Wirtschaftsuniversität Wien

Bachelorarbeit

Die Warteschlangentheorie

Gebraucht der Zeit, sie geht so schnell von hinnen,
Doch Ordnung lehrt Euch Zeit gewinnen.
-Mephistopheles (aus Goethe, Faust.)

Sergey Yurkevich

Betreuer
Univ.Prof.Mag.Dr. Walter Böhm

15. Mai 2018

Inhaltsverzeichnis

1 Einleitung **2**

2 Notation und Definitionen **4**

3 Gesetz von Little **6**

4 Warteschlangenmodelle **10**

4.1 Das M/M/1 Modell 10

4.1.1 Herleitung 11

4.1.2 Eigenschaften 21

4.1.3 Anwendung 23

4.2 Das M/G/1 Modell 24

4.2.1 Herleitung 25

4.2.2 Anwendung 31

4.3 Andere Modelle und Ausblick 33

5 Conclusio **34**

6 Literaturverzeichnis **36**

Zusammenfassung

Diese Arbeit soll einen Einblick in die Theorie der Warteschlangensysteme liefern sowie einige Resultate dieser herleiten und erläutern. Warteschlangen spielen eine essentielle Rolle in der modernen Gesellschaft und Wissen über diese kann bei gezielter Anwendung viel Zeit und Geld sparen.

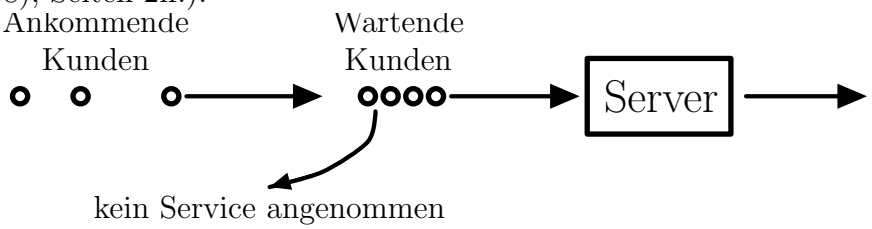
1 Einleitung

Warteschlangen tauchen in der modernen Gesellschaft überraschend oft auf. Sei es im tagtäglichen Leben, wie zum Beispiel als Kunde beim Arzt oder im Geschäft, im betrieblichen Kontext, wie dem Flugverkehr, aber auch in der Informatik, beispielsweise wenn man an Anfragen bearbeitende Server denkt. Eine außer Kontrolle geratene Warteschlange kann zu extremen Wartezeiten führen, welche nicht nur einen Wohlfahrtsverlust für alle beteiligten, sondern auch einen monetären Verlust für Unternehmen bedeuten. Daraus ergibt sich die Notwendigkeit, das Wesen der Warteschlangen zu verstehen, um dieses in der Praxis gezielt anwenden zu können und somit Schäden zu verhindern. Es stellt sich heraus, dass diese Theorie äußerst kompliziert zum Herleiten ist und sehr tiefliegende Resultate aus vielen mathematischen Teilgebieten erfordert. Die Basis für Warteschlangen stellt die Wahrscheinlichkeitstheorie dar, welche wiederum als Anwendung der Maßtheorie gesehen werden kann. Das Thema dieser Arbeit ist ein Paradebeispiel dafür, wie abstrakte Theorie, wie zum Beispiel die der Markov'schen Ketten aus der Wahrscheinlichkeitsrechnung, Anwendung in angewandter Mathematik und in weiterer Folge in Wirtschaft findet.
Obwohl die Notwendigkeit für Kenntnis der Theorie über Warteschlangen gesellschaftlich mehr oder weniger schon immer gegeben war, ist dieses Thema überraschend neu. Erste Resultate stammen von A. K. Erlang im Jahre 1909, als er die notwendige Größe für Telefonvermittlungsanlagen einschätzen wollte. Danach gewann dieses Gebiet im Laufe des 20. Jahrhunderts laufend an Bedeutung. Die solch späte Entwicklung dieser mathematischen Disziplin lässt sich einerseits durch große Abhängigkeit von der Wahrscheinlichkeitstheorie erklären, welche wiederum nur durch Erfindung der Maßtheorie im letzten Jahrhundert große Meilensteine überbrücken konnte. Andererseits wurde erst durch die technologische Entwicklung zuerst durch das Telefon, dann durch den Computer und Computernetzwerke die Nachfrage nach besagter Theorie enorm.

Um mit diesem Werk spannende Resultate tatsächlich zu erreichen und nicht in ewig langen Definitionen und Grundlagen stecken zu bleiben, wird eine gewisse Basis an Vorkenntnissen in der Wahrscheinlichkeitstheorie und Analysis vorausgesetzt. Bei jeder Verwendung von unbewiesenen Formeln wird natürlich auf Quellen verwiesen, die Lücken zudecken und Fragen beantworten sollen, trotzdem ist es sehr ratsam im Vorhinein Grundlagenwissen über die genannten mathematischen Gebiete zu besitzen. Das umfasst Vertrautheit mit solchen Begriffen wie Zufallsvariable, Verteilungsfunktion, Erwartungswert und ähnlichen sowie Erfahrung mit bekannten Wahrscheinlichkeitsverteilungen und Kenntnis deren Eigenschaften.
Das folgende Kapitel stellt essentielle Notation vor, um über Warteschlangenmodelle kommunizieren zu können. Es werden einige grundlegende Definitionen offenbart, die dann in der ganzen späteren Arbeit wesentlich sein werden. Im darauffolgenden Kapitel wird ein zentrales Resultat aus der Theorie der Warteschlangen präsentiert und erläutert, welches so wichtig ist, weil es äußerst allgemein gilt. Als nächstes beschäftigt sich diese Arbeit mit zwei entscheidenden spezifischen Modellen: die wichtigsten Formeln werden hergeleitet, analysiert und in Anwendungsbeispielen erprobt. Schließlich folgt eine Conclusio, worin nochmals die ausschlaggebenden Resultate dieser Arbeit zusammengefasst werden.

2 Notation und Definitionen

Ein Warteschlangensystem ist ein Modell, welches jenen Prozess beschreibt in dem *Kunden* wegen eines Anliegens in ein System eintreffen, warten, bis sie an die Reihe kommen, wenn das nicht gleich passiert, danach eine gewisse Zeit behandelt werden und schließlich das System wieder verlassen. Manchmal betrachtet man auch den allgemeineren Fall, in dem Kunden aufgrund bestimmter Faktoren das System verlassen ohne auf Behandlung zu warten, zum Beispiel, wenn die Warteraumkapazität nicht ausreicht. Hierbei ist „Kunde“ so universell wie möglich zu sehen und bezieht sich keineswegs nur auf Menschen. Das können Geräte sein, die auf Reparatur warten, Flugzeuge, die bald abfliegen sollen, Anfragen an einen Webserver und vieles mehr. Die Bearbeitung der Anliegen der Kunden wird *Service* genannt und wird vom sogenannten *Server* erledigt. Schematisch kann somit so ein Warteschlangenmodell wie folgt dargestellt werden (Gross, Shortle, Thompson, Harris (2008), Seiten 2ff.):



In den meisten Fällen ist das Modell für eine Warteschlange schon durch fünf Charakteristika bestimmt: Art des Ankunftsprozesses, Art des Bearbeitungsprozesses, Anzahl der Server, Kapazität des Systems und Aufrufeprinzip (Gross, Shortle, Thompson, Harris (2008), Seiten 7ff.). Daraus ergibt sich die moderne Kurznotation für Warteschlangenmodelle, die sich dank dem Statistiker und Mathematiker David G. Kendall etabliert hat. So ein Modell wird im Normalfall mit fünf Zeichen, getrennt durch Solidi, angegeben: $A/B/c/K/Z$. Dabei stehen A und B für die Verteilungen der Ankunfts- und Bearbeitungszeiten, c ist eine Konstante, die angibt wie viele Server für die Bedienung der Kunden zur Verfügung stehen, K ist auch konstant und steht für die maximale Kapazität des Systems und Z gibt das Prinzip an nachdem die Kunden aufgerufen werden. Zum Beispiel steht $G/D/4/\infty/FCFS$ für ein Warteschlangenmodell, bei dem der Ankunftsprozess einer beliebigen (**G**eneral) Verteilung folgt (die aber dennoch durch unabhängige und identisch verteilte Zufallsvariablen charakterisiert wird), die Bearbeitungszeit immer konstant (**D**eterministic) ist, genau vier Server die Anfragen der Kunden bearbeiten, beliebig viel Platz (∞) im Warteraum zur Verfügung

steht und die Kunden nach dem Windhundprinzip (**F**irst-**C**ome-**F**irst-**S**erve) aufgerufen werden. Einige Möglichkeiten für die Charakteristika sowie deren kurze Beschreibungen können der Tabelle entnommen werden:

Charakteristik	Symbol	Kurzbeschreibung
Verteilungen von Ankunftszeiten (A) und Bearbeitungszeiten (B)	M	Exponentiell
	D	Deterministisch
	E_k	Erlang mit Parameter k
	PH	Phasen
	G	Beliebige Verteilung
Anzahl von Server	$1, 2, 3, ..., \infty$	
Kapazität des Systems	$1, 2, 3, ..., \infty$	
Aufrufprinzip	FCFS	First Come First Serve
	LCFS	Last Come First Serve
	PR	Priority
	RSS	Random Selection for Service
	GD	General discipline

Es ist üblich die beiden letzten Charakteristika wegzulassen, wenn diese jeweils ∞ und $FCFS$ sind, weil das das häufigste Szenario ist. So bleiben oft nur mehr drei Kennzeichen übrig und das oben angeführte Beispiel hätte somit die Abkürzung $G/D/4$.
Erwähnenswert ist, dass diese Tabelle ganz und gar nicht vollständig ist. Es gibt unzählige Verfeinerungen und autorspezifische Notationen, wie zum Beispiel solche, die Gruppenankünfte von Kunden beachten. Dieser Überblick soll die allgemeine und häufigste Darstellung von Warteschlangenmodellen zeigen und den Leser mit dieser vertraut machen.
Diese Tabelle ist eigentlich selbsterklärend, deshalb wird an dieser Stelle nicht mehr auf diese eingegangen. Der Vollständigkeit halber und um alle mögliche Fragen zu klären, soll auf Gross, Shortle, Thompson, Harris (2008) verwiesen werden.

Im Laufe dieser Arbeit wird laufend neue Notation vorgestellt, um den Leser nicht gleich mit zu vielen Definitionen zu überschütten. An dieser Stelle sollen nun ein paar Kennzahlen definiert werden, welche grundlegend für die Warteschlangentheorie sind; dann soll im nächsten Kapitel eine sehr berühmte Beziehung zwischen diesen erläutert werden.
Man definiert (unter der Annahme, dass diese existieren) λ als die durchschnittliche Ankunftsrate der Kunden und μ als die durchschnittliche Servicerate von einem Server; also pro Zeiteinheit betreten durchschnittlich λ Kunden das System und können $c\mu$ Kunden bearbeitet werden, wenn c die

Anzahl der Server ist. Klarerweise, wird für $\lambda > c\mu$ die Warteschlangengröße gegen unendlich streben (wenn das System unbeschränkte Kapazität aufweist), weshalb im Normalfall $\lambda \leq c\mu$ vorausgesetzt wird. Offenbar ist der Quotient dieser beiden Kennzahlen von Bedeutung und so tauft man $\rho \coloneqq \lambda/(c\mu)$.
Weiters, sei $T(n)$ jene Zufallsvariable, die die Zeit angibt, wie lange der $n-$te Kunde im System verbringt und $T_q(n)$ auch eine, die für die verbrachte Zeit von diesem in der *Warteschlange* steht. Wie sich im Laufe dieser Arbeit auch zeigen wird, handeln viele Resultate in der Warteschlangentheorie nur von sogenannten „Gleichgewichtszuständen". Salopp gesprochen, lässt sich oft zeigen, dass die Verteilung von $T(n)$ für $n \to \infty$ sich einpendelt und so definiert man, in jenen Fällen, in denen sie wohldefiniert ist, die Zufallsvariable $T \coloneqq \lim_{N\to\infty} (1/N) \sum_{n=1}^{N} T(n)$ und analog $T_q \coloneqq \lim_{n\to\infty} (1/N) \sum_{n=1}^{N} T_q(n)$. Dann wird definiert, wieder falls existent, $W \coloneqq \mathbb{E}(T)$ und $W_q \coloneqq \mathbb{E}(T_q)$. Also gibt W_q die durchschnittliche Wartezeit der Kunden in der Schlange an und W die durchschnittlich verbrachte Zeit im System, jeweils nach langer Zeit.
Schließlich soll $N(t)$ die Anzahl der Kunden im System zum Zeitpunkt t repräsentieren und, analog wie die Kennzahlen davor, ist dann $N_q(t)$ die Anzahl der Kunden in der Warteschlange zu diesem Zeitpunkt. Man definiert wieder, falls sie existiert, die Zufallsvariable $N \coloneqq \lim_{t\to\infty} N(t)$ und natürlich auch analog $N_q \coloneqq \lim_{t\to\infty} N_q(t)$. Mithilfe dieser, und nochmals unter der Annahme der Wohldefiniertheit, ist dann $L \coloneqq \mathbb{E}(N)$ und $L_q \coloneqq \mathbb{E}(N_q)$. Mit anderen Worten gibt also L den Erwartungswert für die Anzahl der Kunden im System nach langer Zeit an und L_q, entsprechend, die durchschnittliche Anzahl der Kunden in der Warteschlange.

