Asymptotische Exponentialität und die Approximation von Wartezeitverteilungen für Pattern in zufälligen Zeichenketten

Vom Fachbereich Mathematik der
Universität Hannover
zur Erlangung des Grades
Doktor der Naturwissenschaften
Dr. rer. nat.
genehmigte Dissertation
von

Dipl.-Math. Marcus Reich
geboren am 27. April 1974 in Hamburg

2004

Referent: Prof. Dr. Rudolf Grübel, Universität Hannover
Korreferent: Prof. Dr. F. T. Bruss, Université Libre de Bruxelles
Tag der Promotion: 05. November 2004

Zusammenfassung

Gegenstand dieser Arbeit sind zum einen die Untersuchung der Asymptotik von Warte- und Eintrittszeitverteilungen für Markov-Ketten und Erneuerungsprozesse, zum anderen die Herleitung von Methoden zur Approximation der Wartezeit des sog. *Approximate Pattern Matching Problems*. Sie gliedert sich in zwei Kapitel.

Kapitel 1 untersucht die Asymptotik von Wartezeitverteilungen bei Markov-Ketten und Erneuerungsprozessen. Ein besonderes Augenmerk wird dabei auf die Frage gelegt, unter welchen Bedingungen eine solche Wartezeitverteilung nach entprechender Normierung in Verteilung gegen eine Exponentialverteilung konvergiert. Mehrere Resultate stellen einen Zusammenhang zur Seltenheit der zur Wartezeit gehörenden Ereignisse her, ein Phänomen, das unter dem Namen *Rarity and Exponentiality* Eingang in die Literatur gefunden hat. Die theoretischen Resultate werden durch eine Reihe von Beispielen abgerundet. Zu nennen sind hier Wartezeitverteilungen im Zusammenhang mit der Irrfahrt auf $\mathbb{Z}$, Lebensdauerprozessen, dem Ehrenfestschen Urnenmodell, Irrfahrten auf dem Hyperwürfel $\{0,1\}^N$ und dem *Exact Pattern Matching Problem*.

Kapitel 2 ist der Analyse des *Approximate Pattern Matching Problem* gewidmet, d.h. der Frage nach der Wartezeit, bis in einer zufälligen Zeichenkette ein vorgegebener Pattern erstmals näherungsweise erscheint. Nach der mathematischen Präzisierung dieses Problems gelingt es unter speziellen Voraussetzungen, mit den zuvor hergeleiteten Methoden auch für die Wartezeit dieses Problems asymptotische Exponentialität nachzuweisen. Unter praktischen Gesichtspunkten ist dieses Resultat jedoch insofern unbefriedigend, als dass keine Schranken für den Fehler bei der Approximation durch eine Exponentialverteilung in konkreten Anwendungen angegeben werden können. Der Hauptteil dieses Kapitels ist daher der Herleitung von Näherungen dieser Wartezeitverteilung gewidmet, die zusätzlich die Berechnung solcher Fehlerschranken mit moderatem Rechenaufwand gestatten. Unter diesen Prämissen gelingt die Approximation der Wartezeit durch eine geometrische Verteilung, so wie die Approximation der Anzahl des näherungsweisen Auftauchens des Patterns durch eine Poisson-Verteilung.

Schlagwörter: asymptotische Exponentialität, Wartezeitverteilungen für Markov-Ketten und Erneuerungsprozesse, Pattern Matching.

Abstract

The aim of this dissertation is the analysis of asymptotics of waiting and entrance times in Markov chains and renewal processes as well as the design of methods for the distributional approximation of the waiting time in the so-called *approximate pattern matching problem*. It consists of two chapters.

Chapter 1 deals with the asymptotics of waiting time distributions in Markov chains and renewal processes. It focuses especially on the question under which conditions such a waiting time distribution converges – after an adequate standardisation – to an exponential distribution. In this context, results are derived which establish a connection to the rareness of the respective event, a phenomenon that is known under the label *rarity and exponentiality*. The theoretical results are supplemented by many examples, namely waiting time distributions for random walks on $\mathbb{Z}$, for lifetime processes in renewal theory, for the Ehrenfest urn model, for random walks on the hypercube $\{0,1\}^N$, and for the *exact pattern matching problem*.

Chapter 2 deals with the analysis of the *approximate pattern matching problem*, i.e. how long does it take until a specific pattern occurs for the first time "approximately" in a random string of letters. After this problem is formulated in an exact mathematical manner, we show that the results of chapter 1 lead to a proof of asymptotic exponentiality under special conditions for this waiting time as well. However, under practical aspects this result is unsatisfying because it provides no error bounds for the case that one approximates the waiting time distribution of a specific pattern by an exponential distribution. Hence, the main part of this chapter is dedicated to the design of methods for the approximation of the waiting time distribution for the *approximate pattern matching problem*, which allow the calculation of such error bounds at moderate cost. In this context, results are the distributional approximation of the waiting time by a geometric distribution and the approximation of the number of approximate hits of the pattern by a Poisson distribution.

Keywords: rarity and exponentiality, waiting time distribution for Markov chains and renewal processes, pattern matching.

Inhaltsverzeichnis

Einleitung

Eine endliche Folge von Buchstaben aus einem vorgegebenen Alphabet bezeichnet man als einen Pattern. Die Analyse der Wartezeitverteilung für das Erscheinen eines bestimmten Patterns in einer zufällig über diesem Alphabet generierten Zeichenkette kann inzwischen auf eine lange Tradition in der stochastischen Literatur zurückblicken, die mit einem wachsenden Interesse an solchen Resultaten in diversen Anwendungsgebieten einhergeht. Zu nennen sind hier beispielsweise die Analyse von Gensequenzen in der Molekularbiologie, das Aufspüren von Fehlern in der Qualitätskontrolle oder die Untersuchung von String-Matching-Algorithmen und Internet-Suchmaschinen aus dem Bereich der Informatik. Einen guten Überblick über verschiedene Anwendungsgebiete liefert [Nav01].

Erste erneuerungstheoretische Ansätze zur Gewinnung der Wartezeitverteilung sog. *Runs* – Pattern, für die sämtliche Buchstaben identisch sind – finden sich bereits bei Feller ([Fel68], S. 322ff.). Anfang der 1980er Jahre gelang dann L. J. Guibas und A. M. Odlyzko ([GO81]) erstmals die Bestimmung der Wartezeitverteilung für beliebige Pattern in Zeichenketten, die durch eine Folge von unabhängigen und identisch gleichverteilten Zufallsvariablen erzeugt werden. Sie identifizierten dabei die Selbstüberlappung eines Pattern als die wesentliche Einflussgröße auf die Wartezeitverteilung und leiteten eine geschlossene Formel für die wahrscheinlichkeitserzeugende Funktion dieser Verteilung her. Etwa zur selben Zeit gelang es H. U. Gerber und S. -Y. R. Li ([Li80], [GL81]) mit Hilfe von Martingalmethoden, diese Ergebnisse auch auf den Fall zu verallgemeinern, dass die Zeichenkette durch eine beliebige Verteilung auf dem zu Grunde liegenden Alphabet generiert wird. In den Folgejahren wurden die Resultate dieser Autoren mit verschiedenen anderen Ansätzen bestätigt. Insbesondere sei hier die sog. *Markov Chain Embedding Technique* erwähnt, die auch Gegenstand meiner Diplomarbeit war ([Rei01], vgl. hierzu aber auch [Fu01]).

Alle so gewonnen Resultate weisen jedoch Mängel bezüglich ihrer praktischen Nutzbarkeit in den oben genannten Anwendungsgebieten auf. An dieser Stelle seien die zwei Hauptmängel genannt.

Zum einen stellt das Modell einer durch eine Folge von unabhängigen und identisch verteilten Zufallsvariablen generierten Zeichenkette eine zu starke Vereinfachung der Wirklichkeit dar. Sowohl in DNA-Sequenzen als auch bei Texten aus natürlichen Sprachen bestehen in der Regel starke Abhängigkeiten zwischen den einzelnen Zeichen. Eine natürliche und für die Praxis häufig auch hinreichende Verallgemeinerung stellt hier der Übergang zu Zeichenketten dar, die durch eine Markov-Kette über dem zu Grunde liegenden Alphabet generiert werden. Erste Resultate für die Wartezeit eines Pattern unter dieser Modellannahme finden sich bei Pittenger ([Pit87], Bestimmung des Erwartungswerts der Wartezeit), ihren Abschluss findet diese Erweiterung der Problemstellung bei Robin und Daudin ([RD99]), denen es obendrein gelingt, ihre Resultate mit sehr einfachen erneuerungstheoretischen Argumenten herzuleiten und so den Bogen zurück zu den ersten Ansätzen bei Feller zu schlagen. Dieser häufig auch als *Exact Pattern Matching Problem* bezeichnete Bereich kann damit heute als abgeschlossen betrachtet werden.

Der zweite Kritikpunkt richtet sich gegen die Einschränkung, die Wartezeit für exakt einen Pattern zu bestimmen. In einem Genom erfüllen häufig verschiedene, ähnliche Pattern dieselbe biologische Funktion; bei der Suche in großen Datenbanken ist man häufig auch an Treffern interessiert, die eine gewisse Ähnlichkeit mit dem eingegebenen Suchbegriff aufweisen. Analysiert man die Wartezeit, bis der erste einer ganzen Familie von Pattern in einer zufälligen Zeichenkette erscheint, so spricht man von einem sog. *Compound Pattern Matching Problem*, bei einer Familie von sehr ähnlichen Pattern verwendet man auch den Begriff des *Approximate Pattern Matching Problem*. In der Theorie ist dieses Problem bereits in der Arbeit von Gerber und Li ([GL81]) gelöst worden. Diese Autoren geben ein Gleichungssystem an, mit dem sich das *Compound Pattern Matching Problem* auf das *Exact Pattern Matching Problem* für die beteiligten Pattern zurückführen lässt. (Vgl. auch [RD01].) In der Praxis hat man es jedoch sehr häufig mit extrem großen Patternfamilien zu tun, so dass alle derzeit bekannten Methoden zur exakten Bestimmung der Wartezeitverteilung bei konkreten Problemstellungen am zu großen Rechenaufwand scheitern. Simuliert man andererseits die Wartezeitverteilung für große Familien von hinreichend langen Pattern, so stellt man in nahezu allen Fällen eine starke Ähnlichkeit zu einer geome-

trischen oder Exponentialverteilung fest, so dass der Praktiker in der Regel solche Verteilungsmodelle als hinreichend genaue Näherung verwenden wird.

Diese Beobachtung ist der Ausgangspunkt für die vorliegende Dissertation. Sie gliedert sich in zwei Kapitel.

Kapitel 1 nimmt zunächst keinen Bezug auf das konkrete Problem des *Pattern Matching*. Es beschäftigt sich vielmehr ganz allgemein mit der Analyse der Asymptotik von Warte- und Eintrittszeitverteilungen bei Markov-Ketten und Erneuerungsprozessen. Insbesondere wird die Frage behandelt, unter welchen Voraussetzungen für derartige Wartezeitverteilungen asymptotische Exponentialität vorliegt. Es gelingt, einige Resultate herzuleiten, die einen engen Zusammenhang zur Seltenheit der zur Wartezeit gehörenden Ereignisse herausstellen. Neben den theoretischen Überlegungen werden mehrere Beispiele für dieses Phänomen angeführt, die gleichzeitig seine Grenzen ausloten. Unter anderem sind hier Wartezeitprobleme für die symmetrische Irrfahrt auf $\mathbb{Z}$, Lebensdauerprozesse, das Ehrenfestsche Urnenmodell und Irrfahrten auf dem Hyperwürfel $\{0,1\}^N$ zu nennen. Den Bogen zum *Approximate Pattern Matching Problem* und den Resultaten des folgenden zweiten Kapitels schlägt schließlich der Nachweis der asymptotischen Exponentialität für das *Exact Pattern Matching Problem*.

Kapitel 2 konzentriert sich sodann auf das spezielle Problem des oben angesprochenen *Approximate Pattern Matching*. Es wird eine exakte mathematische Formulierung dieses Problems vorgenommen und es gelingt, die zugehörige Wartezeit als Eintrittszeit für eine Markov-Kette zu interpretieren. Unter speziellen Voraussetzungen gelingt mit den Methoden des ersten Kapitels auch für dieses Problem der Nachweis der asymptotischen Exponentialität. Unter praktischen Gesichtspunkten ist dieses Resultat jedoch unbefriedigend, da es keine Aussage darüber macht, wie gut die Approximation der Wartezeitverteilung durch eine Exponentialverteilung für einen konkreten Pattern ist. Der Hauptteil des zweiten Kapitels ist daher der Herleitung von Resultaten gewidmet, die die Wartezeit des *Approximate Pattern Matching Problems* annähern und gleichzeitig mit moderatem Aufwand zu berechnende Fehlerschranken liefern, mit denen sich die Güte der Näherung beurteilen lässt. Unter diesen Prämissen gelingt die Approximation der Wartezeitverteilung durch eine geometrische Verteilung unter Herleitung eines zu [Ald82] ähnlichen Resultats, sowie die Approximation der Anzahl des näherungsweisen Erscheinens des Patterns durch eine Poisson-Verteilung mit Hilfe der sog. *Chen-Stein-Methode*.

Zur Notation: Sämtliche Bezeichnungen werden dort eingeführt, wo sie zum ersten Mal verwendet werden. Darüber hinaus gibt es einige sehr häufig verwendete Bezeichnungen, die schon an dieser Stelle definiert werden sollen.

Die Indikatorfunktion zu einem Ereignis A wird alternativ mit $\mathbb{1}_A$ oder $\mathbb{1}\{A\}$ bezeichnet.

Häufig ergibt sich die Situation, dass Ereignisse und Zufallsvariablen im Zusammenhang mit einem bestimmten stochastischen Prozess, z.B. $X = (X_n)_{n\in\mathbb{N}_0}$ auf dem Zustandsraum $E = \{i, j, k, ...\}$, betrachtet werden. In einem solchen Zusammenhang bezeichnen $P_i(\,\cdot\,)$, $E_i(\,\cdot\,)$ bzw. $\mathcal{L}_i(\,\cdot\,)$ die bedingte Wahrscheinlichkeit des betrachteten Ereignisses, den bedingten Erwartungswert bzw. die bedingte Verteilung der betrachteten Zufallsvariable unter der Bedingung, dass der Prozess X in i startet, also unter $\{X_0 = i\}$. Ist μ eine Verteilung auf dem Zustandsraum E, so verwenden wir weiter die Bezeichnungen $P_\mu(\,\cdot\,)$, $E_\mu(\,\cdot\,)$ bzw. $\mathcal{L}_\mu(\,\cdot\,)$ für die bedingte Wahrscheinlichkeit, den bedingten Erwartungswert bzw. die bedingte Verteilung unter der Bedingung, dass X der Startverteilung μ genügt. Es ist also zum Beispiel $P_\mu(\,\cdot\,) = \sum_{i\in E} P(\,\cdot\,|X_0 = i)\mu(i)$, E_μ und $\mathcal{L}_\mu$ analog.

Im Zusammenhang mit solchen stochastischen Prozessen wollen wir Wartezeiten betrachten, bis der Prozess einen bestimmten Zustand erreicht, in eine bestimmte Teilmenge des Zustandsraums eintritt, ein bestimmtes Niveau überschreitet, etc. Hier folgen wir der Notation von Aldous und Fill ([AF04], Kapitel 2, S. 1f.) und unterscheiden solche Wartezeiten in Eintritts- und Rückkehrzeiten: Ist beispielsweise die Wartezeit von Interesse, bis der Prozess X in den Zustand i eintritt, so unterscheidet man zwischen der *Eintrittszeit* $T_i := \inf\{n \in \mathbb{N}_0 : X_n = i\}$ und der *Rückkehrzeit* $T_i^+ := \inf\{n \in \mathbb{N} : X_n = i\}$. (Ist $X_0 \neq i$, so fallen in dieser Situation beide Wartezeiten zusammen.)

Im Bereich der stochastischen Konvergenz bezeichne $\xrightarrow{\text{P}}$ Konvergenz in Wahrscheinlichkeit, $\xrightarrow{\text{f.s.}}$ Konvergenz fast sicher und $\xrightarrow{\text{d}}$ Verteilungskonvergenz. $=_d$ bezeichnet die Verteilungsgleichheit zweier Zufallsgrößen.

Sind (a_n) und (b_n) zwei Folgen reeller Zahlen, so ist $a_n = o(b_n)$ genau dann, wenn $\lim_{n\to\infty} a_n/b_n = 0$ gilt. Sind (a_n) und (b_n) Folgen nicht negativer reeller Zahlen, so ist $a_n = \Theta(b_n)$ genau dann, wenn es $K_1, K_2 > 0$ mit $K_1 \leq \liminf_{n\to\infty} a_n/b_n \leq \limsup_{n\to\infty} a_n/b_n \leq K_2$ gibt.

Mein Dank gilt Herrn Prof. Dr. R. Grübel für die Anregung zu dieser Arbeit und die fortwährende Unterstützung bei ihrer Erstellung. Herrn Prof. Dr. F. T. Bruss danke ich für die Übernahme des Korreferats.

Kapitel 1

Asymptotische Exponentialität von Wartezeitverteilungen

In diesem ersten Hauptkapitel wollen wir nun also für eine ganze Reihe von Markov-Ketten und Erneuerungsprozessen Wartezeit- und Eintrittszeitverteilungen analysieren und insbesondere deren asymptotisches Verhalten näher betrachten. Dabei soll uns vor allem interessieren, welches die entscheidenden Aspekte der betrachteten Beispiele sind, die letztlich dazu führen, dass die entsprechenden Wartezeitverteilungen asymptotisch exponentialverteilt ist. Dabei wird sich herausstellen, dass ein enger Zusammenhang zur Seltenheit der Ereignisse, auf die man wartet, besteht. In die Literatur hat dieses Phänomen unter dem Schlagwort *Rarity and Exponentiality* Eingang gefunden (vgl. [Kei79]). Doch obwohl als allgemeines Phänomen durchaus bekannt, sind die Resultate auf diesem Gebiet nach wie vor bruchstückhaft. Selbst bei einfachen Modellen ist bis heute nicht abschließend geklärt, unter welchen Voraussetzungen man für bestimmte Wartezeiten asymptotische Exponentialität erwarten kann.

Zum Auftakt untersuchen wir im ersten Abschnitt als klassisches Beispiel einer Markov-Kette die symmetrische Irrfahrt auf $\mathbb{Z}$ und die Wartezeit, bis diese bei Start in 0 erstmals vom Betrag ein vorgegebenes Niveau N überschreitet. Es stellt sich heraus, dass in diesem Fall mit wachsendem N keine asymptotische Exponentialität vorliegt. Die zur Analyse verwendeten Martingalmethoden können leider nicht dazu beitragen, die entscheidenden Aspekte der Wartezeitverteilung, die zu diesem Verhalten führen, zu erklären.

Daher leiten wir im folgenden zweiten Abschnitt einige allgemeine Resultate für Wartezeitverteilungen in Markov-Ketten her, die es in der Folge ermöglichen, Problemstellungen wie die aus dem ersten Abschnitt in strukturierter Form zu analysieren. Die entscheidende Idee ist dabei, solche Wartezeiten in eine Folge regenerativer Zyklen zwischen dem jeweiligen Durchlaufen des Startzustands ohne zwischenzeitliches Erreichen der Zielmenge und einen abschließenden Teilzyklus bis zum Erreichen der Zielmenge zu zerlegen.

Gerüstet mit diesem Instrumentarium analysieren wir im dritten Abschnitt erneut die Irrfahrt auf $\mathbb{Z}$, nun sowohl für den symmetrischen Fall als auch bei Vorliegen eines Drifts, und zeigen, unter welchen Voraussetzungen für dieses Problem asymptotische Exponentialität zu beweisen ist.

Als zweites Anwendungsbeispiel für die Methoden aus Abschnitt 2 untersuchen wir im vierten Abschnitt mit dem *Residual Age Process* und dem *Current Age Process* zwei wichtige Prozesse der Erneuerungstheorie, für die sich ebenfalls asymptotische Exponentialität nachweisen lässt.

Allen Beispielen der Abschnitte 3 und 4 ist gemeinsam, dass sich in den betrachteten Situationen stets dann asymptotische Exponentialität einstellt, wenn die Wahrscheinlichkeit dafür, innerhalb eines regenerativen Zykluses in die Zielmenge einzutreten, asymptotisch verschwindet. Im fünften Abschnitt wird ein Resultat von Keilson ([Kei66]) für Erneuerungsprozesse präsentiert, das diese Bedingung als hinreichend für asymptotische Exponentialität herausstellt. Anschließend wird dieses Resultat auf eine ganze Familie von Erneuerungsprozessen verallgemeinert.

Als ein wichtiges Anwendungsbeispiel, das sich nun mit diesem verallgemeinerten Resultat behandeln lässt, untersuchen wir im sechsten Abschnitt das sog. *Ehrenfestsche Urnenmodell*. Es gelingt, sehr allgemeine Bedingungen für das Vorliegen von asymptotischer Exponentialität in diesem Modell anzugeben.

Der siebte Abschnitt behandelt als Verallgemeinerung dieses Modells Irrfahrten auf dem Hyperwürfel $\{0,1\}^N$. Auch hier wird unter entsprechenden Voraussetzungen asymptotische Exponentialität von Wartezeitverteilungen für sog. *Nearest Neighbour Random Walks* nachgewiesen, allerdings sind zum Nachweis neue Ideen erforderlich.

Abschnitt 8 leitet schließlich zum Problem des *Pattern Matching* über. Es wird asymptotische Exponentialität für das *Exact Pattern Matching Problem* nachge-

wiesen, wobei im Fall einer unabhängigen und identisch verteilten Zeichenkette ein Resultat von J. Rudander ([Rud96]) präsentiert wird, das sich anschließend auf der Grundlage eines Artikels von S. Robin und J. J. Daudin ([RD99]) auf den Fall einer durch eine Markov-Kette generierten Zeichenkette verallgemeinern lässt.

1.1 Ein einführendes Beispiel: Die symmetrische Irrfahrt auf $\mathbb{Z}$

Als erstes Beispiel für die Analyse der Asymptotik von Eintrittszeitverteilungen betrachten wir die einfache symmetrische Irrfahrt auf $\mathbb{Z}$ mit Start in 0, also die Markov-Kette $X = (X_n)_{n\in\mathbb{N}_0}$ mit $X_0 = 0$ und Übergangsmatrix $P = (p_{ij})_{i,j\in\mathbb{Z}}$ mit $p_{i,i-1} = p_{i,i+1} = 1/2$, $p_{ij} = 0$ sonst.

Für diesen Prozess analysieren wir die Wartezeit, bis X erstmals betragsmäßig das Niveau $N \in \mathbb{N}$ erreicht, d.h. es sei

$$T_N := \inf\{n \in \mathbb{N}_0 : |X_n| \geq N\}.$$

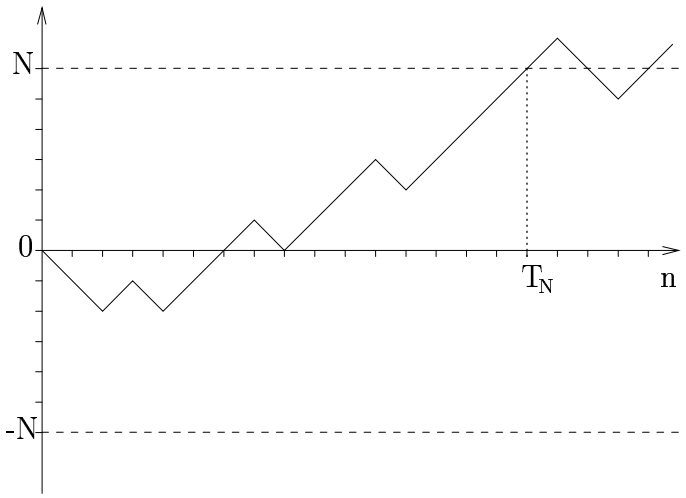

Abbildung 1.1: Graphische Darstellung der Wartezeit T_N.

Offensichtlich strebt ET_N mit $N \to \infty$ gegen ∞. Wenn wir also überhaupt Verteilungskonvergenz für T_N nachweisen möchten, so muss sich diese auf eine aus T_N abgeleitete und in geeigneter Weise skalierte Zufallsgröße beziehen, wie beispielsweise T_N/ET_N. Im Folgenden werden wir mit Martingalmethoden und alternativ auf dem Umweg über die Brownsche Bewegung aufzeigen, wie man für diese Wartezeit Verteilungskonvergenz von T_N/ET_N nachweist und welche Grenzverteilung man erhält.

Dazu bestimmen wir zunächst ET_N. Man zeigt leicht, dass $(X_n^2 - n)_{n\in\mathbb{N}_0}$ ein Martingal und T_N eine Stoppzeit hierzu ist. Also ist auch $\left(X_{n\wedge T_N}^2 - (n \wedge T_N)\right)_{n\in\mathbb{N}_0}$ ein

Martingal (vgl. [Wil91], §10.9, S. 99) und es gilt $EX^2_{n\wedge T_N} = E(n\wedge T_N)$ für alle $n \in \mathbb{N}_0$. Unter Verwendung der Sätze von der monotonen und der majorisierten Konvergenz (es gilt $|X^2_{n\wedge T_N}| \leq N^2$, T_N ist f.s. endlich) erhält man hieraus

$$ET_N = \lim_{n\to\infty} E(n \wedge T_N) = \lim_{n\to\infty} EX^2_{n\wedge T_N} = EX^2_{T_N} = N^2. \tag{1.1}$$

Im nächsten Schritt bestimmen wir nun die wahrscheinlichkeitserzeugende Funktion $g_{T_N}(z) = E\big(z^{T_N}\big)$. Dazu definieren wir für alle $\alpha \in \mathbb{R}$ die Familie von Prozessen $(Y^\alpha_n)_{n\in\mathbb{N}_0}$ durch $Y^\alpha_n := \cosh(\alpha)^{-n}\exp(\alpha X_n)$. Man zeigt leicht, dass für jedes $\alpha \in \mathbb{R}$ dies und damit auch $(Y^\alpha_{n\wedge T_N})_{n\in\mathbb{N}_0}$ ein Martingal ist. Da $\cosh(\alpha) \geq 1$ und $|X_{n\wedge T_N}| \leq N$ gilt, ist letzteres für alle $\alpha \in \mathbb{R}$ beschränkt. Nach dem Vorwärtskonvergenzsatz von Doob ([Wil91], §11.5, S. 109) existiert dann eine Zufallsvariable Y^α_∞ mit $Y^\alpha_{n\wedge T_N} \xrightarrow{\text{f.s.}} Y^\alpha_\infty$. Wegen $P(T_N < \infty) = 1$ stimmen $Y^\alpha_{T_N}$ und Y^α_∞ f.s. überein. Schließlich erhält man wegen der Beschränktheit von $(Y^\alpha_{n\wedge T_N})_{n\in\mathbb{N}_0}$ auch die L_1-Konvergenz von $Y^\alpha_{n\wedge T_N}$ gegen $Y^\alpha_{T_N}$ ([Wil91], §13.6, S. 130). Es gilt also

$$EY^\alpha_{T_N} = \lim_{n\to\infty} EY^\alpha_{n\wedge T_N} = EY^\alpha_0 = 1 \tag{1.2}$$

für alle $\alpha \in \mathbb{R}$.

Wir zerlegen nun $EY^\alpha_{T_N}$ in

$$EY^\alpha_{T_N} = \sum_{x=-N,N} E\left[\cosh(\alpha)^{-T_N}\exp(\alpha X_{T_N})|X_{T_N} = x\right] P(X_{T_N} = x).$$

Aus Symmetriegründen gilt $P(X_{T_N} = N) = P(X_{T_N} = -N) = 1/2$, sowie

$$E\left[\cosh(\alpha)^{-T_N}|X_{T_N} = N\right] = E\left[\cosh(\alpha)^{-T_N}|X_{T_N} = -N\right] = E\cosh(\alpha)^{-T_N}.$$

Also ist $EY^\alpha_{T_N} = \cosh(\alpha N)E\cosh(\alpha)^{-T_N}$, und aus (1.2) folgt

$$E\cosh(\alpha)^{-T_N} = \cosh(\alpha N)^{-1}. \tag{1.3}$$

Für $z \in (0,1]$ sei nun $\alpha_z = \operatorname{arcosh}\big(z^{-1}\big)$. Dann gilt $\cosh(\alpha_z) = \frac{1}{z}$, und man erhält aus (1.3) die wahrscheinlichkeitserzeugende Funktion zu T_N:

$$g_{T_N}(z) = Ez^{T_N} = E\cosh(\alpha_z)^{-T_N} = \cosh(\alpha_z N)^{-1}.$$

Mit Hilfe dieser Funktion können wir nun die Grenzverteilung von T_N/ET_N für $N \to \infty$ bestimmen. Bezeichnet nämlich $\varphi_N(\theta)$, $\theta \in \mathbb{R}$, die charakteristische Funktion zu T_N/ET_N, so gilt

$$\varphi_N(\theta) = g_{T_N}\big(\exp(i\theta/ET_N)\big) = \cosh\left(\alpha_{\exp(i\theta/N^2)}N\right)^{-1},$$

und wegen der Äquivalenz von Verteilungskonvergenz und der punktweisen Konvergenz der charakteristischen Funktionen verbleibt lediglich die Aufgabe, die Grenzfunktion dieses Ausdrucks für $N \to \infty$ zu bestimmen. Unter Verwendung von Hilfssatz A.1 erhält man für $\alpha_{\exp(i\theta/N^2)}N$ mit $N \to \infty$ den Grenzwert $\sqrt{-2i\theta}$, es gilt also

$$\lim_{N\to\infty} \varphi_N(\theta) = \cosh\left(\sqrt{-2i\theta}\right)^{-1}. \tag{1.4}$$

Alternativ kann man dieses Resultat auch mit Hilfe der Brownschen Bewegung erhalten. X_n lässt sich darstellen als

$$X_n =_d \sum_{m=1}^{n} Y_m,$$

wobei $(Y_m)_{m\in\mathbb{N}}$ eine Folge von unabhängigen und identisch verteilten Zufallsvariablen mit $\mathcal{L}(Y_1) = \frac{1}{2}(\delta_{-1} + \delta_1)$ ist. (δ_x bezeichnet das Einpunktmaß in x.) Insbesondere gilt $EY_1 = 0,\ \mathrm{var}(Y_1) = 1$. Dann folgt:

$$\frac{T_N}{ET_N} = \frac{1}{N^2}\inf\{n \in \mathbb{N}_0 \,:\, |X_n| \geq N\} = \inf\left\{t \geq 0 \,:\, \left|\frac{1}{\sqrt{N^2}}\, X_{\lfloor tN^2\rfloor}\right| \geq 1\right\}.$$

Nach dem Satz von Donsker ([Dur96], Theorem 6.6 und Example 6.2, S. 406) gilt dann

$$T_N/ET_N \xrightarrow{\mathrm{d}} \inf\{t \geq 0 \,:\, |W_t| \geq 1\},$$

wobei $(W_t)_{t\geq 0}$ die klassische Brownsche Bewegung bezeichnet. Für diese ist jedoch bekannt, dass die Austrittszeit aus einem symmetrischen Streifen um 0 der Breite 2 die charakteristische Funktion

$$\varphi(\theta) = \cosh\left(\sqrt{-2i\theta}\right)^{-1}$$

besitzt. (Die Laplace-Transformierte findet man beispielsweise in [BS96], II.1, Formel 3.0.1, S. 172; vgl. auch [Yor97], S. 132f.)

Es ist uns also gelungen, das asymptotische Verhalten der Wartezeit T_N für die symmetrische Irrfahrt zu analysieren. Die dabei verwendeten Martingalmethoden muten in diesem Zusammenhang jedoch sehr speziell an und scheinen auf den ersten Blick keine generelle Möglichkeit zu bieten, dieselben Methoden auch auf andere Problemstellungen anzuwenden. Schon der Versuch, auf ähnliche Weise den Fall einer Irrfahrt mit Drift zu untersuchen (vgl. Abschnitt 1.3), stellt einen vor die schwierige Aufgabe, ein für diese Situation geeignetes Martingal zu konstruieren. Darüber hinaus

offenbart dieser Ansatz nichts über die „innere Struktur“ der Wartezeit T_N und das Grenzwertresultat ergibt sich letztlich auf rein technisch-analytischem Wege. Allein der Umweg über die Brownsche Bewegung gewährt einen gewissen Einblick zum höheren Verständnis.

Aus diesem Grund werden wir im folgenden Abschnitt einige einfache Resultate für Eintrittszeitverteilungen bei Markov-Ketten herleiten, mit denen es möglich ist, Problemstellungen wie die hier behandelte in überschaubarer und strukturierter Weise zu analysieren. In der Tat werden wir diese Resultate im dritten Abschnitt dieses Kapitels dazu verwenden, um nochmals die Irrfahrt auf $\mathbb{Z}$ zu untersuchen und neben den jetzt schon bekannten Resultaten einige weiterführende herzuleiten.

Löst man sich von den hier verwendeten Methoden zur Gewinnung der Detailaussagen, so kann man jedoch bereits an diesem Beispiel eine grundsätzliche Vorgehensweise erkennen, die uns noch des Öfteren von Nutzen sein wird: Zur Analyse der Verteilungskonvergenz von T_N/ET_N versucht man zunächst, explizite Ausdrücke für ET_N und die wahrscheinlichkeitserzeugende Funktion g_{T_N} von T_N zu gewinnen und analysiert dann das asymptotische Verhalten der charakteristischen Funktion φ_N von T_N/ET_N mit $N \to \infty$, wobei man die Identität

$$\varphi_N(\theta) = g_{T_N}\big(\exp(i\theta/ET_N)\big) \tag{1.5}$$

ausnutzt. Das zentrale Hilfsmittel ist hier also der Stetigkeitssatz für charakteristische Funktionen.

1.2 Erste theoretische Grundlagen

In diesem Abschnitt werden ein paar grundlegende Resultate vorgestellt, die in der Folge bei der strukturierten Analyse des asymptotischen Verhaltens von Eintrittszeitverteilungen bei einer ganzen Reihe von Markov-Ketten Verwendung finden werden. Insbesondere werden wir im Anschluss an diesen Abschnitt zu einer detaillierten Analyse der Irrfahrt auf $\mathbb{Z}$ in der Lage sein. Im darauf folgenden vierten Abschnitt werden wir außerdem zwei weitere Markov-Ketten mit diesem Instrumentarium untersuchen.

Es sei ganz allgemein $(X_n)_{n\in\mathbb{N}_0}$ eine homogene Markov-Kette auf einem endlichen oder abzählbar unendlichen Zustandsraum E mit Übergangsmatrix P und Start in einem festen Punkt $e \in E$. Ferner sei $\emptyset \neq D \subset E$ mit $e \notin D$. Dann lässt sich die Wartezeit

$$T := \inf\{n \in \mathbb{N}_0 \,:\, X_n \in D\},$$

bis die Markov-Kette erstmals in die Menge D eintritt, darstellen als

$$T = \sum_{k=1}^{M} \gamma_k \tag{1.6}$$

mit

$$\begin{aligned}
\sigma_0 &:= 0,\\
\sigma_k &:= \min\big\{n > \sigma_{k-1} \,:\, X_n \in D \cup \{e\}\big\},\ k \geq 1,\\
\gamma_k &:= \sigma_k - \sigma_{k-1},\ k \geq 1,\\
M &:= \min\big\{k \geq 1 \,:\, X_{\sigma_k} \in D\big\}.
\end{aligned}$$

Dabei weist die Summe (1.6) folgende Eigenschaften auf:

Satz 1.1 *Mit den zuvor eingeführten Bezeichnungen gilt:*

(a) Unter $\{M = m\}$ sind $\gamma_1, ..., \gamma_m$ unabhängig,

(b) $\mathcal{L}(\gamma_k | M = m) = \mathcal{L}(\gamma_1 | X_{\gamma_1} = e)$ für $k = 1, ..., m-1$ und
$\mathcal{L}(\gamma_m | M = m) = \mathcal{L}(\gamma_1 | X_{\gamma_1} \in D)$,

(c) M ist geometrisch verteilt auf $\mathbb{N}$ mit Erfolgswahrscheinlichkeit $P_e(X_{\gamma_1} \in D)$.

Beweis: Mit der starken Markov-Eigenschaft folgt

$$\begin{aligned} P_e(M=m) &= P_e(X_{\sigma_1}=e,...,X_{\sigma_{m-1}}=e, X_{\sigma_m}\in D) \\ &= \prod_{k=1}^{m-1} P(X_{\sigma_k}=e|X_{\sigma_{k-1}}=e)\cdot P(X_{\sigma_m}\in D|X_{\sigma_{m-1}}=e). \end{aligned}$$

Außerdem ergibt sich durch Zerlegung nach dem Wert von σ_k unter Ausnutzung der zeitlichen Homogenität und der starken Markov-Eigenschaft $P(X_{\sigma_k}\in\cdot\,|X_{\sigma_{k-1}}=e)=P_e(X_{\gamma_1}\in\cdot\,)$. Also ist

$$P_e(M=m)=P_e(X_{\gamma_1}=e)^{m-1}P_e(X_{\gamma_1}\in D) \tag{1.7}$$

für $m\geq 1$, d.h. M ist geometrisch verteilt auf $\mathbb{N}$ mit Erfolgswahrscheinlichkeit $P_e(X_{\gamma_1}\in D)$.

Definiert man $j_l:=\sum_{k=1}^{l} i_k$, so erhält man aus der Darstellung der γ_k über die σ_k unter Ausnutzung der schwachen Markov-Eigenschaft und der zeitlichen Homogenität

$$\begin{aligned} &P_e(\gamma_k=i_k,\ k=1,...,m,\ M=m) \\ &= P_e(\sigma_k=j_k,\ k=1,...,m,\ X_{j_k}=e, k=1,...,m-1,\ X_{j_m}\in D) \\ &= \prod_{k=1}^{m-1} P_e(\gamma_1=i_k,\ X_{\gamma_1}=e)\cdot P_e(\gamma_1=i_m,\ X_{\gamma_1}\in D). \end{aligned}$$

Daher ergibt sich aus (1.7)

$$P_e(\gamma_k=i_k,\ k=1,...,m\,|\,M=m)=\prod_{k=1}^{m-1} P_e(\gamma_1=i_k|X_{\gamma_1}=e)\cdot P_e(\gamma_1=i_m|X_{\gamma_1}\in D).$$

Durch Aufsummieren über alle übrigen Komponenten erhält man so

$$P_e(\gamma_k=i_k|M=m)=\begin{cases} P_e(\gamma_1=i_k|X_{\gamma_1}=e), & k<m, \\ P_e(\gamma_1=i_m|X_{\gamma_1}\in D), & k=m. \end{cases}$$

Dies wiederum liefert

$$P_e(\gamma_k=i_k,\ k=1,...,m\,|\,M=m)=\prod_{k=1}^{m} P_e(\gamma_k=i_k|M=m),$$

also die Unabhängigkeit von $\gamma_1,...,\gamma_m$ unter $\{M=m\}$. □

Insbesondere sind also unter $\{M = m\}$ die Zufallsvariablen $\gamma_1, ..., \gamma_m$ unabhängig und darüber hinaus $\gamma_1, ..., \gamma_{m-1}$ identisch verteilt. Es ist uns damit gelungen, die Eintrittszeit T in eine Folge regenerativer Zyklen zu zerlegen, in denen die Markov-Kette in ihren Ausgangszustand e zurückkehrt, ohne vorher die Zielmenge D zu erreichen, und einen abschließenden Teilzyklus, in dem sie von e nach D wandert, ohne noch einmal nach e zurückzukehren. Bei der wahrscheinlichkeitserzeugenden Funktion der Eintrittszeit T schlägt sich dies in der folgenden faktoriellen Zerlegung nieder:

$$g_T(z) = g_{M_0}\big(g_{\gamma_1|X_{\gamma_1}=e}(z)\big) \cdot g_{\gamma_1|X_{\gamma_1}\in D}(z), \tag{1.8}$$

wobei g_{M_0} die wahrscheinlichkeitserzeugende Funktion zu einer auf $\mathbb{N}_0$ geometrisch verteilten Zufallsvariablen mit Erfolgswahrscheinlichkeit $P_e(X_{\gamma_1} \in D)$ bezeichnet. Diese ist wohl bekannt, es ergibt sich somit

$$g_T(z) = \frac{P_e(X_{\gamma_1} \in D)}{1 - P_e(X_{\gamma_1} = e) \cdot g_{\gamma_1|X_{\gamma_1}=e}(z)} \cdot g_{\gamma_1|X_{\gamma_1}\in D}(z). \tag{1.9}$$

Weitere Folgerung:

$$ET = \frac{P_e(X_{\gamma_1} = e)}{P_e(X_{\gamma_1} \in D)} \cdot E[\gamma_1|X_{\gamma_1} = e] + E[\gamma_1|X_{\gamma_1} \in D]. \tag{1.10}$$

Die Darstellung (1.8) bzw. (1.9) bietet also die Möglichkeit, die Wartezeit T als Faltung einer zufälligen Summe mit einer geometrisch verteilten Anzahl von unabhängigen Exkursionen mit Rückkehr zum Ausgangspunkt und einer finalen Exkursion vom Ausgangspunkt in die Menge D zu betrachten. Kennt man die Verteilung dieser beiden Bausteine der Wartezeit T, so ermöglichen es diese Gleichungen, daraus die Verteilung von T selbst zu bestimmen.

Gerade diese Zerlegung der Wartezeit T eignet sich in besonderer Weise zur Untersuchung von T auf asymptotische Exponentialität, stellt sich doch die Exponentialverteilung als Grenzverteilung einer Zufallssumme von regenerativen Zyklen mit einer geometrisch verteilten Anzahl von Summanden ein, wenn die Erfolgswahrscheinlichkeit dieser geometrischen Verteilung gegen Null strebt, wie ein späteres Resultat zeigen wird (vgl. Abschnitt 1.5). Andererseits ist dies nicht die einzige Möglichkeit, um asymptotische Exponentialität nachzuweisen; in Abschnitt 1.7 werden wir mit der Irrfahrt auf dem Hyperwürfel $\{0,1\}^N$ ein Beispiel kennen lernen, bei dem andere Methoden zum Ziel führen.

1.3 Die Irrfahrt auf $\mathbb{Z}$ (Fortsetzung)

Wie bereits angekündigt, werden wir nun also die allgemeinen Ergebnisse des vorangegangenen Abschnitts in diesem und dem folgenden Abschnitt zur Analyse einiger spezieller Eintrittszeitprobleme verwenden.

Als erstes betrachten wir erneut die Irrfahrt $X = (X_n)_{n\in\mathbb{N}_0}$ auf $\mathbb{Z}$ mit Start in 0 und die Wartezeit T_N, bis X erstmals betragsmäßig das Niveau N erreicht. Mit Hilfe der zuvor entwickelten Methoden wollen wir nun eine komplette Analyse dieses Wartezeitproblems durchführen. Dabei beschränken wir uns nicht nur auf den Fall einer symmetrischen Irrfahrt. Vielmehr existiere ein $p \in (0,1)$, so dass für die Übergangsmatrix $P = (p_{ij})_{i,j\in\mathbb{Z}}$ von X gilt

$$p_{ij} = \begin{cases} 1/2, & i = 0,\ j = -1, 1, \\ p, & i > 0,\ j = i-1 \ \text{ oder } \ i < 0,\ j = i+1, \\ 1-p =: q, & i > 0,\ j = i+1 \ \text{ oder } \ i < 0,\ j = i-1. \end{cases}$$

$p = 1/2$ ist offenbar der im ersten Abschnitt analysierte Fall einer symmetrischen Irrfahrt, im Falle $p > 1/2$ spricht man auch von einer Irrfahrt mit Drift nach 0, im Falle $p < 1/2$ analog von einem Drift nach ∞.

Betrachtet man anstelle von X die Markov-Kette $Y = (Y_n)_{n\in\mathbb{N}_0}$ auf dem Zustandsraum $\mathbb{N}_0$ mit Start in 0 und Übergangsmatrix $Q = (q_{ij})_{i,j\in\mathbb{N}_0}$ mit

$$q_{ij} = \begin{cases} 1, & i = 0,\ j = 1, \\ p, & i > 0,\ j = i-1, \\ q, & i > 0,\ j = i+1, \end{cases}$$

so gilt offenbar $(|X_n|)_{n\in\mathbb{N}_0} =_d (Y_n)_{n\in\mathbb{N}_0}$. Für die uns interessierende Wartezeit $T_N := \inf\{n \geq 0 : |X_n| \geq N\}$ gilt also $T_N =_d \inf\{n \geq 0 : Y_n = N\}$, so dass wir uns für die weiteren Betrachtungen auf die Kette Y beschränken können.

Im Sinne von (1.6) lässt sich T_N nun in folgender Weise darstellen:

$$T_N = \sum_{k=1}^{M_N} \gamma_k$$

mit $\sigma_0 := 0$, $\sigma_k := \min\big\{n > \sigma_{k-1} : Y_n \in \{0, N\}\big\}$, $k \geq 1$, $\gamma_k := \sigma_k - \sigma_{k-1}$, $k \geq 1$, $M_N := \min\big\{k \geq 1 : Y_{\sigma_k} = N\big\}$.

Mit Hilfe der zuvor entwickelten Gleichung (1.9) ist es nun möglich, die wahrscheinlichkeitserzeugende Funktion zu T_N zu bestimmen, sobald man $P_0(Y_{\gamma_1} = 0)$, $P_0(Y_{\gamma_1} = N)$ und die wahrscheinlichkeitserzeugenden Funktionen $g_{\gamma_1|Y_{\gamma_1}=0}(z)$, $g_{\gamma_1|Y_{\gamma_1}=N}(z)$ kennt.

Sei dazu $f_{N,i,k} := P(Y_0, ..., Y_{k-1} \in \{1, ..., N-1\}, Y_k = 0 \,|\, Y_0 = i)$. Dann genügen die $f_{N,i,k}$ der Differenzengleichung

$$f_{N,i,k+1} = p \cdot f_{N,i-1,k} + q \cdot f_{N,i+1,k}, \quad i = 1, ..., N-1, \tag{1.11}$$

mit den Randbedingungen $f_{N,0,k} = f_{N,N,k} = 0$ für $k \geq 1$, $f_{N,0,0} = 1$ und $f_{N,i,0} = 0$ für $i = 1, ..., N$. Ist $F_{N,i}(z) := \sum_{k=0}^{\infty} f_{N,i,k} \cdot z^k$ die zugehörige erzeugende Funktion, so liefert (1.11) durch Multiplikation mit z^{k+1} und Summation über k

$$F_{N,i}(z) = pz \cdot F_{N,i-1}(z) + qz \cdot F_{N,i+1}(z) \tag{1.12}$$

mit den Randbedingungen $F_{N,0}(z) = 1$ und $F_{N,N}(z) = 0$. Bei festem z ist dies eine Differenzengleichung 2. Ordnung mit konstanten Koeffizienten und als Lösung ergibt sich

$$F_{N,i}(z) = \left(\frac{p}{q}\right)^i \frac{\lambda_1(z)^{N-i} - \lambda_2(z)^{N-i}}{\lambda_1(z)^N - \lambda_2(z)^N} \tag{1.13}$$

mit

$$\lambda_{1,2}(z) = \frac{1}{2qz} \left(1 \pm \sqrt{1 - 4pqz^2}\right).$$

(Vgl. auch das klassische Ruin-Problem, [Fel68], Kapitel XIV, S. 342ff.)

Man überzeugt sich leicht davon, dass nun die Gleichungen

$$P_0(Y_{\gamma_1} = 0) = F_{N,1}(1) \quad \text{und} \quad g_{\gamma_1|Y_{\gamma_1}=0}(z) = \frac{F_{N,1}(z) \cdot z}{F_{N,1}(1)} \tag{1.14}$$

gelten. ($F_{N,1}(1)$ meint den linksseitigen Grenzwert von $F_{N,1}(z)$ in 1.)

Mit ähnlichen Überlegungen gewinnt man für $H_{N,i}(z) := \sum_{k=0}^{\infty} h_{N,i,k} \cdot z^k$ mit $h_{N,i,k} := P(Y_0, ..., Y_{k-1} \in \{1, ..., N-1\}, Y_k = N \,|\, Y_0 = i)$ die Darstellung

$$H_{N,i}(z) = \frac{\lambda_1(z)^i - \lambda_2(z)^i}{\lambda_1(z)^N - \lambda_2(z)^N}, \tag{1.15}$$

sowie die Gleichungen

$$P_0(Y_{\gamma_1} = N) = H_{N,1}(1) \quad \text{und} \quad g_{\gamma_1|Y_{\gamma_1}=N}(z) = \frac{H_{N,1}(z) \cdot z}{H_{N,1}(1)}. \tag{1.16}$$

Für die wahrscheinlichkeitserzeugende Funktion der Wartezeit T_N ergibt sich somit aus (1.9) die Darstellung

$$g_{T_N}(z) = \frac{H_{N,1}(1)}{1 - F_{N,1}(z) \cdot z} \cdot \frac{H_{N,1}(z) \cdot z}{H_{N,1}(1)}. \tag{1.17}$$

Diese werden wir nun verwenden, um das Verhalten von T_N/ET_N mit $N \to \infty$ zu analysieren. Insbesondere kommt es uns darauf an, die Einflüsse der Zufallssumme von Rückkehrzeiten nach Null und der finalen Exkursion nach N auf das Grenzverhalten getrennt voneinander zu untersuchen. Zunächst betrachten wir dabei nochmals den bereits im ersten Abschnitt behandelten Fall einer symmetrischen Irrfahrt, anschließend dann Irrfahrten mit Drift.

Sei also zunächst $p = 1/2$. Hier vereinfachen sich die beteiligten Terme zu

$$F_{N,1}(z) = \frac{\lambda_1(z)^{N-1} - \lambda_2(z)^{N-1}}{\lambda_1(z)^N - \lambda_2(z)^N}, \quad H_{N,1}(z) = \frac{\lambda_1(z) - \lambda_2(z)}{\lambda_1(z)^N - \lambda_2(z)^N} \tag{1.18}$$

mit $\lambda_{1,2}(z) = (1/z)\big(1 \pm \sqrt{1-z^2}\big)$. Unter Verwendung von $\lambda_1(z)\lambda_2(z) = 1$, also $\log \lambda_1(z) = -\log \lambda_2(z)$ und der Darstellung $\log \lambda_1(z) = \operatorname{arcosh}\big(z^{-1}\big)$ erhält man alternativ

$$F_{N,1}(z) = \frac{\sinh\big((N-1)\operatorname{arcosh}(z^{-1})\big)}{\sinh\big(N\operatorname{arcosh}(z^{-1})\big)}, \quad H_{N,1}(z) = \frac{\sinh\big(\operatorname{arcosh}(z^{-1})\big)}{\sinh\big(N\operatorname{arcosh}(z^{-1})\big)}. \tag{1.19}$$

Aus der Darstellung (1.18) folgert man

$$P_0(Y_{\gamma_1} = 0) = F_{N,1}(1) = \frac{N-1}{N}, \quad P_0(Y_{\gamma_1} = N) = H_{N,1}(1) = \frac{1}{N}. \tag{1.20}$$

(Man verwende die Regel von l'Hospital und $\lambda'_{1,2}(z) \sim \mp 1/\sqrt{1-z^2}$ für $z \uparrow 1$.)

(1.14) und (1.16) liefern

$$E[\gamma_1 | Y_{\gamma_1} = 0] = 1 + \frac{F'_{N,1}(1)}{F_{N,1}(1)}, \quad E[\gamma_1 | Y_{\gamma_1} = N] = 1 + \frac{H'_{N,1}(1)}{H_{N,1}(1)}. \tag{1.21}$$

Hieraus erhalten wir

Lemma 1.2 *Es ist* $E[\gamma_1 | Y_{\gamma_1} = 0] = (2/3)(N+1)$, $E[\gamma_1 | Y_{\gamma_1} = N] = (1/3)(2+N^2)$ *und* $ET_N = N^2$.

Beweis: Sind erst einmal die behaupteten Darstellungen für $E[\gamma_1|Y_{\gamma_1} = 0]$ und $E[\gamma_1|Y_{\gamma_1} = N]$ gezeigt, so ist $ET_N = N^2$ eine einfache Folgerung aus (1.10) in Verbindung mit (1.20).

Zum Beweis der Darstellung von $E[\gamma_1|Y_{\gamma_1} = 0]$: Wegen (1.20) und (1.21) ist hier $F'_{N,1}(1) = \big((N-1)/N\big)\big((2N-1)/3\big)$ zu zeigen. Mit (1.18) ergibt sich

$$\begin{aligned}
F'_{N,1}(1) &= \lim_{z\to 1} \frac{(N-1)/N - F_{N,1}(z)}{1-z} \\
&= \lim_{z\to 1} \frac{(N-1)\big((1+\sqrt{1-z^2})^N - (1-\sqrt{1-z^2})^N\big)}{N(1-z)\big((1+\sqrt{1-z^2})^N - (1-\sqrt{1-z^2})^N\big)} \\
&\quad - \frac{zN\big((1+\sqrt{1-z^2})^{N-1} - (1-\sqrt{1-z^2})^{N-1}\big)}{N(1-z)\big((1+\sqrt{1-z^2})^N - (1-\sqrt{1-z^2})^N\big)} \\
&= \lim_{z\to 1} \frac{(N-1)\big(2N\sqrt{1-z^2} + 2\binom{N}{3}(\sqrt{1-z^2})^3 + \mathcal{O}\big((\sqrt{1-z^2})^5\big)\big)}{N(1-z)\big(2N\sqrt{1-z^2} + \mathcal{O}\big(\sqrt{1-z^2})^3\big)\big)} \\
&\quad - \frac{zN\big(2(N-1)\sqrt{1-z^2} + 2\binom{N-1}{3}(\sqrt{1-z^2})^3 + \mathcal{O}\big((\sqrt{1-z^2})^5\big)\big)}{N(1-z)\big(2N\sqrt{1-z^2} + \mathcal{O}\big(\sqrt{1-z^2})^3\big)\big)} \\
&= \lim_{z\to 1} \frac{2N(N-1)(1-z)\sqrt{1-z^2} + 2\binom{N}{3}(N-1)(\sqrt{1-z^2})^3}{2N^2(1-z)\sqrt{1-z^2} + \mathcal{O}\big((1-z)(\sqrt{1-z^2})^3\big)} \\
&\quad - \frac{2\binom{N-1}{3}Nz(\sqrt{1-z^2})^3 + \mathcal{O}\big((\sqrt{1-z^2})^5\big)}{2N^2(1-z)\sqrt{1-z^2} + \mathcal{O}\big((1-z)(\sqrt{1-z^2})^3\big)} \\
&= \lim_{z\to 1} \frac{2N(N-1) + 2\binom{N}{3}(N-1)(1+z) - 2\binom{N-1}{3}Nz(1+z) + o(1)}{2N^2 + o(1)} \\
&= \frac{N-1}{N} \cdot \frac{2N-1}{3}.
\end{aligned}$$

Mit ähnlichen Umformungen beweist man die Darstellung für $E[\gamma_1|Y_{\gamma_1} = N]$. □

Es gilt also $ET_N = N^2$, und der Anteil der Rückkehrzeiten nach 0 an T_N verhält sich im Mittel wie $(2/3)N^2$ und die finale Exkursion wie $(1/3)N^2$.

Um das Verhalten von T_N/ET_N mit $N \to \infty$ zu analysieren, gehen wir nun wieder zu der zugehörigen charakteristischen Funktion $\varphi_N(\theta)$ über, wobei wir (1.5) und (1.17) verwenden. Es ist dann

$$\varphi_N(\theta) = \frac{1/N}{1 - F_{N,1}\big(\exp(i\theta/N^2)\big)\exp(i\theta/N^2)} \cdot \frac{H_{N,1}\big(\exp(i\theta/N^2)\big)\exp(i\theta/N^2)}{1/N}.$$

Der erste Faktor repräsentiert dabei den Beitrag der Zufallssumme von Rückkehrzeiten nach Null und der zweite Faktor den Beitrag der finalen Exkursion nach N. Als formale Vereinfachung verwenden wir für diese beiden Faktoren die Abkürzungen $\varphi_{N,1}(\theta)$ und $\varphi_{N,2}(\theta)$.

Für den ersten Faktor erhält man bei Verwendung von (1.19) die Darstellung

$$\varphi_{N,1}(\theta) = \frac{A(N,\theta)}{B(N,\theta) - C(N,\theta)}$$

mit

$$\begin{aligned} A(N,\theta) &:= \sinh\big(N\mathrm{arcosh}\big(\exp(-i\theta/N^2)\big)\big)\,, \\ B(N,\theta) &:= N\Big(\sinh\big(N\mathrm{arcosh}\big(\exp(-i\theta/N^2)\big)\big) - \sinh\sqrt{2i\theta}\Big)\,, \\ C(N,\theta) &:= N\Big(\exp(i\theta/N^2)\sinh\big((N-1)\mathrm{arcosh}\big(\exp(-i\theta/N^2)\big)\big) - \sinh\sqrt{2i\theta}\Big)\,. \end{aligned}$$

Mit Hilfssatz A.1 erhält man für $A(\theta,N)$ mit $N\to\infty$ den Grenzwert $\sinh\big(\sqrt{-2i\theta}\big)$. $B(N,\theta)$ ist offenbar ein Differenzenquotient und strebt mit $N\to\infty$ gegen den Wert der Ableitung der Funktion $b(x) := \sinh\big((1/x)\mathrm{arcosh}\big(\exp(-i\theta x^2)\big)\big)$ im Nullpunkt. Führt man die Differentiation aus, so liefert Hilfssatz A.1 hierfür den Wert 0. Ebenso lässt sich der letzte Term $C(N,\theta)$ als Differenzenquotient in Null der Funktion $c(x) := \exp(i\theta x^2)\sinh\big((1-x)(1/x)\mathrm{arcosh}\big(\exp(-i\theta x^2)\big)\big)$ interpretieren, und mit denselben Methoden ergibt sich hier der Grenzwert $-\sqrt{-2i\theta}\cdot\cosh(\sqrt{-2i\theta})$.

