

## **7. Markov-Prozesse und Warteschlangen**

**Diskrete Stochastik**

Prof. Dr. Andreas Vogt



# stochastische Prozesse

Ein stochastischer Prozess beschreibt die Zustände eines zufallsbeeinflussten Systems in aufeinanderfolgenden Zeitpunkten.

**Definition:** Es sei  $(\Omega, P)$  ein Wahrscheinlichkeitsraum.

Für jeden Zeitpunkt  $t \in T \subseteq \mathbb{R}$  beschreibe eine Zufallsvariable  $X_t : \Omega \rightarrow I$  den Zustand eines Systems zum Zeitpunkt  $t$ .

Dann heisst  $(X_t, t \in T)$  oder kurz  $(X_t)$  **stochastischer Prozess mit Zustandsraum  $I$ .**

Wir beschränken uns hier auf Prozesse "mit diskreter Zeit", also  $T = \{0, 1, 2, \dots\}$  und diskretem Zustandsraum, also  $I$  ist eine endliche oder abzählbare Menge.

z.B.  $\mathbb{N}$  oder  $\mathbb{Z}$

Bei vielen Systemen hängt der Folgezustand nur vom aktuellen Zustand ab, und nicht noch von den Zuständen davor. (Oft ist dies zumindest eine gute Approximation.)

Solche Prozesse nennt man **Markov-Ketten**.

**Definition:** Ein stochastischer Prozess  $(X_n, n \in \mathbb{N})$  heisst **Markov-Kette**, falls

$$P(X_n = i_n | X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(X_n = i_n | X_{n-1} = i_{n-1})$$

für alle  $n \in \mathbb{N}$  und alle  $i_0, \dots, i_n \in I$  (bei denen beide Seiten der Gleichung definiert sind, also die Wahrscheinlichkeit der Bedingung von 0 verschieden ist).

Die Wahrscheinlichkeit, dass der Prozess im Zeitpunkt  $n$  im Zustand  $i_n$  ist, wenn man die gesamte Historie des Prozesses kennt, also wenn man kennt, in welchen Zuständen der Prozess in den Zeitpunkten  $0, 1, \dots, n-1$  war, ist genauso gross, wie die Wahrscheinlichkeit, dass der Prozess im Zeitpunkt  $n$  im Zustand  $i_n$  ist, wenn man nur weiss, wo er sich zum Zeitpunkt  $n-1$  befindet.

# stochastische Prozesse

---

## Beispiel:

Wenn ihr Server ohne Probleme läuft, dann hat er mit Wahrscheinlichkeit 0.9 auch am nächsten Tag kein Problem.

Wenn der Server ein Problem hat, dann hat er mit Wahrscheinlichkeit 0.5 am nächsten Tag immer noch ein Problem.

$X_n$  beschreibe den Zustand am Tag  $n$ . Dann ist  $(X_n)$  eine Markov-Kette mit Zustandsraum  $I = \{\text{kein Problem, Problem}\} \stackrel{\wedge}{=} \{1, 2\}$ , wobei wir 1 mit "kein Problem" und 2 mit "Problem" identifizieren.

Die Markov-Eigenschaft ist erfüllt, da gemäss des Modells der Zustand des Servers am nächsten Tag nur von heute und eben nicht auch noch von gestern abhängt.

Im obigen Beispiel ist die Wahrscheinlichkeit, dass sich der Prozess am nächsten Tag in einem gewissen Zustand befindet, nicht nur unabhängig von den bisherigen Zuständen (bis auf heute), sondern auch noch unabhängig vom Tag selber.

(Dies wäre etwa nicht erfüllt, wenn die Wahrscheinlichkeit, dass an einem Sonntag der Server noch Probleme hat, wenn am vorigen Samstag ein Problem hat, z.B. 0.8 betragen würde.)

Solche Markov-Ketten nennt man **homogen**.

# stochastische Prozesse

Wir beschränken uns im Folgenden auf homogene Markov-Ketten (**HMK**):

## Definition:

Eine Markov-Kette  $(X_n)$  heisst **homogen**, wenn die Übergangswahrscheinlichkeiten

$$P(X_n = j | X_{n-1} = i), \quad (i, j \in I)$$

für alle  $n$  gleich sind.

In diesem Fall heisst die Matrix  $\mathbf{P} = (p_{ij})$  mit

$$p_{ij} = P(X_n = j | X_{n-1} = i), \quad (i, j \in I)$$

In Zeile  $i$  und Spalte  $j$  steht also die Wahrscheinlichkeit, dass man aus Zustand  $i$  in Zustand  $j$  wechselt.

## Übergangsmatrix (Ü-Matrix).

### Beispiel:

Wenn ihr Server ohne Probleme läuft, dann hat er mit Wahrscheinlichkeit 0.9 auch am nächsten Tag kein Problem.  
Wenn der Server ein Problem hat, dann hat er mit Wahrscheinlichkeit 0.5 am nächsten Tag immer noch ein Problem.

### Beispiel:

Im Beispiel von vorhin sieht die Übergangsmatrix so aus:

$$\begin{matrix} & \text{keine Probleme} & \text{Probleme} \\ \text{keine Probleme} & (0.9 & 0.1) \\ \text{Probleme} & (0.5 & 0.5) \end{matrix}$$

### Bemerkung:

Die Summe jeder Zeile ist 1: Die Summe der Elemente der  $i$ -ten Zeile ist die Summe aller Wahrscheinlichkeiten, mit denen sich der Prozess aus dem Zustand  $i$  in einen beliebigen Zustand bewegt, und diese Wahrscheinlichkeit ist 1.

# stochastische Prozesse

---

Entscheiden Sie, welche der folgenden stochastischen Prozesse eine Markovkette ist und wenn ja, ob sie homogen ist oder nicht.

- (1) Wir würfeln jeden Tag mit einem Würfel.  $X_n$  bezeichne die Summe aller Würfelwürfe bis einschliesslich Tag  $n$ .

Markov, homogen

- (2) Wir würfeln jeden Tag mit einem Würfel und sonntags sogar zweimal.  $Y_n$  bezeichne die Summe aller Würfelwürfe bis einschliesslich Tag  $n$ .

