

Stochastische Modelle der Populationsbiologie:
Notizen zu Koaleszenten mit multiplen Verschmelzungen

Notizen zu einer Vorlesung an der
Johannes-Gutenberg-Universität Mainz, Winter 2022/23

Matthias Birkner

Vorläufige Version, 17. Januar 2023

Kommentare, Korrekturvorschläge, Hinweise auf (Tipp-)Fehler gerne per Email an
`birkner@mathematik.uni-mainz.de` senden

Inhaltsverzeichnis

3	Koaleszenten mit multiplen Verschmelzungen	2
3.1	Zur Asymptotik des Blockzählprozesses von Λ -Koaleszenten	14

Kapitel 3

Koaleszenten mit multiplen Verschmelzungen

Zur Notation: $f(N) \sim g(N)$ für $N \rightarrow \infty$ bedeutet $\lim_{N \rightarrow \infty} f(N)/g(N) = 1$; $f(N) = O(g(N))$ bedeutet $\limsup_{N \rightarrow \infty} f(N)/g(N) < \infty$ (d.h. es gibt eine Konstante $c < \infty$, so dass $f(N) \leq cg(N)$ für alle N); $f(N) = o(g(N))$ bedeutet $\lim_{N \rightarrow \infty} |f(N)/g(N)| = 0$.

Erinnerung. Wir haben in der Vorlesung Cannings-Modelle und deren Genealogien betrachtet: Für $N \in \mathbb{N}$ sei $\nu^{(N)} = (\nu_1^{(N)}, \dots, \nu_N^{(N)})$ ein austauschbarer Nachkommensanzahlvektor mit $\nu_1^{(N)} + \dots + \nu_N^{(N)} = N$ (der die zufällige Anzahl Nachkommen der Individuen einer Generation beschreibt). Die Genealogie einer n -Stichprobe kodieren wir mittels der Ahnen-Äquivalenzrelation: Stichproben i und j befinden sich in derselben Äquivalenzklassen von $R_g^{(N,n)}$, wenn sie einen gemeinsamen Vorfahren g Generationen in der Vergangenheit besitzen.

Entscheidende Größen sind die Paarverschmelzungswahrscheinlichkeit

$$c_N = \frac{1}{N(N-1)} \sum_{i=1}^N \mathbb{E}[\nu_i^{(N)}(\nu_i^{(N)} - 1)] = \frac{1}{N-1} \mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)} - 1)]$$

und die Tripelverschmelzungswahrscheinlichkeit

$$d_N = \frac{\binom{N}{1\downarrow} \mathbb{E}[(\nu_1^{(N)})_{3\downarrow}]}{\binom{N}{3\downarrow}} = \frac{1}{(N-1)(N-2)} \mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)} - 1)(\nu_1^{(N)} - 2)]$$

Wir haben gesehen: Wenn

$$c_N \rightarrow 0 \quad \text{und} \quad \frac{d_N}{c_N} \rightarrow 0 \quad \text{für } N \rightarrow \infty$$

gilt, so konvergieren die Ahnenverhältnisse einer n -Stichprobe in dieser Schar von Cannings-Modellen basierend auf den (Verteilungen der) Nachkommensvektoren $\nu^{(N)}$ in Verteilung gegen Kingmans n -Koaleszenten, wenn die Zeit in Vielfachen von $1/c_N$ gemessen wird. (Vergleiche z.B. Übungsaufgabe 4.1 oder Satz 1.8 des Skripts [SMPB16].)

In diesem Kapitel betrachten wir die Situation, in der diese Bedingungen verletzt sind, speziell $d_N/c_N \not\rightarrow 0$. Dies kann für Populationen relevant sein, in denen es (z.B. wegen sehr variabler Umweltbedingungen im Zusammenhang mit prinzipiell hohem Reproduktionspotential) gelegentlich sehr hohe Variabilität unter den Nachkommenszahlen gibt.

Mathematisch führt es auf die Theorie der Koaleszenten mit multiplen Verschmelzungen, sogenannte Λ -Koaleszenten, die 1999 unabhängig von Peter Donnelly und Tom Kurtz [DK99], von Jim Pitman [P99] und von Serik Sagitov [S99] eingeführt wurden.

Wir beginnen mit einer Beispielklasse, die das Modell aus Übungsaufgabe 4.2 b) verallgemeinert und die sich ebenso wie jene Übungsaufgabe an dem Artikel [EW06] von Bjarki Eldon und John Wakeley orientiert.

Beobachtung 3.1 (Ein verallgemeinertes Eldon-Wakeley-Modell). $0 < \psi_1 < \dots < \psi_m$, $w_1, \dots, w_m > 0$ mit $w_1 + \dots + w_m = 1$, weiter sei $(\tilde{\nu}_1^{(N)}, \dots, \tilde{\nu}_1^{(N)}) \sim \text{Multinom}(N; 1/N, \dots, 1/N)$.

Im N -ten Populationsmodell sei $\nu^{(N)} = (\nu_1^{(N)}, \dots, \nu_N^{(N)})$

$$(\tilde{\nu}_1^{(N)}, \dots, \tilde{\nu}_1^{(N)}) \quad \text{mit W'keit } 1 - N^{-\gamma}$$

und eine rein zufällige Permutation des folgenden Vektors

$$(\underbrace{[\psi_j N], 0, 0, 0, \dots, 0}_{([\psi_j N]-1)\text{-mal}}, \underbrace{1, 1, 1, 1, 1, \dots, 1}_{(N-[\psi_j N])\text{-mal}}) \quad \text{mit W'keit } w_j N^{-\gamma}, \quad j = 1, \dots, m$$

wobei $0 < \gamma < 1$.

Es ist

$$\begin{aligned} c_N &= \frac{1}{N-1} \mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)} - 1)] = (1 - N^{-\gamma}) \frac{1}{N} + \sum_{j=1}^m w_j N^{-\gamma} \cdot \frac{1}{N} \cdot \frac{1}{N-1} [\psi_j N] ([\psi_j N] - 1) \\ &\sim N^{-\gamma} \sum_{j=1}^m w_j \psi_j^2 \quad \text{für } N \rightarrow \infty \end{aligned}$$

(bzw. $c_N \sim N^{-1}(1 + \sum_{j=1}^m w_j \psi_j^2)$ falls $\gamma = 1$) und

$$\begin{aligned} d_N &= \mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)} - 1)(\nu_1^{(N)} - 2)] \\ &= (1 - N^{-\gamma}) \frac{1}{N^2} + \sum_{j=1}^m w_j N^{-\gamma} \cdot \frac{1}{N} \cdot \frac{1}{(N-1)(N-2)} [\psi_j N] ([\psi_j N] - 1) ([\psi_j N] - 2) \\ &\sim N^{-\gamma} \sum_{j=1}^m w_j \psi_j^3 \end{aligned}$$

Insbesondere konvergiert d_N/c_N nicht nach 0.

Betrachte n -Stichprobe, darin eine vorgegebene k -elementige Teilmenge, $k \geq 2$:

$$\begin{aligned}
p_{n,k}^{(N)} &:= \text{W'keit, dass genau diese } k \text{ der } n \text{ Stichproben in der vorigen Generation einen} \\
&\quad \text{gemeinsamen Vorfahren finden} \\
&= \left(1 - N^{-\gamma}\right) \frac{1}{N^{k-1}} \left(1 - \frac{1}{N}\right) \left(1 - \frac{2}{N}\right) \cdots \left(1 - \frac{n-k}{N}\right) \\
&\quad + \sum_{j=1}^m w_j N^{-\gamma} \cdot \frac{(\lfloor \psi_j N \rfloor)_{k\downarrow} (N - \lfloor \psi_j N \rfloor)_{(n-k)\downarrow}}{(N)_{n\downarrow}} \\
&\sim N^{-\gamma} \sum_{j=1}^m w_j \psi_j^k (1 - \psi_j)^{n-k} \quad \text{für } N \rightarrow \infty
\end{aligned}$$

somit

$$\frac{p_{n,k}^{(N)}}{c_N} \rightarrow \frac{\sum_{j=1}^m w_j \psi_j^k (1 - \psi_j)^{n-k}}{\sum_{j=1}^m w_j \psi_j^2} = \sum_{j=1}^m \psi_j^{k-2} (1 - \psi_j)^{n-k} \frac{w_j \psi_j^2}{\sum_{\ell=1}^m w_\ell \psi_\ell^2} = \int_{[0,1]} x^{k-2} (1-x)^{n-k} \Lambda(dx)$$

mit

$$\Lambda = \sum_{j=1}^m \frac{w_j \psi_j^2}{\sum_{\ell=1}^m w_\ell \psi_\ell^2} \delta_{\psi_j}$$

Die allgemeine Situation sieht dann folgendermaßen aus:

Definition 3.2 (Λ -(n -)Koaleszent). Sei Λ endliches (positives) Maß auf $[0, 1]$,

$$q_{\xi\eta} = \begin{cases} \int_{[0,1]} x^{j-2} (1-x)^{b-j} \Lambda(dx) & |\xi| = b \text{ und } \eta \text{ entsteht aus } \xi \text{ durch Ver-} \\ & \text{schmelzen von } j \text{ versch. Klassen in eine} \\ & \text{einzige (insbes. } \xi \leq \eta, |\eta| = b - j + 1) \\ - \sum_{j=2}^b \binom{b}{j} \int_{[0,1]} x^{j-2} (1-x)^{b-j} \Lambda(dx) & \eta = \xi \text{ und } |\xi| = b \\ = - \int_{[0,1]} \frac{1 - (1-x)^b - bx(1-x)^{b-1}}{x^2} \Lambda(dx) & \\ 0 & \text{sonst} \end{cases}$$

Die zeitkont. Markovkette $(\mathcal{R}_t^{(n)})_{t \geq 0}$ auf $\mathcal{E}_n$ mit Sprungratenmatrix $(q_{\xi\eta})$ heißt Λ -(n -)Koaleszent. (Typischerweise denkt man an $\mathcal{R}_0^{(n)} = \{\{1\}, \dots, \{n\}\}$ zur Modellierung einer n -Stichprobe.)

Bemerkung. 1. Die Wahl $\Lambda = \delta_0$ liefert den Kingman-Koaleszenten.

2. Die Normierung mit $1/c_N$ wie in Beob. 3.1 führt automatisch im Limes auf eine Paarverschmelzungsrate von 1 und damit auf ein W'maß Λ . Im allgemeinen Fall $\Lambda([0, 1]) \neq 1$ kann man durch Zeittransformation zu einem W'maß $\Lambda'(\cdot) = \Lambda(\cdot)/\Lambda([0, 1])$ übergehen.

3. Analog zum Fall des Kingman-Koaleszenten kann man die Realisierungen von $(\mathcal{R}_t^{(n)})_{t \geq 0}$ als gewurzelte Bäume mit n Blättern auffassen. I.A. sind diese Bäume *nicht binär*.

[Skizze an der Tafel]

Bemerkung 3.3 (Münzwurf-Interpretation der Übergangsraten). Wenn es aktuell b Linien gibt, tritt eine gegebene j -Verschmelzung mit Rate

$$\lambda_{b,j} := \int_{[0,1]} x^j (1-x)^{b-j} \frac{\Lambda(dx)}{x^2}$$

auf. Intuitiv also: mit Rate $\frac{1}{x^2} \Lambda(dx)$ finden “ x -Verschmelzungen” statt (bzw. “vorwärts” gedacht: “ x -Reproduktionsereignisse”), jede Ahnenlinie nimmt u.a. mit W’keit x daran teil (in der „Vorwärts“-Interpretation: ein Anteil x der aktuellen Population wird durch die Nachkommen eines Individuums ersetzt). Siehe dazu auch die Poissonprozess-Konstruktion aus Übungsaufgabe 8.1.

