Zufällige binäre Bäume:
Algorithmen, Asymptotik und Statistik

Von der Fakultät für Mathematik und Physik
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des Grades
Doktor der Naturwissenschaften
Dr. rer. nat.
genehmigte Dissertation
von

Dipl.-Math. Florian Dennert
geboren am 3.12.1979 in Hannover

2009

Referent: Prof. Dr. R. Grübel
Korreferent: Prof. Dr. R. Neininger

Tag der Promotion: 5. Februar 2009

Zusammenfassung

Gegenstand dieser Arbeit ist die probabilistische Analyse zweier Algorithmen, die jeweils aus einer Folge von Daten einen binären Baum erzeugen. Dies sind der *digital search tree*- und der *binary search tree*-Algorithmus. Durch die Verwendung zufälliger Daten erhalten wir zufällige binäre Bäume.

Nachdem wir in Kapitel 1 eine Einführung der grundlegenden Begriffe und Konzepte geben, untersuchen wir in Kapitel 2 zunächst das asymptotische Verhalten der Pfadlänge im *digital search tree*. Es stellt sich heraus, dass keine Verteilungskonvergenz vorliegt; jedoch kann eine Familie von Verteilungen gefunden werden, der sich die untersuchte Größe entlang von Teilfolgen asymptotisch nähert.

In Kapitel 3 verwenden wir statistische Methoden, um aus einzelnen Merkmalen eines vorgegebenen Baumes Aufschluss zu erhalten auf seine Gesamtgröße oder auch, von welchem Algorithmus er erzeugt wurde. Man erkennt, dass aufgrund der rekursiven Struktur verschiedener Verteilungsfamilien auf binären Bäumen das in Kapitel 4 weiter untersuchte Teilbaumgrößenprofil eine suffiziente Statistik für die Gesamtheit dieser Familien darstellt. Möchte man zwischen BST- und DST-Algorithmus unterscheiden, stellt sich heraus, dass die *interne Pfadlänge* bereits alle relevante Information hierfür enthält.

Kapitel 4 widmet sich dann dem Teilbaumgrößenprofil und insbesondere dessen Asymptotik. Im Gegensatz zum konventionellen *Knotenprofil* betrachten wir die Anzahl von Knoten im Baum, deren Teilbaum eine bestimmte Größe besitzt. Dies erleichtert insbesondere den rekursiven Zugang zur stochastischen Struktur der vorgegebenen Bäume. Mit Hilfe der Kontraktionsmethode weisen wir für die Verteilung der Anzahl kleiner Teilbäume asymptotisch eine mehrdimensionale Normalverteilung nach. Die großen Knoten fassen wir zu „Eisbergen“ zusammen und erhalten je nach Baumtyp verschiedene Grenzprozesse. Das allgemeine Resultat wird auf verschiedene Verteilungsfamilien auf binären Bäumen angewandt und sogar auf die nicht-binäre Baumstruktur des *Random Recursive Tree* übertragen.

Schlagwörter: Verteilungskonvergenz, Teilbaumgrößenprofil, Kontraktionsmethode

Abstract

This dissertation deals with the probabilistic analysis of two specific algorithms, each of which constructs a binary tree out of a given sequence of data. These are the *digital search tree* and the *binary search tree* algorithm. Since we consider random input the trees are also random.

After the introduction of basic notions and concepts in Chapter 1 we investigate in Chapter 2 the asymptotic behaviour of the pathlength in a *digital search tree*. It turns out that there is no weak convergence. Instead we introduce a family of probability measures which approximates the distribution of interest asymptotically along subsequences.

In Chapter 3 we make use of statistical methods to estimate the total size or to test hypotheses regarding the type of algorithm from a single quantity derived from the given tree. It turns out, that the subtree size profile is sufficient for the whole family of recursively structured distributions on binary trees. For example, if we have to distinguish between a binary search tree and a digital search tree, the pathlength of the given tree will already contain all the information needed.

Chapter 4 then deals with the subtree size profile and in particular with the corresponding asymptotics. In contrast to the well-known *node profile* the subtree size profile counts the number of nodes in a tree which are root to a subtree of a certain size. This facilitates the recursive analysis of the stochastic structure of the given trees. Using the contraction method we prove asymptotic multivariate normality for the subtrees with a small size. The large subtrees are cumulated to "icebergs" for which we obtain different limit processes depending on the type of tree. The result is of some generality and is then applied to different distribution families on binary trees. We carry this over even to the non-binary structure of *random recursive trees*.

Keywords: convergence in distribution, subtree size profile, contraction method

Inhaltsverzeichnis

1 Einführung

Nichtlineare Datenstrukturen sind aus der modernen Informatik nicht mehr wegzudenken. Darunter fallen gerichtete und ungerichtete Graphen ebenso wie binäre und gewöhnliche Bäume. Knuth schreibt in [Knu97, S. 308] dazu, Baumstrukturen seien

> *„the most important nonlinear structures that arise in computer algorithms.“*

Thema dieser Arbeit ist die Analyse bestimmter stochastischer Strukturen auf binären Bäumen. Wir geben zunächst Knuths rekursive Definition eines binären Baumes an:

Ein *binärer Baum* ist eine endliche Menge von Knoten, die entweder leer ist oder aus einer Wurzel und zwei disjunkten binären Bäumen besteht, die wir den linken und rechten Teilbaum nennen (vgl. [Knu97], S. 312). Ist t ein binärer Baum, so schreiben wir $L(t)$ für den linken, $R(t)$ für den rechten Teilbaum.

Aufgrund der Tatsache, dass ein Knoten eines binären Baumes immer zwei Teilbäume besitzt (die leer sein können), kann man den *Rand* ∂t eines binären Baumes t definieren. An jeder Stelle, an der ein leerer binärer Baum als linker oder rechter Teilbaum eines Knotens auftritt, hängen wir einen *externen* Knoten an. Die Menge der externen Knoten eines Baumes bildet seinen Rand. Zur besseren Unterscheidung der externen Knoten von den Elementen von t nennt man letztere daher mitunter *interne* Knoten; ist nur von *Knoten* die Rede, sind immer interne gemeint. (Vgl. Abbildung 1.1)

Betrachtet man einen einzelnen (internen) Knoten v eines binären Baumes t, so kann man v aufgrund der Disjunktheit der Teilbäume eindeutig als Wurzel oder als Element von $L(t)$ bzw. $R(t)$ identifizieren. Dies lässt sich rekursiv fortsetzen, bis man v als Wurzel eines Teilbaumes gefunden hat. Die Anzahl

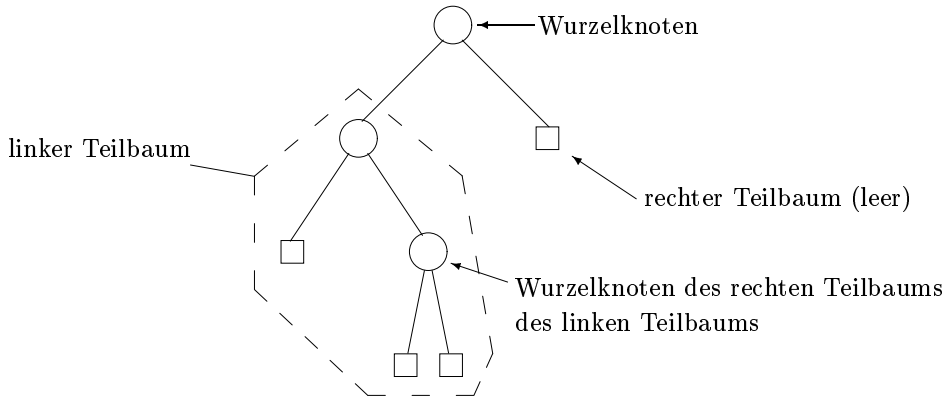

Abbildung 1.1: Veranschaulichung eines binären Baumes. Kreise stehen für interne, Quadrate für externe Knoten.

der Rekursionsschritte, die dafür nötig sind, nennt man die *Tiefe* des Knotens. Die Wurzel von t hat also die Tiefe 0, die Wurzeln von $L(t)$ und $R(t)$ haben jeweils die Tiefe 1 und so fort.

Kodiert man nun einen solchen (Rekursions-)*Abstieg* mit „0“ für links und „1“ für rechts, so erhält man zu v ein Tupel $(v_1,\ldots,v_k) \in \{0,1\}^k$, welches v eindeutig identifiziert. Die Länge k des Tupels ist dabei genau die Tiefe des Knotens v. Diese Darstellung der Knoten als (endliche) Folgen über der Menge $\{0,1\}$ motiviert eine alternative Definition des Begriffs *binärer Baum*, die an die Interpretation der theoretischen Informatik eines solchen Tupels als eine *Zeichenkette* oder ein *Wort* der Länge k über einem Alphabet (hier: $\{0,1\}$) angelehnt ist:

Ein *binärer Baum* ist eine endliche, präfixstabile Teilmenge von $\{0,1\}^*$.

Dabei ist $\{0,1\}^*$ die Menge aller Wörter über dem Alphabet $\{0,1\}$, also

$$\{0,1\}^* := \bigcup_{k=0}^{\infty} \{0,1\}^k,$$

und eine Menge t heißt *präfixstabil*, falls für jedes $v = (v_1,\ldots,v_k) \in t$, $k \in \mathbb{N}$, auch jedes Präfix $(v_1,\ldots,v_j)$, $j = 0,\ldots,k$, ein Element von t ist. Das Tupel

der Länge 0 entspricht dem *leeren Wort*, also dem Wurzelknoten, welchen wir mit $\emptyset$ bezeichnen. Das heißt: $\emptyset$ als Baum ist der leere Baum, der Baum $\{\emptyset\}$ ist derjenige, der nur aus dem Wurzelknoten besteht. Der binäre Baum in Abbildung 1.1 lässt sich also schreiben als $\{\emptyset, (0), (0,1)\}$. In der Informatik trifft man häufig die Schreibweise ϵ als Bezeichner für das leere Wort.

Betrachtet man nun eine beliebige Folge $\theta \in \{0,1\}^{\mathbb{N}}$ von Nullen und Einsen, so ist θ ein *Pfad*, der nacheinander die Knoten $(\theta_1, \ldots, \theta_j)$, $j = 0,1,2,\ldots$, besucht, bis man den Baum verlässt und auf einen externen Knoten trifft. Die Tiefe dieses externen Knotens ist die *Länge* des Pfades θ.

Auf ähnliche Weise ist die *externe Pfadlänge* als die Summe der Tiefen aller externen Knoten definiert. Analog ist die *interne Pfadlänge* die Summe der Tiefen aller internen Knoten. Zum Beispiel beträgt für den binären Baum in Abbildung 1.1 die interne Pfadlänge 3 und die externe 9. Es lässt sich leicht zeigen, dass für einen aus $n \in \mathbb{N}_0$ Knoten bestehenden binären Baum die externe Pfadlänge immer um $2n$ größer ist als die interne Pfadlänge (siehe z.B. [Mah91], S. 83). Unter strukturellen Gesichtspunkten können die beiden Begriffe also austauschbar verwendet werden.

Der *Teilbaum* $t(r)$ eines Knotens $r \in t$ ist diejenige Teilmenge von t, deren Elemente r als Präfix besitzen; die ersten $|r|$ Stellen werden dabei gelöscht, so dass r der Wurzelknoten von $t(r)$ ist. Für den Baum in Abbildung 1.1 gilt beispielsweise $t((0)) = \{\emptyset, (1)\}$.

Algorithmen

Wieder aus der Informatik entlehnt sind zwei wichtige Algorithmen, die aus einem n-Tupel von Daten einen binären Baum erzeugen. Dies sind der *binary search tree*- und der *digital search tree*-Algorithmus (abgekürzt: BST bzw. DST). Beide dienen dem Speichern von Daten und insbesondere dem schnellen Wiederauffinden derselben. Der BST-Algorithmus besitzt dabei eine starke Verwandtschaft zum berühmten Sortieralgorithmus QUICKSORT, und zwar in der nicht-randomisierten Variante, bei der immer das erste Element der zu sortierenden Liste als Pivot dient.

Zwar ist Quicksort ein rekursiv arbeitender Algorithmus, der BST-Algorithmus verfährt jedoch iterativ. Die stochastische Struktur der beiden Algorithmen ist gleich. Wir beschreiben zunächst die Vorgehensweise von BST.

Ausgangspunkt ist eine Folge $(x_1,\ldots,x_n)$ von verschiedenen Datenwerten. t_1 ist der Baum, der nur aus dem Wurzelknoten mit dem Wert x_1 besteht. Aus t_k entsteht nun t_{k+1} durch Einfügen von x_{k+1}. Dazu wird x_{k+1} mit dem Wert des Wurzelknotens (also x_1) verglichen; ist x_{k+1} kleiner, steigen wir nach links ab, bei „größer“ rechts. Nun vergleichen wir x_{k+1} mit dem Wert des dadurch erreichten Knotens und wiederholen diese Prozedur, bis wir auf einen externen Knoten treffen; dort wird x_{k+1} eingefügt.

Da für den Algorithmus nur die absoluten Ränge der Datenwerte entscheidend sind, können wir zu derjenigen Permutation π von $\{1,\ldots,n\}$ übergehen, die von diesen Rängen gebildet wird – kurz: $\pi(k)$ ist der absolute Rang von x_k innerhalb der gegebenen Folge. Abbildung 1.2 zeigt t_9 und die Entstehung von t_{10} zu der Permutation (7,9,4,2,5,8,3,10,1,6).

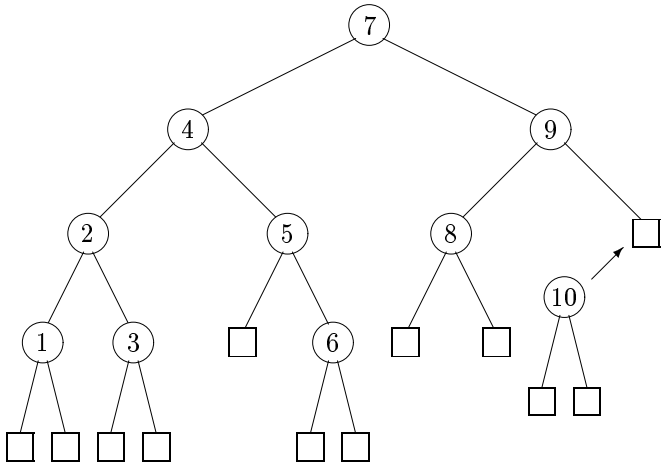

Abbildung 1.2: Funktionsweise des BST-Algorithmus

Der Zusammenhang zu Quicksort lässt sich ebenfalls an dieser Abbildung veranschaulichen: Das Pivot (hier: das erste Element) spaltet die zu sortie-

rende Permutation in die beiden Teilpermutationen (4,2,5,3,1,6) aus Elementen, die kleiner als das Pivot sind, und (9,8,10) aus den übrigen Elementen. Diese Teilpermutationen werden ihrerseits wieder sortiert. In der Sprache der binären Bäume: Setze das Pivot in den Wurzelknoten und bilde aus den verglichen zum Pivot kleineren Elementen den linken, aus den größeren den rechten Teilbaum.

Aus dem erzeugten binären Baum erhält man durch einen sogenannten *in-order*-Durchlauf (d.h. „besuche linken Teilbaum, dann Wurzel, dann rechten Teilbaum“) die Daten in sortierter Reihenfolge zurück.

Im Gegensatz zum BST-Algorithmus findet beim DST-Algorithmus keine Sortierung statt; der DST-Algorithmus dient lediglich dem Speichern und Auffinden von Daten. Ausgehend von einer Folge $(x_n)_{n\in\mathbb{N}}$ von Zahlen aus dem Einheitsintervall erzeugt er eine Folge $(t_n)_{n\in\mathbb{N}_0}$ von binären Bäumen: t_0 ist der leere Baum, t_1 besteht nur aus der Wurzel, die mit x_1 bewertet wird. Wenn der Baum t_n die Werte $x_1, \ldots, x_n$ enthält, so erhalten wir daraus t_{n+1}, indem wir anhand der Binärdarstellung von x_{n+1} im Baum absteigen, d.h. dem Pfad x_{n+1} folgen, bis wir auf einen externen Knoten stoßen. An dieser Stelle wird x_{n+1} abgelegt.

Ein Beispiel: In Abbildung 1.3 sind t_9 und die Entstehung von t_{10} durch das Einfügen von x_{10} zu der Zahlenfolge $(x_n)_{n=1,\ldots,10}$ = (0110, 1011, 0011, 0010, 0111, 0011, 1011, 0001, 0100, 1010) dargestellt.

Sucht man nun nach einem bestimmten Datum, so durchläuft man den Baum, als wolle man dieses Datum einfügen. Trifft man selbiges dabei an, so ist die Suche erfolgreich, trifft man auf einen externen Knoten, so ist das gegebene Datum im Baum nicht vorhanden.

Stochastik und probabilistische Analyse

Betrachten wir nun für festes $n \in \mathbb{N}_0$ die Menge $\mathcal{T}_n$ der binären Bäume mit n Knoten. Der einfachste Fall eines Wahrscheinlichkeitsmaßes auf dieser Menge ist die Gleichverteilung, bzw., da es sich bei $\mathcal{T}_n$ um eine endliche Menge handelt, das Laplace-Experiment auf $\mathcal{T}_n$. Einen Baum, der aus $\mathcal{T}_n$ gleichverteilt

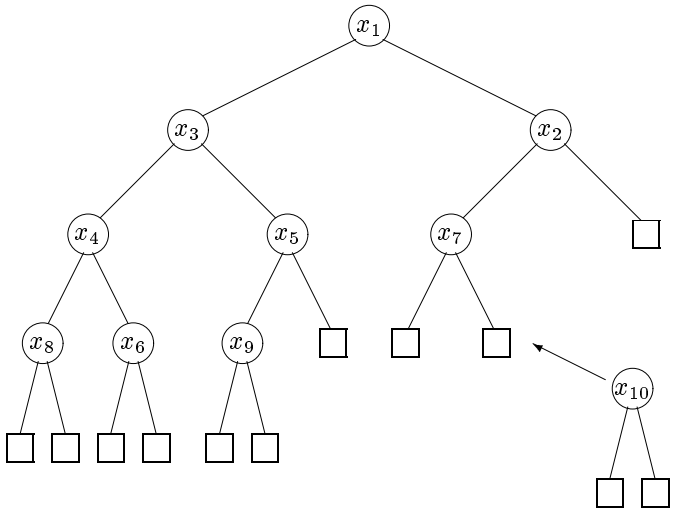

Abbildung 1.3: Funktionsweise des DST-Algorithmus

ausgewählt wird, bezeichnen wir als *Catalan*-Baum. Dieser Baum-Typ spielt jedoch nur in Abschnitt 4.5.4 eine Rolle.

Bei der Analyse von Algorithmen sind Kriterien wie die Laufzeit oder die Gesamtanzahl an benötigten Vergleichen oft direkt abhängig von der Struktur der Eingabedaten. Beispielsweise läuft die oben beschriebene Variante von Quicksort bei bereits sortierten Daten in einer Zeit ab, die proportional zum Quadrat der Anzahl der Eingabedaten ist. Im Mittel erreicht der Algorithmus dagegen für eine typische, unsortierte Eingabe eine Laufzeit proportional zu $n \log n$ (bei n Eingabedaten).

Die Stochastik bildet nun die Struktur dieser Eingabedaten nach und analysiert den Algorithmus auf der Basis einer zufälligen Eingabe. Typisch ist dabei die Verwendung von unabhängigen auf dem Einheitsintervall gleichverteilten Daten. Diesen Zugang zur Analyse von Algorithmen nennt man auch *probabilistische Analyse.*

So ist es mittels der weiter oben vorgestellten Algorithmen möglich, auf kanonische Art und Weise eine Verteilung auf $\mathcal{T}_n$ zu konstruieren. Mehr noch: Durch

das iterative Konzept der Algorithmen erhält man eine Familie $(P_n)_{n\in\mathbb{N}_0}$ von Wahrscheinlichkeitsmaßen, wobei P_n auf $\mathfrak{P}(\mathcal{T}_n)$ definiert ist:

Aus einer Folge $(U_k)_{k\in\mathbb{N}}$ von unabhängigen, auf dem Einheitsintervall gleichverteilten Zufallsvariablen konstruieren wir eine Folge $(T_n)_{n\in\mathbb{N}_0}$ binärer Bäume mittels des DST-Algorithmus und erhalten so die Verteilungsfamilie (P_n^{DST}).

Für den BST-Algorithmus kann man etwas allgemeiner vorgehen. Haben die Daten (U_k) dieselbe stetige Verteilung und sind unabhängig, so hängt (P_n^{BST}) nicht von $\mathcal{L}(U_1)$ ab, da T_n nur von der Folge der Ränge von $U_1,\ldots,U_k$ abhängt. Diese bilden aber stets eine auf der Menge der Permutationen von $\{1,\ldots,n\}$ gleichverteilte Permutation.

Überblick

In Kapitel 2 beschäftigen wir uns mit der Asymptotik der Länge eines Pfades im Digital Search Tree. Die Beobachtung, dass die Wartezeit für das Erreichen einer bestimmten Tiefe eine bestimmte stochastische Struktur aufweist, führt zu einer allgemeineren Betrachtungsweise in einem erneuerungstheoretischen Kontext.

Im darauffolgenden Kapitel wenden wir uns statistischen Fragestellungen zu. Wir setzen die stochastische Struktur eines zufälligen binären Baumes als gegeben voraus. Anhand der Tiefe des externen Knotens zum Pfad $(0,0,\ldots)$ wollen wir mit statistischen Methoden Schätzwerte für die Anzahl der Knoten im ganzen Baum erhalten. Zusätzlich geben wir exakte und approximative Konfidenzschranken für diesen Wert an.

Eine andere Fragestellung, die wir untersuchen, ist die, ob ein gegebener Baum mittels DST- oder BST-Algorithmus erzeugt wurde. Der optimale Test lässt sich dabei so formulieren, dass die Testgröße die interne Pfadlänge des vorgegebenen Baumes ist.

Dies ist die Motivation, das Profil der Teilbaumgrößen in Kapitel 4 näher zu untersuchen, aus dem sich unter anderem auch die interne Pfadlänge bestimmen lässt. Wir klassifizieren die einzelnen Knoten nach der Größe des Teilbaums, dessen Wurzel sie sind. Eine analoge Vorgehensweise ist häufig bei

der Analyse kombinatorischer Strukturen anzutreffen – in diesem Fall bilden die binären Bäume ebendiese kombinatorische Struktur.

Von Interesse ist dabei die Verteilung der Knoten auf diese Klassen, da sie Aufschluss über die Balance des Baumes gibt. Auf den algorithmischen Ursprung aus der Informatik übertragen, lassen sich also Aussagen über das Teilbaumgrößenprofil auf Aussagen über die Laufzeit des betroffenen Algorithmus übertragen.

Hauptanliegen dieses Kapitels ist dann die Untersuchung asymptotischer Fragestellungen.

Literatur

Wann immer spezielle Literatur verwendet wird oder der Vertiefung des Hintergrundes dienlich sein kann, wird sie an Ort und Stelle angeführt. Darüber hinaus seien an dieser Stelle einige Quellen genannt, die sich in ihrem Gebiet als Referenz durchgesetzt haben. Dies ist zuvorderst das Standardwerk von Knuth ([Knu73],[Knu97]) zur grundlegenden Theorie von Such- und Sortieralgorithmen. Zufällige Bäume und insbesondere die stochastische Struktur binärer Bäume werden bei Mahmoud ([Mah91]) ausführlich behandelt. Zieht man einen analytischeren Zugang zur Analyse von Algorithmen vor, so kann das Buch von Sedgewick und Flajolet ([SF96]) als Einstieg dienen. Bezüglich wahrscheinlichkeitstheoretischer Grundlagen sei auf die Werke von Feller ([Fel71]) und Billingsley ([Bil68],[Bil86]) verwiesen, für statistische Grundlagen referenzieren wir Bickel und Doksum ([BD76]).

2 Die DST-Pfadlänge als nicht-homogener Erneuerungsprozess

Es seien $Y_1, Y_2, \ldots$ unabhängige und nicht-negative Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{A}, P)$. Wir betrachten die Y_k als Lebensdauern und definieren die Gesamtlebensdauer S_n der ersten n Komponenten als

$$S_n := \sum_{k=1}^{n} Y_k, \qquad n \in \mathbb{N}.$$

Zusammen mit $S_0 \equiv 0$ ist so S_n der Zeitpunkt der n-ten Erneuerung. Der erste Erneuerungszeitpunkt fällt also mit dem Ausfall der ersten Komponente zusammen.

Nun betrachten wir für $t \geq 0$ die Zahl N_t der Erneuerungen bis einschließlich zum Zeitpunkt t

$$N_t := \sup\{n \in \mathbb{N}_0 : S_n \leq t\}.$$

Der Prozess $(N_t)_{t \geq 0}$ heißt *Erneuerungsprozess*; Abbildung 2.1 zeigt ein Beispiel. An dieser Abbildung kann man auch eine für die Erneuerungstheorie typische Manipulationsmöglichkeit ablesen: Es gilt nämlich für $k \in \mathbb{N}_0$ und $t \geq 0$

$$N_t = k \quad \text{genau dann, wenn} \quad S_k \leq t < S_{k+1},$$

insbesondere erhalten wir durch Vereinigung dieser Ereignisse für $k, k+1, \ldots$

$$P(N_t \geq k) = P(S_k \leq t), \qquad \text{für alle } k \in \mathbb{N}_0 \text{ und alle } t \geq 0. \tag{2.1}$$

In der klassischen Erneuerungstheorie geht man davon aus, dass die Lebensdauern identisch verteilt sind. In diesem Fall nennt man (N_t) einen *homogenen* Erneuerungsprozess.

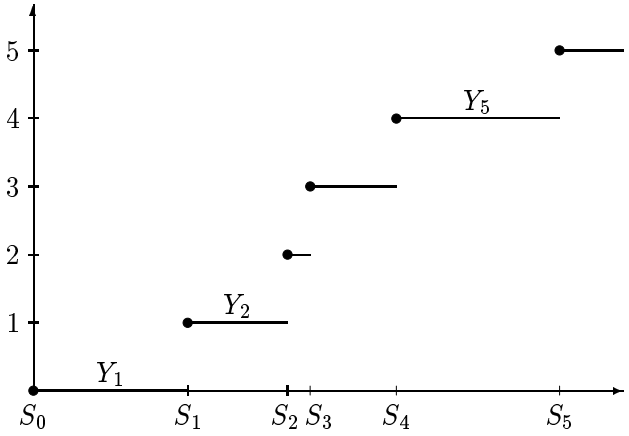

Abbildung 2.1: Ein Pfad eines Erneuerungsprozesses

Wir betrachten nun eine Familie von Lebensdauerverteilungen, die in gewisser Art und Weise exponentiell wachsen. Es wird sich herausstellen, dass im Gegensatz zum klassischen Fall (siehe [Fel71], XI.5) auch nach einer entsprechenden Normierung keine Verteilungskonvergenz vorliegt. Statt dessen stellen sich Fluktuationen in der asymptotischen Verteilung ein, die durch eine Familie von Wahrscheinlichkeitsmaßen modelliert werden können.

Diese *nicht-homogene* Erneuerungstheorie lässt sich auf das Verhalten eines Pfades im digitalen Suchbaum übertragen. Es sei (T_n) eine Folge von binären Bäumen, die vom DST-Algorithmus aus einer Folge (U_n) von unabhängigen, unif(0,1)-verteilten Zufallsvariablen erzeugt werden (vgl. Kapitel 1). Die in diesem Kapitel untersuchte Fragestellung ist die nach der Tiefe, in welcher der nachfolgende Datenwert U_{n+1} eingefügt wird. Bei der beispielhaften Betrachtung des Pfades $(0,0,\ldots)$, also einem „nur links“-Abstieg durch den Baum, erkennt man, dass die stochastische Wartezeit auf einen hier eingefügten Knoten mit zunehmendem n exponentiell wächst. Diesen Zusammenhang werden wir in den nachfolgenden Abschnitten erläutern.

Zu diesem Zweck betrachten wir in Abschnitt 2.1 zunächst ganz allgemein exponentiell wachsende Lebensdauern und stellen fest, dass eine normierte Version des Erneuerungsprozesses in Verteilung gegen eine Familie von Grenzmaßen

konvergiert. Schließlich wenden wir die gewonnenen Resultate in Abschnitt 2.2 auf einen Spezialfall an, der der oben beschriebenen Situation entspricht. In diesem Fall lässt sich die Familie der Grenzmaße mit Hilfe einer Mischung von Wahrscheinlichkeitsmaßen darstellen; dies untersuchen wir in Abschnitt 2.3.

Als erstes wollen wir das exponentielle Wachstum der Lebensdauerverteilungen formalisieren. Dazu sei $(Y_k)_{k\in\mathbb{N}}$ eine Folge von nicht negativen und unabhängigen Zufallsvariablen, für die ein $\alpha > 1$ existiert, so dass

$$\alpha^{-k} Y_k \xrightarrow{\text{distr}} Y_\infty \quad \text{und} \quad \alpha^{-k}\mathbb{E}Y_k \to \mathbb{E}Y_\infty \tag{2.2}$$

für eine Zufallsvariable Y_∞, jeweils mit $k \to \infty$. Außerdem fordern wir, dass alle beteiligten Erwartungswerte existieren, also

$$\mathbb{E}Y_k < \infty \text{ für alle } k \in \mathbb{N} \quad \text{und} \quad \mathbb{E}Y_\infty < \infty. \tag{2.3}$$

Beispiel 2.1 Sei Y_k geometrisch verteilt mit Parameter q^k für alle $k \in \mathbb{N}$ und ein festes $q \in (0,1)$. Dann gilt mit $P(Y_k > j) = (1-q^k)^j$ für $\alpha = q^{-1}$:

$$\begin{aligned} P(\alpha^{-k} Y_k \le x) &= P(Y_k \le \lfloor q^{-k} x \rfloor) \\ &= 1 - (1-q^k)^{\lfloor q^{-k} x \rfloor} \\ &\to 1 - \mathrm{e}^{-x} \end{aligned}$$

und wir haben die erste Hälfte von (2.2) für Y_∞ exponentialverteilt mit Parameter 1. Die zweite Hälfte von (2.2) ist sofort klar, denn es gilt $\mathbb{E}Y_k = q^{-k}$ und $\alpha^{-k}\mathbb{E}Y_k = 1 = \mathbb{E}Y_\infty$, und auch (2.3) ist trivialerweise erfüllt.

2.1 Verteilungsasymptotik des Erneuerungsprozesses

Natürlich strebt der Erneuerungsprozess sowohl bei identisch verteilten als auch unter der Annahme von exponentiell wachsenden Lebensdauern (2.2) mit wachsendem Zeitparameter gegen unendlich. Wir beschäftigen uns daher mit einer normierten Variante, deren Grenzverteilung nicht degeneriert ist: Das

im Falle von exponentiell wachsenden Lebensdauern logarithmische Wachstum des Erneuerungsprozesses legt es nahe, die Verteilung von

$$N_t - \lfloor \log_\alpha t \rfloor$$

zu betrachten. Wie bereits am Anfang des Kapitels erwähnt, liegt für diesen normierten Prozess keine Verteilungskonvergenz vor. Entlang von Teilfolgen, die einer bestimmten Bedingung genügen, erhält man jedoch Konvergenz. Diese Teilfolgenkonvergenz lässt sich dann zusammenfassen zu einer Konvergenzaussage, die unter Zuhilfenahme einer Familie von Wahrscheinlichkeitsmaßen formuliert werden kann.

Eine wichtige Rolle für die Verteilungsasymptotik des Erneuerungsprozesses spielt die Zufallsgröße

$$S_\infty \ := \ \sum_{k=0}^{\infty} \alpha^{-k} Y_{\infty,k}, \tag{2.4}$$

wobei $Y_{\infty,0}, Y_{\infty,1}, \ldots$ unabhängig und identisch $\mathcal{L}(Y_\infty)$-verteilt sind. Setzen wir (2.2) und (2.3) voraus, speziell also die Existenz des Erwartungswertes von Y_∞, so ergibt sich mit dem Satz von der monotonen Konvergenz

$$\mathbb{E}S_\infty \ = \ \sum_{k=0}^{\infty} \alpha^{-k} \mathbb{E}Y_\infty \ = \ \frac{\mathbb{E}Y_\infty}{1-\alpha}.$$

Damit folgt insbesondere, dass S_∞ fast sicher endlich ist; die Summe (2.4) konvergiert also fast sicher.

Darüber hinaus ergibt sich die Verteilung von S_∞ als Grenzwert der Verteilungen der skalierten Partialsummen $\alpha^{-n} S_n$. Eine analoge Konvergenzaussage beweisen wir zunächst für Folgen reeller Zahlen.

Lemma 2.2 Ist $(x_n)_{n\in\mathbb{N}}$ eine Folge reeller Zahlen mit $\lim_{n\to\infty} x_n = x \in \mathbb{R}$ und $\alpha > 1$, so gilt

$$\lim_{n\to\infty} \sum_{k=0}^{n-1} \alpha^{-k} x_{n-k} \ = \ \frac{x}{1-\alpha}.$$

Beweis. Für $n \in \mathbb{N}$ sei $f_n : \mathbb{N}_0 \to \mathbb{R}$ definiert durch

$$f_n(k) := \begin{cases} x_{n-k}, & k = 0, \ldots, n-1, \\ 0 & \text{sonst.} \end{cases}$$

Da (x_n) konvergiert, ist $\sup_{n\in\mathbb{N}} |x_n|$ und damit auch $\sup_{n\in\mathbb{N}} \|f_n\|_\infty$ endlich. Definiert man auf $(\mathbb{N}_0, \mathfrak{P}(\mathbb{N}_0))$ das Maß μ durch $\mu(\{n\}) := \alpha^{-n}$, $n \in \mathbb{N}_0$, so ist μ ein endliches Maß und man erhält mit dem Satz von der majorisierten Konvergenz

$$\lim_{n\to\infty} \sum_{k=0}^{n-1} \alpha^{-k} x_{n-k} = \lim_{n\to\infty} \int_{\mathbb{N}_0} f_n \, d\mu = x \int_{\mathbb{N}_0} d\mu = x \cdot \sum_{k=0}^{\infty} \alpha^{-k} = \frac{x}{1-\alpha}. \quad \square$$

Für die stochastische Version dieser Aussage benötigen wir noch ein Hilfsmittel aus dem Gebiet der *probability metrics*. Es setzt Konvergenz bezüglich der Wasserstein-Metrik und schwache Konvergenz miteinander in Beziehung. Die *Wasserstein-Metrik* ist definiert als

$$d_{\mathrm{W}}(\mu, \nu) := \inf\{\mathbb{E}|X - Y| : \mathcal{L}(X) = \mu, \mathcal{L}(Y) = \nu\},$$

wobei μ und ν zwei Wahrscheinlichkeitsmaße sind. Der Einfachheit halber schreiben wir $d_{\mathrm{W}}(X,Y)$ anstelle von $d_{\mathrm{W}}(\mathcal{L}(X),\mathcal{L}(Y))$. Das folgende Lemma ergibt sich auch als Spezialfall von Theorem 6.3.3 in [Rac91]. Die Argumentation im zweiten Teil des hier angegebenen Beweises kann jedoch im nachfolgenden Satz noch einmal verwendet werden.

Lemma 2.3 Seien X, X_n, $n \in \mathbb{N}$, nichtnegative Zufallsvariablen mit jeweils existierendem Erwartungswert. Dann sind äquivalent, jeweils mit $n \to \infty$,

(i) $d_{\mathrm{W}}(X_n,X) \to 0$,

(ii) $X_n \xrightarrow{\text{distr}} X$ und $\mathbb{E}X_n \to \mathbb{E}X$.

Beweis. (i) $\Rightarrow$ (ii): Zu $\mathcal{L}(X_n)$ und $\mathcal{L}(X)$ existiert aufgrund der Infimumseigenschaft von d_{W} eine Folge (Y_n) von Zufallsvariablen auf demselben Wahrscheinlichkeitsraum mit $\mathcal{L}(Y_n) = \mathcal{L}(X_n)$ für alle $n \in \mathbb{N}$ und $\mathcal{L}(Y) = \mathcal{L}(X)$, für

die nach Voraussetzung gilt

$$\mathbb{E}|Y_n - Y| \;\to\; 0 \quad \text{mit } n \to \infty.$$

Da alle beteiligten Erwartungswerte existieren, folgt die Verteilungskonvergenz aus der Chebyshevschen Ungleichung, die Konvergenz der Erwartungswerte trivialerweise aus der Dreiecksungleichung.

(ii) $\Leftarrow$ (i): Für diesen Beweisschritt benötigen wir zwei Hilfsmittel. Nach dem zweiten Teil von Theorem 5.4 in [Bil68] gilt: Sind $Z,Z_1,Z_2,\ldots$ nichtnegative Zufallsvariablen mit jeweils existierendem Erwartungswert, und gilt $Z_n \xrightarrow{\text{distr}} Z$ sowie $\mathbb{E}Z_n \to \mathbb{E}Z$, jeweils mit $n \to \infty$, so ist $\{Z_n : n \in \mathbb{N}\}$ gleichgradig integrierbar. Weiter besagt ein Standardresultat zur gleichgradigen Integrierbarkeit, dass für eine Folge von Zufallsvariablen $Z_1,Z_2,\ldots$ gleichgradige Integrierbarkeit und Konvergenz in Wahrscheinlichkeit gegen eine weitere Zufallsvariable Z ein notwendiges und hinreichendes Kriterium bilden für die Konvergenz von Z_n gegen Z im Mittel, d.h. $\mathbb{E}|Z_n - Z| \to 0$ (siehe z.B. [Bau74], Satz 20.4).

Nach dem Einbettungssatz von Skorohod lassen sich nun Zufallsvariablen Y und Y_k, $k \in \mathbb{N}$, auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega,\mathfrak{A},P)$ finden, so dass sogar $Y_n \to Y$ mit $n \to \infty$ fast sicher gilt. Insbesondere konvergiert (Y_n) dann in Wahrscheinlichkeit gegen Y, und die beiden Hilfsmittel implizieren $\mathbb{E}|Y_n - Y| \to 0$. Damit folgt sofort

$$d_{\mathrm{W}}(X_n,X) \;\le\; \mathbb{E}|Y_n - Y| \;\to\; 0. \qquad \square$$

Das folgende Lemma ist eine Erweiterung von Lemma 1 aus [DG07]. Der Beweis unterscheidet sich vor allem durch den Gebrauch der Wasserstein-Metrik wesentlich von dem dort gegebenen.

Lemma 2.4 Unter den Bedingungen (2.2) und (2.3) gilt

$$\alpha^{-n} S_n \xrightarrow{\text{distr}} S_\infty$$

sowie $\mathbb{E}\alpha^{-n} S_n \to \mathbb{E}S_\infty$, jeweils mit $n \to \infty$.

Beweis. Die Behauptung ist nach Lemma 2.3 äquivalent zu

$$d_{\mathrm{W}}\left(\alpha^{-n}S_n, S_\infty\right) \;\to\; 0, \qquad n \to \infty.$$

Dies wiederum folgt nach der Dreiecksungleichung für die Wasserstein-Metrik sofort aus den beiden Teilaussagen

$$d_{\mathrm{W}}\left(\alpha^{-n}S_n, \sum_{k=0}^{n-1}\alpha^{-k}Y_{\infty,n-k}\right) \;\to\; 0 \tag{2.5}$$

$$\text{und} \quad d_{\mathrm{W}}\left(S_\infty, \sum_{k=0}^{n-1}\alpha^{-k}Y_{\infty,n-k}\right) \;\to\; 0. \tag{2.6}$$

Wir betrachten zunächst (2.6). Da $Y_{\infty,0}$,$Y_{\infty,1}, \ldots$ unabhängig sind und derselben Verteilung genügen, gilt

$$\begin{aligned} d_{\mathrm{W}}\left(S_\infty, \sum_{k=0}^{n-1}\alpha^{-k}Y_{\infty,n-k}\right) \;&=\; d_{\mathrm{W}}\left(S_\infty, \sum_{k=0}^{n-1}\alpha^{-k}Y_{\infty,k}\right) \\ &\le\; \mathbb{E}\left|\sum_{k=0}^{\infty}\alpha^{-k}Y_{\infty,k} \;-\; \sum_{k=0}^{n-1}\alpha^{-k}Y_{\infty,k}\right| \\ &=\; \sum_{k=n}^{\infty}\alpha^{-k}\mathbb{E}Y_{\infty,0} \end{aligned}$$

nach Definition (2.4) von S_∞. Der verbleibende Ausdruck ist eine Nullfolge in n, da $\alpha > 1$ und $\mathbb{E}Y_{\infty,0}$ endlich ist.

Für den Nachweis von (2.5) benutzen wir eine Variante der Quantilkonstruktion. Sei $(U_n)_{n\in\mathbb{N}}$ eine Folge von unabhängigen, jeweils auf dem Einheitsintervall gleichverteilten Zufallsvariablen. Seien F_k die Verteilungsfunktionen zu Y_k, $k \in \mathbb{N}$, und sei F die Verteilungsfunktion zu Y_∞. Dann definieren wir

$$\tilde{Y}_k := F_k^{-1}(U_k) \qquad \text{und} \qquad \tilde{Y}_{\infty,k} := F^{-1}(U_k) \text{ für } k \in \mathbb{N}.$$

Dabei bezeichnet $G^{-1}(x) := \inf\{y \in \mathbb{R} : G(y) \ge x\}$ die Quantilfunktion zu einer Verteilungsfunktion G. Klar ist, dass dann gilt $\mathcal{L}(Y_k) = \mathcal{L}(\tilde{Y}_k)$ und $\mathcal{L}(Y_{\infty,k}) = \mathcal{L}(\tilde{Y}_{\infty,k})$ für alle $k \in \mathbb{N}$. Aufgrund der Unabhängigkeit von $\{Y_k :$

$k \in \mathbb{N}\}$ bzw. $\{U_k : k \in \mathbb{N}\}$ haben also auch $(Y_1, \ldots, Y_n)$ und $(\tilde{Y}_1, \ldots, \tilde{Y}_n)$ ebenso dieselbe gemeinsame Verteilung wie $(Y_{\infty,1}, Y_{\infty,2}, \ldots)$ und $(\tilde{Y}_{\infty,1}, \tilde{Y}_{\infty,2}, \ldots)$. Daraus folgt nun mit $\tilde{S}_n := \sum_{k=1}^{n} \tilde{Y}_k$, dass gilt

$$\mathcal{L}\big(S_n\big) = \mathcal{L}\big(\tilde{S}_n\big) \tag{2.7}$$

Weiterhin erhält man durch Umkehr der Summationsreihenfolge

$$\alpha^{-n} S_n \;=_{\text{distr}}\; \alpha^{-n} \sum_{k=0}^{n-1} \tilde{Y}_{n-k} \;=\; \sum_{k=0}^{n-1} \alpha^{-k} (\alpha^{-(n-k)} \tilde{Y}_{n-k}), \quad n \in \mathbb{N},$$

und es folgt nach Dreiecksungleichung

$$\begin{aligned}
&d_{\mathrm{W}}\Big(\alpha^{-n} S_n, \sum_{k=0}^{n-1} \alpha^{-k} Y_{\infty,n-k}\Big) \\
&\leq\; \mathbb{E}\Big|\sum_{k=0}^{n-1} \alpha^{-k} \Big(\alpha^{-(n-k)} \tilde{Y}_{n-k}\Big) \;-\; \sum_{k=0}^{n-1} \alpha^{-k} \tilde{Y}_{\infty,n-k}\Big| \\
&\leq\; \sum_{k=0}^{n-1} \alpha^{-k} \mathbb{E}\Big|\alpha^{-(n-k)} \tilde{Y}_{n-k} - \tilde{Y}_{\infty,n-k}\Big| \;=\; \sum_{k=0}^{n-1} \alpha^{-k} x_{n-k},
\end{aligned}$$

mit

$$x_k \;:=\; \mathbb{E}\big|\tilde{Y}_{\infty,k} - \alpha^{-k} \tilde{Y}_k\big|, \qquad k \in \mathbb{N}.$$

Der Wert von x_k hängt von U_k nur über dessen Verteilung ab. Die zum Anfang des Beweises analoge Konstruktion können wir also durchführen mit $U \sim \mathrm{unif}(0,1)$, $Y'_\infty := F^{-1}(U)$ und $Y'_k := F_k^{-1}(U)$. Wir erhalten

$$x_k \;=\; \mathbb{E}|Y'_\infty - \alpha^{-k} Y'_k|.$$

Nach Konstruktion und als Konsequenz aus dem Einbettungssatz von Skorohod konvergiert $\alpha^{-k} Y'_k \to \tilde{Y}'_\infty$ mit $k \to \infty$ fast sicher. Wie im zweiten Teil des Beweises zu Lemma 2.3 benutzen wir die dort verwendeten zwei Hilfsmittel, um auch hier aus den Voraussetzungen (2.2) und (2.3) zu folgern, dass (x_k) eine Nullfolge ist. Mit Lemma 2.2 folgt insgesamt die Behauptung. □

Wir definieren nun eine Familie von Wahrscheinlichkeitsmaßen $\{Q_\eta : \eta \in [0,1]\}$ durch

$$Q_\eta \;:=\; \mathcal{L}\big(\lfloor -\log_\alpha S_\infty + \eta \rfloor\big) \quad \text{für } 0 \le \eta \le 1. \tag{2.8}$$

Der folgende Satz zeigt, dass der normierte Erneuerungsprozess sich dieser Familie asymptotisch nähert. Dabei erhalten wir entlang solcher Folgen (t_n), für die $\{\log_\alpha t_n\}$ konvergiert, sogar Verteilungskonvergenz. $\{x\}$ bezeichnet dabei den *fractional part* einer reellen Zahl x, also $\{x\} := x - \lfloor x \rfloor$.