Markov, nicht homogen

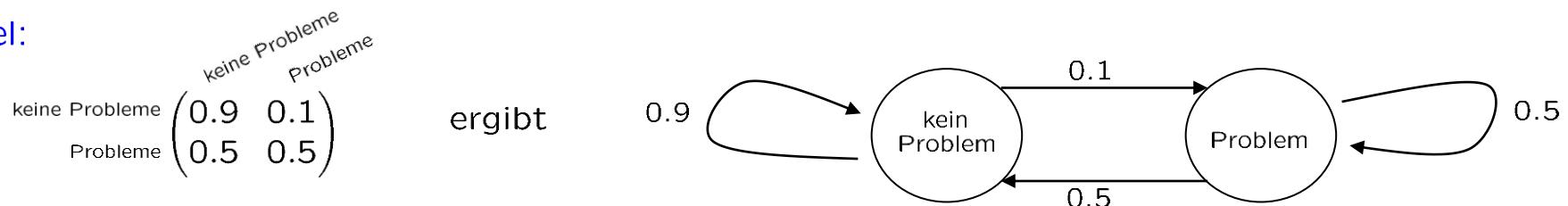
- (3) Wir würfeln jeden Tag mit einem Würfel. Wenn wir dabei an zwei aufeinanderfolgenden Tagen eine 6 würfeln, dann würfeln wir am nächsten Tag nicht.  $Z_n$  bezeichne die Summe aller Würfelwürfe bis einschliesslich Tag  $n$ .

keine Markovkette

# stochastische Prozesse

Markov-Ketten kann man oft übersichtlich mit Hilfe eines **Übergangsgraphen** darstellen, bei dem die Knoten den Zuständen entsprechen, sowie Pfeile mit den entsprechenden Übergangswahrscheinlichkeiten zwischen Zuständen gezeichnet werden.

**Beispiel:**



Am Tag  $n$  habe der Server kein Problem.

Mit welcher Wahrscheinlichkeit hat er am Tag  $n+2$  ein Problem?

$$0.9 \xrightarrow{0.1} 0.1 + 0.1 \xrightarrow{0.5} 0.5$$

Wir berechnen allgemein  $P(X_{n+2} = j | X_n = i)$ .

$$\begin{aligned}
 P(X_{n+2} = j | X_n = i) &= \sum_{k \in I} P(X_{n+2} = j, X_{n+1} = k | X_n = i) = \sum_{k \in I} \frac{P(X_{n+2} = j, X_{n+1} = k, X_n = i)}{P(X_n = i)} = \\
 &= \sum_{k \in I} \frac{P(X_{n+2} = j, X_{n+1} = k, X_n = i)}{P(X_{n+1} = k, X_n = i)} \cdot \frac{P(X_{n+1} = k, X_n = i)}{P(X_n = i)} \\
 &= \sum_{k \in I} P(X_{n+2} = j | X_{n+1} = k, X_n = i) \cdot P(X_{n+1} = k | X_n = i) \\
 &= \sum_{k \in I} P(X_{n+2} = j | X_{n+1} = k) \cdot P(X_{n+1} = k | X_n = i) = \boxed{\sum_{k \in I} p_{kj} \cdot p_{ik}}
 \end{aligned}$$

Wenn  $P$  die Übergangsmatrix bezeichnet, dann ist dies also gerade der  $ij$ -te Eintrag der Matrix  $P^2$ .

# stochastische Prozesse

---

Allgemein erhalten wir für eine beliebige HMK mit Übergangsmatrix  $P$ :

$P(X_{n+m} = j \mid X_n = i)$  ist der  $ij$ -te Eintrag der Matrix  $P^m$ .

Das Verhalten einer Markov-Kette ist also eindeutig durch die Übergangsmatrix  $P$  bestimmt, sofern der Startwert bekannt ist.

Nicht immer ist der Startwert eine feste Zahl, er kann auch zufällig sein.

In unserem Server-Beispiel könnte es etwa so sein, dass man plant, einen solchen Server zum neuen Jahr zu kaufen, wobei es so sein kann, dass ein Prozentsatz von z.B. 1% aller Server gleich zu Beginn auf Grund eines Herstellungsfehlers Probleme hat.

Man hat also eine **Startverteilung** gegeben, d.h.  $P(X_0 = i)$  für alle  $i \in I$ .

(Einen festen Startwert  $X_0 = s$  kann man damit durch  $P(X_0 = s) = 1$  und  $P(X_0 = i) = 0$  für  $i \neq s$  modellieren.)

Den Vektor bestehend aus den Einträgen  $P(X_0 = i)$  für  $i \in I$  bezeichnet man (meist) mit  $\pi_0$ .

Bei gegebenem  $\pi_0$  ist also:

$$P(X_m = j) = \sum_{k \in I} P(X_m = j \mid X_0 = k) \cdot P(X_0 = k) = \sum_{k \in I} (P^m)_{kj} \cdot \pi_0^{(k)},$$

wobei  $\pi_0^{(k)}$  die  $k$ -te Stelle von  $\pi_0$  bezeichnet.

Mit  $\pi_n$  bezeichnet man den Vektor mit den Einträgen  $P(X_n = i)$  für  $i \in I$ .

Damit gilt:  $\pi_n = \pi_0 \cdot P^n$

# stochastische Prozesse

Beispiel:

In unserem Server-Beispiel planen wir, einen solchen Server zum neuen Jahr zu kaufen, wobei erfahrungsgemäss 1% aller Server gleich zu Beginn auf Grund eines Herstellungsfehlers Probleme haben.

		keine Probleme	Probleme
keine Probleme	(0.9 0.1)		
Probleme	(0.5 0.5)		

Wir schauen uns die Wahrscheinlichkeit dafür an, dass der Server am 1., 2., 3, ... Tag Probleme hat.

Für die Startverteilung  $\pi_0$  gilt  $\pi_0 = (0.99, 0.01)$ .

Für die Verteilung  $\pi_1$  nach dem ersten Tag gilt:

$$\pi_1 = \pi_0 \cdot P = (0.99 \cdot 0.9 + 0.01 \cdot 0.5, 0.99 \cdot 0.1 + 0.01 \cdot 0.5) = (0.896, 0.104)$$

Mit welcher Wahrscheinlichkeit hat der Server am 4. Januar Probleme? Und am 31. Dezember?

```
>> pi0=[0.99,0.01]; P=[0.9 0.1;0.5 0.5]; pi0*P^3, pi0*P^364
ans =
0.8434  [0.1566] ←
ans =
0.8333  [0.1667] ←
```

Die Frage, ob der Server an einem konkreten Tag Probleme hat, ist eigentlich gar nicht so wichtig.

Viel wichtiger ist die Frage, an wie vielen Tagen der Server langfristig Probleme hat.

