ist die Wahrscheinlichkeit, daß S den Wert k annimmt, gleich $f^{i*}(k)$, so daß man einen Verzweigungsprozeß als homogene Markoffkette mit den Übergangswahrscheinlichkeiten $p_{ik} = f^{i*}(k)$ und — wegen $Z_0 = 1$ — den Startwahrscheinlichkeiten $p_1 = 1$ und $p_k = 0$ für alle anderen k auffassen kann.

2.2 Homogene Markoffketten

Da in dieser Vorlesung wie auch in der Standardliteratur über dieses Thema fast ausschließlich zeitlich homogene Markoffketten behandelt werden, ist es sinnvoll, die Bezeichnungsweise zu ändern. Unter einer Markoffkette (schlechthin) ist ab sofort eine zeitlich homogene Markoffkette zu verstehen. Falls die Übergangswahrscheinlichkeiten tatsächlich einmal vom Zeitparameter t abhängen, benutzen wir die Bezeichnung „inhomogene Markoffkette“.
Von Beispielen abgesehen, bei denen sich aus dem Zusammenhang heraus ein spezieller Zustandsraum anbietet, wird im folgenden als Zustandsraum $\mathcal{S}$ die Menge der Zahlen $\{1, 2, 3, \ldots\}$ im endlichen und die Menge $\mathbb{N} = \{1, 2, 3, \ldots\}$ der natürlichen Zahlen im abzählbaren Fall verwendet. Die Markoffkette selbst ist als Wahrscheinlichkeit $\mathcal{P}$ auf $\mathcal{Z}_{\mathcal{S}}$ mit den X_t als den Zustandsvariablen zu verstehen und P bezeichnet im folgenden die Matrix der Übergangswahrscheinlichkeiten p_{ik}.

2.2.1 Die Mehrschritt-Übergangswahrscheinlichkeiten

Für $t = 1, 2, \ldots$ heißen die Wahrscheinlichkeiten

$$p_{ik}^t = \mathcal{P}_i(X_t = k) \tag{2.29}$$

für das Erreichen der Zustände k in t Schritten, wenn die Kette im Zustand i startet, die *t-Schritt-Übergangswahrscheinlichkeiten* der Markoffkette. Die Matrix $P^{(t)} = (p_{ik}^t)$ dieser Größen heißt dementsprechend die *t-Schritt-Übergangsmatrix*.
Das Ereignis $(X_t = k)$ ist die Vereinigung aller Pfade $[x_1, \ldots, x_t]$ mit $x_t = k$

$$(X_t = k) = \sum_{x_1, \ldots, x_{t-1}} [x_1, \ldots, x_{t-1}, k] \tag{2.30}$$

und — da diese Pfade paarweise disjunkt sind — die Wahrscheinlichkeit p_{ik}^t die Summe der entsprechenden Wahrscheinlichkeiten $\mathcal{P}_i[x_1, \ldots, x_t]$

$$\begin{aligned} p_{ik}^t = \mathcal{P}_i(X_t = k) &= \sum_{x_1, \ldots, x_{t-1}} \mathcal{P}_i[x_1, \ldots, x_{t-1}, k] \\ &= \sum_{x_1, \ldots, x_{t-1}} p_{ix_1} \cdots p_{x_{t-1}k} \end{aligned} \tag{2.31}$$

Für die Berechnung der t-Schritt-Übergangswahrscheinlichkeiten ergibt sich daraus:

Satz 2.2.1

$$p_{ik}^{t+1} = \sum_{j \in \mathcal{S}} p_{ij} p_{jk}^t \tag{2.32}$$

bzw.

$$P^{(t+1)} = PP^{(t)} \tag{2.33}$$

Beweis: Es ist

$$\begin{aligned} p_{ik}^{t+1} &= \sum_{x_1,\ldots,x_t} \mathcal{P}_i[x_1,\ldots,x_t,k] = \sum_{x_1,\ldots,x_t} p_{ix_1}\cdots p_{x_tk} \\ &= \sum_{x_1} p_{ix_1}\left(\sum_{x_2,\ldots,x_t} p_{x_1x_2}\cdots p_{x_tk}\right) \\ &= \sum_{x_1} p_{ix_1}\sum_{x_2,\ldots,x_t} \mathcal{P}_{x_1}[x_2,\ldots,x_t,k] = \sum_{x_1} p_{ix_1}p_{x_1k}^{t} \end{aligned}$$

Die Gleichung (2.33) stellt dann nichts anderes dar als die Zusammenfassung der Gleichungen (2.32) für die Zustände i und k in Matrixschreibweise. $\diamond$

Aus (2.33) folgt

$$P^{(t)} = \underbrace{P\cdot P\cdots P}_{t\text{-mal}} = P^t \tag{2.34}$$

und wir definieren der Vollständigkeit halber $P^{(0)} := P^0 = I$ als Einheitsmatrix, was nichts anderes bedeutet, als die Wahrscheinlichkeit p_{ik}^0, in Null Schritten von i nach k zu gelangen, für $k = i$ gleich Eins und für $k \neq i$ gleich Null zu setzen.
Aus der Rechenregel

$$P^{m+n} = P^m \cdot P^n$$

für Potenzen von Matrizen resultiert dann noch die bekannte *Gleichung von Chapman und Kolmogoroff*

Satz 2.2.2

$$p_{ik}^{m+n} = \sum_{j\in\mathcal{S}} p_{ij}^m p_{jk}^n \tag{2.35}$$

Zustandswahrscheinlichkeiten. Die Größen

$$a_k^t = \mathcal{P}(X_t = k)$$

nennen wir die *Zustandswahrscheinlichkeiten nach dem t-ten Schritt*. Sind Startwahrscheinlichkeiten $a_i = a_i^0 = \mathcal{P}(X_0 = i)$ vorgegeben, so berechnen sich die übrigen Zustandswahrscheinlichkeiten wegen $p_{ik}^t = \mathcal{P}_i(X_t = k) = \mathcal{P}(X_t = k|X_0 = i)$ mittels der Formel über die totale Wahrscheinlichkeit aus der Gleichung

$$\begin{aligned} a_k^t = \mathcal{P}(X_t = k) &= \textstyle\sum_i \mathcal{P}(X_t = k|X_0 = i)\mathcal{P}(X_0 = i) \\ &= \textstyle\sum_i a_i p_{ik}^t \end{aligned}$$

Die Zeilenvektoren

$$a = (a_1, a_2, \ldots) \quad \text{und} \quad a^t = (a_1^t, a_2^t, \ldots)$$

sind aufgrund ihrer Definitionen stochastische Vektoren. Mit ihnen kann man diese Gleichungen zu

$$a^t = aP^t \tag{2.36}$$

bzw. der Rekursionsformel

$$a^{t+1} = a^t P \tag{2.37}$$

mit $a^0 = a$ zusammenfassen.

2.2.2 Die stationäre Verteilung

Definition 2.2.1 *Ist P die Übergangsmatrix einer Markoffkette, so heißt ein stochastischer Vektor u mit der Eigenschaft $u = uP$ eine* stationäre Verteilung *der Markoffkette.*

Aus Formel (2.37) ersieht man, daß eine Markoffkette, die einmal die Zustandswahrscheinlichkeit $a^t = u$ erreicht hat, darin verharrt. Eine stationäre Verteilung beschreibt daher einen stationären oder Gleichgewichtszustand des durch die Markoffkette beschriebenen Systems.

Berechnung der stationären Verteilung.
Die Berechnung einer stationären Verteilung besteht im Prinzip einfach in der Lösung des linearen Gleichungssystems

$$\sum_{i \in \mathcal{S}} u_i p_{ik} = u_k \tag{2.38}$$

für alle $k \in \mathcal{S}$. Dadurch ist u nur bis auf eine multiplikative Konstante eindeutig bestimmt. (u ist linker Eigenvektor von P zum Eigenwert 1). Für Eindeutigkeit in dieser Hinsicht sorgt die Forderung, daß u ein stochastischer Vektor sein soll, d.h. es muß insbesondere $\sum_k u_k = 1$ gelten.
Bei überschaubaren Zustandsräumen gibt es außerdem noch eine Berechnungsmethode, die auf der Darstellung der Markoffkette als Graph beruht. Wenn man sich den Graphen einer Markoffkette als eine Art elektrisches Netzwerk vorstellt, dessen Bögen die Leitfähigkeiten p_{ik} besitzen und bei dem ein am Knoten i angelegtes „Potential“ u_i in den von i ausgehenden Bögen die Wahrscheinlichkeitsflüsse $u_i p_{ik}$ induziert, so entspricht das Gleichungssystem (2.38) den *Kirchhoffschen Knotenregeln*. Wegen $\sum_k p_{ik} = 1$ läßt sich nämlich die Gleichung $u_i = \sum_k u_k\, p_{ki}$ auch als

$$\sum_k u_i\, p_{ik} = \sum_k u_k\, p_{ki} \tag{2.39}$$

schreiben. Im Rahmen der Netzwerkinterpretation besagt diese Gleichung, daß der *gesamte aus dem Knoten i austretende Wahrscheinlichkeitsfluß gleich dem Wahrscheinlichkeitsfluß ist, der in den Knoten i hineinfließt*. Eine stationäre Verteilung $u = (u_1, u_2, \ldots)$ ist also dadurch charakterisiert, daß sie einen Gleichgewichtsfluß im Markoff-„Netzwerk“ induziert.

Bei relativ kleinen und überschaubaren Graphen ergibt sich aus dieser Interpretation eine elegante Methode zur Berechnung der stationären Verteilung aus dem sog. Schnittsatz. Zerlegt man die Knotenmenge des Graphen wie in Abb. 2.4 durch einen *Schnitt* in zwei disjunkte Teile $\mathcal{C}$ und $\bar{\mathcal{C}} = \{i \in \mathcal{S} \; ; \; i \notin \mathcal{C}\}$, so besagt dieser Satz, daß im Fall eines Gleichgewichts der gesamte Fluß, der die Trennlinie von links nach rechts passiert, gleich dem in der anderen Richtung ist.

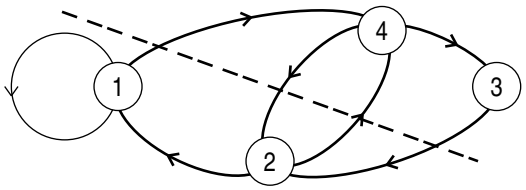


Abbildung 2.4: Der Schnittsatz

Satz 2.2.3 *Ein stochastischer Vektor $u = (u_1, u_2, \ldots)$ ist genau dann eine stationäre Verteilung, wenn für alle Teilmengen $\mathcal{C} \subset \mathcal{S}$ die Gleichung*

$$\sum_{i \in \mathcal{C}} \sum_{k \in \bar{\mathcal{C}}} u_i p_{ik} = \sum_{k \in \bar{\mathcal{C}}} \sum_{i \in \mathcal{C}} u_k p_{ki} \tag{2.40}$$

mit dem Komplement $\bar{\mathcal{C}} = \{i \in \mathcal{S} \; ; \; i \notin \mathcal{C}\}$ der Menge $\mathcal{C}$ in $\mathcal{S}$ gilt.

Beweis:
a) Ist (2.40) für alle Teilmengen erfüllt, dann insbesondere für die einelementigen Mengen $\mathcal{C} = \{i\}$:

$$\sum_{k \neq i} u_i p_{ik} = \sum_{k \neq i} u_k p_{ki}$$

Addiert man auf beiden Seiten dieser Gleichung noch $u_i p_{ii}$, so erhält man die Kirchhoffschen Knotenregeln (2.39) bzw. das Gleichungssystem (2.38).
b) Ist umgekehrt $u = (u_1, u_2, \ldots)$ eine stationäre Verteilung, so erhält man durch Aufsummieren der Gleichungen (2.39)

$$\sum_{i \in \mathcal{C}} \sum_{k \in \mathcal{S}} u_i \, p_{ik} = \sum_{i \in \mathcal{C}} \sum_{k \in \mathcal{S}} u_k \, p_{ki}$$

bzw.

$$\sum_{i\in\mathcal{C}}\sum_{k\in\bar{\mathcal{C}}} u_i p_{ik} + \sum_{i\in\mathcal{C}}\sum_{k\in\mathcal{C}} u_i p_{ik} = \sum_{i\in\mathcal{C}}\sum_{k\in\bar{\mathcal{C}}} u_k p_{ki} + \sum_{i\in\mathcal{C}}\sum_{k\in\mathcal{C}} u_k p_{ki}$$

Durch Umbenennen der Indizes und Vertauschen der Summationsreihenfolge erhält man

$$\sum_{i\in\mathcal{C}}\sum_{k\in\mathcal{C}} u_i p_{ik} = \sum_{k\in\mathcal{C}}\sum_{i\in\mathcal{C}} u_k p_{ki} = \sum_{i\in\mathcal{C}}\sum_{k\in\mathcal{C}} u_k p_{ki}$$

so daß die zweiten Summanden auf beiden Seiten der vorherigen Gleichung entfallen und (2.40) übrig bleibt. ◇

Legt man beim Ehrenfestschen oder beim Bernoulli-Laplace-Diffusionsmodell (s. Abb. 2.3) Schnitte zwischen die Knoten k und $k+1$ mit $k = 0, 1, \ldots, n-1$, so erhält man nach diesem Satz die Gleichungen

$$\begin{aligned} u_0 p_{01} &= u_1 p_{10} \\ u_1 p_{12} &= u_2 p_{21} \\ &\vdots \\ u_{k-1} p_{k-1\,k} &= u_k p_{k\,k-1} \\ &\vdots \end{aligned} \tag{2.41}$$

oder

$$\begin{aligned} u_1 &= \frac{p_{01}}{p_{10}} u_0 \\ u_2 &= \frac{p_{01}p_{12}}{p_{10}p_{21}} u_0 \\ &\vdots \\ u_k &= \frac{p_{01}p_{12}\cdots p_{k-1\,k}}{p_{10}p_{21}\cdots p_{k\,k-1}} u_0 \\ &\vdots \end{aligned}$$

Speziell für das Ehrenfestsche Modell ist

$$u_k = \frac{n(n-1)(n-2)\cdots(n-k+1)}{12\cdots k} u_0 = \left(\begin{array}{c} n \\ k \end{array} \right) u_0$$

Die Größe u_0 ist so zu bestimmen, daß $u = (u_0, u_1, \ldots, u_n)$ ein stochastischer Vektor wird. Es muß

$$1 = \sum_{k=0}^{n} u_k = u_0 \cdot \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) = u_0 \cdot 2^n$$

sein, woraus

$$u_k = \left(\begin{array}{c} n \\ k \end{array} \right) \frac{1}{2^n}$$

folgt.

Beim Bernoulli-Laplace-Modell erhält man

$$u_k = \frac{n^2(n-1)^2(n-2)^2\cdots(n-k+1)^2}{1^2 2^2 \cdots k^2} u_0 = \left(\begin{array}{c} n \\ k \end{array} \right)^2 u_0$$

Mit $u_0 = 1/\left(\begin{array}{c} 2n \\ n \end{array} \right)$ wird

$$u_k = \frac{\left(\begin{array}{c} n \\ k \end{array} \right)^2}{\left(\begin{array}{c} 2n \\ n \end{array} \right)} = \frac{\left(\begin{array}{c} n \\ k \end{array} \right)\left(\begin{array}{c} n \\ n-k \end{array} \right)}{\left(\begin{array}{c} 2n \\ n \end{array} \right)}$$

zur Wahrscheinlichkeitsfunktion einer hypergeometrischen Verteilung, die als solche die Eigenschaft $\sum_k u_k = 1$ besitzt.

Bei der Irrfahrt mit absorbierenden Zuständen (Abb. 2.2) folgt wegen $p_{01} = p_{n\,n-1} = 0$ aus (2.41), daß $u_1 = \ldots = u_{n-1} = 0$. Für die Wahrscheinlichkeiten u_0 und u_1 erhält man lediglich die Bedingung

$$\sum_{k=0}^{n} u_k = u_0 + u_n = 1$$

die diese Zahlen nicht eindeutig festlegt. Jeder stochastische Vektor der Form $(u_0, 0, \ldots, 0, u_n)$ ist eine stationäre Verteilung dieser Markoffkette.

2.2.3 Erreichbarkeit von Zuständen

Definition 2.2.2 *Ein Zustand k heißt vom Zustand i aus erreichbar, wenn es ein $t \geq 1$ gibt, so daß $p_{ik}^t > 0$.*

Diese Eigenschaft wird im folgenden mit $i \rightsquigarrow k$ symbolisiert.
Wegen

$$p_{ik}^t = \sum_{j_1, j_2 \cdots, j_{t-1}} \mathcal{P}_i[j_1 \ldots j_{t-1} k] = \sum_{j_1, j_2 \cdots, j_{t-1}} p_{ij_1} p_{j_1 j_2} \cdots p_{j_{t-1} k}$$

bedeutet $i \rightsquigarrow k$, daß es einen endlichen Pfad vom Zustand i zum Zustand k gibt, der über lauter positive p_{jl} führt.
Bei der in Abb. 2.5 dargestellten Markoffkette ist jeder der Zustände 2, 3 und 4 vom Zustand 1 aus erreichbar. 1 ist aber von keinem dieser Zustände aus erreichbar.

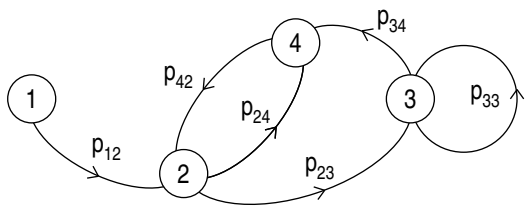


Abbildung 2.5: Zur Erreichbarkeit von Zuständen

Die Relation $\rightsquigarrow$ ist transitiv, d.h., aus $i \rightsquigarrow j$ und $j \rightsquigarrow k$ folgt $i \rightsquigarrow k$. Dies ist aus der obigen Veranschaulichung an Pfaden offensichtlich. Zum formalen Beweis benutzt man die Gleichung (2.35) von Chapman und Kolmogoroff: Es gibt ein $s \geq 1$ und ein $t \geq 1$, so daß $p_{ij}^s > 0$ und $p_{jk}^t > 0$. Daraus folgt

$$p_{ik}^{s+t} = \sum_{j' \in \mathcal{S}} p_{ij'}^s p_{j'k}^t \geq p_{ij}^s p_{jk}^t > 0$$

Definition 2.2.3 *Zwei Zustände i und k heißen miteinander verbunden, in Zeichen $i \approx k$, wenn $i \rightsquigarrow k$ und $k \rightsquigarrow i$ gilt.*

Auch die Relation $\approx$ ist transitiv. Außerdem ist sie symmetrisch: Aus $i \approx k$ folgt $k \approx i$. Sie ist aber *nicht* reflexiv, denn es gilt nicht notwendig $i \approx i$ für alle Zustände i.

Satz 2.2.4 *Der Zustandsraum $\mathcal{S}$ einer homogenen Markoff-Kette läßt sich in paarweise disjunkte Teilmengen*

$$\mathcal{S} = \mathcal{R} + \mathcal{V}_1 + \mathcal{V}_2 + \cdots$$

zerlegen, wobei die $\mathcal{V}_j$ Äquivalenzklassen bezüglich der Relation $\approx$ sind und die Restmenge $\mathcal{R}$ alle die Zustände enthält, die mit keinem anderen Zustand verbunden sind.

Satz 2.2.5 *Sind zwei Zustände i und k miteinander verbunden, so gibt es natürliche Zahlen $r \geq 1$, $s \geq 1$ und eine reelle Zahl $\gamma > 0$ so, daß für alle ganzen Zahlen $t \geq 0$*

$$p_{ii}^{r+s+t} \geq \gamma p_{kk}^t \tag{2.42}$$

und

$$p_{kk}^{r+s+t} \geq \gamma p_{ii}^t \tag{2.43}$$

Beweis: Sind i und k miteinander verbunden, so gibt es gemäß den obigen Definitionen natürliche Zahlen r und s mit $\alpha := p_{ik}^r > 0$ und $\beta := p_{ki}^s > 0$. Zweimalige Anwendung der Gleichung (2.35) von Chapman und Kolmogoroff ergibt für beliebige $t \geq 0$ und $\gamma = \alpha\beta$

$$p_{ii}^{r+s+t} = \sum_j \sum_l p_{ij}^r p_{jl}^t p_{li}^s \geq p_{ik}^r p_{kk}^t p_{ki}^s = \gamma p_{kk}^t$$

und entsprechendes für p_{kk}^{r+s+t}. ◇

2.2.4 Periodische Zustände

Definition 2.2.4 *Ein Zustand k heißt periodisch, wenn es eine natürliche Zahl $t_p > 1$ gibt, so daß $p_{kk}^{t_p} > 0$ und $p_{kk}^t = 0$ für alle Zeitpunkte t, die nicht Vielfache von t_p sind. Die größte Zahl t_p mit dieser Eigenschaft heißt die Periode von k. Gibt es keine solche Zahl, so heißt k aperiodisch.*

Beim Ehrenfestschen Diffusionsmodell kann man zu einem Zustand offensichtlich nur in einer geraden Anzahl von Übergängen wieder zurückkehren, bzw., es ist $p_{kk}^t = 0$ für $t \neq 2s$. Da $p_{kk}^{2s} > 0$ für alle s, ist $t_p = 2$ die Periode von k. Beim Bernoulli-Laplace-Modell sind alle Zustände aperiodisch, da bereits $p_{kk}^1 > 0$.
Ist der Zustand k periodisch und $t_p > 1$ eine Zahl gemäß Definition 2.2.4, so gilt umgekehrt für alle t mit $p_{kk}^t > 0$, daß t ein Vielfaches von t_p oder t_p ein Teiler von t ist. Daraus folgt, daß die Periode eines Zustands k der größte gemeinsame Teiler der Menge $\mathcal{T}_k^+ = \{t > 1 \,;\, p_{kk}^t > 0\}$ ist.