Satz 3.4 (Sagitov¹). Sei Λ ein W’maß auf $[0, 1]$ (o.E., sonst transformiere Zeit mit $1/\Lambda([0, 1])$), es gelte für eine Schar von Cannings-Modellen mit $N \rightarrow \infty$

1) $c_N \rightarrow 0$,

2) $\frac{N}{c_N} \mathbb{P}(\nu_1^{(N)} > Nx) \rightarrow \int_{(x,1]} y^{-2} \Lambda(dy)$ in allen Stetigkeitspunkten x von Λ ,

3) $\frac{1}{N^2 c_N} \mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)} - 1) \nu_2^{(N)}(\nu_2^{(N)} - 1)] \rightarrow 0$ (in der Notation aus (3.1) also $\Phi_2^{(N)}(2, 2)/c_N \rightarrow 0$).

Dann gilt $(\mathcal{R}_{\lfloor t/c_N \rfloor}^{(N,n)})_{t \geq 0} \rightarrow (\mathcal{R}_t^{(n)})_{t \geq 0}$ in Verteilung (auf $D([0, \infty); \mathcal{E}_n)$).

Bemerkung. 1. Dies verallgemeinert den Satz über die Konvergenz gegen den Kingman-Koaleszenten (aus Übungsaufgabe 4.1 bzw. Satz 1.8 aus [SMPB16]): Nehmen wir an, dass $d_N/c_N \rightarrow 0$ gilt. Für $\varepsilon \in (0, 1)$ ist dann

$$\mathbb{P}(\nu_1^{(N)} > N\varepsilon) \leq \frac{1}{(N\varepsilon)^3} \mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)} - 1)(\nu_1^{(N)} - 2)](1 + o(1)) \sim \varepsilon^{-3}(1/N)d_N$$

so dass $(N/c_N) \mathbb{P}(\nu_1 > N\varepsilon) \rightarrow 0$ folgt und Voraussetzung 2) in Satz 3.4 für $\Lambda = \delta_0$ erfüllt ist.

2.) Die Voraussetzungen 1)–3) in Satz 3.4 sind auch notwendig für die Konvergenz der Stichproben-Genealogie gegen den Λ - n -Koaleszenten, siehe [S99]. Speziell verhindert 3), dass es im Grenzübjekt simultane Verschmelzungen von 2 oder mehr Gruppen geben kann.

Sei $a \geq 1, b_1, \dots, b_a \geq 1, b := b_1 + \dots + b_a$

$$\Phi_a^{(N)}(b_1, \dots, b_a) := \frac{(N)_{a\downarrow}}{(N)_{b\downarrow}} \mathbb{E} \left[\prod_{j=1}^a (\nu_j^{(N)})_{b_j\downarrow} \right] \quad (3.1)$$

Dies beschreibt die Wahrscheinlichkeit, dass über eine Generation b Ahnenlinien in a Gruppen der Größen $b_1, \dots, b_a$ verschmelzen. Austauschbarkeit der $\nu_j^{(N)}$ zeigt: $\Phi_a^{(N)}(b_1, \dots, b_a)$ ist invariant unter Permutation der b_i .

In der Notation unterdrücken wir im Folgenden Abhängigkeit von N , d.h. wir schreiben $\Phi_a(b_1, \dots, b_a) := \Phi_a^{(N)}(b_1, \dots, b_a)$.

¹siehe [S99]

Lemma 3.5 (vgl. Möhle-Sagitov [MS01], Lemma 3.3). *Es gilt*

$$\Phi_{a+1}(b_1, \dots, b_a, 1) = \Phi_a(b_1, \dots, b_a) - \sum_{i=1}^a \Phi_a(b_1, \dots, b_{i-1}, b_i + 1, b_{i+1}, b_a) \quad (3.2)$$

und für $a' \geq a$, $b'_i \geq b_i$ ($1 \leq i \leq a$) gilt

$$\Phi_{a'}(b'_1, \dots, b'_{a'}) \leq \Phi_a(b_1, \dots, b_a) \quad (3.3)$$

Bemerkung. Formel (3.2) entspricht der „Stichproben-Konsistenz“: lasse im Geiste die $b + 1$ -te Ahnenlinie weg und zerlege nach den Möglichkeiten, wie sie in das Verschmelzungsmuster der übrigen b Linien eingepasst wird.

Ungleichung (3.3) ist angesichts der Stichproben-Interpretation sehr plausibel: Das Weglassen gewisser Linien verlangt offensichtlich weniger Einschränkungen.

Beweis. Mit Austauschbarkeit gilt

$$\begin{aligned} & (N - a) \mathbb{E} \left[(\nu_1^{(N)})_{b_1 \downarrow} \cdots (\nu_a^{(N)})_{b_a \downarrow} \nu_{a+1}^{(N)} \right] \\ &= \mathbb{E} \left[(\nu_1^{(N)})_{b_1 \downarrow} \cdots (\nu_a^{(N)})_{b_a \downarrow} \underbrace{(\nu_{a+1}^{(N)} + \nu_{a+2}^{(N)} + \cdots + \nu_N^{(N)})}_{\substack{= N - \nu_1^{(N)} - \cdots - \nu_a^{(N)} \\ = N - b - (\nu_1^{(N)} - b_1) - \cdots - (\nu_a^{(N)} - b_a)}} \right] \\ &= (N - b) \mathbb{E} \left[(\nu_1^{(N)})_{b_1 \downarrow} \cdots (\nu_a^{(N)})_{b_a \downarrow} \nu_{a+1}^{(N)} \right] - \sum_{i=1}^a \mathbb{E} \left[(\nu_1^{(N)})_{b_1 \downarrow} \cdots (\nu_i^{(N)})_{(b_i+1) \downarrow} \cdots (\nu_a^{(N)})_{b_a \downarrow} \nu_{a+1}^{(N)} \right] \end{aligned}$$

multipliziere mit $(N)_{a \downarrow} / (N)_{b \downarrow}$ und erhalte (3.2).

Zu (3.3): (3.2) zeigt

$$\begin{aligned} \Phi_a(b_1, \dots, b_a) &\geq \Phi_{a+1}(b_1, \dots, b_a, 1) \quad \text{und} \\ &\geq \max_{i=1}^a \Phi_a(b_1, \dots, b_{i-1}, b_i + 1, b_{i+1}, b_a), \end{aligned}$$

iteriere die erste Ungleichung $a' - a$ mal und die zweite Ungleichung $\sum_1^a b'_i - b_i + \sum_{a+1}^{a'} b'_i - 1$ mal geeignet. $\square$

Beweis von Satz 3.4. Zu zeigen ist

$$p^{(N,n)}(\xi, \eta) = \delta_{\xi, \eta} + c_N q_{\xi \eta} + o(c_N) \quad \text{für } \xi, \eta \in \mathcal{E}_n \quad (3.4)$$

Die Behauptung folgt dann z.B. mit [SMPB16, Lemma 1.6] (vgl. auch Übungsaufgabe 4.1) für Konvergenz im Sinne der endlich-dimensionalen Verteilungen und z.B. mit Übungsaufgabe 7.3 für die Verstärkung zur Konvergenz in Verteilung auf dem Raum der càdlàg-Pfade. (Beachte: Da es zu gegebener Stichprobengröße n nur endlich viele Wahlen von $\xi, \eta \in \mathcal{E}_n$ gibt, ist dann der Fehler in (3.4) tatsächlich uniform in ξ, η klein.)

Voraussetzung 3) und (3.3) aus Lemma 3.5 zeigen: für $b_1, \dots, b_a$ mit mindestens 2 der $b_i \geq 2$ ist

$$(0 \leq) \frac{1}{c_N} \Phi_a(b_1, \dots, b_a) \leq \frac{1}{c_N} \Phi_2(2, 2) \rightarrow 0, \quad (3.5)$$

d.h. $\frac{1}{c_N} p^{(N,n)}(\xi, \eta) \rightarrow 0 = q_{\xi\eta}$, wenn η aus ξ durch Verschmelzung von mind. zwei verschiedenen Gruppen von Äquivalenzklassen entsteht (oder wenn $\xi \neq \eta$)

Voraussetzung 2) ist äquivalent zu

$$\frac{N}{c_N} \mathcal{L}\left(\frac{\nu_1^{(N)}}{N}\right) \xrightarrow{v} y^{-2} \Lambda(dy) \quad \text{auf } (0, 1]$$

(vage Konvergenz von Maßen auf dem lokal kompakten Raum $(0, 1]$), d.h. für stetiges $f : (0, 1] \rightarrow \mathbb{R}$ mit kompaktem Träger gilt

$$\frac{N}{c_N} \mathbb{E}[f(\nu_1^{(N)}/N)] \xrightarrow{N \rightarrow \infty} \int_{[0,1]} f(y)/y^2 \Lambda(dy). \quad (3.6)$$

Sei nun $f : [0, 1] \rightarrow \mathbb{R}$ beschränkt und stetig mit $f(y) = O(y^2)$ für $y \rightarrow 0$. Ein solches f kann man gleichmäßig mit stetigen Funktionen f_m approximieren, die kompakten Träger in $(0, 1]$ haben. Insbesondere gilt dann $\sup_m \sup_{y \in (0,1]} |f_m(y)/y^2| < \infty$ und dominierte Konvergenz zeigt, dass (3.6) auch für dieses f gilt.

Somit gilt für $j \geq 2$

$$\frac{1}{c_N} \Phi_1(j) = \frac{1}{c_N} \frac{N}{(N)_{j\downarrow}} \mathbb{E}[(\nu_1^{(N)})_{j\downarrow}] \sim \frac{N}{c_N} \mathbb{E}[(\nu_1^{(N)}/N)^j] \xrightarrow{N \rightarrow \infty} \int_{[0,1]} y^{j-2} \Lambda(dy).$$

Formel (3.2) in Lemma 3.5 liefert dann für $j \geq 2, a \geq 2$ (wir verwenden in der zweiten Zeile auch (3.3))

$$\begin{aligned} \Phi_a(j, \underbrace{1, \dots, 1}_{a-1}) &= \Phi_{a-1}(j, \underbrace{1, \dots, 1}_{a-2}) - \Phi_{a-1}(j+1, \underbrace{1, \dots, 1}_{a-2}) - (a-2) \Phi_{a-1}(j, 2, \underbrace{1, \dots, 1}_{a-3}) \\ &= \Phi_{a-1}(j, \underbrace{1, \dots, 1}_{a-2}) - \Phi_{a-1}(j+1, \underbrace{1, \dots, 1}_{a-2}) + O(\Phi_2(2, 2)) \\ &= \dots = \sum_{\ell=0}^{a-1} \binom{a-1}{\ell} (-1)^\ell \Phi_1(j+\ell) + O(\Phi_2(2, 2)) \end{aligned} \quad (3.7)$$

(Diese Rekursion entspricht gewissermaßen einem „signierten Pascal-Dreieck“, mit

$$\Phi_a(j) := \Phi_a(j, \underbrace{1, \dots, 1}_{a-1}) \cong x^j (1-x)^{a-1}$$

insbes. mit $\Phi_1(j) \cong x^j$ ist $\Phi_a(j) \cong \sum_{\ell=0}^{a-1} \binom{a-1}{\ell} (-1)^\ell x^{j+\ell} \cong \Phi_1(j+\ell)$.