Insgesamt folgt somit für den ersten Faktor der Grenzwert

$$\lim_{N\to\infty}\varphi_{N,1}(\theta) = \frac{1}{\sqrt{-2i\theta}}\cdot\tanh\big(\sqrt{-2i\theta}\big).$$

Auf demselben Weg erhält man für den zweiten Faktor im Limes

$$\lim_{N\to\infty}\varphi_{N,2}(\theta) = \sqrt{-2i\theta}\cdot\frac{1}{\sinh\big(\sqrt{-2i\theta}\big)},$$

so dass sich insgesamt für T_N/ET_N das aus dem ersten Abschnitt bekannte Grenzwertresultat

$$\lim_{N\to\infty}\varphi_N(\theta) = \cosh(\sqrt{-2i\theta})^{-1}$$

ergibt.

Wie auch schon die berechneten Erwartungswerte verdeutlichen, gelingt es also im symmetrischen Fall weder der Zufallssumme von Rückkehrzeiten nach 0, noch der finalen Exkursion nach N, einen dominierenden Einfluss auf die Wartezeit T_N zu erlangen. Vielmehr tragen beide Anteile zur Grenzverteilung von T_N/ET_N bei.

In diesem Zusammenhang sei noch erwähnt, dass die Interpretation der erhaltenen Grenzverteilungen über Wartezeiten der klassischen Brownschen Bewegung noch weiter verfolgt werden kann. Den Term $\cosh\left(\sqrt{-2i\theta}\right)^{-1}$ hatten wir ja bereits im ersten Abschnitt als die charakteristische Funktion der Wartezeit interpretiert, bis eine Brownsche Bewegung erstmals vom Betrag das Niveau 1 erreicht. Ebenso lassen sich $\left(\sqrt{-2i\theta}\right)^{-1} \cdot \tanh\left(\sqrt{-2i\theta}\right)$ als charakteristische Funktion des letzten Nulldurchgangs vor Erreichen des Niveaus 1 und $\sqrt{-2i\theta} \cdot \sinh\left(\sqrt{-2i\theta}\right)^{-1}$ als charakteristische Funktion der Differenz dieser beiden Zeitpunkte interpretieren. (Vgl. hierzu [Yor97], S. 132f.)

Man sieht also, dass die strukturierte Herangehensweise mit den Methoden aus Abschnitt 2 eine Bestätigung der in Abschnitt 1 hergeleiteten Ergebnisse ermöglicht, wobei nun eine Interpretation der Wartezeitverteilung als Faltung von zwei verschiedenartigen Beiträgen möglich ist. Darüber hinaus werden wir nun zeigen, dass diese Methoden es auch ermöglichen, das asymptotische Verhalten der Wartezeitverteilung bei Irrfahrten mit Drift zu analysieren.

Sei also nun $p \neq q$, das heißt es liegt entweder ein Drift nach 0 ($p > q$) oder ein Drift nach ∞ ($p < q$) vor. Im Fall $p > q$ gilt $\lambda_1(1) = p/q$ und $\lambda_2(1) = 1$, sowie $\lambda_1'(1) = p/\big(q(1-2p)\big)$ und $\lambda_2'(1) = -1/(1-2p)$. Im Fall $p < q$ vertauschen λ_1 und λ_2 ihre Rollen. Aus den Gleichungen (1.14) und (1.16) erhält man somit durch Differentiation der wahrscheinlichkeitserzeugenden Funktionen im Punkte 1 in beiden Fällen die Erwartungswerte

$$E[\gamma_1 | Y_{\gamma_1} = 0] = 1 + \frac{1}{1-2p}\left(N \cdot \frac{1+(p/q)^N}{1-(p/q)^N} - (N-1) \cdot \frac{1+(p/q)^{N-1}}{1-(p/q)^{N-1}}\right)$$

und

$$E[\gamma_1 | Y_{\gamma_1} = N] = 1 + \frac{1}{1-2p}\left(N \cdot \frac{1+(p/q)^N}{1-(p/q)^N} - \frac{1+p/q}{1-p/q}\right).$$

Aus (1.10) erhält man also für den Erwartungswert von T_N

$$\begin{aligned} ET_N &= \frac{p}{q}\frac{1-(p/q)^{N-1}}{1-p/q}\cdot\Bigg[1+\frac{1}{1-2p}\Bigg(N\cdot\frac{1+(p/q)^N}{1-(p/q)^N} \\ &\quad -(N-1)\cdot\frac{1+(p/q)^{N-1}}{1-(p/q)^{N-1}}\Bigg)\Bigg] \\ &\quad +1+\frac{1}{1-2p}\left(N\cdot\frac{1+(p/q)^N}{1-(p/q)^N}-\frac{1+p/q}{1-p/q}\right). \end{aligned}$$

(Für $p \to 1/2$ ergeben sich die aus dem symmetrischen Fall bekannten Grenzwerte.)

Aus dieser Formel für den Erwartungswert ET_N und der bereits bekannten Darstellung im symmetrischen Fall lässt sich zusammenfassend folgendes Lemma ableiten:

Lemma 1.3 *(a) Im symmetrischen Fall $p = q$ ist $ET_N = N^2$ und es gilt*

$$E\Big(\sum_{k=1}^{M_N-1}\gamma_k\Big) = \frac{2}{3}N^2+o(N^2),\quad E(\gamma_{M_N}) = \frac{1}{3}N^2+o(N^2),$$

d.h. die Zufallssumme von Rückkehrzeiten nach Null trägt im Mittel 2/3 und die finale Exkursion nach N 1/3 zur Wartezeit T_N bei.
(b) Im Fall $p > q$, also eines Drifts nach 0, ist $ET_N = \Theta((p/q)^{N-1})$, und es gilt

$$E\Big(\sum_{k=1}^{M_N-1}\gamma_k\Big) = \Theta\big((p/q)^{N-1}\big),\quad E(\gamma_{M_N}) = \Theta(N),$$

d.h. der Beitrag der finalen Exkursion nach N zur Wartezeit T_N verschwindet asymptotisch. Insbesondere gilt $\gamma_{M_N}/ET_N \xrightarrow{d} 0$.
(c) Im Fall $p < q$, also eines Drifts nach ∞, ist $ET_N = \Theta(N)$, und es gilt

$$E\Big(\sum_{k=1}^{M_N-1}\gamma_k\Big) = \Theta(1),\quad E(\gamma_{M_N}) = \Theta(N),$$

d.h. der Beitrag der Zufallssumme von Rückkehrzeiten nach Null zur Wartezeit T_N verschwindet asymptotisch. Insbesondere gilt $\sum_{k=1}^{M_N-1}\gamma_k/ET_N \xrightarrow{d} 0$.

Somit verbleibt bei den Irrfahrten mit Drift nur noch die Aufgabe, den jeweils asymptotisch nicht verschwindenden Anteil der Wartezeit zu untersuchen.

Wir beginnen mit dem Drift nach 0, also $p > q$. Hier ist zu untersuchen, wie sich die mit ET_N normierte Summe von Rückkehrzeiten nach 0 vor der finalen Exkursion

mit $N \to \infty$ verhält. Die charakteristische Funktion zu T_N/ET_N gewinnt man nach wie vor mit (1.5) aus (1.17), den hier interessierenden ersten Faktor bezeichnen wir wiederum mit $\varphi_{N,1}(\theta)$, also

$$\varphi_{N,1}(\theta) = \frac{H_{N,1}(1)}{1 - F_{N,1}\big(\exp(i\theta/ET_N)\big)\exp(i\theta/ET_N)}.$$

Dabei ist

$$H_{N,1}(1) = \frac{1 - p/q}{1 - (p/q)^N}$$

und

$$F_{N,1}\big(\exp(i\theta/ET_N)\big) = \frac{2p\exp(i\theta/ET_N)\cdot\big(1 - B(N,\theta)^{N-1}/A(N,\theta)^{N-1}\big)}{A(N,\theta) - B(N,\theta)^N/A(N,\theta)^{N-1}}$$

mit $A(N,\theta) = 1 + \sqrt{1 - 4pq\exp(2i\theta/ET_N)}$, $B(N,\theta) = 1 - \sqrt{1 - 4pq\exp(2i\theta/ET_N)}$.

Durch entsprechende Reihenentwicklungen erhält man

$$A(N,\theta) = 2p - \frac{4pqi\theta}{2p-1}\big(ET_N\big)^{-1} + o\big((ET_N)^{-1}\big)\,, \tag{1.22}$$

$$B(N,\theta) = 2q + \frac{4pqi\theta}{2p-1}\big(ET_N\big)^{-1} + o\big((ET_N)^{-1}\big)\,, \tag{1.23}$$

und aus $\lim_{N\to\infty} N\log\big(A(N,\theta)/2p\big) = \lim_{N\to\infty} N\log\big(B(N,\theta)/2q\big) = 0$ folgt

$$A(N,\theta)^N \sim (2p)^N\,,\quad B(N,\theta)^N \sim (2q)^N\,. \tag{1.24}$$

Verwendet man $\exp(2i\theta/ET_N) = 1 + 2i\theta/ET_N + o\big((ET_N)^{-1}\big)$, so erhält man für $\varphi_{N,1}(\theta)$ die Darstellung

$$\varphi_{N,1}(\theta) = \frac{1 - p/q}{1 - (p/q)^N}\,ET_N\Big(C(N,\theta) - D(N,\theta) + o(1)\Big)^{-1}$$

mit

$$\begin{aligned} C(N,\theta) &:= ET_N\left(1 - 2p\,\frac{1 - B(N,\theta)^{N-1}/A(N,\theta)^{N-1}}{A(N,\theta) - B(N,\theta)^N/A(N,\theta)^{N-1}}\right)\,,\\ D(N,\theta) &:= 4pi\theta\,\frac{1 - B(N,\theta)^{N-1}/A(N,\theta)^{N-1}}{A(N,\theta) - B(N,\theta)^N/A(N,\theta)^{N-1}}\,. \end{aligned}$$

Trivial sind dabei für $N \to \infty$ die Grenzwerte $D(N,\theta) \to 2i\theta$, sowie $\big((1-p/q)/(1-(p/q)^N)\big)ET_N = 2p/(2p-1)$ für den Vorfaktor. Schreibt man $C(N,\theta)$ als

$$C(N,\theta) = \frac{ET_N\big(A(N,\theta) - 2p\big) + ET_N\big(2p - B(N,\theta)\big)B(N,\theta)^{N-1}/A(N,\theta)^{N-1}}{A(N,\theta) - B(N,\theta)^N/A(N,\theta)^{N-1}}\,,$$

so erhält man unter Verwendung von (1.22) und (1.24) für $N \to \infty$ die Grenzwerte

$$\begin{aligned} A(N,\theta) - B(N,\theta)^N / A(N,\theta)^{N-1} &\to 2p\,, \\ ET_N\big(A(N,\theta) - 2p\big) &\to -4pqi\theta/(2p-1)\,, \\ ET_N\big(2p - B(N,\theta)\big) B(N,\theta)^{N-1} / A(N,\theta)^{N-1} &\to 4p^2/(2p-1)\,. \end{aligned}$$

Insgesamt ergibt sich so

$$\lim_{N\to\infty} \varphi_{N,1}(\theta) \to \frac{1}{1-i\theta}.$$

Bei Drift nach Null ist also die mit ET_N normierte Zufallssumme von Rückkehrzeiten nach Null asymptotisch exponentialverteilt.

Bei Drift nach ∞, also $p < q$, ist es der zweite Faktor der charakteristischen Funktion, der den asymptotisch relevanten Teil zur Grenzverteilung beiträgt. Diesen bezeichnen wir wieder mit $\varphi_{N,2}(\theta)$, also

$$\varphi_{N,2}(\theta) := \frac{H_{N,1}\big(\exp(i\theta/ET_N)\big)\exp(i\theta/ET_N)}{H_{N,1}(1)}.$$

Es gilt nun $H_{N,1}(1) \to 1-p/q$, sowie $ET_N/N \to (1-2p)^{-1}$. Außerdem vertauschen in (1.22) und (1.23) $A(N,\theta)$ und $B(N,\theta)$ ihre Rollen, insbesondere gilt $A(N,\theta) \to 2q$, $B(N,\theta) \to 2p$ und $B(N,\theta)^N \to 0$. Mit Hilfe der Darstellung (1.15) für $H_{N,1}$ folgt somit

$$\varphi_{N,2}(\theta) \sim \exp\big((1-2p)i\theta\big) \cdot (2q)^N / A(N,\theta)^N.$$

Durch Logarithmieren erhält man bei Verwendung der nun gültigen Darstellung (1.23) für $A(N,\theta)$ den Grenzwert $A(N,\theta)^N/(2q)^N \to \exp(-2pi\theta)$, also insgesamt

$$\varphi_{N,2}(\theta) \to \exp(i\theta),$$

d.h. die mit ET_N normierte finale Exkursion nach N strebt im Fall eines Drifts nach ∞ in Verteilung gegen 1.

Fassen wir also noch einmal die Ergebnisse dieses Abschnitts zusammen: Es ging um die Analyse der Verteilung von T_N/ET_N für die Wartezeit T_N, bis die einfache Irrfahrt auf $\mathbb{Z}$ mit Start in 0 betragsmäßig erstmals das Niveau N erreicht. Dabei haben wir besonderes Augenmerk darauf gelegt, den Einfluss der zufälligen Summe von Rückkehrzeiten nach 0 und den Einfluss der finalen Exkursion nach N auf diese Wartezeit getrennt voneinander zu betrachten.

Im Fall einer Irrfahrt mit Drift nach 0 verschwindet der Einfluss der finalen Exkursion nach N mit $N \to \infty$, für die Zufallssumme von Rückkehrzeiten haben wir asymptotische Exponentialität. In Formeln:

$$\sum_{k=1}^{M_N-1} \gamma_k/ET_N \xrightarrow{\mathrm{d}} \mathrm{Exp}(1), \;\; \gamma_{M_N}/ET_N \xrightarrow{\mathrm{d}} 0, \;\; T_N/ET_N \xrightarrow{\mathrm{d}} \mathrm{Exp}(1).$$

Im Fall einer Irrfahrt mit Drift nach ∞ verschwindet der Einfluss der Zufallssumme von Rückkehrzeiten nach 0 mit $N \to \infty$, die finale Exkursion strebt bei Normierung mit ET_N in Verteilung gegen 1. In Formeln:

$$\sum_{k=1}^{M_N-1} \gamma_k/ET_N \xrightarrow{\mathrm{d}} 0, \;\; \gamma_{M_N}/ET_N \xrightarrow{\mathrm{d}} 1, \;\; T_N/ET_N \xrightarrow{\mathrm{d}} 1.$$

Der interessanteste Fall ist der einer symmetrischen Irrfahrt. Hier gelingt es keinem der beiden Anteile, einen dominanten Einfluss auf T_N zu gewinnen. Asymptotisch trägt die Zufallssumme von Rückkehrzeiten nach Null im Mittel 2/3 und die finale Exkursion 1/3 zu T_N bei, in Formeln:

$$E\Big(\sum_{k=1}^{M_N-1} \gamma_k\Big)/ET_N \to \frac{2}{3}, \;\; E(\gamma_{M_N})/ET_N \to \frac{1}{3}.$$

Die erhaltenen Grenzverteilungen lassen sich über Wartezeiten bei der klassischen Brownschen Bewegung interpretieren.

Doch auch wenn der Fall der symmetrischen Irrfahrt auf den ersten Blick als der interessanteste erscheint, stellt er im Zusammenhang mit dem Thema dieser Dissertation nur den entarteten Grenzübergang zwischen dem für uns interessanten Fall der asymptotischen Exponentialität und dem eher unbedeutenden Bereich der Verteilungskonvergenz gegen eine Konstante dar.

In diesem Zusammenhang soll ein Ergebnis unserer Untersuchung der Irrfahrt auf $\mathbb{Z}$ noch einmal deutlich herausgestellt werden: Asymptotische Exponentialität stellt sich hier genau dann ein, wenn das Ereignis, auf das man wartet, – also hier das Erreichen des Niveaus N – „hinreichend“ unwahrscheinlich wird. „Hinreichend“ scheint in diesem Zusammenhang zu bedeuten, dass ein Drift nach 0 vorliegt, so dass ein Erreichen des Niveaus N entsprechend schwierig ist und der Beitrag der finalen Exkursion nach N asymptotisch verschwindet, so dass alleine die Zufallssumme von Rückkehrzeiten in den Ausgangspunkt die Grenzverteilung von T_N bestimmt.

Den Abschluss dieses Abschnitts bilden zwei Abbildungen, die die hergeleiteten Resultate veranschaulichen. Abbildung 1.2 zeigt für den Fall $N = 20$ Realisierungen von Simulationen für $(Y_n)_{n\in\mathbb{N}_0}$ in allen drei betrachteten Situationen. Deutlich ist zu erkennen, dass in der linken Darstellung des symmetrischen Falls sowohl die Zufallssumme von Rückkehrzeiten nach 0 als auch die finale Exkursion einen Beitrag zu T_N leisten, wohingegen im in der Mitte dargestellten Fall eines Drifts nach 0 der erste Anteil und im rechts dargestellten Fall eines Drifts nach ∞ der zweite Anteil eindeutig überwiegen.

Abbildung 1.3 veranschaulicht in derselben Situation ($N = 20$) die hergeleiteten Resultate zur Verteilungskonvergenz anhand von jeweils 2000 Simulationen für T_N/ET_N. Im Fall eines Drifts nach 0 deckt sich der Tail der empirischen Verteilung (hier gelb dargestellt) sehr gut mit dem Tail einer Exp(1)-Verteilung (schwarze Linie). Bei der symmetrischen Irrfahrt zeigt der Tail der empirischen Verteilung (rote Linie) hingegen deutliche Abweichungen zur Exp(1)-Verteilung. Schließlich tendiert der Tail der empirischen Verteilung im Fall eines Drifts nach ∞ (blaue Linie) in Richtung des Tails des Einpunktmaßes in 1 (ebenfalls schwarz eingezeichnet). Die Abweichung erklärt sich aus der Tatsache, dass sich die Verteilungskonvergenz erst mit $N \to \infty$ einstellt, hier jedoch der Fall $N = 20$ simuliert wurde.

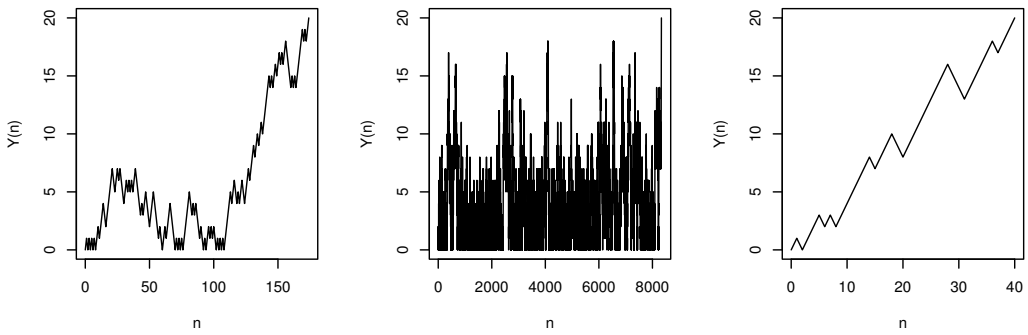

Abbildung 1.2: Drei Realisierungen der Irrfahrt $(Y_n)_{n\in\mathbb{N}_0}$ im Falle $N = 20$. Links symmetrischer Fall ($p = 1/2$), Mitte Drift nach Null ($p = 0.55$), rechts Drift nach ∞ ($p = 1/3$).

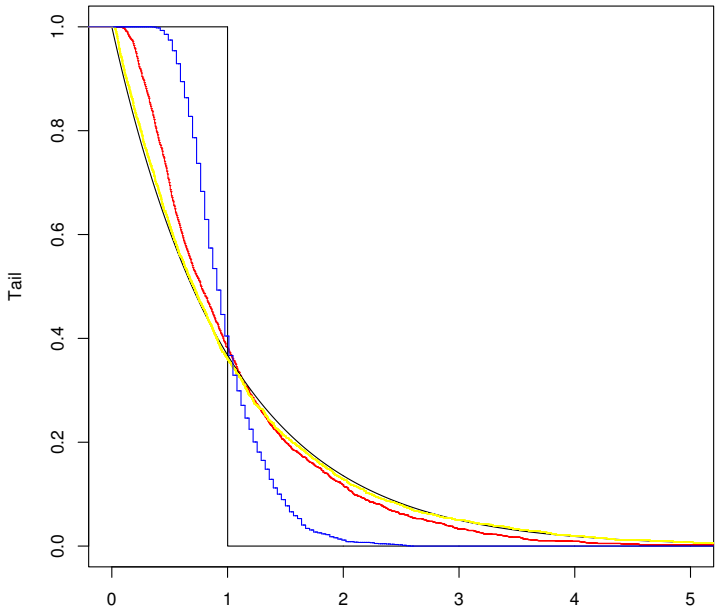

Abbildung 1.3: Graphische Darstellung der Tails zu einer Exp(1)-Verteilung und zum Einpunktmaß in 1 (jeweils schwarz), sowie der Tails zur empirischen Verteilung zu jeweils 2000 Simulationen von T_N/ET_N bei $N = 20$ im symmetrischen Fall $p = 1/2$ (rot), bei Drift nach 0 mit $p = 0.55$ (gelb) und bei Drift nach ∞ mit $p = 1/3$ (blau).

1.4 Lebensdauerprozesse

In diesem Abschnitt werden wir an zwei weiteren Beispielen Eintrittszeitverteilungen bei Markov-Ketten untersuchen, und erneut führen die im zweiten Abschnitt bereitgestellten Methoden zum Ziel.

Wir betrachten zwei klassische Prozesse der Erneuerungstheorie, die in der Literatur unter den Begriffen *Residual Age Process* bzw. *Remaining Lifetime Process* und *Current Age Process* bekannt sind. Diese Prozesse spielen beispielsweise eine wichtige Rolle, wenn es darum geht, ein System von Bauelementen mit zufällig verteilter Lebensdauer stochastisch zu beschreiben. Wird ein solches Bauelement nach seinem Ausfall stets sofort durch ein neues ersetzt, so beschreibt der *Current Age Process* das Alter der aktuell im Gebrauch befindlichen Komponente, der *Residual Age Process* die noch verbleibende Zeitspanne, bis diese Komponente wieder durch eine neue ersetzt werden muss. Mathematisch formuliert geht es um die folgenden beiden Prozesse:

Beim *Residual Age Process* handelt es sich um die Markov-Kette $X = (X_n)_{n\in\mathbb{N}_0}$ mit Start in 0 und der Übergangsmatrix $P_{\mathrm{RA}} = (p_{jk})_{j,k\in\mathbb{N}_0}$,

$$p_{0,k} = p_k,\ p_{k,k-1} = 1,\ k \geq 1,$$

wobei $(p_k)_{k\in\mathbb{N}}$ eine Folge in $(0,1)$ mit $\sum p_k = 1$ bezeichnet. Der *Current Age Process* bezeichnet die Markov-Kette $X = (X_n)_{n\in\mathbb{N}_0}$ mit Start in 0 und der Übergangsmatrix $P_{\mathrm{CA}} = (p_{jk})_{j,k\in\mathbb{N}_0}$,

$$p_{01} = 1,\ p_{k,k+1} = \sum_{j>k} p_j \Big/ \sum_{j\geq k} p_j,\ p_{k,0} = 1 - p_{k,k+1},\ k \geq 1.$$

Auch ohne die erneuerungstheoretische Interpretation erklären sich die Bezeichnungen der Prozesse insofern von selbst, dass ganz offensichtlich der Wert des *Residual Age Process* zur Zeit n der „verbleibenden Lebensdauer“ entspricht, bis der Prozess wieder nach 0 zurückkehrt, während der *Current Age Process* die “aktuelle Lebenszeit“ seit dem letzten Besuch in 0 angibt.

Wir verwenden an dieser Stelle für beide Prozesse dieselben Bezeichnungen, weil sich die folgenden Betrachtungen in Abhängigkeit vom zu Grunde liegenden Prozess nur minimal unterscheiden werden und somit Arbeit gespart werden kann. Wenn es

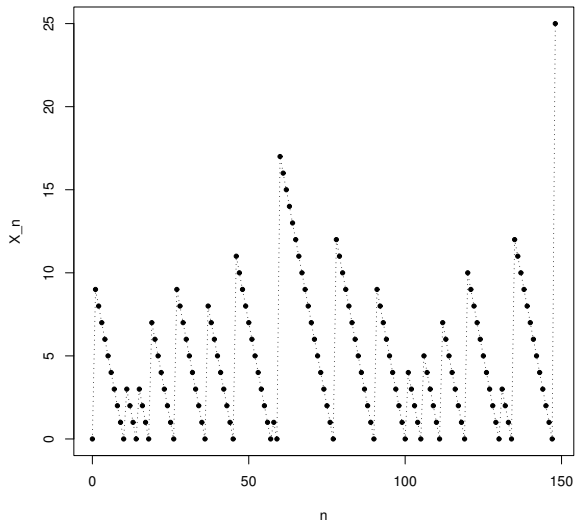

Abbildung 1.4: Beispiel für einen *Remaining Lifetime Process*. $(p_k)_{k\in\mathbb{N}}$ ist hier die Zähldichte einer geometrischen Verteilung auf $\mathbb{N}$ mit Erfolgswahrscheinlichkeit 1/10.

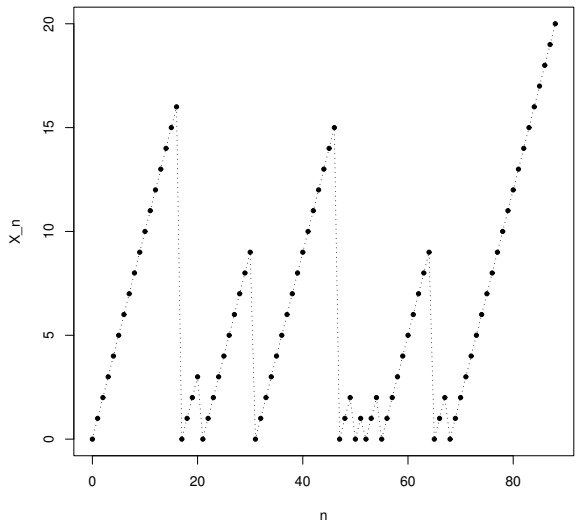

Abbildung 1.5: Beispiel für einen *Current Age Process*. $(p_k)_{k\in\mathbb{N}}$ ist hier die Zähldichte einer geometrischen Verteilung auf $\mathbb{N}$ mit Erfolgswahrscheinlichkeit 1/10.

auf den Unterschied ankommt, ergibt sich der gemeinte Prozess aus den jeweiligen Ausführungen.

Auch hier analysieren wir wieder die Wartezeit, bis der Prozess X erstmals das Niveau N erreicht, d.h. $T_N := \inf\{n \in \mathbb{N}_0 : X_n \geq N\}$. Ähnlich wie bei der Irrfahrt auf $\mathbb{Z}$, wo wir zwischen den verschiedenen Driftmöglichkeiten unterschieden haben, hat bei diesen Prozessen einen entscheidenden Einfluss auf das Grenzverhalten, ob die Verteilung $(p_k)_{k\in\mathbb{N}}$ einen endlichen Erwartungswert $\mu := \sum_{k=1}^{\infty} kp_k$ besitzt oder nicht.

Wir beginnen zunächst mit dem Fall $\mu < \infty$ und betrachten als erstes den *Residual Age Process*. Wieder lässt sich T_N im Sinne von (1.6) darstellen:

$$T_N = \sum_{k=1}^{M_N} \gamma_k$$

mit $\sigma_0 := 0$, $\sigma_k := \min\big\{n > \sigma_{k-1} : X_n \in \{0, N, N+1, ...\}\big\}$, $k \geq 1$, $\gamma_k := \sigma_k - \sigma_{k-1}$, $k \geq 1$, und $M_N := \min\big\{k \geq 1 : X_{\sigma_k} \geq N\big\}$.

Unmittelbar einsichtig sind die Gleichungen $P_0(X_{\gamma_1} \geq N) = \sum_{k=N}^{\infty} p_k$, $P_0(X_{\gamma_1} = 0) = \sum_{k=1}^{N-1} p_k$ und $g_{\gamma_1|X_{\gamma_1}\geq N}(z) = z$, und aus $P_0(\gamma_1 = k, X_{\gamma_1} = 0) = p_{k-1}$ für $2 \leq k \leq N$ und 0 sonst ergibt sich $g_{\gamma_1|X_{\gamma_1}=0}(z) = \big(z\sum_{k=1}^{N-1} p_k z^k\big)/\big(\sum_{k=1}^{N-1} p_k\big)$. Somit ist nach (1.9) die wahrscheinlichkeitserzeugende Funktion von T_N gegeben durch

$$g_{T_N}(z) = \frac{\sum_{k=N}^{\infty} p_k}{1 - z\sum_{k=1}^{N-1} p_k z^k} \cdot z\,. \tag{1.25}$$

Auf die gleiche Weise erhält man für den *Current Age Process* die Gleichungen $P_0(X_{\gamma_1} \geq N) = \sum_{k=N}^{\infty} p_k$, $P_0(X_{\gamma_1} = 0) = \sum_{k=1}^{N-1} p_k$, $g_{\gamma_1|X_{\gamma_1}\geq N}(z) = z^N$, sowie $g_{\gamma_1|X_{\gamma_1}=0}(z) = \big(z\sum_{k=1}^{N-1} p_k z^k\big)/\big(\sum_{k=1}^{N-1} p_k\big)$. Somit erhält man in diesem Fall für die wahrscheinlichkeitserzeugende Funktion von T_N die Darstellung

$$g_{T_N}(z) = \frac{\sum_{k=N}^{\infty} p_k}{1 - z \cdot \sum_{k=1}^{N-1} p_k z^k} \cdot z^N\,. \tag{1.26}$$

Man beachte, dass sich die wahrscheinlichkeitserzeugenden Funktionen in beiden Fällen kaum unterscheiden. Lediglich die letzte Exkursion von 0 nach N hat in

beiden Prozessen einen unterschiedlichen Einfluss auf die Wartezeit T_N. Dabei ist klar, dass diese im *Residual Age Process* stets die Länge 1 und im *Current Age Process* stets die Länge N hat. Interessant ist nun die Frage, ob dieser Unterschied auch zu einem unterschiedlichen Konvergenzverhalten von T_N/ET_N mit $N \to \infty$ führt.

Mit (1.10) gewinnt man für den Erwartungswert ET_N im *Residual Age Process* bzw. im *Current Age Process* die Darstellungen

$$ET_N = \frac{\sum_{k=1}^{N-1}(k+1)p_k}{\sum_{k=N}^{\infty} p_k} + 1 \quad \text{bzw.} \quad ET_N = \frac{\sum_{k=1}^{N-1}(k+1)p_k}{\sum_{k=N}^{\infty} p_k} + N. \tag{1.27}$$

Verwendet man nun wieder (1.5) zusammen mit (1.25) bzw. (1.26) und der abkürzenden Schreibweise

$$\phi_N(\theta) := \frac{\sum_{k=N}^{\infty} p_k}{1 - \exp(i\theta/ET_N)\sum_{k=1}^{N-1} p_k \exp(ik\theta/ET_N)},$$

so ergeben sich für die charakteristische Funktion φ_N von T_N/ET_N die Darstellungen

$$\varphi_N(\theta) = \phi_N(\theta) \cdot \exp(i\theta/ET_N) \quad \text{bzw.} \quad \varphi_N(\theta) = \phi_N(\theta) \cdot \exp(iN\theta/ET_N). \tag{1.28}$$

Man beachte jedoch, dass sich der Erwartungswert ET_N in beiden Fällen gemäß (1.27) unterscheidet; insbesondere wirkt sich dies auch auf die Funktion $\phi_N(\theta)$ aus.

Die genaue Asymptotik des Erwartungswerts ET_N hängt hier natürlich von der speziellen Wahl der Verteilung $(p_k)_{k\in\mathbb{N}}$ ab; unabhängig davon gilt aber für beide Prozesse in jedem Fall $\lim_{N\to\infty} ET_N = \infty$. Der Kehrwert des Faktors $\phi_N(\theta)$ lässt sich darstellen als

$$\begin{aligned}\phi_N(\theta)^{-1} \;=\; & \frac{-i\theta/ET_N}{\sum_{k=N}^{\infty} p_k} \cdot \frac{1 - \exp(i\theta/ET_N)\sum_{k=1}^{\infty} p_k \exp(ik\theta/ET_N)}{0 - i\theta/ET_N} \\ & + \frac{\exp(i\theta/ET_N)\sum_{k=N}^{\infty} p_k \exp(ik\theta/ET_N)}{\sum_{k=N}^{\infty} p_k}.\end{aligned}$$

Für den Vorfaktor des ersten Summanden gilt dabei $\lim_{N\to\infty} \big(-i\theta/ET_N\big)\big/\big(\sum_{k=N}^{\infty} p_k\big) = -i\theta/(1+\mu)$. Der zweite Faktor des ersten Summanden kann als Differenzenquotient der Funktion $f(x) := -i\exp(ix)\sum_{k=1}^{\infty} p_k \exp(ikx)$ in 0 interpretiert werden, folglich ergibt sich hier mit $N \to \infty$ die Ableitung von f in 0, also $1+\mu$. Der zweite Summand strebt schließlich gegen 1, denn

$$\sum_{k=N}^{\infty} \frac{p_k}{\sum_{l=N}^{\infty} p_l} \exp\big(i(k+1)\theta/ET_N\big)$$

lässt sich interpretieren als charakteristische Funktion einer Zufallsvariablen Y_N mit $P\big(Y_N = (k+1)/ET_N\big) = p_k / \sum_{l=N}^{\infty} p_l$ für $k \geq N$. Wegen $EY_N \to 0$ folgt $Y_N \xrightarrow{\mathrm{d}} 0$, also strebt die zugehörige charakteristische Funktion gegen 1.

Insgesamt erhält man so für beide Prozesse

$$\lim_{N\to\infty} \Phi_N(\theta) = \frac{1}{1 - i\theta}. \tag{1.29}$$

Bei $\mu < \infty$ strebt also die zufällige Summe von Rückkehrzeiten nach Null vor der finalen Exkursion nach N bei Normierung mit ET_N gegen eine Exponentialverteilung. Da außerdem beim *Residual Age Process* $\lim_{N\to\infty} \exp(i\theta/ET_N) = 1$ und beim *Current Age Process* $\lim_{N\to\infty} \exp(iN\theta/ET_N) = 1$ gilt (wegen $\mu < \infty$ gilt $N\sum_{k=N}^{\infty} p_K \to 0$ mit $N \to \infty$), strebt in beiden Prozessen die finale Exkursion nach N bei Normierung mit ET_N in Verteilung gegen 0, ist also schließlich vernachlässigbar. Insgesamt erhalten wir damit folgendes Resultat:

Satz 1.4 *Im Fall $\mu < \infty$ gilt für den* Residual Age Process *und für den* Current Age Process

$$\frac{T_N}{ET_N} \xrightarrow{d} \mathrm{Exp}(1). \tag{1.30}$$

Auch hier können wir wieder dasselbe Phänomen beobachten wie bei der Irrfahrt auf $\mathbb{Z}$: Die Wahrscheinlichkeit, dass der Prozess zwischen zwei Besuchen im Nullpunkt tatsächlich das Niveau N erreicht, geht mit $N \to \infty$ so schnell gegen 0, dass der Einfluss der finalen Exkursion auf die Wartezeit T_N asymptotisch verschwindet und in der Folge stellt sich für diese Wartezeit asymptotische Exponentialität ein.

Im Fall $\mu = \infty$ kann man im Allgemeinen keine asymptotische Exponentialität für T_N erwarten. Man erhält hier für unterschiedliche Verteilungen $(p_k)_{k\in\mathbb{N}}$ ein unterschiedliches Grenzverhalten für T_N/ET_N. Wir wollen beispielhaft eine Klasse von Verteilungen $(p_k)_{k\in\mathbb{N}}$ mit $\mu = \infty$ untersuchen, bei der sich keine asymptotische Exponentialität einstellt. Dies ist die Familie der Verteilungen $(p_k)_{k\in\mathbb{N}}$ mit $p_k := ck^{-\alpha}$ mit $\alpha \in (1, 2]$ und $c := \big(\sum_{k=1}^{\infty} k^{-\alpha}\big)^{-1}$. Der Exponent α ist hier also gerade so gewählt, dass $\sum_{k=1}^{\infty} k^{-\alpha}$ endlich ist, man also durch entsprechende Normierung ein Wahrscheinlichkeitsmaß erhält, andererseits aber der Erwartungswert μ nicht mehr existiert.

Natürlich gelten hier nach wie vor die Darstellungen (1.27) und (1.28) für den Erwartungswert und die charakteristische Funktion von T_N/ET_N. Mit Hilfssatz A.2 beweist man:

Lemma 1.5 *Im Fall* $\alpha \in (1,2)$ *gilt für den* Residual Age Process $ET_N \sim \frac{\alpha-1}{2-\alpha} N$, *für den* Current Age Process $ET_N \sim \frac{1}{2-\alpha} N$. *Im Fall* $\alpha = 2$ *gilt für beide Prozesse* $ET_N \sim N \log(N)$.

Für die gesamte Verteilung betrachten wir nun zunächst den *Residual Age Process*. Nach Lemma 1.5 gilt $ET_N \to \infty$, so dass sich unmittelbar $\exp(i\theta/ET_N) \to 1$ ergibt, d.h. die finale Exkursion ist weiterhin asymptotisch vernachlässigbar. Die Untersuchung von $\phi_N(\theta)$ gestaltet sich hingegen ein wenig komplizierter. Es gilt

$$\begin{aligned}
\phi_N(\theta)^{-1} &= \frac{1 - \exp(i\theta/ET_N) \sum_{k=1}^{N-1} p_k \exp(ik\theta/ET_N)}{\sum_{k=N}^{\infty} p_k} \\
&= 1 - \sum_{k=1}^{N-1} \frac{p_k}{\sum_{j=N}^{\infty} p_j} \sum_{m=1}^{\infty} \frac{\big(i\theta(k+1)\big)^m}{m!} \big(ET_N\big)^{-m} \\
&= 1 - \sum_{m=1}^{\infty} \frac{(i\theta)^m}{m!} \cdot \frac{\sum_{k=1}^{N-1} (k+1)^m p_k}{\sum_{j=N}^{\infty} p_j \big(ET_N\big)^m}.
\end{aligned}$$

Sei zunächst $\alpha \neq 2$. Aus Hilfssatz A.2 folgt $\sum_{k=1}^{N-1}(k+1)^m p_k \sim cN^{m+1-\alpha}/(m+1-\alpha)$ und $\sum_{j=N}^{\infty} p_j \sim cN^{(1-\alpha)}/(\alpha-1)$. Damit ergibt sich insgesamt:

$$\frac{\sum_{k=1}^{N-1}(k+1)^m p_k}{\sum_{j=N}^{\infty} p_j \big(ET_N\big)^m} \overset{N\to\infty}{\longrightarrow} \frac{\alpha-1}{m+1-\alpha} \cdot \left(\frac{2-\alpha}{\alpha-1}\right)^m.$$

Insbesondere gilt für genügend großes N

$$\begin{aligned}
\left| \frac{(i\theta)^m}{m!} \cdot \frac{\sum_{k=1}^{N-1}(k+1)^m p_k}{\sum_{j=N}^{\infty} p_j \big(ET_N\big)^m} \right| &\leq \frac{|\theta|^m}{m!} \cdot 2 \cdot \left| \frac{\alpha-1}{m-(\alpha-1)} \cdot \left(\frac{2-\alpha}{\alpha-1}\right)^m \right| \\
&= 2|\alpha-1| \cdot \left| \theta \frac{2-\alpha}{\alpha-1} \right|^m \frac{1}{m!} \frac{1}{m+1-\alpha},
\end{aligned}$$

und da die Reihe hierüber endlich ist, folgt mit dem Satz von der majorisierten Konvergenz

$$\lim_{N\to\infty} \phi_N(\theta)^{-1} = 1 - \sum_{m=1}^{\infty} \frac{\big(i\theta(2-\alpha)/(\alpha-1)\big)^m}{m!} \cdot \frac{\alpha-1}{m+1-\alpha}.$$

Im Falle $\alpha = 2$ ergibt sich ein fundamental anderes Grenzverhalten. Hier liefert Lemma 1.5 zusammen mit Hilfssatz A.2 für $m \geq 2$

$$\frac{\sum_{k=1}^{N-1}(k+1)^m p_k}{\sum_{j=N}^{\infty} p_j \big(ET_N\big)^m} \sim K_m \cdot \log(N)^{-m}, \quad K_m \in \mathbb{R},$$

so dass wir mit $N \to \infty$ den Grenzwert 0 erhalten. Im Fall $m = 1$ ergibt sich jedoch der Grenzwert 1, so dass insgesamt

$$\lim_{N\to\infty} \phi_N(\theta)^{-1} = 1 - i\theta,$$

gilt, also asymptotische Exponentialität.

Mit denselben Methoden kann man auch den *Current Age Process* untersuchen. Hier lautet die charakteristische Funktion zur finalen Exkursion $\exp(i\theta N/ET_N)$. Aus Lemma 1.5 erhält man für den Quotienten ET_N/N mit $N \to \infty$ den (im Fall $\alpha = 2$ uneigentlichen) Grenzwert $(2-\alpha)^{-1}$. Im Gegensatz zum *Residual Age Process* verschwindet die finale Exkursion hier also nicht, sondern strebt mit $N \to \infty$ in Verteilung gegen die Konstante $2 - \alpha$, die mit wachsendem α immer kleiner wird, bis sie für $\alpha = 2$ verschwindet.

Für $\phi_N(\theta)^{-1}$ haben wir wiederum dieselbe Darstellung wie zuvor, und unter Verwendung von Lemma 1.5 und Hilfssatz A.2 und mit denselben Methoden weist man hier mit $N \to \infty$ im Falle $\alpha \in (1, 2)$ den Grenzwert

$$\lim_{N\to\infty} \phi_N(\theta)^{-1} = 1 - \sum_{m=1}^{\infty} \frac{\big(i\theta(2-\alpha)\big)^m}{m!} \cdot \frac{\alpha - 1}{m + 1 - \alpha}$$

und im Falle $\alpha = 2$ den Grenzwert $\lim_{N\to\infty} \phi(\theta)^{-1} = 1 - i\theta$ nach. Insbesondere hat man also für $\alpha = 2$ wieder asymptotische Exponentialität.

Ein Ergebnis dieser Untersuchungen soll hierbei besonders herausgestellt werden: Bei der hier untersuchten Familie von Verteilungen $(p_k)_{k\in\mathbb{N}}$ stoßen wir beim *Residual Age Process* für $\alpha \in (1, 2)$ zum ersten Mal auf die Situation, dass die finale Exkursion mit $N \to \infty$ asymptotisch vernachlässigbar ist, sich aber trotzdem für die Gesamtverteilung der Eintrittszeit keine asymptotische Exponentialität einstellt. Dies allein kann also kein hinreichendes Kriterium für asymptotische Exponentialität sein.

Im nun folgenden Abschnitt werden wir einen Satz von Keilson kennen lernen, der in Situationen wie dieser und der im vorherigen Abschnitt untersuchten ein hinreichendes Kriterium für asymptotische Exponentialität angibt.

Den Abschluss dieses Abschnitts bildet wieder eine Abbildung, die die hergeleiteten Resultate graphisch veranschaulicht. Dargestellt sind für den *Residual Age Process* im Fall $N = 20$ die Tails der empirischen Verteilung für jeweils 5000 Simulationen

von T_N/ET_N im Vergleich zum Tail einer Exp(1)-Verteilung. Für $\alpha = 2$ (gelbe Linie) spiegelt sich die asymptotische Exponentialität in der sehr guten Übereinstimmung mit dem Tail der Exp(1)-Verteilung (schwarze Linie) wieder. Mit fallendem α wird die Abweichung dann immer größer.

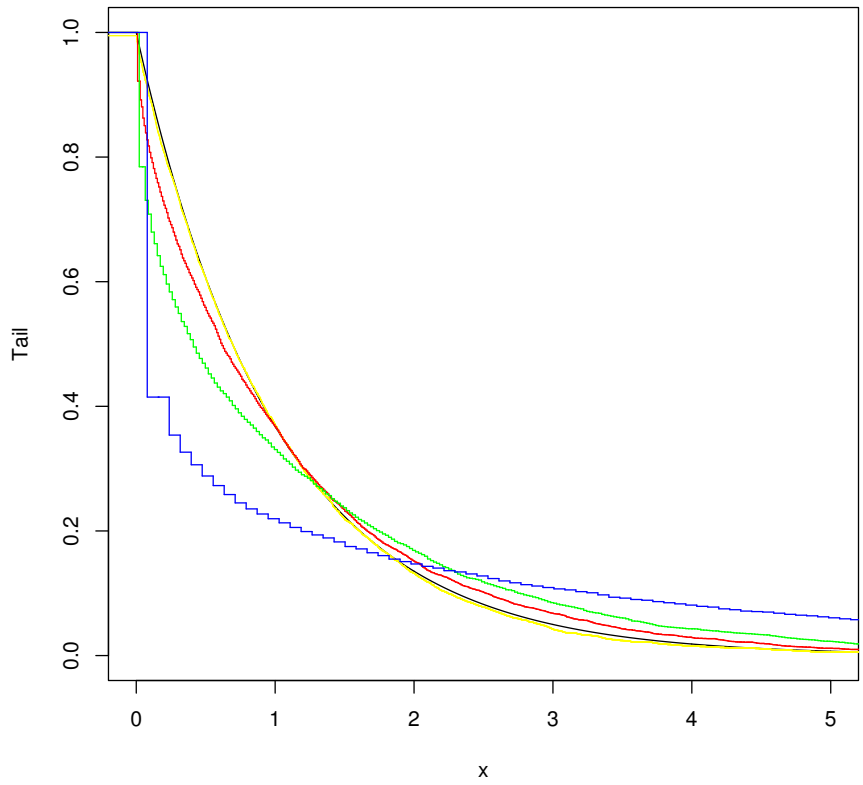

Abbildung 1.6: Graphische Darstellung des Tails zu einer Exp(1)-Verteilung (schwarz), sowie der Tails zur empirischen Verteilung von jeweils 5000 Simulationen von T_N/ET_N für den *Residual Age Process* bei $N = 20$ mit Verteilung $(p_k)_{k\in\mathbb{N}}$, $p_k := ck^{-\alpha}$, $c := \left(\sum_{k=1}^{\infty} k^{-\alpha}\right)^{-1}$, für die Fälle $\alpha = 2$ (gelb), $\alpha = 1.5$ (rot), $\alpha = 1.3$ (grün) und $\alpha = 1.1$ (blau).

1.5 Verallgemeinerung eines Satzes von Keilson

In den beiden vorangegangenen Abschnitten haben wir eine ganze Reihe von Eintrittszeitverteilungen bei Markov-Ketten kennen gelernt, bei denen sich asymptotische Exponentialität einstellt. Allen diesen Beispielen war gemeinsam, dass bei einer Zerlegung der Eintrittszeit in eine Zufallssumme von Rückkehrzeiten in den Ausgangszustand und eine finale Exkursion die mit dem Erwartungswert der Eintrittszeit normierte finale Exkursion asymptotisch verschwindet. Auf der anderen Seite zeigt jedoch das Beispiel des *Residual Age Process* mit $(p_k)_{k\in\mathbb{N}}$, $p_k := ck^{-\alpha}$, $c := \left(\sum_{k=1}^{\infty} k^{-\alpha}\right)^{-1}$, für $\alpha \in (1,2)$, dass dieses Kriterium allein nicht hinreichend ist.

Wir werden nun einen Satz von Keilson kennen lernen, der alle diese Phänomene als Resultat eines allgemeineren Prinzips erklärt. In diesem Zusammenhang nennen wir einen Prozess $(X_t)_{t\in T}$, $T \subset \mathbb{R}_{\geq 0}$, *regenerativ*, wenn eine Folge von Stoppzeiten (Regenerationszeiten) $(\sigma_k)_{k\in\mathbb{N}_0}$ mit Zuwächsen $\gamma_k := \sigma_k - \sigma_{k-1}$, $k \geq 1$, bzgl. der natürlichen Filtration von $(X_t)_{t\in T}$ existiert, so dass gilt

(a) $\left((X_t)_{t\in T\cap(\sigma_n,\infty)}, (\gamma_k)_{k>n}\right)$ und $\left((X_t)_{t\in T\cap[0,\sigma_n]}, \sigma_0, ..., \sigma_n\right)$ sind für alle $n \in \mathbb{N}_0$ stochastisch unabhängig und

(b) $\left((X_t)_{t\in T\cap(\sigma_n,\infty)}, (\gamma_k)_{k>n}\right)$ sind für alle $n \in \mathbb{N}_0$ identisch verteilt.

Insbesondere sind dann die *Zyklen* $(X_{\sigma_n+t})_{0<t\leq\gamma_{n+1}}$, $n \in \mathbb{N}_0$, unabhängig und identisch verteilt. (Vgl. hierzu [Als91], §10, S. 226ff.)

Die Irrfahrt auf $\mathbb{Z}$, der *Residual* und der *Current Age Process* aus den vorangegangenen Abschnitten fallen alle unter diese Modellannahmen. Als Folge von Regenerationszeitpunkten haben wir hier die Nulldurchgänge des betrachteten Prozesses verwendet, die geforderten Bedingungen (a) und (b) sind dann einfache Konsequenzen der Markov-Eigenschaft.

Unter Umständen kann es jedoch auch sinnvoll sein, eine weniger nahe liegende Folge von Regenerationszeiten zu betrachten. Ein solches Beispiel werden wir im folgenden Abschnitt im Zusammenhang mit dem Ehrenfestschen Urnenmodell kennen lernen.

Satz 1.6 (Keilson, [Kei66]) *Es sei $(X_t)_{t\in T}$, $T \subset \mathbb{R}_{\geq 0}$, ein regenerativer Prozess mit Zustandsraum $\mathcal{X}$. Für jedes $N \in \mathbb{N}$ sei $D_N := \{\mathcal{X}_{N,1}, \mathcal{X}_{N,2}\}$ eine Zerlegung des Zustandsraums in zwei disjunkte Teilmengen und T_N die Eintrittszeit in $\mathcal{X}_{N,2}$. (X_t) starte zum Zeitpunkt Null mit einer Regeneration in $\mathcal{X}_{N,1}$, der Erwartungswert $e = E\gamma_1$ der Zeit zwischen zwei Regenerationen sei endlich und die Wahrscheinlichkeit*

q_N, zwischen zwei Regenerationen in $\mathcal{X}_{N,2}$ einzutreten, strebe mit $N \to \infty$ gegen 0. Dann gilt mit $N \to \infty$

$$\frac{q_N}{e} ET_N \to 1 \quad \textit{und} \quad \frac{q_N}{e} T_N \xrightarrow{d} \mathrm{Exp}(1).$$

Eine Konsequenz der Voraussetzungen dieses Satzes ist, dass bei einer Darstellung analog zu (1.6), also

$$T_N = \sum_{k=1}^{M_N - 1} \gamma_{N,k} + \gamma_{N,M_N},$$

der Anteil der finalen Exkursion in Bezug auf die Gesamtwartezeit, also $\gamma_{N,M_N}/ET_N$, asymptotisch verschwindet, was man wie folgt sieht:

Es bezeichne $p_N := 1 - q_N$, μ sei die Verteilung der Regenerationszeit, $\mu_{N,1}$ die Verteilung der Regenerationszeit unter der Bedingung, dass der Prozess stets in $\mathcal{X}_{N,1}$ verbleibt und $\mu_{N,2}$ die Verteilung der Regenerationszeit unter der Bedingung, dass der Prozess zwischen zwei Regenerationen $\mathcal{X}_{N,2}$ besucht. Es gilt dann $\mu = p_N\mu_{N,1} + q_N\mu_{N,2}$.

Mit der Waldschen Gleichung und der Tatsache, dass M_N geometrisch verteilt ist, ergibt sich dann

$$ET_N \geq E(M_N - 1) \int t\, \mu_{N,1}(dt) = \frac{p_N}{q_N} \int t\, \mu_{N,1}(dt),$$

also

$$\frac{E\gamma_{N,M_N}}{ET_N} \leq \frac{\int t\, \mu_{N,2}(dt)}{(p_N/q_N) \int t\, \mu_{N,1}(dt)} = \frac{q_N \int t\, \mu_{N,2}(dt)}{\int t\, \mu(dt) - q_N \int t\, \mu_{N,2}(dt)}.$$

Es gilt nun aber $q_N \int t\mu_{N,2}(dt) \to 0$ (vgl. (11) in [Kei66]), also $E\gamma_{N,M_N}/ET_N \to 0$ und damit $\gamma_{N,M_N}/ET_N \xrightarrow{\mathrm{d}} 0$.

Dass andererseits diese Eigenschaft nicht hinreichend für asymptotische Exponentialität ist, haben wir bereits gesehen. Wir wollen nun kurz die Beispiele der vorangegangenen Kapitel darauf untersuchen, inwieweit sie die Voraussetzungen des Satzes von Keilson erfüllen, bevor wir eine Verallgemeinerung dieses Satzes für Familien regenerativer Prozesse herleiten.

Bei der Irrfahrt auf $\mathbb{Z}$ galt $q_N = H_{N,1}(1) = \big(1 - p/q\big)/\big(1 - (p/q)^N\big)$. Ist $p > q$, so gilt $q_N \to 0$, und da gleichzeitig der Erwartungswert der Regenerationszeit endlich ist, ergibt sich asymptotische Exponentialität. Ist $p < q$, so gilt $q_N \to 1 - p/q$,

die Voraussetzungen des Satzes sind hier also nicht erfüllt. Wie wir gesehen haben, ergibt sich in dieser Situation auch keine asymptotische Exponentialität. Einmal mehr interessant ist der Fall der symmetrischen Irrfahrt $p = q$. Hier hatten wir $H_{N,1}(1) = 1/N$ nachgewiesen, es gilt also $q_N \to 0$. Der Erwartungswert der Regenerationszeit ist hier allerdings nicht endlich, so dass die Voraussetzungen des Satzes nicht erfüllt sind. In der Folge ergibt sich auch in dieser Situation keine asymptotische Exponentialität. Insbesondere haben wir gesehen, dass der Einfluss der finalen Exkursion nicht verschwindet.

Beim *Residual Age Process* und beim *Current Age Process* galt $q_N = \sum_{k=N}^{\infty} p_k$ und der Erwartungswert der Regenerationszeit ist hier $\mu + 1$ mit $\mu = \sum_{k=1}^{\infty} k p_k$. Ist also $\mu < \infty$, so sind die Voraussetzungen des Satzes von Keilson erfüllt und es ergibt sich asymptotische Exponentialität. Ist hingegen $\mu = \infty$, so gilt zwar $q_N \to 0$, aber der Erwartungswert der Regenerationszeit ist nicht mehr endlich, die Voraussetzungen des Satzes sind verletzt. Bei der von uns speziell gewählten Verteilung $(p_k)_{k\in\mathbb{N}}$ mit nicht existierendem Erwartungswert ergibt sich für $\alpha \in (1, 2)$ keine asymptotische Exponentialität, obwohl der Einfluss der finalen Exkursion asymptotisch verschwindet. Andererseits zeigt das Beispiel $\alpha = 2$, dass sich auch asymptotische Exponentialität ergeben kann, wenn die Voraussetzungen des Satzes von Keilson nicht erfüllt sind. Der Satz von Keilson liefert also ein hinreichendes, jedoch kein notwendiges Kriterium für asymptotische Exponentialität.

Manchmal ergibt sich die Situation, dass sich mit N nicht nur die Partition des Zustandsraums $\{\mathcal{X}_{N,1}, \mathcal{X}_{N,2}\}$, sondern der gesamte regenerative Prozess $(X_t)_{t\in T}$ verändert. Ein Beispiel hierfür sind die im folgenden Abschnitt betrachteten Wartezeitverteilungen im Zusammenhang mit dem Ehrenfestschen Urnenmodell. Es ist möglich, das Resultat von Keilson auf diese Situation zu verallgemeinern.

Dazu beweisen wir zunächst den folgenden theoretischen

Hilfssatz 1.7 *Für jeden $N \in \mathbb{N}$ seien M_N, $\gamma_{N,1,k}$, $k \in \mathbb{N}$, und $\gamma_{N,2}$ unabhängige, nicht negative Zufallsvariablen. Dabei sei M_N geometrisch verteilt auf $\mathbb{N}$ mit Erfolgswahrscheinlichkeit q_N, die Folge der $(\gamma_{N,1,k})_{k\in\mathbb{N}}$ sei identisch verteilt, jeweils mit Verteilung $\mu_{N,1}$, und $\gamma_{N,2}$ besitze die Verteilung $\mu_{N,2}$.*

Es seien $p_N := 1 - q_N$, $\mu_N := p_N \mu_{N,1} + q_N \mu_{N,2}$,

$$T'_N := \sum_{k=1}^{M_N - 1} \gamma_{N,1,k}, \quad T''_N := T'_N + \gamma_{N,2}$$

und $(T_N)_{N\in\mathbb{N}}$ *eine Folge von Zufallsvariablen mit*

$$T'_N \le T_N \le T''_N \quad \textit{für alle } N \in \mathbb{N}. \tag{1.31}$$

Gilt dann

(i) $e_N := \int t\,\mu_N(dt) \in (0,\infty)$ *für alle hinreichend großen* $N \in \mathbb{N}$,

(ii) $\lim_{N\to\infty} q_N = 0$ *und*

(iii) $\lim_{c\to\infty} \sup_{N\in\mathbb{N}} \int_{t>ce_N} \frac{t}{e_N}\,\mu_N(dt) = 0$,

so folgt mit $N \to \infty$

$$\frac{q_N}{e_N}\,ET_N \to 1 \quad \textit{und} \quad \frac{q_N}{e_N}\,T_N \xrightarrow{d} \mathrm{Exp}(1).$$

Bemerkung 1.8 Die in (iii) geforderte Eigenschaft bezeichnet man auch als *gleichradige Integrierbarkeit* der auf Erwartungswert 1 normierten Verteilungsfamilie $(\mu_N)_{N\in\mathbb{N}}$.