Satz 2.2.6 *Ist der Zustand i periodisch und sind i und k miteinander verbunden, so ist k periodisch mit der gleichen Periode. Ist i aperiodisch, dann auch k.*

Beweis: **1.** Für $t = 0$ liefert (2.42) $p_{ii}^{r+s} = \gamma > 0$. Daher muß, falls i die Periode t_i besitzt, $r + s$ ein Vielfaches von t_i sein: $r + s = vt_i$.
2. Ist t ein beliebiges Element aus $\mathcal{T}_k^+$, d.h. $p_{kk}^t > 0$, so folgt wiederum aus (2.42), daß $p_{ii}^{r+s+t} > 0$ und somit $r + s + t$ ein Vielfaches von t_i ist: $r + s + t = wt_i$. Aus 1. erhält man $t = (w - v)t_i$, d.h. t_i ist ein Teiler von t.
3. Da das Element t beliebig gewählt war, folgt, daß t_i ein Teiler von $\mathcal{T}_k^+$ ist. Damit ist k periodisch mit einer Periode $t_k \geq t_i$.
4. Vertauscht man in den obigen Argumenten die Zustände i und k, so folgt aus der Periodizität von k diejenige von i mit $t_i \geq t_k$, was zusammen mit $t_k \geq t_i$ aus 3. bedeutet, daß beide Zustände die gleiche Periode besitzen. ◇

2.2.5 Rekurrente und transiente Zustände

Der Zeitpunkt des erstmaligen Eintretens des Zustands k wird durch die Zufallsgröße

$$T_k(\omega) \;=\; \min\{t\,;\; t \geq 1 \text{ und } X_t(\omega) = k\}$$

auf der Menge der Folgen $\omega = (s_1, s_2, \ldots)$ von Zuständen beschrieben. Dabei wird in dem Fall, daß $X_t(\omega) \neq k$ für alle t ist, $T_k(\omega)$ der Wert ∞ zugeordnet, der die Eigenschaft besitzen soll, daß er größer als alle natürlichen Zahlen ist.
Für Zustände i und k und Zeitpunkte $t \geq 1$ sei

$$f_{ik}^t \;=\; \mathcal{P}_i(T_k = t) \tag{2.44}$$

die Wahrscheinlichkeit, daß man — ausgehend vom Zustand i — beim t-ten Übergang zum ersten Mal den Zustand k erreicht und

$$f_{ik} \;=\; \sum_{t\in\mathbb{N}} f_{ik}^t \;=\; \mathcal{P}_i(T_k < \infty) \tag{2.45}$$

die Wahrscheinlichkeit, den Zustand k irgendwann einmal vom Zustand i aus zu erreichen. Da die Ereignisse $(T_k = t)$ paarweise disjunkt sind, ist $f_{ik} \leq 1$ und

$$1 - f_{ik} \;=\; \mathcal{P}_i(T_k = \infty)$$

ist die Wahrscheinlichkeit, daß man den Zustand k von i aus niemals erreicht. Ferner sei

$$\mu_k \;=\; \sum_{t\in\mathbb{N}} t f_{kk}^t \tag{2.46}$$

Falls $f_{kk} = 1$, ist dies der Erwartungswert der Zufallsvariablen T_k bezüglich der Wahrscheinlichkeit $\mathcal{P}_k$, also die mittlere Wartezeit bis zur erstmaligen Wiederkehr des Zustands k und heißt daher auch die *mittlere Wiederkehrzeit des Zustands k*. μ_k kann auch den Wert ∞ annehmen.

Definition 2.2.5 *Ist $f_{kk} = 1$, so heißt der Zustand k rekurrent, ist $f_{kk} < 1$, heißt k transient.*
Ein rekurrenter Zustand k heißt positiv rekurrent, falls $\mu_k < \infty$, und nullrekurrent, falls $\mu_k = \infty$.

Anstelle der Begriffe positiv rekurrent und nullrekurrent sind auch die Bezeichnungen persistent und Nullzustand gebräuchlich.

Das Ereignis $(T_k = t)$ ist die Summe aller Pfade $[j_1, j_2, \ldots, j_{t-1}, k]$ mit $j_\tau \neq k$ für alle $\tau = 1, 2, \ldots, t-1$. Daher ist $f_{ik}^1 = p_{ik}$ und

$$f_{ik}^t = \sum_{j_\tau \neq k} p_{ij_1} p_{j_1 j_2} \cdots p_{j_{t-1}k} \tag{2.47}$$

Zur Charakterisierung der Zustände einer Markoff-Kette durch Eigenschaften der Übergangswahrscheinlichkeiten benötigt man den Zusammenhang zwischen den Wahrscheinlichkeiten f_{ik}^t und den Mehrschritt-Übergangswahrscheinlichkeiten p_{ik}^t. Diesen liefert der

Satz 2.2.7 *Für alle Zustände i, k und alle Zeitpunkte $t \geq 1$ gilt*

$$p_{ik}^t = \sum_{s=1}^{t} f_{ik}^s p_{kk}^{t-s} \tag{2.48}$$

Beweis:
a) Wenn zum Zeitpunkt t der Zustand k eintritt, dann muß er zu einem der Zeitpunkte $s = 1, 2, \ldots, t$ zum ersten Mal eingetreten sein. Daher ist $(X_t = k) \subset (T_k \leq t)$ und

$$\begin{aligned}(X_t = k) &= (X_t = k) \cap (T_k \leq t) \\ &= (X_t = k) \cap \textstyle\sum_{s=1}^{t} (T_k = s) \\ &= \textstyle\sum_{s=1}^{t} (X_t = k) \cap (T_k = s)\end{aligned}$$

mit der Wahrscheinlichkeit

$$p_{ik}^t = \mathcal{P}_i(X_t = k) = \sum_{s=1}^{t} \mathcal{P}_i((X_t = k) \cap (T_k = s))$$

b) Für $s = t$ ist $(T_k = t) \subset (X_t = k)$, weshalb

$$\mathcal{P}_i((X_t = k) \cap (T_k = t)) = \mathcal{P}_i(T_k = t) = f_{ik}^t = f_{ik}^t p_{kk}^0$$

Für $s < t$ ist $(X_t = k) \cap (T_k = s)$ eine Summe von Pfaden der Form $[j_1, \ldots, j_{s-1}, k, l_{s+1}, \ldots, l_{t-1}, k]$, wobei die $j_\sigma \neq k$, weil s der Zeitpunkt des ersten Auftretens von k ist, und die l_τ beliebig aus $\mathcal{S}$ sind. Mit den Formeln (2.47) und (2.29) erhält man daraus

$$\begin{aligned}\mathcal{P}_i((X_t = k) \cap (T_k = s)) &= \sum_{j_\sigma \neq k} \sum_{l_\tau \in \mathcal{S}} p_{ij_1} \cdots p_{j_{s-1}k} p_{kl_{s+1}} \cdots p_{l_{t-1}k} \\ &= \sum_{j_\sigma \neq k} p_{ij_1} \cdots p_{j_{s-1}k} \sum_{l_\tau \in \mathcal{S}} p_{kl_{s+1}} \cdots p_{l_{t-1}k} \\ &= f_{ik}^s p_{kk}^{s-t}\end{aligned}$$

woraus sich zusammen mit a) die Formel (2.48) ergibt. ◇

Eine unmittelbare Folgerung aus diesem Satz ist

Satz 2.2.8 *Der Zustand k ist vom Zustand i aus erreichbar genau dann, wenn $f_{ik} > 0$ ist.*

Beweis: Ist $f_{ik} = 0$, so sind die $f_{ik}^t = 0$ für alle $t \geq 1$. Nach Formel (2.48) ist dann auch für jedes $t \geq 1$ die t-Schritt-Übergangswahrscheinlichkeit $p_{ik}^t = 0$ und das bedeutet, daß k nicht von i aus erreichbar ist.
Ist $f_{ik} > 0$, so gibt es ein t mit $f_{ik}^t > 0$. Wegen

$$0 < f_{ik}^t = \mathcal{P}_i(T_k = t) \leq \mathcal{P}_i(X_t = k) = p_{ik}^t$$

folgt daraus $i \rightsquigarrow k$. ◇

Die rechte Seite von (2.48) unterscheidet sich von der Faltungsformel (1.10) nur dadurch, daß die Summation bei 1 beginnt, weil die f_{ik}^t nur für $t \geq 1$ definiert sind. Wenn wir hilfsweise $f_{ik}^0 := 0$ einführen, ist

$$c_{ik}^t = \sum_{s=0}^{t} f_{ik}^s p_{kk}^{t-s} = \sum_{s=1}^{t} f_{ik}^s p_{kk}^{t-s}$$

ein Faltungsprodukt und für die entsprechenden z-Transformierten

$$\begin{aligned} \hat{c}_{ik}(z) &= \sum_{t=0}^{\infty} c_{ik}^t z^t \\ \hat{f}_{ik}(z) &= \sum_{t=0}^{\infty} f_{ik}^t z^t \\ \hat{p}_{ik}(z) &= \sum_{t=0}^{\infty} p_{ik}^t z^t \end{aligned}$$

gilt die Formel (1.5):

$$\hat{c}_{ik}(z) = \hat{f}_{ik}(z) \cdot \hat{p}_{kk}(z)$$

Nach (2.48) ist $p_{ik}^t = c_{ik}^t$ für alle $t \geq 1$. Da aber stets $c_{ik}^0 = 0$, während $p_{kk}^0 = 1$ und $p_{ik}^0 = 0$ für $i \neq k$, muß man zwei Fälle unterscheiden.

Fall i ≠ k Es ist $p_{ik}^t = c_{ik}^t$ für alle t, damit $\hat{p}_{ik}(z) = \hat{c}_{ik}(z)$ oder

$$\hat{p}_{ik}(z) = \hat{f}_{ik}(z)\hat{p}_{kk}(z) \tag{2.49}$$

Fall i = k Die Potenzreihen unterscheiden sich im 0-ten Term. Es ist $\hat{p}_{kk}(z) = \hat{c}_{kk}(z) + p_{kk}^0$ und daher

$$\hat{p}_{kk}(z) = 1 + \hat{f}_{kk}(z)\hat{p}_{kk}(z) \tag{2.50}$$

Da die f_{ik}^t Werte einer Wahrscheinlichkeitsfunktion sind, gilt stets $0 < \hat{f}_{ik}(z) < 1$ für $0 < z < 1$ und (2.50) läßt sich nach $\hat{p}_{kk}(z)$ auflösen:

$$\hat{p}_{kk}(z) = \frac{1}{1 - \hat{f}_{kk}(z)} \tag{2.51}$$

Diese letzte Gleichung liefert eine Charakterisierung rekurrenter bzw. transienter Zustände durch die Mehrschritt-Übergangswahrscheinlichkeiten p_{kk}^t.

Satz 2.2.9 *Ein Zustand k ist genau dann rekurrent wenn*

$$\sum_{t=0}^{\infty} p_{kk}^t \;=\; \infty \tag{2.52}$$

und dementsprechend genau dann transient, wenn

$$\sum_{t=0}^{\infty} p_{kk}^t \;<\; \infty \tag{2.53}$$

Beweis: Wegen $\lim_{z\to 1} \hat{f}_{kk}(z) = \sum_t f_{kk}^t = f_{kk}$ folgt aus (2.51) im Fall eines transienten Zustands k ($f_{kk} < 1$), daß

$$\lim_{z\to 1} \hat{p}_{kk}(z) = \sum_t p_{kk}^t = \frac{1}{1-f_{kk}} < \infty$$

Ist k rekurrent, so ist $\lim_{z\to 1} \hat{p}_{kk}(z) = \infty$ und damit auch $\sum_t p_{kk}^t = \infty$. ◇

Als unmittelbare Folgerung hat man

Satz 2.2.10 *Ist k ein transienter Zustand, so gilt für alle Zustände i, daß*

$$\sum_{t=0}^{\infty} p_{ik}^t < \infty$$

und insbesondere

$$\lim_{t\to\infty} p_{ik}^t \;=\; 0$$

Beweis: Wegen $0 \le \hat{f}_{ik}(z) \le 1$ folgt aus (2.49) die Ungleichung $\hat{p}_{ik}(z) \le \hat{p}_{kk}(z)$ und daraus

$$\sum_t p_{ik}^t = \lim_{z\to 1} \hat{p}_{ik}(z) \le \lim_{z\to 1} \hat{p}_{kk}(z) < \infty$$

◇

Beispiel: Die eindimensionale Irrfahrt.
Der Zustandsraum ist die Menge der ganzen Zahlen und es sind nur Übergänge zu den unmittelbaren Nachbarzuständen möglich. Ein Sprung nach rechts habe die Wahrscheinlichkeit $p_{i\,i+1} = p$ und ein Sprung nach links $p_{i\,i-1} = q = 1-p$. Offensichtlich kommt man nur nach einer geraden Anzahl von Sprüngen wieder zum Ausgangszustand zurück, so daß $p_{kk}^{2n+1} = 0$ für alle n und alle k. Zur Herleitung der Übergangswahrscheinlichkeit p_{kk}^{2n} überlegt man sich, daß es sich bei $2n$ aufeinanderfolgenden Sprüngen einfach um eine Bernoulli-Versuchsreihe der Länge $2n$ mit der Erfolgswahrscheinlichkeit p

für einen Sprung nach rechts handelt, und daß dabei genau n Sprünge nach rechts und n nach links in irgendeiner Reihenfolge erforderlich sind, um nach $2n$ Schritten wieder zum Ausgangszustand zurückzukommen. Also ist

$$p_{kk}^{2n} = \left(\begin{array}{c} 2n \\ n \end{array} \right) p^n q^n = \frac{(2n)!}{n!\, n!} \left[p(1-p) \right]^n$$

Zur Abschätzung von $\sum_n p_{kk}^{2n}$ benutzen wir die Stirlingsche Formel

$$n! \sim n^{n+1/2} e^{-n} \sqrt{2\pi} \tag{2.54}$$

wobei die Bezeichnung $a_n \sim b_n$ bei reellen Zahlenfolgen für

$$\lim_{n \to \infty} \frac{a_n}{b_n} = 1 \tag{2.55}$$

steht. Ersetzt man die Fakultäten in (2.54) gemäß (2.55), so ist

$$p_{kk}^{2n} \sim \frac{(4p(1-p))^n}{\sqrt{\pi}\sqrt{n}} =: \tilde{p}_n$$

Da die Funktion $g(p) = 4p(1-p)$ an der Stelle $p = 1/2$ ihr Maximum mit $g(1/2) = 1$ annimmt, hat man für $p = 1/2$

$$\sum_n \tilde{p}_n = \sum_n \frac{1}{\sqrt{\pi}\sqrt{n}} = \infty$$

für $p \neq 1/2$ wegen $g(p) < 1$

$$\sum_n \tilde{p}_n = \sum_n \frac{g(p)^n}{\sqrt{\pi}\sqrt{n}} < \infty$$

Da wegen $p_{kk}^{2n} \sim \tilde{p}_n$ die Summe der p_{kk}^{2n} genau dann konvergiert, wenn $\sum_n \tilde{p}_n$ konvergiert, ist der Zustand k transient, wenn $p \neq 1/2$, und rekurrent, wenn $p = 1/2$.

Daß in diesem Beispiel alle Zustände vom gleichen Typ sind, ist kein Zufall, denn es gilt

Satz 2.2.11 *Sind zwei Zustände miteinander verbunden, so sind sie beide entweder transient oder rekurrent.*

Beweis: Seien i und k zwei Zustände mit $i \approx k$. Mit Hilfe des Satzes 2.2.5 erhält man

$$k \text{ rekurrent} \Rightarrow \sum_t p_{kk}^t = \infty \Rightarrow \sum_t p_{kk}^{r+s+t} = \infty \Rightarrow \sum_t p_{ii}^t = \infty \Rightarrow i \text{ rekurrent}$$

und ebenso die umgekehrte Richtung i rekurrent $\Rightarrow$ k rekurrent. Außerdem gilt

$$k \text{ transient} \Rightarrow \sum_t p_{kk}^t < \infty \Rightarrow \sum_t p_{kk}^{r+s+t} < \infty \Rightarrow \sum_t p_{ii}^t < \infty \Rightarrow i \text{ transient}$$

und die umgekehrte Richtung. ◇

Satz 2.2.12 *Eine Markoffkette mit* **endlichem** *Zustandsraum besitzt wenigstens einen rekurrenten Zustand.*

Beweis: Der Zustandsraum sei $\mathcal{S} = \{1, 2, \ldots, n\}$.
Wenn alle Zustände transient wären, müßte nach Satz 5.11 für alle i und k $\lim_{t\to\infty} p_{ik}^t = 0$ und damit auch

$$\lim_{t\to\infty} \sum_{k=1}^{n} p_{ik}^t = 0$$

sein. Dies ist aber nicht möglich, da die Matrizen P^t der t-Schritt-Übergangswahrscheinlichkeiten stochastisch sind und daher $\sum_{k=1}^n p_{ik}^t = 1$ für alle t gilt. ◇

2.2.6 Positive Rekurrenz und Nullrekurrenz

Bei einem rekurrenten Zustand k unterscheidet man nach dem Wert der mittleren Wiederkehrzeit $\mu_k = \sum_{t=1}^{\infty} f_{kk}^t$ zwei Fälle: Ist $\mu_k < \infty$, so heißt k positiv rekurrent, andernfalls nullrekurrent.
Zur Charakterisierung dieser Eigenschaften mittels der t-Schritt-Übergangswahrscheinlichkeiten benötigt man den folgenden Satz, dessen Beweis wegen seiner Länge hier nicht dargestellt wird. Er findet sich im Anhang zum Kapitel XIII von [7].

Satz 2.2.13 *Seien $f_1, f_2, \ldots$ nichtnegative reelle Zahlen mit $\sum_{t=1}^{\infty} = 1$ und $\mu = \sum_{t=1}^{\infty} t\, f_t$, wobei für μ auch der Wert ∞ zugelassen ist.*
Sei ferner $u_0, u_1, u_2, \ldots$ die durch

$$\begin{aligned} u_0 &= 1 \\ u_t &= \sum_{s=1}^{t} f_s u_{t-s} \end{aligned}$$

definierte Folge nichtnegativer reeller Zahlen. Dann gilt

a) *Ist der größte gemeinsame Teiler (GGT) N der Indizes t mit $f_t > 0$ gleich Eins, so ist* $\lim_{t\to\infty} u_t = 1/\mu$ *(mit* $1/\infty = 0$*).*

b) *Ist $N > 1$, so ist $u_t = 0$ für alle nicht durch N teilbaren Indizes t und es gilt* $\lim_{s\to\infty} u_{sN} = N/\mu$ *(mit* $N/\infty = 0$*).*

Der Zusammenhang mit Größen einer Markoffkette ist offensichtlich: f_t entspricht f_{kk}^t, μ ist die mittlere Wiederkehrzeit und u_t entspricht wegen Gleichung (2.48) der t-Schritt-Übergangswahrscheinlichkeit p_{kk}^t. Der GGT N der Folge der f_{kk}^t ist die Periode des Zustands k. Nach Aussage b) des obigen folgt, daß N die größte Zahl mit der Eigenschaft $p_{kk}^t = 0$ für alle t ist, die nicht Vielfache von N sind. Ist umgekehrt N die Periode von k, so ist N wegen $f_{kk}^t \le p_{kk}^t$ die größte Zahl mit $f_{kk}^t = 0$ für Nicht-Vielfache von N bzw. der größte Teiler der Indizes positiver f_{kk}^t.
Daher ergibt sich aus diesem Satz die folgende Charakterisierung nullrekurrenter Zustände:

Satz 2.2.14 *Ein Zustand k ist nullrekurrent genau dann, wenn*

$$\sum_{t=1}^{\infty} p_{kk}^t = \infty \quad \textit{und} \quad \lim_{t\to\infty} p_{kk}^t = 0$$

Beweis:

a) Ist der Zustand k nullrekurrent, so ist er insbesondere rekurrent. Es ist daher $\sum_{t=1}^{\infty} p_{kk}^t = \infty$ und $\sum_{t=1}^{\infty} f_{kk}^t = 1$. Auf die f_{kk}^t kann der obige Satz angewandt werden. Wegen der Nullrekurrenz von k ist die mittlere Wiederkehrzeit ∞ und aus Aussage a) bzw. b) folgt $\lim_{t\to\infty} p_{kk}^t = 0$.

b) Aus $\sum_{t=1}^{\infty} p_{kk}^t = \infty$ folgt, daß k rekurrent ist, und aus $\lim_{t\to\infty} p_{kk}^t = 0$ nach dem obigen Satz, daß die mittlere Wiederkehrzeit unendlich ist.

◇

Wie bei transienten Zuständen gilt

Satz 2.2.15 *Ist k nullrekurrent, so ist für alle $i \in \mathcal{S}$*

$$\lim_{t\to\infty} p_{ik}^t = 0$$

Beweis: Es ist stets $\sum_{t=1}^{\infty} f_{ik}^t \le 1$ und wegen der Nullrekurrenz $\lim_{t\to\infty} p_{kk}^t = 0$.
Zu einem vorgegebenem $\varepsilon > 0$ gibt es daher einen Index t_0 so, daß $\sum_{t>t_0} f_{ik}^t \le \varepsilon/2$ und $p_{kk}^t < \varepsilon/2$ für $t > t_0$.

Für $t > t_1 = 2t_0$ erhält man aus Gleichung (2.48) die folgende Abschätzung

$$\begin{aligned} p_{ik}^t &= \sum_{s=1}^{t_0} f_{ik}^s p_{kk}^{t-s} + \sum_{s=t_0+1}^{t} f_{ik}^s p_{kk}^{t-s} \\ &\leq \max\{p_{kk}^{t-1}, \ldots, p_{kk}^{t-t_0}\} \sum_{s=1}^{t_0} f_{ik}^s \; + \; \sum_{s=t_0+1}^{t} f_{ik}^s \\ &\leq \varepsilon/2 + \varepsilon/2 = \varepsilon \end{aligned}$$

und damit die Aussage der Konvergenz gegen Null. ◇
Ferner gilt

Satz 2.2.16 *Ist i nullrekurrent und $i \approx k$, so ist auch k nullrekurrent.*

Beweis: Satz 2.2.11 liefert bereits, daß k rekurrent ist und aus der Ungleichung (2.42) folgt, daß $\lim_{t\to\infty} p_{kk}^t = 0$, wenn $\lim_{t\to\infty} p_{ii}^t = 0$. ◇

2.2.7 Die Struktur des Zustandsraums

Für die Zerlegung des Zustandsraums $\mathcal{S}$ der Markoffkette nach Satz 2.2.4 ergibt sich aus dem vorhergehenden Abschnitt, daß die Äquivalenzklassen $\mathcal{V}_i$ jeweils aus Zuständen des gleichen Typs bestehen: gleiche Periode und alle positiv rekurrent, alle nullrekurrent oder alle transient. Zur Klassifizierung der Zustände aus der Restklasse $\mathcal{R}$ benötigen wir noch einen weiteren Satz, dessen Beweis hier nicht vorgeführt wird. Die Herleitung findet man grob skizziert in [4].

