Diskrete Ereignissysteme, Kapitel 3

3. Stochastische diskrete Ereignissysteme

3.1 Grundbegriffe der Wahrscheinlichkeitsrechnung

3.2 Stochastische Prozesse in diskreter Zeit

- ➥ Markov-Ketten in diskreter Zeit

3.3 Stochastische Prozesse in kontinuierlicher Zeit

- ➥ Markov-Ketten in kontinuierlicher Zeit
- ➥ Warteschlangen

Material/Folien von Thomas Erlebach. Vielen Dank!

Literatur zu Kapitel 3 der Vorlesung

☞ Schickinger, Steger: **Diskrete Strukturen. Band 2: Wahrscheinlichkeitstheorie und Statistik**.
Springer, Berlin, 2001.
[Kapitel 1–2: Grundlagen, Kapitel 4: Stochastische Prozesse]

☞ Bertsekas, Gallager: **Data Networks**. Second Edition.
Prentice Hall, Upper Saddle River, NJ, 1992.
[Chapter 3: Delay Models in Data Networks]

Weitere Literatur

✩ Kleinrock: **Queueing Systems, Volume 1: Theory**, John Wiley & Sons, 1975.

✩ Kleinrock: **Queueing Systems, Volume 2: Computer Applications**, John Wiley & Sons, 1976.

✩ Gross, Harris: **Fundamentals of Queueing Theory**, Wiley, 1998.

✩ Tanner: **Practical Queueing Analysis**, McGraw-Hill, 1995.

✩ Nelson: **Probability, Stochastic Processes, and Queueing Theory**, Springer, 1995.

✩ . . .

Stochastische Prozesse

- **Bisher:** Das Verhalten von DES war deterministisch.
- **Jetzt:** Einbeziehung von “Unsicherheit” auf der Grundlage stochastischer Prozesse. Unsicherheit bezüglich der Funktion und des Zeitverhaltens.
- **Beispiele:**
  - ❏ Modellierung nur statistisch beschreibbarer Ereignisse, wie Telefonanrufe, Berechnungszeiten von Tasks.
  - ❏ Quantitative Analyse von Verkehrsprozessen (Zahl von Anrufen/Zeit, . . .), Warteschlangen, Computernetzwerken, Rechnerarchitekturen.

3.1 Grundbegriffe

- Menge Ω von Elementarereignissen
- $\Pr[\omega]$ bezeichnet die Wahrscheinlichkeit von $\omega \in \Omega$
- Es muss gelten: $0 \le \Pr[\omega] \le 1$ und $\sum_{\omega\in\Omega} \Pr[\omega] = 1$.
- **Wahrscheinlichkeitsraum:** Ω mit $\Pr[\omega]$ für alle $\omega \in \Omega$
- **diskret**, falls Ω endlich oder abzählbar, sonst **kontinuierlich**.
- **Ereignis**: Teilmenge von Ω
- Wahrscheinlichkeit von $E \subseteq \Omega$: $\Pr[E] = \sum_{\omega\in E} \Pr[\omega]$
- Ereignisse A und B heissen **unabhängig**, falls
  $\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$.

Beispiel 1

Zufallsexperiment: Werfen eines Würfels mit 6 Seiten

- $\Omega = \{❶, ❷, ❸, ❹, ❺, ❻\}$
- $\Pr[❶] = \Pr[❷] = \cdots = \Pr[❻] = \frac{1}{6}$
- E = "gerade Zahl" = $\{❷, ❹, ❻\} \subseteq \Omega$
- $\Pr[E] = \Pr[❷] + \Pr[❹] + \Pr[❻] = \frac{1}{2}$
- F = "durch 3 teilbare Zahl" = $\{❸, ❻\} \subseteq \Omega$
- $\Pr[F] = \Pr[❸] + \Pr[❻] = \frac{1}{3}$
- E und F sind unabhängig, da
  $\Pr[E \cap F] = \Pr[❻] = \frac{1}{6} = \Pr[E] \cdot \Pr[F]$.

Beispiel 2

Zufallsexperiment: Paketübertragung

- Jeder Übertragungsversuch gelingt mit W'keit p.
- Elementarereignis ω_i: Es braucht i Versuche bis zur ersten erfolgreichen Übertragung.
- $\Omega = \{\omega_1, \omega_2, \ldots\}$ abzählbar unendlich
- $\Pr[\omega_1] = p$, $\Pr[\omega_2] = (1-p)p$, $\Pr[\omega_3] = (1-p)^2 p$
- Allgemein: $\Pr[\omega_i] = (1-p)^{i-1} p$
- Es gilt: $\sum_{i=1}^{\infty} (1-p)^{i-1} p = p \sum_{i=0}^{\infty} (1-p)^i = p \frac{1}{1-(1-p)} = 1$

Rechnen mit Wahrscheinlichkeiten

Seien $A, B \subseteq \Omega$ Ereignisse. Es gilt:

- $\Pr[\emptyset] = 0$, $\Pr[\Omega] = 1$
- $\Pr[\bar{A}] = 1 - \Pr[A]$, wobei $\bar{A} := \Omega \setminus A$
- $\Pr[A \cup B] \le \Pr[A] + \Pr[B]$
- $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$

Bedingte Wahrscheinlichkeiten

Seien A, B Ereignisse mit $\Pr[B] > 0$.

Die **bedingte Wahrscheinlichkeit** von A gegeben B ist definiert durch:

$$\Pr[A \mid B] := \frac{\Pr[A \cap B]}{\Pr[B]}$$

Multiplikationssatz. Für Ereignisse $A_1, \ldots, A_n$ mit $\Pr[A_1 \cap \cdots \cap A_n] > 0$ gilt:

$$\Pr[A_1 \cap \cdots \cap A_n] = \Pr[A_1] \cdot \Pr[A_2 \mid A_1] \cdot$$
$$\cdot \Pr[A_3 \mid A_1 \cap A_2] \cdot \ldots \cdot \Pr[A_n \mid A_1 \cap \cdots \cap A_{n-1}]$$

Totale Wahrscheinlichkeit

Satz von der totalen Wahrscheinlichkeit

Seien $A_1, \ldots, A_n$ paarweise disjunkte Ereignisse und $B \subseteq A_1 \cup \cdots \cup A_n$. Dann folgt:

$$\Pr[B] = \sum_{i=1}^{n} \Pr[B \mid A_i] \cdot \Pr[A_i]$$

Bemerkung: Der Satz gilt analog für unendlich viele paarweise disjunkte Ereignisse $A_1, A_2, \ldots$:

$$\Pr[B] = \sum_{i=1}^{\infty} \Pr[B \mid A_i] \cdot \Pr[A_i]$$

Zufallsvariablen

- Eine Abbildung $X : \Omega \to \mathbb{R}$ heisst **Zufallsvariable**. Wir schreiben W_X als Abkürzung für den Wertebereich $X(\Omega)$.
- Falls Ω diskret (endlich oder abzählbar unendlich) ist, heisst auch X diskret. Wir betrachten vorerst nur diskrete Zufallsvariablen
- Die Funktionen $f_X : \mathbb{R} \to [0, 1]$ und $F_X : \mathbb{R} \to [0, 1]$ mit $f_X(x) = \Pr[X = x]$ und $F_X(x) = \Pr[X \leq x]$ heissen **Dichte(funktion)** und **Verteilung(sfunktion)** von X.
- **Erwartungswert**: $\mathbb{E}[X] := \sum_{x \in W_X} x \cdot \Pr[X = x]$
  (falls die Summe konvergiert)
  Varianz: $\mathrm{Var}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
  (falls $\mathbb{E}[X^2]$ und $\mathbb{E}[X]$ existieren)
  Standardabweichung: $\sigma(X) := \sqrt{\mathrm{Var}[X]}$

Rechnen mit Erwartungswert und Varianz

Mit $a, b \in \mathbb{R}$ gilt für die transformierte Zufallsvariable $a \cdot X + b$:

- $\mathbb{E}[a \cdot X + b] = a \cdot \mathbb{E}[X] + b$
- $\mathrm{Var}[a \cdot X + b] = a^2 \cdot \mathrm{Var}[X]$.

Linearität des Erwartungswerts:

Für Zufallsvariablen $X_1, X_2, \ldots, X_n$ und $X := a_1X_1 + \cdots + a_nX_n$ mit $a_1, \ldots, a_n \in \mathbb{R}$ gilt:

$$\mathbb{E}[X] = a_1\mathbb{E}[X_1] + \cdots + a_n\mathbb{E}[X_n]$$

Unabhängigkeit von Zufallsvariablen

- Zufallsvariablen $X_1, \dots, X_n$ heissen **unabhängig**, wenn für alle $(x_1, \dots, x_n) \in W_{X_1} \times \cdots \times W_{X_n}$ gilt:

$$\Pr[X_1 = x_1, \dots, X_n = x_n] = \Pr[X_1 = x_1] \cdot \dots \cdot \Pr[X_n = x_n]$$

- Für unabhängige Zufallsvariablen X und Y gilt:

$$\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$$

Beispiele diskreter Verteilungen (1)

- **Bernoulli-Verteilung** mit Erfolgsw'keit p:

$$\Pr[X = 1] = p, \qquad \Pr[X = 0] = 1 - p$$

Es gilt $\mathbb{E}[X] = p$ und $\mathrm{Var}[X] = p(1-p)$.

- **Binomial-Verteilung** mit Parametern n und p:

$$\Pr[X = k] = \binom{n}{k} p^k (1-p)^{n-k}, \qquad \text{für } 0 \le k \le n$$

Es gilt $\mathbb{E}[X] = np$ und $\mathrm{Var}[X] = np(1-p)$.

- **Poisson-Verteilung** mit Parameter λ:

$$\Pr[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad \text{für } k \in \mathbb{N}_0$$

Es gilt $\mathbb{E}[X] = \lambda$ und $\mathrm{Var}[X] = \lambda$.

