Bedienungstheorie

Vorlesung WS 2007/08

Kapitel 1

Einleitung

Unter Bedienungstheorie versteht man die "Mathematik der Warteschlangen", auf Englisch *queueing theory*. Ein Bedienungssystem besteht immer aus ein oder mehreren Bedienungskanälen (Servern), die von den eintreffenden Kunden in Anspruch genommen werden. Aus den in der Regel zufälligen Ankunfts- und Bedienungszeiten ergeben sich zufällige Wartezeiten für die Kunden ebenso wie zufällige Stillstandszeiten der Server.

1.0.1 Beispiele

Kunden in einem Supermarkt
Check-In am Flughafen
Telefongespräche in einer Leitung
Computerjobs am Server
Nachrichten im WWW
Reparaturaufträge in einer Werkstatt

1.1 Problemstellung

Ein Problem der Bedienungstheorie läßt sich durch folgende Merkmale beschreiben:

- Kundenstrom
  - Zwischenankunftszeiten
  - (Un)geduldige Kunden
  - Kundentypen
  - Einzelkunden oder Blöcke
- Bedienungsanlage
  - Systemkapazität
  - Anzahl der Server
  - Anordnung der Server

  - Wartedisziplin: FCFS, LCFS, RSS, Prioritäten
  - Einzel- oder Parallelservice
  - Bedienungzeiten der Server

Einfache Bedienungssysteme werden durch eine Kurzschreibweise $A/B/X/Y/Z$ charakterisiert. A beschreibt die Verteilung der unabhängigen Zwischenankunftszeiten und B die Verteilung der unabhängigen Bedienungszeiten. Dabei bedeutet

- M Exponentialverteilung (Markov Eigenschaft),
- D deterministisch,
- E_k Erlangverteilung $E(k, .)$,
- H_k Mischung von Exponentialverteilungen,
- G beliebige Verteilung.

$X \in \mathbb{N} \cup \{\infty\}$ beschreibt die Anzahl der parallelen Server und $Y \in \mathbb{N} \cup \{\infty\}$ die Systemkapazität. Z steht für die Wartedisziplin, wobei FCFS (oder FIFO) für *first come first service*, LCFS (oder LIFO) für *last come first service*, RSS für *random selection for service* steht. Die Standardannahmen $Y = \infty$ und $Z =$ FCFS werden nicht angeschrieben.

1.1.1 Beispiel

$G/D/3/5$ bedeutet: Alle Kunden treffen einzeln ein, die Zwischenankunftszeiten sind unabhängig und identisch verteilt. Ihre Verteilung kann beliebig sein. Die Bedienungszeiten sind alle gleich und konstant. Es gibt 3 gleichartige parallele Server. Maximal haben 5 Kunden im System Platz, dh ein Kunde, der bereits 5 Kunden im System vorfindet, geht verloren. Die Kunden werden in der Reihenfolge ihres Eintreffens bedient.

Man sieht, dass durch Verwendung dieser Notation implizit bereits viele Modellannahmen getroffen werden. So können etwa unterschiedliche Kundentypen, ungeduldige Kunden, Abhängigkeiten zwischen verschiedenen Ankunfts- und Bedienungszeiten, allgemeinere Anordnungen von Servern oder gleichzeitige Bedienung von Kunden durch einen Server nicht beschrieben werden. Auch zeitabhängige Ankunfts- und Bedienungszeiten werden nicht erfasst. Interessierende Größen sind je nach konkreter Problemstellung etwa:

- Länge der Warteschlange bzw. Anzahl der Kunden im System,
- Wartezeit eines (typischen) Kunden bzw. seine Systemzeit (Verweilzeit im System),
- Wahrscheinlichkeit für ein leeres System.

Die ersten beiden Punkte beschreiben Zufallsvariablen, hier ist man an der Verteilung interessiert bzw. zumindest an ihrem Erwartungswert. Alle Größen sind im allgemeinen von der Zeit und vom Anfangszustand abhängig. Betrachtet man jedoch das System im Gleichgewicht (nach unendlich langer Zeit), so entstehen zeitunabhängige Größen.

Aufgabe der Bedienungstheorie ist es auch, bei gegebenen Rahmenbedingungen (je nach Problemstellung zB Kundenstrom, Bedienungszeit oder Systemkapazität etc.) eine Bedienungsanlage optimal zu designen. Dabei sind Wartekosten, Stillstandskosten, Kosten für verlorene Kunden, etc. zu berücksichtigen. In einfachen Fällen werden analytische Lösungen möglich sein, oft jedoch wird man auf Abschätzungen, Näherungen oder Simulation angewiesen sein.

1.2 Allgemeine Resultate

In diesem Abschnitt fassen wir einige elementare Resultate zusammen, wie sie für ein $G/G/c$ System gültig sind.

1.2.1 Darstellung eines Warteschlangenproblems

Zeitorientierte Darstellung

Man diskretisiert die Zeit und notiert zu Beginn jedes Zeitintervalls den Zustand des Systems. In unserem Fall wird der Zustand des Systems allein durch die Anzahl der Kunden im System beschrieben. Die zeitorientierte Darstellung eignet sich vor allem dann, wenn nur zu fix vorgegebenen Zeitpunkten Änderungen im Systemzustand eintreten können (Zeittakt).

Ereignisorientierte Darstellung

Man notiert diejenigen Zeitpunkte, in denen ein Ereignis (Änderung im Systemzustand) eintritt, sowie die Art des Ereignisses, etwa Ankunft oder Abgang eines Kunden. Der Systemzustand zum Zeitpunkt t ergibt sich dann aus dem Systemzustand zum Zeitpunkt 0 und den bis t eingetretenen Änderungen. Die ereignisorientierte Darstellung ist im Fall von zufälligen Systemen meist vorzuziehen.

1.2.2 Bezeichnungen

Man bezeichnet mit λ die mittlere Anzahl der je Zeiteinheit eintreffenden Kunden und mit μ die mittlere Bedienungsrate, das ist die mittlere Anzahl der je Zeiteinheit von einem Bedienungskanal fertig bedienten Kunden, wenn dieser immer arbeitet. Die Größe $\rho := \frac{\lambda}{c\mu}$ wird als *Auslastungsgrad*[1] bezeichnet und ist ein Maß für die Verkehrsbelastung des Systems. Falls $\rho > 1$ gilt, treffen im Mittel mehr Kunden ein als das System verlassen. Da Kunden nicht abgewiesen werden, entstehen immer längere Warteschlangen. Ein Gleichgewicht kann in diesem Fall nicht entstehen. Dieses ist nur im Fall $\rho < 1$ möglich. Dann wird nach langer Zeit der Zustand des Systems nicht mehr vom absoluten Wert des Zeitparameters abhängen. Im Fall $\rho = 1$ ist ein Gleichgewicht nur bei deterministischen Zwischenankunfts- und Bedienungszeiten möglich, denn Bedienungsrückstände, wie sie etwa durch ein zufälliges Leerstehen entstehen, können nie aufgeholt werden.

[1] Den Grund dafür werden wir uns später überlegen.

Beispiel

Gegeben ist eine mittlere Ankunftsrate λ und eine mittlere Bedienungsrate μ. Wie groß muss c mindestens gewählt werden, damit ein Gleichgewicht möglich ist?
Antwort: $c = \left[\frac{\lambda}{\mu}\right] + 1$

Wir bezeichnen mit

$N(t)$	Anzahl der Kunden im System zur Zeit t,
$N_q(t)$	Anzahl der Kunden in der Warteschlange zur Zeit t ,
$N_s(t)$	Anzahl der Kunden in den Servern zur Zeit t.

Es gilt

$$N(t) = N_q(t) + N_s(t).$$

Ein Index q bezieht sich stets auf die Warteschlange (queue), ein Index s auf die Server, eine Größe ohne Index bezieht sich stets auf das gesamte System.
Bezeichne weiters N_x die Gleichgewichtsanzahl von Kunden in x und T_x die Zeit, die ein Kunde in x verbringt, wenn sich das System im Gleichgewicht befindet. Interessant sind nun $L_x = \mathbb{E}[N_x]$, die mittleren Anzahlen von Kunden und $W_x = \mathbb{E}[T_x]$ die mittleren Aufenthaltszeiten der Kunden. Es gilt

$$W = W_q + W_s \text{ und } W_s = \frac{1}{\mu}.$$

1.2.3 Die Formeln von Little

Beobachten wir ein System im Gleichgewicht über eine gewisse Zeitspanne τ und addieren die gesamte Zeit, die von Kunden im System verbracht wird, so erhalten wir

$$\int_0^\tau N(t)\ dt.$$

Es werden also im Mittel $L\ \tau = \mathbb{E}[N]\ \tau$ Zeiteinheiten im System verbracht. Notieren wir nun bei jedem im Zeitintervall $[0, \tau[$ eintretenden Kunden k ($k = 1, \ldots, K(\tau)$, wobei $K(\tau)$ der letzte Kunde ist, der vor τ eintrifft) sofort seine Systemzeit $T^{(k)}$, so ergibt sich für das Zeitintervall $[0, \tau[$ eine Gesamtsystemzeit von

$$\sum_{k=1}^{K(\tau)} T^{(k)},$$

im Erwartungswert $\lambda\ \tau\ W$ Zeiteinheiten. Bezeichnet nun $\tilde{T}_0$ die gesamte Systemzeit der Kunden, die zum Zeitpunkt 0 im System sind und $\tilde{T}_\tau$ die Systemzeit der Kunden, die zum Zeitpunkt τ im System sind, nach τ, so gilt

$$L\ \tau + \mathbb{E}\left[\tilde{T}_\tau\right] = \lambda\ \tau\ W + \mathbb{E}\left[\tilde{T}_0\right].$$

Ist nun τ genügend groß, so können nach Division durch τ die Randeffekte (Kunden, die sich zum Zeitpunkt 0 bereits im System befinden, und Systemzeiten, die erst nach dem Zeitpunkt τ beendet werden) vernachlässigt werden.