Satz 2.2.17 *Ist der Zustand i rekurrent und ist k von i aus erreichbar ($i \rightsquigarrow k$), dann gilt*

1) i ist mit k verbunden: $i \approx k$

2) $f_{ik} = f_{ki} = 1$

Für die Struktur des Zustandsraums einer Markoffkette, wie sie in Satz 2.2.4 dargestellt wurde, ergeben sich aus diesem Satz einige genauere Aussagen:

Satz 2.2.18 *Die Restmenge $\mathcal{R}$ enthält keine rekurrenten Zustände.*

Beweis:
a) Ist $i \in \mathcal{R}$, so gibt es wenigstens ein k mit $i \rightsquigarrow k$. Wenn nämlich kein k von i aus erreichbar ist, so muß $p_{ik} = 0$ für alle $k \neq i$ und damit $p_{ii} = 1$ sein. Letzteres bedeutet aber, daß $i \approx i$, d.h. liegt in einer der Äquivalenzklassen $\mathcal{V}_r$ und nicht in $\mathcal{R}$.

b) Angenommen, i wäre rekurrent und $i \rightsquigarrow k$. Dann ist nach obigem Satz $i \approx k$ und i würde deshalb wieder in einer der Äquivalenzklassen $\mathcal{V}_r$ und nicht in $\mathcal{R}$ liegen. $\diamond$

Wenn wir die Menge $\mathcal{R}$ und alle Äquivalenzklassen $\mathcal{V}_j$, die aus transienten Zuständen bestehen, zur Menge $\mathcal{T}$ der *transienten* Zustände zusammenfassen und die übrigen Äquivalenzklassen mit rekurrenten Zuständen in $\mathcal{R}_j$ umbenennen, erhalten wir die Zerlegung

$$\mathcal{S} = \mathcal{T} + \mathcal{R}_1 + \mathcal{R}_2 + \cdots \tag{2.56}$$

für den Zustandsraum. Hier gilt

Satz 2.2.19 *Ist $i \in \mathcal{R}_j$ und $k \notin \mathcal{R}_j$, so ist $p_{ik} = 0$.*

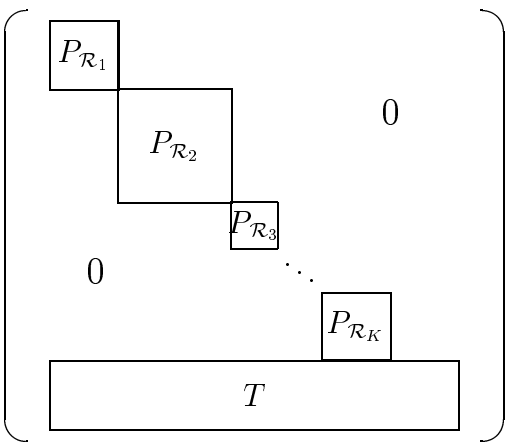
Beweis: Wenn $p_{ik} > 0$, ist $i \rightsquigarrow k$. Da $i \in \mathcal{R}_j$ rekurrent, gilt $i \approx k$ und damit $k \in \mathcal{R}_j$. $\diamond$

Struktur der Übergangsmatrix

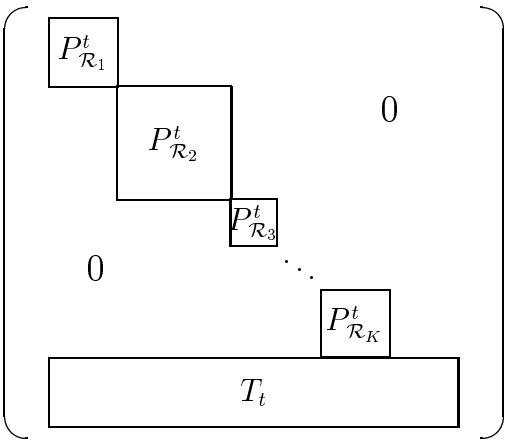
Die allgemeine Struktur der Übergangsmatrix, die sich aus diesem Satz ergibt, läßt sich im Fall eines endlichen Zustandsraums $\mathcal{S} = \{1, 2, \ldots, n\}$ besonders gut veranschaulichen. Hier gibt es nur endlich viele Klassen $\mathcal{R}_1, \ldots, \mathcal{R}_K$ und alle Mengen sind endlich, so daß man durch Umnumerieren die Situation

$$\begin{array}{rcl} \mathcal{R}_1 & = & \{1, \ldots, r_1\} \\ \mathcal{R}_2 & = & \{r_1 + 1, \ldots, r_2\} \\ & \vdots & \\ \mathcal{R}_K & = & \{r_{K-1} + 1, \ldots, r_K\} \\ \mathcal{T} & = & \{r_K, \ldots, n\} \end{array}$$

erreichen kann. Die Matrix P der Übergangswahrscheinlichkeiten hat dann nach den obigen Resultaten die Form



Dabei sind die $P_{\mathcal{R}_j}$ stochastische Matrizen und außerhalb der eingezeichneten Blöcke stehen nur Werte $p_{ik} = 0$. Daher bleibt diese Struktur auch erhalten, wenn man die Übergangsmatrix mit sich selbst multipliziert. Die t-te Potenz ist von der Form



mit den t-ten Potenzen der jeweiligen stochastischen Blockmatrizen.

Satz 2.2.20 *In einer Markoffkette mit* **endlichem** *Zustandsraum gibt es keine nullrekurrenten Zustände.*

Beweis: Angenommem, es gebe einen nullrekurrenten Zustand i. Nach Satz 2.2.16 sind alle Zustände aus der Äquivalenzklasse $\mathcal{R}$, in der i liegt, ebenfalls nullrekurrent. Da die Menge $\mathcal{R}$ endlich ist, wäre nach Satz 2.2.15 für alle $j \in \mathcal{R}$

$$\lim_{t\to\infty} \sum_{k\in\mathcal{R}} p^t_{jk} = \sum_{k\in\mathcal{R}} \lim_{t\to\infty} p^t_{jk} = 0$$

Dies ist aber nicht möglich, denn der Matrixblock $P_{\mathcal{R}}$ bzw. $P^t_{\mathcal{R}}$ ist stochastisch, so daß stets

$$\sum_{k\in\mathcal{R}} p^t_{jk} = 1$$

$\diamond$

2.3 Absorbierende Markoffketten

Beispiele für derartige Markoff-Ketten sind Spiele, die aus mehreren gleichartigen Spielrunden bestehen, wobei die Anzahl der Runden bis zum Gewinn oder Verlust sich erst aus dem Spielverlauf ergibt und nicht von vorne herein determiniert ist. Konkrete Beispiele sind der Tie-Break beim Tennis und das in der Einleitung vorgestellte Problem, beim Roulette ein bestimmtes Zielkapital zu erreichen. Die wesentliche Charakteristik solcher Markoffketten sind *absorbierende Zustände*, bei Spielen z.B. Gewinn oder Verlust, die nicht mehr verlassen werden, wenn man sie einmal erreicht hat.

2.3.1 Absorbierende Zustände

Gegeben eine Markoffkette mit Zustandsraum $\mathcal{S}$ und Übergangswahrscheinlichkeiten p_{ik}.

Definition 2.3.1 *Ein Zustand $k \in \mathcal{S}$ mit $p_{kk} = 1$ heißt ein absorbierender Zustand.*

Ein absorbierender Zustand k hat die folgenden Eigenschaften:

1) Wegen $\sum_{j\in\mathcal{S}} p_{kj} = 1$ ist $p_{kj} = 0$ für alle Zustände $j \neq k$. Aus der Gleichung (2.35) von Chapman und Kolmogoroff folgt daraus weiterhin, daß $p^t_{kk} = 1$ und $p^t_{kj} = 0$ für alle $t = 1, 2, \ldots$ und alle $j \neq k$.

2) Aus $p^t_{kk} = 1$ folgt nach Definition 2.2.4, daß k aperiodisch und nach den Sätzen 2.52 und 2.2.14, daß k ist positiv rekurrent ist.

3) Für alle $t > 1$ gilt $f^t_{kk} = 0$, denn von k aus kann man wegen Eigenschaft 1) nicht in mehr als einem Schritt nach k zurückkommen, ohne diesen Zustand zwischendurch noch einmal zu durchlaufen. Damit ist $f_{kk} = f^1_{kk} = p_{kk} = 1$.

4) Aus $p_{kj} = 0$ folgt $f_{kj} = 0$ für alle $j \neq k$.

Definition 2.3.2 *Eine Markoffkette heißt absorbierend, wenn ihr Zustandsraum $\mathcal{S}$ endlich ist und wenn es eine Teilmenge $\mathcal{R} \subset \mathcal{S}$ gibt, die die folgenden Eigenschaften besitzt:*

1) Die Zustände $k \in \mathcal{R}$ sind absorbierend.

2) Zu jedem Zustand $i \notin \mathcal{R}$ gibt es ein $k \in \mathcal{R}$ mit $i \rightsquigarrow k$.

Die Menge $\mathcal{R}$ heißt der Rand der Markoffkette, die Zustände aus $\mathcal{R}$ heißen Randzustände und die übrigen innere Zustände.

Satz 2.3.1 *Innere Zustände sind transient.*

Beweis: Wenn es einen rekurrenten inneren Zustand i gäbe, von dem nach Definition ein Pfad zu einem Randzustand k führt, so müßte nach Satz 2.2.17 auch $k \rightsquigarrow i$ gelten, was aber bei einem absorbierenden Zustand k wegen $p_{kj} = 0$ für alle $j \neq k$ unmöglich ist. ◇

2.3.2 Die Absorptionswahrscheinlichkeiten

Die wichtigste Aufgabe im Zusammenhang mit absorbierenden Markoffketten ist die Berechnung der Wahrscheinlichkeiten $f_{ik} = \sum_{t=1}^{\infty} f_{ik}^t$, von einem inneren Zustand i aus in den Randzustand k zu gelangen. Bei Spielen ist das die Frage nach der Wahrscheinlichkeit, von einer bestimmten Anfangssituation aus zu gewinnen oder zu verlieren.

Satz 2.3.2 *Für alle Randzustände k und beliebige Zustände i ist*

$$f_{ik} = \lim_{t\to\infty} p_{ik}^t \tag{2.57}$$

Beweis: **1.** Für $i = k$ ist (2.57) wegen $f_{kk} = p_{kk}^t = 1$ erfüllt, ebenso für Randzustände i mit $i \neq k$ wegen $f_{ik} = p_{ik}^t = 1$.
2. Ist i ein innerer Zustand, so ergibt sich aus (2.48) wegen $p_{kk}^s = 1$ für alle s in diesem Fall

$$p_{ik}^t = \sum_{s=1}^{t} f_{ik}^s p_{kk}^{t-s} = \sum_{s=1}^{t} f_{ik}^s$$

und daraus wieder

$$\lim_{t\to\infty} p_{ik}^t = \lim_{t\to\infty} \sum_{s=1}^{t} f_{ik}^s = \sum_{s=1}^{\infty} f_{ik}^s = f_{ik}$$

◇

Diese Beziehung liefert das folgende Berechnungsverfahren für die Absorptionswahrscheinlichkeiten f_{ik}:

Satz 2.3.3 *Ist k ein Randzustand und $i \neq k$, so ist $f_{ik} = 0$ für Randzustände i und*

$$f_{ik} = \sum_{j\in\mathcal{S}} p_{ij} f_{jk} \tag{2.58}$$

für innere Zustände i.

Beweis: Nach der Gleichung (2.35) von Chapman und Kolmogoroff ist

$$p_{ik}^{t+1} = \sum_{j \in \mathcal{S}} p_{ij} p_{jk}^{t}$$

Geht man jetzt auf beiden Seiten dieser Gleichung mit t gegen ∞, so erhält man aus (2.57) die Gleichung (2.58). Dabei wird verwendet, daß man wegen der Endlichkeit der Menge $\mathcal{S}$ auf der rechten Seite die Limesbildung mit der Summation vertauschen darf. $\diamond$

2.3.3 Beispiele

Wir greifen das in der Einleitung erwähnte Roulettespiel auf. Ein Spieler mit einem Startkapital k möchte ein Zielkapital $n(> k)$ durch einige Runden eines Spiels erreichen, bei dem nur die Ausgänge „Gewinn“ und „Verlust“ mit den Wahrscheinlichkeiten p bzw. $q = 1 - p$ möglich sind. Vor jeder Runde setzt er einen bestimmten Betrag aus seinem momentanen Kapitalbestand ein. Im Fall des Gewinns erhält er seinen Einsatz und noch einmal den selben Betrag als Gewinn. Bei Verlust der Runde verliert er seinen Einsatz. Wenn der Spieler vor jeder Runde lediglich auf der Basis seines augenblicklichen Kapitalbestands entscheidet, wieviel er setzen soll, gibt es die folgenden beiden Extremfälle:

Die vorsichtige Strategie. In jeder Runde setzt der Spieler nur eine Kapitaleinheit. Die Markoffkette mit dem jeweiligen Kapitalbestand als Zustand besitzt dann den nachstehenden Graphen, wobei der Rand aus den Zuständen 0 (Bankrott) und n (Zielkapital erreicht) besteht. Die für das Problem bedeutungslosen Schleifen an den Randknoten sind weggelassen.

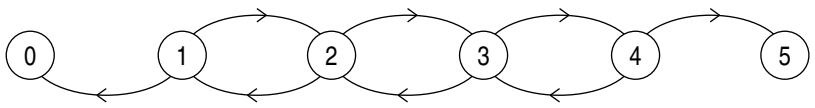


Abbildung 2.6: Die vorsichtige Strategie

Aus jedem inneren Zustand k führen genau zwei Bögen zu den unmittelbaren Nachbarzuständen: $p_{kk+1} = p$ und $p_{kk-1} = q = 1-p$. Die Formel (2.58) lautet

in diesem Fall

$$f_{kn} = p_{kk-1} f_{k-1n} + p_{kk+1} f_{k+1n} = q f_{k-1n} + p f_{k+1n}$$

für $0 < k < n$. Die f_{kn} sind also die Lösungen der Differenzengleichung

$$p x_{k+1} - x_k + q x_{k-1} = 0 \tag{2.59}$$

Die Randbedingungen zur Berechnung der Gewinnwahrscheinlichkeit sind dabei $x_0 = f_{0n} = 0$ und $x_n = f_{nn} = 1$.
Mit dem Ansatz $x_k = z^k$ lautet diese Differenzengleichung

$$z^{k-1} \left(p z^2 - z + q \right) = 0$$

wobei das Polynom in Klammern die Nullstellen $z_1 = 1$ und $z_2 = q/p$ besitzt.

Für $p \neq 1/2$ ist $q \neq p$ bzw. $z_2 \neq 1$, die beiden Nullstellen sind verschieden. In diesem Fall besagt die Theorie der Differenzengleichungen, daß die allgemeine Lösung von der Form

$$x_k = c_1 z_1^k + c_2 z_2^k = c_1 + c_2 \left(\frac{q}{p} \right)^k$$

ist, wobei die Faktoren c_i so gewählt werden müssen, daß die Randbedingungen erfüllt sind. In unserem Fall ergibt das

$$f_{kn} = x_k = \frac{\left(\frac{q}{p} \right)^k - 1}{\left(\frac{q}{p} \right)^n - 1} \tag{2.60}$$

Die Wahrscheinlichkeit, beim Roulette auf diese Weise durch Setzen auf „Rouge“ mit 5 DM Startkapital auf 100 DM zu kommen, ist nach dieser Formel etwa 0.004.

Falls es sich um ein faires Spiel mit $p = q = 1/2$ handelt, hat man eine doppelte Nullstelle $z_1 = z_2 =: z_0 = 1$ und die allgemeine Lösung lautet

$$x_k = c_1 z_0^k + c_2 k z_0^k = c_1 + c_2 k$$

Zusammen mit unseren Randbedingungen ergibt das

$$f_{kn} = x_k = \frac{k}{n} \tag{2.61}$$

d.h., die Gewinnwahrscheinlichkeit ist proportional zur Höhe des Startkapitals.

Die kühne Strategie. Bei der kühnen Strategie setzt der Spieler in jeder Runde aus seinem momentanen Kapitalbestand den Betrag ein, der ihn im Fall des Gewinns am nächsten an das erstrebte Zielkapital bringt. Für das Zielkapital $n = 5$ bei Gewinnwahrscheinlichkeit p besitzt die Markoffkette den in Abb. 2.7 dargestellten Graphen,

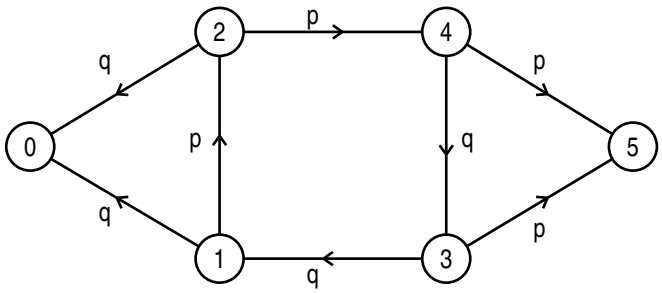


Abbildung 2.7: Die kühne Strategie

in dem die für die Berechnung bedeutungslosen Schleifen an den Randzuständen 0 und 5 weggelassen wurden. Nach (2.58) ist f_{i5} jeweils die Summe der Produkte $p_{ij}f_{j5}$ über alle die Bögen, die aus dem Knoten i herausführen. Das ergibt in diesem Fall für die $x_i = f_{i5}$ das folgende Gleichungssystem:

$$\begin{aligned}
x_0 &= 0 \\
x_5 &= 1 \\
x_4 &= px_5 + qx_3 = p + qx_3 \\
x_3 &= px_5 + qx_1 = p + qx_1 \\
x_2 &= px_4 + qx_0 = px_4 \\
x_1 &= px_2 + qx_0 = px_2
\end{aligned}$$

Aus der letzten Gleichung ergibt sich durch sukzessives Einsetzen

$$\begin{aligned}
x_1 &= px_2 \\
&= p^2x_4 \\
&= p^3 + p^2qx_3 \\
&= p^3 + p^3q + p^2q^2x_1
\end{aligned}$$

und daraus

$$x_1 = \frac{p^3(1+q)}{1-p^2q^2}$$

Speziell für $p = q = 1/2$ hat x_1 den Wert $1/5$.

Wenn man davon ausgehen kann, daß man in einer absorbierenden Markoffkette mit Wahrscheinlichkeit 1 im Rand landet, ist die Wahrscheinlichkeit x_0, das Spiel zu verlieren, gleich $1 - x_5 = 4/5$. Den Beweis für diese Aussage liefert der nächste Satz.

2.3.4 Die mittlere Schrittzahl bis zur Absorption

Für innere Zustände i sei

$$f_{i\mathcal{R}} = \sum_{k\in\mathcal{R}} f_{ik} \tag{2.62}$$

$$= \sum_{k\in\mathcal{R}}\sum_{t=1}^{\infty} f_{ik}^t = \sum_{t=1}^{\infty}\sum_{k\in\mathcal{R}} f_{ik}^t$$

$$=: \sum_{t=1}^{\infty} f_{i\mathcal{R}}^t \tag{2.63}$$

Für diese Größen gilt der

Satz 2.3.4 *Für innere Zustände i ist $f_{i\mathcal{R}} = 1$.*

Beweis: Da die t-Schritt-Übergangsmatrix P^t stochastisch ist, gilt

$$1 = \sum_{k\in\mathcal{S}} p_{ik}^t = \sum_{k\in\mathcal{R}} p_{ik}^t + \sum_{k\notin\mathcal{R}} p_{ik}^t$$

Innere Zustände sind transient, so daß nach Satz 2.2.10

$$\lim_{t\to\infty} \sum_{k\notin\mathcal{R}} p_{ik}^t = \sum_{k\notin\mathcal{R}} \lim_{t\to\infty} p_{ik}^t = 0$$

und deshalb nach (2.57)

$$1 = \lim_{t\to\infty} \sum_{k\in\mathcal{R}} p_{ik}^t = \sum_{k\in\mathcal{R}} \lim_{t\to\infty} p_{ik}^t = \sum_{k\in\mathcal{R}} f_{ik} = f_{i\mathcal{R}}$$

◇

Da nach diesem Satz die Funktion $t \mapsto f_{i\mathcal{R}}^t$ eine Wahrscheinlichkeitsfunktion ist, die die Verteilung der Zufallsvariablen „Anzahl der Schritte von i bis zum Rand“ beschreibt, ist die Größe

$$m_i = \sum_{t=1}^{\infty} t\, f_{i\mathcal{R}}^t \tag{2.64}$$

der Mittelwert dieser Verteilung. Sie heißt dementsprechend die *mittlere Schrittzahl bis zur Absorption*.
Sie ist bisher nur für innere Zustände definiert. Für Randzustände k legt man $m_k := 0$ fest, denn von einem Randzustand aus ist kein Schritt erforderlich, um an den Rand zu gelangen.

Satz 2.3.5 *Für innere Zustände i ist*

$$m_i = 1 + \sum_{k \in \mathcal{S}} p_{ik} m_k \tag{2.65}$$

Beweis: Sei i ein innerer und k ein Randzustand. Für $t \geq 2$ ist nach (2.47)

$$\begin{aligned} f_{ik}^t &= \sum_{j_1,\ldots,j_{t-1} \neq k} p_{ij_1} p_{j_1 j_2} \cdots p_{j_{t-1} k} \\ &= \sum_{j_1 \neq k} p_{ij_1} \left(\sum_{j_2,\ldots,j_{t-1} \neq k} p_{j_1 j_2} \cdots p_{j_{t-1} k} \right) \\ &= \sum_{j_1 \neq k} p_{ij_1} f_{j_1 k}^{t-1} \end{aligned}$$

Da für Randpunkte j_1 mit $j_1 \neq k$ gilt $f_{j_1 k}^{t-1} = 0$, reduziert sich die Summe und man erhält

$$f_{ik}^t = \sum_{j \notin \mathcal{R}} p_{ij} f_{jk}^{t-1}$$

für $t \geq 2$.
Daraus folgt die Gleichung

$$t\, f_{ik}^t = f_{ik}^t + (t-1)\, f_{ik}^t == f_{ik}^t + \sum_{j \notin \mathcal{R}} p_{ij}\, (t-1) f_{jk}^{t-1}$$

die auch für $t = 1$ richtig ist.
Summation über t ergibt

$$\begin{aligned} \sum_{t=1}^{\infty} t\, f_{ik}^t &= \sum_{t=1}^{\infty} f_{ik}^t + \sum_{j \notin \mathcal{R}} p_{ij} \sum_{t=1}^{\infty} (t-1) f_{jk}^{t-1} \\ &= f_{ik} + \sum_{j \notin \mathcal{R}} p_{ij} \sum_{s=1}^{\infty} s\, f_{jk}^s \end{aligned}$$

und anschließende Summation über $k \in \mathcal{R}$

$$\begin{aligned} m_i &= \sum_{t=1}^{\infty} t f_{i\mathcal{R}}^t = \sum_{k \in \mathcal{R}} \sum_{t=1}^{\infty} t\, f_{ik}^t \\ &= \sum_{k \in \mathcal{R}} f_{ik} + \sum_{j \notin \mathcal{R}} p_{ij} \sum_{k \in \mathcal{R}} \sum_{s=1}^{\infty} s\, f_{jk}^s \\ &= f_{i\mathcal{R}} + \sum_{j \notin \mathcal{R}} p_{ij}\, m_j \end{aligned}$$

Daraus folgt schließlich die Behauptung, denn es ist $f_{i\mathcal{R}} = 1$ und wegen $m_k = 0$ für Randzustände k kann die Summe in der letzten Zeile über alle Zustände j erstreckt werden. ◇

Fortsetzung der Beispiele. Für die vorsichtige Strategie lautet (2.65)

$$m_k = 1 + qm_{k-1} + pm_{k+1}$$

oder

$$pm_{k+1} - m_k + qm_{k-1} = -1$$

für $0 < k < n$ mit den Randbedingungen $m_0 = m_n = 0$.
Im Fall $p = q = 1/2$ ist — wie man leicht nachrechnet — die Lösung $m_k = k(n-k)$. Im Fall $p \neq 1/2$ ist die allgemeine Lösung von der Form

$$m_k = \frac{k}{1-2p} + c_1\, 1^k + c_2 \left(\frac{q}{p}\right)^k$$

Mit c_i so, daß die Randbedingungen erfüllt sind, ergibt das

$$m_k = \frac{n}{2p-1}\left[\frac{\left(\frac{q}{p}\right)^k - 1}{\left(\frac{q}{p}\right)^n - 1} - \frac{k}{n}\right]$$

Bei $q/p = 19/18$ dauert ein Spiel bei Startkapital 5 und Zielkapital 100 im Mittel etwa $m_5 \approx 180$ Runden.