Man kann (3.7) beispielsweise induktiv (über a) beweisen. Zeige: $\tilde{\Phi}_a(j)$ erfülle

$$\tilde{\Phi}_a(j) = \tilde{\Phi}_{a-1}(j) - \tilde{\Phi}_{a-1}(j+1)$$

dann gilt für $a \geq 2$:

$$\tilde{\Phi}_a(j) = \sum_{\ell=0}^{a-1} \binom{a-1}{\ell} (-1)^\ell \tilde{\Phi}_1(j+\ell)$$

Für $a = 2$ ist die Formel offensichtlich korrekt. Ind.ann. liefert dann für $a > 2$: $\tilde{\Phi}_{a-1}(j) - \tilde{\Phi}_{a-1}(j+1) = \sum_{\ell=0}^{a-2} (-1)^\ell \binom{a-2}{\ell} (\tilde{\Phi}_1(j+\ell) - \tilde{\Phi}_1(j+1+\ell)) = \tilde{\Phi}_1(j) + \sum_{\ell=1}^{a-2} (-1)^\ell \left(\binom{a-2}{\ell} + \binom{a-2}{\ell-1} \right) \tilde{\Phi}_1(j+\ell) + (-1)^{a-2+1} \tilde{\Phi}_1(j+a-1) = \sum_{\ell=0}^{a-1} \binom{a-1}{\ell} (-1)^\ell \tilde{\Phi}_1(j+\ell).$

Aus (3.7) folgt für $j \geq 2, (a \geq 2)$

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{1}{c_N} \Phi_a(j, 1, \dots, 1) &= \sum_{\ell=0}^{a-1} \binom{a-1}{\ell} (-1)^\ell \lim_{N \rightarrow \infty} \frac{1}{c_N} \Phi_1(j+\ell) \\ &= \sum_{\ell=0}^{a-1} \binom{a-1}{\ell} (-1)^\ell \int_{[0,1]} y^{j+\ell-2} \Lambda(dy) = \int_{[0,1]} y^{j-2} \sum_{\ell=0}^{a-1} \binom{a-1}{\ell} (-1)^\ell y^\ell \Lambda(dy) \\ &= \int_{[0,1]} y^{j-2} (1-y)^{a-1} \Lambda(dy) \end{aligned}$$

Sei $|\xi| = b, \eta$ entstehe aus ξ durch Verschmelzen von j Klassen in eine einzige (insb. $|\eta| = b-j+1 =: a$), somit $\frac{1}{c_N} p^{(N,n)}(\xi, \eta) = \frac{1}{c_N} \Phi_a(j, 1, \dots, 1) \rightarrow q_{\xi\eta}$, d.h. $p^{(N,n)}(\xi, \eta) = c_N q_{\xi\eta} + o(c_N)$.

Da es nur endlich viele mögliche Übergänge gibt, muss schließlich auch der Diagonalterm $\eta = \xi$ stimmen. $\square$

Wir betrachten zunächst eine Familie von Beispielen, die sozusagen Beobachtung 3.1 auf allgemeinere Füße stellt und insbesondere zeigt, dass jedes W -maß Λ auf $[0, 1]$ als Limesobjekt in Satz 3.4 vorkommen kann. Vergleichbare Konstruktionen finden sich beispielsweise in [HM13] oder in [BBE13].

Beispiel 3.6. Sei Λ W -maß auf $(0, 1]$.

1) Falls $\int_{(0,1]} y^{-2} \Lambda(dy) = \infty$ gilt: Sei $Y \sim \Lambda$, setze $Y^{(N)} := \lceil NY \rceil$ und definiere eine Zufallsvariable $\Psi^{(N)}$ mit Werten in $\{2, 3, \dots, N\}$ via

$$\mathbb{P}(\Psi^{(N)} = k) = \frac{1}{\tilde{c}_N} \frac{1}{k(k-1)} \mathbb{P}(Y^{(N)} = k), \quad k = 2, 3, \dots, N$$

mit Normierung

$$\tilde{c}_N := \mathbb{E}[\mathbf{1}(Y^{(N)} \geq 2) / (Y^{(N)}(Y^{(N)} - 1))] \sim \frac{1}{N^2} \int_{(1/N, 1]} \frac{1}{y^2} \Lambda(dy)$$

Ähnlich wie in Beob. 3.1 sei im N -ten Populationsmodell $\nu^{(N)} = (\nu_1^{(N)}, \dots, \nu_N^{(N)})$ eine rein zufällige Permutation des (zufälligen) Vektors

$$\left(\Psi^{(N)}, \underbrace{0, 0, 0, \dots, 0}_{(\Psi^{(N)}-1)\text{-mal}}, \underbrace{1, 1, 1, \dots, 1}_{(N-\Psi^{(N)})\text{-mal}} \right)$$

Dann ist die Paarverschmelzungswahrscheinlichkeit (im N -ten Modell)

$$\begin{aligned} c_N &= \frac{\mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)} - 1)]}{N-1} = \frac{1}{N-1} \frac{1}{N} \sum_{k=2}^N k(k-1) \mathbb{P}(\Psi^{(N)} = k) \\ &= \frac{1}{N(N-1)\tilde{c}_N} \sum_{k=2}^N \mathbb{P}(Y^{(N)} = k) = \frac{1}{N(N-1)\tilde{c}_N} \sim \frac{1}{\int_{(1/N, 1]} y^{-2} \Lambda(dy)} \xrightarrow{N \rightarrow \infty} 0 \end{aligned} \quad (3.8)$$

und

$$\begin{aligned}
\frac{N}{c_N} \mathbb{P}(\nu_1^{(N)} > Nx) &= \frac{1}{c_N} \mathbb{P}(\Psi^{(N)} > Nx) = \frac{1}{c_N} \sum_{k=\lfloor Nx \rfloor + 1}^N \mathbb{P}(\Psi^{(N)} = k) \\
&\sim N^2 \sum_{k=\lfloor Nx \rfloor + 1}^N \frac{1}{k^2} \mathbb{P}(Y^{(N)} = k) = \sum_{k=\lfloor Nx \rfloor + 1}^N \frac{1}{(k/N)^2} \mathbb{P}(\lceil NY \rceil = k) \\
&\xrightarrow{N \rightarrow \infty} \int_{(x,1]} y^{-2} \Lambda(dy)
\end{aligned}$$

sowie $\Phi_2^{(N)}(2, 2) = 0$, d.h. die Bedingungen von Satz 3.4 sind erfüllt.

2) Falls $\int_{[0,1]} y^{-2} \Lambda(dy) < \infty$ gilt: Wir können wie in Beobachtung 3.1 vorgehen: Y habe Verteilung $\nu(dy) := y^{-2} \Lambda(dy) / (\int u^{-2} \Lambda(du))$, setze $\Psi^{(N)} := \lceil NY \rceil$, mit $0 < \gamma < 1$ bilde „Wright-Fisher-Generationen“ mit W'keit $1 - N^{-\gamma}$, Generationen mit $\Psi^{(N)}$ -großer Familie mit W'keit $N^{-\gamma}$.

Es ist

$$\begin{aligned}
c_N &= (1 - N^{-\gamma}) \frac{1}{N} + N^{-\gamma} \sum_{k=2}^N \frac{k(k-1)}{N(N-1)} \mathbb{P}(\Psi^{(N)} = k) \\
&\sim N^{-\gamma} \sum_{k=2}^N \frac{k(k-1)}{N(N-1)} \mathbb{P}\left(\frac{k-1}{N} < Y \leq \frac{k}{N}\right) \sim N^{-\gamma} \int_{(0,1]} y^2 \nu(dy) = N^{-\gamma} \frac{1}{\int u^{-2} \Lambda(du)}
\end{aligned}$$

nach Konstruktion ist stets $\nu_1^{(N)}(\nu_1^{(N)} - 1)\nu_2^{(N)}(\nu_2^{(N)} - 1) \equiv 0$ in Generationen mit $\Psi^{(N)}$ -großer Familie (und im Wright-Fisher-Fall treten asymptotisch ebenfalls keine Mehrfachverschmelzungen in der Stichprobe auf), zudem ist für $x \in (0, 1)$ mit $\Lambda(\{x\}) = 0$

$$\frac{N}{c_N} \mathbb{P}(\nu_1^{(N)} > Nx) = \frac{1}{c_N} N^{-\gamma} \mathbb{P}(\Psi^{(N)} > Nx) \sim \int u^{-2} \Lambda(du) \mathbb{P}(Y > x) = \int_{(x,1]} y^{-2} \Lambda(dy)$$

Wir sehen also bereits, dass jedes W'maß Λ , das kein Atom in 0 hat, in Satz 3.4 vorkommen kann. Im Fall 2) kann man leicht durch die Wahl $\gamma = 1$ und ggfs. einen Vorfaktor (d.h. Familien mit großer Familie treten mit W'keit c/N auf für eine gewisse Konstante c) dem Limesmaß ein Atom in 0 hinzufügen.

Mit etwas mehr Aufwand (da die relevante Zeitskala sozusagen implizit von Verhalten vom $\Lambda(dy)$ in der Nähe der 0 bestimmt wird, vergleiche (3.8)) ist dies auch im Fall 1) möglich, wir gehen hier aber nicht auf die Details ein.

Beispiel 3.7 (Ein Beispiel mit expliziter Modellierung juveniler und adulter Lebensphasen, nach Schweinsberg [So3]). Wir betrachten eine Population, die in jeder Generation aus N Individuen besteht. Der Reproduktionsmechanismus ist zweistufig: Zunächst produziert jedes Individuum $i = 1, \dots, N$ unabhängig eine zufällige Anzahl X_i von potentiellen Nachkommen oder „Juvenilen“, alle gemäß derselben Verteilung mit $\mu := \mathbb{E}[X_i] > 1$. Aus dem Pool von

$$S_N := X_1 + X_2 + \dots + X_N$$

potentiellen Nachkommen (was angesichts des Gesetzes der großen Zahlen typischerweise $\gg N$ ist, da $\mu > 1$) werden N ohne Zurücklegen gezogen. Diese bilden dann die („adulten“) Individuen der nächsten Generation.

Man denke etwa an eine Spezies, in der die Individuen prinzipiell sehr viele Samen oder Larven produzieren können, von denen aber wegen Ressourcenbeschränkungen immer nur eine begrenzte (und in der mathematischen Idealisierung konstante) Anzahl das Erwachsenenstadium erreicht. (Ohne den zweiten Schritt in obigem Mechanismus würde die Gesamtanzahl Individuen über die Generationen zufällig schwanken und typischerweise auch unbegrenzt anwachsen, mathematisch gesehen wäre sie dann ein Galton-Watson-Prozess, was diesen Teil des Titels von [So3] erklärt.)

Zunächst als „Aufwärmbeispiel“ seien alle $X_i \equiv 2$. Dann gilt $c_N = \frac{1}{2^{N-1}}$ und offenbar $d_N = 0$, d.h. die Genealogien von Stichproben konvergieren gegen den Kingman-Koaleszenten mit $\Lambda = \delta_0$, siehe z.B. [SMPB16, Satz 1.8] oder Satz 3.4 und die Bemerkung direkt danach.

Tatsächlich gilt dieselbe Konvergenz (mit entsprechend angepasstem c_N), sobald $\text{Var}[X_i] < \infty$ erfüllt ist, siehe [So3, Thm. 4 (a)].

Dies legt die Frage nahe, was passiert, wenn $\text{Var}[X_i] = \infty$ gilt. Um ein solches Szenario möglichst einfach darzustellen, treffen wir die folgende Annahme über die Verteilungsgewichte der X_i : Es gibt $1 < \alpha < 2$ und $0 < c_X < \infty$, so dass

$$\mathbb{P}(X_1 = k) \sim c_X \alpha k^{-1-\alpha} \quad \text{für } k \rightarrow \infty \quad (3.9)$$

und $\mathbb{P}(X_1 = 0) = 0$. Wegen $\alpha > 1$ ist dann $1 < \mathbb{E}[X_1] < \infty$, wegen $\alpha < 2$ ist aber $\mathbb{E}[X_1^2] = \sum_k k^2 \mathbb{P}(X_1 = k) = \infty$.

(Tatsächlich formuliert [So3] etwas allgemeinere Bedingungen: In [So3] wird $\mathbb{P}(X_1 = 0) > 0$ zugelassen, so dass der – asymptotisch extrem unwahrscheinliche – Fall $S_N < N$ jeweils separat betrachtet werden muss. Weiter wird dort anstelle von (3.9) nur die schwächere Bedingung $\mathbb{P}(X_1 \geq k) \sim c_X k^{-\alpha}$ gefordert, siehe [So3, Formel (II)]. Der Faktor α auf der rechten Seite in (3.9) führt dazu, dass für unsere X_i auch [So3, Formel (II)] wörtlich gilt (lese $C = c_X$ dort), so dass die folgenden Formeln mit denen aus [So3] ohne Umrechnungsfaktor verglichen werden können.)

Wir zeigen: Die Bedingungen von Satz 3.4 sind erfüllt mit

$$c_N \sim c_X \alpha \mu^{-\alpha} \Gamma(2 - \alpha) \Gamma(\alpha) N^{1-\alpha} \quad \text{für } N \rightarrow \infty \quad (3.10)$$

und

$$\Lambda(dx) = \mathbf{1}_{(0,1)}(x) \frac{1}{\Gamma(2 - \alpha) \Gamma(\alpha)} x^{1-\alpha} (1 - x)^{\alpha-1} dx \quad (3.11)$$

d.h. $\Lambda = \text{Beta}(2 - \alpha, \alpha)$.

