Diplomarbeit

Ein universeller Zentraler Grenzwertsatz für den Abstand zweier Kugeln in zufälligen Splitbäumen

Jelena Ryvkina

Johann Wolfgang Goethe-Universität
Fachbereich Informatik und Mathematik
Institut für Mathematik
Frankfurt am Main

Betreuer: Prof. Dr. Ralph Neininger

Mai 2008

Zusammenfassung

In der vorliegenden Arbeit wird ein Modell des zufälligen Splitbaumes untersucht. Dies ist ein verallgemeinertes Modell, das bei passender Wahl der zugehörigen Parameter viele konkrete Suchbäume umfasst.

Das Modell ist in der Arbeit von L. Devroye beschrieben: Nach einem zufallsbasierten Algorithmus werden den Knoten des Baumes Daten in Form von Kugeln hinzugefügt. Tiefe und Höhe sind dabei grundlegende Größen, die die Komplexität von Suchoperationen beschreiben, wenn das Suchbaummodell als Datenstruktur verwendet wird.

Das Augenmerk der Arbeit richtet sich auf eine weitere entscheidende Größe: Den Abstand $\Delta_{I,J}$ zweier rein zufällig gewählter Kugeln im Baum. Aufbauend auf Devroyes Erkenntnissen zum asymptotischen Verhalten der Tiefe der zuletzt eingefügten Kugel im Splitbaum, wird ein neues Resultat erzielt: Ein universeller Zentraler Grenzwertsatz für $\Delta_{I,J}$.

Als Anwendungsbeispiel werden zwei vom allgemeinen Modell abgedeckte Suchbäume betrachtet und der jeweilige Grenzwertsatz für die Abstände aus dem universellen Satz abgeleitet.

Inhaltsverzeichnis

1 Einleitung

Motivation und Vorgehensweise. In der vorliegenden Arbeit wird ein Modell eines zufälligen Splitbaumes analysiert.

Es gibt zahlreiche Untersuchungen zu verschiedenen Baumstrukturen, in welchen insbesondere das Verhalten von Tiefe, Höhe und Knotenabständen untersucht wird. Zum Beispiel behandelt H. M. Mahmoud in seinem Buch Evolution of Random Search Trees [10] viele Suchbaummodelle und weist Sätze zu ihren Eigenschaften nach. Tiefe und Höhe sind dabei grundlegende Größen, die die Komplexität von Suchoperationen beschreiben, wenn ein Suchbaum als Datenstruktur verwendet wird.

Bei der Analyse der verschiedenen Baummodelle wurden oft unterschiedliche Zugänge verwendet, um Resultate in Form von Grenzwertsätzen zu erlangen. Es erweist sich deshalb als sinnvoll, zu generalisierten Baummodellen überzugehen, welche Aussagen zu ganzen Familien von Bäumen universell ermöglichen.

In der Arbeit von L. Devroye [4] wird ein allgemeines Modell eines Splitbaumes eingeführt. Dies ist ein Modell, bei dem den Knoten des Baumes Daten, in Devroyes Darstellung durch Kugeln symbolisiert, zugeordnet werden. Bei passender Wahl der zum Modell gehörigen Parameter enthält der Splitbaum viele der einzeln untersuchten, in der Höhe logarithmisch wachsenden Baumstrukturen der Informatik. Sätze zu diesem allgemeinen Modell liefern somit Resultate für eine Vielzahl von Anwendungen und machen diese Auseinandersetzung im Einzelnen nicht mehr notwendig.

Devroye gelingt es in seiner Arbeit, universelle Ergebnisse für Knotentiefe und Höhe des Splitbaumes zu finden, insbesondere einen Grenzwertsatz für die Tiefe der zuletzt eingefügten Kugel im Baum. Eine Anwendungsmöglichkeit findet sich dabei beispielsweise bei der Erstellung von Datenbanken. Hier ist man stets auf der Suche nach Möglichkeiten, große, zufällige Datenmengen so anzuordnen, dass der spätere Zugriff auf diese Daten in möglichst kurzer Laufzeit des Computers erfolgen kann. Mit Resultaten zum allgemeinen Splitbaummodell lässt sich für viele konkrete Algorithmen vorhersagen, wie sich Laufzeiten bei wachsendem Datenumfang asymptotisch verhalten werden. Die Tiefen und Abstände zwischen den Knoten im Baum stellen dabei die Referenzwerte für die Laufzeiten dar.

In der vorliegenden Arbeit untersuchen wir eine weitere grundlegende Größe des zufälligen Splitbaumes, den Abstand $\Delta_{I,J}$ zwischen zwei zufällig gewählten Elementen im Baum. Diese Größe beschreibt die Komplexität der so genannten finger search Operation im Baum, einer typischen Suchoperation in der Informatik, die zum Beispiel bei Datenbanken durchgeführt wird. Abstände zufälliger Knoten spielen außerdem eine große Rolle bei der Untersuchung und Modellierung

allgemeiner Netzwerke, wie etwa bei den small worlds- oder den preferential attachment Modellen.

Unser Hauptresultat ist ein universeller Zentraler Grenzwertsatz für $\Delta_{I,J}$:

$$\frac{\Delta_{I,J} - \frac{2}{\mu}\log n}{\mu^{-3/2}\sigma\sqrt{2\log n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ für } n \to \infty,$$

wobei μ und σ^2 Parameter des Modells sind. Dieses Ergebnis besagt, dass sich der Abstand zweier rein zufälliger Elemente, logarithmisch normiert, für großen Datenumfang, annähernd wie eine standardnormalverteilte Variable verhält.

Methodisch werden wir, zum Beweis des obigen Resultats, den Zugang und die Ideen Devroyes zur Analyse der Knotentiefe erweitern. Um Devroyes Erkenntnisse auf den Abstand der zufälligen Knoten im Baum anwenden zu können, werden wir ausnutzen, dass sich zwei Kugeln im Modell, sobald sich ihre Wege im Baum trennen, bedingt auf die Kardinalitäten der entsprechenden Teilbäume, unabhängig weiterbewegen. Dies ermöglicht uns, den Abstand der Knoten als Summe unabhängiger Tiefen zu beschreiben.

In den Kapiteln 2, 3 und 4 werden wir den Begriff des Suchbaumes und anschließend das von uns gewählte Modell des zufälligen Splitbaumes beschreiben. Die Variablen, die zur Beschreibung der Tiefe einer Kugel im Baum notwendig sind, werden erläutert.

In Kapitel 5 werden wir einen Grenzwertsatz für den Abstand zweier rein zufälliger Kugeln im Splitbaum nachweisen.

Im 6. Kapitel sollen zwei konkrete Baumstrukturen (binärer Suchbaum, b-närer Suchbaum), die vom generalisierten Suchbaum abgedeckt sind, beispielhaft erläutert werden. Grenzwertsätze für die beiden Modelle werden, anhand der universellen Aussage für den zufälligen Splitbaum, hergeleitet.

Zum Abschluss diskutieren wir in Kapitel 7 unsere Ergebnisse, indem wir die Möglichkeit eines alternativen Zugangs zum Modell erörtern.

Notation. Die Verteilung einer Zufallsvariable X sei hier mit P_X bezeichnet. $\overset{\mathcal{L}}{=}$ bedeute im Folgenden die Verteilungsgleichheit und $\xrightarrow{\mathcal{L}}$ die Verteilungskonvergenz.

Wenn wir das Landau-Symbol $o(\cdot)$ benutzen ist mit $o(1)$ stets eine deterministische Nullfolge gemeint.

Ist A ein Ereignis, so wird $\mathbb{1}_A$ die Indikatorvariable zum Ereignis A bezeichnen. D. h.: $P(\mathbb{1}_A = 1) = P(A),\ P(\mathbb{1}_A = 0) = 1 - P(A)$.

$\mathcal{N}(0,1)$ soll die Standardnormalverteilung bezeichnen, d.h. für eine messbare Menge $A \subseteq \mathbb{R}$ ist $P(N \in A) = (1/\sqrt{2\pi}) \int \mathbf{1}_A(x) \exp(x^2/2) d\lambda(x)$, falls

$N \stackrel{\mathcal{L}}{=} \mathcal{N}(0,1)$ und λ das Lebesgue-Maß auf $\mathbb{R}$ bezeichnet. $\Phi(a) := P(N \leq a)$ soll das Gaußsche Fehlerintegral in a symbolisieren.

$B(n,p)$ bezeichne eine binomialverteilte Zufallsvariable mit Erfolgsparameter $p \in [0,1]$. Das bedeutet, dass die Zufallsvariable die Verteilung $\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \delta_k$ hat. Dabei bezeichne δ_k das Dirac-Maß in k.

Die Beta-Verteilung mit den Parametern $a > 0$ und $m > 0$ wird bei uns mit $Beta(a,m)$ abgekürzt. Ihre Dichte bezüglich des Lebesgue-Maßes ist $\mathbb{1}_{[0,1]}(x) \frac{x^{a-1}(1-x)^{m-1}}{c(a,m)}$, wobei die Normierungskonstante $c(a,m)$ aus einem Quotienten von Gamma-Funktionen besteht: $c = \Gamma(a)\Gamma(m)/\Gamma(a+m)$.

Eine auf dem Intervall $[0,1]$ uniform verteilte Zufallsvariable U, kurz $U \stackrel{\mathcal{L}}{=} unif[0,1]$, hat die λ-Dichte $\mathbb{1}_{[0,1]}(x)$. Ist eine Zufallsvariable uniform auf $\{1,\ldots,n\}$ verteilt, so hat sie die Verteilung $\sum_{k=1}^{n} (1/n)\delta_k$.

Eine geometrisch verteilte Zufallsvariable zum Parameter $q \in (0,1)$ hat die Verteilung $\sum_{k=0}^{\infty} q^k (1-q)\delta_k$.

2 Graphen und Suchbäume

Die folgenden Darstellungen lehnen sich an das Buch „Evolution of random search trees“ von H. M. Mahmoud [10] an.

Graphen. Unter einem *einfachen gerichteten Graphen* G versteht man ein Paar $G = (E, K)$, wobei E die Menge $\{x_1, x_2, \ldots\}$ der *Ecken* oder auch Knoten dieses Graphen bezeichnet, und $K = \{(x_i, x_j) \in E'^2 : E'^2 \subseteq E^2,\ x_i \neq x_j\}$ die Menge aller *Kanten* in dem Graphen darstellt. Die Kanten sind geordnete Paare der Knoten und die Menge K kann im Gegensatz zur Menge E leer sein. Für $k \in K,\ k = (x, y)$ sagen wir, dass die Kante k von Ecke x zur Ecke y verläuft, bzw. in x anfängt und in y endet. Wir sprechen von einem *Pfad* von Knoten x zum Knoten y, wenn für ein $n \in \mathbb{N}$ Ecken $x_2, x_3, \ldots, x_n$ in E und Kanten $(x_1, x_2), \ldots, (x_n, x_{n+1})$ in K existieren, wobei $x_1 = x$ und $x_{n+1} = y$ gilt. Die Länge dieses Pfades ist die Anzahl der Kanten $(x_1, x_2), \ldots, (x_n, x_{n+1})$, sie ist also n. Wir schreiben abkürzend: $x_1 \ldots x_{n+1}$ ist ein Pfad der Länge n zwischen x_1 und x_{n+1}. Existiert zusätzlich die Kante (x_{n+1}, x_1), so nennen wir den Pfad $x_1 \ldots x_{n+1} x_1$ einen *Kreis*. Den *Ausgangsgrad* $d_{out}(x)$ einer Ecke $x \in E$ des Graphen G definieren wir als die Anzahl der in x anfangender Kanten. Analog ist der *Eingangsgrad* $d_{in}(x)$ von x die Anzahl der in x endenden Kanten. Es ist klar, dass

$$\sum_{x \in E} d_{in}(x) = \sum_{x \in E} d_{out}(x) = |E|.$$

Wir nennen einen Graphen *zusammenhängend*, wenn für beliebige zwei Knoten $x, y \in E$ ein Pfad zwischen diesen beiden Knoten existiert.

Bäume. Unter einem *einfachen gerichteten Baum* versteht man einen zusammenhängenden, kreisfreien gerichteten Graphen, bei dem, mit der Ausnahme genau eines Knotens, der *Wurzel*, der innere Grad jedes Knotens 1 ist. Der innere Grad der Wurzel ist 0.

Um uns in einem solchen Baum zu orientieren, wollen wir die Knoten so anordnen, dass wir ihre Position gut beschreiben können. Wir ordnen den Baum nach Stufen. In der 0-ten Stufe platzieren wir die Wurzel. In der ersten Stufe platzieren wir alle Knoten des Graphen, zu denen es von der Wurzel ausgehend eine Kante gibt. In der nächsten Stufe ordnen wir alle Knoten an, zu denen von den Knoten in Stufe 1 ausgehende Kanten existieren. Dabei sollen die Knoten so gruppiert werden, dass jeweils alle Knoten, die eine Kante zum gleichen Knoten in der nächsthöheren Stufe besitzen, nebeneinander positioniert sind. Dieses Verfahren wird fortgesetzt, bis alle Knoten stufenweise angeordnet sind. Es entsteht eine Struktur wie in Abbildung 1. (s. S. 8)

Wenn wir nun auf die Kantenrichtungen verzichten, d.h. die Paare (x_1, x_2) mit den Paaren (x_2, x_1) identifizieren, erhalten wir den eingebetteten *ungerichteten Baum*.

Die Eigenschaften Zusammenhang und Kreisfreiheit bleiben natürlich erhalten. Wenn wir im Weiteren von einem Baum oder auch Suchbaum reden, ist stets ein solcher Baum gemeint.