Bei der kühnen Strategie — s. Abb. 2.7 — ergibt (2.65) das Gleichungssystem

$$\begin{aligned}
m_0 &= 0 \\
m_5 &= 0 \\
m_4 &= 1 + qm_3 + pm_5 = 1 + qm_3 \\
m_3 &= 1 + qm_1 + pm_5 = 1 + qm_1 \\
m_2 &= 1 + qm_0 + pm_4 = 1 + pm_4 \\
m_1 &= 1 + qm_0 + pm_2 = 1 + pm_2
\end{aligned}$$

mit der Lösung

$$m_1 = \frac{(1+p) + (1+q)p^2}{1 - p^2q^2}$$

was speziell für $p = q = 1/2$ eine mittlere Schrittzahl $m_1 = 2$ liefert.

2.3.5 Graphenreduktion

Für einen inneren Zustand i und einen Randzustand k einer absorbierenden Markoffkette ist die Wahrscheinlichkeit f_{ik} gemäß (2.47) durch

$$\begin{aligned} f_{ik} &= \sum_{t=1}^{\infty} f_{ik}^{t} \\ &= \sum_{t=1}^{\infty} \sum_{j_1,\ldots,j_{t-1} \neq k} p_{ij_1} p_{j_1 j_2} \cdots p_{j_{t-1} k} \end{aligned}$$

gegeben, also die Summe der Wahrscheinlichkeiten aller Pfade, die vom Zustand i zum Zustand k führen. Graphenreduktion ist eine Methode zur Berechnung dieser Summe anhand des Graphen der Markoffkette, wobei die einzelnen Rechenschritte in der Zusammenfassung von Gruppen von Bögen zu einem einzelnen mit entsprechender Berechnung der Übergangswahrscheinlichkeiten an diesen Bögen bestehen. Dies wird so lange durchgeführt, bis nur noch ein Bogen vom Knoten i zum Knoten k übrig bleibt, an dem nach Konstruktion dann als Übergangswahrscheinlichkeit die Zahl f_{ik} steht.
Beschreibungen dieses Verfahrens findet man in [6] und [8].

2.4 Ergodische Markoffketten

In diesem Abschnitt wird das Verhalten von Markoffketten für $t \to \infty$ analysiert, die keinen absorbierenden Zustand besitzen.

Definition 2.4.1 *Ein aperiodischer positiv rekurrenter Zustand einer homogenen Markoffkette heißt ergodisch.*

Satz 2.4.1 *Ein aperiodischer Zustand k ist genau dann ergodisch, wenn*

$$\sum_{t=0}^{\infty} p_{kk}^t = \infty \quad \textit{und} \quad \lim_{t\to\infty} p_{kk}^t = u_k$$

Dabei ist $u_k = \frac{1}{\mu_k} > 0$ mit der mittleren Wiederkehrzeit μ_k des Zustands k und für alle Zustände $i \in \mathcal{S}$ gilt

$$\lim_{t\to\infty} p_{ik}^t = f_{ik}\frac{1}{\mu_k}$$

Beweis: Die erste Aussage folgt aus den Sätzen 2.52 und 2.2.13, der Rest sei hier ohne Beweis wiedergegeben. ◇

Definition 2.4.2 *Eine Markoffkette, in der jeder Zustand mit jedem verbunden ist, heißt irreduzibel. Falls die Zustände außerdem noch alle ergodisch sind, heißt die Kette ergodisch.*

Im folgenden wird gezeigt, daß eine ergodische Markoffkette genau eine stationäre Verteilung besitzt und die Zustandswahrscheinlichkeiten gegen diese stationäre Verteilung konvergieren. Zum Beweis dieser Aussagen wird der folgende Hilfsfsatz benötigt:

Satz 2.4.2 *v_n und p_n^t mit $n = 1, 2, \ldots$ und $t = 1, 2, \ldots$ seien nichtnegative reelle Zahlen mit den Eigenschaften*

- $\sum_{n=1}^{\infty} v_n < \infty$
- $0 \le p_n^t \le 1$
- $\lim_{t\to\infty} p_n^t = p_n$

Dann gilt

$$\lim_{t\to\infty} \sum_{n=1}^{\infty} v_n p_n^t = \sum_{n=1}^{\infty} v_n (\lim_{t\to\infty} p_n^t) = \sum_{n=1}^{\infty} v_n p_n \tag{2.66}$$

Beweis:
1. Sei $\nu > 0$ eine Zahl mit $\sum_{n=1}^{\infty} v_n < \nu$. Wegen der Konvergenzvoraussetzungen gibt zu einem vorgegebenen $\varepsilon > 0$ einen Index n_0 so, daß

$$\sum_{n>n_0} v_n < \frac{\varepsilon}{4}$$

und weiter ein t_0 mit

$$|p_n - p_n^t| < \frac{\varepsilon}{2n_0\nu}$$

für alle $n = 1, 2, \ldots, n_0$ und alle $t \geq t_0$.
2. Mit den üblichen Rechenregeln für den Absolutbetrag erhält man die Abschätzung

$$\begin{aligned}\left|\sum_{n=1}^{\infty} v_n p_n \;=\; \sum_{n=1}^{\infty} v_n p_n^t\right| \;&\leq\; \sum_{n=1}^{\infty} v_n |p_n - p_n^t| \\ &=\; \sum_{n=1}^{n_0} v_n |p_n - p_n^t| + \sum_{n>n_0} v_n |p_n - p_n^t|\end{aligned}$$

3. Die erste Summe läßt sich durch

$$\sum_{n=1}^{n_0} (\nu \frac{\varepsilon}{2n_0\nu}) = \frac{n_0\nu\varepsilon}{2n_0\nu} = \frac{\varepsilon}{2}$$

und die zweite wegen $|p_n - p_n^t| \leq |p_n| + |p_n^t| \leq 2$ durch

$$2 \sum_{n>n_0} v_n = \frac{\varepsilon}{2}$$

nach oben abschätzen, so daß für $t \geq t_0$ insgesamt gilt

$$\left|\sum_{n=1}^{\infty} v_n p_n \;=\; \sum_{n=1}^{\infty} v_n p_n^t\right| < \varepsilon$$

◇

Satz 2.4.3 *Die Kehrwerte $u_k = 1/\mu_k$ der mittleren Wiederkehrzeiten der Zustände einer ergodischen Markoffkette besitzen die Eigenschaften*

$$\lim_{t\to\infty} p_{ik}^t = u_k \quad \textit{für alle } i \in \mathcal{S} \tag{2.67}$$

$$u_k > 0 \quad \textit{für alle } k \in \mathcal{S} \tag{2.68}$$

$$\sum_{k\in\mathcal{S}} u_k = 1 \tag{2.69}$$

$$u_k = \sum_{i\in\mathcal{S}} u_i p_{ik} \quad \textit{für alle } k \in \mathcal{S} \tag{2.70}$$

Beweis:
1. Für einen positiv rekurrenten Zustand k ist $0 < \mu_k < \infty$ und nach Satz 2.4.1 $\lim_{t\to\infty} p_{ik}^t = f_{ik}\frac{1}{\mu_k}$, wobei nach Satz 2.2.17 $f_{ik} = 1$, weil die Markoffkette irreduzibel ist. Für die $u_k = \frac{1}{\mu_k}$ sind daher (2.67) und (2.68) erfüllt. Um die beiden anderen Eigenschaften nachzuweisen, benutzen wir die Gleichung (2.35) von Chapman und Kolmogoroff:

$$p_{ik}^{t+1} = \sum_{j\in\mathcal{S}} p_{ij}^t p_{jk}$$

2. Falls der Zustandsraum endlich ist, $\mathcal{S} = \{1, 2, \ldots, n\}$, ist

$$u_k = \lim_{t\to\infty} p_{ik}^{t+1} = \lim_{t\to\infty} \sum_{j\in\mathcal{S}} p_{ij}^t p_{jk} = \sum_{j\in\mathcal{S}} \lim_{t\to\infty} p_{ij}^t p_{jk} = \sum_{j\in\mathcal{S}} u_j p_{jk}$$

und

$$\sum_{k\in\mathcal{S}} u_k = \sum_{k\in\mathcal{S}} \lim_{t\to\infty} p_{ik}^t = \lim_{t\to\infty} \sum_{k\in\mathcal{S}} p_{ik}^t = \lim_{t\to\infty} 1 = 1$$

3. Bei abzählbar unendlichem $\mathcal{S} = \{1, 2, 3, \ldots\}$ darf die Limesbildung und die Summation nicht so ohne weiteres vertauscht werden. Aus der Gleichung von Chapman und Kolmogoroff folgt zunächst für festes, aber beliebiges n, daß

$$p_{ik}^{t+1} \geq \sum_{j=1}^{n} p_{ij}^t p_{jk}$$

und daraus für $t \to \infty$ — weil rechts der Limes unter die (endliche) Summe gezogen werden darf —

$$u_k \geq \sum_{j=1}^{n} u_j p_{jk}$$

Für $n \to \infty$ ergibt das die Ungleichung

$$u_k \geq \sum_{j=1}^{\infty} u_j p_{jk} =: v_k$$

Da bei Summen über nichtnegative Summanden die Summationsreihenfolge beliebig vertauscht werden darf, ist

$$\sum_{k=1}^{\infty} v_k = \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} u_j p_{jk} = \sum_{j=1}^{\infty} u_j \sum_{k=1}^{\infty} p_{jk} = \sum_{j=1}^{\infty} u_j 1 = \sum_{j=1}^{\infty} u_j$$

In Verbindung mit $v_k \leq u_k$ ist die Gleichheit der Summen aber nur möglich, wenn $u_k = v_k$ für alle k und damit (2.70) erfüllt ist.

4. Aus der matriziellen Schreibweise $u = uP$ der Gleichung (2.70) ersieht man sofort, daß $u = uP^t$ für alle $t = 1, 2, 3, \ldots$, bzw.

$$u_k = \sum_{j \in \mathcal{S}} u_j p_{jk}^t \tag{2.71}$$

Aus $\sum_{k=1}^n u_k = \lim_{t\to\infty} \sum_{k=1}^n p_{ik}^t \leq 1$ und $u_k > 0$ folgt weiter für $n \to \infty$

$$0 < s := \sum_{k=1}^{\infty} u_k \leq 1$$

Der Vektor u^* mit den Komponenten $u_k^* = \frac{1}{s} u_k$ ist ein stochastischer Vektor, der ebenfalls die Gleichung (2.71) erfüllt. Nach Satz 2.4.2 folgt daraus

$$u_k^* = \lim_{t\to\infty} \sum_{j\in\mathcal{S}} u_j^* p_{jk}^t = \sum_{j\in\mathcal{S}} u_j^* \lim_{t\to\infty} p_{jk}^t = \sum_{j\in\mathcal{S}} u_j^* u_k = (\sum_{j\in\mathcal{S}} u_j^*) u_k = u_k$$

D.h. u ist bereits ein stochastischer Vektor, bzw. es gilt (2.69). ◇

Satz 2.4.4 *Gibt es bei einer irreduziblen Markoffkette mit aperiodischen Zuständen Zahlen $u_k \geq 0$ mit den Eigenschaften (2.69) und (2.70), dann sind alle Zustände positiv rekurrent.*

Beweis: Angenommen es gebe einen transienten oder nullrekurrenten Zustand, so sind, weil die Markoffkette irreduzibel ist, alle Zustände vom gleichen Typ und es gilt für alle i und k

$$\lim_{t\to\infty} p_{jk}^t = 0$$

woraus mit Satz 2.4.2 und (2.71)

$$u_k = \lim_{t\to\infty} \sum_{j\in\mathcal{S}} u_j p_{jk}^t = \sum_{j\in\mathcal{S}} u_j \lim_{t\to\infty} p_{jk}^t = 0$$

für alle k und damit $\sum_{k\in\mathcal{S}} u_k = 0$ im Widerspruch zur Voraussetzung (2.69) folgen würde. ◇

Satz 2.4.5 *Eine ergodische Markoffkette besitzt genau eine stationäre Verteilung.*

Beweis: Besitzt die Markoffkette außer der in Satz 2.4.3 ermittelten eine weitere stationäre Verteilung mit Komponenten u_k, so gilt — wie im Satz 2.4.4 schon verwendet —

$$u_k = \lim_{t\to\infty} \sum_{j\in\mathcal{S}} u_j p_{jk}^t = \sum_{j\in\mathcal{S}} u_j \lim_{t\to\infty} p_{jk}^t = \sum_{j\in\mathcal{S}} u_j \frac{1}{\mu_k} = \frac{1}{\mu_k}$$

◇

Satz 2.4.6 *Ist (u_k) die stationäre Verteilung einer ergodischen Markoffkette, so gilt*

$$\lim_{t\to\infty} \mathcal{P}(X_t = k) \;=\; u_k$$

Beweis: Sind $a_i = \mathcal{P}(X_0 = i)$ beliebig vorgegebene Startwahrscheinlichkeiten, d.h. nichtnegative reelle Zahlen mit $\sum_{i\in\mathcal{S}} a_i = 1$, so gilt

$$a_k^t := \mathcal{P}(X_t = k) = \sum_{i\in\mathcal{S}} a_i p_{ik}^t$$

und wie bei den beiden vorhergehenden Sätzen erhält man

$$\lim_{t\to\infty} a_k^t = \sum_{i\in\mathcal{S}} a_i (\lim_{t\to\infty} p_{ik}^t) = (\sum_{i\in\mathcal{S}} a_i) u_k = u_k$$

$\diamond$

2.4.1 Simulated Annealing

Simulated Annealing ist ein stochastischer Algorithmus zur Lösung *diskreter Optimierungsprobleme*, d.h. der Berechnung einer globalen Minimalstelle einer Funktion $f : \mathcal{S} \longrightarrow \mathbb{R}$ auf einer *endlichen* Menge $\mathcal{S}$.
Eines der bekanntesten Probleme dieser Art ist das des *travelling salesman*: Gesucht ist die kürzeste Reiseroute, die durch eine gegebene Anzahl von Städten zurück zum Ausgangspunkt führt, wobei jede Stadt genau einmal besucht werden soll.
Bezeichnet man den Start- und Zielort mit 0 und die zu besuchenden Städte mit $1, 2, \ldots, n$ und ist d_{ik} die Länge der Fahrtstrecke von Stadt i nach Stadt k, so ist das Minimum der Funktion

$$f(\pi) = d_{0\pi(1)} + d_{\pi(1)\pi(2)} + \cdots + d_{\pi(n)0} \tag{2.72}$$

auf der Menge $\mathcal{S}$ aller Permutationen $\pi = (\pi(1), \pi(2), \ldots, \pi(n))$ der Zahlen $1, 2, \ldots, n$ zu bestimmen. Die konzeptionell einfachste Methode, nämlich die Berechnung aller Werte $f(\pi)$, ist schon bei relativ kleinen Anzahlen n nicht mehr praktikabel, weil die Berechnung zu lange dauern würde. So gibt es beispielsweise zu $n = 15$ Städten insgesamt $15! = 1\,307\,674\,368\,000$ Permutationen. Wenn man für die Berechnung eines Funktionswerts (2.72) eine tausendstel Sekunde veranschlagt, würde die gesamte Berechnung etwa 41 Jahre dauern.
Eine alternative Methode zur Lösung des Problems ist die iterative Suche nach der Minimalstelle. Dabei gibt man für jedes Element i aus $\mathcal{S}$ eine Menge $\mathcal{S}_i$ von *Nachbarn* vor und durchläuft, ausgehend von einem Startpunkt, eine

Folge von Elementen von $\mathcal{S}$, indem man aus den Nachbarn eines $i \in \mathcal{S}$ ein k mit $f(k) < f(i)$ bestimmt und die Suche in der Menge $\mathcal{S}_k$ fortsetzt. Bei diesem Verfahren landet man aber meist nur in einem *lokalen* Minimum, d.h. einem nichtoptimalen Punkt, in dessen Nachbarschaft aber nur Elemente mit größeren Funktionswerten liegen.
Einen Ausweg aus einer derartigen Sackgasse gäbe es, wenn man den Iterationsalgorithmus so modifizieren könnte, daß *ab und zu* auch einmal zu einem *schlechteren* Nachbarn übergegangen wird, so daß mehrere lokale Minima aufgesucht und verglichen werden können.
Nach diesem Schema geht der nachstehend beschriebene *Simulated-Annealing-Algorithmus*, nach seinem Erfinder auch *Metropolis-Algorithmus* gennant, vor.

Der Metropolis-Algorithmus.
Der Algorithmus entstand aus dem Bestreben, das folgende physikalische Phänomen am Computer zu simulieren: Wird ein Festkörper im Hitzebad bis in den flüssigen Zustand erhitzt, so löst sich die Kristallstruktur auf; die Moleküle werden gegeneinander frei beweglich. Beim anschließenden (allmählichen) Abkühlen wird Energie abgegeben und es stellt sich die Kristallstruktur, d.h. Molekülkonfiguration ein, die dem niedrigsten Energiezustand entspricht. Wenn man jedoch zu schnell abkühlt, friert der Festkörper evtl. in einer anderen Konfiguration ein, die nur ein lokales Energieminimum und damit meist instabil ist.
Als theoretisches Modell für den Abkühlungsvorgang (*annealing*) nimmt man an, daß sich das System „Festkörper“ bei fester absoluter Temperatur T wie eine homogene Markoffkette auf der Menge $\mathcal{S}$ aller möglichen Molekülkonfigurationen s mit der *Boltzmann-Verteilung*

$$u_s(T) = \frac{1}{Z(T)} e^{-\frac{E(s)}{\kappa T}}$$

als stationärer Verteilung verhält. Dabei ist $E(s)$ der Energiezustand der Konfiguration s und

$$Z(T) = \sum_{s \in \mathcal{S}} e^{-\frac{E(s)}{\kappa T}}$$

die Normierungskonstante, die den Vektor $u(T) = (u_s(T))$ zu einem stochastischen Vektor macht.
Wenn es eine eindeutige Konfiguration s_0 mit niedrigstem Energiezustand gibt, d.h. $E(s_0) < E(s)$ für alle $s \neq s_0$, so gilt

$$\lim_{T \to 0} u_s(T) = \begin{cases} 1 & \text{falls } s = s_0 \\ 0 & \text{falls } s \neq s_0 \end{cases}$$

Falls man also so langsam abkühlt ($T \to 0$), daß sich bei festem T jeweils der stationäre Systemzustand einstellt, landet man für $T \to 0$ mit Wahrscheinlichkeit 1 in der energieminimalen Konfiguration s_0.
Zur Simulation dieses Vorgangs benötigt man die Übergangswahrscheinlichkeiten $p_{ik}(T)$ bzw. eine entsprechende Rechenvorschrift. Metropolis et al. (1953) verwendeten das folgende Simulationsschema:

Ist s die momentane (Arbeits-) Konfiguration, so wird durch zufälliges Verschieben eines oder mehrerer Moleküle eine von N *Nachbarkonfiguration* s' *erzeugt* und die Energiedifferenz $\Delta E = E(s') - E(s)$ berechnet.
Ist $\Delta E \leq 0$, d.h. die generierte Konfiguration s' energiemäßig *besser*, so wird sie als neue Arbeitskonfiguration *akzepiert*.
Ist $\Delta E > 0$, so wird ein Bernoulli-Experiment mit der Erfolgswahrscheinlichkeit $p = e^{-\Delta E/\kappa T}$ durchgeführt. Bei Erfolg wird die *schlechtere* Konfiguration s' akzepiert, andernfalls bleibt man bei s.

Dieser Schritt wird bei fester Temperatur ausreichend oft durchgeführt, anschließend wird die Temperatur etwas abgesenkt und weiter nach diesem Schema iteriert.

Mit sinkender Temperatur T wird die Wahrscheinlichkeit p, daß eine schlechtere Konfiguration akzeptiert wird, immmer geringer. Damit wird der Effekt modelliert, daß das System in einem lokalen Energieminimum einfriert, wenn T zu schnell gegen Null geht.

Allgemeines Simulated-Annealing-Verfahren
Sei $\mathcal{S}$ eine endliche Menge, $f : \mathcal{S} \longrightarrow \mathbb{R}$ eine reellwertige Funktion auf $\mathcal{S}$ und

$$\mathcal{S}_0 = \{k \in \mathcal{S} \; ; \; f(k) \leq f(i) \text{ für alle } i \in \mathcal{S}\} \tag{2.73}$$

die Menge der globalen Minimalstellen der Funktion f.
Weiter seien für alle Paare i, k von Elementen aus $\mathcal{S}$ reelle Zahlen a_{ik} und g_{ik} mit den folgenden Eigenschaften gegeben:

V1 Es ist $g_{ik} \geq 0$ und $\sum_{k \neq i} g_{ik} = 1$ für alle i und k mit $i \neq k$

V2 $0 < a_{ik} \leq 1$ für alle i und k.

V3 Zu je zwei Elementen i und k gibt es eine Anzahl m von Punkten $j_1, j_2, \ldots j_m$, so daß (mit $j_0 = i$ und $j_{m+1} = k$ gilt $g_{j_\mu j_{\mu+1}} > 0$ für alle $\mu = 0, 1, \ldots, m$.

V4 Es gibt wenigstens ein Paar i, k mit $f(i) < f(k)$ und $g_{ik} > 0$.

V5 Ist $f(k) \leq f(i)$, so ist $a_{ik} = 1$, andernfalls ist $a_{ik} < 1$.

V6 Für je drei Elemente i, j, k mit $f(i) \leq f(j) \leq f(k)$ gilt $a_{ik} = a_{ij}a_{jk}$.

V7 Für alle i und k gilt $g_{ik} = g_{ki}$

Unter den Voraussetzungen V1 und V2 sind die Größen

$$p_{ik} = \begin{cases} g_{ik}a_{ik} & \text{falls } i \neq k \\ 1 - \sum_{j \neq i} g_{ij}a_{ij} & \text{für } i = j \end{cases} \tag{2.74}$$

nichtnegativ mit $\sum_{k \in \mathcal{S}} p_{ik} = 1$ für alle i, also die Übergangswahrscheinlichkeiten einer homogenen Markoffkette $\mathcal{P}$ mit Zustandsraum $\mathcal{S}$. g_{ik} ist in diesem Zusammenhang als die bedingte Wahrscheinlichkeit zu interpretieren, daß der Zustand k *generiert* wird unter der Bedingung, daß die Kette sich im Zustand i befindet. a_{ik} ist dann die bedingte Wahrscheinlichkeit, daß der Zustand k *akzeptiert* wird unter der Bedingung, daß k von i aus erzeugt wurde.
Mit $g_{ik} = 1/N$ für die Nachbarkonfigurationen k von i und $a_{ik} = 1$ bzw. $= e^{-(f(k)-f(i))/\kappa T}$ ist p_{ik} die Wahrscheinlichkeit, daß der Metropolis-Algorithmus von i nach k wechselt.