Satz 2.5 Gelte (2.2), (2.3) und sei $\mathcal{L}(Y_\infty)$ stetig. Ist $(t_n)_{n\in\mathbb{N}}$ eine Folge reeller Zahlen mit $t_n \to \infty$ und $\{\log_\alpha t_n\} \to \eta$ für ein $\eta \in [0,1]$, so gilt mit $n \to \infty$

$$\mathcal{L}\big(N_{t_n} - \lfloor \log_\alpha t_n \rfloor\big) \;\xrightarrow{\text{w}}\; Q_\eta.$$

Beweis. Wir definieren zwei Abkürzungen: Für $n \in \mathbb{N}$ sei

$$k_n := \lfloor \log_\alpha t_n \rfloor, \quad \text{und} \quad \eta_n := \{\log_\alpha t_n\}.$$

Das erneuerungstheoretische Standardargument (2.1) liefert hier für beliebiges $j \in \mathbb{Z}$ und n so groß, dass $j \ge -k_n$,

$$P\big(N_{t_n} - k_n = j\big) \;=\; P\big(S_{k_n+j} \le t_n\big) \;-\; P\big(S_{k_n+j+1} \le t_n\big).$$

Wir betrachten zunächst nur den ersten Ausdruck auf der rechten Seite. Für das gegebene Ereignis gilt die folgende Äquivalenz:

$$S_{k_n+j} \le t_n \;\iff\; -\log_\alpha\big(\alpha^{-k_n-j} S_{k_n+j}\big) + \eta_n \;\ge\; -\log_\alpha\big(\alpha^{-k_n-j} t_n\big) + \eta_n$$

und auf der rechten Seite bleibt nach Vereinfachen

$$-\log_\alpha\big(\alpha^{-k_n-j} t_n\big) + \eta_n \;=\; k_n + j - \log_\alpha t_n + \eta_n \;=\; j.$$

Nach Lemma 2.4 und dem *continuous mapping theorem* (siehe z.B. [Bil68, S. 31]) konvergiert mit $n \to \infty$

$$-\log_\alpha\big(\alpha^{-k_n-j} S_{k_n+j}\big) \;\xrightarrow{\text{distr}}\; -\log_\alpha S_\infty. \tag{2.9}$$

Um nun links η_n und rechts η hinzuaddieren zu können, ohne dabei die Verteilungskonvergenz zu stören, gebrauchen wir Theorem 4.1 aus [Bil68]. Dieses Theorem besagt, dass für Zufallsvariablen $X,X_1,X_2,\ldots$ und $Z,Z_1,Z_2,\ldots$ mit $X_n \xrightarrow{\text{distr}} X$ und $Z_n \xrightarrow{\text{p}} Z$ gilt, dass $X_n+Z_n \xrightarrow{\text{distr}} X+Z$. Wir verwenden diesen Satz mit $X_n = -\log_\alpha(\alpha^{-k_n-j}S_{k_n+j})$ und $Z_n \equiv \eta_n$. Nach Voraussetzung ist (η_n) eine konvergente Folge reeller Zahlen, insbesondere konvergiert (η_n) also in Wahrscheinlichkeit. Zusammen mit (2.9) folgt aus dem zitierten Theorem

$$-\log_\alpha\big(\alpha^{-k_n-j}S_{k_n+j}\big)+\eta_n \;\xrightarrow{\text{distr}}\; -\log_\alpha S_\infty+\eta \tag{2.10}$$

mit $n\to\infty$. Analoges gilt, wenn wir den Index j um eine Position nach rechts verschieben.

Um nun den Grenzübergang

$$P\big(S_{k_n+j}\le t_n\big) \;\to\; P\big(\log_\alpha S_\infty \ge j\big), \quad n\to\infty,$$

aus der soeben nachgewiesenen Verteilungskonvergenz zu folgern, müssen wir noch zeigen, dass die Verteilungsfunktion von $-\log_\alpha S_\infty+\eta$ in j stetig ist. Dies ist insbesondere der Fall, wenn S_∞ eine stetige Verteilungsfunktion besitzt. Nach Voraussetzung ist $\mathcal{L}(Y_\infty)$ stetig, und aus der Definition (2.4) von S_∞ erhalten wir die Verteilungsgleichheit

$$\mathcal{L}(S_\infty) \;=\; \mathcal{L}(\alpha^{-1}S_\infty)\star\mathcal{L}(Y_\infty).$$

Damit ist $\mathcal{L}(S_\infty)$ eine Faltung mit mindestens einer stetigen Verteilung und damit selbst stetig.

Insgesamt haben wir also

$$\begin{aligned} P\big(N_{t_n}-k_n=j\big) \;\to\; & P\big(-\log_\alpha S_\infty+\eta\ge j\big) \\ & - P\big(-\log_\alpha S_\infty+\eta\ge j+1\big) \\ = \; & P\big(\lfloor -\log_\alpha S_\infty+\eta\rfloor = j\big) \;=\; Q_\eta\big(\{j\}\big) \end{aligned}$$

und damit folgt die Behauptung. □

Wir wollen die Aussage dieses Satzes nun verallgemeinern in dem Sinne, dass die Familie $\{Q_\eta : \eta \in [0,1]\}$ für beliebige Folgen (t_n) mit $t_n \to \infty$ den asymptotischen Ersatz einer Grenzverteilung darstellt. Dazu benötigen wir zunächst noch eine Stetigkeitsaussage.

Lemma 2.6 Ist S_∞ stetig, so ist die Abbildung

$$[0,1] \ni \eta \mapsto Q_\eta = \mathcal{L}\big(\lfloor -\log_\alpha S_\infty + \eta \rfloor\big)$$

stetig im Sinne der Verteilungskonvergenz.

Beweis. Wir müssen zeigen, dass für eine Folge $(\eta_n) \in [0,1]^{\mathbb{N}}$ mit $\eta_n \to \eta$ die schwache Konvergenz

$$Q_{\eta_n} \xrightarrow{\mathrm{w}} Q_\eta$$

erfüllt ist. Sei also (η_n) eine solche Folge. Dann genügt zu zeigen, dass für jedes $j \in \mathbb{Z}$ gilt, dass

$$\lim_{n\to\infty} Q_{\eta_n}\big(\{j\}\big) = Q_\eta\big(\{j\}\big).$$

Sei also $j \in \mathbb{Z}$. Aus dem Beweis zu Satz 2.5 ist bekannt, dass

$$Q_{\eta_n}\big(\{j\}\big) = P\big(-\log_\alpha S_\infty \geq j - \eta_n\big) - P\big(-\log_\alpha S_\infty \geq j + 1 - \eta_n\big).$$

Nach Voraussetzung ist die Verteilung von S_∞ stetig, somit ist die Verteilung von $-\log_\alpha S_\infty$ ebenfalls stetig und es folgt

$$\begin{aligned} &P\big(-\log_\alpha S_\infty \geq j - \eta_n\big) - P\big(-\log_\alpha S_\infty \geq j + 1 - \eta_n\big) \\ &\to P\big(-\log_\alpha S_\infty \geq j - \eta\big) - P\big(-\log_\alpha S_\infty \geq j + 1 - \eta\big) \\ &= Q_\eta\big(\{j\}\big). \end{aligned}$$

□

Der *Totalvariationsabstand* zweier Wahrscheinlichkeitsmaße μ und ν auf einem messbaren Raum $(\Omega, \mathfrak{A})$ ist definiert als

$$d_{\mathrm{TV}}(\mu,\nu) := \sup_{B\in\mathfrak{A}} \big|\mu(B) - \nu(B)\big|.$$

Sind die Wahrscheinlichkeitsmaße auf abzählbaren Mengen wie in diesem Fall auf $\mathbb{Z}$ konzentriert, so gilt

$$d_{\mathrm{TV}}(\mu,\nu) = \frac{1}{2}\sum_{j\in\mathbb{Z}}\Big|\mu(\{j\}) - \nu(\{j\})\Big|.$$

Des Weiteren konvergiert eine Familie von Wahrscheinlichkeitsmaßen, die alle auf derselben abzählbaren Menge konzentriert sind, nach Scheffés Lemma (siehe z.B. [Bil86], S. 218) genau dann in Verteilung, wenn der Totalvariationsabstand zum Grenzmaß im Limes verschwindet. Wir können die Aussage von Satz 2.5 also formulieren als

$$\lim_{n\to\infty} d_{\mathrm{TV}}\Big(\mathcal{L}\big(N_{t_n} - \lfloor\log_\alpha t_n\rfloor\big), Q_\eta\Big) = 0 \tag{2.11}$$

für alle Folgen $(t_n)_{n\in\mathbb{N}}$, für die gilt, dass $t_n \to \infty$ und $\{\log_\alpha t_n\} \to \eta$, jeweils mit $n \to \infty$.

Damit ergibt sich eine neue Formulierung von Satz 2.5, die auf die Beschränkung der Konvergenz entlang bestimmter Teilfolgen verzichtet:

Satz 2.7 Gelte (2.2), (2.3) und sei $\mathcal{L}(Y_\infty)$ stetig. Dann ist

$$\lim_{t\to\infty} d_{\mathrm{TV}}\Big(\mathcal{L}\big(N_t - \lfloor\log_\alpha t\rfloor\big), Q_{\{\log_\alpha t\}}\Big) \;=\; 0.$$

Beweis. Wir schreiben zur Abkürzung

$$\begin{aligned} a(t) \;&:=\; d_{\mathrm{TV}}\Big(\mathcal{L}\big(N_t - \lfloor\log_\alpha t\rfloor\big), Q_{\{\log_\alpha t\}}\Big) \\ \text{und}\quad \tilde{a}_\eta(t) \;&:=\; d_{\mathrm{TV}}\Big(\mathcal{L}\big(N_t - \lfloor\log_\alpha t\rfloor\big), Q_\eta\Big). \end{aligned}$$

Angenommen, es existiert eine Folge $(t_n)_{n\in\mathbb{N}}$ mit $t_n \to \infty$ und $a(t_n) \to c > 0$. Dann existiert dazu eine Teilfolge (t_{n_k}) mit

$$\limsup_{n\to\infty}\{\log_\alpha t_n\} \;=\; \lim_{k\to\infty}\{\log_\alpha t_{n_k}\} \;=:\; \eta.$$

Nach der Formulierung (2.11) der Aussage von Satz 2.5 gilt dann $\tilde{a}_\eta(t_{n_k}) \to 0$. Zusammen mit Lemma 2.6 folgt $a(t_{n_k}) \to 0$. Widerspruch. □

2.2 Asymptotik der Einfügetiefe bei DST

Sei nun $(U_n)_{n\in\mathbb{N}}$ eine Folge unabhängiger, auf dem Einheitsintervall gleichverteilter Zufallsvariablen. Sei $(T_n)_{n\in\mathbb{N}_0}$ die Folge der (zufälligen) DSTs, die von der Folge der U_n erzeugt werden (vgl. Kapitel 1). Wir betrachten für eine feste Folge $\theta \in \{0,1\}^{\mathbb{N}}$ den Pfad entlang eines Baumes T_n, $n \in \mathbb{N}$. Dabei sei $X_n(\theta)$ die Länge des Pfades. Diese Größe wird in der Literatur oft als *unsuccessful search* bezeichnet, da sie Aufschluss über die Zeitkomplexität einer (vergeblichen) Suchoperation gibt (vgl. [Mah91, S. 271]).

Man kann θ auch als eine Intervallschachtelung für den Wert $\sum_{k=1}^{\infty} 2^{-k}\theta_k \in [0,1]$ ansehen; die Intervalle haben die Länge 2^{-k}, $k \in \mathbb{N}$. Fragt man sich, mit welcher Wahrscheinlichkeit U_{n+1} in das Intervall der Länge 2^{-k} fällt, das von $\theta_1, \ldots, \theta_k$ definiert wird, so wird klar, dass für $n \in \mathbb{N}_0$ und $k \leq n$ gilt

$$P\big(X_{n+1}(\theta) = k+1 \mid X_n(\theta) = k\big) \;=\; 2^{-k}. \tag{2.12}$$

Außerdem sieht man an (2.12), dass $\big(X_n(\theta)\big)_{n\in\mathbb{N}}$ eine homogene Markovkette auf dem Zustandsraum $\mathbb{N}_0$ ist: Für die Verteilung von $X_{n+1}(\theta)$ sind bedingt unter $X_n(\theta)$ die Werte von $X_{n-1}(\theta), X_{n-2}(\theta), \ldots, X_0(\theta)$ unerheblich. Diese Markovkette ist ein sog. reiner Geburtsprozess (es gibt nur Zustandswechsel von k nach $k+1$, $k \in \mathbb{N}_0$), daher kann sie über die Verweilzeiten in den einzelnen Zuständen beschrieben werden. Ein Übergang vom Zustand k in den Zustand $k+1$ findet statt, wenn eine der U-Variablen in ein durch θ bestimmtes Intervall der Länge 2^{-k} fällt. Die Wartezeit Y_{k+1} für dieses Ereignis ist damit geometrisch verteilt mit Erfolgswahrscheinlichkeit 2^{-k}, das heißt, es gilt für alle $k \in \mathbb{N}$

$$P\big(Y_k = j\big) \;=\; \big(1 - 2^{-k+1}\big)^{j-1} \cdot 2^{-k+1}, \qquad j \in \mathbb{N}.$$

Die Verweilzeit im Zustand 0 ist dadurch konstant auf 1 festgelegt.

Diese Markovkette erscheint auch als ein Spezialfall des *approximate counting*. Aus der Motivation, beim Zählen von häufig auftretenden Ereignissen Speicherplatz zu sparen, entwickelte zunächst Morris in [Mor78] die Idee, die explizite Zählung jedes einzelnen Ereignisses zu ersetzen durch einen Zähler,

dessen Wert sich nur mit einer vom Zählerstand abhängenden Wahrscheinlichkeit erhöht. Im Fall einer Wahrscheinlichkeit von 2^{-k} bei Zählerstand $k \in \mathbb{N}$ erhalten wir die obige Markovkette. Verallgemeinert wurde dieses Prinzip unter anderem von Flajolet, Kirschenhofer, Louchard und Prodinger in [Fla85], [KP91], [Pro94] und [LP08]. Das Hauptinteresse gilt dort dem Erwartungswert und der Varianz des (zufälligen) Zählerstands nach einer großen Anzahl von Ereignissen. Mit überwiegend analytischen Methoden werden asymptotische Resultate nachgewiesen, die von sehr kleinen periodischen Funktionen Gebrauch machen.

Nun stellen wir den Zusammenhang zu der im vorigen Abschnitt untersuchten Erneuerungstheorie her. Interpretiert man die Verweilzeiten als Lebensdauern, so ist $\big(X_n(\theta)\big)_{n\in\mathbb{N}}$ nichts anderes als der (nur zu diskreten Zeitpunkten beobachtete) Erneuerungsprozess zu den (ebenfalls diskreten) geometrisch verteilten Lebensdauern. Wir überprüfen zunächst die Bedingungen (2.2) und (2.3): Mit $\alpha = 2$ gilt für $k \to \infty$

$$\begin{aligned} P\big(\alpha^{-k} Y_k \geq x\big) &= P\big(Y_k \geq \lfloor \alpha^k x \rfloor\big) \\ &= \left(1 - \frac{2x}{2^k x}\right)^{\lfloor 2^k x \rfloor - 1} \to \mathrm{e}^{-2x} = P\big(Y_\infty \geq x\big) \end{aligned}$$

mit einer mit Parameter 2 exponentialverteilten (und damit trivialerweise stetigen) Zufallsvariable Y_∞. Wegen $\mathbb{E}Y_k = 2^{k-1}$ ist die Bedingung an die Erwartungswerte trivialerweise erfüllt, nach Satz 2.5 gilt also für alle gegen ∞ konvergierenden Folgen $(n_m)_{m\in\mathbb{N}}$ von natürlichen Zahlen mit $\{\log_2 n_m\} \to \eta$ für $m \to \infty$, dass

$$X_{n_m}(\theta) - \lfloor \log_2 n_m \rfloor \xrightarrow{\text{distr}} Q_\eta.$$

Dabei ist wie oben Q_η für $\eta \in [0,1]$ die Verteilung von $\lfloor -\log_2 S_\infty + \eta \rfloor$ und

$$S_\infty =_{\text{distr}} \sum_{k=0}^{\infty} 2^{-k} Y_{\infty,k} \tag{2.13}$$

mit unabhängigen Exp(2)-verteilten Zufallsvariablen $Y_{\infty,k}$, $k \in \mathbb{N}_0$.

Zusätzlich zu der aus Abschnitt 2.1 gefolgerten Verteilungsasymptotik wollen wir eine Oberschranke für den Totalvariationsabstand der beteiligten Wahrscheinlichkeitsmaße angeben. Um dies zu erreichen betrachten wir zunächst die Kolmogorov-Smirnov-Distanz von $2^{-n}S_n$ und S_∞. Dazu zunächst ein Lemma, welches die Handhabung von Kolmogorov-Smirnov-Abständen erleichtert. $\|\cdot\|_\infty$ bezeichne die Supremumsnorm.

Lemma 2.8 (a) Seien X, Y Zufallsvariablen und f_X eine Dichte zu X. Dann gilt

$$d_{\mathrm{KS}}(X, X+Y) \;\leq\; c\|f_X\|_\infty + P(|Y| > c) \quad \forall c > 0.$$

(b) Sind zusätzlich X und Y unabhängig, so ist

$$d_{\mathrm{KS}}(X, X+Y) \;\leq\; \|f_X\|_\infty \mathbb{E}|Y|.$$

(c) Sind f und g Dichten, so folgt

$$\|f \star g\|_\infty \;\leq\; \|f\|_\infty.$$

Beweis. Für eine Zufallsvariable Z mit Dichte f gilt für alle $c \in \mathbb{R}$ und $\varepsilon > 0$

$$P\big(Z \in [c, c+\varepsilon)\big) \;\leq\; \varepsilon\|f\|_\infty. \tag{2.14}$$

Sei nun $c > 0$. Dann ist

$$P(X \leq z - c) \;\leq\; P(X + Y \leq z) + P(|Y| > c)$$

und

$$P(X + Y \leq z) \;\leq\; P(X \leq z + c) + P(|Y| > c).$$

Also gilt

$$P(X \leq z - c) - P(|Y| > c) \leq P(X + Y \leq z) \leq P(X \leq z + c) + P(|Y| > c),$$

und somit folgt

$$\begin{aligned}
&\big|P(X+Y\le z)-P(X\le z)\big| \\
&\le \max\big\{P(X\le z+c)+P(|Y|>c)-P(X\le z), \\
&\qquad\qquad P(X\le z)-P(X\le z-c)+P(|Y|>c)\big\} \\
&= \max\big\{P\big(X\in(z,z+c]\big), P\big(X\in(z-c,z]\big)\big\}+P(|Y|>c).
\end{aligned}$$

Beide Terme im Maximum werden nach (2.14) durch $c\|f_X\|_\infty$ nach oben begrenzt, damit folgt Teil (a). Sind nun X und Y unabhängig, so gilt

$$\begin{aligned}
&\big|P(X\le x)-P(X+Y\le x)\big| \\
&= \big|P(X\le x, Y\ge 0)+P(X\le x, Y<0) \\
&\quad -P(X+Y\le x, Y\ge 0)-P(X+Y\le x, Y<0)\big| \\
&= P(X\le x<X+Y, Y\ge 0)+P(X+Y\le x<X, Y<0) \\
&= \int_{Y\ge 0} P(X\le x<X+y)P^Y(dy)+\int_{Y<0} P(X+y\le x<X)P^Y(dy) \\
&\le \int_{Y\ge 0} |y|\,\|f_X\|_\infty P^Y(dy)+\int_{Y<0}|y|\,\|f_X\|_\infty P^Y(dy) \\
&= \|f_X\|_\infty \mathbb{E}|Y|.
\end{aligned}$$

Damit ist Teil (b) gezeigt. Sind nun f und g Dichten, so gilt für alle $x\in\mathbb{R}$

$$\begin{aligned}
(f\star g)(x) &= \int_{-\infty}^{\infty} f(t)g(x-t)\,dt \\
&\le \int_{-\infty}^{\infty} \|f\|_\infty g(x-t)\,dt = \|f\|_\infty.
\end{aligned}$$

□

Eine hilfreiche Erkenntnis beim Beweis des folgenden Satzes ist, dass sich der natürliche Logarithmus beschränken lässt durch

$$1-\frac{1}{x} \le \log x \le x-1, \quad x>0.$$

Die zweite Ungleichheit ist dabei sofort klar. Für die erste setzen wir zunächst $x=1$ ein und erhalten in diesem Fall Gleichheit; beim Vergleich der Ableitungen stellt man fest, dass $1-1/x$ für $x<1$ schneller wächst als der Logarithmus, für $x>1$ ist es umgekehrt.

Setzen wir nun $1-x$ in diese Ungleichungskette ein und beschränken den Gültigkeitsbereich auf das offene Einheitsintervall, so ergibt sich

$$-\frac{x}{1-x} \;\leq\; \log(1-x) \;\leq\; -x, \quad x \in (0,1).$$

Alle beteiligten Ausdrücke sind negativ, nach Multiplikation mit -1 erhalten wir durch Kehrwertbildung

$$\frac{1}{x} - 1 \;\leq \frac{1}{-\log(1-x)} \;\leq\; \frac{1}{x}, \quad x \in (0,1). \tag{2.15}$$

Satz 2.9 Sei $\alpha_1 := 1$ und $\alpha_n := (-\log(1-2^{-n+1}))^{-1}$ für $n = 2,3,\ldots$. Sei weiter $\zeta_n := 2^n \alpha_n^{-1}$, sowie $\gamma \geq 1$ die eindeutige Lösung von $\gamma e^{1-\gamma} = 1/2$. Dann gilt für alle $n \in \mathbb{N}$

$$d_{\mathrm{KS}}\big(2^{-n} S_n, S_\infty\big) \;\leq\; \big((1+\gamma)\zeta_n n + 2\big)2^{-n}.$$

Beweis. Sei $(Z_k)_{k\in\mathbb{N}}$ eine Folge unabhängiger, jeweils $\mathrm{Exp}(1)$-verteilter Zufallsvariablen. Wir definieren

$$\tilde{Y}_k \;:=\; \lfloor \alpha_k Z_k \rfloor + 1.$$

Dann gilt für $j \in \mathbb{N}$

$$P\big(\tilde{Y}_k \leq j\big) \;=\; P\big(\alpha_k Z_k < j\big) \;=\; P\big(Z_k < -j\log(1-2^{-k+1})\big) \;=\; 1-(1-2^{-k+1})^j,$$

also ist $\tilde{Y}_k \sim \mathrm{Geom}(2^{-k+1})$ und $2^{-n}\sum_{k=1}^n \tilde{Y}_k$ hat dieselbe Verteilung wie $2^{-n}S_n$. Nach Definition (2.4) von S_∞ und aus der Tatsache $2^{-1}Z_k \sim \mathrm{Exp}(2) = \mathcal{L}(Y_\infty)$ folgt weiter, dass $S_\infty =_{\mathrm{distr}} \sum_{k=1}^\infty 2^{-k} Z_k$. Schließlich stellen wir noch fest, dass durch Umkehren der Summationsreihenfolge folgt

$$2^{-n}\sum_{k=1}^n 2^{k-1} Z_k \;=\; 2^{-n}\sum_{k=1}^n 2^{n-k} Z_{n-k+1} \;=_{\mathrm{distr}}\; \sum_{k=1}^n 2^{-k} Z_k.$$

Die folgenden drei Hilfsaussagen beenden nun den Beweis:

$$\phi_1(n) := d_{\mathrm{KS}}\left(2^{-n}\sum_{k=1}^{n}\tilde{Y}_k, 2^{-n}\sum_{k=1}^{n}\alpha_k Z_k\right) \leq n\alpha_n^{-1}$$

$$\phi_2(n) := d_{\mathrm{KS}}\left(2^{-n}\sum_{k=1}^{n}\alpha_k Z_k, 2^{-n}\sum_{k=1}^{n}2^{k-1} Z_k\right) \leq n\alpha_n^{-1}\gamma + 2^{-n},$$

$$\phi_3(n) := d_{\mathrm{KS}}\left(\sum_{k=1}^{n}2^{-k} Z_k, \sum_{k=1}^{\infty}2^{-k} Z_k\right) \leq 2^{-n};$$

denn es gilt

$$\begin{aligned} d_{\mathrm{KS}}\left(2^{-n}S_n, S_\infty\right) &= d_{\mathrm{KS}}\left(2^{-n}\sum_{k=1}^{n}\tilde{Y}_k, \sum_{k=1}^{\infty}2^{-k} Z_k\right) \\ &\leq \phi_1(n) + \phi_2(n) + \phi_3(n) \leq \left((1+\gamma)\zeta_n n + 2\right)2^{-n}. \end{aligned}$$

Aufgrund der Definition von $\tilde{Y}_k$ gilt offensichtlich

$$V_n \;:=\; 2^{-n}\sum_{k=1}^{n}\alpha_k Z_k \;\leq\; 2^{-n}\sum_{k=1}^{n}\tilde{Y}_k \;\leq\; V_n + n2^{-n},$$

und das bedeutet für die beteiligten Verteilungsfunktionen eine punktweise Ungleichungskette in umgekehrter Richtung. Damit lässt sich der maximale Abstand der Verteilungsfunktionen von $2^{-n}\sum_{k=1}^{n}\tilde{Y}_k$ und V_n abschätzen durch

$$\phi_1(n) \leq d_{\mathrm{KS}}\left(V_n, V_n + n2^{-n}\right).$$

V_n ist eine Summe von unabhängigen Zufallsvariablen; sind also f_{V_n}, f_Z und g_n Dichten zu V_n, Z_1 und $2^{-n}\alpha_n Z_n$, so folgt aus $g_n(x) = 2^n\alpha_n^{-1} f_Z(2^n\alpha_n^{-1}x)$, dass

$$\|g_n\|_\infty = \frac{2^n}{\alpha_n}\|f_Z\|_\infty = \zeta_n.$$

Lemma 2.8.(c) impliziert $\|f_{V_n}\|_\infty \leq \|g\|_\infty = \zeta_n$ und Lemma 2.8.(b) liefert

$$\phi_1(n) \;\leq\; \zeta_n n 2^{-n}.$$

Aus (2.15) erhalten wir für $k \in \mathbb{N}$

$$2^{k-1} - 1 \;\leq\; \alpha_k \leq 2^{k-1},$$

und damit ist

$$0 \;\leq\; 2^{-n} \sum_{k=1}^{n} 2^{k-1} Z_k \; - 2^{-n} \sum_{k=1}^{n} \alpha_k Z_k \;\leq\; 2^{-n} \sum_{k=1}^{n} Z_k.$$

Wir wollen Lemma 2.8.(a) anwenden und dafür zunächst die dort auftretenden Wahrscheinlichkeiten abschätzen.

Die momenterzeugende Funktion zu Z_k ist

$$\mathbb{E}\mathrm{e}^{tZ_k} = \int_0^\infty \mathrm{e}^{tx}\mathrm{e}^{-x}\, dx = \frac{1}{1-t}, \quad 0 \leq t < 1,$$

damit gilt nach der Markovschen Ungleichung

$$P\Big(\sum_{k=1}^{n} Z_k \geq n\gamma\Big) \;\leq\; \mathrm{e}^{-nt\gamma}\Big(\frac{1}{1-t}\Big)^n$$

für alle $0 \leq t < 1$. Mit $t = 1 - 1/\gamma$ (wegen $\gamma \geq 1$ ist dies innerhalb des für t zulässigen Bereiches) wird die Oberschranke zu

$$\Big(\frac{\mathrm{e}^{-t\gamma}}{1-t}\Big)^n \;=\; \big(\gamma \mathrm{e}^{-\gamma+1}\big)^n \;=\; 2^{-n},$$

und somit ergibt sich mit demselben Ordnungsargument wie oben und $c = n\gamma$ in Lemma 2.8.(a)

$$\begin{aligned}
\phi_2(n) \;&=\; d_{\mathrm{KS}}\Big(\sum_{k=1}^{n} \alpha_k Z_k, \sum_{k=1}^{n} 2^{k-1} Z_k\Big) \\
&\leq\; d_{\mathrm{KS}}\Big(\sum_{k=1}^{n} \alpha_k Z_k, \sum_{k=1}^{n} \alpha_k Z_k + \sum_{k=1}^{n} Z_k\Big) \\
&\leq\; n\gamma \|f_{\sum \alpha_k Z_k}\|_\infty + 2^{-n} \\
&\leq\; n\gamma \alpha_n^{-1} + 2^{-n};
\end{aligned}$$

denn nach Lemma 2.8.(c) ist die Supremumsnorm der Dichte von $\sum_{k=1}^{n} \alpha_k Z_k$ kleiner gleich der der Dichte von $\alpha_n Z_n$. Dies ist exponentialverteilt mit Parameter α_n^{-1}, und dieser Parameter gibt auch das globale Maximum der zugehörigen Dichte an.

Die Abschätzung von ϕ_3 geschieht schließlich mittels Lemma 2.8.(b): Da die Zufallsvariablen Z_k, $k = 1,2,\ldots$ unabhängig sind, gilt

$$\begin{aligned}\phi_3(n) &= d_{\mathrm{KS}}\Big(\sum_{k=1}^{n} 2^{-k} Z_k, \sum_{k=1}^{n} 2^{-k} Z_k + \sum_{k=n+1}^{\infty} 2^{-k} Z_k\Big) \\ &\leq \|f_{W_n}\|_\infty \mathbb{E}\Big|\sum_{k=n+1}^{\infty} 2^{-k} Z_k\Big|,\end{aligned}$$

wobei f_{W_n} eine Dichte zu $W_n := \sum_{k=1}^{n} 2^{-k+1} Z_k$ ist. Es gilt dabei $\|f_{W_n}\|_\infty \leq \|f_Z\|_\infty = 1$ und

$$\mathbb{E}\Big|\sum_{k=n+1}^{\infty} 2^{-k} Z_k\Big| = \sum_{k=n+1}^{\infty} 2^{-k}\mathbb{E}Z_k = 2^{-n},$$

also $\phi_3(n) \leq 2^{-n}$. □

Bemerkung 2.10 Mittels eines geeigneten numerischen Verfahrens erhält man $\gamma \approx 2.678347$. Die verwendete Oberschranke in Satz 2.9 ist damit für $n \leq 5$ trivial, da sich ein Wert größer als 1 ergibt. Die Beschaffenheit der Folge (ζ_n) ermöglicht so die Abschätzung $(1+\gamma)\zeta_n \leq 15/2$ für $n \geq 6$ und wir erhalten als alternative Formulierung des obigen Satzes

$$d_{\mathrm{KS}}\big(2^{-n} S_n, S_\infty\big) \leq \big(\tfrac{15}{2} n + 2\big) 2^{-n}.$$

Weiterhin ist $15/2n + 2 \leq 8$ für $n \geq 4$. Damit können wir die Aussage des Satzes vereinfachen zu

$$d_{\mathrm{KS}}\big(2^{-n} S_n, S_\infty\big) \leq n 2^{-n+3}.$$

Mit den Abschätzungsmöglichkeiten für den Kolmogorov-Smirnov-Abstand von $2^{-n} S_n$ und S_∞ aus Bemerkung 2.10 nehmen wir nun eine ähnliche Aussage für den Totalvariationsabstand von $\mathcal{L}(X_n(\theta) - \lfloor \log_2 n \rfloor)$ und $Q_{\{\log_2 n\}}$

in Angriff. Für bessere Lesbarkeit schreiben wir wieder $k_n := \lfloor \log_2 n \rfloor$ und $\eta_n := \{\log_2 n\}$. Wieder einmal benötigen wir ein kleines Lemma aus der Analysis:

Lemma 2.11 (a) Für $k \in \mathbb{Z}$ ist

$$\sum_{j=k}^{\infty} j2^{-j} \;=\; (1+k)2^{1-k}.$$

(b) Für $x \in \mathbb{R}_+$ gilt

$$\sum_{j \geq x} j2^{-j} \;\leq\; (1+x)2^{1-x}.$$

(c) Für $|t| < 2$ gilt

$$\prod_{k=1}^{\infty} \frac{1}{1-2^{-k}t} \;\leq\; \Big(\frac{1}{1-t/2}\Big)^2.$$

Beweis. Wir zeigen zunächst Teil (a) mit vollständiger „zweiseitiger“ Induktion: Für $k = 0$ betrachten wir eine geometrisch verteilte Zufallsvariable mit Parameter $1/2$, deren Erwartungswert als die linke Summe geschrieben werden kann. Dass selbiger gleich 2 ist, verifiziert dann den Induktionsanfang.

Weiter gilt

$$\begin{aligned} & \sum_{j=k+1}^{\infty} j2^{-j} \;=\; (2+k)2^{-k} \\ \iff \quad & \sum_{j=k}^{\infty} j2^{-j} \;=\; (2+k)2^{-k} + k2^{-k} \\ & \qquad\qquad\quad = (1+k)2^{1-k}, \end{aligned}$$

und diese Aussage lässt sich für $k \leq 0$ als Induktionsschluss von k auf $k-1$ verwenden wie auch, für $k \geq 0$, als Schluss von k auf $k+1$.

Für Teil (b) definieren wir die Funktion $g : \mathbb{R} \to \mathbb{R}$ durch $g(x) := (1+x)2^{1-x}$. Betrachtet man die Ableitung, sieht man leicht, dass g links von $\log_2(\mathrm{e}/2)$

streng monoton wachsend, rechts davon streng monoton fallend ist. Für $x \geq \log_2(\mathrm{e}/2) \approx 0{,}442$ können wir also Teil (a) mit $k = \lceil x \rceil$ verwenden, und erhalten die Oberschranke durch ein Monotonieargument. Weiter ist $g(0) = g(1) = 2$ und $g(x) > 2$ für $x \in (0,1)$, also folgt die Behauptung auch für diesen Fall.

Zum letzten Teil des Lemmas: Bekannterweise lässt sich der Logarithmus in der Potenzreihenentwicklung um 1 schreiben als

$$\log(1-x) \;=\; -\sum_{j=1}^{\infty} \frac{x^j}{j}, \qquad |x| < 1.$$

Damit ergibt sich

$$\begin{aligned}\prod_{k=1}^{\infty} \frac{1}{1-2^{-k}t} \;&=\; \exp\Bigl(-\sum_{k=1}^{\infty} \log\bigl(1-2^{-k}t\bigr)\Bigr) \\ &=\; \exp\Bigl(\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \frac{(2^{-k}t)^j}{j}\Bigr) \;=\; \exp\Bigl(\sum_{j=1}^{\infty} \frac{t^j}{j} \sum_{k=1}^{\infty} 2^{-kj}\Bigr).\end{aligned}$$

Nun ist der Wert der inneren geometrische Reihe gerade $1/(2^j-1)$ und damit kleiner oder gleich 2^{-j+1}. Damit schätzen wir insgesamt nach oben ab:

$$\begin{aligned}\prod_{k=1}^{\infty} \frac{1}{1-2^{-k}t} \;&\leq\; \exp\Bigl(2\sum_{j=1}^{\infty} \frac{(t/2)^j}{j}\Bigr) \\ &=\; \exp\bigl(-2\log(1-t/2)\bigr) \;=\; \Bigl(\frac{1}{1-t/2}\Bigr)^2. \qquad \square\end{aligned}$$

Satz 2.12 Für $n \in \mathbb{N}$ und $\varepsilon \in (0,1)$ gilt

$$d_{\mathrm{TV}}\bigl(\mathcal{L}(X_n - k_n), Q_{\eta_n}\bigr) \;\leq\; \bigl(53(1-\varepsilon)\log_2 n + 14\bigr)\Bigl(\frac{n}{2}\Bigr)^{-(1-\varepsilon)} \;+\; 6\exp\bigl(-n^{\varepsilon}\bigr).$$

Beweis. Seien $\varepsilon \in (0,1)$ und $n \in \mathbb{N}$ fest. Wir wollen

$$d_{\mathrm{TV}}\bigl(\mathcal{L}(X_n - k_n), Q_{\eta_n}\bigr) \;=\; \sum_{j\in\mathbb{Z}} \bigl|P(X_n - k_n = j) - Q_{\eta_n}(\{j\})\bigr|$$

nach oben begrenzen. Dafür betrachten wir zunächst die beteiligten Ausdrücke einzeln: Es gilt nach (2.1) für $j \geq -k_n$

$$\begin{aligned} P(X_n - k_n = j) &= P(X_n = j + k_n) \\ &= P(S_{k_n+j} \leq n) \; - \; P(S_{k_n+j+1} \leq n), \end{aligned}$$

sowie

$$\begin{aligned} Q_{\eta_n}(\{j\}) &= P(\lfloor -\log_2 S_\infty + \eta_n \rfloor = j) \\ &= P(-\log_2 S_\infty \geq j - \eta_n) \; - \; P(-\log_2 S_\infty \geq j + 1 - \eta_n) \\ &= P(2^{k_n+j} S_\infty \leq n) \; - \; P(2^{k_n+j+1} S_\infty \leq n). \end{aligned}$$

Da die Verteilungen von S_∞ und $2^{-n} S_n$ denselben Kolmogorov-Smirnov-Abstand haben wie die von $2^{-n+k_n+j} S_n$ und $2^{k_n+j} S_\infty$ (und analog mit $j+1$ anstelle von j), lässt sich also abschätzen

$$\begin{aligned} &\big|P(X_n - k_n = j) \; - \; Q_{\eta_n}(\{j\})\big| \\ &\qquad \leq \; \big|P(S_{k_n+j} \leq n) \; - \; P(2^{k_n+j} S_\infty \leq n)\big| \\ &\qquad\quad + \big|P(S_{k_n+j+1} \leq n) \; - \; P(2^{k_n+j+1} S_\infty \leq n)\big| \\ &\qquad \leq \; d_{\mathrm{KS}}\big(2^{-k_n-j} S_{k_n+j}, S_\infty\big) \; + \; d_{\mathrm{KS}}\big(2^{-k_n-j-1} S_{k_n+j+1}, S_\infty\big). \end{aligned}$$

Nun spalten wir die oben angeführte Summe auf. Für diejenigen Summanden, deren Index nicht kleiner als $-\varepsilon k_n$ ist, erhalten wir

$$\begin{aligned} &\sum_{j \geq -\varepsilon k_n} \big|P(X_n - k_n = j) - Q_{\eta_n}(\{j\})\big| \\ &\qquad \leq \sum_{j \geq -\varepsilon k_n} \Big(d_{\mathrm{KS}}\big(2^{-k_n-j} S_{k_n+j}, S_\infty\big) \; + \; d_{\mathrm{KS}}\big(2^{-k_n-j-1} S_{k_n+j+1}, S_\infty\big)\Big), \end{aligned}$$

Indexverschiebung liefert mit Bemerkung 2.10

$$\begin{aligned} \ldots \; &\leq \sum_{j \geq (1-\varepsilon)k_n} \Big(\frac{15}{2} j + 2\Big) 2^{-j} + \Big(\frac{15}{2}(j+1) + 2\Big) 2^{-j-1} \\ &\leq \; \frac{27}{4} \sum_{j \geq (1-\varepsilon)k_n} 2^{-j} \; + \; \frac{45}{4} \sum_{j \geq (1-\varepsilon)k_n} j 2^{-j} \end{aligned}$$

und mit Lemma 2.11.(b) ergibt dies

$$\begin{aligned} \dots &\le 27 \cdot 2^{-(1-\varepsilon)k_n-1} + \frac{45}{4}\big((1-\varepsilon)k_n+1\big)2^{1-(1-\varepsilon)k_n} \\ &\le \big(45(1-\varepsilon)k_n + 72\big)2^{-(1-\varepsilon)k_n-1}. \end{aligned}$$

Für $j < -\varepsilon k_n$ verwenden wir die Dreiecksungleichung. Dazu schätzen wir zunächst den Tail von S_∞ mit Hilfe der Markovschen Ungleichung ab. Nach (2.13) gilt

$$S_\infty \;=_{\text{distr}}\; \sum_{k=0}^{\infty} \tilde{Y}_k$$

mit $\mathcal{L}(\tilde{Y}_k) = \text{Exp}(2^k)$ und $\tilde{Y}_1, \tilde{Y}_2, \dots$ unabhängig; wegen $\mathcal{L}(2^{-k}Y_{\infty,k}) = \mathcal{L}(\tilde{Y}_k)$ bedarf dies keiner weiteren Begründung. Als direkte Konsequenz daraus ist

$$\mathbb{E}\mathrm{e}^{tS_\infty} \;=\; \prod_{k=1}^{\infty} \frac{1}{1-2^{-k}t}, \quad |t| < 2,$$

die momenterzeugende Funktion zu S_∞, also folgt mit Lemma 2.11.(c)

$$P(S_\infty \ge x) \;\le\; \mathrm{e}^{-tx}\Big(\frac{1}{1-t/2}\Big)^2, \quad |t| < 2. \tag{2.16}$$

Weiter gilt

$$\begin{aligned} &\sum_{j<-\varepsilon k_n} \big|P(X_n - k_n = j) - Q_{\eta_n}(\{j\})\big| \\ &\qquad\le\; P(X_n - k_n < -\varepsilon k_n) \;+\; Q_{\eta_n}\big((-\infty, -\varepsilon k_n)\big) \\ &\qquad\le\; P\big(X_n < (1-\varepsilon)k_n\big) + P\big(\lfloor -\log_2 S_\infty + \eta_n \rfloor < -\varepsilon k_n\big). \end{aligned} \tag{2.17}$$

Für den ersten Ausdruck ziehen wir wieder das erneuerungstheoretische Standardargument (2.1) heran:

$$\begin{aligned} P\big(X_n \le (1-\varepsilon)k_n - 1\big) &= P\big(X_n \le \lfloor (1-\varepsilon)k_n - 1 \rfloor\big) \\ &= P\big(S_{\lfloor (1-\varepsilon)k_n - 1\rfloor} \ge n\big) \\ &= P\big(2^{-\lfloor (1-\varepsilon)k_n - 1\rfloor} S_{\lfloor (1-\varepsilon)k_n - 1\rfloor} \ge n 2^{-\lfloor (1-\varepsilon)k_n - 1\rfloor}\big), \end{aligned}$$

dies wiederum ist nicht weiter von $P(S_\infty \geq 2^{-\lfloor(1-\varepsilon)k_n-1\rfloor})$ weg als die entsprechende Kolmogorov-Smirnov-Distanz aus Bemerkung 2.10. Somit gilt

$$\begin{aligned} P\big(X_n \leq (1-\varepsilon)k_n - 1\big) \;&\leq \Big(\frac{15}{2}\big((1-\varepsilon)k_n - 1\big) + 2\Big) \cdot 2^{-\lfloor(1-\varepsilon)k_n-1\rfloor} \\ &\quad + \frac{16}{9}\exp\big(-n2^{-(1-\varepsilon)k_n}\big), \end{aligned}$$

wobei wir (2.16) mit $t = 1/2$ verwendet haben.