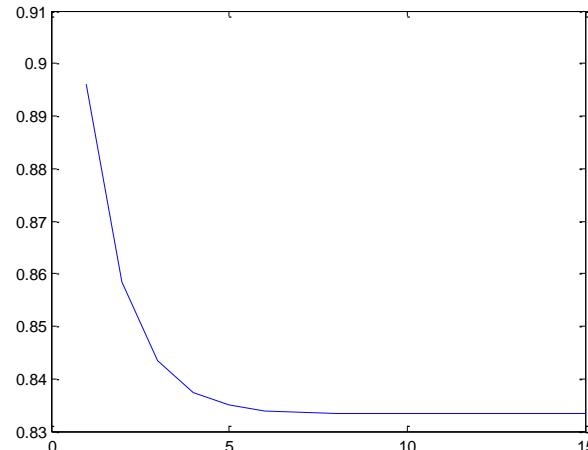
# stochastische Prozesse

Viel wichtiger ist die Frage, an wie vielen Tagen der Server langfristig Probleme hat.

Dazu schauen wir uns mal das Verhalten von  $\pi_n$  über die Zeit an.

$$\begin{array}{cc} \text{keine Probleme} & \text{Probleme} \\ \text{keine Probleme} & \begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix} \\ \text{Probleme} & \end{array}$$

```
i=1:15;
x=zeros(15,1);
pi0=[0.99,0.01];
P=[0.9 0.1;0.5 0.5];
for k=1:15
    a=pi0*P^k;
    x(k)=a(1);
end
plot(i,x);
```



In unserem Server-Beispiel gilt also  $\lim_{n \rightarrow \infty} \pi_n = (0.833, 0.1667)$ .

In der Tat: Wenn  $\pi_n = (0.833, 0.1667) = \left(\frac{5}{6}, \frac{1}{6}\right)$ , dann gilt

$$\pi_{n+1} = \pi_0 \cdot P^{n+1} = \pi_0 \cdot P^n \cdot P = \pi_n \cdot P = \left(\frac{5}{6}, \frac{1}{6}\right) \cdot \begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix} = \left(\frac{5}{6} \cdot \frac{9}{10} + \frac{1}{6} \cdot \frac{1}{2}, \frac{5}{6} \cdot \frac{1}{10} + \frac{1}{6} \cdot \frac{1}{2}\right) = \left(\frac{5}{6}, \frac{1}{6}\right)$$

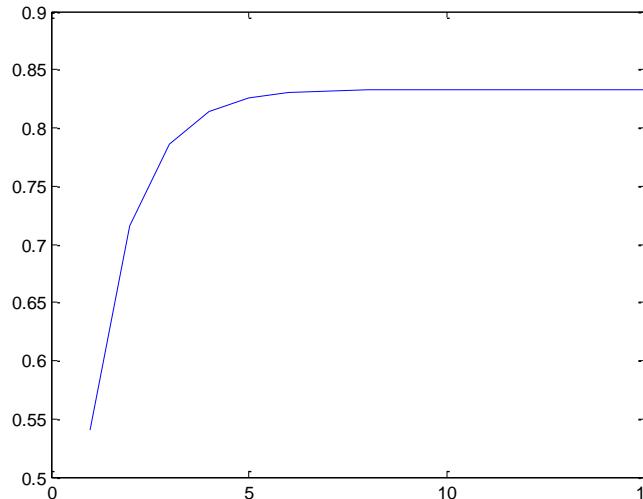
Nach einer gewissen Einschwingphase hat sich die Markov-Kette also auf einen stabilen Zustand eingependelt.

Nach dieser Einschwing-Phase beträgt die Wahrscheinlichkeit dafür, dass der Server an einem beliebigen Tag ein Problem hat, 0.1667.

# stochastische Prozesse

Für einen anderen Startwert  $\pi_0$  erhalten wir

```
i=1:15;  
x=zeros(15,1);  
pi0=[0.1,0.9];  
P=[0.9 0.1;0.5 0.5];  
for k=1:15  
    a=pi0*P^k;  
    x(k)=a(1);  
end  
plot(i,x);
```



keine Probleme      Probleme

$$\begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix}$$

immer noch  $\lim_{n \rightarrow \infty} \pi_n = (0.833, 0.1667)$ .

Die Konvergenz (unabhängig von der Startverteilung) ist kein Zufall, dies gilt für jede für **reguläre** HMK.

## Definition:

Eine HMK mit Ü-Matrix  $P$  heisst **regulär**, wenn ein  $n$  existiert, so dass alle Einträge von  $P^n$  grösser als 0 sind.

# stochastische Prozesse

Dann gilt der folgende wichtige Satz:

Satz:

Zu jeder regulären HMK mit Zustandsraum  $I = \{1, \dots, m\}$  mit Übergangsmatrix  $P$  existiert eine Grenzverteilung (auch stationäre Verteilung oder Gleichgewichtsverteilung genannt)  $\pi^* = (\pi_1^*, \dots, \pi_m^*)$  mit

- (1) Für jede Startverteilung  $\pi_0$  gilt  $\lim_{n \rightarrow \infty} \pi_0 \cdot P^n = \pi^*$ .
- (2)  $\pi^* \cdot P = \pi^*$ .

$$(3) \lim_{n \rightarrow \infty} P^n = \begin{pmatrix} \pi^* \\ \vdots \\ \pi^* \end{pmatrix}. \quad \text{Die Zeilen der Matrix } P^n \text{ konvergieren also gegen die Grenzverteilung } \pi^*.$$

Beispiel:

In unserem Server-Beispiel war jeder Eintrag der Matrix  $P^1 = P$  positiv.

Die HMK ist also regulär, insbesondere existiert also eine von der Startverteilung unabhängige Grenzverteilung (was sich rechnerisch ja schon abgezeichnet hatte).

keine Probleme	Probleme
keine Probleme	$0.9 \quad 0.1$
Probleme	$0.5 \quad 0.5$

# stochastische Prozesse

Wie bestimmt man, im Falle der Existenz, die Gleichgewichtsverteilung  $\pi^*$ ?

