

Masterarbeit

Asymptotische Eigenschaften von Hoppe-Bäumen

von

Kevin Leckey

Matrikelnr: 3489615

Wintersemester 2011/2012

| | |
|--------------------------|---|
| Universität: | J. W. Goethe-Universität |
| Fachbereich: | 12 Informatik und Mathematik Institut für Mathematik |
| Betreuer/Erstkorrekteur: | Prof. Dr. Ralph Neininger |
| Zweitkorrekteur: | Prof. Dr. Anton Wakolbinger |

Inhaltsverzeichnis

| | | |
|----------|--|-----------|
| 1 | Einleitung | 3 |
| | Definition des Hoppe-Baumes | 3 |
| | Asymptotik verallgemeinerter harmonischer Reihen | 6 |
| 2 | Die Tiefe eines Knotens im Hoppe-Baum | 10 |
| | Die Tiefe als Summe unabhängiger Zufallsvariablen | 13 |
| | Erwartungswert, Varianz und Grenzwertsatz für die Tiefe | 13 |
| 3 | Die Höhe des Hoppe-Baumes | 14 |
| | Konzentration der Höhe | 14 |
| | Erwartungswert und Varianz der Höhe | 17 |
| 4 | Die Anzahl der Blätter im Hoppe-Baum | 19 |
| | Erwartete Anzahl der Blätter | 20 |
| | Konzentrationsungleichung für die Anzahl der Blätter | 20 |
| | Starkes Gesetz großer Zahlen für die Anzahl der Blätter | 22 |
| | Varianz der Anzahl der Blätter | 23 |
| | Zentraler Grenzwertsatz für die Anzahl der Blätter | 25 |
| 5 | Die interne Pfadlänge im Hoppe-Baum | 27 |
| 5.1 | Konvergenz der internen Pfadlänge mit Martingalmethoden | 28 |
| | Die Reskalierung der internen Pfadlänge zu einem Martingal | 28 |
| | Erwartungswert der internen Pfadlänge | 28 |
| | Varianz der internen Pfadlänge | 29 |
| | Fast sichere Konvergenz der reskalierten internen Pfadlänge gegen ein $Z_\infty^{(\vartheta)}$ | 30 |
| 5.2 | Die Verteilung des Grenzwertes $Z_\infty^{(\vartheta)}$ | 30 |
| | Analyse des Betrags der charakteristischen Funktion von $Z_\infty^{(\vartheta)}$ | 35 |
| | Existenz einer \mathcal{C}^∞ Dichte für die Verteilung des Grenzwertes $Z_\infty^{(\vartheta)}$ | 38 |
| | Literatur | 39 |

Danksagung

Ich möchte an dieser Stelle allen danken, die mich während meines Studiums unterstützt haben. Mein ganz besonderer Dank gilt Prof. Neininger für die hervorragende Betreuung bei dieser Arbeit und die Förderung und Betreuung während meiner Studienzeit. Auch Henning Sulzbach möchte ich an dieser Stelle danken. Auch wenn er an der Entstehung dieser Arbeit nicht beteiligt war, so war er doch stets eine hilfsbereite und kompetente Anlaufstelle für Fragen jeglicher Art. Ich möchte auch allen weiteren Dozenten und meinen Kommilitonen für die schöne und lehrreiche Studienzeit danken.

Zuletzt möchte ich noch meiner Familie und ganz besonders meinen Eltern danken, die mir dieses Studium ermöglicht und mich durchgängig unterstützt haben.

1 Einleitung

Wir werden uns in dieser Arbeit mit der Analyse einer zufälligen Baumstruktur befassen, welche wir Hoppe-Baum nennen werden. Als ersten Schritt starten wir mit der Definition des Baumes und Zusammenhängen zu dem zufälligen rekursiven Baum, der schon viel untersucht wurde. Man findet einige Ergebnisse zum zufälligen rekursiven Baum zusammengefasst in dem Übersichtsartikel [12] von Smythe und Mahmoud.

Definition des Hoppe-Baumes

Zunächst meinen wir im Kontext dieser Arbeit mit einem Baum einen zusammenhängenden, kreisfreien Graphen mit einem ausgezeichneten Knoten genannt Wurzel. Als Abstand zweier Knoten in einem Baum bezeichnen wir die Anzahl der Kanten des Pfades zwischen den Knoten. Anhand der Wurzel lässt sich der Baum in Level unterteilen: Auf Level $k \in \mathbb{N}$ liegen genau die Knoten, die Abstand k zur Wurzel haben. Ebenso können wir anhand der Wurzel Vorfahren und Nachfahren definieren: Wir sagen für zwei Knoten i, j , $i \neq j$, im Baum, dass i Vorfahr von j ist (und j Nachfahr von i), falls sich i auf dem Pfad von der Wurzel zu j befindet. Wir nennen außerdem i Vater von j (und j Kind von i), falls i ein Vorfahr von j ist und i und j Abstand 1 haben.

Der Hoppe-Baum ist ein zufälliger Baum, dessen Verteilung von einem Parameter $\vartheta \in \mathbb{R}^+$ abhängt. Die Dynamik des Baumes ist hierbei durch die Dynamik der Hoppe-Urne gegeben. In dem Urnenmodell gibt es eine ausgezeichnete schwarze Kugel und weitere farbige Kugeln. Wann immer die schwarze Kugel gezogen wird, wird sie zusammen mit einer farbigen Kugel zurückgelegt, deren Farbe in der Urne noch nicht vertreten ist. Wird eine farbige Kugel gezogen, wird sie zusammen mit einer weiteren Kugel derselben Farbe zurückgelegt. Die Urne startet nun mit nur der schwarzen Kugel und es wird in jedem Schritt eine Kugel zufällig aus der Urne gezogen, wobei die Wahrscheinlichkeit, die schwarze Kugel zu ziehen, proportional zu ϑ ist, alle weiteren Kugeln werden mit einer Wahrscheinlichkeit proportional zu 1 gezogen.

Im Hoppe-Baum übernimmt nun die Wurzel des Baumes die Rolle der schwarzen Kugel und wir erhalten analog zu der Urne folgende Dynamik:

Wir starten den Baum mit der Wurzel, die wir Knoten 1 nennen. Dieser Knoten erhält Gewicht ϑ . Die weitere Verteilung des Baumes wird nun rekursiv über die Anzahl $n \in \mathbb{N}$ der Knoten im Baum definiert. Wir erhalten einen Hoppe-Baum mit n Knoten aus einem Baum mit $n - 1$ Knoten wie folgt: Wähle einen der vorhandenen $n - 1$ Knoten zufällig als Vaterknoten aus, wobei Knoten 1 (also die Wurzel) mit Wahrscheinlichkeit $\frac{\vartheta}{\vartheta+n-2}$ und alle weiteren Knoten mit Wahrscheinlichkeit $\frac{1}{\vartheta+n-2}$ ausgewählt werden. Knoten n wird als Kind des ausgewählten Knotens in den Baum eingefügt. Typischerweise einigen wir uns darauf, dass die Kinder eines Knotens immer von links nach rechts gemäß Einfügezeit sortiert sind, was für diese Arbeit allerdings allein in der Darstellung der Grafiken eine Rolle spielt. In dem Hoppe-Baum hat also die Wurzel Gewicht ϑ und alle weiteren Knoten Gewicht 1, für das Einfügen neuer Knoten wird der Vaterknoten proportional zu den Gewichten gezogen.

Wir benennen die Knoten im Baum gemäß ihrer Einfügezeit, d.h. Knoten i ist der Knoten, der als i -tes in den Baum eingefügt wurde.

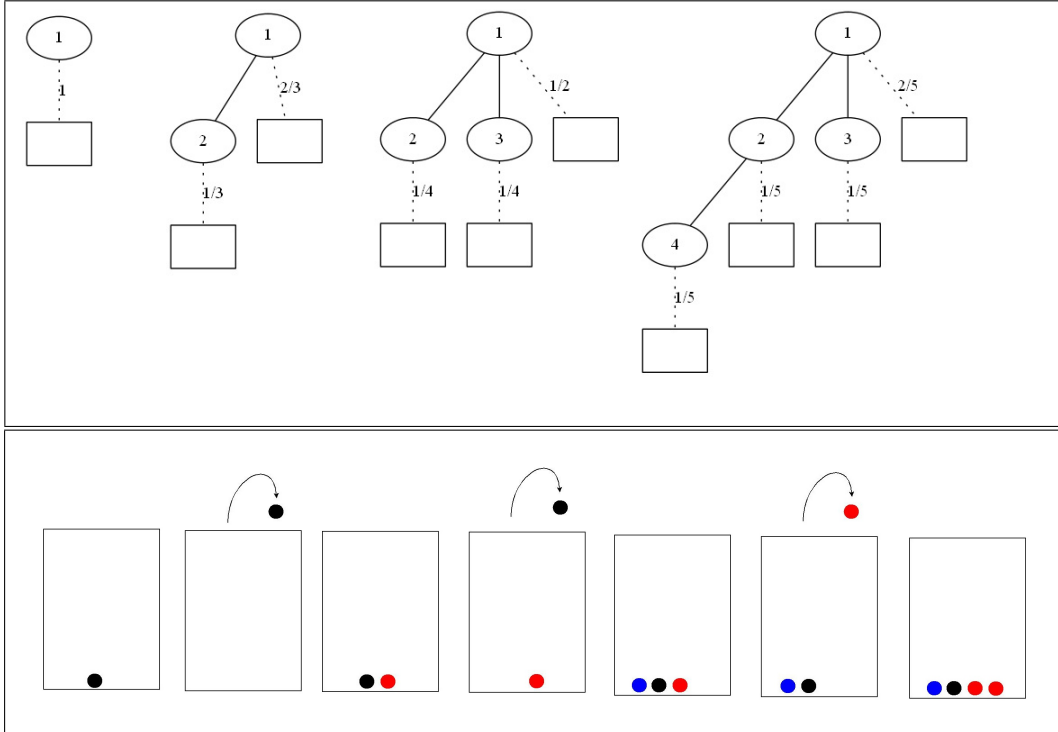


Abbildung 1: Eine mögliche Entwicklung des Hoppe-Baumes bis $n = 4$ Knoten mit Wurzelgewicht $\vartheta = 2$. Darunter die entsprechende Entwicklung der Hoppe-Urne.

Abbildung 1 zeigt ein Beispiel eines Hoppe-Baumes mit Wurzelgewicht $\vartheta = 2$ und $n = 1, \dots, 4$ Knoten. Die rechteckigen Knoten deuten die möglichen Positionen für das Einfügen des nächsten Knotens an, an den Kanten stehen die Wahrscheinlichkeiten für das Einfügen an die entsprechende Stelle. Darunter wird die Entwicklung der entsprechenden Hoppe-Urne abgebildet.

Man beachte, dass die Anzahl der Kinder der Wurzel im Hoppe-Baum genau der Anzahl verschiedener Farben in der Urne entspricht. Bezeichnet K_n diese Anzahl, so erhält man eine leicht zu analysierende Darstellung für K_n , nämlich $K_n = \sum_{i=2}^n \mathbf{1}_{A_i}$, wobei A_i das Ereignis bezeichnet, dass der i -te Knoten Kind der Wurzel ist. Offensichtlich sind A_2, \dots, A_n unabhängig und $P(A_i) = \frac{\vartheta}{\vartheta+i-2}$. Des Weiteren stehen die Größen der Teilbäume an der Wurzel in direkter Beziehung zu der Anzahl von Kugeln entsprechender Farbe in der Urne. Aus der Analyse der Hoppe-Urne von Donnelly und Tavaré in [5] erhalten wir also bereits das folgende Ergebnis für die Größen der Teilbäume.

Theorem 1.1. *Seien $N_n^{(1)}, N_n^{(2)}, \dots$ die Größen der Teilbäume der Kinder der Wurzel, aufsteigend nach Einfügezeit des jeweiligen Kindes nummeriert, wobei wir $N_n^{(k)} = 0$ für $k > K_n$ vereinbaren. Dann gilt für jedes feste $r \in \mathbb{N}$*

$$\left(\frac{N_n^{(1)}}{n}, \frac{N_n^{(2)}}{n}, \dots, \frac{N_n^{(r)}}{n} \right) \xrightarrow{d} \left(B_1, B_2(1 - B_1), \dots, B_r \prod_{i=1}^{r-1} (1 - B_i) \right) \text{ für } n \rightarrow \infty,$$

wobei B_1, \dots, B_r unabhängig und identisch $\text{beta}(1, \vartheta)$ -verteilt sind. Insbesondere gilt also

$$\frac{N_n^{(1)}}{n} \xrightarrow{d} B,$$

wobei B $\text{beta}(1, \vartheta)$ -verteilt ist.

Wir werden im weiteren Verlauf das Wurzelgewicht ϑ oftmals geklammert als oberen Index notieren, an dieser Stelle ist also zu beachten, dass bei $N_n^{(1)}, \dots, N_n^{(K_n)}$ die Zahl (k) im oberen Index *nicht* das Gewicht der Wurzel im Hoppe-Baum bezeichnet, sondern die Nummer des Teilbaumes. Das Gewicht der Wurzel wird an dieser Stelle unterschlagen, um nicht noch mehr Indizes einzuführen.

Bemerkung 1.2. *Man erhält sogar $\frac{N_n^{(1)}}{n} \rightarrow B$ fast sicher, da $\left(\frac{N_n^{(1)}}{\vartheta+n-1}\right)_{n \geq 2}$ ein beschränktes, nichtnegatives Martingal ist.*

Der zufällige rekursive Baum und sein Zusammenhang zum Hoppe-Baum

Den zufälligen rekursiven Baum kann man auf zwei Arten einführen. Zum einen ist ein zufälliger rekursiver Baum mit n Knoten gleichverteilt unter alle rekursiven Bäumen mit n Knoten (vergleiche [12]), auf diese Art der Definition werden wir an dieser Stelle aber nicht näher eingehen. Für uns interessanter ist die äquivalente, auf einer stochastischen Dynamik basierende Definition. Beim Einfügen eines neuen Knotens wird uniform einer der vorhandenen Knoten als Vater für den neuen Knoten ausgewählt. Der zufällige rekursive Baum ist demnach ein Hoppe-Baum zum Parameter $\vartheta = 1$. Alle Ergebnisse aus dieser Arbeit gelten bei Wahl $\vartheta = 1$ also auch für zufällige rekursiven Bäume und sind für zufällige rekursive Bäume bereits bekannt. Das Ziel der Arbeit besteht in erster Linie darin, Ergebnisse über zufällige rekursive Bäume auf die Hoppe-Bäume zu verallgemeinern.

Wir finden aber auch für beliebige $\vartheta \in \mathbb{R}^+$ den zufälligen rekursiven Baum in dem Hoppe-Baum wieder. Besonders für Kapitel 3 und 5 ist die folgende Zerlegung des Hoppe-Baumes sehr interessant: Wir betrachten den ersten Teilbaum des Hoppe-Baumes und den Rest des Baumes getrennt. Mit dem ersten Teilbaum meinen wir den Baum, der aus Knoten 2 und all seinen Nachfahren besteht. In Abbildung 2 ist die Zerlegung für einen Hoppe-Baum mit 10 Knoten angedeutet.

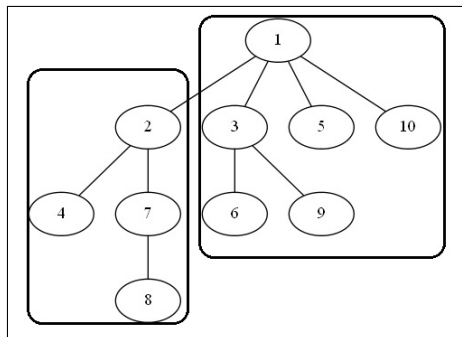


Abbildung 2: Die Zerlegung eines Hoppe-Baumes in den ersten Teilbaum und den Restbaum. Im ersten Teilbaum (linker Rahmen) haben alle Knoten Gewicht 1, im Restbaum hat die Wurzel weiterhin Gewicht ϑ und alle weiteren Knoten Gewicht 1.

Zu beachten bei dieser Zerlegung ist, dass alle Knoten im ersten Teilbaum Gewicht 1 haben.

Fügen wir also einen neuen Knoten in den Teilbaum ein, so wird für diesen uniform aus allen Knoten des Teilbaums ein Vaterknoten ausgewählt. Bedingt auf die Größe des Teilbaums, ist die Teilbaumstruktur die eines zufälligen rekursiven Baumes (natürlich abgesehen von der Nummerierung der Knoten). Der Restbaum in der Zerlegung besteht aus der ursprünglichen Wurzel des Hoppe-Baumes und alle Knoten, die nicht im ersten Teilbaum liegen. Hier hat die Wurzel weiterhin Gewicht ϑ und alle weiteren Knoten Gewicht 1, also ist der Restbaum bedingt auf seine Größe wieder ein Hoppe-Baum zum Parameter ϑ .

Größen des Hoppe-Baumes, die in dieser Arbeit betrachtet werden

Ziel der Arbeit ist es, einen groben Überblick über ein paar charakteristische Größen im Hoppe-Baum zu gewinnen. In Kapitel 2 befassen wir uns mit der Tiefe der Knoten im Hoppe-Baum, also den Abstand der Knoten zur Wurzel. Wir werden eine Darstellung der Tiefe als Summe unabhängiger Zufallsvariablen herleiten, welche die Tiefe zu einer leicht zu analysierenden Größe macht. Danach befassen wir uns mit der Höhe des Hoppe-Baumes, womit das Maximum aller Tiefen gemeint ist. Zentrale Idee bei der Analyse der Höhe wird es sein, die bekannten Ergebnisse für den zufälligen rekursiven Baum auf den Hoppe-Baum zu übertragen, indem wir ihn, wie in der Einleitung bereits angesprochen, geeignet zerlegen. Anschließend werden wir die Anzahl der Blätter im Baum betrachten, wobei wir einen Knoten als Blatt bezeichnen, wenn er keine Kinder besitzt. Die zentrale Idee bei der Analyse der Anzahl der Blätter wird es sein, diese Größe zu einem Martingal zu reskalieren. Das letzte Kapitel handelt dann von der internen Pfadlänge, welche wir als Summe aller Tiefen im Baum definieren. Die interne Pfadlänge kann in gewissem Kontext die Gesamtkomplexität des Baumes beschreiben. Sie wurde insbesondere für zufällige Binärsuchbäume studiert, da die interne Pfadlänge hier die Anzahl der Vergleiche, die der Sortieralgorithmus Quicksort benötigt, ausdrückt (vergleiche [6] oder [9]). Auch hier werden einige Ergebnisse durch Reskalierung zu einem Martingal erzielt, insbesondere folgern wir, dass die interne Pfadlänge nach Reskalierung fast sicher konvergiert. Im zweiten Teil des letzten Kapitels wird die Verteilung des Grenzwertes betrachtet, Hauptergebnis hier ist die Existenz einer unendlich oft differenzierbaren, schnell abfallenden Dichte.

Als Notation werden wir typischerweise die betrachtete Anzahl der Knoten in den unteren Index der Zufallsvariablen schreiben und zudem den Parameter ϑ geklammert im oberen Index notieren. Wir bezeichnen im Hoppe-Baum mit $n \in \mathbb{N}$ Knoten und Wurzelgewicht $\vartheta \in \mathbb{R}^+$ mit $D_n^{(\vartheta)}$ die Tiefe des n -ten Knotens, $H_n^{(\vartheta)}$ die Höhe des Baumes, $L_n^{(\vartheta)}$ die Anzahl der Blätter und $I_n^{(\vartheta)} = \sum_{i=1}^n D_i^{(\vartheta)}$ die interne Pfadlänge.

Wir starten an dieser Stelle mit einem kleinen Lemma, um die Asymptotik zweier Summen zu klären, die im Verlauf der Arbeit öfters auftreten werden. Anschließend werden die Ergebnisse der Arbeit aufgelistet.

Lemma 1.3. Sei $\vartheta > 0$ beliebig. Dann gilt

$$\begin{aligned}\sum_{i=1}^{n-1} \frac{1}{\vartheta+i} &= \Psi(\vartheta+n) - \Psi(\vartheta+1) = \log n + c_\vartheta + o(1), \\ \sum_{i=1}^{n-1} \left(\frac{1}{\vartheta+i} \right)^2 &= d_\vartheta + o(1),\end{aligned}$$

wobei $\Psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}$ die Digamma Funktion bezeichnet und c_ϑ und d_ϑ Konstanten sind, für die Folgendes gilt: Für $\vartheta \in \mathbb{N}$ ist

$$\begin{aligned}c_\vartheta &= \gamma - \sum_{i=1}^{\vartheta} \frac{1}{i}, \\ d_\vartheta &= \frac{\pi^2}{6} - \sum_{i=1}^{\vartheta} \frac{1}{i^2},\end{aligned}$$

wobei $\gamma = \lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{1}{i} - \log n \approx 0,5772156649$ die Euler Konstante bezeichnet. Für beliebige $\vartheta > 0$ gilt

$$\begin{aligned}c_\vartheta &= -\Psi(\vartheta+1) = \gamma - \sum_{i=1}^{\lfloor \vartheta \rfloor} \frac{1}{i} - \varepsilon_\vartheta, \\ d_\vartheta &= \frac{\pi^2}{6} - \sum_{i=1}^{\lfloor \vartheta \rfloor} \frac{1}{i^2} - \delta_\vartheta\end{aligned}$$

mit $\varepsilon_\vartheta \in [0, \frac{1}{\lfloor \vartheta \rfloor}]$, $\delta_\vartheta \in [0, \frac{1}{\lfloor \vartheta \rfloor^2}]$.

