Stochastische Modelle der Populationsbiologie

Skript zu einer Vorlesung an der
Johannes-Gutenberg-Universität Mainz,
Wintersemester 2015/2016

Matthias Birkner

Vorläufige Version, 12. Februar 2016

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

Inhaltsverzeichnis

Kapitel 1

Genealogien und (neutrale) genetische Variabilität

1.1 Cannings[1]-Modelle und Kingman[2]-Koaleszent

Ein (idealisiertes) Populationsmodell:

- feste Populationsgröße: N Individuen in jeder Generation
- nicht-überlappende Generationen
- jedes Individuum hat nur ein „Elter“[3]
- es gibt Zufälligkeit bezüglich der Anzahl der Nachkommen

Die Individuen jeder Generation seien durchnummeriert, sei

$$A_{r,i}^{(N)} := \text{Nr. des Vorfahren (in Gen. } r-1\text{) von Ind. Nr. } i \text{ in Generation } r,$$

somit ist

$$\nu_k^{(N,r)} := \left|\{1 \le i \le N : A_{r+1,i}^{(N)} = k\}\right| \tag{1.1}$$

die Anzahl Nachkommen von Individuum k in Generation r.

Notation. Wir schreiben im Folgenden $[n] := \{1, 2, \ldots, n\}$ für $n \in \mathbb{N}$.

[1] Chris Cannings, The Latent Roots of Certain Markov Chains Arising in Genetics: A New Approach, I. Haploid Models, Advances in Appl. Probability 6 (1974), 260–290.

[2] J.F.C. Kingman, The coalescent, Stochastic Process. Appl. 13 (1982), no. 3, 235–248. Für historischen Hintergrund zum Koaleszenten siehe auch J.F.C. Kingman, Origins of the Coalescent: 1974-1982, Genetics 156:1461–1463, (2000).

[3] Im Jargon der Genetik sind die Individuen „haploid“ – wörtlich angemessen z.B. für Bakterien, mitochondriale genetische Typen, Y-Chromosom. Viele Spezies sind „diploid“, besitzen also zwei Kopien jedes Chromosoms [ggfs. mit Ausnahme der Geschlechtschromosomen], manche Pflanzen sind „polyploid“. Asymptotisch, mit Ersetzung $N \rightsquigarrow 2N$, ist das Modell aber auch für Gene in diploiden Populationen passend.

Beispiel 1.1 (Wright-Fisher-Modell[4], genealogische Version)**.** Die $A_{r,i}^{(N)}$ seien für $i \in [N]$, $r \in \mathbb{Z}$ unabhängig und uniform verteilt auf $[N]$ (d.h. jedem Kind wird rein zufällig ein Elter zugeordnet).

Dann gilt für $m_1, \dots, m_N \in \mathbb{N}_0$ (mit $m_1 + \cdots + m_N = N$)

$$\mathbb{P}\big(\nu_1^{(N,r)} = m_1, \dots, \nu_N^{(N,r)} = m_N\big) = \frac{N!}{m_1! m_2! \cdots m_N!}\Big(\frac{1}{N}\Big)^N,$$

d.h. $\nu^{(N,r)} = \big(\nu_1^{(N,r)}, \cdots, \nu_N^{(N,r)}\big)$ ist (für jede Generation r) Multinom$(N, \frac{1}{N}, \cdots, \frac{1}{N})$-verteilt.

Definition 1.2. $(A_r^{(N)})_{r\in\mathbb{Z}} = \big((A_{r,1}^{(N)}, \dots, A_{r,N}^{(N)})\big)_{r\in\mathbb{Z}}$ heißt *Ahnen-Prozess* eines Cannings-Modells (mit Populationsgröße N), falls gilt

i) die $(A_r^{(N)})_{r\in\mathbb{Z}}$ sind unabhängig und identisch verteilt, $[N]^N$-wertig,

ii) der Nachkommens(anzahl)vektor $\nu^{(N,r)} = \big(\nu_1^{(N,r)}, \dots, \nu_N^{(N,r)}\big)$ aus (1.1) ist austauschbar, d.h.

$$\big(\nu_1^{(N,r)}, \dots, \nu_N^{(N,r)}\big) \overset{d}{=} \big(\nu_{\pi(1)}^{(N,r)}, \dots, \nu_{\pi(N)}^{(N,r)}\big) \quad \text{für jede Permuation } \pi \text{ von } [N]$$

mit $\nu_1^{(N,r)} + \cdots + \nu_N^{(N,r)} = N$.

Beobachtung 1.3. Man kann das Modell leicht um (neutrale) genetische Typen ergänzen: Nehmen wir an, jedes Individuum besitzt einen Typ aus einer Menge E möglicher genetischer Typen („Allele"), z.B. $E = \{0, 1\}$ für eine zwei-Typ-Situation oder $E = \{A, G, C, T\}$ (die „Buchstaben des genetischen Alphabets"). Wir schreiben $t_{r,i}^{(N)}$ für den Typ von Individuum i in Generation r und nehmen (jedenfalls in diesem Kapitel) an, dass der genetische Typ vom Elter übernommen wird (ohne sog. „Mutationen").

Startend von einer Typenbelegung in Generation r_0 ist die Entwicklung des Typenvektors $\big(t_r^{(N)}\big)_{r\ge 0} := \big((t_{r,1}^{(N)}, \dots, t_{r,N}^{(N)})\big)_{r\ge 0}$ dann offenbar rekursiv gegeben durch

$$t_{r,i}^{(N)} = t_{r-1,A_{r,i}^{(N)}}^{(N)} \in E, \quad i \in [N],\ r > r_0.$$

Wir können daraus den Typenzählprozess $X_r^{(N)} = (X_{r,e}^{(N)})_{e\in E}$, $r \ge r_0$ ablesen:

$$X_{r,e}^{(N)} := \sum_{i=1}^{N} \mathbf{1}_{\{t_{r,i}^{(N)} = e\}}$$

gibt an, wieviele Individuen in der r-ten Generation den Typ $e \in E$ besitzen.

Der Typenzählprozess $(X_r^{(N)})_{r=r_0, r_0+1, \dots}$ ist eine Markovkette, wir werden sie später noch genauer studieren. Betrachten wir für den Moment das Wright-Fisher-Modell (Bsp. 1.1) und den Fall $E = \{0, 1\}$. Offenbar genügt es, $X_{r,1}^{(N)}$ zu kennen, denn $X_{r,0}^{(N)} = N - X_{r,1}^{(N)}$. Für $x, y \in \{0, 1, \dots, N\}$ gilt

$$\mathbb{P}\big(X_{r+1,1}^{(N)} = y \,\big|\, X_{r,1}^{(N)} = x\big) = \binom{N}{y}\big(\frac{x}{N}\big)^y\big(1 - \frac{x}{N}\big)^{N-y},$$

[4]Nach Sewall Wright, 1889–1988, und Ronald Fisher, 1890–1962, benannt.

d.h. gegeben $X_{r,1}^{(N)} = x$ ist $X_{r+1,1}^{(N)} \sim \mathrm{Bin}(N, x/N)$. Die entscheidende Beobachtung dazu ist, dass für beliebige $t = (t_1, \dots, t_N), t' = (t'_1, \dots, t'_N) \in \{0,1\}^N$ mit $\sum_{i=1}^N t_i = x$, $\sum_{i=1}^N t'_i = y$ gilt (wir schreiben $I(t) = \{i \in [N] : t_i = 1\}$ und analog $I(t')$)

$$\begin{aligned}
&\mathbb{P}\big(t_{r+1}^{(N)} = t' \,\big|\, t_r^{(N)} = t\big) \\
&= \mathbb{P}\big(A_{r+1,i}^{(N)} \in I(t) \text{ für } i \in I(t') \text{ und } A_{r+1,j}^{(N)} \in [N] \setminus I(t) \text{ für } j \in [N] \setminus I(t')\big) \\
&= \mathbb{P}\Big(\bigcap_{i \in I(t')} \{A_{r+1,i}^{(N)} \in I(t)\} \cap \bigcap_{j \in [N] \setminus I(t')} \{A_{r+1,j}^{(N)} \in [N] \setminus I(t)\}\Big) \\
&= \prod_{i \in I(t')} \mathbb{P}\big(A_{r+1,i}^{(N)} \in I(t)\big) \times \prod_{j \in [N] \setminus I(t')} \mathbb{P}\big(A_{r+1,j}^{(N)} \in [N] \setminus I(t)\big) = \big(\frac{x}{N}\big)^y \big(1 - \frac{x}{N}\big)^{N-y},
\end{aligned}$$

somit

$$\mathbb{P}\big(X_{r+1,1}^{(N)} = y \,\big|\, t_r^{(N)} = t\big) = \sum_{\substack{t' \in \{0,1\}^N, \\ |I(t')| = y}} \mathbb{P}\big(t_{r+1}^{(N)} = t' \,\big|\, t_r^{(N)} = t\big) = \binom{N}{y} \big(\frac{x}{N}\big)^y \big(1 - \frac{x}{N}\big)^{N-y}.$$

Annahme 1.4. Zumeist denken wir uns die bedingte Verteilung von $A_r^{(N)}$, gegeben den Anzahlvektor $\nu^{(N,r-1)}$, als vollkommen symmetrisch, d.h.

- für $m_1, \dots, m_N \in \mathbb{N}_0$ mit $m_1 + \cdots + m_N = N$ und jedes $(a_1, \dots, a_N) \in [N]^N$, das $|\{i : a_i = k\}| = m_k$ für $k = 1, \dots, N$ erfüllt, gilt

$$\mathbb{P}\big(A_{r,1}^{(N)} = a_1, \dots, A_{r,N}^{(N)} = a_N \,\big|\, \nu_1^{(N,r-1)} = m_1, \dots, \nu_N^{(N,r-1)} = m_N\big) = \frac{m_1! m_2! \cdots m_N!}{N!} \quad (1.2)$$

(man denke an eine Urne, die m_1 Kugeln der Farbe 1, …, m_N Kugeln der Farbe N enthält und aus der alle N Kugeln nacheinander ohne Zurücklegen gezogen werden, dann hat jede mögliche beobachtete Farbreihenfolge die Wahrscheinlichkeit (1.2)).

Gelegentlich sind allerdings auch andere Festlegungen nützlich, man beachte, dass für Verteilung des Typenhäufigkeitsprozesses für jede andere Wahl in (1.2) dieselbe ist.

Ahnenverhältnisse Sei $A_{r,i}^{(N)}[k]$ die Nummer des Ahnen vor k Generationen von Individuum Nr. i in Generation r, diese ist offenbar

$$\text{rekursiv bestimmt durch } A_{r,i}^{(N)}[1] = A_{r,i}^{(N)} \text{ und } A_{r,i}^{(N)}[k+1] = A_{r-k, A_{r,i}^{(N)}[k]}^{(N)} \text{ für } k \in \mathbb{N}.$$

Wir betrachte eine Stichprobe von n verschiedenen (zufällig gezogenen) Individuen aus Generation $r = 0$, sagen wir die Individuen Nr. $J_1, \dots, J_n$ mit $\mathbb{P}(J_1 = j_1, \dots, J_n = j_n) = 1/(N)_{n\downarrow}$ für paarweise verschiedene $j_1, \dots, j_n \in [N]$.

Die Verwandtschaftsverhältnisse innerhalb der Stichprobe kodieren wir durch

$$R_k^{(N,n)}, \text{ eine (zufällige) Äquivalenzrelation,}$$

gegeben durch $i \sim_k j$ $(i, j \in [n],\ k = 0, 1, \dots)$, wenn $A_{0,J_i}^{(N)}[k] = A_{0,J_j}^{(N)}[k]$ gilt, d.h. Stichproben i und j haben denselben Ahnen vor k Generationen.

Sei $\mathcal{E}_n := \{\text{Äquivalenzrelationen auf } [n]\}$, wir notieren $\xi \in \mathcal{E}_n$ etwa durch eine (ungeordnete) Liste der Äquivalenzklassen (z.B. $\xi = \{\{1\},\{2,3\}\} \in \mathcal{E}_3$ bedeutet $2 \sim_\xi 3$, $1 \not\sim_\xi 2$, $1 \not\sim_\xi 3$).

Wir schreiben $\xi \preceq \eta$, falls

$$i \sim_\xi j \quad \Longrightarrow \quad i \sim_\eta j \qquad \text{gilt,}$$

d.h. η entsteht aus ξ durch Vereinigung einiger Klassen, ggfs. in mehreren Gruppen (z.B. $\{\{1\},\{2\},\{3,4\},\{5\},\{6,7\},\{8\}\} \preceq \{\{1,2,6,7\},\{3,4,5\},\{8\}\}$).

Offensichtlich ist $i \sim_0 j \iff i = j$, d.h. $R_0^{(N,n)} = \big\{\{1\},\{2\},\dots,\{n\}\big\}$ und es gilt stets $R_k^{(N,n)} \preceq R_{k+1}^{(N,n)}$.

Lemma 1.5. *Für festes $N \geq n$ ist $(R_k^{(N,n)})_{k\in\mathbb{N}_0}$ eine Markovkette mit Werten in $\mathcal{E}_n$. Die Übergangswahrscheinlichkeiten sind gegeben durch*

$$p^{(N,n)}(\xi,\eta) \;:=\; \mathbb{P}\big(R_{k+1}^{(N,n)} = \eta \,\big|\, R_k^{(N,n)} = \xi\big) = \frac{(N)_{a\downarrow}}{(N)_{b\downarrow}} \mathbb{E}\Big[\prod_{j=1}^{a} (\nu_j^{(N)})_{b_j\downarrow}\Big],$$

sofern $\xi \preceq \eta$, wobei η aus $|\eta| = a$ Klassen besteht, ξ aus $b = |\xi| = b_1 + \dots + b_a$ Klassen besteht und η aus ξ durch Verschmelzen von a Gruppen von Klassen in Gruppengrößen $b_1,\dots,b_a$ entsteht (d.h. $\eta = \{C_1,\dots,C_a\}$ und $\xi = \{C_{\alpha\beta} : 1 \leq \alpha \leq a, 1 \leq \beta \leq b_\alpha\}$ mit $C_\alpha = \cup_{\beta=1}^{b_\alpha} C_{\alpha\beta}$ für $\alpha = 1,\dots,a$).

($(x)_{a\downarrow} := x(x-1)\cdots(x-a+1)$ für $x \in \mathbb{R}$, $a \in \mathbb{N}$ ist die a-te fallende Faktorielle von x.)

Beweis. $R_k^{(N,n)} = \xi$ bedeutet, dass es (k Generationen vor der Gegenwart) b verschiedene „aktive Ahnenlinien“ geben muss, d.h. es gibt b paarweise verschiedene Zahlen $i_{\alpha,\beta} \in [N]$, $\beta = 1,\dots,b_\alpha$, $\alpha = 1,\dots,a$, so dass

$$A_{0,J_i}^{(N)}[k] = i_{\alpha,\beta} \text{ für } i \in C_{\alpha\beta}, \;\; \beta = 1,\dots,b_\alpha, \; \alpha = 1,\dots,a$$

gilt. Gegeben dies tritt das Ereignis $\{R_{k+1}^{(N,n)} = \eta\}$ genau dann ein, wenn es paarweise verschiedene $j_1, j_2, \dots, j_a \in [N]$ gibt mit

$$A_{k,i_{\alpha,\beta}}^{(N)} = j_\alpha \text{ für } \beta = 1,\dots,b_\alpha, \; \alpha = 1,\dots,a. \tag{1.3}$$

Wenn wir einen möglichen Wert $(k_1,\dots,k_N) \in \{0,1,\dots,N\}^N$ (mit $k_1 + \dots + k_N = N$) des Nachkommensanzahlvektors $\nu^{(N,-k-1)}$ fixieren, so hat das Ereignis in (1.3), gegeben $\nu^{(N,-k-1)} = (k_1,\dots,k_N)$, gemäß Annahme 1.4 die Wahrscheinlichkeit

$$\begin{aligned}
&\frac{k_{j_1}(k_{j_1}-1)\cdots(k_{j_1}-b_1+1)}{N(N-1)\cdots(N-b_1+1)} \cdot \frac{k_{j_2}(k_{j_2}-1)\cdots(k_{j_2}-b_2+1)}{(N-b_1)(N-b_1-1)\cdots(N-b_1-b_2+1)} \cdots \\
&\quad \cdot \frac{k_{j_a}(k_{j_a}-1)\cdots(k_{j_a}-b_a+1)}{(N-b_1-\dots-b_{a-1})(N-b_1-\dots-b_{a-1}-1)\cdots(N-b_1-b_2-\dots-b_a-1)} \\
&= \frac{1}{(N)_{b\downarrow}} \prod_{\ell=1}^{a} (k_{j_\ell})_{b_\ell\downarrow}
\end{aligned}$$

(denn b_1 aktive Ahnenlinien müssen in Generation $-k-1$ von Ind. j_1 abstammen, b_2 viele von j_2, etc.).