Es gilt daher die Formel von Little[2]

$$L = \lambda \, W. \tag{1.1}$$

Wendet man dieselben Überlegungen für die Warteschlange bzw. die Server (als Teilsysteme) an, so ergibt sich analog

$$L_q = \lambda \, W_q \text{ und } L_s = \lambda \, W_s$$

L_q ist die erwartete Länge der Warteschlange und L_s die mittlere Anzahl der Kunden in den Servern und entspricht der Leistung (offered work load rate) des Bedienungssystems. Für diese gilt nun

$$L_s = \frac{\lambda}{\mu} = c \, \rho.$$

Die Größe $\rho = \frac{L_s}{c}$ entspricht der erwarteten Anzahl von Kunden in einem vorgegebenen Server, also der Wahrscheinlichkeit, dass dieser Server arbeitet. Daher kommt die Bezeichnung Auslastungsgrad für ρ.

[2] John D. C. Little 1961

Kapitel 2

Einfache Markov-Modelle

Definition 2.1. *Ein stochastischer Prozess ist eine Familie von Zufallsvariablen*[1] $(Z(t))_{t \in I}$ *mit einer gemeinsamen Bildmenge* $\mathcal{Z}$*, die als Zustandsraum bezeichnet wird. I wird als Indexmenge bezeichnet. Ist I diskret, so spricht man von einem stochastischen Prozess mit diskreter Zeit, ist* $I \subseteq \mathbb{R}$*, so sagt man, der stochastische Prozess weist kontinuierliche Zeit auf.*

2.1 Poisson-Prozess

Der ankommende Kundenstrom, wobei $A(t)$ die Anzahl der bis zum Zeitpunkt t angekommenen Kunden bezeichnet, kann häufig mit Hilfe eines Poisson[2]-Prozesses modelliert werden.

Definition 2.2. *Ein Poisson-Prozess ist ein stochastischer Prozess* $(A(t))_{t \geq 0}$ *mit Zustandsraum* $\mathbb{N}_0$*,* $A(0) = 0$ *und den Eigenschaften:*

1. *Es gilt* $\Pr[A(t + \Delta t) - A(t) = 1] = \lambda \ \Delta t + o(\Delta t)$*, wobei* λ *eine Konstante ist, die von* $A(t)$ *unabhängig ist, und als Rate des Poisson-Prozesses bezeichnet wird.*
2. *Es gilt* $\Pr[A(t + \Delta t) - A(t) = 0] = 1 - \lambda \ \Delta t + o(\Delta t)$*.*
3. *Für* $a \leq b \leq c \leq d$ *gilt,* $A(b) - A(a)$ *und* $A(d) - A(c)$ *sind unabhängig.*

Ein Poisson-Prozess ist ein Zählprozess. Jede Realisierung beginnt in 0, springt irgendwann nach 1, dann nach 2 usw., und jeder Sprung hat mit Sicherheit die Höhe 1. Man kann den Prozess auch beschreiben, indem man die aufeinanderfolgenden Zeitabstände zwischen zwei Sprüngen, die Zwischenankunftszeiten, angibt.

[1] Eine Zufallsvariable Z ist eine Abbildung von einem Wahrscheinlichkeitsraum Ω in eine Bildmenge $\mathcal{Z}$. Durch Zufallsvariablen werden Beobachtungen von Zufallsexperimenten modelliert.

[2] Siméon Denis Poisson 1781 - 1840, französischer Mathematiker

2.1.1 Verteilung der Zuwächse

Für $t \geq s$ nennt man $A(t) - A(s)$ den Zuwachs im Intervall $]s,t]$. Für $n \in \mathbb{Z}$ bezeichnen wir mit $p_n^{(s)}(t) = \Pr[A(t) - A(s) = n]$, wobei aus den Eigenschaften des Poisson-Prozesses für $n < 0$ folgt $p_n^{(s)}(t) = 0$. Damit gilt nun

$$\begin{aligned} p_n^{(s)}(t+\Delta t) &= \Pr[A(t) - A(s) = n, A(t+\Delta t) - A(s) = n] \\ &\quad + \Pr[A(t) - A(s) = n-1, A(t+\Delta t) - A(s) = n] \\ &\quad + \Pr[A(t) - A(s) = n-2, A(t+\Delta t) - A(s) = n] + \ldots \\ &= \Pr[A(t) - A(s) = n, A(t+\Delta t) - A(t) = 0] \\ &\quad + \Pr[A(t) - A(s) = n-1, A(t+\Delta t) - A(t) = 1] \\ &\quad + \Pr[A(t) - A(s) = n-2, A(t+\Delta t) - A(t) = 2] + \ldots. \end{aligned}$$

Wegen 3. gilt

$$\begin{aligned} p_n^{(s)}(t+\Delta t) &= \Pr[A(t) - A(s) = n] \Pr[A(t+\Delta t) - A(t) = 0] \\ &\quad + \Pr[A(t) - A(s) = n-1] \Pr[A(t+\Delta t) - A(t) = 1] \\ &\quad + \Pr[A(t) - A(s) = n-2] \Pr[A(t+\Delta t) - A(t) = 2] + \ldots, \end{aligned}$$

was wegen 1. und 2.

$$\begin{aligned} p_n^{(s)}(t+\Delta t) &= p_n^{(s)}(t)(1 - \lambda\,\Delta t + o(\Delta t)) \\ &\quad + p_{n-1}^{(s)}(t)(\lambda\,\Delta t + o(\Delta t)) \\ &\quad + p_{n-2}^{(s)}(t)(o(\Delta t)) + o(\Delta t) \end{aligned}$$

ergibt: Division durch Δt und Grenzübergang $\Delta t \to 0$ führt auf die Differentialgleichungen für $t > s$

$$\begin{aligned} \frac{dp_0^{(s)}(t)}{dt} &= -\lambda\, p_0^{(s)}(t), \\ \frac{dp_n^{(s)}(t)}{dt} &= -\lambda\, p_n^{(s)}(t) + \lambda\, p_{n-1}^{(s)}(t) \quad \text{für } n \in \mathbb{N}. \end{aligned}$$

Aus der Definition von $p_n^{(s)}(t)$ folgen die Anfangsbedingungen $p_0^{(s)}(s) = 1$ und $p_n^{(s)}(s) = 0$. Die eindeutige Lösung des Anfangswertproblems mit unendlich (!) vielen Differentialgleichungen ist

$$p_n^{(s)}(t) = \frac{(\lambda(t-s))^n}{n!}\, e^{-\lambda(t-s)},$$

was durch Induktion bewiesen werden kann. Die Zuwächse $A(t) - A(s)$ sind also Poisson-verteilt mit Parameter $\lambda(t-s)$ und, weil sie nur von der Differenz der betrachteten Zeitpunkte abhängen, stationär. Mit $s = 0$ und $A(0) = 0$ ergibt sich, dass $A(t)$ Poisson-verteilt ist mit Parameter λt .

2.1.2 Verteilung der Abstände zwischen aufeinanderfolgenden Sprüngen

Bezeichne $T^{(1)}$ den Zeitpunkt des ersten Sprunges, $T^{(1)}+T^{(2)}$ den Zeitpunkt des 2. Sprunges, usw. Wir berechnen nun die Verteilung von $T^{(1)}$.

$$\begin{aligned}\Pr\left[T^{(1)} \geq t\right] &= \Pr\left[A\left(t\right)=0\right] \\ &= e^{-\lambda t},\end{aligned}$$

dh $T^{(1)}$ ist exponentialverteilt mit Parameter λ. Zu Berechnung der Verteilung von $T^{(k+1)}$ beachten wir, dass der k-te Sprung mit Sicherheit irgendwann stattfindet und damit $\left(dE\left(\tau\right)\right)_{\tau\in\mathbb{R}^+}$ mit $\mathrm{d}E\left(\tau\right) := \left(A\left(\tau\right)=k, A\left(\tau\right)-A\left(\tau-d\tau\right)=1\right)$ ein vollständiges Ereignissystem ist. Wegen der Unabhängigkeit der Zuwächse gilt

$$\begin{aligned}&\Pr\left[A\left(t+\tau\right)-A\left(\tau\right)=0 \mid dE\left(\tau\right)\right] \\ = &\frac{\Pr\left[A\left(t+\tau\right)-A\left(\tau\right)=0, A\left(\tau\right)=k, A\left(\tau\right)-A\left(\tau-d\tau\right)=1\right]}{\Pr\left[A\left(\tau\right)=k, A\left(\tau\right)-A\left(\tau-d\tau\right)=1\right]} \\ = &\Pr\left[A\left(t+\tau\right)-A\left(\tau\right)=0\right]\end{aligned}$$

Wir erhalten daher mit Hilfe der Eigenschaften des Poisson-Prozesses und dem Satz von der totalen Wahrscheinlichkeit

$$\begin{aligned}\Pr\left[T^{(k+1)} \geq t\right] &= \int_0^\infty \Pr\left[T^{(k+1)} \geq t \mid dE\left(\tau\right)\right] \Pr\left[dE\left(\tau\right)\right] \\ &= \int_0^\infty \Pr\left[A\left(t+\tau\right)-A\left(\tau\right)=0 \mid dE\left(\tau\right)\right] \Pr\left[dE\left(\tau\right)\right] \\ &= \int_0^\infty \Pr\left[A\left(t+\tau\right)-A\left(\tau\right)=0\right] \Pr\left[dE\left(\tau\right)\right] \\ &= \int_0^\infty e^{-\lambda t}\ \Pr\left[dE\left(\tau\right)\right] \\ &= e^{-\lambda t}\int_0^\infty \Pr\left[dE\left(\tau\right)\right] \\ &= e^{-\lambda t},\end{aligned}$$

dh auch alle $T^{(k+1)}$ sind exponentialverteilt mit Parameter λ. Weiters sind die $\left(T^{(k)}\right)_{k\in\ \mathbb{N}}$ vollständig unabhängig (Übung).
Es gilt auch die Umkehrung: Sind die Zeitabstände zwischen aufeinanderfolgenden Ereignissen $\left(T^{(k)}\right)_{k\in\mathbb{N}}$ vollständig unabhängig und exponentialverteilt mit demselben Parameter λ, und bezeichnet $A\left(t\right)$ die Anzahl der Ereignisse vor dem Zeitpunkt t, so ist $\left(A\left(t\right)\right)_{t\geq 0}$ ein Poisson-Prozess mit Parameter λ.
Für einen Poisson-Prozess $\left(A\left(t\right)\right)_{t\geq 0}$ gilt außerdem: Wenn man weiß, dass in einem Intervall n Ereignisse eintreten, so sind die Zeitpunkte des Eintretens vollständig unabhängig und gleichverteilt (Übung).