Beweis. Siehe Ende der Einleitung. □

Zentrales Ergebnis für die Tiefe ist Theorem 1.4. Einige Folgerungen sind in Theorem 1.5 aufgelistet. Insbesondere sieht man in 1.5, dass der Parameter ϑ nur geringen Einfluss auf Erwartungswert und Varianz der Tiefe hat.

Theorem 1.4. Sei $D_n^{(\vartheta)}$ die Tiefe des n -ten Knotens im Hoppe-Baum. Dann gilt für jedes $n \geq 2$

$$D_n^{(\vartheta)} \stackrel{d}{=} 1 + \sum_{i=1}^{n-2} B_{\frac{1}{\vartheta+i}},$$

wobei $B_{\frac{1}{\vartheta+1}}, \dots, B_{\frac{1}{\vartheta+n-2}}$ unabhängig sind und $B_{\frac{1}{\vartheta+i}}$ Bernoulli($\frac{1}{\vartheta+i}$)-verteilt ist für jedes $i \in \{1, \dots, n-2\}$.

Theorem 1.5. Für die Tiefe $D_n^{(\vartheta)}$ des n -ten Knotens im Hoppe-Baum gilt

$$\begin{aligned}E[D_n^{(\vartheta)}] &= 1 + \sum_{i=1}^{n-2} \frac{1}{\vartheta+i} = \log n + c_\vartheta + 1 + o(1), \\ \text{Var}(D_n^{(\vartheta)}) &= \sum_{i=1}^{n-2} \frac{1}{\vartheta+i} - \sum_{i=1}^{n-2} \left(\frac{1}{\vartheta+i} \right)^2 = \log n + c_\vartheta - d_\vartheta + o(1), \\ \mathcal{L}(D_n^*) &\xrightarrow{w} \mathcal{N}(0, 1), \\ d_{TV}(\mathcal{L}(D_n^{(\vartheta)}), \Pi_{E[D_n^{(\vartheta)}]}) &= \mathcal{O}\left(\frac{1}{\log n}\right),\end{aligned}$$

wobei c_ϑ, d_ϑ die Konstanten aus Lemma 1.3 sind, $D_n^* = \frac{D_n^{(\vartheta)} - E[D_n^{(\vartheta)}]}{\sqrt{\text{Var}(D_n^{(\vartheta)})}}$ ist, $\mathcal{N}(0, 1)$ die Standardnormalverteilung und $\Pi_{E[D_n^{(\vartheta)}]}$ die Poissonverteilung zum Parameter $E[D_n^{(\vartheta)}]$ bezeichnet.

Für die Analyse der Höhe wird Theorem 1.6 erarbeitet. In Korollar 1.7 sind die Konsequenzen für Erwartungswert und Varianz der Höhe aufgelistet. Wie auch bei der Tiefe hat die Konstante ϑ nur wenig Einfluss auf den Erwartungswert der Höhe.

Theorem 1.6. Sei $H_n^{(\vartheta)}$ die Höhe des Hoppe-Baumes mit n Knoten sowie $M_n = E[H_n^{(1)}] = e \log n - \frac{3}{2} \log \log n + \mathcal{O}(1)$ die erwartete Höhe des zufälligen rekursiven Baumes mit n Knoten. Dann gibt es für jedes $c' < \frac{1}{2e}$ eine Konstante $C = C(c')$, so dass für beliebige $t \geq 2$ gilt

$$P\left(H_n^{(\vartheta)} - M_n \geq t\right) \leq C e^{-c't}.$$

Bezeichnet $K = \sup_{n \in \mathbb{N}} \{|M_n - (e \log n - \frac{3}{2} \log \log n)|\} < \infty$, so gibt es außerdem für jedes $c' < \frac{1}{2e}$ eine Konstante $\tilde{C} = \tilde{C}(c')$ und $\alpha = \frac{c'}{1+c'e}$, sodass für alle $t \geq \frac{2K+1}{1-e\alpha}$ gilt

$$P\left(H_n^{(\vartheta)} - M_n \leq -t\right) \leq \tilde{C} e^{-\alpha t}.$$

Korollar 1.7. Sei $H_n^{(\vartheta)}$ die Höhe des Hoppe-Baumes mit n Knoten sowie $M_n = E[H_n^{(1)}] = e \log n - \frac{3}{2} \log \log n + \mathcal{O}(1)$. Ferner seien $c', C, \tilde{C}, \alpha > 0$ wie in Theorem 1.6 gewählt. Dann gilt

$$|E[H_n^{(\vartheta)}] - M_n| \leq \max \left\{ 2 + \frac{C}{c' e^{2c'}}, \frac{2K+1}{1-e\alpha} + \frac{\tilde{C}}{\alpha \exp\left(\frac{(2K+1)\alpha}{1-e\alpha}\right)} \right\}.$$

Insbesondere gilt $E[H_n^{(\vartheta)}] = e \log n - \frac{3}{2} \log \log n + \mathcal{O}(1)$.

Des Weiteren gilt

$$\text{Var}(H_n^{(\vartheta)}) = \mathcal{O}(1).$$

Zentrale Idee bei der Analyse der Anzahl an Blättern ist die Reskalierung zu einem Martingal, die in Theorem 1.8 angegeben ist. Folgerungen sind unter anderem der Erwartungswert, eine Konzentrationsungleichung und ein starkes Gesetz großer Zahlen. Für die weitere Analyse wird noch die Varianz berechnet und eine Grenzwertsatz gezeigt.

Theorem 1.8. Sei $L_n = L_n^{(\vartheta)}$ die Anzahl an Blättern im Hoppe-Baum mit n Knoten.

Dann liefern $X_1 = 0$ und

$$X_n = (\vartheta + n - 2) \left(L_n - \left(\frac{n-1}{2} + \frac{\vartheta(n-1)}{2(\vartheta+n-2)} \right) \right)$$

für $n \geq 2$ ein Martingal $(X_n)_{n \geq 1}$ bezüglich der natürlichen Filtration $(\mathcal{F}_n)_{n \geq 1}$ gegeben durch $\mathcal{F}_n = \sigma(L_1, \dots, L_n)$ für $n \geq 1$.

Korollar 1.9. Für die Anzahl $L_n = L_n^{(\vartheta)}$ der Blätter im Hoppe-Baum mit n Knoten gilt für $n \geq 2$

$$E[L_n] = \frac{n-1}{2} + \frac{\vartheta(n-1)}{2(\vartheta+n-2)} = \frac{n+\vartheta-1}{2} + \mathcal{O}\left(\frac{1}{n}\right).$$

Theorem 1.10. Für die Anzahl der Blätter $L_n = L_n^{(\vartheta)}$ im Hoppe-Baum mit $n \geq 3$ Knoten gilt für jedes $t > 0$

$$P(|L_n - E[L_n]| \geq t) \leq 2 \exp\left(-\frac{2t^2(\vartheta + n - 2)^2}{\sum_{i=3}^n (\vartheta + 2i - 5)^2}\right).$$

Insbesondere erhalten wir also für jedes $t > 0$ die Schranke

$$P(|L_n - E[L_n]| \geq t) \leq 2 \exp\left(-\frac{t^2}{2n}\right).$$

Theorem 1.11. Für die Anzahl der Blätter $L_n = L_n^{(\vartheta)}$ im Hoppe-Baum mit n Knoten gilt fast sicher

$$\frac{L_n}{n} \rightarrow \frac{1}{2} \quad \text{für } n \rightarrow \infty.$$

Theorem 1.12. Sei $L_n^{(\vartheta)}$ die Anzahl der Blätter im Hoppe-Baum mit n Knoten. Dann gilt

$$\text{Var}\left(L_n^{(\vartheta)}\right) = \frac{\vartheta + n - 1}{12} + \mathcal{O}\left(\frac{1}{n}\right).$$

Theorem 1.13. Sei $L_n^{(\vartheta)}$ die Anzahl der Blätter im Hoppe-Baum mit n Knoten. Dann gilt

$$\frac{L_n^{(\vartheta)} - E[L_n^{(\vartheta)}]}{\sqrt{\text{Var}(L_n^{(\vartheta)})}} \xrightarrow{d} Z,$$

wobei Z standardnormalverteilt ist.

Auch bei der Analyse der internen Pfadlänge starten wir mit der Reskalierung zu einem Martingal. Anschließend bestimmen wir die Varianz der internen Pfadlänge und folgern für das Martingal fast sichere Konvergenz. Im zweiten Teil des Kapitels über die interne Pfadlänge betrachten wir die Verteilung des Grenzwertes genauer und zeigen die Existenz einer unendlich oft differenzierbaren, schnell abfallenden Dichte über die Analyse der charakteristischen Funktion des Grenzwertes.

Theorem 1.14. Sei für $i \geq 1$ $D_i^{(\vartheta)}$ die Tiefe des i -ten Knotens im Hoppe-Baum, sowie $I_n^{(\vartheta)} = \sum_{i=1}^n D_i^{(\vartheta)}$ die interne Pfadlänge des Hoppe-Baumes mit n Knoten. Ferner sei für $n \geq 1$ $\mathcal{F}_n = \sigma(D_1^{(\vartheta)}, \dots, D_n^{(\vartheta)})$ die von $D_1^{(\vartheta)}, \dots, D_n^{(\vartheta)}$ erzeugte σ -Algebra. Dann ist durch

$$Z_n^{(\vartheta)} = \frac{1}{\vartheta + n - 1} I_n^{(\vartheta)} - \sum_{i=1}^{n-1} \frac{1}{\vartheta + i}$$

ein an $(\mathcal{F}_n)_{n \in \mathbb{N}}$ adaptiertes Martingal $(Z_n^{(\vartheta)})_{n \in \mathbb{N}}$ gegeben.

Korollar 1.15. Sei $I_n^{(\vartheta)} = \sum_{i=1}^n D_i^{(\vartheta)}$ die interne Pfadlänge des Hoppe-Baumes mit n Knoten. Dann gilt

$$E[I_n^{(\vartheta)}] = (\vartheta + n - 1) \sum_{i=1}^{n-1} \frac{1}{\vartheta + i} = n \log n + c_\vartheta n + o(n),$$

wobei $c_\vartheta = -\Psi(\vartheta + 1) = \gamma - \sum_{i=1}^{\lfloor \vartheta \rfloor} \frac{1}{i} - \varepsilon_\vartheta$ für ein $\varepsilon_\vartheta \in [0, \frac{1}{\lfloor \vartheta \rfloor}]$ die Konstante aus Lemma 1.3 bezeichnet.

Theorem 1.16. Sei $I_n^{(\vartheta)}$ die interne Pfadlänge des Hoppe-Baumes mit n Knoten sowie $Z_n^{(\vartheta)} = \frac{1}{\vartheta+n-1} I_n^{(\vartheta)} - \sum_{i=1}^{n-1} \frac{1}{\vartheta+i}$ das Martingal aus Theorem 5.1. Dann gilt

$$\text{Var}(Z_n^{(\vartheta)}) = \frac{2}{\vartheta} - \sum_{i=1}^{n-1} \left(\frac{1}{\vartheta+i-1} \right)^2 + o(1).$$

Insbesondere gilt also auch

$$\text{Var}(I_n^{(\vartheta)}) = \left(\frac{2}{\vartheta} - \sum_{i=1}^{n-1} \left(\frac{1}{\vartheta+i-1} \right)^2 \right) (\vartheta+n-1)^2 + o(n^2).$$

Korollar 1.17. Die reskalierte interne Pfadlänge $(Z_n^{(\vartheta)})_{n \geq 1}$ mit $Z_n^{(\vartheta)} = \frac{I_n^{(\vartheta)} - E[I_n^{(\vartheta)}]}{\vartheta+n-1}$ ist ein gleichgradig integrierbares Martingal und konvergiert somit fast sicher und in L_1 gegen eine integrierbare Zufallsvariable $Z_\infty^{(\vartheta)}$.

Theorem 1.18. Für $\vartheta > 0$ sei $Z_\infty^{(\vartheta)}$ der Grenzwert der reskalierten internen Pfadlänge $\left(\frac{I_n^{(\vartheta)} - E[I_n^{(\vartheta)}]}{\vartheta+n-1} \right)_{n \geq 1}$, welche nach Korollar 1.17 fast sicher konvergiert. Dann gilt

$$Z_\infty^{(\vartheta)} \stackrel{d}{=} B \bar{Z}_\infty^{(1)} + (1-B) \bar{\bar{Z}}_\infty^{(\vartheta)} + B \log B + (1-B) \log(1-B) + (\gamma - c_\vartheta) B,$$

wobei $B, \bar{Z}_\infty^{(1)}, \bar{\bar{Z}}_\infty^{(\vartheta)}$ unabhängig sind mit Verteilungen $\mathcal{L}(B) = \text{Beta}(1, \vartheta)$, $\mathcal{L}(\bar{Z}_\infty^{(1)}) = \mathcal{L}(Z_\infty^{(1)})$, $\mathcal{L}(\bar{\bar{Z}}_\infty^{(\vartheta)}) = \mathcal{L}(Z_\infty^{(\vartheta)})$, $\gamma \approx 0,5772156649$ die Euler Konstante ist und c_ϑ die Konstante aus Lemma 1.3 bezeichnet.

Theorem 1.19. Sei $\vartheta > 0$ beliebig und $Z_\infty^{(\vartheta)}$ der Grenzwert der reskalierten internen Pfadlänge $\left(\frac{I_n^{(\vartheta)} - E[I_n^{(\vartheta)}]}{\vartheta+n-1} \right)_{n \geq 1}$, welche nach Korollar 1.17 fast sicher konvergiert. Dann gibt es für die Verteilung $\mathcal{L}(Z_\infty^{(\vartheta)})$ eine unendlich oft differenzierbare, schnell abfallende Dichte f_ϑ .

Beweis von Lemma 1.3. Die Darstellung $\sum_{i=1}^{n-1} \frac{1}{\vartheta+i} = \Psi(\vartheta+n) - \Psi(\vartheta+1)$ und die Asymptotik $\Psi(\vartheta+n) = \log n + o(1)$ folgen direkt aus den Eigenschaften der Digamma Funktion, welche man bei Abramowitz und Stegun in [1, S. 259] findet ($\sum_{i=1}^{n-1} \frac{1}{\vartheta+i} = \Psi(\vartheta+n) - \Psi(\vartheta+1)$ folgt wegen $\Psi(z+1) = \Psi(z) + \frac{1}{z}$). Für die zweite Summe erhalten wir $\sum_{i=1}^{\infty} \left(\frac{1}{\vartheta+i} \right)^2 = d_\vartheta < \infty$ über die Majorante $\left(\frac{1}{\vartheta+i} \right)^2 < \frac{1}{i^2}$. Genauere Ergebnisse für die Konstanten c_ϑ und d_ϑ erhalten wir über die harmonischen Zahlen:

Bekanntermaßen gilt für die harmonischen Zahlen $\mathcal{H}_n = \sum_{i=1}^n \frac{1}{i}$, dass $\mathcal{H}_n = \log n + \gamma + o(1)$ ist, und für die harmonischen Zahlen zweiter Ordnung $\mathcal{H}_n^{(2)} = \sum_{i=1}^n \frac{1}{i^2}$, dass $\mathcal{H}_n^{(2)} = \frac{\pi^2}{6} + o(1)$ ist.

Für $\vartheta \in \mathbb{N}$ erhalten wir die Behauptung direkt aus $\sum_{i=1}^{n-1} \frac{1}{\vartheta+i} = \mathcal{H}_{n+\vartheta-1} - \sum_{i=1}^{\vartheta} \frac{1}{i}$ und

$$\sum_{i=1}^{n-1} \left(\frac{1}{\vartheta+i} \right)^2 = \mathcal{H}_{n+\vartheta-1}^{(2)} - \sum_{i=1}^{\vartheta} \frac{1}{i^2}.$$

Für allgemeine $\vartheta > 0$ erhalten wir die Behauptung durch die Abschätzungen

$$\sum_{i=1}^{n-1} \frac{1}{\lceil \vartheta \rceil + i} \leq \sum_{i=1}^{n-1} \frac{1}{\vartheta+i} \leq \sum_{i=1}^{n-1} \frac{1}{\lfloor \vartheta \rfloor + i} \text{ bzw. } \sum_{i=1}^{n-1} \left(\frac{1}{\lceil \vartheta \rceil + i} \right)^2 \leq \sum_{i=1}^{n-1} \left(\frac{1}{\vartheta+i} \right)^2 \leq \sum_{i=1}^{n-1} \left(\frac{1}{\lfloor \vartheta \rfloor + i} \right)^2$$

und Betrachten des ersten Falls. \square

2 Die Tiefe eines Knotens im Hoppe-Baum

In diesem Abschnitt werden wir uns mit der Tiefe des n -ten Knotens im Hoppe-Baum befassen. Als Tiefe eines Knotens bezeichnen wir seinen Abstand zur Wurzel. Im Folgenden schreiben wir

$D_n^{(\vartheta)}$ für die Tiefe, wobei $n \in \mathbb{N}$ die Anzahl der Knoten im Hoppe-Baum ist und $\vartheta \in \mathbb{R}^+$ das Gewicht der Wurzel im Hoppe-Baum bezeichnet.

Ziel wird es sein, die Tiefe als Summe unabhängiger, bernoulli-verteilter Zufallsvariablen zu schreiben, was die weitere Analyse dann sehr leicht macht. Für den Spezialfall $\vartheta = 1$, also den Fall eines zufälligen rekursiven Baumes, ist eine solche Darstellung bereits bekannt, man findet ein solches Ergebnis bei Dobrow in [4].

In jedem Baum kann man die Tiefe des n -ten Knotens zumindest als Summe von bernoulli-verteilten Zufallsvariablen schreiben, indem man alle Knoten des Baumes durchläuft und zählt, wie viele davon Vorfahren des n -ten Knotens sind. Allerdings müssen die Wahrscheinlichkeiten dafür sowie die Unabhängigkeit noch bestimmt werden.

Als ersten Schritt werden wir also für beliebige Knoten $i, j \in \mathbb{N} \setminus \{1\}$, $i < j$, die Ereignisse betrachten, dass der i -te Knoten Vorfahr des j -ten Knotens ist. Die Wurzel ist natürlich immer Vorfahr jedes weiteren Knotens und wird daher an dieser Stelle ausgeschlossen.

Lemma 2.1. *Sei für $2 \leq i < j \leq n$ $A_{i,j}$ das Ereignis, dass der i -te Knoten im Hoppe-Baum mit $n \in \mathbb{N}$ Knoten Vorfahr des j -ten Knotens ist. Dann gilt:*

- $P(A_{i,j}) = \frac{1}{\vartheta+i-1}$,
- die Ereignisse $A_{2,n}, \dots, A_{n-1,n}$ sind unabhängig.

Beweis. Wir betrachten also für beliebiges $i < j$ das Ereignis $A_{i,j}$. Es ist sofort klar, dass $P(A_{i,i+1}) = \frac{1}{\vartheta+i-1}$ gilt, denn zur Zeit i befinden sich im Hoppe-Baum die Wurzel mit Gewicht ϑ sowie $i-1$ weitere Knoten mit Gewicht 1. Der Knoten i ist zu diesem Zeitpunkt sicherlich ein Blatt im Baum, da er als letzter eingefügt wurde, also muss genau Knoten i als Vaterknoten für Knoten $i+1$ gezogen werden, damit i Vorfahr von $i+1$ ist. Aufgrund der Gewichte der Knoten erhalten wir sofort die Behauptung für $A_{i,i+1}$. Die Situation für allgemeine $A_{i,j}$ ist ähnlich und wird zunächst intuitiv geklärt, bevor wir die Wahrscheinlichkeit formal ausrechnen.

Wieder ist zur Zeit i der Knoten i ein Blatt im Hoppe-Baum. Nun werden $j-i-1$ Knoten bis zur Zeit $j-1$ zufällig eingefügt, diese bilden neue Teilbäume an den ersten i Knoten. Die Größen dieser Teilbäume ist zwar zufällig, allerdings proportional zu den Wahrscheinlichkeiten, an die Knoten $1, \dots, i$ angefügt zu werden. Die Wahrscheinlichkeit, dass der Knoten j dann in den Teilbaum von i eingefügt wird, sollte daher wieder $\frac{1}{\vartheta+i-1}$ sein.