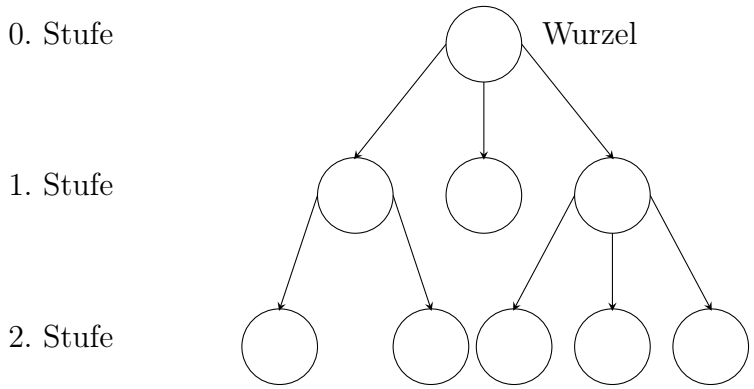

Abbildung 1.
Geordneter, gerichteter, einfacher Baum

Eine wichtige Eigenschaft von Bäumen ist in dem folgenden Satz beschrieben:

Satz 2.0.1. *In einem ungerichteten Baum gibt es zu je beliebigen zwei Knoten genau einen Pfad, der diese beiden Knoten verbindet.*

Beweis: Durch Widerspruch: Da ein Baum per Definition verbunden ist, gibt es für jedes Knotenpaar mindestens einen verbindenden Pfad. Sei nun G ein Baum und x und y zwei Knoten dieses Baumes, die durch mehr als einen Pfad verbunden sind. Seien P_1 und P_2 zwei dieser verbindenden Pfade und unterschiedlich. Da P_1 und P_2 unterschiedlich sind, muss es ausgehend von x einen ersten Knoten geben, nach dem sich die Pfade trennen. Nennen wir diesen v. $v = x$ ist dabei möglich. Da beide Pfade in y münden, muss es auch einen Knoten geben, bei dem sich die beiden Pfade wieder vereinen. Nennen wir diesen Knoten w. $w = y$ ist ebenfalls nicht ausgeschlossen. Also haben P_1 und P_2 die folgenden Formen:

$$P_1 = xx_2 \dots v \underbrace{\dots}_{Q_1} wy_1 \dots y$$

und

$$P_2 = xx_2 \dots v \underbrace{\dots}_{Q_2} wy_1 \dots y.$$

Dabei unterscheiden sich die zwei Pfadstücke zwischen v und w zumindest durch einen Knoten. Wir bezeichnen sie mit Q_1 und Q_2. Ist $Q_2 = q_1 \dots q_i$, dann sei $\overline{Q_2} := q_i \dots q_1$. $\overline{Q_2}$ ist ein Pfad, da in einem ungerichteten Baum (q_j, q_{j+1}) mit (q_{j+1}, q_j) identifiziert ist.

Damit ist der Pfad $vQ_1w\overline{Q_2}v$ ein Kreis und das steht im Widerspruch zur Kreisfreiheit eines Baumes.

□

Da wir nun wissen, dass es in einem Baum zwischen zwei Knoten immer einen eindeutigen verbindenden Pfad gibt, können wir den *Abstand* der Knoten definieren: Der Abstand zweier Knoten im Baum soll die Länge des verbindenden Pfades dieser Knoten sein. Damit hat ein Knoten zu sich selbst den Abstand 0. Ist in einem Baum die Anzahl b der Kanten die einen Knoten mit anderen verbinden für alle Knoten gleich, sprechen wir von einem *Verzweigungsgrad* b.

Bemerkung 1. Ein Baum ist nützlich, um Informationen anzuordnen. In der Anwendung kann es sich zum Beispiel um die Strukturierung einer Datenbank handeln.

Wir werden Information in Form von Kugeln im Baum platzieren. Dabei wird jedem Knoten des Baumes algorithmisch eine bestimmte Anzahl von Kugeln zugeordnet. Der Abstand zweier Kugeln im Baum ist dabei mit dem Abstand der entsprechenden Knoten identifiziert.

3 Das Modell

Um den zufälligen Splitbaum zu beschreiben, greifen wir auf die Darstellung von L. Devroye [4] zurück.

Modell. Beim zufälligen *Splitbaum* gehen wir von einem Suchbaum mit einem Verzweigungsgrad $b > 1$ und unendlich vielen Stufen aus. Auf diesen verteilen wir n Kugeln, die einzeln eingefügt werden und nach und nach von der Wurzel ausgehend die Stufen, an den Kanten enlang, im Baum hinabrutschen. Die Wahl der jeweiligen Kante soll zufällig erfolgen. Nach dem Verteilen der Kugeln werden alle „nutzlosen“ Knoten abgeschnitten.

Folgende Parameter zeichnen das Modell aus:

- s maximale Anzahl Kugeln, die ein Knoten enthalten kann, bis er „aufbricht“ (s. u.),
- s_0 Anzahl Kugeln, die in einem Knoten zurückbleiben, wenn er aufgebrochen ist,
- s_1 Anzahl Kugeln, die in jeden Knoten eingefügt wird, der direkt an einem aufbrechenden Knoten hängt.

Um eine sinnvolle Baumstruktur zu erzeugen, müssen die Parameter den folgenden Ungleichungen genügen:

$$0 < s,\ 0 \leq s_0 \leq s,\ 0 \leq bs_1 \leq s + 1 - s_0.$$

Des Weiteren brauchen wir einen Zufallsvektor $\mathcal{V} = (V_1, \ldots, V_b)$, für den fast sicher $V_i > 0$ für $i = 1, \ldots, b$ und $\sum_{i=1}^{b} V_i = 1$ gelten. Wir wollen die b Knoten, die direkt an einem Knoten u hängen, als *Kinder* des Knotens u bezeichnen. Wir identifizieren sie im Folgenden mit ihrer Nummerierung $i = 1, \ldots, b$. Ein Knoten u soll *Blatt* heißen, wenn er der einzige Knoten, in dem in ihm verwurzelten Teilbaum ist, der Kugeln enthält. Schließlich ordnen wir jedem Knoten u des Baumes eine unabhängige Kopie $\mathcal{V}_u = (V_{1,u}, \ldots, V_{b,u})$ des Vektors $\mathcal{V}$ zu. Die Komponenten $V_{i,u}$ des Vektors $\mathcal{V}_u$ stellen die selbst zufälligen Wahrscheinlichkeiten dar, die jeweils der Kante zwischen u und dem i-ten Kind von u zugeordnet sind. Gemäß dieser Wahrscheinlichkeiten wird sich eine Kugel in unserem Modell den Weg durch den Baum bahnen.

Nun können wir n Kugeln auf den Baum verteilen.

Hinzufügen einer Kugel zum Teilbaum am Knoten u:

a) Ist u kein Blatt, dann wird gemäß der Wahrscheinlichkeiten $\mathcal{V}_u$ eines der Kinder von u ausgesucht, d. h. die Wahl fällt auf Kind i mit Wahrscheinlichkeit $V_{i,u}$. Man wendet den Algorithmus iterativ auf diesen neuen Knoten an.

b) Ist u ein Blatt, das weniger als s Kugeln enthält, so wird die Kugel u hinzugefügt, und man geht zur nächsten Kugel über.

c) Ist u ein Blatt, das bereits genau s Kugeln enthält, so bricht der Knoten auf. Das heißt: s_0 der $s+1$ Kugeln bleiben im Knoten u. Je s_1 Kugel wird jedem Kind von u hinzugefügt. Die verbleibenden $s+1-s_0-bs_1$ Kugeln werden nun, jede einzeln und unabhängig, gemäß der Wahrscheinlichkeiten $V_{i,u}$, an die Kinder von u verteilt. Demnach geht eine Kugel zum i-ten Kind von u mit Wahrscheinlichkeit $V_{i,u}$. Wird die Kapazität s von einem der Knoten dabei überschritten, bricht er ebenfalls auf und man wendet c) auf diesen Knoten an. (Dies kann allerdings nur auftreten, wenn $s_0 = 0$.) Sind auf diese Weise alle $s+1$ Kugeln wieder einem Knoten zugeordnet, geht man zur nächsten Kugel über.

Wir arbeiten folglich gemäß a) für jede Kugel ausgehend von der Wurzel einen zufälligen Pfad im Baum ab, bis wir auf ein Blatt stoßen. Dann wenden wir b) oder c) auf diesen Knoten an. Der Vorgang wird wiederholt bis alle n Kugeln verteilt sind. Ist das geschehen, kappen wir alle Knoten, die sich unterhalb eines Blattes befinden. Alle Blätter enthalten dann bis zu s Kugeln und alle Knoten, die keine Blätter sind, enthalten s_0 Kugeln.

Ein Zentraler Begriff wird bei uns die *Tiefe* einer Kugel sein. Die Tiefe einer Kugel ist die Stufe im Baum, in welcher sich diese Kugel befindet, d. h. es ist der Abstand zur Wurzel. Wir werden die Tiefe der n-ten Kugel mit D_n bezeichnen.

Beispiele.

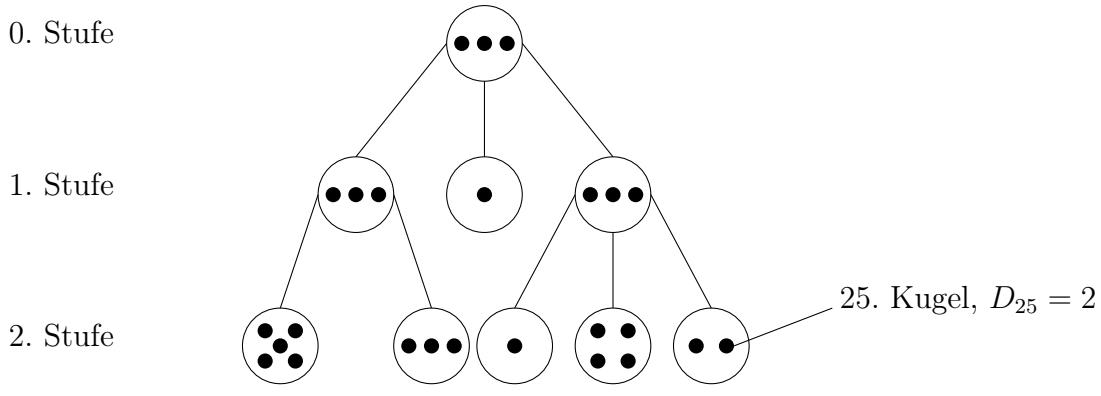

Abbildung 2: Beispiel 1.

$b = 3,\ n = 25,\ s = 5,\ s_0 = 3,\ s_1 = 0$

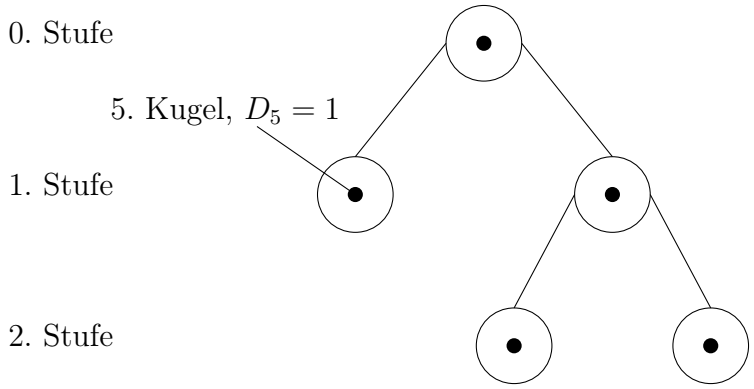

Abbildung 3: Beispiel 2.

$b = 2,\ n = 5,\ s = 1,\ s_0 = 1,\ s_1 = 0$

Bemerkung 2. Wählt man bei Beispiel 2. $\mathcal{V} = (U, 1-U)$, wobei U uniform auf $[0,1]$ verteilt ist, so erhält man einen binären Suchbaum. (s. Abschnitt 6)

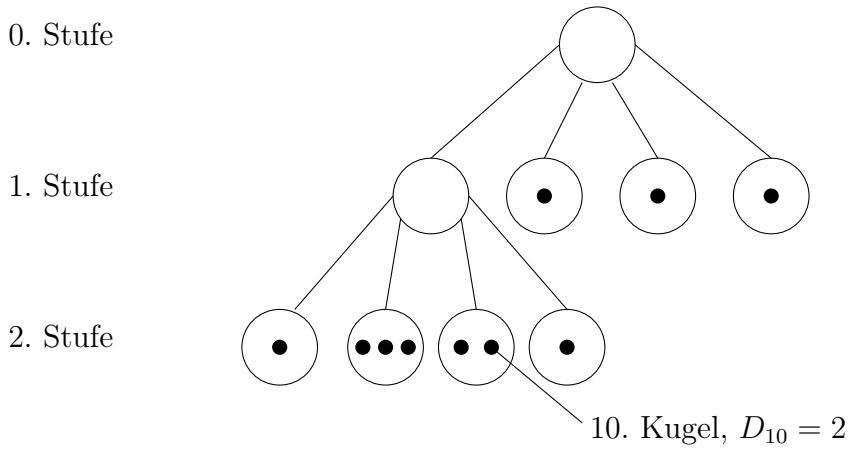

Abbildung 4: Beispiel 3.

$b = 4,\ n = 10,\ s = 3,\ s_0 = 0,\ s_1 = 1$

4 Die Variablen

Die in Abschnitt 5 folgenden Sätze werden in der Anwendung im hohen Maße von der Verteilung von $\mathcal{V}$ abhängen. Um unsere Argumentation zu vereinfachen, wollen wir uns klar machen, dass wir o.B.d.A. davon ausgehen können, dass die Variablen $V_1, \ldots, V_b$ identisch verteilt sind: Wir können eine zufällige, uniform verteilte Permutation $\tau \in S_b$ wählen und jedem Knoten im Baum statt $\mathcal{V}$ eine unabhängige Kopie des Zufallsvektors $\mathcal{V}_\tau = (V_{\tau(1)}, \ldots, V_{\tau(b)})$ zuordnen. Die Abstand-Struktur des Baumes wird sich dabei nicht verändern, da der Übergang zu V_τ lediglich einer Neuanordnung der Baumäste entspricht. Außerdem gilt $V_{\tau(1)} \stackrel{\mathcal{L}}{=} \ldots \stackrel{\mathcal{L}}{=} V_{\tau(b)}$: Sei $i \in \{1, \ldots, b\}$ beliebig, und $A \subseteq [0,1]$ messbar, dann gilt:

$$P(V_{\tau(i)} \in A) = \sum_{j=1}^{b} P(V_{\tau(i)} \in A | \tau(i) = j) P(\tau(i) = j) = \sum_{j=1}^{b} P(V_j \in A) \frac{1}{b}.$$

Dieser Ausdruck hängt nicht von i ab und führt uns auf die folgende Definition:

Definition 4.0.1. *Seien $X_1, \ldots, X_n$ reelle Zufallsvariablen mit den Verteilungen $P_{X_1}, \ldots, P_{X_n}$, dann heißt die Zufallsvariable X mit der Verteilung $(1/n) \sum_{i=1}^{n} P_{X_i}$ Splitter der Variablen $X_1, \ldots, X_n$. Die Verteilung von X nennt man Splitter-Verteilung.*

Mit dieser Definition können wir sagen, dass $V_{\tau(i)}$ die Splitter-Verteilung $(1/b) \sum_{i=1}^{b} P_{V_i}$ der $V_1, \ldots, V_b$ besitzt.

Um später Beweise führen zu können, müssen wir die zufälligen Wahrscheinlichkeiten, mit denen eine bestimmte Kugel die Stufen im Baum passiert, bis sie zu ihrem Platz kommt, beschreiben. Dafür führen wir eine weitere Zufallsvariable $V_S = W \in [0,1]$ ein. Dabei ist S eine Zufallsvariable in $\{1, \ldots, b\}$ mit $P(S = i | \mathcal{V} = (v_1, \ldots, v_b)) = v_i$ für $i = 1, \ldots, b$.