2.1.3 Markov-Eigenschaft der Exponentialverteilung

Ist T exponentialverteilt, dh $\Pr[T < t] = 1 - e^{-\lambda t}$, so gilt für $s, t \geq 0$ die sogenannte Nichtalterungseigenschaft (Markov[3]-Eigenschaft)

$$\Pr[T \geq s + t \mid T \geq s] = \Pr[T \geq t].$$

Die Exponentialverteilung ist die einzige Verteilung, die die Nichtalterungseigenschaft aufweist.

2.1.4 Verallgemeinerungen

Einige andere Ankunftsprozesse können durch Verallgemeinerungen des Poisson-Prozesses beschrieben werden: Bei einem inhomogenen Poisson-Prozess wird $\lambda = \lambda(t)$ zeitabhängig gewählt. Man kann also zeitabhängige zufällige Kundenströme behandeln. Beim Poisson-Prozess mit Mehrfachpunkten gilt für $n \in \mathbb{N}$

$$\Pr[A(t + \Delta t) - A(t) = n] = \lambda_n \, \Delta t + o(\Delta t),$$

es können also mehrere Kunden gleichzeitig eintreffen. Ein Erneuerungsprozess entsteht, wenn die Zwischenankunftszeiten unabhängige und identisch verteilte Zufallsvariablen sind, aber nicht unbedingt exponentialverteilt.

2.2 Markov-Ketten

Definition 2.3. *Ein stochastischer Prozess $(Z(t))_{t \in I}$ heißt Markov-Prozess, wenn für alle $n \in \mathbb{N}$, alle $0 \leq t_1 < \ldots < t_{n+1}$ ($t_i \in I$) und für alle $x_1, \ldots, x_{n+1} \in \mathcal{Z}$ mit $\Pr[Z_{t_n} = x_n, \ldots, Z_{t_1} = x_1] > 0$ gilt*[4]

$$\Pr[\underbrace{Z_{t_{n+1}} = x_{n+1}}_{Zukunft} | \underbrace{Z_{t_n} = x_n}_{Gegenwart}, \underbrace{\ldots, Z_{t_1} = x_1}_{Vergangenheit}] = \Pr[\underbrace{Z_{t_{n+1}} = x_{n+1}}_{Zukunft} | \underbrace{Z_{t_n} = x_n}_{Gegenwart}].$$

Dh. bildlich gesprochen, die Zukunft hängt nur von der Gegenwart, nicht aber von der Vergangenheit ab. Obige Eigenschaft nennt man die *Markov-Eigenschaft* (Gedächtnislosigkeit) des stochastischen Prozesses. Ein Markov-Prozess, bei dem entweder der Zustandsraum oder die Indexmenge diskret sind, heißt **Markov-Kette**. Hat eine Markov-Kette einen endlichen Zustandsraum, so heißt sie **endlich**, sonst **abzählbar**.
Ein **Semi-Markov-Prozess** (Markovscher Erneuerungsprozess) ist ein stochastischer Prozess mit diskretem Zustandsraum, bei dem die Folge der verschiedenen Zustände eine Markovkette mit diskreter Zeit bildet, die Verteilung der Zeitabstände zwischen aufeinanderfolgenden Zustandsänderungen aber beliebig ist.

[3] Andrei Andreyevich Markov, 1856 - 1922, russischer Mathematiker

[4] Ist der Zustandsraum des stochastischen Prozesses nicht diskret, so muss das Ereignis $\{Z_{t_n} = x_n, \ldots, Z_{t_1} = x_1\}$ differenziell interpretiert werden.

2.2.1 Markov-Ketten mit diskreter Zeit

Definition 2.4. *Eine Folge von Zufallsvariablen* $(Z_n)_{n\in\mathbb{N}_0}$ *auf einem abzählbaren Zustandsraum, die die Markov-Eigenschaft erfüllt, heißt eine Markov-Kette mit diskreter Zeit. Der Ausdruck*

$$\Pr[Z_{n+1} = j \mid Z_n = i] =: p_{ij}$$

heißt Übergangswahrscheinlichkeit *von i nach j. (Wenn diese Wahrscheinlichkeit nicht von n abhängt, so nennt man die Markovkette* homogen.*) Die unendliche Matrix der Übergangswahrscheinlichkeiten $(p_{ij}) =: \mathbf{P}$ heißt* Übergangsmatrix. *Der Ausdruck*

$$\Pr[Z_n = j] =: \pi_j^{(n)}$$

bezeichnet die Zustandswahrscheinlichkeit von j zum Zeitpunkt n.
Der Zeilenvektor $\left(\pi_j^{(n)}\right) =: \pi^{(n)}$ *heißt* Zustandsverteilung *zum Zeitpunkt n, $\pi^{(0)}$ heißt* Anfangsverteilung.

Für Markov-Ketten gelten die *Chapman*[5]*-Kolmogorov*[6]*-Gleichungen*

$$\mathbf{P}^m = \mathbf{P}^{m-k}\mathbf{P}^k,$$

wobei $\mathbf{P}^m$ die Matrix der m-Schritt-Übergangswahrscheinlichkeiten mit den Einträgen $p_{ij}^{(m)}$ ist. Für die Zustandsverteilung folgt daraus

$$\pi^{(m)} = \pi^{(0)}\mathbf{P}^m.$$

Bezeichnet man $\mathbf{Q} := \mathbf{P} - \mathbf{I}$, so gilt

$$\pi^{(m)} - \pi^{(m-1)} = \pi^{(m-1)}\mathbf{Q}.$$

2.2.2 Markov-Ketten mit kontinuierlicher Zeit

Wir betrachten eine Markov-Kette mit kontinuierlicher Zeit $(X_t)_{t\geq 0}$ und bezeichnen für $s \leq t$ und Zustände i, j den Ausdruck

$$\Pr\left[X_t = j \mid X_s = i\right] =: p_{ij}\left(s,t\right)$$

als Übergangswahrscheinlichkeit von i nach j im Zeitintervall $[s,t]$. Wir schreiben die Übergangswahrscheinlichkeiten wieder in Übergangsmatrizen $\mathbf{P}\left(s,t\right) := \left(p_{ij}\left(s,t\right)\right)$ zusammen. Die Chapman-Kolmogorov-Gleichungen lauten nun $(r \leq s \leq t)$

$$\mathbf{P}\left(r,t\right) = \mathbf{P}\left(r,s\right)\mathbf{P}\left(s,t\right).$$

Mit $p_j\left(t\right) := \Pr\left[X_t = j\right]$ werden die Zustandwahrscheinlichkeiten bezeichnet, die ebenso in Zeilenvektoren als Zustandsverteilung $\mathbf{p}\left(t\right) := \left(p_j\left(t\right)\right)$ zusammengefasst werden, $\mathbf{p}\left(0\right)$ ist die Anfangsverteilung.

[5] Sydney Chapman, 1888 - 1970, britischer Mathematiker
[6] Andrey Nikolaevich Kolmogorov, 1903 - 1987, russischer Mathematiker

Wir wollen nun die *Differentialgleichungen von Kolmogorov* einführen. Sei $\mathbf{P}$ so beschaffen, dass es eine Matrix $\mathbf{Q} := (q_{ij})$ mit[7]

$$\mathbf{Q} = \lim_{\Delta t \to 0} \frac{\mathbf{P}(t, t+\Delta t) - \mathbf{I}}{\Delta t}$$

gibt, deren Einträge außerhalb der Diagonale positiv sind und deren Zeilensummen 0 sind. Dann gilt

$$p_{ij}(t, t+\Delta t) = \begin{cases} q_{ij}\Delta t + o(\Delta t) & \text{falls} \quad i \neq j \\ (1+q_{ii})\Delta t + o(\Delta t) & \text{falls} \quad i = j \end{cases}.$$

Aus den Chapman-Kolmogorov-Gleichungen folgt

$$\mathbf{P}(r, t+\Delta t) - \mathbf{P}(r,t) = \mathbf{P}(r,t)\left(\mathbf{P}(t, t+\Delta t) - \mathbf{I}\right)$$

und damit die Vorwärtsdifferentialgleichung von Kolmogorov

$$\frac{\partial \mathbf{P}(r,t)}{\partial t} = \mathbf{P}(r,t)\,\mathbf{Q}$$

und aus

$$\mathbf{P}(r+\Delta r, t) - \mathbf{P}(r,t) = \left(\mathbf{I} - \mathbf{P}(r, r+\Delta r)\right)\mathbf{P}(r+\Delta r, t)$$

die Rückwärtsdifferentialgleichung von Kolmogorov

$$\frac{\partial \mathbf{P}(r,t)}{\partial r} = -\mathbf{Q}\mathbf{P}(r,t).$$

Für die Zustandsverteilung folgt aus der Vorwärtsgleichung mit $r = 0$ durch Multiplikation mit $\mathbf{p}(0)$ von links

$$\mathbf{p}'(t) = \mathbf{p}(t)\,\mathbf{Q}.$$

Man nennt $\mathbf{Q}$ den *infinitesimalen Generator* oder die *Intensitätsmatrix* der Markov-Kette mit kontinuierlicher Zeit.

Für den Poisson-Prozess mit Parameter λ gilt

$$\mathbf{Q} = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \dots \\ 0 & -\lambda & \lambda & 0 & \dots \\ 0 & 0 & -\lambda & \lambda & \dots \\ 0 & 0 & 0 & -\lambda & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix}.$$

Hat eine Markov-Kette (X_t) den Zustandsraum $\mathbb{N}_0$ und sind in einem infinitesimalen Zeitintervall nur Sprünge der Höhe ± 1 möglich, so spricht man von einem Geburts- und Todesprozess. $\mathbf{Q}(t)$ hat Tridiagonalgestalt. Gilt außerdem

$$\begin{aligned} \Pr[X_{t+\Delta t} = n+1 \mid X_t = n] &= \lambda_n \Delta t + o(\Delta t) \\ \Pr[X_{t+\Delta t} = n-1 \mid X_t = n] &= \mu_n \Delta t + o(\Delta t) \\ \Pr[X_{t+\Delta t} = n \mid X_t = n] &= 1 - (\lambda_n + \mu_n)\Delta t + o(\Delta t) \end{aligned}$$

[7] Wir betrachten nur den Fall, dass $\mathbf{Q}$ nicht von t abhängt. Der allgemeinere Fall benötigt einige (wenige) Zusatzvoraussetzungen. Für Zwecke der Bedienungstheorie genügt jedoch unser Modell.

so ist der Geburts- und Todesprozess homogen und es gilt

$$\mathbf{Q} = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \dots \\ \mu_1 & -\lambda_1 - \mu_1 & \lambda_1 & 0 & \dots \\ 0 & \mu_2 & -\lambda_2 - \mu_2 & \lambda_2 & \dots \\ 0 & 0 & \mu_3 & -\lambda_3 - \mu_3 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix}.$$

Die Größen λ_n heißen Geburtsraten und μ_n Sterberaten.