Diese Definitionen eingeführt, kann nun das erste, überaus wichtige Resultat der Warteschlangentheorie formuliert und erläutert werden.

3 Gesetz von Little

Die wohl berühmteste Formel in der Theorie der Warteschlangen wurde zum ersten Mal vom amerikanischen Mathematiker J.D.C. Little bewiesen und ist nach ihm benannt. Little wurde am 1. Februar 1928 geboren und ist somit zum Zeitpunkt des Erscheinens dieser Arbeit 90 Jahre alt. Im Jahr 1954 verwendete A. Cobham das später nach Little benannte Gesetz in einem Artikel ohne einen Beweis anzuführen. Diese Beziehung brachte einige philosophische Fragen auf, denn die Aussage dieser hört sich, wenn man über sie eine längere Zeit nachdenkt, absolut plausibel an, der Beweis ist jedoch anschei-

nend gar nicht einfach. So dauerte es 7 Jahre und einige Arbeiten, welche die unbewiesene Formel voraussetzten, bis im Jahr 1961 Little einen formalen Beweis schaffte (Little und Graves (2008), Seiten 97ff.). Dieser verwendete jedoch einige Annahmen über L, λ und W, die, wie sich später in 1972 herausstellte, überflüssig waren. In diesem Jahr stellte nämlich S. Stidham Jr. einen Beweis vor, welcher nur die Existenz dieser drei Kennzahlen erforderte, jedoch schwierige und nicht naheliegende mathematische Theorie ausnutzte. Kurz darauf brachte der selbe Autor einen weiteren, diesmal direkten, Beweis für Little's Theorem und setzte somit mit seiner Arbeit „A Last Word on $L = \lambda W$" Schluss der Diskussion über die Notwendigkeit einer formalen Herleitung (Little und Graves (2008) sowie Stidham (1972)).
Das Gesetz von Little besagt:

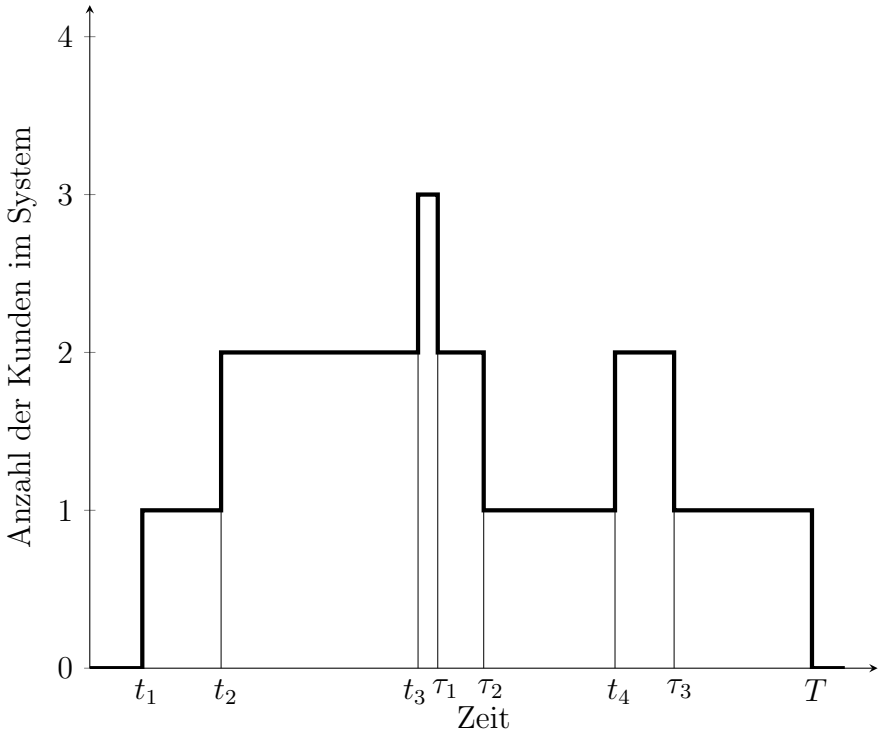
Theorem 3.1 (Gesetz von Little)**.** *Existieren bei einer Warteschlange L, λ und W, so gilt die Beziehung*

$$L = \lambda W. \tag{1}$$

Existieren L_q, λ und W_q, so gilt

$$L_q = \lambda W_q. \tag{2}$$

Der vollständige Beweis für dieses Theorem würde den Rahmen dieser Arbeit verlassen, jedoch soll einerseits eine heuristische Herleitung vorgeführt werden, die dem Leser ohne Zweifel das Gesetz von Little plausibel macht und andererseits der Beweis von Stidham skizziert. (Gross, Shortle, Thompson, Harris (2008), Seiten 10f.).
Es soll angenommen werden, dass zum Zeitpunkt 0 die Warteschlange leer ist; dann betrachtet man den Graphen für die Anzahl der Kunden bis zu einem späteren Zeitpunkt T, an dem das System wieder leer ist. Die Anzahl der Kunden, die das System in diesem Zeitintervall betreten haben, taufen wir N_c. Illustriert man diesen Graphen für das Beispiel $N_c = 4$, so ergibt sich folgendes Bild:



Zu den Zeitpunkten t_1, t_2, t_3 und t_4 kommen die vier Kunden jeweils an und zu den Zeitpunkten τ_1, τ_2, τ_3 und T verlassen sie das System. Daraus ergeben sich die Rechnungen für L und W für diese Zeitperiode:

$$\begin{aligned} L &= \big[1(t_2 - t_1) + 2(t_3 - t_2) + 3(\tau_1 - t_3) + 2(\tau_2 - \tau_1) \\ &\qquad + 1(t_4 - \tau_2) + 2(\tau_3 - t_4) + 1(T - \tau_3)\big]/T \\ &= \frac{\tau_1 + \tau_2 + \tau_3 + T - t_1 - t_2 - t_3 - t_4}{T} =: \frac{A}{T}, \end{aligned}$$

und

$$W = \frac{(\tau_1 - t_1) + (\tau_2 - t_2) + (\tau_3 - t_3) + (T - t_4)}{4} = \frac{A}{N_c}.$$

Vergleicht man nun die beiden Ausdrücke, so fällt auf, dass in beiden A vorkommt. (Im Allgemeinen ist A die Fläche unter der Kurve des oben skizzierten Graphen, was die Bezeichnung (**A**rea) gerechtfertigt aber für diese Herleitung nicht wichtig ist.) Somit ergibt sich die Formel $LT = A = WN_c$ und umgeformt

$$L = W\frac{N_c}{T}.$$

Nun ist aber N_c genau die Anzahl der eintreffenden Kunden bis zum Zeitpunkt T und deshalb spiegelt N_c/T die durchschnittliche Ankunftsrate in

diesem Zeitintervall wider. Die durchschnittliche Ankunftsrate hat den Namen λ und folglich ergibt sich das erste Gesetz von Little. Die zweite Formel kann auf ähnliche Weise plausibel gemacht werden.

Wie schon angemerkt, ist diese heuristische Erklärung ist kein formaler Beweis, es fehlt nämlich das Argument, dass die Beziehung als Limes gegen unendlich erhalten bleibt. Der Beweis von S. Stidham Jr. soll hier auch lakonisch skizziert werden, weil dieser schön zeigt, wie Plausibles formal argumentiert werden kann (Stidham (1972)).
Stidham definiert, wie in der obigen Herleitung, $(t_n)_{n\geq 1}$ als jene Zeitpunkte an denen Kunden eintreffen und dann W_n als jene Zeit, die der $n-$te Kunde im System verbringt. Weiters soll

$$U(t) \coloneqq \sum_{\substack{n \\ t_n \leq t}} W_n \qquad \text{und} \qquad V(t) \coloneqq \sum_{\substack{n \\ t_n + W_n \leq t}} W_n.$$

Es lässt sich dann ganz leicht zeigen, dass

$$U(T) \geq \int_0^T L(t)dt \geq V(T). \tag{3}$$

Eine etwas schwierigere, aber rein kombinatorische, Rechnung liefert, dass

$$\lambda W = \lim_{T\to\infty} U(T)/T.$$

Schafft man es nun zu zeigen, dass $\lim_{T\to\infty} U(T)/T = \lim_{T\to\infty} V(T)/T$, dann folgt wegen Gleichung (3) und $\lim_{T\to\infty} (1/T)\int_0^T L(t)dt = L$ das gewünschte Theorem. Das gelingt Stidham, indem er $V(T)/T$ in drei Teile unterteilt, zeigt dass zwei von diesen für $T \to \infty$ verschwinden und der dritte beliebig nahe $U(T)/T$ annähert.

Die Mächtigkeit des Gesetzes von Little besteht darin, dass die überraschend simple Aussage so extrem allgemein Gültigkeit findet. Es dürfen beliebige Verteilungen der Ankünfte und Bearbeitungen vorliegen, es gibt keine Einschränkungen für die Art des Aufrufeprinzips, für die Anzahl der Server oder für die Kapazität des Systems: das einzige was gefordert wird, ist, dass die Kennzahlen, um die es sich handelt, existieren. Mit der eingeführten Notation könnte man sagen, Little's Gesetz gilt für Modelle $G/G/c/K/G$, wobei c und K beliebig sind (Little und Graves (2008) sowie Stidham (1972)).
Leider ist das Theorem von Little die einzige solch universelle Aussage in der Theorie der Warteschlangen. Um andere Theoreme zu erkunden, muss man die Situation spezifizieren und Annahmen treffen. Genau das ist das Thema des nächsten Kapitels.

4 Warteschlangenmodelle

In dieser Arbeit werden das Modell $M/M/1$ und dessen Verallgemeinerung $M/G/1$ behandelt. Wie im Abschnitt 2 schon angedeutet, bedeutet das erste M, dass die Zeiten zwischen den Ankünften exponentiell verteilt sind und das zweite M, dass die Bearbeitungszeiten auch solch einer Verteilung (jedoch mit anderem Parameter) folgen. Die Praxis zeigt, dass in den meisten Fällen genau die Exponentialverteilung die Zeitintervalle der Ankünfte am besten beschreibt. Das liegt daran, dass der Ankunftsprozess in der Regel einem Poisson Prozess gleicht, was genau die Exponentialverteilung für die Zeitabstände impliziert (Jain, Mohanty, Böhm (2007)). Dank der Gedächtnislosigkeit der Exponentialverteilung kann das $M/M/1$ Modell als Markov Prozess aufgefasst werden und ist dadurch relativ leicht zu analysieren. Um aber die Theorie der Markov-Ketten anwendenden zu können, muss höchst anspruchsvolles Wissen aus diesem Gebiet vorausgesetzt werden, was die Fragestellung nur woanders verschiebt und keinen tiefen Einblick in den Kern des Themas liefert. Deshalb wird in dieser Arbeit eine alternative Herleitung der wichtigsten Formeln des $M/M/1$ Modells angeführt, die keine Theorie über Markov'sche Ketten benötigt. Diese Herleitung ist zwar ziemlich kompliziert, liefert aber noch ein sehr wichtiges Resultat, welches man mit der Theorie von Markov nicht bekommen hätte. Es lohnt sich anzumerken, dass diese Arbeit wahrscheinlich die einzige ist, die dieses Modell tatsächlich ohne besagter Theorie vollständig herleitet.
Für das $M/G/1$ Modell muss jedoch auf die bisher umgangene Theorie verwiesen werden, denn es ist keine andere Möglichkeit bekannt, die wichtigen Resultate für dieses Modell zu beweisen.
Viele grundlegende Ideen und Beweismethoden sind entnommen aus Jain, Mohanty, Böhm (2007), oft wurde jedoch ein anderer Ansatz oder eine andere Herangehensweise gewählt, um die behandelten Warteschlangenmodelle dem Leser so verständlich wie möglich näher zu bringen.