Beispiele diskreter Verteilungen (2)

- **Geometrische Verteilung** mit Parameter p:

$$\Pr[X = k] = (1-p)^{k-1} p \qquad \text{für } k \in \mathbb{N}$$

Der Erwartungswert ist

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1} p = p \sum_{k=1}^{\infty} k(1-p)^{k-1} = p \cdot \frac{1}{p^2} = \frac{1}{p}$$

Die Varianz ist $\frac{1-p}{p^2}$.

Anwendungsbeispiel: Paketübertragung mit Erfolgsw'keit p

$\Rightarrow$ Es sind im Mittel $\frac{1}{p}$ Versuche nötig, bis ein Paket erfolgreich übertragen werden kann.

3.2 Stochastische Prozesse in diskreter Zeit

- dynamisches System $\sim$ zeitliche Folge von Zufallsexperimenten
- Zustand und Verhalten des Systems zur Zeit t wird als Zufallsvariable X_t modelliert. Wir betrachten nur Prozesse mit diskreten Zufallsvariablen X_t (zustandsdiskret).
- stochastischer Prozess: Folge von Zufallsvariablen $(X_t)_{t \in T}$
  - in diskreter Zeit: $T = \mathbb{N}_0$
  - in kontinuierlicher Zeit: $T = \mathbb{R}_0^+$
- Zufallsvariablen X_{t_1} und X_{t_2} können abhängig sein.
- **Markov-Prozesse**: Weiterer Ablauf ist nur vom aktuellen Zustand abhängig, nicht von der Vergangenheit.

Beispiel: Paketübertragung (1)

- Paketübertragung von Rechner *A* zu Rechner *B*.
- Jede Sekunde wird ein Paket übertragen.
- Zufallsvariablen X_t für $t \in \mathbb{N}_0$:

$$X_t = \begin{cases} 1 & \text{falls Übertragung zur Zeit } t \text{ erfolgreich} \\ 0 & \text{sonst} \end{cases}$$

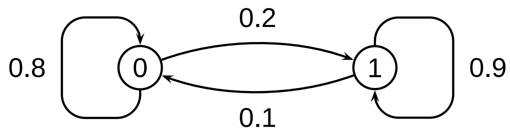
- **Annahme:** Wahrscheinlichkeit für erfolgreiche Übertragung zur Zeit t ist nur abhängig vom Erfolg der Übertragung zur Zeit $t-1$.

Beispiel: Paketübertragung (2)

- W'keit für Erfolg der Übertragung zur Zeit $t+1$:

$$\Pr[X_{t+1} = 0 \mid X_t = 0] = 0.8 \qquad \Pr[X_{t+1} = 0 \mid X_t = 1] = 0.1$$
$$\Pr[X_{t+1} = 1 \mid X_t = 0] = 0.2 \qquad \Pr[X_{t+1} = 1 \mid X_t = 1] = 0.9$$

- Graphische Veranschaulichung durch **Übergangsdiagramm**:

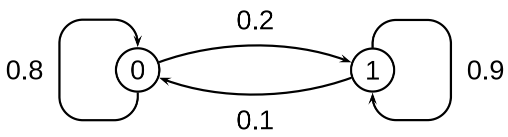


Kante von *a* nach *b* wird mit $\Pr[X_{t+1} = b \mid X_t = a]$ beschriftet.

Ablauf des Systems: **Random Walk** im Übergangsdiagramm.

Beispiel: Paketübertragung (3)

- Übergangsdiagramm



- Alternative Darstellung: **Übergangsmatrix**

$$P = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{pmatrix}$$

Eintrag p_{ij} entspricht $\Pr[X_{t+1} = j \mid X_t = i]$.

Beispiel: Paketübertragung (4)

Entwicklung des Systems über die Zeit:

t	$\Pr[X_t = 0]$	$\Pr[X_t = 1]$	
0	**0**	**1**	Anfangszustand (vorgegeben)
1	0.1	0.9	
2	0.17	0.83	
3	0.219	0.781	$\Pr[X_3 = 1] = 0.17 \cdot 0.2 + 0.83 \cdot 0.9$
4	0.253	0.747	$= 0.781$
	$\vdots$		
1000	0.333	0.667	

Definition: Markov-Kette

Endliche Markov-Kette in diskreter Zeit

über der Zustandsmenge $S = \{0, 1, \dots, n-1\}$:

- **Folge von Zufallsvariablen $(X_t)_{t\in\mathbb{N}_0}$** mit Wertemenge S
- **Startverteilung $q_0 = (q_{00}, q_{01}, \dots, q_{0,n-1})$** mit $q_{0,i} \geq 0$ und $\sum_{i=0}^{n-1} q_{0,i} = 1$.
- **X_{t+1} hängt nur von X_t ab**, d.h. für alle $t > 0$ und alle $I \subseteq \{0, 1, \dots, t-1\}$ und $i, j, s_k \in S$ (für alle $k \in I$) gilt:

$$\Pr[X_{t+1} = j \mid X_t = i, \forall k \in I: \; X_k = s_k] = \Pr[X_{t+1} = j \mid X_t = i]$$

(Falls $S = \mathbb{N}_0$, dann **unendliche Markov-Kette in diskreter Zeit**.)

Zeithomogene Markov-Ketten

- Falls $\Pr[X_{t+1} = j \mid X_t = i]$ für alle $i, j \in S$ unabhängig von t ist, so heisst die Markov-Kette **(zeit)homogen**.
- Wir betrachten (fast) nur zeithomogene Markov-Ketten.
- Für zeithomogene Markov-Ketten sind die Werte

$$p_{ij} := \Pr[X_{t+1} = j \mid X_t = i]$$

eindeutig definiert und ergeben die Übergangsmatrix

$$P = (p_{i,j})_{0 \leq i,j < n}.$$

Ablauf einer Markov-Kette

- Beobachtung einer Markov-Kette von Zeit 0 bis Zeit t_0.
- Möglicher Ablauf: Zustände $x_0, x_1, x_2, \dots, x_{t_0}$.
- Wahrscheinlichkeit für diesen Ablauf (Musterpfad):

$$q_{0,x_0} \cdot \Pr[X_1 = x_1 \mid X_0 = x_0] \cdot \dots \cdot \Pr[X_{t_0} = x_{t_0} \mid X_{t_0-1} = x_{t_0-1}]$$

Beispiel: Paketübertragung mit $P = \begin{pmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{pmatrix}$ und $q_0 = (0.5, 0.5)$:

- Wahrscheinlichkeit für Ablauf $(1, 1, 0, 0, 1, 0)$ ist:

$$0.5 \cdot 0.9 \cdot 0.1 \cdot 0.8 \cdot 0.2 \cdot 0.1 = 0.00072$$

Verweildauer

- Betrachte eine Markov-Kette, die zur Zeit t im Zustand i ist.
- Modelliere die Anzahl der Zeitschritte, die die Kette ab Zeit t im Zustand i bleibt, als Zufallsvariable V_i.
- Es gilt: $\Pr[V_i = k] = p_{ii}^{k-1}(1 - p_{ii})$ und $\Pr[V_i > k] = p_{ii}^k$.
- V_i ist also **geometrisch verteilt**.
- **Beachte:** Die Verweildauer ist unabhängig davon, wie lange die Kette schon im Zustand i war.

Rechnen mit der Übergangsmatrix (1)

- Startverteilung: q_0 (Zeilenvektor mit n Elementen)
- Wahrscheinlichkeitsverteilung auf den Zuständen zur Zeit t:
  $q_t = (q_{t,0}, \ldots, q_{t,n-1})$ mit $q_{t,i} = \Pr[X_t = i]$
- Berechnung von q_{t+1} aus q_t:

$$\begin{aligned} q_{t+1,j} &= \Pr[X_{t+1} = j] \\ &= \sum_{i=0}^{n-1} \Pr[X_t = i] \cdot \Pr[X_{t+1} = j \mid X_t = i] = \sum_{i=0}^{n-1} q_{t,i} \cdot p_{ij} \end{aligned}$$

Geschrieben als Multiplikation Vektor mit Matrix:

$$q_{t+1} = q_t \cdot P$$

Rechnen mit der Übergangsmatrix (2)

- Wir wissen also: $q_{t+1} = q_t \cdot P$.
- Dann muss gelten: $q_1 = q_0 \cdot P$

$$\begin{aligned} q_2 &= q_1 \cdot P = q_0 \cdot P \cdot P = q_0 \cdot P^2 \\ q_3 &= q_2 \cdot P = q_0 \cdot P^2 \cdot P = q_0 \cdot P^3 \\ &\vdots \end{aligned}$$

$$\boxed{q_t = q_0 \cdot P^t}$$

- Ebenso: $\boxed{q_{t+k} = q_t \cdot P^k \text{ für alle } k \geq 0}$
- Der Eintrag in Zeile i und Spalte j von P^k, bezeichnet mit $p_{ij}^{(k)} := (P^k)_{ij}$, gibt die Wahrscheinlichkeit an, in k Schritten von Zustand i nach Zustand j zu gelangen.

Chapman-Kolmogorov-Gleichungen

Kurze Exkursion: Zeitinhomogene Ketten

- $p_{ij}(k,n) := \Pr[X_n = j \mid X_k = i]$, $P(k,n) := (p_{ij}(k,n))_{i,j \in S}$.
- Betrachte Zeitpunkt u mit $k < u < n$:

$$\begin{aligned} p_{ij}(k,n) &= \Pr[X_n = j \mid X_k = i] \\ &= \sum_{r \in S} \Pr[X_n = j \mid X_u = r, X_k = i] \cdot \Pr[X_u = r \mid X_k = i] \\ &= \sum_{r \in S} \Pr[X_n = j \mid X_u = r] \cdot \Pr[X_u = r \mid X_k = i] \\ &= \sum_{r \in S} p_{rj}(u,n) \cdot p_{ir}(k,u) \end{aligned}$$

- $P(k,n) = P(k,u) \cdot P(u,n)$ für $k < u < n$.