Um das formal zu zeigen, definieren wir uns zunächst für $k < l$ neue Zufallsvariablen $\xi_{k,l}$ als Anzahl der Nachfahren des Knotens k bei l Knoten im Hoppe-Baum. Dann gilt

$$P(A_{i,j}) = E[P(A_{i,j}|\xi_{i,j-1})] = E\left[\frac{1 + \xi_{i,j-1}}{\vartheta + j - 2}\right], \quad (1)$$

da i genau dann Vorfahr von j ist, wenn j an i selbst oder einen seiner Nachfahren angefügt wird. Um $P(A_{i,j})$ zu ermitteln, müssen wir also lediglich die Erwartungswerte der $\xi_{i,j}$ ermitteln. Dazu nutzen wir folgende rekursive Beobachtung:

$$\xi_{i,j} = \xi_{i,j-1} + \mathbf{1}_{A_{i,j}}.$$

Der Übergang zu Erwartungswerten und Einsetzen von $P(A_{i,j})$ wie in (1) liefert

$$\begin{aligned}
E[\xi_{i,j}] &= E[\xi_{i,j-1}] + E\left[\frac{1 + \xi_{i,j-1}}{\vartheta + j - 2}\right] \\
&= \frac{\vartheta + j - 1}{\vartheta + j - 2} E[\xi_{i,j-1}] + \frac{1}{\vartheta + j - 2} \\
&= \dots \\
&= \left(\prod_{k=i+2}^j \frac{\vartheta + k - 1}{\vartheta + k - 2}\right) E[\xi_{i,i+1}] + \sum_{k=i+2}^j \left(\prod_{l=k+1}^j \frac{\vartheta + l - 1}{\vartheta + l - 2}\right) \frac{1}{\vartheta + k - 2} \\
&= \frac{\vartheta + j - 1}{\vartheta + i} E[\xi_{i,i+1}] + \sum_{k=i+2}^j \frac{\vartheta + j - 1}{\vartheta + k - 1} \frac{1}{\vartheta + k - 2} \\
&= \frac{\vartheta + j - 1}{\vartheta + i} E[\xi_{i,i+1}] + (\vartheta + j - 1) \sum_{k=i+2}^j \frac{1}{(\vartheta + k - 1)(\vartheta + k - 2)} \\
&= \frac{\vartheta + j - 1}{\vartheta + i} E[\xi_{i,i+1}] + (\vartheta + j - 1) \sum_{k=i+2}^j \left(\frac{1}{\vartheta + k - 2} - \frac{1}{\vartheta + k - 1}\right) \\
&= \frac{\vartheta + j - 1}{\vartheta + i} E[\xi_{i,i+1}] + (\vartheta + j - 1) \left(\frac{1}{\vartheta + i} - \frac{1}{\vartheta + j - 1}\right).
\end{aligned}$$

Nun müssen wir nur noch beachten, dass $E[\xi_{i,i+1}] = P(A_{i,i+1}) = \frac{1}{\vartheta+i-1}$ ist, wie es bereits am Anfang des Beweises besprochen wurde. Einsetzen liefert uns

$$\begin{aligned}
E[\xi_{i,j}] &= \frac{\vartheta + j - 1}{\vartheta + i} \frac{1}{\vartheta + i - 1} + (\vartheta + j - 1) \left(\frac{1}{\vartheta + i} - \frac{1}{\vartheta + j - 1}\right) \\
&= \frac{\vartheta + j - 1}{\vartheta + i} \left(\frac{1}{\vartheta + i - 1} + 1\right) - 1 \\
&= \frac{\vartheta + j - 1}{\vartheta + i - 1} - 1
\end{aligned}$$

Einsetzen des Erwartungswertes $E[\xi_{i,j-1}] = \frac{\vartheta+j-2}{\vartheta+i-1} - 1$ in (1) liefert uns nun die Wahrscheinlichkeit

$$P(A_{i,j}) = E\left[\frac{1 + \xi_{i,j-1}}{\vartheta + j - 2}\right] = \frac{1 + \frac{\vartheta+j-2}{\vartheta+i-1} - 1}{\vartheta + j - 2} = \frac{1}{\vartheta + i - 1}.$$

Nun bleibt für den zweiten Teil des Lemmas noch zu zeigen, dass $A_{2,n}, \dots, A_{n-1,n}$ unabhängig sind. Sei dazu $1 < i_1 < \dots < i_k < n$ beliebig. Die Unabhängigkeit ist leicht einzusehen, denn es gilt offensichtlich $\bigcap_{j=1}^k A_{i_j,n} = A_{i_1,i_2} \cap A_{i_2,i_3} \cap \dots \cap A_{i_k,n}$, da sämtliche Knoten i_1, \dots, i_k nur dann gleichzeitig Vorfahren von Knoten n sind, wenn alle diese Knoten auf dem Pfad von der Wurzel zu n passiert werden und dazu muss $A_{i_1,i_2} \cap A_{i_2,i_3} \cap \dots \cap A_{i_k,n}$ gelten. Die Ereignisse $A_{i_1,i_2}, A_{i_2,i_3}, \dots, A_{i_k,n}$ sind nun aber sicherlich unabhängig, da dort jeweils verschiedene Einfügezeitpunkte betroffen sind. Wir erhalten also

$$\begin{aligned}
P\left(\bigcap_{j=1}^k A_{i_j,n}\right) &= P(A_{i_1,i_2} \cap A_{i_2,i_3} \cap \dots \cap A_{i_k,n}) \\
&= P(A_{i_1,i_2}) \cdot \dots \cdot P(A_{i_k,n}).
\end{aligned}$$

Mit dem ersten Teil des Beweises wissen wir aber bereits, dass $P(A_{i_j,i_{j+1}}) = \frac{1}{\vartheta+i_j-1} = P(A_{i_j,n})$ ist für jedes $1 \leq j \leq k$, somit erhalten wir die Unabhängigkeit der Ereignisse. \square

Nun können wir die Verteilung der Tiefe deutlich leichter beschreiben, indem wir die Anzahl der Vorfahren eines Knotens über Indikatorvariablen zählen, was uns zu folgendem Ergebnis führt.

Korollar 2.2. Sei $D_n^{(\vartheta)}$ die Tiefe des n -ten Knotens im Hoppe-Baum. Dann gilt für jedes $n \geq 2$

$$D_n^{(\vartheta)} \stackrel{d}{=} 1 + \sum_{i=1}^{n-2} B_{\frac{1}{\vartheta+i}},$$

wobei $B_{\frac{1}{\vartheta+1}}, \dots, B_{\frac{1}{\vartheta+n-2}}$ unabhängig sind und $B_{\frac{1}{\vartheta+i}}$ Bernoulli($\frac{1}{\vartheta+i}$)-verteilt ist für jedes $i \in \{1, \dots, n-2\}$.

Beweis. Wieder bezeichnen wir für $i < n$ mit $A_{i,n}$ das Ereignis, dass der i -te Knoten Vorfahr des n -ten Knotens ist. Sicherlich ist $P(A_{1,n}) = 1$, da die Wurzel des Baumes Vorfahr jedes Knotens ist. Wir erhalten durch Abzählen der Vorfahren

$$D_n^{(\vartheta)} = 1 + \sum_{i=2}^{n-1} \mathbf{1}_{A_{i,n}}.$$

Die Unabhängigkeit der Summanden und die Erfolgswahrscheinlichkeiten ergeben sich nun aus Lemma 2.1. \square

Die in dem obigen Korollar geschilderte Darstellung der Tiefe macht es uns nun leicht, Erwartungswert und Varianz dieser Größe exakt zu bestimmen. Außerdem erhalten wir über Lindebergs Version des zentralen Grenzwertsatzes nun auch einen Grenzwertsatz für die Tiefe.

Theorem 2.3. Für die Tiefe $D_n^{(\vartheta)}$ des n -ten Knotens im Hoppe-Baum gilt

$$\begin{aligned} E[D_n^{(\vartheta)}] &= 1 + \sum_{i=1}^{n-2} \frac{1}{\vartheta+i} = \log n + c_\vartheta + 1 + o(1), \\ \text{Var}(D_n^{(\vartheta)}) &= \sum_{i=1}^{n-2} \frac{1}{\vartheta+i} - \sum_{i=1}^{n-2} \left(\frac{1}{\vartheta+i} \right)^2 = \log n + c_\vartheta - d_\vartheta + o(1), \\ \mathcal{L}(D_n^*) &\xrightarrow{w} \mathcal{N}(0, 1), \\ d_{TV}(\mathcal{L}(D_n^{(\vartheta)}), \Pi_{E[D_n^{(\vartheta)}]}) &= \mathcal{O}\left(\frac{1}{\log n}\right), \end{aligned}$$

wobei c_ϑ, d_ϑ die Konstanten aus Lemma 1.3 sind, $D_n^* = \frac{D_n^{(\vartheta)} - E[D_n^{(\vartheta)}]}{\sqrt{\text{Var}(D_n^{(\vartheta)})}}$ ist, $\mathcal{N}(0, 1)$ die Standardnormalverteilung und $\Pi_{E[D_n^{(\vartheta)}]}$ die Poissonverteilung zum Parameter $E[D_n^{(\vartheta)}]$ bezeichnet.

Beweis. Aus Korollar 2.2 wissen wir bereits, dass $D_n^{(\vartheta)} \stackrel{d}{=} 1 + \sum_{i=1}^{n-2} B_{\frac{1}{\vartheta+i}}$ ist für unabhängige $B_{\frac{1}{\vartheta+1}}, \dots, B_{\frac{1}{\vartheta+n-2}}$, wobei $B_{\frac{1}{\vartheta+i}}$ Bernoulli($\frac{1}{\vartheta+i}$) verteilt ist. Also gilt

$$\begin{aligned} E[D_n^{(\vartheta)}] &= 1 + \sum_{i=1}^{n-2} E\left[B_{\frac{1}{\vartheta+i}}\right] = 1 + \sum_{i=1}^{n-2} \frac{1}{\vartheta+i} \\ \text{Var}(D_n^{(\vartheta)}) &= \sum_{i=1}^{n-2} \text{Var}\left(B_{\frac{1}{\vartheta+i}}\right) = \sum_{i=1}^{n-2} \left(\frac{1}{\vartheta+i} - \frac{1}{(\vartheta+i)^2} \right) = \sum_{i=1}^{n-2} \frac{1}{\vartheta+i} - \sum_{i=1}^{n-2} \left(\frac{1}{\vartheta+i} \right)^2. \end{aligned}$$

Für Summen $1 + \sum_{i=1}^{n-2} B_{\frac{1}{\vartheta+i}}$ unabhängiger, beschränkter Zufallsvariablen mit $Var\left(1 + \sum_{i=1}^{n-2} B_{\frac{1}{\vartheta+i}}\right) \rightarrow \infty$ erhält man zusätzlich aus Lindebergs Version des zentralen Grenzwertsatzes, dass gilt

$$\mathcal{L}(D_n^*) \xrightarrow{w} \mathcal{N}(0, 1).$$

Die Asymptotiken von Erwartungswert und Varianz folgen aus Lemma 1.3, die Konvergenz von $\mathcal{L}(D_n^{(\vartheta)})$ gegen die Poissonverteilung erhält man mittels Poissonapproximation (vgl. [3, Gleichung (1.23)]). Man erhält mit $\lambda = E[D_n^{(\vartheta)}] \sim \log n$ für den Totalvariationsabstand

$$d_{TV}(\mathcal{L}(D_n^{(\vartheta)}), \Pi_\lambda) \leq \min\{1, \lambda^{-1}\} \left(1 + \sum_{i=1}^{n-2} \left(\frac{1}{\vartheta+i}\right)^2\right) = \mathcal{O}\left(\frac{1}{\log n}\right).$$

□

3 Die Höhe des Hoppe-Baumes

In diesem Abschnitt werden wir die Höhe des Hoppe-Baumes betrachten. Als Höhe eines Baumes bezeichnen wir das Maximum der Tiefen aller Knoten im Baum. Im Folgenden schreiben wir $H_n^{(\vartheta)} = \max\{D_1^{(\vartheta)}, \dots, D_n^{(\vartheta)}\}$ für die Höhe des Hoppe-Baumes zum Parameter $\vartheta \in \mathbb{R}^+$ mit $n \in \mathbb{N}$ Knoten.

Das Hauptergebnis wird sein, dass $H_n^{(\vartheta)}$ sehr stark um den Erwartungswert $E[H_n^{(1)}] = e \log n - \frac{3}{2} \log \log n + \mathcal{O}(1)$ des zufälligen rekursiven Baumes konzentriert sein wird. Insbesondere werden wir daran sehen, dass der Parameter $\vartheta > 0$ nur sehr geringen Einfluss auf die Höhe im Hoppe-Baum nimmt. Wir werden aus der Konzentration folgern können, dass auch der Erwartungswert $E[H_n^{(\vartheta)}] = e \log n - \frac{3}{2} \log \log n + \mathcal{O}(1)$ sein muss und die Varianz von $H_n^{(\vartheta)}$ beschränkt bleiben wird.

Als Vorgehen in diesem Abschnitt werden wir im Gegenteil zu Abschnitt 2 die Ergebnisse für den zufälligen rekursiven Baum voraussetzen und daraus die Ergebnisse für den Hoppe-Baum folgern. Zentral hierfür ist das folgende Ergebnis für die Konzentration der Höhe des zufälligen rekursiven Baumes, welches man bei Addario-Berry und Ford in [2, Korollar 1.3], findet.

Theorem 3.1. *Für die Höhe $H_n^{(1)}$ eines zufälligen rekursiven Baumes mit n Knoten gilt $E[H_n^{(1)}] = e \log n - \frac{3}{2} \log \log n + \mathcal{O}(1)$. Außerdem gibt es für jedes $c' < \frac{1}{2e}$ eine Konstante $C = C(c')$, die nur von c' abhängt, sodass für jedes $n \geq 1$ und $k \geq 1$ gilt*

$$P(|H_n^{(1)} - E[H_n^{(1)}]| \geq k) \leq C e^{-c'k}.$$

Beweis. Für den Beweis siehe [2, Korollar 1.3].

□

Mit Hilfe von Theorem 3.1 und einer geschickten Zerlegung des Hoppe-Baumes erhalten wir folgendes Ergebnis für beliebige Parameter $\vartheta > 0$.

Theorem 3.2. *Sei $H_n^{(\vartheta)}$ die Höhe des Hoppe-Baumes mit n Knoten sowie $M_n = E[H_n^{(1)}] = e \log n - \frac{3}{2} \log \log n + \mathcal{O}(1)$ die erwartete Höhe des zufälligen rekursiven Baumes*

mit n Knoten. Dann gibt es für jedes $c' < \frac{1}{2e}$ eine Konstante $C = C(c')$, so dass für beliebige $t \geq 2$ gilt

$$P\left(H_n^{(\vartheta)} - M_n \geq t\right) \leq Ce^{-c't}.$$

Bezeichnet $K = \sup_{n \in \mathbb{N}} \{|M_n - (e \log n - \frac{3}{2} \log \log n)|\} < \infty$, so gibt es außerdem für jedes $c' < \frac{1}{2e}$ eine Konstante $\tilde{C} = \tilde{C}(c')$ und $\alpha = \frac{c'}{1+c'e}$, sodass für alle $t \geq \frac{2K+1}{1-e\alpha}$ gilt

$$P\left(H_n^{(\vartheta)} - M_n \leq -t\right) \leq \tilde{C}e^{-\alpha t}.$$

Für den Beweis von Theorem 3.2 brauchen wir zunächst ein Lemma um die Verteilung der Anzahl von Knoten im ersten Teilbaum besser zu verstehen.

Lemma 3.3. Sei $N_n^{(1)}$ die Größe des ersten Teilbaumes im Hoppe-Baum mit n Knoten zum Parameter $\vartheta > 0$. Dann gilt für beliebige $\varepsilon \in (0, 1)$

$$P(N_n^{(1)} \leq \varepsilon n) \leq 3(\vartheta + 1)\varepsilon.$$

Beweis. Für den Fall $\varepsilon > \frac{1}{3}$ ist die Schranke trivial, daher muss die Behauptung nur für $\varepsilon \in (0, \frac{1}{3}]$ gezeigt werden.

Hier ist die Schranke für $n \leq 2$ trivial, da dann $\varepsilon n < 1$ und damit $P(N_n^{(1)} \leq \varepsilon n) = 0$ ist.

Sei also $\varepsilon \in (0, \frac{1}{3}]$ und $n \geq 3$. Zunächst überlegt man sich, dass für beliebige $k \in \{1, \dots, n-1\}$ gilt

$$P(N_n^{(1)} = k) = \binom{n-2}{k-1} \frac{\vartheta(\vartheta+1) \cdots (\vartheta+n-(k+2))(k-1)!}{(\vartheta+1) \cdots (\vartheta+n-2)}.$$

Dies gilt, weil sich im ersten Teilbaum immer mindestens Knoten 2 befindet und man von den verbliebenen Knoten 3 bis n genau $k-1$ weitere benötigt, die in den ersten Teilbaum eingefügt werden. Die Wahrscheinlichkeit, dass bestimmte Knoten $2 < i_1 < \dots < i_{k-1} \leq n$ in den ersten Teilbaum eingefügt werden und die verbliebenen Knoten j , $j \in \{3, \dots, n\} \setminus \{i_1, \dots, i_{k-1}\}$, nicht in den Teilbaum eingefügt werden, beträgt $\frac{\vartheta(\vartheta+1) \cdots (\vartheta+n-(k+2))(k-1)!}{(\vartheta+1) \cdots (\vartheta+n-2)}$. Daraus erhält man

$$P(N_n^{(1)} = k) = \frac{\vartheta}{\vartheta+n-2} \prod_{j=2}^k \frac{n-j}{n-j+\vartheta-1}.$$

Es folgt für $\varepsilon \in (0, \frac{1}{3}]$

$$\begin{aligned} P(N_n^{(1)} \leq \varepsilon n) &= \sum_{k=1}^{\lfloor \varepsilon n \rfloor} \frac{\vartheta}{\vartheta+n-2} \prod_{j=2}^k \frac{n-j}{n-j+\vartheta-1} \\ &\leq \frac{\vartheta(n-2)}{\vartheta+n-2} \sum_{k=1}^{\lfloor \varepsilon n \rfloor} \frac{1}{n-k+\vartheta-1} \\ &\leq \vartheta \frac{\lfloor \varepsilon n \rfloor}{n - \lfloor \varepsilon n \rfloor + \vartheta - 1} \\ &\leq \vartheta \frac{\varepsilon n}{(1-\varepsilon)n-1} \\ &\leq \vartheta \frac{\varepsilon n}{(1-\varepsilon-\frac{1}{3})n} \quad (\text{wegen } n \geq 3) \\ &\leq 3(\vartheta+1)\varepsilon \quad (\text{wegen } \varepsilon \leq \frac{1}{3}). \end{aligned}$$

□

Beweis von Theorem 3.2. Es sei $H_n^{(\vartheta)}$ die Höhe des Hoppe-Baumes mit n Knoten sowie $M_n = E[H_n^{(1)}] = e \log n - \frac{3}{2} \log \log n + \mathcal{O}(1)$ und $c' < \frac{1}{2e}$ beliebig.

Für den ersten Teil der Behauptung machen wir uns zunutze, dass die Höhe $H_n^{(\vartheta)}$ stochastisch kleiner ist als $1 + H_{n-1}^{(1)}$. Der anschauliche Grund hierfür ist, dass die Höhe am größten wird, wenn die Wurzel des Hoppe-Baumes nur ein einziges Kind hat. Der Teilbaum an diesem Kind hat dann $n - 1$ Knoten und weist die Struktur eines zufälligen rekursiven Baumes auf, da jeder Knoten im Teilbaum Gewicht 1 besitzt und somit nachfolgende Knoten beim Einfügen in den Teilbaum ihren Vaterknoten uniform unter den vorhandenen Knoten wählen. Um $1 + H_{n-1}^{(1)}$ stochastisch als obere Schranke zu bekommen kann man alternativ die Tiefen im Hoppe-Baum einzeln betrachten, wir erhalten aus Korollar 2.2, dass $D_j^{(\vartheta)} \stackrel{d}{=} 1 + \sum_{i=1}^{j-2} B_{\frac{1}{\vartheta+i}}$ ist für unabhängige $B_{\frac{1}{\vartheta+1}}, \dots, B_{\frac{1}{\vartheta+j-2}}$, wobei $B_{\frac{1}{\vartheta+i}}$ Bernoulli($\frac{1}{\vartheta+i}$)-verteilt ist.

