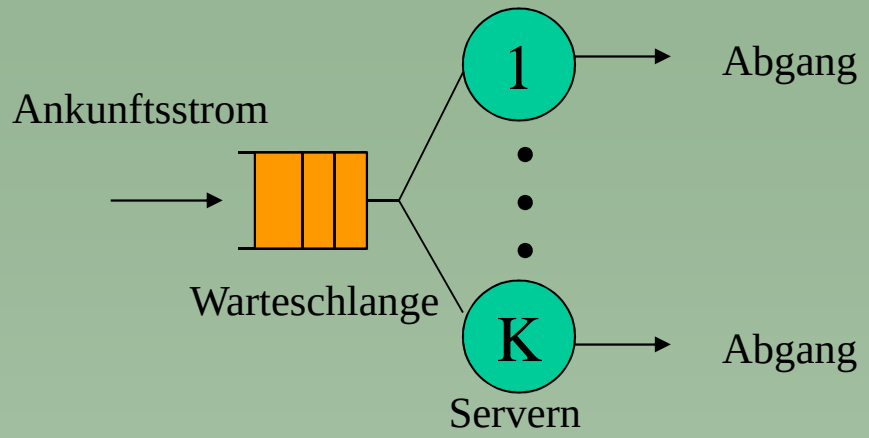
Bedienungstheorie

- Kapitel 1 Einleitung
- Kapitel 2 Einfache Markov-Modelle
- Kapitel 3 Markov-Modelle II
- Kapitel 4 Netzwerke

SS 2005

Kapitel 1. Einleitung

Bild 1



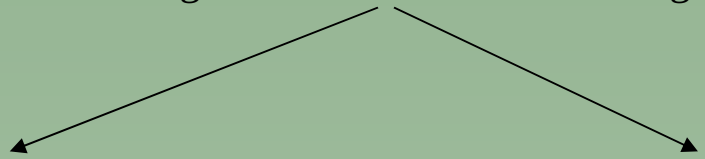
Unter **Bedienungstheorie** versteht man die „Mathematik der Warteschlangen“, auf English ***queueing theory***. Ein Bedienungssystem besteht immer aus ein oder mehreren Bedienungskanälen (Servern), die von den eintreffenden Kunden in Anspruch genommen werden. Aus den in der Regel zufälligen Ankunfts-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

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**
- **Wartedisziplin: FCFS, LCFS, RSS, FFS, Prioritäten**
- **Einzel- oder Parallelservice**
- **Bedienungszeiten**

Dabei bedeutet:

- **FCFS –** für ***first come first served***,
- **LCFS –** für ***last come first served***,
- **RSS** – für ***random selection for service***
- **FFS** – für ***fastest free server selection***

Einfache Bedienungssysteme werden durch eine Kurzschreibweise **A/B/X/Y/Z** charakterisiert (**Kendallsche Notation**):
- **A** beschreibt die Verteilung der unabhängigen Zwischenankunftszeiten
- **B** beschreibt die Verteilung der unabhängigen Bedienungszeiten

Dabei bedeutet:
- **M** Exponentialverteilung (Markov Eigenschaft),
- **D** deterministisch,
- E_k Erlangverteilung
- H_k Mischung von Exponentialverteilungen,
- G beliebige Verteilung