Transientenanalyse

Typische Fragen:

- Wie gross ist die W'keit, nach k Schritten im Zustand j zu sein?
- Wie wahrscheinlich ist es, irgendwann von i nach j zu kommen?
- Wie viele Schritte benötigt die Kette im Mittel, um von i nach j zu gelangen?

Viele dieser Fragen können mit Hilfe der Gleichung

$$q_{t+k} = q_t \cdot P^k \text{ für alle } k \geq 0$$

beantwortet werden!

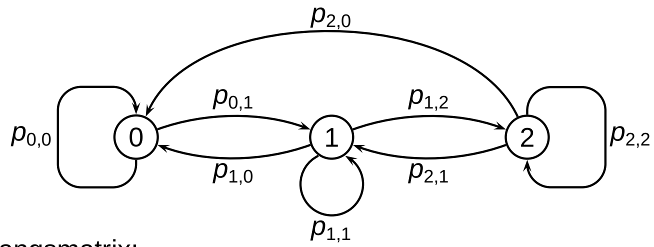
Beispiel (1)

Rechnersystem aus 2 Computern

- In jedem Zeitschritt wird dem System höchstens ein Task übergeben. Dieses Ereignis tritt mit W'keit a auf.
- Der ankommende Task wird nur bearbeitet, wenn mindestens einer der beiden Prozessoren frei ist oder im selben Zeitschritt frei wird.
- Falls ein Prozessor belegt ist, beendet er den Task in jedem Zeitschritt mit W'keit b.

Beispiel (2)

Modellierung als Markov-Kette mit Zustandsmenge $S = \{0, 1, 2\}$:



Übergangsmatrix:

$$P = \begin{pmatrix} 1-a & a & 0 \\ b(1-a) & (1-a)(1-b)+ab & a(1-b) \\ b^2(1-a) & b^2a+2b(1-b)(1-a) & (1-b)^2+2b(1-b)a \end{pmatrix}$$

Beispiel (3)

Für $a = 0.5$ und $b = 0.7$ ergibt sich:

$$P = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.35 & 0.5 & 0.15 \\ 0.245 & 0.455 & 0.3 \end{pmatrix}$$

Sei $q_0 = (1, 0, 0)$.

Fragen und Antworten:

- **W'keit, dass System zur Zeit 3 leer ist?**

  $q_3 = (1, 0, 0) \cdot P^3 = (0.405875, 0.496625, 0.0975)$

  $\Rightarrow \Pr[X_3 = 0] = 0.405875$

Beispiel (4)

- **W'keit, dass System zur Zeit 2 und zur Zeit 3 leer ist?**

$$\Pr[X_3 = 0, X_2 = 0] =$$
$$= \Pr[X_2 = 0] \cdot \Pr[X_3 = 0 \mid X_2 = 0]$$
$$= ((1, 0, 0) \cdot P^2)_0 \cdot p_{0,0} = 0.425 \cdot 0.5 = 0.2125$$

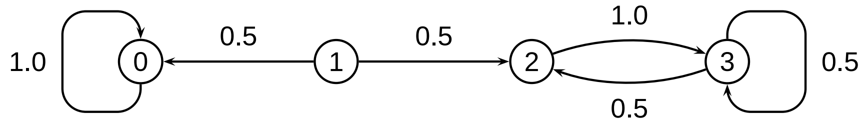
- **W'keit, dass zwischen Zeit 3 und 4 kein Task beendet wird?**

$$\Pr[\text{“keiner fertig zwischen 3 und 4”}]$$
$$= \sum_{j=0}^{2} \Pr[\text{“keiner fertig zwischen 3 und 4”} \mid X_3 = j] \cdot q_{3,j}$$
$$= 1 \cdot q_{3,0} + (1-b) \cdot q_{3,1} + (1-b)^2 \cdot q_{3,2}$$
$$= 1 \cdot 0.405875 + 0.3 \cdot 0.496625 + 0.09 \cdot 0.0975 \approx 0.564$$

Übergangszeit

- **Definition: Übergangszeit** (engl. *hitting time*)
  Zufallsvariable $T_{ij} := \min\{n \geq 1 \mid X_n = j, \text{wenn } X_0 = i\}$
  (falls Zustand j nie erreicht wird, setze $T_{ij} = \infty$)
- $h_{ij} := \mathbb{E}[T_{ij}]$ ist die erwartete Übergangszeit von i nach j.
- $f_{ij} := \Pr[T_{ij} < \infty]$ ist die **Ankunftswahrscheinlichkeit** von i nach j.

Beispiel



- $T_{01} = T_{02} = T_{03} = \infty$
- T_{10} ist 1, falls $X_1 = 0$, und ∞, falls $X_1 = 2$
  ➥ $f_{10} = 0.5$ und $h_{10} = \mathbb{E}[T_{10}] = 0.5 \cdot 1 + 0.5 \cdot \infty = \infty$
- $h_{32} = 0.5 \cdot 1 + 0.5^2 \cdot 2 + 0.5^3 \cdot 3 + \cdots = 0.5 \cdot \sum_{i=1}^{\infty} 0.5^{i-1} \cdot i = 0.5 \cdot \frac{1}{(1-0.5)^2} = 2.$

Berechnung der erwarteten Übergangszeiten

Lemma. Für die erwarteten Übergangszeiten gilt für alle $i, j \in S$

$$\boxed{h_{ij} = 1 + \sum_{k:k\neq j} p_{ik} h_{kj},}$$

falls die Erwartungswerte h_{ij} und h_{kj} existieren.

Für die Ankunftswahrscheinlichkeiten gilt analog

$$\boxed{f_{ij} = p_{ij} + \sum_{k:k\neq j} p_{ik} f_{kj}.}$$

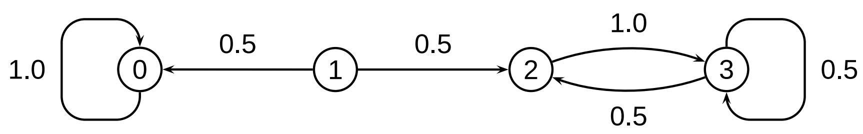
Beweis

Beweis. (nur für $h_{ij} = 1 + \sum_{k:k\neq j} p_{ik} h_{kj}$)

$$\begin{aligned} h_{ij} &= \mathbb{E}[T_{ij}] = \sum_{k\in S} \mathbb{E}[T_{ij} \mid X_1 = k] \cdot p_{ik} \\ &= \mathbb{E}[T_{ij} \mid X_1 = j] \cdot p_{ij} + \sum_{k:k\neq j} \mathbb{E}[T_{ij} \mid X_1 = k] \cdot p_{ik} \\ &= 1 \cdot p_{ij} + \sum_{k:k\neq j} (1 + \mathbb{E}[T_{kj}]) \cdot p_{ik} \\ &= 1 + \sum_{k\neq j} \mathbb{E}[T_{kj}] \cdot p_{ik} = 1 + \sum_{k:k\neq j} p_{ik} h_{kj} \end{aligned}$$

□

Anwendung auf das Beispiel



Wende $h_{ij} = 1 + \sum_{k:k\neq j} p_{ik} h_{kj}$ auf $i, j \in \{2, 3\}$ an:

$h_{22} = 1 + h_{32}$ $\qquad$ $h_{32} = 1 + 0.5 \cdot h_{32}$

$h_{23} = 1$ $\qquad$ $h_{33} = 1 + 0.5 \cdot h_{23}$

Lösen des Gleichungssystems liefert:

$h_{23} = 1$, $h_{33} = 1.5$, $h_{32} = 2$ und $h_{22} = 3$.

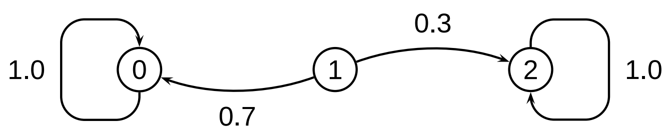
(Analog: $f_{22} = f_{33} = f_{23} = f_{32} = 1$.)

Stationäre Analyse

- Reale dynamische Systeme laufen oft über eine lange Zeit.
- Betrachte Verhalten für $t \to \infty$.
- Wahrscheinlichkeitsverteilung auf den Zuständen der Markov-Kette zur Zeit t ist $q_t = q_0 \cdot P^t$. **Konvergenz?**
- Intuitiv klar: Falls q_t für $t \to \infty$ gegen einen Vektor π konvergiert, so sollte π die Gleichung $\pi = \pi \cdot P$ erfüllen.
- **Definition.** Ein Zustandsvektor π mit $\pi_j \geq 0$ und $\sum_{j \in S} \pi_j = 1$ heisst **stationäre Verteilung** der Markov-Kette mit Übergangsmatrix P, falls $\pi = \pi \cdot P$.
- Die stationäre Verteilung π ist ein Eigenvektor von P zum Eigenwert 1.

(Nicht-)Eindeutigkeit der stat. Verteilung

- Übergangsdiagramm einer Beispiel-Markov-Kette:



- Übergangsmatrix: $P = \begin{pmatrix} 1 & 0 & 0 \\ 0.7 & 0 & 0.3 \\ 0 & 0 & 1 \end{pmatrix}$
- Diese Markov-Kette besitzt mehrere stationäre Verteilungen: zum Beispiel $(1, 0, 0)$ und $(0, 0, 1)$ und $(0.5, 0, 0.5)$
- Ursache: Zustände 0 und 2 sind “absorbierend”.

Irreduzible Markov-Ketten

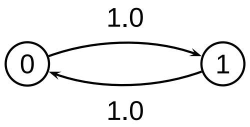
Definition. Eine Markov-Kette heisst **irreduzibel**, wenn es für alle Zustände $i, j \in S$ eine Zahl $n \in \mathbb{N}$ gibt, so dass $p_{ij}^{(n)} > 0$.