Den Tiefen sieht man jetzt sofort an, dass die stochastisch am größten werden, wenn $\vartheta = 0$ ist. Dann ist aber $1 + \sum_{i=1}^{j-2} B_{\frac{1}{i}} \stackrel{d}{=} 1 + D_{j-1}^{(1)}$ und wir erhalten die Aussage für die Höhen, indem wir zum Maximum der Tiefen übergehen.

Außerdem ist $H_{n-1}^{(1)}$ stochastisch kleiner als $H_n^{(1)}$, wir erhalten also, dass $H_n^{(\vartheta)}$ stochastisch kleiner ist als $1 + H_n^{(1)}$. Dies liefert uns

$$\begin{aligned} P\left(H_n^{(\vartheta)} - M_n \geq t\right) &\leq P\left(1 + H_n^{(1)} - M_n \geq t\right) \\ &= P\left(H_n^{(1)} - E[H_n^{(1)}] \geq t - 1\right). \end{aligned}$$

Mit Theorem 3.1 erhalten wir nun die Existenz einer Konstanten $\hat{C} = \hat{C}(c')$, sodass für alle $t \geq 2$ gilt

$$P\left(H_n^{(1)} - E[H_n^{(1)}] \geq t - 1\right) \leq \hat{C}e^{-c'(t-1)} = \hat{C}e^{c'}e^{-c't}.$$

Wählt man $C = \hat{C}e^{c'}$, erhält man somit

$$P\left(H_n^{(\vartheta)} - M_n \geq t\right) \leq Ce^{-c't}.$$

Nun müssen wir noch die Konzentration des linken Tails zeigen. Wir bezeichnen mit $N_n^{(1)}$ wieder die Größe des Teilbaumes des ersten Kindes der Wurzel bei einem Hoppe-Baum mit n Knoten, also die Anzahl der Kinder von Knoten 2 plus den Knoten selbst. Mit $H_{N_n^{(1)}}^{(1)}$ bezeichnen wir die Höhe des Teilbaumes bestehend aus genau diesen $N_n^{(1)}$ Knoten. Da jeder der Knoten dieses Teilbaumes gleiches Gewicht hat, ist der Vaterknoten jedes weiteren Knotens, der in den ersten Teilbaum eingefügt wird, uniform unter allen bestehenden Knoten im ersten Teilbaum verteilt. Wir sind also strukturell in derselben Situation wie bei einem zufälligen rekursiven Baum, d.h. für $N_n^{(1)} = k$ ist $H_{N_n^{(1)}}^{(1)}$ in Verteilung die Höhe eines zufälligen rekursiven Baumes mit k Knoten. Offensichtlich gilt für die Höhe $H_n^{(\vartheta)}$ des Hoppe-Baumes für beliebige $t \in \mathbb{R}$ $\{H_n^{(\vartheta)} \leq t\} \subset \{1 + H_{N_n^{(1)}}^{(1)} \leq t\} \subset \{H_{N_n^{(1)}}^{(1)} \leq t\}$. Dies liefert uns zunächst

$$P(H_n^{(\vartheta)} - M_n \leq -t) \leq P(H_{N_n^{(1)}}^{(1)} - M_n \leq -t).$$

Daraus erhalten wir

$$\begin{aligned} P(H_n^{(\vartheta)} - M_n \leq -t) &\leq P(\{H_{N_n^{(1)}}^{(1)} - M_n \leq -t\} \cap \{N_n^{(1)} \geq e^{-\alpha t} n\}) \\ &\quad + P(\{H_{N_n^{(1)}}^{(1)} - M_n \leq -t\} \cap \{N_n^{(1)} < e^{-\alpha t} n\}), \end{aligned}$$

wobei der Grund für diese spezielle Zerlegung sowie die Wahl $\alpha = \frac{c'}{1+c'e}$ im späteren Verlauf des Beweises klar wird. Da die Höhe eines zufälligen rekursiven Baumes stochastisch monoton in der Anzahl der Knoten ist, können wir im vorderen Term statt $H_{N_n^{(1)}}^{(1)}$ die Höhe eines zufälligen rekursiven Baumes mit $\lceil e^{-\alpha t} n \rceil$ Knoten betrachten und erhalten

$$\begin{aligned} P(H_n^{(\vartheta)} - M_n \leq -t) &\leq P(\{H_{\lceil e^{-\alpha t} n \rceil}^{(1)} - M_n \leq -t\} \cap \{N_n^{(1)} \geq e^{-\alpha t} n\}) \\ &\quad + P(\{H_{N_n^{(1)}}^{(1)} - M_n \leq -t\} \cap \{N_n^{(1)} < e^{-\alpha t} n\}) \\ &\leq \underbrace{P(H_{\lceil e^{-\alpha t} n \rceil}^{(1)} - M_n \leq -t)}_{:=P_1(n,t)} + \underbrace{P(N_n^{(1)} < e^{-\alpha t} n)}_{:=P_2(n,t)}. \end{aligned}$$

Für den ersten Summanden möchten wir Theorem 3.1 von Addario-Berry und Ford verwenden. Es gilt

$$P_1(n, t) = P\left(H_{\lceil e^{-\alpha t} n \rceil}^{(1)} - M_{\lceil e^{-\alpha t} n \rceil} \leq -t + M_n - M_{\lceil e^{-\alpha t} n \rceil}\right)$$

mit $M_n - M_{\lceil e^{-\alpha t} n \rceil} \leq e(\log n - \log(e^{-\alpha t} n)) - \frac{3}{2}(\log \log n - \log \log(\lceil e^{-\alpha t} n \rceil)) + 2K \leq e\alpha t + 2K$ für $t > 0$, wobei $K = \sup_{n \in \mathbb{N}} \{ |M_n - (e \log n - \frac{3}{2} \log \log n)| \} < \infty$ bezeichnet. Somit folgt

$$P_1(n, t) \leq P\left(H_{\lceil e^{-\alpha t} n \rceil}^{(1)} - M_{\lceil e^{-\alpha t} n \rceil} \leq -(1 - e\alpha)t + 2K\right).$$

Für $t \geq \frac{2K+1}{1-e\alpha}$ ist $-(1 - e\alpha)t + 2K \leq -1$ und Theorem 3.1 liefert die Existenz einer Konstanten $C_1 = C_1(c')$ mit

$$P_1(n, t) \leq C_1 e^{-c'((1-e\alpha)t - 2K)} = C_1 e^{2Kc'} e^{-\alpha t},$$

wobei die letzte Gleichheit gilt, da α so gewählt ist, dass $c'(1 - e\alpha) = \alpha$ gilt.

Für den zweiten Summanden gilt nach Lemma 3.3 für beliebige $t > 0$

$$P_2(n, t) = P\left(N_n^{(1)} < e^{-\alpha t} n\right) \leq 3(\vartheta + 1)e^{-\alpha t}.$$

Insgesamt folgt mit der Wahl $\tilde{C} = 3(\vartheta + 1) + C_1 e^{2Kc'}$, dass für alle $t \geq \frac{2K+1}{1-e\alpha}$ gilt

$$P(H_n^{(\vartheta)} - M_n \leq -t) \leq P_1(n, t) + P_2(n, t) \leq \tilde{C} e^{-\alpha t}.$$

□

Korollar 3.4. Sei $H_n^{(\vartheta)}$ die Höhe des Hoppe-Baumes mit n Knoten sowie $M_n = E[H_n^{(1)}] = e \log n - \frac{3}{2} \log \log n + \mathcal{O}(1)$. Ferner seien $c', C, \tilde{C}, \alpha > 0$ wie in Theorem 3.4 gewählt. Dann gilt

$$|E[H_n^{(\vartheta)}] - M_n| \leq \max \left\{ 2 + \frac{C}{c' e^{2c'}}, \frac{2K+1}{1-e\alpha} + \frac{\tilde{C}}{\alpha \exp\left(\frac{(2K+1)\alpha}{1-e\alpha}\right)} \right\}.$$

Insbesondere gilt $E[H_n^{(\vartheta)}] = e \log n - \frac{3}{2} \log \log n + \mathcal{O}(1)$.

Des Weiteren gilt

$$\text{Var}(H_n^{(\vartheta)}) = \mathcal{O}(1).$$

Beweis. Allgemein kann bei starker Konzentration von $H_n^{(\vartheta)}$ um M_n , welche wir in Theorem 3.4 nachgewiesen haben, $|E[H_n^{(\vartheta)}] - M_n|$ nicht allzu groß werden (vergleiche [10, Lemma 4.4] aus Neiningers Vorlesungsskript). Wir erhalten

$$\begin{aligned} E[H_n^{(\vartheta)}] - M_n &\leq E[(H_n^{(\vartheta)} - M_n)^+] \\ &= \int_0^\infty P(H_n^{(\vartheta)} - M_n \geq t) dt. \end{aligned}$$

Nach Theorem 3.2 gibt es Konstanten $c', C > 0$ mit $P(H_n^{(\vartheta)} - M_n \geq t) \leq C e^{-c't}$ für alle $t \geq 2$, also gilt

$$\begin{aligned} E[H_n^{(\vartheta)}] - M_n &\leq 2 + \int_2^\infty C e^{-c't} dt \\ &= 2 + \frac{C}{c' e^{2c'}} < \infty. \end{aligned}$$

Analog erhalten wir auch

$$\begin{aligned} -(E[H_n^{(\vartheta)}] - M_n) &\leq E[(H_n^{(\vartheta)} - M_n)^-] \\ &= \int_0^\infty P(H_n^{(\vartheta)} - M_n \leq -t) dt \\ &\leq \frac{2K+1}{1-e\alpha} + \int_{\frac{2K+1}{1-e\alpha}}^\infty \tilde{C} e^{-\alpha t} dt \\ &= \frac{2K+1}{1-e\alpha} + \frac{\tilde{C}}{\alpha e^{\frac{(2K+1)\alpha}{1-e\alpha}}} < \infty, \end{aligned}$$

wobei wir die Konstanten $K, \tilde{C}, \alpha > 0$ wieder aus Theorem 3.2 erhalten.

Insgesamt folgt also $|E[H_n^{(\vartheta)}] - M_n| \leq \max \left\{ 2 + \frac{C}{c' e^{2c'}}, \frac{2K+1}{1-e\alpha} + \frac{\tilde{C}}{\alpha \exp(\frac{(2K+1)\alpha}{1-e\alpha})} \right\}$ und damit insbesondere wegen $M_n = e \log n - \frac{3}{2} \log \log n + \mathcal{O}(1)$, dass $E[H_n^{(\vartheta)}] = e \log n - \frac{3}{2} \log \log n + \mathcal{O}(1)$ ist. Für die Varianz nutzen wir, dass bekanntermaßen gilt $E[(X - E[X])^2] \leq E[(X - K)^2]$ für alle Konstanten $K \in \mathbb{R}$. Es folgt

$$\begin{aligned} \text{Var}(H_n^{(\vartheta)}) &\leq E[(H_n^{(\vartheta)} - M_n)^2] \\ &= \int_0^\infty P(|H_n^{(\vartheta)} - M_n| \geq \sqrt{t}) dt \\ &\leq \left(\frac{2K+1}{1-e\alpha} + 2 \right)^2 + \int_{\left(\frac{2K+1}{1-e\alpha} + 2\right)^2}^\infty (C + \tilde{C}) e^{-\alpha \sqrt{t}} dt < \infty, \end{aligned}$$

wobei benutzt wurde, dass $\alpha < c'$ ist und damit für hinreichend große t gilt

$$P(|H_n^{(\vartheta)} - M_n| \geq \sqrt{t}) = P(H_n^{(\vartheta)} - M_n \geq \sqrt{t}) + P(H_n^{(\vartheta)} - M_n \leq -\sqrt{t}) \leq (C + \tilde{C}) e^{-\alpha \sqrt{t}}.$$

□

4 Die Anzahl der Blätter im Hoppe-Baum

In diesem Abschnitt werden wir uns mit der Anzahl der Blätter im Hoppe-Baum befassen. Als Blatt im Baum bezeichnet man einen Knoten, der keine Kinder besitzt. Wir werden im Folgenden $L_n^{(\vartheta)}$ für die Anzahl der Blätter im Hoppe-Baum mit $n \in \mathbb{N}$ Knoten und Wurzelgewicht $\vartheta \in \mathbb{R}^+$ schreiben.

Für die Analyse werden wir $L_n^{(\vartheta)}$ zu einem Martingal reskalieren, woraus wir mittels Azuma-Hoeffding Ungleichung zunächst eine Konzentrationsungleichung gewinnen, welche uns auch Gesetze großer Zahlen für $L_n^{(\vartheta)}$ liefert. Da sich $L_n^{(\vartheta)}$ sehr ähnlich zu der Anzahl weißer Kugeln in Friedmanns Urne verhält, lehnt sich die anfängliche Analyse sehr stark an die Analyse von Friedmanns Urne an, wie man sie auch in Neiningers Vorlesungsskript [10, Satz 3.33] findet.

Als zweiten Schritt in der Analyse von $L_n^{(\vartheta)}$ werden wir dann die Varianz bestimmen und mittels eines allgemeinen Vorgehens für Martingale aus dem Buch [7] von Hall und Heyde einen Grenzwertsatz für $L_n^{(\vartheta)}$ beweisen.

Wir starten mit der Reskalierung zum Martingal.

Theorem 4.1. *Sei $L_n = L_n^{(\vartheta)}$ die Anzahl an Blättern im Hoppe-Baum mit n Knoten.*

Dann liefern $X_1 = 0$ und

$$X_n = (\vartheta + n - 2) \left(L_n - \left(\frac{n-1}{2} + \frac{\vartheta(n-1)}{2(\vartheta+n-2)} \right) \right)$$

für $n \geq 2$ ein Martingal $(X_n)_{n \geq 1}$ bezüglich der natürlichen Filtration $(\mathcal{F}_n)_{n \geq 1}$ gegeben durch $\mathcal{F}_n = \sigma(L_1, \dots, L_n)$ für $n \geq 1$.

Beweis. Zunächst überlegt man sich leicht, dass für jedes $n \geq 3$ fast sicher gilt

$$E[L_n | \mathcal{F}_{n-1}] = L_{n-1} \frac{L_{n-1}}{\vartheta + n - 2} + (L_{n-1} + 1) \left(1 - \frac{L_{n-1}}{\vartheta + n - 2} \right) = \frac{(\vartheta + n - 3)L_{n-1}}{\vartheta + n - 2} + 1, \quad (2)$$

da wir beim Einfügen des n -ten Knotens bei L_{n-1} Blättern mit Wahrscheinlichkeit $\frac{L_{n-1}}{\vartheta+n-2}$ ein Blatt als Vaterknoten auswählen und damit die Anzahl der Blätter unverändert bleibt, andernfalls wird ein Blatt hinzugewonnen. Damit lässt sich leicht einsehen, dass $(X_n)_{n \geq 1}$ ein Martingal ist, denn

es gilt für jedes $n \geq 3$ fast sicher

$$\begin{aligned}
E[X_n | \mathcal{F}_{n-1}] &= (\vartheta + n - 2) \left(E[L_n | \mathcal{F}_{n-1}] - \left(\frac{n-1}{2} + \frac{\vartheta(n-1)}{2(\vartheta+n-2)} \right) \right) \\
&\stackrel{(2)}{=} (\vartheta + n - 2) \left(\frac{(\vartheta+n-3)L_{n-1}}{\vartheta+n-2} + 1 - \left(\frac{n-1}{2} + \frac{\vartheta(n-1)}{2(\vartheta+n-2)} \right) \right) \\
&= (\vartheta + n - 3) \left(L_{n-1} + \frac{\vartheta+n-2}{\vartheta+n-3} - \left(\frac{\vartheta+n-2}{\vartheta+n-3} \frac{(n-1)}{2} + \frac{\vartheta(n-1)}{2(\vartheta+n-3)} \right) \right) \\
&= (\vartheta + n - 3) \left(L_{n-1} - \left(-\frac{\vartheta+n-2}{\vartheta+n-3} + \frac{\vartheta+n-2}{\vartheta+n-3} \frac{(n-1)}{2} + \frac{\vartheta(n-2)}{2(\vartheta+n-3)} + \frac{\vartheta}{2(\vartheta+n-3)} \right) \right) \\
&= (\vartheta + n - 3) \left(L_{n-1} - \left(\frac{-2(\vartheta+n-2) + (\vartheta+n-2)(n-1) + \vartheta}{2(\vartheta+n-3)} + \frac{\vartheta(n-2)}{2(\vartheta+n-3)} \right) \right) \\
&= (\vartheta + n - 3) \left(L_{n-1} - \left(\frac{-(\vartheta+n-2) + (\vartheta+n-2)(n-2) + \vartheta}{2(\vartheta+n-3)} + \frac{\vartheta(n-2)}{2(\vartheta+n-3)} \right) \right) \\
&= (\vartheta + n - 3) \left(L_{n-1} - \left(\frac{n-2}{2} + \frac{\vartheta(n-2)}{2(\vartheta+n-3)} \right) \right) \\
&= X_{n-1}.
\end{aligned}$$

Für $n = 2$ ist $L_2 = 1$, da wir hier immer eine Wurzel und ein Blatt haben, somit ist fast sicher

$$E[X_2 | \mathcal{F}_1] = \vartheta \left(1 - \left(\frac{1}{2} + \frac{\vartheta}{2\vartheta} \right) \right) = 0 = X_1.$$

□

Korollar 4.2. Für die Anzahl $L_n = L_n^{(\vartheta)}$ der Blätter im Hoppe-Baum mit n Knoten gilt für $n \geq 2$

$$E[L_n] = \frac{n-1}{2} + \frac{\vartheta(n-1)}{2(\vartheta+n-2)} = \frac{n+\vartheta-1}{2} + \mathcal{O}\left(\frac{1}{n}\right).$$

Beweis. Mit Theorem 4.1 wissen wir bereits, dass $X_1 = 0$ und $X_n = (\vartheta+n-2) \left(L_n - \left(\frac{n-1}{2} + \frac{\vartheta(n-1)}{2(\vartheta+n-2)} \right) \right)$ für $n \geq 2$ ein Martingal liefern. Insbesondere ist also $E[X_n] = E[X_1] = 0$ für alle $n \in \mathbb{N}$. Dies liefert

$$(\vartheta + n - 2) \left(E[L_n] - \left(\frac{n-1}{2} + \frac{\vartheta(n-1)}{2(\vartheta+n-2)} \right) \right) = 0$$

und damit die Behauptung. □

Für die weitere Analyse bestimmen wir zunächst eine Konzentrationsungleichung für die Anzahl der Blätter. Nützlich für Konzentrationsungleichungen bei Martingalen ist hier die Azuma-Hoeffding Ungleichung, welche aus [10, Satz 3.29] zitiert wird.