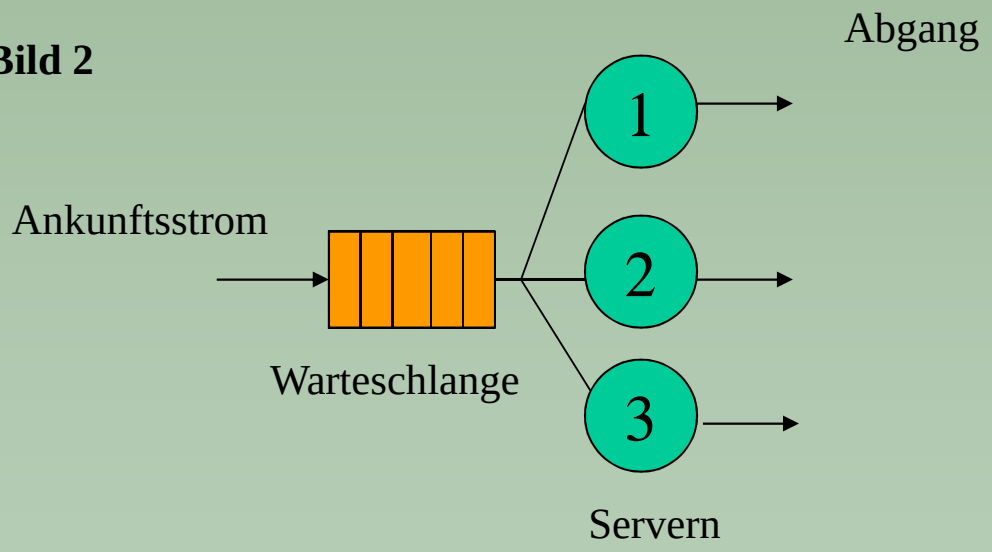
- $X \in \mathbf{N} \cup \{\infty\}$ beschreibt die Anzahl der parallelen Server
- $Y \in \mathbf{N} \cup \{\infty\}$ beschreibt die Systemkapazität,
- Z steht für die Wartedisziplin.

Die Standardannahmen Y=∞ und Z=FCFS werden nicht angeschrieben.

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**, d.h. ein Kunde, der bereits 5 Kunden im System vorfindet, geht verloren. Die Kunden werden in der Reihenfolge ihres Eintreffens bedient.

Bild 2



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.

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 beschreiben. 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.

Bezeichnungen

Man bezeichnet mit

λ die **mittlere Anzahl** der je Zeiteinheit eintreffenden Kunden

μ die **Bedienungsrate**, das ist die mittlere Anzahl der je Zeiteinheiten von einem Bedienungskanal fertig bedienten Kunden, wenn dieser immer arbeitet

$\rho := \lambda/(c\,\mu)$ wird als **Auslastungsgrad** 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.

Falls $\rho < 1$ gilt,
kann ein Gleichgewicht entstehen. Dann wird nach langer Zeit der Zustand des Systems nicht mehr vom absoluten Wert des Zeitparameters abhängen.

Falls $\rho = 1$ gilt,
Ist ein Gleichgewicht nur bei deterministischen Zwischenankunfts- und Bedienungszeiten möglich, denn Bedienungsrückstände, wie sie etwa durch ein zufälliges Leerstehen entstehen, können nie aufgeholt werden.

Beispiel

Für eine Warteschlange mit *c* Server gegeben ist eine mittlere Ankunftsrate λ und eine mittlere Bedienungsrate μ. Wie groß muss *c* mindestens gewählt werden, damit ein Gleichgewicht möglich ist?

Antwort: $c = \left[\frac{\lambda}{\mu} \right] + 1$

Man bezeichnet 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.

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

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.

Man bezeichnet mit

N_x die Gleichgewichtsanzahl von Kunden im Zustand x,

T_x die Zeit, die ein Kunde in x verbringt, wenn sich das System im Gleichgewicht befindet.

$L = \mathrm{E}[N_x]$ die mittleren Anzahlen von Kunden

$W = \mathrm{E}[T_x]$ die mittleren Aufenthaltszeiten der Kunden

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

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 $L\tau = \mathbf{E}[N]\tau$ Zeiteinheiten im System verbracht.

Notieren wir nun bei jedem im Zeitintervall **[0,τ[** eintretenden Kunden **k** (k=1,...,K(τ), wobei K(τ) – der letzte Kunde ist, der vor τ eintrifft) sofort seine Systemzeit $T^{(k)}$, so ergibt sich für das Zeitintervall **[0,τ[** 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

So gilt $$L\,\tau + \mathbf{E}[\tilde{T}_\tau] = \lambda\tau\, W + \mathbf{E}[\tilde{T}_0]$$

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

$$L + \frac{\mathbf{E}[\tilde{T}_\tau]}{\tau} = \lambda\, W + \frac{\mathbf{E}[\tilde{T}_0]}{\tau}, \quad \tau \to \infty \Rightarrow$$

$$L = \lambda\, W$$

(John D. C. Little, 1961)