Wir zeigen zunächst, dass W eine größenverzerrte Version des Splitters der Variablen $V_1, \ldots, V_b$ ist:

Definition 4.0.2. *(Größenverzerrte Version)*
Sei $X : \mathbb{R} \to \mathbb{R}^+$ eine nichtnegative Zufallsvariable mit dem Erwartungswert $E[X] \in (0, \infty)$. X^* heißt größenverzerrte Version von X, wenn für die Verteilung P_{X^*} von X^* folgendes gilt:

$$P_{X^*}(dx) = \frac{x P_X(dx)}{E[X]}, \ x > 0.$$

Diese Gleichheit kann so interpretiert werden, dass X^* größere Werte des Wertebereichs von X bevorzugt. Ist, wie in unserem Fall, der Wertebereich von X das Intervall $[0, 1]$, so gilt:

$$E[X^*] = \int x dP_{X^*}(x) = \int_0^1 \frac{x^2}{E[X]} dP_X(x) = \frac{1}{E[X]}(Var(X) + E[X]^2) \geq E[X].$$

Dass die Zufallsvariable W eine größenverzerrte Version des Splitters von $V_1, \ldots, V_b$ ist, ist anschaulich klar, denn die obige Bedingung impliziert, dass je größer der Wert ist, den eine der V_i annimmt, desto wahrscheinlicher ist es, dass W ebenfalls diesen Wert annimmt. Mit anderen Worten: je größer die Wahrscheinlichkeit, die zu einer Abzweigung gehört, um so größer auch die Chance, dass eine Kugel diese Abzweigung nimmt. Trotzdem wollen wir diese Aussage beweisen und nutzen dazu das folgende, dem Buch von R. Arratia, A. D. Barbour und S. Tavaré [3] entnommene, Korollar:

Korollar 4.0.3. *Ist $X \geq 0$ eine Zufallsvariable mit $E[X] \in (0, \infty)$, dann ist X^* genau dann eine größenverzerrte Version von X, wenn für alle messbaren und beschränkten Funktionen $f : \mathbb{R}^+ \to \mathbb{R}$ gilt:*

$$E[f(X^*)] = \frac{E[Xf(X)]}{E[X]}.$$

Beweis: Sei X^* eine Größenverzerrte von X und die Funktion f wie im Korollar gefordert. Es gilt:

$$E[f(X^*)] = \int f(x) dP_{X^*}(x) = \frac{1}{E[X]} \int x f(x) dP_X(x) = \frac{E[Xf(X)]}{E[X]}$$

nach der Definition von X^*.

Sei nun X^* eine Zufallsvariable, die für alle $f : \mathbb{R}^+ \to \mathbb{R}$, die messbar und beschränkt sind, die Gleichheit $E[f(X^*)] = E[Xf(X)]/E[X]$ erfüllt. Und sei $t > 0$ und $f(x) := \mathbb{1}_{[0,t]}(x)$ für $x \geq 0$. Dann gilt:

$$P(X^* \leq t) = E[\mathbb{1}_{[0,t]}(X^*)] = \frac{E[X\mathbb{1}_{[0,t]}(X)]}{E[X]} = \int_0^t \frac{x}{E[X]} dP_X(x),$$

was der Definition für die Größenverzerrte von X entspricht.

□

Wir wollen im Weiteren den Splitter von $V_1, \ldots, V_b$ mit V bezeichnen und zunächst zeigen, dass für integrierbare Funktionen f die Gleichheit

$$\int f dP_V = (1/b) \sum_{i=1}^{b} \int f dP_{V_i}$$

gilt. Die Gleichheit ist auch für messbare $f \geq 0$ erfüllt.

Algebraische Induktion:
1) Sei A ein Ereignis.

$$\begin{aligned} \int \mathbb{1}_A d(\tfrac{1}{b}\sum_{i=1}^{b} P_{V_i}) &= \tfrac{1}{b}\sum_{i=1}^{b} P_{V_i}(A) \\ \tfrac{1}{b}\sum_{i=1}^{b}\int \mathbb{1}_A dP_{V_i} &= \tfrac{1}{b}\sum_{i=1}^{b} P_{V_i}(A). \end{aligned}$$

2) Seien A_j disjunkte Ereignisse und a_j reelle und positive Konstanten für $1 \leq j \leq n,\ n \in \mathbb{R}$.

$$\begin{aligned} \int \sum_{j=1}^{n} a_j \mathbb{1}_{A_j} d(\tfrac{1}{b}\sum_{i=1}^{b} P_{V_i}) &= \sum_{j=1}^{n} a_j \tfrac{1}{b}\sum_{i=1}^{b} P_{V_i}(A_j) = \tfrac{1}{b}\sum_{i=1}^{b}\sum_{j=1}^{n} a_j P_{V_i}(A_j) \\ \tfrac{1}{b}\sum_{i=1}^{b}\int \sum_{j=1}^{n} a_j \mathbb{1}_{A_j} dP_{V_i} &= \tfrac{1}{b}\sum_{i=1}^{b}\sum_{j=1}^{n} a_j P_{V_i}(A_j). \end{aligned}$$

3) Da man jede nichtnegative messbare Funktion mit Elementarfunktionen wie bei 2) von unten approximieren kann, folgt die Behauptung nach dem Satz von B. Levi für nichtnegative messbare Funktionen.

4) Wegen der Linearität des Integrals und der Zerlegung von Funktionen f in $f^+ := \max\{f, 0\}$ und $f^- := \max\{-f, 0\}$ folgt die Behauptung für beliebige integrierbare Funktionen.

□

Nun wollen wir den Erwartungswert von V berechnen:

$$E[V] = \int v d(\frac{1}{b}\sum_{i=1}^{b} P_{V_i})(v) = \frac{1}{b}\sum_{i=1}^{b}\int v dP_{V_i}(v) = \frac{1}{b}\sum_{i=1}^{b} E[V_i] = \frac{1}{b}E[\underbrace{\sum_{i=1}^{b} V_i}_{=1 \text{ f.s.}}] = \frac{1}{b}.$$

Jetzt können wir zeigen, dass die Bedingung von Korollar 4.0.3 erfüllt ist: Sei $f : \mathbb{R}^+ \to \mathbb{R}$ eine messbare und beschränkte Funktion.

$$\begin{aligned} & E[f(V_S)] \\ = \ & \textstyle\int_{[0,1]^b} E[f(V_S)|\mathcal{V} = (v_1, \ldots, v_b)] dP_{\mathcal{V}}(v_1, \ldots, v_b) \text{ nach dem Faktorisierungslemma,} \\ = \ & \textstyle\int_{[0,1]^b} \sum_{v_S} f(v_S) P(V_S = v_S|\mathcal{V} = (v_1, \ldots, v_b)) dP_{\mathcal{V}}(v_1, \ldots, v_b) \\ = \ & \textstyle\int_{[0,1]^b} \sum_{i=1}^{b} f(v_i) P(V_S = v_i|\mathcal{V} = (v_1, \ldots, v_b)) dP_{\mathcal{V}}(v_1, \ldots, v_b) \\ = \ & \textstyle\int_{[0,1]^b} \sum_{i=1}^{b} f(v_i) v_i dP_{\mathcal{V}}(v_1, \ldots, v_b) \\ = \ & E[\sum_{i=1}^{b} f(V_i) V_i]. \end{aligned}$$

Des Weiteren gilt für $E[Vf(V)]$:

$$E[Vf(V)] = \int vf(v)d(\frac{1}{b}\sum_{i=1}^{b} P_{V_i})(v) = \frac{1}{b}\sum_{i=1}^{b}\int v_i f(v_i)dP_{V_i}(v_i)$$

$$= \frac{1}{b}\sum_{i=1}^{b} E[V_i f(V_i)] = E[V]E[f(V_S)].$$

Damit folgt:

Lemma 4.0.4. *Mit den obigen Bezeichnungen gilt für die Verteilungen der Zufallsvariable W und des Splitters der Variablen $V_1, \ldots, V_b$ die Verteilungsgleichheit*

$$W = V_S \stackrel{\mathcal{L}}{=} V^*.$$

Für den Zentralen Grenzwertsatz, den wir in Abschnitt 5 formulieren wollen, sind zwei, lediglich von b und der Verteilung von $\mathcal{V}$ abhängige, Größen relevant. Sei $P(W \in (0,1)) = 1$. Wir definieren:

$$\mu := E[\log(\frac{1}{W})]$$

und

$$\sigma^2 := Var(\log W).$$

Da wir den Zusammenhang zwischen W und $\mathcal{V}$ kennen, können wir μ und σ^2 berechnen, ohne W zu kennen:

Korollar 4.0.5. *Sind die V_i, $i = 1, \ldots, b$, identisch verteilt mit $V_1 \stackrel{\mathcal{L}}{=} V$, so gilt*

$$\mu = bE[V \log \frac{1}{V}]$$

und

$$\sigma^2 = bE[V \log^2 V] - \mu^2.$$

Beweis:

$E[\log \frac{1}{W}]$

$= \int_{[0,1]^b} E[\log \frac{1}{W}|\mathcal{V} = (v_1, \ldots, v_b)]dP_{\mathcal{V}}(v_1, \ldots, v_b)$ nach dem Faktorisierungslemma,

$= \int_{[0,1]^b} \sum_{v_S}(\log \frac{1}{v_S})P(\log \frac{1}{V_S} = \log \frac{1}{v_S}|\mathcal{V} = (v_1, \ldots, v_b))dP_{\mathcal{V}}(v_1, \ldots, v_b)$

$$= \int_{[0,1]^b} \sum_{i=0}^{b} (\log \tfrac{1}{v_i}) P(\log \tfrac{1}{V_S} = \log \tfrac{1}{v_i} | \mathcal{V} = (v_1, \ldots, v_b)) dP_{\mathcal{V}}(v_1, \ldots, v_b)$$

$$= \int_{[0,1]^b} \sum_{i=0}^{b} (\log \tfrac{1}{v_i}) v_i dP_{\mathcal{V}}(v_1, \ldots, v_b) \text{ nach der Definition von } V_S,$$

$$= E[\sum_{i=0}^{b} (\log \tfrac{1}{V_i}) V_i]$$

$$= bE[(\log \tfrac{1}{V}) V], \text{ da die } V_i \text{ identisch verteilt sind.}$$

Für die zweite Aussage reicht es zu zeigen, dass $E[\log^2 W] = bE[V \log^2 V]$ gilt:

$$E[\log^2 W]$$

$$= \int_{[0,1]^b} E[\log^2 V_S | \mathcal{V} = (v_1, \ldots, v_b)] dP_{\mathcal{V}}(v_1, \ldots, v_b)$$

$$= \int_{[0,1]^b} \sum_{i=0}^{b} (\log^2 v_i) P(\log^2 V_S = \log^2 v_i | \mathcal{V} = (v_1, \ldots, v_b)) dP_{\mathcal{V}}(v_1, \ldots, v_b)$$

$$= \int_{[0,1]^b} \sum_{i=0}^{b} (\log^2 v_i) v_i dP_{\mathcal{V}}(v_1, \ldots, v_b)$$

$$= E[\sum_{i=0}^{b} (\log^2 V_i) V_i]$$

$$= bE[(\log^2 V) V].$$

□

5 Abstand der n-ten und n+1-ten Kugel im zufälligen Suchbaum

5.1 Vorüberlegungen

Um das Modell im Detail beschreiben zu können, müssen wir uns auf einige weitere Bezeichnungen und Zufallsvariablen einigen. Wir hatten schon im dritten Abschnitt D_n, die Tiefe der n-ten Kugel im Baum, angesprochen. Sie hängt natürlich davon ab, ob die besagte Kugel beim Aufbrechen eines Knotens weiter nach unten wandert oder, falls $s_0 > 0$, in dem Knoten mit den anderen $s_0 - 1$ Kugeln bleibt. Wir wollen o.B.d.A. davon ausgehen, dass die Kugel, die gerade im Baum eingefügt wird, immer möglichst weit nach unten transportiert wird. Dies entspricht der Motivation durch Bäume als Datenstrukturen, in die Daten nach und nach eingefügt werden.

Befindet sich eine Kugel im Teilbaum am Knoten u, so hängt ihre Tiefe in diesem Teilbaum auch davon ab, wieviele Kugeln sich insgesammt in dem Teilbaum befinden. Diese Anzahl wollen wir mit $N(u)$ bezeichnen.

Die maximale Tiefe einer Kugel in einem Baum mit insgesammt n Kugeln ist die *Höhe* dieses Baumes und wird von uns H_n genannt. Wir werden später das Verhalten der Höhe ausnutzen, um die Tiefe einer Kugel im Baum abzuschätzen.

Der Abstand der i-ten und j-ten Kugel im Baum bekommt von uns die Bezeichnung $\Delta_{i,j}$. Man kann diesen Abstand in zwei Tiefen D'_i und D'_j zerlegen (vgl. Abbildung 5, S. 19). Betrachtet man nämlich die zurückgelegten Pfade der beiden Kugeln, so gibt es ein Stück, auf dem sich die Kugeln gemeinsam bewegen. Dann erfolgt eine Trennung der Pfade. D'_i bzw. D'_j sind die um 1 verminderten Abstände zwischen dem letzten 'gemeinsamen' Knoten der Kugeln und ihren finalen Positionen im Baum. Gemeinsamer Knoten soll hier bedeuten, dass er auf beiden Pfaden der Kugeln zwischen Wurzel und ihren Standorten im Baum liegt.

Das führt uns auf die Zufallsvariable R, die die Tiefe des letzten gemeinsamen Knotens im Baum beschreiben soll. Das Ereignis $\{R = 0\}$ tritt ein, wenn die Wege der Kugeln sich sofort an der Wurzel trennen. Tritt $\{R = D_i\}$ oder sogar $\{R = D_i = D_j\}$ ein, heißt das, dass die Pfade der Kugeln sich erst am Zielknoten einer der Kugeln trennen oder sogar, dass beide Kugeln im gleichen Knoten gelandet sind.