Theorem 4.3. (Azuma-Hoeffding Ungleichung) Sei $\Delta_1, \dots, \Delta_n$ eine Martingaldifferenzfolge mit $\Delta_i \in [a_i, b_i]$ fast sicher für alle $i = 1, \dots, n$. Dann gilt für alle $t > 0$

$$P \left(\left| \sum_{i=1}^n \Delta_i \right| \geq t \right) \leq 2 \exp \left(- \frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right).$$

Theorem 4.4. Für die Anzahl der Blätter $L_n = L_n^{(\vartheta)}$ im Hoppe-Baum mit $n \geq 3$ Knoten gilt für jedes $t > 0$

$$P(|L_n - E[L_n]| \geq t) \leq 2 \exp\left(-\frac{2t^2(\vartheta + n - 2)^2}{\sum_{i=3}^n (\vartheta + 2i - 5)^2}\right).$$

Insbesondere erhalten wir also für jedes $t > 0$ die Schranke

$$P(|L_n - E[L_n]| \geq t) \leq 2 \exp\left(-\frac{t^2}{2n}\right).$$

Beweis. Wieder verwenden wir das Martingal $(X_n)_{n \geq 1}$ aus Theorem 4.1 mit

$X_n = (\vartheta + n - 2)(L_n - E[L_n])$ für $n \geq 2$. Dies liefert uns eine Martingaldifferenzfolge $(\Delta_i)_{i \geq 2}$ mit $\Delta_2 = X_2 - X_1 = 0$ sowie für $i \geq 3$

$$\begin{aligned} \Delta_i &= X_i - X_{i-1} \\ &= (\vartheta + i - 2)(L_i - E[L_i]) - (\vartheta + i - 3)(L_{i-1} - E[L_{i-1}]) \\ &= L_i - E[L_i] + (\vartheta + i - 3) \underbrace{(L_i - L_{i-1} - E[L_i - L_{i-1}])}_{\in [0,1]}. \end{aligned}$$

Als weitere Abschätzung gilt $L_i \in [1, i-1]$, da es bei i Knoten natürlich auch maximal $i-1$ Blätter geben kann und zumindest der i -te Knoten immer ein Blatt ist, also gilt

$L_i - E[L_i] \in [1 - E[L_i], i - 1 - E[L_i]]$ und damit insgesamt

$$\Delta_i \in [1 - E[L_i] - (\vartheta + i - 3)E[L_i - L_{i-1}], i - 1 - E[L_i] + (\vartheta + i - 3)(1 - E[L_i - L_{i-1}])].$$

Δ_i fällt also fast sicher in ein Intervall der Länge $c_i = \vartheta + 2i - 5$ für $i \geq 3$. Die Azuma Hoeffding Ungleichung (Theorem 4.3) liefert also für jedes $n \geq 3$ und beliebiges $t > 0$

$$P\left(\left|\sum_{i=2}^n \Delta_i\right| \geq t\right) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=3}^n (\vartheta + 2i - 5)^2}\right). \quad (3)$$

Dies liefert uns für jedes $n \geq 3$

$$\begin{aligned} P(|L_n - E[L_n]| \geq t) &= P(|X_n| \geq t(\vartheta + n - 2)) \\ &= P\left(\left|\sum_{i=2}^n \Delta_i\right| \geq t(\vartheta + n - 2)\right) \\ &\stackrel{(3)}{\leq} 2 \exp\left(-\frac{2t^2(\vartheta + n - 2)^2}{\sum_{i=3}^n (\vartheta + 2i - 5)^2}\right). \end{aligned}$$

Um eine etwas vereinfachte (und gröbere) Schranke zu erhalten, macht man nun noch folgende Abschätzungen:

$$\begin{aligned} \exp\left(-\frac{2t^2(\vartheta + n - 2)^2}{\sum_{i=3}^n (\vartheta + 2i - 5)^2}\right) &\leq \exp\left(-\frac{2t^2(\vartheta + n - 2)^2}{n(\vartheta + 2n - 5)^2}\right) \\ &= \exp\left(-\frac{2t^2}{n} \left(\frac{\vartheta + n - 2}{\vartheta + 2n - 5}\right)^2\right) \end{aligned}$$

Nun kann man sich z.B. über Differentiation nach ϑ überlegen, dass für jedes $n \geq 3$ gilt

$\inf_{\vartheta \geq 0} \frac{\vartheta + n - 2}{\vartheta + 2n - 5} = \frac{n-2}{2n-5}$. Dies liefert

$$\frac{\vartheta + n - 2}{\vartheta + 2n - 5} \geq \frac{1}{2}.$$

Damit erhalten wir die weitere Abschätzung

$$\exp\left(-\frac{2t^2(\vartheta + n - 2)^2}{\sum_{i=3}^n (\vartheta + 2i - 5)^2}\right) \leq \exp\left(-\frac{2t^2}{n} \left(\frac{1}{2}\right)^2\right) = \exp\left(-\frac{t^2}{2n}\right).$$

□

Bemerkung 4.5. Der Exponent in Theorem 4.4 ist nicht optimal und lässt sich recht leicht verbessern, indem man eine etwas allgemeinere Form der Azuma-Hoeffding Ungleichung benutzt, die für die Martingaldifferenzfolge nur fordert, dass es für alle $i \in \{1, \dots, n\}$ \mathcal{F}_{i-1} messbare Zufallsvariablen Z_i und Konstanten c_i gibt mit $Z_i \leq D_i \leq Z_i + c_i$ und dann dieselbe Schranke mit $b_i - a_i = c_i$ liefert. Damit schafft man mit $c_i = \vartheta + i - 2$ asymptotisch einen Exponenten von $\frac{-6t^2}{n}$. Allerdings wird die Konstante im Exponenten für die weitere Analyse keine Rolle spielen, deshalb belassen wir es an dieser Stelle mit der etwas schlechteren Schranke.

Ohne weitere Mühe erhalten wir aus Theorem 4.4 die stochastische Konvergenz von $\left(\frac{L_n^{(\vartheta)}}{n}\right)_{n \geq 1}$.

Korollar 4.6. Für die Anzahl der Blätter $L_n = L_n^{(\vartheta)}$ im Hoppe-Baum mit n Knoten gilt

$$\frac{L_n}{n} \xrightarrow{P} \frac{1}{2} \quad \text{für } n \rightarrow \infty.$$

Beweis. Mit Theorem 4.4 erhalten wir für beliebige $\varepsilon > 0$

$$P\left(\left|\frac{L_n - E[L_n]}{n}\right| \geq \varepsilon\right) = P(|L_n - E[L_n]| \geq \varepsilon n) \leq 2 \exp\left(-\frac{\varepsilon^2 n}{2}\right) \xrightarrow{n \rightarrow \infty} 0,$$

also gilt $\frac{L_n - E[L_n]}{n} \xrightarrow{P} 0$. Außerdem wissen wir mit Korollar 4.2, dass gilt

$$\frac{E[L_n]}{n} \xrightarrow{n \rightarrow \infty} \frac{1}{2}.$$

Beides zusammen liefert die Behauptung. □

Wir werden die Aussage aus Korollar 4.6 nun noch verfeinern, indem wir mittels Borel-Cantelli Lemma noch die fast sichere Konvergenz von $\left(\frac{L_n^{(\vartheta)}}{n}\right)_{n \geq 1}$ zeigen.

Theorem 4.7. Für die Anzahl der Blätter $L_n = L_n^{(\vartheta)}$ im Hoppe-Baum mit n Knoten gilt fast sicher

$$\frac{L_n}{n} \rightarrow \frac{1}{2} \quad \text{für } n \rightarrow \infty.$$

Beweis. Sei $\varepsilon > 0$ beliebig.

Für $n \in \mathbb{N}$ sei $A_n^{(\varepsilon)} = \left\{\left|\frac{L_n - E[L_n]}{n}\right| > \varepsilon\right\}$ sowie $A^{(\varepsilon)} = \limsup_{n \rightarrow \infty} A_n^{(\varepsilon)}$. Als ersten Schritt zeigen wir nun, dass $P(A^{(\varepsilon)}) = 0$ ist. Das erhält man aber leicht durch das Borel-Cantelli-Lemma, denn es gilt

$$\sum_{n=1}^{\infty} P(A_n^{(\varepsilon)}) \stackrel{4.4}{\leq} \sum_{n=1}^{\infty} 2 \exp\left(-\frac{\varepsilon^2 n}{2}\right) < \infty.$$

Das Borel-Cantelli-Lemma liefert $P(A^{(\varepsilon)}) = 0$ und somit $\limsup_{n \rightarrow \infty} \left|\frac{L_n - E[L_n]}{n}\right| \leq \varepsilon$ fast sicher.

Da dies für jedes $\varepsilon > 0$ gilt, erhalten wir somit

$$P\left(\lim_{n \rightarrow \infty} \frac{L_n - E[L_n]}{n} = 0\right) = P\left(\bigcap_{m=1}^{\infty} \left\{\limsup_{n \rightarrow \infty} \left|\frac{L_n - E[L_n]}{n}\right| \leq \frac{1}{m}\right\}\right) = 1.$$

Es folgt also $\frac{L_n - E[L_n]}{n} \rightarrow 0$ fast sicher. Mit $\frac{E[L_n]}{n} \rightarrow \frac{1}{2}$ erhalten wir somit die Behauptung. □

Als nächsten Schritt wollen wir noch einen zentralen Grenzwertsatz für die Anzahl der Blätter herleiten. Dazu werden wir zunächst die Varianz bestimmen.

Theorem 4.8. *Sei $L_n^{(\vartheta)}$ die Anzahl der Blätter im Hoppe-Baum mit n Knoten. Dann gilt*

$$\text{Var} \left(L_n^{(\vartheta)} \right) = \frac{\vartheta + n - 1}{12} + \mathcal{O} \left(\frac{1}{n} \right).$$

Beweis. Wir betrachten wieder das Martingal aus Theorem 4.1, das für $n \geq 2$ durch $X_n = (\vartheta + n - 2)(L_n^{(\vartheta)} - E[L_n^{(\vartheta)}])$ gegeben ist und bezeichnen mit $\mathcal{F}_n = \sigma(L_1^{(\vartheta)}, \dots, L_n^{(\vartheta)})$ die Filtrierung. Es gilt für $n \geq 3$

$$X_n = \frac{\vartheta + n - 2}{\vartheta + n - 3} X_{n-1} + (\vartheta + n - 2) \underbrace{(L_n^{(\vartheta)} - L_{n-1}^{(\vartheta)})}_{=: Y_n} - E[L_n^{(\vartheta)} - L_{n-1}^{(\vartheta)}].$$

Y_n ist eine Indikatorvariable und wird genau dann 1, wenn als Vaterknoten für den n -ten Knoten kein Blatt gewählt wird und somit beim Einfügen des n -ten Knotens ein zusätzliches Blatt gewonnen wird. Für die Varianz des Martingals folgt

$$E[X_n^2] = \left(\frac{\vartheta + n - 2}{\vartheta + n - 3} \right)^2 E[X_{n-1}^2] + 2 \frac{(\vartheta + n - 2)^2}{\vartheta + n - 3} E[X_{n-1}(Y_n - E[Y_n])] + (\vartheta + n - 2)^2 \text{Var}(Y_n).$$

Für den mittleren Term erhalten wir wegen $E[X_{n-1}] = 0$

$$\begin{aligned} E[X_{n-1}(Y_n - E[Y_n])] &= E[X_{n-1}Y_n] \\ &= E[X_{n-1}E[Y_n|\mathcal{F}_{n-1}]] \\ &= E \left[X_{n-1} \left(1 - \frac{L_{n-1}^{(\vartheta)}}{\vartheta + n - 2} \right) \right] \\ &= -\frac{1}{(\vartheta + n - 2)(\vartheta + n - 3)} E[X_{n-1}^2] + 0. \end{aligned}$$

Einsetzen ergibt für das zweite Moment

$$\begin{aligned} E[X_n^2] &= \left(\left(\frac{\vartheta + n - 2}{\vartheta + n - 3} \right)^2 - 2 \frac{\vartheta + n - 2}{(\vartheta + n - 3)^2} \right) E[X_{n-1}^2] + (\vartheta + n - 2)^2 \text{Var}(Y_n) \\ &= \frac{(\vartheta + n - 2)(\vartheta + n - 4)}{(\vartheta + n - 3)^2} E[X_{n-1}^2] + (\vartheta + n - 2)^2 \text{Var}(Y_n). \end{aligned}$$

Setzt man $Q_n = \frac{\vartheta + n - 3}{\vartheta + n - 2} E[X_n^2]$ erhält man die Rekursionsvorschrift

$$Q_n = Q_{n-1} + (\vartheta + n - 2)(\vartheta + n - 3) \text{Var}(Y_n)$$

und damit wegen $Q_2 = 0 (= X_2)$, dass $Q_n = \sum_{i=3}^n (\vartheta + i - 3)(\vartheta + i - 2) \text{Var}(Y_i)$ ist. Des Weiteren gilt für den Erwartungswert der Indikatorvariable

$$\begin{aligned} E[Y_i] = E[E[Y_i|\mathcal{F}_{i-1}]] &= E \left[1 - \frac{L_{i-1}^{(\vartheta)}}{\vartheta + i - 2} \right] \\ &= 1 - \frac{E[L_{i-1}^{(\vartheta)}]}{\vartheta + i - 2}. \end{aligned}$$

Mit Korollar 4.2 wissen wir bereits, dass $E[L_n^{(\vartheta)}] = \frac{\vartheta+n-1}{2} + \mathcal{O}(\frac{1}{n})$ ist, somit ist $E[Y_i] = \frac{1}{2} + \mathcal{O}(\frac{1}{i^2})$ und wir erhalten für die Varianz des Indikators

$$\text{Var}(Y_i) = E[Y_i](1 - E[Y_i]) = \frac{1}{4} + \mathcal{O}\left(\frac{1}{i^2}\right).$$

Es folgt

$$\begin{aligned} Q_n &= \sum_{i=3}^n (\vartheta + i - 3)(\vartheta + i - 2)\text{Var}(Y_i) \\ &= \frac{1}{4} \sum_{i=3}^n (\vartheta + i - 3)(\vartheta + i - 2) + \mathcal{O}(n) \\ &= \frac{1}{4} \sum_{i=0}^{n-3} (i^2 + (2\vartheta + 1)i + \vartheta(\vartheta + 1)) + \mathcal{O}(n) \\ &= \frac{(n-3)(n-2)(2n-5)}{24} + \frac{(2\vartheta+1)(n-3)(n-2)}{8} + \mathcal{O}(n) \\ &= \frac{1}{12}n^3 + \left(\frac{\vartheta}{4} - \frac{1}{2}\right)n^2 + \mathcal{O}(n). \end{aligned}$$

Daraus erhalten wir für die Varianz der Anzahl an Blättern

$$\begin{aligned} \text{Var}(L_n^{(\vartheta)}) &= \frac{1}{(\vartheta + n - 2)^2} E[X_n^2] = \frac{1}{(\vartheta + n - 2)(\vartheta + n - 3)} Q_n \\ &= \frac{1}{12}n + \left(\frac{\vartheta}{4} - \frac{1}{2} - \frac{2\vartheta - 5}{12}\right) + \mathcal{O}\left(\frac{1}{n}\right) \\ &= \frac{\vartheta + n - 1}{12} + \mathcal{O}\left(\frac{1}{n}\right). \end{aligned}$$

□

Als nächstes möchten wir einen Zentralen Grenzwertsatz für die Anzahl der Blätter herleiten. Dazu benutzen wir ein allgemeines Verfahren aus [7, Theorem 3.2], um Grenzwertsätze für Martingale zu gewinnen. Zunächst nennen wir $\{S_{n,i}, \mathcal{F}_{n,i} | 1 \leq i \leq k_n, n \geq 1\}$ ein Martingal Dreiecksschema, falls für alle $n \geq 1$ $(S_{n,i})_{1 \leq i \leq k_n}$ ein an $(\mathcal{F}_{n,i})_{1 \leq i \leq k_n}$ adaptiertes Martingal ist. Des Weiteren sagen wir, dass eine Folge $(X_n)_{n \geq 1}$ stabil gegen eine Zufallsvariable X konvergiert, falls für alle beschränkten stetigen Funktionen g und für alle beschränkten Zufallsvariablen W gilt $\lim E[Wg(X_n)] = E[Wg(X)]$. Mit der Wahl $W = 1$ erhält man aus stabiler Konvergenz insbesondere schwache Konvergenz. Mit [7, Theorem 3.2] erhalten wir dann Folgendes.

Theorem 4.9. *Sei $\{S_{n,i}, \mathcal{F}_{n,i} | 1 \leq i \leq k_n, n \geq 1\}$ ein quadratintegrierbares Martingal Dreiecksschema mit Erwartungswert 0, Martingaldifferenzen $X_{n,i} = S_{n,i} - S_{n,i-1}$ und sei η^2 eine fast sicher endliche Zufallsvariable. Es gelte*

$$\begin{aligned} \max_i |X_{n,i}| &\xrightarrow{P} 0, \\ \sum_i X_{n,i}^2 &\xrightarrow{P} \eta^2, \\ E\left[\max_i X_{n,i}^2\right] &\text{ ist in } n \text{ beschränkt,} \end{aligned}$$

und $\mathcal{F}_{n,i} \subset \mathcal{F}_{n+1,i}$ für $1 \leq i \leq k_n$, $n \geq 1$.

Dann gilt $S_{n,k_n} = \sum_i X_{n,i} \xrightarrow{d} Z$ (stabil), wobei die Zufallsvariable Z die charakteristische Funktion $E[\exp(-\frac{1}{2}\eta^2 t^2)]$ hat.

Die Anwendung dieses Theorems auf ein an $(\mathcal{F}_n)_{n \geq 1}$ adaptiertes Martingal $(Y_n)_{n \geq 1}$ erfolgt dann durch die Definitionen $S_{n,i} = \frac{1}{\sqrt{\text{Var}(Y_n)}} Y_i$, $k_n = n$ und $\mathcal{F}_{n,i} = \mathcal{F}_i$.

Theorem 4.10. Sei $L_n^{(\vartheta)}$ die Anzahl der Blätter im Hoppe-Baum mit n Knoten. Dann gilt

$$\frac{L_n^{(\vartheta)} - E[L_n^{(\vartheta)}]}{\sqrt{\text{Var}(L_n^{(\vartheta)})}} \xrightarrow{d} Z,$$

wobei Z standardnormalverteilt ist.

Der Beweis von Theorem 4.10 stützt sich auf Theorem 4.9 von Hall und Heyde. Für die nötigen Eigenschaften des Martingal Dreiecksschemas benötigen wir ein kleines Lemma, welches als Cesàro-Mittel bekannt ist.

Lemma 4.11. Sei $(a_n)_{n \geq 1}$ eine beliebige reellwertige Nullfolge. Dann gilt

$$\frac{1}{n} \sum_{i=1}^n a_i \xrightarrow{n \rightarrow \infty} 0.$$

Beweis. Sei $\varepsilon > 0$ beliebig. Da $(a_n)_{n \geq 1}$ eine Nullfolge ist, gibt es ein $k = k(\varepsilon) \in \mathbb{N}$ mit $|a_n| < \varepsilon$ für alle $n \geq k$. Insbesondere gilt also für alle $n \geq k$

$$\left| \sum_{i=1}^n a_i \right| \leq \sum_{i=1}^k |a_i| + (n-k)\varepsilon.$$

Daraus folgt für den Grenzwert

$$\limsup_{n \rightarrow \infty} \left| \frac{1}{n} \sum_{i=1}^n a_i \right| \leq \lim_{n \rightarrow \infty} \frac{1}{n} \left(\sum_{i=1}^k |a_i| + (n-k)\varepsilon \right) = \varepsilon.$$

Da dies für alle $\varepsilon > 0$ gilt, folgt $\limsup_{n \rightarrow \infty} \left| \frac{1}{n} \sum_{i=1}^n a_i \right| = 0$ und damit die Behauptung. \square

Beweis von Theorem 4.10. Nach Theorem 4.1 ist durch $X_1 = 0$,

$X_n = (\vartheta + n - 2)(L_n^{(\vartheta)} - E[L_n^{(\vartheta)}])$ für $n \geq 2$ und $\mathcal{F}_n = \sigma(L_1^{(\vartheta)}, \dots, L_n^{(\vartheta)})$ ein Martingal gegeben.

Daraus erhalten wir durch $S_{n,i} = \frac{1}{\sqrt{\text{Var}(X_n)}} X_i$ und $\mathcal{F}_{n,i} = \mathcal{F}_i$ ein Martingal Dreiecksschema

$\{S_{n,i}, \mathcal{F}_{n,i} | 1 \leq i \leq n, n \geq 1\}$. Auf dieses Martingal Dreiecksschema möchten wir nun Theorem 4.9 mit $\eta^2 = 1$ anwenden. Als Martingaldifferenzen des Martingal Dreiecksschemas erhalten wir für beliebige $n \geq 2$, $2 \leq i \leq n$

$$X_{n,i} = \frac{1}{\sqrt{\text{Var}(X_n)}} (X_i - X_{i-1}) = \frac{1}{\sqrt{\text{Var}(X_n)}} (L_i^{(\vartheta)} - E[L_i^{(\vartheta)}]) + (\vartheta + i - 3) \underbrace{(L_i^{(\vartheta)} - L_{i-1}^{(\vartheta)})}_{:= Y_i} - E[L_i^{(\vartheta)} - L_{i-1}^{(\vartheta)}],$$

wobei $Y_i = L_i^{(\vartheta)} - L_{i-1}^{(\vartheta)}$ eine Indikatorvariable ist und genau dann 1 wird, wenn als Vaterknoten für den i -ten Knoten kein Blatt ausgewählt wird. Um Theorem 4.9 anwenden zu können, brauchen wir

$$\max_i |X_{n,i}| \xrightarrow{P} 0,$$

$$\sum_{i=2}^n X_{n,i}^2 \xrightarrow{P} \eta^2,$$

$E \left[\max_i X_{n,i}^2 \right]$ ist in n beschränkt.