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

$L_q = \lambda\, W_q$ und $L_s = \lambda\, W_s$

L_q ist die erwartete Länge der Warteschlange und

L_s ist 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 $\rho = L_s / c$ entspricht der erwartete Anzahl von Kunden in einem vorgegebenen Server, also die Wahrscheinlichkeit, dass dieser Server arbeitet. Daher kommt die Bezeichnung Auslastungsgrad für ρ

Kapitel 2. Einfache Markov-Modelle

Ein **stochastischer Prozess** ist eine Familie von Zufallsvariablen $(Z(t))_{t\in I}$ mit einer gemeinsamen Bildmenge 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 $\mathbf{I} \subseteq \mathbf{R}$,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-Prozesses modelliert werden. Ein Poisson-Prozess ist ein stochastischer Prozess $(A(t))_{t\geq 0}$ mit Zustandsraum $\mathbf{N}_0, A(0)=0$ und den Eigenschaften:

1. Es gilt **Pr[*A*(t+Δt)-*A*(t)=1]=λ Δt+*o*(Δt),** 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+Δt)-*A*(t)=0]=1-λ Δt+*o*(Δt),**
3. Für a ≤ b ≤c ≤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.

Verteilung der Zuwächse

Für t≥s nennt man A(t)-A(s) den Zuwachs im Intervall]s,t]. Für n∈ **Z** bezeichnen wir mit

$$p_n^{(s)}(t) = \Pr[\, A(t) - A(s) = n]$$

wobei aus den Eigenschaft des Poisson-Prozesses für n<0 folgt $p_n^{(s)}(t) = 0$

Damit gilt nun

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

Wegen 3. gilt

$$p_n^{(s)}(t+\Delta t) = \Pr[A(t)-A(s)=n]\Pr[A(t+\Delta t)-A(t)=0] + \Pr[A(t)-A(s)=n-1]\Pr[A(t+\Delta t)-A(t)=1] + \Pr[A(t)-A(s)=n-2]\Pr[A(t+\Delta t)-A(t)=2]+\ldots$$

was wegen 1. und 2.

$$p_n^{(s)}(t+\Delta t) = p_n^{(s)}(t)(1-\lambda\Delta t+o(\Delta t)) + p_{n-1}^{(s)}(t)(\lambda\Delta t+o(\Delta t)) + p_{n-2}^{s}(t)(o(\Delta t)+o(\Delta t))$$

ergibt: Division durch **Δt->0** führt auf die Differentialgleichungen für **t>s**

$$\frac{dp_0^{(s)}(t)}{dt} = -\lambda p_0^{(s)}(t), \quad n=0$$

$$\frac{dp_n^{(s)}(t)}{dt} = -\lambda p_n^{(s)}(t)+\lambda p_{n-1}^{(s)}(t), \quad \textit{für } n\in \mathbf{N}$$

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 **λ(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 **λ t**

Verteilung der Abstände zwischen aufeinanderfolgenden Sprüngen

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

$$\Pr[T^{(1)} \geq t] = \Pr[A(t) = 0] = e^{-\lambda t}$$

d.h. $T^{(1)}$ 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 $(dE(\tau))_{\tau \in R^+}$ mit

$$dE(\tau) := (A(\tau) = k, A(\tau) - A(\tau - d\tau) = 1)$$

ein vollständiges Ereignissystem ist.

Wegen der **Unabhängigkeit der Zuwächse** gilt

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

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

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

d.h. auch alle $T^{(k+1)}$ sind exponentialverteilt mit Parameter λ. Weiters sind die $(T^{(k)})_{k \in \mathbf{N}}$ vollständig unabhängig.

Es gibt auch die Umkehrung: Sind die Zeitabstände zwischen aufeinanderfolgenden Ereignissen $(T^{(k)})_{k\in \mathbf{N}}$ **vollständig unabhängig** und **exponentialverteilt** mit demselben Parameter λ, und bezeichnet *A*(*t*) die Anzahl der Ereignisse vor dem Zeitpunkt t, so ist $(A(t))_{t\geq 0}$ ein **Poisson-Prozess mit Parameter λ**

Für einen **Poisson-Prozess** $(A(t))_{t\geq 0}$ gilt außerdem: Wenn man weiß, dass in einem Intervall *n* Ereignisse eintreten, so sind die Zeitpunkte des Eintretens vollständig unabhängig und gleichverteilt.