Beweis. Nach Voraussetzung ist $S_N = X_1 + \dots + X_N \geq N$ stets, wir können also N -mal ohne Zurücklegen aus den potentiellen Nachkommen ziehen. Nach Konstruktion ist $\nu_1^{(N)}$ dann hypergeometrisch verteilt: Ziehe N mal ohne Zurücklegen aus einer Urne, die S_N Kugeln enthält, davon X_1 viele, die zu Individuum 1 gehören. Somit ist

$$\mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)} - 1)] = \mathbb{E}[\mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)} - 1) | X_1, \dots, X_N]] = \mathbb{E}\left[N(N-1) \frac{X_1(X_1 - 1)}{S_N(S_N - 1)}\right]$$

und daher für $N \rightarrow \infty$

$$\begin{aligned}
c_N &= \frac{1}{N-1} \mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)} - 1)] = N \mathbb{E}\left[\frac{X_1(X_1 - 1)}{S_N(S_N - 1)}\right] \\
&\sim N \mathbb{E}\left[\frac{X_1^2}{(X_1 + (N-1)\mu)^2}\right] \sim N c_X \alpha \sum_{k=1}^{\infty} k^{-1-\alpha} \frac{k^2}{(k + N\mu)^2} = c_X \alpha \mu^{-\alpha} N^{1-\alpha} \sum_{k=1}^{\infty} \frac{1}{\mu N} \frac{(k/(\mu N))^{1-\alpha}}{(1 + k/(\mu N))^2} \\
&\sim c_X \alpha \mu^{-\alpha} N^{1-\alpha} \int_0^{\infty} \frac{y^{1-\alpha}}{(y+1)^2} dy = c_X \alpha \mu^{-\alpha} \Gamma(2-\alpha) \Gamma(\alpha) N^{1-\alpha}
\end{aligned}$$

Im ersten Term in der zweiten Zeile haben wir hier $X_2 + X_3 + \dots + X_N$ angesichts des Gesetzes der großen Zahlen durch $\mathbb{E}[X_2 + X_3 + \dots + X_N] = \mu(N-1)$ ersetzt, was die Asymptotik nicht ändert (für eine präzise Rechtfertigung siehe den Beweis von [So3, Lemma 13]). Für das Gleichheitszeichen in der dritten Zeile beachte

$$\int_0^{\infty} \frac{y^{1-\alpha}}{(y+1)^2} dy = \int_0^1 \frac{\left(\frac{x}{1-x}\right)^{1-\alpha}}{\left(1 + \frac{x}{1-x}\right)^2} \frac{1}{(1-x)^2} dx = \int_0^1 x^{1-\alpha} (1-x)^{\alpha-1} dx = \frac{\Gamma(2-\alpha) \Gamma(\alpha)}{\Gamma(2)}$$

wobei wir $y = x/(1-x)$ (somit $x = 1 - 1/(y+1)$ und $dy/dx = 1/(1-x)^2$) substituiert haben.

Weiter ist für $x \in (0, 1)$ und $N \rightarrow \infty$

$$\begin{aligned}
\mathbb{P}\left(\frac{X_1}{S_N} > x\right) &\sim \mathbb{P}\left(\frac{X_1}{X_1 + N\mu} > x\right) = \mathbb{P}\left(X_1 > N\mu \frac{x}{1-x}\right) \sim \sum_{k > N\mu x/(1-x)} c_X \alpha k^{-1-\alpha} \\
&\sim c_X \left(N\mu \frac{x}{1-x}\right)^{-\alpha} = c_X \mu^{-\alpha} N^{-\alpha} \int_x^1 \frac{1}{y^2} \cdot \alpha y^{1-\alpha} (1-y)^{\alpha-1} dy
\end{aligned}$$

für die Gleichung in der zweiten Zeile beachte

$$\begin{aligned}
\frac{d}{dx} \frac{(1-x)^\alpha}{x^\alpha} &= \frac{-\alpha(1-x)^{\alpha-1}x^\alpha - (1-x)^\alpha \alpha x^{\alpha-1}}{x^{2\alpha}} = -\alpha(1-x)^{\alpha-1}x^{-\alpha} (1 + (1-x)x^{-1}) \\
&= -\alpha(1-x)^{\alpha-1}x^{-\alpha-1}
\end{aligned}$$

Wegen $x > 0$ interessieren wir uns hier für die (unwahrscheinliche) Situation, dass X_1 von der Größenordnung N ist (und es gilt dann $S_N \sim X_1 + \mu N$ in Wahrscheinlichkeit wegen des Gesetzes der großen Zahlen, angewendet auf die übrigen Summanden $X_2, \dots, X_N$). Wir ziehen also ohne Zurücklegen aus einer Urne mit mindestens $(x + \mu)N$ vielen Kugeln, in der der Anteil X_1/S_N von Kugeln der uns interessierenden „Farbe“ gleichmäßig für genügend große N mindestens $x/(x + \mu) > 0$ ist. Daher dürfen wir in der folgenden Rechnung $\nu_1^{(N)}/N$ durch X_1/S_N ersetzen (siehe z.B. [So3, Lemma 19] für die dafür verwendeten Konzentrationseigenschaften der hypergeometrischen Verteilungen und den Beweis von [So3, Thm. 4] für die entsprechenden Detailrechnungen) und erhalten

$$\frac{N}{c_N} \mathbb{P}(\nu_1^{(N)} > Nx) \sim \frac{N c_X \mu^{-\alpha} N^{-\alpha} \int_x^1 \frac{1}{y^2} \cdot \alpha y^{1-\alpha} (1-y)^{\alpha-1} dy}{c_X \alpha \mu^{-\alpha} \Gamma(2-\alpha) \Gamma(\alpha) N^{1-\alpha}} = \int_x^1 \frac{1}{y^2} \cdot \frac{y^{1-\alpha} (1-y)^{\alpha-1}}{\Gamma(2-\alpha) \Gamma(\alpha)} dy$$

Um schließlich noch Voraussetzung 3) aus Satz 3.4 nachzuprüfen: Es ist

$$\begin{aligned} \mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)} - 1)\nu_2^{(N)}(\nu_2^{(N)} - 1)] &= \mathbb{E}[\mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)} - 1)\nu_2^{(N)}(\nu_2^{(N)} - 1) \mid X_1, \dots, X_n]] \\ &= (N)_{4\downarrow} \mathbb{E}\left[\frac{X_1(X_1 - 1)X_2(X_2 - 1)}{(S_N)_{4\downarrow}}\right] \\ &\leq N^4 \mathbb{E}\left[\frac{X_1^2}{\max\{X_1^2, N^2\}} \frac{X_2^2}{\max\{X_2^2, N^2\}}\right] = N^4 \left(\mathbb{E}\left[\frac{X_1^2}{\max\{X_1^2, N^2\}}\right]\right)^2 \end{aligned}$$

Weiter ist

$$\mathbb{E}\left[\frac{X_1^2}{\max\{X_1^2, N^2\}}\right] \leq \frac{1}{N^2} \mathbb{E}[X_1^2 \mathbf{1}(X_1 \leq N)] + \mathbb{P}(X_1 > N) \sim \frac{1}{N^2} \sum_{k=1}^N c_X \alpha k^{1-\alpha} + c_X N^{-\alpha} \sim \frac{2c_X}{2-\alpha} N^{-\alpha}$$

und damit für eine Konstante $C < \infty$

$$\frac{\mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)} - 1)\nu_2^{(N)}(\nu_2^{(N)} - 1)]}{N^2 c_N} \leq C N^{2-2\alpha} N^{\alpha-1} = C N^{1-\alpha} \xrightarrow{N \rightarrow \infty} 0$$

□

Bericht 3.8 (Die Familie der $\text{Beta}(2 - \alpha, \alpha)$ -Koaleszenten). Man kann die Randfälle $\alpha = 1$ und $\alpha = 2$ in Beispiel 3.7 hinzunehmen, dies wird in [So3] auch betrachtet. Es gilt $\text{Beta}(2 - \alpha, \alpha) \xrightarrow{w} \delta_0$ für $\alpha \nearrow 2$, in diesem Sinne schließt die Familie in „stetiger Weise“ den Kingman-Koaleszenten mit ein. Für $\alpha > 2$ in (3.9) ist $\text{Var}[X_1] < \infty$ und die Stichproben-Genealogien konvergieren gegen den Kingman-Koaleszenten.

Im Fall $\alpha = 1$ ist $c_N \sim c/\log N$ und der zu $\Lambda = \text{Beta}(1, 1) = \text{Unif}([0, 1])$ gehörende Koaleszent ist der sogenannte Bolthausen-Sznitman-Koaleszent [BS98]. Dieser wurde in [BS98] zur Untersuchung von Fragen der statistischen Physik ungeordneter Systeme eingeführt, tritt aber auch in gewissen Populationsmodellen mit gerichteter Selektion in „natürlicher“ Weise auf.

Wenn man in Beispiel 3.7 ein $\alpha \in (0, 1)$ wählt, so ergibt sich kein Λ -Koaleszent als Limesobjekt (obwohl man natürlich in Definition 3.2 auch $\Lambda = \text{Beta}(2 - \alpha, \alpha)$ mit $0 < \alpha < 1$ einsetzen kann). Tatsächlich ist dann $\mathbb{E}[X_1] = \infty$ und es gibt in jeder Generation derart große Familien, dass $c_N \not\rightarrow 0$ gilt. Daher kann Satz 3.4 nicht greifen, stattdessen gibt es eine zeitdiskrete Limesdynamik (bei der auch mehrere disjunkte Gruppen von Ahnenlinien im selben Schritt verschmelzen können).

Beobachtung 3.9 (Austauschbarkeit und Konsistenzigenschaften der Λ - n -Koaleszenten). Ein Λ - n -Koaleszent ist eine zeitkontinuierliche Markovkette $(\mathcal{R}_t^{(n)})_{t \geq 0}$ auf $\mathcal{E}_n$, deren Sprungraten durch die Ausdrücke in Definition 3.2 gegeben sind. Für eine Permutation σ von $\{1, \dots, n\}$ sei $\sigma(\mathcal{R}_t^{(n)})$ definiert durch

$$i \sim_{\sigma(\mathcal{R}_t^{(n)})} j \iff \sigma^{-1}(i) \sim_{\mathcal{R}_t^{(n)}} \sigma^{-1}(j)$$

$(\mathcal{R}_t^{(n)})_{t \geq 0}$ ist *austauschbar*, d.h.

$$(\mathcal{R}_t^{(n)})_{t \geq 0} \stackrel{d}{=} (\sigma(\mathcal{R}_t^{(n)}))_{t \geq 0} \quad \text{für jede Permutation } \sigma \text{ von } \{1, \dots, n\}$$

Man sieht dies anhand der Form der Sprungraten (diese hängen nur von der Anzahl der beteiligten Klassen ab, vgl. dann auch Übungsaufgabe 2.3 für Bedingungen, unter denen das Bild einer Markovkette wiederum eine Markovkette ist) oder auch anhand der Stichproben-Genealogie-Interpretation aus Satz 3.4 (man kann eine zufällige Stichprobe umnummerieren, ohne ihre Verteilung zu ändern).