2.2.3 Graphische Veranschaulichung von Markov-Ketten

Markov-Ketten mit abzählbarem Zustandsraum werden sehr einfach und übersichtlich durch einen gerichteten gewichteten Graphen, den *Übergangsgraphen*, veranschaulicht. Die Knoten des Graphen entsprechen den Zuständen, die gerichteten Kanten den Übergangswahrscheinlichkeiten bzw Übergangsraten. Zwischen zwei Knoten ist genau dann keine Kante, wenn kein Übergang zwischen ihnen möglich ist.

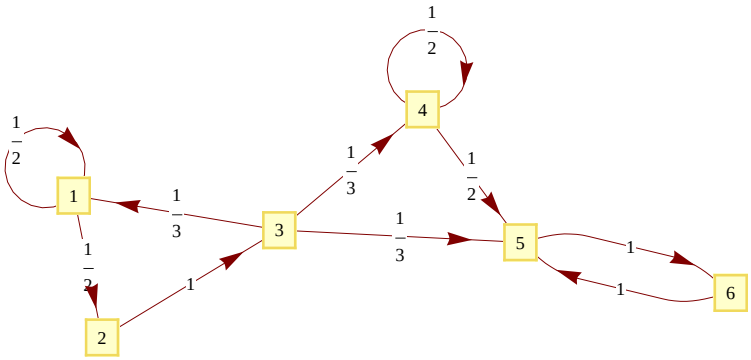

Abbildung 2.1: Ein Graph mit 6 Zuständen

2.2.4 Die eingebettete Markov-Kette

Angenommen, die Modellbildung erfordert die Annahme einer kontinuierlichen Zeit, wir beobachten den Prozess aber nur zu ausgewählten Zeitpunkten, zB immer dann, wenn eine Zustandsänderung eintritt. In manchen Fällen erhalten wir bei geschickter Wahl der Beobachtungszeitpunkte eine Markov-Kette. Man nennt sie die eingebettete Markov-Kette. Handelt es sich zB ursprünglich um einen Markov-Prozess mit kontinuierlicher Zeit, und beobachtet man bei jeder Zustandsänderung, so entsteht immer eine Markovkette mit den Übergangswahrscheinlichkeiten

$$p_{ij} = \begin{cases} \frac{q_{ij}}{-q_{ii}} & \text{für} \quad i \neq j \\ 0 & \text{für} \quad i = j \end{cases}.$$

Die Verweilzeit in jedem Zustand i ist exponentialverteilt mit Parameter $q_i := -q_{ii}$, im Erwartungswert also $1/q_i$.

2.2.5 Langzeitverhalten

Drei Größen kennzeichnen das Langzeitverhalten einer Markov-Kette, ihre Grenzverteilung, ihre stationären Verteilungen und die Eigenschaft der Ergodizität.
Man sagt, eine Markov-Kette mit diskreter Zeit besitzt die *Grenzverteilung* π, wenn für alle i, j

$$\lim_{m\to\infty} p_{ij}^{(m)} = \pi_j$$

gilt bzw.

$$\lim_{m\to\infty} \pi^{(m)} = \pi,$$

dh. wenn nach langer Zeit jeder Zustand, unabhängig vom Anfangszustand, mit einer gewissen Wahrscheinlichkeit auftritt. Die Grenzverteilung erfüllt wegen

$$\pi\mathbf{P} = \lim_{m\to\infty} \pi^{(m)}\mathbf{P} = \lim_{m\to\infty} \pi^{(m+1)} = \pi$$

die Beziehung

$$\pi\mathbf{P} = \pi \quad \text{bzw} \quad \pi\mathbf{Q} = 0 \tag{2.1}$$

und, weil π eine Verteilung ist, auch

$$\pi\ \mathbf{1} = 1.$$

Jede Verteilung, die die Beziehung (2.1) erfüllt, heißt eine *stationäre Verteilung* oder *Gleichgewichtsverteilung* der Markov-Kette. Ist die Markov-Kette endlich, so ist π ein normierter Linkseigenvektor von $\mathbf{P}$ zum Eigenwert 1. Für endliche Markov-Ketten existiert also immer eine stationäre Verteilung.

Bei Markov-Ketten mit kontinuierlicher Zeit ist die Begriffsbildung analog. Man sagt, eine Markov-Kette mit kontinuierlicher Zeit besitzt die *Grenzverteilung* $\mathbf{p}$, wenn für alle i, j

$$\lim_{t\to\infty} p_{ij}(0,t) = p_j$$

gilt bzw $\lim_{t\to\infty} \mathbf{p}(t) = \mathbf{p}$, dh wenn nach langer Zeit jeder Zustand, unabhängig vom Anfangszustand, mit einer gewissen Wahrscheinlichkeit auftritt. Die Grenzverteilung erfüllt wegen

$$\mathbf{p}\mathbf{Q} = \lim_{t\to\infty} \mathbf{p}(t)\,\mathbf{Q} = \lim_{t\to\infty} \mathbf{p}'(t) = 0$$

die Beziehung

$$\mathbf{p}\mathbf{Q} = 0 \tag{2.2}$$

und, weil $\mathbf{p}$ eine Verteilung ist, auch

$$\mathbf{p}\ \mathbf{1} = 1.$$

Jede Verteilung $\mathbf{p}$, die die Beziehung $\mathbf{p}\mathbf{Q} = 0$ erfüllt, heißt eine *stationäre Verteilung* oder *Gleichgewichtsverteilung* der Markov-Kette. Ist die Markov-Kette endlich, so ist $\mathbf{p}$ ein normierter Linkseigenvektor von $\mathbf{Q}$ zum Eigenwert 0. Für endliche Markov-Ketten existiert also immer eine stationäre Verteilung.
Nicht jede stationäre Verteilung ist eine Grenzverteilung. Es gibt Markov-Ketten (auch solche mit kontinuierlicher Zeit), die keine Grenzverteilung aber stationäre Verteilungen besitzen.

Beispiel 2.1. *Man beschreibe die Markov-Ketten mit den Übergangsmatrizen*

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

verbal und mit Hilfe des Übergangsgraphen und bestimme jeweils die Grenzverteilung, falls sie existiert, und die stationären Verteilungen.

Eine Markov-Kette, die in ihrer stationären Verteilung startet, ist *stark stationär*, dh die gemeinsame Verteilung von $X(t_1)$, $X(t_2)$, $\ldots$, $X(t_n)$ ist für alle $t > 0$ dieselbe wie die von $X(t+t_1)$, $X(t+t_2)$, $\ldots$, $X(t+t_n)$. Die Größe π_i bzw p_i kann als der Anteil der Zeit gesehen werden, den die Markov-Kette im Zustand i verbringt.

Im folgenden werden wir Bedingungen angeben, unter denen stationäre Verteilungen oder die Grenzverteilung existieren.
Ein Zustand i einer Markov-Kette mit diskreter Zeit heißt *aperiodisch*, wenn

$$\text{ggT}\left\{n : p_{ii}^{(n)} > 0\right\} = 1$$

gilt.
Bezeichne $f_{ij}^{(n)}$ bzw $f_{ij}(t)$ die Wahrscheinlichkeit, dass die Markov-Kette, die in i startet nach n Schritten bzw zur Zeit t das erste Mal nach 0 in j eintrifft. Ein Zustand i einer Markov-Kette heißt *rekurrent*, wenn

$$\sum_{n=1}^{\infty} f_{ii}^{(n)} = 1 \text{ bzw } \int_0^{\infty} f_{ii}(t) \ dt = 1,$$

dh wenn die Markov-Kette mit Sicherheit nach i zurückkehrt. Ist $\sum_{n=1}^{\infty} f_{ii}^{(n)} < 1$ bzw $\int_0^{\infty} f_{ii}(t) \ dt < 1$, so heißt i *transient*. Ein rekurrenter Zustand i einer Markov-Kette heißt *positiv rekurrent* , wenn

$$\sum_{n=1}^{\infty} n f_{ii}^{(n)} =: m_{ii} < \infty \text{ bzw } \int_0^{\infty} t \ f_{ii}(t) \ dt =: m_{ii} < \infty,$$

dh wenn die mittlere Rückkehrzeit m_{ii} endlich ist. Ein Zustand heißt *null rekurrent*, wenn er rekurrent aber nicht positiv rekurrent ist.
Eine Markov-Kette heißt

aperiodisch		aperiodisch	
rekurrent		rekurrent	
transient	wenn alle Zustände	transient	sind.
positiv rekurrent		positiv rekurrent	
null rekurrent		null rekurrent	

Eine Markov-Kette heißt *irreduzibel*, wenn es für alle i, j ein n bzw t gibt, sodass $p_{ij}^{(n)} > 0$ bzw $p_{ij}(t) > 0$ gilt. Bei einer irreduziblen Markov-Kette weisen stets alle Zustände dieselben Eigenschaften auf.

Eine Markov-Kette heißt *ergodisch*[8], wenn sie irreduzibel, positiv rekurrent und im Fall diskreter Zeit auch aperiodisch ist. In diesem Fall kann aus einer einzigen genügend langen Realisierung des Prozesses schon die gesamte Information über den Prozess gewonnen werden. Man kann hier die Ensemble-Mittelwerte durch die entsprechenden Zeit-Mittelwerte ersetzen.

Satz 2.1. *Ist eine irreduzible Markov-Kette positiv rekurrent, so existiert genau eine stationäre Verteilung π mit $\pi_i = 1/m_{ii}$ bzw $\mathbf{p}$ mit $\mathbf{p}_i = 1/m_{ii}$.*

Satz 2.2. *Hat eine irreduzible Markov-Kette eine stationäre Verteilung π bzw $\mathbf{p}$, so ist sie positiv rekurrent mit $m_{ii} = 1/\pi_i$ bzw $m_{ii} = 1/\mathbf{p}_i$ und die stationäre Verteilung ist eindeutig.*

Endliche irreduzible Markov-Ketten sind also immer positiv rekurrent und haben genau eine stationäre Verteilung.