Markov-Eigenschaft der Exponentialverteilung

Ist *T* exponentialverteilt, d.h.

$$\Pr[T < t] = 1 - e^{-\lambda t}$$

so gilt für $s, t \geq 0$ die sogenannte **Nichtalterungseigenschaft** (Markov-Eigenschaft, Markov Andrei, 1856-1922, russischer Mathematiker)

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

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

Verallgemeinerung

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 $n \in \mathbf{N}$

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

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

2.2 Markov-Ketten

Definition. Ein stochastischer Prozess $(Z(t))_{t\geq 0}$ heißt Markov-Prozess, wenn für alle $n \in \mathbf{N}$ alle $0 < t_1 < t_2 < ... < t_{n+1}$ $(t_i \in I)$ und für alle $x_1,..., x_{n+1} \in I$ mit $\Pr[Z_{t_n} = x_n,..., Z_{t_1} = x_1] > 0$ gilt

$$\Pr[\underbrace{Z_{t_{n+1}} = x_{n+1}}_{\text{Zukunft}} \mid \underbrace{Z_{t_n} = x_n}_{\text{Gegenwart}}, \underbrace{..., Z_{t_1} = x_1}_{\text{Vergangenh eit}}] = \Pr[\underbrace{Z_{t_{n+1}} = x_{n+1}}_{\text{Zukunft}} \mid \underbrace{Z_{t_n} = x_n}_{\text{Gegenwart}}]$$

d.h. wenn, bildlich gesprochen, die Zukunft nur von der Gegenwart, nicht aber von der Vergangenheit abhängt.
Obige Eigenschaft nennt man **Markov-Eigenschaft** (**Gedächtnislosigkeit**) des stochastischen Prozesses.

Definition. Ein Markov-Prozess, bei dem entweder der Zustandsraum oder die Indexmenge diskret sind, heißt **Markov-Kette**.

Definition. Hat eine Markov-Kette einen endlichen Zustandsraum, so heißt sie **endlich**, sonst **abzählbar**.

Definition. Ein **Semi-Markov-Prozess** (**Markovscher Erneuerungsprozess**) ist ein stochastischer Prozess mit diskretem Zustandsraum, bei dem die Folge verschiedenen Zustände eine Markovkette mit diskreter Zeit bildet, die Verteilung der Zeitabstände zwischen aufeinanderfolgenden Zustandsänderungen aber beliebig ist.

Markov-Ketten mit diskreter Zeit

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

Definition. Der Ausdruck

$$\Pr[Z_{n+1} = j \mid Z_n = i, Z_{n-1} = i_{n-1}, \ldots, Z_0 = i_0] = \Pr[Z_{n+1} = j \mid Z_n = i] =: p_{ij}$$

heißt **Übergangswahrscheinlichkeit** von *i* nach *j*.

Definition. Falls die bedingte Wahrscheinlichkeit von *n* nicht abhängt, d.h.

$$\Pr[Z_{n+1} = j \mid Z_n = i] = \Pr[Z_1 = j \mid Z_0 = i]$$

$n \in \mathbf{N}_0$ dann die Kette heißt **homogen**.

Definition. Die unendliche Matrix der Übergangswahrscheinlichkeiten

heißt **Übergangsmatrix**.

$$\mathbf{P} := (p_{ij}) = \begin{pmatrix} p_{11} & p_{12} & \cdots \\ p_{21} & p_{22} & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix}$$

Definition. Für jede Markov-Kette gilt

$$\sum_{j \in I} p_{ij} = 1, \; i \in I$$

dann heißt **P stochastische Matrix**.

Beispiel 1. Gegeben sei eine Folge von unabhängigen identisch verteilten Zufallsvariablen

$(Z_n)_{n \in \mathbf{N}_0}$ dann diese Kette ist eine homogene Markov-Kette.

Beispiel 2(Bernoulli Prozess). Sei $I := \mathbf{N}_0$ und $(Z_n)_{n \in \mathbf{N}_0}$ ein stochastischer Prozess. Wählt man ein Parameter $0<p<1$. Der Ausdruck $\Pr[Z_0 = 0] := 1$ zusammen mit

$$\Pr[Z_{n+1} = j \mid Z_n = i] := \begin{cases} p, & j = i+1 \\ 1-p, & j = i \end{cases}$$

für $i \in \mathbf{N}_0$ bilden ein W-Maß **P** auf einem Ereignisraum Ω und $(Z_n)_{n \in \mathbf{N}_0}$ ist eine homogene Markov-Kette und heißt **Bernoulli Prozess** mit Parameter *p*.