Beweis: Es seien $e_{N,i} := \int t\,\mu_{N,i}(dt)$, $i = 1, 2$. Wir beweisen zunächst

$$\lim_{N\to\infty} \frac{p_N e_{N,1}}{e_N} = 1 \quad \text{und} \quad \lim_{N\to\infty} \frac{q_N e_{N,2}}{e_N} = 0. \tag{1.32}$$

Wegen $e_N = p_N e_{N,1} + q_N e_{N,2}$ reicht es dabei, nur einen dieser Grenzwert nachzuweisen. Für alle $c \in \mathbb{R}$ gilt

$$\begin{aligned} 0 \le \frac{q_N e_{N,2}}{e_N} &= q_N \frac{\int_0^{ce_N} t\,\mu_{N,2}(dt)}{e_N} + q_N \frac{\int_{ce_N}^{\infty} t\,\mu_{N,2}(dt)}{e_N} \\ &\le q_N \cdot c + \sup_{N\in\mathbb{N}} q_N \frac{\int_{ce_N}^{\infty} t\,\mu_{N,2}(dt)}{e_N}, \end{aligned}$$

also wegen $q_N \to 0$

$$0 \le \limsup_{N\to\infty} \frac{q_N e_{N,2}}{e_N} \le \sup_{N\in\mathbb{N}} q_N \frac{\int_{ce_N}^{\infty} t\,\mu_{N,2}(dt)}{e_N}.$$

Für die rechte Seite gilt nun aber

$$\sup_{N\in\mathbb{N}} q_N \frac{\int_{ce_N}^{\infty} t\,\mu_{N,2}(dt)}{e_N} \le \sup_{N\in\mathbb{N}} \frac{\int_{ce_N}^{\infty} t\,\mu_N(dt)}{e_N} = \sup_{N\in\mathbb{N}} \int_{t>ce_N} \frac{t}{e_N}\,\mu_N(dt).$$

Da $c \in \mathbb{R}$ beliebig war, ergibt sich (1.32) mit $c \to \infty$ aus Bedingung (iii).

Es gilt nun

$$\frac{q_N}{e_N} T'_N \leq \frac{q_N}{e_N} T_N \leq \frac{q_N}{e_N} T'_N + \frac{q_N}{e_N} \gamma_{N,2}. \tag{1.33}$$

Wegen $\gamma_{N,2} \sim \mu_{N,2}$ folgt aus (1.32) $(q_N/e_N)\gamma_{N,2} \xrightarrow{\mathrm{P}} 0$, so dass zum Nachweis der asymptotischen Exponentialität lediglich

$$\frac{q_N}{e_N} T'_N \xrightarrow{\mathrm{d}} \mathrm{Exp}(1)$$

zu zeigen bleibt.

Es seien $\phi_N(\theta)$ bzw. $\varphi_N(\theta)$ die charakteristischen Funktionen von $(q_N/e_N)\gamma_{N,1,k}$ bzw. $(q_N/e_N)T'_N$. Wegen der Unabhängigkeit der beteiligten Zufallsvariablen und der Verteilung von M_N gilt dann die Beziehung

$$\varphi_N(\theta) = \frac{q_N}{1 - p_N \phi_N(\theta)}. \tag{1.34}$$

Mit Hilfssatz A.3 folgt für alle $c > 0$:

$$\begin{aligned}
&\frac{p_N}{q_N} \left| \phi_N(\theta) - \left(1 + i\theta \frac{q_N e_{N,1}}{e_N}\right) \right| \\
&\leq \; 2\,\frac{p_N}{q_N} \int_0^\infty \min\left\{ (q_N/e_N)|\theta| t\,,\; (q_N/e_N)^2 \theta^2 t^2 \right\} \mu_{N,1}(dt) \\
&\leq \; \frac{2 p_N q_N}{e_N^2}\,\theta^2 \int_0^{ce_N} t^2\, \mu_{N,1}(dt) + 2p_N|\theta| \int_{ce_N}^\infty \frac{t}{e_N}\, \mu_{N,1}(dt) \\
&\leq \; 2 p_N q_N \theta^2 c^2 + \sup_{N \in \mathbb{N}} 2|\theta| \int_{ce_N}^\infty \frac{t}{e_N}\, \mu_N(dt),
\end{aligned}$$

also wegen $q_N \to 0$ und $p_N \to 1$

$$\limsup_{N\to\infty} \frac{p_N}{q_N} \left| \phi_N(\theta) - \left(1 + i\theta \frac{q_N e_{N,1}}{e_N}\right) \right| \leq \sup_{N \in \mathbb{N}} 2|\theta| \int_{ce_N}^\infty \frac{t}{e_N}\, \mu_N(dt).$$

Nun ist aber $c > 0$ beliebig, so dass sich aus (iii) für festes θ Null als Grenzwert der rechten Seite mit $c \to \infty$ ergibt. Verwendet man dieses Resultat in Verbindung mit (1.32) und (1.34), so ergibt sich

$$\varphi_N(\theta)^{-1} = 1 - i\theta\,\frac{p_N e_{N,1}}{e_N} - \frac{p_N}{q_N} \left(\varphi_N(\theta) - \left(1 + i\theta \frac{q_N e_{N,1}}{e_N}\right) \right) \to 1 - i\theta,$$

also asymptotische Exponentialität.

Nimmt man in (1.33) auf beiden Seiten den Erwartungswert, so liefern die Waldsche Gleichung und die Verteilungsannahme für M_N

$$\frac{p_N}{e_N}\, e_{N,1} \leq \frac{q_N}{e_N}\, ET_N \leq \frac{p_N}{e_N}\, e_{N,1} + \frac{q_N}{e_N}\, e_{N,2}.$$

Mit (1.32) erhält man so auch

$$\frac{q_N}{e_N}\, ET_N \to 1.$$

□

Dieser Hilfssatz liefert uns nun für regenerative Prozesse das folgende Resultat:

Theorem 1.9 *Für jedes $N \in \mathbb{N}$ sei X_N ein ergodischer Prozess in diskreter oder stetiger Zeit auf einem Wahrscheinlichkeitsraum $(\Omega_N, \mathcal{A}_N, P_N)$. $(\sigma_{N,l})_{l\in\mathbb{N}_0}$ mit $\sigma_{N,0} := 0$ sei eine Folge von Regenerationszeiten für X_N und $\gamma_{N,l} := \sigma_{N,l} - \sigma_{N,l-1}$ für $l \geq 1$. $(E_{N,l})_{l\in\mathbb{N}}$ sei für jedes $N \in \mathbb{N}$ eine Folge von $\{X_{\sigma_{N,l-1}+1}, ..., X_{\sigma_{N,l}}\}$-messbaren Ereignissen mit je derselben Wahrscheinlichkeit $q_N := P(E_{N,l})$ für alle $l \in \mathbb{N}$. Schließlich seien $M_N := \inf\{l \in \mathbb{N} : E_{N,l}$ tritt ein$\}$,*

$$T'_N := \sum_{k=1}^{M_N-1} \gamma_{N,l}, \quad T''_N := T'_N + \gamma_{N,M_N}$$

und $(T_N)_{N\in\mathbb{N}}$ eine Folge von Zufallsvariablen mit

$$T'_N \leq T_N \leq T''_N \quad \text{für alle } N \in \mathbb{N}.$$

Gilt dann

(i) $e_N := E\gamma_{N,1} \in (0, \infty)$ *für alle hinreichend großen* $N \in \mathbb{N}$,

(ii) $\lim_{N\to\infty} q_N = 0$ *und*

(iii) $\lim_{c\to\infty} \sup_{N\in\mathbb{N}} \int_{\gamma_{N,1}>ce_N} \frac{\gamma_{N,1}}{e_N}\, dP_N = 0$,

so folgt mit $N \to \infty$

$$\frac{q_N}{e_N}\, ET_N \to 1 \quad \text{und} \quad \frac{q_N}{e_N}\, T_N \xrightarrow{d} \text{Exp}(1).$$

Beweis: Wir beweisen diese Aussage mit dem vorangegangenen Hilfssatz. Man beachte jedoch, dass M_N und die Folge der $(\gamma_{N,l})_{l\in\mathbb{N}}$ i.a. nicht unabhängig sein werden, so dass wir Hilfssatz 1.7 nicht direkt anwenden können.

Wegen der Voraussetzungen für die Ereignisse $(E_{N,l})_{l\in\mathbb{N}}$ ist klar, dass M_N auf $\mathbb{N}$ geometrisch verteilt ist. Es gilt

$$P(M_N = m) = \prod_{k=1}^{m-1} P(E_{N,k}^{\mathrm{c}})\, P(E_{N,m}).$$

Definiert man $j_l := \sum_{k=1}^{l} i_k$, so gilt weiterhin

$$\begin{aligned} & P(\gamma_{N,k} = i_k,\ k = 1, ..., m,\ M_N = m) \\ = \ & P(\sigma_{N,k} = j_k,\ k = 1, ..., m,\ E_{N,k}^{\mathrm{c}},\ k = 0, ..., m-1,\ E_{N,m}) \\ = \ & \prod_{k=1}^{m-1} P(\gamma_{N,k} = i_k,\ E_{N,k}^{\mathrm{c}})\, P(\gamma_{N,m} = i_m,\ E_{N,m}). \end{aligned}$$

Durch Quotientenbildung ergibt sich

$$P(\gamma_{N,k} = i_k,\ k = 1, ..., m \mid M_N = m) = \prod_{k=1}^{m-1} P(\gamma_{N,k} = i_k | E_{N,k}^{\mathrm{c}})\, P(\gamma_{N,m} = i_m | E_{N,m}).$$

Durch Aufsummieren über alle übrigen Komponenten erhält man so

$$P(\gamma_{N,k} | M_N = m) = \begin{cases} P(\gamma_{N,k} = i_k | E_{N,k}^{\mathrm{c}}), & k < m, \\ P(\gamma_{N,m} = i_m | E_{N,m}), & k = m. \end{cases}$$

Dies wiederum liefert

$$P(\gamma_{N,k} = i_k,\ k = 1, ..., m \mid M_N = m) = \prod_{k=1}^{m} P(\gamma_{N,k} = i_k | M_N = m).$$

Die Regenerationszeiten $\gamma_{N,1}, ..., \gamma_{N,m}$ sind also unabhängig unter $\{M_N = m\}$. Definiert man weiter $\mu_{N,1} := \mathcal{L}(\gamma_{N,1} | E_{N,1}^{\mathrm{c}})$ und $\mu_{N,2} := \mathcal{L}(\gamma_{N,1} | E_{N,1})$, so besitzen $\gamma_{N,1}, ..., \gamma_{N,m-1}$ unter $\{M_N = m\}$ die Verteilung $\mu_{N,1}$ und $\gamma_{N,m}$ die Verteilung $\mu_{N,2}$. Die zu $\mu_{N,1}$ bzw. $\mu_{N,2}$ gehörigen charakteristischen Funktionen bezeichnen wir mit $\varphi_{N,1}(\theta)$ bzw. $\varphi_{N,2}(\theta)$.

Damit folgt nun für die charakteristische Funktion von T'_N

$$\begin{aligned} \varphi_{T'_N}(\theta) &= E\exp(i\theta T'_N) = \sum_{m=1}^{\infty} E\Big[\exp\Big(i\theta \sum_{k=1}^{m-1} \gamma_{N,k}\Big)\Big| M_N = m\Big]\, P(M_N = m) \\ &= \sum_{m=1}^{\infty} \varphi_{N,1}^{m-1}(\theta)\, P(M_N = m) = \frac{q_N}{1 - p_N \varphi_{N,1}(\theta)}. \end{aligned}$$

Ebenso zeigt man

$$\varphi_{T_N''}(\theta) = \frac{q_N}{1 - p_N \varphi_{N,1}(\theta)} \, \varphi_{N,2}(\theta).$$

Folgerung:

$$T_N' =_d \sum_{k=1}^{M_N - 1} \gamma_{N,1,k}' \quad \text{und} \quad T_N'' =_d \sum_{k=1}^{M_N - 1} \gamma_{N,1,k}' + \gamma_{N,2}'$$

mit M_N, $\gamma_{N,1,k}'$, $k \in \mathbb{N}$, und $\gamma_{N,2}'$ unabhängig, M_N geometrisch verteilt auf $\mathbb{N}$ mit Erfolgswahrscheinlichkeit q_N, $(\gamma_{N,1,k}')_{k \in \mathbb{N}}$ identisch verteilt, je mit Verteilung $\mu_{N,1}$, $\gamma_{N,2}' \sim \mu_{N,2}$ und $\gamma_{N,1} \sim p_N \mu_{N,1} + q_N \mu_{N,2}$.

Mit Hilfssatz 1.7 ergibt sich aus den Voraussetzungen des Theorems die asymptotische Exponentialität für $(q_N/e_N)T_N'$ und $(q_N/e_N)T_N''$, was diese unmittelbar auch für $(q_N/e_N)T_N$ liefert. □

Aus diesem Theorem erhalten wir nun insbesondere eine Verallgemeinerung des Satzes von Keilson für eine Folge regenerativer Prozesse:

Lemma 1.10 *Für jedes $N \in \mathbb{N}$ sei X_N ein ergodischer regenerativer Prozess in diskreter oder stetiger Zeit auf einem Wahrscheinlichkeitsraum $(\Omega_N, \mathcal{A}_N, P_N)$ mit Zustandsraum $\mathcal{X}_N$. Für jeden dieser Prozesse sei $D_N := \{\mathcal{X}_{N,1}, \mathcal{X}_{N,2}\}$ eine Partition des Zustandsraums in zwei disjunkte Teilmengen. Dabei möge X_N zum Zeitpunkt 0 mit einer Regeneration in $\mathcal{X}_{N,1}$ starten. $(\sigma_{N,l})_{l \in \mathbb{N}_0}$ mit $\sigma_{N,0} := 0$ sei eine Folge von Regenerationszeiten für X_N und $\gamma_{N,l} := \sigma_{N,l} - \sigma_{N,l-1}$ für $l \geq 1$. Schließlich sei*

$$T_N := \inf\big\{n \in \mathbb{N} : X_{N,n} \in \mathcal{X}_{N,2}\big\}.$$

Gilt dann

(i) $e_N := E\gamma_{N,1} \in (0, \infty)$ *für alle hinreichend großen* $N \in \mathbb{N}$,

(ii) $q_N := P\big(Es\ gibt\ ein\ n \in \{1, ..., \sigma_{N,1}\}\ mit\ X_{N,n} \in \mathcal{X}_{N,2}\big) \to 0$ *mit* $N \to \infty$ *und*

(iii) $\lim_{c \to \infty} \sup_{N \in \mathbb{N}} \int_{\gamma_{N,1} > c e_N} \frac{\gamma_{N,1}}{e_N} \, dP_N = 0$,

so folgt mit $N \to \infty$

$$\frac{q_N}{e_N} ET_N \to 1 \quad und \quad \frac{q_N}{e_N} T_N \xrightarrow{d} \mathrm{Exp}(1).$$

Bemerkung 1.11 Es zeigt sich also, dass sich der Satz von Keilson auf eine Familie regenerativer Prozesse verallgemeinern lässt, wenn man zusätzlich die gleichgradige

Integrierbarkeit der auf Erwartungswert 1 normierten Familie von Regenerationszeiten voraussetzt ((iii) in Lemma 1.10).

Mit dieser verallgemeinerten Version des Satzes von Keilson werden wir nun im folgenden Kapitel das sog. Ehrenfestsche Urnenmodell auf asymptotische Exponentialität untersuchen. Die vorangegangenen theoretischen Untersuchungen, insbesondere Theorem 1.9, werden sich dann später als nützlich erweisen, wenn es in Abschnitt 2.2 darum geht, in Spezialfällen asymptotische Exponentialität für das *Approximate Pattern Matching Problem* nachzuweisen.

1.6 Ehrenfestsches Urnenmodell I

Wir werden nun diese Erweiterung des Satzes von Keilson benutzen, um bei einer weiteren Klasse von Markov-Ketten die asymptotische Exponentialität von Eintrittszeitverteilungen zu untersuchen. Hierbei handelt es sich um das sog. *Ehrenfestsche Urnenmodell* (im Folgenden kurz *EF-Modell*): Gegeben seien zwei Urnen, jede dieser beiden enthalte zu Beginn N Kugeln. Zum Zeitpunkt $n \in \mathbb{N}$ wählt man eine dieser $2N$ Kugeln gleichverteilt aus und legt sie in die jeweils andere Urne. Dieses Modell bildet den Ausgangspunkt für gleich zwei verschiedene Markov-Ketten.

Eine Betrachtungsweise ist die Analyse der Anzahl der Kugeln in einer der beiden Urnen. Ist man lediglich hieran interessiert, so hat man es mit einer Markov-Kette auf dem Zustandsraum $\{0, 1, ..., 2N\}$ zu tun. Diese Markov-Kette soll der Gegenstand unserer Betrachtungen in diesem Abschnitt sein.

Darüber hinaus gibt es jedoch auch die Möglichkeit, das EF-Modell detaillierter zu betrachten, und für jede einzelne Kugel festzuhalten, in welcher der beiden Urnen sie sich befindet. Auch in diesem Fall ergibt sich eine Markov-Kette als beschreibender Prozess, jetzt jedoch auf dem Zustandsraum $\{0, 1\}^{2N}$. Bei dieser Markov-Kette spricht man auch anstelle eines EF-Modells von einer *Irrfahrt auf dem Hyperwürfel* $\{0, 1\}^{2N}$. Die Analyse solcher Irrfahrten ist Gegenstand von Abschnitt 1.7. Man beachte insbesondere, dass sich eine Irrfahrt auf dem Hyperwürfel unter gewissen Regularitätsbedingungen durch Zusammenfassung von Zuständen in ein EF-Modell auf den natürlichen Zahlen überführen lässt. (Auch hierzu mehr im folgenden Abschnitt.)

Wir betrachten nun also zunächst das EF-Modell auf den natürlichen Zahlen $\{0, 1, ..., 2N\}$. Es sei $X_{N,n}$ die Anzahl der Kugeln in der ersten der beiden Urnen unmittelbar nach der Ziehung zur Zeit $n \in \mathbb{N}$. $(X_{N,n})_{n\in\mathbb{N}_0}$ ist dann eine Markov-Kette mit der Übergangsmatrix $P_N = (p_{ij}^{(N)})_{i,j=0,...,2N}$, wobei $p_{i,i+1}^{(N)} = (2N - i)/2N$ und $p_{i,i-1}^{(N)} = i/2N$ für alle $i = 0, 1, ..., 2N$ gilt.

Der Zustand N stellt für diese Markov-Kette eine Art zentralen Zustand dar, hier gilt $p_{N,N-1}^{(N)} = p_{N,N+1}^{(N)} = 1/2$. Sobald man sich von diesem Zustand entfernt, zeigt die Markov-Kette eine Tendenz, wieder dorthin zurückzukehren, d.h. die Wahrscheinlichkeit, sich auf den Zustand N zuzubewegen, ist größer als die, sich von ihm zu entfernen. Insofern erkennt man eine Parallele zur Irrfahrt auf $\mathbb{Z}$ mit Drift nach 0.

Im Unterschied zu der damaligen Situation hängt die Stärke des Drifts nach N hier jedoch davon ab, wie weit man sich vom zentralen Zustand entfernt: Je weiter man von N entfernt ist, desto stärker ist der Drift nach N zurück.

Wir werden nun die Wartezeit untersuchen, bis dieser Prozess erstmals eine vorgegebene Abstandsoberschranke vom zentralen Zustand N überschreitet. Man beachte, dass die Übergangswahrscheinlichkeiten des Prozesses selber von N abhängen. Wir haben es also mit einer ganzen Familie von Prozessen zu tun, auf die wir nicht den Satz von Keilson, wohl aber die im vorangegangenen Abschnitt entwickelte Verallgemeinerung anwenden können und werden. Ähnlich wie bei der Irrfahrt auf $\mathbb{Z}$ gehen wir zu einem anderen Prozess über, der zur Analyse dieser Wartezeit besser geeignet ist.

Sei $Y_{N,n} := |X_{N,n} - N|$, $Y_{N,0} := 0$. Auch dies ist eine Markov-Kette, jetzt auf dem Zustandsraum $\{0, ..., N\}$, die Übergangsmatrix lautet $Q_N = (q_{ij}^{(N)})_{i,j=0,...,N}$ mit $q_{01}^{(N)} = 1$, $q_{i,i+1}^{(N)} = (N-i)/2N$ für $i = 1, ..., N-1$ und $q_{i,i-1}^{(N)} = (N+i)/2N$ für $i = 1, ..., N$.

Sei $b_N \in \{1, ..., N\}$ fest vorgegeben. Wir werden nun das asymptotische Verhalten der Eintrittszeit T_N in die Menge $\{b_N, ..., N\}$ untersuchen:

$$T_N := \inf\{n \geq 0 \,:\, Y_{N,n} \geq b_N\} = \inf\{n \geq 0 \,:\, Y_{N,n} = b_N\}.$$

Dabei werden wir die Resultate des vorangegangenen Abschnitts verwenden, um unter bestimmten Voraussetzungen an die Schranke b_N auch hier wieder asymptotische Exponentialität für die normierte Wartezeit T_N/ET_N nachzuweisen. Konkret werden wir das folgende Theorem beweisen:

Theorem 1.12 *Es gelte* $\lim_{N\to\infty} b_N^2/N = \infty$. *Dann folgt:*

$$T_N/ET_N \xrightarrow{d} \mathrm{Exp}(1).$$

Um in dieser Situation die bereits erarbeiteten Resultate verwenden zu können, braucht man eine geeignete Folge von Regenerationszeiten für das EF-Modell. In diesem Fall liefert uns eine auf den ersten Blick etwas ungewöhnliche Idee eine geeignete Folge: Für jedes $N \in \mathbb{N}$ sei τ_N die Rückkehrzeit des Ehrenfestmodells $(Y_{N,n})$ nach 0 *nach dem ersten Besuch in* $\lceil\sqrt{N}\rceil$. Sei weiter $e_N := E\tau_N$ und q_N die Wahrscheinlichkeit, dass der Prozess zwischen zwei solchen Regenerationen das Niveau b_N erreicht.

Theorem 1.12 folgt nun aus Lemma 1.10, sofern wir $q_N \to 0$ und die Bedingung der gleichgradigen Integrierbarkeit (iii) für τ_N nachweisen können. Wir kümmern uns zunächst um die Bestimmung der Wahrscheinlichkeit q_N und des Erwartungswerts e_N.

Eine wohl bekannte und leicht zu überprüfende Eigenschaft des Ehrenfestmodells Y_N ist dessen Reversibilität, d.h. es gilt $\pi_i^{(N)} q_{ij}^{(N)} = \pi_j^{(N)} q_{ji}^{(N)}$ für alle $i, j = 0, ..., N$, wobei $\pi_N := (\pi_i^{(N)})_{i=0,...,N}$ mit $\pi_i^{(N)} = 2^{-2N}\binom{2N}{N+i}(1 + \mathbb{1}\{i \neq 0\})$, $i = 0, ..., N$, die stationäre Verteilung dieser Markov-Kette ist (vgl. [Bré99], S. 76f.). (Die Verteilung π_N ergibt sich, wenn man die $2N$ Kugeln zufällig auf die beiden Urnen verteilt.)

An dieser Stelle können wir nun zwei Resultate aus [AF04] verwenden. Es bezeichne in Übereinstimmung mit den in der Einleitung getroffenen Vereinbarungen $T_{N,a} := \inf\{n \geq 0 : Y_{N,n} = a\}$ bzw. $T_{N,a}^+ := \inf\{n \geq 1 : Y_{N,n} = a\}$ die Eintrittszeit in a bzw. die Rückkehrzeit nach a.

Lemma 1.13 *Sei $1 \leq a \leq N$ und $\tau_{N,a}$ die Rückkehrzeit von $Y_{N,n}$ nach 0 nach dem ersten Besuch in a. Dann gilt*

$$E_0\tau_{N,a} = E_0T_{N,a} + E_aT_{N,0} = \frac{1}{\pi_0^{(N)} P_0(T_{N,a} < T_{N,0}^+)}.$$

(vgl. [AF04], Kapitel 2, Korollar 8 bzw. [Chu67], Abschnitt I.11, Korollar 1, S. 65.)

Lemma 1.14 *Sei $0 < i < a \leq N$. Dann gilt*

$$P_i(T_{N,a} < T_{N,0}) = \frac{\sum_{k=0}^{i-1}(\pi_k^{(N)} p_{k,k+1}^{(N)})^{-1}}{\sum_{k=0}^{a-1}(\pi_k^{(N)} p_{k,k+1}^{(N)})^{-1}}.$$

(vgl. [AF04], Kapitel 5, Proposition 3.(a).)

Verwendet man diese Resultate speziell für $a = a_N := \lceil\sqrt{N}\rceil$, so ergibt sich (für hinreichend großes N) mit Hilfe der trivialen Identitäten $P_0(T_{N,a_N} < T_{N,0}^+) = P_1(T_{N,a_N} < T_{N,0})$ und $q_N = P_{\lceil\sqrt{n}\rceil}(T_{N,b_N} < T_{N,0})$

$$e_N = \sum_{k=0}^{a_N-1} (\pi_k^{(N)} q_{k,k+1}^{(N)})^{-1} \tag{1.35}$$

und

$$q_N = \frac{\sum_{k=0}^{a_N-1}(\pi_k^{(N)} q_{k,k+1}^{(N)})^{-1}}{\sum_{k=0}^{b_N-1}(\pi_k^{(N)} q_{k,k+1}^{(N)})^{-1}}. \tag{1.36}$$

Entscheidend ist nun das folgende

Lemma 1.15 *Sei $C > 0$ beliebig und $c_N := \lceil C\sqrt{N} \rceil$. Dann gilt*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{c_N-1} (\pi_k^{(N)} q_{k,k+1}^{(N)})^{-1} = \sqrt{\pi} \int_0^C \exp(x^2)\, dx.$$

Beweis: Es gilt $\pi_0^{(N)} q_{01}^{(N)} = \binom{2N}{N} 2^{-2N}$ und $\pi_k^{(N)} q_{k,k+1}^{(N)} = \binom{2N}{N+k} 2^{-2N}(1 - k/N)$ für $k \geq 1$. Sei $C > 0$ beliebig. Die lokale Form der Normalapproximation für die Binomialverteilung (vgl. [Mor68], §7.1, S. 52) liefert:

$$\lim_{N\to\infty} \sup_{1\leq k\leq C\sqrt{N}} \left| \frac{\binom{2N}{N+k} 2^{-2n}}{\frac{1}{\sqrt{\pi N}} \exp\left(-k^2/N\right)} - 1 \right| = 0.$$

Zusammen mit $\lim_{N\to\infty} \sup_{1\leq k\leq C\sqrt{N}} (1 - k/N) = 1$ und der elementaren Tatsache, dass für beliebige $a_{N,k}, b_{N,k} \in \mathbb{R}$ mit $\sup_{1\leq k\leq c_N} |a_{N,k}/b_{N,k} - 1| \to 0$ auch $\sup_{1\leq k\leq c_N} |b_{N,k}/a_{N,k} - 1| \to 0$ gilt, ergibt sich

$$\lim_{N\to\infty} \sup_{1\leq k\leq C\sqrt{N}} \left| \frac{(\pi_k^{(N)} q_{k,k+1}^{(N)})^{-1}}{\sqrt{\pi N} \exp(k^2/N)} - 1 \right| = 0.$$

Außerdem folgt mit der Stirlingschen Formel $(\pi_0^{(N)} q_{01}^{(N)})^{-1} \sim \sqrt{\pi N}$. Damit gilt für alle $C > 0$

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{c_N-1} (\pi_k^{(N)} q_{k,k+1}^{(N)})^{-1} &= \lim_{N\to\infty} \sqrt{\pi} \sum_{1/\sqrt{N}\leq k/\sqrt{N}\leq C} \frac{1}{\sqrt{N}} \exp\left((k/\sqrt{N})^2\right) \\ &= \sqrt{\pi} \int_0^C \exp(x^2)\, dx. \end{aligned}$$

□

Mit diesem Lemma erhält man nun vermöge der Darstellung (1.36) die gewünschte Aussage:

Korollar 1.16

$$\lim_{N\to\infty} q_N = 0.$$

Beweis: Erweitert man in der Darstellung (1.36) für q_N Zähler und Nenner mit $1/N$, so ergibt sich für den Zähler aus dem vorangegangenen Lemma

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{a_N-1} (\pi_k^{(N)} q_{k,k+1}^{(N)})^{-1} = \sqrt{\pi} \int_0^1 \exp(x^2)\, dx.$$

Im Nenner verwenden wir die Darstellung $b_N = \lceil C_N\sqrt{N}\rceil$ mit $C_N \to \infty$. Dann folgt hier für alle $C > 0$:

$$\begin{aligned}\liminf_{N\to\infty} \frac{1}{N} \sum_{k=0}^{b_N-1} (\pi_k^{(N)} q_{k,k+1}^{(N)})^{-1} &\geq \liminf_{N\to\infty} \frac{1}{N} \sum_{k=0}^{\lceil C\sqrt{N}\rceil-1} (\pi_k^{(N)} q_{k,k+1}^{(N)})^{-1} \\ &= \sqrt{\pi} \int_0^C \exp(x^2)\, dx.\end{aligned}$$

Mit $C \to \infty$ ergibt sich $\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{b_N-1} (\pi_k^{(N)} q_{k,k+1}^{(N)})^{-1} = \infty$ und damit die Behauptung. □

Eine weitere Folgerung aus Lemma 1.15 ist das asymptotische Verhalten von e_N:

Korollar 1.17

$$e_N \sim N \cdot \sqrt{\pi} \int_0^1 \exp(x^2)\, dx.$$

Beweis: Ergibt sich unmittelbar aus der Darstellung (1.35) mit Lemma 1.15. □

Der Nachweis der Bedingung $q_N \to 0$ ist also gelungen. Verbleibt die Aufgabe, die gleichgradige Integrierbarkeit der normierten Regenerationszeiten $\tau_N / E\tau_N$ nachzuweisen. Hierfür verwenden wir das folgende

Lemma 1.18 *Es seien $Z, Z_1, Z_2, ...$ nicht negative Zufallsvariablen mit $\sup_{N\in\mathbb{N}} EZ_N < \infty$. Weiter gelte $EZ_N \to EZ$ und $Z_N \xrightarrow{d} Z$. Dann ist die Familie der Z_N gleichgradig integrierbar.*

Beweis: Aus $Z_N \xrightarrow{\mathrm{d}} Z$ folgt zunächst, dass auch $EZ < \infty$ gilt (vgl. [Bil79], S. 291, Theorem 25.11). Die gleichgradige Integrierbarkeit der Familie der Z_N ergibt sich dann mit [Bil68], Theorem 5.4, S. 32. □

Dieses Resultat werden wir nun auf $\tau_N / E\tau_N$ anwenden. Die L_1-Beschränktheit erledigt sich hier wegen der Normierung auf Erwartungswert 1 von selbst. Zum Nachweis der Konvergenz in Verteilung und der Konvergenz der Momente betrachten wir nun zunächst das folgende zeitstetige EF-Modell:

Es sei $(Z_{N,t})_{t\geq 0}$ der Markov-Prozess mit Generator $G_N = (g_{ij}^{(N)})_{i,j=-N,...,N}$, wobei $g_{i,i+1}^{(N)} = (N-i)/2$, $g_{i,i-1}^{(N)} = (N+i)/2$ und $g_{i,i}^{(N)} = -N$, jeweils für $-N \leq i \leq N$, und $(W_{N,t})_{t\geq 0}$ gegeben durch $W_{N,t} := (1/\sqrt{N}) Z_{N,t}$. Dann ist $(W_{N,t})_{t\geq 0}$ ebenfalls ein Markov-Prozess mit Zustandsraum $S_N := \{x_{N,i} = i/\sqrt{N} \ : \ i = -N, ..., N\}$.

Sei $\mathcal{X} := \mathcal{C}_0(\mathbb{R}) := \{f : f \text{ ist stetig und beschränkt auf } \mathbb{R}, \lim_{|x|\to\infty} f(x) = 0\}$ und $\mathcal{X}_N := \{g : S_N \to \mathbb{R}\}$, jeweils versehen mit der Norm $|f| := \sup_{x\in\mathbb{R}} |f(x)|$ bzw. $|g| := \sup_{-N\le i\le N} |f(x_{N,i})|$. Wir definieren eine Abbildung $\Pi_N : \mathcal{X} \to \mathcal{X}_N$ durch $\Pi_N f(x_{N,i}) := f(x_{N,i})$ für alle $x_{N,i} \in S_N$. Π_N ist ein durch 1 beschränkter linearer Operator und es gilt $\lim_{N\to\infty} |\Pi_N f| = |f|$. Im Sinne von [RD80], Definition 2.5, S. 711, wird also der Banachraum $\mathcal{X}$ durch die Folge der $(\mathcal{X}_N, \Pi_N)$ approximiert.

Unser Ziel ist es nun nachzuweisen, dass die Folge der Prozesse $(W_{N,t})_{t\ge0}$ in Verteilung gegen einen Ornstein-Uhlenbeck-Prozess (kurz: OU-Prozess) $(U_t)_{t\ge0}$ mit Start in 0, infinitesimaler Varianz 1 und Drift $-x$ konvergiert, d.h. (U_t) hat den infinitesimalen Generator

$$Af(x) := \frac{1}{2}f''(x) - xf'(x).$$

Diese Aussage ergibt sich mit [RD80], Theorem 2.6. Dieses besagt in Verbindung mit [Kur75], S. 630, Theorem 4.29, dass es zum Nachweis dieser Aussage ausreicht, für alle Funktionen f aus dem Kern $\mathcal{C}_0^3(\mathbb{R}) := \{f : f, f', f'', f''' \in \mathcal{C}_0(\mathbb{R})\}$ des infinitesimalen Generators A die folgende Bedingung nachzuweisen:

$$\lim_{N\to\infty} |\Pi_N Af - A_N f_N| = 0, \tag{1.37}$$

wobei A_N den infinitesimalen Generator von $(W_{N,t})_{t\ge0}$ und $f_N := \Pi_N f$ bezeichnet. (Vgl. insb. auch [RD80], Abschnitt 3, S. 712-714.) Es gilt offenbar

$$A_N f_N(x) = \frac{1}{2}(N-\sqrt{N}x)f(x+1/\sqrt{N}) - Nf(x) + \frac{1}{2}(N+\sqrt{N}x)f(x-1/\sqrt{N}), \quad x \in S_N.$$

Für $f \in \mathcal{C}_0^3(\mathbb{R})$ haben wir die Taylorentwicklung

$$f(x \pm 1/\sqrt{N}) = f(x) \pm (1/\sqrt{N})f'(x) + (1/2N)f''(x) + \mathcal{O}(N^{-3/2}) \text{ für } x \in \mathbb{R},$$

denn $|f'''|$ ist endlich. Damit ergibt sich wegen $|x| < \sqrt{N}$ für $x \in S_N$

$$A_N f_N(x) = Af(x) + \mathcal{O}(N^{-1/2}), \quad x \in S_N,$$

also (1.37).

Nach dieser Zwischenbetrachtung kommen wir nun zurück zu unserem eigentlichen Problem.

Da $E\tau_N \sim K \cdot N$ mit $K > 0$ gilt, reicht es offenbar, die verbleibenden beiden Bedingungen aus Lemma 1.18 für τ_N/N anstelle von $\tau_N/E\tau_N$ nachzuweisen. Wir werden

zeigen, dass τ_N/N sowohl in L_1 als auch in Verteilung gegen die entsprechende Regenerationszeit des OU-Prozesses $(U_t)_{t\geq 0}$ konvergiert.

τ_N war die Rückkehrzeit von $(Y_{N,n})_{n\in\mathbb{N}_0}$ nach 0 nach dem ersten Besuch in $\lceil\sqrt{N}\rceil$. Wir führen nun noch die folgenden Bezeichnungen ein: $\tilde{\tau}_N$ sei die Rückkehrzeit von $(|W_{N,t}|)_{t\geq 0}$ nach 0 nach dem ersten Überschreiten des Niveaus 1 und $\bar{\tau}$ die Rückkehrzeit von $(|U_t|)_{t\geq 0}$ nach 0 nach dem ersten Besuch in 1. Außerdem sei

$$\begin{aligned}
\tau_{N,a} &:= \inf\{n \in \mathbb{N}_0 : Y_{N,n} = a\} \quad , a \in \mathbb{N}_0, \\
\tilde{\tau}_{N,a} &:= \inf\{t \geq 0 : |W_{N,t}| = a\} \quad , a\sqrt{N} \in \mathbb{N}_0, \\
\bar{\tau}_a &:= \inf\{t \geq 0 : |U_t| = a\} \qquad , a \geq 0.
\end{aligned}$$

Damit sind alle notwendigen Bezeichnungen getroffen, um nun zu zeigen, dass τ_N/N in L_1 und in Verteilung gegen $\bar{\tau}$ konvergiert.

Lemma 1.19

$$\lim_{N\to\infty} E_0(\tau_N/N) = E_0\bar{\tau}.$$

Beweis: Aus Korollar 1.17 ist bekannt, dass $\lim_{N\to\infty} E_0(\tau_N/N) = \sqrt{\pi}\int_0^1 \exp(x^2)dx$ gilt. Es bleibt nachzuweisen, dass dies dem Erwartungswert der Regenerationszeit $\bar{\tau}$ entspricht.

Wegen der Symmetrie des OU-Prozesses $(U_t)_{t\geq 0}$ bezüglich des Nullpunkts ist klar, dass $E_0\bar{\tau} = E_0\min\{\bar{\tau}_{-1}, \bar{\tau}_1\} + E_{-1}\bar{\tau}_0$ gilt.

Es lässt sich nun leicht zeigen, dass $E_0\min\{\bar{\tau}_{-1}, \bar{\tau}_1\} = (1/2)(E_0\bar{\tau}_1 - E_{-1}\bar{\tau}_0)$ gilt. Zum Nachweis verwendet man beispielsweise die in [BS96], II.7, Formel 2.0.1 (S. 429) und Formel 3.0.1 (S. 434) gegebenen expliziten Darstellungen der Laplace-Transformierten der beteiligten Wartezeiten. Hieraus ergibt sich durch eine einfache Rechnung die Relation

$$E_0\exp\big(-\theta\min\{\bar{\tau}_{-1}, \bar{\tau}_1\}\big) = 2\Big(\big(E_0\exp(-\theta\bar{\tau}_1)\big)^{-1} + E_1\exp(-\theta\bar{\tau}_0)\Big)^{-1}.$$

Differentiation in $\theta = 0$ liefert das gewünschte Resultat.

Es gilt also $E_0\bar{\tau} = (1/2)(E_0\bar{\tau}_1 + E_{-1}\bar{\tau}_0)$. Hier verwendet man nun die in [NRS85], S. 368, gegebene Darstellung für Erwartungswerte $E_a\bar{\tau}_b$ mit $a < b$,

$$E_a\bar{\tau}_b = \Psi(b) - \Psi(a) \text{ mit } \Psi(x) = \sqrt{\pi}\int_0^x \exp(y^2)dy + 2\int_0^x \exp(y^2)\int_0^y \exp(-z^2)dzdy,$$

und erhält so die gewünschte Aussage. □

Lemma 1.20

$$\tau_N/N \xrightarrow{d} \bar{\tau}.$$

Beweis: Offenbar lässt sich bei allen drei Prozessen, $(Y_{N,n})_{n\in\mathbb{N}_0}$, $(|W_{N,t}|)_{t\geq 0}$ und $(|U_t|)_{t\geq 0}$ wegen der Markov-Eigenschaft eine Zerlegung der entsprechenden Wartezeit als Faltung zweier einfacherer Wartezeiten vornehmen:

$$\begin{aligned}
\mathcal{L}_0(\tau_N) &= \mathcal{L}_0(\tau_{N,\lceil\sqrt{N}\rceil}) &&* \mathcal{L}_{\lceil\sqrt{N}\rceil}(\tau_{N,0}),\\
\mathcal{L}_0(\tilde{\tau}_N) &= \mathcal{L}_0(\tilde{\tau}_{N,\lceil\sqrt{N}\rceil/\sqrt{N}}) &&* \mathcal{L}_{\lceil\sqrt{N}\rceil/\sqrt{N}}(\tilde{\tau}_{N,0}),\\
\mathcal{L}_0(\bar{\tau}) &= \mathcal{L}_0(\bar{\tau}_1) &&* \mathcal{L}_1(\bar{\tau}_0).
\end{aligned}$$

Außerdem haben wir bereits die Verteilungskonvergenz des Prozesses $(W_{N,t})_{t\geq 0}$ gegen den OU-Prozess $(U_t)_{t\geq 0}$ nachgewiesen. Hieraus folgt unmittelbar, dass $\mathcal{L}_0(\tilde{\tau}_{N,\lceil\sqrt{N}\rceil/\sqrt{N}})$ gegen $\mathcal{L}_0(\bar{\tau}_1)$ und $\mathcal{L}_{\lceil\sqrt{N}\rceil/\sqrt{N}}(\tilde{\tau}_{N,0})$ gegen $\mathcal{L}_1(\bar{\tau}_0)$ konvergiert. Es ist nämlich beispielsweise

$$P_0(\tilde{\tau}_{N,\lceil\sqrt{N}\rceil/\sqrt{N}} > t) = P_0(\sup_{0\leq s\leq t} |W_{N,s}| < 1)$$

sowie

$$P_0(\bar{\tau}_1 > t) = P_0(\sup_{0\leq s\leq t} |U_s| < 1).$$

Hiermit ergibt sich nun die Verteilungskonvergenz aus der Verteilungskonvergenz der beteiligten Prozesse mit dem Continuous Mapping Theorem (vgl. z.B. [Bil68], S. 30, Theorem 5.1 und auch Kapitel 11), wobei wir verwenden, dass $\mathcal{L}_0(\sup_{0\leq s\leq t}|U_s|)$ für alle $t \geq 0$ eine stetige Verteilungsfunktion besitzt. (Folgt beispielsweise unter Verwendung von [PT79], Theorem 3.1, S. 483.)

Wir zeigen nun, dass auch $\mathcal{L}_0(\tau_{N,\lceil\sqrt{N}\rceil}/N)$ gegen $\mathcal{L}_0(\tau_1)$ konvergiert. Der Prozess $(|W_{N,t}|)_{t\geq 0}$ entsteht aus $(Y_{N,n})_{n\in\mathbb{N}_0}$ durch „Continuization“ (d.h. durch den Übergang von $Y_{N,n}$ zu Y_{N,P_t} mit von Y_N unabhängigem und mit Parameter 1 Poisson-verteiltem P_t), eine zeitliche Umskalierung $t \mapsto Nt$ und eine räumliche Umskalierung $x \mapsto x/\sqrt{N}$. Es gilt also insbesondere: Ist $(P_t)_{t\geq 0}$ ein von $(Y_{N,n})_{n\in\mathbb{N}_0}$ unabhängiger Poisson-Prozess mit Intensität N, so gilt

$$(|W_{N,t}|)_{t\geq 0} =_d \left(\frac{1}{\sqrt{N}} Y_{N,P_t}\right)_{t\geq 0}.$$

Damit folgt

$$\begin{aligned} P_0(\tilde{\tau}_{N,\lceil\sqrt{N}\rceil/\sqrt{N}} > t) &= P_0\Big(\sup_{0\le s\le t}|W_{N,s}| < \lceil\sqrt{N}\rceil/\sqrt{N}\Big) \\ &= P_0\Big(\sup_{0\le s\le t} Y_{N,P_s} < \lceil\sqrt{N}\rceil\Big) \\ &= \sum_{l=0}^{\infty} P_0\Big(\sup_{0\le k\le l} Y_{N,k} < \lceil\sqrt{N}\rceil\Big) P(P_t = l). \end{aligned}$$

Sei nun $\varepsilon > 0$ beliebig. Dann gilt

$$\begin{aligned} P_0(\tilde{\tau}_{N,\lceil\sqrt{N}\rceil/\sqrt{N}} > t) = &\sum_{|l-Nt|\le\varepsilon Nt} P_0\Big(\sup_{0\le k\le l} Y_{N,k} < \lceil\sqrt{N}\rceil\Big) P(P_t = l) + \\ &\sum_{|l-Nt|>\varepsilon Nt} P_0\Big(\sup_{0\le k\le l} Y_{N,k} < \lceil\sqrt{N}\rceil\Big) P(P_t = l). \end{aligned}$$

Da $P_0\big(\sup_{0\le k\le l} Y_{N,k} < \lceil\sqrt{N}\rceil\big)$ fallend in l ist, ergibt sich

$$P_0(\tilde{\tau}_{N,\lceil\sqrt{N}\rceil/\sqrt{N}} > t) \le P_0\big(\tau_{N,\lceil\sqrt{N}\rceil} > (1-\varepsilon)Nt\big)\cdot P(|P_t - Nt| \le \varepsilon Nt) + P(|P_t - Nt| > \varepsilon Nt)$$

sowie

$$P_0(\tilde{\tau}_{N,\lceil\sqrt{N}\rceil/\sqrt{N}} > t) \ge P_0\big(\tau_{N,\lceil\sqrt{N}\rceil} > (1+\varepsilon)Nt\big)\cdot P(|P_t - Nt| \le \varepsilon Nt).$$

Die Chebyshevsche Ungleichung liefert uns $\lim_{N\to\infty} P(|P_t - Nt| \le \varepsilon Nt) = 1$, also erhält man mit $N \to \infty$

$$\limsup_{N\to\infty} P_0\big(\tau_{N,\lceil\sqrt{N}\rceil}/N \le (1-\varepsilon)t\big) \le P_0(\bar{\tau}_1 \le t) \le \liminf_{N\to\infty} P_0\big(\tau_{N,\lceil\sqrt{N}\rceil}/N \le (1+\varepsilon)t\big)$$

für alle $\varepsilon > 0$ und alle $t \ge 0$, wobei man erneut die Stetigkeit der Verteilungsfunktion zu $\mathcal{L}_0(\bar{\tau}_1)$ ausgenutzt hat.

Setzt man nun $t := s/(1-\varepsilon)$ bzw. $t := s/(1+\varepsilon)$, so folgt

$$\limsup_{N\to\infty} P_0\Big(\tau_{N,\lceil\sqrt{N}\rceil}/N \le s\Big) \le P_0\Big(\bar{\tau}_1 \le \frac{s}{1-\varepsilon}\Big) \text{ bzw.}$$

$$\liminf_{N\to\infty} P_0\Big(\tau_{N,\lceil\sqrt{N}\rceil}/N \le s\Big) \ge P_0\Big(\bar{\tau}_1 \le \frac{s}{1+\varepsilon}\Big).$$

Mit $\varepsilon \downarrow 0$ folgt daraus für alle $s > 0$

$$\lim_{N\to\infty} P_0\Big(\tau_{N,\lceil\sqrt{N}\rceil}/N \le s\Big) = P_0(\bar{\tau}_1 \le s),$$

wobei wir nochmals die Stetigkeit der Verteilungsfunktion zu $\mathcal{L}_0(\bar{\tau}_1)$ verwenden.

Analog zeigt man

$$\lim_{N\to\infty} P_{\lceil\sqrt{N}\rceil}\Big(\tau_{N,0}/N \le s\Big) = P_1(\bar{\tau}_0 \le s)$$

für alle $s > 0$, also die Konvergenz von $\mathcal{L}_{\lceil\sqrt{N}\rceil}(\tau_{N,0}/N)$ gegen $\mathcal{L}_1(\bar{\tau}_0)$.

Daraus ergibt sich insgesamt die behauptete Verteilungskonvergenz. □

Somit sind also alle Voraussetzungen von Lemma 1.10 erfüllt und Theorem 1.12 ist nun eine einfache Folgerung.

Das EF-Modell liefert auch ein Beispiel, das zeigt, dass die Bedingung der gleichgradigen Integrierbarkeit (iii) in Lemma 1.10 nicht ersatzlos gestrichen werden kann.

Man betrachte hierzu die Eintrittszeit

$$T_{N,\lceil\sqrt{N}\rceil} := \inf\{n \in \mathbb{N}_0 \,:\, Y_{N,n} = \lceil\sqrt{N}\rceil\}.$$

Wir haben bereits gezeigt, dass

$$ET_{N,\lceil\sqrt{N}\rceil} \sim N\sqrt{\pi}\int_0^1 \exp(x^2)\,dx$$

und

$$\mathcal{L}_0(T_{N,\lceil\sqrt{N}\rceil}/N) \xrightarrow{\mathrm{d}} \mathcal{L}_0(\bar{\tau}_1)$$

gilt. Hierzu ist in [BS96], II.7, Formel 3.0.1 (S. 434) eine explizite Darstellung der Laplace-Transformierten gegeben, die insbesondere zeigt, dass $\mathcal{L}_0(\bar{\tau}_1)$ *keine* Exponentialverteilung ist.

Als Erneuerungsfolge betrachten wir die Rückkehrzeiten des EF-Modells nach 0, also

$$\tau_N := T_0^+ = \min\{n \in \mathbb{N} \,:\, Y_{N,n} = 0\}.$$

Dann gilt

$$q_N = P_1(\tau_{N,\lceil\sqrt{N}\rceil} < \tau_{N,0}) = \frac{\big(\pi_0^{(N)} q_{01}^{(N)}\big)^{-1}}{\sum_{k=0}^{\lceil\sqrt{N}\rceil-1}(\pi_k^{(N)} q_{k,k+1}^{(N)})^{-1}} \sim \frac{\sqrt{N\pi}}{N\sqrt{\pi}\int_0^1 \exp(x^2)\,dx} \to 0.$$

Diese Bedingung allein garantiert also offensichtlich nicht die asymptotische Exponentialität der betrachteten Eintrittszeit.

1.7 Ehrenfestsches Urnenmodell II: Irrfahrten auf dem Hyperwürfel

Der N-dimensionale Hyperwürfel $\{0,1\}^N$ bildet zusammen mit der komponentenweisen Addition modulo 2 eine Abelsche Gruppe mit neutralem Element $(0,...,0)$. Durch

$$(i_1,...,i_n) \mapsto \{1 \leq k \leq N \,:\, i_k = 1\}$$

erhält man eine bijektive Abbildung von $\{0,1\}^N$ auf $G_N := \mathcal{P}(\{1,...,N\})$, die die komponentenweise Addition in die symmetrische Differenz überführt. Insbesondere ist auch (G_N, Δ) eine Abelsche Gruppe mit neutralem Element $\emptyset$, in der jedes Element zu sich selbst invers ist.

Zu einer Funktion $f : G_N \to \mathbb{C}$ definiert man die Fourier-Transformierte $\hat{f} : G_N \to \mathbb{C}$ durch

$$\hat{f}(B) := \sum_{A \in G_N} f(A)(-1)^{\#(A \cap B)},$$

und die zugehörige Umkehrformel ist

$$f(A) = 2^{-N} \sum_{B \in G_N} \hat{f}(B)(-1)^{\#(A \cap B)}.$$

Definiert man die Faltung zweier solcher Funktionen durch

$$(f * g)(A) := \sum_{C \in G_N} f(C) g(A \Delta C),$$

so ergibt sich die einprägsame Faltungsformel

$$(f * g)\hat{}(B) = \hat{f}(B)\hat{g}(B).$$

(Vgl. hierzu [Dia88].)

Wir betrachten nun eine Irrfahrt $(X_n)_{n \in \mathbb{N}_0}$ auf G_N mit Start in $\emptyset$, d.h.

$$X_0 = \emptyset, \quad X_n = \Delta_{k=1}^n Y_k, \; n \in \mathbb{N},$$

mit unabhängigen und identisch verteilten G_N-wertigen Zufallsgrößen Y_k, $k \in \mathbb{N}$. Im Zusammenhang mit dem Ehrenfestschen Urnenmodell lässt sich X_n als die Menge der Kugeln in der ersten der beiden Urnen zur Zeit n interpretieren, wobei sich nun

insgesamt N Kugeln in beiden Urnen befinden. Insofern stellt dieses Modell eine Verallgemeinerung des zuvor betrachteten Modells dar.

Es sei $p : G_N \to \mathbb{C}$, $A \mapsto P(Y_1 = A)$, die Massenfunktion zur Verteilung der Zuwächse (Schrittweiten). Dann liefert unsere Faltungsformel $B \mapsto \hat{p}(B)^n$ als Fourier-Transformierte zur Verteilung von X_n, woraus unter Verwendung der Umkehrformel

$$P_\emptyset(X_n = \emptyset) = 2^{-N} \sum_{B \in G_N} \hat{p}(B)^n \text{ für alle } n \in \mathbb{N} \tag{1.38}$$

folgt.

Wir wollen nun das asymptotische Verhalten der Wartezeit

$$T_\emptyset^+ := \inf\{n \in \mathbb{N} : X_n = \emptyset\}$$

mit $N \to \infty$ untersuchen. Es geht uns also um die Frage, wie lange es dauert, um, ausgehend von der Konstellation, dass sich sämtliche Kugeln in der zweiten Urne befinden, wieder zu dieser Konstellation zurückzukehren. Um im Rahmen dieses Urnenmodells zu bleiben, beschränken wir daher unsere Betrachtungen auf sog. *Nearest Neighbour Random Walks*, deren Schrittweitenverteilung von der Form

$$p(\emptyset) =: p_{N,0}, \quad p(\{k\}) =: p_{N,k}, \; k = 1, ..., N,$$

mit $p_{N,k} \geq 0$, $k = 0, 1, ..., N$, und $\sum_{k=0}^N p_{N,k} = 1$ ist. Die zugehörige Fourier-Transformierte ist gegeben durch

$$\hat{p}(B) = 1 - 2 \sum_{k \in B} p_{N,k}. \tag{1.39}$$

Im Fall $p_{N,0} = 0$ und $p_{N,k} = 1/N$ für alle $k = 1, ..., N$ spricht man auch von einem *symmetrischen Nearest Neighbour Random Walk*; hier wird also jede der Kugeln mit der gleichen Wahrscheinlichkeit ausgewählt. Durch Zusammenfassen von Zuständen mit jeweils der gleichen Anzahl von Kugeln können wir dann zu unserem Modell aus Abschnitt 1.6 übergehen und die Fragestellung mit den dort vorgestellten Methoden behandeln. Bei allgemeinen Schrittweitenverteilungen geht jedoch durch diesen Übergang die Markov-Eigenschaft verloren, so dass wir hier einen neuen Ansatz entwickeln müssen.

Wir definieren

$$\begin{aligned} A_{N,0} &:= \{1 \le k \le N : p_{N,k} = 0\} \quad , A_{N,1} := \{1 \le k \le N : p_{N,k} > 0\}, \\ m_{N,0} &:= \#A_{N,0} \qquad\qquad\qquad , m_{N,1} := \#A_{N,1}, \\ G_{N,0} &:= \{B \in G_N : B \subset A_{N,0}\} \quad , G_{N,1} := G_N \setminus G_{N,0}. \end{aligned}$$

Dann gilt insbesondere $m_{N,0} + m_{N,1} = N$, $\#G_{N,0} = 2^{m_{N,0}}$.

Durch $u_0 := 1$, $u_n := P_\emptyset(X_n = \emptyset)$, $n \in \mathbb{N}$, wird die Erneuerungsfolge zur Verteilung $(f_k)_{k\in\mathbb{N}}$, $f_k := P_\emptyset(T_\emptyset^+ = k)$, $k \in \mathbb{N}$, der Rückkehrzeit $T_\emptyset^+$ gegeben. Ist $p_{N,0} > 0$, so ist das Ereignis $\{X_n = \emptyset\}$ aperiodisch und der diskrete Erneuerungssatz (vgl. [Fel68], S. 313, Theorem 3) liefert $E_\emptyset T_\emptyset^+ = \lim_{n\to\infty} u_n^{-1}$. Aus (1.38) und (1.39) ergibt sich in der konkreten Situation

$$u_n = 2^{-N} \sum_{B\in G_N} \Big(1 - 2\sum_{k\in B} p_{N,k}\Big)^n,$$

und da offenbar nur die Summanden für Mengen $B \in G_{N,0}$ asymptotisch einen Beitrag liefern (für diese ist $\sum_{k\in B} p_{N,k} = 0$), erhalten wir

$$\lim_{n\to\infty} u_n = 2^{-m_{N,1}}, \quad \text{also} \quad E_\emptyset T_\emptyset^+ = 2^{m_{N,1}}.$$

Ist $p_{N,0} = 0$, so ist $\{X_n = \emptyset\}$ periodisch mit Periode 2. Hier folgt aus einer allgemeineren Version des diskreten Erneuerungssatzes ([Fel68], S. 313, Theorem 4) $E_\emptyset T_\emptyset^+ = 2\lim_{n\to\infty} u_{2n}^{-1}$. Nun liefern aber sowohl die Mengen $B \in G_{N,0}$ (für diese ist $\sum_{k\in B} p_{N,k} = 0$) als auch die Komplemente dieser Mengen (für diese ist $\sum_{k\in B} p_{N,k} = 1$) asymptotisch einen Beitrag, so dass sich wiederum

$$E_\emptyset T_\emptyset^+ = 2^{m_{N,1}}$$

ergibt.

Man beachte insbesondere, dass der Erwartungswert der Rückkehrzeit $T_\emptyset^+$ nicht von der konkreten Gestalt der Schrittweitenverteilung abhängt; nur die Anzahl der (mit positiver Wahrscheinlichkeit) möglichen Übergänge ist entscheidend. Dies ist eine Konsequenz aus der Tatsache, dass die Übergangsmatrix der Irrfahrt $(X_n)_{n\in\mathbb{N}_0}$ symmetrisch ist, so dass nur die Gleichverteilung auf den mit positiver Wahrscheinlichkeit besuchten Zuständen als stationäre Verteilung in Frage kommt. Das Resultat ergibt sich dann mit dem Hauptgrenzwertsatz für Markov-Ketten.