Satz 2.4.7 *Unter den Voraussetzungen V1-V5 ist die durch (2.74) gegebene Markoffkette ergodisch.*

Beweis:
1. Wegen $a_{ik} > 0$ (V2) erfüllen die p_{ik} ebenfalls die Voraussetzung V3, d.h. jeder Zustand ist mit jedem verbunden. Die Markoffkette ist daher irreduzibel und die Zustände alle vom gleichen Typ.
2. Da der Zustandsraum endlich ist, gibt es mindestens einen positiv rekurrenten Zustand. Da die Markoffkette irreduzibel ist, sind alle Zustände positiv rekurrent.
3. Für ein Paar i, k gemäß V3 gilt $g_{ik} > 0$ und nach V5 $a_{ik} < 1$, so daß

$$\sum_{j \neq i} g_{ij}a_{ij} < \sum_{j \neq i} g_{ij} = 1$$

und damit $p_{ii} > 0$. Dann sind aber i und in der Folge auch alle Zustände aperiodisch. ◇
Da die Markoffkette ergodisch ist, besitzt sie eine eindeutig bestimmte stationäre Verteilung. Unter den Voraussetzungen V6 und V7 läßt sie sich explizit angeben:

Satz 2.4.8 *Unter den Voraussetzungen V1-V7 besitzt die durch (2.74) gegebene Markoffkette die stationäre Verteilung*

$$u_j = \frac{a_{ij}}{\sum_{k \in \mathcal{S}} a_{ik}} \tag{2.75}$$

wo i irgendein Element der Menge $\mathcal{S}_0$ der globalen Minima ist.

Beweis:
1. Wir zeigen im Punkt 2, daß für die obigen u_j und alle Paare j, k von Zuständen die Gleichung

$$u_j p_{jk} = u_k p_{kj} \tag{2.76}$$

gilt. Durch Summation über j ergibt sich daraus

$$\sum_{j\in\mathcal{S}} u_j p_{jk} = u_k \sum_{j\in\mathcal{S}} p_{kj} = u_k$$

weil $\sum_{j\in\mathcal{S}} p_{kj} = 1$. Da offensichtlich $\sum_{j\in\mathcal{S}} u_j = 1$, erfüllen die u_j die Bedingungen einer stationären Verteilung.
2. Für $j = k$ ist die Gleichung (2.76) trivialerweise richtig. Wir nehmen also $j \neq k$ an und setzen

$$\alpha = \frac{1}{\sum_{l\in\mathcal{S}} a_{il}}$$

Dann ist mit V7

$$u_j p_{jk} = \alpha a_{ij} g_{jk} a_{jk} = \alpha g_{kj} a_{ij} a_{jk} \tag{2.77}$$

Da $f(i) \leq f(l)$ für alle $l \in \mathcal{S}$, brauchen nur zwei Fälle zu unterschieden werden:
1. Fall: $f(i) \leq f(j) \leq f(k)$. Dann ist nach V6 $a_{ij}a_{jk} = a_{ik}$ und nach V5 $a_{kj} = 1$, so daß

$$\alpha g_{kj} a_{ij} a_{jk} = \alpha g_{kj} a_{ik} a_{kj} = u_k p_{kj}$$

2. Fall: $f(i) \leq f(k) < f(j)$. Dann ist nach V5 $a_{jk} = 1$ und nach V6 $a_{ij} = a_{ik}a_{kj}$, so daß wieder

$$\alpha g_{kj} a_{ij} a_{jk} = \alpha g_{kj} a_{ik} a_{kj} = u_k p_{kj}$$

$\diamond$

Zur Durchführung des Algorithmus benötigt man eine Schar von Größen $a_{ik}(c)$ und $g_{ik}(c)$ in Abhängigkeit von einem Kontrollparameter $c > 0$ (der dem Term $1/\kappa T$ beim Metropolis-Algorithmus entspricht. Den Annealing-Effekt erreicht man durch die Voraussetzung

V8 $\lim_{c\to\infty} a_{ik}(c) = 0$, falls $f(i) < f(k)$,

die bei den Größen des Metropolis-Algorithmus offensichtlich erfüllt ist.

Satz 2.4.9 *Sei N die Anzahl der Elemente der Menge $\mathcal{S}_0$ der globalen Minimalstellen. Dann gilt unter der Vorausstzung V8 für die stationären Verteilungen*

$$u_j(c) = \frac{a_{ij}(c)}{\sum_{k\in\mathcal{S}} a_{ik}(c)}$$

mit einem $i \in \mathcal{S}_0$*, daß*

$$\lim_{c\to\infty} u_j(c) = \begin{cases} \frac{1}{N} & \textit{falls } j \in \mathcal{S}_0 \\ 0 & \textit{falls } j \notin \mathcal{S}_0 \end{cases}$$

Beweis:
Für $k \in \mathcal{S}_0$ ist $a_{ik}(c) = 1$ und für $k \notin \mathcal{S}_0$ gilt nach Konstruktion $f(i) < f(k)$, so daß wegen V8 der Nenner von $u_j(c)$ gegen $\sum_{k\mathcal{S}_0} 1 = N$ konvergiert, und der Zähler je nach Index gegen Null oder Eins. $\diamond$

Literatur. Ausführliche Darstellungen, insbesondere zur praktischen Durchführung des Algorithmus und zu den Anwendungen finden sich in [10] und [1].

Kapitel 3

Markoffprozesse

3.1 Markoff-Prozesse mit diskretem Zustandsraum

Wir betrachten jetzt stochastische Prozesse mit kontinuierlichem Zeitparameter t, d.h. $\mathcal{T}$ ist die Menge der reellen Zahlen $t \geq 0$. Der Zustandsraum bleibt jedoch eine endliche oder abzählbare Menge, wobei für allgemeine Darstellungen $\mathcal{S} = \{1, 2, \ldots, n\}$ bzw. $\mathcal{S} = \{1, 2, 3, \ldots\}$ verwendet wird.
Wie im Kapitel über Markoffketten ist unter einem stochastischen Prozeß entweder eine durch die Zeit t indizierte Menge von Zufallsvariablen $X_t : \Omega \longrightarrow \mathcal{S}$ auf einem beliebigem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ oder eine Wahrscheinlichkeit $\mathcal{P}$ auf einer σ-Algebra $\mathcal{Z}_\mathcal{S}$ auf der Menge $\mathcal{S}^\mathcal{T}$ aller Funktionen $s : \mathcal{T} \longrightarrow \mathcal{S}$ zu verstehen.
Im letzteren Fall sind die im folgenden mit X_t bezeichneten Funktionen die Zustandsvariablen (2.2) und $\mathcal{Z}_\mathcal{S}$ die σ-Algebra der Zylindermengen, d.h. die kleinste σ-Algebra, die alle Mengen der Form $(X_t = k)$ enthält.

Definition 3.1.1 *Ein stochastischer Prozess $X = \{X_t; t \geq 0\}$ von Zufallsvariablen $X_t : \Omega \longrightarrow \mathcal{S}$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ mit Werten in einer abzählbaren Menge $\mathcal{S}$ heißt ein Markoff-Prozeß mit dem diskreten Zustandsraum $\mathcal{S}$, wenn für jede streng monoton steigende Folge*

$$0 \leq t_0 < t_1 < t_2 < t_3 < \cdots$$

von Zeitpunkten ohne endlichen Häufungspunkt die Folge

$$X_{t_0}, X_{t_1}, X_{t_2}, X_{t_3}, \ldots$$

eine (i.a. inhomogene) Markoffkette ist.

Ein derartiger Prozess ist also dadurch charakterisiert, daß für die Zeitpunkte t_ν wie oben, beliebige Zustände $k_0, k_1, k_2, \ldots$ und Anzahlen n die Gleichung

$$\begin{aligned} & \mathcal{P}\left(X_{t_n} = k_n, \ldots, X_{t_1} = k_1 | X_{t_0} = k_0\right) \\ = \; & p_{k_0 k_1}(t_0, t_1) p_{k_1 k_2}(t_1, t_2) \cdots p_{k_{n-1} k_n}(t_{n-1}, t_n) \end{aligned} \tag{3.1}$$

mit den Übergangswahrscheinlichkeiten

$$p_{ik}(s, t) = \mathcal{P}\left(X_t = k | X_s = i\right) \tag{3.2}$$

erfüllt ist.
Die Übergangswahrscheinlichkeiten sind dabei wieder nichtnegative reelle Zahlen mit

$$\sum_{k \in \mathcal{S}} p_{ik}(s, t) = 1$$

und genügen der Gleichung von *Chapman und Kolmogoroff*: Für $t_0 < t_1 < t_2$ ist

$$p_{ik}(t_0, t_2) = \sum_{j \in \mathcal{S}} p_{ij}(t_0, t_1) p_{jk}(t_1, t_2) \tag{3.3}$$

Diese Gleichung erhält man durch Anwendung von (3.1) auf

$$(X_{t_2} = k) = \sum_{j \in \mathcal{S}} (X_{t_2} = k) \cap (X_{t_1} = j) = \sum_{j \in \mathcal{S}} (X_{t_2} = k, X_{t_1} = j)$$

Definition 3.1.2 *Ein Markoff-Prozeß $\{X_t\}$ heißt zeitlich homogen, wenn für alle Zeitpunkte $0 \leq s \leq t$ und alle Zustände i, k gilt*

$$p_{ik}(s, t) = p_{ik}(0, t - s)$$

Bei zeitlich homogenen Markoffprozessen hängen die Übergangswahrscheinlichkeiten also nicht vom absoluten Zeitpunkt s ab, ab dem das Übergangsverhalten beobachtet wird, sondern nur von der Länge $t - s$ des Beobachtungszeitintervalls.
Im folgenden werden nur noch zeitlich homogene Markoffprozesse behandelt. Daher schreiben wir der Einfachheit halber $p_{ik}(t - s)$ anstelle von $p_{ik}(0, t - s)$. Zeitlich homogene Markoffprozesse sind analog (3.3) und (3.2) durch die folgenden beiden Bedingungen charakterisiert:
1.) Für alle rellen Zahlen $s \geq 0$ und $t > 0$ und alle Zustände i, k ist

$$\mathcal{P}(X_{s+t} = k | X_s = i) = p_{ik}(t) \tag{3.4}$$

2.) Für jede streng monoton steigende Folge $0 \leq t_0 < t_1 < t_2 < t_3 < \cdots < t_n$ von Zeitpunkten und beliebige Zustände $k_0, k_1, \ldots, k_n$ gilt

$$\begin{aligned} & \mathcal{P}\left(X_{t_n} = k_n, \ldots, X_{t_1} = k_1 | X_{t_0} = k_0\right) \\ = \; & p_{k_0 k_1}(t_1 - t_0) p_{k_1 k_2}(t_2 - t_1) \cdots p_{k_{n-1} k_n}(t_n - t_{n-1}) \end{aligned} \tag{3.5}$$

Die $p_{ik}(t)$ sind dabei nichtnegative reelle Zahlen mit $\sum_{k \in \mathcal{S}} p_{ik}(t) = 1$ für alle i, die den Chapman-Kolmogoroff-Gleichungen

$$p_{ik}(s + t) = \sum_{j \in \mathcal{S}} p_{ij}(s) p_{jk}(t) \tag{3.6}$$

für $s > 0$ und $t > 0$ genügen. Dies folgt aus (3.3) mit $s = t_1 - t_0$ und $t = t_2 - t_1$. Definiert man zusätzlich noch $p_{ii}(0) := 1$ und $p_{ik}(0) = 0$ für $k \neq i$, so ist (3.6) auch für $s = 0$ oder $t = 0$ richtig.

Faßt man die $p_{ik}(t)$ zu der Übergangsmatrix

$$P(t) = \begin{pmatrix} p_{11}(t) & p_{12}(t) & \dots \\ p_{21}(t) & p_{22}(t) & \dots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{3.7}$$

zusammen, so lautet (3.6)

$$P(s+t) = P(s)P(t) \tag{3.8}$$

mit $P(0) = I$, der Einheitsmatrix.

Aus mathematischer Sicht ergeben sich jetzt notwendig zwei Fragen:
1. Gibt es überhaupt einen Markoff-Prozeß, d.h. eine Wahrscheinlichkeit $\mathcal{P}$ auf $\mathcal{Z}_{\mathcal{S}}$ mit der Eigenschaft (3.1), und
2. wie sehen Matrizen $P(t)$ aus, die der Gleichung (3.8) genügen?
Die erste Frage werden wir durch die Einführung des Poissonprozesses im folgenden Abschnitt beantworten, mit der zweiten befaßt sich der übernächste Abschnitt.

3.2 Der Poissonprozeß

Der Poissonprozeß beschreibt das Abzählen der Häufigkeit des Eintretens eines bestimmten Ereignisses, das zu zufälligen Zeitpunkten eintritt. Typische Beispiele für solche Ereignisse sind das Eintreffen von Kunden an einem Bedienungsschalter, das Zählen von Fahrzeugen an einer Meßstelle oder der Anforderungen von Workstations an einen Fileserver.
Sind die Zufallsvariablen

$$0 \leq T_1 \leq T_2 \leq T_3 \leq \cdots$$

die Zeitpunkte, an denen das Ereignis eintritt, so wird die Häufigkeit des Eintretens bis zum Zeitpunkt $t > 0$ durch die Zufallsvariable

$$N_t = \max\{n \ ; \ T_n \leq t\} \tag{3.9}$$

mit dem Wertebereich $\mathcal{S} = \{0, 1, 2, 3, \ldots\}$ beschrieben. Zusätzlich setzen wir noch $N_0 := 0$ fest, d.h. zu Beginn der Beobachtungszeit ist das Ereignis noch nicht eingetreten, und bezeichnen mit

$$N_{(s,t]} := N_t - N_s \tag{3.10}$$

die Häufigkeit des Eintretens im Zeitintervall $(s, t] = \{r; s < r \leq t\}$.

In der Vorlesung Wahrscheinlichkeitsrechnung 1 wurde die Verteilung der Zufallsvariablen N_t aus den fünf von Chintschin postulierten Axiomen hergeleitet. Wir gehen hier einen anderen Weg, indem wir die Eigenschaften der *Zwischenankunftszeiten* $Y_k = T_k - T_{k-1}$ vorgeben:

Voraussetzung: Die Y_k seien stochastisch unabhängige und exponentiell verteilte Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ mit für alle k gleichem Parameter λ.

Die Eintrittszeitpunkte T_n erhält man aus den Zwischenankunftszeiten Y_k über die Formel

$$T_n = Y_1 + Y_2 + \cdots + Y_n \tag{3.11}$$

Satz 3.2.1 *Unter der obigen Voraussetzung besitzt die Verteilung des Vektors $T = (T_1, T_2, \ldots, T_m)$ der ersten m Eintrittszeitpunkte des Ereignisses die Dichte*

$$g_m(t_1, t_2, \ldots, t_m) = \begin{cases} \lambda^m e^{-\lambda t_m} & \textit{falls } 0 < t_1 < \cdots < t_m \\ 0 & \textit{sonst} \end{cases} \tag{3.12}$$

Beweis: Die Verteilung des Zufallsvektors $Y = (Y_1, \ldots, Y_m)$ besitzt die Dichte

$$f_m(y_1, y_2, \ldots, y_m) = f(y_1)f(y_2)\cdots f(y_m)$$

mit

$$f(s) = \begin{cases} \lambda e^{-\lambda s} & \text{falls } s > 0 \\ 0 & \text{falls } s \leq 0 \end{cases}$$

Zwischen den Vektoren T und Y besteht gemäß (3.11) die Beziehung

$$\begin{pmatrix} T_1 \\ T_2 \\ T_3 \\ \vdots \\ T_m \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_m \end{pmatrix}$$

oder

$$T = AY$$

mit $det(A) = 1$. Wegen $Y_1 = T_1$, $Y_2 = T_2 - T_1$, $T_3 = Y_3 - Y_2$, ... besitzt nach dem Transformationssatz für Dichten der Zufallsvektor T die Dichte

$$\begin{aligned} g_m(t_1, t_2, \ldots, t_m) &= f_m(t_1, t_2 - t_1, \ldots, t_m - t_{m-1}) \frac{1}{|det(A)|} \\ &= f(t_1)f(t_2 - t_1)\cdots f(t_m - t_{m-1}) \end{aligned}$$

Falls $t_1 \leq 0$ oder $t_i \leq t_{i-1}$, ist der entsprechende Funktionswert auf der rechten Seite und damit $g_m(t_1, t_2, \ldots, t_m)$ gleich Null.
Falls $0 < t_1 < t_2 < \cdots t_m$, ist nach Definition von f

$$\begin{aligned} g_m(t_1, t_2, \ldots, t_m) &= \lambda e^{-\lambda t_1} \lambda e^{-\lambda(t_2 - t_1)} \cdots \lambda e^{-\lambda(t_m - t_{m-1})} \\ &= \lambda^m e^{-\lambda t_1} e^{\lambda t_1} e^{-\lambda t_2} \cdots e^{\lambda t_{m-1}} e^{-\lambda t_m} \\ &= \lambda^m e^{-\lambda t_m} \end{aligned}$$

◇

Satz 3.2.2 *Unter den Voraussetzungen des Satzes 3.2.1 gilt für die Zuwächse* $N_{(s,t]} = N_t - N_s$*:*
1. $N_{(s,t]}$ *ist Poissonverteilt mit Parameter* $\lambda(t - s)$*.*
2. *Für eine beliebige Anzahl* m *von paarweise disjunkten Zeitintervallen* $(s_1, t_1]$, $(s_1, t_1]$, ..., $(s_m, t_m]$ *sind die Zufallsvariablen* $N_{(s_i,t_i]}$, $i = 1 \ldots m$, *stochastisch unabhängig.*

Beweis: Zu zeigen ist, daß mit der Wahrscheinlichkeitsfunktion

$$f(k;\mu) = e^{-\mu}\frac{\mu^k}{k!}\ ;\ k = 0, 1, 2, \ldots$$

der Poissonverteilung mit Parameter μ, für Zeitpunkte $0 = t_0 < t_1 < \ldots < t_n$ und beliebige nichtnegative Zahlen $k_1, k_2, \ldots, k_n$ die Gleichung

$$\mathcal{P}\left(\bigcap_{\nu=1}^{n}(N_{(t_{\nu-1},t_\nu]} = k_\nu)\right) = \prod_{\nu=1}^{n} f(k_\nu; \lambda(t_\nu - t_{\nu-1})) \tag{3.13}$$

erfüllt ist.
Zur Vereinfachung der Darstellung führen wir den Beweis nur für den Fall $n = 2$; die Vorgehensweise im allgemeinen Fall wird daraus offensichtlich.

Da die Wahrscheinlichkeit der Menge $N_Y \subset \Omega$ der Ergebnisse ω mit der Eigenschaft $Y_k(\omega) \leq 0$ für ein oder mehrere k gleich Null ist, kann man sie bei der Berechnung der Wahrscheinlichkeiten vernachlässigen. Wir nehmen daher vereinfachend an, daß $Y_k(\omega) > 0$ für alle k und alle $\omega \in \Omega$.
Unter dieser Annahme treten die Ereignisse $(N_{(0,t_1]} = k_1)$ und $(N_{(t_1,t_2]} = k_2)$ genau dann ein, wenn für die Ankunftszeiten T_i gilt

$$0 < T_1 < \ldots < T_{k_1} \leq t_1 < T_{k_1+1} < \ldots < T_{k_1+k_2} \leq t_2 < T_{k_1+k_2+1}$$

Mit den Mengen

$$\begin{aligned} B_1 &= \{(y_1, \ldots, y_{k_1}) \in \mathbb{R}^{k_1}\ ;\ 0 < y_1 < \ldots y_{k_1} \leq t_1\} \\ B_2 &= \{(y_{k_1+1}, \ldots, y_{k_1+k_2}) \in \mathbb{R}^{k_2}\ ;\ t_1 < y_{k_1+1} < \ldots y_{k_1+k_2} \leq t_2\} \\ B_3 &= \{y_{k_1+k_2+1} \in \mathbb{R}\ ;\ y_{k_1+k_2+1} > t_2\} \end{aligned}$$

gilt daher für den Zufallsvektor $T = (T_1, \ldots, T_m)$ mit $m = k_1 + k_2 + 1$

$$\begin{aligned} & \mathcal{P}\left((N_{(0,t_1]} = k_1) \cap (N_{(t_1,t_2]} = k_2)\right) \\ = \ & \mathcal{P}(T \in B_1 \times B_2 \times B_3) \\ = \ & \int_{B_1\times B_2\times B_3} g_m(y_1, \ldots, y_m)\, d(y_1, \ldots, y_m) \\ = \ & \int_{B_1\times B_2\times B_3} \lambda^m e^{-\lambda y_m}\, d(y_1, \ldots, y_m) \\ = \ & \lambda^{k_1}\lambda^{k_2}\int_{B_1} 1\, d(y_1, \ldots, y_{k_1}) \int_{B_2} 1\, d(y_{k_1+1}, \ldots, y_{k_1+k_2}) \int_{B_3} \lambda e^{-\lambda y_m}\, dy_m \\ = \ & \lambda^{k_1}\lambda^{k_2}\frac{(t_1-0)^{k_1}}{k_1!}\frac{(t_2-t_1)^{k_2}}{k_2!}e^{-\lambda t_2} \\ = \ & \frac{[\lambda(t_1-0)]^{k_1}}{k_1!}e^{-\lambda(t_1-0)}\frac{[\lambda(t_2-t_1)]^{k_2}}{k_2!}e^{-\lambda(t_2-t_1)} \end{aligned}$$

$\diamond$

Satz 3.2.3 *Unter den Voraussetzungen der beiden vohergehenden Sätze ist der Prozeß N_t ein zeitlich homogener Markoff-Prozeß mit den Übergangswahrscheinlichkeiten*

$$p_{ik}(t) = \begin{cases} e^{-\lambda t}\dfrac{(\lambda t)^{k-i}}{(k-i)!} & falls\ k \geq i \\ 0 & falls\ k < i \end{cases} \tag{3.14}$$

Beweis: Für Zeitpunkte $0 \leq t_0 < t_1 < \ldots t_n$ und Zustände $k_0, k_1, \ldots, k_n$ ist

$$\bigcap_{\nu=0}^{n}(N_{t_\nu} = k_\nu) = (N_{(0,t_0]} = k_0) \cap \bigcap_{\nu=1}^{n}(N_{(t_{\nu-1},t_\nu]} = k_\nu - k_{\nu-1})$$

Daraus folgt unter Berücksichtigung der stochastischen Unabhängigkeit der Variablen $N_{(t_{\nu-1},t_\nu]}$

$$\begin{aligned} & \mathcal{P}\left(N_{t_n} = k_n, N_{t_{n-1}} = k_{n-1}, \ldots, N_{t_1} = k_1 | N_{t_0} = k_0\right) \\ = & \frac{\mathcal{P}\left(\bigcap_{\nu=0}^{n}(N_{t_\nu} = k_\nu)\right)}{\mathcal{P}(N_{t_0} = k_0)} \\ = & \frac{\mathcal{P}\left((N_{(0,t_1]} = k_0) \cap \bigcap_{\nu=1}^{n}(N_{(t_{\nu-1},t_\nu]} = k_\nu - k_{\nu-1})\right)}{\mathcal{P}(N_{(0,t_0]} = k_0)} \\ = & \frac{\mathcal{P}(N_{(0,t_1]} = k_0)\prod_{\nu=1}^{n}\mathcal{P}(N_{(t_{\nu-1},t_\nu]} = k_\nu - k_{\nu-1})}{\mathcal{P}(N_{(0,t_0]} = k_0)} \\ = & \prod_{\nu=1}^{n}\mathcal{P}(N_{(t_{\nu-1},t_\nu]} = k_\nu - k_{\nu-1}) \end{aligned}$$

Da die $N_{(t_{\nu-1},t_\nu]}$ nur Werte ≥ 0 annehmen, ist diese bedingte Wahrscheinlichkeit gleich Null, wenn $k_0 < 0$ oder $k_{\nu-1} > k_\nu$ für irgendein ν. In den anderen Fällen gilt nach (3.13)

$$\begin{aligned} & \mathcal{P}\left(N_{t_n} = k_n, N_{t_{n-1}} = k_{n-1}, \ldots, N_{t_1} = k_1 | N_{t_0} = k_0\right) \\ = & \prod_{\nu=1}^{n} e^{-\lambda(t_\nu - t_{\nu-1})}\frac{[\lambda(t_\nu - t_{\nu-1})]^{k_\nu}}{k_\nu!} \end{aligned}$$

Für $n = 1$ ergibt das die behaupteten Übergangswahrscheinlichkeiten p_{ik} und für allgemeines n die Formel (3.5). $\diamond$

Definition 3.2.1 *Der Markoffprozeß $\{N_t\}$ heißt homogener Poissonprozeß mit Parameter λ.*

Die Verweildauer des Poissonprozesses in einem Zustand.
Befindet sich der Prozeß N_t bereits s Zeiteinheiten im Zustand $k-1$, so ist die Wahrscheinlichkeit, daß er noch mindestens t weitere Zeiteinheiten darin bleibt, durch

$$\begin{aligned}\mathcal{P}(Y_k > s+t | Y_k > s) &= \frac{\mathcal{P}\left((Y_k > s+t) \cap (Y_k > s)\right)}{\mathcal{P}(Y_k > s)} \\ &= \frac{\mathcal{P}(Y_k > s+t)}{\mathcal{P}(Y_k > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} \\ &= e^{-\lambda t} = \mathcal{P}(Y_k > t)\end{aligned}$$

gegeben. D.h. unabhängig vom Zeitpunkt, ab dem der Prozeß beobachtet wird, ist die Wartezeit bis zum nächsten Eintreten des Ereignisses exponentiell verteilt mit Parameter λ. Dies ist die Eigenschaft der *Gedächtnislosigkeit* der Exponentialverteilung bzw. des Poissonprozesses.