Wir wollen uns nun die Zufallsvariable R genauer anschauen, wobei wir uns auf die n-te und $n+1$-te Kugel im Baum beziehen: Sei dazu
$A_k := \{n-\text{te } \textit{und } n+1-\text{te } \textit{Kugel durchlaufen selben Weg im Baum bis Stufe } k\}$.

$$
\begin{aligned}
& P(A_1) \\
= & \int P(A_1|\mathcal{V} = (v_1, \ldots, v_b)) dP_{\mathcal{V}}(v_1, \ldots, v_b) \\
= & \int \sum_{i=1}^{b} P(A_1,\ Kugeln\ gehen\ Weg\ i|\mathcal{V} = (v_1, \ldots, v_b)) dP_{\mathcal{V}}(v_1, \ldots, v_b) \\
= & \int \sum_{i=1}^{b} v_i^2 dP_{\mathcal{V}}(v_1, \ldots, v_b) \\
= & \sum_{i=1}^{b} E[V_i^2] =: q \\
= & \ bE[V^2],\ \text{falls die } V_i \text{ identisch verteilt sind.}
\end{aligned}
$$

Analog errechnen sich $P(A_2) = q^2,\ P(A_3) = q^3$, etc. Also ist R geometrisch verteilt zum Parameter $q = bE[V^2]$.

Sei ein zufälliger Splitbaum mit $n+1$ Kugeln gegeben. Der letzte gemeinsame Knoten soll im Weiteren mit u_R bezeichnet werden. Der erste Knoten auf Stufe $R+1$, den die n-te Kugel passiert, soll u_0 heißen, der entsprechende Knoten zur $n+1$-ten Kugel soll mit v_0 bezeichnet werden.

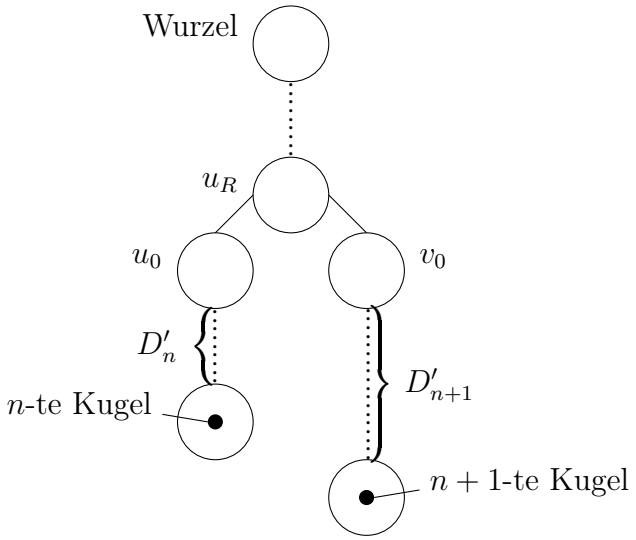

Abbildung 5.

Schließlich bezeichne $A_{j,u}$ das Ereignis, dass sich die j-te Kugel nicht im Knoten u befindet. Damit gilt:

$$\Delta_{n,n+1} = D'_n + D'_{n+1} + \mathbb{1}_{A_{n,u_R}} + \mathbb{1}_{A_{n+1,u_R}}.$$

Bemerkung 3. Die so definierten Tiefe, Höhe und Abstände sind nach Satz 2.0.1 eindeutig.

D'_n und D'_{n+1} sind nicht unabhängig. Später werden wir auf die Variablen $N(u_0)$ und $N(v_0)$ bedingen, um die Teilhöhen bezüglich stochastischer Konvergenz separat betrachten zu können und damit auf die Konvergenz von $\Delta_{n,n+1}$ zu schließen.

5.2 Sätze und Beweise

Die Tiefen D_n und D_{n+1} lassen sich von oben mit Hilfe der Höhe H_n abschätzen. Das werden wir später ausnutzen, um den Beweis zum Gesetz der großen Zahlen (Satz 5.2.3) zu liefern. Dafür wollen wir uns zunächst das Verhalten der Höhe vergegenwärtigen.

Bei einem Baum mit n Kugeln ist die Höhe H_n in Wahrscheinlichkeit asymptotisch höchstens ein Vielfaches von $\log n$. Der folgende Satz von L. Devroye [4] stellt dies dar und liefert außerdem ein Kriterium, unter dem eine untere Schranke γ für Konstanten c existiert, so dass $P(H_n > c \log n) \overset{n\to\infty}{\longrightarrow} 0$ für alle $c > \gamma$ erfüllt ist.

Satz 5.2.1. *Ist V ein Splitter eines zufälligen Splitbaumes und ist $P(V = 1) = 0$, dann existiert eine Konstante c, so dass*

$$\lim_{n\to\infty} P(H_n > c \log n) = 0.$$

Sei $m(t) := E[V^t]$ die Momentenfunktion von V, $M := \lim_{t\to\infty} \log m(t) - tm(t)'/m(t)$. Ist dann $M < -\log b$, so ist die obige Aussage wahr für alle $c > \gamma$. $\gamma \in (0, \infty)$ ist ein lediglich von b und der Verteilung von V abhängiger Parameter, der über

$$\gamma = \inf\{d : \exp(t^*(d))(bm(t^*(d)))^d < 1\}$$

definiert ist. $t^*(d)$ ist dabei Lösung der Gleichung $m(t)'/m(t) = -1/d$.

Bemerkung 4. Wir bedienen uns der Konvention, dass das Infimum einer leeren Menge ∞ ist. Damit ist γ immer wohldefiniert.

Bemerkung 5. Sind die Bedingungen von Satz 5.2.1 erfüllt, so ist die Aussage, nach L. Devroye [5] auch wahr für alle $c < \gamma$ und $H_n/\log n$ konvergiert somit gegen γ in Wahrscheinlichkeit. Da wir jedoch später, für die Beweise der Konvergenzsätze zu den Tiefen, nur die Existenz der oberen Schranke für H_n benötigen, gehen wir nicht näher auf den Beweis zur unteren Schranke von H_n ein. Der Beweis des Satzes 5.2.1 ist dem Anhang A zu entnehmen.

Für den Beweis des folgenden Satzes werden wir die Kardinalitäten der Kugeln in den Teilbäumen abschätzen müssen. Wie wir sehen werden, kann das durch binomialverteilte Zufallsvariablen mit zufälligem Erfolgsparameter geschehen. Wir brauchen ein Werkzeug, um die Masse, die die Verteilungen solcher Variablen messbaren Mengen zuordnen, abschätzen zu können. Dieses finden wir in dem folgenden, der Arbeit von L. Devroye [4] entnommenem, Korollar:

Korollar 5.2.2. *Sei $Z \in [0,1]$ eine Zufallsvariable und $n \in \mathbb{N}$. Dann gilt für $0 < a < n$:*

$$P(B(n,Z) \geq a) \leq P(Z > \frac{a}{2n}) + \left(\frac{e}{4}\right)^{a/2}$$

und

$$P(B(n,Z) \leq a) \leq P(Z < \frac{2a}{n}) + \left(\frac{2}{e}\right)^{a}.$$

Beweis: Unser Vorgehen wird sein, Chernoffschranken auf binomialverteilte Zufallsvariablen anzuwenden und dann zu einem zufälligen Erfolgsparameter überzugehen:
Für eine binomialverteilte Zufallsvariable $B(n,p)$ gelten, wie bei M. Okamoto [12] beschrieben, nach Chernoff die folgenden Schranken:
Ist $q \in [p,1)$, dann gilt

$$P(B(n,p) \geq nq) \leq \left(\left(\frac{p}{q}\right)^{q}\left(\frac{1-p}{1-q}\right)^{1-q}\right)^{n}$$

und ist $q \in (0,p]$, so gilt

$$P(B(n,p) \leq nq) \leq \left(\left(\frac{p}{q}\right)^{q}\left(\frac{1-p}{1-q}\right)^{1-q}\right)^{n}.$$

Setzen wir nun $q = 2p$ (sei für diesen Fall $p < 1/2$, später wird uns nur die Erfolgswahrscheinlichkeit $a/2n < 1/2$ interessieren) bzw. $q = p/2$ erhalten wir die Schranken

$$P(B(n,p) \geq 2np) \leq \left(\left(\frac{1}{2}\right)^{2p}\left(\frac{1-p}{1-2p}\right)^{1-2p}\right)^{n} = \left(\left(\frac{1}{2}\right)^{2p}\left(1+\frac{p}{1-2p}\right)^{1-2p}\right)^{n}$$

$$\leq \left(\left(\frac{1}{2}\right)^{2p} e^{p}\right)^{n} = \left(\frac{e}{4}\right)^{np}$$

bzw.

$$P(B(n,p) \leq np/2) \leq \left(2^{p/2}\left(1-\frac{p/2}{1-p/2}\right)^{1-p/2}\right)^n \leq \left(2^{p/2}e^{-p/2}\right)^n = \left(\sqrt{\frac{2}{e}}\right)^{np}.$$

Kehren wir nun zurück zu der Variable $B(n,Z)$. Sei $a \in (0,n)$ gewählt. Nach dem Satz von der totalen Wahrscheinlichkeit und den obigen Abschätzungen gilt:

$$\begin{aligned}
& P(B(n,Z) > a) \\
= \;& P(B(n,Z) > a,\ Z > a/(2n)) + P(B(n,Z) > a,\ Z \leq a/(2n)) \\
\leq \;& P(Z > a/(2n)) + P(B(n,a/(2n)) > a) \\
\leq \;& P(Z > a/(2n)) + \left(\tfrac{e}{4}\right)^{a/2}.
\end{aligned}$$

Und analog:

$$\begin{aligned}
& P(B(n,Z) \leq a) \\
= \;& P(B(n,Z) \leq a,\ Z < 2a/n) + P(B(n,Z) \leq a,\ Z \geq 2a/n) \\
\leq \;& P(Z < 2a/n) + P(B(n,2a/n) \leq a) \\
\leq \;& P(Z < 2a/n) + \left(\tfrac{2}{e}\right)^{a}.
\end{aligned}$$

□

Nun folgt das schwache Gesetz der großen Zahlen für die Tiefen D_n und D_{n+1}. Wir werden uns dieses Satzes von L. Devroye [4] bedienen, um später den analogen Satz für die Tiefe D'_n und D'_{n+1} zu beweisen.

Satz 5.2.3. *Ist* $E[\log \frac{1}{W}] = \mu,\ \sigma^2 = Var(\log W) \in (0,\infty)$ *und* $P(W \in (0,1)) = 1$*, so gilt*

$$\frac{D_n}{\log n} \to \frac{1}{\mu} \text{ in } Wahrscheinlichkeit,\ für\ n \to \infty$$

und

$$\frac{D_{n+1}}{\log n} \to \frac{1}{\mu} \text{ in } Wahrscheinlichkeit,\ für\ n \to \infty.$$

Beweis: Wir beweisen die erste Aussage des Satzes, die zweite folgt analog. Dazu schauen wir uns den Weg der n-ten Kugel ab der Wurzel w an und konstruieren den zufälligen Pfad $w, u_1, u_2, \ldots$, den die Kugel von da an nimmt. Dieser Pfad hat eine Länge $\leq n-1$. Bei gegebenem u_i und zugehörigem Vektor $\mathcal{V} = (V_1, \ldots, V_b)$ ist V_j die Wahrscheinlichkeit, dass u_{i+1} das j-te Kind von u_i ist.
Wir zeigen, dass $\forall c > 1/\mu$ bzw. $\forall c < 1/\mu$ $P(D_n > c\log n) \to 0$ bzw.

$P(D_n < c \log n) \to 0$ für $n \to \infty$ erfüllt ist.
Sei zunächst $c > 1/\mu$. Für beliebige $\beta,\ k,\ l \geq 0$ gilt

$$\{D_n \leq k + l\} \supseteq \{N(u_k) \leq \beta\} \cap \{H_\beta \leq l\}$$

und damit die Beziehung

$$\{D_n > k + l\} \subseteq \{N(u_k) > \beta\} \cup \{H_\beta > l\}.$$

Also ist

$$P(D_n > k + l) \leq P(N(u_k) > \beta) + P(H_\beta > l).$$

Für unsere Zwecke wählen wir $k = \lfloor c \log n \rfloor,\ \beta = (s_0 + 1)k$, l werden wir später passend bestimmen. Damit $\{N(u_k) > \beta\}$ eintritt, müssen von den n Kugeln im Baum mindestens β den gleichen Weg wie die n-te Kugel durchlaufen. Dies geschieht in jeder Stufe i unabhängig. Wenn man weiß, dass eine Kugel die Stufe i auf jeden Fall passiert, ist die zufällige Wahrscheinlichkeit, dass sie die „richtige“ Abzweigung nimmt, W. Außerdem bleiben in jeder Stufe dieses Weges mindestens s_0 Kugeln zurück.

Des Weiteren gilt, dass $B(B(m, p_1), p_2) \stackrel{\mathcal{L}}{=} B(m, p_1 p_2)$. Für $x \geq 0$ ist $P(B(m - x, p) > \beta) \leq P(B(m, p) > \beta - x)$.

Wir können nun die Anzahl der Kugeln abschätzen: Seien $(W_i)_{i \geq 1}$ unabhängige Kopien von W.

$$\begin{aligned}
& P(N(u_k) > \beta) \\
\leq\ & P\left(B(n, \prod_{i=1}^{k} W_i) > \beta - k s_0\right) \\
\leq\ & P\left(\prod_{i=1}^{k} W_i > k/(2n)\right) + (e/4)^{k/2} \text{ nach Korollar 5.2.2,} \\
=\ & P\left(\sum_{i=1}^{k} \log W_i > \log\left(k/(2n)\right)\right) + o(1) \\
=\ & P\left(\left(\sum_{i=1}^{k} \log W_i + k\mu\right)/k > (\log(k/(2n)) + k\mu)/k\right) + o(1) \\
\leq\ & P\left(\left|\left(\sum_{i=1}^{k} \log W_i + k\mu\right)/k\right| > (\log(k/(2n)) + k\mu)/k\right) + o(1).
\end{aligned}$$

Nach dem Gesetz der Großen Zahlen gilt $\forall \delta > 0$:

$$\lim_{n \to \infty} P\left(\left|\left(\sum_{i=1}^{k} \log W_i + k\mu\right)/k\right| > \delta\right) = 0$$

und da nach unserer Wahl von c

$$\liminf_{n\to\infty} \frac{\log(k/(2n)) + k\mu}{k}$$

$$= \liminf_{n\to\infty} \frac{\log c \log n - \log 2 - \log n}{c \log n} + \mu = -\frac{1}{c} + \mu > 0 \text{ gilt,}$$

folgt unmittelbar:

$$\lim_{n\to\infty} P\left(N(u_k) > \beta\right) = 0$$

Um zu erzwingen, dass auch $P(H_\beta > l)$ gegen 0 konvergiert, wählen wir $l = 2\gamma \log \beta$. γ ist die Konstante aus Satz 5.2.1 (Ist $\gamma = \infty$, so nehmen wir die Konstante aus dem ersten Teil des Satzes). Demnach gilt:

$$\lim_{\beta\to\infty} P(H_\beta > l) = \lim_{n\to\infty} P(H_\beta > l) = 0$$

und damit, da l proportional zu $\log\log n$ ist:

$$\begin{aligned}
& \lim_{n\to\infty} P(D_n > k + l) \\
= & \lim_{n\to\infty} P(D_n > c \log n + \log\log n) \\
= & \lim_{n\to\infty} P\left(D_n / \log n > c + \log\log n / \log n\right) \\
= & \lim_{n\to\infty} P\left(D_n / \log n > c + o(1)\right) \\
= & \ 0 \\
= & \lim_{n\to\infty} P\left(D_n / \log n > c\right) \text{ nach dem Satz von Slutzky.}
\end{aligned}$$