Beispiel 3. Ein Mann fährt entweder mit dem Auto oder mit dem Zug zur Arbeit. Er fährt niemals zweimal hintereinander mit dem Zug; fährt er an einem Tag mit dem Auto, so ist es gleichwahrscheinlich, daß er am nächsten Tag das Auto bzw. Zug nimmt.

Der Zustandsraum des Systems ist {Z,A}, Z=Zug, A=Auto. Dieser stochastische Prozess ist offenbar eine Markov-Kette, da jedes Ergebnis nur vom unmittelbar davorliegenden abhängt. Die Übergangsmatrix dieser Markov-Kette ist

$$\begin{array}{c} \\ Z \\ A \end{array} \begin{array}{c} \begin{array}{cc} Z & A \end{array} \\ \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix} \end{array}$$

In der ersten Zeile der Matrix ist berücksichtigt, daß der Mann, falls er einen Tag mit dem Zug gefahren ist, am nächsten sicher mit dem Auto fährt. In der zweiten Zeile steht, daß er nach einem Tag mit dem Auto am nächsten mit jeweils gleicher Wahrscheinlichkeit das Auto oder den Zug nimmt.

Beispiel 4. In einer Schule sind 200 Jungen und 150 Mädchen. Zu einer Augenuntersuchung wird ein Schüler nach dem anderen zufällig ausgewählt. Es sei Z_n das Geschlecht der *n*-ten untersuchten Person. Der Zustandsraum dieses stochastischen Prozesses ist {m, w}, m=männlich, w=weiblich.
Eine Markov-Kette liegt jedoch nicht vor, denn das Ereignis eines Versuches hängt von den Ereignissen aller vorhergehenden Versuche ab.

Satz. Für eine homogene Markov-Kette $(Z_n)_{n\in \mathbf{N}_0}$ mit Übergangsmatrix **P** gilt

$$\Pr[X_{n+1} = j_1,..., X_{n+m} = j_m \mid X_n = i] = p_{ij_1} p_{j_1 j_2} ... p_{j_{m-1} j_m}$$

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

bezeichnet die **Zustandswahrscheinlichkeit** von *j* zum Zeitpunkt *n*.

Der Zeilevektor $(\pi_j^{(n)}) =: \pi^{(n)}$ heißt **Zustandsverteilung** zum Zeitpunkt *n*

.

$(\pi_j^{(0)}) = \pi^{(0)}$ heißt **Anfangsverteilung**

Bemerkung. Falls $\Pr[Z_0 = j] =: \pi_j^{(0)}$ dann aus dem vorhergehenden Satz gilt

$$\Pr[Z_0 = j_0, Z_1 = j_1,..., Z_m = j_m] = \pi_{j_0}^{(0)} p_{j_0 j_1} p_{j_1 j_2} ... p_{j_{m-1} j_m}$$

Satz. Es sei **P** die Übergangsmatrix einer Markov-Kette $(Z_n)_{n\in \mathbf{N}_0}$ Dann gilt

$$\Pr[X_{n+m} = j \mid X_n = i] = \Pr^m[X_{n+1} = j \mid X_n = i] \text{ oder } \mathrm{P}^{(m)} = \mathrm{P}^m$$

Satz. Für Markov-Ketten gelten die ***Chapman-Kolmogorov-Gleichungen***

$$\Pr[X_{n+m} = j \mid X_0 = i] = \sum_{k \in I} \Pr[X_m = k \mid X_0 = i] \Pr[X_n = j \mid X_0 = k]$$

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

wobei $\mathbf{P}^m$ die Matrix der ***m*-Schritt-Übergangswahrscheinlichkeit** mit den Einträgen

$$p_{ij}^{(m)} = \Pr[X_m = j \mid X_0 = i]$$

Satz. Für die Zustandsverteilung folgt daraus

$$\pi_i^{(m)} = \sum_{j \in I} \pi_j^{(0)} p_{ji}^{(m)}, \; i \in I$$

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

Definition. Eine stochastische Matrix $\mathbf{P}$ heißt ***regulär***, wenn alle Elemente eine Potenz $\mathbf{P}^m$ von $\mathbf{P}$ positiv sind und nicht null sind.