Weiterhin bilden die Λ - n -Koaleszenten eine *konsistente Familie*: Sei $\pi_{n,n-1} : \mathcal{E}_n \rightarrow \mathcal{E}_{n-1}$ die Einschränkung aller Äquivalenzklassen auf $\{1, \dots, n-1\}$, so gilt

$$(\mathcal{R}_t^{(n-1)})_{t \geq 0} \stackrel{d}{=} (\pi_{n,n-1}(\mathcal{R}_t^{(n)}))_{t \geq 0} \quad \text{für jedes } n \in \mathbb{N}$$

Dies folgt wiederum aus Satz 3.4 (wenn man aus einer zufälligen Stichprobe der Größe n z.B. die zuletzt gezogene Stichprobe weglässt, erhält man eine zufällige Stichprobe der Größe $n-1$) oder aus der Form der Sprungraten aus Definition 3.2: Es ist

$$\lambda_{n-1,j} = \lambda_{n,j} + \lambda_{n,j+1} \quad \text{für } 2 \leq j < n$$

(denn $\int x^{j-2}(1-x)^{n-j} \Lambda(dx) + \int x^{j-1}(1-x)^{n-j-1} \Lambda(dx) = \int x^{j-2}(1-x)^{n-1-j}(1-x+x) \Lambda(dx) = \int x^{j-2}(1-x)^{n-1-j} \Lambda(dx)$) und man kann dann wie in Übungsaufgabe 2.3 prüfen, dass $(\pi_{n,n-1}(\mathcal{R}_t^{(n)}))_{t \geq 0}$ eine Markovkette mit den korrekten Sprungraten ist. Obige Gleichung für Stichproben ausgesprochen besagt: Eine Verschmelzung von j Klassen in der $(n-1)$ -Stichprobe entspricht entweder einer j -Verschmelzung in der n -Stichprobe (falls die n -te Stichprobe nicht an der Verschmelzung teilnimmt) oder einer $(j+1)$ -Verschmelzung in der n -Stichprobe (falls die n -te Stichprobe an der Verschmelzung teilnimmt).

Wie im Fall des Kingman-Koaleszenten kann man mittels der Konsistenzeigenschaft einen Λ -Koaleszenten auf ganz $\mathbb{N}$, d.h. einen Markovprozess $(\mathcal{R}_t)_{t \geq 0}$ mit Werten in den Äquivalenzklassen auf $\mathbb{N}$ definieren, so dass für jedes $n \in \mathbb{N}$ dessen Einschränkung auf $\{1, \dots, n\}$ ein Λ - n -Koaleszent ist. Man kann dazu entweder „abstrakte Theorie“, basierend auf Kolmogorovs Erweiterungssatz, verwenden oder z.B. die explizite Konstruktion aus Übungsaufgabe 8.1 ausnutzen.

Bemerkung 3.10 (Pitmans Ansatz). J. Pitman [P99] hat gewissermaßen einen zu Satz 3.4 komplementären Ansatz zur Definition und Konstruktion der Λ -Koaleszenten gewählt, der die Konsistenzeigenschaft aus Beobachtung 3.9 als Ausgangspunkt nimmt: Betrachte eine zeitkontinuierliche Markovkette auf $\bigcup_{n=1}^{\infty} \mathcal{E}_n$, die austauschbar und konsistent ist wie in Beob. 3.9 und zudem als Übergänge nur Verschmelzungen einer Gruppe von Äquivalenzklassen pro Sprung zulässt. Sei $\lambda_{b,j}$ die Rate, mit der eine Teilmenge von j Äquivalenzklassen verschmilzt, wenn es aktuell b Äquivalenzklassen gibt (wegen der Austauschbarkeit hängt dies nicht davon ab, welche j der b Äquivalenzklassen das sind). Die Konsistenzeigenschaft erzwingt dann (wie oben)

$$\lambda_{b,j} = \lambda_{b+1,j} + \lambda_{b+1,j+1}, \quad b \geq j \geq 2$$

Um daraus das zugehörige endliche Maß Λ auf $[0, 1]$ wie in Definition 3.2 zu gewinnen, geht [P99, Section 3] folgendermaßen vor: Setze

$$\mu_{i,j} := \frac{\lambda_{i+j+2,i+2}}{\lambda_{2,2}}, \quad i, j \in \mathbb{N}_0$$

dann gilt $\mu_{i,j} = \mu_{i+1,j} + \mu_{i,j+1}$ für $i, j \in \mathbb{N}_0$ mit Randbedingung $\mu_{0,0} = 1$ (insbesondere $\mu_{i,j} \leq 1$ für alle i, j) und man kann mittels Kolmogorovs Konsistenzsatz eine unendliche, austauschbare Familie $X = (X_1, X_2, \dots)$ von $\{0, 1\}$ -wertigen Zufallsvariablen definieren via

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \mu_{k,n-k} \quad \text{für } x_1, \dots, x_n \in \{0, 1\} \text{ mit } \sum_{j=1}^n x_j = k$$

(beachte: es gilt dann $\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n, X_{n+1} = 0) + \mathbb{P}(X_1 = x_1, \dots, X_n = x_n, X_{n+1} = 1)$ und da die Wahrscheinlichkeit oben nur von der Anzahl der 1er, nicht aber von deren Position abhängt, gilt Austauschbarkeit).

Gemäß dem Satz von de Finetti (z.B. [Kl, Satz 12.26] oder Übungsaufgabe 8.2) gibt es dann eine $[0, 1]$ -wertige Zufallsvariable ξ , so dass

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \mathbb{E}[\xi^k (1 - \xi)^{n-k}] \quad \text{mit } k = \sum_{j=1}^n x_j$$

und demnach leistet $\Lambda = \lambda_{2,2}\mathcal{L}(\xi)$ das Gewünschte.

- Bericht 3.II.** 1. $(\mathcal{R}_t)_{t \geq 0}$ Λ -Koaleszent auf $\mathbb{N}$ mit $\mathcal{R}_0 = \{\{1\}, \{2\}, \{3\}, \dots\}$ (vgl. Diskussion in Beob. 3.9). Eintritt/Herunterkommen aus dem Unendlichen (d.h. $\#\mathcal{R}_t < \infty$ für alle $t > 0$) g.d.w. $\sum_{b=2}^{\infty} 1/\gamma_b < \infty$ mit $\gamma_b := \sum_{j=2}^b (j-1) \binom{b}{j} \lambda_{b,j}$ siehe [Soob] und auch Bericht 3.I8 unten.
2. “dust vs. no dust” (vgl. [P99], speziell Thm. 8 dort): $(\mathcal{R}_t)_{t \geq 0}$ Λ -Koaleszent auf $\mathbb{N}$ mit $\mathcal{R}_0 = \{\{1\}, \{2\}, \{3\}, \dots\}$ (vgl. Diskussion in Beob. 3.9). Falls $\int_{[0,1]} \frac{1}{x} \Lambda(dx) = \infty$, so hat $\mathcal{R}_t$ für jedes $t > 0$ “proper frequencies”, d.h. die asymptotischen Anteile der Äquivalenzklassen (diese existieren f.s. wegen Austauschbarkeit) addieren sich zu 1 und

$$\begin{aligned} & \mathbb{P}(\text{jede Äquivalenzklasse von } \mathcal{R}_t \text{ hat mehr als ein Element}) \\ &= \mathbb{P}(\text{jede Äquivalenzklasse von } \mathcal{R}_t \text{ ist unendlich}) = 1 \end{aligned}$$

Falls dagegen $\int_{[0,1]} \frac{1}{x} \Lambda(dx) < \infty$, so addieren sich für jedes $t \geq 0$ die asymptotischen Anteile der Äquivalenzklassen von $\mathcal{R}_t$ zu < 1 und $\mathcal{R}_t$ besitzt unendlich viele Äquivalenzklassen der Größe 1 (“singletons”). Man kann dies anhand der Poissonkonstruktion aus Aufgabe 8.1 einsehen.

3. Wenn man die Bedingung $\Phi_2(2, 2)/c_N \rightarrow 0$ in Satz 3.4 aufgibt (und geeignet ersetzt), so erhält man die Limesklasse der Ξ -Koaleszenten (mit simultanen multiplen Verschmelzungen), siehe [MSOI], [Soo].

3.I Zur Asymptotik des Blockzählprozesses von Λ -Koaleszenten

Sei Λ ein W’maß auf $[0, 1]$, $(\mathcal{R}_t)_{t \geq 0}$ ein Λ -Koaleszent wie in Definition 3.2 (mit einer gewissen Stichprobengröße n , die wir hier in der Notation nicht immer explizit machen). Der zugehörige Blockzählprozess

$$N_t = \#\mathcal{R}_t, \quad t \geq 0$$

(wir schreiben auch $N^{(n)}$, wenn $\#\mathcal{R}_0 = n$ explizit gemacht werden soll) ist eine zeitkontinuierliche Markovkette auf $\mathbb{N}$ mit Sprungrate von $\ell \rightarrow m$

$$q_{\ell,m} = \binom{\ell}{\ell-m+1} \int x^{\ell-m-1} (1-x)^{m-1} \Lambda(dx), \quad 1 \leq m \leq \ell$$

(vgl. auch Übungsaufgabe 2.3 „Wann ist das Bild einer Markovkette wieder eine?“).

Die Leitfrage dieses Unterkapitels:

Verhalten bei Start in $N_0 = n \gg 1$?

Wir wollen in hier, J. Berestycki, N. Berestycki und V. Limic [BBL10, BBL14] folgend, eine gewisse quantitative Antwort geben. Die Antwort ist mathematisch interessant und sie ist z.B. via die Gesamtlänge, etc. angesichts von Satz 3.4 auch für die (zumindest asymptotische) Untersuchung der Verteilung der genetischen Variabilität in einer (großen) Stichprobe relevant, siehe Satz 3.16 unten.

Die Sprungraten lassen sich mit $B_{\ell,x} \sim \text{Bin}(\ell, x)$ folgendermaßen formulieren: von $\ell \rightarrow \ell - k$ (für $1 \leq k \leq \ell - 1$)

$$\begin{aligned} q_{\ell,\ell-k} &= \mathbf{1}(k=1) \Lambda(\{0\}) \binom{\ell}{2} + \int_{(0,1]} \binom{\ell}{\ell-(k+1)} x^{k+1} (1-x)^{\ell-k-1} \frac{1}{x^2} \Lambda(dx) \\ &= \mathbf{1}(k=1) \Lambda(\{0\}) \binom{\ell}{2} + \int_{(0,1]} \mathbb{P}(B_{\ell,x} - \mathbf{1}(B_{\ell,x} > 0) = k) \frac{1}{x^2} \Lambda(dx), \end{aligned}$$

(beachte: der Integrand ist beschränkt), also ist die Gesamtsprungrate in ℓ

$$\begin{aligned} -q_{\ell,\ell} &= \Lambda(\{0\}) \binom{\ell}{2} + \int_{(0,1]} \sum_{k=1}^{\ell-1} \mathbb{P}(B_{\ell,x} - \mathbf{1}(B_{\ell,x} > 0) = k) \frac{1}{x^2} \Lambda(dx) \\ &= \Lambda(\{0\}) \binom{\ell}{2} + \int \frac{\mathbb{P}(B_{\ell,x} \neq 0, 1)}{x^2} \Lambda(dx) = \int \frac{1 - (1-x)^\ell - \ell x (1-x)^{\ell-1}}{x^2} \Lambda(dx) \end{aligned}$$

Beobachtung 3.12. Erwartete infinitesimale (“abwärts”-)Änderung von N_t , gegeben $N_t = \ell$

$$\begin{aligned} \gamma_\ell &:= \sum_{j=2}^{\ell} (j-1) \binom{\ell}{j} \lambda_{\ell,j} = \Lambda(\{0\}) \binom{\ell}{2} + \int_{(0,1]} \mathbb{E}[B_{\ell,x} - \mathbf{1}(B_{\ell,x} > 0)] \frac{1}{x^2} \Lambda(dx) \\ &= \Lambda(\{0\}) \binom{\ell}{2} + \int_{(0,1]} \frac{(1-x)^\ell - 1 + \ell x}{x^2} \Lambda(dx) \end{aligned}$$

Es ist $\mathbb{E}[N_{t+\delta} - N_t | \mathcal{F}_t] = -\delta \gamma_{N_t} + o(\delta)$ und $N_t + \int_0^t \gamma_{N_s} ds, t \geq 0$ ist ein Martingal.

Sei

$$\psi(q) := \int_{[0,1]} \frac{e^{-qx} - 1 + qx}{x^2} \Lambda(dx) = \frac{q^2}{2} \Lambda(\{0\}) + \int_{(0,1]} \frac{e^{-qx} - 1 + qx}{x^2} \Lambda(dx), \quad q \geq 0 \quad (3.12)$$

(es ist $\psi(q) > 0$ für $q > 0$, denn strikte Konvexität der Exponentialfunktion liefert $e^{-y} - 1 + y > 0$ für $y > 0$).