Für den zweiten Summanden in (2.17) verwenden wir wieder Abschätzung (2.16), diesmal mit $t = 1$:

$$\begin{aligned} P\big(\lfloor -\log_2 S_\infty + \eta_n\rfloor \leq -\varepsilon k_n - 1\big) \;&\leq\; P\big(-\log_2 S_\infty \leq (1-\varepsilon)k_n - \log_2 n\big) \\ &\leq\; P\big(S_\infty \geq n2^{-(1-\varepsilon)k_n}\big) \\ &\leq\; 4\exp\big(-n2^{-(1-\varepsilon)k_n}\big). \end{aligned}$$

Wir fassen zusammen:

$$\begin{aligned} d_{\mathrm{TV}}\big(\mathcal{L}(X_n - k_n), Q_{\eta_n}\big) \;\leq\; &\big(45(1-\varepsilon)k_n + 72\big)2^{-(1-\varepsilon)k_n-1} \\ &+ \Big(\frac{15}{2}\big((1-\varepsilon)k_n - 1\big) + 2\Big) \cdot 2^{-\lfloor(1-\varepsilon)k_n-1\rfloor} \\ &+ \frac{16}{9}\exp\big(-n2^{-(1-\varepsilon)k_n}\big) \\ &+ 4\exp\big(-n2^{-(1-\varepsilon)k_n}\big). \end{aligned} \tag{2.18}$$

Nun ist

$$n\,2^{-(1-\varepsilon)k_n} \;=\; n\,2^{-(1-\varepsilon)\log_2 n + (1-\varepsilon)\{\log_2 n\}} \;=\; n^\varepsilon \cdot 2^{(1-\varepsilon)\{\log_2 n\}},$$

und da der Exponent der Zweierpotenz nicht negativ wird, kann man dies nach unten durch n^ε abschätzen. Damit erhalten wir für die letzten beiden Summanden in (2.18) insgesamt die Oberschranke

$$6\exp\big(-n^\varepsilon\big),$$

grob abgeschätzt, denn $16/9 + 4 \leq 6$. Für die übrigen Summanden stellen wir fest, dass wegen $\{\log_2 n\} \leq 1$ gilt

$$2^{-(1-\varepsilon)k_n} \;=\; n^{-(1-\varepsilon)} \cdot 2^{(1-\varepsilon)\{\log_2 n\}} \;\leq\; \Big(\frac{n}{2}\Big)^{-(1-\varepsilon)},$$

und wir können die ersten beiden Summanden in (2.18) nach oben begrenzen durch

$$\Big(53(1-\varepsilon)k_n \;+\; 14\Big)\cdot\Big(\frac{n}{2}\Big)^{-(1-\varepsilon)}.$$

Die Behauptung folgt. □

Bemerkung 2.13 Die Oberschranke des vorhergegangenen Satzes erscheint auf den ersten Blick etwas unübersichtlich. Für große n ist es jedoch für beliebiges festes ε immer der erste Summand, der langsamer gegen Null strebt. Aus Satz 2.12 kann man also die folgende Aussage erhalten:

Zu $\gamma < 1$ existieren ein $C > 0$ und ein $n_0 \in \mathbb{N}$, so dass für alle $n > n_0$ gilt

$$d_{\mathrm{TV}}\big(\mathcal{L}(X_n - k_n), Q_{\eta_n}\big) \;\leq\; C \cdot \log_2 n \cdot n^{-\gamma}.$$

Im Sinne einer Konvergenzgeschwindigkeit „verliert“ außerdem der Logarithmus gegen die Potenz. Damit folgern wir schließlich

$$d_{\mathrm{TV}}\big(\mathcal{L}(X_n - k_n), Q_{\eta_n}\big) \;=\; \mathcal{O}(n^{-\gamma}), \qquad n \to \infty,$$

für alle $\gamma < 1$.

2.3 Darstellung der Grenzverteilung

Für die im vorhergegangenen Kapitel erhaltene Grenzverteilung wollen wir nun mit Hilfe von charakteristischen Funktionen eine Darstellung der Verteilungsfunktion dieser Grenzverteilung erhalten. Hauptansatzpunkt ist dabei eine Partialbruchzerlegung. Ähnliche Darstellungen erhalten auch Louchard in [Lou87] und Flajolet in [Fla85], diese verwenden jedoch einen wesentlich analytischeren Zugang und arbeiten u.a. mit Mellin-Transformationen.

Lemma 2.14 Seien $b := \prod_{j=1}^{\infty}(1-2^{-j})$ und, für $k \in \mathbb{N}$,

$$a_k \;:=\; \left(b\prod_{j=1}^{k-1}\big(1-2^j\big)\right)^{-1}.$$

Dann gilt für alle $x \in \mathbb{Z}$

$$Q_\eta\big((-\infty,x]\big) \;=\; \sum_{k=1}^{\infty} a_k \exp\big(-2^{k+\eta-1-x}\big).$$

Beweis. Sei zunächst $n \in \mathbb{N}$ fest. Wir beginnen mit einer Partialbruchzerlegung von $\prod_{k=1}^{n}(1-2^{-k}z)^{-1}$. Der Ansatz

$$\prod_{k=1}^{n}\big(1-2^{-k}z\big)^{-1} = \sum_{k=1}^{n} a_{n,k}\big(1-2^{-k}z\big)^{-1} \tag{2.19}$$

führt nach geeigneter Multiplikation und Einsetzen von $z = 2^k$ mit $k = 1,\ldots,n$ auf

$$a_{n,k} = \left(\prod_{j=1}^{k-1}\big(1-2^{j}\big)\prod_{j=1}^{n-k}\big(1-2^{-j}\big)\right)^{-1}.$$

Dieses Vorgehen zur Bestimmung der Koeffizienten in der Partialbruchzerlegung nennt man mitunter auch die „Zuhaltemethode“.

Nach (2.7) ist $\mathcal{L}(S_\infty)$ eine Faltung von Exponentialverteilungen. Die charakteristische Funktion zu $\mathrm{Exp}(2^k)$ lautet $z \mapsto (1-2^{-k}z)^{-1}$, also steht auf der linken Seite von (2.19) die charakteristische Funktion zur Faltung $\mathrm{Exp}(2) \star \ldots \star \mathrm{Exp}(2^n)$. Auf der rechten Seite von (2.19) steht eine Linearkombination charakteristischer Funktionen, die zu einer Mischung von Verteilungen gehört. Dies geht hier jedoch über die übliche Mischung von Wahrscheinlichkeitsmaßen hinaus, da die Koeffizienten $a_{n,k}$ alternieren. Bezüglich der Verteilungen lässt sich (2.19) also nach dem Eindeutigkeitssatz für charakteristische Funktionen schreiben als

$$\mathrm{Exp}(2) \star \mathrm{Exp}(2^2) \star \ldots \star \mathrm{Exp}(2^n) = \sum_{k=1}^{n} a_{n,k}\,\mathrm{Exp}(2^k). \tag{2.20}$$

Wir wollen nun auf beiden Seiten den Grenzübergang $n \to \infty$ betrachten, um daraus

$$\mathcal{L}(S_\infty) = \sum_{k=1}^{\infty} a_k\,\mathrm{Exp}(2^k)$$

zu erhalten. Aus (2.20) folgt sofort

$$P(Y_1 + \ldots + Y_n \geq x) \;=\; \sum_{k=1}^{n} a_{n,k} \exp(-x2^k), \qquad x \geq 0,$$

wobei $Y_k \sim \mathrm{Exp}(2^k)$ für $k \in \mathbb{N}$ und $Y_1, \ldots, Y_n$ unabhängig. Dabei reicht uns für $x \geq 0$ die punktweise Bildung des Limes. Mit der Stetigkeit von unten steht nach dem Grenzübergang links $P(S_\infty \geq x)$, rechts brauchen wir zwei zusätzliche Erkenntnisse. Dass gilt

$$\lim_{n\to\infty} a_{n,k} \;=\; a_k,$$

ist dabei leicht zu sehen. Außerdem brauchen wir ein Argument für die Zulässigkeit der Vertauschung von Limes und Summe. Sei $x > 0$. Dazu definieren wir das Maß ν_x auf $\mathbb{N}$ durch

$$\nu_x(\{k\}) \;:=\; \exp(-x2^k).$$

Mit $f_n : \mathbb{N} \to \mathbb{R},\ k \mapsto 1_{\{k\leq n\}} a_{n,k},\ k \in \mathbb{N}$, gilt $\|f_n\|_\infty \leq b^{-1}$. Da $\nu_x(\mathbb{N}) < \infty$ folgt nach dem Satz von der majorisierten Konvergenz mit $f := \lim_{n\to\infty} f_n$

$$\begin{aligned}\lim_{n\to\infty} \sum_{k=1}^{n} a_{n,k} \exp(-x2^k) \;&=\; \lim_{n\to\infty} \int_{\mathbb{N}} 1_{\{1,\ldots,n\}} f_n \; d\nu_x \\ &=\; \int_{\mathbb{N}} f \; d\nu_x \;=\; \sum_{k=1}^{\infty} a_k \exp(-x2^k).\end{aligned}$$

Damit folgt für $x \in \mathbb{Z}$

$$\begin{aligned} Q_\eta\big((-\infty, x]\big) &= P\big(\lfloor -\log_2 S_\infty + \eta \rfloor \leq x\big) \\ &= P\big(-\log_2 S_\infty < x + 1 - \eta\big) \\ &= P\big(S_\infty > 2^{-x-1+\eta}\big) \qquad (2.21) \\ &= \sum_{k=1}^{\infty} a_k P\big(\tilde{Y}_k > 2^{-x-1+\eta}\big) \\ &= \sum_{k=1}^{\infty} a_k \exp\big(-2^{-x-1+\eta+k}\big). \end{aligned}$$

□

Diese Darstellung können wir verwenden, um die Zähldichte dieses Grenzmaßes zu skizzieren, siehe Abbildung 2.2.

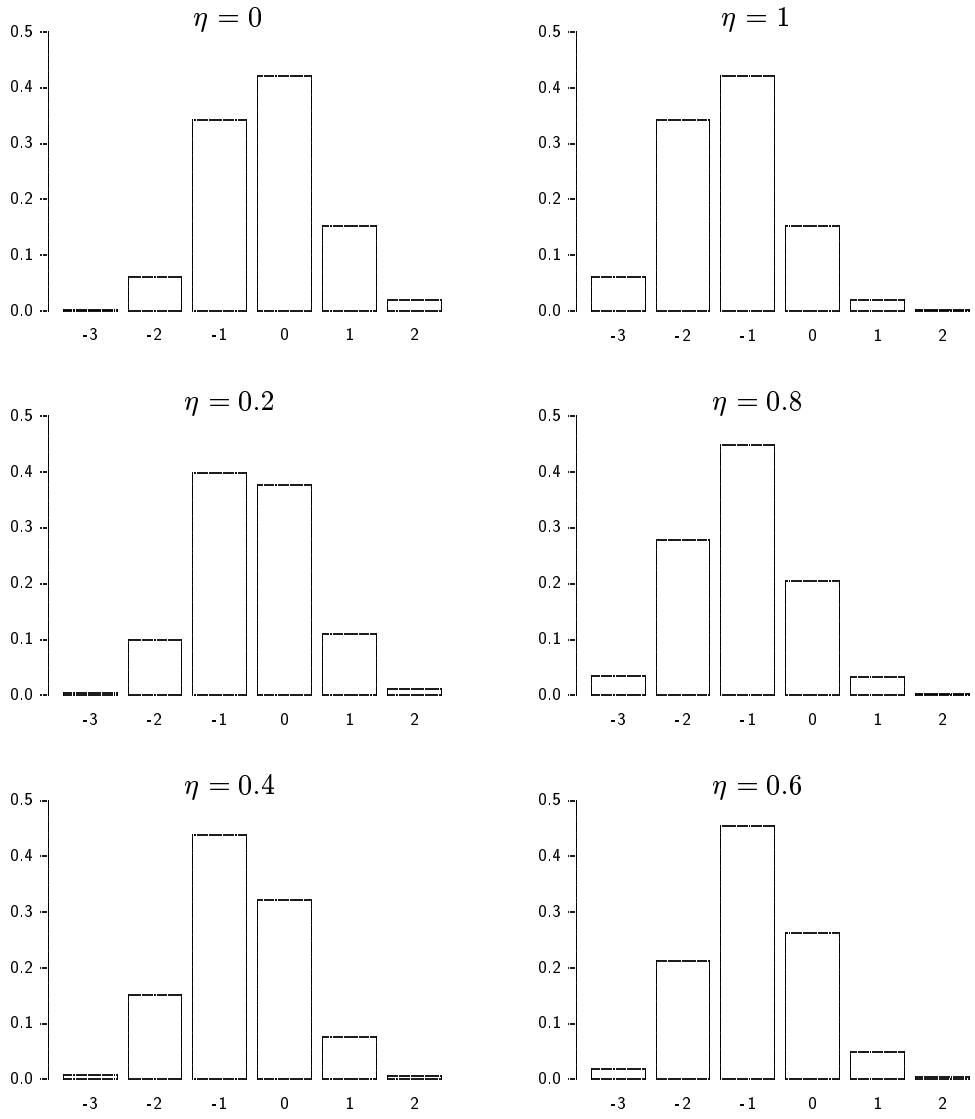

Abbildung 2.2: Massenfunktionen $k \mapsto Q_\eta(\{k\})$ für verschiedene Werte von η.

3 Statistische Konzepte für binäre Bäume

In digitalen Suchbäumen finden sich schnell statistische Fragestellungen und Anwendungen. Eine naheliegende Fragestellung ist zum Beispiel folgende: Angenommen, man beobachtet die Tiefe eines einzelnen Pfades, wie sollte man die Zahl der Knoten im gesamten Baum schätzen? Auf die Problematik des *approximate counting* (vgl. S. 21) übertragen ist dies sozusagen der stochastische Rückschluss von einem Zählerstand auf die Anzahl der gezählten Werte.

Eine weitere Fragestellung aus dem Bereich der Statistik ist dann diese: Ist ein binärer Baum gegeben, kann man anhand seiner Struktur erkennen, ob dieser durch den BST- oder den DST-Algorithmus entstanden ist? Gibt es vielleicht sogar eine einzelne Kenngröße, die die relevante Information für diese Unterscheidung bereits vollständig enthält?

Bei DST-erzeugten binären Bäumen untersuchen wir zunächst die Möglichkeit, das Likelihood-Prinzip auf das Problem des Schätzens der Knotenzahl n anzuwenden. Beobachtet wird dafür die Länge eines vorgegebenen Pfades, deren Verteilung wir im vorhergegangenen Kapitel untersucht haben. Der Schätzer für den Parameter (!) n, der sich daraus ergibt, wird unter asymptotischen Gesichtspunkten wie auch im Vergleich zu anderen Schätzern betrachtet. Im darauffolgenden Abschnitt geben wir Konfidenzschranken für die unbekannte Anzahl der Knoten an. Hier können wir die asymptotischen Schlussfolgerungen in Kapitel 2 verwenden und approximative Konfidenzschranken angeben. Schließlich wenden wir uns dem oben beschriebenen Entscheidungsproblem zu.

Für binäre Bäume, die aus dem BST-Algorithmus hervorgehen, sind die entsprechenden Resultate zum Teil bereits bekannt. Daher beschränken wir uns für den BST-Fall auf einen kurzen Abschnitt zu approximativen Konfidenzintervallen.

3.1 Schätzen der Knotenzahl

In einem zufälligen digitalen Suchbaum mit unbekannter Anzahl n von Knoten wird entlang eines Pfades $\theta \in \{0,1\}^{\mathbb{N}}$ die Tiefe $K := X_n(\theta)$ des externen Knotens beobachtet. Die Knotenzahl soll anhand dieser Beobachtung geschätzt werden.

Aus dem vorigen Kapitel übernehmen wir die Darstellung von $X_n(\theta)$ als Markovkette. Es gilt $P(X_{n+1}(\theta) = k+1 | X_n(\theta) = k) = 2^{-k}$ und daraus folgt

$$\begin{aligned} P\big(X_n(\theta) = k\big) &= 2^k \; P\big(X_{n+1}(\theta) = k+1, X_n(\theta) = k\big) \\ &= 2^k \; P\big(X_n(\theta) < k+1 \le X_{n+1}(\theta)\big) \\ &= 2^k \; \big[P\big(X_{n+1}(\theta) \ge k+1\big) \; - \; P\big(X_n(\theta) \ge k+1\big)\big], \end{aligned}$$

denn (X_n) hat für festes θ nur Sprünge der Höhe 1. Wir benutzen das erneuerungstheoretische Standardargument (2.1), S. 9, und erhalten

$$\begin{aligned} P\big(X_n(\theta) = k\big) &= 2^k \; \big[P(S_{k+1} \le n+1) \; - \; P(S_{k+1} \le n)\big] \\ &= 2^k \; P(S_{k+1} = n+1). \end{aligned}$$

Dabei ist (S_k) die Folge der Partialsummen zu (Y_k), mit $Y_k \sim \mathrm{Geom}(2^{1-k})$ für $k \in \mathbb{N}$ und (Y_k) unabhängig. Insbesondere erkennt man an dieser Darstellung, dass die Verteilung von $X_n(\theta)$ nicht von der Wahl von θ abhängt.

Das Problem der Maximierung von $n \mapsto P(X_n(\theta) = k)$ für festes k ist also äquivalent dazu, das Maximum der Wahrscheinlichkeitsmassenfunktion von S_{k+1} zu finden.

Lemma 3.1 Für jedes $k \in \mathbb{N}$ besitzt die Wahrscheinlichkeitsmassenfunktion von S_k ein eindeutiges Maximum.

Beweis. Der Beweis gliedert sich in zwei Teile. Wir nutzen zunächst das Konzept der diskreten Unimodalität, wie es von Keilson und Gerber in [KG71] entwickelt wird, um zu zeigen, dass für die Folge (p_n) mit

$$p_n := P(S_k = n), \quad n \in \mathbb{N},$$

(mindestens) ein $M \in \mathbb{N}$ existiert, so dass

$$\begin{aligned} p_n &\geq p_{n-1} \quad \text{für } n \leq M, \\ p_{n+1} &\leq p_n \quad \text{für } n \geq M. \end{aligned}$$

In der Notation von [KG71] heißt eine solche Verteilung *(diskret) unimodal*. Darüberhinaus sind solche Verteilungen *streng unimodal*, deren Faltung mit einer beliebigen unimodalen Verteilung wieder unimodal ist. Zwei Resultate aus [KG71] werden nun verwendet:

- Die Klasse der streng unimodalen Verteilungen ist unter der Faltung abgeschlossen.
- Geometrische Verteilungen sind streng unimodal.

Es ist klar, dass hieraus sofort die strenge Unimodalität von S_k gefolgert werden kann.

Nun zum zweiten Teil des Beweises. Wir zeigen, dass es kein $n \geq k$ gibt mit

$$p_n = p_{n+1}.$$

Zerlegen wir nach den Werten der $Y_1, \ldots, Y_k$, so erhalten wir

$$\begin{aligned} p_n &= \sum_{\substack{1 \leq j_1, \ldots, j_k \\ j_1 + \ldots + j_k = n}} P\big(Y_1 = j_1, \ldots, Y_k = j_k\big) \\ &= \sum_{\substack{1 \leq j_2, \ldots, j_k \\ j_2 + \ldots + j_k = n-1}} \prod_{i=2}^{k} 2^{1-i}(1 - 2^{1-i})^{j_i - 1} \\ &= 2^{-k(k-1)/2} \sum_{\substack{1 \leq j_2, \ldots, j_k \\ j_2 + \ldots + j_k = n-1}} 2^{-\sum_{i=2}^{k}(i-1)(j_i-1)} \cdot \prod_{i=2}^{k} (2^{i-1} - 1)^{j_i - 1}. \end{aligned}$$

Die „feinste“ Zweierpotenz in dieser Summe wird dabei für $1 = j_2 = \ldots = j_{k-1}$ und $j_k = n - k + 1$ erreicht, die verbleibende Potenz einer ungeraden Zahl ist wieder ungerade. In der Summe kommen nun nur „gröbere“ Zweierpotenzen hinzu, also ist p_n ein ungerades Vielfaches von

$$2^{-k(k-1)/2-(k-1)(n-k)} = 2^{-(k-1)(n-k/2)} =: N.$$

Analog ist p_{n+1} ein ungerades Vielfaches von

$$2^{-(k-1)(n+1-k/2)} \;=\; 2^{1-k}N$$

und damit können p_n und p_{n+1} nicht gleich sein. □

Die Identifikation dieses eindeutigen Maximums ist jedoch nicht trivial – man rechnet mit einer Maximalstelle in der Größenordnung von 2^K. Es gibt jedoch einen numerisch gut zugänglichen Weg über eine Rekursionsvorschrift für die Wahrscheinlichkeitsmassenfunktion zu $X_n(\theta)$.

Für das Zustandekommen des Ereignisses $\{X_n(\theta) = k\}$ gibt es zwei mögliche vorhergehende Situationen: $\{X_{n-1}(\theta) = k\}$ und $\{X_{n-1}(\theta) = k-1\}$. Nach dem Satz von der totalen Wahrscheinlichkeit ist also

$$\begin{aligned} P(X_n(\theta) = k) &= P(X_n(\theta) = k | X_{n-1}(\theta) = k) \cdot P(X_{n-1}(\theta) = k) \\ &\quad + P(X_n(\theta) = k | X_{n-1}(\theta) = k-1) \cdot P(X_{n-1}(\theta) = k-1) \\ &= (1 - 2^{-k}) P(X_{n-1}(\theta) = k) + 2^{-k+1} P(X_{n-1}(\theta) = k-1), \end{aligned}$$

vergleiche (2.12). Schreiben wir abkürzend $p_{n,k} := P(X_n(\theta) = k)$ so erhalten wir die etwas übersichtlichere Darstellung

$$p_{n,k} \;=\; \big(1 - 2^{-k}\big) p_{n-1,k} + 2^{1-k} p_{n-1,k-1} \tag{3.1}$$

Damit können wir die Maximum-Likelihood-Schätzfunktion $k \mapsto \hat{n}_{\mathsf{ML}}(k)$ berechnen: Folgende `R`-Zeilen erzeugen einen Vektor der ersten `kmax` Werte dieser Funktion.

```
kmax <- 20
nmax <- floor(2^(kmax+.5))

p <- matrix(0,nmax,kmax)
p[1:nmax,1] <- (1/2)^(1:nmax-1)

for (n in 2:nmax)
  p[n,2:kmax] <- (1-2^(-(2:kmax))) * p[n-1,2:kmax]
                   + 2^(1-(2:kmax)) * p[n-1,2:kmax-1]

ml <- which.max(p[,1])
for (k in 2:kmax)
  ml <- c(ml, which.max(p[,k]))

ml
```

Das oben erwähnte exponentielle Wachstum führt jedoch schnell zu einer rechenzeitbedingten Beschränkung in der Wahl von `kmax`. In Tabelle 3.1 sind die Werte des Maximum-Likelihood-Schätzers für einige (kleinere) Werte von k angegeben. Durch die zusätzliche Berechnung des Logarithmus des jeweiligen Schätzwertes kommen wir bereits der Asymptotik auf die Spur.

k	$\hat{n}(k)$	$\log_2 \hat{n}(k)$	k	$\hat{n}(k)$	$\log_2 \hat{n}(k)$	k	$\hat{n}(k)$	$\log_2 \hat{n}(k)$
1	2	1	7	162	7.339850	13	10462	13.35287
2	4	2	8	325	8.344296	14	20925	14.35294
3	8	3	9	652	9.348728	15	41852	15.35301
4	19	4.247928	10	1306	10.35094	16	83706	16.35304
5	39	5.285402	11	2614	11.35204	17	167415	17.35307
6	80	6.321928	12	5230	12.35260	18	334831	18.35307

Tabelle 3.1: Werte des ML-Schätzers für kleine Beobachtungen k

Betrachtet man in Tabelle 3.1 die letzten Werte des logarithmierten Schätzers, so könnte man vermuten, dass $\log_2 \hat{n}(k) - k$ gegen einen festen Wert konvergiert. Wir wollen diese Vermutung mit einigen heuristischen Überlegungen stützen, eine genauere Untersuchung wird Gegenstand von [DG09] sein.

Die Schätzfunktion, deren Werte wir oben numerisch bestimmt haben, lautet

$$\hat{n}_{\mathsf{ML}}(k) \;=\; \operatorname*{argmax}_{n \in \mathbb{N}} P\big(X_n(\theta) = k\big).$$

Damit $\hat{n}_{\mathsf{ML}}(k)$ mit großem k nicht in der Unendlichkeit verschwindet, betrachten wir die normierte Variante

$$\log_2\big(\hat{n}_{\mathsf{ML}}(k)\big) \;-\; k.$$

Dabei gilt wegen der Bijektivität des Logarithmus als Funktion von $\mathbb{R}_{>0}$ nach $\mathbb{R}$

$$\log_2 \operatorname*{argmax}_{n \in \mathbb{N}} P\big(X_n(\theta) = k\big) \;=\; \operatorname*{argmax}_{\nu \in \log_2 \mathbb{N}} P\big(X_{2^\nu}(\theta) = k\big),$$

wobei $\log_2 \mathbb{N} = \{\nu \in \mathbb{R} \,:\, 2^\nu \in \mathbb{N}\}$.

Mit demselben Bijektivitätsargument für die lineare Abbildung $x \mapsto x - k$ verschieben wir nun die Maximalstelle, indem wir die Argumente der argmax-Bildung entsprechend anpassen:

$$\log_2 \hat{n}_{\mathsf{ML}}(k) \;-\; k \;=\; \operatorname*{argmax}_{\nu \in A_k} P\big(X_{2^{\nu+k}} = k\big).$$

Dabei ist $A_k := \log_2 \mathbb{N} - k = \{\nu \in \mathbb{R} \,:\, 2^{\nu+k} \in \mathbb{N}\}$. Diese Mengenfolge ist isoton, denn mit $x \in A_k$ existiert eine natürliche Zahl n mit $x = \log_2 n - k$, und da mit n auch $2n \in \mathbb{N}$ ist, gilt $x = \log_2(2n) - 1 - k = \log_2(2n) - (k+1)$, d.h. $x \in A_{k+1}$. Weiterhin liegt der Grenzwert

$$A \;:=\; \bigcup_{k=1}^{\infty} A_k$$

dieser Mengenfolge dicht in $\mathbb{R}$. Es gilt sogar noch mehr. Zu jedem $a \in \mathbb{R}$ existiert eine Folge $(a_k)_{k\in\mathbb{N}}$ mit $a_k \in A_k$ für alle $k \in \mathbb{N}$ und $\lim_{k\to\infty} a_k = a$: Sei $a \in \mathbb{R}$ beliebig, dann ist $\lceil 2^{k+a} \rceil \in \mathbb{N}$ für alle $k \in \mathbb{N}$ und

$$a_k \;:=\; \log_2\lceil 2^{k+a}\rceil - k \;\in\; A_k$$

leistet das gewünschte, wie man mit der Stetigkeit des Logarithmus leicht einsieht:

$$\lim_{k\to\infty} a_k \;=\; \lim_{k\to\infty}\Big(\log_2\lceil 2^{k+a}\rceil - k\Big) \;=\; \log_2 \lim_{k\to\infty}\left(\frac{\lceil 2^{k+a}\rceil}{2^k}\right) \;=\; \log_2 2^a \;=\; a.$$

So liegt es nahe, zu schließen, dass also

$$\lim_{k\to\infty}\big(\log_2 \hat{n}_{\mathsf{ML}}(k) - k\big) \;=\; \operatorname*{argmax}_{\nu\in\mathbb{R}} \lim_{k\to\infty} P\big(X_{n_k}(\theta) = k\big)$$

mit $n_k := \lfloor 2^{\nu+k} \rfloor$ gilt. Nach Satz 2.5 erhalten wir auf der rechten Seite

$$P\big(X_{n_k}(\theta) = k\big) \;=\; P\big(X_{n_k}(\theta) - \lfloor \log_2 n_k \rfloor = -\lfloor \nu \rfloor\big) \;\to\; Q_{\{\nu\}}\big(-\lfloor \nu \rfloor\big),$$

also insgesamt

$$\lim_{k\to\infty}\big(\log_2 \hat{n}_{\mathsf{ML}}(k) - k\big) \;=\; \operatorname*{argmax}_{\nu\in\mathbb{R}} Q_{\{\nu\}}\big(-\lfloor \nu \rfloor\big) \;\approx\; 0.3530830.$$

Mit diesem rein analytischen Ergebnis für die Schätzfunktion können wir zu einer Konsistenzaussage für den Schätzer selbst gelangen. Dafür stellen wir zunächst fest, dass mit wachsendem n auch die beobachtete Pfadlänge gegen unendlich strebt. Formal:

$$\lim_{n\to\infty} P_n(K > C) \;=\; 1$$

für jedes feste $C > 0$. Mit ζ als Maximalstelle der Funktion $\nu \mapsto Q_{\{\nu\}}(-\lfloor\nu\rfloor)$ gilt also

$$P_n\big(|\log_2 \hat{n}_{\mathsf{ML}} - K \;-\; \zeta| \;>\; \varepsilon\big) \;\to\; 0$$

mit $n \to \infty$ für alle $\varepsilon > 0$. Daraus folgt

$$2^{-K} \cdot \hat{n}_{\mathsf{ML}} \;\xrightarrow{\mathrm{p}}\; 2^{\zeta} \;\approx\; 1{,}2773.$$

Es bleibt die Frage, ob der berechnete Schätzer erwartungstreu ist, und wie sich seine Varianz mit wachsender Knotenzahl verhält.

Wir betrachten zunächst einen weiteren Schätzer für die Knotenzahl, der sich offensichtlich vom Maximum-Likelihood-Schätzer unterscheidet:

$$\hat{n}_0(K) \;:=\; 2^K - 1.$$

Im Kontext des approximate counting ist C_n der (zufällige) Zählerstand nach Zählung von n Objekten. Durch eine leichte Differenz in den Anfangsbedingungen erhalten wir auf den Fall des digitalen Suchbaums bezogen $\mathcal{L}(K|n) = \mathcal{L}(C_{n-1})$. In diesen Kontext übertragen finden wir die folgenden Resultate zu Erwartungswert und Varianz dieses Schätzers in [Fla85]. Dort steht in Proposition 0: $\mathbb{E}2^{C_n} = n + 2$ und $\mathrm{var}(2^{C_n}) = n(n+1)/2$. Der Ansatz ist dabei insofern verschieden, als dass grundsätzlich von einem bekannten n ausgegangen wird.

Unter der Bedingung, dass n der „wahre“ Parameter ist, erhalten wir als Erwartungswert des Schätzers $\hat{n}_0$

$$\mathbb{E}_n \hat{n}_0 \;=\; \mathbb{E}\big(2^{C_{n-1}} - 1\big) \;=\; n,$$

der Schätzer ist also erwartungstreu. Für die Varianz ergibt sich

$$\mathrm{var}_n\big(\hat{n}_0(K)\big) \;=\; \mathrm{var}\big(2^{C_{n-1}}\big) \;=\; \frac{n(n-1)}{2}.$$

Natürlich ist eine Varianz in der Größenordnung vom Quadrat der geschätzten Größe nicht gerade ein Zeichen für die Zuverlässigkeit des Schätzers. Andererseits ist aber ebenso klar, dass sich aus der Beobachtung einer einzelnen Pfadlänge kaum das Aussehen des Baumes in den übrigen Pfaden erraten lässt.

Um nun die Überlegungen zum Schätzen der Knotenzahl abzuschließen, wollen wir die betrachteten Schätzer vergleichen. Aus den vorhergehenden Überlegungen wird dabei klar, dass gilt

$$\frac{\hat{n}_{\mathsf{ML}}(K)}{\hat{n}_0(K)} \;\to\; 2^{\zeta}$$

in Wahrscheinlichkeit, mit wachsender Knotenzahl. Asymptotisch schätzt also der Likelihood-basierte Schätzer bei identischer Beobachtung immer um einen Faktor von $2^{\zeta} \approx 1.27$ größer als der erwartungstreue Schätzer.

3.2 Konfidenzschranken

Beobachtet man wieder die Tiefe K des externen Knotens entlang eines festen Pfades $\theta \in \{0,1\}^{\mathbb{N}}$, so lassen sich Konfidenzschranken für die Gesamtzahl der Knoten angeben. In der Situation des approximate counting lassen sich diese Schranken verwenden, um, basierend auf dem Wert des Zählers, stochastisch präzise Aussagen über die Anzahl der gezählten Ereignisse zu machen.

Zunächst ein paar Vorüberlegungen: Zu einem $\alpha \in (0,1)$ definieren wir

$$\psi_\alpha : \begin{cases} \mathbb{N} \;\to\; \mathbb{N}, \\ k \;\mapsto\; \inf\{n \in \mathbb{N} \,:\, P(S_k \leq n) \;\geq\; 1-\alpha\}, \end{cases}$$

wobei $\mathcal{L}(S_k) = \star_{j=1}^{k} \mathrm{Geom}(2^{1-j})$ wie oben. $\psi_\alpha(k)$ ist also das $(1-\alpha)$-Quantil der Verteilung von S_k. Offensichtlich ist $k \mapsto S_k$ in dem Sinne monoton wachsend, dass für $k,n \in \mathbb{N}$ gilt

$$P(S_k \leq n) \;\geq\; 1-\alpha \;\Longrightarrow\; P(S_{k-1} \leq n) \;\geq\; 1-\alpha,$$

also ist ψ_α (schwach) isoton. Weiter definieren wir

$$\psi_\alpha^{-1}(n) \;:=\; \inf\{k \in \mathbb{N} \,:\, \psi_\alpha(k) \geq n\}$$

in gewisser Analogie von Verteilungs- und Quantilfunktion. Eben diese Analogie inspiriert das folgende

Lemma 3.2 Für alle $k,n \in \mathbb{N}$ gilt

$$\psi_\alpha^{-1}(n) \;\leq\; k \quad \Longleftrightarrow \quad \psi_\alpha(k) \;\geq\; n.$$

Beweis. Seien $n,k \in \mathbb{N}$. Wegen der Isotonie von ψ_α folgt aus $\psi_\alpha^{-1}(n) \leq k$, dass $k \in \{j \in \mathbb{N} : \psi_\alpha(j) \geq n\}$, also ist $\psi_\alpha(k) \geq n$. Sei umgekehrt $\psi_\alpha(k) \geq n$, dann muss auch $\psi_\alpha^{-1}(n) \leq k$ gelten. □

In ihrer Eigenschaft als Quantilfunktion erfüllt $\alpha \mapsto \psi_\alpha(k)$ natürlich die Gleichungen

$$P(S_k \leq \psi_\alpha(k)) \;\geq\; 1-\alpha \qquad \text{und} \qquad P(S_k < \psi_\alpha(k)) \;<\; 1-\alpha. \tag{3.2}$$

Damit konstruieren wir nun die Konfidenzschranken.

Satz 3.3 Beobachtet man entlang des Pfades $\theta \in \{0,1\}^{\mathbb{N}}$ die Tiefe K des externen Knotens, so ist

$$\psi_{1-\alpha}(K) \qquad \text{bzw.} \qquad \psi_\alpha(K+1)$$

eine $(1-\alpha)$-Konfidenzunterschranke bzw. eine $(1-\alpha)$-Konfidenzoberschranke für die Gesamtzahl n der Knoten des Baumes.

Beweis. Es ist für alle $n \in \mathbb{N}$ zu zeigen, dass

$$P_n(n \geq \psi_{1-\alpha}(K)) \;\geq\; 1-\alpha \qquad \text{sowie} \qquad P_n(n \leq \psi_\alpha(K+1)) \;\geq\; 1-\alpha$$

gilt. Für die Unterschranke folgt dabei aus Lemma 3.2

$$\begin{aligned} P_n\big(n \geq \psi_{1-\alpha}(K)\big) &= P_n\big(n > \psi_{1-\alpha}(K) - 1\big) \\ &= P_n\big(\psi_{1-\alpha}^{-1}(n+1) > K\big) \\ &= P_n\big(X_n(\theta) < \psi_{1-\alpha}^{-1}(n+1)\big) \\ &= P_n(S_{\psi_{1-\alpha}^{-1}(n+1)} > n) \\ &= P_n\big(S_{\psi_{1-\alpha}^{-1}(n+1)} \geq n+1\big). \end{aligned}$$

Setzt man in Lemma 3.2 $k = \psi_\alpha(n)$, so erhält man $\psi_\alpha(\psi_\alpha^{-1}(n)) \geq n$, also ergibt sich mit (3.2)

$$\begin{aligned} P_n\big(n \geq \psi_{1-\alpha}(K)\big) \;\geq\; P_n\big(S_{\psi_{1-\alpha}^{-1}(n+1)} \geq \psi_{1-\alpha}(\psi_{1-\alpha}^{-1}(n+1))\big) \\ \geq\; 1 - (1 - (1-\alpha)) \;=\; 1 - \alpha. \end{aligned}$$

Für die Oberschranke erhalten wir analog nach Lemma 3.2

$$\begin{aligned} P_n(n \leq \psi_\alpha(K+1)) &= P_n(K+1 \geq \psi_\alpha^{-1}(n)) \\ &= P_n(X_n(\theta) \geq \psi_\alpha^{-1}(n) - 1) \\ &= P_n(S_{\psi_\alpha^{-1}(n)-1} \leq n). \end{aligned}$$

Setzen wir in Lemma 3.2 $k = \psi_\alpha^{-1}(n) - 1$, so erhalten wir $n \geq \psi_\alpha(\psi_\alpha^{-1}(n) - 1)$, und damit können wir weiter abschätzen

$$P_n(S_{\psi_\alpha^{-1}(n)-1} \leq n) \;\geq\; P_n(S_{\psi_\alpha^{-1}(n)-1} \leq \psi_\alpha(\psi_\alpha^{-1}(n) - 1)) \;\geq\; 1 - \alpha$$

nach Gleichung (3.2). □

Bemerkung 3.4 Zur tatsächlichen Berechnung der beschriebenen Schranken wollen wir zwei verschiedene Ansätze vorstellen.

- So, wie die Funktion ψ definiert wurde, liegt es nahe, die Verteilung von S_k für vorgegebenes k zu untersuchen. Dazu bedienen wir uns der wahrscheinlichkeitserzeugenden Funktionen und einer Computeralgebra-Software wie etwa `Maple`. Es ist

$$z \;\mapsto\; \frac{pz}{1-(1-p)z}, \quad z \in \mathbb{R}$$

die wahrscheinlichkeitserzeugende Funktion zur geometrischen Verteilung auf $\mathbb{N}$ mit Erfolgswahrscheinlichkeit p, somit ergibt sich als erzeugende Funktion zu $\star_{j=1}^{k} \mathrm{Geom}(2^{1-k})$

$$g_k(z) \;=\; \prod_{j=1}^{k} \frac{2^{1-j}z}{1-(1-2^{1-j})z}.$$

Die folgende Befehlsfolge liefert also $\psi_\alpha(k)$ für verschiedene Werte k und α:

```
g := (z,n) -> product((2^(1-i)*z)
                      /(1-(1-2^(1-i))*z),i=1..n):
for a in [0.95, 0.975] do
  for k in [2,3,4,5,6,7] do
    j := 0:
    quant := 0:
    while quant < a do
      j := j+1:
      quant := quant + coeff(series(g(z,k),
                                    z=0, j+1),
                             z,j):
    end do:
    printf("psi_%g(%d) = %d\n", a, k, j):
  end do:
end do:
```

Leider ist der (insbesondere algebraische) Rechenaufwand sehr hoch.

- Als weitere Möglichkeit ergibt sich ein Zugang zur Verteilung von S_k durch das erneuerungstheoretische Standardargument (2.1) zusammen mit der Rekursion (3.1). Wir arbeiten mit dem vorher angegebenen R-Programm weiter und speichern die Werte der Verteilungsfunktion in der Tabelle `pkumul`:

```
pkumul <- matrix(0,nmax,kmax)
pkumul[,1] <- p[,1]
for (k in 2:kmax)
  pkumul[,k] <- pkumul[,k-1] + p[,k]

alpha <- .9
konfober <- which.max(pkumul[,1] <= alpha)
for (k in 2:kmax)
  konfober <- c(konfober,
                which.max(pkumul[,k] <= alpha))
konfober
```

Damit erhält man einen Teil der in Tabelle 3.2 angegebenen Werte, die man natürlich auch, etwa wie in Tabelle 3.3, zu Konfidenzintervallen zusammenfassen kann. Hierbei arbeiten wir zunächst mit *equal tails*. Allerdings ist zu bedenken, dass die Konfidenzintervalle umso kleiner werden, je weniger man die Konfidenzunterschranke berücksichtigt: Nehmen wir beispielsweise das 0.95-Konfidenzintervall, so ist dessen Länge 593 für $k = 7$. Die Konfidenzoberschranke mit denselben Parametern ist aber 538, wir erhalten also ein kleineres Konfidenzintervall, indem wir nur die Oberschranke berücksichtigen. Als Alternative kann man auch Likelihood-basierte Konfidenzintervalle betrachten.

3.3 Approximative Konfidenzschranken (DST)

Aus Abschnitt 2.1 wissen wir, dass $2^{-k}S_k$ mit $k \to \infty$ in Verteilung gegen S_∞ konvergiert. Dies können wir für eine Approximation der Konfidenzschranken verwenden, indem wir für die Funktion ψ anstelle der Quantile von S_k diejenigen von $2^k S_\infty$ verwenden. Sei also

$$\tilde{\psi}_\alpha(k) := \inf\{x > 0 : P(S_\infty \leq 2^{-k}x) \geq 1 - \alpha\},$$

dann ist analog zu Satz 3.3 $\tilde{\psi}_{1-\alpha}(K)$ eine approximative $(1-\alpha)$-Konfidenzunterschranke für den unbekannten Parameter n.

Aus der Definition von $\tilde{\psi}$ ist sofort ersichtlich, dass gilt

$$\tilde{\psi}_\alpha(k) \;=\; 2^k \cdot \tilde{\psi}_\alpha(0).$$

Es genügt also, die Quantile von S_∞ zu bestimmen um daraus durch Multiplikation mit der der Beobachtung entsprechenden Zweierpotenz die Konfidenzober- und -unterschranken zu erhalten.