Satz. Eine irreduzible endliche Markov-Kette besitzt eine **eindeutige stationäre Verteilung** π und es gilt $\pi_j = 1/h_{jj}$ für alle $j \in S$.

Frage: Konvergiert eine irreduzible endliche Markov-Kette immer gegen ihre stationäre Verteilung?

Konvergenz?

Frage: Konvergiert eine irreduzible endliche Markov-Kette immer gegen ihre stationäre Verteilung? **NEIN!**



Diese Kette ist irreduzibel und endlich, aber der Zustandsvektor q_t konvergiert nicht unbedingt für $t \to \infty$:

$q_0 = (1, 0)$, $q_1 = (0, 1)$, $q_2 = (1, 0)$, $q_3 = (0, 1)$, ...

Ursache: Periodizität!

Aperiodische Markov-Ketten

- Die **Periode** eines Zustands $j \in S$ ist die grösste Zahl $\xi \in \mathbb{N}$, so dass gilt:

$$\{n \in \mathbb{N}_0 \mid p_{jj}^{(n)} > 0\} \subseteq \{i \cdot \xi \mid i \in \mathbb{N}_0\}$$

- Ein Zustand mit Periode $\xi = 1$ heisst **aperiodisch**.
- Eine Markov-Kette heisst **aperiodisch**, wenn alle Zustände aperiodisch sind.
- **Nützliche Testbedingung:** Zustand j ist aperiodisch, falls eine der beiden folgenden Bedingungen gilt:
  - $p_{jj} > 0$
  - $\exists n, m \in \mathbb{N}: \ p_{jj}^{(m)}, p_{jj}^{(n)} > 0$ und $\mathrm{ggT}(m, n) = 1$

Ergodische Markov-Ketten

- Irreduzible, aperiodische Markov-Ketten heissen **ergodisch**.

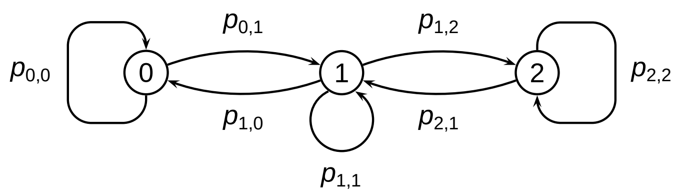
Fundamentalsatz für ergodische Markov-Ketten

Für jede ergodische endliche Markov-Kette gilt unabhängig vom Startzustand

$$\lim_{t \to \infty} q_t = \pi,$$

wobei π die eindeutige stationäre Verteilung der Kette ist.

Beispiel: Rechensystem mit 2 Computern



$$\text{Übergangsmatrix } P = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.35 & 0.5 & 0.15 \\ 0.245 & 0.455 & 0.3 \end{pmatrix}$$

- Kette ist aperiodisch und irreduzibel, also ergodisch.
- Aus $\pi = \pi P$ und $\pi_0 + \pi_1 + \pi_2 = 1$ erhält man die eindeutige stationäre Verteilung: $\pi = (0.399, 0.495, 0.106)$

Beispiel: Paging (1)

Modellierung eines Paging-Systems

- Hauptspeicher eines Rechners mit n logischen Seiten und $m < n$ physikalischen Seiten.
- Zugriff auf logische Seite σ, die nicht im physikalischen Hauptspeicher ist $\Rightarrow$ σ wird von Platte geladen, eine andere Seite wird aus dem physikalischen Hauptspeicher verdrängt.
- Zufallsvariable M_t gibt an, auf welche der n logischen Seiten zur Zeit t zugegriffen wird.
- **Annahme:** M_t unabhängig von t und von Zugriffen in anderen Zeitschritten, also $\Pr[M_t = i] = \beta_i$ für $1 \leq i \leq n$, wobei $\sum_{i=1}^{n} \beta_i = 1$.

Beispiel: Paging (2)

- Betrachte Paging-Strategie LRU (least recently used): Die Seite, die am längsten nicht mehr zugegriffen wurde, wird verdrängt.
- Modell des Paging-Systems: **Markov-Kette** in diskreter Zeit.
- Zustand s_t zur Zeit t: Menge der m im phys. Hauptspeicher befindlichen logischen Seiten nach Zugriff zur Zeit $t-1$.
- Betrachte Spezialfall $m = 2$: Zustand s_t zur Zeit t ist Paar $s_t = (i, j)$, wenn i und j die Seiten im physikalischen Speicher sind und zuletzt auf i zugegriffen wurde (zur Zeit $t-1$).
- Wenn $s_t = (i, j)$, dann $s_{t+1} = \begin{cases} (i, j) & \text{falls } M_t = i \\ (j, i) & \text{falls } M_t = j \\ (k, i) & \text{falls } M_t = k \end{cases}$

Beispiel: Paging (3)

Für $m = 2$ und $n = 3$ erhalten wir folgende Übergangsmatrix P:

	(1, 2)	(2, 1)	(1, 3)	(3, 1)	(2, 3)	(3, 2)
(1, 2)	β_1	β_2	0	β_3	0	0
(2, 1)	β_1	β_2	0	0	0	β_3
(1, 3)	0	β_2	β_1	β_3	0	0
(3, 1)	0	0	β_1	β_3	β_2	0
(2, 3)	β_1	0	0	0	β_2	β_3
(3, 2)	0	0	β_1	0	β_2	β_3

Beispiel: Paging (4)

- Markov-Kette ist irreduzibel und aperiodisch, also ergodisch.
- Durch Lösen des Gleichungssystems $\pi = \pi \cdot P$, $\sum_{(i,j) \in S} \pi_{(i,j)} = 1$ erhält man:

$$\pi_{(i,j)} = \frac{\beta_i \beta_j}{1 - \beta_i}.$$

- Wahrscheinlichkeit, dass im Zustand (i, j) eine Seite nachgeladen werden muss, ist $1 - (\beta_i + \beta_j)$.
- Über lange Zeit ist in jedem Zeitschritt die Wahrscheinlichkeit, dass eine Seite nachgeladen werden muss, gegeben durch:

$$\sum_{(i,j) \in S} (1 - \beta_i - \beta_j) \frac{\beta_i \beta_j}{1 - \beta_i}$$

Vektor-Ketten

- Prozesse, bei denen X_t nicht nur von X_{t-1} abhängt, sondern auch von $X_{t-2}, \ldots, X_{t-k}$, sind keine Markov-Prozesse.
- Sie können aber in Vektor-Ketten mit Markov-Eigenschaft umgewandelt werden.
- **Beispiel:** Der Prozess mit Zustandsmenge $S = \mathbb{Z}$ und $X_t = X_{t-1} - X_{t-2}$ erfüllt nicht die Markov-Bedinung.
- Der Vektorprozess mit Zustandsvektoren $Z_t = (X_t, X_{t-1})^T$ erfüllt die Markov-Bedingung:

$$Z_t = \begin{pmatrix} X_t \\ X_{t-1} \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} X_{t-1} \\ X_{t-2} \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix} \cdot Z_{t-1}$$

3.3 Stoch. Prozesse in kontinuierlicher Zeit

Stochastische Prozesse in kontinuierlicher Zeit

- Oft müssen diskrete Ereignis-Systeme betrachtet werden, bei denen die Ereignisse zu beliebigen Zeitpunkten eintreten können (d.h. in kontinuierlicher Zeit).
- Im Weiteren:
  - Kontinuierliche Zufallsvariablen
  - Markov-Ketten in kontinuierlicher Zeit
  - Warteschlangen

Kontinuierliche Zufallsvariablen

- Einer kontinuierlichen Zufallsvariable X liegt der kontinuierliche Wahrscheinlichkeitsraum $\Omega = \mathbb{R}$ zugrunde.
- X ist definiert durch eine integrierbare **Dichte** (auch: Dichtefunktion) $f_X : \mathbb{R} \to \mathbb{R}_0^+$ mit der Eigenschaft

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$$

- Jeder Dichte f_X kann eine **Verteilung** (auch: Verteilungsfunktion) F_X zugeordnet werden:

$$F_X(x) := \Pr[X \leq x] = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t$$

- $\Pr[a < X \leq b] = \int_{(a,b]} f_X(x)\,\mathrm{d}x = F_X(b) - F_X(a)$

Erwartungswert und Varianz

- Zur Berechnung von Erwartungswert und Varianz einer kontinuierlichen Zufallsvariable X ersetzen wir die Summen aus dem diskreten Fall durch Integrale.
- $\mathbb{E}[X] = \displaystyle\int_{-\infty}^{\infty} t \cdot f_X(t)\,\mathrm{d}t,$

  falls $\int_{-\infty}^{\infty} |t| \cdot f_X(t)\,\mathrm{d}t$ endlich.
- $\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \displaystyle\int_{-\infty}^{\infty} (t - \mathbb{E}[X])^2 \cdot f_X(t)\,\mathrm{d}t,$

  falls $\mathbb{E}[(X - \mathbb{E}[X])^2]$ existiert.
- Kontinuierliche Zufallsvariablen X und Y heissen **unabhängig**, falls $\forall x, y \in \mathbb{R}$: $\Pr[X \leq x, Y \leq y] = \Pr[X \leq x] \cdot \Pr[Y \leq y]$

Beispiele

Beispiele kontinuierlicher Verteilungen

- **Gleichverteilung** auf $[a, b]$:

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{falls } a \le x \le b \\ 0, & \text{sonst} \end{cases}$$

$$\mathbb{E}[X] = \frac{a+b}{2}, \quad \mathrm{Var}[X] = \frac{(b-a)^2}{12}$$

- **Normalverteilung** mit Parametern μ und σ:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\mathbb{E}[X] = \mu, \quad \mathrm{Var}[X] = \sigma^2$$

Exponentialverteilung (1)

Exponentialverteilung mit Parameter $\lambda > 0$

- Dichte $f_X(x) = \begin{cases} \lambda \cdot e^{-\lambda x}, & \text{falls } x \ge 0 \\ 0, & \text{sonst} \end{cases}$
- $\mathbb{E}[X] = \frac{1}{\lambda}, \quad \mathrm{Var}[X] = \frac{1}{\lambda^2}$
- Verteilungsfunktion $F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{falls } x \ge 0 \\ 0, & \text{sonst} \end{cases}$
- **Gutes Modell für:**
  - Dauer von Telefongesprächen
  - Zwischenankunftszeiten von Anfragen
  - Ausführungszeiten von Tasks

Exponentialverteilung (2)

Eigenschaften der Exponentialverteilung

- **Gedächtnislosigkeit:** Für alle $x, y > 0$ gilt:

$$\Pr[X > x + y \mid X > y] = \Pr[X > x]$$

- **Skalierung:** Falls X exponentialverteilt mit Parameter λ, so ist für $a > 0$ die Zufallsvariable $Y := aX$ exponentialverteilt mit Parameter λ/a.
- **Warteproblem:** Falls $X_1, \ldots, X_n$ unabhängig und exponentialverteilt mit Parametern $\lambda_1, \ldots, \lambda_n$, dann ist $X := \min\{X_1, \ldots, X_n\}$ exponentialverteilt mit dem Parameter $\lambda_1 + \cdots + \lambda_n$.