Beobachtung 3.13. $\psi(\ell) \approx \gamma_\ell$ (für $\ell \gg 1$), genauer:

$$\begin{aligned} \left| \frac{\psi(\ell) - \gamma_\ell}{\ell} \right| &\leq \frac{1}{\ell} \Lambda(\{0\}) \frac{\ell^2 - \ell(\ell - 1)}{2} + \int_{(0,1]} \frac{e^{-\ell x} - (1-x)^\ell}{\ell x^2} \Lambda(dx) \\ &= \frac{\Lambda(\{0\})}{2} + \underbrace{\frac{1}{\ell} \int_{(1/2,1]} \frac{e^{-\ell x} - (1-x)^\ell}{x^2} \Lambda(dx)}_{\leq 4\Lambda((1/2,1])} + \int_{(0,1/2]} \frac{e^{-\ell x} - (1-x)^\ell}{\ell x^2} \Lambda(dx) \\ &\leq \Lambda(\{0\}) \frac{1}{2} + \frac{4}{\ell} + C \end{aligned}$$

denn (verwende $\log(1-x) = -\sum_{k=1}^{\infty} x^k/k$)

$$\frac{e^{-\ell x} - (1-x)^\ell}{\ell x^2} = \frac{e^{-\ell x}}{\ell x^2} (1 - e^{\ell x \log(1-x)}) = \frac{e^{-\ell x}}{\ell x^2} \left(1 - \exp\left(-\ell \sum_{k=2}^{\infty} x^k/k\right) \right)$$

und $1 - \exp\left(-\ell \sum_{k=2}^{\infty} x^k/k\right) \leq 1 - \exp(-4\ell x^2) \leq C\ell x^2$ für $x \leq 1/2$.

Bemerkung 3.14. Wenn $\Lambda(dx) = f(x)dx$ und $f(x) \sim Ax^{1-\alpha}$ ($x \downarrow 0$) für ein $A > 0$, $\alpha \in (1, 2)$ gilt („starke α -reguläre Variation“), so ist

$$\psi(q) \sim A \frac{\Gamma(2-\alpha)}{\alpha(\alpha-1)} q^\alpha, \quad q \rightarrow \infty$$

Annahme 3.15. Es gelte

$$\int_1^\infty \frac{1}{\psi(q)} dq < \infty$$

Notiere $N_t^{(n)}$ für den Blockzählprozess mit Start in $N_0^{(n)} = n$, $\tau_k^{(n)} := \inf\{t : N_t^{(n)} \leq k\}$,

$$L_n := \int_0^{\tau_1^{(n)}} N_s^{(n)} ds$$

(Gesamtlänge eines Λ - n -Koaleszenten),

$S_n = \#$ Mutationen in n -Stichprobe an den Blättern eines Λ - n -Koaleszenten, längs dessen Ästen sich mit Rate $\theta/2$ Mutationen ereignen

Satz 3.16 (Siehe Theorem 3 in [BBL14]). *Es gelte Annahme 3.15.*

$$\begin{aligned} \frac{L_n}{\int_1^n \frac{q}{\psi(q)} dq} &\xrightarrow{n \rightarrow \infty} 1 \quad \text{in } W\text{'keit}, \\ \frac{S_n}{\int_1^n \frac{q}{\psi(q)} dq} &\xrightarrow{n \rightarrow \infty} \frac{\theta}{2} \quad \text{in } W\text{'keit}. \end{aligned}$$

Beispiel 3.17. Im Fall $\Lambda(dx) = f(x)dx$ mit $f(x) \sim Ax^{1-\alpha}$ ($x \downarrow 0$) wie in Berkung 3.14 ist $\psi(q) \sim A \frac{\Gamma(2-\alpha)}{\alpha(\alpha-1)} q^\alpha$, also

$$\int_1^n \frac{q}{\psi(q)} dq \sim \frac{\alpha(\alpha-1)}{A\Gamma(2-\alpha)} \int_1^n q^{1-\alpha} dq = \frac{\alpha(\alpha-1)}{A\Gamma(2-\alpha)} \left[\frac{1}{2-\alpha} q^{2-\alpha} \right]_1^n \sim \frac{\alpha(\alpha-1)}{A(2-\alpha)\Gamma(2-\alpha)} n^{2-\alpha}$$

Im Fall $\Lambda(dx) = \delta_0$ ist $\psi(q) = q^2/2$, $\int_1^n \frac{q}{\psi(q)} dq = 2 \int_1^n q^{-1} dq \sim 2 \log n$ und wir gewinnen (erneut) die Konsistenz des Watterson-Schätzers $\widehat{\theta}_W := S_n / \sum_{i=1}^{n-1} i^{-1}$ für θ .

Bericht 3.18. Zum „Abstieg aus dem Unendlichen“ (coming down from infinity):

$$\text{Die Familie } \mathcal{L}(\tau_1^{(n)}), n \in \mathbb{N} \text{ ist straff} \iff \sum_{\ell=2}^{\infty} \frac{1}{\gamma_\ell} < \infty$$

siehe [Soob]. Diese Bedingung ist äquivalent zu Annahme 3.15 (vgl. Beobachtung 3.13).

Sei $v(t)$ gegeben durch

$$v(t) := \inf \left\{ s \geq 0 : \int_s^\infty \frac{1}{\psi(q)} dq < t \right\}$$

(mit Interpretation $\inf \emptyset = +\infty$, $v(t)$ ist wegen Annahme 3.15 tatsächlich endlich für jedes $t > 0$), d.h. $v(t)$ löst

$$\int_{v(t)}^\infty \frac{1}{\psi(q)} dq = t, \quad t \geq 0$$

oder $\frac{d}{dt}v(t) = -\psi(v(t))$, $t > 0$ mit $v(0) = \infty$.

Heuristik: Es ist

$$\mathbb{E}[dN_t | \mathcal{F}_t] = -\gamma_{N_t} dt \approx -\psi(N_t) dt$$

man ersetzt (N_t) durch diese Differentialgleichung (im Jargon ist dies ein „hydrodynamischer Limes“ für eine (Schar von) Markovketten) und verwendet, dass die (divergente) Länge asymptotisch nur in der Nähe der Blätter erzeugt wird.

Sei $v^{(n)}(t)$ Lösung von

$$\frac{d}{dt}v^{(n)}(t) = -\psi(v^{(n)}(t)), \quad v^{(n)}(0) = n$$

d.h. $v^{(n)}(t) = v(t_n + t)$ mit $t_n := \int_n^\infty \frac{1}{\psi(q)} dq = v^{-1}(n) = \inf\{t : v(t) \leq n\}$,

ersetze $\int_0^1 N_s^{(n)} ds \approx \int_0^1 v^{(n)}(s) ds$,

$$\int_0^1 v^{(n)}(s) ds = \int_0^1 v(t_n + t) ds = \int_{v(t_n+1)}^{v(t_n)} \frac{q}{\psi(q)} dq \sim \int_{v(1)}^n \frac{q}{\psi(q)} dq \sim \int_1^n \frac{q}{\psi(q)} dq \quad (3.13)$$

[substituiere $v(t_n + t) = q$, also $dq/dt = v'(t_n + t) = -\psi(v(t_n + t)) = -\psi(q)$, verwende $t_n \rightarrow_{n \rightarrow \infty} 0$, $\int_a^b \frac{q}{\psi(q)} dq < \infty$, $\int_a^\infty \frac{q}{\psi(q)} dq = \infty$ für $0 < a < b < \infty$]

Das zentrale technische Ergebnis ist:

Satz 3.19 (Vgl. Proposition 12 in [BBL10]). *Es gelte Annahme 3.15. Für $t_0 \leq 1$, $a \in (0, 1/2)$ gibt es $c, C \in (0, \infty)$, $n_0 \in \mathbb{N}$ mit*

$$\mathbb{P}_n \left(\sup_{t \leq t_0} \left| \frac{N_t}{v^{(n)}(t)} - 1 \right| \geq ct_0^a \right) \leq Ct_0^{1-2a} \quad \text{gleichmäßig in } n \geq n_0.$$

Wir nehmen von hier ab an:

$$\Lambda(\{0\}) = 0, \quad \Lambda([1-\eta, 1]) = 0 \quad (3.14)$$

für ein $\eta > 0$.

Da $\int_{[1-\eta, 1]} x^{-2} \Lambda(dx) < \infty$ stets gilt und daher „ x -Ereignisse“ mit $x \in [1-\eta, 1]$ nur mit gleichmäßig in n beschränkter Rate geschehen, ändert diese Annahme nichts an der asymptotischen Aussage aus Satz 3.19, siehe [BBL10] für Details. Siehe ebenso [BBL10] für die Behandlung des Falls mit „Kingman-Atom“, d.h. $\Lambda(\{0\}) > 0$.

Beweisskizze (für Satz 3.19). Mit $Lf(\ell) = \sum_m q_{\ell, m}(f(m) - f(\ell))$, dem Generator von (N_t) , sei

$$g(\ell) = (L \log)(\ell) = \sum_{m=1}^{\ell-1} q_{\ell, m}(\log m - \log \ell) = \int_{(0,1]} \mathbb{E} \left[\log \left(\frac{\ell - B_{\ell, x} + \mathbf{1}(B_{\ell, x} > 0)}{\ell} \right) \right] \frac{1}{x^2} \Lambda(dx) \quad (3.15)$$

$$M_t := \log N_t - \log N_0 - \int_0^t g(N_s) ds, \quad t \geq 0$$

ist Martingal (bei Start in $N_0 = n < \infty$) und

$$M_t^2 - \int_0^t (L \log^2(\cdot))(N_s) - 2 \log(N_s)(L \log(\cdot))(N_s) ds$$

ist ebenfalls ein Martingal ($\Gamma(f, g) = Lfg - fLg - gLf$ ist der sogenannte opérateur carré du champ, siehe z.B. [KL99, Lemma 5.1 in Appendix 1], insbesondere für einen reinen Sprunggenerator $Lf(x) = \sum_y q_{xy}[f(y) - f(x)]$ ist $L(f^2)(x) - 2f(x)Lf(x) = \sum_y q_{xy}[f^2(y) - f^2(x) - 2f(x)f(y)] = \sum_y q_{xy}[f(y) - f(x)]^2$ der erw. quadrierte Sprung.)