3.3 Homogene Markoffprozesse

Zur Charakterisierung der Matrizen $P(t) = (p_{ik}(t))$, die Übergangsmatrizen eines zeitlich homogenen Markoffprozesses darstellen, gehen wir von der Voraussetzung aus, daß die Funktionen $p_{ik}(t)$ im Punkt $t = 0$ differenzierbar sind:

$$\begin{aligned} p'_{ii}(0) &= \lim_{t\to 0} \frac{p_{ii}(t) - 1}{t} =: a_{ii} \\ p'_{ik}(0) &= \lim_{t\to 0} \frac{p_{ik}(t) - 0}{t} =: a_{ik} \end{aligned}$$

für $k \neq i$. In Matrizenschreibweise läßt sich diese Eigenschaft zu

$$P'(0) = \lim_{t\to 0} \frac{1}{t} (P(t) - I) =: A$$

zusammenfassen.
Wegen $0 \leq p_{ik}(t) \leq 1$ gilt $a_{ii} \leq 0$ und $a_{ik} \geq 0$ bei $k \neq i$. Aus $\sum_k p_{ik}(t) = 1$ folgt weiter $\sum_k p'_{ik}(0) = \sum_k a_{ik} = 0$.

Definition 3.3.1 *Eine Matrix $A = (a_{ik})$ mit $a_{ik} \geq 0$ für $k \neq i$ und $\sum_{k\neq i} a_{ik} = -a_{ii}$ heißt eine Intensitätsmatrix und ihre Komponenten a_{ik} mit $k \neq i$ sowie $\lambda_i := -a_{ii}$ heißen Intensitäten.*

Die Chapman-Kolmogoroff-Gleichung $P(s + t) = P(s)P(t)$ liefert Differenzierbarkeit von $P(t)$ für alle t:

$$P'(t) = \lim_{s\to 0} \frac{1}{s} (P(s + t) - P(t)) = \lim_{s\to 0} \frac{1}{s} (P(s) - I) P(t) = AP(t)$$

und

$$P'(s) = \lim_{t\to 0} \frac{1}{t} (P(s + t) - P(s)) = P(s) \lim_{t\to 0} \frac{1}{t} (P(t) - I) = P(s)A$$

Satz 3.3.1 *Sei A eine Intensitätsmatrix so, daß es positive relle Zahlen μ und ν mit $0 < \mu \leq -a_{ii} \leq \nu < \infty$ für alle i gibt. Dann besitzen die Kolmogoroffschen Differentialgleichungen*

$$P'(t) = P(t)A \tag{3.15}$$

$$P'(t) = AP(t) \tag{3.16}$$

mit der Anfangsbedingung $P(0) = I$ die Lösung

$$P(t) = e^{tA} = \sum_{n=0}^{\infty} \frac{t^n}{n!} A^n$$

und es gibt einen eindeutig bestimmten zeitlich homogenen Markoffprozeß $\mathcal{P}$ auf $\mathcal{Z}_{\mathcal{S}}$ mit diesen $P(t)$ als Übergangsmatrizen.

Beweis: s. [13]

Die Zustandswahrscheinlichkeiten
Sind die Anfangswahrscheinlichkeiten $p_k = \mathcal{P}(X_0 = k)$ für den Markoffprozeß vorgegeben, so erhält man aus (3.15) ein Differentialgleichungssystem für die Zustandswahrscheinlichkeiten

$$p_k(t) = \mathcal{P}(X_t = k) \tag{3.17}$$

Nach der Formel über die totale Wahrscheinlichkeit ist

$$p_k(t) = \sum_{i \in \mathcal{S}} \mathcal{P}(X_t = k | X_0 = i)\mathcal{P}(X_0 = i) = \sum_{i \in \mathcal{S}} p_i p_{ik}(t)$$

was sich mit den Zeilenvektoren $p = (p_1, p_2, \ldots)$ und $p(t) = (p_1(t), p_2(t), \ldots)$ in der Form

$$p(t) = pP(t)$$

schreiben läßt. Nach (3.15) erhält man für die Ableitung $p'(t)$ dieses Vektors

$$p'(t) = pP'(t) = pP(t)A = p(t)A$$

bzw. für die einzelnen Komponenten

$$\begin{aligned} p'_k(t) &= \sum_{i \in \mathcal{S}} p_i(t) a_{ik} \\ &= \sum_{i \neq k} p_i(t) a_{ik} + p_k(t) a_{kk} \\ &= \sum_{i \neq k} p_i(t) a_{ik} - \sum_{j \neq k} p_k(t) a_{kj} \end{aligned} \tag{3.18}$$

Beispiel: Wie man leicht ausrechnet, ist beim homogenen Poissonprozeß $a_{ii} = -\lambda$, $a_{ii+1} = \lambda$ und $a_{ik} = 0$ für alle anderen k. Mit den Anfangswahrscheinlichkeiten $p_0 = 1$ und $p_k = 0$ für $k > 0$ ergibt sich das Differentialgleichungssystem

$$\begin{aligned} p'_0(t) &= -p_0(t) \\ p'_1(t) &= \lambda p_0(t) - \lambda p_1(t) \\ &\vdots \\ p'_n(t) &= \lambda p_{n-1}(t) - \lambda p_n(t) \\ &\vdots \end{aligned}$$

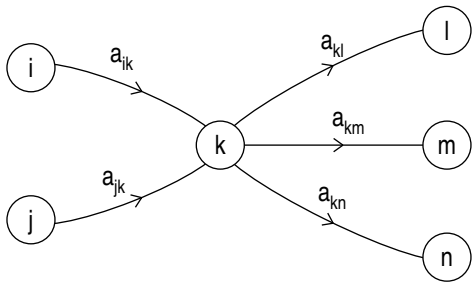


Abbildung 3.1: Der Graph eines Markoffprozesses

Mit den obigen Anfangswerten $p_n(0) = p_n$ ergibt das zunächst $p_0(t) = e^{-\lambda t}$. Mit dem Ansatz $p_n(t) = e^{-\lambda t} q_n(t)$ erhält man die Differentialgleichungen

$$q_n'(t) = \lambda q_{n-1}(t)$$

mit den Lösungen $q_n(t) = (\lambda t)^n / n!$ bzw.

$$p_n(t) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}$$

Der Graph eines Markoffprozesses
Zur Veranschaulichung eines Markoffprozesses, dessen Zustandswahrscheinlichkeiten durch (3.18) gegeben sind, kann man wie bei Markoffketten einen Graphen verwenden, wobei die Knoten des Graphen die Zustände repräsentieren und einer positiven Intensität $a_{ik} > 0$ ein gerichteter Bogen vom Knoten i zum Knoten k entspricht (vgl. Abb. 3.1).
Speziell für den homogenen Poissonprozeß erhält man den in Abb. 3.2 dargestellten Graphen.

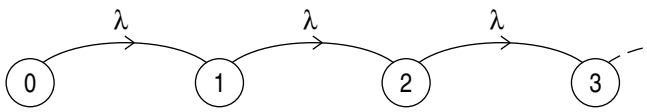


Abbildung 3.2: Der Graph des Poissonprozesses

Der Mechanismus eines Markoffprozesses
Für das Aufstellen des Differentialgleichungssystems (3.18) zu einem vorgegebenen Modell ist es wichtig zu wissen, welche Bedeutung die Intensitäten a_{ik} für den Ablauf eines Markoffprozesses besitzen. Die folgende Darstellung

könnte man natürlich auch als einen mathematischen Satz formulieren. Wir wollen hier darauf verzichten und verweisen auf das Buch [13] von Mathar und Pfeifer und die darin erwähnte weiterführende Literatur. Die nachstehende informelle Beschreibung ist als eine Art Vorschrift zum Erstellen der Differentialgleichungen (3.18) zu verstehen. Sie wird im nächsten Abschnitt an einigen typischen Beispielen verdeutlicht werden.

Voraussetzung: Zu jedem Zustand i gibt es nur endlich viele Zustände k mit $a_{ik} > 0$ und $\mathcal{S}_i \subset \mathcal{S}$ sei die Menge aller dieser „Nachbarzustände".
Mechanismus des Markoffprozesses: Befindet sich der Prozeß zu irgendeinem Zeitpunkt t im Zustand i, so erfolgt der nächste Sprung zu einem anderen Zustand nach dem folgenden Muster: Zum Zeitpunkt t starten zufällige Wartezeiten Y_{ik} für alle $k \in \mathcal{S}_i$. Diese Zufallsvariablen sind stochastisch unabhängig und exponentiell verteilt mit den Parametern a_{ik}. Ist Y_{ij} die kürzeste dieser Wartezeiten, so erfolgt nach der Zeit Y_{ij}, gemessen ab dem Zeitpunkt t, ein Sprung in den Zustand j.

Die Verweildauer $Y_i = \min\{Y_{ik} \ ; \ k \in \mathcal{S}_i\}$ ist aufgrund dieser Vorgaben wieder exponentiell verteilt mit dem Parameter

$$\sum_{k \in \mathcal{S}_i} a_{ik} = \sum_{k \in \mathcal{S}} = -a_{ii} = \lambda_i$$

und die bedingte Wahrscheinlichkeit, daß ein Sprung in den Zustand j erfolgt unter der Bedingung, daß der Zustand i verlassen wird, ist

$$\mathcal{P}(Y_{ij} = \min\{Y_{ik} \ ; \ k \in \mathcal{S}_i\}) = \frac{a_{ij}}{\sum\limits_{k \in \mathcal{S}} a_{ik}} = \frac{a_{ij}}{\sum\limits_{k \in \mathcal{S}_i} a_{ik}} = \frac{a_{ij}}{\lambda_i}$$

Diese beiden Formeln ergeben sich aus den Rechenregeln für exponentiell verteilte Zufallsvariable, wie sie in den Übungen zur Vorlesung Wahrscheinlichkeitsrechnung I behandelt wurden.
Anzumerken bleibt noch, daß die Wahrscheinlichkeit, daß mehrere Wartezeiten zum gleichen Zeitpunkt ablaufen, wegen der absoluten Stetigkeit und der stochastischen Unabhängigkeit der beteiligten Zufallsvariablen gleich Null und diese Ereignisse aus der Sicht der Wahrscheinlichkeitsrechnung somit vernachlässigbar sind.

3.3.1 Markoff-Bedienungssysteme

In einem Bedienungssystem mit einem oder mehreren Schaltern kommen zu zufälligen Zeitpunkten Kunden an. Falls ein Schalter frei ist, wird ein eintreffender Kunde sofort bedient. Andernfalls reiht er sich in einer Warteschlange ein, die nach dem first-in-first-out-Prinzip abgearbeitet wird.

Ist der zeitliche Abstand zwischen dem Eintreffen zweier Kunden exponentiell verteilt mit einem für alle Abstände gleichen Parameter λ, sind die Bedienungszeiten der Kunden stochastisch unabhängige mit einem für alle Kunden gleichen Parameter μ exponentiell verteilte Zufallsvariable und ist m die Anzahl der Bedienungsschalter, so spricht man von einem Bedienungssystem vom Typ M/M/m. Das „M“ wie „Markoff“ steht dabei für den exponentiellen Verteilungstyp bei den Ankünften bzw. Bedienungen.
Die Anzahl N_t der Kunden, die bis zum Zeitpunkt t eintreffen, ist unter diesen Voraussetzungen ein homogener Poissonprozeß und man sagt, daß die Ankünfte einen Poissonstrom bilden.
Als Zustand des Systems zum Zeitpunkt t betrachten wir die Anzahl X_t der Kunden, die sich zum Zeitpunkt t im System befinden, also entweder gerade bedient werden oder in der Warteschlange stehen. Je nachdem der Warteraum eine unbegrenzte Kapazität besitzt oder nur eine endliche Anzahl n von Plätzen aufweist, ist der Zustandsraum gleich $\mathcal{S} = \{0, 1, 2, \ldots\}$ oder eine endliche Menge der Form $\mathcal{S} = \{0, 1, 2, \ldots m + n\}$.
Eine Zustandsänderung kommt dadurch zustande, daß entweder ein neuer Kunde in das System eintritt oder ein Kunde nach Ablauf seiner Bedienungszeit das System verläßt. Ist i der Zustand des Systems zum Zeitpunkt t, so ist die (Rest-)Wartezeit bis zur Ankunft eines neuen Kunden — wie bei der Herleitung des Poissonprozesses angemerkt — $\mathcal{E}(\lambda)$-verteilt, ein Übergang zum Zustand $i + 1$ erfolgt daher mit der Intensität $a_{ii+1} = \lambda$. Die Intensität $\mu_i = a_{ii-1}$, mit der ein Übergang zum Zustand $i - 1$ erfolgt, weil ein Kunde das System verläßt, hängt von der Anzahl m der Bedienungsschalter ab und wird im folgenden für die verschiedenen Fälle einzeln berechnet. Da aus dem Zustand i keine anderen Sprünge möglich sind, ist $a_{ik} = 0$ für $k \neq i, i+1, i-1$.
Bei unbegrenzter Kapazität des Warteraums ergibt sich daraus der in Abb. 3.3 dargestellte Graph,

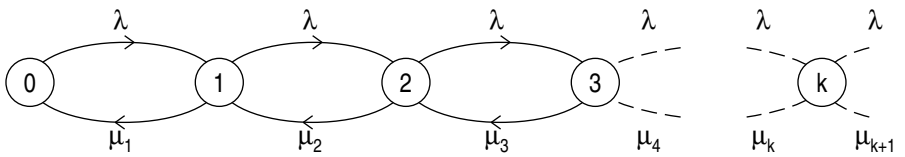


Abbildung 3.3: Der Graph eines Markoff-Bedienungsprozesses

der dem Differentialgleichungssystem

$$\begin{aligned} p_0'(t) &= \mu_1 p_1(t) - \lambda p_0(t) \\ p_1'(t) &= \lambda p_0(t) + \mu_2 p_2(t) - (\lambda + \mu_1) p_1(t) \\ &\vdots \end{aligned}$$

$$\begin{aligned} p_k'(t) &= \lambda p_{k-1}(t) + \mu_{k+1} p_{k+1}(t) - (\lambda + \mu_k) p_k(t) \qquad (3.19) \\ &\vdots \end{aligned}$$

entspricht.
Ist die Kapazität des Warteraums auf n Plätze beschränkt, so können keine neuen Kunden mehr in das System eintreten, wenn sich bereits $m+n$ in ihm befinden. Dies ergibt den in Abb. 3.4 dargestellten Graphen. Die zugehörigen

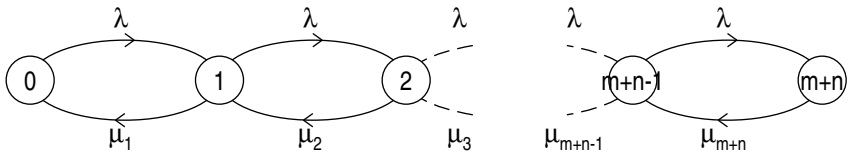


Abbildung 3.4: Der Graph eines Markoff-Bedienungsprozesses

Differentialgleichungen sind für $k = 0, 1, \ldots, m+n-1$ die gleichen wie zuvor. Die letzte Gleichung lautet

$$p_{m+n}'(t) = \lambda p_{m+n-1}(t) - \mu_{m+n} p_{m+n}(t)$$

Spezialfälle

M/M/1: Die Restwartezeit bis zum Ende der Bedienung eines Kunden ist wegen der Gedächtnislosigkeit der Exponentialverteilung genauso wie die gesamte Bedienungszeit $\mathcal{E}(\mu)$-verteilt, die Intensität daher $\mu_i = \mu$ für alle i.

M/M/∞: Dies ist ein Modell für einen großen Parkplatz, wobei unter der Bedienungszeit die Parkdauer zu verstehen ist. Sind gerade n Parkplätze belegt und sind $Y_1, \ldots, Y_n$ die entsprechenden $\mathcal{E}(\mu)$-verteilten Rest-Parkdauern ab dem Beobachtungszeitpunkt, so ist $Y = \min\{Y_1, \ldots, Y_n\}$ die Wartezeit, bis der erste dieser PKWs den Parkplatz verläßt. Da diese Variable $\mathcal{E}(n\mu)$-verteilt ist, gilt $\mu_n = n\mu$.

M/M/m: Falls sich $k \leq m$ Kunden im System befinden, werden sie alle gleichzeitig bedient. Die Wartezeit bis zum Ende der ersten Bedienungszeit ist $\mathcal{E}(k\mu)$-verteilt und $\mu_k = k\mu$. Falls $k > m$, werden m Kunden gleichzeitig bedient und es ist $\mu_k = m\mu$.

3.4 Die stationäre Verteilung eines Markoffprozesses

Sei $\mathcal{P}$ wieder ein zeitlich homogener Markoffprozeß mit Zustandsraum $\mathcal{S} = \mathbb{N}$ oder $\mathcal{S} = \{1, 2, \ldots, n\}$, der Intensitätsmatrix $A = (a_{ik})$ und den Übergangsmatrizen $P(t) = e^{tA}$.
Unter der stationären Verteilung versteht man die Zustandswahrscheinlichkeiten $p_k(t) = \mathcal{P}(X_t = k)$ in der Situation, daß sich das System „eingependelt“ hat. Dieses Einpendeln bedeutet nicht, daß der Zustand X_t zeitlich konstant bleibt. Wie man sich am Geschehen in einem Bedienungssystem klar macht, ist das im allgemeinen unmöglich. Die einzige Art von Stationarität, die bei einem Zufallsprozeß erwartet werden kann, ist die Eigenschaft

$$p_k(t) = p_k \text{ für } t \geq t_0$$

daß die Zustandswahrscheinlichkeiten ab einem bestimmten Zeitpunkt über die Zeit konstant bleiben. Da dies gleichbedeutend damit ist, daß $p_k'(t) = 0$ für diese t gilt, erhält man aus (3.18) die Gleichungen

$$0 = \sum_{i \neq k} p_i a_{ik} - p_k \sum_{j \neq k} a_{kj} \tag{3.20}$$

zusammen mit der Bedingung

$$\sum_{k \in \mathcal{S}} p_k = 1 \tag{3.21}$$

und $p_k \geq 0$ als Charakteristikum einer stationären Verteilung. Wegen $\sum_{j \neq k} a_{kj} = -a_{kk}$ kann man die Gleichungen (3.20) auch zu

$$0 = pA \tag{3.22}$$

mit dem Zeilenvektor $p = (p_1, p_2, \ldots)$ zusammenfassen.

3.4.1 Definition der stationären Verteilung

Definition 3.4.1 *Ein stochastischer Vektor $p = (p_k)$ mit der Eigenschaft (3.20) bzw. (3.22) heißt eine stationäre Verteilung des Markoff-Prozesses $\mathcal{P}$.*

Satz 3.4.1 *Ein stochastischer Vektor p ist genau dann eine stationäre Verteilung des Markoffprozesses, wenn für alle $t \geq 0$ die Gleichung*

$$p = pP(t) \tag{3.23}$$

erfüllt ist.

Beweis: a) Ist p stationäre Verteilung, so ist $pA = 0$ und damit

$$\begin{aligned} pP(t) &= pe^{tA} = p\sum_{m=0}^{\infty}\frac{t^m}{m!}A^m \\ &= \sum_{m=0}^{\infty}\frac{t^m}{m!}pA^m \\ &= pI + tpA + \frac{t^2}{2}(pA)A + \cdots \\ &= p \end{aligned}$$

b) Für $t > 0$ ist die Gleichung $p = pP(t)$ äquivalent zu $p(P(t) - I) = 0$ oder $p\frac{1}{t}(P(t) - I) = 0$. Mit $t \to 0$ ergibt sich daraus $pA = 0$. ◇

$p(t) = p(0)P(t)$ seien die Zustandswahrscheinlichkeiten des Prozesses zur Startverteilung $p(0)$. Ist p eine stationäre Verteilung und gilt für einen Zeitpunkt $t_0 \geq 0$, daß $p(t_0) = p$, so folgt aus dem obigen Satz, daß für alle $t > t_0$

$$p(t) = p(0)P(t) = p(0)P(t_0)P(t - t_0) = p(t_0)P(t - t_0) = pP(t - t_0) = p \tag{3.24}$$

Ein System, das durch diesen Markoffprozeß beschrieben wird, bleibt also im *stationären Zustand* p, wenn es einmal in ihn gelangt ist. Dabei hat das Wort „Zustand“ hier nichts mit Elementen aus dem Zustandsraum zu tun!
Nicht jeder Markoffprozeß besitzt eine stationäre Verteilung. Das typische Gegenbeispiel ist der homogene Poissonprozeß, bei dem die Gleichungen (3.20)

$$\begin{aligned} 0 &= -\lambda p_0 \\ 0 &= \lambda p_0 - \lambda p_1 \\ 0 &= \lambda p_1 - \lambda p_2 \\ &\vdots \end{aligned}$$

mit der Lösung

$$0 = p_0 = p_1 = p_2 = \ldots$$

lauten, was mit der Forderung $\sum_k p_k = 1$ nicht vereinbar ist.