Dazu ist ein Gleichungssystem bestehend aus den folgenden Gleichungen zu lösen:

$$\text{Bedingung } (N): \sum_{i \in I} \pi_i^* = 1$$

Normierung

$$\text{Bedingung } (G): \pi^* = \pi^* \cdot P$$

Gleichgewicht

Beispiel:

$$\begin{matrix} & \text{keine Probleme} & \text{Probleme} \\ \text{keine Probleme} & 0.9 & 0.1 \\ \text{Probleme} & 0.5 & 0.5 \end{matrix}$$

$$(N): \pi_1^* + \pi_2^* = 1$$

$$(G): \pi^* = \pi^* \cdot P$$

$$\left\{ \begin{array}{l} \pi_1^* = 0.9\pi_1^* + 0.5\pi_2^* \\ \pi_2^* = 0.1\pi_1^* + 0.5\pi_2^* \end{array} \right.$$

$$\text{mit der Lösung } \pi_1^* = \frac{5}{6} \text{ und } \pi_2^* = \frac{1}{6}.$$

liefert

$$\pi_1^* + \pi_2^* = 1$$

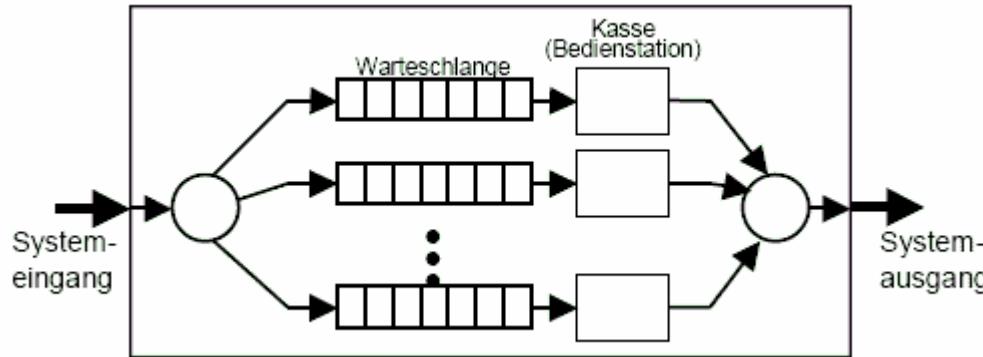
$$0.1\pi_1^* - 0.5\pi_2^* = 0$$

Serie 7, bis Aufgabe 7

# Bediensysteme

Jetzt:

Einführung in Bediensysteme (manchmal: Warteschlangentheorie)



viele Anwendungsgebiete

- Telekommunikationssysteme
- Verkehrssysteme
- Fertigungssysteme
- Rechnernetze
- Computer

# Bediensysteme

---

Man geht in der klassischen Warteschlangentheorie davon aus, dass sich im System (näherungsweise) eine stationäre Verteilung eingestellt hat und bestimmt für die stationäre Verteilung z.B. die folgenden Kenngrößen:

- (1) Verteilung der Zahl  $Q$  der Kunden im System
- (2) Länge der Warteschlange  $N_Q$
- (3) Anzahl  $N_s$  der Kunden, die gerade bearbeitet werden
- (4) Wartezeit  $W_q$  eines einzelnen Kunden in der Schlange (Zeit vom Eintreffen bis zum Beginn der Abarbeitung)
- (5) Verweilzeit  $D$  eines Kunden im System (Wartezeit plus Zeit zur Abarbeitung)

Je nach Annahmen kann nur der Erwartungswert einiger (weniger) der obigen Verteilungen bestimmt werden.

# Bediensysteme: Kendall-Notation

---

Zur Beschreibung von Bediensystemen verwendet man oft die **Kendall-Notation**.

Dabei werden die charakteristischen Größen des Wartesystems in einer definierten Reihenfolge von Buchstaben und Ziffern klassifiziert.

**A|B|s|c|R**

Hierbei steht

- **A** für die Art der **Ankunftsprozesses**
- **B** für die Art des **Bedienvorgangs**
- **s** für die Anzahl der **Bediener (server)**
- **c** für die **Größe des Warteraums (capacity)**
- **R** für die **Reihenfolge der Bedienung**

Meist steht für **A**

- **D** für deterministisch: Die Ankünfte der Kunden findet zu festen (nicht zufälligen) Zeitpunkten statt.  
→ langweiler Fall
- **G** für generelle Annahmen: Über die Ankünfte der Kunden ist gar nichts bekannt.  
→ auch nicht so interessant, da man fast keine Aussagen treffen kann
- **M** für Markov-Eigenschaft: Die Wartezeit auf den nächsten Kunden ist exponentialverteilt.  
Die Exponentialverteilung besitzt bekanntlich die no-memory-property: Die Wartezeit auf den nächsten Kunden hängt nicht davon ab, wie lange man bereits gewartet hat. Das Eintreffen des nächsten Kundens ist also nicht von der Vergangenheit abhängig, und das war bei Markov-Prozessen ja auch so.  
→ man kann viele Aussagen treffen und die Annahme ist in vielen Fällen (zumindest approximativ) erfüllt

# Bediensysteme

**A|B|s|c|R**

Hierbei steht

- **A** für die Art der **Ankunftsprozesses**
- **B** für die Art des **Bedienvorgangs**
- **s** für die Anzahl der Bediener (**server**)
- **c** für die Grösse des Warteraums (**capacity**)
- **R** für die **Reihenfolge** der Bedienung

Meist steht für **B**

- **D** für deterministisch: Die Dauer zur Abfertigung eines Kundens ist nicht zufällig.  
→ langweiler Fall
- **G** für generelle Annahmen: Über die Dauer zur Abfertigung eines Kundens ist gar nichts bekannt.  
→ auch nicht so interessant, da man fast keine Aussagen treffen kann
- **M** für Markov-Eigenschaft: Die Dauer der Abfertigung eines Kundens ist exponentialverteilt.  
→ man kann viele Aussagen treffen und die Annahme ist in vielen Fällen (zumindest approximativ) erfüllt

Für **s** steht

- eine ganze Zahl  $\geq 1$
- oder  $\infty$

Für **c** steht

- eine ganze Zahl  $\geq 0$
- oder  $\infty$

Meist steht für **R**

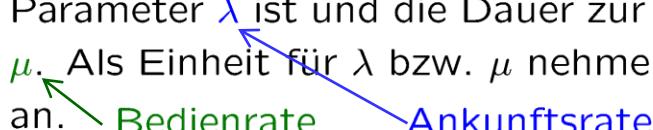
- **FCFS** für **first come first serve**
- **LCFS** für **last come first serve**
- **SIRO** für **service in random order**

Der Standardfall **FCFS** wird meist nicht notiert, ebenso  $c = \infty$ . Bei  $s = \infty$  entfällt die Angabe von  $c$  immer, weil kein Warteraum benötigt wird.

# Bediensysteme

Für den Fall **M|M|1|0** leiten wir die Grenzverteilung her und geben im Anschluss für **M|M|s|c** nur die entsprechende Formeln an (die man analog herleiten könnte).

Wir gehen also davon aus, dass die Wartezeit auf einen Kunden exponentialverteilt mit einem Parameter  $\lambda$  ist und die Dauer zur Abfertigung eines Kundens ist exponentialverteilt mit Parameter  $\mu$ . Als Einheit für  $\lambda$  bzw.  $\mu$  nehme wir Stunden an. Pro Stunde kommen also im Schnitt  $\lambda$  Kunden an.