4.1 Das M/M/1 Modell

Folgende Annahmen sollen getroffen werden:

(i) Der Ankunftsprozess unterliegt einer Poisson Verteilung mit Parameter λ.

(ii) Die Servicezeiten sind unabhängige und identisch verteilte Zufallsvariablen, welche der Exponentialverteilung mit Parameter μ folgen.

(iii) Servicezeiten sind unabhängig von Ankunftszeiten.

(iv) Im Warteraum ist beliebig viel Platz.

(v) Es gibt nur einen Server.

(vi) Kunden werden nach dem Windhundprinzip (auch bekannt unter First-Come First-Serve - FCFS - Prinzip) aufgerufen.

(vii) Am Anfang ist niemand in der Warteschlange.

Bemerkung. Die letzten beiden Bedingungen sind nicht zwingend notwendig, sie machen die Argumentation etwas einfacher, einige Brechungen leichter und die Formeln schöner. Weil jedoch trotzdem der Sinn der Herleitung der $M/M/1$ Warteschlangen gefasst wird und die Annahme eines bei Eröffnung leeren Wartezimmers realistisch ist, beschränkt sich diese Arbeit nur auf die Lösung dieses Modells.

4.1.1 Herleitung

Weil die Ankünfte unabhängig Poisson verteilt sind, wissen wir, dass dadurch die Zeitabstände T zwischen zwei aufeinanderfolgenden Ankünften zwingendermaßen exponentialverteilt mit dem Parameter λ und auch unabhängig sind (Forbes, Evans, Hastings, Peacock (2011)). Wir setzen

$$\begin{aligned} f_T(t) &\coloneqq \lambda e^{-\lambda t}, \qquad t \geq 0, \\ f_S(s) &\coloneqq \mu e^{-\lambda s}, \qquad s \geq 0, \end{aligned}$$

für die Dichtefunktionen von jeweils den Zwischenankunftszeiten T und den Servicezeiten S.
Das Ziel ist es jenen Ausdruck zu finden, der die Wahrscheinlichkeit beschreibt, dass nach t Zeit sich genau n Kunden im System befinden. Mit anderen Worten suchen wir $P_n(t)$ nach folgender Definition:

$$\begin{aligned} X(t) \quad &: \quad \text{Anzahl der Kunden im System zum Zeitpunkt t} \\ P_n(t) \quad &: \quad \mathbb{P}(X(t) = n) \end{aligned}$$

Um $P_n(t)$ in den Griff zu bekommen, brauchen wir eine Gleichung, welche diesen Term beschreibt. Dafür formulieren wir und beweisen das Lemma:

Lemma 4.1. *Für das oben definierte $P_n(t)$ gilt die Differenzen- und Differentialgleichung*

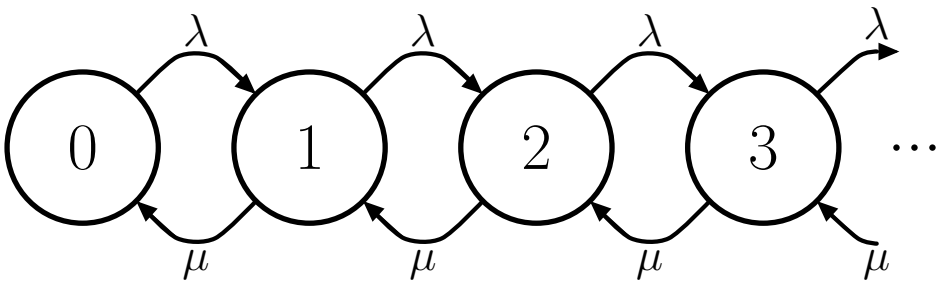
$$\frac{dP_n(t)}{dt} = \mu P_{n+1}(t) + \lambda P_{n-1}(t) - (\lambda + \mu) P_n(t), \;\; n \geq 1, \tag{4}$$

mit der Anfangsbedingung

$$\frac{dP_0(t)}{dt} = \mu P_1(t) - \lambda P_0(t).$$

Es sollen zwei unterschiedliche Beweise vorgestellt werden. Der erste erklärt anschaulich, warum die Gleichung (4) stimmt, der zweite ist eine formale Herleitung der Formel aus den Annahmen über die Warteschlange und den Definitionen der Verteilungen.

Beweis. Betrachte das Flussdiagramm in Abbildung:



Jeder Kreis steht für jenen Zustand, in dem die Anzahl der Kunden im System der Zahl im Kreis entspricht. Die Pfeile symbolisieren den Wechsel von einem Zustand in einen anderen, wobei die Raten, mit denen das passiert, genau λ beziehungsweise μ sind, was aus den Annahmen (i), (ii) und (v) über die Warteschlange ersichtlich ist.
Für $n \geq 1$ ist die Übergangsrate *in* den Zustand n gegeben durch den Strom aus den Zuständen $n-1$ und $n+1$ und somit gleich $\mu P_{n+1} + \lambda P_{n-1}$ (Annahmen (iii) und (iv)). Die Ausgangsrate *aus* dem Zustand n ist dann auch leicht zusehen und ist gegeben durch $(\mu + \lambda)P_n$. Für $n = 0$ ist die Situation ähnlich: die Eingangsrate beträgt μP_1 und die Ausgangsrate ist λP_0. Nun repräsentiert $\frac{dP_n(t)}{dt}$ genau die Veränderungsrate des Flusses im Zustand n zum Zeitpunkt t, also die Differenz zwischen der Eingangsrate und Ausgangsrate. Das liefert uns für $n \geq 1$ wie gewünscht

$$\frac{dP_n(t)}{dt} = \mu P_{n+1}(t) + \lambda P_{n-1}(t) - (\lambda + \mu)P_n(t),$$

und für $n = 0$ auch wie gewollt

$$\frac{dP_0(t)}{dt} = \mu P_1(t) - \lambda P_0(t).$$

□

Um das Lemma rigoros zu begründen, soll diese heuristische Erklärung durch einen formalen Beweis gestützt werden.

Beweis. Für $n \geq 1$ und Δt klein betrachte $P_n(t + \Delta t)$. Der Kunstgriff hier ist es diese Funktion, die eine Wahrscheinlichkeit widerspiegelt mit Termen darzustellen, die maximal ein Ereignis im Zeitraum $(t, t + \Delta t)$ beschreiben:

Die anderen Ereignisse sind für kleine Δt verschwindend klein. Formal sieht das wie folgt aus:

$$\begin{aligned}
P_n(t+\Delta t) = P_{n+1}(t)\mathbb{P}(&\textbf{keine} \text{ Ankunft, und } \textbf{ein} \text{ Service} \\
&\text{im Zeitraum } (t, t+\Delta t)) \\
+ P_n(t)\mathbb{P}(&\textbf{keine} \text{ Ankunft, und } \textbf{kein} \text{ Service} \\
&\text{im Zeitraum } (t, t+\Delta t)) \\
+ P_n(t)\mathbb{P}(&\textbf{eine} \text{ Ankunft, und } \textbf{ein} \text{ Service} \\
&\text{im Zeitraum } (t, t+\Delta t)) \\
+ P_{n-1}(t)\mathbb{P}(&\textbf{eine} \text{ Ankunft, und } \textbf{kein} \text{ Service} \\
&\text{im Zeitraum } (t, t+\Delta t)) \\
+ &\text{ Terme, die } \textbf{mehr als ein} \text{ gleiches Ereignis} \\
&\text{im Zeitraum } (t, t+\Delta t) \text{ berücksichtigen}
\end{aligned}$$

Weil sowohl die Ankünfte, als auch die Service Erledigungen Poisson verteilt mit Parameter λ, beziehungsweise μ sind, ist die Wahrscheinlichkeit dass genau eine Ankunft im Zeitraum $(t, t+\Delta t)$ passiert gleich $\lambda\Delta t + o(\Delta t)$. Dabei steht $o(\Delta t)$ für die klein-o-Notation von Landau und besagt, dass dieser Term für sehr kleine Δt schneller gegen Null strebt als Δt; mit anderen Worten $\lim_{\Delta t \to 0} o(\Delta t)/\Delta t = 0$. Analog ist die Wahrscheinlichkeit, dass genau eine Service Erledigung im behandelten Zeitintervall durchgeführt wird gleich $\mu\Delta t + o(\Delta t)$. Für keine Ankunft, bzw. keine Erledigung erhält man $1 - \lambda\Delta t + o(\Delta t)$ bzw. $1 - \mu\Delta t + o(\Delta t)$ als Wahrscheinlichkeiten. Terme, die mehr als ein gleiches Ereignis berücksichtigen sind nach der Definition der Poisson Verteilung von Ordnung $o(\Delta t)$. Weil die Ereignisse unabhängig voneinander sind, können wir die **und**-Verknüpfung mit Multiplikation ersetzen und erhalten

$$\begin{aligned}
P_n(t+\Delta t) = &P_{n+1}(t)\left[(1-\lambda\Delta t + o(\Delta t))(\mu\Delta t + o(\Delta t))\right] \\
&+ P_n(t)\left[(1-\lambda\Delta t + o(\Delta t))(1-\mu\Delta t + o(\Delta t))\right] \\
&+ P_n(t)\left[(\lambda\Delta t + o(\Delta t))(\mu\Delta t + o(\Delta t))\right] \\
&+ P_{n-1}(t)\left[(\lambda\Delta t + o(\Delta t))(1-\mu\Delta t + o(\Delta t))\right] \\
&+ o(\Delta t).
\end{aligned}$$

Ausmultiplizieren und Gruppieren gibt, bei Verwendung der Eigenschaft, dass für kleine Δt, $(\Delta t)^2$ von Ordnung $o(\Delta t)$ ist, für $n \geq 1$:

$$P_n(t+\Delta t) = P_{n+1}(t)\mu\Delta t + P_{n-1}(t)\lambda\Delta t + P_n(t)(1-\lambda\Delta t - \mu\Delta t) + o(\Delta t). \quad (5)$$

Für $n = 0$ ändert sich wenig: man muss bedenken, dass wenn niemand in der Warteschlange ansteht, auch niemand bedient werden kann und dass dieser

Zustand nicht durch das Ankommen eines Kunden passieren konnte. Man erhält in diesem Sinne eine ähnliche Formel für $P_0(t + \Delta t)$, die sich auch vereinfachen lässt:

$$P_0(t + \Delta t) = P_1(t)\mu\Delta t + P_0(t)(1 - \lambda\Delta t) + o(\Delta t). \tag{6}$$

Nun schreiben wir die Gleichungen (5) und (6) um und dividieren durch Δt:

$$\frac{P_n(t + \Delta t) - P_n(t)}{\Delta t} = \mu P_{n+1}(t) + \lambda P_{n-1}(t) - (\lambda + \mu)P_n(t) + \frac{o(\Delta t)}{\Delta t}, \quad n \geq 1$$
$$\frac{P_0(t + \Delta t) - P_0(t)}{\Delta t} = \mu P_1(t) - \lambda P_0(t) + \frac{o(\Delta t)}{\Delta t}.$$

Bei $\Delta t \to 0$ steht links jeweils der Differentialquotient nach t, rechts verschwindet $\frac{o(\Delta t)}{\Delta t}$ nach Definition und uns bleibt wie gewollt:

$$\frac{dP_n(t)}{dt} = \mu P_{n+1}(t) + \lambda P_{n-1}(t) - (\lambda + \mu)P_n(t), \;\; n \geq 1$$
$$\frac{dP_0(t)}{dt} = \mu P_1(t) - \lambda P_0(t).$$

□

Es stellt sich heraus, dass dieses System von Differenzen- und Differentialgleichungen sehr kompliziert zu lösen ist. Wir folgen der teilweise vereinfachten Herleitung aus Jain, Mohanty, Böhm (2007) und werden dafür einige Definitionen sowie tiefliegende Resultate aus unterschiedlichen Teilgebieten der Mathematik benötigen. Dazu zählen die Theorie von erzeugenden Funktionen, die Laplace Transformation sowie Kenntnisse komplexen Analysis. Die Laplace Transformation soll nun kurz erklärt werden:

Theorem 4.1 (Laplace Transformation)**.** *Gegeben sei $f : [0, \infty) \to \mathbb{C}$, mit der Bedingung, dass positive Konstanten C, s_0 existieren, sodass für alle $t > 0$ die Ungleichung $|f(t)| \leq Ce^{s_0 t}$ gilt. Dann existiert für alle $\theta \in \mathbb{C}$ die sogenannte Laplace Transformation von $f(t)$:*

$$\mathcal{L}\{f\}(\theta) = f^*(\theta) \coloneqq \int_0^\infty f(t)e^{-\theta t}dt.$$

Darüber hinaus ist die Transformation bis auf Nullmengen eindeutig.