Nun sei $c < 1/\mu$, k und β verbleiben wie oben. Bezieht sich $N(\cdot)$ auf einen Baum, in dem genau $n-1$ Kugeln eingefügt sind, so gilt

$$\{D_n < k\} \subseteq \{N(u_k) = 0\}$$

und damit

$$P(D_n < k) \leq P(N(u_k) = 0).$$

Diese Wahrscheinlichkeit schätzen wir ab. Damit im Teilbaum am Knoten u_k keine Kugeln landet, darf keine der $n-1$ Kugeln von der Wurzel aus den eindeutig durch $\prod_{i=1}^{k} W_i$ beschriebenen Weg zu u_k nehmen. Dabei bleiben von den Kugeln,

deren Weg über $w, u_1, \ldots, u_{k-1}$ führt, höchstens ks auf der Strecke.

$$\begin{aligned}
&P(N(u_k) = 0) \\
&\leq P\left(B(n-1, \prod_{i=1}^{k} W_i) = 0\right) \\
&\leq P\left(B(n-1, \prod_{i=1}^{k} W_i) - ks \leq 0\right) \\
\leq P\left(\prod_{i=1}^{k} W_i \leq 2ks/(n-1)\right) &+ (2/e)^{ks} \text{ nach Korollar 5.2.2} \\
= P\left(\sum_{i=1}^{k} \log W_i \leq \log\left(2ks/(n-1)\right)\right) &+ o(1) \\
= P\left(\left(\sum_{i=1}^{k} \log W_i + k\mu\right)/k \leq \left(\log\left(2ks/(n-1)\right) + k\mu\right)/k\right) &+ o(1).
\end{aligned}$$

Wir hatten $c < 1/\mu$ gewählt und damit gilt

$$\limsup_{n\to\infty} \frac{\log\left(2ks/(n-1)\right) + k\mu}{k} = -\frac{1}{c} + \mu < 0.$$

Wir können, wie oben, das Gesetz der großen Zahlen anwenden und schlussfolgern, dass $P(N(u_k) = 0)$ und somit $P(D_n/\log n < c)$ für $n \to \infty$ gegen 0 konvergieren muss. Daraus folgt die Behauptung.

□

Für D_n und D_{n+1} formulieren wir nun Zentrale Grenzwertsätze. Der Beweis ist der Arbeit von L. Devroye [4] entnommen.

Satz 5.2.4. *In dem zufälligen Splitbaum gelten, bei Voraussetzungen wie in Satz 5.2.3, für D_n und D_{n+1} Grenzwertsätze der folgenden Form:*

$$\frac{D_n - (1/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{\log n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ für } n \to \infty$$

und

$$\frac{D_{n+1} - (1/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{\log n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ für } n \to \infty.$$

Beweis: Seien $(W_i)_{i\geq 1}$ wieder eine Folge unabhängiger Kopien von W. Wir nutzen aus, dass für $\sum \log W_i$ der ZGW-Satz gilt, d.h.:

$$\frac{\sum_{i=0}^{k} \log W_i + k\mu}{\sqrt{k}\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ für } n \to \infty.$$

Wir gehen von einem Baum mit n Kugeln aus. Sei $a \in \mathbb{R}$ beliebig, $k = (1/\mu)\log n + a\sqrt{\log n}$, $\beta = (s_0+1)k$ und $l = 2\gamma \log \beta$. Wir verwenden viele Überlegungen, die wir schon im Beweis zum Satz 5.2.3 auf Seite 22 verwendet haben:

$$
\begin{aligned}
& P(D_n \geq k+l) \\
& \leq P\left(N(u_k) \geq \beta\right) + P(H_\beta \geq l) \\
& \leq P\left(B(n, \prod_{i=0}^{k} W_i) \geq \beta - k s_0\right) + o(1) \text{ nach Satz 5.2.1,} \\
& \leq P\left(\sum_{i=0}^{k} \log W_i \geq \log(k/(2n))\right) + (e/4)^{k/2} + o(1) \text{ nach Korollar 5.2.2,} \\
& \leq P\left(\left(\sum_{i=0}^{k} \log W_i + k\mu\right)/\sqrt{k\sigma^2} \geq (\log(k/(2n)) + k\mu)/\sqrt{k\sigma^2}\right) + o(1).
\end{aligned}
$$

Es gilt:

$$
\begin{aligned}
\lim_{n\to\infty} \frac{\log(k/(2n)) + k\mu}{\sqrt{k}\sigma} &= \lim_{n\to\infty} \frac{-\log n + \log n + \mu a\sqrt{\log n}}{\sigma\sqrt{(1/\mu)\log n + a\sqrt{\log n}}} \\
&= \lim_{n\to\infty} \frac{\mu^{3/2} a}{\sigma} \frac{\sqrt{\log n}}{\sqrt{\log n + a\mu\sqrt{\log n}}} = \frac{\mu^{3/2} a}{\sigma}
\end{aligned}
$$

und damit:

$$
\begin{aligned}
& P(D_n \geq k+l) \\
\leq \quad & P\left(\left(\sum_{i=0}^{k} \log W_i + k\mu\right)/\sqrt{k\sigma^2} + o(1) \geq \mu^{3/2} a/\sigma\right) + o(1) \\
\overset{n\to\infty}{\longrightarrow} \quad & 1 - \Phi\left(\mu^{3/2} a/\sigma\right) \text{ nach dem ZGWS und dem Satz von Slutzky.}
\end{aligned}
$$

Andererseits gilt:

$$\begin{aligned}
& P(D_n < k) \\
\leq\ & P(N(u_k) = 0) \\
\leq\ & P\left(B(n-1, \prod_{i=0}^{k} W_i) < ks\right) \\
\leq\ & P\left(\sum_{i=0}^{k} \log W_i < \log(2ks/(n-1))\right) + (2/e)^{ks} \text{ nach Korollar 5.2.2,} \\
\leq\ & P\left(\left(\sum_{i=0}^{k} \log W_i + k\mu\right)/\sqrt{k\sigma^2} < \tfrac{\log(2ks/(n-1))+k\mu}{\sqrt{k\sigma^2}}\right) + o(1) \\
\leq\ & P\left(\left(\sum_{i=0}^{k} \log W_i + k\mu\right)/\sqrt{k\sigma^2} + o(1) < \mu^{3/2}a/\sigma\right) + o(1) \\
\overset{n\to\infty}{\longrightarrow}\ & \Phi\left(\mu^{3/2}a/\sigma\right) \text{ nach dem Satz von Slutzky.}
\end{aligned}$$

Sei N standardnormalverteilt. Nach unserer Wahl von k und l gilt:
$\lim_{n\to\infty}(k+l-(1/\mu)\log n)/\sqrt{\log n} = \lim_{n\to\infty}(k-(1/\mu)\log n)/\sqrt{\log n} = a$. Damit haben wir gezeigt:

$$1)\ \limsup_{n\to\infty} P\left(\frac{D_n - (1/\mu)\log n}{\sqrt{\log n}} \geq a\right) \leq P(N \geq \frac{\mu^{3/2}a}{\sigma})$$

$$2)\ \limsup_{n\to\infty} P\left(\frac{D_n - (1/\mu)\log n}{\sqrt{\log n}} < a\right) \leq P(N < \frac{\mu^{3/2}a}{\sigma})$$

$$\Longrightarrow \liminf_{n\to\infty} P\left(\frac{D_n - (1/\mu)\log n}{\sqrt{\log n}} \geq a\right) \geq P(N \geq \frac{\mu^{3/2}a}{\sigma}).$$

Das ist gleichbedeutend mit:

$$1)\ \limsup_{n\to\infty} P\left(\frac{\mu^{3/2}}{\sigma}\frac{D_n - (1/\mu)\log n}{\sqrt{\log n}} \geq a\right) \leq P(\frac{\mu^{3/2}}{\sigma}N \geq \frac{\mu^{3/2}a}{\sigma}) = P(N \geq a)$$

$$2)\ \liminf_{n\to\infty} P\left(\frac{\mu^{3/2}}{\sigma}\frac{D_n - (1/\mu)\log n}{\sqrt{\log n}} \geq a\right) \geq P(N \geq a).$$

Es folgt die Behauptung:

$$\frac{D_n - (1/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{\log n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ für } n \to \infty.$$

□

Wir können jetzt unsere ersten Resultate, das schwache Gesetz der großen Zahlen und den Zentralen Grenzwertsatz für die Teiltiefen D'_n und D'_{n+1}, formulieren und beweisen:

Satz 5.2.5. *Ist* $E[\log \frac{1}{W}] = \mu,\ \sigma^2 = Var(\log W) \in (0,\infty)$ *und* $P(W \in (0,1)) = 1$*, so gilt:*

$$\frac{D'_n}{\log n} \to \frac{1}{\mu} \text{ in } Wahrscheinlichkeit,\ für\ n \to \infty$$

und

$$\frac{D'_{n+1}}{\log n} \to \frac{1}{\mu} \text{ in } Wahrscheinlichkeit,\ für\ n \to \infty.$$

Beweis: Per Definition gilt $D'_n = D_n - R - \mathbb{1}_{A_{n,u_R}}$. Sei $\varepsilon > 0$. Nach dem Satz von der totalen Wahrscheinlichkeit gilt:

$$\begin{aligned}
&P(D'_n \le (\tfrac{1}{\mu} - \varepsilon)\log n)\\
&= P(D_n - R - \mathbb{1}_{A_{n,u_R}} \le (\tfrac{1}{\mu} - \varepsilon)\log n)\\
&= P(D_n - R - \mathbb{1}_{A_{n,u_R}} \le (\tfrac{1}{\mu} - \varepsilon)\log n,\ D_n \le (\tfrac{1}{\mu} - \tfrac{\varepsilon}{2})\log n,\ \mathbb{1}_{A_{n,u_R}} + R \ge \tfrac{\varepsilon}{2}\log n)\\
&\quad + P(D_n - R - \mathbb{1}_{A_{n,u_R}} \le (\tfrac{1}{\mu} - \varepsilon)\log n,\ D_n \le (\tfrac{1}{\mu} - \tfrac{\varepsilon}{2})\log n,\ \mathbb{1}_{A_{n,u_R}} + R < \tfrac{\varepsilon}{2}\log n)\\
&\quad + P(D_n - R - \mathbb{1}_{A_{n,u_R}} \le (\tfrac{1}{\mu} - \varepsilon)\log n,\ D_n > (\tfrac{1}{\mu} - \tfrac{\varepsilon}{2})\log n,\ \mathbb{1}_{A_{n,u_R}} + R \ge \tfrac{\varepsilon}{2}\log n)\\
&\le 2P(D_n \le (\tfrac{1}{\mu} - \tfrac{\varepsilon}{2})\log n) + P(\mathbb{1}_{A_{n,u_R}} + R \ge \tfrac{\varepsilon}{2}\log n).
\end{aligned}$$

Nach Satz 5.2.3 und da R geometrisch verteilt ist und somit fast sicher kleiner ∞ ist, folgt:

$$P(D'_n \le (\frac{1}{\mu} - \varepsilon)\log n) \overset{n\to\infty}{\longrightarrow} 0.$$

Außerdem gilt:

$$\begin{aligned}
& P(D'_n \ge (1/\mu + \varepsilon)\log n)\\
\le\ & P(D_n \ge (1/\mu + \varepsilon)\log n)\\
\overset{n\to\infty}{\longrightarrow}\ & 0 \text{ nach Satz 5.2.3.}
\end{aligned}$$

□

Satz 5.2.6. *Im zufälligen Splitbaum gelten, bei Voraussetzungen wie im Satz 5.2.5, für* D'_n *und* D'_{n+1} *Grenzwertsätze der folgenden Form:*

$$\frac{D'_n - (1/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{\log n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ für } n \to \infty$$

und

$$\frac{D'_{n+1} - (1/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{\log n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ für } n \to \infty.$$

Beweis: Sei $x \in \mathbb{R}$.

$$\begin{aligned} & P(\tfrac{D'_n - (1/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{\log n}} \leq x) \\ = \quad & P(\tfrac{D_n - (1/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{\log n}} - \tfrac{R+\mathbb{1}_{A_{n,u_R}}}{\sigma\mu^{-3/2}\sqrt{\log n}} \leq x) \\ \overset{n\to\infty}{\longrightarrow} \quad & \Phi(x) \text{ nach Satz 5.2.4 und dem Satz von Slutzky} \end{aligned}$$

und da R geometrisch verteilt ist und somit für beliebiges $\varepsilon > 0$ für wachsendes n die Wahrscheinlichleit $P(R/\sqrt{\log n} \geq \varepsilon)$ gegen 0 geht.

□

Wir haben gezeigt, dass sich die Tiefen D'_n und D'_{n+1} logarithmisch verhalten. Damit können wir, da der Abstand der n-ten und $n+1$-ten Kugel bis auf eine Abweichung von höchstens 2 die Summe dieser zwei Tiefen ist, den Zentralen Grenzwertsatz für den Abstand der beiden zuletzt eingefügten Kugeln formulieren.