Ereignissysteme in kontinuierlicher Zeit

Diskrete Ereignissysteme in kontinuierlicher Zeit

- In vielen Systemen ist es unnatürlich, Ereignisse nur zu diskreten Zeitpunkten zuzulassen:
  - Ankunft von Paketen in einem Router
  - Auftreten von Anfragen an einen Server
- Um stochastische Prozesse in kontinuierlicher Zeit zu modellieren, können wieder Markov-Ketten verwendet werden:
  - Zustandsübergänge nicht nur zu diskreten Zeitpunkten zulassen, sondern exponentialverteilte Aufenthaltsdauern annehmen!

Markov-Ketten in kontinuierlicher Zeit (1)

Endliche Markov-Kette in kontinuierlicher Zeit

über der Zustandsmenge $S = \{0, 1, \ldots, n-1\}$:

- **Folge von Zufallsvariablen** $(X(t))_{t \in \mathbb{R}_0^+}$ mit Wertemenge S
- **Startverteilung** $q(0) = (q_0(0), q_1(0), \ldots, q_{n-1}(0))$ mit $q_i(0) \geq 0$ und $\sum_{i=0}^{n-1} q_i(0) = 1$.
- **Markov-Bedingung:** Für alle $k \in \mathbb{N}_0$ und beliebige $0 \leq t_0 < t_1 < \cdots < t_k < t$ und $s, s_0, \ldots, s_k \in S$ gilt:

$$\Pr[X(t) = s \mid X(t_k) = s_k, X(t_{k-1}) = s_{k-1}, \ldots, X(t_0) = s_0]$$
$$= \Pr[X(t) = s \mid X(t_k) = s_k]$$

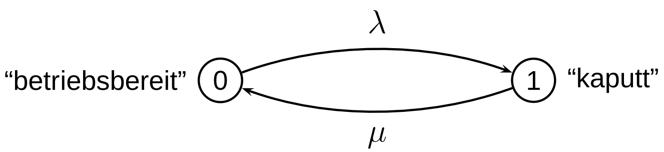
($S = \mathbb{N}_0 \Rightarrow$ **unendliche Markov-Kette in kontinuierlicher Zeit**.)

Markov-Ketten in kontinuierlicher Zeit (2)

- **Bemerkung:** Aus der Markov-Bedingung (Gedächtnislosigkeit) für die Markov-Kette kann man folgern, dass die Aufenthaltsdauern in den Zuständen exponentialverteilt sein müssen.
- Falls $\Pr[X(t+u) = j \mid X(t) = i] = \Pr[X(u) = j \mid X(0) = i]$ für alle $i, j \in S$ und $t, u \in \mathbb{R}_0^+$, so heisst die Markov-Kette **zeithomogen**.
- Wir betrachten im Folgenden ausschliesslich zeithomogene Markov-Ketten.

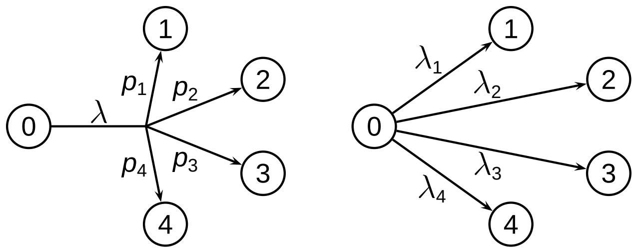
Beispiel

Beispiel einer Markov-Kette in kontinuierlicher Zeit:



- Aufenthaltsdauer in Zustand 0 ist exponentialverteilt mit Parameter λ.
- Aufenthaltsdauer in Zustand 1 ist exponentialverteilt mit Parameter μ.

Zustände mit mehreren Nachfolgern (1)



Gleichwertige Sichtweisen:

① Zustand 0 hat Aufenthaltsdauer exponentialverteilt mit Parameter λ. Wenn Zustand 0 verlassen wird, so werden die Nachfolger mit Wahrscheinlichkeit p_1, p_2, p_3, p_4 ausgewählt, $p_1 + p_2 + p_3 + p_4 = 1$.

② Es werden gleichzeitig vier Werte zufällig bestimmt gemäss Exponentialverteilungen mit Parameter λ_1 bis λ_4, wobei $\lambda_i = \lambda \cdot p_i$ für $i = 1, 2, 3, 4$. Der kleinste Wert “gewinnt”.

Zustände mit mehreren Nachfolgern (2)

Allgemein bedeutet das:

- Jeder Zustand $i \in S$ hat eine exponentialverteilte Aufenthaltsdauer mit Parameter ν_i.
- Wenn Zustand $i \in S$ verlassen wird, so wird mit Wahrscheinlichkeit p_{ij} der Nachfolgezustand $j \in S$ angenommen, wobei $p_{i,i} = 0$ und $\sum_{j \in S} p_{i,j} = 1$.
- Die **Übergangsrate** von Zustand i nach j ist als $\nu_{i,j} := \nu_i \cdot p_{i,j}$ definiert.
- Es gilt für $i \in S$: $\sum_{j \in S} \nu_{i,j} = \nu_i$.

Aufenthaltswahrscheinlichkeiten (1)

Bestimmung der Aufenthaltswahrscheinlichkeiten

- Startverteilung $q(0)$: $q_i(0) = \Pr[X(0) = i]$ für $i \in S$
- Verteilung zur Zeit t: $q_i(t) = \Pr[X(t) = i]$ für $i \in S$
- Die Änderung der Aufenthaltswahrscheinlichkeiten kann durch **Differentialgleichungen** für alle $i \in S$ beschrieben werden:

$$\boxed{\underbrace{\frac{d}{dt} q_i(t)}_{\text{Änderung}} = \underbrace{\sum_{j:j \neq i} q_j(t) \cdot \nu_{j,i}}_{\text{Zufluss}} - \underbrace{q_i(t) \cdot \nu_i}_{\text{Abfluss}}}$$

Aufenthaltswahrscheinlichkeiten (2)

$$\frac{d}{dt} q_i(t) = \sum_{j:j \neq i} q_j(t) \cdot \nu_{j,i} - q_i(t) \cdot \nu_i$$

- Lösung dieser Differentialgleichungen ist meist aufwändig.
- ➥ Betrachte Verhalten des Systems für $t \to \infty$.
- ➥ Falls die Aufenthaltswahrscheinlichkeiten gegen eine stationäre Verteilung konvergieren, so muss $\frac{d}{dt} q_i(t) = 0$ gelten.
- ➥ Man erhält für $t \to \infty$ ein lineares Gleichungssystem, das von einer stationären Verteilung π erfüllt werden muss:

$$0 = \sum_{j:j \neq i} \pi_j \cdot \nu_{j,i} - \pi_i \cdot \nu_i, \quad \text{für alle } i \in S$$

Irreduzible Markov-Ketten

- Ein Zustand j ist von i aus erreichbar, wenn es ein $t \geq 0$ gibt mit $\Pr[X(t) = j \mid X(0) = i] > 0$.
- Eine Markov-Kette, in der jeder Zustand von jedem anderen aus erreichbar ist, heisst **irreduzibel**.

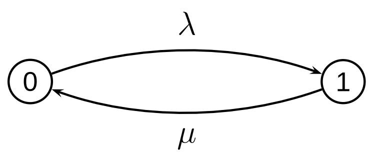
Satz. Für irreduzible Markov-Ketten existieren die Grenzwerte

$$\pi_i := \lim_{t \to \infty} q_i(t)$$

für alle $i \in S$ und ihre Werte sind unabhängig von $q(0)$.

Berechnung der stationären Verteilung

Im Beispiel:



Gleichungssystem:

$$0 = \mu \cdot \pi_1 - \lambda \cdot \pi_0$$
$$0 = \lambda \cdot \pi_0 - \mu \cdot \pi_1$$

Zusammen mit $\pi_0 + \pi_1 = 1$ erhält man:

$$\pi_0 = \frac{\mu}{\lambda + \mu} \qquad \pi_1 = \frac{\lambda}{\lambda + \mu}$$

Warteschlangen (1)

Warteschlangentheorie

- Besonders wichtige Anwendung von Markov-Ketten mit kontinuierlicher Zeit.
- Systeme mit Servern, die Jobs abarbeiten
- **Ankunftszeiten** der Jobs und **Bearbeitungsdauern** auf den Servern werden als Zufallsvariablen modelliert.
- Jobs, die ankommen, wenn alle Server belegt sind, werden in eine Warteschlange eingefügt.
- Ein freiwerdender Server wählt einen neuen Job aus der Warteschlange zur Bearbeitung aus (hier: FCFS, "first come, first serve," aber andere Strategien denkbar).