Somit ist

$$\begin{aligned} p^{(N,n)}(\xi,\eta) &= \mathbb{P}\big(R_{r+1}^{(N,n)} = \eta \,\big|\, R_r^{(N,n)} = \xi\big) \\ &= \sum_{\substack{j_1,\dots,j_a \\ \text{paarw. versch.}}} \sum_{\substack{k_1,\dots k_N=1 \\ k_1+\dots+k_N=N}}^{N} \frac{1}{(N)_{b\downarrow}} \prod_{\ell=1}^{a} (k_{j_\ell})_{b_\ell\downarrow} \mathbb{P}\big(\nu^{(N,-k-1)} = (k_1,\dots,k_N)\big) \\ &= \sum_{\substack{j_1,\dots,j_a \\ \text{paarw. versch.}}} \mathbb{E}\Big[\prod_{\ell=1}^{a} (\nu_{j_\ell}^{(N,-k-1)})_{b_\ell\downarrow}\Big] = \frac{(N)_{a\downarrow}}{(N)_{b\downarrow}} \mathbb{E}\Big[\prod_{j=1}^{a} (\nu_j^{(N)})_{b_j\downarrow}\Big], \end{aligned}$$

wobei wir in der letzten Gleichung die Austauschbarkeit des Nachkommensanzahlvektors ausnutzen (und in der Notation $\nu_j^{(N)} = \nu_j^{(N,-k-1)}$ die Zeitabhängigkeit unterdrücken – nach Voraussetzung hängt die Verteilung von $\nu^{(N,-k-1)}$ nicht von k ab, und bei der Berechnung der Erwartungswerte kommt es nur auf die Verteilung an).

□

Typischerweise wird bei großem N die ein-Schritt-Übergangswahrscheinlichkeit $p^{(N,n)}(\xi,\eta)$ für alle $\eta \neq \xi$ sehr klein sein: Betrachten wir z.B. den Fall $n = 2$, so ist nach Lemma 1.5

$$c_N := p^{(N,2)}\big(\{\{1\},\{2\}\},\{1,2\}\big) = \frac{1}{N-1}\mathbb{E}\big[\nu_1(\nu_1-1)\big]$$

(für das Wright-Fisher-Modell ergibt sich $c_N = 1/N$). $\tau_1^{(N,2)} := \inf\{k \in \mathbb{N}_0 : 1 \sim_k 2\}$, die Anzahl Generationen in die Vergangenheit, bis die beiden Stichproben ihren ersten gemeinsamen Vorfahren finden, ist $\sim \mathrm{geom}(c_N)$ und sofern $c_N \to 0$ für $N \to \infty$ gilt, folgt

$$\mathbb{P}\big(R_{\lfloor t/c_N\rfloor}^{(N,2)} \neq \{1,2\}\big) = \mathbb{P}\big(\tau_1^{(N,2)} > \lfloor t/c_N\rfloor\big) = (1-c_N)^{\lfloor t/c_N\rfloor} \underset{N\to\infty}{\longrightarrow} e^{-t} \quad \text{für jedes } t \geq 0.$$

Dies legt nahe, den Prozess der reskalierten Ahnenverhältnisse $(R_{\lfloor t/c_N\rfloor}^{(N,n)})_{t\geq 0}$ zu betrachten. Für dazu notwendige Techniken schieben wir den folgenden Exkurs ein.

Ein Exkurs zum Poissonprozess und zu zeitkontinuierlichen Markovketten

Sei $c_N \longrightarrow_{N\to\infty} 0$ eine Nullfolge, $\lambda > 0$, $Z_i^{(N)}$, $i \in \mathbb{N}$ u.i.v. $\sim \mathrm{Ber}(\lambda c_N)$ (wir betrachten o.E. nur so große N, dass $\lambda c_N \leq 1$), seien

$$T_0^{(N)} := 0, \quad T_\ell^{(N)} := \inf\big\{i > T_{\ell-1}^{(N)} : Z_i^{(N)} = 1\big\},\ \ell \in \mathbb{N}$$

($T_\ell^{(N)}$ ist der Zeitpunkt des ℓ-ten Erfolgs in der Münzwurffolge $(Z_i^{(N)})_{i\in\mathbb{N}}$), dann sind

$$\tau_\ell^{(N)} := T_\ell^{(N)} - T_{\ell-1}^{(N)},\ \ell \in \mathbb{N}$$

u.i.v., $\tau_\ell^{(N)} \sim \mathrm{geom}(\lambda c_N)$, d.h. $\mathbb{P}(\tau_\ell^{(N)} = j) = c_N\lambda(1-c_N\lambda)^{j-1}$ für $j \in \mathbb{N}$ und für $x \geq 0$ gilt

$$\mathbb{P}(c_N\tau_\ell^{(N)} > x) = \mathbb{P}\big(\tau_\ell^{(N)} > \lfloor \frac{x}{c_N}\rfloor\big) = \big(1-c_N\lambda\big)^{\lfloor x/c_N\rfloor} \underset{N\to\infty}{\longrightarrow} e^{-\lambda x},$$

d.h. $c_N \tau_\ell^{(N)} \xrightarrow[N\to\infty]{d} \mathrm{Exp}(\lambda)$ (Übung: Beweisen Sie diese Aussagen).

Sei weiter

$$M_k^{(N)} := \big|\{1 \le i \le k : Z_i^{(N)} = 1\}\big| = \max\{\ell \in \mathbb{N}_0 : T_\ell^{(N)} \le k\},$$

offenbar gilt für $0 \le k_0 < k_1 < \cdots < k_m$

$$M_{k_1}^{(N)} - M_{k_0}^{(N)}, M_{k_2}^{(N)} - M_{k_1}^{(N)}, \ldots, M_{k_m}^{(N)} - M_{k_{m-1}}^{(N)} \quad \text{sind unabhängig}$$

und für $0 \le k < k'$ ist $M_{k'}^{(N)} - M_k^{(N)} \sim \mathrm{Bin}\big(k' - k, c_N\lambda\big)$, somit gilt für $0 \le t < t'$

$$M_{\lfloor t'/c_N \rfloor}^{(N)} - M_{\lfloor t/c_N \rfloor}^{(N)} \xrightarrow[N\to\infty]{d} \mathrm{Pois}\big(\lambda(t'-t)\big).$$

(Übung: Beweisen Sie diese Aussagen).

Dies lädt ein, folgendes Limesobjekt zu betrachten: Sei $\tau_1, \tau_2, \ldots$ u.i.v., $\tau_\ell \sim \mathrm{Exp}(\lambda)$, $T_0 := 0, T_\ell := \tau_1 + \cdots + \tau_\ell$, $\ell \in \mathbb{N}$,

$$M_t := \max\{i \in N_0 : T_i \le t\}, \quad t \in [0, \infty)$$

der stochastische Prozess $(M_t)_{t\ge 0}$ heißt *Poissonprozess* mit Rate λ. (Beachte: die Definition ist so eingerichtet, dass $t \mapsto M_t$ rechtsstetig ist, man sagt auch: $(M_t)_t$ hat rechtsstetige Pfade.)

Aus obigen Überlegungen folgt für jedes $m \in \mathbb{N}$

$$\big(c_N\tau_1^{(N)}, \ldots, c_N\tau_m^{(N)}\big) \xrightarrow[N\to\infty]{d} \big(\tau_1, \ldots, \tau_m\big),$$

$$\big(c_N T_1^{(N)}, \ldots, c_N T_m^{(N)}\big) \xrightarrow[N\to\infty]{d} \big(T_1, \ldots, T_m\big)$$

somit ergibt sich für $t_1 < t_2 < t_m$, $k_1, \ldots, k_m \in \mathbb{N}_0$

$$\begin{aligned}
&\mathbb{P}\big(M_{\lfloor t_1/c_N \rfloor}^{(N)} = k_1, \ldots, M_{\lfloor t_m/c_N \rfloor}^{(N)} = k_m\big) = \mathbb{P}\big(T_{k_1}^{(N)} \le \lfloor t_1/c_N \rfloor < T_{k_1+1}^{(N)}, \ldots, T_{k_m}^{(N)} \le \lfloor t_m/c_N \rfloor < T_{k_m+1}^{(N)}\big) \\
&\xrightarrow[N\to\infty]{d} \mathbb{P}\big(T_{k_1} \le t_1 < T_{k_1+1}, \ldots, T_{k_m} \le t_m < T_{k_m+1}\big) = \mathbb{P}\big(M_{t_1} = k_1, \ldots, M_{t_m} = k_m\big),
\end{aligned}$$

d.h. die Folge von stochastischen Prozessen $(M_{\lfloor t/c_N \rfloor}^{(N)})_{t\ge 0}$ konvergiert gegen den Prozess $(M_t)_{t\ge 0}$ im Sinne der endlich-dimensionalen Verteilungen.

Aus diesen Beobachtungen folgt

$$\begin{aligned}
&\mathbb{P}\big(M_{t_1} - M_{t_0} = j_1, M_{t_2} - M_{t_1} = j_2 \ldots, M_{t_m} - M_{t_{m-1}} = j_m\big) \\
&= \lim_{N\to\infty} \mathbb{P}\big(M_{\lfloor t_1/c_N \rfloor}^{(N)} - M_{\lfloor t_0/c_N \rfloor}^{(N)} = j_1, \ldots, M_{\lfloor t_m/c_N \rfloor}^{(N)} - M_{\lfloor t_{m-1}/c_N \rfloor}^{(N)} = j_m\big) \\
&= \lim_{N\to\infty} \mathbb{P}\big(M_{\lfloor t_1/c_N \rfloor}^{(N)} - M_{\lfloor t_0/c_N \rfloor}^{(N)} = j_1\big) \times \cdots \times \mathbb{P}\big(M_{\lfloor t_m/c_N \rfloor}^{(N)} - M_{\lfloor t_{m-1}/c_N \rfloor}^{(N)} = j_m\big) \\
&= \prod_{i=1}^{m} e^{-\lambda(t_i - t_{i-1})} \frac{(\lambda(t_i - t_{i-1}))^{j_i}}{j_i!},
\end{aligned}$$

d.h. die Inkremente eines Poissonprozesses (M_t) sind Poissonverteilt [der Parameter ist $\lambda\times$ die Länge des betrachteten Zeitintervalls] und Inkremente über jeweils disjunkte Zeitintervalle sind unabhängig. Diese beiden Eigenschaften charakterisieren den Poissonprozess [ggfs. mit Forderung der Rechtsstetigkeit].

Der Parameter λ kann als Sprungrate interpretiert werden in dem Sinne, dass für ein (kurzes) Zeitintervall $(t, t+h]$ die Wahrscheinlichkeit, einen Sprung in diesem Zeitintervall zu sehen, $\approx \lambda \times$ Intervalllänge ist, genauer

$$\mathbb{P}\big(M_{t+h} = k+1 \,\big|\, M_t = k\big) = \mathbb{P}\big(M_{t+h} - M_t = 1\big) = e^{-\lambda h}\frac{\lambda h}{1!} = \lambda h + O(h^2) \quad \text{für } h \downarrow 0.$$

Zu allgemeinen zeitkontinuierlichen Markovketten Sei E endliche Menge, $\widehat{p} = (\widehat{p}(x,y))_{x,y\in E}$ stochastische Matrix (d.h. $\widehat{p}(x,y) \geq 0$, $\sum_{y\in E}\widehat{p}(x,y) = 1$ für alle $x \in E$), $\widehat{X} = (\widehat{X}_n)_{n\in\mathbb{N}_0}$ (zeitdiskrete, homogene) $\widehat{p}$-Markovkette (d.h. $\mathbb{P}(\widehat{X}_0 = x_0, \widehat{X}_1 = x_1, \ldots, \widehat{X}_n = x_n) = \mathbb{P}(\widehat{X}_0 = x_0)\widehat{p}(x_0,x_1)\times\cdots\times\widehat{p}(x_{n-1},x_n)$ für $x_0, x_1, \ldots, x_n \in E$).

Sei $(M_t)_{t\geq 0}$ Poissonprozess mit Rate $\lambda > 0$, unabhängig von $\widehat{X}$,

$$X_t := \widehat{X}_{M_t}, \quad t \in [0,\infty),$$

so ist [$\widehat{p}^m$ bezeichne die m-te Potenz von $\widehat{p}$, I die $E\times E$-Identitätsmatrix]

$$\begin{aligned}
p_t(x,y) &:= \mathbb{P}\big(X_t = y \,\big|\, X_0 = x\big) = \sum_{m=0}^{\infty}\mathbb{P}\big(M_t = m, X_t = y \,\big|\, X_0 = x\big)\\
&= \sum_{m=0}^{\infty} e^{-\lambda t}\frac{(\lambda t)^m}{m!}\widehat{p}^m(x,y) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\frac{(\lambda t)^n}{n!}\frac{(\lambda t)^m}{m!}\big((-I)^n\widehat{p}^m\big)(x,y)\\
&= \sum_{\ell=0}^{\infty}\frac{t^\ell}{\ell!}\lambda^\ell\sum_{m=0}^{\ell}\binom{\ell}{m}\big(\widehat{p}^m(-I)^{\ell-m}\big)(x,y) = \sum_{\ell=0}^{\infty}\frac{t^\ell}{\ell!}\big(\lambda(\widehat{p}-I)\big)^\ell(x,y) = \sum_{\ell=0}^{\infty}\frac{t^\ell}{\ell!}Q^\ell(x,y) = \big(e^{tQ}\big)(x,y)
\end{aligned}$$

[wobei die Matrix $Q = (q_{x,y})_{x,y\in E}$ Einträge $q_{x,y} = \lambda(\widehat{p}(x,y) - \delta_{xy})$ besitzt; obige Reihe konvergiert, denn $\max_{x,y\in E}|Q^n_{xy}| \leq (|E|\max_{x,y\in E}|Q_{x,y}|)^n$; beachte auch, dass $\widehat{p}^m$ und I^n kommutieren] und analoge Rechnungen, die die Unabhängigkeit der Zuwächse von $(M_t)_{t\geq 0}$ ausnutzen, zeigen

$$\mathbb{P}\big(X_{t_1} = x_1, \ldots, X_{t_n} = x_n \,\big|\, X_0 = x_0\big) = \prod_{i=1}^{n} p_{t_i - t_{i-i}}(x_{i-1}, x_i)$$

für $0 = t_0 < t_1 < \cdots < t_n$, $x_0, x_1, \ldots, x_n \in E$.