Satz 2.3. *(Konvergenzsatz) Für eine irreduzible, positiv rekurrente und aperiodische (also ergodische) Markov-Kette mit diskreter Zeit ist die eindeutige stationäre Verteilung auch Grenzverteilung.*

Satz 2.4. *(Konvergenzsatz) Für eine irreduzible und positiv rekurrente Markov-Kette mit kontinuierlicher Zeit ist die eindeutige stationäre Verteilung stets auch Grenzverteilung.*

Hat die eingebettete Markov-Kette die Grenzverteilung π, so gilt für die Grenzverteilung $\mathbf{p}$ der Markov-Kette mit kontinuierlicher Zeit

$$\mathbf{p}_i = \frac{\pi_i/q_i}{\sum_j \pi_j/q_j},$$

sofern die Reihe $\sum_j \pi_j/q_j$ konvergiert. Dafür reicht zB aus, dass die mittleren Verweilzeiten beschränkt sind.

2.3 Geburts- und Todesprozesse im Gleichgewicht

Wir betrachten einen Geburts- und Todesprozess mit infinitesimalem Generator

$$\mathbf{Q} = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \dots \\ \mu_1 & -\lambda_1 - \mu_1 & \lambda_1 & 0 & \dots \\ 0 & \mu_2 & -\lambda_2 - \mu_2 & \lambda_2 & \dots \\ 0 & 0 & \mu_3 & -\lambda_3 - \mu_3 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix},$$

wobei für alle n gilt $\lambda_n > 0$ und $\mu_n > 0$. Der Prozess ist irreduzibel. Wir berechnen die stationäre Verteilung. Dazu schreiben wir das unendlich große Gleichungssystem

$$\mathbf{pQ} = 0$$

[8] Der Begriff *ergodisch* wird bei Markov-Ketten etwas anders verwendet als bei stationären stochastischen Prozessen.

ausführlich auf und erhalten

$$\begin{aligned} -\lambda_0 p_0 + \mu_1 p_1 &= 0 \\ \lambda_0 p_0 - (\lambda_1 + \mu_1)\, p_1 + \mu_2 p_2 &= 0 \\ &\dots \\ \lambda_{n-2} p_{n-2} - (\lambda_{n-1} + \mu_{n-1})\, p_{n-1} + \mu_n p_n &= 0 \\ &\dots \ . \end{aligned}$$

Dieses Gleichungssystem kann sehr einfach und anschaulich direkt aus dem Übergangsgraphen abgelesen werden, indem man sich das *stochastische Gleichgewicht* überlegt. Die obigen Gleichungen können als Gleichgewicht zwischen dem Fluss in einen Zustand und dem Fluss aus einem Zustand interpretiert werden (*global balance equations*).

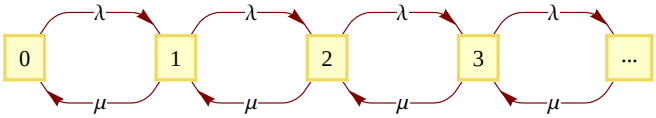

Abbildung 2.2: Geburts- und Todesprozess

Es führt uns durch schrittweises Einsetzen auf die Formel

$$p_n = p_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}, \tag{2.3}$$

die durch vollständige Induktion bewiesen wird (siehe Vorlesung). Soll $\mathbf{p}$ eine Verteilung werden, so muss

$$\sum_{n=0}^{\infty} p_n = p_0 \left(1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right) = 1$$

gelten, was die Konvergenz der Reihe $\sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}$ erfordert. In diesem Fall gilt

$$p_0 = \left(1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right)^{-1} > 0$$

und eine stationäre Verteilung für den Geburts- und Todesprozess existiert und ist eindeutig.

2.3.1 Berechnungsmethoden für die Gleichgewichtsverteilung am Beispiel des $M/M/1$ Modells

Hier sind die Zwischenankunftszeiten (Parameter λ) und die Bedienungszeiten (Parameter μ) exponentialverteilt und es gibt einen Server. Die mittlere Zeit zwischen zwei aufeinanderfolgenden Ankünften ist gleich $1/\lambda$ und die mittlere Bedienungszeit $1/\mu$. Es handelt

sich um einen Geburts- und Todesprozess mit $\lambda_i = \lambda$ und $\mu_i = \mu$ für alle i.
Die Gleichgewichtsverteilung wird nun durch Lösung von

$$\begin{aligned} -\lambda p_0 + \mu p_1 &= 0 \\ \lambda p_0 - (\lambda + \mu) p_1 + \mu p_2 &= 0 \\ &\dots \\ \lambda p_{n-2} - (\lambda + \mu) p_{n-1} + \mu p_n &= 0 \\ &\dots \end{aligned}$$

bestimmt.

Iterative Methode

Diese haben wir bereits beim allgemeinen Geburts- und Todesprozess kennengelernt. Wir erhalten für alle $n \in \mathbb{N}_0$

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0.$$

Wie im allgemeinen Fall ergibt sich mit $\rho = \lambda/\mu$

$$\begin{aligned} p_0 &= \left(1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda}{\mu}\right)^{-1} \\ &= \left(1 + \sum_{n=1}^{\infty} \rho^n\right)^{-1}. \end{aligned}$$

Im Fall $\rho < 1$ gibt es genau eine Gleichgewichtsverteilung, nämlich

$$p_n = \rho^n (1 - \rho).$$

Die Anzahl der Kunden im System ist also geometrisch verteilt mit Parameter ρ.

Erzeugende Funktionen bzw. z-Transformation

Das Gleichungssystem, bei dem es sich ja um Differenzengleichungen handelt, kann auch durch erzeugende Funktionen gelöst werden. Dazu wird die (Wahrscheinlichkeits-) erzeugende Funktion $P(z) := \sum_{n=0}^{\infty} p_n z^n$ mit $z \in \mathbb{C}$ und $|z| \leq 1$ eingeführt. Für diese wird aus dem Gleichungssystem eine Formel gefunden, aus der dann durch Reihenentwicklung die Koeffizienten p_n gewonnen werden. Die Methode soll nun am $M/M/1$ Modell vorgeführt werden. Das Gleichungssystem umgeschrieben für ρ lautet

$$\begin{aligned} p_{n+1} &= (\rho + 1) p_n - \rho p_{n-1} \quad n \geq 1 \\ p_1 &= \rho p_0. \end{aligned}$$

Daraus folgt durch Multiplikation mit z^n und Summation über n

$$\begin{aligned} z^{-1} p_{n+1} z^{n+1} &= (\rho + 1) p_n z^n - z \rho p_{n-1} z^{n-1} \\ z^{-1} \sum_{n=1}^{\infty} p_{n+1} z^{n+1} &= (\rho + 1) \sum_{n=1}^{\infty} p_n z^n - z\rho \sum_{n=1}^{\infty} p_{n-1} z^{n-1} \\ z^{-1} (P(z) - p_1 z - p_0) &= (\rho + 1)(P(z) - p_0) - z\rho P(z) \end{aligned}$$

und daraus mit Hilfe der zweiten Gleichung $p_1 = \rho p_0$ und Auflösen nach $P(z)$

$$P(z) = \frac{p_0}{1 - z\rho}.$$

Da wegen der Definition von $P(z)$ die Beziehung $P(1) = 1$ gilt, ist $\rho \neq 1$ nötig, und es folgt $p_0 = 1 - \rho$ und damit $\rho < 1$ und wir erhalten

$$P(z) = \frac{1 - \rho}{1 - z\rho} \quad \rho < 1,\ |z| \leq 1.$$

Indem man nun $P(z)$ in eine Potenzreihe entwickelt, erhält man in diesem Fall sehr schnell $p_n = \rho^n (1 - \rho)$. Häufig muss man sich damit begnügen, interessante Größen aus der erzeugenden Funktion direkt herzuleiten, zB die erwartete Anzahl L von Kunden im System

$$L = \sum_{n=1}^{\infty} n p_n = \sum_{n=1}^{\infty} n p_n z^{n-1}\Big|_{z=1} = \frac{d}{dz} \sum_{n=0}^{\infty} p_n z^n\Big|_{z=1} = P'(1)$$

In unserem Fall ergibt sich $L = \frac{\rho}{1-\rho}$.

Operatoren

Die Gleichgewichtsverteilung kann auch mit Hilfe von Operatoren gefunden werden, die auf Folgen angewendet werden. Ist etwa $(a_0, a_1, a_2, \ldots)$ eine Folge, so wird durch $Da_n = a_{n+1}$ $(n \in \mathbb{N})$ der Shift-Operator definiert, der die Folge $(a_0, a_1, a_2, \ldots)$ auf die Folge $(a_1, a_2, a_3, \ldots)$ abbildet (einfache Shift). Man beachte, dass $D^2 = D \circ D$ eine Folge $(a_0, a_1, a_2, \ldots)$ auf die Folge $(a_2, a_3, \ldots)$ abbildet (doppelte Shift). Eine lineare Differenzengleichung hat nun die Form

$$c_n a_n + c_{n+1} a_{n+1} + \ldots + c_{n+k} a_{n+k} = 0,$$

was sich mittels Shift-Operator als

$$\left(\sum_{i=0}^{k} c_{n+i} D^i \right) a_n = 0$$

schreibt, weil $D^m a_n = a_{n+m}$ für alle n, m gilt. Ist nun r eine Wurzel (Nullstelle) des Operator - Polynoms $\sum_{i=0}^{k} c_{n+i} D^i$, dh

$$\sum_{i=0}^{k} c_{n+i} D^i = (D - r) \sum_{i=0}^{k-1} \tilde{c}_{n+i} D^i,$$

so gilt

$$\begin{aligned}\left(\sum_{i=0}^{k} c_{n+i}D^i\right) r^n &= (D-r)\left(\sum_{i=0}^{k-1} \tilde{c}_{n+i}D^i\right) r^n \\ &= (D-r)\sum_{i=0}^{k-1} \tilde{c}_{n+i}r^{n+i} \\ &= D\sum_{i=0}^{k-1} \tilde{c}_{n+i}r^{n+i} - r\sum_{i=0}^{k-1} \tilde{c}_{n+i}r^{n+i} \\ &= \sum_{i=0}^{k-1} \tilde{c}_{n+i}r^{n+i+1} - \sum_{i=0}^{k-1} \tilde{c}_{n+i}r^{n+i+1} \\ &= 0,\end{aligned}$$