Warteschlangen (2)

- Beispielanwendung: Paketverzögerung in Datennetzen (Pakete = Jobs), Antwortzeiten von Tasks in Rechenzentren, ...
- Interessante Grössen wie
  - durchschnittliche Anzahl Jobs im System
  - durchschnittliche Verzögerung (Antwortzeit, Systemzeit, Aufenthaltsdauer) der Jobs

  werden in Abhängigkeit von der **Ankunftsrate** (mittlere Anzahl ankommender Jobs pro Zeiteinheit) und den **Bearbeitungsdauern** analysiert, wobei das System über lange Zeit betrachtet wird.

Kendall-Notation $X/Y/m/\ldots$

- X steht für die Verteilung der **Zwischenankunftszeiten** (Zeiten zwischen zwei ankommenden Jobs).
- Y steht für die Verteilung der reinen **Bearbeitungszeiten** (d.h. ohne Wartezeit) der Jobs auf dem Server.
- Die Zwischenankunftszeiten und Bearbeitungszeiten sind **unabhängige** Zufallsvariablen.
- m steht für die **Anzahl der Server**.
- Die Verteilungen für X und Y werden angegeben als:
  - **"D"** für feste Dauer (engl. *deterministic*)
  - **"M"** für exponentialverteilt (engl. *memoryless*)
  - **"G"** für beliebige Verteilung (engl. *general*)

Der Poisson-Prozess

- Im Fall von exponentialverteilten Zwischenankunftszeiten (mit Parameter λ) ist der Ankunftsprozess der Jobs ein **Poisson-Prozess mit Rate λ**.
- Die Anzahl ankommender Jobs in einem Intervall der Länge τ ist nämlich Poisson-verteilt mit Rate $\lambda\tau$:

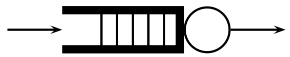
$$\Pr[\alpha(t+\tau) - \alpha(t) = n] = e^{-\lambda\tau}\frac{(\lambda\tau)^n}{n!}, \quad \text{für } n = 0, 1, 2, \ldots$$

$$\mathbb{E}[\alpha(t+\tau) - \alpha(t)] = \lambda \cdot \tau$$

- Poisson-Prozesse sind ein gutes Modell für die Ankunft von Paketen, Anfragen, Telefongesprächen, Jobs, etc., die von **vielen unabhängigen und ähnlichen Benutzern** erzeugt werden.

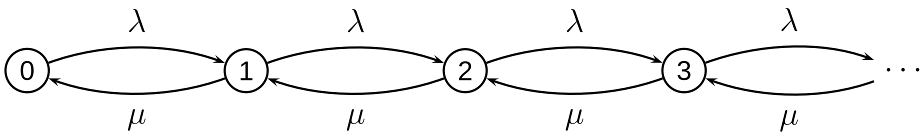
M/M/1-Warteschlangen

Die M/M/1-Warteschlange



- Zwischenankunftszeiten und Bearbeitungszeiten exponential-verteilt mit Parameter λ (Ankunftsrate) bzw. μ (Bedienrate).
- Definition: **Verkehrsdichte** $\rho := \dfrac{\lambda}{\mu}$
- Modellierung als Markov-Kette in kontinuierlicher Zeit:
  - Zustand: Anzahl Jobs im System (Warteschlange + Server)
  - Zustandsmenge $S = \mathbb{N}_0$
  - Übergangsrate von i nach $i+1$ ist λ.
  - Übergangsrate von $i > 0$ nach $i-1$ ist μ.

M/M/1: Stationäre Verteilung (1)



Gleichungssystem für stationäre Verteilung π:

$$\begin{aligned} 0 &= \mu \cdot \pi_1 - \lambda\pi_0 \\ 0 &= \lambda \cdot \pi_{k-1} + \mu \cdot \pi_{k+1} - (\lambda+\mu)\pi_k \quad \text{für alle } k \geq 1 \end{aligned}$$

Umformen liefert:

$$\mu \cdot \pi_{k+1} - \lambda \cdot \pi_k = \mu \cdot \pi_k - \lambda \cdot \pi_{k-1} = \cdots = \mu \cdot \pi_1 - \lambda \cdot \pi_0 = 0$$

$$\Rightarrow \mu \cdot \pi_k - \lambda \cdot \pi_{k-1} = 0 \Rightarrow \pi_k = \rho \cdot \pi_{k-1} \Rightarrow \pi_k = \rho^k \cdot \pi_0$$

M/M/1: Stationäre Verteilung (2)

Wir wissen: $\pi_k = \rho^k \cdot \pi_0$ für alle $k \geq 0$

- Falls $\rho \geq 1$, ist $\pi = (0, 0, \ldots)$ die einzige Lösung. Das System **konvergiert nicht**, die Warteschlange wächst ins Unendliche.
- Falls $\rho < 1$, so rechnen wir:

$$1 = \sum_{k=0}^{\infty} \pi_k = \pi_0 \cdot \sum_{k=0}^{\infty} \rho^k = \pi_0 \cdot \frac{1}{1-\rho} \Rightarrow \pi_0 = 1 - \rho$$

Das System **konvergiert** gegen eine stationäre Verteilung π mit $\pi_k = (1-\rho)\rho^k$ für alle $k \geq 0$.

Die mittlere **Auslastung** des Servers ist $1 - \pi_0 = \rho$.

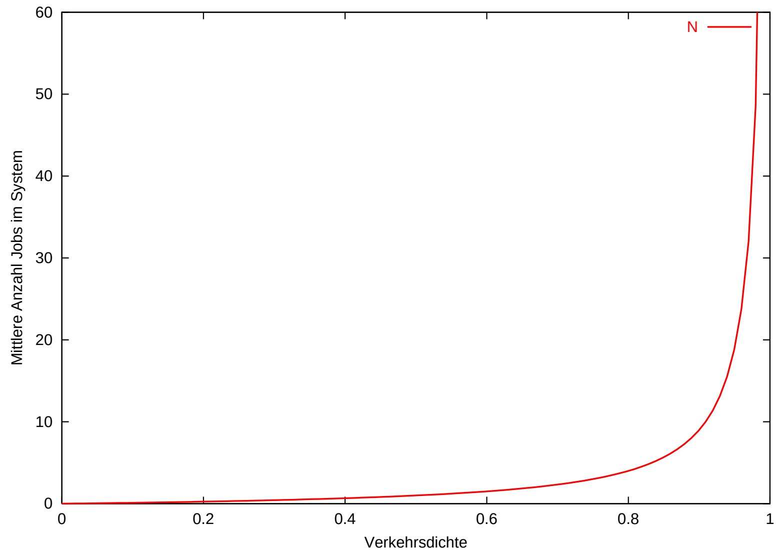
M/M/1: Anzahl Jobs im System (1)

- Sei N der Erwartungswert der Anzahl der Jobs im System (Warteschlange + Server).
- In der stationären Verteilung ergibt sich:

$$
\begin{aligned}
N &= \sum_{k=0}^{\infty} k \cdot \pi_k = \sum_{k=0}^{\infty} k(1-\rho)\rho^k = (1-\rho)\rho \sum_{k=0}^{\infty} k\rho^{k-1} \\
&= (1-\rho)\rho \frac{1}{(1-\rho)^2} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}
\end{aligned}
$$

Die Varianz der Anzahl Jobs im System ist $\frac{\rho}{(1-\rho)^2}$.

M/M/1: Anzahl Jobs im System (2)



Little's Law – Definitionen

- $N(t)$:= Anzahl Jobs im System (Warteschlange + Server) zur Zeit t.
- $\alpha(t)$:= Anzahl Jobs, die in [0,t] angekommen sind.
- T_i := Antwortzeit des i-ten Jobs (Wartezeit + Bearbeitungszeit).

Berechne **Durchschnittswerte** bis zur Zeit t:

$$N_t := \frac{1}{t}\int_0^t N(\tau)\,d\tau, \qquad \lambda_t := \frac{\alpha(t)}{t}, \qquad T_t := \frac{\sum_{i=1}^{\alpha(t)} T_i}{\alpha(t)}$$

Betrachte **Grenzwerte** für $t \to \infty$:

$$N := \lim_{t\to\infty} N_t, \qquad \lambda := \lim_{t\to\infty} \lambda_t, \qquad T := \lim_{t\to\infty} T_t,$$

Little's Law – Mittelwerte über die Zeit

Formel von Little

Falls die Grenzwerte

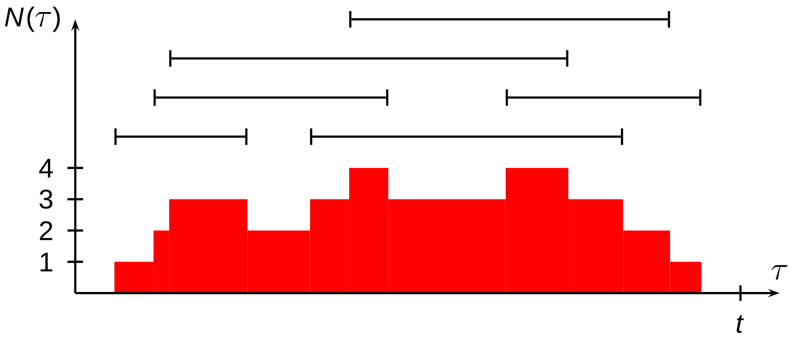
$$N := \lim_{t\to\infty} N_t, \qquad \lambda := \lim_{t\to\infty} \lambda_t, \qquad T := \lim_{t\to\infty} T_t,$$

existieren und auch $\lim_{t\to\infty} \frac{\beta(t)}{t}$ existiert und gleich λ ist, wobei $\beta(t)$ die Anzahl der in $[0, t]$ beendeten Jobs ist, so gilt:

$$\boxed{N = \lambda \cdot T}$$

Bemerkung: Die Formel von Little gilt auch für andere Server-Strategien als FCFS.