Satz 5.2.7. *(ZGWS)*
Ist $\Delta_{n,n+1}$ der Abstand der n-ten und $n+1$ Kugeln in einem zufälligem Splitbaum und sind $E[\log \frac{1}{W}] = \mu$, $\sigma^2 = Var(\log W) \in (0,\infty)$ mit $P(W \in (0,1)) = 1$, so gilt:

$$\frac{\Delta_{n,n+1} - (2/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{2\log n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ für } n \to \infty.$$

Zum Beweis dieses Satzes benötigen wir noch die folgenden Lemmata:

Lemma 5.2.8. *Seien $(X_n)_{n\geq 1}$ und $(Y_n)_{n\geq 1}$ zwei unabhängige Folgen von Zufallsvariablen und seien X und Y standardnormalverteilt mit*

$$X_n \xrightarrow{\mathcal{L}} X \text{ und } Y_n \xrightarrow{\mathcal{L}} Y \text{ für } n \to \infty,$$

dann gilt:

$$\frac{1}{\sqrt{2}}X_n + \frac{1}{\sqrt{2}}Y_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ für } n \to \infty.$$

Beweis: Seien $\varphi_{X_n}(t)$ und $\varphi_{Y_n}(t)$ die charakteristischen Funktionen der Zufalls-

variablen X_n und Y_n für $n \in \mathbb{N}$. Es gilt:

$$\begin{aligned}
& \varphi_{(1/\sqrt{2})(X_n+Y_n)}(t) \\
= \quad & E[\exp((1/\sqrt{2})(X_n+Y_n)it)] \\
= \quad & E[\exp(\tfrac{1}{\sqrt{2}}X_n it)\exp(\tfrac{1}{\sqrt{2}}Y_n it)] \\
= \quad & E[\exp(\tfrac{1}{\sqrt{2}}X_n it)]E[\exp(\tfrac{1}{\sqrt{2}}Y_n it)], \text{ da die Zufallsvariablen unabhängig sind,} \\
\overset{n\to\infty}{\longrightarrow} \quad & \varphi_X(t/\sqrt{2})\varphi_Y(t/\sqrt{2}) \text{ nach dem Stetigkeitssatz von Lévy,} \\
= \quad & \exp(t^2/4)\exp(t^2/4) \\
= \quad & \exp(t^2/2),
\end{aligned}$$

was die charakteristische Funktion einer standardnormalverteilten Zufallsvariable ist. Mit dem Stetigkeitssatz von Lévy und der Eindeutigkeit charakteristischer Funktionen folgt die Behauptung.

□

Lemma 5.2.9. *Sind μ und $\sigma \in (0,\infty)$ wie oben definiert und gilt $P(W \in (0,1)) = 1$, so gilt für alle $\varepsilon \in (0,1/2)$ und alle $k \in [\varepsilon n, n]$:*

$$\frac{D_k - (1/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{\log n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ für } n \to \infty.$$

Beweis: Für wachsendes n gilt:

$$\frac{D_k - (1/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{\log n}} = \underbrace{\sqrt{\frac{\log k}{\log n}}}_{\to 1} \frac{D_k - (1/\mu)\log k}{\sigma\mu^{-3/2}\sqrt{\log k}} + \underbrace{\frac{(1/\mu)\log(k/n)}{\sigma\mu^{-3/2}\sqrt{\log n}}}_{\to 0}.$$

Da $\frac{D_k-(1/\mu)\log k}{\sigma\mu^{-3/2}\sqrt{\log k}}$ nach Satz 5.2.4 in Verteilung gegen eine Standardnormalverteilte Zufallsvariable konvergiert, folgt die Behauptung nach dem Satz von Slutzky.

□

Nun zum Beweis des Satzes 5.2.7:
Beweis: Um später das Lemma 5.2.8 anwenden zu können wollen wir im ersten Schritt des Beweises auf die Kardinalitäten $N(u_0)$ und $N(v_0)$ der Teilbäume an den Knoten u_0 und v_0 bedingen und die daraus resultierende Unabhängigkeit der Variablen verwenden. Wir nutzen dabei aus, dass die Tiefen zweier verschiedener Bäume unabhängig sind. u_0 und v_0 sollen weiterhin die ersten Knoten auf Stuffe $R+1$ auf den Pfaden zu der n-ten und $n+1$-ten Kugel sein. (vgl. Abbildung 5, S.19).

Also: Sei $\varepsilon \in (0,1/2)$ beliebig und $I_\varepsilon = [\varepsilon n, n]$ ein Intervall und $x \in \mathbb{R}$ beliebig. Nach dem Satz von der totalen Wahrscheinlichkeit gilt:

$$P\left((\Delta_{n,n+1} - (2/\mu)\log n)/(\sigma\mu^{-3/2}\sqrt{2\log n}) \le x\right)$$

$$= P\left(\frac{\Delta_{n,n+1}-(2/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{2\log n}} \le x | N(u_0), N(v_0) \in I_\varepsilon\right) P(N(u_0), N(v_0) \in I_\varepsilon)$$

$$+P\left(\frac{\Delta_{n,n+1}-(2/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{2\log n}} \le x | \{N(u_0), N(v_0) \in I_\varepsilon\}^c\right) P(\{N(u_0), N(v_0) \in I_\varepsilon\}^c).$$

Es ist

$$P\left(\frac{\Delta_{n,n+1} - (2/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{2\log n}} \le x | N(u_0), N(v_0) \in I_\varepsilon\right) =$$

$$P\left(\frac{D'_n + D'_{n+1} + \mathbb{1}_{A_{n,u_R}} + \mathbb{1}_{A_{n+1,u_R}} - (2/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{2\log n}} \le x | N(u_0), N(v_0) \in I_\varepsilon\right) =$$

$$P\left(\frac{1}{\sqrt{2}}\frac{\overline{D_{N(u_0)'}} - (1/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{\log n}} + \frac{1}{\sqrt{2}}\frac{\overline{\overline{D_{N(v_0)'}}} - (1/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{\log n}} + \frac{\mathbb{1}_{A_{n,u_R}} + \mathbb{1}_{A_{n+1,u_R}}}{\sigma\mu^{-3/2}\sqrt{\log n}} \le x\right),$$

wobei $\overline{D_{N(u_0)'}}$ und $\overline{\overline{D_{N(v_0)'}}}$ Tiefen zweier unabhängiger Bäume mit $N(u_0)'$ bzw. mit $N(v_0)'$ Kugeln sind. $N(u_0)'$ und $N(v_0)'$ sind aus I_ε und somit gilt nach Lemma 5.2.9:

$$\frac{\overline{D_{N(u_0)'}} - (1/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{\log n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ für } n \to \infty$$

und

$$\frac{\overline{\overline{D_{N(v_0)'}}} - (1/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{\log n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ für } n \to \infty.$$

$\frac{\mathbb{1}_{A_{n,u_R}} + \mathbb{1}_{A_{n+1,u_R}}}{\sigma\mu^{-3/2}\sqrt{\log n}}$ konvergiert fast sicher gegen 0. Damit folgt mit Lemma 5.2.9 und dem Satz von Slutzky:

$$P\left(\frac{\Delta_{n,n+1} - (2/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{2\log n}} \le x | N(u_0), N(v_0) \in I_\varepsilon\right) \overset{n\to\infty}{\longrightarrow} \Phi(x).$$

Es bleiben die restlichen Wahrscheinlichkeiten in der Zerlegung abzuschätzen. Wir werden ε beliebig klein werden lassen und dadurch die gewünschten Konvergenzen erlangen. Dazu definieren wir:

$$\psi^+(\varepsilon) := \limsup_{n\to\infty} P(N(u_0), N(v_0) \in I_\varepsilon),\ \psi^-(\varepsilon) := \liminf_{n\to\infty} P(N(u_0), N(v_0) \in I_\varepsilon)$$

Nach den obigen Überlegungen gilt für alle $\varepsilon \in (0,1/2)$:

$$\limsup_{n\to\infty} P\left(\frac{\Delta_{n,n+1} - (2/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{2\log n}} \le x\right) \le \psi^+(\varepsilon)\Phi(x) + (1 - \psi^-(\varepsilon))$$

und

$$\liminf_{n\to\infty} P\left(\frac{\Delta_{n,n+1} - (2/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{2\log n}} \leq x\right) \geq \psi^-(\varepsilon)\Phi(x).$$

Um den Beweis zu beenden, zeigen wir: $\lim_{\varepsilon\downarrow 0} \psi^{\pm}(\varepsilon) = 1$. Dazu wiederum genügt es zu zeigen, dass

$$\forall \delta > 0\ \exists \varepsilon > 0 \text{ mit } \limsup_{n\to\infty} P(N(u_0), N(v_0) \leq \varepsilon n) \leq \delta$$

erfüllt ist.

Sei also $\delta > 0$ beliebig. Und seien $(W_i)_{i\geq 1}$ unabhängig und identisch verteilt mit $P_{W_1} = P_{V_S}$. $\overline{W} := \prod_{i=1}^{R+1} W_i$ ist das Produkt der Gewichte auf dem Pfad zum Knoten u_0. Da wir gefordert hatten, dass $W_1 > 0$ fast sicher erfüllt ist, gilt auch selbiges für $\overline{W}$. Damit existiert ein $\varepsilon > 0$, mit $P(\overline{W} \leq 2\varepsilon) > \delta/6$.

R ist geometrisch verteilt und insbesondere gilt $R < \infty$ fast sicher. Damit existiert ein $L \in \mathbb{N}$ mit $P(R \geq L) < \delta/6$. Wir benutzen erneut den Satz von der totalen Wahrscheinlichkeit:

$$\begin{aligned}
& P(N(u_0) \leq \varepsilon n) \\
= \ & P(N(u_0) \leq \varepsilon n | R < L,\ \overline{W} > 2\varepsilon) P(R < L,\ \overline{W} > 2\varepsilon) \\
& + P(N(u_0) \leq \varepsilon n,\ \{R < L,\ \overline{W} > 2\varepsilon\}^c) \\
\leq \ & P(N(u_0) \leq \varepsilon n | R < L,\ \overline{W} > 2\varepsilon) + P(R \geq L) + P(\overline{W} \leq 2\varepsilon) \\
\leq \ & P(N(u_0) \leq \varepsilon n | R < L,\ \overline{W} > 2\varepsilon) + 2\delta/6, \text{ nach der Wahl von } \varepsilon \text{ und } L.
\end{aligned}$$

Gegeben $R < L$ und $W > 2\varepsilon$ gilt $N(u_0) \geq B(n-1, 2\varepsilon) - Ls_0$ im stochastischen Sinne, denn: $B(n-1, 2\varepsilon)$ schätzt die Anzahl Kugeln, die den selben Weg wie die n-te Kugel gehen könnten, von unten ab. Ls_0 schätzt die Anzahl Kugeln, die dabei in jeder Stuffe zurückbleiben müssten, von oben ab. Es folgt:

$$\begin{aligned}
& P(N(u_0) \leq \varepsilon n | R < L,\ \overline{W} > 2\varepsilon) \\
\leq \ & P(B(n-1, 2\varepsilon) \leq \varepsilon n + Ls_0) \\
= \ & P(B(n-1, 2\varepsilon) - 2\varepsilon(n-1) \leq -\varepsilon n + \varepsilon + Ls_0) \\
\leq \ & \exp(\tfrac{-2(\varepsilon n + \varepsilon + Ls_0)^2}{n}) \text{ nach Höffding und für alle } n \text{ hinreichend groß}, \\
\overset{n\to\infty}{\longrightarrow} \ & 0.
\end{aligned}$$

Wählt man also n groß genug, gilt $P(N(u_0) \leq \varepsilon n | R < L,\ \overline{W} > 2\varepsilon) \leq \delta/6$ und folglich $\limsup_{n\to\infty} P(N(u_0) \leq \varepsilon n) \leq \delta/2$. Daraus folgt:

$$\limsup_{n\to\infty} P(N(u_0), N(v_0) \leq \varepsilon n) \leq \delta,$$

und damit die Behauptung des Satzes.

□

Wir wollen nun zu unserem Zielresultat, dem Zentralen Grenzwertsatz für den Abstand zwischen zwei uniform gewählten Kugeln kommen. Seien dazu I und J zwei unabhängige, auf der Menge $\{1,\dots,n\}$ uniform verteilte Zufallsvariablen. Sie stellen die zufällige Wahl zweier Kugeln im Splitbaum dar. Die Wahrscheinlichkeit, dass $\{I=J\}$ eintritt, geht für die wachsende Anzahl von Kugeln gegen 0:

$$P(I=J)=\sum_{i=1}^{n}P(I=i,J=i)=\sum_{i=1}^{n}P(I=i)P(J=i)=n/n^2=1/n.$$

Sind die zwei Kugeln gewählt, so kann man von dem Baum mit n Kugeln, zu einem Baum mit $\max(I,J)$ Kugeln übergehen, ohne dass sich der Abstand zwischen den beiden Kugeln ändert. Es gilt, dass
$\lim_{n\to\infty}P(\max(I,J)>m)=\lim_{n\to\infty}P(\min(I,J)>m)=1\ \forall m\in\mathbb{N}$, denn:

$$\begin{aligned}P(\max(I,J)>m)&=2P(I>m,\ J\le m)+P(I>m,\ J>m)\\&=2\frac{n-m}{n}\frac{m}{n}+\left(\frac{n-m}{n}\right)^2=\frac{n^2-m^2}{n^2}.\end{aligned}$$

Eine analoge Rechnung lässt sich auch für $\min(I,J)$ durchführen. Wir können also davon ausgehen, dass mit wachsendem n auch die zwei zufälligen Kugeln immer später im Baum eingefügt werden.

Wir formulieren nun unser Zielresultat, den Zentralen Grenzwertsatz für den Abstand zweier zufälliger Kugeln im zufälligen Splitbaum:

Satz 5.2.10. *Ist mit $\Delta_{I,J}$ der Abstand zwischen zwei rein zufälligen Kugeln im Baum mit insgesammt n Kugeln bezeichnet und sind $E[\log\frac{1}{W}]=\mu$, $\sigma^2=Var(\log W)\in(0,\infty)$ mit $P(W\in(0,1))=1$, so gilt der folgende Grenzwertsatz:*

$$\frac{\Delta_{I,J}-(2/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{2\log n}}\xrightarrow{\mathcal{L}}\mathcal{N}(0,1)\text{ für }n\to\infty.$$

Beweis: Man kann $\Delta_{I,J}$ auf folgende Weise darstellen:

$$\Delta_{I,J}=D'_{\max(I,J)}+D'_{\min(I,J)}+\mathbb{1}_{A_{I,u_R}}+\mathbb{1}_{A_{J,u_R}}.$$

Wir gehen in Analogie zum Beweis von Satz 5.2.7 vor: Sei $\varepsilon\in(0,1/2)$, $x\in\mathbb{R}$ und $I_\varepsilon=[\varepsilon n,n]$ ein Intervall. Es gilt:

$$\begin{aligned}&P\left(\tfrac{\Delta_{I,J}-(2/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{2\log n}}\le x\right)\\&=P\left(\tfrac{\Delta_{I,J}-(2/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{2\log n}}\le x|\max(I,J),\min(I,J)\in I_\varepsilon\right)P(\max(I,J),\min(I,J)\in I_\varepsilon)\\&+P\left(\tfrac{\Delta_{I,J}-(2/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{2\log n}}\le x|\{\max(I,J),\min(I,J)\in I_\varepsilon\}^c\right)P(\{\max(I,J),\min(I,J)\in I_\varepsilon\}^c).\end{aligned}$$

Mit der obigen Zerlegung gilt:

$$P\left(\frac{\Delta_{I,J} - (2/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{2\log n}} \leq x | \max(I,J), \min(I,J) \in I_\varepsilon\right) =$$

$$P\left(\frac{D'_{\max(I,J)} + D'_{\min(I,J)} + \mathbb{1}_{A_{I,u_R}} + \mathbb{1}_{A_{J,u_R}} - (2/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{2\log n}} \leq x | \max(I,J), \min(I,J) \in I_\varepsilon\right)$$

$$\stackrel{n\to\infty}{\longrightarrow} \Phi(x),$$

denn: Gegeben $\max(I,J), \min(I,J) \in I_\varepsilon$ gelten für $D_{\max(I,J)}$ und $D_{\min(I,J)}$ nach Lemma 5.2.9 die Grenzwertsätze

$$\frac{D_{\max(I,J)} - (1/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{\log n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ und } \frac{D_{\min(I,J)} - (1/\mu)\log n}{\sigma\mu^{-3/2}\sqrt{\log n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Man kann wie beim Beweis des Satzes 5.2.6 argumentieren und erhält Grenzwertsätze der gleichen Form auch für die Teiltiefen $D'_{\max(I,J)}$ und $D'_{\min(I,J)}$. Bedingt man weiter, wie im Beweis des Satzes 5.2.7, auf die Kardinalitäten der Kugeln in den Teilbäumen, in denen sich die $\max(I,J)$-te und $\min(I,J)$-te Kugel befinden, erhält man einen Ausdruck mit zwei unabhängigen Tiefen. Nach den Lemmata 5.2.9 und 5.2.8 und dem Satz von Slutzky folgt die Konvergenz gegen $\Phi(x)$.