Dieses Theorem ist eine fundamentale Aussage in der Theorie der Differentialgleichungen, welcher wir uns ohne hier einen Beweis anzugeben, bedienen werden. Der Beweis, zentrale Eigenschaften, wie zum Beispiel die Linearität, und Listen mit Transformationen bekannter Funktionen können unter Preuß (2002) sowie EqWorld (2005) gefunden werden.
Im Speziellen sind für uns folgende Formeln vom besonderen Interesse:

Lemma 4.2. *Es gilt*

$$\mathcal{L}\left\{e^{-at}f(t)\right\}(\theta) = F^*(\theta + a) \tag{7}$$

und für $n \in \mathbb{N}$

$$f^*(\theta) = (\sqrt{\theta^2 - a^2} + \theta)^{-n} \Leftrightarrow f(t) = na^{-n}t^{-1}I_n(at), \tag{8}$$

wobei $I_n(x)$ *die modifizierte Bessel-Funktion erster Art ist und wird beschrieben durch den Ausdruck*

$$I_n(x) \coloneqq \sum_{j=0}^{\infty} \frac{(x/2)^{n+2j}}{j!(n+j)!}.$$

An dieser Stelle ist es schon möglich das wichtigste Resultat dieses Abschnittes zu verraten:

Satz 4.1. *Die Lösung der Gleichung* (4) *ist gegeben durch*

$$\begin{aligned} P_n(t) = e^{-(\lambda+\mu)t}\Big(&\rho^{\frac{n}{2}} I_n(2\sqrt{\lambda\mu}t) + \rho^{\frac{n-1}{2}} I_{n+1}(2\sqrt{\lambda\mu}t) \\ &+ (1-\rho)\rho^n \sum_{j=n+2}^{\infty} \rho^{-\frac{j}{2}} I_j(2\sqrt{\lambda\mu}t)\Big). \end{aligned} \tag{9}$$

Doch, wie schon erwähnt, um diesen Satz zu beweisen, ist mehr Vorarbeit benötigt. Die Fakten über die Bessel-Funktion, welche gebraucht werden, sind im folgenden Lemma zusammengefasst.

Lemma 4.3. *Für die modifizierte Bessel-Funktionen erster Art gelten:*

$$\begin{aligned} I_{-n}(z) &= I_n(z) \\ \frac{2n}{z} I_n(z) &= I_{n-1}(z) - I_{n+1}(z) \\ I_n(z) &\sim \frac{e^z}{\sqrt{2\pi z}}, \end{aligned}$$

wobei die letzte Gleichung bedeutet, dass für fixe n *und* $z \to \infty$ *der Quotient der Funktionen gegen* 1 *strebt.*

Wir geben hier keinen Beweis an, diese Eigenschaften sind gut bekannt und lassen sich zum Beispiel in Jain, Mohanty, Böhm (2007) finden. Stattdessen schreiten wir fort und definieren weiter.

Definition 4.1. *Definiere die auf $|z| < 1$ holomorphe (also komplex differenzierbare) Funktion wie folgt:*

$$P(z,t) \coloneqq \sum_{n=0}^{\infty} P_n(t) z^n,$$

und die (dank Theorem 4.1 existierende) Laplace Transformation von dieser:

$$P^*(z,\theta) \coloneqq \int_0^{\infty} e^{-\theta t} P(z,t) dt.$$

Nun können wir zur Lösung der Gleichung (4) schreiten. Der Plan ist wie folgt:
Klarerweise ist das Ziel die Koeffizienten von z^n der Funktion $P(z,t)$ herauszufinden. Wir bemerken, dass wegen der Gleichung (4) diese erzeugende Funktion eine Differentialgleichung erfüllen muss, welche wir explizit angeben können. Mithilfe der Laplace Transformation wird diese in die transformierte Form äquivalent umgeschrieben und dann unter Anwendung eines Tricks gelöst. Dann müssen wir diese Lösung in eine Potenzreihe entwickeln und schließlich zurück transformieren. Das ermöglicht das Ablesen der gesuchten Koeffizienten.

Beweis vom Satz 4.1. Von der Gleichung (4) ausgehend, multiplizieren wir sie mit z^n, addieren über alle $n \geq 1$ auf und erhalten

$$\begin{aligned}\sum_{n=1}^{\infty} \frac{dP_n(t)}{dt} z^n &= \mu \sum_{n=1}^{\infty} P_{n+1}(t) z^n + \lambda \sum_{n=1}^{\infty} P_{n-1}(t) z^n - (\lambda+\mu) \sum_{n=1}^{\infty} P_n(t) z^n \\ &= \mu \frac{P(z,t) - P_1(t) z - P_0(t)}{z} + \lambda z P(z,t) - (\lambda+\mu)(P(z,t) - P_0(t)).\end{aligned}$$

Die linke Seite der Gleichung gibt, nach dem Vertauschen von Summe und Ableitung, sowie unter Verwendung der Anfangsbedingung aus Lemma 4.1:

$$\begin{aligned}\sum_{n=1}^{\infty} \frac{dP_n(t)}{dt} z^n &= \frac{\partial P(z,t)}{\partial t} - \frac{dP_0(t)}{dt} \\ &= \frac{\partial P(z,t)}{\partial t} + \lambda P_0(t) - \mu P_1(t).\end{aligned}$$

So erhalten wir nach dem Vereinfachen die partielle Differentialgleichung erster Ordnung

$$z \frac{\partial P(z,t)}{\partial t} = (1-z)\big((\mu - \lambda z) P(z,t) - \mu P_0(t)\big), \tag{10}$$

welche nun gelöst werden soll. Dabei bereitet uns $P_0(t)$ die Schwierigkeiten und so verwenden wir die Laplace Transformation, um unser Ziel zu erreichen. Die Transformation der rechten Seite ist dank der Linearität in t einfach; bei der linken Seite müssen wir partiell integrieren:

$$\begin{aligned}\mathcal{L}\Big\{z\frac{\partial P(z,t)}{\partial t}\Big\} &= z\int_0^\infty e^{-\theta t}\Big(\frac{\partial P(z,t)}{\partial t}\Big)dt \\ &= z\Big(e^{-\theta t}P(z,t)\Big|_{t=0}^{\infty} + \theta\int_0^\infty e^{-\theta t}P(z,t)dt\Big) \\ &= z(-1+\theta P^*(z,\theta)),\end{aligned}$$

wobei der letzte Schritt verwendet (vii), also, dass am Anfang niemand in der Schlange wartet, mit anderen Worten $P(z,0) = P_(0) = 1$. So erhalten wir

$$z(\theta P^*(z,\theta)-1) = (1-z)\big((\mu-\lambda z)P^*(z,\theta) - \mu P_0^*(\theta)\big),$$

und umgeformt

$$P^*(z,\theta) = \frac{z-\mu(1-z)P_0^*(\theta)}{(\lambda+\mu+\theta)z-\mu-\lambda z^2}. \tag{11}$$

Mit einem Trick soll $P_0^*(\theta)$ bestimmt werden. Da die linke Seite von der obigen Gleichung auf $z < 1$ holomorph (insbesondere stetig) ist, muss es die rechte auch sein. Wir uns aber gleich davon überzeugen, dass der Nenner genau eine Nullstelle in dieser Kreisscheibe hat und so, damit trotzdem kein Pol entsteht, muss der Zähler in diesem Punkt auch verschwinden. So extrahiert man $P_0^*(\theta)$.
Seien $z_1(\mu,\lambda,\theta)$ und $z_2(\mu,\lambda,\theta)$ die Nullstellen vom Nenner aus der Gleichung (11), wobei wir der Übersichtlichkeit halber diese kurz z_1 und z_2 bezeichnen. Es gilt wegen der positiven Diskriminante für $\theta > 0$:

$$\begin{aligned}z_1 &= \frac{\lambda+\mu+\theta-\sqrt{(\lambda+\mu+\theta)^2-4\lambda\mu}}{2\lambda} > 0 \\ z_2 &= \frac{\lambda+\mu+\theta+\sqrt{(\lambda+\mu+\theta)^2-4\lambda\mu}}{2\lambda} > z_1 > 0\end{aligned}$$

Weiters, aus der Formel von Vieta bekommen wir gleich die äquivalenten Bedingungen an die Nullstellen, mit denen es sich oft leichter rechnen lässt:

$$z_1+z_2 = \frac{\lambda+\mu+\theta}{\lambda} \qquad \text{und} \qquad z_1z_2 = \frac{\mu}{\lambda}. \tag{12}$$

Nun bemerken wir aber, dass die obige Gleichungen ein erstaunliches Geheimnis bergen:

$$\begin{aligned} z_1 + z_2 = \frac{\lambda+\mu+\theta}{\lambda} &= 1 + z_1 z_2 + \frac{\theta}{\lambda} > 1 + z_1 z_2 \\ \iff \qquad (1-z_1)(1-z_2) &< 0. \end{aligned}$$

Weil $0 < z_1 < z_2$ gilt, muss z_1 eine Nullstelle sein, die in der Einheitskreisscheibe $|z| < 1$ liegt, und wir erhalten aus den obigen Überlegungen, dass $z_1 - \mu(1-z_1)P_0^*(\theta) = 0$ sein muss. Das führt zu

$$P_0^*(\theta) = \frac{z_1}{\mu(1-z_1)}. \tag{13}$$

Die Laplace Transformierte Gleichung gelöst, bleiben uns noch zwei Schritte: $P^*(z,\theta)$ muss in eine Potenzreihe entwickelt und anschließend zurück transformiert werden.
Setzt man die Gleichung (13) in die Gleichung (11) ein, so und vereinfacht das Ergebnis es ergibt sich folgende Rechnung:

$$\begin{aligned} P^*(z,\theta) &= \frac{z - \frac{z_1(1-z)}{1-z_1}}{(\lambda+\mu+\theta)z - \mu - \lambda z^2} \\ &= \frac{z - z_1}{\lambda(1-z_1)(z-z_1)(z_2-z)} = \frac{1}{\lambda z_2(1-z_1)} \cdot \frac{1}{1-\frac{z}{z_2}}. \end{aligned}$$

Drückt man nun hier z_1 durch z_2, λ und μ mithilfe der Gleichung (12) aus und entwickelt das Ergebnis als geometrische Reihe, so ergibt sich für die Koeffizienten von z in $P^*(z,\theta)$:

$$P_n^*(\theta) = \frac{1}{\lambda}\rho^{n+1} \sum_{j=n+1}^{\infty} (\rho z_2)^{-j}, \tag{14}$$

wobei $\rho \coloneqq \lambda/\mu$.
Uns bleibt das Rücktransformieren von $P_n^*(\theta)$, im Speziellen müssen wir die inverse Laplace Transformation von z_2^{-n} berechnen. Dafür nehmen wir die explizite Formel für z_2 und wenden nacheinander die beiden in Lemma 4.2 beschriebenen Eigenschaften an:

$$\begin{aligned} \mathcal{L}^{-1}\left\{z_2^{-n}\right\} &= \mathcal{L}^{-1}\left\{\left(\frac{\lambda+\mu+\theta+\sqrt{(\lambda+\mu+\theta)^2-4\lambda\mu}}{2\lambda}\right)^{-n}\right\} \\ &= e^{-(\lambda+\mu)t}(2\lambda)^n \mathcal{L}^{-1}\left\{\left(\theta+\sqrt{\theta^2-4\lambda\mu}\right)^{-n}\right\} \\ &= e^{-(\lambda+\mu)t}(2\lambda)^n n(2\sqrt{\lambda\mu})^{-n} t^{-1} I_n(2\sqrt{\lambda\mu}t) \\ &= e^{-(\lambda+\mu)} n \rho^{n/2} t^{-1} I_n(2\sqrt{\lambda\mu}t). \end{aligned}$$