Somit

$$\mathbb{E}_n[M_t^2] = \int_0^t \mathbb{E}_n[\tilde{g}(N_s)] ds \leq C't$$

mit

$$\tilde{g}(\ell) = \int_{(0,1]} \mathbb{E} \left[\left(\log \frac{\ell - B_{\ell, x} + \mathbf{1}(B_{\ell, x} > 0)}{\ell} \right)^2 \right] \frac{1}{x^2} \Lambda(dx)$$

Wir benötigen für den Beweis von Satz 3.19 folgende technische Abschätzungen/Beobachtungen:

Lemma 3.20. *Es gilt*

1)

$$\begin{aligned}\mathbb{E}[B_{\ell,x} - \mathbf{1}(B_{\ell,x} > 0)] &= \ell x - 1 + (1-x)^\ell, \\ \mathbb{E}[(B_{\ell,x} - \mathbf{1}(B_{\ell,x} > 0))^2] &= -\ell x - \ell x^2 + \ell^2 x^2 + 1 - (1-x)^\ell \leq C\ell^2 x^2 \quad \text{für } x \in (0, 1-\eta)\end{aligned}$$

Für die 2. Zeile beachte:

$$\begin{aligned}\text{Var}(B_{\ell,x} - \mathbf{1}(B_{\ell,x} > 0)) &= \text{Var}(B_{\ell,x}) + \text{Var}(\mathbf{1}(B_{\ell,x} > 0)) - 2\text{Cov}(B_{\ell,x}, \mathbf{1}(B_{\ell,x} > 0)) \\ &= \ell x(1-x) + (1 - (1-x)^\ell)(1-x)^\ell - 2\underbrace{(\mathbb{E}[B_{\ell,x}\mathbf{1}(B_{\ell,x} > 0)] - \ell x(1 - (1-x)^\ell))}_{=\ell x} \\ &= \ell x(1-x) + (1 - (1-x)^\ell)(1-x)^\ell - 2\ell x(1-x)^\ell = \ell x(1-x) + (1-x)^\ell(1 - (1-x)^\ell - 2\ell x),\end{aligned}$$

also

$$\begin{aligned}\mathbb{E}[(B_{\ell,x} - \mathbf{1}(B_{\ell,x} > 0))^2] &= \text{Var}(B_{\ell,x} - \mathbf{1}(B_{\ell,x} > 0)) + (\mathbb{E}[B_{\ell,x} - \mathbf{1}(B_{\ell,x} > 0)])^2 \\ &= \ell x(1-x) + (1-x)^\ell(1 - (1-x)^\ell - 2\ell x) + (\ell x - 1)^2 + 2(\ell x - 1)(1-x)^\ell + (1-x)^{2\ell} \\ &= \ell x - \ell x^2 + (1-x)^\ell - (1-x)^{2\ell} - 2\ell x(1-x)^\ell + \ell^2 x^2 - 2\ell x + 1 + 2\ell x(1-x)^\ell - (1-x)^\ell + (1-x)^{2\ell} \\ &= -\ell x - \ell x^2 + \ell^2 x^2 + 1 - (1-x)^\ell\end{aligned}$$

und

$$1 - \ell x - (1-x)^\ell \leq \begin{cases} c\ell^2 x^2, & x \in (0, 1/\ell], \\ 0, & x > 1/\ell. \end{cases}$$

2) Für $x \in (0, 1-\eta]$

$$\begin{aligned}& \left| \mathbb{E} \left[\log \left(\frac{\ell - B_{\ell,x} + \mathbf{1}(B_{\ell,x} > 0)}{\ell} \right) \right] - \frac{1 - (1-x)^\ell - \ell x}{\ell} \right| \\ &= \left| \mathbb{E} \left[\log \left(\frac{\ell - B_{\ell,x} + \mathbf{1}(B_{\ell,x} > 0)}{\ell} \right) + \frac{B_{\ell,x} - \mathbf{1}(B_{\ell,x} > 0)}{\ell} \right] \right| \\ &\leq \frac{C}{\ell^2} \mathbb{E}[(B_{\ell,x} - \mathbf{1}(B_{\ell,x} > 0))^2] + (1 + \log \ell) \mathbb{P}(|B_{\ell,x}/\ell - x| \geq \eta/2) \leq C' x^2\end{aligned}$$

(verwende $\log(1-x) - x \leq C_\eta x^2$ für $x \in [0, 1-\eta]$, Teil 1) und elementare Abschätzungen für große Abweichungen für die Binomialverteilung)

3) Analog ist für $x \in (0, 1-\eta]$

$$\begin{aligned}& \left| \mathbb{E} \left[\left(\log \frac{\ell - B_{\ell,x} + \mathbf{1}(B_{\ell,x} > 0)}{\ell} \right)^2 \right] \right| \\ &\leq \frac{C}{\ell^2} \mathbb{E}[(B_{\ell,x} - \mathbf{1}(B_{\ell,x} > 0))^2] + (\log \ell)^2 \mathbb{P}(|B_{\ell,x}/\ell - x| \geq \eta/2) \leq C' x^2\end{aligned}$$

(verwende $|\log(1-x)| \leq C_\eta x$ für $x \in [0, 1-\eta]$ und große Abweichungen wie oben).

Nach obigem (Definition von g aus (3.15), Lemma 3.20, 2) und der Definition von γ_ℓ aus Beobachtung 3.12) ist

$$g(\ell) \approx -\frac{\gamma_\ell}{\ell} \approx -\frac{\psi(\ell)}{\ell}, \quad \text{genauer } |g(\ell) + \gamma_\ell/\ell| \leq C \quad \text{und mit } h(\ell) := g(\ell) + \psi(\ell)/\ell \text{ auch}$$

$$|h(\ell)| = |g(\ell) + \psi(\ell)/\ell| \leq |g(\ell) + \gamma_\ell/\ell| + |\gamma_\ell - \psi(\ell)|/\ell \leq C'$$

(verwende Beobachtung 3.13 für den zweiten Summanden auf der rechten Seite).

Es gilt

$$\mathbb{P}_n \left(\sup_{t \leq t_0} |M_t| \geq t_0^a \right) \leq \frac{1}{t_0^{2a}} \mathbb{E}_n \left[\sup_{t \leq t_0} M_t^2 \right] \leq \frac{1}{t_0^{2a}} \mathbb{E}_n [M_t^2] \leq 4C' t_0^{1-2a}$$

(Markov-, dann Doob's L^2 -Ungleichung).

Sei $v^{(n)}(t)$ Lösung von $v^{(n)}(0) = n$, $\frac{d}{dt} v^{(n)}(t) = -\psi(v^{(n)}(t))$, d.h.

$$\log v^{(n)}(t) - \log n = \int_0^t \frac{d}{ds} (\log v^{(n)})(s) ds = - \int_0^t \frac{\psi(v^{(n)}(s))}{v^{(n)}(s)} ds$$

Starte in $N_0 = n (< \infty)$. Es ist

$$\begin{aligned} \log \left(\frac{N_t}{v^{(n)}(t)} \right) &= M_t + \int_0^t g(N_s) + \frac{\psi(v^{(n)}(s))}{v^{(n)}(s)} ds \\ &= M_t + \int_0^t \frac{\psi(v^{(n)}(s))}{v^{(n)}(s)} - \frac{\psi(N_s)}{N_s} ds + \int_0^t h(N_s) ds \end{aligned}$$

wegen $\sup_{t \leq t_0} \left| \int_0^t h(N_s) ds \right| \leq C t_0$ und (Definition von M_t sowie der DGL. für $v^{(n)}$) also ($t_0 \leq 1$ so dass $t_0 \leq t_0^a$ für $0 < a < 1$, zudem $\int_0^t h(N_s) ds \leq C't$)

$$\mathbb{P}_n \left(\sup_{t \leq t_0} \left| \log \left(\frac{N_t}{v^{(n)}(t)} \right) + \int_0^t \frac{\psi(N_s)}{N_s} - \frac{\psi(v^{(n)}(s))}{v^{(n)}(s)} ds \right| \geq C_1 t_0^a \right) \leq C_2 t_0^{1-2a}$$

mit geeign. Konstanten $C_1, C_2 \in (0, \infty)$.

Lemma 3.21 zeigt, dass $\log \left(\frac{N_t}{v^{(n)}(t)} \right)$ und $\frac{\psi(N_s)}{N_s} - \frac{\psi(v^{(n)}(s))}{v^{(n)}(s)}$ dasselbe Vorzeichen haben, Lemma 3.22 liefert dann (mit $c_1, c_2, C \in (0, \infty)$ geeignet)

$$\mathbb{P}_n \left(\sup_{t \leq t_0} \left| \frac{N_t}{v^{(n)}(t)} - 1 \right| \geq c_1 t_0^a \right) \leq \mathbb{P}_n \left(\sup_{t \leq t_0} \left| \log \left(\frac{N_t}{v^{(n)}(t)} \right) \right| \geq c_2 t_0^a \right) \leq C t_0^{1-2a}$$

□

Lemma 3.21 (Siehe [BBL10, Lemma 9]). $q \mapsto \frac{\psi(q)}{q}$ ist nicht-fallend,

$$\frac{d}{dq} \left(\frac{\psi(q)}{q} \right) = \frac{\psi'(q)q - \psi(q)}{q^2} = \frac{1}{q^2} \int_{[0,1]} \frac{1 - (qx+1)e^{-qx}}{x^2} \Lambda(dx) \geq 0$$

Denn mit $f(x) := (qx + 1)e^{-qx}$, also $f'(x) = qe^{-qx} - q(qx + 1)e^{-qx} = -q^2xe^{-qx} \leq 0$ ist $\sup_{x \geq 0} f(x) = f(0) = 1$; weiter ist $\lim_{x \downarrow 0} \frac{1 - (qx+1)e^{-qx}}{x^2} = q^2/2$.

Lemma 3.22 (Siehe [BBL10, Lemma 10]). Seien $\varphi_1, \varphi_2 : [0, t_0] \rightarrow \mathbb{R}$ càdlàg-Funktionen mit

$$\sup_{0 \leq t \leq t_0} \left| \varphi_1(t) + \int_0^t \varphi_2(s) ds \right| \leq c, \quad \varphi_1(t) \neq 0 \Rightarrow \varphi_1(t)\varphi_2(t) > 0$$

dann gilt

$$\sup_{0 \leq t \leq t_0} \left| \int_0^t \varphi_2(s) ds \right| \leq c \quad \text{und insbes.} \quad \sup_{0 \leq t \leq t_0} |\varphi_1(t)| \leq 2c$$

Beweis. $\varphi(t) := \int_0^t \varphi_2(s) ds$ ist stetig, und $\varphi_1(t) > 0$ (bzw. < 0) $\Rightarrow$ wachsend an der Stelle $t+$ (bzw. fallend). $t_* := \min\{t : |\varphi(t)| > c\} \wedge t_0$. Angenommen, $t_* < t_0$: Stetigkeit liefert $|\varphi(t_*)| = c$, betr. o.E. den Fall $\varphi(t_*) = c$, also

$$\varphi(t_* + \epsilon) > c \quad \text{für alle } \epsilon > 0 \text{ und gen. klein.}$$

Wenn $\varphi_1(t) < 0$ auf $(t_*, t_* + \epsilon)$, so müsste dort auch $\varphi_1(t) < 0$ und somit $\varphi_2(t_* + \epsilon) < c$ sein, ein Widerspruch. Demnach existiert $\tilde{t} \in (t_*, t_* + \epsilon)$ mit $\varphi_1(\tilde{t}) \geq 0$, dann gälte allerdings $\varphi_1(\tilde{t}) + \varphi(\tilde{t}) > c$ im Widerspruch zur Annahme. Also muss $t_* = t_0$ sein. $\square$

Beweis von Satz 3.16. Sei $s > 0$. Wir zeigen zunächst, dass für jedes $\varepsilon > 0$

$$\mathbb{P}_n \left(\left| \frac{\int_0^s N_t dt}{\int_0^s v^{(n)}(t) dt} - 1 \right| > \varepsilon \right) \xrightarrow{n \rightarrow \infty} 0 \quad (3.16)$$

gilt.

Es ist (vgl. (3.13) in der Heuristik vor Satz 3.16)

$$\int_0^s v^{(n)}(t) dt = \int_0^s v(t + t_n) dt = \int_{v(t_n+s)}^{v(t_n)} \frac{q}{\psi(q)} dq = \int_{v(t_n+s)}^n \frac{q}{\psi(q)} dq$$

(substituiere $v(t_n + t) = q$, also $dq/dt = v'(t_n + t) = -\psi(v(t_n + t)) = -\psi(q)$)

(3.12) zeigt $\psi(q) \leq Cq^2$ für eine Konstante $C < \infty$, somit gilt $\int_a^n \frac{q}{\psi(q)} dq \rightarrow \infty$ (für beliebiges $a < \infty$) für $n \rightarrow \infty$. Da $t_n \downarrow 0$ (und somit $v(t_n + s) \nearrow v(s) < \infty$) gilt, kann man $\delta_n \downarrow 0$ so wählen, dass

$$\frac{\int_{\delta_n}^s v^{(n)}(t) dt}{\int_0^{\delta_n} v^{(n)}(t) dt} \xrightarrow{n \rightarrow \infty} 0$$

gilt: Sei dazu $f(x) := \int_1^x \frac{q}{\psi(q)} dq$. $[1, \infty) \ni x \mapsto f(x)$ ist stetig und streng monoton wachsend ($\psi(q) > 0$) mit $\lim_{x \rightarrow \infty} f(x) = \infty$ (denn (3.12) zeigt, dass $\psi(q) \leq Cq^2$ für eine Konstante $C < \infty$). Daher existiert die Umkehrfunktion $f^{-1} : [0, \infty) \rightarrow [1, \infty)$ (und ist ebenfalls stetig und strikt wachsend).