Wir verwenden die Darstellung von $\mathcal{L}(S_\infty)$ als Pseudo-Mischung von Exponentialverteilungen aus Abschnitt 2.3. Daraus ergibt sich

$$P(S_\infty \leq x) \;=\; 1 - P(S_\infty > x) \;=\; 1 - \sum_{j=1}^{\infty} a_j \; P(\tilde{Y}_j > x),$$

wobei $\mathcal{L}(\tilde{Y}_j) = \mathrm{Exp}(2^j)$ für alle $j \in \mathbb{N}$ und a_j definiert wie in Abschnitt 2.3. Mit $\mathcal{L}(2^j\tilde{Y}_j) = \mathrm{Exp}(1)$ folgt weiter

$$P(S_\infty \le x) \;=\; 1 - \sum_{j=1}^{\infty} a_j P(2^j\tilde{Y}_j > 2^j x) \;=\; 1 - \sum_{j=1}^{\infty} a_j \exp\bigl(-2^j x\bigr).$$

Da die Beträge der Koeffizienten a_j sehr schnell sehr klein werden, lässt sich die Verteilungsfunktion gut auf numerischem Wege bestimmen. Eine graphische Darstellung der Verteilungsfunktion findet sich in Abbildung 3.1, die numerischen Werte dazu stehen in Tabelle 3.5. Einige approximative Werte für die Konfidenzschranken stehen in Tabelle 3.4. (Die Tabellen finden sich auf den Seiten 61 bis 63.) Man sieht leicht, dass bereits bei kleinen Werten die Asymptotik sehr gut greift.

3.4 Approximative Konfidenzschranken (BST)

Für binäre Bäume, die aus dem BST-Algorithmus hervorgehen, gibt es eine Interpretation der Pfadlänge eines vorgegebenen Pfades $\theta \in \{0,1\}^{\mathbb{N}}$ als Anzahl von Rekorden. Sind $U_1,U_2,\ldots$ unabhängige, auf dem Einheitsintervall gleichverteilte Daten, $r_1 := 1$ und, für $k \in \mathbb{N}$,

$$R_k \;:=\; U_{r_k} \quad \text{mit} \quad r_{k+1} \;:=\; \min\{j > r_k : U_j < R_k\},$$

dann ist $R_1 > R_2 > \ldots$ eine absteigende Folge von Rekorden. Dies ist genau die Folge der Datenwerte, die entlang des Pfades $(0,0,\ldots)$ besucht werden. Beschränkt man sich auf die ersten n Datenwerte, so ist die Länge dieses Pfades die Anzahl der absteigenden Rekorde in der Folge $U_1,\ldots,U_n$. Einen analogen Zugang findet man ebenso für einen beliebigen Pfad θ. In der Tat ergibt sich (im Gegensatz zum Geburtsprozess bei DST) in diesem Fall für die Verteilung der Pfadlänge eine Faltung von Bernoulliverteilungen. Unabhängig von θ ist stets für $n \in \mathbb{N}$

$$\mathcal{L}(X_n(\theta)) \;=\; \star_{k=1}^{n} \mathrm{Ber}\bigl(\tfrac{1}{k}\bigr).$$

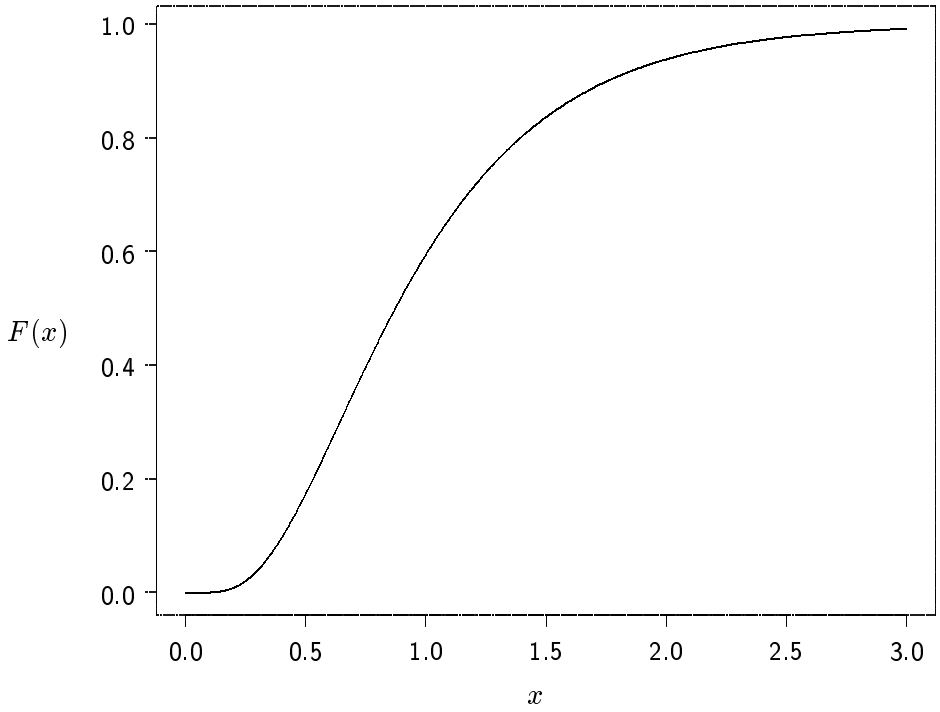

Abbildung 3.1: Verteilungsfunktion F von S_∞

Der Suche nach einem Schätzer für die Gesamtzahl an Daten ausgehend von einer vorgegebenen Anzahl von Rekorden haben sich bereits Moreno Rebollo et al. in [MBCB96] gewidmet. Sie schlagen drei verschiedene Schätzer vor, deren Asymptotik von Cramer in [Cra00] untersucht wird. Wir beschäftigen uns daher in diesem Kapitel für den BST-Fall nur mit approximativen Konfidenzschranken.

Ist nun X_n die Anzahl der Rekorde in $U_1, \ldots, U_n$, so ergibt sich aus dem

zentralen Grenzwertsatz, dass mit $n \to \infty$ gilt

$$\frac{X_n - \log n}{\sqrt{\log n}} \xrightarrow{\text{distr}} Z,$$

mit einer standardnormalverteilten Zufallsvariable Z (vgl. z.B. [ABN98, S. 12]). Dieses Resultat ist auch im Zusammenhang mit dem BST-Algorithmus wohlbekannt. Daraus können wir approximative Konfidenzschranken ableiten.

Satz 3.5 Zu $\alpha \in (0,1)$ sei q_α das α-Quantil der Standardnormalverteilung. Weiter sei

$$\tilde{\psi}_\alpha(k) := \exp\Big(\big(\sqrt{k + (q_\alpha/2)^2} - q_\alpha/2\big)^2 \Big), \quad k \in \mathbb{N}.$$

Beobachtet man entlang eines festen Pfades θ die Pfadlänge K, so ist

$$\tilde{\psi}_\alpha(K) \quad \text{bzw.} \quad \tilde{\psi}_{1-\alpha}(K)$$

eine approximative Konfidenzober- bzw. -unterschranke zum Niveau $1-\alpha$ für die Anzahl der Knoten im gesamten Baum, approximativ in dem Sinne, dass mit wachsendem Parameter n das Konfidenzniveau im Grenzübergang $n \to \infty$ eingehalten wird.

Beweis. Wir rechnen nach. Da $\sqrt{k+x^2} > x$ für alle $k \in \mathbb{N}$ und alle $x \in \mathbb{R}$, gilt für $x > -\sqrt{\log n}$

$$\begin{aligned} P_n\Big(n \geq \exp\Big(\big(\sqrt{K+x^2} - x\big)^2 \Big) \Big) &= P_n\Big(\sqrt{\log n} + x \geq \sqrt{K+x^2} \Big) \\ &= P_n\big((\sqrt{\log n} + x)^2 \geq K + x^2 \big) \\ &= P_n\big(K - \log n \leq 2x\sqrt{\log n} \big). \end{aligned}$$

Nun ist $\mathcal{L}(K|n) = \mathcal{L}(X_n(\theta))$, also folgt mit der Verteilungsasymptotik

$$\ldots = P_n\Big(\frac{X_n(\theta) - \log n}{\sqrt{\log n}} \leq 2x \Big) \to \Phi(2x), \quad \text{mit } n \to \infty.$$

Dabei bezeichnet Φ die Verteilungsfunktion der Standardnormalverteilung. Setzen wir nun $x = q_\alpha/2$ so ergibt sich durch Übergang zum Komplementärereignis die Behauptung für die Unterschranke. Mit $x = q_{1-\alpha}/2$ folgt die Behauptung für die Oberschranke. □

Bemerkung 3.6 Die Bedingung, die wir zu Beginn des Beweises angenommen haben, ist für interessante Unterschranken trivialerweise erfüllt, da hier das entsprechende Quantil positiv ist. Wie groß muss nun n sein, damit man auch bei negativen Quantilen, wie sie sich bei der Oberschranke ergeben, diese Approximation verwenden kann? Nennt man $n_{\min}$ denjenigen Wert von n, ab dem $q_\alpha > -2\sqrt{\log n}$ erfüllt ist, so ergeben sich folgende Werte

α	0.90	0.95	0.975	0.99	0.999
$n_{\min}$	2	2	2	4	11

Für Beobachtungen $K \geq 4$ kann man also problemlos mit $\alpha \geq 1/100$ arbeiten, ab $K \geq 11$ sogar auf $\alpha = 1/1000$ reduzieren.

Einige Werte dieser Schranke sind in Tabelle 3.6, S. 63, zusammengefasst. Im Vergleich mit den entsprechenden DST-Werten wird deutlich, dass der stochastische Unterschied in der Balance zwischen BST und DST erheblich ist. Führt im DST-Fall eine Erhöhung des beobachteten Wertes um 1 noch zu ungefähr einer Verdopplung des Wertes der Oberschranke, ist dies bei BST der Faktor 4.

3.5 Entscheidungsprobleme

Wie zu Beginn des Kapitels angekündigt, wollen wir zu einem gegebenen Baum untersuchen, mit welchem Algorithmus – DST oder BST – dieser Baum entstanden ist. Sei dazu t ein binärer Baum ohne Datenwerte. Mit

$$H_0 : t \text{ entstand durch den BST-Algorithmus,}$$
$$H_1 : t \text{ entstand durch den DST-Algorithmus,}$$

haben wir einen Test mit einfacher Hypothese und einfacher Alternative vorliegen. Nach dem Lemma von Neyman-Pearson ist damit die Testgröße für einen optimalen Test durch den Quotienten der Zähldichten gegeben, denn die beteiligten Wahrscheinlichkeitsmaße sind beide diskret und besitzen denselben Träger, nämlich die Menge der binären Bäume mit n Knoten.

Ist nun T_n ein durch den BST- oder den DST-Algorithmus aus einer Folge $U_1,\ldots,U_n$ von unabhängigen, unif(0,1)-verteilten Variablen erzeugter binärer Baum, so sind linker und rechter Teilbaum von T_n bedingt unabhängig unter der Knotenzahl im linken Teilbaum. Das bedeutet

$$\begin{aligned} P\big(T_n = t \big| \#L(T_n) = \#L(t)\big) \;&=\; P\big(L(T_n) = L(t)\big|\#L(T_n) = \#L(t)\big) \\ &\quad\cdot\; P\big(R(T_n) = R(t)\big|\#L(T_n) = \#L(t)\big). \end{aligned} \tag{3.3}$$

Für den DST-Algorithmus sieht man dies wie folgt: Ausgangspunkt sind n Eingabedaten $U_1,\ldots,U_n$, die unabhängig sind und derselben unif(0,1)-Verteilung genügen. Seien nun $U_{j_1},\ldots,U_{j_k}$, $2 \le j_1 < j_2 < \ldots < j_k \le n$, genau diejenigen k Datenwerte, die kleiner als $1/2$ sind, d.h. deren erstes Bit „0“ ist. Dann sind $U_{j_1},\ldots,U_{j_k}$ gleichverteilt auf $(0,1/2)$ und unabhängig von den übrigen Datenwerten. Insbesondere hängen also linker und rechter Teilbaum nur über ihre jeweilige Größe voneinander ab. Bedingen wir unter dieser Größe, so erhalten wir entsprechend Unabhängigkeit. Weiterhin ist für die Struktur des linken Teilbaums das erste Bit nicht von Bedeutung, dieses können wir durch Multiplikation mit 2 löschen. Dadurch erhalten wir für $2U_{j_1},\ldots,2U_{j_k}$ wieder die vorausgesetzte unif(0,1)-Verteilung, also einen DST-verteilten Baum mit k Knoten. Für die Verteilung des rechten Teilbaums reicht ein Symmetrieargument.

Um auch für den BST-Fall die bedingte Unabhängigkeit zu erklären, gehen wir von einer zufälligen Permutation Π aus, die gleichverteilt aus der Menge $\mathbb{S}_n$ der Permutationen von $\{1,\ldots,n\}$ ausgewählt wird. Aus Kapitel 1 wissen wir bereits, dass sich damit die Verteilung der BST-erzeugten Bäume nicht ändert. Die Größe des linken Teilbaums wird durch das erste Element der Permutation bestimmt. Streichen wir aus Π diejenigen Werte, die kleiner oder gleich Π_1 sind, so erhalten wir die (immernoch zufällige) Permutation Π_-.

Analog ergibt sich Π_+. Für die bedingte Unabhängigkeit der Teilbäume ist nun zu zeigen, dass für alle Permutationen $\pi \in \mathbb{S}_n$ und alle $k = 1, \ldots, n$ gilt

$$P(\Pi_- = \pi_-,\, \Pi_+ = \pi_+ \mid \Pi_1 = k) \;=\; P(\Pi_- = \pi_- | \Pi_1 = k) P(\Pi_+ = \pi_+ | \Pi_1 = k).$$

Dies lässt sich leicht nachrechnen. Es ist

$$\begin{aligned} P(\Pi_- = \pi_-,\, \Pi_+ = \pi_+ \mid \Pi_1 = k) \;&=\; \frac{\#\{\tau \in \mathbb{S}_n : \tau_- = \pi_-, \tau_+ = \pi_+, \tau_1 = k\}}{\#\{\tau \in \mathbb{S}_n : \tau_1 = k\}} \\ &=\; \frac{\binom{n-1}{k-1}}{(n-1)!} \;=\; \frac{1}{(k-1)!}\frac{1}{(n-k)!}, \end{aligned}$$

denn es gibt $\binom{n-1}{k-1}$ Möglichkeiten, die Positionen der Elemente von τ_- in τ festzulegen. Andererseits gibt es bei festem k genau $(k-1)!$ Permutationsmöglichkeiten für die Elemente von τ_- bzw. $(n-k)!$ für τ_+. Damit folgt

$$\begin{aligned} P(\Pi_- = \pi_- | \Pi_1 = k) \;&=\; \frac{\#\{\tau \in \mathbb{S}_n : \tau_- = \pi_-, \tau_1 = k\}}{(n-1)!} \\ &=\; \frac{\binom{n-1}{k-1}(n-k)!}{(n-1)!} \;=\; \frac{1}{(k-1)!}, \end{aligned}$$

und analog

$$P(\Pi_+ = \pi_+ | \Pi_1 = k) \;=\; \frac{1}{(n-k)!}.$$

Insgesamt erhalten wir die behauptete bedingte Unabhängigkeit.

Weiterhin sieht man an dieser Rechnung, dass $\mathcal{L}(\Pi_- | \Pi_1 = k) = \mathrm{unif}(\mathbb{S}_{k-1})$ gilt. Der linke Teilbaum hat also bedingt unter seiner Knotenzahl k dieselbe Verteilung wie ein BST-erzeugter Baum mit k Knoten. Damit ergibt sich

$$P\big(L(T_n) = L(t) \big| \#L(T_n) = \#L(t)\big) \;=\; P\big(T_{\#L(t)} = L(t)\big). \tag{3.4}$$

Analog ergibt sich diese Aussage auch für $R(T_n)$.

Nun können wir die bedingte Unabhängigkeit der Teilbäume wie folgt ausnutzen.

Lemma 3.7

$$P(T_n = t) \;=\; \prod_{u \in t} P\big(\#L(T_n(u)) = \#L(t(u))\big).$$

Beweis. Rekursiv über die Anzahl der Knoten. Für $n = 1$ ist die Aussage trivialerweise richtig, gelte sie also für alle natürlichen Zahlen kleiner als n. Multipliziert man (3.3) mit $P(\#L(T_n) = \#L(t))$, so erhält man nach Gleichung (3.4)

$$P(T_n = t) \;=\; P\big(\#L(T_n) = \#L(t)\big) \cdot P\big(T_{\#L(t)} = L(t)\big) \cdot P\big(T_{\#R(t)} = R(t)\big).$$

Wegen $\#L(t) + \#R(t) = n - 1$ sind insbesondere $\#L(t)$ und $\#R(t)$ beide kleiner oder gleich $n - 1$, und für die äußeren Wahrscheinlichkeiten gilt die Induktionsvoraussetzung, d.h.

$$\begin{aligned} P(T_n = t) \;=\; & \prod_{u \in L(t)} P\big(\#L(T_n(u)) = \#L(t(u))\big) \\ & \cdot P\big(\#L(T_n) = \#L(t)\big) \\ & \cdot \prod_{u \in R(t)} P\big(\#L(T_n(u)) = \#L(t(u))\big). \end{aligned}$$

Zieht man die „großen“ Produkte zusammen, fehlt nur noch der Faktor für den Wurzelknoten, der jedoch in der Mitte der drei Faktoren steht; insgesamt folgt die Behauptung. □

Bis hierhin konnten wir gleiche strukturelle Eigenschaften von BST- und DST-erzeugten Bäumen ausnutzen, daher war keine Unterscheidung der Wahrscheinlichkeitsmaße nötig. Betrachtet man jedoch die einzelnen Faktoren im vorangegangenen Lemma, so ist schnell klar, dass diese vom erzeugenden Algorithmus abhängen. Von hier an unterscheiden wir also die Wahrscheinlichkeitsmaße durch einen selbsterklärenden Index.

Um die Lesbarkeit weiter zu erhöhen, führen wir noch die folgenden Abkürzungen ein. Zu $u \in t$ sei

$$k(u) := \#t(u), \qquad l(u) := \#L(t(u)), \quad \text{und} \quad r(u) := \#R(t(u)).$$

Der BST-bezogene Teil des nachfolgenden Lemmas ist bereits als Theorem 6.1 in [SF96] zu finden.

Lemma 3.8 Ist t ein binärer Baum mit n Knoten, so gilt

$$P^{\mathrm{BST}}(T_n = t) \;=\; \prod_{u \in t} \frac{1}{k(u)}, \qquad P^{\mathrm{DST}}(T_n = t) \;=\; n! \cdot 2^{n - \sum_{u \in t} k(u)} \cdot \prod_{u \in t} \frac{1}{k(u)}.$$

Beweis. Nach Lemma 3.7 ist die entscheidende Größe die jeweilige Wahrscheinlichkeit des Ereignisses

$$\big\{ \#L(T_n) = k \big\}, \qquad k = 0, \ldots, n-1.$$

Im BST-Fall heißt dies, dass genau k der Datenwerte $U_2, \ldots, U_n$ kleiner sind als U_1, dass also U_1 den absoluten Rang $k+1$ besitzt. Da die Permutationen der Ränge der $U_1, \ldots, U_n$ alle gleichwahrscheinlich sind, folgt

$$P^{\mathrm{BST}}\big(\#L(T_n) = k\big) \;=\; \frac{1}{n}, \qquad \text{für alle } k = 0, \ldots, n-1,$$

$\#L(T_n)$ ist also gleichverteilt auf $\{0, \ldots, n-1\}$, und damit ist die erste Gleichheit bereits klar.

Im DST-Fall ist $\#L(T_n)$ gerade die Anzahl der Datenwerte, die in das Intervall $[0, \frac{1}{2})$ fallen, so man den ersten Wert (die Wurzel) außer Acht lässt. Damit ergibt sich sofort, dass $\#L(T_n)$ einer Binomialverteilung genügt: Ein Experiment mit Erfolgswahrscheinlichkeit $\frac{1}{2}$ wird $n-1$ mal wiederholt. Damit können wir Lemma 3.7 benutzen und erhalten

$$P^{\mathrm{DST}}(T_n = t) \;=\; \prod_{u \in t} \binom{k(u) - 1}{l(u)} 2^{k(u)-1}.$$

Wir schauen uns das Produkt der Binomialkoeffizienten etwas genauer an. Offensichtlich gilt für alle $u \in t_n$

$$k(u) - 1 \;=\; l(u) + r(u),$$

also ist

$$\prod_{u\in t}\binom{k(u)-1}{l(u)} = \prod_{u\in t}\frac{\big(k(u)-1\big)!}{l(u)!\big(k(u)-1-l(u)\big)!} = \prod_{u\in t}\frac{k(u)!}{l(u)!r(u)!}\cdot\frac{1}{k(u)}.$$

Nun ist jeder Knoten $u \in t$ entweder linker oder rechter Nachfolger eines anderen Knotens $u' \in t$ – mit Ausnahme der Wurzel. Das bedeutet, dass jedes $k(u')!$ sich genau entweder als $l(u)!$ oder als $r(u)!$ wiederfindet, solange u' nicht der Wurzelknoten ist. Leere Teilbäume spielen wegen $0! = 1$ hier keine Rolle. Damit vereinfacht sich das erste Teilprodukt zu

$$\prod_{u\in t}\frac{k(u)!}{l(u)!r(u)!} \;=\; (\#t)! \;=\; n!$$

und es folgt die Behauptung. □

Als Dichtequotienten erhalten wir damit

$$\frac{P^{\mathrm{DST}}(T_n = t)}{P^{\mathrm{BST}}(T_n = t)} \;=\; n!\cdot 2^{n-\sum_{u\in t}k(u)}.$$

Für die Summe der Teilbaumknotenzahlen gibt es nun verschiedene Vereinfachungs- und Umformungsmöglichkeiten. Wie oft ein einzelner Knoten in der Summe gezählt wird, d.h., in wievielen Teilbäumen ein Knoten vorkommt, hängt direkt von seiner Tiefe im Baum ab. Es gilt also

$$\sum_{u\in t}k(u) \;=\; \sum_{u\in t}\big(h(u)+1\big) \;=\; n+\sum_{u\in t}h(u),$$

wobei $h(u)$ die Tiefe des Knotens u bezeichnet. (Vergleiche dazu auch Abschnitt 4.1.2 im folgenden Kapitel.)

Da die Transformation $x \mapsto 2^{-x}$ monoton ist, kann man also bemerkenswerterweise die Pfadlänge des vorgegebenen Baumes als Testgröße in einem optimalen Test BST gegen DST verwenden.

Satz 3.9 Der optimale Test φ für H_0 gegen H_1 zum Niveau α ist von der Form

$$\varphi(t) \;=\; \begin{cases} 1, & \text{falls } P(t) < c(t) \\ \gamma, & \text{falls } P(t) = c(t) \\ 0, & \text{falls } P(t) > c(t). \end{cases}$$

Dabei bezeichnet $P(t)$ die interne Pfadlänge von t und der kritische Wert $c(t)$ ist das α-Quantil der Verteilung von $P(T_n)$ mit $n = \#t$ unter BST.

Beweis. Folgt aus dem Lemma von Neyman-Pearson und den vorhergegangenen Überlegungen. □

Über die Asymptotik der Verteilung der Pfadlänge im BST-erzeugten Baum kann man auch einen Test konstruieren, der das geforderte Niveau zumindest asymptotisch einhält.

Für die Pfadlänge P_n eines binären Baumes gilt unter H_0

$$\frac{P_n - 2n\log n}{n} \;\xrightarrow{P\text{-f.s.}}\; Z$$

mit einer quadratisch integrierbaren und nicht degenerierten Zufallsvariable Z (Theorem 7.3 in [Mah00]). Damit können wir zu einem vorgegebenen Testniveau $\alpha \in (0,1)$ eine Folge von kritischen Werten angeben, so dass der Test das Testniveau asymptotisch einhält.

Sei dazu q_α das α-Quantil der Grenzverteilung, dann erfüllt $C_n = nq_\alpha + 2n\log n$ diese Bedingung. Denn es gilt

$$\begin{aligned} P^{\mathrm{BST}}(P_n \le nq_\alpha + 2n\log n) \;&=\; P^{\mathrm{BST}}\Big(\frac{P_n - 2n\log n}{n} \le q_\alpha\Big) \\ &\to\; P^{\mathrm{BST}}(Z \le q_\alpha) \;=\; \alpha. \end{aligned}$$

$k\backslash\alpha$	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99
2	2	2	2	2	5	6	7	8
3	3	3	3	3	12	14	17	20
4	4	5	6	7	26	31	36	43
5	8	9	11	13	54	65	76	90
6	14	17	21	26	110	133	155	184
7	28	34	42	52	223	268	312	371
8	54	68	83	104	448	538	627	745
9	108	136	166	208	898	1079	1258	1493
10	214	271	331	415	1797	2160	2518	2989

Tabelle 3.2: Werte von $\psi_\alpha(k)$ für verschiedene α und k (DST)

	k	1	2	3	4	5	6	7	8	9
$\alpha = 0.1$	von	1	2	3	6	11	21	42	83	166
	bis	5	12	26	65	133	268	538	1079	2160
$\alpha = 0.05$	von	1	2	3	5	9	17	34	68	136
	bis	6	14	31	76	155	312	627	1258	2518

Tabelle 3.3: DST-Konfidenzintervalle zum Niveau 0.9 und 0.95

$k\backslash\alpha$	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99
2	0.8	1.1	1.3	1.6	7.0	8.4	9.8	11.7
3	1.7	2.1	2.6	3.2	14.1	16.9	19.7	23.4
4	3.3	4.2	5.2	6.5	28.1	33.8	39.4	46.8
5	6.7	8.4	10.3	13	56.2	67.6	78.8	93.5
6	13.3	16.9	20.7	25.9	112.5	135.1	157.6	187.0
7	26.7	33.8	41.3	51.9	224.9	270.3	315.1	374.0
8	53.4	67.6	82.6	103.7	449.8	540.5	630.2	748.1
9	106.7	135.2	165.2	207.4	899.7	1081.1	1260.5	1496.1
10	213.4	270.4	330.4	414.9	1799.4	2162.2	2520.9	2992.3

Tabelle 3.4: Approximative DST-Konfidenzschranken mittels $\tilde{\psi}$.

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002
0.1	0.0003	0.0005	0.0008	0.0012	0.0017	0.0023	0.0032	0.0042	0.0054	0.0068
0.2	0.0084	0.0103	0.0124	0.0148	0.0175	0.0204	0.0236	0.0271	0.0309	0.0349
0.3	0.0392	0.0438	0.0486	0.0538	0.0591	0.0648	0.0706	0.0768	0.0831	0.0897
0.4	0.0964	0.1034	0.1106	0.1179	0.1254	0.1331	0.1410	0.1489	0.1571	0.1653
0.5	0.1737	0.1821	0.1907	0.1994	0.2081	0.2169	0.2258	0.2347	0.2436	0.2526
0.6	0.2617	0.2707	0.2798	0.2889	0.2980	0.3071	0.3162	0.3253	0.3344	0.3434
0.7	0.3524	0.3614	0.3703	0.3792	0.3881	0.3969	0.4057	0.4144	0.4230	0.4316
0.8	0.4401	0.4486	0.4570	0.4653	0.4735	0.4817	0.4898	0.4978	0.5057	0.5136
0.9	0.5214	0.5291	0.5367	0.5442	0.5516	0.5590	0.5662	0.5734	0.5805	0.5875
1.0	0.5944	0.6012	0.6080	0.6146	0.6212	0.6276	0.6340	0.6403	0.6465	0.6526
1.1	0.6587	0.6646	0.6705	0.6762	0.6819	0.6875	0.6930	0.6985	0.7038	0.7091
1.2	0.7143	0.7194	0.7244	0.7294	0.7342	0.7390	0.7438	0.7484	0.7530	0.7575
1.3	0.7619	0.7662	0.7705	0.7747	0.7788	0.7829	0.7869	0.7908	0.7947	0.7985
1.4	0.8022	0.8059	0.8095	0.8130	0.8165	0.8199	0.8233	0.8266	0.8299	0.8330
1.5	0.8362	0.8393	0.8423	0.8452	0.8482	0.8510	0.8538	0.8566	0.8593	0.8620
1.6	0.8646	0.8672	0.8697	0.8722	0.8746	0.8770	0.8793	0.8816	0.8839	0.8861
1.7	0.8883	0.8904	0.8925	0.8946	0.8966	0.8986	0.9005	0.9024	0.9043	0.9062
1.8	0.9080	0.9097	0.9115	0.9132	0.9149	0.9165	0.9181	0.9197	0.9213	0.9228
1.9	0.9243	0.9257	0.9272	0.9286	0.9300	0.9313	0.9327	0.9340	0.9352	0.9365
2.0	0.9377	0.9389	0.9401	0.9413	0.9424	0.9436	0.9447	0.9457	0.9468	0.9478
2.1	0.9489	0.9499	0.9508	0.9518	0.9527	0.9537	0.9546	0.9554	0.9563	0.9572
2.2	0.9580	0.9588	0.9596	0.9604	0.9612	0.9620	0.9627	0.9634	0.9642	0.9649
2.3	0.9655	0.9662	0.9669	0.9675	0.9682	0.9688	0.9694	0.9700	0.9706	0.9712
2.4	0.9717	0.9723	0.9728	0.9734	0.9739	0.9744	0.9749	0.9754	0.9759	0.9764
2.5	0.9768	0.9773	0.9777	0.9782	0.9786	0.9790	0.9794	0.9798	0.9802	0.9806
2.6	0.9810	0.9814	0.9817	0.9821	0.9825	0.9828	0.9831	0.9835	0.9838	0.9841
2.7	0.9844	0.9847	0.9850	0.9853	0.9856	0.9859	0.9862	0.9865	0.9867	0.9870
2.8	0.9872	0.9875	0.9877	0.9880	0.9882	0.9885	0.9887	0.9889	0.9891	0.9893
2.9	0.9895	0.9898	0.9900	0.9902	0.9903	0.9905	0.9907	0.9909	0.9911	0.9913
3.0	0.9914	0.9916	0.9918	0.9919	0.9921	0.9923	0.9924	0.9926	0.9927	0.9928
3.1	0.9930	0.9931	0.9933	0.9934	0.9935	0.9937	0.9938	0.9939	0.9940	0.9941
3.2	0.9943	0.9944	0.9945	0.9946	0.9947	0.9948	0.9949	0.9950	0.9951	0.9952
3.3	0.9953	0.9954	0.9955	0.9956	0.9957	0.9957	0.9958	0.9959	0.9960	0.9961
3.4	0.9961	0.9962	0.9963	0.9964	0.9964	0.9965	0.9966	0.9967	0.9967	0.9968
3.5	0.9968	0.9969	0.9970	0.9970	0.9971	0.9971	0.9972	0.9973	0.9973	0.9974
3.6	0.9974	0.9975	0.9975	0.9976	0.9976	0.9977	0.9977	0.9978	0.9978	0.9978
3.7	0.9979	0.9979	0.9980	0.9980	0.9980	0.9981	0.9981	0.9982	0.9982	0.9982
3.8	0.9983	0.9983	0.9983	0.9984	0.9984	0.9984	0.9985	0.9985	0.9985	0.9986
3.9	0.9986	0.9986	0.9986	0.9987	0.9987	0.9987	0.9987	0.9988	0.9988	0.9988

Tabelle 3.5: Werte der Verteilungsfunktion von S_∞

$k \backslash \alpha$	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99
2	1.6	1.7	1.9	2.3	122.8	421.4	1470.0	7830.3
3	2.3	2.8	3.3	4.3	486.8	1820.0	6776.3	38545.3
4	3.8	4.7	6.0	8.4	1831.0	7403.9	29317.4	177779.7
5	6.3	8.5	11.4	17.1	6648.6	28901.9	121267.9	781938.7
6	11.0	15.6	22.3	35.7	23528.5	109410.3	484854.9	3315627.6
7	19.7	29.7	44.5	76.1	81624.0	404356.1	1887181.5	13652649.6
8	36.3	57.6	90.7	164.8	278679.6	1465642.3	7186104.5	54872959.0
9	68.1	113.9	187.9	361.3	938993.0	5227270.3	26866337.3	216091558.1
10	130.0	228.6	394.8	801.3	3128834.7	18389457.0	98885454.7	836204884.5

Tabelle 3.6: Approximative BST-Konfidenzschranken mittels $\tilde{\psi}_\alpha$

4 Das Teilbaumgrößenprofil

Die Untersuchung von Profilen und Kenngrößen hat eine große Tradition bei der Analyse diskreter Strukturen. Statt eine Struktur direkt zu betrachten, beschreibt man ausgesuchte Charakteristika und deren Verhalten unter bestimmten Voraussetzungen. Meist lassen die erhaltenen Ergebnisse dann wieder Rückschlüsse auf die eingangs untersuchte Struktur zu.

Das konventionelle Profil eines (binären) Baumes bildet eine nichtnegative ganze Zahl k ab auf die Anzahl U_k der externen bzw. internen Knoten auf Höhe k. Man unterscheidet das externe Knotenprofil und das interne Knotenprofil. Im Falle des binären Baumes gehen diese auseinander hervor, weshalb meistens nur von einem *Knotenprofil* oder hier insbesondere vom *konventionellen* Profil die Rede ist.

Das konventionelle Profil hat insbesondere bei der iterativen Betrachtung des Wachstums eines binären Baumes große Vorteile: Beim Einfügen eines Knotens in einen Baum in Tiefe k ändert sich im internen Profil ausschließlich U_k, im externen Profil wächst U_{k+1} um 2, während U_k um 1 abnimmt. Natürlich muss bei dieser Betrachtungsweise die Einfügetiefe k (zumindest ihrer stochastischen Struktur nach) bekannt sein; die Verteilung von k hängt jedoch meist direkt von den Werten des Profils ab.

Wir wollen nun die Perspektive auf den Baum ändern und ein weiteres Profil einführen: Dazu betrachten wir zu einer natürlichen Zahl j die Anzahl $k_{n,j}$ der Knoten, deren Teilbaum aus genau j Knoten besteht. Der erste Index n bezieht sich dabei auf die Gesamtzahl der (internen) Knoten im Baum. Das so erhaltene Profil $k_n = (k_{n,1}, k_{n,2}, \ldots)$ nennen wir *Teilbaumgrößenprofil* oder *subtree size profile*.

Wie schon in Kapitel 1 angekündigt, interessieren wir uns vor allem für das Verhalten dieses Profils bei immer größer werdenden Bäumen. Dabei ergeben sich verschiedene „interessante Bereiche“ dieses Profils. Bei der Untersuchung des Anfangsstückes, also von $(k_{n,1},\ldots,k_{n,d})$ für festes $d \in \mathbb{N}$ und wachsendes n, sprechen wir von *Mäusen*. Die Asymptotik der Mäuse wird in Abschnitt 4.3.1 für den Fall des binären Suchbaums behandelt. Betrachten wir umgekehrt die „großen Knoten“, so sprechen wir mitunter von *Eisbergen*. Ein zentrales Resultat dieses Kapitels ist Satz 4.17, ein funktionaler Grenzwertsatz für stochastische Prozesse, die sich aus der Betrachtung dieser Eisberge ergeben.

4.1 Elementare Eigenschaften

Wir wollen zunächst ein paar elementare Eigenschaften festhalten. Da der Wurzelknoten stets der einzige Knoten ist, dessen Teilbaum aus allen Knoten des Baumes besteht, gilt

$$k_{n,n} \equiv 1, \qquad \text{für alle } n \in \mathbb{N}.$$

Größere Teilbäume als den des Wurzelknotens kann es freilich nicht geben, also

$$k_{n,j} \equiv 0, \qquad \text{für alle } n,j \in \mathbb{N} \text{ mit } j > n.$$

Aus diesem Grund ignorieren wir von Zeit zu Zeit die „leeren“ Komponenten und schreiben $(k_{n,1},\ldots,k_{n,n})$ anstelle von $(k_{n,1},k_{n,2},\ldots)$.

Zählt man alle Knoten zusammen, so erhält man die Gesamtknotenzahl n; dies ist auch der Fall, wenn wir die Menge der Knoten zerlegen nach der jeweiligen Größe ihres Teilbaums. Folglich gilt

$$\sum_{j=1}^{\infty} k_{n,j} = \sum_{j=1}^{n} k_{n,j} = n.$$

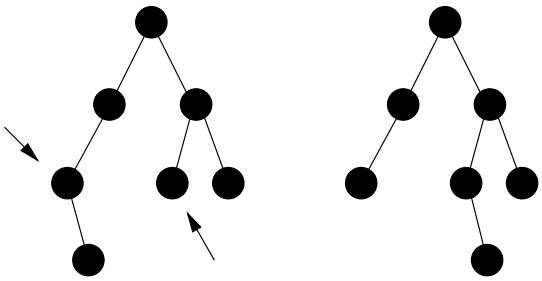

Abbildung 4.1: Vertauschen von zwei Teilbäumen in gleicher Tiefe. Zum linken Baum gehört das Teilbaumgrößenprofil (3,1,2,0,0,0,1); rechts ergibt sich nach Vertauschen der markierten Teilbäume (3,2,0,1,0,0,1).

4.1.1 Abgrenzung zum konventionellen Profil

Die wesentlichen Unterschiede zwischen Teilbaumgrößenprofil und konventionellem Knotenprofil lassen sich insbesondere an solchen Operationen auf Bäumen festmachen, unter denen diese Profile invariant sind. Ein Beispiel für diese Invarianz ist die Vertauschung von linkem und rechtem Teilbaum eines ausgewählten Knotens. Dabei ändern sich nämlich weder konventionelles Knotenprofil noch Teilbaumgrößenprofil.

Anders sieht es aus, wenn man zwei Teilbäume vertauscht, deren Wurzelknoten dieselbe Tiefe besitzen. Man überlegt sich leicht, dass das konventionelle Profil hiervon unbeeinflusst bleibt; denn die Anzahl der Knoten auf einem bestimmten Level ändert sich nicht. Das Teilbaumgrößenprofil ist unter dieser Operation jedoch im Allgemeinen nicht invariant. Ein Beispiel zeigt Abbildung 4.1.

Es gibt jedoch eine Möglichkeit, Teilbäume so zu vertauschen, dass sich das Teilbaumgrößenprofil nicht ändert: Die Teilbäume müssen nur dieselbe Anzahl von Knoten haben.

Lemma 4.1 Sei t_n ein binärer Baum mit n Knoten und $u,v \in t_n$ mit $\#t(u) = \#t(v)$. Dann gilt: durch Vertauschen von $t(u)$ und $t(v)$ ändert sich das Teilbaumgrößenprofil von t_n nicht.

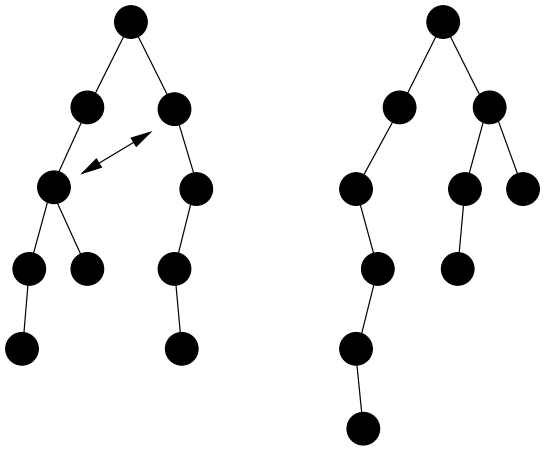

Abbildung 4.2: Vertauschen von zwei Teilbäumen mit gleicher Knotenzahl. Dass die konventionellen Profile der beiden Bäume verschieden sind, sieht man bereits an der unterschiedlichen Gesamthöhe.

Beweis. Wir stellen zunächst fest, dass sich für keinen Knoten außerhalb von $t(u)$ und $t(v)$ die Knotenzahl seines Teilbaums ändert: da $t(u)$ und $t(v)$ dieselbe Knotenzahl haben, gilt dies auch für alle Ahnen von u bzw. v.

Die Beiträge von $t(u)$ und $t(v)$ zum Teilbaumgrößenprofil ändern sich jedoch ebenfalls nicht durch das Vertauschen. Insgesamt bleibt das Profil also konstant. □

Mit diesem Lemma haben wir einen weiteren Unterschied zum konventionellen Profil gefunden: Dieses ist nämlich im Allgemeinen nicht invariant unter der beschriebenen Vertauschungsoperation. Ein Beispiel hierfür liefert Abbildung 4.2.

4.1.2 Interne und externe Pfadlänge

Wir schreiben $V(t)$ für die externe Pfadlänge eines binären Baumes t, $P(t)$ für die interne. Bereits in Kapitel 1 wurde erwähnt, dass stets

$$V(t) \;=\; P(t) + 2 \cdot \#t$$

gilt. Weiterhin ist klar, dass mit $h(u)$ definiert als die Höhe des Knotens u, ∂t als die Menge der externen Knoten von t gilt

$$V(t) \;=\; \sum_{v \in \partial t} h(v) \quad \text{und} \quad P(t) \;=\; \sum_{u \in t} h(u).$$

Es ist leicht zu sehen, dass man die Pfadlänge aus dem externen oder internen Knotenprofil berechnen kann. Desweiteren lässt sich die Pfadlänge aber auch aus dem Teilbaumgrößenprofil berechnen. Für die Beweise in diesem und im nächsten Teilabschnitt nutzen wir die Abkürzung t^* für die Menge $t \setminus \{\emptyset\}$ der Knoten von t ohne den Wurzelknoten.

Lemma 4.2

$$P(t) \;=\; \sum_{j=1}^{n-1} j \cdot k_{n,j}.$$

Beweis. Wir betrachten einen einzelnen Knoten $u \in t^*$. Wie oft wird bei der Berechnung der internen Pfadlänge nun die Strecke von u zu seinem Vorfahr gezählt? Genau so oft, wie $t(u)$ Knoten besitzt. Damit folgt

$$P(t) = \sum_{u \in t} h(u) = \sum_{u \in t^*} \#t(u) = \sum_{j=1}^{n-1} j k_{n,j}. \qquad \square$$

4.1.3 Wiener-Index

Für einen zusammenhängenden Graphen bezeichnet der Wiener-Index die Summe aller Abstände zwischen ungeordneten Knotenpaaren. Die Größe geht

auf den Chemiker H. Wiener zurück, der damit Korrelationen zwischen bestimmten Eigenschaften organischer Substanzen und deren Molekülstrukturen nachweisen konnte. Siehe dazu auch [Nei02].