dh $a_n = r^n$ ist eine Lösung der Differenzengleichung. Alle Lösungen der Differenzengleichung sind Linearkombinationen der Gestalt $a_n = \sum_i \alpha_i r_i^n$, wobei die r_i Wurzeln des Operator - Polynoms sind.
Wir wenden nun die Methode auf unsere Rekursionsgleichung

$$p_{n+2} - (\rho+1)\, p_{n+1} + \rho p_n = 0 \quad n \geq 0$$

mit den Nebenbedingungen $p_1 = \rho p_0$ und $\sum_{n=0}^{\infty} p_n = 1$ an.
Mittels Shift - Operator schreibt sich die Rekursionsgleichung als

$$\begin{aligned}0 &= \left(D^2 - (\rho+1)\, D + \rho\right) p_n \\ &= (D-1)\,(D-\rho)\, p_n,\end{aligned}$$

was auf $p_n = \alpha + \beta\rho^n$ führt. Aus der Nebenbedingung $\sum_{n=0}^{\infty} p_n = 1$ folgt $\alpha = 0$ (die Nebenbedingung $p_1 = \rho p_0$ ist damit automatisch erfüllt) und ebenso $\rho < 1$ und $p_n = \rho^n\,(1-\rho)$.
Sind die p_n berechnet, so können alle interessierenden Kenngrößen des Bedienungssystems im Gleichgewicht berechnet werden, zB die erwartete Anzahl von Kunden im System

$$\begin{aligned}L &= \mathbb{E}\,[N] = \sum_{n=1}^{\infty} n p_n = (1-\rho)\sum_{n=1}^{\infty} n\rho^n \\ &= (1-\rho)\,\rho\frac{d}{d\rho}\sum_{n=0}^{\infty}\rho^n \\ &= (1-\rho)\,\rho\frac{d}{d\rho}\frac{1}{1-\rho} \\ &= \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}.\end{aligned}$$

Ebenso ist auch die erwartete Anzahl von wartenden Kunden interessant,

$$L_q = \mathbb{E}\,[N_q] = \sum_{n=1}^{\infty}(n-1)\, p_n = \frac{\rho}{1-\rho} - (1-p_0) = \frac{\rho^2}{1-\rho}.$$

Interessant ist auch die erwartete Anzahl von Kunden in der Warteschlange, wenn diese nicht leer ist ($\tilde{L}_q = \frac{1}{1-\rho}$ Übung).

2.3.2 Wartezeiten beim $M/M/1$ Modell

Aus der Formel von Little können die erwartete Systemzeit W und die erwartete Wartezeit W_q in der Warteschlange gewonnen werden, nämlich

$$W = \frac{L}{\lambda} = \frac{1}{\mu - \lambda} \text{ und } W_q = \frac{L_q}{\lambda} = \frac{\lambda}{\mu\,(\mu - \lambda)}.$$

Ist das System im Gleichgewicht, so kann die Verteilung der (Zufallsvariable) Wartezeit T_q eines (virtuellen) Kunden, der zum Zeitpunkt 0 eintrifft, durch folgende Überlegungen gewonnen werden.
Wegen der Gedächtnislosigkeit der Exponentialverteilung gilt für $n \in \mathbb{N}_0$ bei einer FCFS Strategie

$$\Pr\left[T_q \le t \mid N = n\right] = 1 - \sum_{k=0}^{n-1} e^{-\mu t} \frac{(\mu t)^k}{k!}.$$

Daraus folgt mit dem Satz von der totalen Wahrscheinlichkeit

$$\begin{aligned}
\Pr\left[T_q \le t\right] &= \sum_{n=0}^{\infty} \Pr\left[T_q \le t \mid N = n\right] p_n \\
&= 1 - \sum_{n=1}^{\infty} \left(\sum_{k=0}^{n-1} e^{-\mu t} \frac{(\mu t)^k}{k!} \right) p_n \\
&= 1 - (1-\rho) \sum_{n=1}^{\infty} \left(\sum_{k=0}^{n-1} e^{-\mu t} \frac{(\mu t)^k}{k!} \right) \rho^n \\
&= 1 - (1-\rho)\, e^{-\mu t} \sum_{k=0}^{\infty} \frac{(\mu t)^k}{k!} \left(\sum_{n=k+1}^{\infty} \rho^n \right) \\
&= 1 - (1-\rho)\, e^{-\mu t} \sum_{k=0}^{\infty} \frac{(\mu t)^k}{k!} \frac{\rho^{k+1}}{1-\rho} \\
&= 1 - \rho e^{-\mu t} \sum_{k=0}^{\infty} \frac{(\rho \mu t)^k}{k!} \\
&= 1 - \rho e^{-(\mu - \lambda) t}.
\end{aligned}$$

Ein Kunde muss also mit Wahrscheinlichkeit $1 - \rho$ gar nicht warten und, wenn er warten muss, eine exponentialverteilte Zeit mit Parameter $\mu - \lambda$. Man beachte, dass die Annahme von exponentialverteilten Bedienungszeiten wesentlich in die Berechnung eingegangen ist. Da die Zwischenankunftszeiten exponentialverteilt sind, gilt diese Verteilung für die Wartezeit eines jeden Kunden, der in ein Bedienungssystem im Gleichgewicht eintritt. Daraus könnte natürlich wieder die mittlere Wartezeit gewonnen werden, die wir ja bereits mit der Formel von Little berechnet haben. Man kann aber auch die Wahrscheinlichkeit angeben, dass ein Kunde länger als einen vorgegebenen Wert warten muss - es könnte ja sein, dass dann zusätzliche Kosten anfallen.

2.3.3 Markov-Modelle mit unendlicher Population

Das $M/M/c$ Modell

Dabei handelt es sich um ein System mit c gleichartigen Servern. Das mathematische Modell ist ein Geburts- und Todesprozess mit den Raten

$$\begin{aligned} \lambda_n &= \lambda \\ \mu_n &= \begin{cases} n\mu & \text{falls} \quad 1 \leq n \leq c \\ c\mu & \text{falls} \quad n \geq c \end{cases} . \end{aligned}$$

Die Gleichgewichtsverteilung

$$p_n = \begin{cases} \frac{\lambda^n}{\mu^n n!} p_0 & \text{falls} \quad 1 \leq n \leq c \\ \frac{\lambda^n}{\mu^n c! c^{n-c}} p_0 & \text{falls} \quad n \geq c \end{cases}$$

ergibt sich aus der Formel (2.3), was sich für $\lambda <= \mu c$ normieren läßt mit

$$p_0 = \left(\sum_{n=0}^{c-1} \frac{\lambda^n}{\mu^n n!} + \frac{\lambda^c}{\mu^c c! \left(1 - \frac{\lambda}{\mu c}\right)} \right)^{-1} .$$

Für die erwartete Länge der Warteschlange gilt

$$L_q = \frac{\lambda^{c+1}}{c\mu^{c+1} c! \left(1 - \frac{\lambda}{\mu c}\right)^2} p_0 .$$

Daraus können nun mittels der allgemeinen Resultate L, W_q und W berechnet werden.
Auch für das $M/M/c$ Modell im Gleichgewicht kann die Verteilung der (Zufallsvariable) Wartezeit T_q eines (virtuellen) Kunden, der zum Zeitpunkt 0 eintrifft, durch folgende Überlegungen gewonnen werden.
Wegen der Gedächtnislosigkeit der Exponentialverteilung gilt für $n \in \mathbb{N}_0$ bei einer FCFS Strategie

$$\Pr\left[T_q \leq t \mid N = n\right] = \begin{cases} 1 & \text{für} \quad n < c \\ 1 - \sum_{k=0}^{n-c} e^{-c\mu t} \frac{(c\mu t)^k}{k!} & \text{für} \quad n \geq c \end{cases} .$$

Daraus folgt mit dem Satz von der totalen Wahrscheinlichkeit

$$\begin{aligned}
\Pr[T_q \le t] &= \sum_{n=0}^{\infty} \Pr[T_q \le t \mid N = n]\, p_n \\
&= 1 - \sum_{n=c}^{\infty} \left(\sum_{k=0}^{n-c} e^{-c\mu t} \frac{(c\mu t)^k}{k!} \right) p_n \\
&= 1 - \sum_{n=c}^{\infty} \left(\sum_{k=0}^{n-c} e^{-c\mu t} \frac{(c\mu t)^k}{k!} \right) \frac{\lambda^n}{\mu^n c! c^{n-c}} p_0 \\
&= 1 - \frac{e^{-c\mu t} c^c}{c!} p_0 \sum_{k=0}^{\infty} \frac{(c\mu t)^k}{k!} \left(\sum_{n=c+k}^{\infty} \frac{\lambda^n}{\mu^n c^n} \right) \\
&= 1 - \frac{e^{-c\mu t} c^c}{c!} p_0 \sum_{k=0}^{\infty} \frac{(c\mu t)^k}{k!} \frac{\left(\frac{\lambda}{\mu c}\right)^{c+k}}{1 - \frac{\lambda}{\mu c}} \\
&= 1 - \frac{e^{-c\mu t}}{c! \left(1 - \frac{\lambda}{\mu c}\right)} \left(\frac{\lambda}{\mu}\right)^c p_0 \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} \\
&= 1 - \frac{\left(\frac{\lambda}{\mu}\right)^c e^{-(c\mu - \lambda)t}}{c! \left(1 - \frac{\lambda}{\mu c}\right)} p_0.
\end{aligned}$$

Interessant ist auch die bedingte Wahrscheinlichkeit

$$\Pr[T_q > t \mid T_q > 0] = e^{-(c\mu - \lambda)t}$$

Ein Kunde muss also mit Wahrscheinlichkeit

$$\frac{\left(\frac{\lambda}{\mu}\right)^c}{c! \left(1 - \frac{\lambda}{\mu c}\right)} p_0$$

gar nicht warten und, wenn er warten muss, so muss er eine exponentialverteilte Zeit mit Mittelwert $(c\mu - \lambda)^{-1}$ warten.
Man beachte, dass alle Formeln für das $M/M/c$ Modell für $c = 1$ in die entsprechenden Formeln für das $M/M/1$ Modell übergehen.