Die Matrix Q heißt die *Sprungratenmatrix* (auch: *Ratenmatrix* oder Q-Matrix) der zeitkontinuierlichen Markovkette X, sie hat die Eigenschaften

$$q_{x,y} \geq 0 \text{ für } x \neq y, \qquad \sum_{y\in E, y\neq x} q_{x,y} = -q_{x,x}.$$

Zur Interpretation der Einträge von Q als Sprungraten: Für $x \neq y$ und $h \downarrow 0$ ist

$$\mathbb{P}\big(X_{t+h} = y \,\big|\, X_t = x\big) = \big(e^{hQ}\big)(x,y) = Q^0(x,y) + hQ^1(x,y) + O(h^2) = hq_{x,y} + O(h^2).$$

Lemma 1.6. *E endliche Menge, $Q = (q_{xy})_{x,y\in E}$ Sprungratenmatrix, $X^{(N)}$, $N \in \mathbb{N}$ zeitdiskrete E-wertige Markovketten mit Übergangsmatrix*

$$p^{(N)}(x,y) = \delta_{x,y} + c_N q_{xy} + o(c_N), \quad x,y \in E,$$

wo $c_N \to 0$ für $N \to \infty$ und $X_0^{(N)} = x_0 \in E$. Dann konvergieren die (zeitlich reskalierten) Prozesse $(X^{(N)}_{\lfloor t/c_N \rfloor})_{t\geq 0}$ für $N \to \infty$ gegen die zeitkontinuierliche Markovkette X mit Sprungratenmatrix Q (im Sinne der endlich-dimensionalen Verteilungen).

Beweis. Wir schreiben die Übergangsmatrix von $X^{(N)}$ als

$$p^{(N)} = I + c_N Q_N$$

mit $Q_N := c_N^{-1}(p^{(N)} - I)$, somit nach Voraussetzung $Q_N \underset{N\to\infty}{\longrightarrow} Q$ (eintrags-weise).

$$\begin{aligned}(I + c_N Q_N)^{\lfloor c_N^{-1} t\rfloor} &= \sum_{k=0}^{\lfloor c_N^{-1}t\rfloor} \binom{\lfloor c_N^{-1}t\rfloor}{k} c_N^k Q_N^k = \sum_{k=0}^{\lfloor c_N^{-1}t\rfloor} c_N^k \frac{\lfloor c_N^{-1}t\rfloor(\lfloor c_N^{-1}t\rfloor - 1)\cdots(\lfloor c_N^{-1}t\rfloor - k + 1)}{k!} Q_N^k \\ &\to \sum_{k=0}^{\infty} \frac{t^k}{k!} Q^k = e^{tQ} \qquad \text{für } N \to \infty.\end{aligned}$$

Beachte: Für $k \in \mathbb{N}$ gilt

$$c_N^k \frac{\lfloor c_N^{-1}t\rfloor(\lfloor c_N^{-1}t\rfloor - 1)\cdots(\lfloor c_N^{-1}t\rfloor - k + 1)}{k!} Q_N^k \underset{N\to\infty}{\longrightarrow} \frac{t^k}{k!} Q^k$$

(eintrags-weise) und die Beträge der Einträge der Matrix auf der linken Seite sind für genügend großes N

$$\leq t^k (2\max\{|Q(x,y)| : x,y \in E\})^k / k!,$$

so dass Grenzwert und Summation vertauscht werden können. Somit

$$\begin{aligned}\mathbb{P}\big(X^{(N)}_{\lfloor c_N^{-1}t_1\rfloor} = x_1, \ldots, X^{(N)}_{\lfloor c_N^{-1}t_n\rfloor} = x_n\big) &= \prod_{j=1}^{n} (I + c_N Q_N)^{\lfloor c_N^{-1}(t_j - t_{j-1})\rfloor}(x_{j-1}, x_j) \\ &\underset{N\to\infty}{\longrightarrow} \prod_{j=1}^{n} (e^{(t_j - t_{j-1})Q})(x_{j-1}, x_j) = \mathbb{P}\big(X_{t_1} = x_1, \ldots, X_{t_n} = x_n\big).\end{aligned}$$

□

Siehe auch Übungsblatt 1 für weitere Eigenschaften zeitkontinuierlicher Markovketten und insbesondere Aufg. 1.4 für eine stärkere Form der Konvergenz (nämlich als zufälliger Pfad) als die in Lemma 1.6 formulierte Aussage.

Zurück zu den Ahnenverhältnissen in stochastischen Populationsmodellen (und ihrer Reskalierung)

Zur Motivation betrachten wir zunächst wieder das Wright-Fisher-Modell, dort hatten wir gesehen: Für eine Stichprobe der Größe $n = 2$ ist die „korrekte“ Zeitskala der Genealogie [Vielfache von] N, denn die Paarverschmelzungsw'keit ist $p^{(N,2)}\big(\{\{1\},\{2\}\},\{1,2\}\big) = \frac{1}{N}$ und somit die Zeit, bis die Stichprobe ihren ersten gemeinsamen Vorfahren findet, $\sim$ geom$(1/N)$.

Für $n = 3$ sieht die Übergangsmatrix $p^{(N,3)}(\cdot,\cdot)$ von $R^{(N,3)}$ folgendermaßen aus:

	$\{\{1\},\{2\},\{3\}\}$	$\{\{1,2\},\{3\}\}$	$\{\{1,3\},\{2\}\}$	$\{\{1\},\{2,3\}\}$	$\{\{1,2,3\}\}$
$\{\{1\},\{2\},\{3\}\}$	$1-3\frac{1}{N}+2\frac{1}{N^2}$	$\frac{1}{N}(1-\frac{1}{N})$	$\frac{1}{N}(1-\frac{1}{N})$	$\frac{1}{N}(1-\frac{1}{N})$	$\frac{1}{N^2}$
$\{\{1,2\},\{3\}\}$	0	$1-\frac{1}{N}$	0	0	$\frac{1}{N}$
$\{\{1,3\},\{2\}\}$	0	0	$1-\frac{1}{N}$	0	$\frac{1}{N}$
$\{\{1\},\{2,3\}\}$	0	0	0	$1-\frac{1}{N}$	$\frac{1}{N}$
$\{\{1,2,3\}\}$	0	0	0	0	1

d.h.

$$p^{(N,3)} = I + \frac{1}{N}\begin{pmatrix} -3 & 1 & 1 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} + \frac{1}{N^2}\begin{pmatrix} 2 & -1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} =: I + \frac{1}{N}Q + \frac{1}{N^2}R.$$

Dies passt zu den Voraussetzungen von Lemma 1.6 (mit $c_N = 1/N$), wegen

$$p^{(N,3)}\big(\{\{1\},\{2\},\{3\}\},\{1,2,3\}\big) = \frac{1}{N^2} \ll \frac{1}{N} - \frac{1}{N^2} = p^{(N,3)}\big(\{\{1\},\{2\},\{3\}\},\{\{1,2\},3\}\big)$$

sind im Grenzwert (nach Beschleunigung der Zeit mit dem Faktor N) nur noch Paarverschmelzungen sichtbar.

Die allgemeine Struktur des Grenzwerts (für beliebige Stichprobengröße n) ist folgende:

Definition 1.7. Die zeitkontinuierliche Markovkette $(R_t^{(n)})_{t\geq 0}$ auf $\mathcal{E}_n$ mit Sprungratenmatrix

$$q_{\xi\eta} = \begin{cases} 1 & \text{falls } \eta \text{ aus } \xi \text{ durch Verschmelzung von genau zwei Klassen entsteht,} \\ -\binom{|\xi|}{2} & \text{falls } \eta = \xi, \\ 0 & \text{sonst} \end{cases} \tag{1.4}$$

heißt Kingmans (n-)Koaleszent.

Zumeist betrachten wir den Startzustand $R_0^{(n)} = \big\{\{1\},\{2\},\dots,\{n\}\big\}$. Wir können den Pfad $(R_t^{(n)})_{t\geq 0}$ als Baum interpretieren, dessen Blätter mit $1,\dots,n$ markiert sind:

Zu den Zeitpunkten $0 = \tau_n^{(n)} < \tau_{n-1}^{(n)} < \cdots < \tau_2^{(n)} < \tau_1^{(n)}$, wo $\tau_k^{(n)} := \inf\{t \geq 0 : |R_t^{(n)}| \leq k\}$, verschmelzen jeweils zwei Zweige. Für Stichproben $i, j \in [n]$ können wir den genealogischen Abstand von i und j, $\inf\{t \geq 0 : i \sim_{R_t^{(n)}} j\}$ aus dem Baum ablesen.

`vgl. Bild an der Tafel`

Betrachten wir ein allgemeines Cannings-Populationsmodell und dessen Ahnen-Prozess $(A_r^{(N)})_{r\in\mathbb{Z}}$ wie in Def 1.2 (und Annahme 1.4 gelte), mit zugehörigen Nachkommen(zahl)vektoren $\nu^{(N,r)} = \big(\nu_1^{(N,r)},\cdots,\nu_N^{(N,r)}\big)$.

$R_k^{(N,n)}$ sei die Äquivalenzrelation, die die Verwandtschaftsverhältnisse einer zufälligen n-Stichprobe $J_1,\dots,J_n$ (zur Zeit $r = 0$ gezogen) vor k Generationen beschreibt:

$$i \sim_k j \iff A_{0,J_i}^{(N)}[k] = A_{0,J_j}^{(N)}[k] \quad \begin{array}{l}\text{(d.h. Stichproben } i \text{ und } j \text{ haben denselben} \\ \text{Ahnen vor } k \text{ Generationen)}\end{array}$$

[Erinnerung: Lemma 1.5 beschreibt die Übergangsmatrix der Markovkette $(R_k^{(N,n)})_{k\in\mathbb{N}_0}$.]

Sei

$$c_N := \frac{1}{N(N-1)} \sum_{i=1}^{N} \mathbb{E}\big[\nu_i^{(N)}(\nu_i^{(N)}-1)\big] = \frac{1}{N-1}\mathbb{E}\big[\nu_1^{(N)}(\nu_1^{(N)}-1)\big] \tag{1.5}$$

die Paarverschmelzungwahrscheinlichkeit (über eine Generation) [im Ahnen-Prozess eines Cannings-Modells mit Nachkommensvektor $\nu^{(N)}$].

Beachte: $N = \mathbb{E}\big[\nu_1^{(N)} + \cdots + \nu_N^{(N)}\big] = N\mathbb{E}[\nu_1^{(N)}]$, also $\mathbb{E}[\nu_1^{(N)}] = 1$ (Austauschbarkeit), somit ist $\mathbb{E}\big[\nu_1^{(N)}(\nu_1^{(N)}-1)\big] = \mathbb{E}\big[(\nu_1^{(N)})^2\big] - \big(\mathbb{E}[\nu_1^{(N)}]\big)^2$ und wir können alternativ schreiben

$$c_N = \mathrm{Var}[\nu_1^{(N)}]/(N-1).$$

Sei weiter

$$d_N := \frac{(N)_{1\downarrow}}{(N)_{3\downarrow}}\mathbb{E}\big[(\nu_1^{(N)})_{3\downarrow}\big] = \frac{1}{(N-1)(N-2)}\mathbb{E}\big[\nu_1^{(N)}(\nu_1^{(N)}-1)(\nu_1^{(N)}-2)\big] \tag{1.6}$$

(mit Lemma 1.5 ist $d_N = p^{(N,3)}\big(\{\{1\},\{2\},\{3\}\},\{1,2,3\}\big)$ die W'keit, eine Dreifachverschmelzung über eine Generation zu beobachten).

Satz 1.8. *Betrachte eine Schar von mit der Populationsgröße N parametrisierten (Ahnenprozessen von) Cannings-Modellen.*

Sei $n \in \mathbb{N}_0$, $(R_k^{(N,n)})_{k\in\mathbb{N}_0}$ die Markovkette, die die Verwandtschaftsverhältnisse in einer n-Stichprobe im N-ten Modell beschreibt.

Wenn $c_N \to 0$ und $d_N/c_N \to 0$ gilt, so konvergiert

$$(R^{(n)}_{\lfloor t/c_N\rfloor})_{t\geq 0} \longrightarrow \big(R_t^{(N,n)}\big)_{t\geq 0} \quad \text{für } N\to\infty.$$

Bemerkung 1.9. 1. Satz 1.8 zeigt die "Robustheit" des Kingman-Koaleszenten: Es kommt nicht auf die Details der Nachkommensverteilungen an (sofern die Bedingungen $d_N/c_N \to 0$, $c_N \to 0$ erfüllt sind).

Es kommt nur auf eine Größe an: den Zeitreskalierungsfaktor $1/c_N$. $1/c_N$ heißt auch die "effektive Populationsgröße" (dieser Schar von Cannings-Modellen; für das Wright-Fisher-Modell mit Populationsgröße N gilt tatsächlich $1/c_N = N$).

2. Die Bedingungen $c_N \to 0$ und $d_N/c_N \to 0$ sind auch notwendig für die Konvergenz von $(R^{(N,n)}_{\lfloor t/c_N\rfloor})_{t\geq 0}$ gegen den Kingman-Koaleszenten.

Dies ist zumindest intuitiv sehr plausibel: Wenn $\liminf_{N\to\infty} d_N/c_N > 0$ gilt, so gibt es für ein etwaiges Limesobjekt eine nicht-verschwindende Wahrscheinlichkeit, in der Genealogie einer Stichprobe der Größe $n = 3$ eine Dreifachverschmelzung zu beobachten, was für den Kingman-3-Koaleszenten unmöglich ist.

3. Wir beweisen die in Satz 1.8 formulierte Konvergenz im Sinne der endlich-dimensionalen Verteilungen; tatsächlich gilt auch Konvergenz in Verteilung auf dem Pfadraum $D([0,\infty),\mathcal{E}_n)$.

Beweis von Satz 1.8. Fixiere n. Gemäß Lemma 1.6 müssen wir zeigen, dass für $\xi,\eta \in \mathcal{E}_n$ gilt

$$p^{(N,n)}(\xi,\eta) = \delta_{\xi,\eta} + c_N q_{\xi\eta} + o(c_N) \tag{1.7}$$

[da $|\mathcal{E}_n| < \infty$ ist dann der Fehler gleichmäßig klein, d.h. wir zeigen, dass $\lim_{N\to\infty} \max_{\xi,\eta\in\mathcal{E}_n} \big|p^{(N,n)}(\xi,\eta) - \delta_{\xi,\eta} - c_N q_{\xi\eta}\big|/c_N = 0$ gilt].

Wir lassen im folgenden die oberen Indizes bei $\nu_k^{(N,r)}$, etc. weg und schreiben knapper ν_k.