Wir wählen eine Folge $m(n) \nearrow \infty$ so, dass

$$\frac{m(n) + f(m(n))}{f(n)} \xrightarrow{n \rightarrow \infty} 0$$

gilt, beispielsweise $m(n) := \min \{ \lfloor \sqrt{f(n)} \rfloor, \lfloor f^{-1}(\sqrt{f(n)}) \rfloor \}$, und setzen

$$\delta_n = t_{m(n)} - t_n$$

Damit ist

$$\int_0^{\delta_n} v^{(n)}(t) dt = \int_{m(n)}^n \frac{q}{\psi(q)} dq = f(n) - f(m(n)) = f(n)(1 + o(1))$$

und

$$\int_{\delta_n}^s v^{(n)}(t) dt = \int_{v(s)}^{m(n)} \frac{q}{\psi(q)} dq = O(f(m(n))) = o(f(n))$$

zudem gilt

$$\frac{m(n)}{\int_0^{\delta_n} v^{(n)}(t) dt} \xrightarrow{n \rightarrow \infty} 0$$

Da $\delta_n \rightarrow 0$ liefert Satz 3.19

$$\mathbb{P}_n \left(\sup_{t \leq \delta_n} \left| \frac{N_t}{v^{(n)}(t)} - 1 \right| \geq \frac{\varepsilon}{2} \right) \xrightarrow{n \rightarrow \infty} 0$$

(insbesondere auch $\mathbb{P}_n(N_{\delta_n} > (1 - \varepsilon/2)v^{(n)}(\delta_n)) = \mathbb{P}_n(N_{\delta_n} > (1 - \varepsilon/2)m(n)) \xrightarrow{n \rightarrow \infty} 0$), zusammen mit obigem folgt (3.16).

Da Annahme 3.15 impliziert, dass der zugehörige Λ -Koaleszent aus dem Unendlichen absteigt (anders gewendet: $(N_t)_{t \geq 0}$ besitzt ein Eintrittsgesetz bei ∞), vgl. Bericht 3.18, ist für jedes $s > 0$ die Familie $\mathbb{P}_n(N_s \in \cdot)$, $n \in \mathbb{N}$ straff und daher auch die Familie $\mathbb{P}_n(\int_s^{\tau_1^{(n)}} N_t dt \in \cdot)$, somit

$$\mathbb{P}_n \left(\frac{\int_s^{\tau_1^{(n)}} N_t dt}{\int_0^n v^{(n)}(t) dt} > \frac{\varepsilon}{2} \right) \xrightarrow{n \rightarrow \infty} 0$$

Zusammen mit (3.16) folgt die erste Aussage von Satz 3.16.

Da $L_n = \int_0^{\tau_1^{(n)}} N_t dt \sim \int_0^n v^{(n)}(t) dt$ in Wahrscheinlichkeit und S_n , gegeben L_n $\text{Pois}(L_n \theta/2)$ -verteilt ist, gilt mit Chebychev-Ungleichung

$$P_n(|S_n - L_n \theta/2| > \varepsilon L_n | L_n) \leq \frac{\theta/2}{\varepsilon^2 L_n \theta/2} \xrightarrow{n \rightarrow \infty} 0 \quad \text{in Wahrscheinlichkeit}$$

Zusammen mit der erste Aussage von Satz 3.16 impliziert dies die zweite. □

Frequenzspektrum unter Λ -Koaleszenten

Betrachte einen Λ - n -Koaleszenten, längs dessen Kanten (bis zum MRCA) sich mit Rate $\theta/2$ Mutationen (gemäß dem infinite sites-Modell) ereignen

S_n = Gesamtanzahl Mutationen

Frequenzspektrum:

$\xi_i^{(n)}$ = # Mutationen, die in i der n Blätter sichtbar sind ($i = 1, \dots, n-1$)

Für $\Lambda = \text{Beta}(2 - \alpha, \alpha)$ mit $\alpha \in (1, 2)$ gilt

$$\begin{aligned} \frac{S_n}{n^{2-\alpha}} &\longrightarrow \frac{\theta}{2} \frac{\alpha(\alpha-1)\Gamma(\alpha)}{2-\alpha} \\ \frac{\xi_i^{(n)}}{n^{2-\alpha}} &\longrightarrow \frac{\theta}{2} \alpha(\alpha-1)^2 \frac{\Gamma(i+\alpha-2)}{i!} \end{aligned}$$

für $n \rightarrow \infty$

(J. Berestycki, N. Berestycki, J. Schweinsberg, Beta-coalescents and continuous stable random trees, Ann. Probab. 35, 1835–1887 (2007); dasselbe Resultat mit geänderten Konstanten, wenn nur $\Lambda(dx) = f(x)dx$ und $f(x) \sim cx^{1-\alpha}$ für $x \downarrow 0$ vorausgesetzt wird)

$$\frac{\xi_i^{(n)}}{S_n} = \frac{\xi_i^{(n)}/n^{2-\alpha}}{S_n/n^{2-\alpha}} \longrightarrow \frac{\theta}{2} \alpha(\alpha-1)^2 \frac{\Gamma(i+\alpha-2)}{i!} \bigg/ \frac{\theta}{2} \frac{\alpha(\alpha-1)\Gamma(\alpha)}{2-\alpha}$$

für $n \gg 1$ also typischerweise

$$\frac{\xi_i^{(n)}}{S_n} \approx \frac{(2-\alpha)\Gamma(i+\alpha-2)}{i!\Gamma(\alpha-1)} =: a_i$$

Verteilung und Momente von $S_n, \xi_i^{(n)}$ sind nicht explizit bekannt, aber man kann mittels einer Rekursionsformel $\mathbb{E}[S_n], \mathbb{E}[\xi_i^{(n)}]$ numerisch bestimmen und somit auch

$$\frac{\mathbb{E}[\xi_i^{(n)}]}{\mathbb{E}[S_n]} =: b_i$$

(M.B., J. Blath, B. Eldon, Statistical properties of the site-frequency spectrum associated with Lambda-coalescents, Genetics 195, no. 3 (2013), 1037-1053.

(Erinnerung: Für den Kingman-Koaleszenten gilt $\mathbb{E}[\xi_i^{(n)}] = \theta/i, \mathbb{E}[S_n] = \theta \sum_{i=1}^{n-1} 1/i \sim \theta \log n$)

Einar Árnason, Mitochondrial Cytochrome *b* DNA Variation in the High-Fecundity Atlantic Cod: Trans-Atlantic Clines and Shallow Gene Genealogy, Genetics 166, 1871–1885 (2004)

berichtet genetische Variabilität an einem 250 Bp langen Stück des mitochondrialen Genoms in einer Stichprobe von $n = 1278$ atlantischen Kabeljaus:

$S = 39$ segregierende Positionen

i	1	2	3	4	5	6	7	8	9	10	11+
ξ_i	17	4	4	2	4	1	0	1	0	0	6
Km	5.0	2.5	1.7	1.3	1.0	0.8	0.7	0.6	0.6	0.5	24.2
Sb_i	18.5	4.6	2.3	1.5	1.0	0.8	0.6	0.5	0.4	0.4	8.4
Sa_i	20.3	4.9	2.4	1.5	1.0	0.8	0.6	0.5	0.4	0.3	5.8

Km.: Erwartung unter Kingman-Koaleszent,

Sb_i : Erwartung unter Beta($2 - \alpha, \alpha$)-Koaleszent ($\alpha = 1.48$),

Sa_i : Asymptotische Formel für Beta($2 - \alpha, \alpha$)-Koaleszent ($\alpha = 1.48$),

(Es gibt allerdings in den Daten Widersprüche zum IMS-Modell.)

Literaturverzeichnis

- [Bo9] N. Berestycki, Recent Progress in Coalescent Theory, *Ensaio Matemáticos*, Vol. 16, 1–193, (2009).
- [B] L. Breiman, *Probability*, Addison-Wesley, 1968.
- [C74] C. Cannings, The Latent Roots of Certain Markov Chains Arising in Genetics: A New Approach, I. Haploid Models, *Advances in Appl. Probability* 6, 260–290, (1974).
- [C] K.L. Chung, *Markov chains with stationary transition probabilities*, 2nd ed., Springer 1967.
- [EK] S.N. Ethier, T.G. Kurtz, *Markov processes: characterization and convergence*, Wiley, 1986.
- [KW] G. Kersting, A. Wakolbinger, *Stochastische Prozesse*, Birkhäuser, 2014.
- [KL] A. Klenke, *Wahrscheinlichkeitstheorie*, 4. Aufl., Springer 2020.
- [N] J. Norris, *Markov chains*, Cambridge University Press 1997.
- [RW] L.C.G. Rogers, D. Williams, *Diffusions, Markov processes and martingales*, Band I und II, Wiley, 1994.
- [St2] M. Birkner, Skript zur Stochastik II, WS 2014/15
- [St3] M. Birkner, Skript zur Stochastik III, SS 2015
- [St2] M. Birkner, Notizen zur Stochastik II, JGU Mainz, WS 2021/22.
- [T84] Simon Tavaré, Line-of-descent and genealogical processes, and their applications in population genetics models, *Theoretical Population Biology* 26 (2), 119–164, (1984).
- [W31] Sewall Wright, Evolution in Mendelian populations, *Genetics* 16, 97–159, 1931.
- [BBL10] Julien Berestycki, Nathanaël Berestycki, Vlada Limic, The Λ -coalescent speed of coming down from infinity, *Ann. Probab.* 38, no. 1, 207–233, (2010).
- [BBL14] Julien Berestycki, Nathanaël Berestycki, Vlada Limic, Asymptotic sampling formulae for Λ -coalescents, *Ann. Inst. H. Poincaré* 50, no. 3, 715–731, (2014)

- [SMPB16] Matthias Birkner, *Stochastische Modelle der Populationsbiologie*, Notizen zu einer Vorlesung an der JGU Mainz, WS 15/16, <https://www.staff.uni-mainz.de/birkner/SMPB1516/smpb1516.pdf>
- [BBE13] Matthias Birkner, Jochen Blath, Bjarki Eldon, An ancestral recombination graph for diploid populations with skewed offspring distribution, *Genetics* 193, no. 1, 255–290, (2013).
- [BS98] Erwin Bolthausen, Alain-Sol Sznitman, On Ruelle’s Probability Cascades and an Abstract Cavity Method, *Comm. Math. Physics* 197, no. 2, 247–276, (1998).
- [DK99] Peter Donnelly, Thomas G. Kurtz, Particle representations for measure-valued population models, *Ann. Probab.* 27, no. 1, 166–205, (1999).
- [EW06] Bjarki Eldon and John Wakeley, Coalescent processes when the distribution of offspring number among individuals is highly skewed, *Genetics* 172, 2621–2633, (2006).
- [HM13] Thierry Huillet, Martin Möhle, On the extended Moran model and its relation to coale-scents with multiple collisions, *Theor. Popul. Biol.* 87, 5–14, (2013).
- [KL99] Claude Kipnis, Claudio Landim, *Scaling Limits of Interacting Particle Systems*, Springer, 1999.
- [MS01] Martin Möhle, Serik Sagitov, A classification of coalescent processes for haploid exchange-able population models, *Ann. Probab.* 29, 1547–1562 (2001).
- [P99] Jim Pitman, Coalescents with multiple collisions, *Ann. Probab.* 27, no. 4, 1870–1902 (1999).
- [S99] Serik Sagitov, The general coalescent with asynchronous mergers of ancestral lines, *J. Appl. Prob.* 36, no. 4, 1116–1125, (1999).
- [S00] Jason Schweinsberg, Coalescents with Simultaneous Multiple Collisions *Electronic Journal of Probability* 5, paper 12, (2000).
- [Soob] Jason Schweinsberg, A Necessary and Sufficient Condition for the Lambda-Coalescent to Come Down from Infinity, *Electronic Communications in Probability* 5, p. 1–11 (2000)
- [So3] Jason Schweinsberg, Coalescent processes obtained from supercritical Galton-Watson pro-cesses, *Stoch. Proc. Appl.* 106, no. 1, 107–139, (2003).