Als Nächstes wenden wir uns nun der gesamten Verteilung der Rückkehrzeit $T_\emptyset^+$ zu. Ziel soll es dabei sein, Bedingungen für die Schrittweitenverteilung anzugeben, unter denen $T_\emptyset^+ / E_\emptyset T_\emptyset^+$ asymptotisch exponentialverteilt ist.

Sei $U(z)$ die erzeugende Funktion zur Erneuerungsfolge $(u_n)_{n\in\mathbb{N}}$, d.h.

$$U(z) = \sum_{n=0}^{\infty} u_n z^n = 2^{-N} \sum_{n=0}^{\infty} \sum_{B\in G_N} \Big(1 - 2\sum_{k\in B} p_{N,k}\Big)^n z^n.$$

Wegen $|1 - 2\sum_{k\in B} p_{N,k}| \leq 1$ erhalten wir für $0 < z < 1$

$$U(z) = 2^{-N} \sum_{B\in G_N} \Big(1 - \Big(1 - 2\sum_{k\in B} p_{N,k}\Big) z\Big)^{-1}.$$

Ist $F(z) := \sum_{k=1}^{\infty} f_k z^k$ die wahrscheinlichkeitserzeugende Funktion der Rückkehrzeit $T_\emptyset^+$, so gilt bekanntermaßen $U(z) = (1 - F(z))^{-1}$ ([Fel68], Kapitel XIII.3, Theorem 1, S. 311). Bezeichnet man weiter mit $\varphi_N(\theta)$ die charakteristische Funktion der auf Erwartungswert 1 normierten Rückkehrzeit nach $\emptyset$, so gilt also

$$\varphi_N(\theta) = F\big(\exp(i\theta/E_\emptyset T_\emptyset^+)\big) = \frac{U\big(\exp(i\theta 2^{-m_{N,1}})\big) - 1}{U\big(\exp(i\theta 2^{-m_{N,1}})\big)}.$$

Um für $T_\emptyset^+ / E_\emptyset T_\emptyset^+$ mit $N \to \infty$ asymptotische Exponentialität nachzuweisen, reicht es also, für $\theta \neq 0$

$$\lim_{N\to\infty} U\big(\exp(i\theta 2^{-m_{N,1}})\big) = 1 - \frac{1}{i\theta} \tag{1.40}$$

zu zeigen. (Für $\theta = 0$ gilt ohnehin $\varphi_N(0) = F(1) = 1$.)

Wir treffen nun die folgende Modellannahme: Es existiere eine Folge nicht negativer reeller Gewichte $(w_k)_{k\in\mathbb{N}_0}$ derart, dass mit $W_N := \sum_{k=0}^{N} w_k$ gilt:

$$p_{N,k} = w_k / W_N\,,\ k = 0, ..., N\,,\ N \in \mathbb{N}.$$

Im Folgenden bezeichne $(\omega_{N,k})_{k=1}^{m_{N,1}}$ für jedes $N \in \mathbb{N}$ die Folge der von Null verschiedenen, aufsteigend geordneten Gewichte ohne w_0.

Unter diesen Modellannahmen liefert der folgende Satz hinreichende Bedingungen für die asymptotische Exponentialität von $T_\emptyset^+ / E_\emptyset T_\emptyset^+$. Diese Bedingungen sind zunächst rein technischer Natur und in erster Linie durch die verwendete Beweistechnik

motiviert. Sicherlich können sie auch noch an der einen oder anderen Stelle abgeschwächt werden. Andererseits sind sie hinreichend allgemein, um viele interessante Schrittweitenverteilungen abzudecken, vgl. Beispiele 1.22 und 1.23. Zumindest die erste dieser Bedingungen ist auch notwendig für das Vorliegen von asymptotischer Exponentialität, wie das anschließende Lemma 1.24 zeigt.

Satz 1.21 *Es gelte*

(i) $\lim_{N\to\infty} W_N = \infty$,

(ii) *es existiert ein* $\kappa \in (0, \frac{1}{2})$ *mit* $\liminf_{N\to\infty} \sum_{k=1}^{\lfloor \kappa m_{N,1} \rfloor} \omega_{N,k}/W_N > 0$,

(iii) *für* $c := \frac{1}{2}\left(\left((1-\kappa)/\kappa\right)^{\kappa} + \left((1-\kappa)/\kappa\right)^{\kappa-1}\right)$ *gilt* $\lim_{N\to\infty} c^{m_{N,1}} W_N/\omega_{N,1} = 0$,

(iv) $\sum_{N\in A_1} \left(w_N/W_N\right)^2 < \infty$ *mit* $A_1 := \bigcup_{N\in\mathbb{N}} A_{N,1}$.

Dann folgt

$$T_{\emptyset}^{+}/E_{\emptyset}T_{\emptyset}^{+} \xrightarrow{d} \mathrm{Exp}(1).$$

Beweis: Wegen (i) gilt $\lim_{N\to\infty} m_{N,1} = \infty$ und damit insbesondere

$$\begin{aligned}
&\lim_{N\to\infty} 2^{-N} \sum_{B\in G_{N,0}} \Big(1 - \big(1 - 2\sum_{k\in B} p_{N,k}\big) \exp\left(i\theta 2^{-m_{N,1}}\right)\Big)^{-1} \\
&= \lim_{N\to\infty} 2^{-m_{N,1}} \Big(1 - \exp\left(i\theta 2^{-m_{N,1}}\right)\Big)^{-1} = -\frac{1}{i\theta}.
\end{aligned}$$

Daher reicht der Nachweis von

$$\begin{aligned}
&\lim_{N\to\infty} 2^{-N} \sum_{B\in G_{N,1}} \Big(1 - \big(1 - 2\sum_{k\in B} p_{N,k}\big) \exp\left(i\theta 2^{-m_{N,1}}\right)\Big)^{-1} \\
&= \lim_{N\to\infty} 2^{-m_{N,1}} \sum_{\substack{B\subset A_{N,1} \\ B\neq\emptyset}} \Big(1 - \big(1 - 2\sum_{k\in B} p_{N,k}\big) \exp\left(i\theta 2^{-m_{N,1}}\right)\Big)^{-1} = 1
\end{aligned}$$

zum Beweis von (1.40) aus, wobei sich die erste Gleichheit aus der Tatsache ergibt, dass die Summanden zu jeweils $2^{m_{N,0}}$ Mengen $B \in G_{N,1}$, deren Durchschnitte mit $A_{N,1}$ identisch sind, übereinstimmen.

Wir betrachten einen Vektor $(Q_{N,k})_{k\in A_{N,1}}$, der aus $m_{N,1}$ unabhängigen und identisch $\mathrm{Bin}(1, \frac{1}{2})$-verteilten Zufallsvariablen besteht. Es sei $Q_N := \sum_{k\in A_{N,1}} Q_{N,k}$, $R_N := \sum_{k\in A_{N,1}} p_{N,k} Q_{N,k}$ und $\Psi_N : [0,1] \to \mathbb{C}$ mit

$$\Psi_N(r) := \Big(1 - \big(1 - 2r\big) \exp(i\theta 2^{-m_{N,1}})\Big)^{-1}.$$

Dann lässt sich

$$2^{-m_{N,1}} \sum_{\substack{B\subset A_{N,1}\\ B\neq\emptyset}} \Big(1-\big(1-2\sum_{k\in B} p_{N,k}\big)\exp\big(i\theta 2^{-m_{N,1}}\big)\Big)^{-1}$$

schreiben als EY_N mit

$$Y_N := \Psi_N\big(R_N\big)\mathbb{1}\{Q_N>0\}.$$

Da wir uns nur für den Grenzübergang $N\to\infty$ interessieren, können wir im Folgenden wegen $\lim_{N\to\infty} m_{N,1}=\infty$ voraussetzen, dass der Realteil von $\exp(i\theta 2^{-m_{N,1}})$ positiv ist. Dann gilt

$$|\Psi_N(r)|\le 1 \text{ für } r\ge 1/2 \text{ sowie } |\Psi_N(r)|\le 1/(2a) \text{ für } 1/2>r\ge a>0. \tag{1.41}$$

Wir zerlegen nun EY_N in

$$EY_N = EY_N\mathbb{1}\{Q_N\le \kappa m_{N,1}\} + EY_N\mathbb{1}\{Q_N>\kappa m_{N,1}\}.$$

Da Q_N mit den Parametern $m_{N,1}$ und $1/2$ binomialverteilt ist, haben wir dann wegen (1.41) und Hilfssatz A.5 für den ersten Summanden die Abschätzung

$$EY_N 1\{Q_N\le\kappa m_{N,1}\} \le \frac{W_N}{2\omega_{N,1}} P(Q_N\le\kappa m_{N,1}) \le \frac{1}{2}\,\frac{W_N c^{m_{N,1}}}{\omega_{N,1}},$$

und diese Obergrenze strebt wegen Bedingung (iii) gegen 0.

Für den zweiten Summanden zeigen wir zunächst, dass R_N f.s. gegen $1/2$ konvergiert. Es ist

$$R_N = \sum_{k\in A_{N,1}} p_{N,k}Q_k = \sum_{k\in A_{N,1}} \frac{w_k Q_k}{W_N}.$$

Die Voraussetzung (iv) liefert

$$\sum_{N\in A_1} \frac{\operatorname{var}(w_N Q_N)}{W_N^2} = \frac{1}{4}\sum_{N\in A_1}\frac{w_N^2}{W_N^2} < \infty.$$

Voraussetzung (i) garantiert, dass $W_N\uparrow\infty$. Dann liefert [Loè63], S. 238, A., dass

$$\frac{\sum_{k\in A_{N,1}} w_k(Q_k-1/2)}{W_N} \xrightarrow{\text{f.s.}} 0$$

gilt und wegen

$$\frac{1}{2}\frac{\sum_{k\in A_{N,1}} w_k}{W_N} = \frac{1}{2}\frac{W_N - w_0}{W_N} \to \frac{1}{2}$$

folgt $R_N \xrightarrow{\text{f.s.}} 1/2$.

Als unmittelbare Konsequenz ergibt sich $\Psi_N(R_N) \xrightarrow{\text{f.s.}} 1$. Außerdem folgt aus den Betrachtungen zum ersten Summanden, dass $P(Q_N > \kappa m_{N,1})$ gegen 1 strebt. Schließlich folgt aus $Q_N > \kappa m_{N,1}$, dass $R_N \geq \sum_{k=1}^{\lfloor \kappa m_{N,1} \rfloor} \omega_{N,k}/W_N$ gilt. Wegen Bedingung (ii) gibt es dann ein $\varepsilon > 0$ und ein $N_0 \in \mathbb{N}$, so dass $R_N \geq \varepsilon$ für alle $N \geq N_0$ gilt und somit nach (1.41) $|\Psi_N(R_N)| \leq 1/(2\varepsilon)$ ist. So erhält man insgesamt mit dem Satz von der majorisierten Konvergenz

$$\lim_{N\to\infty} EY_N \mathbb{1}\{Q_N > \kappa m_{N,1}\} = 1.$$

□

Es folgen nun zwei Beispiele für Klassen von positiven Folgen $(w_k)_{k\in\mathbb{N}_0}$, die den Voraussetzungen von Satz 1.21 genügen. Beide können problemlos auf den Fall verallgemeinert werden, dass einzelne Folgenglieder verschwinden, sofern nur $\lim_{N\to\infty} \#\{0 \leq k \leq N : w_k > 0\} = \infty$ gilt, und die positiven Folgenglieder in diese Klasse fallen.

Beispiel 1.22 Die Voraussetzungen von Satz 1.21 sind für positive, beschränkte Folgen $(w_k)_{k\in\mathbb{N}_0}$ mit $\liminf_{k\to\infty} w_k > 0$ und $\lim_{N\to\infty} W_N/N = a > 0$ erfüllt. Ein Spezialfall hiervon sind konvergente Folgen $(w_k)_{k\in\mathbb{N}_0}$ mit positivem Grenzwert. Insbesondere liefert jede konstante Folge die asymptotische Exponentialität für *symmetrische Nearest Neighbour Random Walks*, also solche, bei denen sämtliche Übergänge dieselbe Wahrscheinlichkeit haben.

Zum Nachweis der einzelnen Voraussetzungen: (i) folgt aus $W_N \sim aN$. Beachtet man, dass bei positiven Folgengliedern w_k aus $\liminf_{k\to\infty} w_k > 0$ folgt, dass $\liminf_{N\to\infty} \omega_{N,1} > 0$ gilt, so ergibt sich (ii) für $\kappa = 1/4$ aus

$$\liminf_{N\to\infty} \sum_{k=1}^{\lfloor N/4 \rfloor} \frac{\omega_{N,k}}{W_N} \geq \liminf_{N\to\infty} \frac{\lfloor N/4 \rfloor\, \omega_{N,1}}{W_N} \geq \frac{1}{4a} \liminf_{N\to\infty} \omega_{N,1} > 0.$$

(iii) folgt für das zu $\kappa = 1/4$ gehörige c aus

$$\frac{c^N W_N}{\omega_{N,1}} \sim \frac{c^N Na}{\omega_{N,1}} \leq c^N N \frac{a}{\liminf_{N\to\infty} \omega_{N,1}} \to 0.$$

(iv) folgt schließlich wegen der Beschränktheit der w_k aus

$$\sum_{N=1}^{\infty}\Big(\frac{w_N}{W_N}\Big)^2 \le \sum_{N=1}^{\infty}\Big(\frac{N}{W_N}\Big)^2\Big(\frac{\sup w_k}{N}\Big)^2 < \infty.$$

Beispiel 1.23 Die Voraussetzungen von Satz 1.21 sind ebenfalls erfüllt für Folgen $(w_k)_{k\in\mathbb{N}_0}$ der Bauart $w_k = k^{-\alpha}$ mit $\alpha \in (0,1)$, $w_0 \ge 0$ beliebig. Auch hier weisen wir wieder die einzelnen Voraussetzungen nach: (i) ist trivial. Voraussetzung (ii) ist beispielsweise mit $\kappa = 1/4$ erfüllt, es gilt bei Verwendung von $\int_1^{n+1} x^{-\alpha}dx \le \sum_{k=1}^{n} k^{-\alpha} \le 1 + \int_1^n x^{-\alpha}dx$, $n \in \mathbb{N}$:

$$\sum_{k=1}^{\lfloor N/4\rfloor}\frac{\omega_{N,k}}{W_N} = \frac{\sum_{k=\lceil 3N/4\rceil}^{N} k^{-\alpha}}{\sum_{k=1}^{N} k^{-\alpha}} \ge 1 - \frac{1+\int_1^{3N/4} x^{-\alpha}dx}{\int_1^{N+1} x^{-\alpha}dx} \to 1-(3/4)^{1-\alpha} > 0.$$

Zu Voraussetzung (iii): Für das zu $\kappa = 1/4$ gehörige c gilt

$$\frac{c^N W_N}{\omega_{N,1}} = N^{\alpha}c^N\sum_{k=1}^{N}k^{-\alpha} \le N^{1+\alpha}c^N \to 0.$$

Schließlich ist auch Voraussetzung (iv) erfüllt:

$$\sum_{N=1}^{\infty}\Big(\frac{w_N}{W_N}\Big)^2 \le \sum_{N=1}^{\infty}\Big(\frac{N^{-\alpha}}{N\cdot N^{-\alpha}}\Big)^2 = \sum_{N=1}^{\infty}\frac{1}{N^2} < \infty.$$

Lemma 1.24 *Die Bedingung (i)* $\lim_{N\to\infty} W_N = \infty$ *in Satz 1.21 ist notwendig für die asymptotische Exponentialität von* $T_\emptyset^+/E_\emptyset T_\emptyset^+$.

Beweis: Angenommen es gilt $\lim_{N\to\infty} m_{N,1} < \infty$. In diesem Fall bleibt $E_\emptyset T_\emptyset^+ = 2^{m_{N,1}}$ beschränkt, so dass die Grenzverteilung von $T_\emptyset^+/E_\emptyset T_\emptyset^+$ nur diskrete Werte annehmen kann. Dann kann sich jedoch keine asymptotische Exponentialität einstellen. Sei also $\lim_{N\to\infty} m_{N,1} = \infty$. Angenommen nun gilt $\limsup_{N\to\infty} W_N = c < \infty$. Dann ist $\liminf_{N\to\infty} p_{N,k} > 0$ für alle mit $k \in A_1 = \bigcup_{N\in\mathbb{N}} A_{N,1}$. Damit ergibt sich

$$\lim_{N\to\infty} P(T_\emptyset^+/E_\emptyset T_\emptyset^+ \le 2/2^{m_{N,1}}) = \lim_{N\to\infty} P(T_\emptyset^+ \le 2) \ge \lim_{N\to\infty}\sum_{k\in A_{N,1}} p_{N,k}^2 > 0,$$

im Widerspruch dazu, dass eine Exponentialverteilung eine stetige Verteilungsfunktion besitzt, die an der Stelle Null den Wert Null hat. □

Mit ähnlichen Argumenten kann man auch die Eintrittszeit in andere Zustände untersuchen. Sei z.B. $M := \{1,...,N\}$ und

$$T_M^+ := \inf\{n \in \mathbb{N} : X_n = M\},$$

d.h. bei Start in $\emptyset$ wäre T_M^+ die Wartezeit zum Durchqueren des Hyperwürfels. Im Zusammenhang mit unserem EF-Modell lässt sich dies auch als die Zeit interpretieren, bis sämtliche Kugeln von einer Urne in die andere gewandert sind.

Die Umkehrformel liefert in diesem Fall

$$P_\emptyset(X_n = M) = 2^{-N} \sum_{B \in G_N} (-1)^{\#B} \Big(1 - 2 \sum_{k \in B} p_{N,k}\Big)^n.$$

Die zugehörige erzeugende Funktion lautet entsprechend

$$P_{\emptyset M}(z) = \sum_{n=0}^{\infty} P_\emptyset(X_n = M) z^n = 2^{-N} \sum_{B \in G_N} (-1)^{\#B} \Big(1 - \big(1 - 2 \sum_{k \in B} p_{N,k}\big) z\Big)^{-1}.$$

Verwendet man, dass $P_B(X_n = B)$ unabhängig von $B \in G_N$ ist, so erhält man durch allgemeine erneuerungstheoretische Betrachtungen (vgl. [Fel68], S. 316f.) die Beziehung

$$F_{\emptyset M}(z) U(z) = P_{\emptyset M}(z),$$

wobei $F_{\emptyset M}(z)$ die wahrscheinlichkeitserzeugende Funktion zu $\mathcal{L}_\emptyset(T_M^+)$ ist.

Wir wollen nun ein Analogon zu Satz 1.21 für $\mathcal{L}_\emptyset(T_M^+)$ beweisen. Dabei ist klar, dass nun sämtliche Übergangswahrscheinlichkeiten $p_{N,k}$ mit $k \geq 1$ positiv sein müssen, damit überhaupt die Chance besteht, den Zustand M zu erreichen. Dementsprechend verändern sich die Anforderungen an die Folge der Gewichte w_k für $k \geq 1$. Insbesondere erübrigt sich nun die Zerlegung der Menge G_N in $G_{N,0}$ und $G_{N,1}$; $m_{N,0}$ ist nun 0, $m_{N,1}$ entspricht N.

Bezeichnet also $\varphi_N(\theta)$ die charakteristische Funktion von $\mathcal{L}_\emptyset(T_M^+/2^N)$, so ist

$$\varphi_N(\theta) = \frac{P_{\emptyset M}\big(\exp(i\theta 2^{-N})\big)}{U\big(\exp(i\theta 2^{-N})\big)}.$$

Wir beweisen nun den folgenden

Satz 1.25 *Es gelte*

(i) $\lim_{N\to\infty} W_N = \infty$,

(ii) *es existiert ein* $\kappa \in (0, \frac{1}{2})$ *mit* $\liminf_{N\to\infty} \sum_{k=1}^{\lfloor \kappa N \rfloor} \omega_{N,k}/W_N > 0$,

(iii) *für* $c := \frac{1}{2}\left(\big((1-\kappa)/\kappa\big)^\kappa + \big((1-\kappa)/\kappa\big)^{\kappa-1}\right)$ *gilt* $\lim_{N\to\infty} c^N W_N/\omega_{N,1} = 0$,

(iv) $\sum_{N=1}^{\infty} \big(w_N/W_N\big)^2 < \infty$.

Dann folgt

$$\mathcal{L}_{\emptyset}\big(T_M^+/2^N\big) \xrightarrow{d} \mathrm{Exp}(1).$$

Beweis: Das Verhalten von $U\big(\exp(i\theta 2^{-N})\big)$ haben wir bereits im Beweis von Satz 1.21 untersucht, es gilt

$$\lim_{N\to\infty} U\big(\exp(i\theta 2^{-N})\big) = 1 - \frac{1}{i\theta}.$$

Bleibt also zu zeigen, dass

$$\lim_{N\to\infty} P_{\emptyset M}\big(\exp(i\theta 2^{-N})\big) = -\frac{1}{i\theta}$$

ist. Wegen

$$\lim_{N\to\infty} 2^{-N}\big(1-\exp(i\theta 2^{-N})\big)^{-1} = -\frac{1}{i\theta}$$

reduziert sich dies wiederum auf den Nachweis von

$$\lim_{N\to\infty} 2^{-N} \sum_{\substack{B\in G_N\\ B\neq\emptyset}} (-1)^{\#B}\Big(1-\big(1-2\sum_{k\in B} p_{N,k}\big)\exp\big(i\theta 2^{-N}\big)\Big)^{-1} = 0.$$

Auch hier betrachten wir wieder einen Vektor $(Q_{N,k})_{k=1,\dots,N}$ von N unabhängigen und identisch $\mathrm{Bin}(1,\frac{1}{2})$-verteilten Zufallsvariablen; es sei $Q_N := \sum_{k=1}^N Q_{N,k}$, $R_N := \sum_{k=1}^N p_{N,k}Q_{N,k}$ und $\Psi_N : [0,1] \to \mathbb{C}$ mit

$$\Psi_N(r) = \big(1-(1-2r)\exp(i\theta 2^{-N})\big)^{-1}.$$

Dann lässt sich hier

$$2^{-N} \sum_{\substack{B\in G_N\\ B\neq\emptyset}} (-1)^{\#B}\Big(1-\big(1-2\sum_{k\in B} p_{N,k}\big)\exp\big(i\theta 2^{-N}\big)\Big)^{-1}$$

schreiben als EY_N, wobei nun

$$Y_N = \Psi_N(R_N)\mathbb{1}\{Q_N>0\}(-1)^{Q_N}.$$

Wir zerlegen wieder EY_N in

$$EY_N = EY_N\mathbb{1}\{Q_N \le \kappa N\} + EY_N\mathbb{1}\{Q_N > \kappa N\}.$$

Für den ersten Summanden argumentieren wir nun, dass

$$EY_N \mathbb{1}\{Q_N \leq \kappa N\} \leq E|Y_N| \mathbb{1}\{Q_N \leq \kappa N\}$$

gilt. Für den rechten Term haben wir bereits im Beweis zu Satz 1.21 mit $N \to \infty$ den Grenzwert 0 nachgewiesen.

Beim zweiten Term können wir wieder verwenden, dass $\Psi_N(R_N)$ f.s. gegen 1 konvergiert und ebenso, dass $P(Q_N > \kappa N)$ gegen 1 strebt und $\Psi_N(R_N)$ für $Q_N > \kappa N$ bei hinreichend großem N beschränkt ist. Wir erhalten

$$\begin{aligned} & E\Psi_N(R_N)\mathbb{1}\{Q_N > \kappa N\}(-1)^{Q_N} \\ = \;& E\Psi_N(R_N)\mathbb{1}\{Q_N > \kappa N,\, Q_N \text{ gerade}\} - \\ & E\Psi_N(R_N)\mathbb{1}\{Q_N > \kappa N,\, Q_N \text{ ungerade}\} \\ \to \;& \frac{1}{2} - \frac{1}{2} = 0. \end{aligned}$$

□

1.8 Das *Exact Pattern Matching Problem*

Zum Abschluss dieses Kapitels über asymptotische Exponentialität bei Wartezeitverteilungen und gleichzeitig als Übergang zur folgenden Analyse des *Approximate Pattern Matching Problems* werden wir nun noch die asymptotische Exponentialität für das *Exact Pattern Matching Problem* nachweisen. Im Fall einer unabhängigen und identisch verteilten Zeichenkette wurde dieses Resultat in [Rud96] gezeigt (vgl. Satz 1.27). Mit den in [RD99] bereitgestellten Resultaten gelingt die Verallgemeinerung auf den Fall einer durch eine Markov-Kette erzeugten Zeichenkette (Satz 1.28).

Es sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von unabhängigen und identisch verteilten Zufallsvariablen über einem endlichen Alphabet $\Sigma := \{\sigma_1, ..., \sigma_s\}$, $s \in \mathbb{N}$. Ein solche Folge bezeichnet man auch als zufällige Zeichenkette über Σ. Um Trivialitäten zu vermeiden, sei $s \geq 2$ und $p_{\sigma_i} := P(X_1 = \sigma_i) > 0$ für alle $i = 1, ..., s$.

Eine endliche Folge $A = a_1a_2...a_N$ über Σ bezeichnet man als einen *Pattern*. Für die Länge eines Patterns verwenden wir die Schreibweise $|A|$. Σ^N bezeichnet die Menge aller Pattern über Σ der Länge N, $\Sigma^* := \bigcup_{N\in\mathbb{N}} \Sigma^N$. Für einen Pattern $A = a_1a_2...a_N$ verwenden wir häufig die Kurzschreibweisen $A_{1:N}$ oder $A = a_{1:N}$. Für $1 \leq i \leq j \leq N$ bezeichnet dementsprechend $A_{i:j}$ den Teilpattern $a_ia_{i+1}...a_{j-1}a_j$. Ebenso bezeichnet für eine Zeichenkette $(X_n)_{n\in\mathbb{N}}$ der Term $X_{i:j}$ die endliche Teilfolge $X_iX_{i+1}...X_{j-1}X_j$ $(1 \leq i \leq j)$. Ist $i > j$, so sei $A_{i:j}$ bzw. $X_{i:j}$ grundsätzlich der sog. *leere Pattern* $\emptyset$.

Es sei nun $(A_N)_{N\in\mathbb{N}}$ eine Folge von Pattern über Σ mit $|A_N| = N$, $A_N := a_{N,1:N}$. T_N sei die Wartezeit, bis der Pattern A_N erstmals in der zufälligen Zeichenkette $(X_n)_{n\in\mathbb{N}}$ erscheint, d.h.

$$T_N := \inf\{n \geq N \,:\, X_{n-N+1:n} = A_N\}.$$

Für die zugehörige wahrscheinlichkeitserzeugende Funktion

$$g_{T_N}(z) := \sum_{n=N}^{\infty} P(T_N = n)z^n$$

ist dann bekannt, dass sie sich in der folgenden Weise darstellen lässt:

$$g_{T_N}(z) = \frac{\mathbb{P}(A_N)z^N}{\mathbb{P}(A_N)z^N + (1-z)\Phi_{A_N}(z)} \tag{1.42}$$

mit

$$\Phi_B(z) := \sum_{j=0}^{r-1} \chi_B(j)\mathbb{P}(B_{r-j+1:r})z^j, \quad \mathbb{P}(B) := \prod_{i=1}^{r} p_{b_i}, \quad \chi_B(j) := \mathbb{1}\{B_{j+1:r} = B_{1:r-j}\},$$

für beliebige Pattern $B = b_1 b_2 ... b_r$. (Vgl. [GL81], Theorem 4.1, S. 106 oder [Rei01].) $\Phi_B(z)$ bezeichnet man auch als *Überlappungspolynom* des Patterns B.

Weiter ist dann $\varphi_N(\theta) := g_{T_N}(\exp(i\theta/ET_N))$ die charakteristische Funktion zu T_N/ET_N. Um die Verteilungskonvergenz von T_N/ET_N gegen eine Exponentialverteilung nachzuweisen, ist es mithin ausreichend,

$$\lim_{N\to\infty} \mathbb{P}(A_N)^{-1} \frac{1-\exp(i\theta/ET_N)}{\exp(i\theta N/ET_N)} \Phi_{A_N}\big(\exp(i\theta/ET_N)\big) = -i\theta$$

zu zeigen. Der Ausdruck auf der linken Seite lässt sich darstellen als

$$(-i\theta)\cdot\exp(-i\theta N/ET_N)\cdot\frac{1-\exp(i\theta/ET_N)}{0-i\theta/ET_N}\cdot\frac{\Phi_{A_N}\big(\exp(i\theta/ET_N)\big)}{\Phi_{A_N}(1)}, \tag{1.43}$$

wobei wir $ET_N = \Phi_{A_N}(1)/\mathbb{P}(A_N)$ verwenden. (Vgl. [GL81], S. 106, (30) oder [Rei01], oder man differenziere die wahrscheinlichkeitserzeugende Funktion in $z = 1$.)

Wegen $\chi_{A_N}(0) = 1$ gilt $\Phi_{A_N}(1) \geq 1$ und damit $ET_N \geq 1/\mathbb{P}(A_N)$. Deshalb folgt $ET_N/N \to \infty$. Damit ergibt sich für den zweiten Faktor in (1.43) mit $N \to \infty$ der Grenzwert $\exp(0) = 1$. Der dritte Faktor strebt als Differenzenquotient gegen $\exp'(0)$, also ebenfalls gegen 1. Der vierte Faktor ist schließlich von der Bauart $\sum_{k=0}^{N-1} c_{N,k}\cdot\exp(i\theta k/ET_N)$ mit $c_{N,k} \in [0,1]$, $\sum_{k=0}^{N-1} c_{N,k} = 1$.

Lemma 1.26 *Seien $c_{N,k} \in [0,1]$, $k = 0, ..., N-1$ mit $\sum_{k=0}^{N-1} c_{N,k} = 1$. Dann gilt*

$$\lim_{N\to\infty} \sum_{k=0}^{N-1} c_{N,k}\cdot\exp(i\theta k/ET_N) = 1.$$

Beweis: Es gilt

$$\begin{aligned}
\left|1 - \sum_{k=0}^{N-1} c_{N,k} \exp(i\theta k/ET_N)\right| &\leq \sum_{k=0}^{N-1} c_{N,k} |1 - \exp(i\theta k/ET_N)| \\
&\leq \sum_{k=0}^{N-1} c_{N,k} \max_{l=0}^{N-1} |1 - \exp(i\theta l/ET_N)| \\
&= \max_{l=0}^{N-1} |1 - \exp(i\theta l/ET_N)| \\
&\leq \max_{l=0}^{N-1} \max_{\{y \in \mathbb{C} : |y| \leq 1\}} |\exp(y)| \cdot |0 - i\theta l/ET_N| \\
&\leq \max_{\{y \in \mathbb{C} : |y| \leq 1\}} |\exp(y)| \cdot |\theta| \cdot N/ET_N \;\to\; 0.
\end{aligned}$$

□

Damit haben wir die asymptotische Exponentialität von T_N nachgewiesen:

Satz 1.27 *Es sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von unabhängigen und identisch verteilten Zufallsvariablen über einem endlichen Alphabet Σ und $(A_N)_{N\in\mathbb{N}}$ eine Folge von Pattern über Σ mit $|A_N| = N$. Dann gilt für die Wartezeit T_N, bis der Pattern A_N erstmals in der zufälligen Zeichenkette $(X_n)_{n\in\mathbb{N}}$ erscheint, mit $N \to \infty$:*

$$T_N/ET_N \xrightarrow{d} \mathrm{Exp}(1).$$

(Vgl. [Rud96], S. 51, Theorem 13.2.)

Dieses Resultat lässt sich auch problemlos auf den Fall übertragen, bei dem die zufällige Zeichenkette X nicht durch eine Folge von unabhängigen, identisch verteilten Zufallsvariablen, sondern durch eine irreduzible und aperiodische Markov-Kette generiert wird, d.h. $(X_n)_{n\in\mathbb{N}_0}$ ist eine Markov-Kette über Σ mit Übergangsmatrix $P = (p(\sigma_i, \sigma_j))_{i,j=1,\ldots,s}$ mit $p(\sigma_i, \sigma_j) = P(X_1 = \sigma_j | X_0 = \sigma_i)$.

Die stationäre Verteilung von X bezeichnen wir mit $\pi := \big(\pi(\sigma_i)\big)_{i=1,\ldots,s}$. Der Einfachheit halber gehen wir hier davon aus, dass die Markov-Kette X stationär ist, d.h. π ist zugleich die Startverteilung für X: $P(X_0 = \,\cdot\,) = \pi(\,\cdot\,)$. Eine Verallgemeinerung auf beliebige Startverteilungen ist möglich, soll hier aber nicht genauer untersucht werden.

Unter diesen Voraussetzungen ist die wahrscheinlichkeitserzeugende Funktion von T_N gegeben durch [RD99], S. 183, Theorem 2. Bezeichnet man mit $g_{\sigma_i\sigma_j}(z)$ die

wahrscheinlichkeitserzeugende Funktion von $\tau^+_{\sigma_j} := \inf\{n \in \mathbb{N} : X_n = \sigma_j\}$ bei Start in σ_i (also unter der Bedingung $\{X_0 = \sigma_i\}$), so gilt wegen [Fel68], S. 317, (5.3)

$$\sum_{n=1}^{\infty} P(X_n = a_{N,1} | X_0 = a_{N,N}) z^n = g_{a_{N,N}a_{N,1}}(z) / \big(1 - g_{a_{N,1}a_{N,1}}(z)\big).$$

Das zitierte Ergebnis aus [RD99] lässt sich dann in Anlehnung an die Notation von (1.42) darstellen als

$$g_{T_N}(z) = g_{N,1}(z) \cdot g_{N,2}(z)$$

mit

$$\begin{aligned} g_{N,1}(z) &= \pi(a_{N,1}) \cdot \frac{1 - g_{a_{N,1}a_{N,1}}(z)}{1 - z}, \\ g_{N,2}(z) &= \frac{\mathbb{P}(A_N) z^N}{\mathbb{P}(A_N) z^{N-1} g_{a_{N,N}a_{N,1}}(z) + \big(1 - g_{a_{N,1}a_{N,1}}(z)\big)\Phi_{A_N}(z)} \end{aligned}$$

sowie

$$\Phi_B(z) := \sum_{j=0}^{r-1} \chi_B(j) \mathbb{P}(B_{r-j:r}) z^j, \ \mathbb{P}(B) := \prod_{i=1}^{r-1} p(b_i, b_{i+1}), \ \chi_B(j) := \mathbb{1}\{B_{j+1:r} = B_{1:r-j}\}$$

für beliebige Pattern $B = b_1 b_2 ... b_r$. (Um Trivialitäten zu vermeiden, setzen wir dabei voraus, dass $\mathbb{P}(A_N) > 0$ gilt.)

Auch hier liefert $\varphi_N(\theta) := g_{T_N}\big(\exp(i\theta/ET_N)\big)$ die charakteristische Funktion zu T_N/ET_N. Die Verteilungskonvergenz gegen eine Exponentialverteilung wäre gezeigt, wenn man für diese Funktion mit $N \to \infty$ den Grenzwert $(1 - i\theta)^{-1}$ nachweisen könnte.

Den Erwartungswert ET_N erhalten wir wieder durch Differenzieren von $g_{T_N}(z)$ in 1. Es gilt $g_{N,1}(1) = g_{N,2}(1) = 1$, sowie $g'_{N,1}(1) = E_{a_{N,1}} \tau^+_{a_{N,1}} (\tau^+_{a_{N,1}} - 1)/(2E_{a_{N,1}} \tau^+_{a_{N,1}})$ und $g'_{N,2}(1) = E_{a_{N,1}} \tau^+_{a_{N,1}} \Phi_{A_N}(1)/\mathbb{P}(A_N) + 1 - E_{a_{N,N}} \tau^+_{a_{N,1}}$ (man stelle die Ableitung jeweils als Grenzwert des Differenzenquotienten dar und verwende die Regel von l'Hospital), also

$$ET_N = \frac{E_{a_{N,1}} \tau^+_{a_{N,1}}}{\mathbb{P}(A_N)} \Phi_{A_N}(1) + 1 - E_{a_{N,N}} \tau^+_{a_{N,1}} + \frac{E_{a_{N,1}} \tau^+_{a_{N,1}} (\tau^+_{a_{N,1}} - 1)}{(2E_{a_{N,1}} \tau^+_{a_{N,1}})}. \tag{1.44}$$

Wie zuvor gilt also wieder $ET_N/N \to \infty$. Dies sowie $g_{N,1}\big(\exp(i\theta/ET_N)\big) \to 1$ liefert

$$g_{T_N}\big(\exp(i\theta/ET_N)\big)^{-1} \sim 1 + \frac{1 - g_{a_{N,1}a_{N,1}}\big(\exp(i\theta/ET_N)\big)}{\mathbb{P}(A_N)} \Phi_{A_N}\big(\exp(i\theta/ET_N)\big).$$

Wegen

$$\lim_{N\to\infty} \frac{1 - g_{a_{N,1}a_{N,1}}\big(\exp(i\theta/ET_N)\big)}{0 - i\theta/ET_N} = E_{a_{N,1}}\tau^+_{a_{N,1}}$$

ist deshalb zum Nachweis der asymptotischen Exponentialität lediglich der Nachweis von

$$\lim_{N\to\infty} \frac{E_{a_{N,1}}\tau^+_{a_{N,1}}}{\mathbb{P}(A_N)ET_N}\,\Phi_{A_N}(\exp(i\theta/ET_N)) = 1$$

erforderlich. Dieser Grenzwert ergibt sich tatsächlich bei Verwendung der expliziten Darstellung (1.44) für den Erwartungswert ET_N mit Hilfe von Lemma 1.26.

Wir haben also folgende Verallgemeinerung von Satz 1.27:

Satz 1.28 *Es sei $(X_n)_{n\in\mathbb{N}_0}$ eine irreduzible und aperiodische Markov-Kette über einem endlichen Alphabet Σ und $(A_N)_{N\in\mathbb{N}}$ eine Folge von Pattern über Σ mit $|A_N| = N$ und $\mathbb{P}(A_N) > 0$. Dann gilt für die Wartezeit T_N, bis der Pattern A_N erstmals in der zufälligen Zeichenkette $(X_n)_{n\in\mathbb{N}_0}$ erscheint, mit $N \to \infty$:*

$$T_N/ET_N \xrightarrow{d} \mathrm{Exp}(1).$$

Motiviert durch dieses Resultat sowie die allgemeinen Resultate zur asymptotischen Exponentialität von Wartezeitverteilungen soll es im nun folgenden Kapitel darum gehen, das *Approximate Pattern Matching Problem* ebenfalls auf ein eventuell vorhandenes asymptotisch-exponentielles Verhalten hin zu untersuchen.

Kapitel 2

Das *Approximate Pattern Matching Problem*

In diesem Kapitel geht es um die Untersuchung eines weiteren *Pattern Matching Problems*, des so genannten *Approximate Pattern Matching Problems* (im Folgenden kurz *APMP*): Wie lange dauert es, bis in einer zufälligen Zeichenkette ein fest vorgegebener Pattern erstmals näherungsweise erscheint? Fragestellungen dieser Art spielen eine wichtige Rolle in vielen Anwendungsgebieten, beispielsweise in der mathematischen Biologie, der Qualitätskontrolle, bei Internet-Suchmaschinen oder der Laufzeitanalyse von *Pattern Matching* Algorithmen. (Für diese und weitere Anwendungen siehe [Nav01] und [Wit03].)

Die exakte Verteilung der Wartezeit hängt sowohl vom Pattern als auch von der mathematischen Präzisierung des Begriffs „näherungsweise“ ab und wird in der Regel eine sehr komplizierte Gestalt haben, die bei langen Pattern nicht mehr mit vertretbarem Rechenaufwand zugänglich ist. Andererseits zeigen experimentelle Untersuchungen, dass es eine Reihe von Möglichkeiten gibt, diese Wartezeitverteilung gerade bei sehr langen Pattern durch einfachere und wohlbekannte Verteilungen zu approximieren. So zeigt eine graphische Darstellung einer solchen Wartezeitverteilung nach Normierung auf Erwartungswert 1 häufig eine starke Ähnlichkeit zum entsprechenden Graphen einer Exp(1)-Verteilung. Eine solche Approximation trägt allerdings nicht der Tatsache Rechnung, dass es sich bei der Wartezeitverteilung um eine diskrete Verteilung handelt. Möchte man diesen Aspekt berücksichtigen, so bietet sich zur Approximation als diskretes Analogon zur Exponentialverteilung die geometrische

Verteilung – die klassische diskrete Wartezeitverteilung – an. Dies ist in der Tat eine von Praktikern häufig verwendete Approximation des eigentlichen Problems (vgl. Abbildung 2.1). Eine weitere Möglichkeit ist, die Anzahl des näherungsweisen Auftretens des Patterns durch eine geeignete Poisson-Verteilung anzunähern. Alle diese Näherungsverfahren wollen wir in diesem Kapitel untersuchen, und zwar insbesondere unter den (diesem sehr anwendungsorientierten Problem angemessenen) Aspekten ihres praktischen Nutzens und ihrer effizienten Umsetzbarkeit für konkrete Problemstellungen.

Im ersten Abschnitt werden wir zunächst als Grundlage unserer weiteren Untersuchungen einen Abstandsbegriff für Pattern einführen. Wir entscheiden uns dabei für den geläufigsten Begriff der *Editierdistanz* oder *Edit Distance*. Es werden einige einfache Eigenschaften bewiesen und eine algorithmische Berechnungsmethode vorgestellt, wobei wir im Wesentlichen [Gus97] folgen.

Nach dieser Vorarbeit können wir im zweiten Abschnitt die Methoden des vorangegangenen Kapitels verwenden, um unter speziellen Voraussetzungen für das *APMP* asymptotische Exponentialität nachzuweisen. Unter praktischen Gesichtspunkten ist dieses Resultat allerdings wenig zufrieden stellend, da es im Fall endlicher Pattern keine Aussage über die Güte der Approximation der Wartezeitverteilung durch eine Exponentialverteilung macht.

Im dritten Abschnitt beschäftigen wir uns daher allgemein mit dem Problem, wie sich Eintrittszeitverteilungen für Markov-Ketten inklusive Fehlerabschätzungen approximieren lassen. Als Abstandsbegriff für Verteilungen führen wir den Totalvariationsabstand ein. In diesem Zusammenhang erweist sich eine Exponentialverteilung als Grenzverteilung nun als nutzlos und wird durch eine geometrische Verteilung ersetzt.

Nachdem wir die entsprechenden theoretischen Resultate hergeleitet haben, untersuchen wir im vierten Abschnitt, wie diese unter praktischen Gesichtspunkten im Fall des *APMP* zu beurteilen sind. Basierend auf der algorithmischen Berechnungsmethode für den Totalvariationsabstand zweier Pattern definieren wir dazu eine neue Markov-Kette, den sog. *Spaltenprozess*, mit dem sich die Wartezeit des *APMP* als Eintrittszeit interpretieren lässt. Wichtige theoretische und praxisrelevante Eigenschaften dieses Prozesses werden nachgewiesen.

Mit seiner Hilfe ist es nun möglich, die Wartezeit des *APMP* unter Verwendung der

Resultate aus Abschnitt 3 mit vertretbarem Rechenaufwand durch eine geometrische Verteilung zu approximieren. In Abschnitt 5 wird ein weiteres, speziell auf das *APMP* zugeschnittenes Resultat nachgewiesen. Die so gewonnenen Aussagen eignen sich hervorragend dazu, in einem Computerprogramm umgesetzt zu werden (vgl. Anhang C).

Um die erzielten Resultate zu veranschaulichen und ihren praktischen Nutzen zu unterstreichen, werden im sechsten Abschnitt zwei Beispiele für mögliche Anwendungen vorgestellt, zum einen ein Beispiel aus dem Bereich der Gensequenzierung und zum anderen das sog. *Monkey Typing Shakespeare* Problem.

Im siebten Abschnitt wird schließlich eine zweite Möglichkeit aufgezeigt, Patternwartezeiten zu approximieren und gleichzeitig Fehlerabschätzungen zu erhalten. Das Verfahren fußt auf der so genannten *Chen-Stein-Methode*. Zur Approximation der Anzahl des näherungsweisen Auftretens des Patterns wird eine Poisson-Verteilung verwendet.

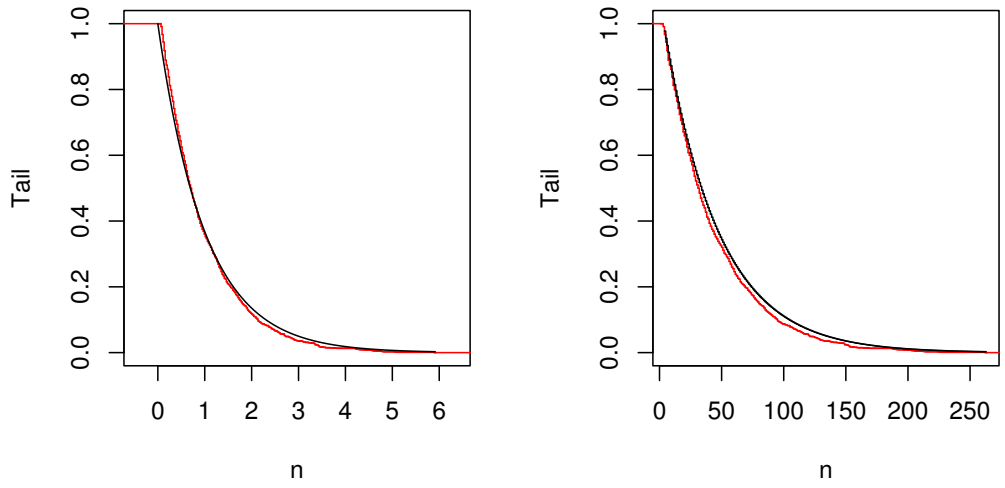

Abbildung 2.1: Simulation der Wartezeit, bis der Pattern *aggct* erstmals mit maximal einem Fehler in der Editierdistanz (vgl. Abschnitt 2.1) in einer unabhängigen und identisch über dem Alphabet $\{a, g, c, t\}$ gleichverteilten Zeichenkette auftaucht. Dargestellt sind für 1000 Stichprobenwerte dieser Wartezeitverteilung in der linken Abbildung der Tail der empirischen Verteilungsfunktion für die mit dem Stichprobenmittelwert normierten Simulationen (rot) und der Tail einer Exp(1)-Verteilung (schwarz), in der rechten Abbildung der Tail der empirischen Verteilungsfunktion für die eigentlichen Wartezeiten (rot) und der Tail einer geometrischen Verteilung, deren Erwartungswert dem Stichprobenmittelwert entspricht (schwarz). (Vgl. auch Beispiel 2.22.)

2.1 Die Editierdistanz und das *Approximate Pattern Matching Problem*

Das Ziel dieses Kapitels ist es, die Wartezeit zu approximieren, bis ein Pattern in einer zufälligen Zeichenkette erstmals „näherungsweise“ auftaucht, und so müssen wir zunächst diesen Begriff mathematisch präzisieren. Es muss uns darum gehen, den Abstand zwischen Pattern messen zu können, um dadurch zu beurteilen, wie gut ein Pattern einen anderen approximiert. Zu den folgenden Ausführungen vgl. [Gus97].

Einer der am weitesten verbreiteten Abstandsbegriffe für Pattern ist die sog. Editierdistanz oder *Edit Distance* (siehe [Wit03]). Die Grundidee ist dabei, den Abstand zwischen zwei Pattern zu messen, indem man zählt, wie viele Editieroperationen nötig sind, um einen Pattern in den anderen zu überführen. Erlaubte Editieroperationen sind dabei das Einfügen eines Buchstabens in den ersten Pattern (Abkürzung I wie „insert“), das Löschen eines Buchstabens im ersten Pattern (Abkürzung D wie „delete“) und das Austauschen eines Buchstabens des ersten Patterns gegen einen anderen (Abkürzung R wie „replace“). Bezeichnet man zusätzlich noch mit M die Nichtoperation einer Übereinstimmung zwischen zwei Buchstaben der beteiligten Wörter (M wie „match“), so lässt sich mit Hilfe dieser vier Buchstaben auf sehr kompakte Art und Weise eine Folge von Operationen angeben, die einen Pattern in einen anderen überführt.

Beispiel 2.1 Überführung des Wortes SICHERUNG in das Wort WERTUNG:

R	*D*	*D*	*D*	*M*	*M*	*I*	*M*	*M*	*M*
S	I	C	H	E	R		U	N	G
W				E	R	T	U	N	G

Wir tauschen also im Wort SICHERUNG das S gegen ein W aus, löschen I, C und H, haben dann eine Übereinstimmung bei E und R, fügen ein T ein und haben wiederum Übereinstimmungen bei U, N und G.

Definition 2.2 Eine Folge von Editieroperationen I, D und R, die einen Pattern in einen anderen überführt, heißt *Edit Transcript* für diese beiden Pattern.

Es ist klar, dass es eine Vielzahl von Edit Transcripts gibt, um einen Pattern in einen anderen zu überführen. Beispielsweise kann man jederzeit eine R-Operation durch eine Kombination aus einer D- und einer I-Operation ersetzen.

Definition 2.3 Die *Editierdistanz* oder *Edit Distance* zwischen zwei Pattern ist die minimale Anzahl von Editieroperationen – Einfügen (I), Löschen (D) und Austauschen (R) – die man benötigt, um den ersten Pattern in den zweiten zu überführen. Ein *Edit Transcript* heißt *minimal*, wenn es aus einer minimalen Anzahl von Editieroperationen besteht.

Aufgrund der Definition sieht es zunächst so aus, als würde die *Edit Distance* von der Reihenfolge der beiden beteiligten Pattern abhängen. In Wirklichkeit handelt es sich jedoch um einen symmetrischen Begriff, wie der folgende Satz zeigt:

Satz 2.4 *Gegeben sei ein endliches Alphabet* Σ. *Dann ist die* Edit Distance *eine Metrik auf der Menge* Σ^* *aller Pattern über* Σ.

Beweis: Natürlich ist die *Edit Distance* zwischen einem beliebigen Pattern über Σ und sich selbst 0, da man keine Operationen anzuwenden braucht. Ebenso ist klar, dass für zwei nicht identische Pattern die *Edit Distance* mindestens 1 beträgt, da ohne Anwendung mindestens einer der erlaubten Operationen der erste Pattern nicht in den zweiten überführt werden kann. Die Symmetrie ergibt sich leicht aus der Überlegung, dass es zu jedem *Edit Transcript*, das den ersten Pattern in den zweiten Pattern überführt, genau ein *inverses Edit Transcript* mit der gleichen Anzahl an Operationen gibt, das den zweiten Pattern in den ersten überführt. Man tausche einfach I- und D-Operationen gegeneinander aus und führe R-Operationen mit der umgekehrten Ersetzung durch. Die Dreiecksungleichung ergibt sich schließlich aus folgender Überlegung: Angenommen man hat drei Pattern sowie ein minimales *Edit Transcript*, um Pattern 1 in Pattern 2 zu überführen, und ein minimales *Edit Transcript*, um Pattern 2 in Pattern 3 zu überführen. Dann erhält man durch Hintereinanderausführen dieser beiden ein *Edit Transcript*, das Pattern 1 in Pattern 3 überführt, und dessen Anzahl an Operationen gerade der Summe der *Edit Distances* von Pattern 1 zu Pattern 2 und Pattern 2 zu Pattern 3 entspricht. Folglich ist die minimale Anzahl von Operationen, um Pattern 1 in Pattern 3 zu überführen, kleiner oder gleich dieser Summe. □

Die hier angegebene Definition der *Edit Distance* scheint für die praktische Berechnung in konkreten Situationen wenig geeignet, schließlich muss man das Minimum über die Anzahl der Operationen aller denkbaren *Edit Transcripts* bilden. Es gibt jedoch eine recht einfache Methode, die *Edit Distance* zwischen zwei Pattern algo-

rithmisch mit Hilfe einer sog. *dynamischen Programmierung* zu berechnen.

Im Folgenden bezeichnen wir die *Edit Distance* zwischen zwei Pattern $A = a_{1:N}$ und $B = b_{1:M}$ mit $\mathrm{ed}(A, B)$. Weiterhin bezeichnen wir mit $d(i, j) := \mathrm{ed}(A_{1:i}, B_{1:j})$, $0 \leq i \leq N$, $0 \leq j \leq M$, also den Abstand der Präfixe von A und B der Länge i bzw. j. (In diesem Zusammenhang sei noch einmal an die Bezeichnungen aus Abschnitt 1.8 erinnert, insbesondere daran, dass $A_{i:j}$ für $i > j$ stets den leeren Pattern $\emptyset$ bezeichnet.) Offenbar ist $\mathrm{ed}(A, B) = d(N, M)$.

Die Idee der dynamischen Programmierung ist nun, $\mathrm{ed}(A, B)$ nicht direkt zu berechnen, sondern sämtliche der Werte $d(i, j)$ mit Hilfe einer einfachen Rekursion zu bestimmen. Den Anfang dieser Rekursion bilden dabei die offensichtlichen Werte $d(i, 0) = i$, $i = 0, 1, ..., N$, $d(0, j) = j$, $j = 0, 1, ..., M$, und die Rekursionsformel ergibt sich aus elementaren Überlegungen zu

$$d(i, j) = \min\big\{d(i-1, j) + 1,\ d(i, j-1) + 1,\ d(i-1, j-1) + \delta(i, j)\big\}, \qquad (2.1)$$

wobei $\delta(i, j) = \mathbb{1}\{a_i \neq b_j\}$ gilt. Ein Beweis dieser Rekursionsformel findet sich beispielsweise in [Gus97], Abschnitt 11.3.1, S. 218f.

Die Berechnung der *Edit Distance* mit Hilfe dieser Rekursion lässt sich recht übersichtlich in einem Schema darstellen:

Beispiel 2.5 Berechnung der *Edit Distance* zwischen den Wörtern SICHERUNG und WERTUNG mit Hilfe der dynamischen Programmierung.

	∅	W	E	R	T	U	N	G
∅	0	1	2	3	4	5	6	7
S	1	1	2	3	4	5	6	7
I	2	2	2	3	4	5	6	7
C	3	3	3	3	4	5	6	7
H	4	4	4	4	4	5	6	7
E	5	5	4	5	5	5	6	7
R	6	6	5	4	5	6	6	7
U	7	7	6	5	5	5	6	7
N	8	8	7	6	6	6	5	6
G	9	9	8	7	7	7	6	**5**

In diesem Schema steht in Zeile i und Spalte j der Eintrag $d(i, j)$. Die erste Zeile und die erste Spalte dieses Schemas sind dabei fest vorgegeben, die übrigen Einträge

ergeben sich gemäß der oben angegebenen Rekursionsformel. Der Eintrag in der unteren rechten Ecke ist dann die *Edit Distance* zwischen den beiden Worten, die hier 5 beträgt. Insbesondere ist also das in Beispiel 2.1 angegeben *Edit Transcript* minimal.

Im Übrigen ist es möglich, durch einen sog. *Traceback* aus diesem Schema sämtliche *Edit Transcripts* mit minimaler Operationsanzahl zu gewinnen; siehe hierzu [Gus97], Abschnitt 11.3.3, S. 221-223. Verschiedene Verallgemeinerungen dieses Abstandsbegriffs sind möglich, siehe auch hierzu [Gus97], Kapitel III.

Damit haben wir einen Abstandsbegriff für Pattern eingeführt. Wie können wir diesen nun verwenden, um das Ereignis zu beschreiben, dass in einer zufälligen Zeichenkette ein vorgegebener Pattern zu einem bestimmten Zeitpunkt das erste Mal näherungsweise erscheint?

Sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsgrößen mit Werten in Σ (unsere zufällige Zeichenkette). Mit $A = a_1a_2...a_N$ bezeichnen wir unseren vorgegebenen Pattern über Σ. Weiter sei $0 \leq k \leq N$. Dann sagen wir, dass A in der Zeichenkette X zur Zeit n näherungsweise mit einem Abstand kleiner oder gleich k in der *Edit Distance* auftaucht, wenn es einen Suffix S von $X_{1:n}$ gibt, so dass $\mathrm{ed}(A, S) \leq k$ ist. Die uns interessierende Wartezeit T des *APMP* ist dann der Zeitpunkt, zu dem dies das erste Mal geschieht, d.h.

$$T := \min\big\{n \in \mathbb{N} \,:\, \text{Es gibt ein } m \in \{1, 2, ..., n+1\}, \text{ so dass } \mathrm{ed}(A, X_{m:n}) \leq k\big\}.$$

Bezeichnen wir für einen Pattern A und eine zufällige Zeichenkette X mit

$$\mathrm{msed}(A, X_{1:n}) := \min\big\{\mathrm{ed}(A, X_{m:n}) \,:\, m = 1, ..., n+1\big\}$$

die *Minimal Suffix Edit Distance*, so lässt sich T auch schreiben als

$$T = \min\big\{n \in \mathbb{N} \,:\, \mathrm{msed}(A, X_{1:n}) \leq k\big\}.$$

Ist klar, um welchen Pattern A und welche Zeichenkette X es sich handelt, so schreiben wir kurz $\mathrm{msed}(n)$ für $\mathrm{msed}(A, X_{1:n})$.