3.4.2 Berechnung der stationären Verteilung

Die Gleichung (3.20) ist formal identisch mit mit der Gleichung (2.65) für die stationäre Verteilung einer Markoffkette. Deshalb gilt auch hier der *Schnittsatz*:

Satz 3.4.2 *Ein stochastischer Vektor $p = (p_1, p_2, \ldots)$ ist genau dann eine stationäre Verteilung, wenn für alle Teilmengen $\mathcal{C} \subset \mathcal{S}$ die Gleichung*

$$\sum_{i\in\mathcal{C}} \sum_{k\in\bar{\mathcal{C}}} p_i a_{ik} = \sum_{k\in\bar{\mathcal{C}}} \sum_{i\in\mathcal{C}} p_k a_{ki}$$

mit dem Komplement $\bar{\mathcal{C}} = \{i \in \mathcal{S}\ ;\ i \notin \mathcal{C}\}$ der Menge $\mathcal{C}$ in $\mathcal{S}$ gilt.

Der Beweis verläuft analog zu dem des Satzes 2.2.3.

Der Schnittsatz liefert ein einfaches Berechnungsverfahren für die stationäre Verteilung des allgemeinen Geburts- und Todesprozesses mit dem Prozeßdiagramm

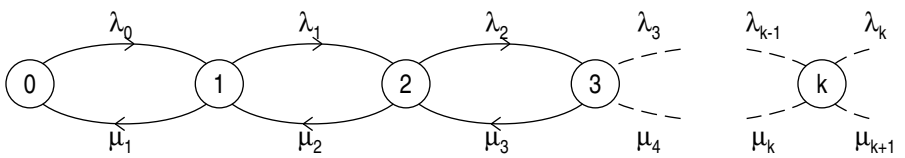


Abbildung 3.5: Der Graph eines Markoff-Bedienungsprozesses

Teilt man den Zustandsraum durch Schnitte zwischen den Knoten $k-1$ und k in die Mengen $\mathcal{C} = \{0, 1, \ldots, k-1\}$ und $\bar{\mathcal{C}} = \{k, k+1, \ldots\}$, so liefert der Schnittsatz die Gleichungen

$$\lambda_{k-1} p_{k-1} = \mu_k p_k$$

bzw.

$$p_k = \frac{\lambda_{k-1}}{\mu_k} p_{k-1}$$

woraus sich rekursiv die Gleichungen

$$p_k = \frac{\lambda_0 \lambda_1 \cdots \lambda_{k-1}}{\mu_1 \mu_2 \cdots \mu_k} p_0 =: \rho_k p_0 \tag{3.25}$$

für $k \geq 1$ ergeben. Mit $\rho_0 := 1$ ist auch der Fall $k = 0$ mit enthalten.
Die Bedingung $\sum_k p_k = 1$ und damit die Existenz einer stationären Verteilung ist offensichtlich nur dann gegeben, wenn $\rho := \sum_{k=0}^{\infty} \rho_k < \infty$ und man erhält in diesem Fall

$$p_k = \frac{\rho_k}{\rho}$$

Markoff-Bedienungssysteme
Bei Markoff-Bedienungssystemen erfolgen die Ankünfte unabhängig vom Zustand des Systems gemäß einem Poissonstrom mit Parameter λ, so daß $\lambda_k = \lambda$ für alle k. Die verschiedenen Systeme unterscheiden sich bei den Bedienungsintensitäten μ_k.

M/M/1 Wegen $\mu_k = \mu$ ist $\rho_k = \rho^k$ mit $\rho = \frac{\lambda}{\mu}$. Falls $\rho < 1$, konvergiert $\sum_k \rho^k = \frac{1}{1-\rho}$ und man erhält

$$p_k = (1-\rho)\rho^k$$

Da $1/\lambda$ die mittlere Zwischenankunftszeit und $1/\mu$ die mittlere Bedienungsdauer ist, sind die Kehrwerte λ bzw. μ die mittlere Anzahl von Ankünften bzw. die mittlere Anzahl von bedienten Kunden pro Zeiteinheit, und die Bedingung $\rho < 1$ besagt, daß sich ein stationärer Systemzustand nur dann einstellen kann, wenn die Zahl der ankommenden Kunden *im Mittel kleiner* ist als die Anzahl der Kunden, die in der gleichen Zeit bedient werden.

M/M/∞ Mit $\mu_k = k\mu$ ist $\rho_k = \frac{1}{k!}\rho^k$, wobei wie oben $\rho = \lambda/\mu$. Da hier $\sum_{k=0}^{\infty} \rho^k = e^{\rho}$, gibt es für alle Werte der Parameter die stationäre Verteilung

$$p_k = e^{-\rho}\frac{\rho^k}{k!}$$

also die Poissonverteilung mit Parameter ρ.

M/M/m Wegen $\mu_k = k\mu$ für $k \leq m$ und $\mu_{m+k} = m\mu$ ist $\rho_k = \frac{\rho^k}{k!}$ für $k \leq m$ und

$$\rho_{m+k} = \frac{\rho^m}{m!}\left(\frac{\lambda}{m\mu}\right)^k$$

wieder mit $\rho = \lambda/\mu$. Konvergenz und damit eine stationäre Verteilung ist für den Fall

$$\frac{\lambda}{m\mu} < 1$$

gegeben.

3.4.3 Das M/M/1-System im stationären Zustand

In diesem Abschnitt werden die wichtigsten Kenngrößen des M/M/1-Systems bei unbegrenzter Kapazität des Warteraums berechnet. Die Darstellung ist — wie schon bei den letzten Abschnitten — aus der Sicht der Wahrscheinlichkeitstheorie nicht rigoros mathematisch, sondern soll die üblichen Berechnungsmethoden bei Warteschlangenproblemen demonstrieren.
Wir nehmen an, daß ein M/M/1-System im stationären Zustand vorliegt. Die Anzahl K_S der Kunden, die sich zu einem beliebigen aber festen Zeitpunkt im System befinden, ist dann eine Zufallsvariable mit der Verteilung

$$\mathcal{P}(K_S = k) = p_k = (1-\rho)\rho^k$$

wobei wie zuvor schon definiert $\rho = \lambda/\mu < 1$.
Die *mittlere Anzahl k_S der Kunden im System* ist der Erwartungswert dieser Zufallsvariablen:

$$\begin{aligned} k_S &= \mathcal{E}(K_S) = \sum_{k=0}^{\infty} k p_k = (1-\rho)\sum_{k=1}^{\infty} k\rho^k \\ &= (1-\rho)\rho\sum_{k=1}^{\infty} k\rho^{k-1} \\ &= (1-\rho)\rho\frac{d}{d\rho}\sum_{k=0}^{\infty}\rho^k = (1-\rho)\rho\frac{d}{d\rho}\frac{1}{1-\rho} \\ &= (1-\rho)\rho\frac{1}{(1-\rho)^2} = \frac{\rho}{(1-\rho)} \\ &= \frac{\lambda}{\mu-\lambda} \end{aligned}$$

Die Anzahl K_W der Kunden in der *Warteschlange* ist

$$K_W = \begin{cases} 0 & \text{falls } K_S \leq 1 \\ K_S - 1 & \text{falls } K_S > 1 \end{cases}$$

mit der Verteilung $\mathcal{P}(K_W = 0) = p_0 + p_1$ und $\mathcal{P}(K_W = k) = p_{k+1}$ für $k > 0$.
Daraus erhält man die *mittlere Länge der Warteschlange k_W* zu

$$\begin{aligned} k_W &= \sum_{k=1}^{\infty} k\mathcal{P}(K_W = k) = \sum_{k=1}^{\infty} k p_{k+1} \\ &= (1-\rho)\rho^2\sum_{k=1}^{\infty} k\rho^{k-1} \\ &= \frac{\rho^2}{1-\rho} = \frac{\lambda}{\mu}\frac{\lambda}{\mu-\lambda} \end{aligned}$$

Um die mittlere Zeit zu berechnen, die ein Kunde in der Warteschlange des Systems verbringt, geht man wie folgt vor: Sei K_S die Anzahl der Kunden, die sich bereits im System befinden, wenn der betrachtete Kunde ankommt. Ist $K_S = 0$, so ist seine Wartezeit $W = 0$. Ist $K_S > 0$, so ist $W = B_1 + B_2 + \cdots + B_{K_S}$, wobei B_1 die Restbedienungszeit des Kunden ist, der gerade bedient wird, und die übrigen B_i die Bedienungszeiten der Kunden, die noch in der Warteschlange stehen.
Es ist also $W = G(B, K_S)$ mit $B = (B_1, B_2, \ldots)$ und der Funktion G mit $G(b, 0) = 0$ und $G(b, n) = b_1 + \cdots + b_n$ für $n > 0$. Die Zufallsvariable K_S hängt nur von den Ankunfts- und Bedienungszeiten der früher eingetroffenen Kunden ab und ist somit von den Bedienungszeiten $B_1, B_2, \ldots$ der jetzt

im System befindlichen bzw. später eintreffenden Kunden stochastisch unabhängig. Daher gilt für die Verteilungsfunktion $F_W(t) = \mathcal{P}(W \leq t)$ der Verteilung von W nach den im ersten Kapitel erarbeiteten Rechenregeln für Zufallsvariable der Form $W = G(B, K_S)$

$$\begin{aligned} F_W(t) &= \sum_{n=0}^{\infty} \mathcal{P}(W \leq t | K_S = n)\mathcal{P}(K_S = n) \\ &= \sum_{n=0}^{\infty} \mathcal{P}(G(B, K_S) \leq t | K_S = n)p_n \\ &= \sum_{n=0}^{\infty} \mathcal{P}(G(B, n) \leq t)p_n \\ &= \sum_{n=0}^{\infty} F_{G(B,n)}(t)p_n \end{aligned}$$

mit den Verteilungsfunktionen $F_{G(B,n)}$ der Zufallsvariablen $G(B, n)$.
Durch Differentiation nach der Variablen t erhält man daraus für die Dichten

$$f_W(t) = \sum_{n=0}^{\infty} f_{G(B,n)}(t)p_n$$

und durch Integration

$$\begin{aligned} \mathcal{E}(W) &= \int_0^{\infty} t f_W(t)dt \\ &= \sum_{n=0}^{\infty} \int_0^{\infty} t f_{G(B,n)}(t)dt \; p_n \\ &= \sum_{n=0}^{\infty} \mathcal{E}(G(B, n))p_n \end{aligned}$$

Wegen

$$\mathcal{E}(G(B, n)) = \mathcal{E}(B_1 + \cdots B_n) = \mathcal{E}(B_1) + \cdots + \mathcal{E}(B_n) = n\frac{1}{\mu}$$

ergibt sich schließlich für die *mittlere Wartezeit eines Kunden* der Ausdruck

$$\mathcal{E}(W) = \sum_{n=0}^{\infty} n\frac{1}{\mu}p_n = \frac{1}{\mu}k_S = \frac{1}{\mu}\frac{\lambda}{\mu - \lambda}$$

Die Verweildauer V eines Kunden setzt sich zusammen aus der Wartezeit W und seiner Bedienungszeit B_0, so daß sich die *mittlere Verweildauer* zu

$$\mathcal{E}(V) = \mathcal{E}(W) + \mathcal{E}(B_0) = \frac{1}{\mu}\frac{\lambda}{\mu - \lambda} + \frac{1}{\mu} = \frac{1}{\mu - \lambda}$$

berechnet.

3.4.4 Existenz einer stationären Verteilung

Aus den obigen Beispielen ersieht man, daß nicht in jedem Fall eine stationäre Verteilung existiert, sondern bestimmte Voraussetzungen erfüllt sein müssen. Für einen *endlichen* Zustandsraum $\mathcal{S}$ ergeben sich hinreichende Bedingungen für die Existenz aus der folgenden Aussage über die Übergangswahrscheinlichkeiten $p_{ik}(t)$ des Prozesses:

Satz 3.4.3 *Der Zustandsraum des Markoffprozesses sei endlich — $\mathcal{S} = \{1, 2, \ldots, n\}$ — und es gebe einen Zeitpunkt $t^* > 0$ so, daß $p_{ik}(t^*) > 0$ für alle Zustände i und k aus $\mathcal{S}$. Dann gibt es Zahlen $p_k^* > 0$ mit*

$$\lim_{t\to\infty} p_{ik}(t) = p_k^* \tag{3.26}$$

für alle Zustände i und jeden Zustand k.

Die Eigenschaft (3.26) zieht die folgenden Konsequenzen nach sich:

- $p^* = (p_1^*, p_2^*, \ldots, p_n^*)$ ist ein stochastischer Vektor! Da man bei endlich vielen Summanden das Summenzeichen und lim vertauschen darf, ist

$$\sum_{k\in\mathcal{S}} p_k^* = \sum_{k\in\mathcal{S}} \lim_{t\to\infty} p_{ik}(t) = \lim_{t\to\infty} \sum_{k\in\mathcal{S}} p_{ik}(t) = \lim_{t\to\infty} 1 = 1$$

- Ist p eine beliebige Startverteilung und $p(t) = pP(t)$, so gilt $\lim_{t\to\infty} p(t) = p^*$, denn

$$\begin{aligned}\lim_{t\to\infty} p_k(t) &= \lim_{t\to\infty} \sum_{k\in\mathcal{S}} p_i p_{ik}(t) \\ &= \sum_{k\in\mathcal{S}} p_i \lim_{t\to\infty} p_{ik}(t) \\ &= \left(\sum_{k\in\mathcal{S}} p_i\right) p_k^* = p_k^*\end{aligned}$$

- p^* ist eine stationäre Verteilung! Aus der Gleichung

$$p(s+t) = pP(s+t) = pP(s)P(t) = p(s)P(t)$$

und $\lim_{s\to\infty} p(s+t) = \lim_{s\to\infty} p(s) = p^*$ erhält man $p^* = p^*P(t)$ und daraus nach Satz 3.4.1 die Behauptung.

- p^* ist die einzige stationäre Verteilung. Ist $\tilde{p}$ eine weitere stationäre Verteilung und $p(t) = \tilde{p}P(t)$, so gilt einmal nach (3.24) $p(t) = \tilde{p}$ für alle t und zum anderen $\lim_{t\to\infty} p(t) = p^*$, was zusammen $\tilde{p} = p^*$ ergibt.

Beweis des Satzes 3.4.3
a) Sei $\delta = \min_{i,k\in\mathcal{S}} p_{ik}(t^*)$. Da das Minumum wegen $\mathcal{S} = \{1, 2, \ldots, n\}$ nur über eine endliche Anzahl positiver Summanden gebildet wird, ist $\delta > 0$ und zusammen mit $1 = \sum_{k=1}^{n} p_{ik}(t^*) \geq n\delta$ folgt weiter

$$0 \leq 1 - n\delta < 1$$

b) Sei im folgenden k ein beliebiger aber fester Zustand aus $\mathcal{S}$ und

$$\begin{aligned} b_k(t) &= \min_{i\in\mathcal{S}} p_{ik}(t) \\ B_k(t) &= \max_{i\in\mathcal{S}} p_{ik}(t) \end{aligned}$$

Dann ist

$$b_k(t^*) > 0 \tag{3.27}$$

und für alle Zustände i

$$b_k(t) \leq p_{ik}(t) \leq B_k(t) \tag{3.28}$$

Im folgenden wird gezeigt, daß die Funktion $t \longmapsto b_k(t)$ monoton nichtfallend, die Funktion $t \longmapsto B_k(t)$ monoton nichtsteigend und $\lim_{t\to\infty}(B_k(t) - b_k(t)) = 0$ ist. Daraus folgt, daß beide Funktionen für $t \to \infty$ den gleichen Limes p_k^* besitzen, der wegen (3.27) positiv und wegen (3.28) auch der Limes von $t \longmapsto p_{ik}(t)$ sein muß.
c) Monotonie der Funktionen b_k und B_k.
Für $\Delta t > 0$ gilt nach der Gleichung von Chapman und Kolmogoroff

$$\begin{aligned} b_k(t + \Delta t) &= \min_i p_{ik}(t + \Delta t) \\ &= \min_i \sum_j p_{ij}(\Delta t) p_{jk}(t) \\ &\geq \min_i \sum_j p_{ij}(\Delta t) b_k(t) \\ &= 1 \cdot b_k(t) = b_k(t) \end{aligned}$$

und analog zeigt man, daß $B_k(t + \Delta t) \leq B_k(t)$.
d) Abschätzung von $B_k(t) - b_k(t)$, 1.Teil
Für zwei beliebige aber für den Augenblick festgehaltene Zustände i und m seien

$$\begin{aligned} \mathcal{S}_+ &= \{j \in \mathcal{S} \,;\, p_{ij}(t^*) \geq p_{mj}(t^*)\} \\ \mathcal{S}_- &= \{j \in \mathcal{S} \,;\, p_{ij}(t^*) < p_{mj}(t^*)\} \end{aligned}$$

Dann ist

$$\begin{aligned} 0 &= \sum_j p_{ij}(t^*) - \sum_j p_{mj}(t^*) \\ &= \sum_j (p_{ij}(t^*) - p_{mj}(t^*)) \\ &= \sum_{j\in\mathcal{S}_+} (p_{ij}(t^*) - p_{mj}(t^*)) + \sum_{j\in\mathcal{S}_-} (p_{ij}(t^*) - p_{mj}(t^*)) \end{aligned}$$

oder

$$\sum_{j\in\mathcal{S}_+} (p_{ij}(t^*) - p_{mj}(t^*)) = -\sum_{j\in\mathcal{S}_-} (p_{ij}(t^*) - p_{mj}(t^*))$$

wobei die Summanden der ersten Summe alle ≥ 0 und die der zweiten Summe alle negativ sind.
Für $t > t^*$ folgt aus dieser Zerlegung dann die Abschätzung

$$\begin{aligned} &\sum_j (p_{ij}(t^*) - p_{mj}(t^*))\, p_{jk}(t-t^*) \\ = &\sum_{j\in\mathcal{S}_+} (p_{ij}(t^*) - p_{mj}(t^*))\, p_{jk}(t-t^*) + \sum_{j\in\mathcal{S}_-} (p_{ij}(t^*) - p_{mj}(t^*))\, p_{jk}(t-t^*) \\ \leq &\sum_{j\in\mathcal{S}_+} (p_{ij}(t^*) - p_{mj}(t^*))\, B_k(t-t^*) + \sum_{j\in\mathcal{S}_-} (p_{ij}(t^*) - p_{mj}(t^*))\, b_k(t-t^*) \\ = &\sum_{j\in\mathcal{S}_+} (p_{ij}(t^*) - p_{mj}(t^*))\, B_k(t-t^*) - \sum_{j\in\mathcal{S}_+} (p_{ij}(t^*) - p_{mj}(t^*))\, b_k(t-t^*) \\ = &\sum_{j\in\mathcal{S}_+} (p_{ij}(t^*) - p_{mj}(t^*))\, (B_k(t-t^*) - b_k(t-t^*)) \\ = &\left[\sum_{j\in\mathcal{S}_+} p_{ij}(t^*) - \sum_{j\in\mathcal{S}_+} p_{mj}(t^*)\right] (B_k(t-t^*) - b_k(t-t^*)) \end{aligned}$$

Die Menge $\mathcal{S}_+$ bestehe aus w und damit die Menge $\mathcal{S}_-$ aus $n-w$ Elementen. Dann gilt $\sum_{j\in\mathcal{S}_+} p_{mj}(t^*) \geq w\delta$ bzw. $-\sum_{j\in\mathcal{S}_+} p_{mj}(t^*) \leq -w\delta$ und entsprechend $\sum_{j\in\mathcal{S}_+} p_{ij}(t^*) = 1 - \sum_{j\in\mathcal{S}_-} p_{ij}(t^*) \leq 1-(n-w)\delta$, wodurch man die Abschätzung

$$\begin{aligned} &\sum_j (p_{ij}(t^*) - p_{mj}(t^*))\, p_{jk}(t-t^*) \\ \leq &\left[1-(n-w)\delta - w\delta\right] (B_k(t-t^*) - b_k(t-t^*)) \\ = &\; q\,(B_k(t-t^*) - b_k(t-t^*)) \end{aligned}$$

mit $q = 1-n\delta$ erhält.
e) Abschätzung von $B_k(t) - b_k(t)$, 2.Teil

Für $t > t^*$ folgt aus der Gleichung von Chapman und Kolmogoroff und Teil d), daß

$$\begin{aligned} B_k(t) - b_k(t) &= \max_i p_{ik}(t) \; - \; \min_m p_{mk}(t) \\ &= \max_{i,m} \left(p_{ik}(t) - p_{mk}(t)\right) \\ &= \max_{i,m} \sum_j \left(p_{ij}(t^*) - p_{mj}(t^*)\right) p_{jk}(t - t^*) \\ &\leq q \left(B_k(t - t^*) - b_k(t - t^*)\right) \end{aligned}$$

Für $t > Nt^*$ liefert die iterierte Anwendung dieser Ungleichung

$$B_k(t) - b_k(t) \leq q^N \left(B_k(t - t^*) - b_k(t - t^*)\right) \leq q^N$$

so daß wegen $0 \leq q < 1$ die Differenzen $B_k(t) - b_k(t)$ mit $t \to \infty$ gegen Null konvergieren. $\diamond$

Um ein einfaches Kriterium dafür zu finden, wann die Voraussetzung $p_{ik}(t^*) > 0$ des Satzes 3.4.3 erfüllt ist, zerlegt man das Intervall $(0, t^*]$ in m disjunkte Teilintervalle der Länge $\Delta t = t^*/m$. Durch mehrfache Anwendung der Gleichung von Chapman und Kolmogoroff auf $p_{ik}(t^*)$ erhält man eine Summe von Pfadwahrscheinlichkeiten der Form

$$p_{ij_1}(\Delta t) p_{j_1 j_2}(\Delta t) \cdots p_{j_{m-1} k}(\Delta t) \tag{3.29}$$

wobei über alle Pfade der Länge m des Prozeßdiagramms summiert wird, die von i über Zwischenknoten j_μ nach k führen. Wenn $p_{ik}(t^*) > 0$ ist, müssen wenigstens eines der Produkte (3.29) und damit alle seine Faktoren positiv sein. Da für kleine Δt aber $p_{jl}(\Delta t) \approx a_{jl} \Delta t$ bei $j \neq l$, müssen alle $a_{j_\mu j_{\mu+1}}$ mit $j_\mu \neq j_{\mu+1}$ ebenfalls positiv sein. Diese Überlegung liefert den

Satz 3.4.4 *Die Voraussetzungen des Satzes 3.4.3 sind erfüllt, wenn im Prozeßdiagramm jeder Zustand von jedem aus über Bögen mit positiven a_{ij} erreichbar oder — anders ausgedrückt — wenn die eingebettete Markoffkette irreduzibel ist.*

Eine entsprechende Aussage läßt sich auch für Markoffprozesse mit abzählbarem Zustandsraum beweisen. Man muß in diesem Fall aber noch zusätzlich fordern, daß alle Zustände der eingebetteten Markoffkette positiv rekurrent sind, was bei irreduziblen Markoffketten mit endlichem Zustandsraum automatisch gegeben ist.