Das, zusammen mit der Gleichung (14) und den Eigenschaften aus Lemma 4.3, liefert schließlich

$$P_n(t) = e^{-(\lambda+\mu)t}\Big(\rho^{\frac{n}{2}} I_n(2\sqrt{\lambda\mu}t) + \rho^{\frac{n-1}{2}} I_{n+1}(2\sqrt{\lambda\mu}t) + (1-\rho)\rho^n \sum_{j=n+2}^{\infty} \rho^{-\frac{j}{2}} I_j(2\sqrt{\lambda\mu}t)\Big),$$

und damit auch den vollendeten Beweis. □

Auch wenn dieses Resultat bemerkenswert ist, so ist es höchstwahrscheinlich nicht zufriedenstellend. Der Ausdruck ist sehr kompliziert und unübersichtlich, obwohl wir ja absichtlich das Modell einfach halten wollten. Nun, die gute Nachricht ist, dass die Komplexität der Lösung verschwindet, sobald der Warteschlange Zeit gelassen wird um sich „einzupendeln“. Anders ausgedrückt wird der Ausdruck $P_n(t)$ für t gegen unendlich viel schöner, nämlich:

Satz 4.2. *Für die $M/M/1$ Warteschlange mit Parameter λ und μ, sodass $\rho \coloneqq \lambda/\mu < 1$, gilt:*

$$P_n \coloneqq \lim_{t\to\infty} P_n(t) = (1-\rho)\rho^n.$$

Beweis. Wir gehen von dem gerade gewonnenen Ausdruck für $P_n(t)$ aus und versuchen den Limes in Griff zu bekommen. Um das zu erfolgreich zu meistern, müssen wir die unendliche Summe von Bessel-Funktionen loswerden. Der Trick hierbei ist es zu bemerken, dass die Summanden nicht von n abhängen, was dazu führt, dass $P_n(t)$ und $P_{n+1}(t)$ sehr viele gleiche Terme haben. Diese wollen wir effektiv eliminieren und betrachten deshalb für $n \geq 0$

$$\begin{aligned}
P_{n+1}(t) - \rho P_n(t) &= e^{-(\lambda+\mu)t}\Big(\rho^{\frac{n}{2}} I_{n+2}(2\sqrt{\lambda\mu}t) - \rho^{\frac{n+2}{2}} I_n(2\sqrt{\lambda\mu}t) \\
&\qquad - (1-\rho)\rho^{n+1}\rho^{-\frac{n+2}{2}} I_{n+2}(2\sqrt{\lambda\mu}t)\Big) \\
&= e^{-(\lambda+\mu)t}\Big(-\rho^{\frac{n+2}{2}} I_n(2\sqrt{\lambda\mu}t) + \rho^{\frac{n+2}{2}} I_{n+2}(2\sqrt{\lambda\mu}t)\Big) \\
&= e^{-(\lambda+\mu)t}\rho^{\frac{n+2}{2}}\Big(I_{n+2}(2\sqrt{\lambda\mu}t) - I_n(2\sqrt{\lambda\mu}t)\Big) \\
&= -e^{-(\lambda+\mu)t}\rho^{\frac{n+2}{2}}(n+1)\frac{I_{n+1}(2\sqrt{\lambda\mu}t)}{\sqrt{\lambda\mu}t}.
\end{aligned}$$

Der letzte Schritt verwendet die zweite Beziehung in *Lemma* 4.3.
Lässt man jetzt t groß werden, so sieht man, dank der dritten Formel im gerade erwähnten Lemma, dass der Abstand von $P_{n+1}(t)$ und $\rho P_n(t)$ zwingend

gegen 0 gehen muss:

$$\begin{aligned}\lim_{t\to\infty}|P_{n+1}(t)-\rho P_n(t)| &= \lim_{t\to\infty}\frac{\rho^{\frac{n+2}{2}}(n+1)}{\sqrt{\lambda\mu}t}e^{-(\lambda+\mu)t}I_{n+1}(2\sqrt{\lambda\mu}t)\\ &= \lim_{t\to\infty}\frac{\rho^{\frac{n+2}{2}}(n+1)}{\sqrt{\lambda\mu}t}e^{-(\lambda+\mu)t}\frac{e^{\sqrt{2\lambda\mu}t}}{\sqrt{2\sqrt{\lambda\mu}t}}\\ &= \lim_{t\to\infty}\frac{\rho^{\frac{n+2}{2}}(n+1)}{\sqrt{\lambda\mu}t\sqrt{2\sqrt{\lambda\mu}t}}e^{-(\sqrt{\lambda}-\sqrt{\mu})^2t}=0,\end{aligned}$$

aus dem Grund, dass der Koeffizient von t in der Exponentialfunktion, also $-(\sqrt{\lambda}-\sqrt{\mu})^2$, stets kleiner oder gleich 0 ist.
Nun wollen wir daraus folgern, dass $P_{n+1}=\rho P_n$, doch um das zu behaupten, müssen wir sicherstellen, dass der Limes existiert. Das erledigt für uns eigentlich die Theorie der Markov-Ketten, doch weil wir diese in diesem Kapitel gezielt umgehen wollen, müssen wir wieder tricksen.
Aus der Anfangsbedingung in Lemma 4.1 und weil zum Zeitpunkt $t=0$ niemand in der Schlange wartet, folgern wir

$$\int_0^t \mu P_1(y)-\lambda P_0(y)dy=\int_0^t\frac{dP_0(y)}{dy}dy=P_0(t)-1.$$

Umformen und anschließendes Betrachten von $t\to\infty$ führt zu

$$\lim_{t\to\infty}P_0(t)=1-\mu\int_0^\infty \rho P_0(y)-P_1(y)dy=1-\mu\int_0^\infty e^{-(\lambda+\mu)y}\rho\frac{I_1(2\sqrt{\lambda\mu}y)}{\sqrt{\lambda\mu}y}dy$$

Dieses Integral können wir aber in Griff bekommen, indem wir bemerken, dass es nichts anderes als die Laplace Transformation der Funktion $a^{-1}t^{-1}I_1(at)$ in der neuen Unbekannten $\theta=\lambda+\mu$ ist, wobei $a=2\sqrt{\lambda\mu}$. Für diese haben wir ja eine Formel und so ergibt sich:

$$\begin{aligned}\lim_{t\to\infty}P_0(t) &= 1-2\lambda\mathcal{L}\Big\{(2\lambda\mu)^{-1}t^{-1}I_1(2\sqrt{\lambda\mu}t)\Big\}(\lambda+\mu)\\ &= 1-2\lambda(\sqrt{(\lambda+\mu)^2-4\lambda\mu}+\lambda+\mu)^{-1}\\ &= 1-\frac{\lambda}{\mu}=1-\rho,\end{aligned}$$

wobei wir beim Vereinfachen verwenden, dass $\lambda\le\mu$.
Nun ist wegen $\lim_{t\to\infty}|P_{n+1}(t)-\rho P_n(t)|=0$ und Induktion klar, dass P_n für jedes $n\ge 0$ existiert und es gilt $P_{n+1}=\rho P_n$. Das führt dann nach dem Verschieben vom Index und n Mal Anwenden auf $P_n=\rho^nP_0=\rho^n(1-\rho)$. □

4.1.2 Eigenschaften

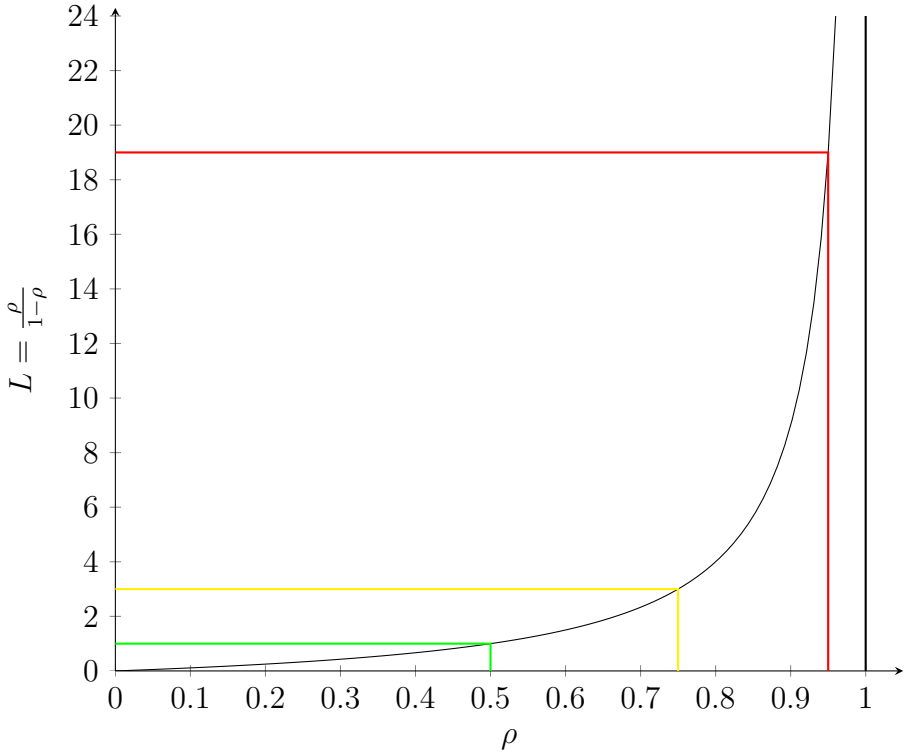
Tatsächlich ist, wie man sieht, das Ergebnis also eine Wahrscheinlichkeitsverteilung für $\rho < 1$, weil dann die Summe über alle natürlichen Zahlen n genau 1 ergibt. Dieses Resultat ist sehr bemerkenswert, denn es besagt, dass die Wahrscheinlichkeiten für die Anzahl von Kunden im System sich mit der Zeit einpendeln. Das würde die Theorie der Ketten von Markov direkt verraten, hier wurde aber den Beweis ohne Verwendung dieser geführt. Diese Verteilung wird stationäre Lösung genannt und ihr Aussehen, sowie die Auswirkungen sollen nun beschrieben werden.
Die Eigenschaften der resultierenden Verteilung sind wohl bekannt unter dem Namen der geometrischen Wahrscheinlichkeitsverteilung. Bevor die wichtigsten Folgerungen niedergeschrieben werden, soll an die passende und, wegen obigen Überlegungen, wohldefinierte Notation erinnert werden:

$$\begin{aligned} X &: \lim_{t\to\infty} X(t) \\ L &: \text{Erwatungswert von } X. \end{aligned}$$

L gibt also den Erwartungswert für die Anzahl der Kunden im *System* nach langer Zeit an.
Der eigentlich bekannte Erwartungswert dieser geometrischen Verteilung ist leicht nachgerechnet:

$$L = \mathbb{E}(X) = \sum_{n=0}^{\infty} n\rho^n(1-\rho) = (1-\rho)\cdot\frac{\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho}. \tag{15}$$

Diese Gleichung wird viel anschaulicher, wenn man sie graphisch darstellt:



Man erblickt gleich, wie rasant die erwartende Anzahl von Kunden im System bei $\rho \nearrow 1$ zunimmt. Für $\lambda/\mu = 0.5$ ist diese gleich 1, bei dem Bearbeitungsgrad von 0.75 erwartet man schon durchschnittlich 3 Kunden im System und für $\rho = 0.95$ ist das System im Erwartungswert mit 19 Personen in der Praxis schon überfüllt.