Es gilt für beliebige $n \geq 2$ und $i \leq n$

$$|X_{n,i}| = \left| \frac{1}{\sqrt{\text{Var}(X_n)}} \underbrace{(L_i^{(\vartheta)} - E[L_i^{(\vartheta)}])}_{\in [-i, i] \subset [-n, n]} + (\vartheta + i - 3) \underbrace{(Y_i - E[Y_i])}_{\in [-1, 1]} \right| \leq \frac{2n + \vartheta + 3}{\sqrt{\text{Var}(X_n)}},$$

somit erhalten wir wegen $\text{Var}(X_n) \sim \frac{n^3}{12}$ (Theorem 4.8)

$$\max_{2 \leq i \leq n} |X_{n,i}| \leq \frac{2n + \vartheta + 3}{\sqrt{\text{Var}(X_n)}} \xrightarrow{n \rightarrow \infty} 0 \text{ f.s.}$$

und damit auch insbesondere die stochastische Konvergenz. Ebenso erhalten wir daraus

$$\left| E \left[\max_i X_{n,i}^2 \right] \right| \leq \left(\frac{2n + \vartheta + 3}{\sqrt{\text{Var}(X_n)}} \right)^2,$$

was insbesondere in n beschränkt ist. Zu zeigen bleibt also nur noch die zweite Eigenschaft für die Martingaldifferenzen. Hier gilt für beliebige $n \geq 2$

$$\begin{aligned} \sum_{i=2}^n X_{n,i}^2 &= \frac{1}{\text{Var}(X_n)} \sum_{i=2}^n (L_i^{(\vartheta)} - E[L_i^{(\vartheta)}])^2 + 2(\vartheta + i - 3)(Y_i - E[Y_i]) \sum_{i=2}^n (L_i^{(\vartheta)} - E[L_i^{(\vartheta)}]) \\ &+ \frac{1}{\text{Var}(X_n)} \sum_{i=2}^n (\vartheta + i - 3)^2 (Y_i - E[Y_i])^2. \end{aligned}$$

$\underbrace{\hspace{15em}}_{:=T_1(n)} \quad \underbrace{\hspace{15em}}_{:=T_2(n)} \quad \underbrace{\hspace{15em}}_{:=T_3(n)}$

Nun müssen wir die Terme $T_1(n), T_2(n), T_3(n)$ etwas genauer betrachten. Für den ersten Term gilt

$$|T_1(n)| \leq \frac{n^3}{\text{Var}(X_n)} \frac{1}{n} \sum_{i=2}^n \left(\frac{L_i^{(\vartheta)} - E[L_i^{(\vartheta)}]}{i} \right)^2,$$

wobei $\left(\left(\frac{L_i^{(\vartheta)} - E[L_i^{(\vartheta)}]}{i} \right)^2 \right)_{i \geq 2}$ nach Theorem 4.7 fast sicher eine Nullfolge ist. Nach Lemma 4.11

konvergiert also $\frac{1}{n} \sum_{i=2}^n \left(\frac{L_i^{(\vartheta)} - E[L_i^{(\vartheta)}]}{i} \right)^2$ fast sicher gegen 0. Aus Theorem 4.8 erhält man

$\frac{n^3}{\text{Var}(X_n)} \rightarrow 12$ und damit insgesamt $T_1(n) \rightarrow 0$ fast sicher. Für den zweiten Term folgt analog

$$|T_2(n)| \leq 2 \frac{n^3}{\text{Var}(X_n)} \frac{1}{n} \sum_{i=2}^n \underbrace{\left| \frac{(\vartheta + i - 3)(Y_i - E[Y_i])i}{n^2} \right|}_{\leq \vartheta + 3} \left| \frac{L_i^{(\vartheta)} - E[L_i^{(\vartheta)}]}{i} \right|$$

und damit mit denselben Argumenten wie beim ersten Term $T_2(n) \rightarrow 0$ fast sicher. Für den Term $T_3(n)$ erinnern wir uns daran, dass wir beim Bestimmen der Varianz von $L_n^{(\vartheta)}$ (Theorem 4.8) gesehen haben, dass $E[Y_i] = \frac{1}{2} + \mathcal{O}\left(\frac{1}{i^2}\right)$ ist. Damit erhalten wir

$$\begin{aligned} T_3(n) &= \frac{1}{\text{Var}(X_n)} \sum_{i=2}^n (\vartheta + i - 3)^2 \left(Y_i - \frac{1}{2} + \left(\frac{1}{2} - E[Y_i] \right) \right)^2 \\ &= \frac{1}{\text{Var}(X_n)} \sum_{i=2}^n (\vartheta + i - 3)^2 \underbrace{\left(Y_i - \frac{1}{2} \right)^2}_{=\frac{1}{4}} + \frac{1}{\text{Var}(X_n)} 2 \underbrace{\sum_{i=2}^n (\vartheta + i - 3)^2 \left(\frac{1}{2} - E[Y_i] \right) \left(Y_i - \frac{1}{2} \right)}_{=\mathcal{O}(n) \text{ f.s.}} \\ &\quad + \frac{1}{\text{Var}(X_n)} \sum_{i=2}^n (\vartheta + i - 3)^2 \underbrace{\left(\frac{1}{2} - E[Y_i] \right)^2}_{=\mathcal{O}(1)}. \end{aligned}$$

Es folgt $T_3(n) = \frac{1}{\text{Var}(X_n)} \sum_{i=2}^n \frac{1}{4} (\vartheta + i - 3)^2 + o(1)$ fast sicher, wobei für die Summe gilt

$$\sum_{i=2}^n \frac{1}{4} (\vartheta + i - 3)^2 = \frac{1}{4} \sum_{i=1}^n i^2 + \mathcal{O}(n^2) = \frac{n^3}{12} + \mathcal{O}(n^2)$$

und damit insgesamt wegen $\text{Var}(X_n) \sim \frac{n^3}{12}$ (Theorem 4.8) $T_3(n) \rightarrow 1$ fast sicher. Also gilt insgesamt $\sum_{i=2}^n X_{n,i}^2 \rightarrow 1$ fast sicher und damit auch stochastisch. Theorem 4.9 liefert

$$\frac{X_n}{\sqrt{\text{Var}(X_n)}} = S_{n,n} \xrightarrow{d} Z(\text{stabil}),$$

wobei Z die charakteristische Funktion $E[\exp(-\frac{1}{2}t^2)] = \exp(-\frac{1}{2}t^2)$ hat und damit standardnormalverteilt ist. Wegen $\frac{X_n}{\sqrt{\text{Var}(X_n)}} = \frac{L_n^{(\vartheta)} - E[L_n^{(\vartheta)}]}{\sqrt{\text{Var}(L_n^{(\vartheta)})}}$ erhalten wir auch die (stabile) Verteilungskonvergenz für $L_n^{(\vartheta)}$ und damit die Behauptung. □

5 Die interne Pfadlänge im Hoppe-Baum

Wir werden nun die interne Pfadlänge im Hoppe-Baum betrachten, welche in gewissem Sinne Auskunft über die Komplexität oder Gesamtgröße des Baumes liefert. Mit der internen Pfadlänge bezeichnet man die Summe der Tiefen jedes Knotens im Baum, wir definieren also $I_n^{(\vartheta)} = \sum_{i=1}^n D_i^{(\vartheta)}$ als interne Pfadlänge des Hoppe-Baumes mit $n \in \mathbb{N}$ Knoten und Wurzelgewicht $\vartheta \in \mathbb{R}^+$, wobei die einzelnen Tiefen $D_i^{(\vartheta)}$ bereits in Kapitel 2 studiert wurden. Die Idee für das Vorgehen im ersten Teil des Abschnitts wurde bereits 1989 von Régnier [11] für die Analyse des Binärsuchbaums (und damit Quicksort) verwendet und 1991 von Mahmoud [8] auf zufällige rekursive Bäume angewandt, um die interne Pfadlänge auf Konvergenz zu untersuchen. Wie auch in Kapitel 4 wird dazu $I_n^{(\vartheta)}$ zu einem Martingal reskaliert, welches allerdings anders als in Kapitel 4 eine beschränkte Varianz haben wird, was uns nach dem Martingalkonvergenzsatz fast sichere und L_1 Konvergenz liefert. Im zweiten Teil des Abschnitts werden wir diesen Grenzwert mit den Methoden von Fill und Janson aus [6] analysieren, was uns die Existenz einer unendlich oft differenzierbaren, schnell abfallenden Dichte für dessen Verteilung liefern wird.

5.1 Konvergenz der internen Pfadlänge mit Martingalmethoden

Wir starten mit der Reskalierung von $I_n^{(\vartheta)}$ zu einem Martingal.

Theorem 5.1. *Sei für $i \geq 1$ $D_i^{(\vartheta)}$ die Tiefe des i -ten Knotens im Hoppe-Baum, sowie $I_n^{(\vartheta)} = \sum_{i=1}^n D_i^{(\vartheta)}$ die interne Pfadlänge des Hoppe-Baumes mit n Knoten. Ferner sei für $n \geq 1$ $\mathcal{F}_n = \sigma(D_1^{(\vartheta)}, \dots, D_n^{(\vartheta)})$ die von $D_1^{(\vartheta)}, \dots, D_n^{(\vartheta)}$ erzeugte σ -Algebra. Dann ist durch*

$$Z_n^{(\vartheta)} = \frac{1}{\vartheta + n - 1} I_n^{(\vartheta)} - \sum_{i=1}^{n-1} \frac{1}{\vartheta + i}$$

ein an $(\mathcal{F}_n)_{n \in \mathbb{N}}$ adaptiertes Martingal $(Z_n^{(\vartheta)})_{n \in \mathbb{N}}$ gegeben.

Beweis. Wir setzen $Z_n := Z_n^{(\vartheta)}$.

Die Adaptiertheit von $(Z_n)_{n \in \mathbb{N}}$ ist klar, da $(I_n^{(\vartheta)})_{n \in \mathbb{N}}$ an $(\mathcal{F}_n)_{n \in \mathbb{N}}$ adaptiert ist. Wir beobachten zunächst für beliebiges $n \geq 1$, dass fast sicher gilt

$$\begin{aligned} E[I_{n+1}^{(\vartheta)} | \mathcal{F}_n] &= E[I_n^{(\vartheta)} + D_{n+1}^{(\vartheta)} | \mathcal{F}_n] \\ &= I_n^{(\vartheta)} + E[D_{n+1}^{(\vartheta)} | \mathcal{F}_n] \\ &= I_n^{(\vartheta)} + 1 \cdot \frac{\vartheta}{\vartheta + n - 1} + \sum_{i=2}^n (D_i^{(\vartheta)} + 1) \frac{1}{\vartheta + n - 1}, \end{aligned}$$

da der $(n+1)$ -te Knoten Tiefe 1 hat, wenn er Kind der Wurzel im Hoppe-Baum wird, sowie bei gegebenen $D_1^{(\vartheta)}, \dots, D_n^{(\vartheta)}$ Tiefe $D_i^{(\vartheta)} + 1$ hat, wenn er Kind des i -ten Knoten im Hoppe-Baum wird.

Dies liefert uns fast sicher

$$E[I_{n+1}^{(\vartheta)} | \mathcal{F}_n] = I_n^{(\vartheta)} + \frac{1}{\vartheta + n - 1} \underbrace{\sum_{i=2}^n D_i^{(\vartheta)}}_{=I_n^{(\vartheta)}} + \frac{\vartheta}{\vartheta + n - 1} + \frac{n-1}{\vartheta + n - 1} = \frac{\vartheta + n}{\vartheta + n - 1} I_n^{(\vartheta)} + 1, \quad (4)$$

wobei $\sum_{i=2}^n D_i^{(\vartheta)} = I_n^{(\vartheta)}$ gilt, da die Tiefe $D_1^{(\vartheta)}$ der Wurzel gleich 0 ist.

Damit lässt sich nun leicht einsehen, dass $(Z_n)_{n \in \mathbb{N}}$ ein Martingal ist, denn es gilt fast sicher für beliebiges $n \geq 1$

$$\begin{aligned} E[Z_{n+1} | \mathcal{F}_n] &= \frac{1}{\vartheta + n} E[I_{n+1}^{(\vartheta)} | \mathcal{F}_n] - \sum_{i=1}^n \frac{1}{\vartheta + i} \\ &\stackrel{(4)}{=} \frac{1}{\vartheta + n} \left(\frac{\vartheta + n}{\vartheta + n - 1} I_n^{(\vartheta)} + 1 \right) - \sum_{i=1}^n \frac{1}{\vartheta + i} \\ &= Z_n. \end{aligned}$$

□

Als erstes Korollar aus Theorem 5.1 erhalten wir den Erwartungswert der internen Pfadlänge.

Korollar 5.2. *Sei $I_n^{(\vartheta)} = \sum_{i=1}^n D_i^{(\vartheta)}$ die interne Pfadlänge des Hoppe-Baumes mit n Knoten. Dann gilt*

$$E[I_n^{(\vartheta)}] = (\vartheta + n - 1) \sum_{i=1}^{n-1} \frac{1}{\vartheta + i} = n \log n + c_\vartheta n + o(n),$$

wobei $c_\vartheta = -\Psi(\vartheta + 1) = \gamma - \sum_{i=1}^{\lfloor \vartheta \rfloor} \frac{1}{i} - \varepsilon_\vartheta$ für ein $\varepsilon_\vartheta \in [0, \frac{1}{\lfloor \vartheta \rfloor}]$ die Konstante aus Lemma 1.3 bezeichnet.

Beweis. Mit dem Martingal $Z_n^{(\vartheta)} = \frac{1}{\vartheta+n-1} I_n^{(\vartheta)} - \sum_{i=1}^{n-1} \frac{1}{\vartheta+i}$ aus Theorem 5.1 erhalten wir

$$E \left[\frac{1}{\vartheta+n-1} I_n^{(\vartheta)} - \sum_{i=1}^{n-1} \frac{1}{\vartheta+i} \right] = E[Z_1^{(\vartheta)}] = 0,$$

da $I_1^{(\vartheta)} = D_1^{(\vartheta)} = 0$ ist, und damit durch Auflösen nach $E[I_n^{(\vartheta)}]$ die Behauptung. Die Asymptotik für den Erwartungswert erhält man aus Lemma 1.3. \square

Als nächsten Schritt wollen wir die fast sichere Konvergenz von $(Z_n^{(\vartheta)})_{n \geq 1}$ für allgemeines $\vartheta > 0$ nachweisen. Dazu werden wir zunächst die Varianz von $Z_n^{(\vartheta)}$ ermitteln um zu sehen, dass diese beschränkt bleibt. Anschließend erhalten wir dann aus dem Martingalkonvergenzsatz die gewünschte Aussage über die Konvergenz von $(Z_n^{(\vartheta)})_{n \geq 1}$.

Theorem 5.3. Sei $I_n^{(\vartheta)}$ die interne Pfadlänge des Hoppe-Baumes mit n Knoten sowie $Z_n^{(\vartheta)} = \frac{1}{\vartheta+n-1} I_n^{(\vartheta)} - \sum_{i=1}^{n-1} \frac{1}{\vartheta+i}$ das Martingal aus Theorem 5.1. Dann gilt

$$\text{Var}(Z_n^{(\vartheta)}) = \frac{2}{\vartheta} - \sum_{i=1}^{n-1} \left(\frac{1}{\vartheta+i-1} \right)^2 + o(1).$$

Insbesondere gilt also auch

$$\text{Var}(I_n^{(\vartheta)}) = \left(\frac{2}{\vartheta} - \sum_{i=1}^{n-1} \left(\frac{1}{\vartheta+i-1} \right)^2 \right) (\vartheta+n-1)^2 + o(n^2).$$

Beweis. Es bezeichne wieder $D_n := D_n^{(\vartheta)}$ die Tiefe des n -ten Knotens im Hoppe-Baum, sowie $I_n := I_n^{(\vartheta)} = \sum_{i=1}^n D_i$ die interne Pfadlänge, $\mathcal{F}_n = \sigma(D_1, \dots, D_n)$ und $Z_n := Z_n^{(\vartheta)}$. Es gilt

$$Z_n = \frac{I_n - E[I_n]}{\vartheta+n-1} = \frac{\vartheta+n-2}{\vartheta+n-1} \underbrace{\frac{I_{n-1} - E[I_{n-1}]}{\vartheta+n-2}}_{=Z_{n-1}} + \frac{D_n - E[D_n]}{\vartheta+n-1}.$$

Wir erhalten also

$$E[Z_n^2] = \left(\frac{\vartheta+n-2}{\vartheta+n-1} \right)^2 E[Z_{n-1}^2] + 2 \frac{\vartheta+n-2}{(\vartheta+n-1)^2} E[Z_{n-1}(D_n - E[D_n])] + \left(\frac{1}{\vartheta+n-1} \right)^2 \text{Var}(D_n). \quad (5)$$

Die Varianz der Tiefe D_n haben wir bereits in Abschnitt 2 ermittelt, es bleibt also nur noch den mittleren Term zu bestimmen:

$$\begin{aligned} E[Z_{n-1}(D_n - E[D_n])] &= E[Z_{n-1}D_n] - \underbrace{E[Z_{n-1}]E[D_n]}_{=0} \\ &= E[Z_{n-1}E[D_n|\mathcal{F}_{n-1}]] \\ &= E \left[Z_{n-1} \left(\frac{\vartheta}{\vartheta+n-2}(D_1+1) + \sum_{i=2}^{n-1} \frac{1}{\vartheta+n-2}(D_i+1) \right) \right] \\ &= E \left[Z_{n-1} \left(\frac{1}{\vartheta+n-2} I_{n-1} + 1 \right) \right] \quad (\text{weil } D_1 = 0 \text{ ist}) \\ &= E \left[Z_{n-1} \left(Z_{n-1} + \frac{E[I_{n-1}]}{\vartheta+n-2} + 1 \right) \right] \\ &= E[Z_{n-1}^2] + 0. \end{aligned}$$

Setzt man dies nun in (5) ein, erhält man folgende rekursive Gleichung für das zweite Moment

$$E[Z_n^2] = \frac{(\vartheta + n - 2)(\vartheta + n)}{(\vartheta + n - 1)^2} E[Z_{n-1}^2] + \left(\frac{1}{\vartheta + n - 1} \right)^2 \text{Var}(D_n).$$

Diese Rekursion lässt sich dann etwas leichter lösen, indem man mit $Q_n := \frac{\vartheta+n-1}{\vartheta+n} E[Z_n^2]$ substituiert. Nun muss lediglich das Ergebnis $\text{Var}(D_n) = \sum_{i=2}^{n-1} \frac{1}{\vartheta+i-1} - \sum_{i=2}^{n-1} \left(\frac{1}{\vartheta+i-1} \right)^2$ aus Theorem 2.3 eingesetzt werden und die Rekursionsgleichung gelöst werden. Man erhält

$$\begin{aligned} Q_n &= \frac{2}{\vartheta} - \sum_{i=1}^{n-1} \left(\frac{1}{\vartheta+i-1} \right)^2 - \frac{2}{\vartheta+n-1} - \frac{1}{\vartheta+n} \sum_{i=2}^{n-1} \left(\frac{1}{\vartheta+i-1} - \left(\frac{1}{\vartheta+i-1} \right)^2 \right) \\ &= \frac{2}{\vartheta} - \sum_{i=1}^{n-1} \left(\frac{1}{\vartheta+i-1} \right)^2 + o(1) \end{aligned}$$

und damit die Behauptung, da $\text{Var}(Z_n) = E[Z_n^2] \sim Q_n$ ist.

Die Varianz der internen Pfadlänge erhält man nun aus

$$E[Z_n^2] = \left(\frac{1}{\vartheta+n-1} \right)^2 \text{Var}(I_n).$$

□

Mit Theorem 5.3 wissen wir nun also insbesondere, dass die Varianz von $Z_n^{(\vartheta)}$ beschränkt bleibt. Da wir bereits am Anfang des Kapitels gezeigt haben, dass $(Z_n^{(\vartheta)})_{n \geq 1}$ ein Martingal ist, können wir nun durch den Martingalkonvergenzsatz schließen, dass $(Z_n^{(\vartheta)})_{n \geq 1}$ fast sicher konvergiert.

Korollar 5.4. *Die reskalierte interne Pfadlänge $(Z_n^{(\vartheta)})_{n \geq 1}$ mit $Z_n^{(\vartheta)} = \frac{I_n^{(\vartheta)} - E[I_n^{(\vartheta)}]}{\vartheta+n-1}$ ist ein gleichgradig integrierbares Martingal und konvergiert somit fast sicher und in L_1 gegen eine integrierbare Zufallsvariable $Z_\infty^{(\vartheta)}$.*

Beweis. In Theorem 5.1 wurde gezeigt, dass $(Z_n^{(\vartheta)})_{n \geq 1}$ ein Martingal ist. Mit dem Ergebnis $c := \sup_{n \geq 1} \{E[(Z_n^{(\vartheta)})^2]\} < \infty$ aus Theorem 5.3 erhalten wir nach der Chebyshev Ungleichung $P(|Z_n^{(\vartheta)}| > t) \leq \frac{c}{t^2}$ für beliebige $t > 0$ und $n \geq 1$. Für $\varepsilon > 0$ erhält man mit $K = \frac{c}{\varepsilon}$, dass für beliebige $n \geq 1$ gilt

$$E \left[|Z_n^{(\vartheta)}| \cdot \mathbf{1}_{\{|Z_n^{(\vartheta)}| \geq K\}} \right] \leq \int_K^\infty \frac{c}{t^2} dt = \frac{c}{K} = \varepsilon.$$

Es folgt, dass $(Z_n^{(\vartheta)})_{n \geq 1}$ gleichgradig integrierbar ist. Mittels Martingalkonvergenzsätzen erhalten wir die Konvergenzen. □

5.2 Die Verteilung des Grenzwertes $Z_\infty^{(\vartheta)}$

Wir haben gesehen, dass $\left(\frac{I_n^{(\vartheta)} - E[I_n^{(\vartheta)}]}{\vartheta+n-1} \right)_{n \geq 1}$ fast sicher gegen eine integrierbare Zufallsvariable $Z_\infty^{(\vartheta)}$ konvergiert und wollen nun die Verteilung dieses Grenzwertes näher analysieren. Dazu werden wir die Existenz einer unendlich oft differenzierbaren, schnell abfallenden Dichte mit dem Verfahren von Fill und Janson aus [6] nachweisen.