Bezeichnet man den Wiener-Index zu einem binären Baum t mit $\mathrm{WI}(t)$, so lässt sich dieser wie folgt berechnen:

Lemma 4.3

$$\mathrm{WI}(t) \;=\; \#t \cdot \sum_{u \in t^*} \#t(u) \;-\; \sum_{u \in t^*} \big(\#t(u)\big)^2.$$

Beweis. Für zwei Knoten $u,v \in t$ sei $g(u,v)$ die maximale Tiefe eines gemeinsamen Vorfahr von u und v. Wir zeigen zunächst

$$\sum_{u,v \in t} g(u,v) \;=\; \sum_{u \in t^*} \big(\#t(u)\big)^2.$$

Induktion: Bei $\#t = 1$ sind beide Seiten Null. Sei also $\#t = n$ und die Aussage für kleinere Bäume bewiesen. t zerfällt in $L(t)$ und $R(t)$, und da $g(u,v)$ nur dann verschieden von Null ist, falls u und v beide aus $L(t)$ oder beide aus $R(t)$ stammen, gilt

$$\sum_{u,v \in t} g(u,v) \;=\; \sum_{u,v \in L(t)} g(u,v) \;+\; \sum_{u,v \in R(t)} g(u,v).$$

Jetzt schreiben wir $g^*(u,v)$, wenn wir die Tiefe bezüglich $L(t)$ bzw. $R(t)$ meinen, und erhalten nach Induktionsvoraussetzung

$$\begin{aligned}
\sum_{u,v \in t} g(u,v) &= \sum_{u,v \in L(t)} \big(g^*(u,v)+1\big) \;+\; \sum_{u,v \in R(t)} \big(g^*(u,v)+1\big) \\
&= \sum_{u \in L(t)^*} \big(\#t(u)\big)^2 \;+\; \big(\#L(t)\big)^2 \;+\; \sum_{u \in R(t)^*} \big(\#t(u)\big)^2 \;+\; \big(\#R(t)\big)^2 \\
&= \sum_{u \in L(t)} \big(\#t(u)\big)^2 \;+\; \sum_{u \in R(t)} \big(\#t(u)\big)^2 \\
&= \sum_{u \in t^*} \big(\#t(u)\big)^2.
\end{aligned}$$

Der Abstand d zweier Knoten ist $d(u,v) = h(u) + h(v) - 2g(u,v)$, also folgt für den Wiener-Index

$$\begin{aligned} 2 \cdot \mathrm{WI}(t) &= 2 \sum_{\{u,v\} \subset t} d(u,v) \;=\; \sum_{u,v \in t} d(u,v) \\ &= \sum_{u,v \in t} h(u) \;+\; \sum_{u,v \in t} h(v) \;-\; 2 \sum_{u,v \in t} g(u,v) \\ &= 2\#t \cdot \sum_{u \in t} h(u) \;-\; 2 \sum_{u \in t^*} \big(\#T(u)\big)^2. \end{aligned}$$

Die Beziehung $\sum_{u \in t} h(u) = \sum_{u \in t^*} \#t(u)$ wurde bereits in Abschnitt 4.1.2 nachgewiesen. □

Natürlich kann man so den Wiener-Index auch als Funktion des Teilbaumgrößenprofils schreiben: Mit $n = \#t$ und $(k_{n,1}, k_{n,2}, \ldots)$ als Teilbaumgrößenprofil zu t gilt

$$\mathrm{WI}(t) \;=\; n \cdot \sum_{j=1}^{n-1} j \cdot k_{nj} \;-\; \sum_{j=1}^{n-1} j^2 \cdot k_{nj} \;=\; \sum_{j=1}^{n-1} j(n-j) k_{nj}.$$

4.1.4 Beziehungen zwischen den Größen

Offensichtlich lassen sich alle in diesem Abschnitt eingeführten Profile und Kenngrößen für einen gegebenen Baum eindeutig bestimmen. Die gesamten Beziehungen sind in Abbildung 4.3 dargestellt. Ein Pfeil von A nach B bedeutet dabei, dass sich die Größe B aus der Größe A eindeutig bestimmen lässt. Ein fehlender Pfeil heißt, dass diese Bestimmung im allgemeinen nicht möglich ist. Dabei liefern Abbildungen 4.1 und 4.2 bereits die nötigen Gegenbeispiele für die Nicht-Beziehung zwischen Teilbaumgrößen- und konventionellem Profil. Dass sich der Wiener-Index nicht aus letzteren berechnen lässt, sieht man ebenfalls an Abbildung 4.1: die externen Profile der beiden Bäume sind gleich (somit auch die internen), die Wiener Indizes sind jedoch verschieden – nämlich 52 und 50.

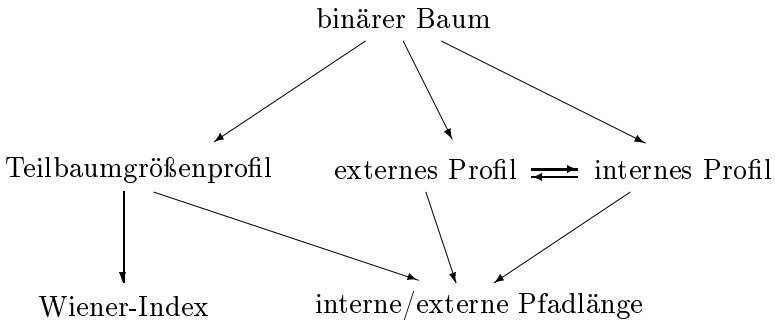

Abbildung 4.3: Beziehungen zwischen Profilen und Kenngrößen

4.2 Rekursive Verteilungsfamilien

Nun konstruieren wir auf der Menge $\mathcal{T}_n$ der binären Bäume mit n Knoten eine Verteilung Q_n. Ausgangspunkt ist eine Familie $(\mu_n)_{n\in\mathbb{N}}$ von diskreten Verteilungen, wobei für alle $n \in \mathbb{N}$ gilt, dass μ_n auf $\{0,1,\ldots,n-1\}$ konzentriert ist. Wir erhalten eine Zufallsgröße T_n mit der Verteilung Q_n wie folgt: Ist I_n eine Zufallsvariable mit der Verteilung μ_n, so interpretieren wir I_n als die Knotenzahl des linken Teilbaums von T_n. Dadurch ist die Zahl der Knoten im rechten Teilbaum auf $n-1-I_n$ festgelegt. Dazu fordern wir, dass, bedingt unter I_n, linker und rechter Teilbaum von T_n unabhängig sein sollen (vgl. Abschnitt 3.5). Außerdem habe der linke Teilbaum die Verteilung Q_{I_n}, der rechte die Verteilung Q_{n-1-I_n}. Da I_n und $n-1-I_n$ stets echt kleiner als n sind, reicht nun eine Anfangsbedingung um (Q_n) vollständig festzulegen. Dazu setzen wir – vollkommen kanonisch – Q_0 als auf den leeren Baum konzentrierte Verteilung. Zur Veranschaulichung siehe auch Abbildung 4.4.

Die oben beschriebene Rekursion können wir formalisieren. Ähnlich wie aus der Summe von unabhängigen Zufallsvariablen auf Verteilungsebene die Faltung wird, schreiben wir $\circ$ für die Verknüpfung von zwei binären Bäumen zu einem neuen Baum. An einen neuen Wurzelknoten wird der links vom Operator stehende Baum links, der rechts stehende rechts angehängt. Es gilt beispielsweise

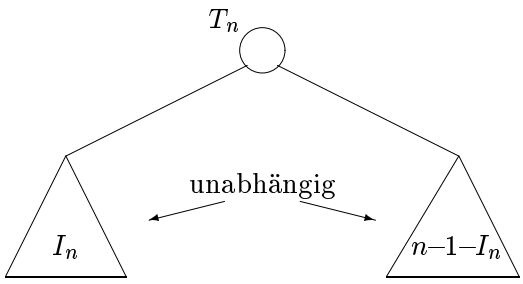

Abbildung 4.4: Veranschaulichung der Konstruktion von T_n als Q_n-verteilte Zufallsgröße

für einen binären Baum t stets

$$t \;=\; L(t) \circ R(t).$$

Sei $\circledcirc$ der entsprechende Operator auf Verteilungsebene, dann lautet die oben beschriebene Rekursionsvorschrift

$$Q_n \;=\; Q_{I_n} \;\circledcirc\; Q_{n-1-I_n}, \qquad n \in \mathbb{N},\; I_n \sim \mu_n. \tag{4.1}$$

Beispiel 4.4 (a) Ist $I_n = n-1$ fast sicher für alle $n \in \mathbb{N}$, so sind alle rechten Teilbäume leer. Damit ist Q_n auf die nach links absteigenden lineare Liste konzentriert.

(b) Ist $P(I_n = n-1) = P(I_n = 0) = 1/2$ für alle $n \in \mathbb{N}$, so erhalten wir ebenfalls eine lineare Liste. Beim Abstieg wird jedoch „per Münzwurf“ entschieden, ob es links oder rechts weitergeht.

(c) Ist I_n auf $\{0,1,\ldots,n-1\}$ gleichverteilt, so erhalten wir für Q_n die Verteilung, die der Binary Search Tree-Algorithmus erzeugt. Dieser Verteilungsfamilie wenden wir uns im nächsten Abschnitt zu.

Die Gleichheit (4.1) lässt sich direkt auf das – nun zufällige – Teilbaumgrößenprofil $(K_{n,1},K_{n,2},\ldots)$ übertragen. Die Verknüpfungsoperation $\circ$ bedeutet für das Teilbaumgrößenprofil die Addition der Profile zu $L(T_n)$ und $R(T_n)$ sowie

einem Vektor $(0,\ldots,0,1) \in \mathbb{R}^n$, der den (neuen) Wurzelknoten repräsentiert. Wir erhalten

$$K_n \;=_{\text{distr}}\; K_{I_n} \,+\, K'_{n-1-I_n} \,+\, \delta_n, \quad n \in \mathbb{N}, \tag{4.2}$$

mit $\mathcal{L}(K') = \mathcal{L}(K_n)$ für alle $n \in \mathbb{N}$, $\delta_n = (0,\ldots,0,1) \in \mathbb{R}^n$ und $I_n = \#L(T_n)$, wobei $(K_n)_{n\in\mathbb{N}}$, $(K'_n)_{n\in\mathbb{N}}$ und $(I_n)_{n\in\mathbb{N}}$ unabhängig sind. Auch diese Rekursion lässt sich (einfacher) auf der Ebene der Verteilungen ausdrücken. Schreiben wir $\nu_n = \mathcal{L}(K_n)$ so gilt

$$\nu_n \;=\; \nu_{I_n} \,\star\, \nu_{n-1-I_n} \,\star\, \delta_{(0,\ldots,0,1)}, \qquad n \in \mathbb{N}.$$

4.3 Binary Search Tree

Im Sinne dieser rekursiven Verteilungsfamilien ergibt sich nun aus der Konstruktion des BST-Baumes die Verteilung von I_n. Die Größe des linken Teilbaums ist genau die Zahl der $U_2,\ldots,U_n$, die kleiner als U_1 sind. Ist $R_{n,1}$ der absolute Rang von U_1 in $\{U_1,\ldots,U_n\}$, so gilt $I_n = R_{n,1} - 1$ und $R_{n,1}$ ist auf $\{1,\ldots,n\}$ gleichverteilt. Folglich ergibt sich für $\mathcal{L}(I_n)$ die Gleichverteilung auf $\{0,\ldots,n-1\}$. Zusammen mit der beschriebenen Unabhängigkeitsstruktur führt umgekehrt also die Verteilung von I_n zwangsläufig auf einen binären Baum, dessen Verteilung der eines vom BST-Algorithmus erzeugten Baumes entspricht. Für den Nachweis der bedingten Unabhängigkeit der Teilbäume sei hier abermals auf Abschnitt 3.5 verwiesen.

Die Rekursion (4.2) ermöglicht nun die Herleitung einer Rekursion für den Erwartungswertvektor des Teilbaumgrößenprofils. Dass dieser existiert, sieht man leicht an $K_{n1} + \ldots + K_{nn} = n < \infty$. Durch Zerlegen nach dem Wert von I_n erhält man aus (4.2) für den Vektor der Erwartungswerte

$$\begin{aligned} a_n := \mathbb{E}K_n \;&=\; \mathbb{E}K_{I_n-1} + \mathbb{E}K'_{n-I_n} + \delta_n \\ &= \sum_{i=1}^{n} P(I_n = i)\big[\mathbb{E}K_{i-1} + \mathbb{E}K'_{n-i}\big] + \delta_n \end{aligned}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} a_{i-1} + \sum_{i=1}^{n} a_{n-i}\right) + \delta_n$$
$$= \frac{2}{n}\sum_{i=1}^{n} a_{i-1} + \delta_n. \tag{4.3}$$

Mit vollständiger Induktion können wir daraus ableiten, dass mit $n \in \mathbb{N}$ gilt $\mathbb{E}K_{nn} = 1$ und, für $j < n$,

$$\mathbb{E}K_{nj} \;=\; \frac{2(n+1)}{(j+1)(j+2)}, \quad 1 \le j \le n-1; \tag{4.4}$$

dabei ist wegen $K_{nn} = 1$, $n \in \mathbb{N}$, fast sicher die erste Aussage klar. Dies ist für $n = 1$ auch der Induktionsanfang. Sei nun $1 \le j \le n-1$. Nach Induktionsvoraussetzung und (4.3) gilt also

$$\begin{aligned} a_{nj} &= \frac{2}{n}\sum_{i=1}^{n} a_{i-1,j} + \delta_{nj} \\ &= \frac{2}{n}\left(a_{jj} + \sum_{i=j+1}^{n-1} a_{ij}\right) \\ &= \frac{2}{n}\left(1 + \sum_{i=j+1}^{n-1} \frac{2(i+1)}{(j+1)(j+2)}\right) \\ &= \frac{2}{n}\left(1 + \frac{2}{(j+1)(j+2)}\left[\frac{n(n+1)}{2} - \frac{(j+1)(j+2)}{2}\right]\right) \\ &= \frac{2(n+1)}{(j+1)(j+2)}, \end{aligned}$$

und dies schließt die Induktion.

Aus diesen Erwartungwerten (4.4) lassen sich bereits verschiedene asymptotische Eigenschaften des Teilbaumgrößenprofils ablesen. Zum Beispiel erhalten wir für eine Folge $j(n) < n$ mit $j(n)/\sqrt{n} \to \infty$

$$EK_{n,j(n)} \;=\; \frac{2n+2}{j(n)^2 + 3j(n) + 2} \;\to\; 0 \quad \text{mit } n \to \infty.$$

Da es sich ausschließlich um nichtnegative Zufallsvariablen handelt, impliziert dies

$$K_{n,j(n)} \xrightarrow{\mathrm{p}} 0 \quad \text{mit } n \to \infty. \tag{4.5}$$

Dabei bezeichnet $\xrightarrow{\mathrm{p}}$ die Konvergenz in Wahrscheinlichkeit. Feng, Mahmoud und Panholzer haben dies in [FMP08] mit Hilfe von erzeugenden Funktionen gezeigt.

Obige Aussage gilt jedoch nicht in „kumulierter“ Form: Betrachtet man die Anzahl der Teilbäume mit einer proportional zu n wachsenden Mindestanzahl von Knoten, d.h. für festes $\alpha \in (0,1)$ jene Teilbäume, die mehr als αn Knoten besitzen, so erhält man mit $X_{n,\alpha} := \sum_{j=\lceil \alpha n \rceil}^{n} K_{nj}$

$$\begin{aligned} EX_{n,\alpha} &= 1 + \sum_{j=\lceil \alpha n \rceil}^{n-1} \frac{2(n+1)}{(j+1)(j+2)} \\ &= 1 + 2(n+1) \sum_{j=\lceil \alpha n \rceil}^{n-1} \Big(\frac{1}{j+1} - \frac{1}{j+2} \Big) \\ &= 1 + 2(n+1) \Big(\frac{1}{\lceil \alpha n \rceil + 1} - \frac{1}{n+1} \Big) \;=\; \frac{2(n+1)}{\lceil \alpha n \rceil + 1} - 1. \end{aligned}$$

Also folgt, mit $n \to \infty$, $EX_{n,\alpha} \to 2/\alpha - 1 > 1$. Die Untersuchung der Asymptotik der Verteilung von $X_{n,\alpha}$ ist Gegenstand von Abschnitt 4.4. Den stochastischen Prozess $\alpha \mapsto X_{n,\alpha}$ bezeichnen wir im Weiteren auch als *Eisberg-Prozess*.

Natürlich liefert Gleichung (4.4) auch die Erwartungswerte von Wiener-Index und interner wie externer Pfadlänge. Wir nennen H_n, $n \in \mathbb{N}$, die n-te harmonische Zahl, also $H_n := \sum_{j=1}^{n} 1/j$. Mit $(W_n)_{n\in\mathbb{N}}$ als Folge der Wiener Indizes, $(P_n)_{n\in\mathbb{N}}$ als Folge der internen Pfadlängen folgt

$$\begin{aligned} \mathbb{E}P_n &= \sum_{j=1}^{n-1} j \cdot \mathbb{E}K_{nj} \;=\; \sum_{j=1}^{n-1} j \cdot \frac{2(n+1)}{(j+1)(j+2)} \\ &= 2(n+1) \sum_{j=1}^{n-1} \Big(\frac{2}{j+2} - \frac{1}{j+1} \Big) \end{aligned}$$

$$= 2(n+1)\left[H_{n+1} - \frac{3}{2} + \frac{1}{n+1} - \frac{1}{2}\right]$$
$$= 2(n+1)H_n + 2 - 4(n+1) + 2 \;=\; 2(n+1)H_n - 4n,$$

sowie

$$\mathbb{E}W_n \;=\; n \cdot \mathbb{E}P_n - \sum_{j=1}^{n-1} j^2 \cdot \mathbb{E}K_{nj}.$$

Mit

$$\sum_{j=1}^{n-1} j^2 \cdot \frac{1}{(j+1)(j+2)} = \sum_{j=1}^{n-1}\Big(1 + \frac{1}{j+1} - \frac{4}{j+2}\Big)$$
$$= n - 1 - 3\Big(H_{n+1} + \frac{3}{2}\Big) + \frac{1}{2} - \frac{1}{n+1}$$
$$= n - 3H_{n+1} + 4 - \frac{1}{n+1}$$

folgt also

$$\mathbb{E}W_n = 2n(n+1)H_n - 4n^2 - 2n(n+1) + 6(n+1)H_{n+1} - 8(n+1) + 2$$
$$= 2n^2 H_n + 8nH_n + 6H_n - 6n^2 - 10n.$$

Diese Erwartungswerte wurden bereits von Neininger in [Nei02] ermittelt. Er verwendet dazu eine Rekursion, die direkt auf Wiener-Index und interne Pfadlänge eingeht.

4.3.1 Asymptotik des Anfangsstückes (Mäuse)

Wir wollen nun zeigen, dass, für festes $d \in \mathbb{N}$, das Anfangsstück $(K_{n1}, \ldots, K_{nd})$ des Teilbaumgrößenprofils mit wachsendem n asymptotisch mehrdimensional normalverteilt ist. Aufgrund der rekursiven Struktur des Profils bietet sich dafür die Kontraktionsmethode an. Diese Methode wurde bei der Analyse von Algorithmen erstmals 1991 von Rösler im Zusammenhang mit QUICKSORT in [Rös91] angewendet und hat sich seither als wichtiges Werkzeug in diesem Gebiet etabliert. Sie wurde von Rösler selbst wie auch von Rachev und Rüschendorf unabhängig weiterentwickelt und 2001 von Neininger in [Nei01] für

multivariate Asymptotik erweitert. Ausgehend von einer Rekursionsgleichung wie (4.2) erhält man nach geeigneter Normierung durch den Grenzübergang $n \to \infty$ eine Rekursionsgleichung für die Grenzverteilung. Liegt hierbei eine Kontraktion vor, so ist deren Fixpunkt nach dem Banachschen Fixpunktsatz unter gewissen Voraussetzungen eindeutig.

In der Form, wie Rösler die Kontraktionsmethode entwirft, ist diese allerdings nicht geeignet, Konvergenz gegen Normalverteilungen nachzuweisen; denn die Fixpunktgleichung, die sich im Limes ergibt, ist bezüglich der von Rösler verwendeten L_2-Metrik keine Kontraktion. Auch Neininger verwendet in [Nei01] noch ausschließlich die Konvergenz in L_2.

Ein Ausweg aus diesem Dilemma ist die Wahl einer anderen Metrik. Hierzu legten Rachev und Rüschendorf 1995 den Grundstein, vgl. [RR95]. In [NR04] findet sich schließlich das Resultat, welches wir hier verwenden wollen.

Sei nun $Y_n := (K_{n1}, \dots, K_{nd})$, $d \in \mathbb{N}$ weiterhin fest. Dann gilt wie oben

$$Y_n \;=_{\text{distr}}\; Y_{I_n} + Y'_{n-1-I_n} + b_n, \quad n \in \mathbb{N}. \tag{4.6}$$

Hier ist wieder (Y'_n) eine unabhängige Kopie von (Y_n) und $b_n = (\delta_{1n}, \dots, \delta_{dn})$. Für die Normierung suchen wir zusätzlich zum bereits bekannten Erwartungswertvektor die Kovarianzmatrix.

Lemma 4.5 Es gibt ein $\Lambda \in \mathbb{R}^{d\times d}$ so, dass für alle $n \geq 2d+2$ gilt

$$\Sigma_n \;:=\; \big(\operatorname{cov}(Y_{nk}, Y_{nj})\big)_{k,j=1,\dots,d} \;=\; (n+1)\cdot\Lambda.$$

Beweis. Sei $k \geq j$ und $h_n := \mathbb{E}Y_{nk}Y_{nj}$. Dann liefert die Rekursion (4.6)

$$\begin{aligned} h_n &= \mathbb{E}\big(Y_{I_n,k} + Y'_{n-1-I_n,k} + \delta_{nk}\big)\big(Y_{I_n,j} + Y'_{n-1-I_n,j} + \delta_{nj}\big) \\ &= \frac{1}{n}\sum_{i=0}^{n-1} \mathbb{E}\big[Y_{i,k}Y_{i,j} + Y_{i,k}Y'_{n-1-i,j} + Y_{i,k}\delta_{nj} \\ &\qquad + Y'_{n-1-i,k}Y_{i,j} + Y'_{n-1-i,k}Y'_{n-1-i,j} + Y'_{n-1-i,k}\delta_{nj} \\ &\qquad + \delta_{nk}Y_{i,j} + \delta_{nk}Y'_{n-1-i,j} + \delta_{nk}\delta_{nj}\big] \end{aligned}$$

$$= \frac{2}{n}\sum_{i=0}^{n-1} h_i \;+\; \frac{2}{n}\sum_{i=0}^{n-1} \mathbb{E}Y_{i,k}Y'_{n-1-i,j}$$
$$+ \delta_{nj}\frac{2}{n}\sum_{i=0}^{n-1}\mathbb{E}Y_{i,k} \;+\; \delta_{nk}\frac{2}{n}\sum_{i=0}^{n-1}\mathbb{E}Y_{i,j} \;+\; \delta_{nk}\delta_{nj}.$$

Der erste Summand in der letzten Zeile ist Null, da für $n = j$ wegen $k \geq j$ alle $Y_{i,k} = 0$ sind für $i = 0, \ldots, j-1$. Wegen der Unabhängigkeit von Y und Y' folgt weiter

$$nh_n \;=\; 2\sum_{i=0}^{n-1} h_i \;+\; 2\sum_{i=0}^{n-1} a_{i,k}a_{n-1-i,j} \;+\; 2\delta_{nk}\sum_{i=0}^{n-1} a_{i,j} \;+\; n\delta_{nk}\delta_{nj}.$$

Subtraktion von $(n-1)h_{n-1}$ und Division durch n liefern also

$$h_n = \frac{n+1}{n}h_{n-1} \;+\; \frac{2}{n}\sum_{i=0}^{n-2} a_{i,k}\big(a_{n-1-i,j} - a_{n-2-i,j}\big)$$
$$+ \frac{2}{n}(\delta_{nk} - \delta_{n-1,k})\sum_{i=0}^{k-1} a_{i,j} \;+\; \delta_{nk}\delta_{nj} \;-\; \frac{n-1}{n}\delta_{n-1,k}\delta_{n-1,j}.$$

Damit folgt

$$\begin{aligned}
\sigma_{kj}^{(n)} &= h_n - a_{nk}a_{nj} \\
&= \frac{n+1}{n}\big(h_{n-1} - a_{n-1,k}a_{n-1,j}\big) + \frac{n+1}{n}a_{n-1,k}a_{n-1,j} - a_{nk}a_{nj} \\
&\quad + \frac{2}{n}\sum_{i=0}^{n-2} a_{i,k}\big(a_{n-1-i,j} - a_{n-2-i,j}\big) \\
&\quad + \frac{2}{n}(\delta_{nk} - \delta_{n-1,k})\sum_{i=0}^{k-1} a_{i,j} + \delta_{nk}\delta_{nj} - \frac{n-1}{n}\delta_{n-1,k}\delta_{n-1,j} \\
&= \frac{n+1}{n}\sigma_{kj}^{(n-1)} + r_{kj}^{(n)} \qquad (4.7)
\end{aligned}$$

für ein geeignetes $R_n = (r_{kj}^{(n)})_{k,j=1,\ldots,d} \in \mathbb{R}^{d\times d}$. Wir betrachten dieses R_n etwas

genauer:

$$\begin{aligned} r_{kj}^{(n)} &= \frac{n+1}{n} a_{n-1,k} a_{n-1,j} - a_{nk} a_{nj} \\ &\quad + \frac{2}{n} \sum_{i=0}^{n-2} a_{i,k} (a_{n-1-i,j} - a_{n-2-i,j}) + \tilde{r}_{kj}^{(n)}, \end{aligned} \tag{4.8}$$

wobei $\tilde{r} = 0$ wegen $n > 2d$ und $k \leq d$. Die Summe lässt sich mit Hilfe der geschlossenen Form (4.4) für den Erwartungswert des Teilbaumgrößenprofils zerlegen. Dabei ist

$$a_{i,k} = \begin{cases} 0 & \text{für } i < k \\ 1 & \text{für } i = k \\ \frac{2(i+1)}{(k+1)(k+2)} & \text{für } i > k. \end{cases}$$

und

$$a_{n-1-i,j} - a_{n-2-i,j} = \begin{cases} 0 & \text{für } i > n-1-j \\ 1 & \text{für } i = n-1-j \\ \frac{2}{j+1} - 1 & \text{für } i = n-2-j \\ \frac{2}{(j+1)(j+2)} & \text{für } i < n-2-j. \end{cases}$$

und es folgt

$$\begin{aligned} &\sum_{i=0}^{n-2} a_{i,k} (a_{n-1-i,j} - a_{n-2-i,j}) \\ &= \sum_{i=k+1}^{n-3-j} \frac{4(i+1)}{(k+1)(k+2)(j+1)(j+2)} \\ &\qquad + \frac{2}{(j+1)(j+2)} + \frac{2(n-j-1)}{(k+1)(k+2)} \Big(\frac{2}{j+1} - 1 \Big) + \frac{2(n-j)}{(k+1)(k+2)}, \end{aligned}$$

wobei die letzten drei Summanden diejenigen für $i = k$, $i = n-2-j$ und $i = n-1-j$ sind. Die Summe darf Null werden, falls sie keinen Summanden besitzt, also $k+1 = n-3-j+1$ gilt, d.h. $k+j = n-3$. Für kleinere n wird die Rechnung an dieser Stelle falsch. Wegen $k,j \leq d$ führt dieser Schritt auf die Voraussetzung $n \geq 2d+2$ im Lemma.

Wir kürzen $g(k,j) := ((k+1)(k+2)(j+1)(j+2))^{-1}$ ab und erhalten

$$\begin{aligned}
\ldots &= g(k,j) \cdot \Big[2\big((n-j-2)(n-j-1) - (k+1)(k+2)\big) + 2(k+1)(k+2) \\
&\qquad + 4(n-j-1)(j+1) - 2(n-j-1)(j+1)(j+2) \\
&\qquad + 2(n-j)(j+1)(j+2)\Big] \\
&= g(k,j)\Big[2(n-j-1)(n-j-2) + 4(n-j-1)(j+2) + 2(j+1)(j+2)\Big] \\
&= g(k,j) \cdot 2n(n+1).
\end{aligned}$$

Für die ersten beiden Summanden in (4.8) gilt schließlich

$$\begin{aligned}
\frac{n+1}{n} a_{n-1,k} a_{n-1,j} - a_{nk} a_{nj} &= \frac{n+1}{n} \cdot \frac{2n}{(k+1)(k+2)} \cdot \frac{2n}{(j+1)(j+2)} \\
&\quad - \frac{2(n+1)}{(j+1)(j+2)} \cdot \frac{2(n+1)}{(k+1)(k+2)} \\
&= g(k,j) \cdot \Big[4n(n+1) - 4(n+1)^2\Big] \\
&= -4(n+1) g(k,j),
\end{aligned}$$

also folgt

$$r_{k,j}^{(n)} \;=\; g(k,j) \cdot \Big[-4(n+1) + \tfrac{2}{n} \cdot 2n(n+1)\Big] \;=\; 0 \quad \text{für } n \geq 2d+2.$$

Damit folgt die Behauptung; denn wählt man beispielsweise

$$\Lambda \;:=\; \frac{1}{2d+3} \Sigma_{2d+2},$$

(dies zeigt die Behauptung für $n = 2d+2$) so folgt aus (4.7) für $n \geq 2d+3$

$$\sigma_{kj}^{(n)} \;=\; \frac{n+1}{2d+3} \sigma_{kj}^{(2d+2)} + \sum_{i=2d+3}^{n} \frac{n+1}{i+1} \cdot r_{kj}^{(i)} \;=\; (n+1) \cdot \Lambda_{kj}. \qquad \square$$

Sei nun $\mathbb{M}_3^d$ die Menge der Wahrscheinlichkeitsmaße μ auf $\mathbb{R}^d$ mit endlichem dritten Moment, also

$$\int |x|^3 \, \mu(dx) \;<\; \infty.$$

Wir betrachten die Transformation

$$T : \mathbb{M}_3^d \to \mathbb{M}_3^d, \qquad \mu \mapsto \mathcal{L}\big(\sqrt{U} \cdot X + \sqrt{1-U} \cdot X'\big),$$

mit $\mathcal{L}(X) = \mathcal{L}(X') = \mu$, $U \sim \text{unif}(0,1)$ und U,X,X' unabhängig. Als Abbildung auf $\mathbb{M}_3^d$ ist T jedoch keine Kontraktion; denn da beispielsweise alle Normalverteilungen $\mu = \mathrm{N}_d(0,\Sigma)$ für beliebige Kovarianzmatrix Σ die Gleichung $T\mu = \mu$ erfüllen (dies ist ein Vorgriff auf den Beweis zu Satz 4.9), gibt es mehr als einen Fixpunkt.

Wir können jedoch die Menge der in Frage kommenden Wahrscheinlichkeitsmaße, d.h. den Definitions- und Wertebereich von T so einschränken, dass T eine Kontraktion bzgl. der Zolotarev-Metrik ζ ist (vgl. [RR95, S. 264]) und einen eindeutigen Fixpunkt besitzt. Für

$$\mathbb{M}_3^d(b,\Sigma) := \{\mu \in \mathbb{M}_3^d : \mathbb{E}X = b, \operatorname{cov}(X) = \Sigma \text{ für } X \sim \mu\}$$

ist die Einschränkung von T auf $\mathbb{M}_3^d(0,\Sigma)$

$$\tilde{T} : \mathbb{M}_3^d \to \mathbb{M}_3^d, \qquad \mu \mapsto \tilde{T}(\mu)$$

wohldefiniert.

Wir transformieren also Y_n so, dass die transformierte Verteilung aus $\mathbb{M}_3^d$ ist. Dazu sei $X_n := \Sigma_n^{-1/2}(Y_n - a_n)$, mit $a_n \in \mathbb{R}^d$ der Erwartungswertvektor zu Y_n. (Zum linearalgebraischen Hintergrund insbesondere symmetrischer und positiv definiter Matrizen verweisen wir auf [Str80, Kap. 6].) Wir schreiben die Rekursion (4.6) für Y_n zu einer Rekursion für X_n um:

$$\begin{aligned} X_n &=_{\text{distr}} \big(\Sigma_n^{-1/2}\Sigma_{I_n}^{1/2}\big)X_{I_n} + \big(\Sigma_n^{-1/2}\Sigma_{n-1-I_n}^{1/2}\big)X'_{n-1-I_n} \\ &\qquad + \Sigma_n^{-1/2}\big(a_{I_n} + a_{n-1-I_n} - a_n\big) \\ &= A_n^1 X_{I_n} + A_n^2 X'_{n-1-I_n} + v_n, \end{aligned}$$

mit $A_n^1 = \Sigma_n^{-1/2}\Sigma_{I_n}^{1/2}$, $A_n^2 = \Sigma_n^{-1/2}\Sigma_{n-1-I_n}^{1/2}$ und $v_n = \Sigma_n^{-1/2}(a_{I_n} + a_{n-1-I_n} - a_n)$. Lässt man nun informell $n \to \infty$ gehen, so konvergieren A_n^1 und A_n^2 gegen $\sqrt{U}\mathrm{Id}_d$ und $\sqrt{1-U}\mathrm{Id}_d$ und man erhält die Verteilungsgleichung

$$X_\infty =_{\text{distr}} \sqrt{U} \cdot X_\infty + \sqrt{1-U} \cdot X'_\infty,$$

wobei U, X_∞, X'_∞ unabhängig sind mit $\mathcal{L}(X'_\infty) = \mathcal{L}(X_\infty)$. Aus der Matrizenmultiplikation wird hier ein Skalarprodukt, da nur Einheitsmatrizen beteiligt sind. Dies ist genau die Formulierung des Fixpunktes der oben definierten Abbildung $\tilde{T}$.

Damit dieser Schluss über die heuristische Argumentation hinaus Gültigkeit erlangen kann, führen wir Theorem 4.1 aus [NR04] an. Wir formulieren einen Spezialfall[1] dieses Theorems in unserer Notation

Satz 4.6 (Neininger & Rüschendorf) Sei (X_n) definiert wie oben mit existierendem dritten Moment. Wir nehmen an

$$\big(A_1^{(n)},A_2^{(n)},b^{(n)}\big) \xrightarrow{L_3} \big(A_1^*, A_2^*, b^*\big), \tag{4.9}$$

$$\mathbb{E}\Big(\|A_1^*\|_{\mathrm{op}}^3 + \|A_2^*\|_{\mathrm{op}}^3 \Big) < 1 \tag{4.10}$$

$$\begin{aligned} \mathbb{E}1_{\{I_n \le k\} \cup \{I_n = n\}} \|A_1^{(n)}\|_{\mathrm{op}}^3 &\to 0, \quad \text{und} \\ \mathbb{E}1_{\{n-1-I_n \le k\} \cup \{n-1-I_n = n\}} \|A_2^{(n)}\|_{\mathrm{op}}^3 &\to 0, \quad n \to \infty, \end{aligned} \tag{4.11}$$

für alle $k \in \mathbb{N}$. Dann konvergiert X_n gegen X in dem Sinne, dass

$$\zeta_3(X_n,X) \to 0 \quad \text{mit } n \to \infty,$$

wobei $\mathcal{L}(X) \in \mathbb{M}_3^d(0,\mathrm{Id}_d)$ gegeben ist als der eindeutige Fixpunkt der Verteilungsgleichung

$$X =_{\mathrm{distr}} A_1^* X + A_2^* X' + b^*,$$

mit $(A_1^*,A_2^*,b^*),X,X'$ unabhängig und $\mathcal{L}(X) = \mathcal{L}(X')$.

Die Voraussetzungen für die Konvergenzaussage dieses Satzes weisen wir im Folgenden nach. Zunächst zu (4.11):

Lemma 4.7 Für beliebiges $k \in \mathbb{N}$ gilt mit $n \to \infty$

$$\mathbb{E}\big[1_{\{I_n \le k\}} \|A_n^1\|_{\mathrm{op}}^3 + 1_{\{n-1-I_n \le k\}} \|A_n^2\|_{\mathrm{op}}^3\big] \to 0, \tag{4.12}$$

[1] Je nach den spezifischen Eigenschaften der Rekursion kann man die „3“ in $\mathbb{M}_3^d$ zu einem s, $1 < s \le 3$ verallgemeinern. Hier beschränken wir uns auf den Fall $s = 3$. Außerdem kann die rechte Seite der Rekursion aus mehr als zwei Summanden bestehen – auch hier beschränken wir uns auf zwei.

Beweis. Sei $k \in \mathbb{N}$ beliebig. Dann ist

$$C_1 \;:=\; \max\Big\{ \Big\| \Sigma_1^{1/2} \Big\|_{\mathrm{op}}^3, \Big\| \big(\tfrac{1}{2}\Sigma_2\big)^{1/2} \Big\|_{\mathrm{op}}^3, \ldots, \Big\| \big(\tfrac{1}{k}\Sigma_k\big)^{1/2} \Big\|_{\mathrm{op}}^3 \Big\}$$

endlich und von n unabhängig. Ist n groß genug, so ist $(\frac{1}{n+1}\Sigma_n)^{-1/2} = \Lambda^{-1/2}$ und $C_2 := \|\Lambda^{-1/2}\|_{\mathrm{op}}$ ist ebenfalls von n unabhängig. Es gilt also für $C := C_1^3 \cdot C_2^3$

$$\begin{aligned}
&\mathbb{E}\big[1_{\{I_n \le k\}} \|A_n^1\|_{\mathrm{op}}^3\big] \\
&= \mathbb{E}\Big[1_{\{I_n \le k\}} \Big\| \sqrt{\frac{I_n+1}{n+1}} \cdot \Big(\frac{1}{n+1}\Sigma_n\Big)^{-1/2} \Big(\frac{1}{I_n+1}\Sigma_{I_n}\Big)^{1/2} \Big\|_{\mathrm{op}}^3\Big] \\
&\le \mathbb{E}\Big[1_{\{I_n \le k\}} \sqrt{\frac{I_n+1}{n+1}} \cdot \Big\|\Lambda^{-1/2}\Big\|_{\mathrm{op}}^3 \cdot \Big\| \Big(\frac{1}{I_n+1}\Sigma_{I_n}\Big)^{1/2} \Big\|_{\mathrm{op}}^3\Big] \\
&\le C \cdot \sqrt{\frac{k+1}{n+1}} \cdot \mathbb{E}1_{\{I_n \le k\}} \;=\; C \cdot \Big(\frac{k+1}{n+1}\Big)^{3/2} \;\to\; 0.
\end{aligned}$$

Die Konvergenz des zweiten Summanden folgt analog. □

Außerdem müssen mit $n \to \infty$ auch die beteiligten Skalierungsmatrizen in L_3 konvergieren, vgl. (4.9):

Lemma 4.8 Mit $n \to \infty$ gilt bezüglich der Konvergenz in L_3

$$(A_n^1, A_n^2) \to (\sqrt{U}\mathrm{Id}_d, \sqrt{1-U}\mathrm{Id}_d) \quad \text{und} \quad v_n \to 0.$$

Beweis. Zu zeigen ist

$$\mathbb{E}\|A_n^1 - \sqrt{U}\mathrm{Id}_d\|_{\mathrm{op}}^3 \to 0 \quad \text{mit } n \to \infty.$$

Wegen $(I_n/n, (n-1-I_n)/n) \to (U, 1-U)$ folgt daraus bereits der erste Teil der Behauptung. Sei $U \sim \mathrm{unif}(0,1)$ und $I_n := \lfloor nU \rfloor$, dann ist $I_n \sim \mathrm{unif}\{0, \ldots, n-1\}$ und $(I_n+1)/(n+1)$ konvergiert mit $n \to \infty$ fast sicher gegen U, insbesondere folgt aufgrund der Beschränktheit von I_n/n und U, dass

$\mathbb{E}\|\sqrt{(I_n+1)/(n+1)}\mathrm{Id}_d - \sqrt{U}\mathrm{Id}_d\|_{\mathrm{op}}^3 \to 0$. Nach Dreiecksungleichung reicht also zu zeigen

$$\mathbb{E}\Big\|\,\Sigma_n^{-1/2}\cdot\Sigma_{I_n}^{1/2} - \sqrt{\frac{I_n+1}{n+1}}\cdot\mathrm{Id}_d\Big\|_{\mathrm{op}}^3 \to 0 \quad \text{ mit } n\to\infty.$$

Wir zerlegen nach dem Wert von I_n: Für $I_n > 2d+1$ gilt nämlich nach Lemma 4.5

$$\Sigma_n^{-1/2}\cdot\Sigma_{I_n}^{1/2} \;=\; \sqrt{\frac{I_n+1}{n+1}}\cdot\Big(\underbrace{\frac{1}{n+1}\Sigma_n}_{=\Lambda}\Big)^{-1/2}\Big(\underbrace{\frac{1}{I_n+1}\Sigma_{I_n}}_{=\Lambda}\Big)^{1/2} \;=\; \sqrt{\frac{I_n+1}{n+1}}\cdot\mathrm{Id}_d,$$

und für die Konvergenz ist nichts zu zeigen. Für die Fälle $I_n \le 2d+1$ verwenden wir wieder die Dreiecksungleichung und dann (4.12) mit $k = 2d+1$

$$\begin{aligned}&\mathbb{E}1_{\{I_n\le 2d+1\}}\|A_n^1 - \sqrt{(I_n+1)/(n+1)}\cdot\mathrm{Id}_d\|_{\mathrm{op}}^3\\ &\qquad\le\; \mathbb{E}1_{\{I_n\le 2d+1\}}\|A_n^1\|_{\mathrm{op}}^3 \;+\; \mathbb{E}1_{\{I_n\le 2d+1\}}\|\sqrt{(I_n+1)/(n+1)}\mathrm{Id}_d\|_{\mathrm{op}}^3\end{aligned}$$

und beide Summanden gehen gegen Null.

Zuletzt bleibt noch $v_n \to 0$ in L_3 zu zeigen. Bei $n > 2d+1$ ist $\{I_n \le d\}\cap\{n-1-I_n \le d\}$ fast sicher leer und für $1\le j\le d$ gilt also

$$\begin{aligned}&\mathbb{E}\Big|\,\big(\Sigma_n^{1/2}v_n\big)_j\Big|^3 \;=\; \mathbb{E}|a_{I_n,j} + a_{n-1-I_n,j} - a_{nj}|^3\\
&=\; \mathbb{E}1_{\{I_n>d,\ n-1-I_n>d\}}\left|\frac{2(I_n+1)}{(j+1)(j+2)} + \frac{2(n-I_n)}{(j+1)(j+2)} - \frac{2(n+1)}{(j+1)(j+2)}\right|^3\\
&\quad + \mathbb{E}1_{\{I_n\le d,\ n-1-I_n>d\}}\left|\delta_{jd}\delta_{I_n,d} + \frac{2(n-I_n)}{(j+1)(j+2)} - \frac{2(n+1)}{(j+1)(j+2)}\right|^3\\
&\quad + \mathbb{E}1_{\{I_n>d,\ n-1-I_n\le d\}}\left|\frac{2(I_n+1)}{(j+1)(j+2)} + \delta_{jd}\delta_{n-1-I_n,d} - \frac{2(n+1)}{(j+1)(j+2)}\right|^3\\
&\le\; 0 + \Big(\frac{2(d+1)}{(j+1)(j+2)}\Big)^3\cdot\frac{d+1}{n} + \Big(\frac{2(d+1)}{(j+1)(j+2)}\Big)^3\cdot\frac{d+1}{n} \;\to\; 0.\end{aligned}$$

Insbesondere ist also $\|\Sigma_n^{1/2}v_n\|^3$ beschränkt und für $n > 2d$ gilt

$$\mathbb{E}\|v_n\|^3 \;=\; (n+1)^{-3/2}\cdot\mathbb{E}\|\Lambda^{-1/2}\Sigma_n^{1/2}v_n\|^3 \to 0. \qquad \square$$

Schließlich bleibt noch die Kontraktionsbedingung. Es gilt

$$\mathbb{E}\sqrt{U}^3 + \mathbb{E}\sqrt{1-U}^3 \;=\; \frac{4}{5} \;<\; 1.$$

Dies zeigt (4.10), da $A_1^* = \sqrt{U}\mathrm{Id}_d$ und $A_2^* = \sqrt{1-U}\mathrm{Id}_d$.

Bei vielen Anwendungen der Kontraktionsmethode erhält man zwar Konvergenz gegen eine Grenzverteilung – diese ist aber nur implizit als Fixpunkt einer Transformation gegeben. Im vorliegenden Fall ist es jedoch leicht möglich, diese Verteilung explizit anzugeben:

Satz 4.9 Mit den obigen Bezeichnungen und $n \to \infty$ gilt

$$\mathcal{L}\big(\Sigma_n^{-1/2}(Y_n - a_n)\big) \;\xrightarrow{\mathrm{w}}\; \mathrm{N}_d(0,\mathrm{Id}_d).$$

Beweis. Wir rechnen zunächst nach, dass $\mathrm{N}_d(0,\mathrm{Id}_d) \in \mathbb{M}_3^d(0,\mathrm{Id}_d)$ Fixpunkt von T ist: Bedingt unter $U = u \in (0,1)$ gilt

$$\begin{aligned}\mathcal{L}\big(X|U=u\big) &= \mathcal{L}\big(\sqrt{u}\cdot X' + \sqrt{1-u}\cdot X''|U=u\big)\\ &= \mathcal{L}\big(\sqrt{u}\cdot X'\big) \star \mathcal{L}\big(\sqrt{1-u}\cdot X''\big)\\ &= \mathrm{N}(0,u\cdot\mathrm{Id}_d) \star \mathrm{N}(0,(1-u)\cdot\mathrm{Id}_d) \;=\; \mathrm{N}(0,\mathrm{Id}_d)\end{aligned}$$

und dies hängt nicht von u ab; damit ist $\mathrm{N}_d(0,\mathrm{Id}_d)$ tatsächlich der gesuchte Fixpunkt von T.