Es folgen zwei wichtige Aussagen im Zusammenhang mit der *Minimal Suffix Edit Distance*:

Satz 2.6 *Seien $A = a_1a_2...a_N$ ein Pattern, X eine Zeichenkette und $n \in \mathbb{N}_0$. Dann gilt*

$$\mathrm{msed}(A, X_{1:n+1}) \in \mathrm{msed}(A, X_{1:n}) + \{-1, 0, 1\}$$

Beweis: Sei $1 \leq m \leq n+1$ mit $\mathrm{msed}(A, X_{1:n}) = \mathrm{ed}(A, X_{m:n})$. Löscht man in $X_{m:n+1}$ zunächst X_{n+1} und verwendet man dann ein minimales *Edit Transcript*, das $X_{m:n}$ in A überführt, so erhält man ein *Edit Transcript*, das $X_{m:n+1}$ in A überführt und nur eine Operation mehr benötigt. Folglich gilt $\mathrm{msed}(A, X_{1:n+1}) \leq \mathrm{msed}(A, X_{1:n}) + 1$. Analog zeigt man $\mathrm{msed}(A, X_{1:n}) \leq \mathrm{msed}(A, X_{1:n+1}) + 1$. □

Satz 2.7 *Seien $A = a_1a_2...a_N$ ein Pattern, X eine Zeichenkette und $k, n \in \mathbb{N}_0$. Dann gilt für alle $m \leq n$ mit $m < N - k$ oder $m > N + k$*

$$\mathrm{ed}(A, X_{n-m+1:n}) > k.$$

Beweis: Ist $m < N - k$, so ist die Länge von $X_{n-m+1:n}$ kleiner als $N - k$. Um $X_{n-m+1:n}$ in A zu überführen, sind also mindestens $k+1$ Einfügungen nötig, also ist die *Edit Distance* größer als k. Ist $m > N + k$, so ist die Länge von $X_{n-m+1:n}$ größer als $N + k$, mithin sind nun zumindest $k+1$ Löschungen notwendig, um $X_{n-m+1:n}$ in A zu überführen und die *Edit Distance* ist wiederum größer als k. □

Wegen $\mathrm{msed}(A, X_{1:n}) \leq \mathrm{ed}(A, \emptyset) = N$ gilt für alle $n \geq 2N$

$$\mathrm{msed}(A, X_{1:n}) = \mathrm{msed}(A, X_{n-2N+1:n}),$$

und $\mathrm{msed}(A, X_{1:n}) = k$ genau dann, wenn es ein $m \in \{N - k, ..., N + k\}$ gibt, so dass $\mathrm{ed}(A, X_{n-m+1:n}) = k$ gilt. Um also in einer konkreten Situation die Wartezeit T zu bestimmen, reicht es aus, zu jedem Zeitpunkt $n \in \mathbb{N}$ das Teilstück $X_{n-2N+1:n}$ zu betrachten. Den zugehörigen Prozess $(Y_n)_{n\in\mathbb{N}}$ mit $Y_0 = \emptyset$ (leerer Pattern) und

$$Y_n := \begin{cases} X_{1:n} & , n < 2N, \\ X_{n-2N+1:n} & , n \geq 2N, \end{cases}$$

bezeichnet man auch als *Snake Chain* (der Länge $2N$) zur Zeichenkette X. (Vgl. [Bré99], S. 90.)

Im nun folgenden Abschnitt wollen wir die Methoden aus Kapitel 1 verwenden, um unter speziellen Bedingungen auch für das *APMP* asymptotische Exponentialität nachzuweisen.

2.2 Asymptotische Exponentialität für das *Approximate Pattern Matching Problem*

Bis zu diesem Zeitpunkt haben wir noch keine Voraussetzungen über die Beschaffenheit der zufälligen Zeichenkette X getroffen. Die beiden Standardmodelle in der mathematischen Biologie zur Modellierung einer DNA-Sequenz sind zum einen eine Folge von unabhängigen und identisch verteilten Zufallsgrößen, zum anderen eine Markov-Kette. Im Hinblick auf Anwendungen in diesem Gebiet werden wir hier und im Folgenden (wie schon in Abschnitt 1.8) diese beiden Modelle betrachten.

Sei also Σ ein endliches Alphabet mit $\#\Sigma \geq 2$. Im ersten Modell ist X eine unabhängige und identisch verteilte Folge von Zufallsgrößen $(X_n)_{n\in\mathbb{N}}$ über Σ mit Verteilung $p_\sigma := P(X_1 = \sigma)$ für alle $\sigma \in \Sigma$, wobei wir, um triviale Fälle zu vermeiden, voraussetzen, dass $p_\sigma > 0$ für alle $\sigma \in \Sigma$ gilt. Im zweiten Modell ist X eine homogene, aperiodische und positiv rekurrente Markov-Kette $(X_n)_{n\in\mathbb{N}_0}$ mit Zustandsraum Σ. Im diesem Fall sei die zugehörige Übergangsmatrix $P := \big(p(\sigma_i,\sigma_j)\big)_{\sigma_i,\sigma_j\in\Sigma}$ mit $p(\sigma_i,\sigma_j) := P(X_1 = \sigma_j | X_0 = \sigma_i)$, die Startverteilung bezeichnen wir mit $\nu := \big(\nu(\sigma)\big)_{\sigma\in\Sigma}$, $\nu(\sigma) := P(X_0 = \sigma)$, und die stationäre Verteilung mit $\pi := \big(\pi(\sigma)\big)_{\sigma\in\Sigma}$.

Ziel dieses Abschnitte soll es nun sein, mit den im ersten Kapitel dieser Arbeit bereitgestellten Methoden unter speziellen Voraussetzungen asymptotische Exponentialität für das *APMP* nachzuweisen. Der Einfachheit halber beschränken wir unsere Betrachtungen in diesem Abschnitt auf den Fall, dass $X = (X_n)_{n\in\mathbb{N}}$ eine unabhängige und identisch verteilte Zeichenkette über Σ ist.

Es sei $A = a_1a_2... \in \Sigma^\infty$ ein unendlich langer Pattern über Σ und $A_N := a_{1:N}$ für jedes $N \in \mathbb{N}$. Außerdem sei $(k_N)_{N\in\mathbb{N}}$ mit $0 \leq k_N \leq N$ eine Folge von Editierdistanzen. Ziel ist der Nachweis der asymptotischen Exponentialität für T_N/ET_N mit

$$T_N := \inf\{n \in \mathbb{N} : \operatorname{msed}(A_N, X_{1:n}) \leq k_N\}.$$

Zum Nachweis wollen wir Theorem 1.9 des ersten Kapitels verwenden. Als Folge von Regenerationszeiten für X wählen wir hier für jedes $N \in \mathbb{N}$ das überlappungsfreie Auftreten eines speziell gewählten Patterns S_N in X, wobei wir $|S_N|$ als monoton

wachsend und $|S_N| \leq N - k_N$ voraussetzen. Es seien also:

$$\begin{aligned}
\sigma_{N,0} &:= \inf\{n \geq |S_N| : X_{n-|S_N|+1:n} = S_N\}, \\
\sigma_{N,l} &:= \inf\{n \geq \tau_{N,l-1} + |S_N| : X_{n-|S_N|+1:n} = S_N\},\ l \geq 1, \\
\gamma_{N,l} &:= \sigma_{N,l} - \sigma_{N,l-1},\ l \geq 1.
\end{aligned}$$

Aus den Voraussetzungen für die Zeichenkette X folgt, dass es sich bei den Zeitpunkten $(\sigma_{N,l})_{l\geq 0}$ tatsächlich um Regenerationszeitpunkte handelt.

Nach Satz 2.7 ist $\mathrm{ed}(A_N, X_{n-m+1:n}) > k_N$, falls $m < N - k_N$ oder $m > N + k_N$ gilt. Anders ausgedrückt:

$$\mathrm{msed}(A_N, X_{1:n}) \leq k_N \ \Leftrightarrow\ \exists m \in \{N - k_N, ..., N + k_N\} : \mathrm{ed}(A_N, X_{n-m+1:n}) \leq k_N.$$

Daher definieren wir für $l \geq 1$

$$\begin{aligned}
E_{N,l} := \big\{&\exists n \in \{\sigma_{N,l-1} + 1, ..., \sigma_{N,l}\} \exists m \in \{N - k_N, ..., N + k_N\} : \\
&\mathrm{ed}(A_N, X_{n-m+1:n}) \leq k_N\big\},
\end{aligned}$$

d.h. $E_{N,l}$ ist das Ereignis, dass der Pattern A_N zwischen den Erneuerungszeitpunkten $\sigma_{N,l-1}$ und $\sigma_{N,l}$ mit einem Fehler kleiner oder gleich k_N erscheint.

Seien schließlich $M_N := \inf\{l \in \mathbb{N} : E_{N,l} \text{ tritt ein}\}$ und

$$\tilde{T}_N := \inf\{n \geq \sigma_{N,0} : \mathrm{msed}(A_N, X_{1:n}) \leq k_N\}.$$

Dann gilt:

$$\sigma_{N,0} + \sum_{l=1}^{M_N - 1} \gamma_{N,l} \leq \tilde{T}_N \leq \sigma_{N,0} + \sum_{l=1}^{M_N} \gamma_{N,l}.$$

Bemerkung 2.8 $\sigma_{N,0}$ hat zwar die gleiche Verteilung wie $\gamma_{N,l}$ für $l \geq 1$, definiert man allerdings

$$E_{N,0} := \big\{\exists n \in \{1, ..., \sigma_{N,0}\} \exists m \in \{N - k_N, ..., N + k_N\} \ :\ \mathrm{ed}(A_N, X_{n-m+1:n}) \leq k_N\big\},$$

so ist $P(E_{N,0}) \neq P(E_{N,l})$ für $l \geq 1$. Dies ist der Grund, weshalb wir an dieser Stelle die asymptotische Exponentialität zunächst nur für $\tilde{T}_N / E\tilde{T}_N$ beweisen. Anschließend werden wir jedoch sehen, dass sich hieraus unmittelbar dieselbe Aussage für T_N / ET_N ergibt.

Um nun asymptotische Exponentialität für $\tilde{T}_N/E\tilde{T}_N$ mit Theorem 1.9 zu folgern, müssen folgende Bedingungen erfüllt sein:

(i) Für alle $N, l \in \mathbb{N}$ ist $E_{N,l}$ messbar bzgl. $\{X_{\sigma_{N,l-1}+1}, ..., X_{\sigma_{N,l}}\}$ und $P(E_{N,l})$ ist in Abhängigkeit von l konstant.

(ii) $q_N := P(E_{N,1}) \to 0$ mit $N \to \infty$.

(iii) $(\gamma_{N,1}/E\gamma_{N,1})_{N\in\mathbb{N}}$ ist gleichgradig integrierbar.

(iv) $\sigma_{N,0}/E\tilde{T}_N \xrightarrow{\mathrm{P}} 0$.

Damit die erste dieser Bedingungen erfüllt ist, stellen wir an die Folge der (Regenerations-)Pattern S_N folgende Forderung:

(B1) Ist S_N ein Subpattern eines beliebigen Pattern B (kurz: $S_N \subset B$), so ist stets $\mathrm{ed}(A_N, B) > k_N$.

Als unmittelbare Folgerung ergibt sich:

$$\begin{aligned} E_{N,l} \;=\; \big\{&\exists n \in \{\sigma_{N,l-1}+1, ..., \sigma_{N,l}\} \exists m \in \{N-k_N, ..., N+k_N\} : \\ &\mathrm{ed}(A_N, X_{n-m+1:n}) \le k_N \text{ und } n-m+1 > \tau_{N,l-1} - |S_N| + 1\big\}. \end{aligned}$$

Sei für $s \ge |S_N|$

$$\begin{aligned} C_s \;:=\; \big\{&x_{1:s} \in \Sigma^s \,:\, x_{s-|S_N|+1:s} = S_N \text{ und es ex. ein Subpattern } B \\ &\text{von } S_N \circ x_{1:s} \text{ mit } \mathrm{ed}(A_N, B) \le k_N\big\}, \end{aligned}$$

wobei für zwei Pattern $A = a_1...a_n$ und $B = b_1...b_m$ die Verknüpfung $A \circ B$ den Pattern $a_1...a_n b_1...b_m$ bezeichnet. Dann gilt für alle $l \ge 1$: $E_{N,l}$ tritt ein $\Leftrightarrow$ $\big(X_{\tau_{N,l-1}+1}, ..., X_{\tau_{N,l}}\big) \in C_{\sigma_{N,l}}$. Damit ist (i) gezeigt.

Im Zusammenhang mit Bedingung (ii) fordern wir für die Folge der (Regenerations-)Pattern S_N und die Folge der Editierdistanzen k_N des Weiteren:

(B2) Es gelte $\lim_{N\to\infty} P(\sigma_{N,0} > N - k_N - |S_N|) = 0$.

Ist diese Forderung erfüllt, so ergibt sich unmittelbar Bedingung (ii), was man wie folgt sieht: Mit

$$\begin{aligned} \sigma_{N,0} &= \inf\big\{n \ge |S_N| \,:\, X_{n-|S_N|+1:n} = S_N\big\} \text{ und} \\ \sigma_N &:= \inf\big\{n \ge 1 \,:\, \mathrm{msed}(A_N, S_N \circ X_{1:n}) \le k_N\big\} \end{aligned}$$

gilt

$$q_N = P(\sigma_N \leq \sigma_{N,0}).$$

Wegen Satz 2.7 gilt nun einerseits

$$\mathrm{ed}(A_N, X_{n-m+1:n}) > k_N, \text{ falls } m < N - k_N,$$

wegen (B1) andererseits

$$\mathrm{ed}(A_N, X_{n-m+1:n}) > k_N, \text{ falls } S_N \text{ ein Subpattern von } X_{n-m+1:n} \text{ ist},$$

also folgt

$$\sigma_N > N - k_N - |S_N|.$$

Daher ist $q_N \leq P(\sigma_{N,0} > N - k_N - |S_N|)$. (B2) liefert nun für die rechte Seite mit $N \to \infty$ den Grenzwert 0.

Nun zur Bedingung (iii), der gleichgradigen Integrierbarkeit von $(\gamma_{N,1}/E\gamma_{N,1})_{N\in\mathbb{N}}$.

Angenommen $|S_N|$ bleibt beschränkt, d.h. es existiert ein $K \in \mathbb{R}$ mit $|S_N| \leq K$ für alle $N \in \mathbb{N}$. Nun gibt es aber nur endlich viele Pattern S_N über Σ mit $|S_N| \leq K$. Daher liegen der betrachteten Familie von normierten Wartezeiten nur endlich viele verschiedene Verteilungen mit endlichem Erwartungswert zu Grunde und die gleichgradige Integrierbarkeit ergibt sich trivialerweise.

Daher gelte $|S_N| \to \infty$ mit $N \to \infty$. Nach Satz 1.27 gilt dann mit $N \to \infty$

$$\gamma_{N,1}/E\gamma_{N,1} \xrightarrow{\mathrm{d}} \mathrm{Exp}(1).$$

Die gleichgradige Integrierbarkeit ergibt sich nun unmittelbar aus Lemma 1.18, wobei die Konvergenz der Momente trivialerweise erfüllt ist (alle beteiligten Größen haben Erwartungswert 1).

Schließlich der Nachweis der vierten Bedingung. Definieren wir $\hat{T}_N := \tilde{T}_N - \sigma_{N,0}$, so folgt aus den bisher gezeigten Bedingungen (i) bis (iii) mit Theorem 1.9, dass $\hat{T}_N/E\hat{T}_N \xrightarrow{\mathrm{d}} \mathrm{Exp}(1)$ gilt und $(q_N E\hat{T}_N)/E\sigma_{N,0}$ mit $N \to \infty$ gegen 1 strebt. Wegen $\lim_{N\to\infty} q_N = 0$ ergibt sich $E\sigma_{N,0}/E\hat{T}_N \to 0$. Wegen $E\hat{T}_N = E\tilde{T}_N - E\sigma_{N,0}$ folgt dann aber auch $E\sigma_{N,0}/E\tilde{T}_N \to 0$, woraus sich mit der Markovschen Ungleichung sofort $\sigma_{N,0}/E\tilde{T}_N \xrightarrow{\mathrm{P}} 0$ ergibt.

Damit haben wir die asymptotische Exponentialität von $\tilde{T}_N/E\tilde{T}_N$ nachgewiesen. Unser eigentliches Ziel war jedoch der Nachweis dieser Aussage für T_N/ET_N. Zur

Erinnerung:

$$\begin{aligned} T_N &= \inf\{n \geq 1 \quad\quad : \mathrm{msed}(A_N, X_{1:n}) \leq k_N\}, \\ \tilde{T}_N &= \inf\{n \geq \sigma_{N,0} \; : \mathrm{msed}(A_N, X_{1:n}) \leq k_N\}. \end{aligned}$$

Wegen $\mathrm{msed}(A_N, X_{1:n}) > k_N$ für $n < N - k_N$ ist

$$T_N = \inf\{n \geq N - k_N \; : \mathrm{msed}(A_N, X_{1:n}) \leq k_N\}.$$

Daher gilt im Fall $\sigma_{N,0} \leq N - k_N$: $T_N = \tilde{T}_N$. Mit (B2) ergibt sich hieraus

$$P(T_N = \tilde{T}_N) \geq P(\sigma_{N,0} \leq N - k_N) \geq P(\sigma_{N,0} \leq N - k_N - |S_N|) \stackrel{N\to\infty}{\longrightarrow} 1.$$

Weiter gilt offenbar

$$ET_N \leq E\tilde{T}_N \leq E\sigma_{N,0} + ET_N.$$

Die erste Ungleichung ergibt sich unmittelbar aus der Definition von T_N und $\tilde{T}_N$, die zweite erhält man durch Betrachtung des Post-$\sigma_{N,0}$-Prozesses. Wegen $E\sigma_{N,0}/E\tilde{T}_N \to 0$ ergibt sich daraus

$$\lim_{N\to\infty} E\tilde{T}_N/ET_N = 1.$$

Nun ist

$$\left|\frac{T_N}{ET_N} - \frac{\tilde{T}_N}{E\tilde{T}_N}\right| = \mathbb{1}\{T_N = \tilde{T}_N\}\frac{\tilde{T}_N}{E\tilde{T}_N}\left|\frac{E\tilde{T}_N}{ET_N} - 1\right| + \mathbb{1}\{T_N \neq \tilde{T}_N\}\frac{\tilde{T}_N}{E\tilde{T}_N}\left|\frac{E\tilde{T}_N}{ET_N}\frac{T_N}{\tilde{T}_N} - 1\right|.$$

Wegen $\mathbb{1}\{T_N = \tilde{T}_N\} \stackrel{\text{f.s.}}{\longrightarrow} 1$, $\tilde{T}_N/E\tilde{T}_N \stackrel{\text{d}}{\longrightarrow} \mathrm{Exp}(1)$, $E\tilde{T}_N/ET_N \to 1$ und $T_N/\tilde{T}_N \in [0,1]$ für alle $N \in \mathbb{N}$ ergibt sich

$$\left|\frac{T_N}{ET_N} - \frac{\tilde{T}_N}{E\tilde{T}_N}\right| \stackrel{\text{P}}{\longrightarrow} 0.$$

Mit [Bil68], Theorem 4.1, S. 25 folgt hieraus dann

$$\frac{T_N}{ET_N} \stackrel{\text{d}}{\longrightarrow} \mathrm{Exp}(1).$$

Wir fassen unsere Ergebnisse in einem Theorem zusammen:

Theorem 2.9 *Es sei Σ ein endliches Alphabet, $X = (X_n)_{n\in\mathbb{N}}$ eine unabhängige und identisch verteilte Zeichenkette über Σ, $A = a_1a_2... \in \Sigma^\infty$ ein unendlich langer Pattern über Σ und $A_N := a_{1:N}$ für jedes $N \in \mathbb{N}$. Weiter sei $(k_N)_{N\in\mathbb{N}}$ eine Folge natürlicher Zahlen mit $0 \leq k_N \leq N$.*

Wir definieren

$$T_N := \inf\big\{n \geq 1 \,:\, \mathrm{msed}(A_N, X_{1:n}) \leq k_N\big\}.$$

Schließlich sei $(S_N)_{N\in\mathbb{N}}$ eine Folge von Pattern über Σ mit folgenden Eigenschaften:

(i) Für alle (hinreichend großen) $N \in \mathbb{N}$ ist $|S_N|$ monoton wachsend und es gilt $|S_N| \leq N - k_N$.

(ii) Für alle (hinreichend großen) $N \in \mathbb{N}$ gilt: Ist $S_N \subset B$, so ist $\mathrm{ed}(A_N, B) > k_N$.

(iii) Für $\sigma_{N,0} := \inf\{n \geq |S_N| : X_{n-|S_N|+1:n} = S_N\}$ gilt $\lim_{N\to\infty} P(\sigma_{N,0} > N - k_N - |S_N|) = 0$.

Dann folgt mit $N \to \infty$

$$\frac{T_N}{ET_N} \xrightarrow{d} \mathrm{Exp}(1).$$

Wir wollen nun zeigen, dass die Voraussetzungen dieses Satzes insbesondere erfüllt sind, wenn es für jedes $N \in \mathbb{N}$ einen Buchstaben $\sigma_N \in \Sigma$ gibt, der in A_N „gleichmäßig hinreichend selten“ vorkommt. Im Folgenden bezeichne

$$m := \big(\min_{\sigma\in\Sigma} p_\sigma\big)^{-1}.$$

Satz 2.10 *Es sei Σ ein endliches Alphabet, $X = (X_n)_{n\in\mathbb{N}}$ eine unabhängige und identisch verteilte Zeichenkette über Σ, $A = a_1a_2... \in \Sigma^\infty$ ein unendlich langer Pattern über Σ und $A_N := a_{1:N}$ für jedes $N \in \mathbb{N}$. Weiter sei $(k_N)_{N\in\mathbb{N}}$ eine Folge natürlicher Zahlen mit $0 \leq k_N \leq N$.*

Wir definieren

$$T_N := \inf\big\{n \geq 1 \,:\, \mathrm{msed}(A_N, X_{1:n}) \leq k_N\big\}.$$

Schließlich sei $(L_N)_{N\in\mathbb{N}}$ eine monoton wachsende Folge natürlicher Zahlen mit folgenden Eigenschaften:

(i) Für alle $N \in \mathbb{N}$ gilt $k_N < L_N \leq N - k_N$

(ii) Es gibt ein $N_0 \in \mathbb{N}$*, so dass für alle* $N \geq N_0$ *ein Buchstabe* $\sigma_N \in \Sigma$ *mit der Eigenschaft existiert, dass er in jedem Subpattern von* A_N *der Länge* L_N *weniger als* $(L_N - k_N)$*-mal vorkommt.*

(iii) Es gibt ein $N_1 \in \mathbb{N}$ *und ein* $\alpha \in (0,1)$*, so dass* $L_N \leq \alpha \log_m N$ *für alle* $N \geq N_1$ *gilt.*

Dann folgt mit $N \to \infty$

$$\frac{T_N}{ET_N} \xrightarrow{d} \text{Exp}(1).$$

Beweis: O.B.d.A. ist $N_1 = N_0$. Wir beweisen diesen Satz mit Theorem 2.9. Für jedes $N \in \mathbb{N}$ sei S_N ein σ_N-Run der Länge L_N.

Die erste Bedingung aus Theorem 2.9 ist dann automatisch erfüllt.

Als nächstes beweisen wir, dass aus $S_N \subset B$ stets $\text{ed}(A_N, B) > k_N$ folgt. Sei dazu B ein beliebiger Pattern mit $S_N \subset B$. Dann gibt es zu jedem Edit Transcript, dass B in A_N überführt einen (nicht eindeutig festgelegten) Subpattern C von A_N, so dass durch dieses Edit Transcript gerade S_N in C überführt wird. Sei $c := |C|$. Wir unterscheiden drei Fälle:

(1) $c < L_N - k_N$ oder $c > L_N + k_N$. Dann sind mehr als k_N D- oder I-Operationen nötig, um S_N in C zu überführen, also gilt $\text{ed}(S_N, C) \geq k_N + 1$.

(2) $L_N - k_N \leq c \leq L_N$. Dann sind mindestens $L_N - c$ D-Operationen nötig, um S_N in C zu überführen. Außerdem kommt der Buchstabe σ_N in C höchstens $(L_N - k_N - 1)$-mal vor, und da jede Editieroperation die Anzahl der σ_N's in einem Pattern um höchstens 1 verändert, sind also zusätzlich noch mindestens $c - (L_N - k_N - 1)$ weitere Operationen nötig, um S_N in C zu überführen. Insgesamt ist also auch in diesem Fall $\text{ed}(S_N, C) \geq L_N - c + c - (L_N - k_N - 1) = k_N + 1$.

(3) $L_N < c \leq L_N + k_N$. In diesem Fall betrachten wir das inverse Edit Transcript, dass C in S_N überführt. Dieses umfasst auf jeden Fall $c - L_N$ D-Operationen. Außerdem kommt σ_N in jedem Subpattern der Länge L_N von A_N maximal $(L_N - k_N - 1)$-mal vor. Löscht man nun in A_N einen Buchstaben, so kann sich diese Zahl um maximal 1 erhöhen. Löscht man also $c - L_N$ Buchstaben in $C \subset A_N$, so kommt in dem resultierenden Pattern der Länge L_N der Buchstabe σ_N maximal $(L_N - k_N - 1) + (c - L_N) = (c - k_N - 1)$-mal vor. Also sind noch mindestens $L_N - (c - k_N - 1)$ weiter Operationen notwendig. Damit gilt auch in diesem letzten Fall $\text{ed}(S_N, C) \geq (c - L_N) + L_N - (c - k_N - 1) = k_N + 1$.

Damit ist $\mathrm{ed}(C, S_N) > k_N$ für jedes $C \subset A_N$ und damit auch $\mathrm{ed}(A_N, B) > k_N$ für jedes B mit $S_N \subset B$.

Schließlich zur dritten Bedingung aus Theorem 2.9. Bekanntlich gilt

$$E\tau_{N,0} = \Phi_{S_N}(1)/\mathbb{P}(S_N).$$

Wegen $\Phi_{S_N}(1) \leq \sum_{k=0}^{\infty}(\max_{\sigma\in\Sigma} p_\sigma)^k$ existiert ein $K \in \mathbb{R}$ mit $E\tau_{N,0} \leq K/\mathbb{P}(S_N)$ für alle $N \in \mathbb{N}$. Nun gilt offenbar $\mathbb{P}(S_N) \geq m^{-L_N}$, und mit der Markovschen Ungleichung folgt für hinreichend großen N

$$\begin{aligned} P(\sigma_{N,0} > N - k_N - |S_N|) &\leq \frac{E\sigma_{N,0}}{N - k_N - L_N} \leq \frac{K/\mathbb{P}(S_N)}{N - k_N - L_N} \\ &\leq \frac{Km^{L_N}}{N - k_N - L_N} \leq \frac{Km^{\alpha \log_m N}}{N - k_N - L_N} \\ &\leq \frac{KN^\alpha}{N - 2L_N} \leq \frac{KN^\alpha}{N - 2\alpha \log_m N} \to 0. \end{aligned}$$

Damit folgt die asymptotische Exponentialität von T_N/ET_N aus Theorem 2.9. □

Als Spezialfall ergibt sich

Lemma 2.11 *Es sei Σ ein endliches Alphabet, $X = (X_n)_{n\in\mathbb{N}}$ eine unabhängige und identisch verteilte Zeichenkette über Σ, $A = a_1a_2... \in \Sigma^\infty$ ein unendlich langer Pattern über Σ und $A_N := a_{1:N}$ für jedes $N \in \mathbb{N}$. Weiter sei $(k_N)_{N\in\mathbb{N}}$ eine monoton wachsende Folge natürlicher Zahlen mit $0 \leq k_N \leq N$.*

Wir definieren

$$T_N := \inf\left\{n \geq 1 : \mathrm{msed}(A_N, X_{1:n}) \leq k_N\right\}.$$

Schließlich gelte:

(i) Es existiert ein Buchstabe $\sigma^* \in \Sigma$, der in A nicht vorkommt

(ii) Es gibt ein $N_0 \in \mathbb{N}$ und ein $\alpha \in (0, 1)$, so dass $k_N + 1 \leq \alpha \log_m N$ für alle $N \geq N_0$ gilt.

Dann folgt mit $N \to \infty$

$$\frac{T_N}{ET_N} \xrightarrow{d} \mathrm{Exp}(1).$$

Bemerkung 2.12 Bedingung (ii) ist insbesondere für konstante bzw. beschränkte Folgen $(k_N)_{N\in\mathbb{N}}$ erfüllt.

Beweis: Verwendet Satz 2.10 mit $L_N := k_N + 1$ und $\sigma_N \equiv \sigma^*$ für alle $N \in \mathbb{N}$. □

Das nun folgende Beispiel zeigt eine Anwendung dieses Satzes bei der Approximation von Wartezeitverteilungen:

Beispiel 2.13 Wir wollen hier eine Fragestellung betrachten, die man in vielen einführenden Werken zur Stochastik finden kann und die unter dem Stichwort *Monkey Typing Shakespeare* Eingang in die Literatur gefunden hat. (Vgl. [Wil91], Abschnitt 4.9, S. 44f.)

Die am häufigsten verwendete Formulierung des *Monkey Typing Shakespeare* Phänomens lautet etwa wie folgt: "Wenn ein Affe unendlich lange auf einer Schreibmaschine herumtippt, so wird er irgendwann einmal die gesammelten Werke von Shakespeare verfasst haben." Im englischsprachigen Raum wird die Redewendung *Monkey Typing Shakespeare* auch häufig als Bild gebraucht, wenn man ausdrücken möchte, dass ein bestimmtes Ereignis, so unwahrscheinlich es auch sein mag, doch irgendwann eintreten kann.

Unter der Modellannahme, dass der Affe durch sein Herumtippen auf der Schreibmaschinentastatur eine unabhängige Zeichenkette erzeugt, ist diese Aussage tatsächlich richtig; es ist nicht weiter schwierig zu beweisen, dass die Wahrscheinlichkeit des betrachteten Ereignisses 1 beträgt. Dazu bringen wir die Werke von Shakespeare in eine (wie auch immer geartete) Reihenfolge, so dass wir einen Textpattern der Gesamtlänge N erhalten. F sei das Ereignis, dass unser Affe durch sein zufälliges Herumtippen diesen Textpattern niemals produziert. Wir unterteilen nun die Folge der vom Affen getippten Zeichen in aufeinander folgende Blöcke der Länge N und bezeichnen mit G_i, $i \in \mathbb{N}$, das Ereignis, dass der i-te dieser Blöcke nicht dem gesuchten Textpattern entspricht. Offensichtlich sind sämtliche Ereignisse G_i unabhängig und haben dieselbe von 1 verschiedene Wahrscheinlichkeit. Weiter bezeichnen wir mit $G := \bigcap_{i\in\mathbb{N}} G_i$ das Ereignis, dass unser Affe niemals in einem der Blöcke die Werke Shakespeares verfasst. Dann gilt $P(G) \leq \prod_{i\in\mathbb{N}} P(G_i) = 0$. Nun ist aber F in G enthalten und daher auch $P(F) = 0$, so dass schließlich die Wahrscheinlichkeit des Gegenereignisses, dass der Affe doch irgendwann einmal Shakespeares Werke verfasst, 1 beträgt. Man beachte dabei, dass dieses Resultat unabhängig davon ist,

welche Reihenfolge man bei Shakespeares Werken zu Grunde legt, ebenso wie von der angenommenen Verteilung, mit der der Affe die Tasten der Schreibmaschine drückt, solange nur jeder Buchstabe aus Shakespeares Werken mit positiver Wahrscheinlichkeit vorkommt.

In der Praxis kann man den Affen allerdings immer nur für eine begrenzte Zeitspanne bei seiner Arbeit beobachten, und so wollen wir hier im Gegensatz zu dem obigen theoretischen Resultat unter dem Begriff *Monkey Typing Shakespeare* die realitätsbezogenere Frage diskutieren, wie lange der Affe braucht, um zumindest ein berühmtes Shakespeare-Zitat, beispielsweise „*To Be Or Not To Be*“ oder eine hierzu ähnliche Buchstabenfolge zu tippen. In dieser Situation legt Lemma 2.11 die Vermutung nahe, dass die Wartezeit, bis der Pattern $TOBEORNOTTOBE$ in einer unabhängigen und identisch verteilten Zeichenkette über den Alphabet der 26 Buchstaben $\{A, B, ..., Z\}$ erstmals mit maximal einem Fehler in der *Edit Distance* erscheint, bei Normierung mit dem zugehörigen Erwartungswert näherungsweise Exp(1)-verteilt sein wird. Allerdings ist keine Aussage über die Güte dieser Approximation möglich. Wir werden später wieder auf dieses Beispiel zurückkommen.

Ähnliche Fragestellungen wie die des vorangegangenen Beispiels tauchen auch im Zusammenhang mit Gensequenzierungen in der mathematischen Biologie auf. Die theoretischen Antworten, die die hier vorgestellten Resultate auf solche Fragestellungen zu geben vermögen, sind für den an konkreten Nährungen der Wartezeitverteilung für endliche Pattern interessierten Anwender unbefriedigend, solange sie nicht durch entsprechende Fehlerabschätzungen erweitert werden können. Beispielsweise ist es sonst nicht möglich, Quantile der Wartezeitverteilung abzuschätzen.

Dieser Missstand ist der Antrieb für die folgenden Bemühungen, Wartezeitapproximationen für das *APMP* unter möglichst allgemeinen Voraussetzungen inklusive Fehlerabschätzungen herzuleiten. Ausgangspunkt unserer Untersuchungen ist dabei die Interpretation der entsprechenden Wartezeit als Eintrittszeit für Markov-Ketten. Im folgenden Abschnitt werden wir deshalb ein Resultat herleiten, das es ermöglicht, solche Eintrittszeitverteilungen inklusive einer entsprechenden Fehlerabschätzung zu approximieren.

2.3 Eintrittszeitapproximation durch eine geometrische Verteilung

In diesem Abschnitt behandeln wir in allgemeiner Form das Problem, Eintrittszeitverteilungen in Markov-Ketten zu approximieren und gleichzeitig Fehlerabschätzungen für die gewonnene Näherung zu erhalten, mit denen sich deren Güte beurteilen lässt. In den folgenden Abschnitten wenden wir die erzielten Resultate dann auf das *APMP* an.

Es sei $X = (X_n)_{n\in\mathbb{N}_0}$ eine homogene, aperiodische und irreduzible Markov-Kette auf einem endlichen Zustandsraum $E = \{i, j, k, ...\}$ mit Übergangsmatrix $P = \left(p_{ij}\right)_{i,j\in E}$ und stationärer Verteilung $\pi = (\pi_i)_{i\in E}$. Für eine Teilmenge $\emptyset \neq D \subset E$ sei

$$T_D := \inf\{n \in \mathbb{N}_0 : X_n \in D\}$$

die Wartezeit, bis X in D eintritt.

Wenn wir die Wartezeit T_D durch eine andere Verteilung approximieren und Fehlerschranken für die Approximation angeben wollen, so müssen wir als erstes einen Abstandsbegriff für Verteilungen einführen. Ein üblicher Abstandsbegriff ist in diesem Zusammenhang die sog. *Total Variation Distance*.

Definition 2.14 Es seien $(\Omega, \mathcal{A})$ ein messbarer Raum und P, Q zwei Wahrscheinlichkeitsmaße auf $\mathcal{A}$. Dann ist der *Totalvariationsabstand* oder auch die *Total Variation Distance* zwischen P und Q definiert durch

$$\mathrm{d_{TV}}(P, Q) := \sup_{A\in\mathcal{A}} |P(A) - Q(A)|.$$

Für abzählbares Ω gilt $\mathrm{d_{TV}}(P, Q) = \frac{1}{2}\sum_{\omega\in\Omega} |P(\{\omega\}) - Q(\{\omega\})|$. Weitere wichtige Eigenschaften dieses Abstandsbegriffs findet man in Anhang B.

Es ist klar, dass die Exponentialverteilung bei Verwendung des Totalvariationsabstands nun nicht mehr geeignet ist, um (diskrete) Eintrittszeitverteilungen in Markov-Ketten zu approximieren, da der Totalvariationsabstand zwischen einer diskreten und einer absolut-stetigen Verteilung wie der Exponentialverteilung stets 1 beträgt. An ihre Stelle tritt hier, quasi als diskretes Analogon, die geometrische Verteilung.

Es lässt sich aber sogar noch mehr zeigen: Betrachtet man eine ganze Folge $(D_N)_{N\in\mathbb{N}}$ von Zielmengen für die Markov-Kette X oder noch allgemeiner eine ganze Folge von Markov-Ketten $(X_N)_{N\in\mathbb{N}}$ mit zugehörigen Zielmengen $(D_N)_{N\in\mathbb{N}}$, so folgt aus den Tatsachen, dass $\lim_{N\to\infty} ET_{D_N} = \infty$ gilt und dass der Totalvariationsabstand der Wartezeitverteilung T_{D_N} zu einer geometrischen Verteilung mit gleichem Erwartungswert für $N \to \infty$ gegen 0 strebt, sofort die asymptotische Exponentialität für T_{D_N}, wie das folgende Lemma zeigt:

Lemma 2.15 *Für jedes $N \in \mathbb{N}$ habe T_N eine diskrete Verteilung auf $\mathbb{N}$ und X_N eine geometrische Verteilung auf $\mathbb{N}$ mit $EX_N = ET_N$. Gilt dann*

$$(i)\ \lim_{N\to\infty} ET_N = \infty\,, \quad (ii)\ \lim_{N\to\infty} \mathrm{d_{TV}}\big(\mathcal{L}(T_N), \mathcal{L}(X_N)\big) = 0\,,$$

so folgt $T_N/ET_N \xrightarrow{d}$ Exp(1).

Beweis: Für alle $x > 0$ gilt

$$P(T_N/ET_N \le x) = \big(P(T_N/ET_N \le x) - P(X_N/ET_N \le x)\big) + P(X_N/ET_N \le x).$$

Die wesentliche Idee ist nun, dass wegen Voraussetzung (i) $X_N/ET_N \xrightarrow{\mathrm{d}}$ Exp(1) gilt. Daher strebt der zweite Summand in der obigen Darstellung gegen $1-\exp(-x)$. Für den ersten Summanden folgt mit Hilfe von Voraussetzung (ii)

$$\begin{aligned} \big|P(T_N/ET_N \le x) - P(X_N/ET_N \le x)\big| &\le \sum_{k=1}^{\lfloor xET_N\rfloor} \big|P(T_N = k) - P(X_N = k)\big| \\ &\le 2 \cdot \mathrm{d_{TV}}\big(\mathcal{L}(T_N), \mathcal{L}(X_N)\big) \to 0. \end{aligned}$$

Also ist $\lim_{N\to\infty} P(T_N/ET_N \le x) = 1 - \exp(-x)$. □

Das wesentliche Hilfsmittel unserer Betrachtungen ist im Folgenden eine Verteilung ρ auf E, die folgende Bedingung erfüllt:

$$\rho(\,\cdot\,) = P_\rho(X_n \in\ \cdot\ | T_D > n) \text{ für alle } n \in \mathbb{N}. \tag{2.2}$$

In diesem Zusammenhang stellt sich natürlich die Frage, ob eine solche Verteilung stets existiert und wenn ja, wie man sie bestimmen kann. Um diese Fragen zu beantworten, treffen wir folgende Bezeichnungen. Das Komplement von D in E, also $E\setminus D$, bezeichnen wir mit D^{c}. Weiter sei $P_{D^{\mathrm{c}}}$ die Einschränkung von P auf D^{c}, d.h. die

Matrix, die man aus P erhält, wenn man sämtliche Zeilen und Spalten eliminiert, die zu Zuständen aus D gehören.

P_{D^c} ist eine substochastische Matrix. Jede quadratische Matrix mit nicht negativen Komponenten (insbesondere also jede substochastische Matrix) hat die Eigenschaft, dass der Spektralradius zugleich ein Eigenwert dieser Matrix ist und es dazu einen (vom Nullvektor verschiedenen) Linkseigenvektor mit nicht negativen Komponenten gibt. (Folgt aus [BR97], Kapitel 1, Theorem 1.7.3, S. 35.)

Ist also η_0 der Spektralradius von P_{D^c}, so gibt es einen Vektor v mit nicht negativen Komponenten, so dass $vP_{D^c} = \eta_0 v$ gilt. Weiter sei $|v| > 0$ die Summe über die Komponenten von v. Damit definieren wir nun $\rho(i) := v(i)/|v|$ für $i \in D^c$ bzw. $\rho(i) = 0$ sonst. Diese Verteilung ρ auf E, die offenbar stets existiert, hat nun unter sehr schwachen Voraussetzungen die gewünschte Eigenschaft (2.2):

Lemma 2.16 *Der Spektralradius von P_{D^c} sei größer 0. Dann erfüllt die zuvor definierte Verteilung ρ die Bedingung (2.2).*

Beweis: Für $i \in D^c$ gilt

$$P_\rho(X_n = i,\, T_D > n) \;=\; \rho_{D^c} \cdot (P_{D^c})^n \cdot \mathbf{1}_i \;=\; \eta_0^n \cdot \rho_{D^c} \cdot \mathbf{1}_i \;=\; \eta_0^n\, \rho(i) \quad \text{und}$$

$$P_\rho(T_D > n) \;=\; \rho_{D^c} \cdot (P_{D^c})^n \cdot \mathbf{1} \;=\; \eta_0^n \cdot \rho_{D^c} \cdot \mathbf{1} \;=\; \eta_0^n,$$

wobei ρ_{D^c} die Einschränkung von ρ auf D^c bezeichnet, $\mathbf{1} := (1, ..., 1)^t$ ist und $\mathbf{1}_i$ der Vektor, dessen i-te Komponente 1 und alle übrigen 0 sind. Durch Quotientenbildung folgt die Aussage für alle $i \in D^c$. Für $i \in D$ ist sie trivial. □

Die Voraussetzung $\eta_0 > 0$ ist nicht besonders stark und wird in den von uns im Folgenden betrachten Situationen in der Regel erfüllt sein. Ist $\eta_0 = 0$, so gilt $P_\rho(T_D > 1) = \eta_0 = 0$, d.h. bei Startverteilung ρ tritt die Markov-Kette mit Wahrscheinlichkeit 1 sofort in die Zielmenge ein. Immer, wenn wir im Folgenden direkt oder indirekt von einer Startverteilung ρ mit Eigenschaft (2.2) Gebrauch machen, setzen wir stillschweigend voraus, dass die Bedingung $\eta_0 > 0$ erfüllt ist.

Bemerkung 2.17 Ist P_{D^c} sogar irreduzibel, so entspricht ρ der sog. *quasi-stationären Verteilung* von X bzgl. D, definiert durch

$$\rho(i) := \lim_{n \to \infty} P_\lambda(X_n = i | T_D > n)$$

für alle $i \in E$ und alle Startverteilungen λ auf E. In dieser Situation existiert also für die Markov-Kette X eine stationäre Verteilung auf D^c unter der Bedingung, dass X noch nicht in D eingetreten ist. Unter den hier betrachteten allgemeinen Bedingungen ist die Existenz einer solchen quasi-stationären Verteilung i.A. nicht mehr gewährleistet. Was wir aber im Folgenden benötigen, ist nicht die Existenz einer quasi-stationären Verteilung an sich, sondern lediglich die Existenz einer Verteilung ρ, die die Bedingung (2.2) erfüllt. (Siehe hierzu auch die Betrachtungen in [Ald82]. Die Existenz der quasi-stationären Verteilung im irreduziblen Fall ist eine Konsequenz aus dem *Perron-Frobenius-Theorem* ([Bré99], Kapitel 6, Theorem 1.1, S. 197); entscheidend ist, dass in dieser Situation $|\eta_0| > |\eta|$ für alle weiteren Eigenwerte η von P_{D^c} gilt.)

Der nun folgende Satz zeigt, warum die Verteilung ρ das entscheidende Hilfsmittel ist, um die Wartezeit T_D durch eine geometrische Verteilung zu approximieren. Sie stellt für dieses Vorhaben die „optimale“ Startverteilung dar:

Satz 2.18 *Sei $\eta_0 > 0$. Dann ist T_D unter der Startverteilung ρ geometrisch verteilt auf $\mathbb{N}$, d.h. für $n \in \mathbb{N}$ gilt*

$$P_\rho(T_D = n) = (1-p)^{n-1}p \text{ mit } p = P_\rho(T_D = 1) = \big(E_\rho T_D\big)^{-1}.$$

Beweis: Die Voraussetzung $\eta_0 > 0$ garantiert zunächst einmal die Existenz von ρ. Wir zeigen nun, dass die Verteilung von T_D unter ρ gedächtnislos ist, d.h. es gilt

$$P_\rho(T_D = n+m \mid T_D > n) = P_\rho(T_D = m) \text{ für alle } n, m \in \mathbb{N}. \tag{2.3}$$

Mit (2.2), der zeitlichen Homogenität und der Markov-Eigenschaft folgt nämlich

$$\begin{aligned} P_\rho(T_D = m)P_\rho(T_D > n) &= \sum_{i \notin D} P_i(T_D = m)\rho(i)P_\rho(T_D > n) \\ &= \sum_{i \notin D} P_i(T_D = m)P_\rho(X_n = i, T_D > n) \\ &= \sum_{i \notin D} P(T_D = m+n | X_n = i, T_D > n)P_\rho(X_n = i, T_D > n) \\ &= P_\rho(T_D = m+n), \end{aligned}$$

und damit (2.3).

Da P_{D^c} eine substochastische Matrix ist, ist der Spektralradius η_0 kleiner oder gleich 1. Nun gilt für alle $n \in \mathbb{N}$

$$P_\rho(T_D > n) = \rho_{D^c} \cdot \left(P_{D^c}\right)^n \cdot \mathbf{1} = \eta_0^n \left(\rho_{D^c} \cdot \mathbf{1}\right) = \eta_0^n.$$

Wäre $\eta_0 = 1$, so wäre $P_\rho(T_D < \infty) = 0$ im Widerspruch zur Irreduzibilität der Markov-Kette X. Also ist $\eta_0 < 1$. Dann gilt insbesondere

$$P_\rho(T_D = 1) = 1 - P_\rho(T_D > 1) = 1 - \eta_0 > 0.$$

Daraus und aus der Gedächtnislosigkeit ergibt sich dann aber sofort, dass T_D unter ρ auf $\mathbb{N}$ geometrisch verteilt sein muss. (Vgl. beispielsweise [Fel68], Abschnitt XIII.9, S. 328f.) □

Startet X also in ρ, so ist die Eintrittszeit T_D geometrisch verteilt. Das nun folgende Theorem behandelt den Fall, dass X mit einer beliebigen Verteilung λ auf E startet. Insbesondere werden Fehlerschranken angegeben, wie stark sich dann die Verteilung von T_D von einer geometrischen Verteilung auf $\mathbb{N}$ mit Parameter $P_\rho(T_D = 1)$ unterscheidet.

Theorem 2.19 *Für jede Verteilung λ auf E und jede nicht leere Teilmenge D von E gilt*

$$\mathrm{d_{TV}}\left(\mathcal{L}_\lambda(T_D), \mathcal{L}_\rho(T_D)\right) \leq \mathrm{d_{TV}}(\rho, \lambda).$$

Beweis: Sei $A \subset \mathbb{N}$ beliebig. Wendet man Hilfssatz B.4 auf $f(X) := \mathbb{1}\{T_D \in A\}$ an, so folgt

$$|P_\lambda(T_D \in A) - P_\rho(T_D \in A)| \leq \mathrm{d_{TV}}(\rho, \lambda).$$

Da die rechte Seite nicht von der Wahl der Teilmenge A abhängt, folgt hieraus

$$\mathrm{d_{TV}}\left(\mathcal{L}_\lambda(T_D), \mathcal{L}_\rho(T_D)\right) \leq \mathrm{d_{TV}}(\rho, \lambda).$$

□

Mit Hilfe dieses Theorems ist es also möglich, die Wartezeit T_D unter der Startverteilung λ durch eine geometrische Verteilung zu approximieren.

Eine notwendige Voraussetzung ist dabei die Kenntnis der Verteilung ρ. Diese lässt sich mit numerischen Methoden bei Markov-Ketten mit hinreichend kleinem Zustandsraum recht leicht gewinnen, da es sich letztlich um einen auf Länge 1 normierten Linkseigenvektor zum Perron-Eigenwert (d.h. zum betragsgrößten und reellwertigen Eigenwert) von P_{D^c} handelt.

In der Regel wird die Fehlerschranke $\mathrm{d_{TV}}(\rho, \lambda)$ jedoch keine gute Approximation liefern. Eine Ausnahme stellt hier der Fall dar, dass D „klein“ ist im Vergleich zu E, so dass T_D unter ρ mit hoher Wahrscheinlichkeit „groß“ ist. Entspricht die Startverteilung λ dann der stationären Verteilung π, so wird kaum ein Unterschied zwischen der Verteilung ρ mit $\rho(\,\cdot\,) = P_\rho(X_n \in \,\cdot\, | T_D > n)$ und der Verteilung π mit $\pi(\,\cdot\,) = P_\pi(X_n \in \,\cdot\,)$ bestehen, da die Bedingung $T_D > n$ nur eine geringe Einschränkung darstellt.

Der nun folgende Satz zeigt eine Möglichkeit auf, wie man auch für beliebige Verteilungen λ zu guten Näherungen für $\mathcal{L}_\lambda(T_D)$ gelangen kann.

Satz 2.20 *Sei λ eine beliebige Verteilung auf E, $n_0 \in \mathbb{N}$ und λ_{n_0} definiert durch $\lambda_{n_0}(\,\cdot\,) := P_\lambda(X_{n_0} \in \,\cdot\, | T_D > n_0)$. Dann gilt für alle $n \in \mathbb{N}$*

$$\begin{aligned} P_\lambda(T_D > n_0 + n) &\geq \big(P_\rho(T_D > n) - \mathrm{d_{TV}}(\rho, \lambda_{n_0})\big) P_\lambda(T_D > n_0) \quad \textit{und} \\ P_\lambda(T_D > n_0 + n) &\leq \big(P_\rho(T_D > n) + \mathrm{d_{TV}}(\rho, \lambda_{n_0})\big) P_\lambda(T_D > n_0). \end{aligned}$$

Beweis: Aufgrund der Homogenität von X gilt

$$\begin{aligned} P_\lambda(T_D > n_0 + n) &= P_\lambda(T_D > n_0 + n | T_D > n_0) P_\lambda(T_D > n_0) \\ &= P_{\lambda_{n_0}}(T_D > n) P_\lambda(T_D > n_0). \end{aligned}$$

Die Aussage des Satzes folgt nun mit Hilfssatz B.4, angewendet auf die Funktion $f(X) := \mathbb{1}(T_D > n)$. □

Mit diesem Satz erhält man immer dann eine brauchbare Approximation für $\mathcal{L}_\lambda(T_D)$, wenn bereits für ein kleines $n_0 \in \mathbb{N}$ der Abstand $\mathrm{d_{TV}}(\rho, \lambda_{n_0})$ sehr klein wird, andererseits aber $P_\lambda(T_D > n_0)$ noch nahe bei 1 liegt. In diesem Zusammenhang weisen wir auch nochmal darauf hin, dass für irreduzibles P_{D^c} λ_{n_0} mit $n_0 \to \infty$ bei beliebigem λ gegen ρ konvergiert.

Bemerkung 2.21 Ähnliche Methoden und Resultate findet man in [Ald82] [Ald83]. Dieser Artikel war zugleich der Anstoß zur Herleitung der hier vorgestellten Resultate. David J. Aldous behandelt dort das Problem der Approximation der Eintrittszeitverteilung durch eine Exponentialverteilung im Fall einer zeitstetigen Markov-Kette. Insbesondere zeigt Aldous, dass sich der Fehler bei der Approximation des Tails der Eintrittszeitverteilung durch den Tail einer Exponentialverteilung mit Parameter

$(E_\pi T_D)^{-1}$ durch eine universelle Funktion $\Psi\big(E_\pi T_D/\tau\big)$ beschränken lässt, die allein vom Quotienten aus $E_\pi T_D$ und einer „Mixing Time“

$$\tau := \inf\big\{t \geq 0 : \mathrm{d_{TV}}\big(\mathcal{L}_i(X_t), \pi\big) \leq (2e)^{-1} \text{ für alle } i \in E\big\}$$

abhängt. Dabei gilt $\Psi : \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ und $\lim_{x\to\infty} \Psi(x) = 0$.

Der Vorteil dieses Resultats liegt darin, dass es keinerlei Bezug auf eine Verteilung ρ nimmt. Sowohl die approximierende Exponentialverteilung als auch die Fehlerschranke hängen lediglich von der stationären Verteilung π und der Geschwindigkeit, mit der sich die Markov-Kette ihrer stationären Verteilung nähert, ab.

Der Autor dieser Dissertation hat jedoch Zweifel an diesem Resultat, auch wenn er es in seiner Allgemeinheit nicht widerlegen kann. In jedem Fall ist der bei Aldous gelieferte Beweis für dieses Resultat fehlerhaft und lässt sich auch nicht „in offensichtlicher Weise“ reparieren. Der Fehler liegt in der von Aldous hergeleiteten Ungleichung (2.11) ([Ald82], S. 308). Das folgende Gegenbeispiel mag dies unterstreichen.

Gegenbeispiel zu Formel (2.11) in [Ald82]: Es sei (X_t) eine zeitstetige Markov-Kette auf $\{0,1\}$ mit Generatormatrix

$$G = \begin{pmatrix} -g_0 & g_0 \\ g_1 & -g_1 \end{pmatrix}.$$

Als Übergangsmatrix ergibt sich

$$P(t) = \begin{pmatrix} \dfrac{g_0}{g_0+g_1}\, e^{-(g_0+g_1)t} + \dfrac{g_1}{g_0+g_1} & -\dfrac{g_0}{g_0+g_1}\, e^{-(g_0+g_1)t} + \dfrac{g_0}{g_0+g_1} \\ -\dfrac{g_1}{g_0+g_1}\, e^{-(g_0+g_1)t} + \dfrac{g_1}{g_0+g_1} & \dfrac{g_1}{g_0+g_1}\, e^{-(g_0+g_1)t} + \dfrac{g_0}{g_0+g_1} \end{pmatrix}, \quad t > 0.$$

Die stationäre Verteilung ist $\pi = \big(g_1/(g_0+g_1), g_0/(g_0+g_1)\big)$. Sei $D := \{1\}$. Unter $\{Z_0 = 0\}$ ist T_D exponentialverteilt mit Parameter g_0, unter $\{Z_0 = 1\}$ ist die Verteilung von T_D das Einpunktmaß in 0.

Damit erhalten wir

$$E_\pi T_D = \pi(0) E_0 T_D + \pi(1) E_1 T_D = \pi(0) E_0 T_D = \frac{g_1/g_0}{g_0+g_1}.$$

τ ist das Infimum über alle $t \geq 0$ mit

$$(1/2) \cdot \Big(|p_{i0}(t) - \pi(0)| + |p_{i1}(t) - \pi(1)| \Big) \leq (2e)^{-1}, \quad i = 0, 1.$$

Ist $g_0 > g_1$, so ergibt sich

$$\tau = \frac{1}{g_0 + g_1} \left(1 - \log \left(\frac{g_0 + g_1}{2g_0} \right) \right).$$

Sei nun in der Notation von [Ald82] $a := \log(E_\pi T_D / \tau)$ und $s := (1+a)\tau$. In [Ald82] (2.11) wird nun behauptet, dass für jede Verteilung λ auf E

$$|P_\lambda(s \leq T_D \leq 2s | T_D \geq s) - P_\pi(T_D \leq s)| \leq e^{-a} \tag{2.4}$$

gilt.

Es ist

$$P_\pi(T_D \leq s) = \pi(0)P_0(T_D \leq s) + \pi(1)P_1(T_D \leq s) = 1 - \pi(0)e^{-g_0 s} = 1 - \frac{g_1}{g_0 + g_1} e^{-g_0 s},$$

und für jede Startverteilung λ mit $\lambda(0) > 0$ gilt

$$\begin{aligned} P_\lambda(s \leq T_D \leq 2s | T_D \geq s) &= \frac{P_\lambda(s \leq T_D \leq 2s)}{P_\lambda(T_D \geq s)} \\ &= \frac{\lambda(0)P_0(s \leq T_D \leq 2s) + \lambda(1)P_1(s \leq T_D \leq 2s)}{\lambda(0)P_0(T_D \geq s) + \lambda(1)P_1(T_D \geq s)} \\ &= P_0(s \leq T_D \leq 2s | T_D \geq s) \;=\; P_0(T_D \leq s) \\ &= 1 - e^{-g_0 s}, \end{aligned}$$

wobei wir die Gedächtnislosigkeit der Exponentialverteilung verwendet haben. Somit ist

$$|P_\lambda(s \leq T_D \leq 2s | T_D \geq s) - P_\pi(T_D \leq s)| = \frac{g_0}{g_0 + g_1} e^{-g_0 s}$$

und (2.4) wird zu

$$\frac{g_0}{g_0 + g_1} e^{-g_0 s} \leq e^{-a}.$$

Wählt man aber beispielsweise $g_0 = 50$ und $g_1 = 1$, so erhält man unmittelbar ein Gegenbeispiel zu dieser Ungleichung.

2.4 Der Spaltenprozess

Mit den im vorangegangenen Abschnitt entwickelten Methoden ist es nun möglich, die Verteilung der Eintrittszeit einer Markov-Kette in eine Teilmenge ihres Zustandsraums durch eine geometrische Verteilung zu approximieren und dabei gleichzeitig Abschätzungen für den gemachten Fehler zu erhalten. Wollen wir diese Methoden auf das *APMP* anwenden, so müssen wir nun die zugehörige Wartezeit als eine solche Eintrittszeit interpretieren.

Ein Möglichkeit bietet dabei die im ersten Abschnitt eingeführte *Snake Chain* $(Y_{N,n})_{n\in\mathbb{N}}$ der Länge $2N$. Diese ist sowohl im Fall einer unabhängigen und identisch verteilten Zeichenkette, als auch im Fall einer durch eine Markov-Kette erzeugten Zeichenkette selbst wieder eine Markov-Kette über den Zustandsraum $\mathcal{X}_N := \bigcup_{m=0}^{2N} \Sigma^m$. Die zugehörige Übergangsmatrix ergibt sich in offensichtlicher Weise aus der Verteilung bzw. der Übergangsmatrix von X. Definiert man nun

$$D_N := \big\{B \in \mathcal{X}_N \,:\, \mathrm{msed}(A_N, B) \leq k_N\big\},$$

so gilt für die Wartezeit T_N des *APMP*

$$T_N = \inf\big\{n \in \mathbb{N} \,:\, \mathrm{msed}(A_N, X_{1:n}) \leq k_N\big\} = \inf\big\{n \in \mathbb{N} \,:\, Y_{N,n} \in D_N\big\}.$$

Leider erweist sich jedoch die *Snake Chain* im Zusammenhang mit praktischen Berechnungen in konkreten Beispielen als völlig ungeeignet. Zur Verwendung der hergeleiteten Resultate ist nämlich die Kenntnis der Verteilung ρ erforderlich. Unter praktischen Gesichtspunkten sind diese also nur dann anwendbar, wenn die Berechnung von ρ mit vertretbarem Rechenaufwand durchgeführt werden kann. Dies setzt insbesondere einen „hinreichend kleinen“ Zustandsraum der betrachteten Markov-Kette voraus.