Beweisidee zur Formel von Little



Annahme: $N(0) = 0$ und $N(t) = 0$ für unendlich viele, beliebig grosse t. Dann gilt:

$$\underbrace{\frac{\alpha(t)}{t}}_{\xrightarrow{t\to\infty} \lambda} \cdot \underbrace{\frac{1}{\alpha(t)} \cdot \sum_{i=1}^{\alpha(t)} T_i}_{\xrightarrow{t\to\infty} T} = \underbrace{\frac{1}{t} \int_0^t N(\tau)\, d\tau}_{\xrightarrow{t\to\infty} N}$$

Little's Law – Stochastische Variante

Betrachte den Ablauf des Systems als stochastischen Prozess mit gegebener Startverteilung. Falls die Grenzwerte

$$\overline{N} = \lim_{t\to\infty} \mathbb{E}[N(t)], \qquad \overline{T} = \lim_{i\to\infty} \mathbb{E}[T_i], \qquad \lambda = \lim_{t\to\infty} \frac{\mathbb{E}[\alpha(t)]}{t}$$

existieren, so gilt:

$$\boxed{\overline{N} = \lambda \cdot \overline{T}}$$

Bemerkung 1: Die Formel von Little gilt für beliebige Verteilungen der Zwischenankunftszeiten und der Bearbeitungszeiten.

Bemerkung 2: Meist gilt $N = \overline{N}$ und $T = \overline{T}$ mit W'keit 1.

Little's Law und M/M/1-Warteschlangen

Mit $N = \frac{\rho}{1-\rho}$ und der Formel von Little erhalten wir:

- $T = \frac{1}{\lambda} N = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu - \lambda}$,

  wobei T die mittlere **Antwortzeit** eines Jobs im Gleichgewichtszustand des Systems ist.

- $W = T - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)} = \frac{\rho}{\mu - \lambda}$,

  wobei W die mittlere **Wartezeit** (ohne Bearbeitungszeit) eines Jobs im Gleichgewichtszustand des Systems ist.

- $N_Q = \lambda W = \frac{\rho^2}{1-\rho}$, wobei N_Q die mittlere Anzahl Jobs in der Warteschlange ist.

Anwendungen der Formel von Little (1)

Pakete in einem Datennetz

- Pakete werden an n Knoten in einem Netz mit Ankunftsraten $\lambda_1, \lambda_2, \ldots, \lambda_n$ erzeugt.
- Jedes Paket wird im Netz zu seiner Zieladresse geleitet und dort aus dem Netz entfernt.
- Sei N die durchschnittliche Zahl von Paketen im Netz.
- Mit der Formel von Little lässt sich die durchschnittliche Paketverzögerung berechnen als:

$$T = \frac{N}{\sum_{i=1}^{n} \lambda_i}$$

Anwendungen der Formel von Little (2)

Ein geschlossenes Warteschlangensystem

- System mit K Servern und Platz für $N \geq K$ Jobs
- System sei immer voll ($N(t) = N$). Wenn ein Job abgearbeitet ist und das System verlässt, kommt sofort ein neuer Job an.
- Alle K Server bearbeiten durchgehend Jobs.
- Mittlere Bearbeitungszeit ist $\overline{X}$.

Bestimmung der **mittleren Antwortzeit T**:

- $N = \lambda T$ (Formel angewendet auf ganzes System)
- $K = \lambda \overline{X}$ (Formel angewendet auf K Server)

⇨ $T = \frac{N}{\lambda} = \frac{N \cdot \overline{X}}{K}$

Anwendungen der Formel von Little (3)

Variante des Systems:

- Jobs kommen mit Rate λ an.
- Jobs werden abgewiesen, wenn bereits N Jobs im System sind.

Analyse des **Anteils abgewiesener Jobs:**

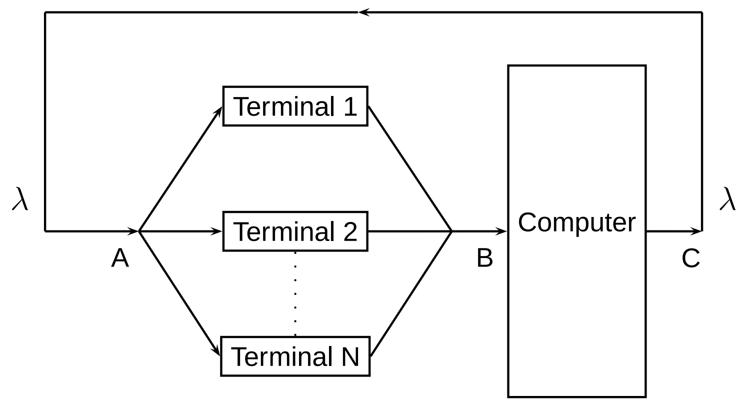
- $\overline{K}$ = mittlere Anzahl aktiver Server
- β = Anteil abgewiesener Jobs
- Formel von Little ⇨ $\overline{K} = (1 - \beta)\lambda\overline{X}$
- Also: $\beta = 1 - \frac{\overline{K}}{\lambda\overline{X}} \geq 1 - \frac{K}{\lambda\overline{X}}$ (untere Schranke für β)

Durchsatzanalyse für Time-Sharing (1)

- System mit N Terminals, die mit einem Time-Sharing Computer verbunden sind.
- Benutzer an einem Terminal verhalten sich folgendermassen:
  1. nachdenken (im Mittel R Sekunden)
  2. an den Computer einen Job abschicken, der im Mittel Ausführungszeit P hat
  3. auf die Beendigung des Jobs warten
  4. System verlassen
- Im Computer werden die Jobs in einer Warteschlange eingereiht und von einer CPU abgearbeitet.
- **Ziel:** maximal erreichbaren Durchsatz λ abschätzen!

Durchsatzanalyse für Time-Sharing (2)

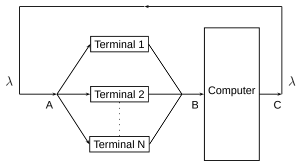
Schematische Darstellung des Systems:



Annahme: freiwerdende Terminals werden sofort wieder belegt

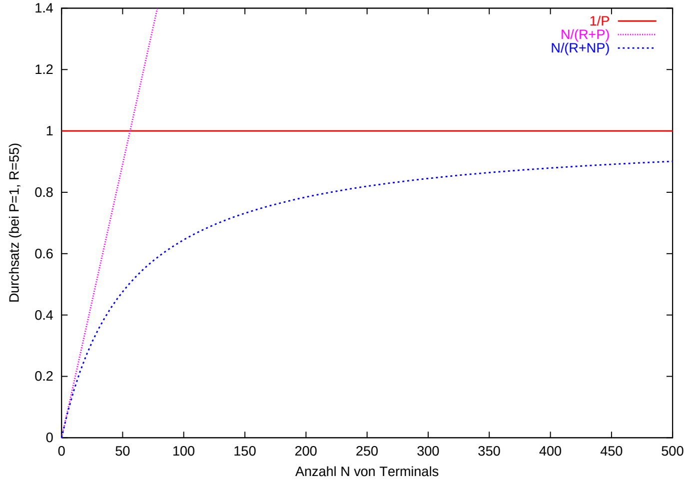
⇨ immer genau N Benutzer im System

Durchsatzanalyse für Time-Sharing (3)



- Es gilt $\lambda = \frac{N}{T}$, wobei T = mittlere Aufenthaltszeit. (Formel von Little angewendet auf System zwischen A und C)
- Es gilt $T = R + D$, wobei $D \in [P, N \cdot P]$ die mittlere Zeit vom Abschicken eines Jobs bis zu seiner Erledigung ist.
- Somit gilt: $\frac{N}{R+NP} \leq \lambda \leq \frac{N}{R+P}$
- Klar: $\lambda \leq \frac{1}{P}$, da mittlere Ausführungszeit P ist.

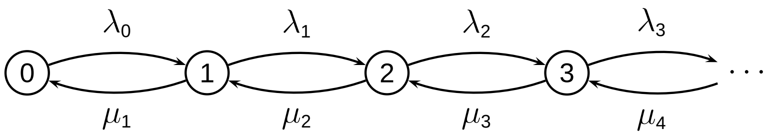
Durchsatzanalyse für Time-Sharing (4)



Maximal erzielbarer Durchsatz erfüllt $\frac{N}{R+NP} \leq \lambda \leq \min\left\{\frac{N}{R+P}, \frac{1}{P}\right\}$.

Birth-and-Death Prozesse

- Verallgemeinerung der Markov-Kette der M/M/1-Warteschlange:



- Gleichungssystem für den Gleichgewichtszustand:

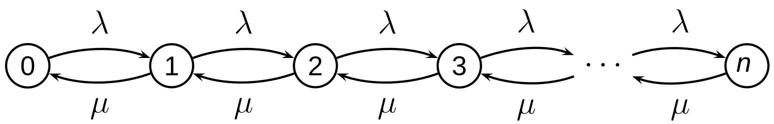
$$\begin{aligned} 0 &= \lambda_{k-1}\pi_{k-1} + \mu_{k+1}\pi_{k+1} - (\lambda_k + \mu_k)\pi_k \text{ für } k \geq 1, \\ 0 &= \mu_1\pi_1 - \lambda_0\pi_0 \end{aligned}$$

- Auflösen liefert $\pi_k = \pi_0 \cdot \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}$ für $k \geq 1$.
- Mit $\sum_{i=0}^{\infty} \pi_i = 1$ ergibt sich $\pi_0 = \frac{1}{1 + \sum_{k\geq 1} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}$.