Schließlich liefert Theorem 4.1 in [NR04] die Konvergenz von X_n gegen den Fixpunkt von T bezüglich der Zolotarev-Metrik, diese wiederum impliziert die behauptete schwache Konvergenz. □

Innerhalb des vorhergegangenen Beweises haben wir gesehen, dass die Folge der Σ_n ab einem (von d abhängenden) n_0 konstant bleibt. Daher ergibt sich das folgende

Korollar 4.10 Mit $n \to \infty$ und Λ_d definiert wie oben gilt

$$\mathcal{L}\Big(\frac{Y_n - a_n}{\sqrt{n}}\Big) \;\xrightarrow{\mathrm{w}}\; \mathrm{N}_d(0,\Lambda_d).$$

Bemerkung 4.11 Die Matrizen $\Lambda_1, \Lambda_2, \ldots$ sind konsistent in dem Sinne, dass es eine unendliche Matrix Λ gibt, so dass für alle $d \in \mathbb{N}$ die linke obere $d \times d$-Teilmatrix von Λ gerade Λ_d ist. Die Berechnung ist rekursiv mittels (4.7) möglich. Für $d = 4$ ergibt sich

$$\Lambda_4 \;=\; \begin{pmatrix} \frac{2}{45} & -\frac{1}{45} & -\frac{1}{105} & -\frac{1}{210} \\ -\frac{1}{45} & \frac{23}{420} & -\frac{11}{840} & -\frac{13}{1890} \\ -\frac{1}{105} & -\frac{11}{840} & \frac{24}{525} & -\frac{7}{900} \\ -\frac{1}{210} & -\frac{13}{1890} & -\frac{7}{900} & \frac{2}{55} \end{pmatrix}.$$

Wir haben nun für beliebiges $k \in \mathbb{N}$ die Konvergenz des Anfangsstückes des Teilbaumgrößenprofils gegen eine mehrdimensionale Normalverteilung nachgewiesen. Dieses Resultat ermöglicht insbesondere Aussagen über die asymptotische Unabhängigkeitsstruktur der Komponenten des Teilbaumgrößenprofils bzw. entsprechender Linearkombinationen seiner Komponenten.

Die Untersuchung der eindimensionalen Konvergenz von einzelnen Komponenten findet sich wohl erstmals bei Devroye in [Dev91] mit einem Argument über eine Version des zentralen Grenzwertsatzes für m-abhängige Zufallsvariablen. Auch Feng, Mahmoud und Panholzer betrachten nur die Konvergenz einzelner Komponenten. Sie benutzen in [FMP08] erzeugende Funktionen und großes analytisches Werkzeug um einen „Phasenübergang“ für die Konvergenz der Verteilung von $K_{n,j(n)}$ zu beschreiben: Für $j(n)/\sqrt{n} \to 0$ ist $K_{n,j(n)}$ asymptotisch normal, für $j(n)/\sqrt{n} \to \infty$ gehen die Komponenten gegen Null. (Vergleiche die entsprechende Bemerkung auf Seite 76.) Geht nun $j(n)$ so gegen unendlich, dass $j(n)/\sqrt{n} \to c \in \mathbb{R}$, so ist $K_{n,j(n)}$ asymptotisch Poisson-verteilt mit Parameter $2/c^2$.

In [Fuc07] wird darüberhinaus nachgewiesen, dass

$$d_{\mathrm{TV}}\big(K_{n,j}, \mathrm{Po}(\mathbb{E}K_{n,j})\big) \;\to\; 0,$$

für jede Folge $j = j_n$ mit $j_n < n$ und $j_n \to \infty$. Wie in der vorher zitierten Arbeit werden auch hier erzeugende Funktionen und analytische Methoden verwendet.

4.4 Allgemeine Asymptotik für Eisberge

Zu Beginn des vorhergegangenen Abschnitts haben wir beobachtet, dass die Zahl der Knoten mit einer bestimmten Größe k gegen Null konvergiert, wenn $k = k(n)$ mit $n \to \infty$ schnell genug wächst. Wir haben außerdem gezeigt, dass dies nicht auf die gemeinsame Verteilung der Komponenten dieses Endstückes verallgemeinert werden kann.

In diesem Abschnitt widmen wir uns der Anzahl der Knoten, deren Teilbaum einen bestimmten Mindestanteil an der Gesamtknotenzahl hat.

Ziel dieses Abschnitts ist ein funktionaler Grenzwertsatz, der basierend auf einer Art Kontraktionsmethode für stochastische Prozesse Verteilungskonvergenz beweist. Dieses Resultat werden wir in Abschnitt 4.5 auf verschiedene Verteilungen auf binären Bäumen anwenden können.

4.4.1 Ein Raum von Wahrscheinlichkeitsmaßen

Sei $\mathfrak{C}$ die Menge der Funktionen von $(0,1)$ nach $\mathbb{N}_0$, die monoton fallend, stetig von rechts und ν-integrierbar sind für das Maß ν, welches durch die Lebesgue-Dichte

$$\nu(dx) \;=\; x dx, \qquad x \in (0,1),$$

gegeben ist. Auf $\mathfrak{C}$ definieren wir mittels ν den Abstand

$$\mathfrak{d}(f,g) \;:=\; \int |f-g|\, d\nu \;=\; \int_0^1 t|f(t)-g(t)|\, dt,$$

für $f,g \in \mathfrak{C}$.

Lemma 4.12 $(\mathfrak{C},\mathfrak{d})$ ist ein separabler, vollständiger, metrischer Raum.

Beweis. Wir zeigen zunächst, dass $\mathfrak{d}$ eine Metrik auf $\mathfrak{C}$ ist. Seien also $f,g,h \in \mathfrak{C}$. Klar ist $\mathfrak{d}(g,f) = \mathfrak{d}(f,g) \geq 0 = \mathfrak{d}(f,f)$ und aus der Dreiecksungleichung für den gewöhnlichen Betrag ergibt sich sofort

$$\mathfrak{d}(f,g) \;\leq\; \mathfrak{d}(f,h) + \mathfrak{d}(h,g).$$

Seien nun $f,g \in \mathfrak{C}$ mit $\mathfrak{d}(f,g) = 0$. Dann ist $f(t) = g(t)$ für alle $t \in (0,1) \setminus N$ und N ist eine Lebesgue-Nullmenge. Wir zeigen $N = \emptyset$ und damit $f = g$. Angenommen $N \neq \emptyset$. Sei also $t \in N$ und $|f(t) - g(t)| > 0$. Mit f und g ist auch $t \mapsto |f(t) - g(t)|$ stetig von rechts, also existiert eine rechtsseitige Umgebung U_t zu t mit $|f(t) - g(t)| > 0$ für alle $t \in U_t$. Wegen $\ell(U_t) > 0$ kann aber nicht $U_t \subset N$ gelten und es wäre $\mathfrak{d}(f,g) > 0$. Widerspruch.

Nun zur Vollständigkeit: Bekannterweise ist der Maßraum (L_1, ν) insbesondere vollständig, wenn ν die Lebesguedichte $\nu(dx) = xdx$ besitzt. Dabei werden solche Funktionen, die sich nur auf einer ℓ-Nullmenge unterscheiden, durch eine Interpretation über Äquivalenzklassen miteinander identifiziert. Sei also $(f_n)_{n\in\mathbb{N}}$ eine Cauchyfolge in $(\mathfrak{C},\mathfrak{d})$ und $(\bar{f}_n)_{n\in\mathbb{N}} \subset L_1$ die Folge der zugehörigen Äquivalenzklassen. Dann existiert eine Klasse $\bar{f}$ als Grenzwert der Folge $(\bar{f}_n)$ und es gilt

$$\int |\bar{f}_n - \bar{f}|\, d\nu \;\to\; 0 \quad \text{mit } n \to \infty.$$

Daraus folgt die Existenz einer Teilfolge (f_{n_k}) mit $f_{n_k} \in \bar{f}_{n_k}$, die ν-fast überall gegen ein $\tilde{f} \in \bar{f}$ konvergiert. Da alle f_{n_k} ν-fast überall monoton fallen und den Wertebereich $\mathbb{N}_0$ haben, gilt dies auch für $\tilde{f}$. Damit ist die Menge N der Sprungstellen von $\tilde{f}$ höchstens abzählbar und mit

$$f(t) \;:=\; \begin{cases} \tilde{f}(t) & \text{falls } t \notin N, \\ \tilde{f}(t+) & \text{sonst,} \end{cases}$$

ist schließlich f stetig von rechts. Damit ist $f \in \mathfrak{C}$ und bezüglich der Metrik $\mathfrak{d}$ auch Limes der Folge (f_n):

$$\mathfrak{d}(f_n,f) = \int_0^1 t|f_n(t) - f(t)|\, dt = \int\limits_{(0,1)\setminus N} t|f_n(t) - \tilde{f}(t)|\, dt = \int |\bar{f}_n - \bar{f}|\, d\nu \to 0$$

mit $n \to \infty$. Somit ist $(\mathfrak{C},\mathfrak{d})$ vollständig.

Zuletzt weisen wir nach, dass $(\mathfrak{C},\mathfrak{d})$ separabel ist. Dazu betrachten wir die Menge

$$B \;:=\; \bigcup_{n=1}^{\infty} B_n \quad \text{mit} \quad B_n := \Big\{ \sum_{k=0}^{n} 1_{(0,a_k)} \;:\; a_0 \geq \ldots \geq a_n \in \mathbb{Q} \cap [0,1] \Big\}.$$

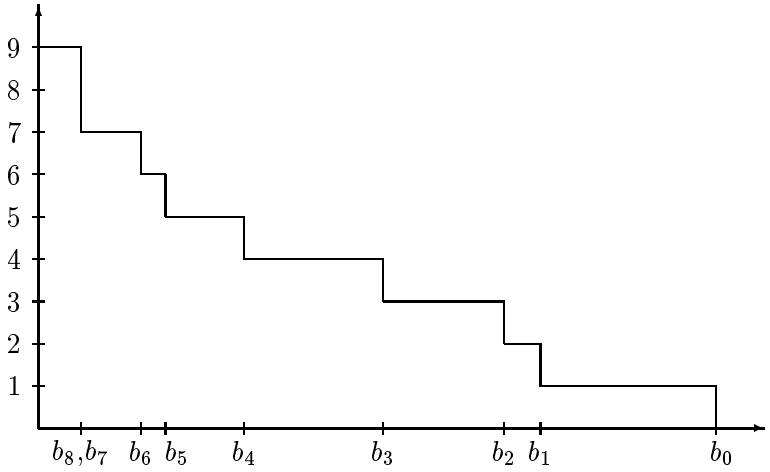

Abbildung 4.5: Veranschaulichung der Darstellung einer Funktion f aus $\mathfrak{C}$ und ihrer Sprungstellen $b_0,\ldots,b_8$.

Da jedes B_n abzählbar ist, ist auch B als abzählbare Vereinigung abzählbarer Mengen abzählbar. Wir zeigen nun, dass sich $\mathfrak{C}$ als Abschluss von B ergibt. Dabei ist $B \subset \mathfrak{C}$ leicht zu sehen, spätestens aber anhand der folgenden Konstruktion einleuchtend: Sei $f \in \mathfrak{C}$ und die Folge $(b_n)_{n\in\mathbb{N}_0}$ definiert durch

$$b_n \;:=\; \inf\{t \in [0,1] : f(t) \leq n\},$$

für $n \in \mathbb{N}_0$, dann ist

$$f \;=\; \sum_{k=0}^{\infty} 1_{(0,b_k)}.$$

Der Vorstellung behilflich ist dabei hoffentlich Abbildung 4.5.

Insbesondere erhält man auf diese Weise umgekehrt für jede monoton gegen Null fallende Folge $(b_n)_{n\in\mathbb{N}_0} \in [0,1]^{\mathbb{N}_0}$ ein (eindeutiges) $f \in \mathfrak{C}$, solange die zusätzliche Bedingung

$$\left(\int_0^1 t f(t)\,dt \;=\right) \quad \frac{1}{2}\sum_{k=0}^{\infty} b_k^2 \;<\; \infty$$

erfüllt ist, die sich aus der geforderten ν-Integrierbarkeit von f ergibt.

Wir wählen die Folgen $(b_{nk})_{k\in\mathbb{N}}$ so, dass $b_{nk} \in \mathbb{Q}$ und b_{nk} schwach isoton gegen b_n geht mit $k \to \infty$, wobei insbesondere $|b_n - b_{nk}| < 2^{-k}$ gelten soll. Dies lässt sich leicht über die jeweiligen Binärdarstellungen der b_n bewerkstelligen. Definiert man nun

$$f_n \;:=\; \sum_{k=0}^{n} 1_{(0,b_{kn})}, \quad n \in \mathbb{N},$$

so ist[2] $f_n \in B_n$, und es verbleibt zu zeigen, dass $\mathfrak{d}(f,f_n) \to 0$ mit $n \to \infty$. Wir erhalten

$$\begin{aligned}
\mathfrak{d}(f,f_n) \;&=\; \mathfrak{d}\Big(\sum_{k=0}^{\infty} 1_{(0,b_k)},\, \sum_{k=0}^{n} 1_{(0,b_{kn})}\Big) \\
&=\; \int_0^1 t\Big|\sum_{k=0}^{\infty} 1_{(0,b_k)}(t) - \sum_{k=0}^{n} 1_{(0,b_{kn})}(t)\Big|\,dt \\
&=\; \int_0^1 t\Big|\sum_{k=n+1}^{\infty} 1_{(0,b_k)}(t) \;+\; \sum_{k=0}^{n} 1_{[b_{kn},b_k)}(t)\Big|\,dt \\
&=\; \sum_{k=n+1}^{\infty} \int_0^{b_k} t\,dt \;+\; \sum_{k=0}^{n} \int_{b_{kn}}^{b_k} t\,dt \\
&=\; \frac{1}{2}\sum_{k=n+1}^{\infty} b_k^2 \;+\; \frac{1}{2}\sum_{k=0}^{n}\big(b_k^2 - b_{kn}^2\big).
\end{aligned}$$

Der erste Summand ist wegen

$$\frac{1}{2}\sum_{k=0}^{\infty} b_k^2 \;=\; \int_0^1 t\sum_{k=0}^{\infty} 1_{(0,b_k)}(t)\,dt \;=\; \int_0^1 t f(t)\,dt < \infty$$

summierbar und konvergiert daher gegen Null. Für den zweiten Summanden erhalten wir mit $b_k^2 - b_{kn}^2 = (b_k + b_{kn})(b_k - b_{kn}) \le 2 \cdot 2^{-n}$

$$\frac{1}{2}\sum_{k=0}^{n}\big(b_k^2 - b_{kn}^2\big) \;\le\; \sum_{k=0}^{n} 2^{-n} \;=\; (n+1)2^{-n} \to 0$$

mit $n \to \infty$. Es folgt die Behauptung. □

[2] Die Indizes der b_{nk} sind absichtlich „vertauscht“.

Nun zurück zur Stochastik: Sei $\mathbb{M}$ die Menge der Wahrscheinlichkeitsmaße P auf $\mathfrak{C}$ mit der Bedingung

$$\int_{\mathfrak{C}} \int_0^1 t|x(t)|\,dt\,P(dx) \;<\; \infty.$$

Als σ-Algebra betrachten wir auf Funktionenräumen grundsätzlich die von den Projektionen $t \mapsto f(t)$, $t \in (0,1)$, $f \in \mathfrak{C}$, erzeugte σ-Algebra. Auf $\mathbb{M}$ definieren wir den Abstand der Maße $Q, Q' \in \mathbb{M}$ als

$$d(Q,Q') \;:=\; \inf\{\mathbb{E}\mathfrak{d}(X,Y) \;:\; X \sim Q, Y \sim Q'\}.$$

Lemma 4.13 $(\mathbb{M}, d)$ ist ein vollständiger metrischer Raum. Konvergenz bezüglich d impliziert schwache Konvergenz.

Beweis. Der Beweis ist eine Zusammenstellung von Resultaten aus [Rac91]. Allerdings soll der besseren Nachvollziehbarkeit halber ein wenig auf die dort verwendete Terminologie eingegangen werden. Alle Bezeichnungen, die in diesem Beweis eingeführt werden, verlieren mit dem Beweisendezeichen ihre Gültigkeit oder nehmen wieder ihre alte Bedeutung an.

Zentraler Ansatzpunkt ist das *Monge-Kantorovich Mass Transference Problem*. Es soll dabei eine bestimmte Masse von einem Maßraum (U_1,P_1) auf einen Maßraum (U_2,P_2) transferiert werden, wobei eine Kostenfunktion $c : U_1 \times U_2 \to \mathbb{R}_+$ die Transportkosten festlegt. Nennt man nun $\mathscr{P}^{(P_1,P_2)}$ den Raum der Wahrscheinlichkeitsmaße auf $U_1 \times U_2$, die die Randverteilungen P_1 und P_2 besitzen, so geht es darum, die von c abhängenden Gesamtkosten

$$\kappa_c(P) := \int c(x,y)P(dx,dy), \quad P \in \mathscr{P}^{(P_1,P_2)},$$

zu minimieren.

Auf diese Weise erhalten wir auf einem separablen, metrischen Raum U im Spezialfall $U_1 = U_2 = U$ einen metrischen Raum $\mathscr{P}(U)$ von Wahrscheinlichkeitsmaßen auf U, ausgestattet mit der Metrik

$$\hat{\kappa}_c(P_1,P_2) \;:=\; \inf\{\kappa_c(P) \;:\; P \in \mathscr{P}^{(P_1,P_2)}\}.$$

Wählt man die oben auf $\mathfrak{C}$ definierte Metrik $\mathfrak{d}$ als Kostenfunktion c, so erhält man entlang der Berechnungen auf Seite 123 in [Rac91]

$$\lambda(x) \;=\; \int_0^1 t|x(t)|dt, \qquad x \in U,$$

und damit $\mathbb{M} = \mathscr{P}_\lambda(U)$ als Raum der Wahrscheinlichkeitsmaße P auf U, für die

$$\int \lambda(x)P(dx) \;<\; \infty.$$

Nach Theorem 6.3.3 gilt, dass $(\mathscr{P}_\lambda(U),\mathring{\kappa}_c)$ vollständig ist, wenn der zugrundeliegende Raum U separabel und vollständig ist. Dabei ist $\mathring{\kappa}_c$ eine weitere zum Monge-Kantorovich-Problem gehörende Metrik. Theorem 6.1.1 sagt aus, dass $\mathring{\kappa}_c$ und $\hat{\kappa}_c$ übereinstimmen, falls es sich bei der Kostenfunktion c um eine Metrik auf U handelt.

In unserem Fall ist der vollständige, separable, metrische Raum $(\mathfrak{C},\mathfrak{d})$ und man sieht leicht, dass die Metrik d mit $\hat{\kappa}_c$ übereinstimmt. Damit folgt der erste Teil der Behauptung.

Darüberhinaus geben wir einen Auszug aus Theorem 6.3.1 an: Es gilt für $P,\ P_1,P_2,\ldots \in \mathscr{P}_\lambda(U)$, dass die Konvergenz von P_n gegen P bezüglich d äquivalent ist zu

$$P_n \xrightarrow{\mathrm{w}} P \quad \text{und} \quad \lim_{C\to\infty} \sup_{n\in\mathbb{N}} \int \lambda(x) 1_{\{\lambda(x)>C\}} P_n(dx) = 0,$$

siehe dazu auch Seite 123 in [Rac91]. □

4.4.2 Voraussetzungen für die Konvergenz

Wir nennen eine Verteilung μ ein *kontrahierendes Split-Maß*, falls gilt

(i) μ ist auf das Einheitsintervall konzentriert: $\mu\big([0,1]\big) = 1$,

(ii) μ ist symmetrisch: Hat U Verteilung μ, so gilt $\mathcal{L}(U) = \mathcal{L}(1-U)$,

(iii) für die Kontraktionsbedingung: $\int x^2\mu(dx) < \frac{1}{2}$.

Beispiel 4.14 a) Ist $U \sim \text{unif}(0,1)$, so gilt $\mathbb{E}U^2 = \frac{1}{3}$ sowie $P(1-U \le x) = P(U \le x) = x$ für $x \in (0,1)$, also ist unif(0,1) ein kontrahierendes Split-Maß.

b) Die symmetrische Beta-Verteilung mit Parameter $k+1$, $k \in \mathbb{N}_0$ hat die Dichte

$$f(x) = \frac{(2k+1)!}{(k!)^2} x^k (1-x)^k, \qquad x \in [0,1].$$

Man sieht leicht, dass diese Verteilung symmetrisch ist. Weiter gilt

$$\int_0^1 x^2 f(x)\, dx = \ldots = \frac{k+2}{2(2k+3)} < \frac{1}{2} \text{ für alle } k \in \mathbb{N}_0.$$

Also ist auch diese Verteilung ein kontrahierendes Split-Maß. Für $k=0$ fällt diese Verteilung übrigens mit der vorher betrachteten Gleichverteilung zusammen.

c) Ist U eine Zufallsvariable mit $P(U=0) = P(U=1) = 1/2$, so gilt zwar $\mathcal{L}(U) = \mathcal{L}(1-U)$, allerdings ist $\mathbb{E}U^2 = 1/2$. Also ist die Verteilung von U *kein* kontrahierendes Split-Maß.

Das folgende Lemma zeigt insbesondere, dass die unter c) genannte Verteilung die einzige symmetrische Verteilung auf dem Einheitsintervall ist, die nicht die geforderte Kontraktionseigenschaft besitzt.

Lemma 4.15 Sind Teile (i) und (ii) obiger Definition erfüllt, so sind äquivalent:

a) $\int u^\gamma \, \mu(du) < 1/2$ für ein $\gamma > 1$,

b) $\int u^\gamma \, \mu(du) < 1/2$ für alle $\gamma > 1$,

c) $\mu(\{0,1\}) < 1$.

Beweis. Für $u \in (0,1)$ und $\gamma > 1$ gilt $u^\gamma < u$, also ist wegen $\mathbb{E}U = \mathbb{E}(1-U) = \frac{1}{2}$

$$\int u^\gamma \, \mu(du) < \int u \, \mu(du) = \mathbb{E}U = \frac{1}{2},$$

solange die Masse nicht in den Randpunkten konzentriert ist. Damit folgen bereits alle behaupteten Äquivalenzen. □

4.4.3 Die charakterisierende Rekursion

Auf $\mathbb{M}$ definieren wir nun in Abhängigkeit von μ die Transformation $T_\mu : \mathbb{M} \to \mathbb{M}$ durch

$$Q \;\mapsto\; \mathcal{L}\big(1 + 1_{\{U \geq \cdot\}} X_{\frac{\cdot}{U}} + 1_{\{1-U \geq \cdot\}} X'_{\frac{\cdot}{1-U}}\big),$$

wobei $\mathcal{L}(X) = \mathcal{L}(X') = Q$, $U \sim \mu$, und U, X, X' unabhängig. Für jedes $\alpha \in (0,1)$ ergibt sich also mit $Y \sim T_\mu Q$

$$Y_\alpha \;=_{\text{distr}}\; 1 + 1_{\{U \geq \alpha\}} X_{\frac{\alpha}{U}} + 1_{\{1-U \geq \alpha\}} X'_{\frac{\alpha}{1-U}},$$

mit $X,X' \sim Q$ unabhängig. Wegen $\alpha/U \geq \alpha$ und der vorausgesetzten Antitonie der Pfade ist $X_{\alpha/U} \leq X_\alpha$ und damit

$$\mathbb{E}\int_0^1 tY_t \; dt \;\leq\; \frac{1}{2} + 2 \cdot \mathbb{E}\int_0^1 tX_t \; dt \;<\; \infty,$$

somit führt T_μ nicht aus $\mathbb{M}$ heraus.

Lemma 4.16 Ist μ ein kontrahierendes Split-Maß, so ist T_μ eine Kontraktion bezüglich d.

Beweis. Seien $Q, Q' \in \mathbb{M}$ und $\varepsilon > 0$, dann existieren aufgrund der Infimumseigenschaft $X \sim Q$ und $Y \sim Q'$ mit

$$\mathbb{E}\int_0^1 t|X_t - Y_t|dt \;\leq\; d(Q,Q') + \varepsilon.$$

Seien $(X',Y') =_{\text{distr}} (X,Y)$, $U \sim \mu$ und $U, (X,Y), (X',Y')$ unabhängig. Dazu sei, für alle $t \in (0,1)$,

$$\begin{aligned} R_t \;&:=\; 1 + 1_{\{U \geq t\}} X_{\frac{t}{U}} + 1_{\{1-U \geq t\}} X'_{\frac{t}{1-U}}, \\ S_t \;&:=\; 1 + 1_{\{U \geq t\}} Y_{\frac{t}{U}} + 1_{\{1-U \geq t\}} Y'_{\frac{t}{1-U}}. \end{aligned}$$

Dann ist $\mathcal{L}(R) = T_\mu Q$, $\mathcal{L}(S) = T_\mu Q'$, und somit gilt

$$\begin{aligned} d(T_\mu Q, T_\mu Q') &\leq \mathbb{E}\int_0^1 t|R_t - S_t|dt \\ &\leq \int_0^1 t \cdot \mathbb{E}1_{\{U \geq t\}} \left|X_{\frac{t}{U}} - Y_{\frac{t}{U}}\right| dt \\ &\qquad + \int_0^1 t \cdot \mathbb{E}1_{\{1-U \geq t\}} \left|X'_{\frac{t}{1-U}} - Y'_{\frac{t}{1-U}}\right| dt. \end{aligned}$$

Wegen $\mathcal{L}(U) = \mathcal{L}(1-U)$ sind die beiden Summanden offensichtlich identisch. Wir erhalten durch Zerlegen nach dem Wert von U

$$\begin{aligned} d(TQ, TQ') &\leq 2\int_0^1 t\mathbb{E}1_{\{U \geq t\}} |X_{\frac{t}{U}} - Y_{\frac{t}{U}}| \; dt \\ &= 2\int_0^1 t \int 1_{\{u \geq t\}} |X_{\frac{t}{u}} - Y_{\frac{t}{u}}| \; \mu(du) \; dt. \end{aligned}$$

Mit Fubini und der anschließenden Substitution $s := t/u$ im inneren Integral wird dies zu

$$\begin{aligned} 2\int_0^1 t \int 1_{\{u \geq t\}} |X_{\frac{t}{u}} - Y_{\frac{t}{u}}| \; \mu(du) \; dt &= 2\int \int_0^u t|X_{\frac{t}{u}} - Y_{\frac{t}{u}}| \; dt \; \mu(du) \\ &= 2\int u^2 \int_0^1 s|X_s - Y_s| \; ds \; \mu(du) \\ &= 2\int u^2 \mu(du) \cdot \big(d(Q,Q') + \varepsilon\big). \end{aligned}$$

Da $\varepsilon > 0$ beliebig war, ergibt sich insgesamt

$$d(T_\mu Q, T_\mu Q') \leq K \cdot d(Q, Q'),$$

mit einem $K < 1$. □

Wir können nun die Vollständigkeit von $\mathbb{M}$ ausnutzen, die wir im vorigen Abschnitt nachgewiesen haben. Als Kontraktion besitzt T_μ einen eindeutigen Fixpunkt, den wir Q nennen. Damit kommen wir zum Hauptresultat dieses Kapitels.

Satz 4.17 Sei $(I_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsvariablen. Dabei gelte $P(I_n \in \{0,\dots,n-1\}) = 1$ für $n \in \mathbb{N}$ sowie $\lim_{n\to\infty} P(I_n = j) = 0$ für alle $j \in \mathbb{N}$. Außerdem konvergiere $\mathcal{L}(I_n/n)$ mit $n \to \infty$ schwach gegen $\mu \neq \frac{1}{2}(\delta_0 + \delta_1)$. Sei weiter $(X_n)_{n\in\mathbb{N}_0}$ eine von (I_n) unabhängige Folge stochastischer Prozesse auf $\mathfrak{C}$. Genügt dann $(X_n)_{n\in\mathbb{N}_0}$ der Rekursion $X_0 \equiv 0$, $X_1 \equiv 1$ und, für $n > 1$, $\alpha \in (0,1]$

$$X_{n,\alpha} \;=_{\text{distr}}\; 1 + 1_{\{I_n>\alpha n\}} X_{I_n,\frac{\alpha n}{I_n}} + 1_{\{n-1-I_n>\alpha n\}} X'_{n-1-I_n,\frac{\alpha n}{n-1-I_n}}, \qquad (4.13)$$

mit (X'_n), (X_n), (I_n) unabhängig und $\mathcal{L}(X_j) = \mathcal{L}(X'_j)$ für alle $j \in \mathbb{N}_0$, so folgt $\lim_{n\to\infty} d(X,X_n) = 0$, insbesondere gilt also $\mathcal{L}(X_n) \xrightarrow{\text{w}} Q$.

Beweis. Sei $\varepsilon > 0$. Aufgrund der Infimumseigenschaft der Metrik d existieren Zufallsgrößen X und X_i, $i \in \mathbb{N}$, mit

$$\int_0^1 t\,\mathbb{E}\big|X_{i,t} - X_t\big|\,dt \;\le\; d(X_i,X) + \varepsilon.$$

Weiterhin erlaubt die Skorohod-Einbettung die Konstruktion von $(I_n)_{n\in\mathbb{N}}$ und $U \sim \mu$ so, dass I_n/n sogar fast sicher gegen U konvergiert. Definiert man nun

$$\begin{aligned} Y_\alpha &:= 1 + 1_{\{U>\alpha\}} X_{\frac{\alpha}{U}} + 1_{\{1-U>\alpha\}} X'_{\frac{\alpha}{1-U}}, \quad \alpha \in (0,1], \quad \text{und} \\ Y_{n,\alpha} &:= 1 + 1_{\{I_n>n\alpha\}} X_{I_n,\frac{\alpha n}{I_n}} + 1_{\{n-1-I_n>\alpha n\}} X'_{n-1-I_n,\frac{\alpha n}{n-1-I_n}}, \quad \alpha \in (0,1], \end{aligned}$$

für $n > 1$, dazu $Y_0 \equiv 0$ und $Y_1 \equiv 1$, so gilt einerseits $Y =_{\text{distr}} X$; denn Y hat die Verteilung $T_\mu Q = Q$. Andererseits ist $Y_n =_{\text{distr}} X_n$ für alle $n \in \mathbb{N}_0$ nach der Voraussetzung (4.13) des Satzes. Also folgt

$$\begin{aligned} d(X_n,X) \;&\le\; \int_0^1 t\,\mathbb{E}\big|Y_{n,t} - Y_t\big|\,dt \\ &\le\; \int_0^1 t\Big[\mathbb{E}\big|\,1_{\{I_n>nt\}} X_{I_n,\frac{nt}{I_n}} - 1_{\{U>t\}} X_{\frac{t}{U}}\big| \\ &\qquad + \mathbb{E}\big|\,1_{\{n-1-I_n>nt\}} X_{n-1-I_n,\frac{nt}{n-1-I_n}} - 1_{\{1-U>t\}} X_{\frac{t}{1-U}}\big|\Big]\,dt \end{aligned}$$

Wir betrachten zunächst nur den ersten Erwartungswert. Der zweite wird dann analog behandelt; insbesondere ist I_n in dem Sinne „asymptotisch symmetrisch“, dass mit $I_n/n \to U$ folgt $(n-1-I_n)/n \to 1-U$, und U und $1-U$ haben dieselbe Verteilung.

Mit der Dreiecksungleichung zerlegen wir das Problem der Konvergenz:

$$\begin{aligned}\left|1_{\{I_n>nt\}}X_{I_n,\frac{nt}{I_n}} - 1_{\{U>t\}}X_{\frac{t}{U}}\right| \;\leq\; & \left|1_{\{I_n>nt\}}X_{I_n,\frac{nt}{I_n}} - 1_{\{I_n>nt\}}X_{\frac{nt}{I_n}}\right| \\ & + \left|1_{\{I_n>nt\}}X_{\frac{nt}{I_n}} - 1_{\{I_n>nt\}}X_{\frac{t}{U}}\right| \\ & + \left|1_{\{I_n>nt\}}X_{\frac{t}{U}} - 1_{\{U>t\}}X_{\frac{t}{U}}\right|.\end{aligned} \tag{4.14}$$

Hier offenbart sich bereits das weitere Vorgehen: Während das Integral über dem ersten Summanden durch die Abstände $d(X_i,X)$, $i < n$, abgeschätzt werden kann, konvergieren der zweite und dritte Summand (und dadurch auch deren Integrale) gegen Null. Einerseits, weil der Grenzprozess X stetig ist, andererseits, weil I_n/n nach Voraussetzung gegen U konvergiert.

Wir beginnen mit dem letzten Summanden. Es ist mit Fubini I

$$\int_0^1 t\,\mathbb{E}\left|1_{\{I_n>nt\}}X_{\frac{t}{U}} - 1_{\{U>t\}}X_{\frac{t}{U}}\right| dt \;=\; \mathbb{E}\int_0^1 t\left(X_{\frac{t}{U}}\cdot 1_{\{t \text{ zwischen } I_n/n \text{ und } U\}}\right) dt$$

Da die Pfade von X antiton sind, folgt wegen $t/U \geq t$, dass $X_{t/U} \leq X_t$ für alle $t \in (0,1)$ gilt, also lässt sich der Integrand nach erneuter Anwendung von Fubini I durch $t\,\mathbb{E}X_t$ nach oben beschränken. Wegen $\mathcal{L}(X) \in \mathbb{M}$ ist diese Majorante integrierbar, und nach dem Satz von der majorisierten Konvergenz folgt somit, dass dieser Summand wegen $I_n/n \to U$ verschwindet.

Für die Bearbeitung des zweiten Summanden stellen wir zunächst fest, dass wegen $X_\alpha \in \mathbb{N}_0$ mit Wahrscheinlichkeit 1 gilt, dass X nur abzählbar viele Sprünge hat, somit also ℓ-fast überall stetig ist. Mit Fubini I und der Substitution $s := nt/I_n$ gilt

$$\int_0^1 t\,\mathbb{E}\left|1_{\{I_n>nt\}}X_{\frac{nt}{I_n}} - 1_{\{I_n>nt\}}X_{\frac{t}{U}}\right| dt \;=\; \mathbb{E}1_{\{I_n>0\}}\int_0^{I_n/n} t\left|X_{\frac{nt}{I_n}} - X_{\frac{t}{U}}\right| dt.$$

Sollte das Argument α von X_α je größer als 1 werden, nehmen wir in Übereinstimmung mit Wertebereich und Monotonie stets $X_\alpha = 0$ an. Wegen $I_n/n \to U$

ist $(I_n s/(nU))_n$ eine fast sicher gegen s konvergierende Folge. Wie beim vorigen Summanden lässt sich der Integrand leicht majorisieren: Mit der Antitonie von X folgt wegen $nt/I_n \geq t$ und $t/U \geq t$, dass sich das Integral durch

$$2 \int_0^1 t\, X_t \, dt \;<\; \infty$$

nach oben beschränken lässt. Mit majorisierter Konvergenz und der zuvor beobachteten Stetigkeitseigenschaft folgt also auch für diesen Summanden die Konvergenz gegen Null.

Der Schlüssel zur Konvergenz der Verteilung von X_n gegen Q ist jedoch die Untersuchung des ersten Summanden in (4.14). Wir verwenden Fubini I, zerlegen nach dem Wert i von I_n und substituieren $s := nt/i$. Damit erhalten wir

$$\begin{aligned}
&\int_0^1 t\,\mathbb{E}\big|\, 1_{\{I_n > nt\}} X_{I_n, \frac{nt}{I_n}} - 1_{\{I_n > nt\}} X_{\frac{nt}{I_n}} \big| \; dt \\
&= \int_0^1 t \sum_{i=0}^{n-1} P(I_n = i) 1_{\{i > nt\}} \mathbb{E}\big| X_{i, \frac{nt}{i}} - X_{\frac{nt}{i}} \big| \; dt \\
&= \sum_{i=0}^{n-1} P(I_n = i) \Big(\frac{i}{n}\Big)^2 \int_0^1 s\,\mathbb{E}|X_{i,s} - X_s| \; ds \\
&\leq \sum_{i=0}^{n-1} P(I_n = i) \Big(\frac{i}{n}\Big)^2 \big(d(X_i, X) + \varepsilon\big)
\end{aligned}$$

Mit $a_n := d(X_n, X)$ ergibt sich insgesamt

$$a_n \;\leq\; 2 \sum_{i=0}^{n-1} P(I_n = i) \Big(\frac{i}{n}\Big)^2 (a_i + \varepsilon) + \mathcal{O}(1) \qquad \text{mit } n \to \infty.$$

Es verbleibt nun zu zeigen, dass $\limsup_{n\to\infty} a_n = 0$. Wir stellen zunächst fest, dass sich a_n asymptotisch grob abschätzen lässt durch

$$a_n \;\leq\; 2\,\mathbb{E}\Big(\frac{I_n}{n}\Big)^2 \big(\sup_{i<n} a_i + \varepsilon\big) + \mathcal{O}(1)$$

also ist die Folge (a_n) beschränkt. Sei $a := \limsup_{n\to\infty} a_n$, dann existiert ein $n_1 \in \mathbb{N}$ so, dass $a_n \leq a + \varepsilon$ für alle $n > n_1$. Außerdem folgt aus $I_n/n \to U$, dass $\mathbb{E}(I_n/n)^2 \to \mathbb{E}U^2$, da alle beteiligten Zufallsvariablen auf [0,1] beschränkt sind. Weiterhin ist $\mathbb{E}U^2 < 1/2$ nach Lemma 4.15. Also existiert ein $\xi < 1$ und ein $n_2 \in \mathbb{N}$ mit

$$2\,\mathbb{E}\Big(\frac{I_n}{n}\Big)^2 \;\leq\; \xi \qquad \text{für alle } n > n_2.$$

Wir setzen $n_0 := \max\{n_1,n_2\}$ und erhalten für $n > n_0$

$$\begin{aligned} a_n \;\leq\; & 2\sum_{i=0}^{n_0-1} P(I_n = i)(a_i + \varepsilon) \\ & + 2\sum_{i=n_0}^{n-1} P(I_n = i)\Big(\frac{i}{n}\Big)^2 (a_i + \varepsilon) \;+\; \mathcal{O}(1). \end{aligned}$$

Betrachtet man nun auf beiden Seiten den Limes superior, so verschwindet die erste Summe, da $\lim_{n\to\infty} P(I_n = j) = 0$ für alle $j \in \mathbb{N}$ vorausgesetzt war. In der zweiten Summe können wir $a_i \leq a+\varepsilon$ abschätzen; die verbleibende Summe begrenzen wir anschließend durch $\mathbb{E}(I_n/n)^2$ nach oben. Insgesamt folgt also

$$a \;\leq\; \xi(a + 2\varepsilon), \qquad \text{d.h.} \qquad a \;\leq\; \frac{2\xi}{1-\xi}\cdot\varepsilon,$$

und da $\varepsilon > 0$ beliebig war, folgt die Behauptung. □

4.4.4 Darstellung und Eigenschaften der Grenzverteilung

Wie sieht nun dieser Grenzprozess aus? Wir können eine Art „Grenzbaum“ konstruieren, an dem wir die Beschaffenheit der Verteilung des Grenzprozesses veranschaulichen wollen.

Seien U_v, $v \in \{0,1\}^*$, unabhängige Zufallsvariablen mit der Verteilung μ. Wir bewerten einen „unendlichen binären Baum“, also die Menge $\{0,1\}^*$ aller möglichen Knoten, mit der Bewertungsfunktion ϕ. Dabei sei $\phi(v)$ definiert als das Produkt der Variablen

$$U_{v'} \quad \text{bzw.} \quad 1 - U_{v'},$$

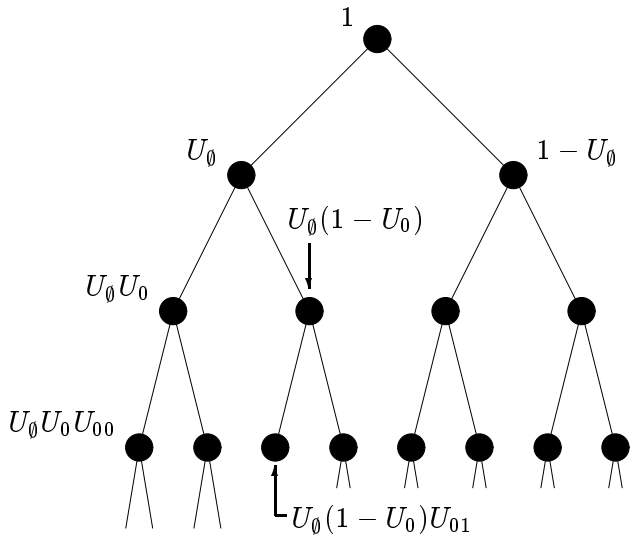

Abbildung 4.6: Veranschaulichung der Gewichtsfunktion ϕ.

bei einem Abstieg in v' nach links, bzw. rechts, wobei v' die Knoten von der Wurzel bis zum direkten Vorfahr von v durchläuft. So erhalten wir

$$\phi(v) \;:=\; \prod_{j=0}^{|v|-1} U_{v|_j}^{1-v_j}(1-U_{v|_j})^{v_j}, \quad v \in \{0,1\}^*. \tag{4.15}$$

Dabei bezeichnet $v|_j$ die Einschränkung von v auf dessen erste j Komponenten. Abbildung 4.6 zeigt eine Veranschaulichung von ϕ.

Wir betrachten die zufällige Funktion $X : (0,1) \to \mathbb{N}$, die einer Zahl $\alpha \in (0,1)$ die Anzahl der Knoten mit einer Bewertung größer als α zuordnet:

$$X(\alpha) \;:=\; \#\{v \in \{0,1\}^* : \phi(v) > \alpha\}.$$

Nach Konstruktion ist X stetig von rechts und monoton fallend, also gilt vorbehaltlich der Integrierbarkeitsbedingung $\mathcal{L}(X) \in \mathbb{M}$. Dies wird im Beweis des folgenden Satzes nachgewiesen.

Satz 4.18

$$\mathcal{L}(X) \;=\; Q.$$

Beweis. Wir zeigen zunächst, dass X für alle $\alpha \in (0,1)$ die Verteilungsrekursion

$$X_\alpha \;=_{\text{distr}}\; 1 + 1_{\{U \geq \alpha\}} X_{\alpha/U} + 1_{\{1-U \geq \alpha\}} X'_{\alpha/(1-U)} \tag{4.16}$$

erfüllt, wobei $U \sim \mu$, $\mathcal{L}(X) = \mathcal{L}(X')$, und X,X',U unabhängig sind.

Man überlegt sich leicht, dass für den Beitrag des linken Teilbaums zu X_α gilt

$$X_\alpha^L := \#\{v \in \{0,1\}^* : v_1 = 0, \phi(v) > \alpha\} = \#\{v \in \{0,1\}^* : \tilde{\phi}(v) > \frac{\alpha}{U_\emptyset}\}.$$

Dabei ist $\tilde{\phi}$ die Bewertungsfunktion des linken Teilbaums, also

$$\tilde{\phi}(v) \;=\; \prod_{j=1}^{|v|-1} U_{v|_j}^{1-v_j} (1 - U_{v|_j})^{v_j}.$$

Aufgrund der iid-Eigenschaft der U_v haben X_α^L und $X_{\alpha/U_\emptyset}$ bedingt unter $U_\emptyset$ dieselbe Verteilung. Ist dabei $U_\emptyset < \alpha$ sind beide Null. Die analoge Argumentation für den rechten Teilbaum belegt damit die obige Rekursion.