Um die Gesamtkonvergenz zu beweisen, bleibt zu zeigen, dass

$$\limsup_{\varepsilon\downarrow 0} \limsup_{n\to\infty} P(\{\max(I,J), \min(I,J) \in I_\varepsilon\}^c) = 0,$$

und dafür zeigen wir, dass folgendes erfüllt ist:

$\forall\delta > 0\, \exists\varepsilon > 0$ mit $P(\{\max(I,J), \min(I,J) \in I_\varepsilon\}^c) < \delta$ für alle n hinreichend groß.

Sei also $\delta > 0$ beliebig. Es gilt:

$$\begin{aligned}
& P(\{\max(I,J), \min(I,J) \in I_\varepsilon\}^c) \\
= \;& P(\max(I,J) \in I_\varepsilon,\ \min(I,J) \notin I_\varepsilon) + P(\max(I,J) \notin I_\varepsilon,\ \min(I,J) \notin I_\varepsilon) \\
\leq \;& P(\min(I,J) \notin I_\varepsilon) + P(I \notin I_\varepsilon)^2 \\
= \;& 2P(\min(I,J) \notin I_\varepsilon | I = \min(I,J)) \underbrace{P(I = \min(I,J))}_{=1/2} + P(I \notin I_\varepsilon)^2 \\
\leq \;& \sum_{k=1}^{\lfloor \varepsilon n \rfloor} 1/n + \left(\sum_{k=1}^{\lfloor \varepsilon n \rfloor} 1/n\right)^2 \\
\leq \;& \varepsilon + \varepsilon^2.
\end{aligned}$$

Wählen wir also ε so, dass $\varepsilon + \varepsilon^2 < \delta$ erfüllt ist, folgt die Behauptung.

□

6 Beispiele: Der Binärer Suchbaum und der b-näre Suchbaum

6.1 Der Binäre Suchbaum

Der Baum hat einen Verzweigungsgrad von 2, und wir gehen von n Kugeln aus. Ein Knoten des Baumes enthält immer genau eine Kugel.

Wir greifen auf die Darstellung von L. Devroye [4] zurück:

Modell. Jeder Kugel i ist eine unabhängige und auf $[0,1]$ uniformverteilte Zufallsvariable U_i zugeordnet. Auf diese Weise erhalten wir eine Folge von n zufälligen Werten. Die erste Kugel wird in der Wurzel platziert. Die zweite wandert zum linken Kind der ersten, falls $U_1 < U_2$ und in das rechte sonst. Dieser Vorgang wird rekursiv wiederholt, bis jede Kugel einen Knoten zugewiesen bekommen hat.

Um den binären Suchbaum in das verallgemeinerte Modell einzugliedern, wählen wir $b = 2,\ s = s_0 = 1,\ s_1 = 0$ und $\mathcal{V} = (U, 1-U)$, wobei $U \stackrel{\mathcal{L}}{=} unif[0,1]$. Wir wenden den allgemeinen Algorithmus an, um die n Kugeln zu verteilen.

Der Splitter V von $\mathcal{V}$ ist wieder uniform verteilt, da für $x_1, x_2 \in [0,1],\ x_2 \geq x_1$ folgende Gleichheit erfüllt ist:

$$1/2P(U \in [x_1, x_2]) + 1/2P(1-U \in [x_1, x_2]) = x_2 - x_1.$$

Damit hat $V_S = W$ die Verteilung $P_W(dx) = 2xdx$ und $\mu = 1/2,\ \sigma^2 = 1/4$. Nach Satz 5.2.7 folgt für uniform gewählte Kugeln I, J:

$$\frac{\Delta_{I,J} - 4\log n}{2\sqrt{\log n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ für } n \to \infty.$$

Dieses Resultat wurde von H. M. Mahmoud und R. Neininger [11] über das binäre Suchbaummodell direkt gezeigt.

6.2 Der b-näre Suchbaum

Dieser Suchbaum ist eine Verallgemeinerung des binären Suchbaummodells und lässt einen Verzweigungsgrad von $b \geq 2$ zu. Außerdem kann nun ein Knoten bis zu $b-1$ Kugeln enthalten. Dieses Modell war zunächst dazu gedacht, durch die größeren Verzweigung und Kapazität der Knoten die Verbindungswege zwischen den Daten (Kugeln) zu verkleinern.

Darstellungen dieses Suchbaummodells finden sich bei L. Devroye [4] und H. M. Mahmoud [10].

Modell. Um den Baum zu erzeugen, wird auch hier jeder der n Kugeln eine unabhängige, auf $[0,1]$ uniformverteilte Zufallsvariable zugeordnet. Ist $n \leq b-1$, so werden alle Kugeln in der Wurzel des Baumes platziert. Ist $n > b-1$, so werden $b-1$ zufällig ausgewählte Kugeln der Wurzel hinzugefügt. Die entsprechenden, der Größe nach sortierten Zufallsvariablen $U^{(1)}, \ldots, U^{(b-1)}$ liefern eine Unterteilung des Intervalls $[0,1]$ in b Abschnite $[0, U^{(1)}), \ldots, [U^{(b-1)}, 1]$. Die restlichen $n-b+1$ Kugeln werden nun wie folgt auf die b Kinder des Wurzelknotens aufgeteilt: Alle Kugeln, deren zugehörige Zufallsvariablen im i-ten der Intervalle liegen, gehen zum i-ten Kind über. Werden dabei einem oder mehreren Knoten mehr als $b-1$ Kugeln zugewiesen, so wird der Algorithmus rekursiv auf diese Knoten angewendet.

Die Verteilung der $n-b+1$ Kugeln auf die b Knoten der 1. Stufe des Baumes entspricht einer multinomialverteilten Zufallsvariable mit dem zufälligen Parameter $(n-b+1, V_1, \ldots, V_b)$. V_i ist die zufällige Länge des i-ten Intervalls. Da $U_{(k)}$ als Ordnungsstatistiken von b uniformverteilten Zufallsvariablen nach G. Kersting [8] Beta-verteilt zum Parameter $(k, b-k)$ sind, ist V_1, da $P(V_1 \leq a) = P(U_{(1)} \leq a)$ gilt, $Beta(1, b-1)$ verteilt. Damit folgt für den Splitter V der V_i:

$$1/b\,(P(V_1 \leq a) + \ldots + P(V_{b-1} \leq a))$$

$$= 1/b\,\big(P(U_{(1)} \leq a) + P(U_{(2)} - U_{(1)} \leq a) + \ldots + P(1 - U_{(b-1)} \leq a)\big)$$

$$= P(U_{(1)} \leq a).$$

Das bedeutet, dass V in diesem Fall ebenfalls $Beta(1, b-1)$ verteilt ist.

Um nun also wieder zum generalisierten Modell zu wechseln und den b-näre Suchbaum einzugliedern, wählen wir $b = b,\ s = s_0 = b-1,\ s_1 = 0$. Der Vektor $\mathcal{V} = (V_1, \ldots, V_b)$ soll aus Variablen bestehen, die der Verteilung nach den Intervalllängen im obigen Modell entsprechen und V soll wieder den Splitter dieser Variablen bezeichnen. Um μ und σ^2 zu bestimmen, nutzen wir das folgende Korollar. Der Beweis findet sich bei M. Abramowitz und I. A. Stegun [1] und M. Sibuya [13].

Korollar 6.2.1. *(Eigenschaften der Beta-Verteilung)*
Sei $\psi(u) = \Gamma'(u)/\Gamma(u) = (\log \Gamma(u))'$, wobei Γ die Gamma-Funktion bezeichnet und sei des Weiteren γ die Eulerische Konstante, dann gilt:

$$\psi(n) = -\gamma + \sum_{i=0}^{n-1} \frac{1}{i} \text{ für natürliche } n \geq 2$$

und

$$\psi'(u) = \sum_{i=0}^{\infty} \frac{1}{(u+n)^2} \text{ für } u \in \mathbb{R}.$$

Ist nun X eine Beta(a,m)-verteilte Zufallsvariable, dann gilt:

$$E[X\log(1/X)] = \frac{a}{a+m}(\psi(a+1+m) - \psi(a+1))$$

und

$$E[X\log^2(X)] = \frac{a}{a+m}\left((\psi(a+1+m) - \psi(a+1))^2 - \psi'(a+1) + \psi'(a+1+m)\right).$$

V ist nach den obigen Überlegungen im Fall des b-nären Suchbaumes $Beta(1, b-1)$-verteilt. Damit ergeben sich die Werte $\mu = \sum_{i=2}^{b} 1/i$ und $\sigma^2 = \sum_{i=2}^{b} 1/i^2$. Ist nun mit $\mathcal{H}_n$ die n-te Harmonische Zahl bezeichnet, ergibt sich für den Abstand zweier zufälliger Kugeln im b-nären Suchbaum der folgende Grenzwertsatz:

$$\frac{\Delta_{I,J} - (1/(\mathcal{H}_b - 1))\log n}{(\mathcal{H}_b - 1)^{-3/2}\sqrt{\log n \sum_{i=2}^{b} 1/i^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ für } n \to \infty.$$

7 Diskussion: Ein anderer Zugang

In den Beweisen der Sätze 5.2.3 und 5.2.4 haben wir die Tiefen der Kugeln durch die Anzahlen der Kugeln in den entsprechenden Teilbäumen abgeschätzt und erhielten Ausdrücke der Form $\sum \log(1/W_i)$. Man könnte alternativ auch die Tiefen direkt berechnen und die Erneuerungstheorie als Zugang zum Modell wählen. D_n bzw. D'_n würden sich dabei als Eintrittszeiten des um einen messbaren Fehler verschobenen Prozesses $\sum \log(1/W_i)$ erweisen:

Sei ein zufälliger Splitbaum mit $n+1$ Kugeln gegeben und sei

$$\overline{B_{n,k}} :\stackrel{\mathcal{L}}{=} B(n - k(s_0 + (b-1)s_1) - s_1, \prod_{i=1}^{k} W_i).$$

Dann ist die Anzahl der Kugeln, die bis Stufe k den selben Pfad wie die $n+1$-te Kugel gehen, $\overline{B_{n,k}} + s_1$, denn:

1. Stufe: $B(n - s_0 - (b-1)s_1 - s_1, W_1) + s_1$
2. Stufe: $B(n - 2s_0 - 2(b-1)s_1 - s_1, W_1W_2) + s_1$
$\vdots$
k. Stufe: $B(n - ks_0 - k(b-1)s_1 - s_1, \prod_{i=1}^{k} W_i) + s_1$.

Damit ist

$$D_{n+1} \stackrel{\mathcal{L}}{=} \min\{k \geq 1 | \overline{B_{n,k}} + s_1 < s\},$$

da die Kugel in dem ersten Knoten landet, in dem von allen Kugeln, die den gleichen Weg wie die n-te gehen, bisher weniger als s Kugeln gelandet sind. Man kann zeigen, dass

$$D_{n+1} \stackrel{\mathcal{L}}{=} \min\{k \geq 1 | \sum_{i=1}^{k} \log(\frac{1}{W_i}) + R_{n,k} > \log n - \log s\}.$$

Dabei wäre $R_{n,k}$ der oben erwähnte Fehler, der hinreichend schnell stochastisch gegen 0 konvergiert. Damit könnte man auf Sätze der Erneuerungstheorie, wie bei G. Alsmeyer [2] zu finden, zurückgreifen, um das asymptotische Verhalten der Tiefen zu beschreiben.

Ähnlich wäre dann auch die Herangehensweise an den Abstand zweier Knoten im Baum:
Sei G eine Zufallsvariable und beschreibe die Wahrscheinlichkeit, die der Abzweigung des gemeinsamen Weges der n-ten und $n+1$-ten Kugel in Stufe 1 des Baumes zugeordnet ist. Das bedeutet:

$$P(G = v_j|\mathcal{V} = (v_1, \ldots, v_b))$$

$$= P(Kugeln\ gehen\ Weg\ j|\ Kugeln\ gehen\ gemeinsamen\ Weg, \mathcal{V} = (v_1, \ldots, v_b))$$

$$= \frac{v_j^2}{\sum\limits_{i=1}^{b} v_i^2}.$$

Seien nun $(G_i)_{i\geq 1}$ unabhängige Kopien von G. $W_0^{(1)}$, $W_0^{(2)}$ seien die Zufallsvariablen, die die Wahrscheinlichkeiten beschreiben, mit denen die n-te und $n+1$-te Kugeln die Pfade in Stufe $R+1$ passieren. Es gilt: $P(W_0^{(1)} = W_0^{(2)}) = 0$. Des Weiteren sollen $(W_i^{(1)})_{i\geq 1}$, $(W_i^{(2)})_{i\geq 1}$ unabhängige Folgen unabhängiger Zufallsvariablen sein. Ihre Verteilungen sollen der von $W = V_S$ entsprechen. Wenn man die letzten beiden Kugeln nicht berücksichtigt, ist die Anzahl der Kugeln im Teilbaum am Knoten u_R

$$N(u_R) \stackrel{\mathcal{L}}{=} B(n - Rs_0 - (R-1)(b-1)s_1 - bs_1 - 1, \prod_{i=1}^{R} G_i) + s_1$$

und

$$N(u_0) \stackrel{\mathcal{L}}{=} B(N(u_R) - s_0 - bs_1, W_0^{(1)}) + s_1,$$
$$N(v_0) \stackrel{\mathcal{L}}{=} B(N(u_R) - s_0 - bs_1, W_0^{(2)}) + s_1.$$

Damit gilt für die Tiefen der Teilbäume:

$$D'_n \stackrel{\mathcal{L}}{=} \min\{k \geq 0 | B(N(u_0) - k(s_0 - (b-1)s_1) - s_1, \prod_{i=1}^{k} W_i^{(1)}) + s_1 < s\}$$

und

$$D'_{n+1} \stackrel{\mathcal{L}}{=} \min\{k \geq 0 | B(N(v_0) - k(s_0 - (b-1)s_1) - s_1, \prod_{i=1}^{k} W_i^{(2)}) + s_1 < s\},$$

was uns schließlich wieder auf Erneuerungsprozesse mit Erneuerungspunkten $\sum \log(1/W_i)$ führt.