  
Bedienrate      Ankunftsrate

Wir haben einen Server und keinen Platz im Warteraum.

Das System kann also zwei verschiedene Zustände annehmen:

0: kein Kunde wird bearbeitet      1: ein Kunde wird bearbeitet

Wird ein Kunde gerade bearbeitet, so werden weitere Kunden abgewiesen.

Wir wählen ein Zeitintervall  $h$  (in der Einheit Stunden), welches so klein ist, dass in diesem Zeitintervall höchstens ein Kunde ankommt und höchstens ein Kunde abgearbeitet wird.

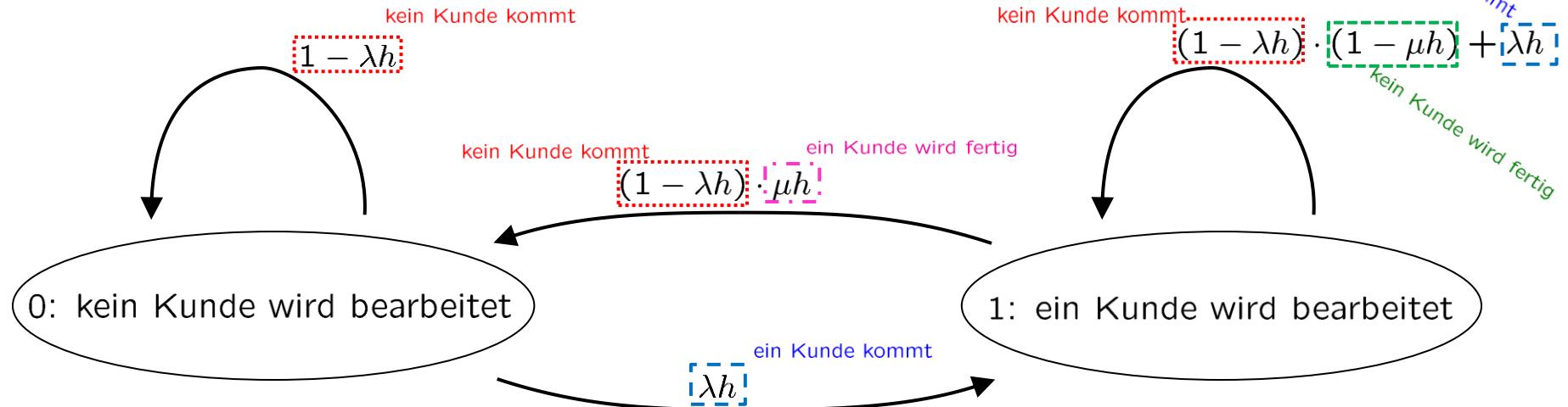
Streng genommen existiert so ein  $h$  nicht. In jedem beliebig kleinen Zeitintervall können im Prinzip beliebig viele Kunden ankommen. Die Wahrscheinlichkeit dafür ist nur vernachlässigbar klein, wenn  $h$  klein genug ist. Und wir bilden am Schluss sowieso den Limes für  $h$  gegen 0.

Da im Zeitintervall  $h$  also entweder ein Kunde ankommt oder eben nicht, muss die Wahrscheinlichkeit für das Ankommen eines Kundens  $\lambda \cdot h$  betragen, damit tatsächlich im Mittel  $\lambda$  Kunden pro Stunde ankommen.

Analog beträgt die Wahrscheinlichkeit, dafür dass ein Kunde, der gerade bearbeitet wird, im Intervall  $h$  fertiggestellt wird  $\mu \cdot h$ .

# Bediensysteme

Wir erhalten als Übergangsdiagramm für die Zeitpunkte  $h, 2h, 3h, \dots$ :



Und Übergangsmatrix:  $P = \begin{pmatrix} 1 - \lambda h & \lambda h \\ (1 - \lambda h) \cdot \mu h & (1 - \lambda h) \cdot (1 - \mu h) + \lambda h \end{pmatrix}$

Wir berechnen davon die stationäre Verteilung, also den Vektor  $\pi^* = (p_0, p_1)$  mit  $\pi^* \cdot P = \pi^*$  und  $p_0 + p_1 = 1$ . Dies ist die erste Gleichung von  $\pi^* \cdot P = \pi^*$ .

Dies ist die erste Gleichung von  $\pi^* \cdot P = \pi^*$ , bei der  $p_1 = 1 - p_0$  eingesetzt wurde.

$$(1 - \lambda h)p_0 + (1 - \lambda h) \cdot \mu h \cdot (1 - p_0) = p_0 \quad \text{liefert } p_0 = \frac{(1 - \lambda h) \cdot \mu h}{(1 - \lambda h) \cdot \mu h + \lambda h}$$

Wir erhalten  $p_0 = \frac{(1 - \lambda h) \cdot \mu}{(1 - \lambda h) \cdot \mu + \lambda} \xrightarrow{h \rightarrow 0} \frac{\mu}{\mu + \lambda}$ , und somit  $p_1 = 1 - p_0 = \frac{\lambda}{\mu + \lambda}$ .

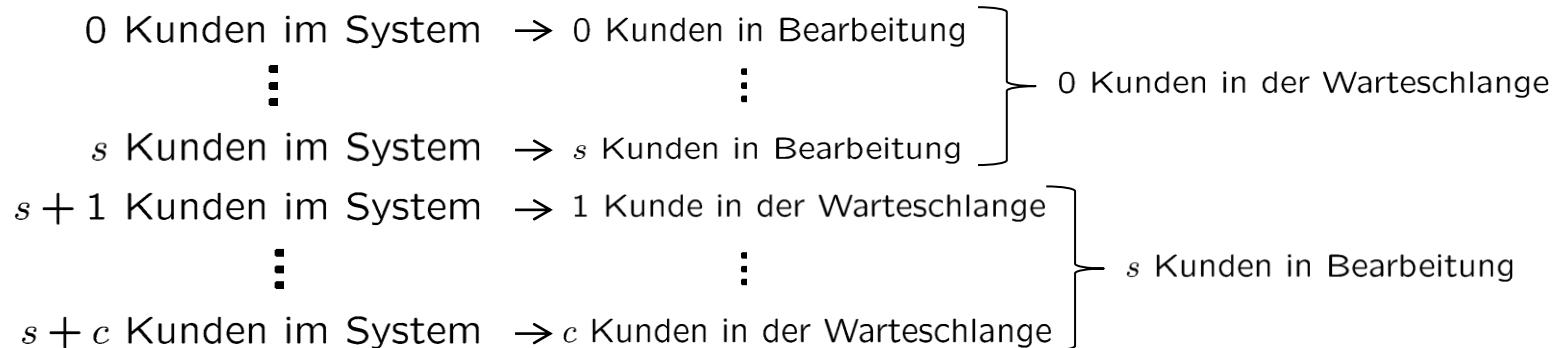
Die Wahrscheinlichkeit (nach einer Einschwingphase), dass der Server belegt ist, beträgt also  $\frac{\lambda}{\lambda+\mu}$ .