Beispiel 5. Die stochastische Matrix $A = \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}$ ist ***regulär***, da zumindest eine Potenz existiert, deren Elemente positive und nicht null sind

$$A^2 = \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{pmatrix}$$

Beispiel 6. Ein Psychologe untersucht das Verhalten von Mäusen, die einem bestimmten Fütterungsschema unterzogen werden, und macht folgende Beobachtungen: Bei einem bestimmten Versuch gingen 80% der Mäuse, die im vorhergehenden Experiment nach rechts gingen, wieder nach rechts; 60% der Mäuse, die zuerst nach links gingen, gehen nach rechts. Im ersten Versuch ging die Hälfte der Mäuse nach rechts. Was ist zu erwarten für den (i) zweiten und (ii) dritten Versuch?

Lösung. Die Zustände des Systems sind R (rechts) und L (links), und als Übergangsmatrix erhalten wir

$$\mathbf{P} = \begin{array}{c} \\ R \\ L \end{array}\begin{array}{c} \begin{array}{cc} R & L \end{array} \\ \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix} \end{array}$$

Als Anfangsverteilung haben wir p=(0.5,0.5). Um die Verteilung im zweiten Versuch zu erhalten, müssen wir p und **P** multiplizieren:

$$(0.5,0.5)\begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix} = (0.7,0.3)$$

Also gehen 70% der Mäuse nach rechts und 30% nach links. Das Ergebnis des dritten Versuchs erhalten wir, indem wir das des zweiten mit **P** multiplizieren:

$$(0.7,0.3)\begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix} = (0.74,0.26)$$

Also gehen im dritten Versuch 74% der Mäuse nach rechts und 26% nach links.

Langzeitverhalten

Drei Größen kennzeichnen das Langzeitverhalten einer Markov-Kette, ihre Grenzverteilung, ihre stationären Verteilungen und die Eigenschaft der Ergodizität.

Definition. 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.

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

d.h. 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 $\pi\mathbf{P} = \pi$

und weil $\boldsymbol{\pi}$ eine Verteilung ist, auch $\pi\mathbf{1} = 1$

Definition. Jede Verteilung, die die Beziehung $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ erfüllt, heißt **stationäre Verteilung** oder **Gleichgewichtsverteilung** der Markov-Kette.

Bemerkung. Ist die Markov-Kette endlich, so ist $\boldsymbol{\pi}$ ein normierter Linkseigenvektor von P zum Eigenwert 1. Für endliche Markov-Ketten existiert also immer eine stationäre Verteilung.

Beispiel 7. (Beispiel 6) Was ist zu erwarten für den tausendsten Versuch?

Lösung. Wir nehmen an, daß die Verteilung im tausendsten Versuch gleich der stationären Verteilung der Markov-Kette ist. Also aus dem Gleichungssystem

$$\begin{cases} 0.8\pi_1 + 0.6\pi_2 = \pi_1 \\ 0.2\pi_1 + 0.4\pi_2 = \pi_2 \\ \pi_1 + \pi_2 = 1 \end{cases}$$

erhält man also $\boldsymbol{\pi}$**={0.75,0.25},** d.h. in diesem Versuch gehen 75% der Mäuse nach rechts und 25% nach links.

Markov-Ketten mit kontinuierlicher Zeit

Definition. 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

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

als **Übergangswahrscheinlichkeit** von i nach j im Zeitintervall [s,t].

Wir schreiben die Übergangswahrscheinlichkeiten wieder in **Übergangsmatrizen**

$\mathbf{P}(s,t) := (p_{ij}(s,t))$ zusammen.

Satz. Die **Chapman-Kolmogorov-Gleichung** lauten nun ($r \leq s \leq t$)

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

Definition. Eine Markov-Kette mit kontinuierlicher Zeit $(X_t)_{t\geq 0}$ heißt **homogen**, wenn

$$p_{ij}(r, r+s) = \Pr[X_{r+s} = j \mid X_r = i] = \Pr[X_s = j \mid X_0 = i] = p_{ij}(s)$$