Eine weitere sehr interessante Kennzahl ist L_q, welche den Erwartungswert von Kunden in der *Schlange* nach langer Zeit angeben soll. Weil der Fall mit nur einem Server betrachtet wird, lässt sich dieser Wert auch leicht auf folgende Weise bestimmen:

$$L_q = \sum_{n=1}^{\infty}(n-1)P_n = L - \rho = \frac{\rho^2}{1-\rho}. \tag{16}$$

Die Wahrscheinlichkeit, dass langer Zeit sich mindestens ein Kunde im System befindet ist in der Praxis auch oft spannend und lässt sich einfach ablesen:

$$\mathbb{P}(\text{Die Warteschlange ist nicht leer im stationären Zustand}) = 1 - P_0 = \rho$$

Schließlich erhält man durch Nachrechnen oder Nachschlagen der Formel für die geometrische Verteilung, zum Beispiel auf der Seite 398 in Jain, Mohanty,

Böhm (2007), auch die Varianz von X:

$$\mathrm{Var}(X) = \sigma^2 = \frac{\rho}{1-\rho} + \frac{\rho^2}{1-\rho^2} \tag{17}$$

Die oben angeführten Formeln sind eher aus der Managementsicht relevant. Um die Sicht der Kunden zu berücksichtigen greift man am besten auf die Formel von Little (Theorem 3.1) zurück und erhält für die durchschnittliche Wartezeit in der Warteschlange W_q:

$$W_q = \frac{1}{\lambda} L_q = \frac{1}{\lambda} \frac{\rho^2}{1-\rho} \tag{18}$$

4.1.3 Anwendung

Das gewonnene Wissen über $M/M/1$ Warteschlangen lässt sich nun an einem Beispiel gut veranschaulichen. Die Situation in einem Blumengeschäft soll modelliert werden mit folgenden Annahmen: Im Durchschnitt besuchen 15 Kunden den Laden stündlich und die Verkäuferin schafft es durchschnittlich 20 Kunden in einer Stunde zu bedienen. Sowohl die Ankünfte als auch die Bedienungen sind unabhängig voneinander und Poisson verteilt. Das Geschäft öffnet um $8^{\underline{00}}$ Uhr in der Früh.
Die Angabe impliziert also gleich, dass es sich um eine $M/M/1$ Schlange handelt, $\mu = 20$, $\lambda = 15$ und damit $\rho = 3/4$.

a) *Wie hoch ist die Wahrscheinlichkeit, dass um $8^{\underline{15}}$ Uhr sich genau ein Kunde im Geschäft befindet?*
Gesucht ist also $P_1(1/4)$ und man erhält nach dem Einsetzen in Gleichung (9):

$$\begin{aligned} P_1(10) = e^{-35/4}\Big(&\sqrt{0.75} I_1(\sqrt{300}/2) + I_2(\sqrt{300}/2) \\ &+ 0.75(1-0.75)\sum_{j=3}^{\infty} 0.75^{-\frac{j}{2}} I_j(\sqrt{300}/2)\Big) \\ \approx\ & 31.33\%^{1}. \end{aligned}$$

b) *Wie hoch ist die Wahrscheinlichkeit, dass sich zu einem fixen Zeitpunkt am Nachmittag eine Schlange von 7 Personen oder mehr bildet?*

[1] Dieser Wert wurde mit Wolfram Mathematica ausgerechnet.

Weil es sich hier um einen Zeitpunkt handelt, welcher genügend weit in der Zukunft liegt, kann man die stationäre Lösung verwenden:

$$\sum_{n=8}^{\infty} P_n = \sum_{n=8}^{\infty} \rho^n(1-\rho) = (1-0.75)\frac{0.75^8}{1-0.75} \approx 10.01\%,$$

wobei ab 8 summiert werden muss, weil bei einer einzigen Verkäuferin 7 Personen in der Warteschlange gleichbedeutend mit 8 Kunden im System ist.

c) *Wie lange ist die zu erwartende Schlange im Blumengeschäft?*
Gesucht ist L_q, für welches schon in Gleichung (16) eine Formel gefunden wurde:

$$L_q = \frac{0.75^2}{1-0.75} = 2.25.$$

Also es warten durchschnittlich 2,25 Personen in der Schlange.

d) *Wie lange wartet ein Kunde durchschnittlich bis er bedient wird?*
Gesucht ist also W_q, wofür die Gleichung (18) eine Formel liefert:

$$W_q = \frac{1}{15} \cdot \frac{0.75^2}{1-0.25} = 0.15 \text{ h} = 9 \text{ min}.$$

Die durchschnittliche Wartezeit in der Schlange beträgt also 9 Minuten.

Beim dem im letzten Kapitel besprochenen Modell wurden zwei restriktive Annahmen getroffen: Die Verteilungen sowohl der Ankünfte von Kunden als auch die der Bearbeitungszeiten wurden Poisson beziehungsweise exponentiell vorausgesetzt. Auch wenn diese Annahmen in der Praxis oft erfüllt sind, gibt es Situationen welche allgemeinere Modelle benötigen. Im nächsten Kapitel soll nun ein nicht Markov'sches Modell vorgestellt, erörtert und gelöst werden. Ironischerweise wird die Theorie der Ketten von Markov genau hier essentiell, denn das Umgehen dieser mit mathematischen Tricks wird nicht mehr möglich sein.

4.2 Das M/G/1 Modell

Dieser Abschnitt handelt vom $M/G/1$ Modell. Wie schon im Kapitel 2 beschrieben, steht diese Notation für ein 1-Server-Warteschlangensystem, bei dem der Ankunftsprozess Poisson verteilt ist und aber für die Verteilung der Bearbeitungszeiten keine restriktiven Annahmen, außer der Unabhängigkeit von einander, herrschen. So gesehen ist das eine direkte Verallgemeinerung

des im letzten Abschnitt gelösten $M/M/1$ Systems. Weil nun nicht mehr beide Zufallsprozesse gedächtnislos seien müssen, spricht man hier von einem *nicht Markov'schen* Modell. Es stellt sich heraus, dass das Fehlen der Annahmen über die Verteilung der Bearbeitungszeiten die Herleitung komplizierter macht und zwingt die Theorie der Markov'schen Ketten zu benutzen. Das bedeutet, im Kontrast zum letzten Kapitel, dass nur die stationäre Lösung exakt berechnet werden kann, diese hat trotzdem in der Praxis enorme Bedeutung und ist berühmt unter dem Namen *Pollaczek-Khinchin Formel*. Mithilfe des Wissens über die Ketten von Markov kann gezeigt werden, dass $\lim_{t\to\infty} P_n(t)$ existiert und es wird gleich dieser Grenzwert betrachtet und explizit gelöst. Genau ist diese Herleitung im nächsten Abschnitt erläutert.

4.2.1 Herleitung

Um das $M/G/1$ Modell in Griff zu bekommen, müssen wir einen Markov Prozess innerhalb dieses stochastischen Prozesses finden. Betrachten wir nur $X(t + \Delta t)$, also die Anzahl der Kunden im System zum Zeitpunkt $t + \Delta t$, so merkt man schnell: das ist bei dem $M/G/1$ Modell im Allgemeinen keine Markov Kette. Das liegt daran, dass diese Anzahl der Kunden nicht nur von jener zum Zeitpunkt t abhängt, sondern, weil die Verteilung der Bearbeitungszeiten beliebig ist, noch von der Zeit wie lange der aktuelle Kunde bereits serviert wird. Um dieses Problem zu lösen, betrachtet man nur jene Zeitpunkte, bei denen Kunden das System verlassen. An diesen Zeitpunkten ist die Bearbeitungszeit des neuen Kunden gleich Null und so ist die Anzahl der Kunden im System zu diesen besonderen Zeitpunkten nur abhängig von der Anzahl der Kunden bei dem letzten solchen Zeitpunkt und der Anzahl im Bearbeitungszeitraum neu dazu gekommenen Kunden: jedenfalls unabhängig von der Verteilung der Bearbeitungszeit.
Um das oben beschriebene exakt zu fassen benötigen wir wieder neue Notation:

C_m	:	Der m-te Kunde
X_m	:	Anzahl der Kunden im System beim Abgang von C_m
X_0	:	Anzahl der Kunden im System am Anfang
γ_m	:	Anzahl der Ankünfte während der Bearbeitungszeit von C_m

Der Einfachheit halber werden wir annehmen, dass die Verteilungsfunktion von der Bearbeitungszeit $F_T(t)$ absolut stetig bezüglich des Lebesgue-Maßes ist, denn dann, berufend auf den Satz von Radon-Nikodým, können wir annehmen, dass diese eine Wahrscheinlichkeitsdichtefunktion $f_T(t)$ besitzt. Die Resultate aus diesem Kapitel können auch ohne dieser Annahme hergeleitet werden, doch dann muss man mit Laplace–Stieltjes Integralen arbeiten,

was eindeutig über den Rahmen dieser Arbeit gehen würde. Die meisten in der Praxis auftauchenden Verteilungsfunktionen für Zeitintervalle besitzen eine Wahrscheinlichkeitsdichtefunktion, deshalb ist diese Einschränkung keine große. Weiters fordern wir, dass der Erwartungswert und die Varianz von $F_T(t)$ endlich sind und nennen den ersteren $\frac{1}{\mu}$, um Analogien zum Model im vorigen Kapitel zu sehen. Die Varianz bekommt den naheliegenden Namen σ^2.
Um die Verteilung von γ besser zu verstehen, definieren wir

$$\alpha_n \coloneqq \mathbb{P}(\gamma_m = n)$$
$$\alpha(z) \coloneqq \mathbb{E}(z^{\gamma_m}) = \sum_{j=0}^{\infty} \alpha_j z^j.$$

also α_n beschreibt die Wahrscheinlichkeit, dass die Anzahl der Ankünfte von Kunden während der Bearbeitungszeit des m-ten Kunden gleich n ist. Weil die γ_m alle unabhängige und identisch verteilte Zufallsvariablen sind, sieht man ein, dass α_n tatsächlich von m unabhängig ist. $\alpha(z)$ ist einfach die dazugehörende erzeugende Funktion.
Die Ankünfte bilden einen Poisson Prozess mit Parameter λ, deshalb ist die bedingte Wahrscheinlichkeit für γ_m zum Zeitpunkt t nach Definition gegeben durch

$$\mathbb{P}(\gamma_m = n | T = t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!},$$

was für die Wahrscheinlichkeiten α_n folgende Gleichung impliziert:

$$\alpha_n = \int_0^{\infty} \frac{(\lambda t)^n e^{-\lambda t}}{n!} f_T(t) dt.$$

Betrachtet man nun $\alpha(z)$ und vertauscht, legitimerweise dank dem Satz von Fubini, Summe und Integral erhält man

$$\begin{aligned}
\alpha(z) &= \sum_{j=0}^{\infty} \alpha_j z^j = \sum_{j=0}^{\infty} \int_0^{\infty} \frac{(\lambda t)^j e^{-\lambda t}}{j!} z^j f_T(t) dt \\
&= \int_0^{\infty} \left(\sum_{j=0}^{\infty} \frac{(\lambda t z)^j}{j!} \right) e^{-\lambda t} f_T(t) dt \\
&= \int_0^{\infty} e^{\lambda t z} e^{-\lambda t} f_T(t) dt = \int_0^{\infty} e^{\lambda t (z-1)} f_T(t) dt.
\end{aligned}$$

Damit können wir nun überraschend schön den Erwartungswert von γ_m berechnen:

$$\begin{aligned}\mathbb{E}(\gamma_m) = \alpha'(1) &= \frac{\partial}{\partial z}\Big(\int_0^\infty e^{\lambda t(z-1)} f_T(t)dt\Big)\Big|_{z=1} \\ &= \int_0^\infty \frac{\partial}{\partial z} e^{\lambda t(z-1)}\Big|_{z=1} f_T(t)dt = \lambda \int_0^\infty t f_T(t)dt \\ &= \lambda \frac{1}{\mu} =: \rho. \end{aligned} \tag{19}$$

Für die Varianz müssen wir ein wenig mehr arbeiten, schaffen diese aber auf eine ähnliche Art zu bestimmen:

$$\begin{aligned}\mathrm{Var}(\gamma_m) &= \alpha''(1) + \alpha'(1) - \big(\alpha'(1)\big)^2 \\ &= \int_0^\infty \frac{\partial^2}{\partial z} e^{\lambda t(z-1)}\Big|_{z=1} f_T(t)dt + \rho - \rho^2 \\ &= \lambda^2 \int_0^\infty t^2 f_T(t)dt + \rho - \rho^2 \\ &= \lambda^2\big(\sigma^2 + \frac{1}{\mu^2}\big) + \rho - \rho^2 = \lambda^2\sigma^2 + \rho. \end{aligned} \tag{20}$$

Nun können wir eine leichte, aber für dieses Kapitel zentrale, Rekursion für X_m niederschreiben. Dazu unterscheiden wir ob C_m das System leer oder nicht verlässt. Im zweiten Fall hinterlässt C_{m+1} im System einerseits einen Kunden weniger als sein Vorgänger - weil er ja abgeht - andererseits kommen genau γ_{m+1} neue Kunden während seiner Bearbeitungszeit dazu. Das liefert:

$$X_{m+1} = X_m - 1 + \gamma_{m+1}, \quad X_m > 0.$$

Der erste Fall ist noch einfacher zu beschreiben: Wenn C_m das System leer verlassen hat, so wird C_{m+1} genau jene Kunden hinterlassen, welche in seiner Bearbeitungszeit angekommen sind, also γ_{m+1}:

$$X_{m+1} = \gamma_{m+1}, \quad X_m = 0.$$

Diese zwei Gleichungen lassen sich kompakt schreiben als

$$X_{m+1} = X_m - \delta(X_m) + \gamma_{m+1}, \tag{21}$$

wobei

$$\delta(X_m) = \begin{cases} 0, & \text{falls } X_m = 0 \\ 1, & \text{falls } X_m > 0. \end{cases}$$