# Bediensysteme M|M|s|c

Ähnliche Überlegungen kann man auch im Fall von mehr als einem Server (also  $s > 1$ ) mit Warte-  
kapazität (also  $c > 0$ ) anstellen.

Es ergeben sich dann die folgenden Ergebnisse:

Ein  $M|M|s|c$ -System kann sich den  $s + c + 1$  Zuständen



befinden.

Für alle  $\lambda, \mu$  ergibt sich als Gleichgewichtsverteilung:

$$\begin{aligned}
 p_0 &= \left( \sum_{i=0}^{s-1} \frac{1}{i!} \left(\frac{\lambda}{\mu}\right)^i + \frac{1}{s!} \left(\frac{\lambda}{\mu}\right)^s \underbrace{\sum_{i=s}^{s+c} \left(\frac{\lambda}{s\mu}\right)^{i-s}}_{-1} \right) \\
 p_i &= p_0 \cdot \frac{1}{i!} \left(\frac{\lambda}{\mu}\right)^i \quad \text{für } 0 < i < s \qquad = \begin{cases} c+1, & \lambda = s\mu \\ (1 - (\lambda/(s\mu))^{c+1})/(1 - \lambda/(s\mu)), & \text{sonst} \end{cases} \\
 p_i &= p_0 \cdot \frac{1}{s!} \left(\frac{\lambda}{\mu}\right)^s \left(\frac{\lambda}{s\mu}\right)^{i-s} \quad \text{für } s \leq i \leq s + c
 \end{aligned}$$

$p_i$  gibt also die Wahrscheinlichkeit  
für  $i$  Kunden im System an  
(nach der Einschwingphase)

# Bediensysteme M|M|s|c

## Beispiel:

Ein Autobahnparkplatz wird entworfen. Es stehen nach der Fertigstellung 5 Parkplätze zur Verfügung. Man geht von einer Aufenthaltsdauer der Gäste von durchschnittlich 15 Minuten aus (exponentielle Verteilung). Autofahrer, die einen komplett besetzten Parkplatz vorfinden fahren weiter zur nächsten Parkgelegenheit. Es fahren im Mittel stündlich 10 Autos den Parkplatz an.

- (1) Wie hoch ist die Wahrscheinlichkeit, einen komplett leeren Parkplatz vorzufinden?
- (2) Wie hoch ist die Wahrscheinlichkeit, einen vollen Parkplatz vorzufinden?

## Lösung:

Es handelt sich um ein M|M|5|0-System.

Es ist  $\lambda = 10$ ,  $\mu = 4$ ,  $s = 5$  und  $c = 0$ .

$$p_0 = \left( \sum_{i=0}^{s-1} \frac{1}{i!} \left(\frac{\lambda}{\mu}\right)^i + \frac{1}{s!} \left(\frac{\lambda}{\mu}\right)^s \left[ \sum_{i=s}^{s+c} \left(\frac{\lambda}{s\mu}\right)^{i-s} \right]^{-1} \right)$$
$$p_i = \pi_0 \cdot \frac{1}{i!} \left(\frac{\lambda}{\mu}\right)^i \quad \text{für } 0 < i < s$$
$$p_i = p_0 \cdot \frac{1}{s!} \left(\frac{\lambda}{\mu}\right)^s \left(\frac{\lambda}{s\mu}\right)^{i-s} \quad \text{für } s \leq i \leq s+c$$

$\lambda = s\mu$

$$(1) p_0 = \left( \sum_{i=0}^{s-1} \frac{1}{i!} \left(\frac{10}{4}\right)^i + \frac{1}{5!} \left(\frac{10}{4}\right)^5 \right)^{-1} = 0.0857$$

$$(2) p_5 = p_0 \cdot \frac{1}{5!} \left(\frac{10}{4}\right)^5 = 0.0697$$

```
p=0;
for i=0:5
    p=p+1/factorial(i)*(10/4)^i;
end
1/p
```

# Bediensysteme M|M|s|c

Neben der Verteilung für die Anzahl der Kunden im System (die wir auf der vorigen Folie angegeben haben), sind auch noch die erwartete Länge  $E(N_Q)$  der Warteschlange, die erwartete Anzahl der Kunden  $E(N_s)$ , die gerade bearbeitet werden, die erwartete Wartezeit  $E(W_Q)$  eines einzelnen Kunden bis zur Bedienung und die erwartete Verweilzeit  $E(D)$  eines Kunden im System interessant.

Für die erwartete Warteschlangenlänge ergibt sich:  $E(N_Q) = \sum_{j=s+1}^{s+c} (j-s)p_j$

Erklärung:

Die Warteschlange hat die Länge  $(j - s)$ , wenn  $j$  Kunden im System sind und dafür beträgt die Wahrscheinlichkeit  $p_j$ .

Eine (nicht allzu schwere) Rechnung ergibt:

$$E(N_Q) = \begin{cases} \frac{p_0 \cdot \left(\frac{\lambda}{\mu}\right)^s \cdot \frac{\lambda}{s\mu} \cdot (1+c) \cdot c}{2 \cdot s!} & \text{falls } \lambda = s\mu \\ \frac{p_0 \cdot \left(\frac{\lambda}{\mu}\right)^s \cdot \frac{\lambda}{s\mu} \left[1 - \left(\frac{\lambda}{s\mu}\right)^{c+1} - \left(1 - \frac{\lambda}{s\mu}\right) \cdot (c+1) \cdot \left(\frac{\lambda}{s\mu}\right)^c\right]}{s! \cdot \left(1 - \frac{\lambda}{s\mu}\right)^2} & \text{falls } \lambda \neq s\mu \end{cases}$$

# Bediensysteme M|M|s|c

Für die erwartete Anzahl an gerade bedienten Kunden ergibt sich:  $E(N_s) = \sum_{j=0}^{s-1} jp_j + s \sum_{j=s}^{s+c} p_j$

Erklärung:

Wenn  $0 \leq j < s$  Kunden im System sind (wofür die Wahrscheinlichkeit  $p_j$  ist), werden  $j$  Kunden bedient. Wenn mindestens  $s$  Kunden im System sind (wofür die Wahrscheinlichkeit  $\sum_{j=s}^{s+c} p_j$  ist), werden  $s$  Kunden bedient.