Gerade dieses Kriterium wird von der *Snake Chain* jedoch überhaupt nicht erfüllt, schon in einfachen Beispielen nimmt der Zustandsraum riesige Größenordnungen an. Beispielsweise beträgt die Größe des Zustandsraums im Beispiel 2.13 des Pattern $TOBEORNOTTOBE$ (Länge 13) über dem Alphabet der 26 Buchstaben $\#\mathcal{X}_{13} = \#\big(\bigcup_{m=1}^{26} \Sigma^m\big) = \sum_{m=1}^{26} 26^m$. Auch wenn von jedem dieser Zustände ausgehend nur 26 Übergänge mit positiver Wahrscheinlichkeit möglich sind, die Übergangsmatrix als nur dünn besetzt ist, zwingt allein schon die Aufgabe, diese Übergangsmatrix

zu bestimmen, jeden handelsüblichen PC problemlos in die Knie. Hinzu kommt außerdem noch die Aufgabe, für jeden dieser Zustände zu überprüfen, ob er in die Zielmenge fällt oder nicht.

Wenn wir also praktisch nutzbare Resultate erzielen wollen, so müssen wir uns an dieser Stelle von der *Snake Chain* trennen und stattdessen ein neues Konzept entwickeln, das diesen praktischen Anforderungen Rechnung trägt.

Als Ausgangspunkt betrachten wir dazu die Frage, wie man die *Minimal Suffix Edit Distance* zwischen einem vorgegebenem Pattern A und einer zufälligen Zeichenkette X zu einem Zeitpunkt n ökonomisch berechnen kann. Unter praktischen Gesichtspunkten erscheint dies ja zunächst recht aufwendig, muss doch die Editierdistanz zwischen A und sämtlichen Suffixen von $X_{1:n}$ bestimmt werden. Doch wie schon bei der *Edit Distance* zwischen zwei Pattern, kommt uns auch hier wieder die dynamische Programmierung zu Hilfe. Die *Minimal Suffix Edit Distance* zwischen A und X zur Zeit n lässt sich durch eine ganz ähnliche Rekursion berechnen. Die eigentliche Rekursionsformel (2.1) ist dabei sogar exakt dieselbe, es ändert sich lediglich die Initialisierung der ersten Zeile zu $d(0,n) = 0$, $n \in \mathbb{N}_0$. $d(N,n)$ ist dann die *Minimal Suffix Edit Distance* zwischen A und X zur Zeit n. (Vgl. hierzu [Gus97], Abschnitt 11.6.5, S. 229f.)

Beispiel 2.22 Gegeben sei der Pattern $A := aggct$ über dem Alphabet $\Sigma := \{a, g, c, t\}$. Dann könnte für eine Realisierung der Zeichenkette $(X_n)_{n\in\mathbb{N}}$ das Schema zur Bestimmung der *Minimal Suffix Edit Distance* beispielsweise folgendermaßen aussehen, wobei in der i-ten Zeile und n-ten Spalte jeweils der Wert $d(i,n) := \mathrm{msed}(A_{1:i}, X_{1:n})$ steht:

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X_n	$\emptyset$	g	a	c	t	c	c	g	a	t	c	g	a	g	c	t	t
$\emptyset$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1
g	2	1	1	1	2	2	2	1	1	1	2	1	1	0	1	2	2
g	3	2	2	2	2	3	3	2	2	2	2	2	2	1	1	2	...
c	4	3	3	2	3	2	3	3	3	3	2	3	3	2	1	2	
t	5	4	4	3	**2**	3	3	4	4	3	3	3	4	3	2	**1**	

Insbesondere hat die Wartezeit T des *APMP* für $k = 2$ den Wert 4, für $k = 1$ den Wert 15. (Vgl. zu diesem Beispiel auch Abbildung 2.1.)

Für die weiteren Betrachtungen wollen wir einige Eigenschaften dieses Schemas zur Bestimmung der *Minimal Suffix Edit Distance* zwischen einem Pattern und einer Zeichenkette festhalten.

Satz 2.23 *Es sei* $(d(i,n))_{i=0:N}^{n\in\mathbb{N}_0}$ *das Schema zur Bestimmung der* Minimal Suffix Edit Distance *zwischen einem Pattern und einer Zeichenkette. Dann gilt:*

(a) $d(i,n) \in \{0,...,N\}$ *für alle* $i = 0,...,N$, $n \in \mathbb{N}_0$.

(b) $d(i,n+1) \in d(i,n) + \{-1,0,1\}$ *für alle* $i = 0,...,N$, $n \in \mathbb{N}_0$.

(c) $d(i+1,n) \in d(i,n) + \{-1,0,1\}$ *für alle* $i = 0,...,N-1$, $n \in \mathbb{N}_0$.

(d) $d(i+1,n+1) \in d(i,n) + \{0,1\}$ *für alle* $i = 0,...,N-1$, $n \in \mathbb{N}_0$.

Beweis: (a) ist trivial. (b), (c), (d) beweisen wir durch vollständige Induktion. Aus $d(0,0) = d(0,1) = 0$ und $d(1,0) = 1$ sowie $d(1,1) = \min\{d(0,1)+1, d(1,0)+1, d(0,0)+\delta(1,1)\}$ mit $\delta(1,1) \in \{0,1\}$ folgt $d(1,1) \in \{0,1\}$. Damit sind die Aussagen für $i = n = 0$ klar. Nun angenommen, (b), (c) und (d) sind bereits für alle Paare (i,n) mit $i \in \{0,...,N\}$, $n \in \{0,...,m-1\}$, sowie für alle (i,n) mit $i \in \{0,...,j-1\}$, $n = m$ gezeigt. Dann sind nun zu zeigen:

$$\begin{aligned} d(j,m+1) &\in d(j,m) + \{-1,0,1\}, \\ d(j+1,m) &\in d(j,m) + \{-1,0,1\}, \\ d(j+1,m+1) &\in d(j,m) + \{0,1\}. \end{aligned}$$

Ist $d(j,m+1) \notin d(j,m) + \{-1,0,1\}$, so ist wegen der Induktionsvoraussetzung zwangsläufig $d(j,m) = d(j-1,m) - 1$ und $d(j,m+1) = d(j-1,m) + 1$. Dann folgt aber aus der Rekursionsformel (2.1)

$$\begin{aligned} d(j,m+1) &= \min\left\{d(j-1,m+1)+1, d(j,m)+1, d(j-1,m)+\delta(j,m+1)\right\} \\ &= \min\left\{d(j-1,m+1)+1, d(j-1,m), d(j-1,m)+\delta(j,m+1)\right\} \\ &\leq d(j-1,m) \end{aligned}$$

im Widerspruch zu $d(j, m+1) = d(j-1, m) + 1$.

Ist $j = N$, so ist an dieser Stelle der Induktionsschritt bereits abgeschlossen. Anderenfalls zeigt man auf ganz ähnliche Weise $d(j+1, m) \in d(j, m) + \{-1, 0, 1\}$. Bleibt noch Aussage (d):

$$d(j+1, m+1) = \min\big\{ \underbrace{d(j, m+1) + 1}_{\in d(j,m)+\{0,1,2\}}, \underbrace{d(j+1, m) + 1}_{\in d(j,m)+\{0,1,2\}}, \underbrace{d(j, m) + \delta(j+1, m+1)}_{\in d(j,m)+\{0,1\}} \big\},$$

also $d(j+1, m+1) \in d(j, m) + \{0, 1\}$. □

Dieser Satz verallgemeinert insbesondere Satz 2.6.

Das Schema $\big(d(i, n)\big)_{i=0:N}^{n \in \mathbb{N}_0}$ bietet uns nun die Möglichkeit, eine neue Markov-Kette als Ersatz für die *Snake Chain* zu konstruieren. Dazu betrachten wir den stochastischen Prozess, der die zeitliche Entwicklung der Spalten dieses Schemas beschreibt: Der sog. *Spaltenprozess* $D = (D_n)_{n \in \mathbb{N}_0}$ zu einem Pattern A und einer Zeichenkette X ist für alle $n \in \mathbb{N}_0$ definiert durch

$$D_n := \big(d(i, n)\big)_{i=0:N}.$$

Für die Konstruktion der neuen Markov-Kette müssen wir nun wieder zwischen unseren beiden Modellannahmen für die Verteilung der Zeichenkette X unterscheiden.

Satz 2.24 *Sei X eine Folge von unabhängigen und identisch verteilten Zufallsgrößen über Σ. Dann ist der Spaltenprozess $D = (D_n)_{n \in \mathbb{N}_0}$ eine Markov-Kette. Betrachtet man nur das Teilstück $(D_n)_{n \geq 2N}$, so ist dies eine positiv rekurrente und stationäre Markov-Kette.*

Beweis: Aus der Rekursionsformel (2.1) folgt, dass D_n nur von D_{n-1} und X_n abhängt, oder anders gesagt: Es existiert eine Funktion f mit

$$D_n = f(X_n, D_{n-1}).$$

Nach [Bré99], Kapitel 2, Theorem 2.1, S. 58, ist $(D_n)_{n \in \mathbb{N}_0}$ dann eine Markov-Kette.

Um zu zeigen, dass $(D_n)_{n \geq 2N}$ positiv rekurrent ist, begeben wir uns auf die Ebene der Zeichenkette X. Für jedes $i \in \{0, 1, ..., N\}$ ist $d(i, n) = \mathrm{msed}(A_{1:i}, X_{1:n})$. In den Ausführungen im Anschluss an Satz 2.7 haben wir gezeigt, dass $\mathrm{msed}(A_{1:i}, X_{1:n}) = \mathrm{msed}(A_{1:i}, X_{n-2i+1:n})$ gilt. Also ist $D_n = \big(d(i, n)\big)_{i=0:N}$ eine Funktion von $X_{n-2N+1:N}$. Für $n \geq 2N$ gehört zu jeder Realisierung von D_n also (mindestens) eine Realisierung

des Suffix $X_{n-2N+1:n}$ der Zeichenkette, die D_n erzeugt. So kann die Markov-Kette $(D_n)_{n\geq 2N}$ offenbar stets in höchstens $2N$ Schritten in jeden ihrer Zustände zurückkehren, indem in der Zeichenkette X eines der erzeugenden Suffixe erscheint.

Aus dieser Überlegung ergibt sich auch sofort die Stationarität von $(D_n)_{n\geq 2N}$, denn zu jedem Zeitpunkt ist die Wahrscheinlichkeit, dass D_n einen bestimmten Zustand annimmt, gleich und entspricht der Wahrscheinlichkeit, dass $X_{n-2N+1:n}$ eines der erzeugenden Suffixe ist. □

Im Fall einer unabhängigen und identisch verteilten Zeichenkette X ist der zu A gehörende Spaltenprozess D also eine Markov-Kette, deren Zustandsraum wir im Folgenden mit Δ bezeichnen. Ist X hingegen eine Markov-Kette, so trifft dies in der Regel nicht mehr zu, wie das folgende Gegenbeispiel zeigt.

Beispiel 2.25 Sei $A = 121$ über dem Alphabet $\Sigma := \{1, 2, 3\}$. Die Zeichenkette X sei eine Markov-Kette mit Übergangsmatrix $P = (p_{ij})_{i,j=1,2,3}$, die zugehörige Startverteilung sei $\nu = (\nu_1, \nu_2, \nu_3)$. Der Zustandsraum des Spaltenprozesses $D = (D_n)_{n\in\mathbb{N}_0}$ findet man zusammen im dem zugehörigen Übergangsgraphen in Abbildung 2.2. Insbesondere sind $d_0 := (0, 1, 2, 3)$, $d_1 := (0, 0, 1, 2)$ und $d_2 := (0, 1, 1, 2)$ Elemente des Zustandsraums von D. Nun gilt

$$\begin{aligned} P(D_3 = d_2 | D_2 = d_2, D_1 = d_2, D_0 = d_0) &= \sum_{i=1}^{3} \nu_i p_{i2} p_{22} p_{22} \quad \text{und} \\ P(D_3 = d_2 | D_2 = d_2, D_1 = d_1, D_0 = d_0) &= \sum_{i=1}^{3} \nu_i p_{i1} p_{13} p_{22}. \end{aligned}$$

In der Regel werden diese beiden bedingten Wahrscheinlichkeiten nicht gleich sein, was zeigt, dass D keine Markov-Kette sein kann.

Es gilt jedoch folgendes

Lemma 2.26 *Sei $(Z_n)_{n\in\mathbb{N}}$ eine homogene Markov-Kette mit endlichem Zustandsraum F und X_0 eine von $(Z_n)_{n\in\mathbb{N}}$ unabhängige Zufallsvariable mit Werten in einer endlichen Menge E. Weiter sei $f : E \times F \to E$ eine surjektive Funktion, und $(X_n)_{n\in\mathbb{N}}$ sei definiert durch*

$$X_{n+1} := f(X_n, Z_{n+1}).$$

Hat nun f die Eigenschaft, dass aus $f(e_1, f_1) = f(e_2, f_2)$ stets $f_1 = f_2$ folgt, so ist $(X_n)_{n\in\mathbb{N}_0}$ eine homogene Markov-Kette.

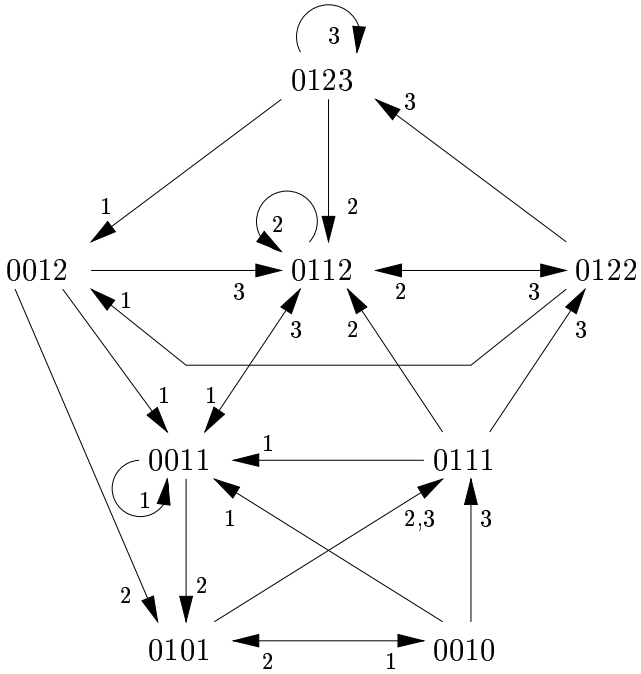

Abbildung 2.2: Graphische Darstellung des Zustandsraums und des zugehörigen Übergangsgraphen für den Spaltenprozess $(D_n)_{n\in\mathbb{N}_0}$ zum Pattern $A = 121$ über dem Alphabet $\Sigma := \{1, 2, 3\}$. Die Zahlen an den Pfeilspitzen geben an, durch das Erscheinen welches Buchstabens in der Zeichenkette der jeweilige Übergang ausgelöst wird. (Vgl. Beispiel 2.25.)

Beweis: Durch die angegebene Eigenschaft von f ist gewährleistet, dass zu jedem $e \in E$ die zweite Komponente sämtlicher Urbilder von e unter f eindeutig bestimmt ist. Sei also $g : E \to F$ die Funktion, die einem $e \in E$ die zweite Komponente eines Urbildes (aller Urbilder) von e unter f zuordnet. Dann gilt

$$\begin{aligned}
&P\big(X_{n+1} = e_{n+1} | X_n = e_n, ..., X_1 = e_1, X_0 = e_0\big) \\
=\ & P\big(f(e_n, Z_{n+1}) = e_{n+1} | f(e_{n-1}, Z_n) = e_n, ..., f(e_0, Z_1) = e_1, X_0 = e_0\big) \\
=\ & P\big(Z_{n+1} = g(e_{n+1}) | Z_n = g(e_n), ..., Z_1 = g(e_1), X_0 = e_0\big) \\
=\ & P\big(Z_{n+1} = g(e_{n+1}) | Z_n = g(e_n)\big).
\end{aligned}$$

Andererseits gilt

$$
\begin{aligned}
&P\big(X_{n+1} = e_{n+1} | X_n = e_n\big) \;=\; P\big(f(e_n, Z_{n+1}) = e_{n+1}\big) | f(X_{n-1}, Z_n) = e_n\big) \\
&= \; P\big(Z_{n+1} = g(e_{n+1}) | Z_n = g(e_n), \underbrace{f(X_{n-1}, g(e_n)) = e_n}_{\in \sigma\{X_0, Z_1, \ldots, Z_{n-1}\}}\big) \\
&= \; P\big(Z_{n+1} = g(e_{n+1}) | Z_n = g(e_n)\big),
\end{aligned}
$$

$(X_n)_{n\in\mathbb{N}_0}$ ist also eine Markov-Kette. □

Insbesondere ergibt sich also, dass $(D_n)_{n\in\mathbb{N}_0}$ eine homogene Markov-Kette ist, sofern durch D_n stets der Zustand der Markov-Kette X zur Zeit n eindeutig festgelegt ist. In vielen Fällen ist dies tatsächlich gewährleistet. Sollte dies jedoch wie im obigen Beispiel nicht der Fall sein, so helfen wir uns, indem wir zu einem neuen Prozess $(\tilde{D}_n)_{n\in\mathbb{N}_0}$ mit

$$\tilde{D}_n := (X_n, D_n)$$

übergehen. Der Einfachheit halber bezeichnen wir auch diesen Prozess als Spaltenprozess zu X und A, der zugehörige Zustandsraum sei $\tilde{\Delta}$.

Satz 2.27 *Sei X eine homogene, aperiodische und irreduzible Markov-Kette über Σ. Dann ist auch der Spaltenprozess $\tilde{D} = (\tilde{D}_n)_{n\in\mathbb{N}_0}$ eine Markov-Kette. Betrachtet man nur das Teilstück $(\tilde{D}_n)_{n\geq 2N}$, so ist dies eine positiv rekurrente Markov-Kette.*

Beweis: Durch den Übergang von $(D_n)_{n\in\mathbb{N}_0}$ zu $(\tilde{D}_n)_{n\in\mathbb{N}_0}$ bekommt die surjektive Funktion f, die $\tilde{D}_{n+1}$ aus $\tilde{D}_n$ und X_{n+1} generiert, die in Lemma 2.26 definierte Eigenschaft, also ist $(\tilde{D}_n)_{n\in\mathbb{N}_0}$ nun tatsächlich eine homogene Markov-Kette. Die positive Rekurrenz ergibt sich wie zuvor im Fall einer unabhängig und identisch verteilten Zeichenkette. □

Die Stationarität des Spaltenprozesses geht in diesem zweiten Modell natürlich verloren. An ihre Stelle tritt eine Aussage über die Geschwindigkeit, mit der sich die Verteilung der Markov-Kette $\tilde{D}$ ihrer stationären Verteilung annähert, die wir im Folgenden mit $\tilde{\mu}$ bezeichnen. Für $c > 0$ sei

$$
\begin{aligned}
\tau_X(c) &:= \min\big\{n \in \mathbb{N} \,:\, \mathrm{d_{TV}}\big(\mathcal{L}_\sigma(X_n), \pi\big) \leq c \text{ für alle } \sigma \in \Sigma\big\}, \\
\tau_{\tilde{D}}(c) &:= \min\big\{n \in \mathbb{N} \,:\, \mathrm{d_{TV}}\big(\mathcal{L}_{\tilde{d}}(\tilde{D}_n), \tilde{\mu}\big) \leq c \text{ für alle } \tilde{d} \in \tilde{\Delta}\big\}.
\end{aligned}
$$

Dann gilt der folgende Satz:

Satz 2.28 *Für alle $c > 0$ gilt*

$$\tau_{\tilde{D}}(c) \leq \tau_X\big(c \cdot (\#\Sigma \cdot p_{\max})^{1-2N}\big) + 2N.$$

Hierbei ist $p_{\max} := \max_{\sigma_1,\sigma_2 \in \Sigma} p(\sigma_1, \sigma_2)$.

Beweis: Sei $Y = (Y_n)_{n \in \mathbb{N}_0}$ die *Snake Chain* zu X der Länge $2N$. $(Y_n)_{n \geq 2N}$ ist dann eine homogene Markov-Kette auf Σ^{2N} mit stationärer Verteilung

$$\pi_Y\big((x_1, ..., x_{2N})\big) = \pi(x_1) \prod_{i=2}^{2N} p(x_{i-1}, x_i) \text{ für alle } (x_1, ..., x_{2N}) \in \Sigma^{2N}.$$

Für $n \geq 2N$ und $\sigma \in \Sigma$ gilt

$$\begin{aligned}
\mathrm{d_{TV}}\big(\mathcal{L}(Y_n|X_0 = \sigma), \pi_Y\big) &= \frac{1}{2} \sum_{x \in \Sigma^{2N}} |P(Y_n = x|X_0 = \sigma) - \pi_Y(x)| \\
&= \frac{1}{2} \sum_{x \in \Sigma^{2N}} |P_\sigma(X_{n-2N+1} = x_1) - \pi(x_1)| \cdot \prod_{i=2}^{2N} p(x_{i-1}, x_i) \\
&\leq (\#\Sigma \cdot p_{\max})^{2N-1} \mathrm{d_{TV}}\big(\mathcal{L}_\sigma(X_{n-2N+1}), \pi\big).
\end{aligned}$$

$\tilde{D}_n$ ist eine Funktion von Y_n. Deshalb folgt mit Hilfssatz B.3 und der Tatsache, dass Y_n unabhängig von D_0 ist:

$$\begin{aligned}
\tau_{\tilde{D}}(c) &\leq \min\big\{n \geq 2N \,:\, \mathrm{d_{TV}}\big(\mathcal{L}_{\tilde{d}}(\tilde{D}_n), \tilde{\mu}\big) \leq c,\ \forall \tilde{d} \in \tilde{\Delta}\big\} \\
&\leq \min\big\{n \geq 2N \,:\, \mathrm{d_{TV}}\big(\mathcal{L}(Y_n|X_0 = \sigma), \pi_Y\big) \leq c,\ \forall \sigma \in \Sigma\big\} \\
&\leq \min\big\{n \geq 2N \,:\, (\#\Sigma \cdot p_{\max})^{2N-1} \mathrm{d_{TV}}\big(\mathcal{L}_\sigma(X_{n-2N+1}), \pi\big) \leq c,\ \forall \sigma \in \Sigma\big\} \\
&= \min\big\{n \in \mathbb{N} \,:\, (\#\Sigma \cdot p_{\max})^{2N-1} \mathrm{d_{TV}}\big(\mathcal{L}_\sigma(X_n), \pi\big) \leq c,\ \forall \sigma \in \Sigma\big\} + 2N \\
&= \tau_X\big(c(\#\Sigma \cdot p_{\max})^{1-2N}\big) + 2N.
\end{aligned}$$

□

Bemerkung 2.29 Gilt $\mathrm{d_{TV}}\big(\mathcal{L}_\sigma(X_n), \pi\big) \leq \kappa \cdot \rho^n$ für alle $\sigma \in \Sigma$, $n \in \mathbb{N}$ mit $0 < \rho < 1$ und $\kappa > 0$, so ist

$$\tau_X(c) \leq \lceil \log_\rho\big(c/\kappa\big) \rceil \quad \text{und} \quad \tau_D(c) \leq \left\lceil \log_\rho\left(\frac{c \cdot (\#\Sigma \cdot p_{\max})^{1-2N}}{\kappa}\right) \right\rceil + 2N.$$

Unter beiden Modellannahmen für die Zeichenkette X ist der zum $APMP$ gehörende Spaltenprozess D bzw. $\tilde{D}$ also eine Markov-Kette. Wie schon bei der *Snake Chain* kann man auch hier wieder die Wartezeit T des $APMP$ als Eintrittszeit des zugehörigen Spaltenprozesses interpretieren: Definiert man für $0 \leq k \leq N$

$$\Delta_k := \{d = (d_0, ..., d_N) \in \Delta \mid d_N \leq k\},$$

so gilt im Fall einer unabhängigen und identisch verteilten Zeichenkette X

$$T = \min\{n \in \mathbb{N} \mid D_n \in \Delta_k\},$$

wobei die Markov-Kette $D = (D_n)_{n \in \mathbb{N}_0}$ im Zustand $d^* = (0, 1, ..., N)$ startet. Analog geht man im Fall einer durch eine Markov-Kette erzeugten Zeichenkette X vor.

Der entscheidende Vorteil des Spaltenprozesses gegenüber der *Snake Chain* liegt nun darin, dass sein Zustandsraum wesentlich kleiner ist. Eine triviale Oberschranke lässt sich dabei sofort angeben: Da $d(0, n)$ stets Null ist und für die Komponenten von D_n stets $d(i+1, n) \in d(i, n) + \{-1, 0, 1\}$ gilt, ist offenbar $\#\Delta \leq 3^N$ sowie $\#\tilde{\Delta} \leq \#\Sigma \cdot 3^N$.

Bevor wir den Spaltenprozess nun verwenden, um die Wartezeit des $APMP$ durch eine geometrische Verteilung zu approximieren, wollen wir an dieser Stelle die praxisrelevanten Eigenschaften dieses Prozesses genauer untersuchen. Dabei interessiert uns insbesondere, wie man für ein konkretes *Pattern Matching Problem* den Zustandsraum des Spaltenprozesses und die zugehörige Übergangsmatrix bestimmen kann, und ob es möglich ist, die Größe dieses Zustandsraums durch theoretische Überlegungen a priori noch besser abzuschätzen. Aus dem Zustandsraum Δ von $(D_n)_{n \in \mathbb{N}_0}$ lässt sich der Zustandsraum $\tilde{\Delta}$ von $(\tilde{D}_n)_{n \in \mathbb{N}_0}$ sehr einfach gewinnen, insbesondere gilt $\#\tilde{\Delta} \leq \#\Sigma \cdot \#\Delta$. Im Folgenden beschränken wir deshalb unsere Betrachtungen auf Δ.

Zunächst zur ersten Frage. Wir wissen, dass auf jeden Fall der Startzustand $d^* = (0, 1, ..., N)$ in Δ liegt. Außerdem ist uns der durch (2.1) beschriebene Übergangsmechanismus der Markov-Kette $(D_n)_{n \in \mathbb{N}_0}$ bekannt. Mit diesen beiden Komponenten ist es recht einfach, den gesamten Zustandsraum Δ mit Hilfe einer sog. Tiefensuche zu gewinnen. Ein Pseudocode-Programmtext hierfür könnte etwa wie folgt aussehen:

```
step<-function(d,x)
  {liefert als Rueckgabewert den Zustand, der sich aus dem
   Zustand d bei Erscheinen des Buchstabens x in der
   Zeichenkette mit der Rekursion (2.1) ergibt.}

tiefensuche<-function(stateset,state)
  {stateset<-union(stateset,{state})
   for x from ersterBuchstabedesAlphabets
            to letzterBuchstabedesAlphabets
     {newstate<-step(state,x)
      if (newstate not in stateset)
        stateset<-tiefensuche(stateset,newstate)
      else next}
   return(stateset)}

print(tiefensuche(emptyset,d*))
```

Man beachte, dass diese rekursive Bestimmung des Zustandsraums Δ in der Theorie zwar sehr einfach ist, bei der praktischen Programmierung aber schnell zu Speicherplatzproblemen führen kann, da bei jedem Aufruf der Funktion `tiefensuche` die bisherigen Ergebnisse im Hauptspeicher bleiben müssen. Praktisch würde man Δ also durch eine andere Programmstruktur bestimmen (vgl. Anhang C), aber die Grundidee ist hier auf sehr einfache Art und Weise dargelegt.

Ist erst einmal der Zustandsraum bestimmt, so ist es auch ein Leichtes, aus der Rekursion (2.1) die Übergangsmatrix herzuleiten. Offensichtlich gibt es von jedem Zustand aus höchstens so viele Übergänge, wie Buchstaben im Alphabet Σ vorhanden sind, ebenso wie bei der *Snake Chain* ist also auch hier die Übergangsmatrix dünn besetzt.

Beispiel 2.30 Sei X eine unabhängige und identisch verteilte Zeichenkette über dem Alphabet $\Sigma := \{1, 2, 3\}$. Die zugehörige Verteilung sei gegeben durch $P(X_1 = 1) = 1/6$, $P(X_1 = 2) = 1/3$ und $P(X_1 = 3) = 1/2$. Über Σ betrachten wir den Pattern $A = 121$. Dann ist der Zustandsraum Δ des zugehörigen Spaltenprozesses gegeben durch

$$\begin{aligned}\Delta \;=\; \big\{&(0,1,2,3),(0,0,1,2),(0,1,1,2),(0,1,2,2),\\ &(0,0,1,1),(0,1,1,1),(0,1,0,1),(0,0,1,0)\big\}.\end{aligned}$$

Die dazu gehörige Übergangsmatrix ergibt sich aus der Verteilung der Zeichenkette und dem zugehörigen Übergangsgraphen (vgl. Abbildung 2.2) zu

$$Q = \begin{pmatrix} 1/2 & 1/6 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 1/6 & 0 & 1/3 & 0 \\ 0 & 0 & 1/3 & 1/2 & 1/6 & 0 & 0 & 0 \\ 1/2 & 1/6 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 1/6 & 0 & 1/3 & 0 \\ 0 & 0 & 1/3 & 1/2 & 1/6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5/6 & 0 & 1/6 \\ 0 & 0 & 0 & 0 & 1/6 & 1/2 & 1/3 & 0 \end{pmatrix}.$$

Bei der zweiten Frage nach einer a priori Abschätzung für die Größe des Zustandsraums Δ hilft die folgende Überlegung weiter: Wir betrachten das Gitter $\mathbb{Z}^2$. Von jedem Punkt $(i,j) \in \mathbb{Z}^2$ mögen drei Übergänge möglich sein: $(i,j) \to (i+1, j+1)$ (Abkürzung u wie up), $(i,j) \to (i+1, j)$ (Abkürzung e wie equal) und $(i,j) \to (i+1, j-1)$ (Abkürzung d wie down). Sei $\mathcal{X}_N$ die Menge aller Pfade der Länge N mit Start in 0, die man auf diese Weise beschreiten kann (vgl. Abbildung 2.3).

Jedem Zustand in Δ können wir nun über die Differenz aufeinander folgender Komponenten in eindeutiger Weise einen solchen Pfad in $\mathcal{X}_N$ zuordnen:

$$\Delta \ni d = (d_0, d_1, ..., d_N) \mapsto \big((0,0), (1, d_1 - d_0), (2, d_2 - d_1), ..., (N, d_N - d_{N-1})\big) \in \mathcal{X}_N.$$

Da diese Abbildung injektiv ist, ergibt sich die zuvor angegebene triviale Oberschranke nun aus der Überlegung, dass wir $\#\Delta$ durch $\#\mathcal{X}_N = 3^N$ nach oben abschätzen können. Eine weitere Reduzierung erhalten wir, wenn wir berücksichtigen, dass die Komponenten von D_n stets nicht negativ sind. Daher ist $\#\Delta$ kleiner oder gleich der Anzahl aller Pfade in $\mathcal{X}_N$, die die Linie $(\,\cdot\,, -1)$ niemals treffen.

Offenbar gilt das folgende

Lemma 2.31 (Spiegelungsprinzip) *Die Menge aller Pfade von* $(0,0)$ *nach* (N,c), $c \in \{0, ..., N\}$, *die die Linie* $(\,\cdot\,, -1)$ *mindestens einmal treffen, lässt sich bijektiv auf die Menge aller Pfade abbilden, die von* $(0,-2)$ *nach* (N,c) *führen. (Vgl. Abbildung 2.4.)*

Um dieses Spiegelungsprinzip zur Abschätzung von $\#\Delta$ gewinnbringend anwenden zu können, benötigen wir

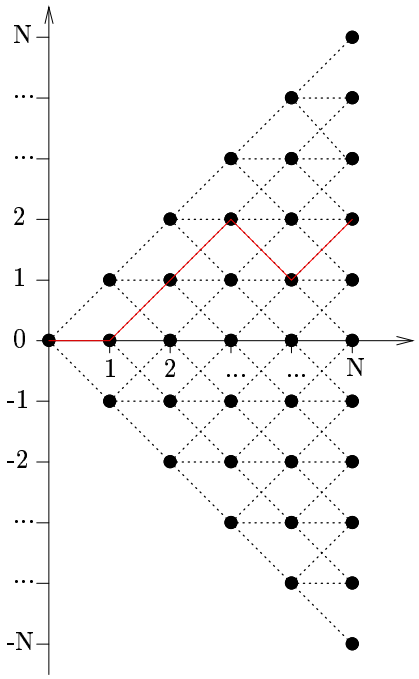

Abbildung 2.3: Veranschaulichung der Zustände in Δ durch Pfade im Gitter $\mathbb{Z}^2$.

Lemma 2.32 *Sei $c_1 \leq c_2 \leq c_1 + N$. Dann ist die Anzahl der Pfade von $(0, c_1)$ nach (N, c_2) gegeben durch $a(N, c_1, c_2)$ mit*

$$a(N, c_1, c_2) = \sum_{k=0}^{\frac{N-(c_2-c_1)}{2}} \frac{N!}{(2k)! \left(\frac{N-(c_2-c_1)-2k}{2}\right)! \left(\frac{N+(c_2-c_1)-2k}{2}\right)!},$$

falls $N - (c_2 - c_1)$ gerade, bzw.

$$a(N, c_1, c_2) = \sum_{k=1}^{\frac{N-(c_2-c_1)+1}{2}} \frac{N!}{(2k-1)! \left(\frac{N-(c_2-c_1)-(2k-1)}{2}\right)! \left(\frac{N+(c_2-c_1)-(2k-1)}{2}\right)!},$$

falls $N - (c_2 - c_1)$ ungerade.

Beweis: Zu jedem Pfad von $(0, c_1)$ nach (N, c_2) sei $(\#u, \#e, \#d)$ der Vektor der Anzahlen der Übergänge u, e, d. Dabei gilt offenbar $\#u, \#e, \#d \geq 0$, $\#u + \#e + \#d = N$ und $\#u - \#d = c_2 - c_1$. Daraus ergibt sich

$$2 \cdot \#d + \#e = N - (c_2 - c_1). \tag{2.5}$$

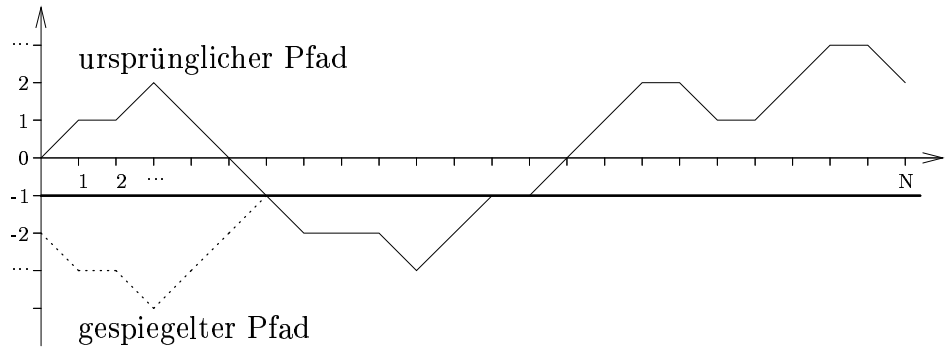

Abbildung 2.4: Graphische Darstellung des Spiegelungsprinzips.

Die Anzahl der Lösungen $(\#e, \#d) \in \mathbb{N}_0^2$ dieser Gleichung beträgt

$$\begin{aligned} &\frac{1}{2}\big(N-(c_2-c_1)\big)+1, \quad \text{im Falle } N-(c_2-c_1) \text{ gerade,} \\ &\frac{1}{2}\big(N-(c_2-c_1)+1\big), \quad \text{im Falle } N-(c_2-c_1) \text{ ungerade.} \end{aligned}$$

Verbleibt die Frage, auf wie viele Arten man diese insgesamt N Operationen u, e, d für eine spezielle Lösung von (2.5) anordnen kann. Antwort:

$$\binom{N}{e}\binom{N-e}{d}\binom{N-e-d}{u} = \binom{N}{e}\binom{N-e}{d}.$$

Im Falle $N-(c_2-c_1)$ gerade ergibt sich also als Anzahl der Pfade $(0,c_1)$ nach (N,c_2)

$$\sum_{k=0}^{\frac{N-(c_2-c_1)}{2}} \binom{N}{2k}\binom{N-2k}{\frac{N-(c_2-c_1)}{2}-k} = \sum_{k=0}^{\frac{N-(c_2-c_1)}{2}} \frac{N!}{(2k)!\left(\frac{N-(c_2-c_1)-2k}{2}\right)!\left(\frac{N+(c_2-c_1)-2k}{2}\right)!}.$$

Analog im Fall $N-(c_2-c_1)$ ungerade. □

Sei $\mathcal{C} \subset \mathbb{Z}$. Dann sei die Menge aller Pfade von (i_1,c_1) nach $(i_2,\mathcal{C})$ die Menge aller Pfade, die in (i_1,c_1) starten und in einem Punkte (i_2,c_2) mit $c_2 \in \mathcal{C}$ enden. Aus Lemma 2.32 ergibt sich nun

Lemma 2.33 *Sei* $\mathcal{C} := \{c \in \mathbb{Z} \mid 0 \le c \le N\}$. *Dann ist die Anzahl der Pfade von* $(0,0)$ *nach* $(N,\mathcal{C})$, *die* $(\,\cdot\,,-1)$ *nicht treffen, gleich* $b(N)$ *mit*

$$b(N) = \sum_{k=0}^{\frac{N}{2}} \frac{N!}{(2k)!\left(\left(\frac{N}{2}-k\right)!\right)^2} + \sum_{k=1}^{\frac{N}{2}} \frac{N!}{(2k-1)!\left(\frac{N}{2}-k\right)!\left(\frac{N}{2}-k+1\right)!},$$

falls N gerade, bzw.

$$b(N) = \sum_{k=1}^{\frac{N+1}{2}} \frac{N!}{(2k-1)!\left(\left(\frac{N}{2}-\frac{2k-1}{2}\right)!\right)^2} + \sum_{k=0}^{\frac{N-1}{2}} \frac{N!}{(2k)!\left(\frac{N-1}{2}-k\right)!\left(\frac{N+1}{2}-k\right)!},$$

falls N ungerade.

Beweis: Mit Hilfe des Spiegelungsprinzips ergibt sich

$$\begin{aligned}
b(N) &= \sum_{c=0}^{N} \#(\text{Pfade von } (0,0) \text{ nach } (N,c) \text{ die } (\cdot,-1) \text{ nicht treffen}) \\
&= \sum_{c=0}^{N} \left[\#(\text{Pfade von } (0,0) \text{ nach } (N,c)) - \#(\text{Pfade von } (0,-2) \text{ nach } (N,c))\right] \\
&= \sum_{c=0}^{N} \#(\text{Pfade von } (0,0) \text{ nach } (N,c)) - \sum_{c=0}^{N-2} \#(\text{Pfade von } (0,-2) \text{ nach } (N,c)) \\
&= \sum_{c=0}^{N} \#(\text{Pfade von } (0,0) \text{ nach } (N,c)) - \sum_{c=0}^{N-2} \#(\text{Pfade von } (0,0) \text{ nach } (N,c+2)) \\
&= \sum_{c=0}^{N} \#(\text{Pfade von } (0,0) \text{ nach } (N,c)) - \sum_{c=2}^{N} \#(\text{Pfade von } (0,0) \text{ nach } (N,c)) \\
&= \sum_{c=0}^{1} \#(\text{Pfade von } (0,0) \text{ nach } (N,c)).
\end{aligned}$$

Es gilt also $b(N) = a(N,0,0) + a(N,0,1)$ mit den entsprechenden Darstellungen aus Lemma 2.32. □

Aus dem anfangs beschriebenen Zusammenhang zwischen dem Zustandsraum Δ und den Pfaden in $\mathcal{X}_N$ ergibt sich also

Satz 2.34 *Sei $A = a_1 a_2 ... a_N$ ein Pattern der Länge N. Dann gilt für den Zustandsraum Δ des Spaltenprozesses $D = (D_n)_{n\in\mathbb{N}_0}$*

$$\#\Delta \leq b(N).$$

Diese Obergrenze können wir nochmals reduzieren. Dazu machen wir folgende Beobachtung: $d(N,n)$ ist genau dann Null, wenn der Pattern A in der Zeichenkette X zum Zeitpunkt n (fehlerfrei) erscheint. Daraus ergibt sich mit Satz 2.7, dass es in Δ nur einen einzigen Zustand $d = (d_0, ..., d_N)$ mit $d_N = 0$ geben kann, nämlich denjenigen, der durch das Erscheinen von A in X erzeugt wird. Weiterhin gilt

$d(i,n) = \text{msed}(A_{1:i}, X_{1:n})$, also sind für alle Zustände $d \in \Delta$ mit $d_i = 0$ bereits sämtliche Komponenten $d_0, ..., d_{i-1}$ durch das Erscheinen von $A_{1:i}$ in X eindeutig festgelegt. Aus dieser Überlegung ergibt sich unsere nächste Abschätzung:

Satz 2.35 *Sei $A = a_1 a_2 ... a_N$ ein Pattern der Länge N. Dann gilt für den Zustandsraum Δ des Spaltenprozesses $D = (D_n)_{n \in \mathbb{N}_0}$*

$$\#\Delta \leq c(N)$$

mit $c(N) = 2 + \sum_{h=1}^{N-1} b(h)$.

Beweis: Aus der obigen Überlegung ergibt sich, dass von allen Pfaden von $(0,0)$ nach $(N, \mathcal{C})$, die $(\cdot, -1)$ niemals und $(\cdot, 0)$ letztmalig in $(h,0)$ treffen, nur solche mit demselben Anfangsstück $(0,0)$, ..., $(h,0)$ einem Zustand in Δ entsprechen können. Es gilt also

$$\begin{aligned}
\#\Delta \;&\leq\; 1 + \sum_{c=1}^{N} \sum_{h=0}^{N-1} \#(\text{Pfade von } (h+1,1) \text{ nach } (N,c) \text{ die } (\cdot,0) \text{ nicht treffen}) \\
&=\; 2 + \sum_{c=1}^{N} \sum_{h=0}^{N-2} \#(\text{Pfade von } (h+1,1) \text{ nach } (N,c) \text{ die } (\cdot,0) \text{ nicht treffen}) \\
&=\; 2 + \sum_{h=0}^{N-2} \sum_{c=1}^{N} \#(\text{Pfade von } (0,1) \text{ nach } (N-h-1,c) \text{ die } (\cdot,0) \text{ nicht treffen}) \\
&=\; 2 + \sum_{h=1}^{N-1} \sum_{c=1}^{h+1} \#(\text{Pfade von } (0,1) \text{ nach } (h,c) \text{ die } (\cdot,0) \text{ nicht treffen}) \\
&=\; 2 + \sum_{h=1}^{N-1} \sum_{c=0}^{h} \#(\text{Pfade von } (0,0) \text{ nach } (h,c) \text{ die } (\cdot,-1) \text{ nicht treffen}) \\
&=\; 2 + \sum_{h=1}^{N-1} b(h).
\end{aligned}$$

□

Es ist also möglich, die Größe des Zustandsraums Δ durch $b(N)$ bzw. $c(N)$ nach oben abzuschätzen. Alle diese Abschätzungen machen jedoch keinerlei Gebrauch von der konkreten Gestalt des Pattern A, z.B. von dessen Überlappungseigenschaften oder Ähnlichem. Deshalb werden die tatsächlichen Zustandsräume in konkreten Anwendungssituationen eher noch wesentlich kleiner sein.

Die Tabellen auf den folgenden beiden Seiten geben empirisch gewonnene Werte für verschiedene Patternlängen und Alphabete wieder. Die Daten deuten darauf

hin, dass sich die Umfänge der Zustandsräume in etwa von der Größenordnung $C \cdot 2^N$, $C \in \mathbb{R}$, entwickeln. Im Gegensatz dazu zeigen die theoretischen Oberschranken eher ein Verhalten wie $C \cdot 3^N$, $C \in (0, 1)$ (vgl. Abbildung 2.5).

Unabhängig von den exakten Werten für $\#\Delta$ können wir festhalten, dass es uns gelungen ist, das $APMP$ mit Hilfe des Spaltenprozesses als Eintrittszeitproblem für eine Markov-Kette zu formulieren, deren Zustandsraum hinreichend klein ist, um daraus in praxisrelevanten Beispielen mit vertretbarem Aufwand konkrete Näherungen für die zugehörige Wartezeit zu berechnen.

	#Σ	1			2			3		
	2	3	3	3.0	4	4	4.0	4	4	4.0
	3	4	4	4.0	8	7	7.5	9	8	8.7
	4	5	5	5.0	16	11	13.4	21	16	18.1
	5	6	6	6.0	32	18	23.4	46	30	35.9
	6	7	7	7.0	64	28	40.1	100	53	70.0
	7	8	8	8.0	128	41	67.9	216	84	134.0
N	8	9	9	9.0	256	61	113.7	461	136	252.1
	9	10	10	10.0	512	87	188.4	974	214	466.5
	10	11	11	11.0	1024	127	309.5	2043	329	852.4
	11	12	12	12.0	2048	179	504.6	4259	512	1539.1
	12	13	13	13.0	4096	258	816.4	8829	777	2749.1
	13	14	14	14.0	8192	361	1312.4	18214	1164	4862.5
	14	15	15	15.0	16384	517	2096.8	37420	1766	8522.7

Tabelle 2.1: Maximal-, Minimal- und Durchschnittswerte für die Größe des Zustandsraums Δ bei allen Pattern der Länge N über einem Alphabet der Größe $\#\Sigma = 1, 2, 3$.

	#Σ	4			5			6		
	2	4	4	4.0	4	4	4.0	4	4	4.0
	3	9	8	8.8	9	8	8.8	9	8	8.8
	4	22	16	19.5	22	16	19.8	22	16	20.1
	5	49	30	41.2	49	30	42.7	49	30	43.4
	6	105	57	86.7	106	57	92.9	108	57	95.3
N	7	230	102	178.6	236	102	198.8	241	102	207.0
	8	505	187	360.3	519	187	418.5	526	187	444.8
	9	1098	326	712.8	1133	326	865.8	1144	326	942.5
	10	2358	587	1386.3	2456	587	1762.5	2499	587	1973.2
	11	5021	1010	2653.0	5301	1010	3528.5	5502	1010	4073.1
	12	10619	1690	5005.1	11359	1803	6956.6			

Tabelle 2.2: Maximal-, Minimal- und Durchschnittswerte für die Größe des Zustandsraums Δ bei allen Pattern der Länge N über einem Alphabet der Größe $\#\Sigma = 4, 5, 6$.

	#Σ	7			8			9		
	2	4	4	4.0	4	4	4.0	4	4	4.0
	3	9	8	8.9	9	8	8.9	9	8	8.9
	4	22	16	20.2	22	16	20.3	22	16	20.4
	5	49	30	43.8	49	30	44.1	49	30	44.4
N	6	108	57	96.9	108	57	98.1	108	57	99.0
	7	241	102	211.7	241	102	215.1	241	102	217.8
	8	540	187	458.8	543	187	468.6	543	187	476.2
	9	1172	326	983.1	1212	326	1010.2	1215	326	1030.9
	10	2582	587	2088.1	2635	587	2162.0	2717	587	2217.7

Tabelle 2.3: Maximal-, Minimal- und Durchschnittswerte für die Größe des Zustandsraums Δ bei allen Pattern der Länge N über einem Alphabet der Größe $\#\Sigma = 7, 8, 9$.

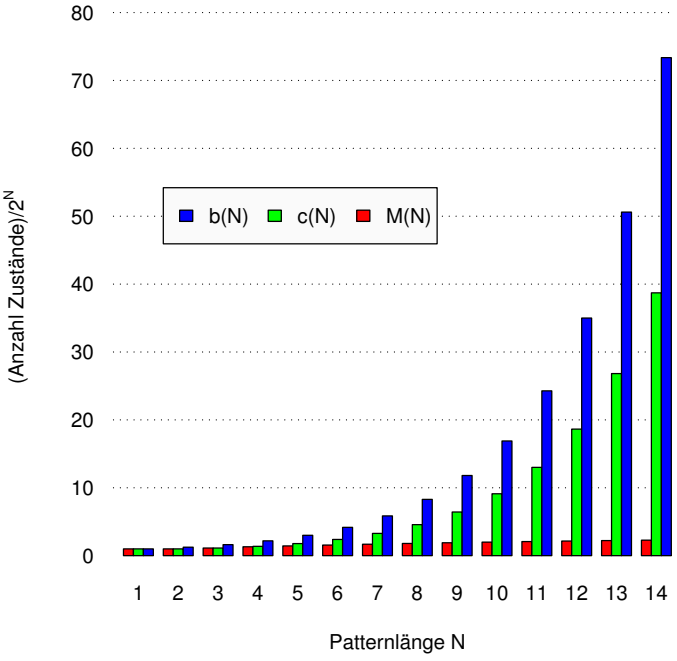

Abbildung 2.5: Darstellung der maximalen Größe $M(N)$ des Zustandsraums Δ bei $\#\Sigma = 3$ und der Oberschranken $b(N)$ und $c(N)$ bei Pattern A der Länge N.

2.5 Approximation des *Approximate Pattern Matching Problems* durch eine geometrische Verteilung

Mit Hilfe des Spaltenprozesses ist es nun also möglich, die Wartezeit, bis ein Pattern in einer zufälligen Zeichenkette erstmals „näherungsweise“ erscheint, durch eine geometrische Verteilung zu approximieren und dabei zugleich Abschätzungen für den gemachten Fehler zu erhalten. Wir unterscheiden wieder die Fälle, dass die zufällige Zeichenkette X durch eine Folge von unabhängigen und identisch verteilten Zufallsgrößen oder durch eine Markov-Kette erzeugt wird.

Ist $X = (X_n)_{n\in\mathbb{N}}$ eine Folge von unabhängigen und identisch verteilten Zufallsgrößen über einem endlichen Alphabet Σ mit $p_\sigma > 0$ für alle $\sigma \in \Sigma$, $A = a_1a_2...a_N$ ein endlicher Pattern über Σ, $D = (D_n)_{n\in\mathbb{N}_0}$ der zugehörige Spaltenprozess auf dem Zustandsraum Δ, $k \in \{0, 1, ..., N\}$ und $\Delta_k := \big\{d = (d_0, ..., d_N) \in \Delta : d_N \leq k\big\}$, so kann man, wie gezeigt,

$$T := \inf\big\{n \in \mathbb{N} : \mathrm{msed}(A, X_{1:n}) \leq k\big\}$$

als Eintrittszeit der Markov-Kette D in Δ_k interpretieren:

$$T = \inf\{n \in \mathbb{N} : D_n \in \Delta_k\},$$

wobei D der Startverteilung $P(D_0 = d^*) = 1$ mit $d^* = (0, 1, ..., N)$ genügt.

Ähnliches gilt, wenn $X = (X_n)_{n\in\mathbb{N}_0}$ eine homogene, aperiodische und positiv rekurrente Markov-Kette auf Σ mit Übergangsmatrix P, Startverteilung ν und stationärer Verteilung π ist. In dieser Situation haben wir gezeigt, dass der (modifizierte) Spaltenprozess $\tilde{D} = (\tilde{D}_n)_{n\in\mathbb{N}_0}$ mit $\tilde{D}_n := (X_n, D_n)$ eine Markov-Kette bildet. Der zugehörige Zustandsraum war $\tilde{\Delta}$. Für $k \in \{0, 1, ..., N\}$ definieren wir $\tilde{\Delta}_k := \big\{\tilde{d} := (x, d_0, d_1, ..., d_N) \in \tilde{\Delta} : d_N \leq k\big\}$. T ist dann

$$T = \inf\{n \in \mathbb{N} : \tilde{D}_n \in \tilde{\Delta}_k\},$$

wobei $\tilde{D}$ der auf (Σ, d^*) konzentrierten Startverteilung auf $\tilde{\Delta}$ mit $P\big(\tilde{D}_0 = (\sigma, d^*)\big) := \nu(\sigma)$ genügt.

In der Regel wird der Totalvariationsabstand zwischen der Startverteilung und der Verteilung ρ, unter der T geometrisch verteilt ist, recht groß sein, so dass man darauf

angewiesen ist, mit Hilfe von Satz 2.20 eine gute Näherung für die Verteilung von T zu gewinnen.

Im Fall einer unabhängigen und identisch verteilten Zeichenkette gibt es darüber hinaus aber eine weitere Methode, um zu guten Ergebnissen zu gelangen. In dieser Situation hatten wir nachgewiesen, dass D bereits nach $2N$ Zeitschritten stationär wird. Dies, zusammen mit der im Zusammenhang mit Satz 2.20 begründeten Tatsache, dass die stationäre Verteilung μ von D in der Regel eine verhältnismäßig gute Näherung für ρ darstellen wird, motiviert uns zu der folgenden Betrachtung:

Es gilt $P_\lambda(D_{2N} = d) = \mu(d)$ für alle $d \in \Delta$ und jede Startverteilung λ auf Δ. Daraus folgt für alle $m \in \mathbb{N}$

$$P_\mu(T = m) = P_\lambda(D_{2N}, ..., D_{2N+m-1} \notin \Delta_k,\ D_{2N+m} \in \Delta_k). \tag{2.6}$$

Wir definieren nun T^* durch

$$T^* := \inf\{n \geq 2N : D_n \in \Delta_k\}.$$

Dann gilt

$$T^* = 2N + m \quad \Leftrightarrow \quad D_{2N}, ..., D_{2N+m-1} \notin \Delta_k,\ D_{2N+m} \in \Delta_k,$$

so dass sich mit (2.6)

$$P_\mu(T = m) = P_\lambda(T^* = 2N + m) \tag{2.7}$$

für alle $m \in \mathbb{N}$ und jede Startverteilung λ auf Δ ergibt.

Offensichtlich gilt nun $(T > 2N + m) \Rightarrow (T^* > 2N + m)$, also ist

$$P_\lambda(T^* > 2N + m) \geq P_\lambda(T > 2N + m). \tag{2.8}$$

Umgekehrt gilt $(T^* > 2N + m, T \geq 2N) \Rightarrow (T > 2N + m)$, und damit

$$\begin{aligned} P_\lambda(T > 2N + m) &\geq P_\lambda(T^* > 2N + m, T \geq 2N) \\ &= P_\lambda(T^* > 2N + m) + P_\lambda(T \geq 2N) \\ &\quad - P_\lambda(T^* > 2N + m \text{ oder } T \geq 2N), \end{aligned}$$

also

$$P_\lambda(T > 2N + m) \geq P_\lambda(T^* > 2N + m) - P_\lambda(T < 2N). \tag{2.9}$$

(2.7), (2.8) und (2.9) liefern dann zusammen

$$P_\mu(T > m) - P_\lambda(T < 2N) \leq P_\lambda(T > 2N + m) \leq P_\mu(T > m). \tag{2.10}$$

So erhält man für die Wartezeit T des *Approximate Pattern Matching Problems* das folgende Theorem.

Theorem 2.36 *Unter allen getroffenen Bezeichnungen und Voraussetzungen sei* $k \in \{0, 1, ..., N\}$, λ^* *das auf* d^* *konzentrierte Einpunktmaß auf* Δ *und* $T := \inf\{n \in \mathbb{N} : D_n \in \Delta_k\}$. *Dann gilt für alle* $m \in \mathbb{N}$

$$\alpha(m) \leq P_{\lambda^*}(T > 2N + m) \leq \beta(m),$$

mit

$$\begin{aligned} \alpha(m) &= P_\rho(T > m) - \mathrm{d_{TV}}(\rho, \mu) - P_{\lambda^*}(T < 2N), \\ \beta(m) &= P_\rho(T > m) + \mathrm{d_{TV}}(\rho, \mu). \end{aligned}$$

Beweis: Folgt aus Theorem 2.19 zusammen mit (2.10). □

Wie geht man also in der Praxis vor, um die Wartezeit T des *Approximate Pattern Matching Problems* näherungsweise zu bestimmen?

In beiden hier beschriebenen Situationen bestimmt man zunächst mit Hilfe einer Tiefensuche den Zustandsraum und die Übergangsmatrix des Spaltenprozesses D bzw. $\tilde{D}$, die wir hier mit Q bzw. $\tilde{Q}$ bezeichnen wollen.

Im Fall einer unabhängigen und identisch verteilten Zeichenkette X beginnt man mit dem Startvektor $x_0 := \mathbf{1}_{d^*}$ ($\mathbf{1}_{d^*}(d^*) = 1$ und 0 sonst) und erhält durch $2N$ Matrix-Vektor-Multiplikation

$$x_{n+1} = x_n \cdot Q$$

die stationäre Verteilung $\mu = x_0 \cdot Q^{2N}$. Schränkt man x_0 auf $(\Delta_k)^\mathrm{c}$ ein, so gewinnt man ebenso durch $2N - 1$ Matrix-Vektor-Multiplikationen $P_{\lambda^*}(T < 2N) = 1 - P_{\lambda^*}(T > 2N - 1) = 1 - x_0 \cdot Q^{2N-1}_{(\Delta_k)^\mathrm{c}} \cdot \mathbf{1}$ mit $\mathbf{1} := (1, ..., 1)^\mathrm{t}$. Die Verteilung ρ erhält man schließlich als auf Länge 1 normierten Linkseigenvektor zum betragsgrößten und reellwertigen Perron-Eigenwert η_0 von $Q_{(\Delta_k)^\mathrm{c}}$. Hierfür stehen diverse numerische Methoden zur Verfügung, beispielsweise Matrixiterationsmethoden oder die inverse Iteration nach Wieland.