Das $M/M/c/K$ Modell

Dabei handelt es sich um ein System mit c gleichartigen Servern mit einer Systemkapazität K ($K \ge c$), dh Kunden, die bereits K Kunden im System vorfinden, werden abgewiesen. Das mathematische Modell ist ein Geburts- und Todesprozess mit den Raten

$$\begin{aligned}
\lambda_n &= \begin{cases} \lambda & \text{falls} \quad 0 \le n < K \\ 0 & \text{falls} \quad n \ge K \end{cases} \\
\mu_n &= \begin{cases} n\mu & \text{falls} \quad 1 \le n \le c \\ c\mu & \text{falls} \quad n \ge c \end{cases} .
\end{aligned}$$

Dieser Prozess hat nur die Zustände $\{0, 1, \ldots, K\}$. Die Gleichgewichtsverteilung

$$p_n = \begin{cases} \frac{\lambda^n}{\mu^n n!} p_0 & \text{falls} \quad 1 \leq n \leq c \\ \frac{\lambda^n}{\mu^n c! c^{n-c}} p_0 & \text{falls} \quad c \leq n \leq K \end{cases}$$

ergibt sich wieder aus der Formel (2.3). Sie existiert in jedem Fall, wobei

$$p_0 = \left(\sum_{n=0}^{c-1} \frac{\lambda^n}{\mu^n n!} + \frac{\lambda^c \left(1 - \left(\frac{\lambda}{\mu c}\right)^{K-c+1}\right)}{\mu^c c! \left(1 - \frac{\lambda}{\mu c}\right)} \right)^{-1} \qquad \text{für } \lambda \leq \mu c \text{ bzw}$$

$$p_0 = \left(\sum_{n=0}^{c-1} \frac{\lambda^n}{\mu^n n!} + \frac{\lambda^c}{\mu^c c!} (K - c + 1) \right)^{-1} \qquad \text{für } \lambda = \mu c.$$

Man beachte, dass die Formel für $K \to \infty$ in die betreffende Formel für das $M/M/c$ Modell übergeht.
Die Formel für die erwartete Länge der Warteschlange lautet für $\lambda \neq \mu c$

$$L_q = \frac{\lambda^{c+1} p_0 \left(1 - \left(\frac{\lambda}{\mu c}\right)^{K-c+1} - \left(1 - \frac{\lambda}{\mu c}\right)(K - c + 1)\left(\frac{\lambda}{\mu c}\right)^{K-c}\right)}{c\mu^{c+1} c! \left(1 - \frac{\lambda}{\mu c}\right)^2}.$$

Die entsprechende Formel für $\lambda = \mu c$ wird durch die Regel von de L'Hospital gewonnen. Die erwarteten Wartezeiten werden wieder über die Formel von Little berechnet, allerdings muss man beachten, dass die mittlere Anzahl von eintreffenden Kunden je Zeiteinheit hier gleich $\lambda(1 - p_K)$ ist, weil ja nicht jeder eintreffende Kunde tatsächlich in das System eintritt. Wir erhalten daher

$$W_q = \frac{L_q}{\lambda(1 - p_K)} \text{ und } W = W_q + \frac{1}{\mu},$$

sowie

$$L = W\lambda(1 - p_K) = L_q + \frac{\lambda(1 - p_K)}{\mu}.$$

Auch in diesem Fall kann die Verteilung der Wartezeit berechnet werden.

Das $M/M/c/c$ Modell - Erlangsches Verlustsystem

Dieses Modell wurde bereits 1917 von Erlang[9] für eine Telefonvermittlung verwendet. Dabei werden die ankommenden Telefongespräche durch einen Poisson-Prozess modelliert und die Sprechzeiten als exponentialverteilt angenommen. Die von ihm hergeleiteten Formeln dienten dazu, die Telefonzentrale richtig zu dimensionieren. Seine sogenannten *Erlangschen Verlustformeln* lauten

$$p_n = \frac{\lambda^n}{\mu^n n!} \left(\sum_{k=0}^{c} \frac{\lambda^k}{\mu^k k!} \right)^{-1} \qquad \text{für } 0 \leq n \leq c.$$

[9] Agner Krarup Erlang, 1878 - 1929, dänischer Ingenieur und Mathematiker

Mit der Wahrscheinlichkeit

$$p_c = \frac{\lambda^c}{\mu^c c!} \left(\sum_{k=0}^{c} \frac{\lambda^k}{\mu^k k!} \right)^{-1}$$

geht der Kunde verloren, weil er keine freie Telefonleitung vorfindet. Die große Bedeutung dieser Formeln liegt darin, dass sie sogar für ein $M/G/c/c$ Modell gelten, in dem die mittlere Bedienungszeit $1/\mu$ ist. Erlang nahm bei seinen Untersuchungen die Bedienungszeit als konstant an, sein Beweis für die Formeln war aber nicht ganz korrekt.

Das $M/M/\infty$ Modell

In diesem Modell wird jeder Kunde sofort bedient, oft ist es gut geeignet, um eine Selbstbedienungsanlage zu beschreiben. Mathematisch handelt es sich um einen Geburts- und Todesprozess mit $\lambda_n = \lambda$ und $\mu_n = n\mu$. Damit gilt

$$p_n = \frac{\lambda^n}{\mu^n n!} \exp\left(-\frac{\lambda}{\mu}\right).$$

Die Anzahl der Kunden im System ist Poisson-verteilt mit Parameter $\frac{\lambda}{\mu}$. Auch diese Formel gilt sogar für ein $M/G/\infty$ Modell.

Zustandsabhängiges Service

Kunden treffen nach einem Poisson-Prozess mit Rate λ in ein System mit einem Server ein, der negativ exponential verteilt mit unterschiedlicher Intensität arbeitet. Sind weniger als k Kunden im System, so arbeitet der Server langsam mit Rate μ_l. Bei mindestens k Kunden arbeitet er schnell mit Rate μ_s. Die Kapazität des System sei unbeschränkt. Wir erhalten die Raten

$$\begin{aligned} \lambda_n &= \lambda, \\ \mu_n &= \begin{cases} \mu_l & \text{falls} \quad 1 \le n < k \\ \mu_s & \text{falls} \quad n \ge k \end{cases}. \end{aligned}$$

Im Fall $\lambda/\mu_s < 1$ exisitiert die Gleichgewichtsverteilung und für sie gilt

$$p_n = \begin{cases} \frac{\lambda^n}{\mu_l^n} p_0 & \text{falls} \quad 1 \le n < k \\ \frac{\lambda^n}{\mu_l^{k-1} \mu_s^{n-k+1}} p_0 & \text{falls} \quad n \ge k \end{cases}$$

mit

$$p_0 = \begin{cases} \left(\frac{1-(\lambda/\mu_l)^k}{1-\lambda/\mu_l} + \frac{(\lambda/\mu_l)^{k-1}\lambda/\mu_s}{1-\lambda/\mu_s} \right)^{-1} & \text{falls} \quad \lambda/\mu_l \neq 1 \\ \left(k + \frac{\lambda/\mu_s}{1-\lambda/\mu_s} \right)^{-1} & \text{falls} \quad \lambda/\mu_l = 1 \end{cases}.$$

Verallgemeinerungen dieses Modells liegen nahe und sind genauso zu behandeln: Man kann statt einem c Server einsetzen, diese können je nach Anzahl der vorhandenen Kunden (Aufträge) mehrere verschiedene Arbeitsintensitäten aufweisen. Man kann auch bei Vorhandensein von k_1, k_2,.. Kunden einen 2., 3.,.. Server einsetzen.

Ungeduldige Kunden

Zumindest drei Typen von ungeduldigen Kunden sind denkbar. Solche, die sich bei einer gewissen Länge der Warteschlange nur ungern anstellen (*balking*), dh für jeden Systemzustand n gibt es eine *Warte* wahrscheinlichkeit w_n, mit der ein eintreffender Kunde in das System eintritt. Ein nächster Typ von ungeduldigen Kunden stellt sich zuerst an, verläßt aber, wenn er zu lange warten muss, das System ohne Bedienung. Diese Art von Ungeduld (*reneging*) kann man dadurch modellieren, dass in jedem Systemzustand n jeder Kunde mit Rate r_n das System verläßt. Wenn es für jeden Server eine eigene Warteschlange gibt, tritt eine dritte Art von Ungeduld auf, nämlich das Wechseln zwischen den Warteschlangen (*jockeying*).

2.3.4 Markov-Modelle mit endlicher Population

Der Poisson-Prozess für die eintreffenden Kunden ist nur dann ein gutes Modell, wenn es ein sehr großes Reservoir von potentiellen Kunden vorliegt. Liegt (etwa in einem kleinen Ort oder wenn die "Kunden" zu reparierende Maschinen in einer Firma sind) nur eine beschränkte Menge M von möglichen Kunden vor, so geht auch diese Größe in die Berechnung ein.