Wir hatten in Lemma 1.5 gesehen: Wenn η aus $|\eta| = a$ Klassen besteht, ξ aus $b = |\xi| = b_1 + \cdots + b_a$ Klassen besteht und η aus ξ durch Verschmelzen von a Gruppen von Klassen in Gruppengrößen $b_1, \ldots, b_a$ entsteht (d.h. $\eta = \{C_1, \ldots, C_a\}$ und $\xi = \{C_{\alpha\beta} : 1 \le \alpha \le a, 1 \le \beta \le b_\alpha\}$ mit $C_\alpha = \cup_{\beta=1}^{b_\alpha} C_{\alpha\beta}$ für $\alpha = 1, \ldots, a$), so ist

$$\begin{aligned} p^{(N,n)}(\xi,\eta) &= \mathbb{P}\big(\exists\, i_1, \ldots, i_a \in [N] \text{ paarw. versch } : A_{0,k}^{(N)}[1] = i_\alpha \text{ für } k \in C_\alpha, \alpha = 1, \ldots, a\big) \\ &= \frac{(N)_{a\downarrow}}{(N)_{b\downarrow}} \mathbb{E}\Big[\prod_{j=1}^{a} (\nu_j)_{b_j\downarrow}\Big]. \end{aligned}$$

Wir schreiben (zur Abkürzung)

$$I_j = A_{0,j}^{(N)}[1],\ j = 1, \ldots, n.$$

Nach Konstruktion (vgl. Def. 1.2 und Ann. 1.4) sind $I_1, I_2, \ldots, I_n$ zufällige Züge (ohne Zurücklegen, mit Beachtung der Reihenfolge) aus den von ν beschriebenen Nachkommenszahlen, d.h.

$$\mathbb{P}\big(I_1 = i_1, \ldots, I_n = i_n\big) = \frac{1}{(N)_{n\downarrow}} \mathbb{E}\Big[\nu_{i_1}\big(\nu_{i_2} - \mathbf{1}_{i_2=i_1}\big)\big(\nu_{i_3} - \mathbf{1}_{i_3=i_1} - \mathbf{1}_{i_3=i_2}\big)\cdots\big(\nu_{i_n} - \sum_{\ell=1}^{n-1} \mathbf{1}_{i_n=i_\ell}\big)\Big].$$

Dann ist $c_N = \mathbb{P}(I_1 = I_2)$ und $d_N = \mathbb{P}(I_1 = I_2 = I_3)$, d.h. n. Vor.

$$\lim_{N\to\infty} \frac{\mathbb{P}(I_1 = I_2 = I_3)}{\mathbb{P}(I_1 = I_2)} = 0. \tag{1.8}$$

Vorbemerkung:

$$\mathbb{E}\big[\nu_2 f(\nu_1)\big] \le \frac{N}{N-1} \mathbb{E}\big[f(\nu_1)\big] \quad \text{gilt für jedes } f : \{0, 1, \ldots, N\} \to \mathbb{R}_+, \tag{1.9}$$

denn wegen Austauschbarkeit ist

$$(N-1)\mathbb{E}\big[\nu_2 f(\nu_1)\big] = \sum_{j=2}^{N} \mathbb{E}\big[\nu_j f(\nu_1)\big] = \mathbb{E}\big[(N - \nu_1) f(\nu_1)\big] \le N\mathbb{E}\big[f(\nu_1)\big].$$

Markov-Ungleichung und Voraussetzung liefern (für jedes $\epsilon > 0$)

$$\mathbb{P}(\nu_1 > \epsilon N) \le \frac{1}{(\epsilon N)_{3\downarrow}} \mathbb{E}\big[(\nu_1)_{3\downarrow}\big] = \frac{1}{\epsilon^3 N^3} o\big(N\mathbb{E}[(\nu_1)_{2\downarrow}]\big) = \epsilon^{-3} o\big(c_N/N\big). \tag{1.10}$$

Demnach

$$\begin{aligned} \mathbb{E}\big[(\nu_1)_{2\downarrow}(\nu_2)_{2\downarrow}\big] &\le \epsilon N \mathbb{E}\big[(\nu_1)_{2\downarrow}\nu_2 \mathbf{1}(\nu_2 \le \epsilon N)\big] + N^2 \mathbb{E}\big[(\nu_1)_{2\downarrow}\mathbf{1}(\nu_2 > \epsilon N)\big] \\ &\le \epsilon N \mathbb{E}\big[(\nu_1)_{2\downarrow}\nu_2\big] + N^3 \mathbb{E}\big[\nu_1 \mathbf{1}(\nu_2 > \epsilon N)\big] \\ &\overset{(1.9)}{\le} \epsilon N \frac{N}{N-1} \mathbb{E}\big[(\nu_1)_{2\downarrow}\big] + N^3 \frac{N}{N-1} \mathbb{P}(\nu_2 > \epsilon N), \end{aligned}$$

Zusammen mit (1.10) folgt (für jedes $\epsilon > 0$)

$$\limsup_{N\to\infty} \frac{\mathbb{E}\big[(\nu_1)_{2\downarrow}(\nu_2)_{2\downarrow}\big]}{N\mathbb{E}\big[(\nu_1)_{2\downarrow}\big]} \le \epsilon + \limsup_{N\to\infty} \frac{N\mathbb{P}(\nu_1 > \epsilon N)}{c_N} = \epsilon,$$

also

$$\lim_{N\to\infty} \frac{\mathbb{P}(I_1 = I_2 \neq I_3 = I_4)}{\mathbb{P}(I_1 = I_2)} = \lim_{N\to\infty} \frac{(N)_{2\downarrow}\mathbb{E}\big[(\nu_1)_{2\downarrow}(\nu_2)_{2\downarrow}\big]}{(N)_{4\downarrow}} \cdot \frac{(N)_{2\downarrow}}{N\mathbb{E}\big[(\nu_1)_{2\downarrow}\big]} = 0. \tag{1.11}$$

Sei nun $\xi = \{C_{11}, C_{12}, C_2, \ldots, C_a\}$, $\eta = \{C_1, \ldots, C_a\}$ mit $C_1 = C_{11} \cup C_{12}$. Es gilt (mit Lemma 1.5)

$$p^{(N,n)}(\xi,\eta) = \mathbb{P}\big(\{I_1 = I_2\} \cap \{I_m \neq I_1, m = 3, \ldots, a+1\} \cap \{I_\ell \neq I_m, 3 \le \ell < m \le a+1\}\big), \tag{1.12}$$

also $p^{(N,n)}(\xi,\eta) \le \mathbb{P}(I_1 = I_2)$, andererseits

$$\begin{aligned} p^{(N,n)}(\xi,\eta) \quad &\ge \quad \mathbb{P}\big(I_1 = I_2\big) - \mathbb{P}\big(\{I_1 = I_2\} \cap \{\exists\, 3 \le m \le a+1 : I_m = I_1\}\big) \\ &\qquad - \mathbb{P}\big(\{I_1 = I_2\} \cap \{\exists\, 3 \le \ell < m \le a+1 : I_\ell = I_m \neq I_1\}\big). \end{aligned}$$

Beachte

$$\mathbb{P}\big(\{I_1 = I_2\} \cap \{\exists\, 3 \le m \le a+1 : I_m = I_1\}\big) \quad \le \quad (a-1)\mathbb{P}\big(I_1 = I_2 = I_3\big) = o\big(\mathbb{P}(I_1 = I_2)\big)$$

wegen (1.8) und beachte

$$\begin{aligned} \mathbb{P}\big(\{I_1 = I_2\} \cap \{\exists\, 3 \le \ell < m \le a+1 : I_\ell = I_m \neq I_1\}\big) \quad &\le \quad \binom{a-1}{2}\mathbb{P}\big(I_1 = I_2 \neq I_3 = I_4\big) \\ &= \quad o\big(\mathbb{P}(I_1 = I_2)\big) \end{aligned}$$

mit (1.11), d.h. $p^{(N,n)}(\xi,\eta) = c_N + o(c_N)$ gilt in diesem Fall. [Analoge Rechnungen zeigen $p^{(N,n)}(\xi,\eta) = c_N + o(c_N)$, falls $\xi = \{C_1, \ldots, C_{j-1}, C_{j1}, C_{j2}, C_{j+1}, \ldots, C_a\}$ und $C_j = C_{j1} \cup C_{j2}$.]

Falls $|\xi| = b$ und η' aus ξ durch Verschmelzung von drei oder mehr Klassen oder durch Verschmelzung von mindestens zwei disjunkten Gruppen von Klassen entsteht, so ist

$$\begin{aligned} p^{(N,n)}(\xi,\eta') &\le \mathbb{P}\big(\exists\, j,k,\ell \in [b] \text{ paarw. versch. } : I_j = I_k = I_\ell\big) \\ &\qquad + \mathbb{P}\big(\exists\, j,k,\ell,m \in [b] \text{ paarw. versch. } : I_j = I_k, I_\ell = I_m, I_j \neq I_\ell\big) \\ &\le \binom{b}{3}\mathbb{P}\big(I_1 = I_2 = I_3\big) + \binom{b}{4}\mathbb{P}\big(I_1 = I_2 \neq I_3 = I_4\big) \end{aligned}$$

also $p(\xi,\eta') = o(c_N)$ in diesem Fall.

Schließlich ist wegen $\sum_{\eta\in\mathcal{E}_n} p^{(N,n)}(\xi,\eta) = 1$ und $|\mathcal{E}_n| < \infty$

$$p^{(N,n)}(\xi,\xi) = 1 - \binom{|\xi|}{2}c_N + o(c_N),$$

d.h. (1.7) gilt. □

Beobachtung 1.10. 1. (Die Zeit bis zum jüngsten gemeinsamen Vorfahren) Aus der Struktur der Sprungratenmatrix (1.4) folgt

$$\tau_1^{(n)} = (\tau_{n-1}^{(n)} - \tau_n^{(n)}) + (\tau_{n-2}^{(n)} - \tau_{n-1}^{(n)}) + \cdots + (\tau_1^{(n)} - \tau_2^{(n)}) \stackrel{d}{=} S_n + S_{n-1} + \cdots + S_2, \qquad n \geq 2,$$

wobei die S_k unabhängige exponentialverteilte Zufallsvariablen mit Parameter $\binom{k}{2}$ sind, somit

$$\mathbb{E}[\tau_1^{(n)}] = \sum_{k=2}^{n} E[S_k] = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\sum_{k=2}^{n}\Big(\frac{1}{k-1} - \frac{1}{k}\Big) = 2\Big(1 - \frac{1}{n}\Big)$$

und $1 = \mathbb{E}[\tau_1^{(2)}] \leq \mathbb{E}[\tau_1^{(n)}] < \lim_{m\to\infty} \mathbb{E}[\tau_1^{(m)}] = 2$.

Für ein Populationsmodell (aus der von uns betrachteten Schar) mit Populationsgröße N bedeutet dies, das der jüngste gemeinsame Vorfahre der heute lebenden Population im Mittel vor etwa $2/c_N$ Generationen gelebt hat.

Weiter ist

$$\begin{aligned}\mathrm{Var}[\tau_1^{(n)}] &= \sum_{k=2}^{n} \mathrm{Var}[S_k] = \sum_{k=2}^{n} \binom{k}{2}^{-2} = \sum_{k=2}^{n} \frac{4}{k^2(k-1)^2} = \sum_{k=2}^{n} \Big\{4\Big(\frac{1}{k^2} + \frac{1}{(k-1)^2}\Big) + 8\Big(\frac{1}{k} - \frac{1}{k-1}\Big)\Big\} \\ &= \Big\{8\sum_{k=1}^{n-1}\frac{1}{k^2}\Big\} - 4 + \frac{4}{n^2} + \frac{8}{n} - 8 = \Big\{8\sum_{k=1}^{n-1}\frac{1}{k^2}\Big\} - 4\Big(1 - \frac{1}{n}\Big)\Big(3 + \frac{1}{n}\Big),\end{aligned}$$

insbesondere

$$1 = \mathrm{Var}[\tau_1^{(2)}] \leq \mathrm{Var}[\tau_1^{(n)}] < \lim_{n\to\infty} \mathrm{Var}[\tau_1^{(n)}] = 8\frac{\pi^2}{6} - 12 \approx 1.16.$$

Der wesentliche Beitrag zur Gesamtvarianz kommt also von der letzten Verschmelzungszeit S_2.

2. (Teilstichproben-Konsistenz) Sei $\pi_{n,n-1} : \mathcal{E}_n \to \mathcal{E}_{n-1}$ die Einschränkung aller Äquivalenzklassen auf $[n-1]$, so gilt

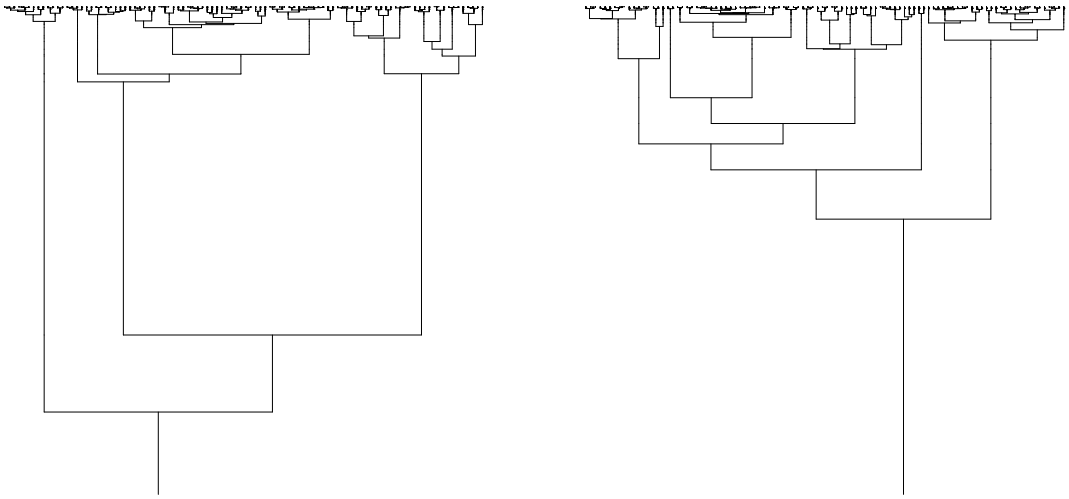
$$\big(\pi_{n,n-1}(R_t^{(n)})\big)_{t\geq 0} \stackrel{d}{=} \big(R_t^{(n-1)}\big)_{t\geq 0}.$$

Dies folgt aus der Form der Sprungraten oder auch aus der Tatsache, dass für die Approximanten (wie in Satz 1.8) nach Konstruktion $\pi_{n,n-1}(R_k^{(N,n)}) = R_k^{(N,n-1)}$ (realisierungsweise) gilt.

3. (Invarianz der Verteilung unter Permutation der Stichprobennummern) Für eine Permutation σ von $[n]$ und $\xi = \{C_1, \ldots, C_a\} \in \mathcal{E}_n$ sei $\sigma(\xi) = \{\sigma(C_1), \ldots, \sigma(C_a)\}$ die Äquivalenzrelation, die man erhält, indem man die Elemente der Blöcke von ξ gemäß σ umnummeriert. Es gilt

$$\big(\sigma(R_t^{(n)})\big)_{t\geq 0} \stackrel{d}{=} \big(R_t^{(n)}\big)_{t\geq 0}.$$

Dies folgt aus der Symmetrie der Sprungraten oder auch aus der Tatsache, dass für die Approximanten (wie in Satz 1.8) nach Konstruktion $\big(\sigma(R_k^{(N,n)})\big)_{k\in\mathbb{N}_0} \stackrel{d}{=} \big(R_k^{(N,n)}\big)_{k\in\mathbb{N}_0}$ gilt. [Man sagt auch, dass $R_t^{(n)}$ eine austauschbare zufällige Äquivalenzrelation ist.]



Zwei Realisierungen des Kingman-100-Koaleszenten (wobei die Blätter jeweils so sortiert wurden, dass der Baum überschneidungsfrei zu zeichnen ist)

Bericht. Mit Beob. 1.10, 2. und Kolmogorovs Erweiterungssatz ist es möglich, den Kingman-Koaleszenten $(R_t)_{t\geq 0}$ mit Stichprobengröße $n = \infty$ als Markovprozess auf $\mathcal{E} := \{\text{Äquivalenzrelationen auf } \mathbb{N}\}$ mit Startwert $R_0 = \{\{1\},\{2\},\dots\}$ zu definieren mit der Eigenschaft $\big(\pi_{\infty,n}(R_t)\big)_{t\geq 0} \stackrel{d}{=} \big(\sigma(R_t^{(n)})\big)_{t\geq 0}$ für jedes $n \in \mathbb{N}$.

Beob. 1.10, 1. zeigt, dass $\mathbb{E}[\tau_1^{(\infty)}] = 2 < \infty$, d.h. auch eine „unendlich große Stichprobe" findet f.s. in endlicher Zeit ihren ersten gemeinsamen Vorfahren. Obwohl $|R_0| = \infty$ ist, gilt $|R_t| < \infty$ für jedes $t > 0$ fast sicher. [Man sagt auch, dass der Kingman-Koaleszent „aus dem Unendlichen herabsteigt."]