Mit μ, ρ und $P_{\lambda^*}(T < 2N)$ hat man dann alle benötigten Parameter, um für beliebiges $m \in \mathbb{N}$ die Näherungsschranken $\alpha(m)$ und $\beta(m)$ aus Theorem 2.36 zu berechnen. $P_\rho(T > m)$ ist einfach der Tail der geometrischen Verteilung von T unter ρ, die zugehörige Erfolgswahrscheinlichkeit erhält man durch die Vektor-Matrix-Vektor-Multiplikation $P_\rho(T = 1) = \rho \cdot Q \cdot \mathbf{1}_{\Delta_k}$ ($\mathbf{1}_{\Delta_k}(d) = 1$ für $d \in \Delta_k$ und 0 sonst).

Ist man in der Situation, dass die Zeichenkette X durch eine Markov-Kette generiert wird, versucht man mit Hilfe von Satz 2.20 eine geeignete Näherung für die Verteilung von T zu bestimmen. λ entspricht hier der zuvor angegebenen, auf (Σ, d^*) konzentrierten Startverteilung λ^* von $\tilde{D}$. Für jedes $n_0 \in \mathbb{N}$ erhält man durch n_0 Matrix-Vektor-Multiplikationen den Vektor $x_{n_0} := \lambda^* \cdot \left(\tilde{Q}_{(\tilde{\Delta}_k)^c}\right)^{n_0}$. Dann ist $P_{\lambda^*}(T > n_0) = x_{n_0} \cdot \mathbf{1}$ ($\mathbf{1} := (1, ..., 1)^t$) und $\lambda_{n_0} = x_{n_0}/P_{\lambda^*}(T > n_0)$. ρ erhält man wie zuvor als auf Länge 1 normierten Linkseigenvektor zum Perron-Eigenwert η_0 von $\tilde{Q}_{(\tilde{\Delta}_k)^c}$, $P_\rho(T > n)$ ist der Tail der geometrischen Verteilung von T unter ρ. Der Parameter n_0 ist in dieser Situation frei wählbar. Man sollte ihn so bestimmen, dass Satz 2.20 möglichst gute Näherungen liefert.

2.6 Zwei Beispiele

Die praktische Anwendbarkeit der hergeleiteten Verfahren wollen wir in diesem Abschnitt anhand zweier Beispiele unterstreichen. Unter den verschiedenen Anwendungsgebieten des *APMP* ist die Molekularbiologie wohl das bei Weitem wichtigste, und so wollen wir als erstes ein Beispiel aus diesem Bereich (speziell: aus dem Bereich der Gen-Sequenzierung) betrachten.

Durch die zunehmend vollständige Sequenzierung des genetischen Codes einer inzwischen unüberschaubaren Anzahl von Lebewesen ist es mittlerweile ohne größeren Aufwand möglich, sich über das Internet beliebig viele „Rohdaten“ zu beschaffen.

Als Beispiel betrachten wir hier die komplette Gensequenz des Ti-Plasmids „Agrobacterium tumefaciens plasmid pTi-SAKURA“. Auch wenn es für die weiteren Betrachtungen keinerlei Wichtigkeit besitzt, soll dem auf dem Gebiet der Bakteriologie unbewanderten Leser eine gewisse Vorstellung davon vermittelt werden, worum es sich hierbei handelt.

Agrobacterium bildet eine Gattung aerober, beweglicher Stäbchen innerhalb der Familie Rhizobiaceae. Ihr natürlicher Lebensraum ist die Rhizosphäre von Pflanzen. Sie kommen weltweit im Erdboden in einer Häufigkeit von bis zu 500 Bakterien pro Gramm Boden vor. Optimale Lebensbedingungen finden sie in einem Temperaturbereich von 20 bis 28 Grad Celsius. Agrobacterium tumefaciens ist eine von 8 Spezies innerhalb dieser Gattung. Dieses Bakterium ist pflanzenpathogen (d.h. krankheitserregend) mit einem für Pflanzen onkogenen (d.h. krebserzeugenden) Potential. Mit Hilfe des großen Tumor-induzierenden Ti-Plasmids (ca. 200 kb) können A. tumefaciens Bakterien Pflanzenzellen zu autonom wuchernden Tumorzellen transformieren. Der Wirtsbereich von A. tumefaciens ist sehr weit und umfasst 640 Pflanzenarten aus 93 Familien. A. tumefaciens hat eine besondere Bedeutung für das *Genetic Engineering* von Pflanzen erlangt. Das Bakterium erwies sich als sehr geeignetes natürliches Vehikel für gentechnische Veränderungen.

(Die hier vorgestellten Fakten sind [ZKB97] entnommen. Den vollständigen genetischen Code des Ti-Plasmids „Agrobacterium tumefaciens plasmid pTi-SAKURA“ findet man beispielsweise unter [BI04].)

Insgesamt umfasst der genetische Code dieses Ti-Plasmids 206479 Basenpaare. In unserem mathematischen Modell wollen wir davon ausgehen, dass es sich hierbei

entweder um eine Realisation einer unabhängigen und identisch verteilten Folge von Zufallsgrößen oder um eine Markov-Kette über dem Alphabet $\{a, g, c, t\}$ handelt. Die zugehörige Verteilung bzw. Übergangsmatrix approximieren wir dabei durch die jeweiligen empirischen Werte, die man aus der Sequenz des Ti-Plasmids erhält.

Als Beispiel für eine interessierende Gensequenz wählen wir willkürlich den Pattern

$$agcttcgcaa$$

der Länge 10 und fragen nach der Wartezeit, bis diese Sequenz erstmals mit einem vorgegebenen Editierdistanz-Fehler in der Zeichenkette erscheint.

Zur (approximativen) Beantwortung dieser Frage verwenden wir im Fall einer unabhängigen und identisch verteilten Zeichenkette Theorem 2.36, im Fall einer Markov-Kette Satz 2.20, wobei hier noch der Zeitpunkt n_0 frei wählbar ist. Natürlich ist die Bestimmung des Zustandsraums Δ bzw. $\tilde{\Delta}$ in einer solchen Situation nicht mehr ohne Hilfe eines Rechners möglich. In Anhang C befindet sich der Quellcode eines `C`-Programms, das Theorem 2.36 für den Fall einer unabhängigen und identisch verteilten Zeichenkette praktisch nutzbar macht. Natürlich ist es auch nicht weiter schwierig, ein entsprechendes Programm für den Fall einer Markov-Kette als Grundlage der Zeichenkette zu schreiben, das entsprechend die für Satz 2.20 wichtigen Kenngrößen berechnet.

In den Tabellen 2.4 und 2.5 findet man die Resultate, die diese beiden Programme für den hier betrachteten Pattern liefern. Generell können wir dabei festhalten, dass wir in beiden Fällen sehr brauchbare Approximationen der Wartezeit T des *APMP* erhalten, wenn der zugelassene Fehler in der Editierdistanz „hinreichend“ klein ist im Vergleich zur Länge des Patterns. (Es ist klar, dass man bei einem so kurzen Pattern und großen Editierdistanzen, also einer großen Zielmenge, keine gute Näherung von T durch eine geometrische Verteilung erwarten darf.)

Aus allen Ergebnissen wollen wir uns ein Beispiel herausgreifen: Die Zeichenkette sei durch eine Markov-Kette generiert, die zugehörige Übergangsmatrix findet man in Tabelle 2.5. Der zum Zielpattern gehörige Zustandsraum umfasst 1249 Zustände, für die Schranke n_0 erweist sich ein Wert von 20 als gute Wahl für brauchbare Ergebnisse. Ist T die Wartezeit, bis der Zielpattern erstmals mit einem maximalen Editierdistanz-Fehler von 1 in der Zeichenkette erscheint, so liefert das Programm die Werte $E_\rho T = P_\rho(T = 1)^{-1} \approx 1.1514 \cdot 10^4$, $\mathrm{d_{TV}}(\rho, \lambda_{20}) \approx 1.6979 \cdot 10^{-6}$ und $P_{\lambda^*}(T >$

$n_0 = 20) \approx 0.998998$. Mit Hilfe von Satz 2.20 erhalten wir also die Approximation

$$\alpha(n) \leq P(T > 20 + n) \leq \beta(n)$$

mit

$$\begin{aligned} \alpha(n) &= \left((1 - 1/(1.1514 \cdot 10^4))^n - 1.6979 \cdot 10^{-6}\right) \cdot 0.998998 \quad \text{und} \\ \beta(n) &= \left((1 - 1/(1.1514 \cdot 10^4))^n + 1.6979 \cdot 10^{-6}\right) \cdot 0.998998. \end{aligned}$$

Die Wartezeit für den betrachteten Pattern bei vorgegebener Editierdistanzschranke 1 ist also näherungsweise geometrisch verteilt mit Erwartungswert $E_\rho T \approx 1.1514 \cdot 10^4$. Insbesondere ist zu erwarten, dass der Pattern in etwa $206479/11514 \approx 18$ Mal überlappungsfrei im genetischen Code des Ti-Plasmids (mit höchstens einem Fehler) auftauchen wird.

Führt man allerdings eine komplette Analyse des Ti-Plasmids durch, so erhält man insgesamt 40 Positionen, an denen die Sequenz auftaucht, davon 37 überlappungsfreie Positionen. Die von uns betrachtete Sequenz scheint also (signifikant?) häufiger im betrachteten Ti-Plasmid vorzukommen als erwartet. Kann man hierzu auch eine exakte mathematische Angabe machen?

Wenn die Wartezeit T bis zum ersten Erscheinen der Sequenz näherungsweise geometrisch verteilt ist, so ist die Wartezeit T_{37} bis zum 37. überlappungsfreien Erscheinen näherungsweise negativ binomialverteilt mit derselben Erfolgswahrscheinlichkeit. Wie groß ist die Wahrscheinlichkeit, dass dieses 37. überlappungsfreie Auftreten in der Markov-Kette vor der Position 206479 geschieht? Antwort:

$$P(T_{37} \leq 206479) \approx 5.336 \cdot 10^{-5}.$$

Das ist eine verhältnismäßig kleine Wahrscheinlichkeit, es spricht also vieles dafür, dass es sich bei der von uns willkürlich gewählten Sequenz nicht um „sinnloses Rauschen“ innerhalb des Ti-Plasmids handelt, sondern eher um ein Teilstück einer Gensequenz, die hierin signifikant häufig auftaucht.

```
Wartezeitanalyse fuer Pattern

Pattern: a g c t t c g c a a
Laenge des Pattern: 10

Alphabet: a g c t
Verteilung auf dem Alphabet: 0.220 0.282 0.278 0.220

Groesse des Zustandsraums: 1220
Gesamtlaufzeit des Programms: 0.25 Sek.

Tabelle der EWs und Fehlerschranken:
```

eddist	E_{rho}(T)	tvd(rho,mu)	P(T<2*Patternlaenge)
0	1.13358816e+006	4.47119618e-006	8.82150767e-006
1	2.00921469e+004	2.15128593e-004	5.21896152e-004
2	9.15577387e+002	4.03855359e-003	1.20175526e-002
3	8.35004988e+001	3.62854405e-002	1.32272186e-001
4	1.39146770e+001	1.79100748e-001	6.14951975e-001
5	4.12593814e+000	5.49974008e-001	9.81621812e-001
6	1.96928635e+000	9.32934171e-001	9.99987239e-001
7	1.39273807e+000	9.99843686e-001	1.00000000e+000

Tabelle 2.4: Wartezeitanalyse für den Pattern *agcttcgcaa* im Modell einer unabhängigen und identisch verteilten Zeichenkette auf dem Alphabet $\{a, g, c, t\}$. Es wird die empirische Verteilung des Ti-Plasmids „Agrobacterium tumefaciens plasmid pTi-SAKURA“ zu Grunde gelegt.

```
Wartezeitanalyse fuer Pattern

Pattern: a g c t t c g c a a
Laenge des Pattern: 10

Alphabet: a g c t
Uebergangsmatrix:  0.261 0.250 0.223 0.266
                   0.263 0.243 0.319 0.175
                   0.220 0.345 0.238 0.197
                   0.127 0.283 0.331 0.259

Groesse des Zustandsraums: 1249
n_0: 20
Gesamtlaufzeit des Programms: 0.30 Sek.

Tabelle der EWs und Fehlerschranken:

eddist E_{rho}(T)        tvd(rho,lam_{n_0}) P(T>n_0)
============================================================
 0     5.32265325e+005   5.23886112e-008    9.99979297e-001
 1     1.15142395e+004   1.69791647e-006    9.98997709e-001
 2     6.22525640e+002   1.99121583e-005    9.80626990e-001
 3     6.49350460e+001   1.63376738e-004    8.18431455e-001
 4     1.18838266e+001   1.15624943e-003    2.94716801e-001
 5     3.75768014e+000   3.49917723e-003    8.37993915e-003
 6     1.87330566e+000   6.83199626e-004    2.30398396e-006
 7     1.34952767e+000   7.03422976e-001    4.73177669e-012
```

Tabelle 2.5: Wartezeitanalyse für den Pattern *agcttcgcaa* im Modell einer Markov-Kette auf dem Alphabet $\{a, g, c, t\}$. Es wird die empirische Übergangsmatrix des Ti-Plasmids „Agrobacterium tumefaciens plasmid pTi-SAKURA“ zu Grunde gelegt.

Als zweites Anwendungsbeispiel wollen wir noch einmal auf das in Beispiel 2.13 eingeführte *Monkey Typing Shakespeare* Phänomen zurückkommen.

Dort hatten wir die Wartezeit T betrachtet, bis ein Affe durch zufälliges Tippen auf einer Schreibmaschine erstmals näherungsweise des Shakespeare-Zitat

$$TOBEORNOTTOBE$$

schreiben wird. Mit unserem damaligen Kenntnisstand war es uns allerdings nur möglich, die qualitative Aussage zu machen, dass T nach Normierung auf Erwartungswert 1 näherungsweise Exp(1)-verteilt ist, ohne jedoch weitere Aussagen über die Güte dieser Approximation angeben zu können.

Mit Hilfe der im vorangegangenen Abschnitt hergeleiteten Resultate ist es nun möglich, konkrete Fehlerschranken anzugeben, wenn man T durch eine geometrische Verteilung approximiert.

Mit den schon im Gensequenzierungsbeispiel verwendeten Programmen untersuchen wir insgesamt vier Modellannahmen. In den ersten drei Modellen nehmen wir an, dass es sich bei der vom Affen erzeugten Zeichenkette um eine unabhängige und identisch verteilte Zeichenkette über dem Alphabet der 26 Buchstaben $\{A, B, ..., Z\}$ handelt, und dass wir warten, bis in dieser Zeichenkette erstmals der Pattern

$$TOBEORNOTTOBE$$

mit einem vorgegebenen Fehler k in der *Edit Distance* erscheint. Dabei machen wir verschiedene Annahmen über die zu Grunde liegende Verteilung.

In Modell Nr. 1 gehen wir davon aus, dass wir es mit einem vollkommen willkürlich agierenden Affen zu tun haben, der jeden der 26 Buchstaben mit derselben Wahrscheinlichkeit wählt. Wir betrachten also die Gleichverteilung auf dem Alphabet.

Modell Nr. 2 untersucht ein Phänomen, das auch häufig unter Lottospielern anzutreffen ist, wenn sie versuchen, einen Lottoschein „rein zufällig“ auszufüllen. In der Regel führt dies dazu, dass ein solcher Spieler eher Felder in der Mitte des Spielscheins ankreuzen wird als am Rand (vgl. [HR98]). Dasselbe Verhalten wollen wir unserem Affen unterstellen. Die auf einer $QWERTY$-Tastatur weit außen liegenden Tasten Q, A, Z, W, I, M, L, O, P wird er eher selten, und zwar jeweils mit der Wahrscheinlichkeit $1/53$ wählen. Die weiter innen liegenden Tasten E, S, X, R, D, C, U, K, N wählt er doppelt so häufig mit der Wahrscheinlichkeit $2/53$. Die schon

recht mittig gelegenen Tasten F, T, Y, J, B, V werden mit der Wahrscheinlichkeit $3/53$ gewählt und schließlich die zentralen Tasten G, H mit der größten Wahrscheinlichkeit $4/53$.

In Modell Nr. 3 haben wir es mit einem „verhältnismäßig intelligenten“ Affen zu tun, der die Buchstaben gemäß der Häufigkeit wählt, mit der sie tatsächlich in Shakespeares „Hamlet“ vorkommen. Dazu haben wir aus diesem Werk alle Leer- und Sonderzeichen sowie alle Regieanweisungen entfernt, so dass schließlich nur noch ein langer Textstring von 120.019 Buchstaben übrig blieb. Anschließend haben wir dann die empirische Verteilung der Buchstaben bestimmt.

All diesen Modellen ist gemeinsam, dass wir im zu Grunde liegenden Alphabet alle 20 nicht im Zielpattern vorkommenden Buchstaben zu einem neuen Buchstaben zusammenfassen können, ohne dass sich an der Wartezeitverteilung etwas ändert.

Das Modell Nr. 4, das wir betrachten wollen, geht schließlich davon aus, dass die Zeichenkette des Affen nicht unabhängig und identisch verteilt erzeugt wird, sondern durch eine Markov-Kette. Als zu Grunde liegende Übergangsmatrix verwenden wir dabei die empirische Übergangsmatrix, die sich aus Shakespeares „Hamlet“ extrahieren lässt. Dadurch erhalten wir eine relativ guten Annäherung an die englische Sprache. So beträgt beispielsweise die Übergangswahrscheinlichkeit von Q nach U 1, abstruse Kombinationen kommen gar nicht erst vor. Und da unser Zielpattern selbst aus einigen der häufigsten englischen Wörter besteht (die Wahrscheinlichkeit für den Übergang von B nach E beträgt beispielsweise fast 34%), kann man erwarten, dass es unter diesen Annahmen tatsächlich weniger lang dauern sollte, bis der Pattern in der Zeichenkette erscheint. Alles in allem haben wir es in diesem Modell mit einem Affen zu tun, der näherungsweise über Kenntnisse der englischen Sprache verfügt.

Für die ersten drei Modelle können wir die in Theorem 2.36 angegebene Approximation der Wartezeit berechnen. Im vierten Modell, das auf einer Markov-Kette beruht, sind wir hingegen gezwungen, eine Näherung gemäß Satz 2.20 zu versuchen, wobei hier der Zeitpunkt n_0 noch frei wählbar ist. Es stellt sich heraus, dass beispielsweise das Doppelte der Länge des Zielpatterns, also $n_0 = 26$ gute Approximationen liefert.

Die Resultate der Approximationen unter den verschiedenen Modellannahmen sind in den folgenden Tabellen wiedergegeben. In allen Modellen bekommen wir wiederum ausgezeichnete Approximationen der gesuchten Wartezeitverteilungen, wenn der zugelassene Editierdistanzfehler im Verhältnis zur Patternlänge klein ist.

Darüber hinaus bestätigen die Resultate die intuitive Vermutung, dass die Wartezeit T im Mittel um so kleiner ausfallen wird, je besser das verwendete Modell die englische Sprache modelliert. (Man vergleiche die Reduzierung des Erwartungswerts der Wartezeit in den Tabellen 2.6, 2.8 und 2.9.)

Bemerkung: Der an diesem Thema interessierte Leser sollte einmal den Suchbegriff „Monkey Typing Shakespeare“ in eine Internet-Suchmaschine eingeben. Er wird eine reiche Auswahl an mehr oder weniger informativen und häufig sehr unterhaltsamen Treffern erhalten.

```
Wartezeitanalyse fuer Pattern

Pattern: T O B E O R N O T T O B E
Laenge des Pattern: 13

Alphabet: B E N O R T X
Verteilung auf dem Alphabet:
0.038 0.038 0.038 0.038 0.038 0.038 0.769

Groesse des Zustandsraums: 17221
Gesamtlaufzeit des Programms: 18.88 Sek.

Tabelle der EWs und Fehlerschranken:
```

eddist	E_{rho}(T)	tvd(rho,mu)	P(T<2*Patternlaenge)
0	2.48115158e+018	1.60178173e-014	3.15303339e-014
1	4.18054201e+015	1.79678035e-014	3.39728246e-014
2	1.61234648e+013	6.82426086e-013	8.91398066e-013
3	1.08589895e+011	9.61456329e-011	1.34204203e-010
4	1.15934597e+009	8.54581659e-009	1.30488716e-008
5	1.88942955e+007	4.86564312e-007	8.31512435e-007
6	4.67800124e+005	1.78672174e-005	3.49317502e-005
7	1.80340357e+004	4.17252203e-004	9.46776813e-004
8	1.15022259e+003	5.86877450e-003	1.56027055e-002
9	1.31299493e+002	4.63635883e-002	1.38541670e-001
10	2.68021472e+001	2.07169654e-001	5.53612036e-001

Tabelle 2.6: Wartezeitanalyse für den Pattern *TOBEORNOTTOBE* im Modell Nr. 1 einer unabhängigen und identisch verteilten Zeichenkette mit Gleichverteilung auf dem zu Grunde liegenden Alphabet.

```
Wartezeitanalyse fuer Pattern

Pattern: T O B E O R N O T T O B E
Laenge des Pattern: 13

Alphabet: B E N O R T X
Verteilung auf dem Alphabet:
0.057 0.038 0.038 0.019 0.038 0.057 0.755

Groesse des Zustandsraums: 17221
Gesamtlaufzeit des Programms: 18.03 Sek.

Tabelle der EWs und Fehlerschranken:
```

eddist	E_{rho}(T)	tvd(rho,mu)	P(T<2*Patternlaenge)
0	6.69669562e+018	7.27833826e-015	1.43218770e-014
1	8.99298949e+015	8.62616355e-015	1.62092562e-014
2	2.83039769e+013	3.83795811e-013	5.72875081e-013
3	1.59836630e+011	6.50171287e-011	1.03602793e-010
4	1.47437497e+009	6.70877554e-009	1.16011388e-008
5	2.14115470e+007	4.29668955e-007	8.25458855e-007
6	4.86089555e+005	1.70947946e-005	3.76331385e-005
7	1.75902780e+004	4.22046693e-004	1.08136460e-003
8	1.07260491e+003	6.20318004e-003	1.85289105e-002
9	1.19012145e+002	5.04690346e-002	1.66124160e-001
10	2.40555110e+001	2.25565647e-001	6.27647495e-001

Tabelle 2.7: Wartezeitanalyse für den Pattern *TOBEORNOTTOBE* im Modell Nr. 2 einer unabhängigen und identisch verteilten Zeichenkette mit einer Verteilung auf dem zu Grunde liegenden Alphabet, die Buchstaben in der Mitte einer *QWERTY*-Tastatur gegenüber Buchstaben am Rand bevorzugt.

```
Wartezeitanalyse fuer Pattern

Pattern: T O B E O R N O T T O B E
Laenge des Pattern: 13

Alphabet: B E N O R T X
Verteilung auf dem Alphabet:
0.015 0.117 0.062 0.087 0.058 0.093 0.568

Groesse des Zustandsraums: 17221
Gesamtlaufzeit des Programms: 19.38 Sek.

Tabelle der EWs und Fehlerschranken:
```

eddist	E_{rho}(T)	tvd(rho,mu)	P(T<2*Patternlaenge)
0	2.06628015e+015	1.15050002e-014	1.86517468e-014
1	4.23058998e+012	2.42368340e-012	3.20243831e-012
2	2.19105739e+010	4.35799708e-010	6.39619913e-010
3	2.19584242e+008	4.03468049e-008	6.62254832e-008
4	3.82517075e+006	2.13375363e-006	3.94672204e-006
5	1.10781794e+005	6.68012955e-005	1.41731457e-004
6	5.33067051e+003	1.23323451e-003	3.07780861e-003
7	4.44471218e+002	1.29510067e-002	3.85170904e-002
8	6.71403124e+001	7.66124438e-002	2.46124024e-001
9	1.75808159e+001	2.70671158e-001	6.98585042e-001
10	6.85420607e+000	5.94163636e-001	9.68820089e-001

Tabelle 2.8: Wartezeitanalyse für den Pattern *TOBEORNOTTOBE* im Modell Nr. 3 einer unabhängigen und identisch verteilten Zeichenkette mit einer Verteilung auf dem zu Grunde liegenden Alphabet, die der empirischen Verteilung der Buchstaben in Shakespeares „Hamlet“ entspricht.

```
Wartezeitanalyse fuer Pattern

Pattern: T O B E O R N O T T O B E
Laenge des Pattern: 13

Alphabet: A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
Uebergangsmatrix: 26x26

Groesse des Zustandsraums: 146251
n_0: 26
Gesamtlaufzeit des Programms: 10129.78 Sek.

Tabelle der EWs und Fehlerschranken:

eddist E_{rho}(T)        tvd(rho,lam_{n_0}) P(T>n_0)
============================================================
 0     6.74354224e+014   3.50423618e-012    1.00000000e+000
 1     1.98504161e+012   3.49766587e-012    1.00000000e+000
 2     1.37165840e+010   3.12522786e-012    9.99999999e-001
 3     1.67942225e+008   4.18437842e-011    9.99999909e-001
 4     3.31860796e+006   9.57544793e-010    9.99995250e-001
 5     1.02949142e+005   1.59095669e-008    9.99841469e-001
 6     5.07781801e+003   1.04121740e-006    9.96649893e-001
 7     4.18758969e+002   1.58614676e-005    9.57812959e-001
 8     6.13890822e+001   7.25905873e-005    7.27396155e-001
 9     1.57652423e+001   3.25845260e-004    2.51761443e-001
10     6.00659708e+000   4.39180806e-004    1.66102916e-002
```

Tabelle 2.9: Wartezeitanalyse für den Pattern *TOBEORNOTTOBE* im Modell Nr. 4 einer Markov-Kette auf dem zu Grunde liegenden Alphabet, wobei die Übergangswahrscheinlichkeiten den empirischen Werten entsprechen, die man aus Shakespeares „Hamlet“ gewinnt.

2.7 Poisson-Approximation für das *Approximate Pattern Matching Problem*

In diesem Kapitel wollen wir zum Abschluss eine weitere Methode vorstellen, mit der es möglich ist, Wartezeiten für Pattern in zufälligen Zeichenketten zu approximieren und gleichzeitig eine Abschätzung für den dabei gemachten Fehler zu bekommen.

Das wesentliche Hilfsmittel ist die so genannte Chen-Stein-Methode. Die Grundidee ist dabei, dass eine Summe der Gestalt $\sum_{n=1}^{m} \mathbb{1}_{A_n}$ näherungsweise mit Parameter $\lambda := \sum_{n=1}^{m} P(A_n)$ Poisson-verteilt ist, sofern die Ereignisse A_n sehr selten und näherungsweise unabhängig sind. Die Chen-Stein-Methode formuliert diesen heuristischen Ansatz aus und gibt explizite Fehlerschranken an, wie gut die gewonnene Approximation ist.

Theorem 2.37 (Chen-Stein-Methode) *Es sei $(A_n)_{n\in\mathbb{N}}$ eine Folge von Ereignissen, $I \subset \mathbb{N}$, $W := \sum_{n\in I} \mathbb{1}_{A_n}$ und Z eine Poisson-verteilte Zufallsvariable mit Parameter $\lambda = EW < \infty$. Für jedes $n \in I$ wählen wir eine „Abhängigkeitsumgebung“ J_n mit $n \in J_n$. Weiter definieren wir*

$$\begin{aligned} b_1 &:= \sum_{n\in I} \sum_{m\in J_n} P(A_n)P(A_m), \\ b_2 &:= \sum_{n\in I} \sum_{n\neq m\in J_n} P(A_n \cap A_m), \\ b_3 &:= \sum_{n\in I} E\left|E\left[\mathbb{1}_{A_n} - P(A_n)|\sigma\{A_m : m \notin J_n\}\right]\right|. \end{aligned}$$

Dann gilt

$$\mathrm{d_{TV}}\big(\mathcal{L}(W), \mathcal{L}(Z)\big) \leq 2(b_1 + b_2 + b_3)$$

und

$$|P(W = 0) - e^{-\lambda}| \leq \frac{1 - e^{-\lambda}}{\lambda}\,(b_1 + b_2 + b_3).$$

Beweis: Siehe [AGG89], Theorem 1, S. 11. Vgl. auch [AGG90]. □

Dieses Resultat werden wir nun auf das *APMP* anwenden. Der Einfachheit halber werden wir uns dabei auf den Fall einer unabhängigen und identisch verteilten Zeichenkette beschränken.

Es sei also wie zuvor Σ ein endliches Alphabet und $A = a_1 a_2 ... a_N \in \Sigma^N$ ein Pattern der Länge N über Σ. $X = (X_n)_{n \in \mathbb{N}}$ sei eine Folge von unabhängigen und identisch verteilten Zufallsgrößen über Σ (die zufällige Zeichenkette) und

$$T := \inf \big\{ n \in \mathbb{N} : \mathrm{msed}(A, X_{1:n}) \leq k \big\},$$

wobei $0 \leq k \leq N$ eine vorgegebene Fehlerschranke ist.

Wir definieren nun

$$S_n := \mathbb{1}\big\{ \mathrm{msed}(A, X_{1:n}) \leq k \big\}, \quad n \in \mathbb{N}.$$

(Der Einfachheit halber setzen wir außerdem $S_n := 0$ für $n \leq 0$.) Offenbar besteht dann folgender Zusammenhang zwischen T und diesen Indikatorfunktionen:

$$T > m \iff \sum_{n=1}^{m} S_n = 0,$$

und diese Darstellung ermöglicht uns die Anwendung der Chen-Stein-Methode. Mit ihr lässt sich nun zeigen, dass die Summe $\sum_{n=1}^{m} S_n$ näherungsweise Poisson-verteilt ist mit Parameter $\lambda = \sum_{n=1}^{m} P(\mathrm{msed}(A, X_{1:n}) \leq k)$.

Eine Voraussetzung für eine gute Approximation ist dabei allerdings, dass die zu den beteiligten Indikatorfunktionen gehörigen Ereignisse sehr selten sind. Dies ist bei den S_n zwar schon der Fall, die Ergebnisse lassen sich aber durchaus noch verbessern (bei gleichzeitigem Anwachsen des Rechenaufwands), wenn man das obige Konzept noch ein wenig verallgemeinert.

Anstelle der Indikatorfunktionen S_n betrachten wir nun für $l \in \mathbb{N}_0$

$$S_n^{(l)} := S_n \cdot \prod_{u=n-l}^{n-1} (1 - S_u).$$

$S_n^{(l)}$ zeigt also an, dass in der Zeichenkette X zur Zeit n der Pattern A mit maximal k Editierdistanzfehlern erschienen ist, dass dies aber zugleich zu den vorangegangenen l Zeitpunkten nicht der Fall war. Offenbar gilt immer noch

$$T > m \iff \sum_{n=1}^{m} S_n^{(l)} = 0,$$

und das oben entwickelte Konzept ergibt sich als Spezialfall $l = 0$.

Man bezeichnet den Übergang von S_n zu $S_n^{(l)}$ als so genannte *Declumping Technique*. Diese Bezeichnung kommt wie folgt zustande: Die zu den S_n gehörenden Ereignisse treten häufig sehr geballt („in Klumpen“) auf, denn ist zur Zeit n der Abstand $\mathrm{msed}(n)$ kleiner oder gleich k, so besteht eine recht hohe Wahrscheinlichkeit, dass dies in unmittelbarer Umgebung noch einmal der Fall ist, da sich $\mathrm{msed}(n)$ in jedem Zeitschritt betragsmäßig um höchstens 1 verändern kann (vgl. Satz 2.6). Durch den Übergang von S_n zu $S_n^{(l)}$ sorgt man dafür, dass ein solches erneutes Auftreten des zugehörigen Ereignisses frühestens nach l Zeiteinheiten wieder geschehen kann, man „entklumpt“ also sozusagen die Folge der S_n. Den Parameter $l \in \mathbb{N}_0$ bezeichnet man auch als *Declumpingtiefe*.

Für die nun folgenden Betrachtungen halten wir einige Eigenschaften der Indikatorfunktionen S_n bzw. $S_n^{(l)}$ fest:

Lemma 2.38 *Für die Indikatorfunktionen* $S_n = \mathbb{1}\{\mathrm{msed}(n) \leq k\}$ *gilt:*

(a) $n < N - k \implies S_n = 0$.

(b) $n_1, n_2 \geq N + k \implies S_{n_1}, S_{n_2}$ *sind identisch verteilt.*

(c) $n_1 \leq n_2 - (N + k) \implies S_{n_1}, S_{n_2}$ *sind unabhängig.*

Beweis: (a) Unmittelbare Folgerung aus Satz 2.7.
(b) $\mathrm{msed}(n)$ ist genau dann kleiner oder gleich k, wenn ein $m \in \{N - k, ..., N + k\}$ existiert, so dass $\mathrm{ed}(A, X_{n-m+1:n}) = k$ ist. Für alle $n \geq N + k$ ist nun stets $n - m + 1$ größer als 0, d.h. in diesem Fall hängt die Indikatorfunktion immer von einem gleichlangen Teilstück der unabhängigen und identisch verteilten Zeichenkette X ab.
(c) Es ist

$$\begin{aligned} S_{n_1} &= \mathbb{1}\{\exists m \in \{N - k, ..., N + k\} : \mathrm{ed}(A, X_{n_1-m+1:n_1}) = k\}, \\ S_{n_2} &= \mathbb{1}\{\exists m \in \{N - k, ..., N + k\} : \mathrm{ed}(A, X_{n_2-m+1:n_2}) = k\}. \end{aligned}$$

Ist nun $n_1 \leq n_2 - (N + k)$, so beziehen sich S_{n_1} und S_{n_2} auf disjunkte Teilabschnitte der unabhängigen Zeichenkette X. □

Lemma 2.39 *Für die Indikatorfunktionen* $S_n^{(l)} = \prod_{u=n-l}^{n-1}(1 - S_u)S_n$, $l \in \mathbb{N}_0$, *gilt:*

(a) $n < N - k \implies S_n^{(l)} = 0$.

(b) $n_1, n_2 \geq N + k + l \implies S_{n_1}^{(l)}, S_{n_2}^{(l)}$ *sind identisch verteilt.*

(c) $n_1 \leq n_2 - (N+k+l) \implies S_{n_1}^{(l)}, S_{n_2}^{(l)}$ *sind unabhängig.*

(d) $n \leq N-k+l \implies ES_n^{(l)} = P(T=n)$.

Beweis: (a) Hier ist nach dem vorangegangenen Lemma der in $S_n^{(l)}$ enthaltene Faktor S_n gleich 0.
(b) $S_n^{(l)}$ nimmt Bezug auf die Zufallsvariablen S_{n-l}, S_{n-l+1},...,S_n. Für $n \geq N+k+l$ ist $n-l \geq N+k$, mithin sind die beteiligten S-Variablen identisch verteilt, also auch die $S^{(l)}$-Variablen.
(c) $S_{n_1}^{(l)} = \prod_{u=n_1-l}^{n_1-1}(1-S_u)S_{n_1}$, $S_{n_2}^{(l)} = \prod_{u=n_2-l}^{n_2-1}(1-S_u)S_{n_2}$. Ist nun $n_1 \leq n_2-(N+k+l)$, so ist der Abstand zwischen den beteiligten S-Variablen größer gleich $N+k$, also sind nach dem vorangegangenen Lemma die jeweils beteiligten S-Variablen unabhängig und somit auch die $S^{(l)}$-Variablen.
(d) $S_n^{(l)} = \prod_{u=n-l}^{n-1}(1-S_u)S_n$. Für $n \leq N-k+l$ ist $n-l \leq N-k$. Da nun $S_n = 0$ für $n < N-k$ gilt, ist im hier betrachteten Bereich also sogar $S_n^{(l)} = \prod_{u=1}^{n-1}(1-S_u)S_n$.
□

Aus der Chen-Stein-Methode erhalten wir nun das folgende

Theorem 2.40 *Es sei* $A = a_1a_2...a_N \in \Sigma^N$, $k \in \{0,1,...,N\}$ *und* $T := \inf\{n \in \mathbb{N} : \mathrm{msed}(A, X_{1:n}) \leq k\}$. *Ferner sei* $l \in \mathbb{N}_0$ *und* $\lambda_m := \sum_{n=1}^m ES_n^{(l)}$. *Dann gilt*

$$\left|P(T>m) - \exp(-\lambda_m)\right| \leq \frac{1-e^{-\lambda_m}}{\lambda_m}(b_1+b_2+b_3)$$

mit

$$\begin{aligned} b_1 &:= \sum_{n_1=1}^m \sum_{n_2 \in J_{n_1}^{(l)}} ES_{n_1}^{(l)} ES_{n_2}^{(l)}, \\ b_2 &:= \sum_{n_1=1}^m \sum_{n_1 \neq n_2 \in J_{n_1}^{(l)}} ES_{n_1}^{(l)} S_{n_2}^{(l)}, \\ b_3 &:= \sum_{n_1=1}^m E\left|E\left[S_{n_1}^{(l)} - ES_{n_1}^{(l)} | \sigma\{S_{n_2}^{(l)} : n_2 \notin J_{n_1}^{(l)}\}\right]\right|. \end{aligned}$$

Dabei ist $J_{n_1}^{(l)}$ *eine von der Declumpingtiefe abhängige, frei wählbare „Abhängigkeitsumgebung“ um* n_1.

Bemerkung 2.41 Gilt für alle $n_2 \notin J_{n_1}^{(l)}$, dass $S_{n_1}^{(l)}$ und $S_{n_2}^{(l)}$ unabhängig sind, so ist offensichtlich $b_3 = 0$. Es bietet sich also an $J_{n_1}^{(l)}$ möglichst so zu wählen, dass dies der

Fall ist. Insbesondere ist diese Bedingung nach Lemma 2.39.(c) erfüllt für

$$J_{n_1}^{(l)} = \big\{n_1 - (N+k+l) + 1, ..., n_1 + (N+k+l) - 1\big\}$$

und es ergibt sich

Theorem 2.42 *Es sei* $A = a_1 a_2 ... a_N \in \Sigma^N$, $k \in \{0, 1, ..., N\}$ *und* $T := \inf\big\{n \in \mathbb{N} : \mathrm{msed}(A, X_{1:n}) \leq k\big\}$. *Ferner sei* $l \in \mathbb{N}_0$ *und* $\lambda_m := \sum_{n=1}^{m} ES_n^{(l)}$. *Dann gilt*

$$\big|P(T > m) - \exp(-\lambda_m)\big| \leq \frac{1 - e^{-\lambda_m}}{\lambda_m}\,(b_1 + b_2)$$

mit

$$b_1 := \sum_{n_1=1}^{m} \sum_{n_2=n_1-(N+k+l)+1}^{n_1+(N+k+l)-1} ES_{n_1}^{(l)} ES_{n_2}^{(l)}, \quad b_2 := \sum_{n_1=1}^{m} \sum_{\substack{n_2=n_1-(N+k+l)+1 \\ n_2 \neq n_1}}^{n_1+(N+k+l)-1} ES_{n_1}^{(l)} S_{n_2}^{(l)}.$$

Bemerkung 2.43 Auf den ersten Blick hat es den Anschein, als müsste man für den Parameter λ_m und die Konstanten b_1 und b_2 mit wachsendem m eine immer größer werdende Anzahl von Erwartungswerten berechnen, doch in Wirklichkeit ist nur eine sehr begrenzte Anzahl dieser Werte notwendig. Was die Erwartungswerte bei λ_m und der Konstanten b_1 betrifft, so haben wir bereits in Lemma 2.39 gezeigt, dass sämtliche $S_n^{(l)}$ für $n \geq N+k+l$ identisch verteilt sind, und dass sie für $n < N-k$ den Wert 0 haben; es sind also lediglich die Erwartungswerte

$$ES_n^{(l)}, \quad n \in \{N-k, ..., N+k+l\}$$

zu berechnen. Bei den gemischten Erwartungswerten in der Konstanten b_2 ist zu beachten, dass für $n_1, n_2 \geq N+k+l$ sämtliche Paare $S_{n_1}^{(l)} S_{n_2}^{(l)}$ mit demselben Abstand von n_1 und n_2 identisch verteilt sind, sich mithin für den Erwartungswert $ES_{n_1}^{(l)} S_{n_2}^{(l)}$ stets derselbe Wert ergibt. Insofern ist hier also die Berechnung der Erwartungswerte

$$ES_{n_1}^{(l)} S_{n_2}^{(l)}, \quad n_1, n_2 \in \{1, ..., 2(N+k+l)-1\}, \quad |n_1 - n_2| \leq (N+k+l) - 1, \quad n_1 \neq n_2,$$

ausreichend. Weiter ist noch zu beachten, dass $S_n^{(l)} = 0$ für $n < N-k$ gilt. Außerdem ist offensichtlich $S_{n_1}^{(l)} S_{n_2}^{(l)} = 0$ für $|n_1 - n_2| \leq l$. Für die Konstante b_2 sind also lediglich die gemischten Erwartungswerte

$$ES_{n_1}^{(l)} S_{n_2}^{(l)}, \ n_1, n_2 \in \{N-k, ..., 2(N+k+l)-1\}, \ |n_1 - n_2| \in \{l+1, ..., (N+k+l)-1\},$$

zu berechnen. Diese verbleibenden Werte kann man nun sehr schnell mit Hilfe von Standardmethoden für Markov-Ketten aus der Übergangsmatrix des Spaltenprozesses herleiten.

Bemerkung 2.44 Im Unterschied zu den Fehlerabschätzungen aus Kapitel 2.5 haben wir hier also keine universelle Schranke für den Fehler bei der Abschätzung von $P(T > m)$, sondern vielmehr eine in m wachsende „Fehlerfunktion". Für die praktische Anwendbarkeit der Methode ist von entscheidender Bedeutung, dass diese Fehlerfunktion nur sehr langsam wächst.

Beispiel 2.45 Um die Methode zu verdeutlichen, wollen wir das Gensequenzierungsbeispiel des Ti-Plasmids „Agrobacterium tumefaciens plasmid pTi-SAKURA" aus Abschnitt 2.6 noch einmal aufgreifen.

Wir approximieren die Wartezeit, bis der Pattern *agcttcgcaa* erstmals mit höchstens einem Fehler in der Editierdistanz in einer zufälligen, unabhängig und identisch verteilten Zeichenkette über dem Alphabet $\{a, g, c, t\}$ erscheint. Als Verteilung auf dem Alphabet legen wir dabei wieder die empirische Verteilung der Basen zu Grunde.

Die entsprechende Wartezeitverteilung haben wir schon einmal in Abschnitt 2.6 approximiert, die entsprechenden Kenngrößen entnimmt man der Tabelle 2.4. Hier wollen wir diese Verteilung nun nochmals mit der Chen-Stein-Methode approximieren. Dabei verwenden wir die Declumpingtiefen $l = 0, 1, 2$. Die Ergebnisse entnimmt man den folgenden Abbildungen.

Abbildung 2.6 zeigt die Approximation und die zugehörigen Fehlerschranken für die Declumpingtiefen $l = 0$ und $l = 1$. Der entsprechende Graph für $l = 2$ ist nicht mehr von dem für $l = 1$ zu unterscheiden und wurde deshalb nicht mit dargestellt.

Abbildung 2.7 zeigt nochmals die Approximationen und die zugehörigen Fehlerschranken für die Declumpingtiefen $l = 0, 1$ und zusätzlich auch $l = 2$, wobei wir diesmal eine halb-logarithmische Darstellungsweise verwenden, um die Unterschiede deutlicher herauszustellen. Es zeigt sich, dass eine größere Declumpingtiefe nicht immer auch automatisch zu besseren Approximationen führen muss. So ist die Verbesserung beim Übergang von $l = 0$ zu $l = 1$ zwar frappierend, beim Übergang von $l = 1$ zu $l = 2$ verschlechtert sich die Approximation aber wieder ein wenig.

Abbildung 2.8 zeigt schließlich einen Ausschnitt aus Abbildung 2.7. Zusätzlich wurden auch noch die Approximationen aus Tabelle 2.4 mit eingezeichnet. Es ist deutlich zu erkennen, dass diese (im oberen Tailbereich) deutlich besser sind. Dies liegt letztlich daran, dass wir in Abschnitt 2.5 eine konstante Fehlerschranke hergeleitet haben, während die Chen-Stein-Methode in Abhängigkeit vom Zeitpunkt wachsende

Fehlerschranken liefert.

Dieses Beispiel wie auch eine Reihe weiterer Simulationen scheinen darauf hinzudeuten, dass die in Abschnitt 2.5 hergeleitete Approximation der Wartezeit häufig bessere Resultate liefert als die hier vorgestellte Chen-Stein-Methode.

Abschließend kann man also sagen, dass es uns gelungen ist, zwei Approximationsmethoden herzuleiten, mit denen es möglich ist, mit moderatem Rechenaufwand in Beispielen mit Patternlängen von praxisrelevanter Größenordnung unter Einsatz eines Rechners brauchbare Näherungen für die Wartezeit des *Approximate Pattern Matching Problems* zu berechnen. Es soll aber auch betont werden, dass dadurch die Resultate zur asymptotischen Exponentialität aus Kapitel 1 und insbesondere auch aus Abschnitt 2.2 durchaus nicht überflüssig werden, da es nicht (zumindest nicht offensichtlich) möglich ist nachzuweisen, dass die in den Näherungsmethoden hergeleiteten Fehlerschranken bei festen Editierdistanzen mit wachsender Patternlänge gegen 0 streben (was zumindest im Fall der ersten Näherungsmethode nach Lemma 2.15 die asymptotische Exponentialität implizieren würde).

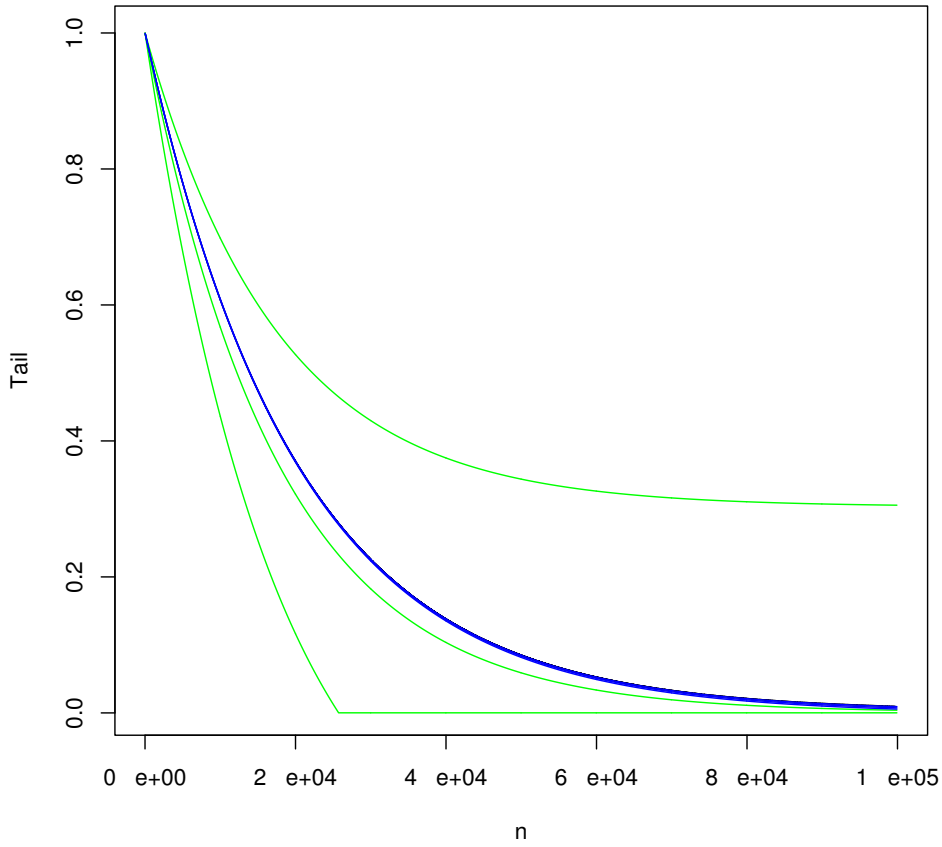

Abbildung 2.6: Approximation und Fehlerschranken für die Verteilung der Wartezeit des Patterns *agcttcgcaa*. Dargestellt sind die Declumpingtiefen $l = 0$ (grün) und $l = 1$ (blau). Im Fall $l = 1$ sind die Fehlerschranken mit bloßem Auge nicht mehr von der eigentlichen Approximation zu unterscheiden.

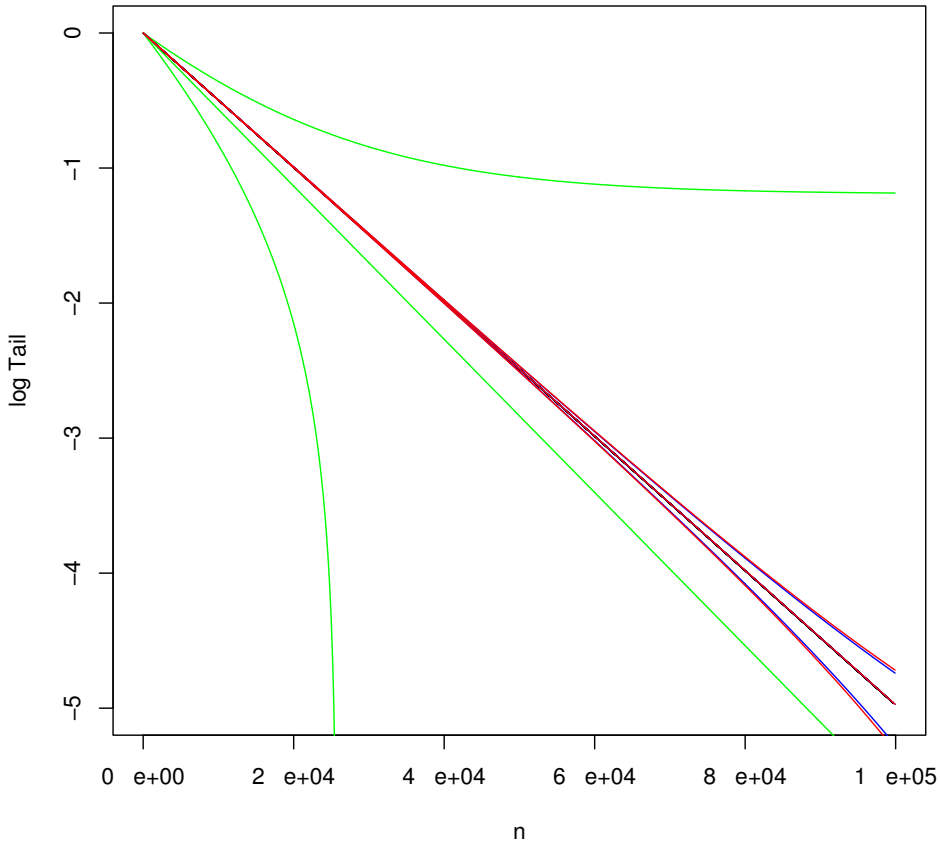

Abbildung 2.7: Approximation und Fehlerschranken für die Verteilung der Wartezeit des Patterns *agcttcgcaa*. Dargestellt sind die Declumpingtiefen $l = 0$ (grün), $l = 1$ (blau) und $l = 2$ (rot). Zur besseren Unterscheidung wurde eine halb logarithmische Darstellungsweise gewählt. In den Fällen $l = 1$ und $l = 2$ sind die Approximationen nicht mehr voneinander unterscheidbar.

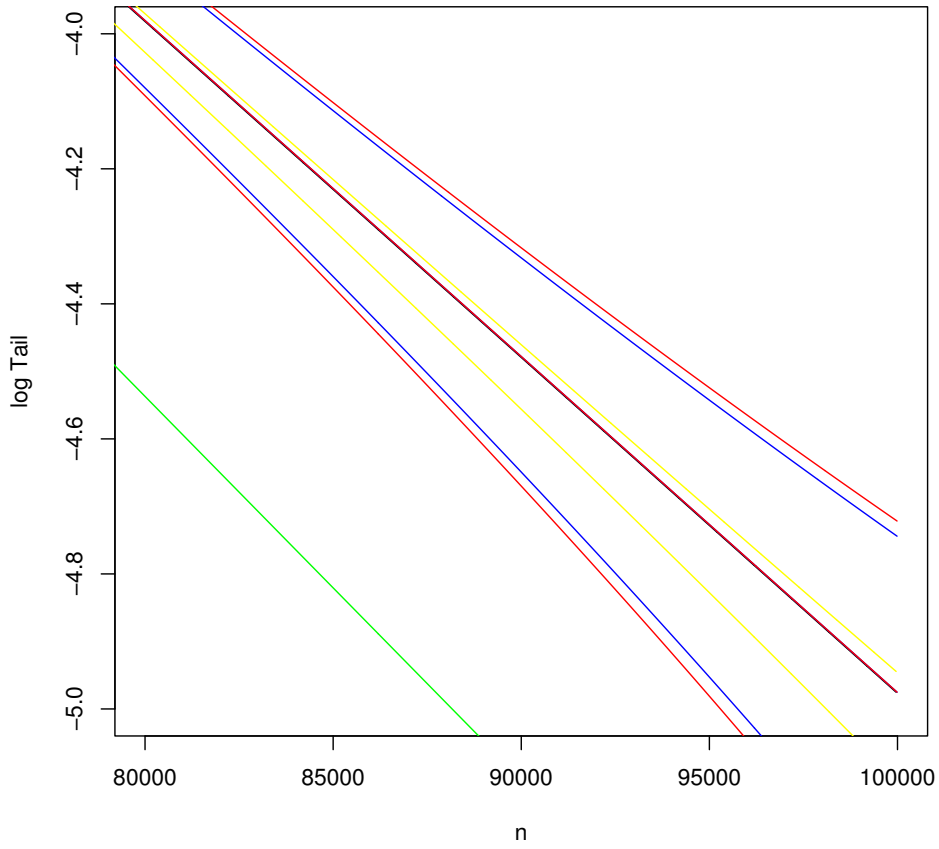

Abbildung 2.8: Ein Ausschnitt aus Abbildung 2.7. Zusätzlich wurden noch die Fehlerschranken der Approximation der Verteilung gemäß Abschnitt 2.5 eingezeichnet (gelb). Im linken unteren Bildbereich ist die Approximation im Fall $l = 0$ zu erkennen, die zugehörigen Fehlerschranken liegen außerhalb des Bildbereichs. In den Fällen $l = 1$, $l = 2$ und bei der Approximation nach Abschnitt 2.5 sind die Werte der Approximation bei dieser Auflösung nicht voneinander unterscheidbar.

Anhang A

Technische Hilfssätze

Dieser Abschnitt enthält einige technische Hilfssätze, die innerhalb der Dissertation Verwendung finden, deren Beweis an der betreffenden Stelle jedoch den Fluss der Darstellung stören würde.

Hilfssatz A.1 *Für $x \to 0$ gilt:*

$$\frac{1}{x}\operatorname{arcosh}\big(\exp(-i\theta x^2)\big) = \sqrt{-2i\theta} + o(x).$$

Für die Funktion $f(x) := \operatorname{arcosh}\big(\exp(-i\theta x^2)\big)$ gilt also insbesondere $\lim_{x\to 0} f'(x) = \sqrt{-2i\theta}$ und $\lim_{x\to 0} f''(x) = 0$.