Wie schon im ersten Absatz dieses Abschnittes angedeutet, sieht man nun hier explizit, dass X_{m+1} nur von Ankunftsprozessen im Zeitraum ab dem Abgang des letzten Kunden abhängig ist. Weil wir annehmen, dass die Ankünfte Poisson verteilt sind, folgt damit, dass die Folge $\{X_m\}_{m\geq 0}$ tatsächlich eine Markov Kette bildet. Weiters, kann überprüft werden, dass diese Markov Kette zwei zentrale Eigenschaft aufweist: sie ist irreduzibel und positiv rekurrent. In dieser Arbeit wird darauf verzichtet diese Eigenschaften zu definieren und zu beweisen, denn diese bauen auf tiefgründigen mathematischen Erkenntnissen und Theorien auf, die zu kompliziert und technisch für den gestellten Zweck wären. Es soll auch nur angemerkt werden, dass diese Charakterisierung der Markov Kette impliziert, dass der Grenzwert der Folge $\{X_m\}_{m\geq 0}$ - die so-genannte stationäre Lösung - tatsächlich existiert. Die Details zu diesen, ohne Beweis angeführten Behauptungen finden sich in Bhat (2008) auf den Seiten 77ff. Wir halten somit aber nur fest, dass die folgende Notation gerechtfertigt ist:

$$\begin{aligned} P_m(n) &:= \mathbb{P}(X_m = n) \\ \Pi^{(n)} &:= \lim_{m\to\infty} \mathbb{P}(X_m = n) \end{aligned}$$

also $P_m(n)$ soll die Wahrscheinlichkeit dafür angeben, dass beim Abgang des m-ten Kunden noch genau n andere im System sind. Diesen Wert explizit auszurechnen ist bei dem $M/G/1$ Modell leider zu optimistisch, deshalb betrachten wir die Gleichgewichtslösung $\Pi^{(n)}$ und versuchen diese in Griff zu bekommen. Wie schon im letzten Kapitel, ist der effektivste Ansatz dafür die Potenzreihe zu definieren und mit dieser weiterzuarbeiten. Also:

$$\begin{aligned} G_m(z) &:= \mathbb{E}(z^{X_m}) = \sum_{j=0}^{\infty} P_m(j) z^j \\ \Pi(z) &:= \lim_{m\to\infty} G_m(z) = \sum_{j=0}^{\infty} \Pi^{(j)} z^j. \end{aligned}$$

Nun folgt aus der Gleichung (21) und der Multiplikativität des Erwartungswertes, dass

$$\begin{aligned} G_{m+1}(z) &= \mathbb{E}(z^{X_m - \delta(X_m) + \gamma_{m+1}}) \\ &= \mathbb{E}(z^{X_m - \delta(X_m)}) \mathbb{E}(z^{\gamma_{m+1}}) \\ &= \mathbb{E}(z^{X_m - \delta(X_m)}) \alpha(z). \end{aligned}$$

Es gilt aber auch nach Definition von δ, dass

$$\begin{aligned}\mathbb{E}(z^{X_m-\delta(X_m)}) &= \mathbb{P}(X_m=0)+\sum_{j=1}^{\infty}\mathbb{P}(X_m=j)z^{j-1}\\ &= \mathbb{P}(X_m=0)+\frac{G_m(z)-\mathbb{P}(X_n=0)}{z}.\end{aligned}$$

Eingesetzt in die obige Gleichung erhält man folgende Rekursionsgleichung für $G_m(z)$:

$$G_{m+1}(z)=\alpha(z)\Big(\mathbb{P}(X_m=0)+\frac{G_m(z)-\mathbb{P}(X_m=0)}{z}\Big).$$

Diese kann nun leider nicht exakt gelöst werden, was man aber schon machen kann, ist den Limes $n\to\infty$ zu betrachten um die Beziehung zu erhalten:

$$\Pi(z)=\alpha(z)\Big(\Pi(0)+\frac{\Pi(z)-\Pi(0)}{z}\Big),$$

und nach $\Pi(z)$ umgeformt

$$\Pi(z)=\frac{(z-1)\alpha(z)}{z-\alpha(z)}\Pi(0). \tag{22}$$

Weil sowohl $\Pi(z)$ als auch $\alpha(z)$ wahrscheinlichkeitserzeugende Funktionen sind, wissen wir, dass $\Pi(1)=\alpha(1)=1$ und die Gleichung (19) verrät uns, dass $\alpha'(1)=\rho$. Betrachtet nun in der obigen Formel $z\to 1$ so erhält man mithilfe der Regel von de l'Hospital, dass

$$1=\lim_{m\to\infty}\frac{\alpha(z)+(z-1)\alpha'(z)}{1-\alpha'(z)}\Pi(0)=\frac{1}{1-\rho}\Pi(0),$$

woraus natürlich folgt, dass $\Pi(0)=1-\rho$. Eingesetzt in die Gleichung (22), erhält man die berühmte Pollaczek-Khinchin Formel:

$$\Pi(z)=\frac{(z-1)\alpha(z)}{z-\alpha(z)}(1-\rho). \tag{23}$$

Will man nun wissen, wie groß der Erwartungswert für die Länge der Warteschlange im stationären Zustand beim Abgang eines Kunden ist, so sucht man eigentlich $\Pi'(1)$. Die logarithmische Ableitung von der Gleichung (23) liefert

$$\frac{\Pi'(z)}{\Pi(z)}=\frac{1}{z-1}+\frac{\alpha'(z)}{\alpha(z)}-\frac{1-\alpha'(z)}{z-\alpha(z)}.$$

Setzt man nun einfach $z = 1$ ein, um auf der linken Seite das gesuchte $\Pi'(1)$ zu erhalten, sieht man gleich ein, dass die rechte Seite dieser Gleichung leider das undefinierte $\infty + \rho - \infty$ liefert. Deshalb betrachten wir wieder den Limes:

$$\lim_{z\to 1}\Big(\frac{1}{z-1}+\frac{\alpha'(z)}{\alpha(z)}-\frac{1-\alpha'(z)}{z-\alpha(z)}\Big) = \rho + \lim_{z\to 1}\Big(\frac{1}{z-1}-\frac{1-\alpha'(z)}{z-\alpha(z)}\Big)$$
$$= \rho + \lim_{z\to 1}\frac{\alpha'(z)(z-1)-\alpha(z)+1}{(z-1)(z-\alpha(z))} = \rho + \lim_{z\to 1}\frac{\alpha''(z)(z-1)}{z-\alpha(z)+(z-1)(1-\alpha'(z))}$$
$$= \rho + \lim_{z\to 1}\frac{\alpha'''(z)(z-1)+\alpha''(z)}{2-2\alpha'(z)-(z-1)(\alpha''(z))} = \rho + \frac{\alpha''(1)}{2-2\rho} = \rho + \frac{\rho^2+\lambda^2\sigma^2}{2(1-\rho)}.$$

Bei dieser nicht schwierigen aber langwierigen Rechnung wurde zwei Mal die Regel von de l'Hospital angewandt, sowie verwendet, dass $\alpha(1) = 1$, $\alpha'(1) = \rho$ und $\alpha''(1) = \rho^2 + \lambda^2\sigma^2$, was aus der Gleichung (20) bekannt ist. Zusammen mit der Formel darüber und der Tatsache, dass $\Pi(1) = 1$ erhalten wir also

$$L := \Pi'(1) = \rho + \frac{\rho^2+\lambda^2\sigma^2}{2(1-\rho)}. \tag{24}$$

Die Formel, welche in der Praxis aber die meiste Bedeutung hat, ist der Erwartungswert für die Länge der *Warteschlange* zu einem beliebigen Zeitpunkt, weit weg in der Zukunft, der im letzten Kapitel L_q getauft wurde. Es lässt sich beweisen, dass dieser gleich dem Erwartungswert für die Länge der Warteschlange zu *Ankunftszeiten* ist, welchen wir mithilfe der der obigen Formel finden. Dazu definiert man N_q: der Gleichgewichtszustand für die Anzahl der Kunden in der Warteschlange zu Ankunftszeitpunkten, sowie analog zu $\Pi^{(n)}$ die Wahrscheinlichkeit, dass ein gerade angekommener Kunde eine Schlange der Länge n auffindet und nennt diese $P^{(n)}$. Es lässt sich ebenso beweisen, dass $\Pi^{(n)} = P^{(n)}$, dazu soll aber in dieser Arbeit auf Jain, Mohanty, Böhm (2008) Seiten 92ff. verwiesen werden. Hat man diese Beziehung, so folgt direkt

$$L_q := \mathbb{E}(N) = \sum_{n=1}^{\infty}(n-1)P^{(n)} = \sum_{n=1}^{\infty}(n-1)\Pi^{(n)} = \sum_{n=1}^{\infty}n\Pi^{(n)} - (1-\Pi^{(0)}).$$

Der erste Term rechts ist genau der Erwartungswert den wir in der Gleichung (24) ausgerechnet haben, von diesem wird und der zweite Term abgezogen, der gleich ρ ist. Es folgt somit eine andere Form der Pollaczek-Khinchin Formel:

$$L_q = \mathbb{E}(N_q) = \frac{\rho^2+\lambda^2\sigma^2}{2(1-\rho)}. \tag{25}$$

Wie schon im vorigen Kapitel, kann mithilfe der Formel von Little die durchschnittliche Wartezeit für Kunden bestimmt werden. Wir bezeichnen diese wieder mit W_q und erhalten somit nach Anwendung von Theorem 3.1

$$W_q = \frac{1}{\lambda} L_q = \frac{\rho^2 + \lambda^2 \sigma^2}{2\lambda(1-\rho)}. \tag{26}$$

Die Formeln für diese Kennzahlen hergeleitet, soll nun anhand von einigen Beispielen deren Wichtigkeit illustriert werden.

4.2.2 Anwendung

Da die Pollaczek-Khinchin Formel als Verallgemeinerung einiger Resultate aus dem letzten Kapitel gesehen werden kann, überprüfen wir zuerst, ob sie für den Fall des $M/M/1$ Modells tatsächlich dasselbe Ergebnis liefert. Ist die Bearbeitungszeit exponentiell verteilt mit Parameter μ, so gilt bekanntlich, dass $\sigma^2 = \frac{1}{\mu^2}$ (Forbes, Evans, Hastings, Peacock (2011)) und somit tatsächlich wie schon in der Gleichung (16):

$$L_q = \frac{\rho^2 + \lambda^2 \frac{1}{\mu^2}}{2(1-\rho)} = \frac{\rho^2}{1-\rho}.$$

Spannend wird es nun, wenn man andere Verteilungen für die Bearbeitungszeit voraussetzt. Eine naheliegende Annahme wäre dabei gewiss jene der deterministischen Verteilung, also des $M/D/1$ Modells. Das bedeutet wir wollen voraussetzten, dass eine Maschine die Kundenanfragen bearbeitet und für jeden Kunden eine konstante Zeit von genau $\frac{1}{\mu}$ Zeiteinheiten benötigt. Das impliziert natürlich, dass μ Kunden pro Zeiteinheit bearbeitet werden und, weil die Bearbeitungszeit stets dieselbe bleibt, dass $\sigma^2 = 0$. Wir erhalten somit für das $M/D/1$ Modell:

$$L_q = \frac{1}{2} \frac{\rho^2}{1-\rho}$$
$$W_q = \frac{1}{2} \frac{\rho^2}{\lambda(1-\rho)}.$$

Die Warteschlange ist also bei konstanter Bearbeitungszeit im Durchschnitt halb so lang und die Wartezeit in dieser halb so kurz als wenn diese exponentiell verteilt wäre. Diese Tatsache ist höchst erstaunlich. Würde die Blumenverkäuferin aus dem Beispiel im letzten Abschnitt für jeden Kunden *genau* 3 Minuten brauchen und nicht nur durchschnittlich, so wäre die

Wartezeit lediglich 4.5 Minuten und nicht 9. Allgemein, schreibt man die Gleichung (26) um, zu

$$W_q = \frac{\rho^2}{2\lambda(1-\rho)} + \sigma^2 \frac{\lambda}{2-2\rho},$$

sieht man gleich, dass, ceteris paribus, je größer die Standartabweichung der Servicezeiten, desto länger die Wartezeit. Dieses, unter dem Namen Wartezeitparadoxon bekannte Phänomen, ist sehr kontraintuitiv, denn man möchte erwarten, dass die Wartezeit nur von der durchschnittlichen Servicezeit und nicht von deren Varianz abhängt. Akzeptiert man jedoch diese Tatsache durch gründliches Nachdenken über den Sachverhalt oder durch das Herleiten der Formel, scheint es andererseits durchaus beachtlich, dass das Wissen des ersten Moments und des zweiten zentralen Moments tatsächlich schon die durchschnittliche Wartezeit verrät. In der Praxis hat diese Tatsache enorme Bedeutung: oft reicht es die Abweichungen vom Durchschnitt bei den Servicezeiten, also deren Varianz, zu vermindern, um merkliche Wohlfahrtsgewinne bei Warteschlangen zu erreichen.