Sei $\xi_i^{(n)}$, $i = n, n-1, \dots, 1$ der Zustand des n-Koaleszenten zum ersten Zeitpunkt, zu dem i Klassen existieren, d.h. $\xi_i^{(n)} = R^{(n)}_{\tau_i^{(n)}}$ (mit $\tau_i^{(n)} := \inf\{t \geq 0 \colon |R_t^{(n)}| \leq i\}$ wie oben). [$(\xi_i^{(n)})_{i=n,n-1,\dots,1}$ heißt die *Skelettkette* des Kingman-n-Koaleszenten, sie ist (offenbar) eine Markovkette.]

Proposition 1.11. *Für $\xi \in \mathcal{E}_n$ mit i Klassen der Größen $\lambda_1, \dots, \lambda_i \in \mathbb{N}$ (mit $\lambda_1 + \dots + \lambda_i = n$) gilt*

$$\mathbb{P}(\xi_i^{(n)} = \xi) = c_{n,i} w(\xi) \quad \text{mit } w(\xi) = \lambda_1! \cdots \lambda_i!, \ c_{n,i} = \frac{i!}{n!} \frac{(n-i)!(i-1)!}{(n-1)!}. \tag{1.13}$$

Beispiel. Betrachte $n = 9$, $i = 3$, es ist $c_{9,3} = \frac{3!}{9!}\frac{6!2!}{8!} = 1/1\,693\,440$.

λ_i	3-3-3	4-3-2	5-2-2	4-4-1	5-3-1	6-2-1	7-1-1
w	216	288	480	576	720	1440	5040

Wir sehen: die Verteilung hat mehr Gewicht auf „unbalanzierten Aufteilungen."

Beweis von Prop. 1.11. Rückwärtsinduktion über i: Für $i = n$ gilt $\mathbb{P}(\xi_n^{(n)} = \{\{1\},\dots,\{n\}\}) = 1$ mit $\lambda_1 = \dots = \lambda_n = 1$, und $c_{n,n} = w(1,\dots,1) = 1$.

$i \to i-1$: Es ist

$$\mathbb{P}(\xi_{i-1}^{(n)} = \eta \,|\, \xi_i^{(n)} = \xi) = \begin{cases} \dfrac{1}{\binom{i}{2}}, & \text{falls } \eta \text{ aus } \xi \text{ durch Verschmelzung eines Paars von Klassen entsteht,} \\ 0, & \text{sonst.} \end{cases}$$

Sei $\eta \in \mathcal{E}_n$, $|\eta| = i-1$, Klassengrößen $\tilde{\lambda}_1, \dots, \tilde{\lambda}_{i-1}$.

$$\begin{aligned} \mathbb{P}(\xi_{i-1}^{(n)} = \eta) &= \frac{2}{i(i-1)} \sum_{\xi : \xi \prec \eta} \mathbb{P}(\xi_i^{(n)} = \xi) \\ &= \frac{2}{i(i-1)} \sum_{\ell=1}^{i-1} \sum_{m=1}^{\tilde{\lambda}_\ell - 1} \frac{1}{2} \binom{\tilde{\lambda}_\ell}{m} c_{n,i} \tilde{\lambda}_1! \cdots \tilde{\lambda}_{\ell-1}! m! (\tilde{\lambda}_\ell - m)! \tilde{\lambda}_{\ell+1}! \cdots \tilde{\lambda}_{i-1}! \\ &= \frac{c_{n,i}}{i(i-1)} \sum_{\ell=1}^{i-1} \sum_{m=1}^{\tilde{\lambda}_\ell - 1} w(\eta) = \frac{c_{n,i} w(\eta)}{i(i-1)} \sum_{\ell=1}^{i-1} \sum_{m=1}^{\tilde{\lambda}_\ell - 1} 1 \\ &= \frac{c_{n,i} w(\eta)}{i(i-1)} (n - (i-1)) = c_{n,i-1} w(\eta) \end{aligned}$$

Für das 2. Gleichheitszeichen verwenden wir die Induktionsannahme und zerlegen gemäß der „aufgespaltenen“ Klasse: die ℓ-te Klasse hat $\tilde{\lambda}_\ell$ Elemente, zerlege in 2 Teilmengen d. Größen m und $\tilde{\lambda}_\ell - m$, es gibt $\frac{1}{2}\binom{\tilde{\lambda}_\ell}{m}$ mögliche Wahlen; der Faktor $\frac{1}{2}$ entsteht, weil die Klassen in η als ungeordnet aufgefasst werden. □

Korollar 1.12. *1. Sei σ eine uniform verteilte Permutation von $\{1, \dots, i\}$, u.a. von $\xi^{(n)}$, $M_i = |C_{i,\sigma(i)}^{(n)}|$ mit $\xi_i^{(n)} = \{C_{i,1}^{(n)}, \dots, C_{i,i}^{(n)}\}$. Dann ist*

$$(M_1, \dots, M_i) \text{ uniform verteilt auf } \big\{(m_1, \dots, m_i) \in \mathbb{N}^i : m_1 + \cdots + m_i = n\big\}.$$

2. Sei $\xi = \{A_1, \dots, A_{i-1}\} \in \mathcal{E}_n$ mit $|A_j| = \lambda_j$. $\mathscr{L}(\xi_i^{(n)} \,|\, \xi_{i-1}^{(n)} = \xi)$ kann folgendermaßen beschrieben werden:

- *wähle A_j mit W'keit $\frac{\lambda_j - 1}{n-i+1}$ ($j \in \{1, \dots, i-1\}$), dann wähle k uniform aus $\{1, \dots, \lambda_j - 1\}$*
- *spalte A_j uniform in zwei Teile der Größen k und $\lambda_j - k$.*

Bemerkung. Für $i = 2$ zeigt Kor. 1.12 die „uniforme Aufspaltung“ der Stichprobe in zwei älteste Familien.

Beweis von Kor. 1.12. 1. Seien $m_1, \dots, m_i$ mit $m_1 + \cdots + m_i = n$ gegeben. Nach Prop. 1.11 hat jede Realisierung des (zufällig) geordneten Vektors $(C_{i,\sigma(1)}^{(n)}, \dots, C_{i,\sigma(i)}^{(n)})$, die mit den geforderten Größen m_j verträglich ist, die W'keit $\frac{1}{i!} c_{n,i} m_1! \cdots m_i!$, somit

$$\begin{aligned} \mathbb{P}\big((M_1, \dots, M_i) &= (m_1, \dots, m_i)\big) = \binom{n}{m_1 \ \dots \ m_i} \frac{1}{i!} c_{n,i} m_1! \cdots m_i! \\ &= \frac{(n-i)!(i-1)!}{(n-1)!} = \frac{1}{\binom{n-1}{i-1}}, \end{aligned}$$

denn es gibt $\binom{n}{m_1 \dots m_i}$ Partitionen, die bezgl. der Größe der Klassen in Frage kommen,

jede hat n. Prop. 1.11 dieselbe W'keit $c_{n,i} m_1! \cdots m_i!$,

die W'keit, dass zuf. Perm. σ geg. Ordnung liefert, ergibt nochmals einen Faktor $\frac{1}{i!}$.

(Beachte auch $\#\{\{(m_1, \dots, m_i) \in \mathbb{N}^i : m_1 + \cdots + m_i = n\}\} = \binom{n-1}{i-1}$: n Kugeln in i (nummerierte) Schachteln legen, so dass keine Schachtel leer ist: $n-1$ mögl. Plätze für $i-1$ "Trennwände".)

2. ξ entstehe aus η durch Aufteilen von A_j in 2 Teile der Größen k und $\lambda_j - k$.

$$\begin{aligned}
\mathbb{P}(\xi_i^{(n)} = \xi \,|\, \xi_{i-1}^{(n)} = \eta) &= \frac{\mathbb{P}(\xi_{i-1}^{(n)} = \eta \,|\, \xi_i^{(n)} = \xi)\mathbb{P}(\xi_i^{(n)} = \xi)}{\mathbb{P}(\xi_{i-1}^{(n)} = \eta)} \\
&= \frac{\frac{1}{\binom{i}{2}} c_{n,i} \lambda_1! \cdots \lambda_{j-1}! k! (\lambda_j - k)! \lambda_{j+1}! \cdots \lambda_{i-1}!}{c_{n,i-1} \lambda_1! \cdots \lambda_{j-1}! \lambda_j! \lambda_{j+1}! \cdots \lambda_{i-1}!} = \frac{1}{\binom{i}{2}} \cdot \frac{i(i-1)}{n-i+1} \frac{1}{\binom{\lambda_j}{k}} \\
&= \frac{\lambda_j - 1}{n-i+1} \cdot \frac{1}{\lambda_j - 1} \cdot 2 \frac{1}{\binom{\lambda_j}{k}}
\end{aligned}$$

($\frac{\lambda_j - 1}{n-i+1} \widehat{=}$ Wahl von A_j; $\frac{1}{\lambda_j - 1} \widehat{=}$ Wahl von k; $2\frac{1}{\binom{\lambda_j}{k}} \widehat{=}$ Wahl der Zerlegung von A_j – beachte: Faktor 2, da die Klassen "ungeordnet" angegeben werden) □

Korollar 1.13. *Betrachte eine Teilstichprobe der Grösse n in einem Kingman-m-Koaleszenten, mit $m > n$. Dann erfüllt die Wahrscheinlichkeit des Ereignisses $E_{m,n}$, dass der jüngste gemeinsame Vorfahre (jgV) der n-Stichprobe mit der Wurzel des m-Koaleszenten übereinstimmt,*

$$\mathbb{P}(E_{m,n}) \to \frac{n-1}{n+1} \quad \text{für } m \to \infty.$$

Beweis. Wir betrachten $\xi_2^{(n)}$, die erste Aufspaltung des Koaleszenten von der Wurzel aus betrachtet. Diese resultiert in einer Aufspaltung der trivialen Partition $\{\{1, \dots, m\}\}$ in eine Äquivalenzrelation mit genau zwei Klassen der Grössen $m - X$ und X, wobei X nach Kor. 1.12 auf $[m-1]$ uniform verteilt ist. Falls der jgV der n-Stichprobe nicht mit der Wurzel übereinstimmt, so müssen die Ahnenlinien aller n Individuen der Stichprobe alle entweder in dem Block der Grösse $m - X$ oder in dem Block der Grösse X liegen.

Das erste Ereignis hat Wahrscheinlichkeit $\frac{(m-X)_{n\downarrow}}{(m)_{n\downarrow}}$ und das zweite $\frac{(X)_{n\downarrow}}{(m)_{n\downarrow}}$.

Wir erhalten

$$\begin{aligned}
\mathbb{P}(E_{mn,}) &= 1 - \mathbb{P}\big((E_{mn,})^c\big) = 1 - \sum_{k=1}^{m-1} \Big[\frac{(m-k)_{n\downarrow}}{(m)_{n\downarrow}} + \frac{(k)_{n\downarrow}}{(m)_{n\downarrow}} \Big] \underbrace{\mathbb{P}(X = k)}_{\frac{1}{m-1}} \\
&\underset{m\to\infty}{\longrightarrow} 1 - \int_0^1 (x^n + (1-x)^n)\, dx \\
&= 1 - \Big[\frac{1}{n+1} x^{n+1} \Big]_0^1 - \Big[\frac{1}{n+1} (1-x)^{n+1} (-1) \Big]_0^1 = 1 - \frac{2}{n+1}
\end{aligned}$$

für $m \to \infty$ durch Konvergenz der Riemann-Summe gegen das Riemann-Integral. □

Lemma 1.14. *Seien $X_1, X_2, \dots, X_n$ unabhängig, X_i sei exponentialverteilt mit Parameter λ_i und $\lambda_1 < \lambda_2 < \dots < \lambda_n$. Dann hat $X := X_1 + \dots + X_n$ die Dichte*

$$f_X(t) = \sum_{j=1}^{n} \lambda_j \exp(-\lambda_j t) \prod_{k=1,k\neq j}^{n} \frac{\lambda_k}{\lambda_k - \lambda_j},$$

insbesondere ist (mit $a_j := \prod_{k=1,k\neq j}^{n} \frac{\lambda_k}{\lambda_k-\lambda_j}$)

$$\mathbb{P}(X > t) = \sum_{j=1}^{n} a_j \exp(-\lambda_j t), \quad t \geq 0.$$

Beweisskizze. Die Formel für die Dichte kann man beispielsweise per Induktion durch sukzessive Faltung mit der Exponentialdichte beweisen, für den Induktionsschritt beachten wir

$$\begin{aligned}
&\int_0^t \sum_{j=1}^{n-1} \lambda_j \exp(-\lambda_j s) \prod_{k=1,k\neq j}^{n-1} \frac{\lambda_k}{\lambda_k - \lambda_j} \times \lambda_n \exp(-\lambda_n(t-s))\, ds \\
&= \sum_{j=1}^{n-1} \lambda_j \lambda_n \prod_{k=1,k\neq j}^{n-1} \frac{\lambda_k}{\lambda_k - \lambda_j} \times e^{-\lambda_n t} \int_0^t e^{(\lambda_n-\lambda_j)s}\, ds = \sum_{j=1}^{n-1} \lambda_j \lambda_n \prod_{k=1,k\neq j}^{n-1} \frac{\lambda_k}{\lambda_k - \lambda_j} \times \frac{e^{-\lambda_j t} - e^{-\lambda_n t}}{\lambda_n - \lambda_j} \\
&= \sum_{j=1}^{n-1} \lambda_j e^{-\lambda_j t} \prod_{k=1,k\neq j}^{n} \frac{\lambda_k}{\lambda_k - \lambda_j} - \lambda_n e^{-\lambda_n t} \sum_{j=1}^{n-1} \frac{\lambda_j}{\lambda_n - \lambda_j} \prod_{k=1,k\neq j}^{n-1} \frac{\lambda_k}{\lambda_k - \lambda_j}.
\end{aligned}$$

Dann verwenden wir die Identität

$$\sum_{j=1}^{n-1} \frac{\lambda_j}{\lambda_n - \lambda_j} \prod_{k=1,k\neq j}^{n-1} \frac{\lambda_k}{\lambda_k - \lambda_j} = -\prod_{k=1}^{n-1} \frac{\lambda_k}{\lambda_k - \lambda_n},$$

die (man dividiere beide Seiten durch $\lambda_1\lambda_2\cdots\lambda_{n-1}$ und sortiere Terme) äquivalent ist zu

$$\sum_{j=1}^{n} \prod_{k=1,k\neq j}^{n} \frac{1}{\lambda_k - \lambda_j} = 0. \tag{1.14}$$

Sei $\ell_j(x) := \prod_{k=1,k\neq j}^{n} \frac{x-\lambda_k}{\lambda_j-\lambda_k}$ (das j-te Lagrange-Polynom zu $\lambda_1, \dots, \lambda_n$), $\ell_1(x)+\ell_2(x)+\dots+\ell_n(x)$ ist ein Polynom in x vom Grad $n-1$, das (mindestens) an den n verschiedenen Stellen $\lambda_1, \dots, \lambda_n$ den Wert 1 annimmt (denn $\ell_j(\lambda_i) = \delta_{ji}$), daher gilt $\ell_1(x) + \ell_2(x) + \dots + \ell_n(x) \equiv 1$. Die linke Seite von (1.14) ist $(-1)^{n-1}$ mal der Koeffizient von x^{n-1} in diesem Polynom.

Alternativ beachte man, dass für $\zeta \in \mathbb{R}$ ist $\mathbb{E}[e^{i\zeta X_j}] = \int_0^\infty e^{i\zeta x} \lambda_j e^{-\lambda_j x}\, dx = \frac{\lambda_j}{\lambda_j - i\zeta}$, also $\mathbb{E}[e^{i\zeta X}] = \prod_{j=1}^{n} \frac{\lambda_j}{\lambda_j - i\zeta} =: \varphi_1(\zeta)$ gilt, während $\int_0^\infty e^{i\zeta x} \sum_{j=1}^{n} a_j \lambda_j e^{-\lambda_j x}\, dx = \sum_{j=1}^{n} \frac{a_j \lambda_j}{\lambda_j - i\zeta} =: \varphi_2(z)$ und es *ist* $\varphi_2 = \varphi_1$ (φ_2 ist die Partialbruchzerlegungs-Darstellung von φ_1), denn beide sind meromorph auf $\mathbb{C}$ mit jeweils einfachen Polen bei $\zeta = -i\lambda_1, \dots, -i\lambda_n$und $\lim_{|z|\to\infty} \varphi_{1/2}(z) = 0$, $\lim_{z\to -i\lambda_j} \frac{\varphi_1(z)}{\lambda_j - iz} = \lambda_j \prod_{k=1,k\neq j}^{n} \frac{\lambda_k}{\lambda_k - \lambda_j} = \lim_{z\to -i\lambda_j} \frac{\varphi_2(z)}{\lambda_j - iz}$.