Beweis: Wegen $\operatorname{arcosh}(z) = \log\big(z + \sqrt{z^2-1}\big)$ gilt

$$\frac{1}{x}\operatorname{arcosh}\big(\exp(-i\theta x^2)\big) = \frac{1}{x}\log\big(\exp(-i\theta x^2) + \sqrt{\exp(-2i\theta x^2) - 1}\big).$$

Verwendet man nun die Potenzreihenentwicklungen von $\exp(z)$, $\sqrt{1+z}$ und $\log(1+z)$, so ergibt sich

$$\begin{aligned}
\frac{1}{x}\operatorname{arcosh}\big(\exp(-i\theta x^2)\big) &= \frac{1}{x}\log\big(1 - i\theta x^2 + \mathcal{O}(x^4) + \sqrt{-2i\theta}x\sqrt{1+\mathcal{O}(x^2)}\big) \\
&= \frac{1}{x}\log\big(1 + \sqrt{-2i\theta}x - i\theta x^2 + \mathcal{O}(x^3)\big) \\
&= \frac{1}{x}\big(\sqrt{-2i\theta}x - i\theta x^2 - (1/2)(-2i\theta x^2) + \mathcal{O}(x^3)\big) \\
&= \sqrt{-2i\theta} + o(x).
\end{aligned}$$

□

Hilfssatz A.2 *Es seien $\alpha \in (1,2)$ und $N, m \in \mathbb{N}$. Dann gilt*

$$\sum_{k=N}^{\infty} \frac{1}{k^\alpha} \sim (\alpha-1)^{-1} \cdot N^{1-\alpha} \quad \textit{und} \quad \sum_{k=1}^{N-1} \frac{(k+1)^m}{k^\alpha} \sim (m+1-\alpha)^{-1} \cdot N^{m+1-\alpha}.$$

Weiterhin gelten diese Formeln auch für den Fall $\alpha = 2$, nur für $\alpha = 2$, $m = 1$ ergibt sich

$$\sum_{k=1}^{N-1} \frac{k+1}{k^2} \sim \log(N).$$

Beweis: Für alle $\alpha \in (1,2]$ gilt

$$\int_N^\infty \frac{1}{x^\alpha}\,dx \le \sum_{k=N}^{\infty} \frac{1}{k^\alpha} \le \int_N^\infty \frac{1}{x^\alpha}\,dx + N^{-\alpha},$$

woraus sich unmittelbar die erste Formel ergibt. Außerdem gilt für $m \ge 2$

$$\int_1^N \frac{(x+1)^m}{x^\alpha}\,dx \ge \sum_{k=1}^{N-1} \frac{(k+1)^m}{k^\alpha} \ge \int_1^{N-1} \frac{(x+1)^m}{x^\alpha}\,dx + 2^m,$$

was in diesem Fall auch die zweite Formel liefert. Für $m = 1$ erhalten wir schließlich

$$\int_1^N \frac{x+1}{x^\alpha}\,dx \le \sum_{k=1}^{N-1} \frac{k+1}{k^\alpha} \le \int_1^{N-1} \frac{x+1}{x^\alpha}\,dx + 2,$$

was je nach dem, ob $\alpha \in (1,2)$ oder $\alpha = 2$ gilt, die entsprechende Formel liefert. □

Hilfssatz A.3 *Für alle $y \in \mathbb{R}$ gilt*

$$|e^{iy} - (1+iy)| \le 2 \cdot \min\{|y|, y^2\}.$$

Beweis: Wegen $|\overline{z}| = |z|$ reicht es, den Fall $y \ge 0$ zu betrachten. Es gilt

$$e^{iy} = 1 + i\int_0^y e^{is}\,ds = 1 + iy + i\int_0^y (e^{is} - 1)\,ds,$$

also

$$|e^{iy} - (1+iy)| \le \int_0^y |e^{is} - 1|\,ds \le 2y.$$

Außerdem erhält man mit partieller Integration

$$e^{iy} = 1 + i\int_0^y e^{is}\,ds = 1 + iy + i^2\int_0^y (y-s)e^{is}\,ds,$$

also

$$|e^{iy} - (1+iy)| \le \int_0^y |(y-s)e^{is}|\,ds = \int_0^y (y-s)\,ds = \frac{y^2}{2} \le 2y^2.$$

□

Hilfssatz A.4 *Sei X_n binomialverteilt mit den Parametern n und $1/2$. Dann gilt für alle $\lambda, t > 0$:*

$$P(X_n \leq \lambda) \leq e^{t\lambda}\Big(\frac{1+e^{-t}}{2}\Big)^n.$$

Beweis: Aus Symmetriegründen und wegen $t > 0$ gilt

$$\begin{aligned} P(X_n \leq \lambda) &= P(X_n \geq n-\lambda) = P\big(\exp(tX_n) \geq \exp(t(n-\lambda))\big) \\ &= \int \mathbb{1}\{\exp(tX_n) \geq \exp(t(n-\lambda))\}\, dP \\ &\leq \exp\big(t(\lambda-n)\big) \int \exp(tX_n)\, dP \\ &= \exp\big(t(\lambda-n)\big)\Big(\frac{e^t+1}{2}\Big)^n \\ &= e^{t\lambda}\Big(\frac{1+e^{-t}}{2}\Big)^n. \end{aligned}$$

□

Hilfssatz A.5 *Es sei X_n binomialverteilt mit den Parametern n und $1/2$. Es sei $0 < \kappa < 1/2$ und*

$$c := \frac{1}{2}\left(\Big(\frac{1-\kappa}{\kappa}\Big)^{\kappa} + \Big(\frac{1-\kappa}{\kappa}\Big)^{\kappa-1}\right).$$

Dann ist $c \in (1/2, 1)$ und für alle $n \in \mathbb{N}$ gilt

$$P(X_n \leq \kappa n) \leq c^n.$$

Beweis: Nach Hilfssatz A.4 gilt für alle $n \in \mathbb{N}$ und alle $t > 0$

$$P(X_n \leq \kappa n) \leq c(t)^n$$

mit

$$c(t) := \frac{1}{2}\left(\exp\big(t\kappa\big) + \exp\big(t(\kappa-1)\big)\right).$$

Ist $0 < \kappa < 1/2$, so wird $c(t)$ minimal für $t = \log\big((1-\kappa)/\kappa\big)$ und der Minimalwert

$$c := \frac{1}{2}\left(\Big(\frac{1-\kappa}{\kappa}\Big)^{\kappa} + \Big(\frac{1-\kappa}{\kappa}\Big)^{\kappa-1}\right)$$

liegt für alle $0 < \kappa < 1/2$ im Intervall $(1/2, 1)$. □

Anhang B

Der Totalvariationsabstand

Definition B.1 Es seien $(\Omega, \mathcal{A})$ ein messbarer Raum und P, Q zwei Wahrscheinlichkeitsmaße auf $\mathcal{A}$. Dann ist der *Totalvariationsabstand* oder auch die *Total Variation Distance* zwischen P und Q definiert durch

$$\mathrm{d_{TV}}(P, Q) := \sup_{A \in \mathcal{A}} |P(A) - Q(A)|.$$

Hilfssatz B.2 *Es sei Ω abzählbar und $\mathcal{A} = \mathcal{P}(\Omega)$. Dann gilt*

$$\begin{aligned}
\mathrm{d_{TV}}(P, Q) &= \sum_{\omega \in \Omega} \big(P(\{\omega\}) - Q(\{\omega\})\big)^+ \\
&= \sum_{\omega \in \Omega} \big(P(\{\omega\}) - Q(\{\omega\})\big)^- \\
&= \frac{1}{2} \sum_{\omega \in \Omega} |P(\{\omega\}) - Q(\{\omega\})| \\
&= 1 - \sum_{\omega \in \Omega} \min\big\{P(\{\omega\}), Q(\{\omega\})\big\} \\
&= \min \mathbb{P}(X \neq Y),
\end{aligned}$$

wobei das Minimum über alle Zufallsgrößen $X, Y : \Omega' \to \Omega$ auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega', \mathcal{A}', \mathbb{P})$ mit $\mathcal{L}(X) = P$ und $\mathcal{L}(Y) = Q$ gebildet wird.

Beweis: Sei $A \in \mathcal{A}$ und $B := \{\omega \in \Omega : P(\{\omega\}) \geq Q(\{\omega\})\}$. Dann gilt

$$\begin{aligned} P(A) - Q(A) &= \sum_{\omega \in A} \big(P(\{\omega\}) - Q(\{\omega\})\big) \leq \sum_{\omega \in A \cap B} \big(P(\{\omega\}) - Q(\{\omega\})\big) \\ &\leq \sum_{\omega \in B} \big(P(\{\omega\}) - Q(\{\omega\})\big) = \sum_{\omega \in \Omega} \big(P(\{\omega\}) - Q(\{\omega\})\big)^+. \end{aligned}$$

Wegen der Symmetrie der rechten Seite in P und Q folgt dann für alle $A \in \mathcal{A}$

$$\big|P(A) - Q(A)\big| \leq \sum_{\omega \in \Omega} \big(P(\{\omega\}) - Q(\{\omega\})\big)^+.$$

Bildet man schließlich noch auf der linken Seite das Supremum über alle $A \in \mathcal{A}$, so erhält man die erste Gleichung. Die zweite bis vierte Gleichung sind lediglich triviale Umformungen. Zur letzten Gleichung: Für $A \in \mathcal{A}$ gilt

$$\begin{aligned} \big|P(A) - Q(A)\big| &= \big|\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)\big| \\ &= \big|\mathbb{P}(X \in A, Y \notin A) + \mathbb{P}(X \in A, Y \in A) - \mathbb{P}(Y \in A)\big| \\ &= \big|\mathbb{P}(X \in A, Y \notin A) - \mathbb{P}(X \notin A, Y \in A)\big| \\ &\leq \max\big\{\mathbb{P}(X \in A, Y \notin A), \mathbb{P}(X \notin A, Y \in A)\big\} \\ &\leq \mathbb{P}(X \neq Y). \end{aligned}$$

Bildet man wieder auf der linken Seite das Supremum über alle $A \in \mathcal{A}$ und rechts das Infimum über alle zulässigen Zufallsgrößen X, Y, so folgt $\mathrm{d_{TV}}(P, Q) \leq \inf \mathbb{P}(X \neq Y)$. Gleichheit erhält man für die folgende gemeinsame Verteilung von X und Y: Seien $\omega, \omega_1, \omega_2 \in \Omega$ und $R(\{\omega\}) := \min\big\{P(\{\omega\}), Q(\{\omega\})\big\}$. Dann sei

$$\mathbb{P}(X = \omega, Y = \omega) = R(\{\omega\})$$

und

$$\mathbb{P}(X = \omega_1, Y = \omega_2) = \frac{\big(P(\{\omega_1\}) - R(\{\omega_1\})\big)\big(Q(\{\omega_2\}) - R(\{\omega_2\})\big)}{1 - \sum_{\omega \in \Omega} R(\{\omega\})}, \quad \omega_1 \neq \omega_2.$$

□

Hilfssatz B.3 *Es seien $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $(E, \mathcal{B})$, $(E', \mathcal{B}')$ messbare Räume, $X, Y : \Omega \to E$ Zufallsgrößen und $f : E \to E'$ eine messbare Abbildung. Dann gilt*

$$\mathrm{d_{TV}}\big(\mathcal{L}(f(X)), \mathcal{L}(f(Y))\big) \leq \mathrm{d_{TV}}\big(\mathcal{L}(X), \mathcal{L}(Y)\big).$$

Beweis: Es gilt

$$\begin{aligned} \mathrm{d_{TV}}\big(\mathcal{L}(f(X)),\mathcal{L}(f(Y))\big) &= \sup_{B'\in\mathcal{B}'} |P^{f(X)}(B') - P^{f(Y)}(B')| \\ &= \sup_{B'\in\mathcal{B}'} |P^{X}(f^{-1}(B')) - P^{Y}(f^{-1}(B'))| \\ &\leq \sup_{B\in\mathcal{B}} |P^{X}(B) - P^{Y}(B)| \\ &= \mathrm{d_{TV}}\big(\mathcal{L}(X),\mathcal{L}(Y)\big). \end{aligned}$$

□

Sei nun $X = (X_n)_{n\in\mathbb{N}_0}$ eine homogene, irreduzible und aperiodische Markov-Kette mit endlichem Zustandsraum E und stationärer Verteilung π.

Hilfssatz B.4 *Es seien μ, λ Verteilungen auf E und $f : E^{\mathbb{N}_0} \to \mathbb{R}_+$ eine Abbildung. Dann gilt:*

$$\big|E_\mu\big(f(X)\big) - E_\lambda\big(f(X)\big)\big| \leq \max_{i\in E} E_i\big(f(X)\big) \cdot \mathrm{d_{TV}}(\mu,\lambda).$$

Beweis:

$$\begin{aligned} |E_\mu\big(f(X)\big) - E_\lambda\big(f(X)\big)| &= \Big|\sum_{i\in E} E_i\big(f(X)\big)\mu(i) - \sum_{i\in E} E_i\big(f(X)\big)\lambda(i)\Big| \\ &\leq \max\Big\{ \sum_{\substack{i\in E\\ \mu(i)\geq\lambda(i)}} E_i\big(f(X)\big)\big(\mu(i)-\lambda(i)\big), \sum_{\substack{i\in E\\ \mu(i)<\lambda(i)}} E_i\big(f(X)\big)\big(\lambda(i)-\mu(i)\big)\Big\} \\ &\leq \max_{i\in E} E_i\big(f(X)\big) \cdot \sup_{A\subset E} |\mu(A) - \lambda(A)| = \max_{i\in E} E_i\big(f(X)\big) \cdot \mathrm{d_{TV}}(\mu,\lambda). \end{aligned}$$

□

Ein häufig verwendetes Maß dafür, wie weit die Verteilung der Markov-Kette X zur Zeit n noch von der stationären Verteilung entfernt ist, ist

$$d(n) := \max_{i\in E} \mathrm{d_{TV}}\big(\mathcal{L}_i(X_n),\pi\big), \quad n\in\mathbb{N}. \tag{B.1}$$

$d(n)$ besitzt folgende Eigenschaften:

Hilfssatz B.5 *(a) $d(n)$ ist monoton fallend.*

(b) Für alle $n,m\in\mathbb{N}$ gilt $d(n+m) \leq 2d(n)d(m)$.

(c) Für jede Verteilung λ auf E und alle $n \in \mathbb{N}$ gilt $\mathrm{d_{TV}}\big(\mathcal{L}_\lambda(X_n), \pi\big) \leq d(n)$.

Beweis: Seien $n, m \in \mathbb{N}$. Dann gilt:

$$\begin{aligned}
d(n+m) &= \max_{i\in E} \mathrm{d_{TV}}\big(\mathcal{L}_i(X_{n+m}), \pi\big) = \max_{i\in E} \frac{1}{2} \sum_{j\in E} \big|P_i(X_{n+m} = j) - \pi(j)\big| \\
&= \max_{i\in E} \frac{1}{2} \sum_{j\in E} \Big| \sum_{k\in E} P_k(X_n = j) P_i(X_m = k) - \pi(j) \Big| \\
&\leq \max_{i\in E} \frac{1}{2} \sum_{j,k\in E} P_i(X_m = k) \big|P_k(X_n = j) - \pi(j)\big| \\
&= \max_{i\in E} \sum_{k\in E} P_i(X_m = k) \underbrace{\frac{1}{2} \sum_{j\in E} \big|P_k(X_n = j) - \pi(j)\big|}_{\leq d(n)} \\
&\leq d(n).
\end{aligned}$$

Damit ist die Monotonie von $d(n)$ nachgewiesen. Weiter gilt

$$\begin{aligned}
d(n+m) &= \max_{i\in E} \frac{1}{2} \sum_{j\in E} \Big| \sum_{k\in E} \big(P_k(X_n = j) - \pi(j)\big) P_i(X_m = k) \Big| \\
&= \max_{i\in E} \frac{1}{2} \sum_{j\in E} \Big| \sum_{k\in E} \big(P_k(X_n = j) - \pi(j)\big) \big(\pi(k) + P_i(X_m = k) - \pi(k)\big) \Big| \\
&\leq \frac{1}{2} \sum_{j\in E} \underbrace{\big|P_\pi(X_n = j) - \pi(j)\big|}_{=0} + \\
&\quad \max_{i\in E} \frac{1}{2} \sum_{j,k\in E} \big|P_k(X_n = j) - \pi(j)\big| \cdot \big|P_i(X_m = k) - \pi(k)\big| \\
&\leq \max_{l\in E} \mathrm{d_{TV}}\big(\mathcal{L}_l(X_n), \pi\big) \cdot 2 \cdot \max_{i\in E} \frac{1}{2} \sum_{k\in E} \big|P_i(X_m = k) - \pi(k)\big| \\
&= 2d(n)d(m).
\end{aligned}$$

Damit ist auch die zweite Eigenschaft nachgewiesen. Schließlich gilt für eine beliebige Verteilung λ auf E

$$\begin{aligned}
\mathrm{d_{TV}}\big(\mathcal{L}_\lambda(X_n), \pi\big) &= \mathrm{d_{TV}}\Big(\sum_{i\in E} \mathcal{L}_i(X_n)\lambda(i), \pi \Big) \leq \sum_{i\in E} \mathrm{d_{TV}}\big(\mathcal{L}_i(X_n), \pi\big)\lambda(i) \\
&\leq \max_{i\in E} \mathrm{d_{TV}}\big(\mathcal{L}_i(X_n), \pi\big) = d(n).
\end{aligned}$$

□

Die meisten dieser Resultate stammen aus [AF04] und [Ald82].

Anhang C

Quellcode

Im Folgenden ist der Quellcode eines C-Programms wiedergegeben, das die Aussage von Theorem 2.36 aus Abschnitt 2.5 formalisiert. Mit diesem Programm ist es möglich, im Fall einer unabhängigen und identisch verteilten Zeichenkette bei verschiedenen Editierdistanzen die Wartezeit des *Approximate Pattern Matching Problems* durch eine geometrische Verteilung zu approximieren.

1. Einbindung der verwendeten Header-Dateien.

```
1  #include <stdio.h>
2  #include <stdlib.h>
3  #include <math.h>
4  #include <time.h>
```

2. Definition der innerhalb des Programms verwendeten Strukturen: Es werden drei Strukturen für Vektoren festgelegt, deren Komponenten vom Typ `integer`, `char` oder `unsigned character` sind. Ein solcher Vektor wird innerhalb der Struktur durch den Zeiger auf seine Startadresse und die Anzahl seiner Komponenten definiert.

```
5  struct ivector
6  {int *p;
7   int length;};
8
9  struct ucvector
10 {unsigned char *p;
11  int length;};
12
13 struct dvector
14 {double *p;
15  int length;};
```

Außerdem wird eine Struktur für Matrizen definiert. Da in der Regel extrem schwach besetzte Matrizen verwendet werden, werden diese in Speicherplatz sparender Form definiert. Innerhalb der Struktur wird eine Matrix durch folgende Komponenten festgelegt: Zunächst haben wir ein Paar von zwei Vektoren. Der erste gibt die jeweiligen Positionen der (von Null verschiedenen) Komponenten innerhalb einer Zeile an, der zweite den Wert der entsprechenden Komponente. Hinzu kommt ein dritter Vektor, der angibt, an welchen Positionen der ersten beiden Vektoren jeweils die Zeilenumbrüche der Matrix vorzunehmen sind. Schließlich kommen noch die Anzahl der Zustände, also die Dimension der Matrix, sowie die Größe des beim *Approximate Pattern Matching Problems* verwendeten Alphabets hinzu.

```
16  struct tmatrix
17  {int *p;                /* Die (P)ostionen der Uebergaenge */
18   double *w;             /* Die zugehoerigen Uebergangs(w)ahrscheinlichkeiten */
19   int *lb;               /* Die Zeilenumbrueche ((l)ine(b)reaks) */
20   int numberofstates;
21   int alphabetsize;};
```

3. Prädefinition der im Anschluss an das Hauptprogramm definierten Unterfunktionen.

```
22  double distvectonum (struct ivector);
23  void numtodistvec (double, struct ivector);
24  void schritt (struct ivector, unsigned char, struct ucvector);
25  double numschritt (double, unsigned char, struct ucvector);
26  void transptmatrix(struct tmatrix, struct tmatrix);
27  void lproduct (struct tmatrix, struct dvector, struct dvector);
28  double tvd (struct dvector, struct dvector);
29  void indexx(int n, int *vector, int *index);
30  void dindexx(int n, double *vector, int *index);
```

4. Beginn des Hauptprogramms:

```
22  int main(int argc, char *argv[])
23  {
```

5. Definition der Variablen des *Approximate Pattern Matching Problems*. Dies ist der Teil, den der Anwender an sein jeweiliges Problem anpassen kann. Beispielhaft sind hier die Daten für das Modell Nr. 1 des *Monkey Typing Shakespeare* Problems wiedergegeben.

5a. Definition des Zielpattern (hier *TOBEORNOTTOBE*) unter Verwendung des Alphabets $\{1, 2, 3, 4, ...\}$. Da der Pattern selbst nur von 6 Buchstaben des Alphabets $\{A, B, ..., Z\}$ Gebrauch macht, fassen wir alle nicht verwendeten Buchstaben zu einem zusammen und verwenden hier das Alphabet $\{1, 2, ..., 7\}$ mit den Entspre-

chungen 1 = B, 2 = E, 3 = N, 4 = O, 5 = R, 6 = T, 7 = Rest.

```
24  unsigned char      PATTERN[] = {6,4,1,2,4,5,3,4,6,6,4,1,2};
```

5b. Definition der Verteilung auf dem Alphabet $\{1, 2, 3, 4, ...\}$. Hier beispielhaft der Fall der Gleichverteilung.

```
25  double        ALPHABETDIST[] = {0.03846154,0.03846154,0.03846154,0.03846154
26                                   0.03846154,0.03846154,0.76923076};
```

5c. Definition des Pfads der Ausgabedatei, in der die Ergebnisse des Programms abgelegt werden.

```
27  char          dateiname[100] = "C:/Daten/ergebnis.txt";
```

5d. Festlegung, ob in der Ausgabedatei alle Daten ('a') oder nur die wichtigsten ('w') gespeichert werden sollen.

```
28  char                    wahl = 'w';
```

6. Definition der übrigen im Programm verwendeten Variablen.

```
29  struct ucvector pattern;
30  struct dvector alphabetdist;
31  struct ivector maxdist;
32  clock_t prgstart,prgbreak1,prgbreak2,prgende;
33  int alphabetsize;
34  int sizeofstateset=1;
35  double *Stateset, **stateset;
36  int *Statetree, **statetree;
37  int maxsizeofstateset=10000;
38  struct ivector tempdistvector;
39  double actualfather,actualstate,newstate;
40  int actualrow;
41  int hit;
42  int i,j,k;
43  struct dvector setofstates;
44  int *index, *index2;
45  int *eddistvector, *anzahlen;
46  double *temp;
47  struct tmatrix tm, itm, redtm, iredtm;
48  struct dvector pi, pineu, rho, rhoneu, vec, vecneu;
49  double fehler,l;
50  struct dvector tvdschranke,EW,vorzeitig;
51  FILE *datei;
```

7. Zuweisung des Pattern und der Verteilung auf dem Alphabet.

```
52  pattern.length=sizeof(PATTERN)/sizeof(unsigned char);
53  pattern.p=PATTERN;
54  alphabetdist.length=alphabetsize=sizeof(ALPHABETDIST)/sizeof(double);
55  alphabetdist.p=ALPHABETDIST;
```

8. Definition des Vektors `maxdist`. Dieser legt fest, für welche Editierdistanzen der Erwartungswert der Wartezeit und die zugehörigen Fehlerschranke berechnet werden sollen. Voreingestellt ist hier die Berechnung dieser Werte für alle Editierdistanzen von 0 bis zur Patternlänge−2.

```
56  maxdist.length=pattern.length-2;
57  maxdist.p=(int *)malloc(maxdist.length*sizeof(int));
58  for(i=0;i<=maxdist.length-1;i++) maxdist.p[i]=i;
```

Ende der Variablendefinition, Beginn der eigentlichen Berechnungen.

9. Bestimmung des Zustandsraums. `Stateset` ist eine Matrix. In der ersten Spalte werden die Zustände abgelegt, in der zweiten Spalte steht die Position des Vaterzustands, der in der Tiefensuche auf diesen Zustand geführt hat, in der dritten Spalte wird festgehalten, wie viele der möglichen Übergänge von diesem Zustand aus innerhalb der Tiefensuche bereits abgearbeitet worden sind, und in den folgenden Spalten stehen die Positionen der Zustände, auf die die jeweiligen Übergänge geführt haben. `stateset` ist ein Zeigervektor, der es ermöglicht, auf `Stateset` wie auf eine Matrix mit Spalten- und Zeilenindex zuzugreifen. `Statetree` ist eine Baumstruktur, die die schnellste Möglichkeit bietet zu prüfen, ob ein neuer Zustand bereits im Zustandsraum enthalten ist oder nicht. `statetree` dient wiederum dem bequemen Zugriff auf `Statetree` mit Zeilen- und Spaltenindizies.

```
60  prgstart=clock();
61  Stateset=(double *)malloc(maxsizeofstateset*(3+alphabetsize)*sizeof(double));
62  stateset=(double **)malloc(maxsizeofstateset*sizeof(double*));
63  Statetree=(int *)malloc(maxsizeofstateset*2*sizeof(int));
64  statetree=(int **)malloc(maxsizeofstateset*sizeof(int *));
65  for (i=0;i<=maxsizeofstateset-1;i++) stateset[i]=&Stateset[i*(3+alphabetsize)];
66  for (i=0;i<=maxsizeofstateset-1;i++) statetree[i]=&Statetree[i*2];
67  for (i=0;i<=2*maxsizeofstateset-1;i++) Statetree[i]=-1;
```

9a. Initialisierung von `Stateset`, erster Zustand ist `(0,1,2,...,pattern.length)`.

```
68  tempdistvector.p=(int *)malloc((pattern.length+1)*sizeof(int));
69  for (i=0;i<=pattern.length;i++) tempdistvector.p[i]=i;
70  tempdistvector.length=pattern.length+1;
71  stateset[0][0]=distvectonum(tempdistvector); stateset[0][1]=-1; stateset[0][2]=1;
72  actualfather=newstate=0; actualstate=stateset[0][0]; actualrow=0;
```

9b. Berechnung des gesamten Zustandsraums, ausgehend vom Startzustand durch eine Tiefensuche.

```
73  while (!((abs(actualfather+1)<=0.1)&&(stateset[actualrow][2]>alphabetsize)))
74  {
```

Wenn vom aktuellen Zustand noch ein Übergang möglich ist, so führe diesen aus, sonst springe zum Vater dieses Zustands zurück.

```
75    if (stateset[actualrow][2]<=alphabetsize+0.1)
76       newstate=numschritt(actualstate,(unsigned char)stateset[actualrow][2],pattern);
77    else
78      {actualrow=actualfather; actualstate=stateset[actualrow][0];
79       actualfather=stateset[actualrow][1]; continue;};
```

Prüfe: Ist der neue Zustand bereits im Zustandsraum enthalten?

```
80    i=0;hit=0;
81    while(1)
82      {if (abs(stateset[i][0]-newstate)<=0.1) {hit=1;break;}
83       if (stateset[i][0]>newstate+0.1)
84         {if(statetree[i][0]==-1) {statetree[i][0]=sizeofstateset;break;}
85          else i=statetree[i][0];}
86       else
87         {if(statetree[i][1]==-1) {statetree[i][1]=sizeofstateset;break;}
88          else i=statetree[i][1];}
89      }
```

Wenn ja, so führe den nächsten Übergang aus. Wenn nein, so füge den neuen Zustand zum Zustandsraum hinzu und mache ihn zum aktuellen Zustand.

```
90    if (hit)
91      {stateset[actualrow][2+(int)stateset[actualrow][2]]=i;
92       stateset[actualrow][2]++; continue;}
93    else
94      {sizeofstateset++;
95       stateset[actualrow][2+(int)stateset[actualrow][2]]=sizeofstateset-1;
96       stateset[actualrow][2]++;
97       i=actualrow; actualrow=sizeofstateset-1;
98       stateset[actualrow][0]=newstate; stateset[actualrow][1]=i; stateset[actualrow][2]=1;
99       actualfather=i; actualstate=newstate;
```

Um nicht unnötig Speicherplatz zu belegen, ist die maximale Größe des Zustandsraums auf 10000 Zustände voreingestellt. Sollte diese Oberschranke überschritten werden, so wird sie verdoppelt und der Zustandsraum reallokiert.

```
100      if (sizeofstateset==maxsizeofstateset)
101        {maxsizeofstateset+=maxsizeofstateset;
102         Stateset=(double *)realloc(Stateset,maxsizeofstateset*(3+alphabetsize)
103                                                              *sizeof(double));
104         stateset=(double **)realloc(stateset,maxsizeofstateset*sizeof(double *));
105         Statetree=(int *)realloc(Statetree,maxsizeofstateset*2*sizeof(int));
106         statetree=(int **)realloc(statetree,maxsizeofstateset*sizeof(int *));
107         for (i=0;i<=maxsizeofstateset-1;i++) stateset[i]=&Stateset[i*(3+alphabetsize)];
108         for (i=0;i<=maxsizeofstateset-1;i++) statetree[i]=&Statetree[i*2];
109         for (j=maxsizeofstateset;j<=2*maxsizeofstateset-1;j++) Statetree[j]=-1;
110        };
111    };
```

```
112   }
113
114   prgbreak1=clock();
115   printf("%4.2f Sek. Bestimmung des Zustandsraums\n",
116           (float)(prgbreak1-prgstart) / CLOCKS_PER_SEC);
```

9c. Aus der oben bestimmten Tabelle `Stateset` wird die Menge der Zustände extrahiert und im Vektor `setofstates` abgelegt.

```
117   setofstates.p=(double *)malloc(sizeofstateset*sizeof(double));
118   setofstates.length=sizeofstateset;
119   for (i=0;i<=sizeofstateset-1;i++) setofstates.p[i]=stateset[i][0];
```

9d. Für diese Zustände wird die jeweilige Editierdistanz zum Zielpattern bestimmt und im Vektor `eddist` festgehalten.

```
120   eddistvector=(int *)malloc(sizeofstateset*sizeof(int));
121   for(i=0;i<=sizeofstateset-1;i++)
122     {numtodistvec(setofstates.p[i],tempdistvector);
123      eddistvector[i]=tempdistvector.p[pattern.length];
124     }
```

In der i-ten Komponente des Vektors `anzahlen` wird festgehalten, wie viele Zustände mit Editierdistanz i zum Zielpattern im Zustandsraum vorhanden sind.

```
125   anzahlen=(int *)malloc((pattern.length+1)*sizeof(int));
126   for(i=0;i<=pattern.length;i++) anzahlen[i]=0;
127   for(i=0;i<=sizeofstateset-1;i++) anzahlen[eddistvector[i]]++;
```

9e. Die Zustände werden fallend nach Ihrer Editierdistanz sortiert.

```
128   index=(int*)malloc(sizeofstateset*sizeof(int));
129   for(i=0;i<=sizeofstateset-1;i++) index[i]=i;
130   index2=(int*)malloc(sizeofstateset*sizeof(int));
131   for(i=0;i<=sizeofstateset-1;i++) index2[i]=i;
132   indexx(sizeofstateset,eddistvector,index);
133   for(i=0;i<=sizeofstateset-1;i++) index2[i]=index[sizeofstateset-1-i];
134   temp=(double*)malloc(sizeofstateset*sizeof(double));
135   for(i=0;i<=sizeofstateset-1;i++) temp[i]=setofstates.p[index2[i]];
136   for(i=0;i<=sizeofstateset-1;i++) setofstates.p[i]=temp[i];
137   indexx(sizeofstateset,index2,index);
138   free(temp);
```

9f. Für die sortierten Zustände wird die Übergangsmatrix bestimmt.

```
139   tm.p=(int*)malloc(sizeofstateset*alphabetsize*sizeof(int));
140   tm.w=(double*)malloc(sizeofstateset*alphabetsize*sizeof(double));
141   tm.lb=(int*)malloc((sizeofstateset+1)*sizeof(int));
142   tm.numberofstates=sizeofstateset; tm.alphabetsize=alphabetsize;
143
144   k=0;
145   for(i=0;i<=sizeofstateset-1;i++)
146     {for(j=1;j<=alphabetsize;j++)
```

```
147         {tm.p[k]=index[(int)stateset[index2[i]][2+j]]; tm.w[k]=alphabetdist.p[j-1]; k++;}
148       tm.lb[i]=i*alphabetsize;
149     }
150   tm.lb[sizeofstateset]=sizeofstateset*alphabetsize;
151
152   free(Stateset); free(stateset); free(Statetree); free(statetree); free(index);
153
154   prgbreak2=clock();
155   printf("%4.2f Sek. Extrahieren der Uebergangsmatrix\n",
156          (float)(prgbreak2-prgbreak1) / CLOCKS_PER_SEC);
```

10. Die Übergangsmatrix wird transponiert.

```
157   itm.p=(int*)malloc(sizeofstateset*alphabetsize*sizeof(int));
158   itm.w=(double*)malloc(sizeofstateset*alphabetsize*sizeof(double));
159   itm.lb=(int*)malloc((sizeofstateset+1)*sizeof(int));
160   itm.numberofstates=sizeofstateset; itm.alphabetsize=alphabetsize;
161
162   transptmatrix(tm,itm);
```

11. Bestimmung der stationären Verteilung `pi` der Übergangsmatrix. Die zugehörige Markov-Kette ist nach `2*pattern.length` Übergängen stationär.

```
163   pi.p=(double *)malloc(sizeofstateset*sizeof(double)); pi.length=sizeofstateset;
164   pineu.p=(double *)malloc(sizeofstateset*sizeof(double)); pineu.length=sizeofstateset;
165   pi.p[0]=1; for (i=1;i<=pi.length-1;i++) pi.p[i]=0;
166   for (i=1;i<=2*pattern.length+4;i++)
167     {lproduct(itm,pi,pineu); for (j=0;j<=pi.length-1;j++) pi.p[j]=pineu.p[j];}
168   free(pineu.p);
169
170   prgbreak1=clock();
171   printf("%4.2f Sek. Bestimmung von pi\n",(float)(prgbreak1-prgbreak2) / CLOCKS_PER_SEC);
```

12. Initialisierung der Vektoren `EW` und `tvdschranke`. In der i-ten Komponente wird zur Editierdistanz `maxdist.p[i]` der jeweilige Erwartungswert der Wartezeit $E_\rho(T)$ bzw. die Fehlerschranke $\mathrm{d_{TV}}(\rho,\pi)$ festgehalten. Außerdem Initialisierung des Vektors `vorzeitig`. In der i-ten Komponente wird zur Editierdistanz `maxdist.p[i]` die Wahrscheinlichkeit $P_{\lambda^*}(T < 2 \cdot \texttt{pattern.length})$ gespeichert.

```
172   tvdschranke.p=(double *)malloc(maxdist.length*sizeof(double));
173   tvdschranke.length=maxdist.length;
174   EW.p=(double *)malloc(maxdist.length*sizeof(double));
175   EW.length=maxdist.length;
176   vorzeitig.p=(double *)malloc(maxdist.length*sizeof(double));
177   vorzeitig.length=maxdist.length;
178   for (i=0;i<=vorzeitig.length-1;i++) vorzeitig.p[i]=0;
```

13. Für alle in `maxdist.p` festgelegten Editierdistanzen werden die reduzierte Übergangsmatrix `redtm` ($P_{\Delta_k^c}$), die Verteilung `rho` (ρ), der Erwartungswert der Wartezeit `EW.p` ($E_\rho(T)$), die Fehlerschranke `tvdschranke.p` ($\mathrm{d_{TV}}(\rho,\pi)$) und die Wahrschein-

lichkeit, vorzeitig in der Zielmenge zu landen, `vorzeitig.p`, bestimmt.

```
179   for (i=0;i<=maxdist.length-1;i++)
180     {printf("Editierdistanz %i\n",i);
```

13a. Initialisierung und Bestimmung der reduzierten Übergangsmatrix.

```
181       redtm.numberofstates=0;
182       for(j=maxdist.p[i]+1;j<=pattern.length;j++) redtm.numberofstates+=anzahlen[j];
183       redtm.alphabetsize=alphabetsize;
184       redtm.p=(int*)malloc(redtm.numberofstates*redtm.alphabetsize*sizeof(int));
185       redtm.w=(double*)malloc(redtm.numberofstates*redtm.alphabetsize*sizeof(double));
186       redtm.lb=(int*)malloc((redtm.numberofstates+1)*sizeof(int));
187
188       for(j=0;j<=redtm.numberofstates*redtm.alphabetsize-1;j++)
189         {redtm.p[j]=tm.p[j]; redtm.w[j]=tm.w[j];}
190       for(j=0;j<=redtm.numberofstates;j++) redtm.lb[j]=tm.lb[j];
191       for(j=0;j<=redtm.numberofstates*redtm.alphabetsize-1;j++)
192          if(redtm.p[j]>redtm.numberofstates-1)
193            {redtm.p[j]=0; redtm.w[j]=0;}
```

13b. Transponieren der reduzierten Übergangsmatrix.

```
194       iredtm.p=(int*)malloc(redtm.numberofstates*redtm.alphabetsize*sizeof(int));
195       iredtm.w=(double*)malloc(redtm.numberofstates*redtm.alphabetsize*sizeof(double));
196       iredtm.lb=(int*)malloc((redtm.numberofstates+1)*sizeof(int));
197       iredtm.numberofstates=redtm.numberofstates; iredtm.alphabetsize=iredtm.alphabetsize;
198
199       transptmatrix(redtm,iredtm);
```

13c. Bestimmung von $P_{\lambda^*}(T < 2 \cdot \texttt{pattern.length})$ durch Iteration.

```
200       vec.p=(double *)malloc(redtm.numberofstates*sizeof(double));
201       vec.length=iredtm.numberofstates;
202       vecneu.p=(double *)malloc(redtm.numberofstates*sizeof(double));
203       vecneu.length=iredtm.numberofstates;
204
205       vec.p[0]=1; for (j=1;j<=vec.length-1;j++) vec.p[j]=0;
206
207       for(j=1;j<=2*pattern.length-1;j++)
208         {lproduct(iredtm,vec,vecneu); for(k=0;k<=vec.length-1;k++) vec.p[k]=vecneu.p[k];}
209       for(j=0;j<=vec.length-1;j++) vorzeitig.p[i]+=vec.p[j];
210       vorzeitig.p[i]=1-vorzeitig.p[i];
```

13d. Bestimmung von `rho` als auf Länge 1 normierten Linkseigenvektor zum Perron-Eigenwert der reduzierten Übergangsmatrix durch Iteration.

```
211       rho.p=(double *)malloc(redtm.numberofstates*sizeof(double));
212       rho.length=iredtm.numberofstates;
213       rhoneu.p=(double *)malloc(redtm.numberofstates*sizeof(double));
214       rhoneu.length=iredtm.numberofstates;
215
216       for (j=0;j<=rho.length-1;j++) rho.p[j]=1;
217       fehler=1;
```

```
218       while(fehler>=pow(10,-14))
219         {lproduct(iredtm,rho,rhoneu);
220          l=0; for(j=0;j<=rhoneu.length-1;j++) l+=rhoneu.p[j];
221          for(j=0;j<=rhoneu.length-1;j++) rhoneu.p[j]=rhoneu.p[j]/l;
222          fehler=tvd(rhoneu,rho);
223          for(j=0;j<=rho.length-1;j++) rho.p[j]=rhoneu.p[j];
224         }
```

13e. Reallokierung von `rho` auf die gleiche Länge wie `pi` und Berechnung von $d_{TV}(\rho, \pi)$.

```
225       rho.p=(double *)realloc(rho.p,sizeofstateset*sizeof(double));
226       rhoneu.p=(double *)realloc(rhoneu.p,sizeofstateset*sizeof(double));
227       rho.length=sizeofstateset;
228       for(j=redtm.numberofstates;j<=sizeofstateset-1;j++)
229          rho.p[j]=0;
230
231       tvdschranke.p[i]=tvd(pi,rho);
```

13f. Berechnung des Erwartungswertes $E_\rho(T)$ als Kehrwert von $P_\rho(T = 1)$ (T ist unter ρ geometrisch verteilt). Dabei werden die zu addierenden Werte aus numerischen Gründen der Größe nach sortiert.

```
232       lproduct(itm,rho,rhoneu);
233       EW.p[i]=0;
234       index=(int*)malloc((sizeofstateset-redtm.numberofstates)*sizeof(int));
235       for(j=0;j<=(sizeofstateset-redtm.numberofstates)-1;j++) index[j]=j;
236       dindexx((sizeofstateset-redtm.numberofstates),rho.p+redtm.numberofstates,index);
237       for(j=0;j<=(sizeofstateset-redtm.numberofstates)-1;j++)
238          EW.p[i]+=rhoneu.p[redtm.numberofstates+index[j]];
239       EW.p[i]=pow(EW.p[i],-1);
240       free(index);
```

13g. Freigabe aller allokierten Vektoren.

```
241       free(rho.p); free(rhoneu.p); free(redtm.p); free(redtm.w);
242       free(redtm.lb); free(iredtm.p); free(iredtm.w); free(iredtm.lb);
243      }
244
245    printf("\n");
246    prgbreak2=clock();
247    printf("%4.2f Sek. Bestimmung der EWs und Fehlerschranken\n",
248            (float)(prgbreak2-prgbreak1) / CLOCKS_PER_SEC);
249    prgende=clock();
250    printf("%4.2f Sek. Gesamtlaufzeit\n",
251            (float)(prgende-prgstart) / CLOCKS_PER_SEC);
```

14. Ausgabe der Ergebnisse.

```
252  printf("\nEnde der Berechnungen\n"); printf("Daten wurden gespeichert\n");
253   datei=fopen(dateiname,"w");
254   fprintf(datei,"Wartezeitanalyse fuer Pattern\n\n");
```

```
255   fprintf(datei,"Pattern: ");
256   for(i=0;i<=pattern.length-1;i++) fprintf(datei,"%i ",pattern.p[i]); fprintf(datei,"\n");
257   fprintf(datei,"Laenge des Pattern: %i\n\n",pattern.length);
258   fprintf(datei,"Alphabet: ");
259   for(i=0;i<=alphabetdist.length-1;i++) fprintf(datei,"%i ",i+1);fprintf(datei,"\n");
260   fprintf(datei,"Verteilung auf dem Alphabet: ");
261   for(i=0;i<=alphabetdist.length-1;i++) fprintf(datei,"%.3f ",(float)alphabetdist.p[i]);
262   fprintf(datei,"\n\n");
263   fprintf(datei,"Groesse des Zustandsraums: %i\n",sizeofstateset);
264   fprintf(datei,"Gesamtlaufzeit des Programms: %.2f Sek.\n\n",
265           (float)(prgende-prgstart) / CLOCKS_PER_SEC);
266   if(wahl=='a')
267   {fprintf(datei,Uebergangsmatrix:\n\n");
268    for(i=0;i<=tm.numberofstates-1;i++)
269      {fprintf(datei,"%i\t|",i);
270       for(j=tm.lb[i];j<tm.lb[i+1];j++) fprintf(datei,"%i\t",(int)tm.p[j]);
271       fprintf(datei,"\t|",i);
272       for(j=tm.lb[i];j<tm.lb[i+1];j++) fprintf(datei,"%.3f\t",(float)tm.w[j]);
273       fprintf(datei,"\n");
274      }
275    fprintf(datei,"\npi:\n\n");
276    for(i=0;i<=pi.length-1;i++) fprintf(datei,"%i\t|%.8f\n",i,pi.p[i]);
277    fprintf(datei,"\n");
278   }
279   fprintf(datei,"Tabelle der EWs und Fehlerschranken:\n\n");
280   fprintf(datei,"eddist\tEW\t\ttvd(rho,mu)\tP(T<2*Patternl"ange)\n");
281   fprintf(datei,"============================================================\n");
282   for(i=0;i<=maxdist.length-1;i++)
283      fprintf(datei,"%i\t%.8e\t%.8e\t%.8e\n",maxdist.p[i],EW.p[i],tvdschranke.p[i],
284                                                              vorzeitig.p[i]);
285   fclose(datei);
286   printf("fertig\n");
287   getchar();
288   return 0;
289  }
```

Ende des Hauptprogramms.

15. Definition der Hilfsfunktionen:

15a. Definition der Hilfsfunktion `distvectonum`. Wandelt einen Zustand des Spaltenprozesses in eine Ternärzahl um. Umkehrfunktion zu `numtodistvec`.

```
290  double distvectonum (struct ivector distvec)
291  {int i;
292   double erg=0;
293
294   for (i=0;i<=distvec.length-2;i++) erg+=(distvec.p[i+1]-distvec.p[i]+1)*pow(3,i);
295   return(erg+pow(3,distvec.length-1));
296  }
```

15b. Definition der Hilfsfunktion `numtodistvec`. Wandelt einen Ternärzahl in einen Zustand des Spaltenprozesses um. Umkehrfunktion zu `distvectonum`.

```
297  void numtodistvec (double num, struct ivector distvec)
298  {int i=1;
299
300   distvec.p[0]=0;
301   for(i=1;i<=distvec.length-1;i++)
302     {distvec.p[i]=distvec.p[i-1]+(int)(num-3*floor(num/3))-1; num=floor(num/3);}
303  }
```

15c. Definition der Hilfsfunktion `schritt`. Ausführen eines Übergangsschritts im Spaltenprozess bei Vorliegen des Zustands `distvec` und Erscheinen des Buchstabens `letter` in der Zeichenkette.

```
304  void schritt (struct ivector distvec, unsigned char letter, struct ucvector pattern)
305  {int x=0,y=0,i,a,b,c;
306
307   for(i=0;i<=distvec.length-2;i++)
308      {a=x+1;b=distvec.p[i+1]+1;c=distvec.p[i]+((pattern.p[i]==letter)?0:1);
309       y = (a<=b) ? ((a<=c)?a:c) : ((b<=c)?b:c); /* y=min(a,b,c) */
310       distvec.p[i]=x; x=y;
311      }
312   distvec.p[distvec.length-1]=x;
313  }
```

15d. Definition der Hilfsfunktion `numschritt`. Erfüllt dieselbe Funktion wie `schritt`, bloß auf der Ebene der Ternärdarstellung der Zustände des Spaltenprozesses. Die Schritte `numtodistvec` - `schritt` - `distvectonum` werden laufzeitoptimiert zusammengefasst.

```
314  double numschritt (double num, unsigned char letter, struct ucvector pattern)
315  {int i,x1=0,x2,y1=1,y2;
316   double erg=0;
317
318   for(i=1;i<=pattern.length;i++)
319      {x2=x1+(int)(num-3*floor(num/3)); num=floor(num/3); x1+=(pattern.p[i-1]==letter)?0:1;
320       y2=(y1<=x1)? ((y1<=x2)?y1:x2) : ((x1<=x2)?x1:x2);
321       erg+=(y2-y1+2)*pow(3,i-1); x1=x2-1; y1=y2+1;
322      }
323   return(erg+pow(3,pattern.length));
324  }
```

15e. Definition der Hilfsfunktion `transpmatrix`. Transponieren von Übergangsmatrizen.

```
325  void transptmatrix(struct tmatrix tm, struct tmatrix itm)
326  {int i,j,k;
327   int *anzahlen;
328
```

```
329   anzahlen=(int*)malloc(tm.numberofstates*sizeof(int));
330   for (i=0;i<=tm.numberofstates-1;i++) anzahlen[i]=0;
331   for (i=0;i<=tm.numberofstates*tm.alphabetsize-1;i++) anzahlen[tm.p[i]]++;
332
333   for (i=0;i<=tm.numberofstates*tm.alphabetsize-1;i++) itm.p[i]=-1;
334   for (i=0;i<=tm.numberofstates*tm.alphabetsize-1;i++) itm.w[i]=0;
335   for (i=0;i<=tm.numberofstates;i++) itm.lb[i]=0;
336
337   itm.lb[0]=0;
338   for (i=1;i<=tm.numberofstates;i++) for (j=0;j<i;j++) itm.lb[i]+=anzahlen[j];
339
340   for (i=0;i<=tm.numberofstates-1;i++) for(j=tm.lb[i];j<tm.lb[i+1];j++)
341        {k=itm.lb[tm.p[j]]; while (itm.p[k]!=-1) k++; itm.p[k]=i; itm.w[k]=tm.w[j];}
342
343   free(anzahlen);
344  }
```

15f. Definition der Hilfsfunktion `lproduct`. Matrix-Vektor-Produkt.

```
345  void lproduct (struct tmatrix tm, struct dvector v, struct dvector u)
346  {int i,j;
347
348   for (i=0;i<=v.length-1;i++) u.p[i]=0;
349   for (i=0;i<=v.length-1;i++) for (j=tm.lb[i];j<tm.lb[i+1];j++) u.p[i]+=tm.w[j]*v.p[tm.p[j]];
350  }
```

15g. Definition der Hilfsfunktion `tvd`. Berechnung des Totalvariationsabstands zwischen zwei Wahrscheinlichkeitsvektoren.

```
351  double tvd (struct dvector u, struct dvector v)
352  {int i;
353   double erg=0;
354   int *index;
355   struct dvector ergv;
356
357   ergv.p=(double *)malloc(u.length*sizeof(double));
358   ergv.length=u.length;
359   for(i=0;i<=u.length-1;i++) ergv.p[i]=(u.p[i]>v.p[i])?(u.p[i]-v.p[i]):(v.p[i]-u.p[i]);
360   index=(int*)malloc(u.length*sizeof(int));
361   for(i=0;i<=u.length-1;i++) index[i]=i;
362   dindexx(u.length,ergv.p,index);
363   for(i=0;i<=u.length-1;i++) erg+=ergv.p[index[i]];
364   free(ergv.p);free(index);
365   return(0.5*erg);
366  }
```

15h. Definition der Hilfsfunktion `indexx`. Bestimmung des Indexvektors `index`, der dafür sorgt, dass die Komponenten des `int`-Vektors `vector[index[]]` in aufsteigender Reihenfolge sind. Im Prinzip eine Variante von `quicksort`.

```
367  void indexx(int n, int *vector, int *index)
368  {int pivot=0,L=1,R=n-1,swap,zufall;
```

```
369   int i;
370
371   if (n==1) return;
372
373   srand(time(NULL)); zufall=rand()%(n-1);
374   swap=index[0];index[0]=index[zufall];index[zufall]=swap;
375
376   while(L<R)
377     {if(vector[index[L]]<=vector[index[pivot]]) L++;
378      else if(vector[index[R]]>vector[index[pivot]]) R--;
379      else {swap=index[L];index[L]=index[R];index[R]=swap;}
380     }
381   if(vector[index[L]]>vector[index[pivot]])
382     {swap=index[pivot];index[pivot]=index[L-1];index[L-1]=swap;pivot=L-1;}
383   else
384     {swap=index[pivot];index[pivot]=index[L];index[L]=swap;pivot=L;}
385
386   if(0<pivot-1) indexx(pivot, vector, index);
387   if(pivot+1<n-1) indexx(n-1-pivot, vector, index+pivot+1);
388 }
```

15i. Definition der Hilfsfunktion `dindexx`. Bestimmung des Indexvektors `index`, der dafür sorgt, dass die Komponenten des `double`-Vektors `vector[index[]]` in aufsteigender Reihenfolge sind. Im Prinzip eine Variante von `quicksort`.

```
389 void dindexx(int n, double *vector, int *index)
390 {int pivot=0,L=1,R=n-1,swap,zufall;
391  int i;
392
393  if (n==1) return;
394
395  srand(time(NULL)); zufall=rand()%n;
396  swap=index[0];index[0]=index[zufall];index[zufall]=swap;
397
398  while(L<R)
399    {if(vector[index[L]]<=vector[index[pivot]]) L++;
400     else if(vector[index[R]]>vector[index[pivot]]) R--;
401     else {swap=index[L];index[L]=index[R];index[R]=swap;}
402    }
403  if(vector[index[L]]>vector[index[pivot]])
404    {swap=index[pivot];index[pivot]=index[L-1];index[L-1]=swap;pivot=L-1;}
405  else
406    {swap=index[pivot];index[pivot]=index[L];index[L]=swap;pivot=L;}
407
408  if(0<pivot-1) dindexx(pivot, vector, index);
409  if(pivot+1<n-1) dindexx(n-1-pivot, vector, index+pivot+1);
410 }
```

Index

Literaturverzeichnis

[AF04] Aldous, D., Fill, J. (2004). *Reversible Markov Chains and Random Walks on Graphs*. Vorversion, verfügbar über die Internetseite von David Aldous: http://www.stat.berkeley.edu/~aldous/index.html. In Vorbereitung.

[Ald82] Aldous, D. J. (1982). *Markov chains with almost exponential hitting times*. Stochastic Process. Appl. **13**, 305–310.

[Ald83] Aldous, D. J. (1983). *Corrigenda: "Markov chains with almost exponential hitting times"*. Stochastic Process. Appl. **14**, 107.

[Als91] Alsmeyer, G. (1991). *Erneuerungstheorie*. Teubner Skripten zur Mathematischen Stochastik. [Teubner Texts on Mathematical Stochastics]. B. G. Teubner, Stuttgart. Analyse stochastischer Regenerationsschemata. [Analysis of stochastic regeneration schemes].

[AGG89] Arratia, R., Goldstein, L., Gordon, L. (1989). *Two moments suffice for Poisson approximations: the Chen-Stein method*. Ann. Probab. **17**, 9–25.

[AGG90] Arratia, R., Goldstein, L., Gordon, L. (1990). *Poisson approximation and the Chen-Stein method*. Statist. Sci. **5**, 403–434. With comments and a rejoinder by the authors.

[BR97] Bapat, R. B., Raghavan, T. E. S. (1997). *Nonnegative matrices and applications*, Band 64 aus *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge.

[Bil68] Billingsley, P. (1968). *Convergence of probability measures*. John Wiley & Sons Inc., New York.

[Bil79] Billingsley, P. (1979). *Probability and measure*. John Wiley & Sons, New York-Chichester-Brisbane. Wiley Series in Probability and Mathematical Statistics.

[BI04] Bioinformatics Institute, E. (2004). Internetseite des http://www.ebi.ac.uk/genomes/plasmid.html.

[BS96] Borodin, A. N., Salminen, P. (1996). *Handbook of Brownian motion—facts and formulae*. Probability and its Applications. Birkhäuser Verlag, Basel.

[Bré99] Brémaud, P. (1999). *Markov chains*, Band 31 aus *Texts in Applied Mathematics*. Springer-Verlag, New York. Gibbs fields, Monte Carlo simulation, and queues.

[Chu67] Chung, K. L. (1967). *Markov chains with stationary transition probabilities*. Second edition. Die Grundlehren der mathematischen Wissenschaften, Band 104. Springer-Verlag New York, Inc., New York.

[Dia88] Diaconis, P. (1988). *Group representations in probability and statistics*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 11. Institute of Mathematical Statistics, Hayward, CA.

[Dur96] Durrett, R. (1996). *Probability: theory and examples*. Second Auflage. Duxbury Press, Belmont, CA.

[Fel68] Feller, W. (1968). *An introduction to probability theory and its applications. Vol. I*. Third edition. John Wiley & Sons Inc., New York.

[Fu01] Fu, J. C. (2001). *Distribution of the scan statistic for a sequence of bistate trials*. J. Appl. Probab. **38**, 908–916.

[GL81] Gerber, H. U., Li, S.-Y. R. (1981). *The occurrence of sequence patterns in repeated experiments and hitting times in a Markov chain*. Stochastic Process. Appl. **11**, 101–108.

[GO81] Guibas, L. J., Odlyzko, A. M. (1981). *String overlaps, pattern matching, and nontransitive games*. J. Combin. Theory Ser. A **30**, 183–208.

[Gus97] Gusfield, D. (1997). *Algorithms on strings, trees, and sequences*. Cambridge University Press, Cambridge. Computer science and computational biology.

[HR98] Henze, N., Riedwyl, H. (1998). *How to win more. Strategies for increasing a lottery win.*. Natick, MA: A. K. Peters. x, 149 p. .

[Kei66] Keilson, J. (1966). *A limit theorem for passage times in ergodic regenerative processes*. Ann. Math. Statist. **37**, 866–870.

[Kei79] Keilson, J. (1979). *Markov chain models—rarity and exponentiality*, Band 28 aus *Applied Mathematical Sciences*. Springer-Verlag, New York.

[Kur75] Kurtz, T. G. (1975). *Semigroups of conditioned shifts and approximation of Markov processes*. Ann. Probability **3**, 618–642.

[Li80] Li, S.-Y. R. (1980). *A martingale approach to the study of occurrence of sequence patterns in repeated experiments*. Ann. Probab. **8**, 1171–1176.

[Loè63] Loève, M. (1963). *Probability theory*. Third edition. D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto, Ont.-London.

[Mor68] Morgenstern, D. (1968). *Einführung in die Wahrscheinlichkeitsrechnung und mathematische Statistik*. Zweite, verbesserte Auflage. Die Grundlehren der mathematischen Wissenschaften, Band 124. Springer-Verlag, Berlin.

[Nav01] Navarro, G. (2001). *A guided tour to approximate string matching*. ACM Computing Surveys **33**, 31–88.

[NRS85] Nobile, A. G., Ricciardi, L. M., Sacerdote, L. (1985). *Exponential trends of Ornstein-Uhlenbeck first-passage-time densities*. J. Appl. Probab. **22**, 360–369.

[PT79] Pitt, L. D., Tran, L. T. (1979). *Local sample path properties of Gaussian fields*. Ann. Probab. **7**, 477–493.

[Pit87] Pittenger, A. O. (1987). *Hitting times of sequences*. Stochastic Process. Appl. **24**, 225–240.

[Rei01] Reich, M. (2001). *Wartezeitverteilungen für Muster in zufälligen Zeichenketten*. Diplomarbeit, Universität Hannover.

[RD99] Robin, S., Daudin, J. J. (1999). *Exact distribution of word occurrences in a random sequence of letters*. J. Appl. Probab. **36**, 179–193.

[RD01] Robin, S., Daudin, J.-J. (2001). *Exact distribution of the distances between any occurrences of a set of words*. Ann. Inst. Statist. Math. **53**, 895–905.

[RD80] Rosenkrantz, W. A., Dorea, C. C. Y. (1980). *Limit theorems for Markov processes via a variant of the Trotter-Kato theorem*. J. Appl. Probab. **17**, 704–715.

[Rud96] Rudander, J. (1996). *On the first occurence of a given pattern in a semi-Markov process*. Uppsala Dissertations in Mathematics.

[Wil91] Williams, D. (1991). *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge.

[Wit03] Wittkewitz, J. (2003). *Leicht daneben – Unscharfes Suchen: Methoden und Ausblicke*. iX, Magazin für professionelle Informationstechnik , 104–107.

[Yor97] Yor, M. (1997). *Some aspects of Brownian motion. Part II*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel. Some recent martingale problems.

[ZKB97] ZKBS, Z. K. f. d. B. S. (1997). *Stellungnahme der ZKBS zur Einstufung von Agrobacterium tumefaciens*. Internetseite des Robert Koch Instituts: http://www.rki.de/GENTEC/ZKBS/ZKBS.HTM?ALLGSTELL/97/AGRO_T07.HTM&1.

Lebenslauf

Name	Marcus Reich
geboren	am 27.04.1974 in Hamburg
Familienstand	ledig

Bildungsweg

1980 - 1984	Grundschule am Öjendorfer Damm, Hamburg
1984 - 1985	Matthias-Claudius-Gymnasium, Hamburg
1985 - 1986	Orientierungsstufe Suderburg
1986 - 1993	Herzog-Ernst-Gymnasium, Uelzen

Wehrdienst

1993 - 1995	Ausbildung zum Reserveoffizier beim Panzerartilleriebataillon 85 in Lüneburg, Lehrgänge in Idar-Oberstein und Hannover

Studium

1995 - 2001	Studium an der Universität Hannover, zunächst Mathematik und Physik für das Lehramt an Gymnasien, ab dem dritten Semester Mathematik Diplom mit Nebenfach Betriebswirtschaftslehre, Abschluss als „*Diplom-Mathematiker*“

Hochschultätigkeit

1997 - 2001	studentische Hilfskraft an den Instituten für Angewandte Mathematik und für Mathematische Stochastik der Universität Hannover
2001-	wissenschaftlicher Mitarbeiter des Instituts für Mathematische Stochastik der Universität Hannover