Eine (nicht allzu schwere) Rechnung ergibt:  $E(N_s) = \frac{\lambda}{\mu} \cdot (1 - p_{s+c})$

Für die erwartete Wartezeit eines Kundens bis zur Bedienung ergibt sich:  $E(W_Q) = \frac{E(N_Q)}{\lambda \cdot (1 - p_{s+c})}$

Erklärung:

Dies folgt aus der in (fast) beliebigen System gültigen [Formel von Little](#), welche besagt, dass die erwartete Anzahl von Kunden in einem System gleich dem Produkt ihrer durchschnittlichen Ankunftsrate und ihrer erwartete Verweildauer im System ist.

Für die erwartete Verweilzeit eines Kundens im System ergibt sich:  $E(D) = E(W_Q) + \frac{1}{\mu}$

Erklärung: Summe aus erwarteter Zeit bis zur Bedienung und erwarteter Dauer der Bedienung

# Bediensysteme M|M|s|∞

Es gibt Situationen, wo die Grösse des Warteraumes praktisch unendlich ist.

Für **M|M|s|∞** Systeme ergeben sich die folgenden Resultate:

Für  $\lambda \geq s\mu$  existiert keine Gleichgewichtsverteilung.

Intuitiv kommen dann mehr Anfragen rein, als von den  $s$  Servern zusammen bearbeitet werden können. Die Anzahl der Kunden im System geht also im Laufe der Zeit gegen  $\infty$  und pendelt sich nicht ein.

Für  $\lambda < s\mu$  existiert eine Gleichgewichtsverteilung, und zwar

$$p_0 = \left( \sum_{i=0}^{s-1} \frac{1}{i!} \left( \frac{\lambda}{\mu} \right)^i + \frac{1}{s!} \left( \frac{\lambda}{\mu} \right)^s \frac{s}{s - \frac{\lambda}{\mu}} \right)^{-1}$$
$$p_i = p_0 \cdot \frac{1}{s!} \left( \frac{\lambda}{\mu} \right)^i \quad \text{für } i = 1, 2, \dots$$

$p_i$  gibt also die Wahrscheinlichkeit für  $i$  Kunden im System an (nach der Einschwingphase)

Für die bereits diskutierten Leistungsmasse erhalten wir in diesem Fall ( $\lambda < s\mu$ ):

Für die erwartete Warteschlangenlänge ergibt sich:  $E(N_Q) = C_s \left( \frac{\lambda}{\mu} \right) \cdot \frac{\frac{\lambda}{\mu}}{s - \frac{\lambda}{\mu}}$

wobei  $C_s \left( \frac{\lambda}{\mu} \right) = \sum_{i=s}^{\infty} p_i = \left( \frac{\lambda}{\mu} \right)^s \frac{1}{s!} \cdot \frac{s}{s - \frac{\lambda}{\mu}} \cdot p_0$  die Wahrscheinlichkeit angibt, dass alle Server besetzt sind.

Für die erwartete Anzahl an gerade bedienten Kunden ergibt sich:  $E(N_s) = \frac{\lambda}{\mu}$

Für die erwartete Wartezeit eines Kundens bis zur Bedienung ergibt sich:  $E(W_Q) = \frac{C_s \left( \frac{\lambda}{\mu} \right)}{\mu s - \lambda}$

Für die erwartete Verweilzeit eines Kundens im System ergibt sich:  $E(D) = E(W_Q) + \frac{1}{\mu}$

# Bediensysteme M|M|s|∞

Beispiel:

Im Schnitt benötigt ein Flugzeug zwei Minuten zum landen (exponentialverteilt), und im Schnitt kommen 27 Flugzeuge pro Stunde an (poissonverteilt).

Wie viele Landebahnen werden benötigt, damit die Wahrscheinlichkeit, dass ein Flugzeug wartet, 10% nicht übersteigt?

Wir berechnen für  $s = 1, 2, 3, \dots$  die Wahrscheinlichkeit, dass ein Flugzeug warten muss. Dabei ist es sinnvoll, dies als M|M|s|∞ Warteschlange zu modellieren, da im Prinzip beliebig viele Flugzeuge warten können.

Es ist  $\lambda = 27$  und  $\mu = 30$ .

$$p_0 = \left( \sum_{i=0}^{s-1} \frac{1}{i!} \left(\frac{\lambda}{\mu}\right)^i + \frac{1}{s!} \left(\frac{\lambda}{\mu}\right)^s \frac{s}{s - \frac{\lambda}{\mu}} \right)^{-1} = \frac{1}{1 + \frac{\lambda}{\mu} \frac{1}{1 - \frac{\lambda}{\mu}}} = 0.1$$

Für  $s = 1$  ergibt sich:  $C_s \left(\frac{\lambda}{\mu}\right) = \left(\frac{\lambda}{\mu}\right)^s \frac{1}{s!} \cdot \frac{s}{s - \frac{\lambda}{\mu}} \cdot p_0 = 0.9$

$$p_0 = \left( \sum_{i=0}^{s-1} \frac{1}{i!} \left(\frac{\lambda}{\mu}\right)^i + \frac{1}{s!} \left(\frac{\lambda}{\mu}\right)^s \frac{s}{s - \frac{\lambda}{\mu}} \right)^{-1} = \frac{1}{1 + \frac{\lambda}{\mu} + \frac{1}{2} \left(\frac{\lambda}{\mu}\right)^2 \frac{2}{2 - \frac{\lambda}{\mu}}} = 0.3793$$

Für  $s = 2$  ergibt sich:  $C_s \left(\frac{\lambda}{\mu}\right) = \left(\frac{\lambda}{\mu}\right)^s \frac{1}{s!} \cdot \frac{s}{s - \frac{\lambda}{\mu}} \cdot p_0 = 0.2793$

$$p_0 = \left( \sum_{i=0}^{s-1} \frac{1}{i!} \left(\frac{\lambda}{\mu}\right)^i + \frac{1}{s!} \left(\frac{\lambda}{\mu}\right)^s \frac{s}{s - \frac{\lambda}{\mu}} \right)^{-1} = \frac{1}{1 + \frac{\lambda}{\mu} + \frac{1}{2} \left(\frac{\lambda}{\mu}\right)^2 \frac{2}{2 - \frac{\lambda}{\mu}} + \frac{1}{6} \left(\frac{\lambda}{\mu}\right)^3 \frac{3}{3 - \frac{\lambda}{\mu}}} = 0.4035$$

Für  $s = 3$  ergibt sich:  $C_s \left(\frac{\lambda}{\mu}\right) = \left(\frac{\lambda}{\mu}\right)^s \frac{1}{s!} \cdot \frac{s}{s - \frac{\lambda}{\mu}} \cdot p_0 = 0.07$

Es werden also 3 Landebahnen benötigt.