Um dies zu zeigen, werden wir für die charakteristische Funktion $\phi_\vartheta(t) = E[\exp(itZ_\infty^{(\vartheta)})]$ beweisen, dass es für jeden Exponent $p > 0$ eine Konstante $C_p^{(\vartheta)}$ gibt, sodass $|\phi_\vartheta(t)| \leq C_p^{(\vartheta)} |t|^{-p}$ für alle

$t \in \mathbb{R}$ ist. Damit werden wir dann durch die Fourier Inversionsformel die Existenz der Dichte zeigen können.

Als ersten wichtigen Schritt für die Analyse von $\phi_\vartheta(t)$ beobachten wir zunächst, dass der Grenzwert $Z_\infty^{(\vartheta)}$ der Gleichung aus dem folgenden Theorem genügt.

Theorem 5.5. Für $\vartheta > 0$ sei $Z_\infty^{(\vartheta)}$ der Grenzwert der reskalierten internen Pfadlänge $\left(\frac{I_n^{(\vartheta)} - E[I_n^{(\vartheta)}]}{\vartheta + n - 1}\right)_{n \geq 1}$, welche nach Korollar 5.4 fast sicher konvergiert. Dann gilt

$$Z_\infty^{(\vartheta)} \stackrel{d}{=} B\bar{Z}_\infty^{(1)} + (1 - B)\bar{\bar{Z}}_\infty^{(\vartheta)} + B \log B + (1 - B) \log(1 - B) + (\gamma - c_\vartheta)B,$$

wobei $B, \bar{Z}_\infty^{(1)}, \bar{\bar{Z}}_\infty^{(\vartheta)}$ unabhängig sind mit Verteilungen $\mathcal{L}(B) = \text{Beta}(1, \vartheta)$, $\mathcal{L}(\bar{Z}_\infty^{(1)}) = \mathcal{L}(Z_\infty^{(1)})$, $\mathcal{L}(\bar{\bar{Z}}_\infty^{(\vartheta)}) = \mathcal{L}(Z_\infty^{(\vartheta)})$, $\gamma \approx 0,5772156649$ die Euler Konstante ist und c_ϑ die Konstante aus Lemma 1.3 bezeichnet.

Beweis. Wir bezeichnen mit $N_n^{(1)}$ die Anzahl der Knoten im ersten Teilbaum des Hoppe-Baumes mit n Knoten. Die grundlegende Beobachtung für den Beweis ist nun folgende: Zerlegen wir den Hoppe-Baum in den ersten Teilbaum mit Knoten 2 als Wurzel und den verbleibenden Baum, so ist der erste Teilbaum für $N_n^{(1)} = k$ in Verteilung ein zufälliger rekursiver Baum mit k Knoten, da alle Knoten in dem Teilbaum Gewicht 1 haben und damit bei Einfügen eines neuen Knoten in den Teilbaum uniform ein Vaterknoten gewählt wird. Außerdem ist für $N_n^{(1)} = k$ der verbleibende Baum in Verteilung wieder ein Hoppe-Baum mit $n - k$ Knoten, da hier die Wurzel die ursprüngliche Wurzel des Baumes ist und damit Gewicht ϑ hat. Abgesehen von den Anzahl der Knoten ist die Struktur der Teilbäume dieser Zerlegung natürlich unabhängig. In Abbildung 3 ist die Zerlegung für einen Baum mit 10 Knoten dargestellt. Bildet man also die internen Pfadlängen

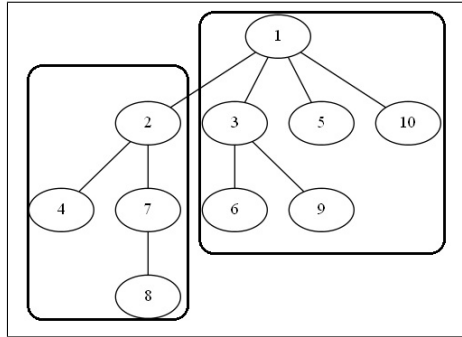


Abbildung 3: Die Zerlegung eines Hoppe-Baumes in den ersten Teilbaum und den Restbaum. Die interne Pfadlänge im ersten Teilbaum beträgt 4, im Restbaum ist die interne Pfadlänge 7. Berücksichtigt man für die 4 Knoten im ersten Teilbaum, dass ihre tatsächliche Tiefe im Gesamtbaum um 1 größer ist, erhält man die interne Pfadlänge des gesamten Baumes als $4 + 4 + 7 = 15$.

beider Teilbäume getrennt und berücksichtigt, dass jede Tiefe im ersten Teilbaum tatsächlich um 1 größer ist, wenn man den Knoten im kompletten Baum statt im ersten Teilbaum betrachtet (im ersten Teilbaum betrachtet wird die Kante von der Wurzel des Hoppe-Baumes zu Knoten 2 nicht berücksichtigt), so erhält man für die interne Pfadlänge des gesamten Hoppe-Baumes

$$I_n^{(\vartheta)} \stackrel{d}{=} \bar{I}_{N_n^{(1)}}^{(1)} + N_n^{(1)} + \bar{I}_{n - N_n^{(1)}}^{(\vartheta)},$$

wobei $(\bar{I}_n^{(1)})_{n \geq 1}$, $(\bar{I}_n^{(\vartheta)})_{n \geq 1}$ und $N_n^{(1)}$ unabhängig sind mit $\mathcal{L}((\bar{I}_n^{(1)})_{n \geq 1}) = \mathcal{L}((I_n^{(1)})_{n \geq 1})$ und $\mathcal{L}((\bar{I}_n^{(\vartheta)})_{n \geq 1}) = \mathcal{L}((I_n^{(\vartheta)})_{n \geq 1})$. Dies liefert

$$\begin{aligned} \frac{I_n^{(\vartheta)} - E[I_n^{(\vartheta)}]}{\vartheta + n - 1} &\stackrel{d}{=} \frac{N_n^{(1)}}{\vartheta + n - 1} \frac{\bar{I}_{N_n^{(1)}}^{(1)} - E[\bar{I}_{N_n^{(1)}}^{(1)} | N_n^{(1)}]}{N_n^{(1)}} + \left(1 - \frac{N_n^{(1)}}{\vartheta + n - 1}\right) \frac{\bar{I}_{n-N_n^{(1)}}^{(\vartheta)} - E[\bar{I}_{n-N_n^{(1)}}^{(\vartheta)} | N_n^{(1)}]}{\vartheta + n - N_n^{(1)} - 1} \\ &\quad + \frac{N_n^{(1)}}{\vartheta + n - 1} + \frac{E[\bar{I}_{N_n^{(1)}}^{(1)} | N_n^{(1)}] + E[\bar{I}_{n-N_n^{(1)}}^{(\vartheta)} | N_n^{(1)}] - E[I_n^{(\vartheta)}]}{\vartheta + n - 1}, \end{aligned}$$

und damit wegen $Z_n^{(\vartheta)} = \frac{I_n^{(\vartheta)} - E[I_n^{(\vartheta)}]}{\vartheta + n - 1}$

$$Z_n^{(\vartheta)} \stackrel{d}{=} \frac{N_n^{(1)}}{\vartheta + n - 1} \bar{Z}_{N_n^{(1)}}^{(1)} + \left(1 - \frac{N_n^{(1)}}{\vartheta + n - 1}\right) \bar{Z}_{n-N_n^{(1)}}^{(\vartheta)} + g_n \left(\frac{N_n^{(1)}}{n}\right),$$

wobei wieder $(\bar{Z}_n^{(1)})_{n \geq 1}$, $(\bar{Z}_n^{(\vartheta)})_{n \geq 1}$ und $N_n^{(1)}$ unabhängig sind, $\mathcal{L}((\bar{Z}_n^{(1)})_{n \geq 1}) = \mathcal{L}((Z_n^{(1)})_{n \geq 1})$, $\mathcal{L}((\bar{Z}_n^{(\vartheta)})_{n \geq 1}) = \mathcal{L}((Z_n^{(\vartheta)})_{n \geq 1})$ ist, sowie $g_n : \{\frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}\} \rightarrow \mathbb{R}$ gegeben ist durch

$$g_n(k) = \frac{n}{\vartheta + n - 1} k + \frac{E[I_{nk}^{(1)}] + E[I_{n(1-k)}^{(\vartheta)}] - E[I_n^{(\vartheta)}]}{\vartheta + n - 1}.$$

Mit Theorem 1.1 und Bemerkung 1.2 wissen wir bereits, dass $\left(\frac{N_n^{(1)}}{n}\right)_{n \geq 1}$ fast sicher gegen eine Beta(1, ϑ)-verteilte Zufallsvariable B konvergiert, welche wiederum fast sicher im Intervall $(0, 1)$ liegt. Insbesondere gilt also $N_n^{(1)} \rightarrow \infty$, $n - N_n^{(1)} \rightarrow \infty$ fast sicher und damit $\bar{Z}_{N_n^{(1)}}^{(1)} \rightarrow \bar{Z}_\infty^{(1)}$, $\bar{Z}_{n-N_n^{(1)}}^{(\vartheta)} \rightarrow \bar{Z}_\infty^{(\vartheta)}$ fast sicher wegen der Konvergenz der reskalierten internen Pfadlänge aus Korollar 5.4. Ebenso gilt $Z_n^{(\vartheta)} \rightarrow Z_\infty^{(\vartheta)}$ fast sicher, also bleibt nur noch das asymptotische Verhalten der Funktion g_n zu bestimmen. Aus Korollar 5.2 erhalten wir, dass für den Erwartungswert der Pfadlänge für beliebige $\vartheta > 0$ gilt $E[I_n^{(\vartheta)}] = n \log n + c_\vartheta n + r_\vartheta(n)$, wobei für den Restterm gilt $\frac{r_\vartheta(n)}{n} \rightarrow 0$ für $n \rightarrow \infty$. Dies liefert

$$\begin{aligned} g_n(k) &= \frac{n}{\vartheta + n - 1} k + \frac{nk \log(nk) + n(1-k) \log(n(1-k)) - n \log n}{\vartheta + n - 1} \\ &\quad + \frac{c_1 nk + c_\vartheta n(1-k) - c_\vartheta n}{\vartheta + n - 1} + \frac{r_1(nk) + r_\vartheta(n(1-k)) - r_\vartheta(n)}{\vartheta + n - 1} \\ &= \frac{n}{\vartheta + n - 1} k \log k + \frac{n}{\vartheta + n - 1} (1-k) \log(1-k) + \frac{n}{\vartheta + n - 1} (c_1 - c_\vartheta + 1)k \\ &\quad + \frac{r_1(nk) + r_\vartheta(n(1-k)) - r_\vartheta(n)}{\vartheta + n - 1}. \end{aligned}$$

Setzt man $k = \frac{N_n^{(1)}}{n}$, so konvergiert der letzte Bruch mit den Resttermen fast sicher gegen 0 und wir erhalten insgesamt wegen $\frac{N_n^{(1)}}{n} \rightarrow B$ fast sicher, dass fast sicher gilt

$$g_n \left(\frac{N_n^{(1)}}{n}\right) \rightarrow B \log B + (1-B) \log(1-B) + (c_1 - c_\vartheta + 1)B.$$

Setzt man alles zusammen, folgt wegen $c_1 + 1 = \gamma$ (siehe Lemma 1.3) die Behauptung. \square

Aus Theorem 5.5 können wir nun einen Zusammenhang der charakteristischen Funktionen von $Z_\infty^{(\vartheta)}$ und $Z_\infty^{(1)}$ gewinnen, der zentral für die weitere Analyse sein wird.

Korollar 5.6. Für die charakteristische Funktion $\phi_\vartheta(t) = E[\exp(itZ_\infty^{(\vartheta)})]$ gilt für jedes $t \in \mathbb{R}$

$$|\phi_\vartheta(t)| \leq \int_0^1 |\phi_1(xt)| |\phi_\vartheta((1-x)t)| \vartheta(1-x)^{\vartheta-1} dx.$$

Beweis. Mit Theorem 5.5 erhalten wir

$$E[\exp(itZ_\infty^{(\vartheta)})] = E[\exp(it(B\bar{Z}_\infty^{(1)} + (1-B)\bar{\bar{Z}}_\infty^{(\vartheta)} + f(B)))],$$

wobei die Zufallsvariablen unabhängig sind, B die Dichte $\vartheta(1-x)^{\vartheta-1}\mathbf{1}_{[0,1]}(x)$ der Beta(1, ϑ)-Verteilung hat und $f(B) = B \log B + (1-B) \log(1-B) + (\gamma - d_\vartheta)B$ ist. Dies liefert

$$\begin{aligned} |\phi_\vartheta(t)| &= \left| \int_0^1 E[\exp(it(x\bar{Z}_\infty^{(1)} + (1-x)\bar{\bar{Z}}_\infty^{(\vartheta)} + f(x)))] \vartheta(1-x)^{\vartheta-1} dx \right| \\ &\leq \int_0^1 |\phi_1(xt)| |\phi_\vartheta((1-x)t)| \underbrace{|\exp(itf(x))|}_{=1} \vartheta(1-x)^{\vartheta-1} dx \end{aligned}$$

und damit die Behauptung. \square

Nun werden wir für den Anfang zunächst zeigen, dass wir zumindest für $\vartheta = 1$ eine Konstante $C_{0.5}^{(1)}$ finden, sodass $|\phi_1(t)| \leq C_{0.5}^{(1)}|t|^{-\frac{1}{2}}$ ist. Ist dies geschafft, so folgern wir aus Korollar 5.6 zunächst für jedes $\vartheta > 0$ die Existenz einer Konstante $C_{0.5}^{(\vartheta)}$ mit $|\phi_\vartheta(t)| \leq C_{0.5}^{(\vartheta)}|t|^{-\frac{1}{2}}$. Danach werden wir, ebenfalls mit Hilfe von Korollar 5.6, daraus die Existenz von Konstanten $C_p^{(\vartheta)}$ zu beliebigen Exponenten $p > 0$ und beliebigen $\vartheta > 0$ mit $|\phi_\vartheta(t)| \leq C_p^{(\vartheta)}|t|^{-p}$ folgern.

Lemma 5.7. Sei $\phi_1(t) = E[\exp(itZ_\infty^{(1)})]$. Dann existiert eine Konstante $C_{0.5}^{(1)}$, sodass für alle $t \in \mathbb{R}$ gilt

$$|\phi_1(t)| \leq C_{0.5}^{(1)}|t|^{-\frac{1}{2}}.$$

Insbesondere gibt es dann auch für jedes $0 < p \leq \frac{1}{2}$ Konstanten $C_p^{(1)} \leq \max\{C_{0.5}^{(1)}, 1\}$, sodass für alle $t \in \mathbb{R}$ gilt

$$|\phi_1(t)| \leq C_p^{(1)}|t|^{-p}.$$

Der Beweis wird analog zu dem Beweis [6, Lemma 2.4] von Fill und Janson geführt, dazu brauchen wir folgendes Lemma.

Lemma 5.8. Sei h eine zweimal stetig differenzierbare Funktion auf dem offenen Intervall (a, b) mit

$$h'(x) \geq c > 0 \text{ und } h''(x) \geq 0 \text{ für alle } x \in (a, b).$$

Dann gilt für alle $t > 0$

$$\left| \int_a^b e^{ith(x)} dx \right| \leq \frac{2}{ct}.$$

Beweis. Den Beweis findet man in [6, Lemma 2.2]. \square

Beweis von Lemma 5.7. Wir bezeichnen für beliebige $x, y \in \mathbb{R}$ mit $h_{x,y} : (0, 1) \rightarrow \mathbb{R}$ die Funktion $h_{x,y}(u) = ux + (1-u)y + u \log u + (1-u) \log(1-u) + u$ und erhalten damit nach Theorem 5.5 $\phi_1(t) = E[\exp(it h_{\bar{Z}_\infty^{(1)}, \bar{\bar{Z}}_\infty^{(1)}}(U))]$, wobei U auf $[0, 1]$ gleichverteilt ist. Unser Ziel wird

es sein, die stärkere Aussage $E[e^{ith_{x,y}(U)}] \leq 2\sqrt{2}|t|^{-\frac{1}{2}}$ für alle $x, y \in \mathbb{R}$ zu zeigen, woraus wir durch die Glättungsregel direkt die Behauptung gewinnen.

Seien $x, y \in \mathbb{R}$ beliebig. Dann gilt für die Funktion $h_{x,y}$

$$\begin{aligned} h'_{x,y}(u) &= x - y + 1 + \log u - \log(1 - u), \\ h''_{x,y}(u) &= \frac{1}{u} + \frac{1}{1 - u} = \frac{1}{u(1 - u)} \geq 4. \end{aligned}$$

Da $h'_{x,y}$ auch negativ werden kann, lässt sich Lemma 5.8 noch nicht direkt anwenden, aber wir erhalten als einzige Nullstelle

$$h'_{x,y}(u) = 0 \iff u = 1 - (e^{y-x-1} + 1)^{-1} =: \alpha_{x,y} \in (0, 1).$$

Sei $\gamma > 0$ passend gewählt (die optimale Wahl ergibt sich aus dem Beweis) und seien $a = \alpha_{x,y} + \gamma t^{-\frac{1}{2}}$, $b = 1$ (ist $a \geq 1$ folgt braucht man nur den trivialen Fall betrachten, der im Anschluss besprochen wird). Dann gilt für alle $u \in (a, b)$

$$h'_{x,y}(u) = \int_{\alpha_{x,y}}^u h''_{x,y}(z) dz \geq 4(u - \alpha_{x,y}) \geq 4\gamma t^{-\frac{1}{2}}$$

und Lemma 5.8 liefert für die auf (a, b) eingeschränkte Funktion $h_{x,y}$ für $t > 0$

$$\left| \int_a^b e^{ith_{x,y}(u)} du \right| \leq \frac{2}{4\gamma t^{-\frac{1}{2}} t} = \frac{1}{2\gamma} t^{-\frac{1}{2}}.$$

Da außerdem trivialerweise gilt $\left| \int_{\alpha_{x,y}}^a e^{ith_{x,y}(u)} du \right| \leq a - \alpha_{x,y} = \gamma t^{-\frac{1}{2}}$, erhalten wir insgesamt

$$\left| \int_{\alpha_{x,y}}^1 e^{ith_{x,y}(u)} du \right| \leq \left(\gamma + \frac{1}{2\gamma} \right) t^{-\frac{1}{2}}.$$

Wählt man nun $\gamma = \frac{1}{\sqrt{2}}$, so erhält man $\sqrt{2}t^{-\frac{1}{2}}$ als Schranke. Für das Integral auf $(0, \alpha_{x,y})$ kann man analog vorgehen, sofern man die Funktion $g_{x,y}(u) = h_{x,y}(1 - u)$ betrachtet und mit dem Integral

$$\int_{1-\alpha_{x,y}}^1 e^{itg_{x,y}(u)} du = \int_0^{\alpha_{x,y}} e^{ith_{x,y}(u)} du$$

arbeitet. Man erhält insgesamt für alle $t > 0$

$$\left| \int_0^1 e^{ith_{x,y}(u)} du \right| \leq 2\sqrt{2}t^{-\frac{1}{2}}.$$

Wegen $\left| \int_0^1 e^{ith_{x,y}(u)} du \right| = \left| \int_0^1 e^{-ith_{x,y}(u)} du \right|$ gilt das Ergebnis auch für $t < 0$ mit der Schranke $2\sqrt{2}|t|^{-\frac{1}{2}}$ und wir erhalten insgesamt für beliebige $x, y, t \in \mathbb{R}$

$$\left| E \left[e^{ith_{x,y}(U)} \right] \right| \leq 2\sqrt{2}|t|^{-\frac{1}{2}}$$

und damit

$$|\phi_1(t)| \leq E \left| E \left[\exp(ith_{\bar{Z}_\infty^{(1)}, \bar{\bar{Z}}_\infty^{(1)}}(U)) \middle| \bar{Z}_\infty^{(1)}, \bar{\bar{Z}}_\infty^{(1)} \right] \right| \leq 2\sqrt{2}|t|^{-\frac{1}{2}}.$$