Die Formel für den Verteilungsschwanz von X ergibt sich durch entsprechendes Integrieren der Dichte. □

1.1.1 Beispiel: Die Beobachtungen von Dorit et al, 1995

Robert L. Dorit, Hiroshi Akashi und Walter Gilbert berichten in Absence of Polymorphism at the ZFY Locus on the Human Y Chromosome, *Science* 268, 1183–1185 (1995) die Ergebnisse einer genetischen Studie:

- Weltweite[5] Stichprobe von 38 Männern (*homo sapiens*)
- Ein 729 Basenpaare langes, nicht-kodierendes Stück des Y-Chromosoms (das 3. Intron des ZFY-Gens) wurde für jede Stichprobe sequenziert
- Es wurden keinerlei Mutationen gefunden: Alle 38 Stichproben identisch
- Inter-spezies-Vergleich mit Schimpanse, Gorilla, Orang-Utan (und Pavian als "outgroup") zeigt, dass am betrachteten Lokus Mutationen vorkommen können
- Molekulare Uhr-Annahme und auf Fossilien beruhende Annahmen über die Zeit seit der Aufspaltung von der Vorfahren von Mensch und Schimpanse bzw. Orang-Utan ergeben geschätzte Rate von (fixierten) Mutationen

  $1{,}35 \times 10^{-3}$ Mutationen pro Basenpaar pro Million Jahre

Was können wir angesichts dieser Beobachtungen über die Zeit bis zum jüngsten gemeinsamen Vorfahren der gezogenen 38 Y-Chromosomen (und damit implizit auch über den jgV aller heute lebenden Männer) sagen?

Wir verwenden den Kingman-Koaleszenten als Modell der Genealogie.

A-priori-Verteilung Ohne Berücksichtigung der Beobachtungen würden wir annehmen, dass

$$T_{\text{jgV}} \stackrel{d}{=} S_{38} + S_{37} + \cdots + S_2$$

wo T_{jgV} die Zeit (in Koaleszenten-Zeiteinheiten) bis zum jüngsten gemeinsamen Vorfahren der 38 gezogenen Männer, die S_k unabhängig mit $S_k \sim \text{Exp}\left(\binom{k}{2}\right)$,

$$1 \text{ Koaleszenten-Zeiteinheit } \widehat{=}\ N_{\text{eff}} \times g \text{ Jahre}$$

mit $N_{\text{eff}}\ldots$ effektive Populationsgröße (für Männer), $g\ldots$ Generationslänge (in Jahren), also

a-priori-Verteilung: $\mathscr{L}(T_{\text{jgV}}) = \overset{38}{\underset{k=2}{\ast}} \text{Exp}\left(\binom{k}{2}\right)$, d.h. mit Lemma 1.14 ist die Dichte

$$f_{\text{a-pri}} = \sum_{i=2}^{38} \binom{i}{2} \exp\left(-\binom{i}{2} t\right) \prod_{j=2, j\neq i}^{38} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}}$$

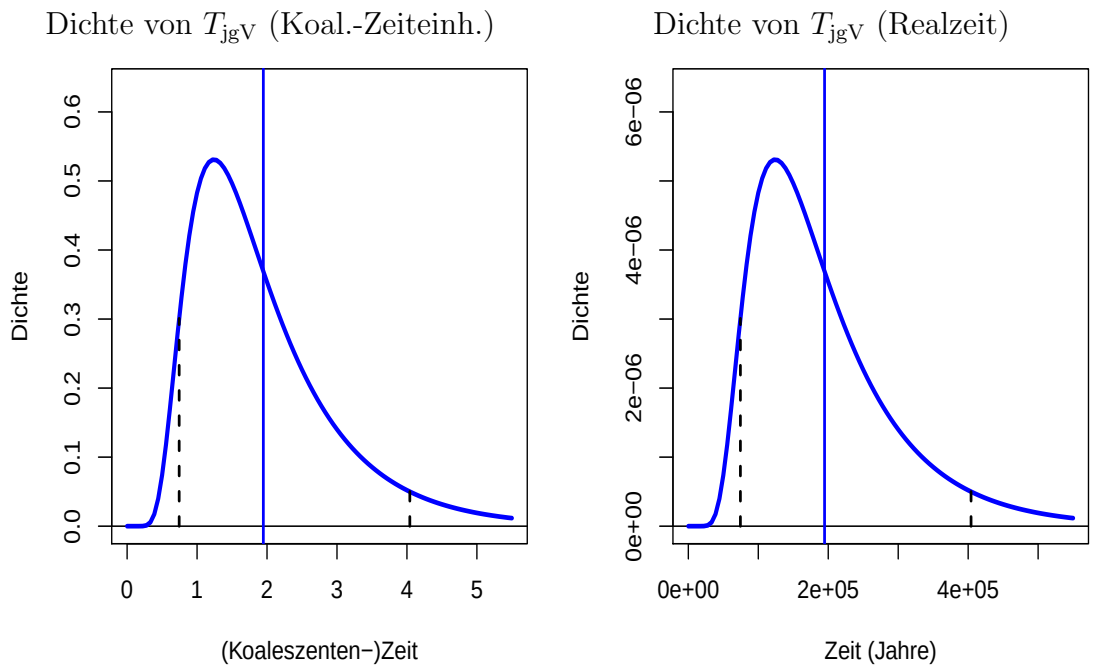
[5] Loc. cit., S. 1184: "Human DNA samples were obtained from male volunteers who donated hair follicle samples or from cell lines provided by L. L. Cavalli-Sforza and K. K. Kidd. Geographic origins were determined by interview. Whenever possible, geographic origins of parents and grandparents were also ascertained. The samples are grouped by continent of origin, and the number of individuals is given in parentheses. Africa: Nigeria* (1), Ivory Coast (1), Tanzania (1), Southern Africa (2), Algeria (1), Central African Republic* (2), African American (2); Americas: Mexico (2), Guatemala (1), Peru* (1), Argentina (1), Native American (2); Asia: China* (2), Korea (1), Japan* (2), Taiwan (2), Indonesia (1), India (1); Europe/Middle East: Ireland* (1), Belgium (1), Italy* (1), Spain (1), Russia* (2), Poland* (1), Saudi Arabia* (1), Turkey (1); South Pacific: Melanesia (1), New Guinea* (1), Australia* (1). (*) Indicates samples where the 3'-most zinc-finger exon was also sequenced."

und

$$\mathbb{E}[T_{\mathrm{jgV}}] = \sum_{k=2}^{38} \frac{2}{k(k-1)} = 2\big(1 - \tfrac{1}{38}\big) = \frac{37}{19} \doteq 1{,}947,$$

5%-Quantil von T_{jgV}: $q_{0,05} \doteq 0{,}744$, 95%-Quantil: $q_{0,95} \doteq 4{,}041$.

Mit Annahmen $N_{\mathrm{eff}} = 5.000$, $g = 20$ Jahre übersetzt sich dies zu MW $\doteq$ 195.000 Jahre, $q_{0,05} \doteq$ 74.000 Jahre, $q_{0,95} \doteq$ 404.000 Jahre.



A-posteriori-Verteilung Sei S_k die Länge des Zeitintervalls (in Koaleszenten-Zeiteinheiten) während dessen die Genealogie der Stichprobe aus k Linien bestand, M_k die Anzahl Mutationen, die während dieses Intervalls in der Genealogie auftreten.

Gegeben

$$S_k = t \text{ ist } M_k \text{ Poisson-verteilt mit Parameter } tk\frac{\theta}{2}, \text{ d.h.} \tag{1.15}$$

$$\mathbb{P}\big(M_k = m \,|\, S_k = t\big) = \exp\Big(-tk\frac{\theta}{2}\Big)\frac{\big(tk\frac{\theta}{2}\big)^m}{m!} \quad \text{wobei} \quad \theta = 2N_{\mathrm{eff}} \times g \times \mu$$

mit N_{eff} effektive Populationsgröße, g Generationslänge (in Jahren), μ Mutationsrate der betrachteten Region im Genom (pro Jahr) (und gegeben $S_2, \ldots, S_n$ sind $M_2, \ldots, M_n$ unabhängig).

Eine heuristische Begründung für (1.15) ist folgende (wir werden dies im weiteren Verlauf der Vorlesung noch genauer betrachten): Angesichts Satz 1.8 entsprechen t Koaleszenten-Zeiteinheiten im Populationsmodell mit Populationsgröße N etwa t/c_N Generationen und somit etwa $tg/c_N = tgN_{\mathrm{eff}}$ Jahren. Wenn wir annehmen, dass pro Jahr (unabhängig von allem anderen) eine Mutation mit der (sehr kleinen) Wahrscheinlichkeit μ auftritt, so ist die Verteilung der Anzahl Mutationen, die wir auf einem Stück der Genealogie dieser Länge sehen,

$$\mathrm{Bin}(tgN_{\mathrm{eff}}, \mu) \approx \mathrm{Poi}(tgN_{\mathrm{eff}}\mu) = \mathrm{Poi}(t\theta/2).$$

Da gegeben $S_k = t$ der Teil der Genealogie, währenddessen k Linien existieren, aus k Stücken von je t Koaleszenten-Zeiteinheiten besteht, ist

$$\mathbb{P}(M_k = m \,|\, S_k = t) = \underbrace{\mathrm{Poi}(t\theta/2) * \cdots * \mathrm{Poi}(t\theta/2)}_{k \text{ mal}} = \mathrm{Poi}(tk\theta/2).$$

(Die Normierung des Mutationsparameters als $\theta/2$ hat historische Gründe und sorgt auch dafür, dass manche Formeln „schöner“ aussehen.)

Frage: Wie ist $S_{38} + \cdots + S_2$ verteilt, gegeben dass $M_{38} + \cdots + M_2 = 0$? $S_k \sim \mathrm{Exp}\big(\binom{k}{2}\big)$, $\mathscr{L}(M_k \,|\, S_k = t) = \mathrm{Poi}(tk\theta/2)$, dann ist

$$\begin{aligned}\mathbb{P}(M_k = m) &= \int_0^\infty \binom{k}{2} \exp\big(-\tbinom{k}{2}t\big) e^{-tk\theta/2} \frac{(tk\theta/2)^m}{m!}\, dt \\ &= \binom{k}{2} \frac{(k\theta/2)^m}{m!} \int_0^\infty t^m \exp\Big(-\big(\tbinom{k}{2} + k\theta/2\big)t\Big)\, dt = \frac{k-1}{k-1+\theta}\Big(\frac{\theta}{k-1+\theta}\Big)^m,\end{aligned}$$

(wir substituieren $u = \big(\binom{k}{2} + k\theta/2\big)t$ und nutzen $\int_0^\infty u^m e^{-u}\, du = \Gamma(m+1) = m!$) d.h. $\mathscr{L}(M_k) = \mathrm{Geom}\big(\frac{k-1}{k-1+\theta}\big)$ — man könnte dies alternativ auch über ein „konkurrierende Raten“-Argument einsehen. Weiter ist damit

$$\mathbb{P}(S_k \le t \,|\, M_k = 0) = \frac{k-1+\theta}{k-1} \int_0^t \binom{k}{2} \exp\big(-\tbinom{k}{2}s\big) e^{-sk\theta/2}\, ds.$$

$$\mathscr{L}(S_k \,|\, M_k = 0) = \mathrm{Exp}\Big(\frac{k(k-1+\theta)}{2}\Big).$$

Bedingt auf $M_2 = \cdots = M_{38} = 0$ sind $S_2, \ldots, S_{38}$ (weiterhin) unabhängig.

Demnach: Verteilung der Zeit bis zum jüngsten gemeinsamen Vorfahren (in Koaleszenten-Zeiteinheiten), bedingt auf $M := M_2 + \cdots + M_{38} = 0$ ist

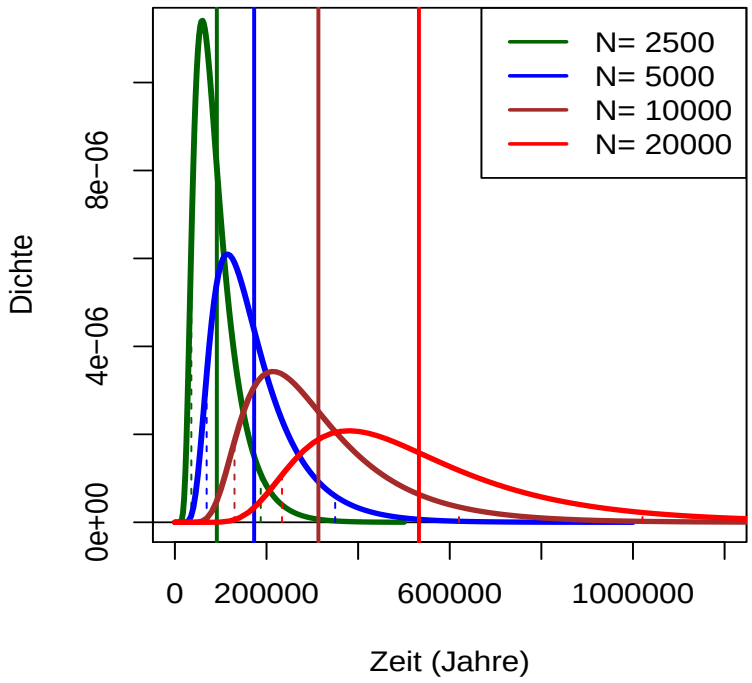
$$T_{\mathrm{jgV}}\big|_{\{M=0\}} \stackrel{d}{=} S'_{38} + S'_{37} + \cdots + S'_2$$

mit S'_k u.a., $S'_k \sim \mathrm{Exp}\big(\frac{k(k-1+\theta)}{2}\big)$, d.h. $\mathscr{L}(T_{\mathrm{jgV}} \,|\, M = 0) = \overset{38}{\underset{k=2}{*}} \mathrm{Exp}\big(\frac{k(k-1+\theta)}{2}\big)$. Die Dichte von $\mathscr{L}(T_{\mathrm{jgV}} \,|\, M = 0)$ (and damit auch den Erwartungswert und die Verteilungsfunktion) können wir wiederum mit Lemma 1.14 bestimmen.

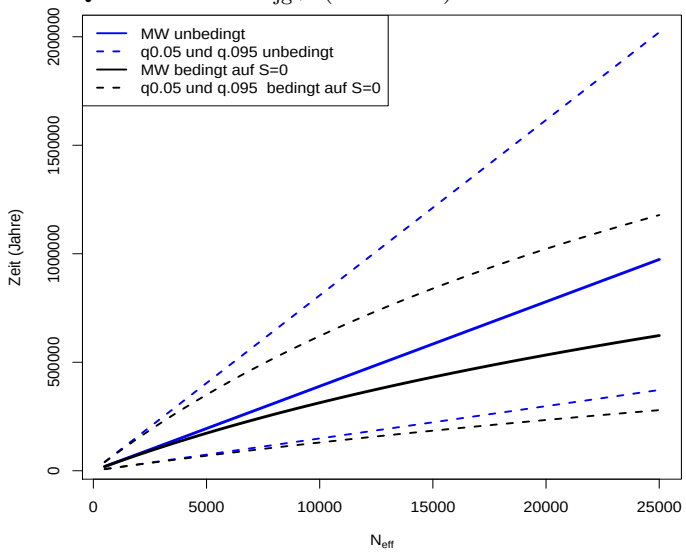
Wir fixieren $g = 20\mathrm{a}$, $\mu = 729 \times 1{,}35 \cdot 10^{-9}\mathrm{a}^{-1} \doteq 0{,}98 \cdot 10^{-6}\mathrm{a}^{-1}$ (aus Dorit et al (1995), diese Werte waren auch in der Literatur unstrittig), so hängt die bedingte Verteilung von T_{jgV} (und nicht nur ihre „Übersetzung in Realzeit“) vom Parameter N_{eff} ab.