Damit ist $\mathcal{L}(X)$ ein Fixpunkt der Abbildung T. Können wir also $\mathcal{L}(X) \in \mathbb{M}$ zeigen, so ist der Beweis erbracht. Die Pfadeigenschaften wurden oben bereits erwähnt.

Für alle $\alpha \in (0,1)$ gilt

$$X_\alpha \;=\; \sum_{v \in \{0,1\}^*} 1_{\{\phi(v) > \alpha\}}$$

und damit nach monotoner Konvergenz

$$\mathbb{E}X_\alpha \;=\; \sum_{v \in \{0,1\}^*} P\big(\phi(v) > \alpha\big).$$

Wir zerlegen nach der Tiefe $|v|$ von v. Der Wurzelknoten hat die Bewertung $1 > \alpha$ für alle $\alpha \in (0,1)$; unter Berücksichtigung von $v_j \in \{0,1\}$ und der

Unabhängigkeit der U_v, $v \in \{0,1\}^*$, folgt nun wegen $1 - U_v =_{\text{distr}} U_v$

$$\begin{aligned}
\mathbb{E}X_\alpha &= 1 + \sum_{k=1}^{\infty} \sum_{v \in \{0,1\}^k} P\Big(\prod_{j=0}^{k-1} U_{v|_j}^{1-v_j} (1 - U_{v|_j})^{v_j} > \alpha \Big) \\
&= 1 + \sum_{k=1}^{\infty} \sum_{v \in \{0,1\}^k} P\Big(\prod_{j=0}^{k-1} U_{v|_j} > \alpha \Big) \\
&= 1 + \sum_{k=1}^{\infty} 2^k \cdot P\Big(\sum_{j=0}^{k-1} \log U_{\tilde{v}|_j} > \log \alpha \Big),
\end{aligned}$$

mit $\tilde{v} = (0,0,\ldots)$.

Wir verwenden die Markovsche Ungleichung. Zur Vereinfachung der Schreibweise gehen wir zu $U_0, U_1, U_2, \ldots$ anstelle von $U_\emptyset, U_0, U_{00}, \ldots$ über. Für alle $t \geq 0$ gilt demnach

$$\begin{aligned}
P\Big(\sum_{j=0}^{k-1} \log U_j > \log \alpha \Big) \;&\leq\; \mathrm{e}^{-t \log \alpha} \cdot \big(\mathbb{E} \exp(t \, \log U_0) \big)^k. \\
&=\; \alpha^{-t} \cdot \big(\mathbb{E} U_0^t \big)^k
\end{aligned}$$

Laut Lemma 4.15 erhalten wir mit $t = 3/2$ für den Erwartungswert von U_0^t einen Wert $\kappa < \frac{1}{2}$. Damit folgt

$$\begin{aligned}
\mathbb{E}X_\alpha \;&\leq\; 1 + \alpha^{-3/2} \sum_{k=1}^{\infty} (2\kappa)^k \\
&=\; 1 + \alpha^{-3/2} \cdot 2\kappa \cdot \frac{1}{1 - 2\kappa} \;<\; \infty.
\end{aligned}$$

Schließlich gilt also

$$\begin{aligned}
\int_0^1 t \, \mathbb{E}X_t \; dt \;&\leq\; \int_0^1 t \; dt + \int_0^1 t^{-1/2} \frac{2\kappa}{1 - 2\kappa} \; dt \\
&=\; \frac{1}{2} + 2 \cdot \frac{2\kappa}{1 - 2\kappa} \;<\; \infty,
\end{aligned}$$

und damit $\mathcal{L}(X) \in \mathbb{M}$. □

Bemerkung 4.19 Die Größe, deren Existenz wir soeben nachgewiesen haben, taugt auch als Maß für die Balance eines Baumes. In einem unbalancierten Baum häufen sich die Knoten typischerweise bei einem größeren Wert α als bei einem ausgeglichenen Baum. Interessanterweise hängt jedoch $\int t\mathbb{E}X_t\ dt$ nur über $\mathbb{E}U^2$ (bei $U \sim \mu$) von μ ab. Es gilt

$$\int_0^1 t\mathbb{E}X_t\, dt \;=\; \frac{1}{2(1-2\mathbb{E}U^2)}.$$

Dies sehen wir wie folgt: Nennen wir $\chi := \int_0^1 t\,\mathbb{E}X_t\ dt$, so folgt durch Einsetzen der Rekursion unter Ausnutzung der Symmetrie von μ

$$\begin{aligned}
\chi \;=\; \int_0^1 t\,\mathbb{E}X_t\ dt \;&=\; \int_0^1 t\,\left(1+2\,\mathbb{E}1_{\{U>t\}}X_{\frac{t}{U}}\right)\ dt\\
&= \frac{1}{2} + 2\int_0^1 t\int_t^1 \mathbb{E}X_{\frac{t}{u}}\ \mu(du)\ dt\\
&= \frac{1}{2} + 2\int_0^1 \int_0^u t\,\mathbb{E}X_{\frac{t}{u}}\ dt\ \mu(du)\\
&= \frac{1}{2} + 2\int_0^1 u^2 \int_0^1 s\,\mathbb{E}X_s\ ds\ \mu(du)\\
&= \frac{1}{2} + 2\int u^2\mu(du)\cdot\chi.
\end{aligned}$$

Auflösen nach χ liefert obige Gleichheit. Offensichtlich kann χ nicht kleiner als 1 werden; die „beste“ Balance sollte also ein Algorithmus mit $\mathbb{E}U^2 = 1/4$ (und damit $\operatorname{var} U = 0$) erreichen.

Zum Ende dieses Abschnitts wollen wir noch nachweisen, dass das zweite Moment $\mathbb{E}X_\alpha^2$ von X_α für jedes $\alpha \in (0,1)$ existiert.

Lemma 4.20 Für $\alpha \in (0,1)$ gilt

$$\mathbb{E}X_\alpha^2 \;<\; \infty.$$

Beweis. Wir benutzen wieder die Darstellung

$$X_\alpha \;=\; \sum_{v\in\{0,1\}^*} 1_{\{\phi(v)>\alpha\}}.$$

Wie bei jeder Summe entstehen beim Quadrieren verschiedene Typen von Summanden. Für $u,v \in \{0,1\}^*$ kann man drei Möglichkeiten unterscheiden:

1. $u = v$, d.h. summiert über diesem Typus erhalten wir exakt X_α.

2. u ist im Teilbaum $T(v)$ mit der Wurzel v enthalten (oder umgekehrt). In diesem Fall ist

   $$1_{\{\phi(u)>\alpha\}} \cdot 1_{\{\phi(v)>\alpha\}} = 1_{\{\phi(u)>\alpha\}} \quad \left(\text{oder } 1_{\{\phi(u)>\alpha\}} \cdot 1_{\{\phi(v)>\alpha\}} = 1_{\{\phi(v)>\alpha\}}\right)$$

   außerdem lässt sich, zerlegt nach der Tiefe k von u, auf $k-1$ verschiedene Arten ein entsprechender Ahn v finden. Der Beitrag dieser Summanden ist also

   $$2 \cdot \sum_{k=1}^{\infty} (k-1) 2^k 1_{\{U_1 \cdots U_k > \alpha\}},$$

   wobei $U_1, U_2, \ldots$ iid mit Verteilung μ.

3. Die Teilbäume $T(u)$ und $T(v)$ sind disjunkt. Sei m die maximale Tiefe eines gemeinsamen Vorfahr r von u und v, k die relative Tiefe von u im Teilbaum zu r, j entsprechend für v. Für u gibt es damit 2^k Möglichkeiten, für v unter Berücksichtigung der Disjunktheit der Teilbäume noch $2^j - 2^{\max\{j-k,0\}}$; denn ist $j \leq k$, so „sperrt“ u genau einen Knoten (nämlich den jeweiligen Ahnen, vgl. 2.), ist $j > k$, so sind 2^{j-k} Knoten durch Nachkommen von u „besetzt“. Summiert man diesen Typus von Summanden, so erhält man also

   $$\sum_{m=0}^{\infty} 2^m \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} 2^k \left(2^j - 2^{(j-k)^+}\right) 1_{\{U_1 \cdots U_m \cdot V_1 \cdots V_k > \alpha\}} \cdot 1_{\{U_1 \cdots U_m \cdot V_1' \cdots V_j' > \alpha\}},$$

   mit $U_1, \ldots, U_m, V_1, \ldots, V_k, V_1', \ldots, V_j'$ iid mit Verteilung μ.

Diese drei Typen addieren wir nun zusammen. Dabei zerlegen wir im dritten Fall nach dem Wert von $U_1 \cdots U_m$; zu diesem Zweck bezeichne ν die Verteilung

von $U_1 \cdots U_m$

$$\mathbb{E}X_\alpha^2 \;=\; \mathbb{E}X_\alpha \;+\; 2 \cdot \sum_{k=1}^{\infty} (k-1) 2^k P\big(U_1 \cdots U_k > \alpha\big)$$
$$+ \sum_{m=0}^{\infty} 2^m \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} 2^k \left(2^j - 2^{(j-k)^+}\right) \int P\big(V_1 \cdots V_k > \tfrac{\alpha}{u}\big) P\big(V_1' \cdots V_j' > \tfrac{\alpha}{u}\big)\, \nu(du),$$

Wie im Beweis zu Satz 4.17 verwenden wir die Markovsche Ungleichung um

$$P\big(U_1 \cdots U_k > \alpha\big) \;\leq\; \alpha^{-3/2} \cdot \kappa^k$$

mit einem $\kappa < 1/2$ zu erhalten. Damit erhalten wir für das Integral als Oberschranke

$$\int_0^1 u^3 \;\nu(du) \cdot \alpha^{-3} \kappa^{k+j} \;=\; \big(\mathbb{E}U_1^3\big)^m \alpha^{-3} \kappa^{k+j},$$

denn das Integral ist nichts anderes als das m-fache Produkt des dritten Momentes der Verteilung μ. Nach Lemma 4.15 gilt nun, dass $\kappa' := \mathbb{E}U_1^3 < 1/2$ gilt und somit folgt

$$\mathbb{E}X_\alpha^2 \;\leq\; \mathbb{E}X_\alpha \;+\; 2\alpha^{-3/2} \sum_{k=1}^{\infty} (k-1)(2\kappa)^k$$
$$+ \alpha^{-3} \sum_{m=0}^{\infty} (2\kappa')^m \sum_{k=1}^{\infty} (2\kappa)^k \sum_{j=1}^{\infty} (2\kappa)^j,$$

da alle beteiligten Summen konvergieren, ist somit der Beweis erbracht. □

4.5 Eisberge bei binären Bäumen

Die allgemeine Formulierung von Satz 4.17 ermöglicht die Anwendung auf eine Vielzahl von rekursiven Verteilungen. In diesem Abschnitt werden wir sowohl die bereits bekannten Algorithmen BST und DST unter diesem Gesichtspunkt untersuchen als auch das sog. *Median-of-*$(2k+1)$*-Quicksort.*

Etwas aus der Reihe fällt die Klasse der Catalan-Bäume (vgl. Kapitel 1). Hier ist der zentrale Satz wertlos. Aus den in Erklärungen, die wir in Abschnitt 4.5.4 anführen werden, wird dabei klar, dass der Eisberg-Prozess mit wachsender Knotenzahl fast sicher überall unbeschränkt ist. Um diesem Dilemma zu entgehen, betrachten wir die einzelnen Knotengrößen. In diesem Fall ergibt sich eine Grenzverteilung, die eine Darstellung des asymptotischen Teilbaumgrößenprofils des Catalan-Baums im Sinne eines Erneuerungsprozesses motiviert.

4.5.1 Binary Search Tree

Aus Beispiel 4.4 ist bekannt, dass die Knotenzahl I_n des linken Teilbaums der Gleichverteilung auf $\{0,\dots,n-1\}$ genügt. Als Grenzverteilung von I_n/n ergibt sich damit die Gleichverteilung auf (0,1), diese ist nach Beispiel 4.14 ein kontrahierendes Split-Maß. Schließlich gilt $P(I_n = k) \leq 1/n \to 0$ mit $n \to \infty$ für alle $k \in \mathbb{N}$ und damit sind die Voraussetzungen für Satz 4.17 erfüllt.

Über die Aussage dieses Satzes hinaus wollen wir nun zeigen, dass wir für den (gegen Null konvergierenden) Abstand der Verteilungen von X_n und X bzgl. d im BST-Fall sogar eine konkrete Oberschranke angeben können.

Dazu berechnen wir zunächst den Erwartungswert des Grenzprozesses, also die Funktion $\alpha \mapsto \mathbb{E}X_\alpha$. Die Anwendung des Erwartungswertoperators auf die Rekursionsgleichung für ein X, welches der Grenzverteilung genügt, liefert

$$\mathbb{E}X_\alpha \;=\; 1 + \mathbb{E}1_{\{U\geq\alpha\}}X_{\alpha/U} + \mathbb{E}1_{\{1-U\geq\alpha\}}X_{\alpha/(1-U)}, \quad \alpha \in (0,1).$$

Wir erhalten durch Zerlegen nach dem Wert von U und unter Ausnutzung der Symmetrie $\mathcal{L}(U) = \mathcal{L}(1-U)$

$$\mathbb{E}X_\alpha \;=\; 1 + 2\int_\alpha^1 \mathbb{E}X_{\alpha/u}\; du.$$

Man rechnet leicht nach, dass diese Rekursion von $\mathbb{E}X_\alpha = 2/\alpha - 1$ erfüllt wird:

Es gilt

$$1+2\int_\alpha^1 \Big(\frac{2u}{\alpha}-1\Big)\,du \;=\; 1+\frac{2}{\alpha}\big(1-\alpha^2\big)-2(1-\alpha)$$
$$=\; \frac{2}{\alpha}-1 \;=\; \mathbb{E}X_\alpha. \qquad (4.17)$$

In Abschnitt 4.5.1.2 berechnen wir außerdem die Varianzfunktion des Grenzprozesses.

4.5.1.1 Konvergenzgeschwindigkeit

Wir konstruieren den Baum T_n, auf dessen Teilbaumgrößenprofil sich X_n bezieht, wie folgt: Sei $\tilde{I}_\emptyset := n$ und für $v \in \{0,1\}^*$ sei

$$\mathcal{L}\big(\tilde{I}_{(v,0)}|\tilde{I}_v\big) = \mathrm{unif}\{0,\dots,\tilde{I}_v-1\} \quad \text{und} \quad \tilde{I}_{(v,1)} := \tilde{I}_v - \tilde{I}_{(v,0)} - 1.$$

Ist $\tilde{I}_v = 0$, so setzen wir auch alle $I_{v'}$ mit v' Nachkomme von v konstant Null.

Dann gibt $\tilde{I}_v$ jeweils die Größe des Teilbaums $T_n(v)$ an und nach Konstruktion erzeugen alle $v \in \{0,1\}^*$ mit $\tilde{I}_v \geq 1$ einen binären Baum mit n Knoten. Die bedingte Unabhängigkeit der Teilbäume ist erfüllt, da $\tilde{I}_v$ mit $|v| = k$ von allen $\tilde{I}_{v|_j}$, $j = 0,\dots,k-1$, nur durch $\tilde{I}_{v|_{k-1}}$ abhängt. Ebenso hängen natürlich die nachfolgenden „Knoten“ oder Teilbaumgrößen nur durch $\tilde{I}_v$ vom bisherigen Geschehen ab. Bedingt unter $\tilde{I}_v$ sind sie unabhängig.

Ausgehend von einer Folge U_v, $v \in \{0,1\}^*$, von unabhängigen und jeweils auf $[0,1]$ gleichverteilten Zufallsvariablen wählen wir nun

$$I_v \;:=\; n\cdot\phi_n(v), \qquad \text{mit}$$
$$\phi_n(v) \;:=\; \frac{1}{n}\lfloor\dots\lfloor nU_\emptyset^{1-v_1}(1-U_\emptyset)^{v_1}\rfloor\dots U_{v|_{k-1}}^{1-v_k}(1-U_{v|_{k-1}})^{v_k}\rfloor,$$

wobei $v \in \{0,1\}^*, |v| = k$. Dann ist wegen $\phi_n(\emptyset) = 1$ die Anfangsbedingung $I_\emptyset = n$ erfüllt und für $v \in \{0,1\}^*$ mit $|v| = k$ gilt

$$I_{(v,0)} \;=\; n\cdot\frac{1}{n}\lfloor I_v\cdot U_v\rfloor,$$

also

$$\mathcal{L}\big(I_{(v,0)}|I_v\big) \;=\; \mathrm{unif}\{0,\ldots,I_v-1\}$$

und

$$I_{(v,1)} \;=\; \lfloor I_v\cdot(1-U_v)\rfloor \;=\; I_v+\lfloor -I_vU_v\rfloor \;=\; I_v-I_{(v,0)}-1,$$

fast sicher. Damit gilt $I_v =_{\mathrm{distr}} \tilde{I}_v$ für alle $v\in\{0,1\}^*$, ebenso bleibt die oben beschriebene Unabhängigkeitsstruktur erhalten. Also bilden diejenigen $v\in\{0,1\}^*$ mit $I_v\geq 1$ einen BST-verteilten Baum mit n Knoten.

Bemerkung 4.21 Ein interessantes kombinatorisches Resultat, das sich hieraus ergibt, ist

$$\#\big\{v\in\{0,1\}^* \,:\, \phi_n(v)\geq\tfrac{1}{n}\big\} \;=\; n \qquad \text{fast sicher.}$$

Im Beweis des folgenden Satzes wird diese Erkenntnis jedoch keine Rolle mehr spielen.

Das kumulierte Endstück $X_{n,\alpha}$ lässt sich nun schreiben als die Anzahl der Knoten $v\in T_n$, für die $I_v>n\alpha$, d.h. $\phi_n(v)>\alpha$, erfüllt ist.

Satz 4.22 Für $n\in\mathbb{N}$ gilt

$$d(X,X_n) \;\leq\; \frac{3H_n}{n}-\frac{5}{n}+\frac{3H_n}{n^2}.$$

Insbesondere konvergiert $\mathcal{L}(X_n)$ mit $n\to\infty$ schwach gegen Q.

Beweis. Wir stellen zunächst fest, dass für alle $v\in\{0,1\}^*$ gilt

$$\phi_n(v) \;\leq\; \phi(v). \tag{4.18}$$

Vergleiche dazu auch die Definition (4.15) von ϕ auf Seite 101. Nach der bereits bekannten Konstruktion lassen sich die beteiligten Verteilungen darstellen als

$$\begin{aligned} X_\alpha &= \#\big\{v\in\{0,1\}^* \,:\, \phi(v)>\alpha\big\} \\ \text{und } X_{n,\alpha} &= \#\big\{v\in\{0,1\}^* \,:\, \phi_n(v)>\alpha \text{ und } \phi_n(v)\geq\tfrac{1}{n}\big\}. \end{aligned}$$

Offensichtlich ist nach (4.18) also $X_{n,\alpha} \leq X_\alpha$ für alle $n \in \mathbb{N}$ und alle $\alpha \in (0,1)$. Somit folgt

$$\begin{aligned}\lim_{n\to\infty} d(X_n,X) &\leq \lim_{n\to\infty} \int_0^1 t\,\mathbb{E}|X_{n,t} - X_t|\,dt \\ &= \lim_{n\to\infty} \int_0^1 t\,\big(\mathbb{E}X_t - \mathbb{E}X_{n,t}\big)\,dt.\end{aligned}$$

Beide beteiligten Erwartungswerte haben wir bereits bestimmt: Den von X_t auf Seite 108 und den von $X_{n,\alpha}$ in Abschnitt 4.3. Wir berechnen jedoch der Einfachheit halber zunächst die Integrale: Es gilt

$$\int_0^1 t\,\mathbb{E}X_t\,dt = \int_0^1 t\cdot\Big(\frac{2}{t} - 1\Big)\,dt = \frac{3}{2},$$

sowie

$$\begin{aligned}\int_0^1 t\,\mathbb{E}X_{n,t}\,dt &= \int_0^1 t\cdot\Big(\frac{2(n+1)}{\lceil nt\rceil + 1} - 1\Big) \\ &= 2(n+1)\sum_{k=1}^{n}\int_{\frac{k-1}{n}}^{\frac{k}{n}} \frac{t\,dt}{\lceil nt\rceil + 1} - \frac{1}{2} \\ &= 2(n+1)\sum_{k=1}^{n}\int_{\frac{k-1}{n}}^{\frac{k}{n}} \frac{t\,dt}{k+1} - \frac{1}{2} \\ &= 2(n+1)\sum_{k=1}^{n}\frac{2k-1}{2n^2(k+1)} - \frac{1}{2} \\ &= \ldots = \frac{3}{2} - \frac{3H_n}{n} + \frac{5}{n} - \frac{3H_n}{n^2}.\end{aligned}$$

Damit erhalten wir die behauptete Oberschranke für den betrachteten Abstand:

$$d(X,X_n) \leq \frac{3H_n}{n} - \frac{5}{n} + \frac{3H_n}{n^2}. \qquad \square$$

4.5.1.2 Eigenschaften der Grenzverteilung

Aus früheren Abschnitten wissen wir bereits, dass die Erwartungswertfunktion der Grenzverteilung Q

$$\alpha \mapsto \frac{2}{\alpha} - 1 = \mathbb{E}X_\alpha, \quad \alpha \in (0,1), \tag{4.19}$$

mit $X \sim Q$, lautet. Auch die Varianz lässt sich bestimmen. Wir benutzen dafür die bekannte Formel

$$\Psi(\alpha) := \operatorname{var} X_\alpha = \operatorname{var}\big(\mathbb{E}[X_\alpha|U]\big) + \mathbb{E}\operatorname{var}[X_\alpha|U]. \tag{4.20}$$

Das U, unter dem wir hier bedingen, ist dabei das U aus der Verteilungsrekursion (4.16). Daraus erkennt man nun, dass für $\alpha > \frac{1}{2}$ jeweils nur einer der beiden Summanden

$$1_{\{U\geq\alpha\}}X_{\alpha/U} \quad \text{und} \quad 1_{\{1-U\geq\alpha\}}X_{\alpha/(1-U)}$$

verschieden von Null sein kann. Sei also zunächst $\alpha > \frac{1}{2}$, dann erhält man durch Anwenden des bedingten Erwartungswertes auf die Verteilungsrekursion und mit den Rechenregeln für bedingte Erwartungswerte

$$\mathbb{E}[X_\alpha|U] = 1 + 1_{\{U\geq\alpha\}}\mathbb{E}[X_{\alpha/U}|U] + 1_{\{1-U\geq\alpha\}}\mathbb{E}[X_{\alpha/(1-U)}|U].$$

Nach (4.19) folgt also

$$\mathbb{E}[X_\alpha|U] = 1 + 1_{\{U\geq\alpha\}}\Big(\frac{2U}{\alpha} - 1\Big) + 1_{\{1-U\geq\alpha\}}\Big(\frac{2(1-U)}{\alpha} - 1\Big). \tag{4.21}$$

Wegen $\alpha > \frac{1}{2}$ sind die Bedingungen der Indikatorfunktionen disjunkt, daher fällt beim Quadrieren ein gemischter Term weg:

$$\begin{aligned}\mathbb{E}[X_\alpha|U]^2 &= 1 + 1_{\{U\geq\alpha\}}\Big(\frac{2U}{\alpha} - 1\Big)^2 + 1_{\{1-U\geq\alpha\}}\Big(\frac{2(1-U)}{\alpha} - 1\Big)^2 \\ &\quad + 2\cdot 1_{\{U\geq\alpha\}}\Big(\frac{2U}{\alpha} - 1\Big) + 2\cdot 1_{\{1-U\geq\alpha\}}\Big(\frac{2(1-U)}{\alpha} - 1\Big) \\ &= 2\cdot\mathbb{E}[X_\alpha|U] - 1 + 1_{\{U\geq\alpha\}}\Big(\frac{2U}{\alpha} - 1\Big)^2 \\ &\quad + 1_{\{1-U\geq\alpha\}}\Big(\frac{2(1-U)}{\alpha} - 1\Big)^2 \end{aligned} \tag{4.22}$$

Die Berechnung des zweiten Momentes dieses bedingten Erwartungswertes erfolgt nun durch Zerlegen nach dem Wert von U und unter Ausnutzung der Verteilungssymmetrie $\mathcal{L}(U) = \mathcal{L}(1-U)$:

$$\begin{aligned}
\mathbb{E}\big(\mathbb{E}[X_\alpha|U]^2\big) &= 2\mathbb{E}X_\alpha - 1 + 2\cdot\int_\alpha^1 \Big(\frac{2u}{\alpha}-1\Big)^2 du \\
&= 2\cdot\Big(\frac{2}{\alpha}-1\Big) - 1 + \frac{\alpha}{3}\bigg(\Big(\frac{2}{\alpha}-1\Big)^3 - 1\bigg) \\
&= \frac{4}{\alpha} - 3 + \frac{\alpha}{3}\Big(\frac{8}{\alpha^3} - \frac{12}{\alpha^2} + \frac{6}{\alpha} - 2\Big) \\
&= -1 - \frac{2\alpha}{3} + \frac{8}{3\alpha^2}. \qquad (4.23)
\end{aligned}$$

Damit erhalten wir für den ersten Summanden in (4.20)

$$\begin{aligned}
\operatorname{var}\big(\mathbb{E}[X_\alpha|U]\big) &= \mathbb{E}\big(\mathbb{E}[X_\alpha|U]\big)^2 - \big(\mathbb{E}\mathbb{E}[X_\alpha|U]\big)^2 \\
&= -1 - \frac{2\alpha}{3} + \frac{8}{3\alpha^2} - \Big(\frac{2}{\alpha}-1\Big)^2 \\
&= -\frac{2\alpha}{3} - 2 + \frac{4}{\alpha} - \frac{4}{3\alpha^2}.
\end{aligned}$$

Für die bedingte Varianz gilt nach Definition von Ψ und unter Ausnutzung der Verteilungsrekursion

$$\operatorname{var}[X_\alpha|U] = 1_{\{U\geq\alpha\}}\Psi\Big(\frac{\alpha}{U}\Big) + 1_{\{1-U\geq\alpha\}}\Psi\Big(\frac{\alpha}{1-U}\Big),$$

also erhalten wir durch Zerlegung nach U und unter erneuter Verwendung der Symmetrie von $\mathcal{L}(U)$

$$\mathbb{E}\big(\operatorname{var}[X_\alpha|U]\big) = 2\cdot\int_\alpha^1 \Psi\Big(\frac{\alpha}{u}\Big)\,du = 2\alpha\cdot\int_\alpha^1 \frac{1}{u^2}\Psi(u)\,du,$$

im letzten Schritt wurde naheliegend substituiert.

Insgesamt folgt damit für $\alpha > \frac{1}{2}$ die folgende Darstellung:

$$\Psi(\alpha) = 2\alpha\cdot\int_\alpha^1 \frac{1}{u^2}\Psi(u)\,du - \frac{2\alpha}{3} - 2 + \frac{4}{\alpha} - \frac{4}{3\alpha^2}. \qquad (4.24)$$

Lemma 4.23 Für $\alpha > \frac{1}{2}$ ist Ψ unendlich oft stetig differenzierbar, und es gilt $\Psi(1) = 0$.

Beweis. Folgt sofort aus (4.24) und Lemma 4.20. □

Mit diesem Lemma können wir also durch einmaliges Differenzieren zusammen mit der Randbedingung eine alternative Darstellung für Ψ erhalten. Der besseren Lesbarkeit geschuldet schreiben wir $\Psi'(\alpha)$ für die Ableitung nach α. Die Produktregel für die Differentiation zusammen mit dem Hauptsatz der Differentialrechnung liefert

$$\begin{aligned}\Psi'(\alpha) &= 2 \cdot \int_\alpha^1 \frac{1}{u^2}\Psi(u)\, du - 2\alpha \cdot \frac{1}{\alpha^2}\Psi(\alpha) - \frac{2}{3} - \frac{4}{\alpha^2} + \frac{8}{3\alpha^3} \\ &= -\frac{1}{\alpha}\Psi(\alpha) + \frac{2}{\alpha} - \frac{8}{\alpha^2} + \frac{4}{\alpha^3},\end{aligned}$$

dabei haben wir das Integral durch die entsprechende Umformulierung von (4.24) ersetzt. Diese Differentialgleichung (mit der Randbedingung aus Lemma 4.23) wird durch

$$\Psi(\alpha) = 2 - \frac{8\log\alpha}{\alpha} + \frac{2}{\alpha} - \frac{4}{\alpha^2}$$

gelöst.

Nun betrachten wir $\alpha \in (0{,}1)$. Durch Quadrieren von (4.21) kommt in (4.22) der Summand

$$2 \cdot 1_{\{\alpha \le U \le 1-\alpha\}} \Big(\frac{2U}{\alpha} - 1\Big)\Big(\frac{2(1-U)}{\alpha} - 1\Big)$$

hinzu, damit müssen wir – falls $\alpha \le \frac{1}{2}$ – in (4.23)

$$\begin{aligned}2 \cdot \int_\alpha^{1-\alpha} \Big(\frac{2u}{\alpha} - 1\Big)\Big(\frac{2(1-u)}{\alpha} - 1\Big)\, du &= 4 \cdot \int_\alpha^{1/2} \Big(\frac{4u - 4u^2}{\alpha^2} - \frac{2}{\alpha} + 1\Big)\, du \\ &= \frac{4\alpha}{3} + 2 - \frac{4}{\alpha} + \frac{4}{3\alpha^2}\end{aligned}$$

addieren und erhalten somit, für $\alpha \in (0,1)$,

$$\operatorname{var}\big(\mathbb{E}[X_\alpha|U]\big) \;=\; 1_{\{\alpha>\frac{1}{2}\}}\Big(-\frac{2\alpha}{3}-2+\frac{4}{\alpha}-\frac{4}{3\alpha^2}\Big) \;+\; 1_{\{\alpha\le\frac{1}{2}\}}\frac{2\alpha}{3}.$$

Damit folgt wie oben

$$\Psi(\alpha) \;=\; \begin{cases} 2\alpha\cdot\displaystyle\int_\alpha^1 \frac{1}{u^2}\Psi(u)\,du - \frac{2\alpha}{3} - 2 + \frac{4}{\alpha} - \frac{4}{3\alpha^2}, & \alpha > \frac{1}{2} \\ 2\alpha\cdot\displaystyle\int_\alpha^1 \frac{1}{u^2}\Psi(u)\,du + \frac{2\alpha}{3}, & \alpha \le \frac{1}{2}. \end{cases} \tag{4.25}$$

Lemma 4.24 Ψ ist für $\alpha \neq \frac{1}{2}$ unendlich oft stetig differenzierbar und stetig in $\alpha = \frac{1}{2}$.

Beweis. Die Stetigkeit in $\frac{1}{2}$ lässt sich leicht nachrechnen; denn das Integral hat für $\alpha+$ und $\alpha-$ denselben Wert und kann daher unbeachtet bleiben. Hier, wie auch bei der Differenzierbarkeitsaussage wird wieder Lemma 4.20 verwendet. □

Analog zum Vorgehen für den Fall $\alpha > \frac{1}{2}$ wird differenziert, die Differentialgleichung gelöst unter der Randbedingung $\Psi(\frac{1}{2}) = \Psi(\frac{1}{2}+)$. Wir erhalten insgesamt den

Satz 4.25 Für $\alpha \in (0,1)$ ist

$$\operatorname{var} X_\alpha \;=\; \begin{cases} 2 + \dfrac{2}{\alpha} - \dfrac{4}{\alpha^2} - \dfrac{8\log\alpha}{\alpha}, & \alpha > \frac{1}{2} \\ \dfrac{8\log 2 - 5}{\alpha}, & \alpha \le \frac{1}{2}. \end{cases}$$

In Abbildung 4.7 ist die Varianzfunktion skizziert.

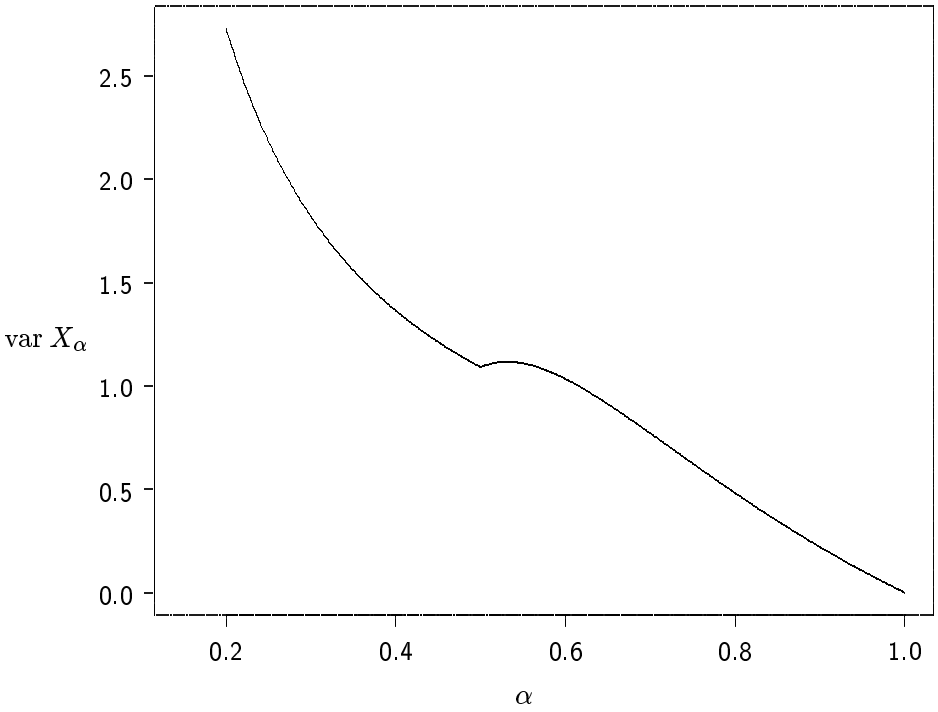

Abbildung 4.7: Varianzfunktion $\Psi : \alpha \mapsto \text{var } X_\alpha$ im BST-Fall

4.5.2 Digital Search Tree

Im Digital Search Tree-Algorithmus werden die Datenwerte nach ihrer Binärdarstellung in den binären Baum einsortiert. Geht man von gleichverteilten Daten aus dem Einheitsintervall aus, so landen solche, die kleiner als $\frac{1}{2}$ sind, links, die anderen rechts und in Folge dessen ist I_n binomialverteilt mit Parametern $n-1$ und $\frac{1}{2}$.

Nach dem Gesetz der großen Zahlen folgt $\frac{1}{n} I_n \to \frac{1}{2}$, die Grenzverteilung μ ist das Einpunktmaß in $\frac{1}{2}$, also sind die Voraussetzungen für Satz 4.17 erfüllt. Da die von U abhängenden Indikatorvariablen nun aufgrund der Beschaffenheit von μ deterministisch sind, ist auch der Grenzprozess deterministisch. Wir

behaupten

$$X^{\mathrm{DST}}(\alpha) \;=\; -1 + \sum_{k=1}^{\infty} 1_{[2^{-k},2^{-k+1})}(\alpha)\cdot 2^k, \quad \alpha > 0,$$

und weisen dies wie folgt nach: Einsetzen der rechten Seite in die Rekursion liefert

$$\begin{aligned}
&1 + 1_{\{U>\alpha\}} X_{\frac{\alpha}{U}} + 1_{\{1-U>\alpha\}} X'_{\frac{\alpha}{1-U}} \\
&= 1 + 2\cdot 1_{\{\alpha<1/2\}}\left(-1 + \sum_{k=1}^{\infty} 1_{[2^{-k},2^{-k+1})}(2\alpha)2^k\right) \\
&= -1 + 2\cdot 1_{[\frac{1}{2},1)}(\alpha) + \sum_{k=1}^{\infty} 1_{[2^{-k-1},2^{-k})}(\alpha)2^{k+1} \\
&= -1 + \sum_{k=1}^{\infty} 1_{[2^{-k},2^{-k+1})}(\alpha)\cdot 2^k \;=\; X(\alpha).
\end{aligned}$$

Von der ersten zur zweiten Zeile benutzen wir, dass $\alpha < \frac{1}{2}$ bereits von $\alpha \in [2^{-k-1},2^{-k})$ impliziert wird, so $k \in \mathbb{N}$ gilt; im letzten Schritt wurde der Summationsbereich verschoben. X ist also Fixpunkt der Rekursion und es bleibt lediglich der Nachweis, dass $\mathcal{L}(X) \in \mathbb{M}$ gilt. Die Pfadeigenschaften stehen außer Frage, weiter gilt mit Fubini I und dem Satz von der monotonen Konvergenz

$$\mathbb{E}\int_0^1 tX_t\,dt \;=\; -\frac{1}{2} + \sum_{k=1}^{\infty} 2^k\cdot\frac{1}{2}\Big(\big(2^{-k+1}\big)^2 - \big(2^{-k}\big)^2\Big) \;=\; 1,$$

und damit ist die obige Behauptung bewiesen. Es überrascht nicht, dass folglich nach Bemerkung 4.19 der DST-Algorithmus unter allen binären Bäumen, die aus der Konstruktion über rekursive Verteilungsfamilien nach Abschnitt 4.2 hervorgehen, die bestmögliche Balance erreicht.

4.5.3 Median-of-$(2k+1)$-Quicksort

Der Algorithmus des Median-of-$(2k+1)$-Quicksort versucht, den gewöhnlichen Quicksort-Algorithmus dadurch zu verbessern, dass nicht das erste auftretende

Element in den Wurzelknoten gesetzt wird, sondern der Median der ersten $(2k+1)$ Elemente. Somit entscheidet der absolute Rang desjenigen Elementes, welches unter den ersten $(2k+1)$ Elementen den Rang $k+1$ hat, über die Einordnung in linken und rechten Teilbaum. Mit $k=0$ erhält man den BST-Algorithmus.

Um die Verteilung von I_n zu bestimmen, nutzen wir aus, dass die absoluten Ränge von n unabhängigen, auf dem Einheitsintervall gleichverteilten Zufallsvariablen eine zufällige Permutation von $\{1,\ldots,n\}$ bilden; dabei ist jede Permutation wieder gleich wahrscheinlich. Für $I_n = j$ muss das k-kleinste Element der ersten $2k+1$ genau das j-kleinste von allen sein. Stellen wir uns eine Urne mit n Kugeln vor, von denen $2k+1$ markiert sind, so erhalten wir $I_n = j$ genau dann, wenn beim sukzessiven Ziehen gerade im $(j+1)$-ten Zug die $(k+1)$-te markierte Kugel zum Vorschein kommt. Dafür müssen unter den ersten j Kugeln genau k markierte sein und im letzten Zug eine weitere markierte Kugel gezogen werden. Somit gilt

$$P(I_n = j) \;=\; \frac{\binom{n-2k-1}{j-k}\binom{2k+1}{k}}{\binom{n}{j}} \cdot \frac{k+1}{n-j}.$$

In diesem Fall ist I_n+1 negativ hypergeometrisch verteilt mit den Parametern n, $2k+1$ und $k+1$ (vgl. [JK69, S. 309]). Grenzverteilung für I_n/n ist damit die symmetrische Betaverteilung mit Parameter $k+1$: Wir zeigen

$$\frac{P(I_n/n \;=\; \lfloor nx \rfloor/n)}{1/n} \;\to\; \frac{(2k+1)!}{(k!)^2} x^k (1-x)^k, \quad \text{für alle } x \in [0{,}1]. \qquad (4.26)$$

In den Randpunkten ist nichts zu zeigen, sei also $x \in (0{,}1)$. Wir verwenden die Notation $x^{\underline{k}}$ für die absteigende Potenz, d.h. $x^{\underline{k}} := x(x-1)\cdots(x-k+1)$. Es gilt

$$\begin{aligned} n \cdot P(I_n = \lfloor nx \rfloor) \;&=\; \frac{n\binom{\lfloor nx \rfloor}{k}\binom{n-\lfloor nx \rfloor - 1}{k}}{\binom{n}{2k+1}} \\ &=\; \frac{n\lfloor nx \rfloor^{\underline{k}} \lfloor n(1-x) \rfloor^{\underline{k}} (2k+1)!}{k!k!n^{\underline{2k+1}}} \\ &=\; \frac{(2k+1)!}{(k!)^2} \cdot \frac{\lfloor nx \rfloor^{\underline{k}} \lfloor n(1-x) \rfloor^{\underline{k}}}{(n-1)^{\underline{2k}}}, \end{aligned}$$

und der zweite Faktor konvergiert gegen $x^k(1-x)^k$. Dass daraus bereits die behauptete Konvergenz gegen die Beta-Verteilung folgt, sieht man wie folgt. Es ist mit einer geeigneten Substitution

$$\begin{aligned} P(I_n/n \le x) &= \sum_{j=0}^{\lfloor nx \rfloor} P(I_n = j) \\ &= \int_0^{\lceil nx \rceil} P(I_n = \lfloor y \rfloor)\, dy \\ &= \int_0^{\lceil nx \rceil / n} nP(I_n/n = \lfloor nz \rfloor / n)\, dz. \end{aligned}$$

Ein nahezu triviales Majorisierungsargument führt zusammen mit (4.26) auf

$$P(I_n/n \le x) \to \int_0^x \frac{(2k+1)!}{(k!)^2} x^k (1-x)^k.$$

Mit Satz 4.17 erhalten wir also die Konvergenz der Eisberg-Prozesse auch im Median-of-$(2k+1)$-Quicksort-Fall.

Für die Balance ergibt sich aus $\mathbb{E}U^2 = (k+2)/(4k+6)$ der Wert $\chi = 1 + 1/(2k+2)$. Vergleiche auch Beispiel 4.14 und Bemerkung 4.19. Für großes k nähern wir uns also mehr und mehr der Balance des DST-Baumes an.

4.5.3.1 Erwartungswert des Grenzprozesses

Wie bei den BST-Eisbergen kann man auch in diesem Fall den Erwartungswert aus der Rekursion bestimmen. Dies wird mit wachsendem k jedoch zunehmend komplizierter. Für $k = 1$ erhält man

$$\mathbb{E}X_\alpha = \frac{12}{7\alpha} + \frac{2\alpha^6}{7} - 1, \tag{4.27}$$

für größere k wollen wir einen Algorithmus angeben, der auf eine explizite Differentialgleichung für die Erwartungswertfunktion des Grenzprozesses führt.