Für beliebige Exponenten $p \in (0, \frac{1}{2}]$ erhält man die Behauptung mit $C_p^{(1)} = \max\{1, C_{0.5}^{(1)}\}$, weil hier für $|t| \leq 1$ trivialerweise gilt $|\phi_1(t)| \leq 1 \leq C_p^{(1)}|t|^{-p}$ und für $|t| > 1$ gilt $|\phi_1(t)| \leq C_{0.5}^{(1)}|t|^{-\frac{1}{2}} \leq C_p^{(1)}|t|^{-p} \leq C_p^{(1)}|t|^{-\frac{1}{2}}$. \square

Lemma 5.9. Sei $\vartheta > 0$ beliebig und $\phi_\vartheta(t) = E[\exp(itZ_\infty^{(\vartheta)})]$. Dann gibt es für jedes $0 < p \leq \frac{1}{2}$ eine Konstante $C_p^{(\vartheta)}$, sodass für alle $t \in \mathbb{R}$ gilt

$$|\phi_\vartheta(t)| \leq C_p^{(\vartheta)} |t|^{-p}.$$

Beweis. Sei $0 < p \leq \frac{1}{2}$ beliebig. Nach Lemma 5.7 gibt es eine Konstante $C_p^{(1)}$ mit $|\phi_1(t)| \leq |t|^{-p}$ für alle $t \in \mathbb{R}$. Mit Korollar 5.6 erhalten wir

$$\begin{aligned} |\phi_\vartheta(t)| &\leq \int_0^1 \underbrace{|\phi_1(xt)|}_{\leq C_p^{(1)} x^{-p} |t|^{-p}} \underbrace{|\phi_\vartheta((1-x)t)|}_{\leq 1} \vartheta (1-x)^{\vartheta-1} dx \\ &\leq |t|^{-p} C_p^{(1)} \vartheta \underbrace{\int_0^1 x^{-p} (1-x)^{\vartheta-1} dx}_{< \infty}, \end{aligned}$$

wobei das verbliebene Integral ein beta-Integral ist, was für Exponenten $-p, (\vartheta-1) > -1$ sicherlich endlich ist. Es gibt also eine Konstante $C_p^{(\vartheta)}$, die höchstens $C_p^{(1)} \vartheta \int_0^1 x^{-p} (1-x)^{\vartheta-1} dx$ ist, sodass für alle $t \in \mathbb{R}$ gilt

$$|\phi_\vartheta(t)| \leq C_p^{(\vartheta)} |t|^{-p}.$$

□

Theorem 5.10. Sei $\vartheta > 0$ beliebig und $\phi_\vartheta(t) = E[\exp(itZ_\infty^{(\vartheta)})]$. Dann gibt es für jedes $p > 0$ eine Konstante $C_p^{(\vartheta)}$, sodass für alle $t \in \mathbb{R}$ gilt

$$|\phi_\vartheta(t)| \leq C_p^{(\vartheta)} |t|^{-p}.$$

Der Beweis für Theorem 5.10 ergibt sich induktiv aus Lemma 5.9 und den folgenden zwei Lemmata.

Lemma 5.11. Sei $\vartheta > 0$ beliebig und sei $0 < p_1 < 1$ so gewählt, dass es für p_1 und $p_2 := \min\{p_1, \frac{\vartheta}{2}\}$ Konstanten $C_{p_1}^{(1)}$ und $C_{p_2}^{(\vartheta)}$ gibt mit $|\phi_1(t)| \leq C_{p_1}^{(1)} |t|^{-p_1}$, $|\phi_\vartheta(t)| \leq C_{p_2}^{(\vartheta)} |t|^{-p_2}$ für alle $t \in \mathbb{R}$. Dann gibt es eine Konstante $C_{p_1+p_2}^{(\vartheta)}$, sodass für alle $t \in \mathbb{R}$ gilt

$$|\phi_\vartheta(t)| \leq C_{p_1+p_2}^{(\vartheta)} |t|^{-(p_1+p_2)}.$$

Beweis. Sei $\vartheta > 0$ beliebig und p_1, p_2 wie im Lemma beschrieben. Nach Korollar 5.6 gilt

$$\begin{aligned} |\phi_\vartheta(t)| &\leq \int_0^1 \underbrace{|\phi_1(xt)|}_{\leq C_{p_1}^{(1)} (x|t|)^{-p_1}} \underbrace{|\phi_\vartheta((1-x)t)|}_{\leq C_{p_2}^{(\vartheta)} ((1-x)|t|)^{-p_2}} \vartheta (1-x)^{\vartheta-1} dx \\ &\leq |t|^{-(p_1+p_2)} C_{p_1}^{(1)} C_{p_2}^{(\vartheta)} \vartheta \int_0^1 x^{-p_1} (1-x)^{\vartheta-p_2-1} dx, \end{aligned}$$

wobei das verbliebene beta-Integral endlich ist, da $-p_1 > -1$ und $\vartheta - p_2 - 1 \geq \frac{\vartheta}{2} - 1 > -1$ sind. Somit existiert eine Konstante $C_{p_1+p_2}^{(\vartheta)}$, die höchstens $C_{p_1}^{(1)} C_{p_2}^{(\vartheta)} \vartheta \int_0^1 x^{-p_1} (1-x)^{\vartheta-p_2-1} dx$ ist, sodass für alle $t \in \mathbb{R}$ gilt

$$|\phi_\vartheta(t)| \leq C_{p_1+p_2}^{(\vartheta)} |t|^{-(p_1+p_2)}.$$

□

Lemma 5.12. Sei $\vartheta > 0$ beliebig und sei $p > 1$ so gewählt, dass es Konstanten $C_p^{(1)}$ und $C_p^{(\vartheta)}$ gibt mit $|\phi_1(t)| \leq C_p^{(1)}|t|^{-p}$ und $|\phi_\vartheta(t)| \leq C_p^{(\vartheta)}|t|^{-p}$ für alle $t \in \mathbb{R}$. Dann gibt es für $p' = p + \min\{1, \vartheta\}$ eine Konstante $C_{p'}^{(\vartheta)}$, sodass für alle $t \in \mathbb{R}$ gilt

$$|\phi_\vartheta(t)| \leq C_{p'}^{(\vartheta)}|t|^{-p'}.$$

Beweis. Sei $\vartheta > 0$ und $p > 1$ wie im Lemma beschrieben.

Weil $|\phi_\vartheta(t)|, |\phi_1(t)| \leq 1$ sind, gilt nach Voraussetzung für alle $t \in \mathbb{R}$, dass $|\phi_\vartheta(t)| \leq \min\{C_p^{(\vartheta)}|t|^{-p}, 1\}$ und $|\phi_1(t)| \leq \min\{C_p^{(1)}|t|^{-p}, 1\}$ sind.

Damit erhalten wir zunächst mit Korollar 5.6 für alle $t \geq 2(\max\{C_p^{(1)}, C_p^{(\vartheta)}\})^{\frac{1}{p}}$:

$$\begin{aligned} |\phi_\vartheta(t)| &\leq \int_0^1 |\phi_1(xt)| |\phi_\vartheta((1-x)t)| \vartheta (1-x)^{\vartheta-1} dx \\ &\leq \int_0^1 \min\{C_p^{(1)}|xt|^{-p}, 1\} \min\{C_p^{(\vartheta)}|(1-x)t|^{-p}, 1\} \vartheta (1-x)^{\vartheta-1} dx \\ &= \underbrace{\int_0^{(C_p^{(1)})^{\frac{1}{p}} t^{-1}} C_p^{(\vartheta)} |(1-x)t|^{-p} \vartheta (1-x)^{\vartheta-1} dx}_{:=T_1} \\ &\quad + \underbrace{\int_{(C_p^{(1)})^{\frac{1}{p}} t^{-1}}^{1-(C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1}} C_p^{(1)} |xt|^{-p} C_p^{(\vartheta)} |(1-x)t|^{-p} \vartheta (1-x)^{\vartheta-1} dx}_{:=T_2} \\ &\quad + \underbrace{\int_{1-(C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1}}^1 C_p^{(1)} |xt|^{-p} \vartheta (1-x)^{\vartheta-1} dx}_{:=T_3}. \end{aligned}$$

Wir analysieren nun die drei Terme T_1 , T_2 und T_3 getrennt und zeigen für jeden Term, dass er höchstens $Const \cdot |t|^{-p'}$ mit $p' = p + \min\{1, \vartheta\}$ ist. Wir erhalten für den ersten Term

$$\begin{aligned} T_1 &= C_p^{(\vartheta)} \vartheta |t|^{-p} \int_0^{(C_p^{(1)})^{\frac{1}{p}} t^{-1}} \underbrace{(1-x)^\vartheta}_{\leq 1} \underbrace{(1-x)^{-(p+1)}}_{\leq \left(1-(C_p^{(1)})^{\frac{1}{p}} t^{-1}\right)^{-(p+1)}} dx \\ &\leq C_p^{(\vartheta)} \vartheta \left(1 - (C_p^{(1)})^{\frac{1}{p}} t^{-1}\right)^{-(p+1)} |t|^{-p} (C_p^{(1)})^{\frac{1}{p}} t^{-1}. \end{aligned}$$

Da wir uns in dem Fall $t \geq 2(\max\{C_p^{(1)}, C_p^{(\vartheta)}\})^{\frac{1}{p}}$ befinden, ist $\left(1 - (C_p^{(1)})^{\frac{1}{p}} t^{-1}\right)^{-(p+1)} \leq 2^{p+1}$ und wir erhalten mit der Konstante $\gamma_1 := 2^{p+1} C_p^{(\vartheta)} (C_p^{(1)})^{\frac{1}{p}} \vartheta$, dass gilt

$$T_1 \leq \gamma_1 |t|^{-(p+1)}.$$

Analog erhalten wir auch für den dritten Term

$$\begin{aligned}
T_3 &= C_p^{(1)} |t|^{-p} \int_{1-(C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1}}^1 \underbrace{x^{-p}}_{\leq \left(1-(C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1}\right)^{-p}} \vartheta(1-x)^{\vartheta-1} dx \\
&\leq C_p^{(1)} \left(1-(C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1}\right)^{-p} |t|^{-p} \int_{1-(C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1}}^1 \vartheta(1-x)^{\vartheta-1} dx \\
&= C_p^{(1)} \left(1-(C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1}\right)^{-p} |t|^{-p} \left((C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1}\right)^{\vartheta}.
\end{aligned}$$

Wieder erhalten wir wegen $t \geq 2(\max\{C_p^{(1)}, C_p^{(\vartheta)}\})^{\frac{1}{p}}$, dass $\left(1-(C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1}\right)^{-p} \leq 2^p$ ist. Also gilt mit $\gamma_3 = 2^p C_p^{(1)} (C_p^{(\vartheta)})^{\frac{\vartheta}{p}}$

$$T_3 \leq \gamma_3 |t|^{-(p+\vartheta)}.$$

Für den verbleibenden Term gilt

$$\begin{aligned}
T_2 &= C_p^{(1)} C_p^{(\vartheta)} \vartheta |t|^{-2p} \int_{(C_p^{(1)})^{\frac{1}{p}} t^{-1}}^{1-(C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1}} x^{-p} (1-x)^{-(p+1-\vartheta)} dx \\
&= C_p^{(1)} C_p^{(\vartheta)} \vartheta |t|^{-2p} \left(\int_{(C_p^{(1)})^{\frac{1}{p}} t^{-1}}^{\frac{1}{2}} x^{-p} (1-x)^{-(p+1-\vartheta)} dx + \int_{\frac{1}{2}}^{1-(C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1}} x^{-p} (1-x)^{-(p+1-\vartheta)} dx \right).
\end{aligned}$$

Schätzt man im ersten Integral $(1-x)^{-(p+1-\vartheta)} \leq (1-x)^{-(p+1)} \leq 2^{p+1}$ und im zweiten Integral $x^{-p} \leq 2^p \leq 2^{p+1}$ ab, erhält man

$$T_2 \leq 2^{p+1} C_p^{(1)} C_p^{(\vartheta)} \vartheta |t|^{-2p} \left(\int_{(C_p^{(1)})^{\frac{1}{p}} t^{-1}}^{\frac{1}{2}} x^{-p} dx + \int_{\frac{1}{2}}^{1-(C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1}} (1-x)^{-(p+1-\vartheta)} dx \right).$$

Nun ist nach Voraussetzung $p > 1$, somit erhalten wir für das erste Integral

$$\int_{(C_p^{(1)})^{\frac{1}{p}} t^{-1}}^{\frac{1}{2}} x^{-p} = \left[-\frac{1}{p-1} x^{-(p-1)} \right]_{x=(C_p^{(1)})^{\frac{1}{p}} t^{-1}}^{\frac{1}{2}} \leq \frac{1}{p-1} \left((C_p^{(1)})^{\frac{1}{p}} t^{-1} \right)^{-(p-1)}.$$

Für das zweite Integral gilt im Falle $\vartheta \leq 1$, dass $p+1-\vartheta > 1$ ist und damit

$$\int_{\frac{1}{2}}^{1-(C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1}} (1-x)^{-(p+1-\vartheta)} dx = \left[\frac{1}{p-\vartheta} (1-x)^{-(p-\vartheta)} \right]_{x=\frac{1}{2}}^{1-(C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1}} \leq \frac{1}{p-\vartheta} \left((C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1} \right)^{-(p-\vartheta)}.$$

Im Falle $\vartheta > 1$ gilt

$$\int_{\frac{1}{2}}^{1-(C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1}} (1-x)^{-(p+1-\vartheta)} dx \leq \int_{\frac{1}{2}}^{1-(C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1}} (1-x)^{-p} dx \leq \frac{1}{p-1} \left((C_p^{(\vartheta)})^{\frac{1}{p}} t^{-1} \right)^{-(p-1)}.$$

Wir erhalten also für den zweiten Term

$$T_2 \leq \gamma_2 \left(|t|^{-(p+1)} + |t|^{-(p+\min\{1, \vartheta\})} \right)$$

mit $\gamma_2 = 2^{p+1} C_p^{(1)} C_p^{(\vartheta)} \vartheta \frac{1}{p-1} \max \left\{ (C_p^{(1)})^{-\frac{p-1}{p}}, (C_p^{(\vartheta)})^{-\frac{p-\min\{1,\vartheta\}}{p}} \right\}$.

Insgesamt erhalten wir somit für alle $t \geq 2(\max\{C_p^{(1)}, C_p^{(\vartheta)}\})^{\frac{1}{p}}$, dass gilt

$$|\phi_\vartheta(t)| \leq \gamma_1 |t|^{-(p+1)} + \gamma_2 \left(|t|^{-(p+1)} + |t|^{-(p+\min\{1,\vartheta\})} \right) + \gamma_3 |t|^{-(p+\vartheta)}$$

und damit im Fall $|t| > 1$

$$|\phi_\vartheta(t)| \leq (\gamma_1 + 2\gamma_2 + \gamma_3) |t|^{-(p+\min\{1,\vartheta\})}.$$

Da im Fall $|t| \leq 1$ gilt, dass $|\phi_\vartheta(t)| \leq 1 \leq |t|^{-(p+\min\{1,\vartheta\})}$ ist, erhalten wir mit

$\gamma := \max\{1, \gamma_1 + 2\gamma_2 + \gamma_3\}$, dass für alle $t \geq 2(\max\{C_p^{(1)}, C_p^{(\vartheta)}\})^{\frac{1}{p}}$ gilt

$$|\phi_\vartheta(t)| \leq \gamma |t|^{-(p+\min\{1,\vartheta\})}.$$

Nun gilt für $0 < t < 2(\max\{C_p^{(1)}, C_p^{(\vartheta)}\})^{\frac{1}{p}}$ trivialerweise mit $\gamma' := \left(2(\max\{C_p^{(1)}, C_p^{(\vartheta)}\})^{\frac{1}{p}} \right)^{p+\min\{1,\vartheta\}}$, dass $|\phi_\vartheta(t)| \leq 1 \leq \gamma' |t|^{-(p+\min\{1,\vartheta\})}$ ist.

Wir erhalten also mit $C_{p+\min\{1,\vartheta\}}^{(\vartheta)} = \max\{\gamma, \gamma'\}$, dass für alle $t > 0$ gilt

$$|\phi_\vartheta(t)| \leq C_{p+\min\{1,\vartheta\}}^{(\vartheta)} |t|^{-(p+\min\{1,\vartheta\})}.$$

Die Schranke gilt ebenfalls für $t < 0$, da $\phi_\vartheta(-t) = \overline{\phi_\vartheta(t)}$ ist und damit $|\phi_\vartheta(t)| = |\phi_\vartheta(-t)|$. □

Beweis von Theorem 5.10. Sei $\vartheta > 0$ beliebig. Zu zeigen ist, dass es für jedes $p > 0$ eine Konstante $C_p^{(\vartheta)}$ gibt mit $|\phi_\vartheta(t)| \leq C_p^{(\vartheta)} |t|^{-p}$ für alle $t \in \mathbb{R}$. Für $0 < p \leq \frac{1}{2}$ erhalten wir dies aus Lemma 5.9. Wendet man nun mehrfach Lemma 5.11 an, erhält man das Ergebnis auch für alle $p < 1 + \min\{1, \frac{\vartheta}{2}\}$ und durch iteratives Anwenden von Lemma 5.12 erhält man das Ergebnis auch für $p > 1$. □

Theorem 5.13. Sei $\vartheta > 0$ beliebig und $Z_\infty^{(\vartheta)}$ der Grenzwert der reskalierten internen Pfadlänge $\left(\frac{I_n^{(\vartheta)} - E[I_n^{(\vartheta)}]}{\vartheta + n - 1} \right)_{n \geq 1}$, welche nach Korollar 5.6 fast sicher konvergiert. Dann gibt es für die Verteilung $\mathcal{L}(Z_\infty^{(\vartheta)})$ eine unendlich oft differenzierbare, schnell abfallende Dichte f_ϑ .

Beweis. Die Existenz der Dichte sowie die Differenzierbarkeit folgt unmittelbar aus Theorem 5.10, wie es auch von Fill und Janson in [6] geschildert wird. Man erhält die Dichte aus der charakteristischen Funktion ϕ_ϑ über die Fourier Inversionsformel

$$f_\vartheta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi_\vartheta(t) dt,$$

wobei $e^{-itx} \phi_\vartheta(t)$ wegen Theorem 5.10 integrierbar ist. Ebenso ist für jedes $k \geq 0$ auch $t^k \phi_\vartheta(t)$ wegen Theorem 5.10 integrierbar und wir erhalten alle Ableitungen aus

$$f_\vartheta^{(k)}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (-it)^k e^{itx} \phi_\vartheta(t) dt.$$

□

Literatur

- [1] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964.
- [2] L. Addario-Berry and K. Ford. Poisson-dirichlet branching random walks. 2010.
- [3] A. D. Barbour, L. Holst, and S. Janson. *Poisson approximation*, volume 2 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1992. Oxford Science Publications.
- [4] R. P. Dobrow. On the distribution of distances in recursive trees. *J. Appl. Probab.*, 33(3):749–757, 1996.
- [5] P. Donnelly and S. Tavaré. The ages of alleles and a coalescent. *Adv. in Appl. Probab.*, 18(1):1–19, 1986.
- [6] J. A. Fill and S. Janson. Smoothness and decay properties of the limiting Quicksort density function. In *Mathematics and computer science (Versailles, 2000)*, Trends Math., pages 53–64. Birkhäuser, Basel, 2000.
- [7] P. Hall and C. C. Heyde. *Martingale limit theory and its application*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1980. Probability and Mathematical Statistics.
- [8] H. M. Mahmoud. Limiting distributions for path lengths in recursive trees. *Probab. Engrg. Inform. Sci.*, 5(1):53–59, 1991.
- [9] C. J. H. McDiarmid and R. B. Hayward. Large deviations for Quicksort. *J. Algorithms*, 21(3):476–507, 1996.
- [10] R. Neininger. Vorlesungsskript: Stochastische Konzentrationsungleichungen. 2011.
- [11] M. Régnier. A limiting distribution for quicksort. *RAIRO Inform. Théor. Appl.*, 23(3):335–343, 1989.
- [12] R. T. Smythe and H. M. Mahmoud. A survey of recursive trees. *Teor. Ĭmovĭr. Mat. Stat.*, (51):1–29, 1994.

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbstständig und nur unter Benutzung der angegebenen Literatur und Hilfsmittel angefertigt habe. Wörtlich übernommene Sätze und Satzteile sind als Zitate belegt, andere Anlehnungen hinsichtlich Aussage und Umfang unter den Quellenangaben kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen und ist nicht veröffentlicht.

Kevin Leckey