Beispiel 1: Beschränkter Warteraum

M/M/1-Warteschlange mit nur *n* Plätzen

Neue Jobs werden abgewiesen, wenn bereits n Jobs im System sind.
Es ergibt sich der folgende Birth-and-Death Prozess:

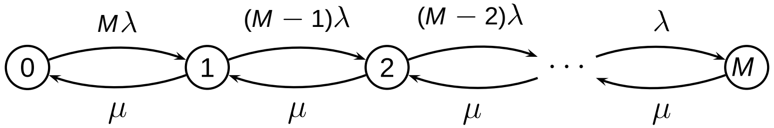


Wir erhalten:

$$\begin{aligned} \pi_k &= \rho^k \cdot \pi_0 \quad \text{für } 1 \leq k \leq n \\ \pi_0 &= \frac{1}{\sum_{i=0}^{n} \rho^i} = \begin{cases} \frac{1}{n+1} & \text{für } \rho = 1 \\ \frac{1-\rho}{1-\rho^{n+1}} & \text{sonst} \end{cases} \end{aligned}$$

Beispiel 2: Beschränkte Benutzerzahl

Anfragesystem mit M Terminals und einem Server



Wir erhalten:

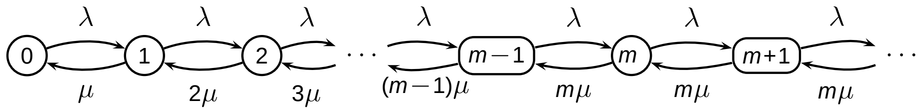
$$\pi_k = \pi_0 \cdot \prod_{i=0}^{k-1} \frac{\lambda(M-i)}{\mu} \quad \text{für } 1 \le k \le M$$

$$\pi_0 = \frac{1}{\sum_{k=0}^{M} \left(\frac{\lambda}{\mu}\right)^k \cdot M^{\underline{k}}}$$

wobei $M^{\underline{k}} := M(M-1)(M-2)\cdot \ldots \cdot (M-k+1)$.

Das M/M/m-System (1)

System mit einer Queue und m Servern (z.B. Call-Center)



Wir erhalten mit $\rho := \frac{\lambda}{m\mu} < 1$:

$$\pi_k = \begin{cases} \pi_0 \cdot \frac{\lambda^k}{\mu^k \cdot k!} = \pi_0 \cdot \frac{(\rho m)^k}{k!} & \text{für } 1 \le k \le m \\ \pi_0 \cdot \frac{\lambda^k}{\mu^k \cdot m! \cdot m^{k-m}} = \pi_0 \cdot \frac{\rho^k m^m}{m!} & \text{für } k \ge m \end{cases}$$

$$\pi_0 = \frac{1}{1 + \sum_{k=1}^{m-1} \frac{(\rho m)^k}{k!} + \sum_{k=m}^{\infty} \frac{\rho^k m^m}{m!}} = \frac{1}{\sum_{k=0}^{m-1} \frac{(\rho m)^k}{k!} + \frac{(\rho m)^m}{m!(1-\rho)}}$$

Das M/M/m-System (2)

Die Wahrscheinlichkeit P_Q, dass ein ankommender Job in der Warteschlange des M/M/m-Systems warten muss, ist also:

$$P_Q = \sum_{k=m}^{\infty} \pi_k = \sum_{k=m}^{\infty} \frac{\pi_0 \rho^k m^m}{m!}$$

$$= \frac{\pi_0 (\rho m)^m}{m!} \sum_{k=m}^{\infty} \rho^{k-m} = \frac{\pi_0(\rho m)^m}{m!(1-\rho)}$$

$$\Rightarrow \boxed{P_Q = \frac{(\rho m)^m / (m!(1-\rho))}{\sum_{k=0}^{m-1} \frac{(\rho m)^k}{k!} + \frac{(\rho m)^m}{m!(1-\rho)}} \quad \left(\text{für } \rho = \frac{\lambda}{m\mu} < 1\right)}$$

Diese Formel wird nach A.K. Erlang (1878-1929) die **Erlang C-Formel** genannt.

Das M/M/m-System (3)

Nun können wir aus P_Q weitere Grössen ableiten:

- Für N_Q (erwartete Anzahl von Jobs in der Warteschlange) erhalten wir:

$$N_Q = \sum_{k=1}^{\infty} k \cdot \pi_{m+k} = \sum_{k=1}^{\infty} k \cdot \pi_0 \cdot \frac{\rho^{m+k} m^m}{m!}$$

$$= \pi_0 \frac{\rho^m m^m}{m!} \sum_{k=1}^{\infty} k \rho^k = \pi_0 \frac{\rho^{m+1} m^m}{m!(1-\rho)^2}$$

$$= \frac{P_Q m!(1-\rho)}{\rho^m m^m} \cdot \frac{\rho^{m+1} m^m}{m!(1-\rho)^2} = P_Q \cdot \frac{\rho}{1-\rho}$$

Das M/M/m-System (4)

- Für W (mittlere Wartezeit in der Queue) erhalten wir mit der Formel von Little:

$$W = \frac{N_Q}{\lambda} = P_Q \cdot \frac{\rho}{\lambda(1-\rho)} = \frac{\rho P_Q}{\lambda(1-\rho)}$$

- Die mittlere Antwortzeit T ist dann:

$$T = W + \frac{1}{\mu} = \frac{\rho P_Q}{\lambda(1-\rho)} + \frac{1}{\mu} = \frac{P_Q}{m\mu - \lambda} + \frac{1}{\mu}$$

- Erneute Anwendung der Formel von Little liefert die mittlere Anzahl N von Jobs im System:

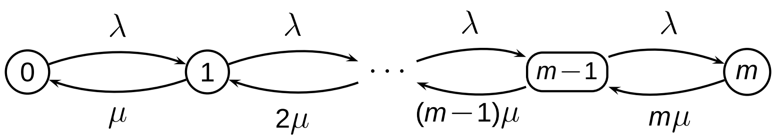
$$N = \lambda T = \frac{\lambda P_Q}{m\mu - \lambda} + \frac{\lambda}{\mu} = \frac{\rho P_Q}{1-\rho} + m\rho$$

Das M/M/m/m-System (1)

- System mit m Servern und maximal m Jobs im System.
- Jobs, die ankommen, wenn alle m Server belegt sind, werden abgewiesen.
- Klassisches Modell für Analyse von Leitungsvermittlung im Telefonnetz:
  - Ankunftsrate von Telefongesprächen λ.
  - Gesprächsdauern exponentialverteilt mit Parameter μ.
  - Kapazität für m gleichzeitige Telefongespräche.
  - Anzahl Benutzer ist viel grösser als m.

Das M/M/m/m-System (2)

Modellierung als Birth-and-Death Prozess:



Wir erhalten:

$$\pi_k = \pi_0 \cdot \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} \quad \text{für } 1 \leq k \leq m$$

Mit $\sum_{k=0}^{m} \pi_k = 1$ ergibt sich:

$$\pi_0 = \frac{1}{\sum_{k=0}^{m} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}}$$

Das M/M/m/m-System (3)

Die Blockierungswahrscheinlichkeit (W'keit, dass ein neu ankommender Job abgewiesen wird), ist damit:

$$\boxed{\pi_m = \frac{\left(\frac{\lambda}{\mu}\right)^m \frac{1}{m!}}{\sum_{k=0}^{m} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}}}$$

Diese Formel ist als **Erlang B-Formel** bekannt.

Bemerkung: Die Erlang B-Formel gilt auch für M/G/m/m-Systeme (d.h. wenn die Bearbeitungszeiten (Gesprächsdauern) Erwartungswert $1/\mu$ haben, aber ansonsten beliebig verteilt sind).

Warteschlangen-Netzwerke

- Warteschlangen-Netzwerke sind Graphen, bei denen die Knoten Warteschlangen-Systeme darstellen (z.B. M/M/1), und gerichtete Kanten die Jobs von einem Knoten zum nächsten führen.
- Man unterscheidet zwischen *offenen* und *geschlossenen* Warteschlangen-Netzwerken:
  - Offene Netzwerke erlauben, dass Jobs von aussen zum Netzwerk dazustossen oder das Netzwerk verlassen.
  - Bei geschlossenen Netzwerken sind die Jobs im Netzwerk gefangen; die Anzahl der Jobs im Netzwerk bleibt deshalb konstant.

Burke's Theorem

- Gegeben ein M/M/m ($m = 1, \dots, \infty$) System mit Ankunftsrate λ. Wir nehmen an, dass das System im stationären Zustand gestartet wird. Dann ist der Ausgangsprozess (der Prozess, der das System verlässt) auch ein Poisson-Prozess mit Rate λ.
- Dank Burke's Theorem kann man direkt Warteschlangen-Netzwerke analysieren.
- Allerdings muss man vereinfachend annehmen, dass die Servicezeit eines Jobs beim betreten jedes weiteren Warteschlangen-Systems wieder unabhängig ist.
- Wenn man diese vereinfachende Annahme nicht trifft, kann bisher schon ein einfaches Tandem-System (zwei M/M/1 Systeme in Serie) nicht analysiert werden.

Jackson's Theorem für offene Netze

- Jobs kommen bei Knoten j als Poisson-Prozess mit Rate r_j von aussen an.
- Jobs verlassen Knoten i mit Wahrscheinlichkeit p_{ij} Richtung Knoten j, oder verlassen das Netzwerk mit Wahrscheinlichkeit $p_{i,exit}$, wobei $p_{i,exit} + \sum_{\forall j} p_{ij} = 1$.
- Dann ist der gesamte Ankunftsprozess bei Knoten j gegeben durch:

$$\lambda_j = r_j + \sum_{\forall i} \lambda_i p_{ij}$$

- Die Lösung dieses linearen Gleichungssystems ergibt direkt die gesamten Ankunftsraten λ_j.
- Geschlossene Netze sind etwas komplexer...

Simulation

- Kompliziertere und realistischere Warteschlangen-Systeme und -Netzwerke werden in der Regel simuliert. Eine vereinfachte Analyse (z.B. M/M/1) kann aber schon einen ersten Eindruck vermitteln.
- Zum Thema Simulation verweise ich auf die Vorlesung Informatik 2 von Prof. Mattern (Skript ab Seite 225).
- "And now for something completely different..."