Dazu sei $\phi_k(x) := \mathbb{E}X_\alpha^{(k)}$, und $f_k(x)$ die Dichte der Beta(k+1,k+1)-Verteilung. Mit

$$r_n(x) := \int_x^1 t^{-n}\phi_k(t)\ dt \qquad \text{gilt} \qquad r_n'(x) = -x^{-n}\phi_k(x), \tag{4.28}$$

und damit erhalten wir aus der Rekursion für den Grenzprozess mit einer Substitution $t = x/u$

$$\begin{aligned}
\phi_k(x) &= 1 + 2\int_x^1 f_k(u)\phi_k(x/u)\ du \\
&= 1 + \frac{2(2k+1)!}{k!^2}\int_x^1 \frac{x}{t^2}\Big(\frac{x}{t}\Big)^k\Big(1-\frac{x}{t}\Big)^k\phi_k(t)\ dt \\
&= 1 + \frac{2(2k+1)!}{k!^2}\int_x^1 x^{k+1}t^{-(k+2)}\sum_{j=0}^{k}\binom{k}{j}(-1)^j t^{-j}x^j\phi_k(t)\ dt \\
&= 1 + \frac{2(2k+1)!}{k!}\sum_{j=0}^{k}\frac{(-1)^j}{j!(k-j)!}\cdot x^{k+j+1}r_{k+j+2}(x).
\end{aligned}$$

Lemma 4.26 Für $l = 0,1,\ldots,k$ gilt

$$\phi_k^{(l)}(x) = \delta_{l0} + \frac{2(2k+1)!}{k!}\sum_{j=0}^{k}\frac{(-1)^j}{j!(k-j)!}\cdot\frac{(k+j+1)!}{(k+j+1-l)!}x^{k+j+1-l}r_{k+j+2}(x).$$

Beweis. Induktion. Für $l = 0$ ist nichts zu zeigen. Wir betrachten nun zunächst einen später benötigten Ausdruck. Mit einer einfachen kombinatorischen Gleichheit folgt

$$\begin{aligned}
\sum_{j=0}^{k}\binom{k}{j}\binom{k+j+1}{l}(-1)^j &= \sum_{j=0}^{k}\binom{k}{j}\sum_{i=0}^{l}\binom{j}{i}\binom{k+1}{l-i}(-1)^j \\
&= \sum_{i=0}^{l}\binom{k+1}{l-i}\binom{k}{i}\sum_{j=0}^{k}\binom{k-i}{k-j}(-1)^j.
\end{aligned}$$

Die innere Summe ist genau $(-1)^k$, falls $i = k$, und 0 sonst, daher ergibt sich insgesamt Null, solange $l < k$. Dies gilt natürlich auch noch nach Multiplikation von $l!/k!$, also folgt

$$\sum_{j=0}^{k} \frac{(-1)^j}{j!(k-j)!} \cdot \frac{(k+j+1)!}{(k+j+1-l)!} = \begin{cases} 0 & \text{für } 0 \leq l < k \in \mathbb{N}. \\ (-1)^k & \text{für } l = k. \end{cases} \tag{4.29}$$

Nach Produktregel und (4.28) gilt

$$\left(x^{k+j+1-l} r_{k+j+2}(x)\right)' = (k+j+1-l)x^{k+j+1-(l+1)} - x^{-(l+1)}\phi_k(x),$$

setzt man dies in die Induktionsvoraussetzung ein, so erhält man mit (4.29)

$$\frac{d}{dx}\phi_k^{(l)}(x) = \frac{2(2k+1)!}{k!} \sum_{j=0}^{k} \frac{(-1)^j}{j!(k-j)!} \frac{(k+j+1)!}{(k+j+1-(l+1))!} \cdot x^{k+j+1-(l+1)} r_{k+j+2}(x)$$

und dies zeigt die Behauptung. □

Dieses Lemma legt das weitere Vorgehen nahe: Die $(k+1)$-te Ableitung von ϕ_k verlässt das obige Schema und die beteiligten r-Terme lassen sich durch geschickte Linearkombination früherer Ableitungen eliminieren. Wir betrachten $\phi_k^{(k+1)}(x)$, welches wir aus Lemma 4.26 (mit $l = k$) erhalten:

$$\begin{aligned} \phi_k^{(k+1)}(x) &= \left(\phi_k^{(k)}(x)\right)' \\ &= \frac{2(2k+1)!}{k!} \sum_{j=0}^{k} \frac{(-1)^j}{j!(k-j)!} \frac{(k+j+1)!}{(j+1)!} \\ &\qquad \cdot \left((j+1)x^j r_{k+j+2}(x) - x^{-(k+1)}\phi_k(x)\right) \\ &= \frac{2(2k+1)!}{k!} \sum_{j=0}^{k} \frac{(-1)^j (k+j+1)!}{j!^2 (k-j)!} x^j r_{k+j+2}(x) \\ &\quad + \frac{2(2k+1)!}{k!}(-1)^{k+1} x^{-(k+1)} \phi_k(x), \end{aligned}$$

nach (4.29). Sei nun $(a_0,\ldots,a_k)$ so definiert, dass

$$\sum_{l=0}^{k} \frac{a_l}{(k+j+1-l)!} \;=\; j! \qquad \text{für alle } j=0,\ldots,k.$$

Man überlegt sich leicht, dass dieses Gleichungssystem für jedes $k \in \mathbb{N}$ eine eindeutige Lösung besitzt. Mit Lemma 4.26 folgt dann

$$\begin{aligned}
&a_0 x^{-(k+1)}\big(\phi_k(x)-1\big) + \sum_{l=1}^{k} a_l x^{-(k+1-l)} \phi_k^{(l)}(x) \\
&= \sum_{l=0}^{k} a_l \frac{(2k+1)!}{k!} \sum_{j=0}^{k} \frac{(-1)^j}{j!(k-j)!} \frac{(k+1+j)!}{(k+1+j-l)!} x^j r_{k+j+2}(x) \\
&= \frac{(2k+1)!}{k!} \sum_{j=0}^{k} \frac{(-1)^j (k+j+1)!}{j!(k-j)!} x^j r_{k+j+2}(x) \sum_{l=0}^{k} \frac{a_l}{(k+j+1-l)!} \\
&= \phi_k^{(k+1)}(x) + \frac{2(2k+1)!}{k!}(-1)^k x^{-(k+1)} \phi_k(x).
\end{aligned}$$

Definiert man nun $\tilde{a}_0 := a_0 + (-1)^{k+1}\frac{2(2k+1)!}{k!}$, $\tilde{a}_l := a_l$ für $l=1,\ldots,k$, sowie $\tilde{a}_{k+1} := -1$, und multipliziert die gesamte Gleichung mit x^{k+1}, so ergibt sich $\phi_k(x)$ als Lösung der gewöhnlichen Differentialgleichung $(k+1)$-ter Ordnung

$$\sum_{l=0}^{k+1} \tilde{a}_l x^l \phi_k^{(l)}(x) \;=\; a_0, \quad \phi_k(1)=1,\ \phi_k^{(l)}(1)=0,\ \text{für } l=1,\ldots,k+1.$$

Beispiel 4.27 a) Für $k=1$ lautet das LGS für die a_l:

$$\frac{a_0}{2} + a_1 = 1 \quad \text{und} \quad \frac{a_0}{6} + \frac{a_1}{2} = 1,$$

also folgt $a_0=-6$ und $a_1=4$. Die Differentialgleichung

$$6\phi_1(x) + 4x\phi_1'(x) - x^2\phi_1''(x) = -6$$

führt mit den entsprechenden Randbedingungen auf die am Anfang dieses Abschnitts angegebene Erwartungswertfunktion (4.27).

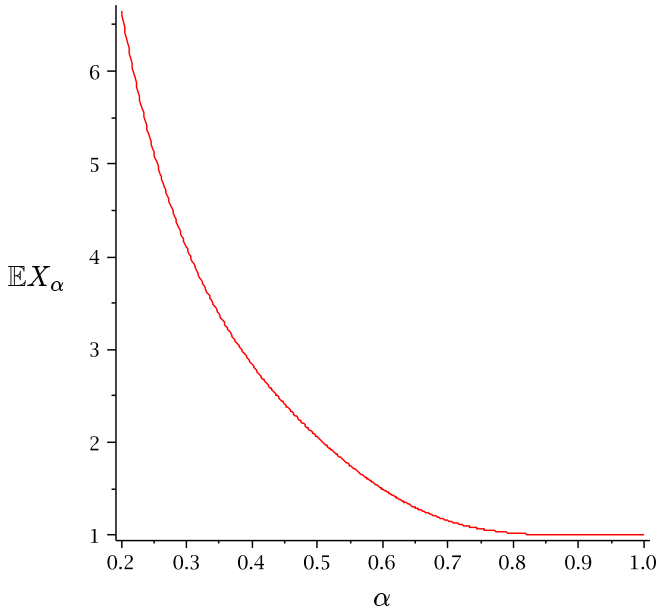

Abbildung 4.8: $\alpha \mapsto \mathbb{E}X_\alpha$ im Fall des Median-of-11-Algorithmus ($k = 5$)

b) Für $k = 5$ erhält man als Lösung des LGS

$$(a_0, \ldots, a_5) = (-332640, 181440, -45360, 6720, -630, 36),$$

eine numerische Lösung der Differentialgleichung ist in Abbildung 4.8 zu sehen.

4.5.4 Catalan-Bäume

Als *Catalan-Baum* bezeichnen wir zu einem vorgegebenen $n \in \mathbb{N}$ einen auf der Menge der binären Bäume mit n Knoten gleichverteilten Baum (vgl. Kapitel 1). Es ist bekannt, dass es

$$C_n = \frac{1}{n+1}\binom{2n}{n}$$

binäre Bäume mit n Knoten gibt; C_n heißt *n-te Catalansche Zahl*. Ist also T_n ein Catalan-Baum und t_n ein binärer Baum mit n Knoten, so gilt

$$P(T_n = t_n) \;=\; \frac{1}{C_n}.$$

Für die Größe I_n des linken Teilbaumes erhalten wir damit die folgende Verteilung:

$$P(I_n = k) \;=\; \frac{C_k C_{n-1-k}}{C_n}, \qquad k = 0, \ldots, n-1,$$

denn auf dem Laplaceschen Wahrscheinlichkeitsraum der binären Bäume mit n Knoten zählen wir, wieviele davon einen linken Teilbaum der Größe k besitzen. Da ein binärer Baum aber durch seinen linken und rechten Teilbaum vollständig beschrieben ist, ergeben sich genau $C_k C_{n-1-k}$ viele Bäume mit n Knoten, die $I_n = k$ erfüllen.

Auch im Catalan-Fall liegt die bereits bekannte Unabhängigkeitsstruktur vor: Bedingt unter I_n sind linker und rechter Teilbaum unabhängig. Dies lässt sich leicht nachrechnen.

Damit können wir auch in diesem Fall die Verteilungsrekursion für das Teilbaumgrößenprofil aufstellen:

$$K_n \;=_{\text{distr}}\; K_{I_n} \;+\; K'_{n-1-I_n} + \delta_n,$$

mit (K_n), (K'_n), I_n unabhängig, und I_n mit oben angegebener Verteilung.

Lemma 4.28 Für den Erwartungswert gilt für $k \leq n$

$$\mathbb{E}K_{n,k} \;=\; (n-k+1) \cdot \frac{C_k C_{n-k}}{C_n}.$$

Beweis. Vollständige Induktion. Für $n = 1$ ist $K_{1,1} \equiv 1$, gelte die Behauptung also für alle $n' < n$. Wir rechnen die Rekursion nach: es gilt mit $a_{n,k} := \mathbb{E}K_{n,k}$

$$a_{n,k} \;=\; \mathbb{E}K_{I_n,k} + \mathbb{E}K_{n-1-I_n,k} + \delta_{n,k},$$

die rechten Erwartungswerte zerlegen wir nach dem Wert von I_n und nutzen dabei die Verteilungssymmetrie $I_n =_{\text{distr}} n-1-I_n$ aus. Dies ergibt $a_{n,n}=1$ wie behauptet und, für $k<n$,

$$a_{n,k} = 2\sum_{i=0}^{n-1} a_{i,k} P(I_n = i).$$

Nach Induktionsvoraussetzung sind alle Summanden bis zu $i=k$ Null, außerdem können wir wegen $i \le n-1$ die $a_{i,k}$ durch die behaupteten Werte ersetzen. Damit folgt mit einer Indexverschiebung $i \mapsto i+k$

$$\begin{aligned} a_{n,k} &= 2\sum_{i=k}^{n-1}(i-k+1)\cdot\frac{C_k C_{i-k}}{C_i}\cdot\frac{C_i C_{n-1-i}}{C_n} \\ &= 2\cdot\frac{C_k C_{n-k}}{C_n}\sum_{i=0}^{n-k-1}(i+1)\cdot\frac{C_i C_{n-k-1-i}}{C_{n-k}} \\ &= (n-k+1)\cdot\frac{C_k C_{n-k}}{C_n}, \end{aligned}$$

denn die Summe ist gerade der Erwartungswert von $I_{n-k}+1$, nämlich $(n-k+1)/2$. Damit folgt die Behauptung. □

Aus diesem Erwartungswertvektor können wir wie vorher im BST-Fall bereits elementare asymptotische Eigenschaften des Teilbaumgrößenprofils im Catalan-Fall ablesen. Dabei ist die Asymptotik der Catalan-Zahlen selbst von ausschlaggebender Bedeutung. Mit $n \to \infty$ gilt

$$C_n \sim \frac{4^n}{\sqrt{\pi n^3}}.$$

In diesem Zusammenhang ist mit $a_n \sim b_n$ $(n\to\infty)$ gemeint, dass mit $n\to\infty$ gilt $a_n/b_n \to 1$.

Da für $k \ge n/2$ nicht $K_{n,k} > 1$ gelten kann, können wir aus der Asymptotik der Catalanzahlen sofort die asymptotische Verteilung der „großen“ Knoten bestimmen:

Lemma 4.29 Für jedes $k \in \mathbb{N}_0$ gilt

$$K_{n,n-k} \xrightarrow{\text{distr}} Z_k, \qquad \text{mit } P(Z_k = 1) = 1 - P(Z_k = 0) = 4^{-k}\binom{2k}{k}.$$

Beweis. Der Fall $k = 0$ ist trivial. Sei $k \in \mathbb{N}$. Ist n groß genug ($n \geq 2k$), so nimmt $K_{n,n-k}$ nur noch die Werte 0 und 1 an. Der Erwartungswert liefert für letzteren Fall die Wahrscheinlichkeit. Dabei gilt nach Lemma 4.28

$$\begin{aligned} P(K_{n,n-k} = 1) \;=\; \mathbb{E}K_{n,n-k} \;&=\; (k+1) \cdot \frac{C_k C_{n-k}}{C_n} \\ &\sim\; (k+1)C_k \cdot \frac{4^{n-k}\sqrt{\pi n^3}}{\sqrt{\pi(n-k)^3}4^n} \\ &\to\; (k+1)C_k \cdot 4^{-k} \;=\; 4^{-k}\binom{2k}{k}. \end{aligned}$$

□

Wie vorher bei der Untersuchung der Eisberg-Prozesse sind die großen Knoten natürlich nicht asymptotisch unabhängig. Wir können die Asymptotik der gemeinsamen Verteilung dieser Knotenzahlen angeben.

Satz 4.30 Sei $k \in \mathbb{N}$, dann gilt, mit $n \to \infty$

$$\big(K_{n,n-1}, K_{n,n-2}, \ldots, K_{n,n-k}\big) \xrightarrow{\text{distr}} X,$$

und die Verteilung von X ist gegeben wie folgt: Für ein $x = (x_1, \ldots, x_k) \in \{0,1\}^k$ seien $0 =: j_0 < j_1 < \ldots < j_m \leq k$ genau die Positionen in x an denen eine 1 steht, außerdem $\Delta_i := j_i - j_{i-1}$, $i = 1, \ldots, k$. Dann gilt

$$P\big(X = x\big) \;=\; 2^{m-2j_m}\Big(\prod_{i=1}^{m} C_{\Delta_i - 1}\Big) \cdot \Big(1 - \frac{1}{2}\sum_{i=0}^{k-j_m-1} C_i 4^{-i}\Big).$$

Beweis. Sei $X_n := (K_{n,n-1}, \ldots, K_{n,n-k})$ und sei $x = (x_1, \ldots, x_k) \in \{0,1\}^k$. Wir betrachten den Wurzelknoten. Ist nun x_{j_1} die erste Position, an der in x eine „1“ auftaucht, so bedeutet dies, dass der linke oder der rechte Teilbaum des Wurzelknotens genau $n - j_1$ Knoten besitzen muss. Bedingt unter dieser

Knotenanzahl sind dessen Teilbäume unabhängig, wir können dieselbe Überlegung also rekursiv wiederholen. Schließlich ergibt die Betrachtung desjenigen Knotens zu dem der Teilbaum mit $n - j_m$ Knoten gehört, dass dieser keinen Teilbaum mit mehr als $n - k$ Knoten besitzen darf. Dies führt auf

$$\begin{aligned} P(X_n = x) \;=\; P\Big(& I_n = n - j_1 \text{ oder } n - 1 - I_n = n - j_1, \\ & I_{n-j_1} = n - j_2 \text{ oder } n - j_1 - 1 - I_{n-j_1} = n - j_2, \\ & \vdots \\ & I_{n-j_{m-1}} = n - j_m \text{ oder } n - j_{m-1} - 1 - I_{n-j_{m-1}} = n - j_m, \\ & I_{n-j_m} < n - k \text{ und } n - j_m - 1 - I_{n-j_m} < n - k \Big). \end{aligned}$$

Symmetrie und Unabhängigkeit der I-Variablen ermöglichen die Umformung

$$\begin{aligned} & P(X_n = x) \\ &= 2^m \Bigg(\prod_{i=1}^{m} P(I_{n-j_{i-1}} = \underbrace{j_i - j_{i-1}}_{=\Delta_i} - 1) \Bigg) \cdot \Big(1 - 2P(I_{n-j_m} \le k - j_m - 1) \Big) \\ &= 2^m \Bigg(\prod_{i=1}^{m} \frac{C_{\Delta_i - 1} C_{n-j_{i-1}-\Delta_i}}{C_{n-j_{i-1}}} \Bigg) \cdot \Big(1 - 2 \sum_{i=0}^{k-j_m-1} \frac{C_i C_{n-j_m-1-i}}{C_{n-j_m}} \Big). \end{aligned}$$

Innerhalb des Produktes kürzen sich nun wegen $n - j_{i-1} - \Delta_i = n - j_i$ jeweils der zweite Faktor des Zählers gegen den Nenner des Faktors mit dem nächsthöheren Laufindex i. Es verbleibt außer den $C_{\Delta_i - 1}$ also nur der Faktor

$$\frac{C_{n-j_m}}{C_n} \;\sim\; \frac{4^{n-j_m}\sqrt{\pi n^3}}{4^n \sqrt{\pi (n-j_m)^3}} \;\to\; 4^{-j_m}.$$

Ebenso ergibt sich das asymptotische Verhalten der Summanden zu

$$\frac{C_{n-j_m-1-i}}{C_{n-j_m}} \;\to\; 4^{-(i+1)}.$$

Insgesamt folgt

$$\lim_{n\to\infty} P(X_n = x) \;=\; 2^{m-2j_m} \Bigg(\prod_{i=1}^{m} C_{\Delta_i - 1} \Bigg) \cdot \Big(1 - \frac{1}{2} \sum_{i=0}^{k-j_m-1} C_i 4^{-i} \Big) \;=\; P(X = x)$$

wie behauptet. □

Wir können X als eine Art Erneuerungsprozess interpretieren: Seien $Y_1, Y_2, \ldots$ unabhängige und identisch verteilte Zufallsvariablen mit

$$P(Y_1 = j) \;=\; 2^{1-2j} C_{j-1}, \qquad j \in \mathbb{N}.$$

Dass dies ein Wahrscheinlichkeitsmaß ist, kann man mit Hilfe der erzeugenden Funktion der Catalan-Zahlen

$$\sum_{j=0}^{\infty} z^j C_j \;=\; \frac{1-\sqrt{1-4z}}{2z}$$

nachweisen. Durch Einsetzen von $z = 1/4$ erhält man $\sum_{j=0}^{\infty} 2^{-2j} C_j = 2$ und damit

$$\sum_{j=1}^{\infty} P(Y_1 = j) \;=\; \sum_{j=0}^{\infty} 2^{-2j-1} C_j \;=\; \frac{1}{2} \sum_{j=0}^{\infty} 2^{-2j} C_j \;=\; 1.$$

Die Verteilung von Y_1 nennen wir *Catalan-Verteilung*. Zu diesen „Lebensdauern“ Y_j sei nun $S_l := \sum_{j=1}^{l} Y_j$ die Gesamtlebensdauer der ersten l Komponenten und $N = (N_m)_{m \in \mathbb{N}}$ mit $N_0 := 0$ und $N_j := \sup\{l : S_l \leq j\}$ der zugehörige (diskrete) Erneuerungsprozess. Für eine Folge von „Sprungstellen“ für N, also ein $x \in \{0,1\}^k$ erhalten wir dann mit den Bezeichnungen des vorhergehenden Satzes

$$\begin{aligned}
&P(N|_{\{1,\ldots,k\}} \text{ springt genau in } x) \\
&= P(Y_1 = \Delta_1, \ldots, Y_m = \Delta_m, Y_{m+1} > k - j_m) \\
&= \left(\prod_{i=1}^{m} P(Y_1 = \Delta_i) \right) \cdot \big(1 - P(Y_1 \leq k - j_m)\big) \\
&= \left(\prod_{i=1}^{m} 2^{1-2\Delta_i} C_{\Delta_i - 1} \right) \cdot \left(1 - \sum_{i=1}^{k-j_m} 2^{1-2i} C_{i-1} \right) \\
&= 2^{m-2j_m} \left(\prod_{i=1}^{m} C_{\Delta_i - 1} \right) \cdot \left(1 - \frac{1}{2} \sum_{i=0}^{k-j_m-1} C_i 4^{-i} \right) \\
&= P(X = x).
\end{aligned}$$

Das bedeutet insbesondere für den „unendlichen Catalan-Baum“, dass sich dieser wie folgt konstruieren lässt: Ausgehend von der Wurzel wählen wir zufällig

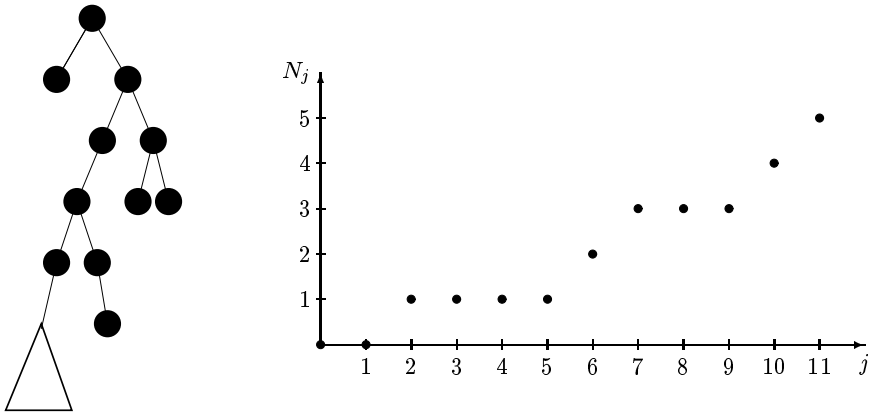

Abbildung 4.9: Anfangsstück eines unendlichen Catalan-Baumes und der daraus resultierende Erneuerungsprozess. Die Sprungstellen sind in diesem Fall $x = (0,1,0,0,0,1,1,0,0,1)$, die Lebensdauern damit $Y_1 = 2$, $Y_2 = 4$, $Y_3 = 1$, $Y_4 = 3$ und $Y_5 = 1$.

(je mit Wahrscheinlichkeit 1/2) aus, ob der linke oder der rechte Teilbaum endlich sein soll, und setzen dessen Größe auf $Y_1 - 1$. Der Teilbaum selbst ist also auf der Menge der binären Bäume mit $Y_1 - 1$ Knoten gleichverteilt. Für den anderen (den unendlichen) Teilbaum wählen wir wieder, ob es „links oder rechts weitergeht“ und setzen die Größe des „Sackgassen-Stückes“ auf $Y_2 - 1$. Die Prozedur wird unendlich oft wiederholt.

Damit haben wir über die Asymptotik des Teilbaumgrößenprofils eine Erklärung für den bekannten Sachverhalt gefunden, dass ein „unendlicher Catalan-Baum“ genau einen unendlichen Pfad besitzt, von dem endliche Catalan-Bäume abzweigen. Dies wird zum Beispiel bei Janson in [Jan02] gezeigt.

Zum Ende dieses Abschnitts wollen wir noch nachweisen, dass die Anwendung des zentralen Resultates für die „Eisberge“ an dessen Voraussetzungen scheitert; denn die Grenzverteilung von I_n/n ist gerade die Gleichverteilung auf $\{0,1\}$, vgl. Beispiel 4.14. Dies wollen wir zeigen, in dem wir die Varianz von I_n betrachten.

Lemma 4.31 Für $n \in \mathbb{N}$ gilt

$$\mathbb{E}I_n(I_n+1) \;=\; \frac{n(n+1)}{2} - 2^{n-2}\frac{(n+1)!}{1\cdot 3\cdots(2n-1)}.$$

Beweis. Mittels des Verhältnisses aufeinanderfolgender Catalan-Zahlen

$$\frac{C_{n-1}}{C_n} = \frac{n+1}{2(2n-1)}$$

erhalten wir

$$\begin{aligned}
\mathbb{E}I_n(I_n+1) &= \sum_{i=0}^{n-1} i(i+1)\frac{C_i C_{n-1-i}}{C_n} \\
&= \sum_{i=1}^{n-1} i(i+1)\frac{C_{i-1}C_{n-2-(i-1)}}{C_{n-1}}\frac{C_{n-1}}{C_n}\frac{C_i}{C_{i-1}} \\
&= \frac{n+1}{2(2n-1)}\cdot 2\cdot \sum_{i=1}^{n-1} i(2i-1)\cdot P(I_{n-1}=i-1) \\
&= \frac{n+1}{2n-1}\mathbb{E}(I_{n-1}+1)(2I_{n-1}+1) \\
&= \frac{2(n+1)}{2n-1}\mathbb{E}I_{n-1}(I_{n-1}+1) \;+\; \frac{n(n+1)}{2(2n-1)}.
\end{aligned}$$

Wir nennen $q_n := \mathbb{E}I_n(I_n+1)$ und erhalten die Rekursion

$$q_n \;=\; \frac{2(n+1)}{2n-1}q_{n-1} \;+\; \frac{n(n+1)}{2(2n-1)}$$

mit der Anfangsbedingung $q_1 = 0$; denn $I_1 \equiv 0$. Nun zeigen wir die Aussage des Lemmas mit vollständiger Induktion. Der Induktionsanfang besteht darin, nachzurechnen, dass die rechte Seite des Lemmas für $n = 1$ Null ergibt. Ist die Aussage bis $n-1$ gezeigt, so folgt aus der obigen Rekursion nach Induktionsvoraussetzung

$$\begin{aligned}
q_n &= \frac{2(n+1)}{2n-1}\left(\frac{(n-1)n}{2} - 2^{n-3}\frac{n!}{1\cdot 3\cdots(2n-3)}\right) + \frac{n(n+1)}{2(2n-1)} \\
&= \frac{n(n+1)}{2(2n-1)}\Big(2(n-1)+1\Big) \;-\; 2^{n-2}\frac{(n+1)!}{1\cdot 3\cdots(2n-1)},
\end{aligned}$$

und die Behauptung folgt. □

Hauptansinnen dieses Lemmas ist natürlich die Berechnung der Varianz von I_n. Es gilt dabei

$$\begin{aligned}\operatorname{var} I_n &= \mathbb{E}I_n^2 - \big(\mathbb{E}I_n\big)^2 \\ &= \mathbb{E}I_n\,(I_n+1) - \mathbb{E}I_n\,\big(1+\mathbb{E}I_n\big) \\ &= \frac{n(n+1)}{2} - 2^{n-2}(n+1)!\frac{2^n n!}{(2n)!} - \frac{n-1}{2}\Big(1+\frac{n-1}{2}\Big) \\ &= \Big(\frac{n+1}{2}\Big)^2 - \frac{4^{n-1}}{C_n}.\end{aligned}$$

Daraus lässt sich nun die Asymptotik der Verteilung von I_n/n ableiten:

Satz 4.32 Mit $n \to \infty$ gilt

$$\frac{1}{n}\, I_n \xrightarrow{\text{distr}} Z,$$

und Z ist gleichverteilt auf $\{0,1\}$.

Beweis. Wir betrachten die Varianz der linken Seite. Es gilt

$$\operatorname{var}\frac{I_n}{n} = \frac{1}{n^2}\operatorname{var} I_n = \frac{1}{4}\frac{(n+1)^2}{n^2} - \frac{4^{n-1}}{n^2 C_n}.$$

Der erste Summand konvergiert gegen $1/4$, der zweite aufgrund des asymptotischen Verhaltens der Catalanschen Zahlen gegen Null. Außerdem ist I_n/n auf das Einheitsintervall konzentriert, die Familie der Verteilungen $\mathcal{L}(I_n/n)$ ist also trivialerweise straff. Die einzige Verteilung auf dem Einheitsintervall mit der Varianz $1/4$ ist jedoch oben genannte Gleichverteilung auf den Randpunkten. (Dies ist die maximale Varianz, die eine Verteilung auf dem Einheitsintervall überhaupt besitzen kann.) Mit einem Standardargument zur Straffheit folgt die Behauptung. □

4.6 Random Recursive Tree

Bei der Untersuchung asymptotischer Strukturen von binären Bäumen trifft man häufig auf analoge Argumentationen zwischen dem *Binary Search Tree*

und dem *Random Recursive Tree*. Diese Analogie gründet sich zumeist auf die Ähnlichkeit der beteiligten erzeugenden Funktionen. In diesem Abschnitt nutzen wir Ähnlichkeiten in der rekursiven Struktur dieser beiden stochastischen Objekte aus.

Allerdings ist ein Random Recursive Tree gar kein binärer Baum. Wir betrachten die Menge der nicht-planaren Bäume mit n Knoten, die eine surjektive Gewichtsfunktion mit Werten in $\{1,\ldots,n\}$ besitzen derart, dass die Gewichte entlang eines jeden Pfades von der Wurzel zu einem externen Knoten eine streng isotone Zahlenfolge bilden. Wählen wir T_n gleichverteilt aus dieser Menge, so nennen wir T_n einen Random Recursive Tree.

Alternativ kann man solch einen Random Recursive Tree auch iterativ erzeugen. Dazu beginnt man mit T_1, welcher nur aus dem mit „1“ gelabelten Wurzelknoten besteht. Nun gelangt man von T_n zu T_{n+1} durch zufällige gleichverteilte Auswahl eines Knotens aus T_n, an den der Knoten mit dem Wert $n+1$ angehängt wird.

Bezeichnet man den Wert des Knotens, an den im Schritt von T_n zu T_{n+1} der neue Knoten angehängt wird, mit a_n, so beschreibt die Folge $(a_k)_{k=1,\ldots,n-1}$ den entstandenen Baum T_n vollständig. Außerdem ist (a_k) die einzige Folge, die auf diesen Baum führt.

Auf diese Weise kann man einen Random Recursive Tree mit n Knoten beschreiben als ein Tupel $(a_1,\ldots,a_{n-1})$ mit der Eigenschaft $a_j \leq j$ für alle $j = 1,\ldots,n-1$. Dabei bedeutet $a_j = i$, dass der Knoten mit dem Wert $j+1$ an den Knoten mit dem Wert i angehängt wird. (Der Wurzelknoten hat stets den Wert 1). Wie oben betrachten wir ein einzelnes Tupel, welches auf der Menge der zulässigen Tupel gleichverteilt sein soll.

Um für diesen Algorithmus eine Verteilungsrekursion wie für die binären Bäume zu erhalten, zerlegt man in den Teilbaum des Wurzelknotens mit dem kleinsten Label und den Rest. Dabei hat der ausgewählte Teilbaum eine Knotenzahl, die auf $\{1,\ldots,n-1\}$ gleichverteilt ist (siehe z.B. [Mah91]). Die Eisberge genügen also der Verteilungsrekursion

$$X_{n,\alpha} =_{\text{distr}} 1 + 1_{\{I_n>\alpha n\}} X_{I_n,\frac{n}{I_n}\alpha} + 1_{\{n-I_n>\alpha n\}} X'_{n-I_n,\frac{n}{n-I_n}\alpha} - 1_{\{n-I_n>\alpha n\}},$$

wobei wieder X' eine unabhängige Kopie von X und I_n auf $\{1,\dots,n-1\}$ gleichverteilt und von X und X' unabhängig ist. Diese Rekursion lässt sich auch aus der Teilbaumgrößenprofilrekursion für diesen Baumtyp ableiten: Für $j \in \mathbb{N}$ gilt

$$K_{n,j} \;=_{\text{distr}}\; K_{I_n,j} + K'_{n-I_n,j} - 1_{\{n-I_n=j\}} + \delta_{nj},$$

nämlich finden sich Teilbäume der Größe j entweder links, oder als j-Teilbäume des übrigen Baumes – dann allerdings darf nicht der um den linken Teilbaum beschnittene Baum selbst j Knoten haben. Bei $n = j$ sind alle Summanden für die Teilbäume Null, also wird mit dem Kronecker-Delta korrigiert. An diesen beiden Rekursionen kann man die am Anfang des Abschnitts beschriebene Ähnlichkeit der rekursiven Strukturen gut erkennen.

Nicht direkt aus Satz 4.17 aber mit einer vollkommen analogen Argumentation erhalten wir auch hier die Konvergenz des Eisbergprozesses gegen diejenige (eindeutige) Verteilung Q, welche die folgende Rekursion erfüllt:

$$X_\alpha \;=_{\text{distr}}\; 1_{\{U>\alpha\}} X_{\frac{\alpha}{U}} + 1_{\{1-U>\alpha\}} X'_{\frac{\alpha}{1-U}} + 1_{\{1-U\le\alpha\}},$$

$X, X' \sim Q$ etc. und $U \sim \text{unif}(0,1)$.

Auch hier können wir uns den „Limes-Baum“ analog zum BST-Fall vorstellen: einziger Unterschied ist, dass die Knoten, deren Pfad mit einem Schritt nach rechts endet, nur als „Verbinder“ fungieren. Dies ist eine Variante der Knuthschen Korrespondenz zwischen Bäumen und binären Bäumen: Die direkten Nachfolger des Wurzelknotens im nicht-binären Baum sind die Knoten (0), (1,0), (1,1,0),.... Diese Beschreibung wird rekursiv auf den gesamten Baum ausgedehnt.

Im Eisberg-Prozess werden nun diese „Verbinder“ nicht mitgezählt. Mit diesem Argument folgt sofort, dass Erwartungswert und Varianz existieren müssen, da (bei entsprechender Konstruktion) $X_\alpha^{\text{RRT}} \le X_\alpha^{\text{BST}}$ für alle $\alpha \in (0,1]$ erfüllt ist. Wieder einmal lösen wir eine Rekursionsgleichung auf und erhalten für den Random Recursive Tree

$$\mathbb{E}X_\alpha = \frac{1}{\alpha} \qquad \text{und} \qquad \operatorname{var} X_\alpha = \begin{cases} \frac{1}{\alpha} - \frac{1}{\alpha^2} - \frac{2\log\alpha}{\alpha}, & \alpha \ge \frac{1}{2}, \\ \frac{2\log 2 - 1}{\alpha}, & \alpha < \frac{1}{2}. \end{cases}$$

4.7 Ausblick

Für die Untersuchung des Teilbaumgrößenprofils gibt es – abgesehen vom reinen kombinatorischen Interesse – mehrere Motivationen. Flajolet, Gourdon und Martinez führen beispielsweise in [FGM97] an, dass man aus Erkenntnissen über das Teilbaumgrößenprofil den Speicherplatz für die Implementierung nichtlinearer Datenstrukturen optimieren kann. Devroye schreibt in [Dev91], dass die Untersuchung sogenannter *local counter*, wie beispielsweise der Anzahl der Knoten ohne Nachkommen, zu einem gewissen Grad Aufschluss geben kann über die Struktur des Baumes.

In Kapitel 3 haben wir eine weitere (indirekte) Motivation dadurch erhalten, dass die Pfadlänge als Testgröße für einen optimalen Test zur Unterscheidung von BST und DST taugt. Diese Idee lässt sich insoweit ausbreiten, dass das Teilbaumgrößenprofil nach dem Neyman-Kriterium eine suffiziente Statistik für die Familie der in Abschnitt 4.2 beschriebenen rekursiven Verteilungsfamilien bildet.

Bei der Analyse der Asymptotik sind wir zunächst vom BST-Fall ausgegangen. Die „Mäuse“, also die Anfangsstücke des Profils, sind asymptotisch mehrdimensional normalverteilt (Satz 4.9), die Endstücke verschwinden, wenn man sie einzeln betrachtet. Über die Kumulierung der großen Knoten haben wir im Sinne des ersten Absatzes dieses Kapitels (S. 65) eine weitere „Kenngröße“ entwickelt und ihr asymptotisches Verhalten beschrieben. Wie erhofft, lässt diese Untersuchung wieder den Rückschluss auf den Baum selbst und dessen asymptotisches Verhalten zu.

Leider entzog sich der Bereich „zwischen Mäusen und Eisbergen“ einer Betrachtung mit stochastischen Methoden. Hier haben Feng, Mahmoud und Panholzer in [FMP08] und Fuchs in [Fuc07] asymptotische Poissonverteilung der einzelnen Komponenten des Teilbaumgrößenprofils nachgewiesen. Dabei kommen im Wesentlichen erzeugende Funktionen und analytische Methoden zum Einsatz.

Das Verhalten dieser Komponenten lässt sich vielleicht auch aus einer Betrachtung des Eisbergprozesses für sehr kleine Argumente begründen.

Da die Asymptotik der Eisberge abgesehen von wenigen zusätzlichen Voraussetzungen nur auf einer Verteilungsrekursion basiert, lässt sich das Resultat (Satz 4.17) auf verschiedene Strukturen übertragen. DST, Median-of-$(2k+1)$-Quicksort und Random Recursive Tree sind dabei nur ein paar Beispiele – weitere findet man im großen Gebiet der stochastischen Algorithmen zuhauf.

Literaturverzeichnis

[ABN98] Barry C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *Records.* John Wiley & Sons Inc., New York, 1998.

[Bau74] Heinz Bauer. *Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie.* Walter de Gruyter & Co., Berlin, second edition, 1974.

[BD76] Peter J. Bickel and Kjell A. Doksum. *Mathematical statistics. Basic Ideas and Selected Topics.* Holden-Day Inc., San Francisco, Calif., 1976.

[Bil68] Patrick Billingsley. *Convergence of Probability Measures.* John Wiley & Sons Inc., New York, 1968.

[Bil86] Patrick Billingsley. *Probability and Measure.* John Wiley & Sons Inc., New York, second edition, 1986.

[Cra00] Erhard Cramer. Asymptotic estimators of the sample size in a record model. *Statist. Papers*, 41(2):159–171, 2000.

[Dev91] Luc Devroye. Limit laws for local counters in random binary search trees. *Random Structures Algorithms*, 2(3):303–315, 1991.

[DG07] Florian Dennert and Rudolf Grübel. Renewals for exponentially increasing lifetimes, with an application to digital search trees. *Ann. Appl. Probab.*, 17(2):676–687, 2007.

[DG09] Florian Dennert and Rudolf Grübel. Estimation of search tree size and approximate counting: a likelihood approach. 2009. In Vorbereitung.

[Fel71] William Feller. *An Introduction to Probability Theory and its Applications. Vol. II.* Second edition. John Wiley & Sons Inc., New York, 1971.

[FGM97] Philippe Flajolet, Xavier Gourdon, and Conrado Martínez. Patterns in random binary search trees. *Random Structures Algorithms*, 11(3):223–244, 1997.

[Fla85] Philippe Flajolet. Approximate counting: a detailed analysis. *BIT*, 25(1):113–134, 1985.

[FMP08] Qunqiang Feng, Hosam M. Mahmoud, and Alois Panholzer. Phase changes in subtree varieties in random recursive and binary search trees. *SIAM J. Discrete Math.*, 22(1):160–184, 2008.

[Fuc07] Michael Fuchs. Subtree sizes in recursive trees and binary search trees: Berry-Esseen bound and Poisson approximation. *to appear*, 2007+.

[Jan02] Svante Janson. Ideals in a forest, one-way infinite binary trees and the contraction method. In *Mathematics and computer science, II (Versailles, 2002)*, Trends Math., pages 393–414. Birkhäuser, Basel, 2002.

[JK69] Norman L. Johnson and Samuel Kotz. *Distributions in Statistics: Discrete Distributions.* Houghton Mifflin Co., Boston, Mass., 1969.

[KG71] J. Keilson and H. Gerber. Some results for Discrete Unimodality. *Journal of the American Statistical Association*, 66(334):386–389, 1971.

[Knu73] Donald E. Knuth. *The Art of Computer Programming. Volume 3: Sorting and Searching.* Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., 1973.

[Knu97] Donald E. Knuth. *The Art of Computer Programming. Volume 1: Fundamental Algorithms.* Addison-Wesley Publishing Co., Reading, Mass.-London-Amsterdam, third edition, 1997.

[KP91] Peter Kirschenhofer and Helmut Prodinger. Approximate counting: an alternative approach. *RAIRO Inform. Théor. Appl.*, 25(1):43–48, 1991.

[Lou87] Guy Louchard. Exact and asymptotic distributions in digital and binary search trees. *RAIRO Inform. Théor. Appl.*, 21(4):479–495, 1987.

[LP08] Guy Louchard and Helmut Prodinger. Generalized approximate counting revisited. *Theoret. Comput. Sci.*, 391(1-2):109–125, 2008.

[Mah91] Hosam M. Mahmoud. *Evolution of Random Search Trees.* John Wiley & Sons, Inc., New York, 1991.

[Mah00] Hosam M. Mahmoud. *Sorting. A Distribution Theory.* Wiley-Interscience, New York, 2000.

[MBCB96] J. L. Moreno Rebollo, I. Barranco Chamorro, F. López Blázquez, and T. Gómez Gómez. On the estimation of the unknown sample size from the number of records. *Statist. Probab. Lett.*, 31(1):7–12, 1996.

[Mor78] Robert Morris. Counting large numbers of events in small registers. *Commun. ACM*, 21(10):840–842, 1978.

[MR87] Makoto Maejima and Svetlozar T. Rachev. An ideal metric and the rate of convergence to a self-similar process. *Ann. Probab.*, 15(2):708–727, 1987.

[Nei01] Ralph Neininger. On a multivariate contraction method for random recursive structures with applications to Quicksort. *Random Structures Algorithms*, 19(3-4):498–524, 2001.

[Nei02] Ralph Neininger. The Wiener index of random trees. *Combin. Probab. Comput.*, 11(6):587–597, 2002.

[NR04] Ralph Neininger and Ludger Rüschendorf. A general limit theorem for recursive algorithms and combinatorial structures. *Ann. Appl. Probab.*, 14(1):378–418, 2004.

[NR06] Ralph Neininger and Ludger Rüschendorf. A survey of multivariate aspects of the contraction method. *Discrete Math. Theor. Comput. Sci.*, 8(1):31–56 (electronic), 2006.

[Pro94] Helmut Prodinger. Approximate counting via Euler transform. *Math. Slovaca*, 44(5):569–574, 1994.

[Rac91] Svetlozar T. Rachev. *Probability Metrics and the Stability of Stochastic Models.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Ltd., Chichester, 1991.

[Rös91] Uwe Rösler. A limit theorem for "Quicksort". *RAIRO Inform. Théor. Appl.*, 25(1):85–100, 1991.

[RR95] Svetlozar T. Rachev and Ludger Rüschendorf. Probability metrics and recursive algorithms. *Adv. in Appl. Probab.*, 27(3):770–799, 1995.

[RR98] Svetlozar T. Rachev and Ludger Rüschendorf. *Mass Transportation Problems. Vol. I.* Springer-Verlag, New York, 1998. Theory.

[RR01] Uwe Rösler and Ludger Rüschendorf. The contraction method for recursive algorithms. *Algorithmica*, 29(1-2):3–33, 2001. Average-case analysis of algorithms (Princeton, NJ, 1998).

[SF96] Robert Sedgewick and Philippe Flajolet. *An Introduction to the Analysis of Algorithms.* Addison Wesley, 1996.

[Str80] Gilbert Strang. *Linear Algebra and its Applications.* Academic Press, New York, second edition, 1980.

Danksagung

Mein besonderer Dank gilt meinem Doktorvater Herrn Prof. Dr. Grübel für die intensive Betreuung während der Entstehung dieser Arbeit.

Ebenso möchte ich mich bei allen Mitarbeiterinnen und Mitarbeitern des Instituts bedanken für die allzeit angenehme Arbeitsatmosphäre.

Außerdem sei Herrn Prof. Dr. Neininger für die Übernahme des Korreferats sehr herzlich gedankt.