Ein überaus spannendes Beispiel für die gerade erarbeitete Theorie findet sich im Buch von Bhat (2008) auf den Seiten 93ff. und soll hier vorgestellt werden:
In einer Autowerkstatt arbeitet ein Mechaniker. Historische Daten verraten statistisch, dass für die Wahrscheinlichkeiten der Anzahl neuer Ankünfte während der Bearbeitungsperiode eines Autos folgendes gilt:

$$\begin{aligned} \mathbb{P}(\text{keine neue Ankunft}) &= 50\%, \\ \mathbb{P}(\text{eine neue Ankunft}) &= 30\%, \\ \mathbb{P}(\text{zwei neue Ankünfte}) &= 20\%. \end{aligned}$$

Unter der (plausiblen) Annahme, dass die Ankünfte Poisson verteilt sind, können wir mit unserem hergeleiteten Wissen diese Situation modellieren und interessante Kennzahlen extrahieren. Weil keine sonstigen Informationen über den Mechaniker und seine Bearbeitungszeiten vorliegen, handelt es sich hier klarerweise um eine $M/G/1$ Warteschlange. Die Angabe impliziert weiters unmittelbar, dass:

$$\alpha_0 = 0.5 \qquad \alpha_1 = 0.3 \qquad \alpha_2 = 0.2,$$

und somit ist $\alpha(z) = 0.5 + 0.3z + 0.2z^2$. Man erhält auch $\rho = \alpha'(1) = 0.7$. Eingesetzt in die Pollaczek-Khinchin Formel (Gleichung (23)) ergibt sich für

$\Pi(z)$ folgender vereinfachter Ausdruck:

$$\Pi(z) = 0.3\frac{(z-1)(0.5+0.3z+0.2z^2)}{-0.5+0.3z-0.2z^2} = \frac{15+9z+6z^2}{50-20z}.$$

Diese erzeugende Funktion liefert gleich sämtliche interessante Informationen über den zu modellierenden Sachverhalt:

$$\begin{aligned} L &= \Pi'(1) = \frac{41}{30} \approx 1.367, \\ L_q &= L - \rho = \frac{2}{3} \approx 0.667, \\ W_q &= \frac{L_q}{0.7} = \frac{20}{21} \approx 0.952 \text{ Bearbeitungszeiteinheiten.} \end{aligned}$$

Dabei wurde für die durchschnittliche Wartezeit W_q die durchschnittliche Bearbeitungszeit als eine Zeiteinheit verwendet.
Wir haben es also geschafft die zu erwartende Anzahl von Kunden im System und in der Warteschlange auszurechnen, sowie die durchschnittliche Verweildauer der Kunden in der Schlange. Die letzte Kennzahl offenbart, dass bei diesem Beispiel ein Kunde im Erwartungswert in der Warteschlange fast genau so lange verbringt wie er danach bedient wird.

4.3 Andere Modelle und Ausblick

Die im obigen Kapitel besprochenen Modelle lassen sich erheblich verallgemeinern und sollten eigentlich nur einen kurzen Einblick in das Thema der Warteschlangentheorie liefern.
Eine nicht schwierige und naheliegende Verallgemeinerung des $M/M/1$ Modells ist jene zum $M/E_k/1$ oder zum $E_k/M/1$. E_k steht für die Erlang Verteilung, welche nach A. K. Erlang benannt ist. Diese beschreibt die Summe von k unabhängigen identisch und exponentiell verteilten Zufallsvariablen und ist in der Warteschlangentheorie besonders wichtig, weil ein Bearbeitungsprozess oft als Summe von solchen unabhängigen Prozessen gesehen werden kann. Weil die Gedächtnislosigkeit der Exponentialfunktion bei Summenbildung bestehen bleibt, ist bei diesen Modellen die Markoveigenschaft automatisch erfüllt und so ist die Herleitung dieser Verallgemeinerung nicht sonderlich kompliziert.
In dieser Arbeit wurde für den Ankunftsprozess immer nur die Poisson Verteilung angenommen, denn, wie schon erwähnt, ist das in den meisten Fällen gerechtfertigt. Es existieren dafür aber auch allgemeinere Modelle, abgesehen vom $E_k/M/1$, um diesen Sachverhalt abzudecken. Zum Beispiel beschreibt

das Modell $M^{[X]}/\cdot/1$ zwar immer noch einen Poisson verteilten Ankunftsprozess, bei welchem jedoch die Tatsache beachtet wird, dass Kunden manchmal in Gruppen ankommen können. Die Größe dieser Gruppen wird mit der Zufallsvariable X beschrieben. Will man noch allgemeinere Modelle, so existieren Theorie und Resultate für die $G/M/1$ und sogar $G/G/1$ Warteschlangenmodelle, also jene, bei denen der Ankunftsprozess einer beliebigen Wahrscheinlichkeitsverteilung folgt. An dieser Stelle muss jedoch angemerkt werden, dass je weniger man über eine Warteschlange annimmt, desto schwieriger wird die Herleitung und desto unüberschaubarer die Formeln. Extrem wird das bei dem $G/G/1$ Modell, bei welchem man schließlich nur mehr auf Abschätzungen angewiesen ist (Gross, Shortle, Thompson, Harris (2008) sowie Seiten 168ff. in Bhat (2008)).
Ein weiterer Aspekt für Verallgemeinerung ist, dass in dieser Arbeit nur die Situation eines einzelnen Servers betrachtet wurde, was in der Praxis oft anders gehandhabt wird. Es lohnt sich deshalb Warteschlangenmodelle der Form $\cdot/\cdot/c$ anzuschauen, wobei c eine Konstante ist, welche die Anzahl der verfügbaren Server angibt (Gross, Shortle, Thompson, Harris (2008)).
Ein anderes Phänomen, das in der Praxis manchmal auftritt, in dieser Arbeit aber nicht behandelt werden konnte, ist dass sich Kunden in eine viel zu lange Schlange oft gar nicht anstellen. Das kann zum Beispiel daran liegen, dass die Warteraumkapazität nicht ausreicht um mehr als eine bestimmte Anzahl von Kunden zu beherbergen. Ist diese Anzahl gleich K, so ist das geeignete Modell eines der Form $\cdot/\cdot/\cdot/K$.
Schließlich wurde angenommen, dass die Kunden nach dem Windhundprinzip aufgerufen werden. Dieses Prinzip FCFS ist zwar in der Praxis in den meisten Fällen gegeben, doch gibt es Fälle, wie zum Beispiel der Lageraufbau, die andere Annahmen benötigen. So existieren Modelle der Form $\cdot/\cdot/\cdot/\cdot/Z$, wobei Z für FCFS, LCFS, RSS, PR oder GD stehen kann. Diese Abkürzungen stehen für First-Come-First-Serve, Last-Come-First-Serve, Random Selection for Service, Priority und General Descipline und sind hoffentlich selbsterklärend (Gross, Shortle, Thompson, Harris (2008)).

5 Conclusio

Heutzutage ist jedermann an Warteschlangen gewohnt, denn sie entstehen in allen möglichen Situationen im alltäglichen Leben. Die wenigsten können sich jedoch vorstellen welch schwierige mathematische Theorie sich hinter diesen Produkten des gesellschaftlichen Zusammenlebens birgt. Um der Wirklichkeit aber einen Schritt voraus zu sein, muss man diese Theorie beherrschen und in der Praxis anwenden. Diese Arbeit zeigte die Herleitung der markante-

sten Resultate aus der Warteschlangentheorie und deren Wichtigkeit anhand einiger Anwendungsbeispiele.
Das Gesetz von Little ist wahrscheinlich das wichtigste Ergebnis in der Theorie der Warteschlangen. Dessen Geltung, Simplizität und Allgemeinheit sind nicht nur sehr überraschend, sondern auch enorm wichtig für die Anwendung in der Realität. Die Beziehung $L_q = \lambda W_q$ lässt erstaunlich einfach aus der durchschnittlichen Länge einer Warteschlange auf die durchschnittliche Wartezeit in dieser schließen und ist so gesehen die Brücke zwischen der Sicht des Systembesitzers zur der Sicht der Kunden.
Das $M/M/1$ Modell konnte soweit analysiert werden, dass die Verteilung der Kunden in einem System zu jedem Zeitpunkt berechnet wurde. Auch wenn die gefundene Formel sich als höchst kompliziert herausstellt, ist diese trotzdem explizit und kann angewandt werden, um den Sachverhalt einer Warteschlange zu verstehen. Viel schönere Formeln ergaben sich, wenn man die Zeit gegen unendlich laufen ließ, oder, realistischer gesprochen, sich lange nach Entstehen der Warteschlange die Situation anschaute. So macht die Gleichung (15) und ihr zugehöriger Graph Vieles klar in Fragen nach der Anzahl der Kunden im System in Abhängigkeit von ihrer Ankunftsrate und der Bearbeitungsrate des Servers.
Die Verallgemeinerung auf das Modell der $M/G/1$ Warteschlange erlaubte keinen Einblick in beliebige Zeitpunkte nach Entstehung mehr, sondern handelte nur von Gleichgewichtszuständen. Dabei ergab erneut ein erstaunlich einfacher Ausdruck für die durchschnittliche Größe der Warteschlange. Diese, unter dem Namen Pollaczek-Khinchin Formel, berührte Beziehung findet nicht nur außerordentlich oft Anwendung in der Realität, sondern erklärt auch formal das bekannte Wartezeitparadoxon. Dessen Kern und somit ein fundamentales Resultat aus der Warteschlangentheorie formulierte unwissend schon Goethe als er schrieb: «Ordnung lehrt Euch Zeit gewinnen».
Schließlich wurde im kleinen Ausblick dieser Arbeit gezeigt, dass das 20. Jahrhundert weitreichende Fortschritte und tiefliegende Erkenntnisse für die Welt der Warteschlangen bedeutet hat. Die sich ständig entwickelnde Gesellschaft und Technologie zwingen die Mathematik mitzuhalten und andauernd neue Weisheit zu schaffen, um Wohlfahrtsverluste zu verhindern. Die Entstehung der Warteschlangentheorie ist hierfür ein ideales Beispiel.

6 Literaturverzeichnis

C. Forbes, M. Evans, N. Hastings, B. Peacock (2011), *Statistical Distributions*, John Wiley & Sons.

D. Gross, J.F. Shortle, J.M. Thompson, C.M. Harris (2008), *Fundamentals of Queueing Theory*, Wiley-Interscience New York, USA.

EqWorld (2005), A.D. Polyanin, http://eqworld.ipmnet.ru/, http://eqworld.ipmnet.ru/en/auxiliary/inttrans/LapInv3.pdf

J.D.C. Little und S.C. Graves (2008), *Little's Law-Published*, Springer Science + Business Media.

J.L. Jain, S.G. Mohanty, W. Böhm (2007), *A Course on Queueing Models*, Chapman and Hall/CRC.

K. Jänich (2003), *Funktionentheorie*, Springer.

N. Bhat (2008), *An Introduction to Queueing Theory*, Birkhäuser Basel.

S. Stidham, Jr. (1972), *Technical Note—A Last Word on* $L = \lambda W$, Operations Research 22(2):417-421.

W. Preuß (2002), *Funktionaltransformationen: Fourier-, Laplace- und Z-Transformation*, Carl-Hanser-Verlag.