Satz 3.4.5 *Die Diagonalelemente der Intensitätsmatrix seien durch* $0 < \mu \leq -a_{ii} \leq \nu < \infty$ *beschränkt und die eingebettete Markoffkette sei irreduzibel und besitze positiv rekurrente und aperiodische Zustände. Dann gibt es eine eindeutig bestimmte stationäre Verteilung* $p^* = (p_1^*, p_2^*, \ldots)$ *mit*

$$\lim_{t \to \infty} p_{ik}(t) = p_k^*$$

für alle Zustände i *und* k.

3.5 Bevölkerungsentwicklungen

In diesem Abschnitt wird die Entwicklung einer (abstrakten) Population betrachtet, deren Bestand sich durch Geburts- und Todesfälle zu zufälligen Zeitpunkten ändert. Die Master-Gleichungen für die Zustandswahrscheinlichkeiten $p_n(t) = \mathcal{P}(X_t = n)$ werden aus allgemeinen Annahmen über das Auftreten und die Häufigkeit der Geburts- und Todesfälle hergeleitet und in dem speziellen Fall des zeitkontinuierlichen Verzweigungsprozesses die erzeugenden Funktionen der Zustandswahrscheinlichkeiten berechnet.

3.5.1 Die $o(t)$-Notation

Zur Herleitung der Mastergleichungen wird bei der hier betrachteten Art von Prozessen eine spezielle Notation zur Beschreibung des Verhaltens einer Funktion $f(t)$ für $t \to 0$ verwendet.
Wir nennen eine Funktion $f : [0, \infty) \longrightarrow \mathbb{R}$ *von kleinerer Ordnung als* t, wenn

$$\lim_{t \to 0} \frac{f(t)}{t} = 0$$

Um anzuzeigen, daß eine Funktion f bzw. Teile von ihr diese Eigenschaft besitzen, schreibt man

$$f(t) = o(t) \text{ bzw. } f(t) = f_0(t) + o(t)$$

Für eine Funktion $f(t) = a + bt + o(t)$ gilt daher

$$f(0) = \lim_{t \to 0} f(t) = a$$

und

$$f'(0) = \lim_{t \to 0} \frac{1}{t}(f(t) - a) = \lim_{t \to 0}(b + \frac{o(t)}{t}) = b$$

Die $o(t)$-Notation gestattet es, komplexe Formeln zu vereinfachen, indem alle Terme, die von kleinerer Ordnung als t sind, zu einem Term $o(t)$ zusammengefaßt werden.
Sind z.B. $f_1(t), f_2(t), \ldots, f_n(t)$ vom Typ $o(t)$ und $g_1(t), g_2(t), \ldots, g_n(t)$ Funktionen, die bei $t = 0$ beschränkt sind, so ist, wie man sich leicht überlegt,

$$g(t) = f_1(t)g_1(t) + f_2(t)g_2(t) + \cdots + f_n(t)g_n(t) = o(t)$$

Ebenso ist

$$\begin{aligned} & (a + bt + o(t))(A + Bt + o(t)) \\ = \; & aA + (aB + bA)t + \Big[bBt^2 + (a + A)o(t) + (b + B)to(t) + o(t)o(t)\Big] \\ = \; & aA + (aB + bA)t + o(t) \end{aligned}$$

3.5.2 Der allgemeine Geburts- und Todesprozeß

Wir betrachten eine abstrakte Population bestehend aus identischen Mitgliedern und der Anzahl der Mitglieder zum Zeitpunkt t als Zustandsgröße. Zustandsänderungen erfolgen zu zufälligen Zeitpunkten, wobei wir Zunahmen als *Geburten* und Abnahmen als *Todesfälle* bezeichnen wollen. $G_{t,\Delta t}$ sei die Anzahl der Geburten, $T_{t,\Delta t}$ die Anzahl der Todesfälle und $D_{t,\Delta t} = G_{t,\Delta t} - T_{t,\Delta t}$ die Nettobestandsänderung im Zeitintervall $(t, t+\Delta t]$. Über die Verteilungen der Zufallsvariablen $G_{t,\Delta t}$ und $T_{t,\Delta t}$ treffen wir die folgenden Annahmen:

$$G_{t,\Delta t} \text{ und } T_{t,\Delta t} \text{ sind stochastisch unabhängig} \tag{3.30}$$

$$\mathcal{P}(G_{t,\Delta t} = 1 | X_t = n) = \lambda_n \Delta t + o(\Delta t) \text{ mit } \lambda_0 = 0 \tag{3.31}$$

$$\mathcal{P}(G_{t,\Delta t} > 1) = o(\Delta t) \tag{3.32}$$

$$\mathcal{P}(T_{t,\Delta t} = 1 | X_t = n) = \mu_n \Delta t + o(\Delta t) \text{ mit } \mu_0 = 0 \tag{3.33}$$

$$\mathcal{P}(T_{t,\Delta t} > 1) = o(\Delta t) \tag{3.34}$$

Aus (3.31) und (3.32) folgt

$$\begin{aligned} \mathcal{P}(G_{t,\Delta t} > 1 | X_t = n) &= \frac{\mathcal{P}((G_{t,\Delta t} > 1) \cap (X_t = n))}{\mathcal{P}(X_t = n)} \\ &\leq \frac{1}{\mathcal{P}(X_t = n)} \mathcal{P}(G_{t,\Delta t} > 1) = o(t) \end{aligned}$$

und

$$\begin{aligned} \mathcal{P}(G_{t,\Delta t} = 0 | X_t = n) &= 1 - \mathcal{P}(G_{t,\Delta t} > 0 | X_t = n) \\ &= 1 - \mathcal{P}(G_{t,\Delta t} = 1 | X_t = n) - \mathcal{P}(G_{t,\Delta t} > 1 | X_t = n) \\ &= 1 - \lambda_n \Delta t + o(\Delta t) \end{aligned}$$

Analog erhält man für die Variable $T_{t,\Delta t}$ aus (3.33) und (3.34)

$$\mathcal{P}(T_{t,\Delta t} > 1 | X_t = n) = o(\Delta t)$$

und

$$\mathcal{P}(T_{t,\Delta t} = 0 | X_t = n) = 1 - \mu_n \Delta t + o(\Delta t)$$

Für die Variable $D_{t,\Delta t}$ ergeben sich daraus die folgenden Eigenschaften:

$$\mathcal{P}(D_{t,\Delta t} = 0 | X_t = n) = 1 - (\lambda_n + \mu_n)\Delta t + o(\Delta t) \tag{3.35}$$

Beweis:

$$\begin{aligned}
& \mathcal{P}(D_{t,\Delta t}=0|X_t=n)=\mathcal{P}\left(\sum_{k\geq 0}(G_{t,\Delta t}=k)\cap(T_{t,\Delta t}=k)\right)\\
= \;& \mathcal{P}\left((G_{t,\Delta t}=0)\cap(T_{t,\Delta t}=0)\right)+\mathcal{P}\left((G_{t,\Delta t}=1)\cap(T_{t,\Delta t}=1)\right)\\
& +\mathcal{P}\left(\sum_{k\geq 2}(G_{t,\Delta t}=k)\cap(T_{t,\Delta t}=k)\right)\\
= \;& \mathcal{P}\left(G_{t,\Delta t}=0)\right)\mathcal{P}\left(T_{t,\Delta t}=0)\right)+\mathcal{P}\left(G_{t,\Delta t}=1)\right)\mathcal{P}\left(T_{t,\Delta t}=1)\right)\\
& +\mathcal{P}\left(B\right)\\
= \;& (1-\lambda_n\Delta t+o(\Delta t))(1-\mu_n\Delta t+o(\Delta t))+(\lambda_n\Delta t+o(\Delta t))(\mu_n+o(\Delta t))\\
& +\mathcal{P}\left(B\right)\\
= \;& 1-(\lambda_n+\mu_n)\Delta t+o(\Delta t)+\mathcal{P}\left(B\right)
\end{aligned}$$

wobei wegen $B\subset(G_{t,\Delta t}>1)$ auch $\mathcal{P}\left(B\right)=o(\Delta t)$ ist. $\diamond$

$$\mathcal{P}(D_{t,\Delta t}=1|X_t=n)=\lambda_n\Delta t+o(\Delta t) \tag{3.36}$$

Beweis:

$$\begin{aligned}
& \mathcal{P}(D_{t,\Delta t}=1|X_t=n)\\
= \;& \mathcal{P}\left(\sum_{k\geq 0}(G_{t,\Delta t}=k+1)\cap(T_{t,\Delta t}=k)\right)\\
= \;& \mathcal{P}\left((G_{t,\Delta t}=1)\cap(T_{t,\Delta t}=0)\right)+\mathcal{P}\left(\sum_{k\geq 1}(G_{t,\Delta t}=k+1)\cap(T_{t,\Delta t}=k)\right)\\
= \;& \mathcal{P}\left(G_{t,\Delta t}=1)\right)\mathcal{P}\left(T_{t,\Delta t}=0)\right)+\mathcal{P}\left(B\right)\\
= \;& (\lambda_n\Delta t+o(\Delta t))(1-\mu_n\Delta t+o(\Delta t))+\mathcal{P}\left(B\right)\\
= \;& \lambda_n\Delta t+o(\Delta t)+\mathcal{P}\left(B\right)
\end{aligned}$$

wobei wegen $B\subset(G_{t,\Delta t}>1)$ auch $\mathcal{P}\left(B\right)=o(\Delta t)$ ist. $\diamond$
Analog berechnet man

$$\mathcal{P}(D_{t,\Delta t}=-1|X_t=n)=\mu_n\Delta t+o(\Delta t) \tag{3.37}$$

und schließlich gilt wegen $(|D_{t,\Delta t}|>1)\subset(G_{t,\Delta t}>1)\cup(T_{t,\Delta t}>1)$ auch noch

$$\mathcal{P}(|D_{t,\Delta t}|>1)=o(\Delta t) \tag{3.38}$$

Mit Hilfe dieser Formeln kann man die Master-Gleichungen für die Zustandswahrscheinlichkeiten $p_n(t)$ auf die folgende Weise herleiten: Für $n\geq 1$ ist

$$p_n(t+\Delta t) \;=\; \mathcal{P}(X_{t+\Delta t}=n)$$

$$
\begin{aligned}
&= \mathcal{P}((X_{t+\Delta t} = n) \cap (|X_t - n| \leq 1)) \\
&\quad + \mathcal{P}((X_{t+\Delta t} = n) \cap (|X_t - n| > 1)) \\
&= \mathcal{P}(\sum_{k=-1}^{1} (X_{t+\Delta t} = n) \cap (X_t = n - k)) + \mathcal{P}(B) \\
&= \mathcal{P}(\sum_{k=-1}^{1} (D_{t,\Delta t} = k) \cap (X_t = n - k)) + \mathcal{P}(B) \\
&= \sum_{k=-1}^{1} \mathcal{P}(D_{t,\Delta t} = k | X_t = n - k) \mathcal{P}(X_t = n - k)) + \mathcal{P}(B)
\end{aligned}
$$

Wegen $B \subset (|D_{t,\Delta t}| > 1)$ ist $\mathcal{P}(B) = o(\Delta t)$ und unter Verwendung von (3.35) – (3.37) erhält man

$$
\begin{aligned}
p_n(t + \Delta t) &= (\lambda_{n-1}\Delta t + o(\Delta t))p_{n-1}(t) \\
&\quad + (1 - (\lambda_n + \mu_n)\Delta t + o(\Delta t))p_n(t) \\
&\quad + (\mu_{n+1}\Delta t + o(\Delta t))p_{n+1}(t) + o(\Delta t)
\end{aligned}
$$

Bringt man $1 \cdot p_n(t)$ auf die linke Seite der Gleichung und dividiert beide Seiten durch Δt, so erhält man

$$
\begin{aligned}
\frac{p_n(t + \Delta t) - p_n(t)}{\Delta t} &= \lambda_{n-1}p_{n-1}(t) + \mu_{n+1}p_{n+1}(t) \\
&\quad - (\lambda_n + \mu_n)p_n(t) + \frac{o(\Delta t)}{\Delta t}
\end{aligned}
$$

und im Limes für $\Delta t \to 0$

$$
p_n'(t) = \lambda_{n-1}p_{n-1}(t) + \mu_{n+1}p_{n+1}(t) - (\lambda_n + \mu_n)p_n(t) \tag{3.39}
$$

Für $n = 0$ erhält man auf die selbe Weise

$$
p_0'(t) = \mu_1 p_1(t) - \lambda_0 p_0(t) \tag{3.40}
$$

3.5.3 Der zeitkontinuierliche Verzweigungsprozeß

Im Spezialfall $\lambda_n = n\lambda$ und $\mu_n = n\mu$ kann man die Gleichungen (3.39) und (3.40) explizit lösen. Dieser Ansatz entspricht einer Population, in der die einzelnen Mitglieder unabhängig voneinander mit Wahrscheinlichkeit $\lambda\Delta t + o(\Delta t)$ in einem Zeitintervall der Länge Δt einen Nachkommen erzeugen und mit Wahrscheinlichkeit $\mu\Delta t + o(\Delta t)$ aus der Population ausscheiden. Vergleicht man diese Situation mit der in Abschnitt 1.3 beschriebenen, so sieht

man, daß es sich um einen Verzweigungsprozeß handelt, bei dem die Zeit nicht in Generationen sondern kontinuierlich gemessen wird. Die Master-Gleichungen für diesen Spezialfall des Geburts- und Todesprozesses lauten

$$\begin{aligned} p_0'(t) &= \mu p_1(t) \\ p_n'(t) &= (n-1)\lambda p_{n-1}(t) + (n+1)\mu p_{n+1}(t) - n(\lambda+\mu)p_n(t) \end{aligned} \quad (3.41)$$

Die Annahme, daß die Population zum Zeitpunkt $t = 0$ mit einem Mitglied startet, ergibt die Anfangsbedingungen

$$p_1(0) = 1 \text{ und } p_n(0) = 0 \text{ für } n \neq 1 \quad (3.42)$$

Zur Berechnung der Zustandswahrscheinlichkeiten $p_n(t)$ benutzen wir die *erzeugenden Funktionen*

$$f(z,t) = \sum_{n=0}^{\infty} p_n(t) z^n \quad (3.43)$$

für $0 \leq z \leq 1$.
Multipliziert man die Gleichungen (3.41) mit z^n und summiert über n, so erhält man

$$\begin{aligned} \sum_{n=0}^{\infty} p_n'(t) z^n &= \sum_{n=0}^{\infty} (n+1)\mu p_{n+1}(t) z^n \\ &\quad + \sum_{n=1}^{\infty} (n-1)\lambda p_{n-1}(t) z^n \\ &\quad - \sum_{n=1}^{\infty} n(\lambda+\mu) p_n(t) z^n \\ &= [\mu - (\lambda+\mu)z + \lambda z^2] \sum_{n=1}^{\infty} p_n(t) n z^{n-1} \end{aligned}$$

also die partielle Differentialgleichung

$$\frac{\partial f}{\partial t}(z,t) - [\mu - (\lambda+\mu)z + \lambda z^2] \frac{\partial f}{\partial z}(z,t) = 0 \quad (3.44)$$

die wegen $\sum_{n=0}^{\infty} p_n(0) z^n = z$ mit der Randbedingung

$$f(z,0) = z \quad (3.45)$$

zu lösen ist.
Zur Lösung geht man bei diesem Typ — lineare partielle Differentialgleichung erster Ordnung — nach der Charakteristikenmethode vor.

Ist $t \longmapsto z(t)$ eine Kurve, für die die Funktion $t \longmapsto f(z(t),t) = C$ als Funktion von t konstant ist, so gilt

$$\frac{d}{dt}\left(f(t,z(t)\right) = \frac{\partial f}{\partial t}(z(t),t) + \frac{\partial f}{\partial z}(z(t),t)\frac{dz}{dt}(t) = 0$$

und im Vergleich mit (3.44) sieht man, daß die Lösungen f dieser Differentialgleichung dadurch charakterisiert sind, daß für alle Lösungen $z(t)$ der gewöhnlichen Differentialgleichung

$$z' = -\mu + (\lambda + \mu)z - \lambda z^2 \tag{3.46}$$

$f(z(t),t) = C = const.$

(3.46) ist eine Riccatische Differentialgleichung der Form $z' = a + bz + cz^2$ mit $a + b + c = 0$. Zur Lösung müssen zwei Fälle unterschieden werden.

Fall $\lambda \neq \mu$
Hier lautet sie

$$z(t) = \frac{C + \mu E(t)}{C + \lambda E(t)} \tag{3.47}$$

mit

$$E(t) = e^{(\mu-\lambda)t}$$

Eine Funktion von t und z, die auf dieser Lösung $z(t)$ zeitlich konstant ist, erhält man durch Auflösen der Gleichung (3.47) nach der Konstanten C:

$$C = \frac{\lambda z - \mu}{1 - z}E(t) =: u(z,t)$$

Man rechnet leicht nach, daß $f(z,t) = g(u(z,t))$ mit einer beliebigen reellen differenzierbaren Funktion die gleichen Eigenschaften wie u besitzt, und die Theorie der partiellen Differentialgleichungen besagt, daß man auf diese Weise alle Lösungen erhält. Die allgemeine Lösung von (3.44) lautet also

$$f(z,t) = g\left(\frac{\lambda z - \mu}{1 - z}E(t)\right)$$

und die Funktion g wird dadurch festgelegt, daß die Randbedingung

$$f(z,0) = g\left(\frac{\lambda z - \mu}{1 - z}\right) = z$$

erfüllt sein muß. $z = g(y)$ ist die Umkehrfunktion von

$$y = \frac{\lambda z - \mu}{1 - z}$$

und man erhält sie, wenn man die letzte Gleichung nach z auflöst, was

$$z = \frac{y+\mu}{y+\lambda}$$

ergibt. Die endgültige Lösung von (3.44) und (3.45) lautet damit

$$\begin{aligned} f(z,t) &= g\left(\frac{\lambda z-\mu}{1-z}E(t)\right) \\ &= \frac{(\lambda z-\mu)E(t)+\mu(1-z)}{(\lambda z-\mu)E(t)+\lambda(1-z)} \end{aligned}$$

Die Zustandswahrscheinlichkeiten ließen sich jetzt durch die Entwicklung der Funktion $f(z,t)$ in eine Potenzreihe nach der Variablen z bestimmen. Im Rahmen der Verzweigungsprozesse interessiert aber eigentlich nur die Wahrscheinlichkeit $a = \lim_{t\to\infty} p_0(t)$, daß die Population irgendwann einmal ausstirbt, und die Entwicklung der mittleren Größe $\mathcal{E}X_t$ der Population, die man beide mit Hilfe der erzeugenden Funktion bestimmen kann.
Es gilt

$$p_0(t) = f(0,t) = \frac{\mu(1-E(t))}{\lambda-\mu E(t)} \longrightarrow \begin{cases} \frac{\mu}{\lambda} & \text{falls } \lambda > \mu \\ 1 & \text{falls } \lambda < \mu \end{cases}$$

und

$$\mathcal{E}X_t = \frac{\partial f}{\partial z}(1,t) = e^{(\lambda-\mu)t} \longrightarrow \begin{cases} \infty & \text{falls } \lambda > \mu \\ 0 & \text{falls } \lambda < \mu \end{cases}$$

Im Fall $\lambda < \mu$ besitzt die Zufallsvariable $W =$ *Wartezeit bis zum Aussterben der Population* wegen

$$(W \le t) = (X_t = 0)$$

die Verteilungsfunktion

$$F(t) = \mathcal{P}(W \le t) = p_0(t)$$

für $t \ge 0$ und sie besitzt den Erwartungswert

$$\mathcal{E}W = \int_0^\infty t f(t)dt = \int_0^\infty (1-F(t))dt = \frac{1}{\lambda}\log\left(\frac{\mu}{\mu-\lambda}\right)$$

Fall $\mu = \lambda$
Hier lautet die Riccatische Differentialgleichung

$$z' = -\lambda(z-1)^2$$

mit der Lösung

$$z = \frac{1}{C + \lambda t} + 1$$

woraus sich auf dem gleichen Weg wie oben

$$f(z,t) = \frac{z - \lambda t(z-1)}{1 - \lambda t(z-1)}$$

mit

$$p_0(t) = \frac{\lambda t}{1 + \lambda t} \longrightarrow 1$$

und

$$\mathcal{E} X_t = 1$$

ergibt. Die Dichte der Verteilung der Zufallsvariable W ist

$$p_0'(t) = \frac{\lambda}{(1 + \lambda t)^2}$$

mit dem Erwartungswert

$$\mathcal{E} W = \int_0^\infty \frac{\lambda t}{(1 + \lambda t)^2} dt = \infty$$

Kapitel 4

Anhang

Literaturverzeichnis

[1] E. Aarts, *Simulated Annealing and Boltzmann Machines*, Wiley, Chichester, 1989

[2] K.B. Athreya, P.E. Ney, *Branching Processes*, Springer, Berlin, 1972

[3] A. T. Bharucha-Reid. *Elements of the Theory of Markov Processes and their Applications*. McGraw-Hill, New York, 1960

[4] Kai Lai Chung. *Elementare Wahrscheinlichkeitstheorie und stochastische Prozesse*. Springer, Berlin, 1978

[5] Kai Lai Chung. *Markov Chains With Stationary Transition Probabilities*. Springer, Berlin, 1967

[6] Arthur Engel. *Wahrscheinlichkeitsrechnung und Statistik*. Band 2, Klett, Stuttgart, 1980

[7] William Feller. *An Introduction to Probability Theory and Its Applications*. Vol. 1, Wiley, New York, 1968

[8] Ronald A. Howard. *Dynamic Probabilistic Systems. Volume 1: Markov Models*. Wiley, New York, 1971

[9] David G. Kendall, *Branching Processes since 1873*, J. London Math. Soc. 41 (1966), 385–406

[10] P.J.M. Laarhoven, E.H.L. Aarts, *Simulated Annealing: Theory and Applications*, Reidel, Dordrecht, 1988

[11] Hans Lahres. *Einführung in die diskreten Markoff-Prozesse und ihre Anwendungen*. Vieweg, Braunschweig, 1964

[12] Wolfgang Luh. *Mathematik für Naturwissenschaftler I.* Akademische Verlagsgesellschaft Wiesbaden, 1978

[13] Rudolf Mathar, Dietmar Pfeifer. *Stochastik für Informatiker*. Teubner, Stuttgart, 1990

[14] B.A. Sewastjanow, *Verzweigungsprozesse*, Oldenbourg, München, 1975