A Anhang: Beweis des Satzes zur Höhe des zufälligen Splitbaumes

Wir wollen einige Eigenschaften von der Momentenfunktion $m(t)$, von M und von $t^*(d)$ referieren, um auf diese beim Beweis des Satzes 5.2.1 zurückgreifen zu können. Das folgende Korollar ist der Arbeit von L. Devroye [4] entnommen.

Korollar A.0.1. *Eigenschaften der Momentenfunktion*

a) Die Funktion $m(t)$ ist von $m(0) = 1$ bis $P(V = 1)$ monoton fallend.

b) Eine Lösung $t^* = t^*(d)$ der Gleichung $m(t)'/m(t) = -1/d$ existiert, falls $-1/E[\log V] < d < -1/v_\infty$, wobei v_∞ der äußerste rechte Punkt des Trägers von V ist.

c) $t^*/d + \log m(t^*)$ ist monoton fallend in d und nimmt Werte zwischen 0 und M an. $M = -\infty$ ist dabei zugelassen.

Bemerkung. Bei uns ist stets $P(V = 1) = 0$ erfüllt, die Momentenfunktion fällt also von 1 auf 0.

Satz 5.2.1. *Ist V ein Splitter eines zufälligen Splitbaumes und ist $P(V = 1) = 0$, dann existiert eine Konstante c, so dass*

$$\lim_{n\to\infty} P(H_n > c \log n) = 0.$$

Sei $m(t) := E[V^t]$ die Momentenfunktion von V, $M := \lim_{t\to\infty} \log m(t) - tm(t)'/m(t)$. Ist dann $M < -\log b$, so ist die obige Aussage wahr für alle $c > \gamma$. $\gamma \in (0, \infty)$ ist ein lediglich von b und der Verteilung von V abhängiger Parameter, der über

$$\gamma = \inf\{d : \exp(t^*(d))(bm(t^*(d)))^d < 1\}$$

definiert ist. $t^*(d)$ ist dabei Lösung der Gleichung $m(t)'/m(t) = -1/d$.

Beweis: Wir gehen von einem Baum mit n Kugeln aus und werden zeigen, dass $P(H_n \geq (c + 3\varepsilon) \log n) \to 0$ für $n \to \infty$, beliebiges $\varepsilon > 0$ und passendes c. Sei dazu $c \in \mathbb{R}^+$ zunächst beliebig, $k = \lfloor c \log n \rfloor$, $k' = \lfloor \varepsilon \log n \rfloor$ und $l = k'(s_1 + 1)$. Sei weiter
$Z_k := B(n, \prod_{i=1}^{k} V^{(i)}) + B(s_1, \prod_{i=2}^{k} V^{(i)}) + B(s_1, \prod_{i=3}^{k} V^{(i)}) + \ldots + B(s_1, V^{(k)}) + s_1$.
Die $V^{(i)}$, $1 \leq i \leq k$ sollen dabei unabhängig und identisch, wie V verteilt sein. Sie entsprechen den Wahrscheinlichkeiten der einzelnen Kanten eines Pfades der Länge k. (Wir gehen davon aus, dass die Komponenten des Vektors $\mathcal{V}$ identisch und splitter-verteilt sind.)

Für einen Knoten u auf Stufe k des Baumes läßt sich $N(u)$, die Anzahl der Kugeln im zugehörigen Teilbaum, durch Z_k abschätzen, denn: Bei Z_k nehmen wir

an, dass alle n Kugeln von der Wurzel aus nur den Pfad zum Knoten u zurücklegen müssen und dann in dem betrachteten Teilbaum ankommen können. Das entspricht dem ersten Term von Z_k. Außerdem sind in jedem Knoten zwischen Wurzel und u irgendwann s_1 Kugeln platziert worden. Auch diese Kugeln lassen wir die entsprechenden kürzeren Pfade zu Knoten u zurücklegen. Das entspricht den anderen „binomialverteilten“ Termen von Z_k. Beim Aufbrechen des Elternknotens von u sind genau s_1 Kugeln in u platziert worden. Diese Anzahl stellt den letzten Term von Z_k dar. Im stochastischen Sinne gilt also:

$$N(u) \leq Z_k$$

Wir schätzen nun Z_k ab: Die $k - k'$ Produkte $\prod_{i=2}^{k} V^{(i)}, \prod_{i=3}^{k} V^{(i)}, \ldots, \prod_{i=k-k'+1}^{k} V^{(i)}$ sind stochastisch kleiner oder gleich dem Produkt $\prod_{i=k-k'+1}^{k} V^{(i)}$.

Die $s_1(k'-1)$ Variablen $B(s_1, \prod_{i=k-k'+2}^{k} V^{(i)}), B(s_1, \prod_{i=k-k'+3}^{k} V^{(i)}), \ldots, B(s_1, V^{(k)})$ nehmen Werte kleiner oder gleich s_1 an. Damit folgt:

$$N(u) \leq Z_k \leq B(n, \prod_{i=1}^{k} V^{(i)}) + B(s_1(k-k'+1), \prod_{i=k-k'+1}^{k} V^{(i)}) + k's_1.$$

Es gibt insgesammt b^k Knoten auf Stufe k. Da die zu den b^k Pfaden gehörigen Folgen von Splittern identisch verteilt sind, folgt mit der Subadditivität:

$$P(H_n \geq k + 3l) \leq b^k P(Z_k \geq 3l).$$

Damit $\{Z_k \geq 3l\}$ eintritt, muss mindestens eines der Ereignisse $\{B(n, \prod_{i=1}^{k} V^{(i)}) \geq l\}$, $\{B(s_1(k-k'+1), \prod_{i=k-k'+1}^{k} V^{(i)}) \geq l\}$ und $\{k's_1 \geq l\}$ eintreten. Damit ist die folgende Ungleichung erfüllt:

$$\begin{aligned} & P(Z_k \geq 3l) \\ \leq\ & P(B(n, \prod_{i=1}^{k} V^{(i)}) \geq l) \\ & + P\left(B(s_1(k-k'+1), \prod_{i=k-k'+1}^{k} V^{(i)}) \geq l\right) \\ & + P(k's_1 \geq l). \end{aligned}$$

Nach der Wahl von $l = k'(s_1 + 1)$ ist der letzte Term 0.

Für eine binomialverteilte Zufallsvariable $B(n, p)$ und $t, s > 0$ gilt nach der Markov-Ungleichung:

$P(B(n,p) > s) \leq E[\exp(tB(n,p))]\exp(-ts) = (1-p+p\exp(t))^n \exp(-ts)$.
Sei $Z := \prod_{i=1}^{k} V^{(i)}$, dann gilt:

$$\begin{aligned} & P(B(n,Z) \geq l|Z) \\ \leq\ & E[(1-Z+Ze^t)^n|Z]e^{-tl} \\ \leq\ & E[(1+\tfrac{nZ(e^t-1)}{n})^n|Z]e^{-tl} \\ \leq\ & E[\exp(nZ(e^t-1))|Z]e^{-tl}. \end{aligned}$$

Wählen wir nun t so, dass $\exp(t) = l/(nZ)$, dann folgt:

$$\begin{aligned} & P(B(n,Z) \geq l|Z) \\ \leq\ & E[\exp(l(1+\log(nZ)-\log l)-nZ)|Z] \\ \leq\ & E[\exp(l-nZ+l\log(nZ))|Z]. \end{aligned}$$

Damit errechnen wir für die gesuchte Abschätzung:

$$\begin{aligned} & P(B(n,Z) \geq l) \\ =\ & \int_0^1 P(B(n,Z) \geq l|Z=z)dP_Z(z) \\ \leq\ & \int_0^1 E[\exp(l-nZ+l\log(nZ))|Z=z]dP_Z(z) \\ =\ & \int_0^1 \exp(l-nz+l\log(nz))dP_Z(z) \\ =\ & \int\limits_{\{nz\leq z^*\}} \exp(l-nz+l\log(nz))dP_Z(z) \\ & + \int\limits_{\{nz>z^*\}} \exp(l-nz+l\log(nz))dP_Z(z) \text{ für ein } z^* \in (0,1), \\ \leq\ & \exp(l+l\log(z^*)) + P(nZ > z^*) \text{ für } n \text{ hinreichend groß}, \\ \leq\ & (ez^*)^l + E[Z^t](n/z^*)^t \end{aligned}$$

nach der Markov-Ungleichung und für nun wieder ein beliebiges $t > 0$.

Wir schätzen analog den Term $P(B(s_1(k-k'+1), \prod_{i=k-k'+1}^{k} V^{(i)}) \geq l)$ ab:

$$\begin{aligned} & P(B(s_1(k-k'+1), \prod_{i=k-k'+1}^{k} V^{(i)}) \geq l) \\ \leq \; & (ez^{**})^l + E[(Z')^{t'}](s_1(k-k'+1)/z^{**})^{t'}, \end{aligned}$$

wobei hier $Z' = \prod_{i=k-k'+1}^{k} V^{(i)}$, $z^{**} \in (0,1)$ und $t' > 0$.

Nach der Wahl $z^* = z^{**} = b^{-2k/l}/e$ setzen wir nun die zwei Abschätzungen zusammen und erhalten:

$$\begin{aligned} & b^k P(Z_k \geq 3l) \\ \leq \; & b^k (ez^*)^l + b^k E[Z^t](n/z^*)^t + b^k (ez^{**})^l + b^k E[(Z')^{t'}](s_1(k-k'+1)/z^{**})^{t'} \\ = \; & b^k b^{-2k} + b^k E[(V^k)^t](nb^{2k/l}e)^t + b^k b^{-2k} + b^k (s_1(k-k'+1)b^{2k/l}e)^{t'} E[(V^{k'})^{t'}] \\ = \; & 2b^{-k} + b^k m(t)^k (nb^{2k/l}e)^t + b^k (s_1(k-k'+1)b^{2k/l}e)^{t'} m(t')^{k'} \end{aligned}$$

Der erste Summand ist für wachsendes n eine Nullfolge.

Nach der obigen Wahl von k und l entspricht k/l etwa $c/(\varepsilon(s_1+1))$ und k'/k etwa ε/c. Damit ist der dritte Term der Abschätzung von der Größenordnung

$$(s_1((c-\varepsilon)\log n + 1)b^{2c/(\varepsilon(s_1+1))}e)^{t'}(bm(t')^{\varepsilon/c})^k \text{ für festes } t'.$$

Damit dieser Ausdruck ebenfalls eine Nullfolge darstellt, wählen wir t' so groß, dass $bm(t')^{\varepsilon/c} < 1$ erfüllt ist. Dies ist nach Korollar A.0.1 a) möglich, da wir $P(V=1) = 0$ gefordert haben.

Der zweite Term entspricht dem Ausdruck

$$(b^{2c/(\varepsilon(s_1+1))}e)^t n^t (bm(t))^k \text{ für festes } t.$$

Das bedeutet, dass wir, um die Aussagen des Satzes zu beweisen, t und c so wählen müssen, dass $n^t(bm(t))^{c\log n} = (e^t(bm(t))^c)^{\log n} \to 0$ für $n \to \infty$ erfüllt ist. Für die erste Aussage des Satzes wählen wir t so groß, dass $bm(t) < 1$ gilt. Dann ist noch c hinreichend groß zu wählen, so dass auch $e^t(bm(t))^c < 1$ erfüllt ist.

Für die zweite Aussage des Satzes ist es notwendig die Bedingung

$$e^t(bm(t))^c < 1 \iff c\log(bm(t)) + t < 0 \iff \log(m(t)) + \frac{t}{c} < -\log b$$

genauer zu betrachten. Durch eine einfache Extremwertberechnung ergibt sich, dass für festes c der Ausdruck $c\log(bm(t)) + t$ minimal für t^*, die Lösung der

Gleichung $m'(t)/m(t) = -1/c$, ist. Korollar A.0.1 b) liefert ein Kriterium, wann solche Lösungen existieren.

Wenn wir nun t durch $t^*(c)$ ersetzen und $c > \gamma$ wählen, ist durch die Bedingung $M < -\log b$, die Definition von γ und Korollar A.0.1 c) sichergestellt, dass die Bedingung $\log(m(t^*)) + \frac{t^*}{c} < -\log b$ erfüllt ist. Damit folgt die Behauptung.

□

Literatur

[1] M. Abramowitz und I. A. Stegun. (1970) *Handbook of mathematical tables*. Dover Publications.

[2] G.Alsmeyer. (1991) *Erneuerungstheorie*. B.G.Teubner Stuttgart.

[3] R. Arratia, A. D. Barbour, und S. Tavaré. (2003) *Logarithmic combinatorial structures: a probabilistic approach*. European Mathematic Society Publishing House.

[4] L. Devroye. (1998) Universal Limit Laws for depth in random trees. *SIAM J. Comput.* 28, 409-432.

[5] L. Devroye. (1987) Branching processes in the analysis of the heights of trees. *Acta Inform.* 26, 272-298.

[6] L. Devroye und R. Neininger. (2004) Distances and finger search in random binary search trees. *SIAM J. Comput.* 3, 647-658.

[7] O. Forster. (2001) *Analysis 1*. Vieweg Verlag.

[8] G. Kersting. (2003) *Zufallsvariablen und Wahrscheinlichkeiten. Eine elementare Einführung in die Stochastik*. Skript zur Vorlesung.

[9] A. Klenke. (2006) *Wahrscheinlichkeitstheorie*. Springer Verlag Berlin Heidelberg.

[10] H. M. Mahmoud. (1992) *Evolution of random search trees*. John Wiley & sons Inc.

[11] H. M. Mahmoud und R. Neininger. (2003) Distributions of distances in random binary search trees. *Ann. Appl. Probab.* 13, 253-276.

[12] M. Okamoto. (1958) Some inequalities relating to the partial sum of binomial probabilities. *Ann. Math. Statist.* 10, 29-35.

[13] M. Sibuya. (1979) Generalized hypergeometric, digamma and trigamma distributions. *Ann. Inst. Statist. Math.* 31, 373-390.