Reparatur

Ein einfaches mathematisches Modell mit c Servern und Systemkapazität M (dh $M > c$) ist ein Geburts- und Todesprozess mit den Raten

$$\lambda_n = \begin{cases} (M-n)\,\lambda & \text{falls} \quad 0 \le n < M \\ 0 & \text{falls} \quad n \ge M \end{cases},$$

$$\mu_n = \begin{cases} n\mu & \text{falls} \quad 1 \le n \le c \\ c\mu & \text{falls} \quad n \ge c \end{cases}.$$

Dieser Prozess hat den endlichen Zustandsraum $\{0, 1, \ldots, M\}$. Welches konkrete Beispiel könnte dadurch modelliert werden? Die Gleichgewichtsverteilung erfüllt

$$p_n = \begin{cases} \binom{M}{n} \frac{\lambda^n}{\mu^n} p_0 & \text{falls} \quad 1 \le n \le c \\ \binom{M}{n} \frac{n!}{c!c^{n-c}} \frac{\lambda^n}{\mu^n} p_0 & \text{falls} \quad c \le n \le M \end{cases}$$

mit

$$p_0 = \left(\sum_{n=0}^{c} \binom{M}{n} \frac{\lambda^n}{\mu^n} + \sum_{n=c+1}^{M} \binom{M}{n} \frac{n!}{c!c^{n-c}} \frac{\lambda^n}{\mu^n} \right)^{-1}.$$

Reparatur und Reserve

In einer Firma werden stets M Maschinen benötigt, die unabhängig voneinander arbeiten. Ist eine davon defekt, so wird dafür wenn möglich sofort eine Reservemaschine eingesetzt, von denen es Y gibt. Jede Maschine hat auch nach einer Reparatur eine negativ exponential verteilte Lebensdauer mit Erwartungswert λ. Defekte Maschinen kommen in

die Reparatur, wo an maximal $c \leq Y$ Maschinen gleichzeitig Reparaturen durchgeführt werden können. Die Reparatur einer Maschine dauert eine negativ exponential verteilte Zeitspanne mit Mittelwert μ. Die Größen M, λ, μ sind vorgegeben. Ziel ist es, dass der Prozess nicht länger als d% der Zeit mit weniger als M Maschinen auskommen muss.
Wir bezeichnen mit n den Zustand des Systems, bei dem n Maschinen defekt sind. Dann hat der Prozess den endlichen Zustandsraum $\{0, 1, \ldots, M + Y\}$ und es liegt ein Geburts- und Todesprozess vor mit den Raten

$$\lambda_n = \begin{cases} M\lambda & \text{falls} \quad 0 \leq n \leq Y \\ (M + Y - n)\,\lambda & \text{falls} \quad Y \leq n \leq M + Y \\ 0 & \text{falls} \quad n \geq M + Y \end{cases},$$

$$\mu_n = \begin{cases} n\mu & \text{falls} \quad 1 \leq n \leq c \\ c\mu & \text{falls} \quad n \geq c \end{cases}.$$

Die Gleichgewichtsverteilung erfüllt

$$p_n = \begin{cases} \frac{M^n}{n!}\frac{\lambda^n}{\mu^n}p_0 & \text{falls} \quad 0 \leq n \leq c \\ \frac{M^n}{c^{n-c}c!}\frac{\lambda^n}{\mu^n}p_0 & \text{falls} \quad c \leq n \leq Y \\ \frac{M^Y M!}{(M-n+Y)!c!c^{n-c}}\frac{\lambda^n}{\mu^n}p_0 & \text{falls} \quad Y \leq n \leq M + Y \end{cases},$$

wobei man p_0 durch Normierung erhält. Will man erreichen, dass höchstens 5% der Zeit weniger als M Maschinen intakt sind, so muss

$$\sum_{n=Y+1}^{Y+M} p_n < d$$

gelten. Ob man mehr Reservemaschinen oder mehr Reparaturkapazität einsetzt, hängt von konkreten Kostenüberlegungen ab.
In einem anderen Fall können aber Reparatur und Reservemaschinen so teuer sein, dass man besser mehr Systemausfälle einkalkuliert.

Aufgabe

Was ergibt sich im obigen Modell für $c > Y$? Leiten Sie die entsprechenden Formeln her. Beantworten Sie die konkrete Fragestellung für $\lambda = 0.1$, $\mu = 1$, $d = 0.05$ und $M = 5$ bzw. $M = 50$.

Auch in diesen Modellen kann die Verteilung der Wartezeit berechnet werden.

2.4 Zeitabhängiges Verhalten

Die Berechnung des zeitabhängigen Verhaltens einer Markov-Kette erfordert die Lösung der Vorwärtsdifferentialgleichungen von Kolmogorov. Im allgemeinen handelt es sich dabei um ein System von unendlich vielen linearen Differentialgleichungen. Eine Lösung ist daher normalerweise schwierig. Eine Ausnahme bilden die Modelle mit endlichem Zustandsraum.

Ein weiterer Zugang, um zeitabhängiges Verhalten zu analysieren, ist die sogenannte *Busy-period analysis*. Dabei wird die Verteilung der Länge einer Aktivitätsperiode berechnet. Das ist die Zeitspanne zwischen dem Eintreffen eines Kunden in ein leeres System und dem ersten Zeitpunkt nachher, zu dem ein Kunde ein leeres System zurückläßt. Die Verteilung der Länge der Aktivitätsperiode kann durch geringe Modifikation der Vorwärtsdifferentialgleichungen von Kolmogorov gewonnen werden. Nach der Aktivitätsperiode steht das System eine exponentialverteilte Zeitspanne leer. Dann beginnt die nächste Aktivitätsperiode, die dieselbe Verteilung hat wie die vorige.

2.4.1 Modelle mit endlichem Zustandsraum

Dadurch werden alle Systeme mit endlicher Systemkapazität und alle Modelle mit endlicher Population erfasst. Es handelt sich stets um Markov-Ketten mit kostanten Übergangsraten. Die Vorwärtsdifferentialgleichungen von Kolmogorov sind homogene lineare Differentialgleichungssysteme 1. Ordnung mit konstanten Koeffizienten mit Randbedingungen und Anfangsbedingungen, die sich aus der Anfangsverteilung ergeben. Sie können entweder über die Eigenwerte der Koeffizientenmatrix oder mittels Laplace[10]-Transformation gelöst werden.

Das $M/M/1/1$ Modell

In diesem Fall lauten die Vorwärtsdifferentialgleichungen von Kolmogorov

$$\mathbf{p}'(t) = \mathbf{p}(t)\,\mathbf{Q} \text{ mit } \mathbf{Q} = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix},$$

dh

$$\begin{aligned} p_0'(t) &= -\lambda p_0(t) + \mu p_1(t) \\ p_1'(t) &= \lambda p_0(t) - \mu p_1(t), \end{aligned}$$

wobei $p_0(t) + p_1(t) = 1$ gelten muss. Als Lösung ergibt sich für $t \geq 0$

$$\begin{aligned} p_0(t) &= p_0(0)\ e^{-(\lambda+\mu)t} + \frac{\mu}{\lambda+\mu}\left(1 - e^{-(\lambda+\mu)t}\right) \\ p_1(t) &= p_1(0)\ e^{-(\lambda+\mu)t} + \frac{\lambda}{\lambda+\mu}\left(1 - e^{-(\lambda+\mu)t}\right). \end{aligned}$$

Lösung und Grenzverteilung exisiteren für alle $\lambda, \mu > 0$. Ist $\mathbf{p}(0) = \mathbf{p} = \left(\frac{\mu}{\lambda+\mu}, \frac{\lambda}{\lambda+\mu}\right)$, dh startet der Prozess in der Gleichgewichtsverteilung $\mathbf{p}$, so gilt $\mathbf{p}(t) = \mathbf{p}$, dh der Prozess bleibt in der Gleichgewichtsverteilung. Sonst nähert sich die Verteilung $\mathbf{p}(t)$ exponentiell der Gleichgewichtsverteilung.

[10] Pierre-Simon Laplace 1749 - 1827, französischer Mathematiker

2.4.2 Modelle mit unendlichem Zustandsraum

Das $M/M/1/\infty$ Modell

In diesem Fall lauten die Vorwärtsdifferentialgleichungen von Kolmogorov

$$\mathbf{p}'(t) = \mathbf{p}(t)\,\mathbf{Q} \text{ mit } \mathbf{Q} = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \dots \\ \mu & -\lambda-\mu & \lambda & 0 & \dots \\ 0 & \mu & -\lambda-\mu & \lambda & \dots \\ 0 & 0 & \mu & -\lambda-\mu & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix},$$

dh für $n > 0$

$$\begin{aligned} p_0'(t) &= -\lambda p_0(t) + \mu p_1(t) \\ p_n'(t) &= \lambda p_{n-1}(t) - (\lambda+\mu)\, p_n(t) + \mu p_{n+1}(t). \end{aligned}$$

Dabei handelt es sich um ein System von unendlich vielen linearen Differentialgleichungen. Es wurde erst 1954 gelöst. Die Lösung läßt sich mit Hilfe von Besselfunktionen I_n darstellen. Der Beweis verläuft über erzeugende Funktionen und deren Laplace-Transformierte.

2.4.3 Das $M/M/\infty$ Modell

In diesem Fall lauten die Vorwärtsdifferentialgleichungen von Kolmogorov

$$\mathbf{p}'(t) = \mathbf{p}(t)\,\mathbf{Q} \text{ mit } \mathbf{Q} = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \dots \\ \mu & -\lambda-\mu & \lambda & 0 & \dots \\ 0 & 2\mu & -\lambda-2\mu & \lambda & \dots \\ 0 & 0 & 3\mu & -\lambda-3\mu & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix},$$

dh für $n > 0$

$$\begin{aligned} p_0'(t) &= -\lambda p_0(t) + \mu p_1(t) \\ p_n'(t) &= \lambda p_{n-1}(t) - (\lambda+n\mu)\, p_n(t) + (n+1)\,\mu p_{n+1}(t). \end{aligned}$$

Als Anfangsbedingung wählen wir $\mathbf{p}(0) = (1,0,0,\dots)$. Auch hier handelt es sich um ein System von unendlich vielen linearen Differentialgleichungen. Die Lösungsmethode kommt aber ohne Laplace Transformation aus.
Sei $P(z,t) := \sum_{n=0}^{\infty} p(t)\, z^n$ die erzeugende Funktion von $p(t)$, die für $z \in \mathbb{C}$ mit $|z| \le 1$ existiert. Durch Einsetzen in die obigen Differentialgleichungen erhält man eine partielle Differentialgleichung mit Anfangsbedingung für $P(z,t)$, deren Lösung

$$\begin{aligned} P(z,t) &= \exp\left(\frac{\lambda}{\mu}(z-1)\left(1-e^{-\mu t}\right)\right) \\ &= \exp\left(\frac{\lambda}{\mu}\left(1-e^{-\mu t}\right)z\right)\exp\left(-\frac{\lambda}{\mu}\left(1-e^{-\mu t}\right)\right) \end{aligned}$$

ist. Entwickeln in eine Potenzreihe ergibt die gesuchten zeitabhängigen Wahrscheinlichkeiten

$$p_n(t) = \frac{1}{n!}\left(\frac{\lambda}{\mu}\left(1 - e^{-\mu t}\right)\right)^n \exp\left(-\frac{\lambda}{\mu}\left(1 - e^{-\mu t}\right)\right).$$

Auch hier ergibt sich für $t \to \infty$ die Grenzverteilung.