N_{eff}	EW	$q_{0,05}$	$q_{0,95}$
2.500	91.519	35.851	187.369
5.000	173.007	69.263	349.909
10.000	313.234	130.095	620.279
20.000	532.785	233.853	1.020.819

Dichte von T_{jgV} (Realzeit)



MW und Quantile von T_{jgV} (Realzeit) als Funktion von N_{eff}



Wir sehen insbesondere: Die Verteilung von T_{jgV} hängt in nicht-linearer Weise von N_{eff} ab.

Literaturbericht Die ursprüngliche Studie erschien in Robert L. Dorit, Hiroshi Akashi und Walter Gilbert, Absence of Polymorphism at the ZFY Locus on the Human Y Chromosome, *Science* 268, 1183–1185 (1995), siehe auch die Diskussionsbeiträge ("technical comments") dazu in *Science* 272, 1356–1362 (1996) von Y.-X. Fu und W.-H. Li, von P. Donnelly, S. Tavaré, D.J. Balding und R.C. Griffiths, von G. Weiss und A. von Haeseler und von J. Rogers, P.B. Samollow und A.G. Comuzzie, die insbesondere eine etwas unpräzise Anwendung der Koaleszenten-Theorie von Dorit et al korrigierten. Unsere

Darstellung fußt in weiten Teilen auf Kapitel 8.1 in J. Wakeley, *Coalescent Theory: An Introduction*, Roberts & Company, 2008.

1.2 Vorwärtsdynamik der Typenverteilung und Wright-Fisher-Diffusion

Wir betrachten in diesem Kapitel Modelle, die das Phänomen der „genetischen Drift“ (auch „Gendrift“ genannt), die Tatsache, dass die genetische Zusammensetzung einer endlichen Population sich im Lauf der Zeit aufgrund von Zufälligkeit in den Nachkommenszahlen ändert[6], beschreiben.

Beobachtung 1.15. Nehmen wir an, dass es in der Population verschiedene genetische Typen aus einer Menge E möglicher Typen (oder „Allele“) gibt, sagen wir $E = \{1, 2, \dots, d\}$.

Wir hatten in Beob. 1.3 den Typenvektor $\left(t_r^{(N)}\right)_{r\geq 0} := \left((t_{r,1}^{(N)}, \dots, t_{r,N}^{(N)})\right)_{r\geq 0}$ ($t_{r,i}^{(N)}$ ist der Typ von Individuum i in Generation r) betrachtet, er ist rekursiv gegeben durch

$$t_{r,i}^{(N)} = \sum_{k\in[N]} t_{r-1,k}^{(N)} \mathbf{1}_{\{A_{r,i}^{(N)}=k\}} \in E, \quad i \in [N],\ r > r_0.$$

Wir können daraus den Typenzählprozess $X_r^{(N)} = (X_{r,e}^{(N)})_{e\in E}$, $r \geq r_0$ ablesen:

$$X_{r,e}^{(N)} := \sum_{i=1}^{N} \mathbf{1}_{\{t_{r,i}^{(N)}=e\}}$$

gibt an, wieviele Individuen in der r-ten Generation den Typ $e \in E$ besitzen.

$(X_r^{(N)})_{r\in\mathbb{N}_0}$ *ist* eine Markovkette, mit Def. 1.2 und Ann. 1.4 (die „Indexnummern“ der Kinder werden rein zufällig vergeben, wir können daher für die Verteilung o.E. annehmen, dass gegeben $\left(X_{r,1}^{(N)}, \dots, X_{r,d}^{(N)}\right) = (x_1, \dots, x_d)$ in Generation r die Individuen $1, \dots, x_1$ den Typ 1, die Individuen $x_1 + 1, \dots, x_1 + x_2$ den Typ 2, etc. besitzen) gilt

$$\begin{aligned}
&\mathscr{L}\Big(\big(X_{r+1,1}^{(N)}, X_{r+1,2}^{(N)}, \dots, X_{r+1,d}^{(N)}\big) \,\Big|\, \big(X_{r,1}^{(N)}, \dots, X_{r,d}^{(N)}\big) = (x_1, \dots, x_d)\Big) \\
&= \mathscr{L}\Big(\big(\sum_{i_1=1}^{x_1} \nu_{i_1}^{(N,r)}, \sum_{i_2=x_1+1}^{x_1+x_2} \nu_{i_2}^{(N,r)}, \dots, \sum_{i_d=x_1+\dots+x_{d-1}+1}^{x_1+\dots+x_d} \nu_{i_d}^{(N,r)}\big)\Big)
\end{aligned}$$

Wir betrachten im Folgenden Nachkommenszahlvektoren von Cannings-Modellen $\nu^{(N)}$ wie in Def. 1.1 bzw. Def. 1.2 mit (nur) 2 verschiedene Typen (sagen wir $E = \{0, 1\}$ und

$$X_{t+1} = \sum_{i=1}^{X_t} \nu_i^{(N,t)}, \quad t \in \mathbb{N}$$

[6] Sewall Wright [W31, p. 106] schreibt: “Merely by chance one or the other of the allelomorphs may be expected to increase its frequency in a given generation and in time the proportions may drift a long way from the original values.”
(Gewissermaßen ist die Tatsache, dass wir in den Genealogien von Stichproben Verschmelzungen beobachten – wie im vorigen Kapitel 1.1 – ebenfalls ein Ausdruck der genetischen Drift, man sieht das Phänomen aber expliziter, wenn man die Dynamik der Typenverteilung in der „Vorwärtsrichtung“ betrachtet.)

(mit einem Startwert $X_0 \in \{0, 1, \dots, N\}$). Wir interpretieren

$$X_t = \text{Anz. Typ-1-Individuen in Generation } t.$$

Definition 1.16. Diese Markovkette nennen wir den Typenanzahlprozess im 2-Typ Cannings-Modell mit Nachkommensvektor(-verteilung) $\mathscr{L}(\nu^{(N)})$.

Erinnerung. Im WF-Modell (s. Bsp. 1.1 und Diskussion nach Beob. 1.3) hatten wir gesehen:

$$\mathbb{P}\big(X^{(N)}_{r+1,1} = y \,\big|\, X^{(N)}_{r,1} = x\big) = \binom{N}{y}\big(\frac{x}{N}\big)^y\big(1 - \frac{x}{N}\big)^{N-y}.$$

Beispiel 1.17 (Moran-Modell (Skelettkette))**.** Im sog. Moran-Modell (mit Populationsgröße N) ist die Verteilung des Nachkommensvektors $\nu = (\nu_1, \dots, \nu_N)$ die einer rein zufällig gewählten Permutation von

$$\big(2, 0, \underbrace{1, 1, \dots, 1}_{(N-2)\text{-mal}}\big),$$

d.h. ein zufällig gewähltes Individuum hat 2 Nachkommen, ein anderes, zufällig gewähltes Individuum hat 0 Nachkommen und alle anderen genau einen (man kann dies auch so interpretieren, dass ein Individuum einen Nachkommen hat, ein anderes stirbt und alle übrigen weiterleben).

$$\mathbb{P}(X_{t+1} = x \pm 1 \,|\, X_t = x) = \frac{x(N-x)}{N(N-1)}, \quad x \in \{1, 2, \dots, N-1\}$$

(und $\mathbb{P}(X_{t+1} = 0 \,|\, X_t = 0) = \mathbb{P}(X_{t+1} = N \,|\, X_t = N) = 1, \mathbb{P}(X_{t+1} = x \,|\, X_t = x) = \frac{x(x-1)+(N-x)(N-x-1)}{N(N-1)}$).

Es ist (in der Notation von Satz 1.8) $c_N = \frac{2}{N(N-1)}$ und $d_N = 0$, insbesondere sind die Voraussetzungen von Satz 1.8 hier erfüllt.

Beobachtung 1.18. Es gilt (mit Austauschbarkeit von $\nu(N)$)

$$\begin{aligned}
\mathbb{E}\big[X_{t+1} \,\big|\, X_t = x\big] &= \sum_{i=1}^{x} \mathbb{E}[\nu_i^{(N)}] = x, \\
\mathbb{E}\big[X^2_{t+1} \,\big|\, X_t = x\big] &= \sum_{i,j=1}^{x} \mathbb{E}[\nu_i^{(N)}\nu_j^{(N)}] = x\mathbb{E}[(\nu_1^{(N)})^2] + x(x-1)\mathbb{E}[\nu_1^{(N)}\nu_2^{(N)}].
\end{aligned}$$

Es ist

$$\begin{aligned}
\mathbb{E}[\nu_1^{(N)}\nu_2^{(N)}] = \mathbb{E}[\nu_1^{(N)}\nu_3^{(N)}] = \dots = \mathbb{E}[\nu_1^{(N)}\nu_N^{(N)}] &= (N-1)^{-1}\mathbb{E}[\nu_1^{(N)}(\nu_2^{(N)} + \dots + \nu_N^{(N)})] \\
&= (N-1)^{-1}\mathbb{E}[\nu_1^{(N)}(N - \nu_1^{(N)})] = N/(N-1) - (N-1)^{-1}\mathbb{E}[(\nu_1^{(N)})^2] \\
&= 1 - (N-1)^{-1}\mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)} - 1)]
\end{aligned}$$

(d.h.

$$\mathrm{Cov}[\nu_1^{(N)}, \nu_2^{(N)}] = -\frac{1}{N-1}\mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)} - 1)] = -c_N$$

mit c_N aus (1.5), wegen $\nu_1^{(N)} + \dots + \nu_N^{(N)} = N$ ist $\mathrm{Cov}[\nu_1^{(N)}, \nu_2^{(N)}] < 0$).

Für die Varianz ergibt sich:

$$\begin{aligned}\operatorname{Var}\big[X_{t+1}\,\big|\,X_t = x\big] &= x\mathbb{E}[(\nu_1^{(N)})^2] + x(x-1)\mathbb{E}[\nu_1^{(N)}\nu_2^{(N)}] - x^2\\ &= x\mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)}-1)] + x + (x^2-x)\Big(1 - \frac{1}{N-1}\mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)}-1)]\Big) - x^2\\ &= \mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)}-1)]\Big(x - \frac{x^2-x}{N-1}\Big) = \frac{\mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)}-1)]}{N-1}x(N-x).\end{aligned}$$

Die erwartete *Stichprobenheterozygotie*

$$\mathbb{E}\Big[2\frac{X_t}{N}\big(1-\frac{X_t}{N}\big)\Big]$$

ist die Wahrscheinlichkeit, in einer zufälligen Stichprobe der Größe zwei (mit Zurücklegen) zwei unterschiedliche genetischen Typen vorzufinden.

Lemma 1.19. *Für die erwartete Stichprobenheterozygotie gilt*

$$\mathbb{E}\Big[2\frac{X_t}{N}\big(1-\frac{X_t}{N}\big)\Big] = (1-c_N)\mathbb{E}\Big[2\frac{X_{t-1}}{N}\big(1-\frac{X_{t-1}}{N}\big)\Big] = \cdots = (1-c_N)^t\mathbb{E}\Big[2\frac{X_0}{N}\big(1-\frac{X_0}{N}\big)\Big].$$

Beweis. Nach Beob. 1.18 ist

$$\begin{aligned}\mathbb{E}\big[X_{t+1}(N-X_{t+1})\,\big|\,X_t = x\big] &= Nx - \operatorname{Var}\big[X_{t+1}\,\big|\,X_t = x\big] - x^2\\ &= x(N-x) - c_N x(N-x) = (1-c_N)x(N-x)\end{aligned}$$

und somit $E\big[X_{t+1}(N-X_{t+1})\big] = \mathbb{E}\Big[E\big[X_{t+1}(N-X_{t+1})\,\big|\,X_t\big]\Big] = (1-c_N)E\big[X_t(N-X_t)\big]$. □

Sei

$$T_{\text{fix}} := \inf\{t \in \mathbb{N}_0 : X_t = 0 \text{ oder } X_t = N\}$$

die Zeit, in Generationen gemessen, bis einer der beiden Typen in der Population fixiert (und der andere demnach verschwunden) ist.

Bemerkung. Sofern

$$\operatorname{Var}[\nu_1^{(N)}] > 0$$

gilt, was wir stets annehmen, damit die Typenentwicklung nicht-trivial ist (aus $\operatorname{Var}[\nu_1^{(N)}] = 0$ folgt $\nu_i^{(N)} \equiv 1$), ist

$$T_{\text{fix}} < \infty \quad \text{f.s.}$$

Man kann dies mit einem Markovketten-Argument leicht einsehen: Die Kette (X_r) hat nur 0 und N als absorbierende Zustände und für jedes $x \in \{1,2,\ldots,N-1\}$ gilt $\mathbb{P}_x(X_N = 0) > 0$ (das Ereignis $\{\nu_1^{(N)} + \cdots + \nu_x^{(N)} < x\}$ hat positive Wahrscheinlichkeit). Wir könnten natürlich auch analog argumentieren, dass $\mathbb{P}_x(X_N = N) > 0$ gilt.

Alternativ können wir beobachten, dass

$$M_t := X_t^2 - c_N \sum_{s=0}^{t-1} X_s(N - X_s)$$

ein Martingal ist (Beob. 1.18 zeigt $\mathbb{E}\big[X_{t+1}^2 - X_t^2\,\big|\,X_t\big] = c_N X_t(N-X_t)$). Als nach oben (durch N^2) beschränktes Martingal konvergiert M_t, andererseits ist $\{T_{\text{fix}} = \infty\} \subset \{M_t \to -\infty\}$.

Lemma 1.20. *Es gilt* $\mathbb{P}_x(X_{T_{\text{fix}}} = N) = \dfrac{x}{N}$.

Beweis. Nach Beob. 1.18 ist

$$\sum_{y=0}^{N} y\mathbb{P}_x(X_1 = y) = x \quad \text{für jedes } x \in \{0, 1, \dots, N\}$$

sei $f(x) := \mathbb{P}_x(X_{T_{\text{fix}}} = N)$, Zerlegung nach dem ersten Schritt für die Markovkette $(X_t)_{t=0,1,\dots}$ zeigt

$$f(x) = \sum_{y=0}^{N} \mathbb{P}_x(X_1 = y) f(y), \quad x \in \{1, 2, \dots, N-1\}$$

und für die Randwerte gilt (offenbar) $f(0) = 0$, $f(N) = 1$. Nach obigem löst auch $\tilde{f}(x) := \frac{x}{N}$ dieses Gleichungssystem mit denselben Randwerten, wegen der Eindeutigkeit der Lösung (verwende etwa ein Randminimumprinzip, z.B. [Kl, Satz 19.6]) gilt $f(x) = \frac{x}{N}$.

Alternativ: Obiges bedeutet

$$\mathbb{E}\big[X_{t+1} \,\big|\, \mathscr{F}_t^{(N)}\big] = X_t$$

(mit $\mathscr{F}_t^{(N)} = \sigma(X_s, s \le t)$, nach Beob. 1.18 zusammen mit der Markov-Eigenschaft), d.h. $X = (X_t)_{t=0,1,\dots}$ ist ein (beschränktes) Martingal, 0 und N sind absorbierende Zustände für X. Also gilt (mit dem optional sampling-Satz)

$$x = \mathbb{E}_x[X_0] = \mathbb{E}_x[X_{T_{\text{fix}}}] = N \cdot \mathbb{P}_x(X_{T_{\text{fix}}} = N) + 0 \cdot \mathbb{P}_x(X_{T_{\text{fix}}} = 0).$$

□

Bemerkung. Dasselbe Argument funktioniert bei allgemeinem Typenraum: Die Wahrscheinlichkeitkeit, dass ein gewisser Typ fixiert, ist gleich seiner Anfangsfrequenz. Dazu vergröbern wir die Typenmenge in „Fokaltyp“ und Rest, dann haben wir es wieder mit einer 2 Typ-Situation zu tun.