## Bediensysteme M|M|s| $\infty$

Für die Wartezeit  $W_Q$  in der Schlange bis zur Bedienung gilt:

$$P(W_Q > t) = C_s \left( \frac{\lambda}{\mu} \right) e^{-(su - \lambda)t}.$$

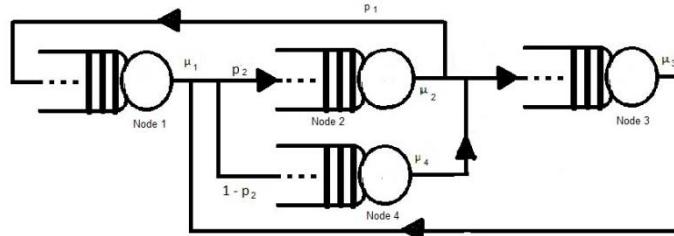
In der Literatur finden sich (ziemlich lange) Formeln für  $P(D > t)$ , also die Wahrscheinlichkeit, dass jemand mindestens die Zeit  $t$  im System verbringt, auch für den Fall eines M|M|s|c Systems.

# Bediensysteme

Damit haben wir einige wenige Aspekte der sehr reichhaltigen Warteschlangentheorie behandelt.

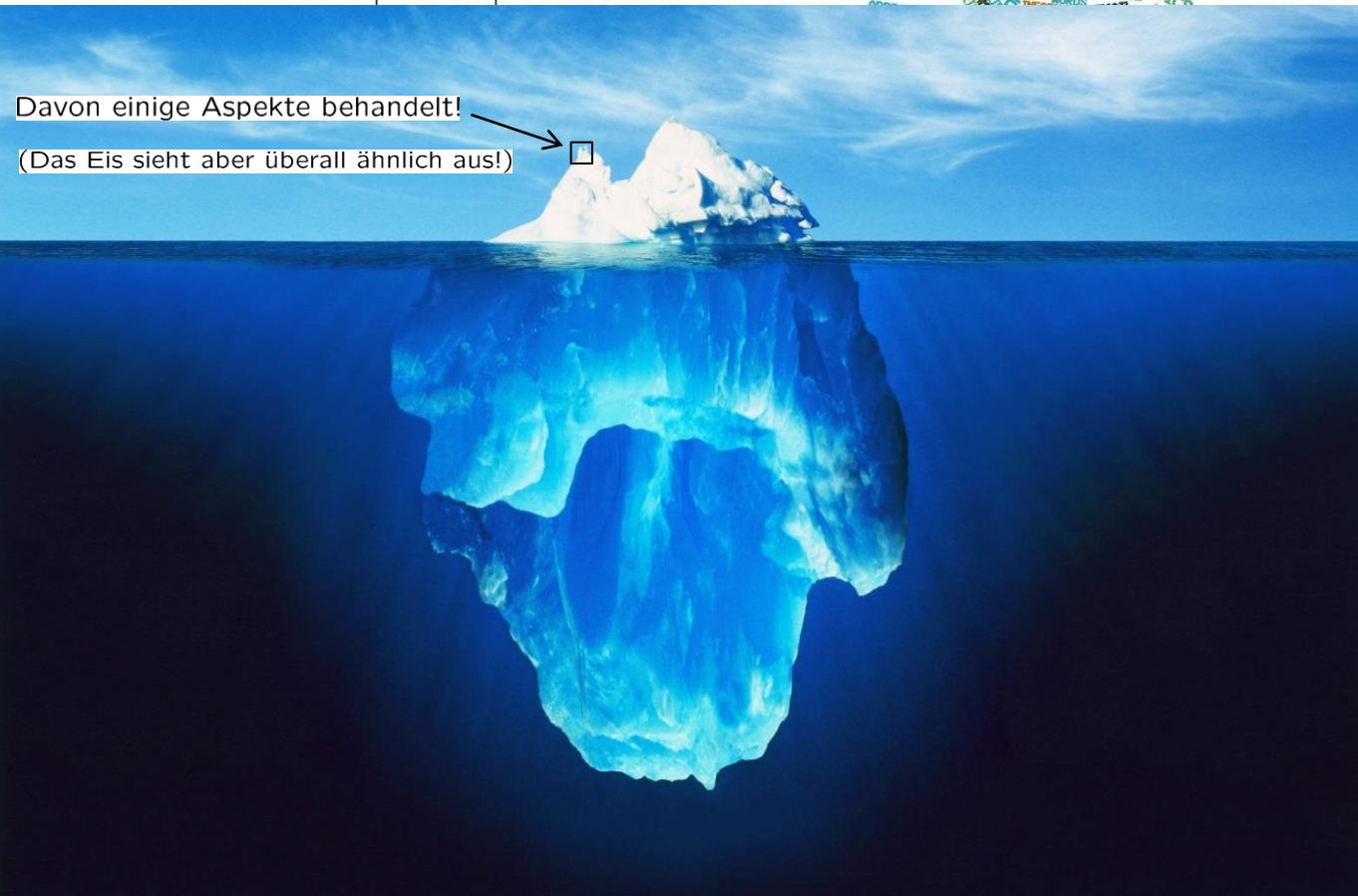
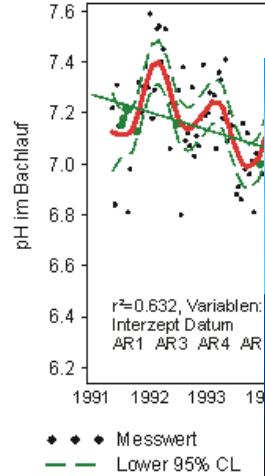
Viele Aspekte können im Rahmen dieses Moduls nicht thematisiert werden:

- andere Verteilungen
- andere Abarbeitungstypen neben FCFS, etwa *Processor Sharing*, wo ankommende Aufträge gleichzeitig (d.h. es wird schnell zwischen den Aufgaben hin- und hergesprungen) bearbeitet werden
- Warteschlangennetze, wo viele Warteschlangensysteme (auch ggfs. rückgekoppelt) verbunden sind



<http://www.ee.cityu.edu.hk/~zukerman/classnotes.pdf> enthält sehr viel zusätzliches Material.

# diskrete Stochastik



```
repeat s - 1 times:  
    x ← x2 mod n  
    if x = 1 then return composite  
    if x = n - 1 then do next WitnessLoop  
return composite  
return probably prime
```