Beobachtung 1.21. *Für das 2-Typ (zeitdiskrete) Moran-Modell gilt für*

$$\lim_{N\to\infty} \frac{\mathbb{E}_{k_N}[T_{\text{fix}}]}{N^2 H(p)} = 1 \tag{1.16}$$

sofern $k_N/N \to p \in (0,1)$, *wo* $H(p) = -p\log(p) - (1-p)\log(1-p)$.

Beweis. $f_N(k) := \mathbb{E}_k[T_{\text{fix}}]$ löst (Zerlegung nach dem ersten Schritt):

$$f_N(k) = 1 + \frac{k(N-k)}{N(N-1)} f_N(k+1) + \frac{k(N-k)}{N(N-1)} f_N(k-1) + \frac{k(k-1)+(N-k)(N-k-1)}{N(N-1)} f_N(k)$$

für $k = 2, 3, \dots, N-1$ mit Randwerten $f_N(0) = f_N(N) = 0$ oder äquivalent

$$\frac{k(N-k)}{N(N-1)}\big(f_N(k+1) - f_N(k)\big) = \frac{k(N-k)}{N(N-1)}\big(f_N(k) - f_N(k-1)\big) - 1$$

Somit erfüllt $g_k := f_N(k+1) - f_N(k)$ [wir unterdrücken die N-Abhängigkeit von g_k in der Notation]

$$g_k = g_{k-1} - \frac{N(N-1)}{k(N-k)} = g_{k-1} - \frac{N-1}{k} - \frac{N-1}{N-k}, \quad k = 1, 2, \dots, N-1$$

d.h.

$$g_k = g_0 - (N-1)\sum_{j=1}^{k}\Big(\frac{1}{j} + \frac{1}{N-j}\Big) = g_0 - (N-1)\big(H_k + H_{N-1} - H_{N-k-1}\big), \quad k = 1, 2, \dots, N-1$$

wo

$$H_m := 1 + \frac{1}{2} + \cdots + \frac{1}{m} = \log(m) + \gamma + O\big(\frac{1}{m}\big)$$

die m-te harmonische Zahl[7] ist (und wir setzen $H_0 := 0$).

Folglich

$$f_N(k) = \underbrace{f_0}_{=0} + \sum_{j=0}^{k-1} g_j = g_0 + \sum_{j=1}^{k-1} g_j = kg_0 - (N-1)\sum_{j=1}^{k-1}\Big(H_j + H_{N-1} - H_{N-j-1}\Big)$$

und wegen $f_N(N) = 0$ ist also

$$g_0 = \frac{(N-1)^2}{N} H_{N-1} + \frac{N-1}{N} H_{N-1} = (N-1)H_{N-1}$$

und

$$f_N(k) = (N-1)H_{N-1} - (N-1)\sum_{j=1}^{k-1}\Big(H_j - H_{N-j-1}\Big)$$

Für $k = k_N$ mit $k_N/N \to p \in (0,1)$ ergibt sich

$$\begin{aligned}
f_N(k_N) &= -(N-1)\int_1^{Np} \log(x) - \log(N-x)\,dx + O(N\log N) \\
&= -N\int_1^{Np} \log(x) - \log(N-x)\,dx + O(N\log N) \\
&= -N\Big[x\log(x) + (N-x)\log(N-x)\Big]_{x=1}^{x=Np} + O(N\log N) \\
&= -N\Big(Np\log(Np) + (N-Np)\log(N-Np) - N\log N\Big) + O(N\log N) \\
&= N^2\big(-p\log(p) - (1-p)\log(1-p)\big) + O(N\log N).
\end{aligned}$$

□

[7] Für die explizite Asymptotik siehe z.B. M. Abramowitz und I.A. Stegun, *Handbook of mathematical functions*, Dover publications, 9. Aufl., 1970, 6.3.18 beachte $\psi(z) = \Gamma'(z)/\Gamma(z)$ $(= \frac{d}{dz}\log(\Gamma(z))$, die Digamma-Funktion, erfüllt $\psi(n) + \gamma = \sum_{k=1}^{n-1} k^{-1}$ nach 6.3.2); für ein Argument „von Hand“ beachte (vgl. z.B. Heuser, *Lehrbuch der Analysis 1*, Aufg. 88.5) $d_m := \sum_{k=1}^{m} \frac{1}{k} - \int_1^m \frac{1}{x}\,dx$ erfüllt $d_m \geq 0$ (Integralvergleich) und $d_m - d_{m+1} = -\frac{1}{m+1} + \int_m^{m+1} \frac{1}{x}\,dx = \int_m^{m+1} \frac{m+1-x}{x(m+1)}\,dx \in (0, \frac{1}{m(m+1)})$, d.h. $d_m \searrow \gamma$ (die Euler-Mascheroni-Konstante) und $d_m - \gamma \leq \sum_{\ell=m}^{\infty}(d_\ell - d_{\ell-1}) \leq C/m$]

Die Zerlegung nach dem ersten Schritt wie im Beweis von Beob. 1.21 gilt auch allgemein. Sei X_t der Typenanzahlprozess in einem 2-Typ Cannings-Modell wie in Def. 1.16 mit Nachkommensvektor $\mathscr{L}(\nu^{(N)})$. Stets ist

$$\mathbb{E}_k[T_{\text{fix}}] = 1 + \sum_{j=0}^{N} \mathbb{P}_k(X_1 = j)\mathbb{E}_j[T_{\text{fix}}]$$

(dies verwendet nur die Markoveigenschaft). Allerdings ist das resultierende lineare Gleichungssystem in $N-1$ Unbekannten i.A. nicht explizit lösbar (das Moran-Modell hat hier eine besonders angenehme Struktur mit einer Koeffizientenmatrix in Tridiagonalform; beispielsweise für das Wright-Fisher-Modell ist die Koeffizientenmatrix voll besetzt).

Ein prinzipiell gangbarer Weg liegt in der Heuristik, dass $\mathbb{E}_k[T_{\text{fix}}]$ für großes N in „genügend glatter" Weise von $p := \frac{k}{N}$ abhängen sollte, und dann eine Taylorenwicklung von $\mathbb{E}_{pN}[T_{\text{fix}}]$ als Funktion von p anzusetzen.

Satz 1.22 (*)**.** *Betrachte eine Schar von (Typenhäufigkeitsprozessen von) Cannings-Modellen, mit der Populationsgröße N parametrisiert. Falls*

$$\operatorname{Var}[\nu_1^{(N)}] \to \sigma^2 \in (0, \infty)$$

und

$$\sup_N \mathbb{E}[(\nu_1^{(N)})^4] < \infty$$

gilt (dies impliziert $c_N = \frac{\operatorname{Var}[\nu_1^{(N)}]}{N-1} \sim \sigma^2/N$ und $d_N = O(1/N^2)$, d.h. insbesondere auch sind die Voraussetzungen von Satz 1.8 erfüllt), so ist

$$\lim_{N\to\infty} c_N \frac{\mathbb{E}_{k_N}[T_{\text{fix}}]}{2H(p)} = 1 \tag{1.17}$$

mit $H(p) = -p\log(p) - (1-p)\log(1-p)$.

Beweisskizze/-heuristik. Zerlegung nach dem ersten Schritt zeigt

$$\mathbb{E}_x[T_{\text{fix}}] = 1 + \sum_{y=0}^{N} \mathbb{P}_x(X_1 = y)\mathbb{E}_y[T_{\text{fix}}], \quad x = 1, 2, \ldots, N-1$$

(mit Randwerten $\mathbb{E}_0[T_{\text{fix}}] = \mathbb{E}_N[T_{\text{fix}}] = 0$).

Nehmen wir an, es gibt eine genügend glatte Funktion $f : [0,1] \to \mathbb{R}_+$ mit

$$f\Big(\frac{x}{N}\Big) = \mathbb{E}_x[T_{\text{fix}}],$$

so gilt (Taylor-Entwicklung von f um $\frac{x}{N}$)

$$\begin{aligned}
f\Big(\frac{x}{N}\Big) &= 1 + \sum_{y=0}^{N} \mathbb{P}_x(X_1 = y) f\Big(\frac{y}{N}\Big) \\
&= 1 + \sum_{y=0}^{N} \mathbb{P}_x(X_1 = y)\Big[f\Big(\frac{x}{N}\Big) + \Big(\tfrac{y-x}{N}\Big) f'\Big(\frac{x}{N}\Big) + \frac{1}{2}\Big(\tfrac{y-x}{N}\Big)^2 f''\Big(\frac{x}{N}\Big)\Big] + R(N,x) \\
&= 1 + f\Big(\frac{x}{N}\Big) + \frac{1}{N} f'\Big(\frac{x}{N}\Big)\mathbb{E}_x[X_1 - x] + \frac{1}{2}\frac{1}{N^2} f''\Big(\frac{x}{N}\Big)\operatorname{Var}_x[X_1 - x] + R(N,x) \\
&= 1 + f\Big(\frac{x}{N}\Big) + \frac{c_N}{2}\frac{x(N-x)}{N^2} f''\Big(\frac{x}{N}\Big) + R(N,x)
\end{aligned}$$

mit Restterm

$$R(N,x) = \sum_{y=0}^{N} \mathbb{P}_x(X_1 = y)\frac{1}{6}\big(\tfrac{y-x}{N}\big)^3 f'''\big(\zeta(\tfrac{x}{N},\tfrac{y}{N})\big)$$

($\zeta(\frac{x}{N},\frac{y}{N})$ ist eine Zahl zwischen $\frac{x}{N}$ und $\frac{y}{N}$).

Nun ist

$$\sum_{y=0}^{N} \mathbb{P}_x(X_1 = y)\big(\frac{y-x}{N}\big)^3 = \frac{1}{N^3}\mathbb{E}\Big[\big(\sum_{i=1}^{x}(\nu_i^{(N)} - 1)\big)^3\Big]$$

und die Voraussetzung $d_N/c_N \to 0$ impliziert, dass (geeignet skalierte) dritte Momente gegenüber der Varianz c_N vernachlässigt werden können, d.h. die Annahme

$$R(N,x) = o(c_N)$$

ist zumindest plausibel.

Schreibe $\frac{x}{N} = p$, also erfüllt f näherungsweise

$$f''(p) = -\frac{2}{c_N}\frac{1}{p(1-p)}, \quad 0 < p < 1 \tag{1.18}$$

mit den Randbedingungen $f(0) = f(1) = 0$. Man sieht nun leicht, dass eine explizite Lösung von (1.18) für $p \in (0,1)$ näherungsweise gegeben ist durch

$$f(p) - \frac{2}{c_N}\big(p\log(p) + (1-p)\log(1-p)\big),$$

denn

$$f'(p) = \frac{2}{c_N}(\log(p) + 1 - \log(1-p) - 1) = -\frac{2}{c_N}(\log(p) - \log(1-p)),$$

und

$$f''(p) = \big(-\frac{2}{c_N}(\log(p) - \log(1-p))\big)' = -\frac{2}{c_N}\Big(\frac{1}{p} + \frac{1}{1-p}\Big).$$

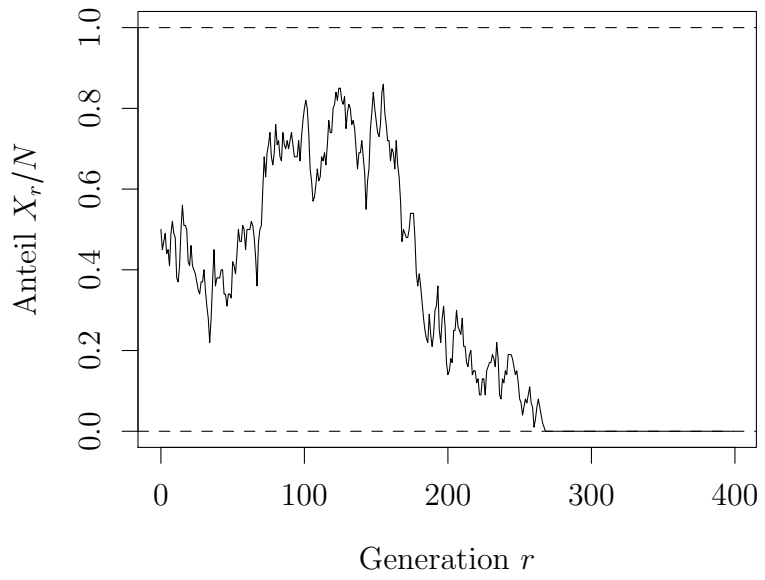
Unten arbeiten wir diese Heuristik zu einem vollständigen Beweis aus. □

Der Satz zeigt insbesondere, dass für $X_0 = N/2$ die erwartete Zeit bis zur Absorption von entweder a oder A gegeben ist durch

$$\mathbb{E}_{N/2}[T_{\text{fix}}] \approx -\frac{2}{c_N}\big(1/2\log(1/2) + 1/2\log(1/2)\big) = \frac{2\log(2)}{c_N} \approx 1.39 \cdot N$$

Generationen.

Abbildung 1.1: Anteilsprozess des Wright-Fisher-Modells im 2 Typ-Fall (eine Realisierung für $N = 100$)



1.2.1 Die (neutrale 2 Typ-)Wright-Fisher-Diffusion*

Sei $X = X^{(N)} = (X_r^{(N)})_{r\in\mathbb{N}_0}$ der Typenzählprozess in einem 2-Typ Cannings-Modell mit Nachkommensvektor(-verteilung) $\mathscr{L}(\nu^{(N)})$ (vgl. Def. 1.16). Für großes N liegt es nahe, anstelle der absoluten Zahlen des Anteil von Typ 1 zu betrachten, angesichts Lemma 1.19 sollte die „relevante“ Zeitskala $1/c_N$ (mit $c_N = \mathbb{E}[\nu_1^{(N)}(\nu_1^{(N)}-1)]/(N-1) = \mathrm{Var}[\nu_1^{(N)}]/(N-1)$) sein. Wir betrachten daher

$$Z_t^{(N)} := X_{\lfloor t/c_N\rfloor}^{(N)}/N, \quad t \geq 0. \tag{1.19}$$

Der stochastische Prozess $Z^{(N)} = (Z_t^{(N)})_{t\geq 0}$ hat (für jedes $N \in \mathbb{N}$) Werte in $[0,1]$, die Pfade sind zwar nicht stetig, aber für großes N sind die Sprünge typischerweise sehr klein, nämlich $O(1/N)$.

Beobachtung 1.23. Beob. 1.18 liefert für $z \in \frac{1}{N}\mathbb{Z}\cap[0,1]$

$$\begin{aligned}\mathbb{E}\big[Z_{t+c_N}^{(N)} \,\big|\, Z_t^{(N)} = z\big] &= z,\\ \mathrm{Var}\big[Z_{t+c_N}^{(N)} \,\big|\, Z_t^{(N)} = z\big] &= \mathbb{E}\big[\big(Z_{t+c_N}^{(N)} - z\big)^2 \,\big|\, Z_t^{(N)} = z\big] = c_N z(1-z)\end{aligned}$$

Wir werden sehen: Unter den Bedingungen von Satz 1.8 (nämlich $c_N \to 0$, $d_N/c_N \to 0$ mit $d_N = \frac{1}{(N-1)(N-2)}\mathbb{E}\big[\nu_1^{(N)}(\nu_1^{(N)}-1)(\nu_2^{(N)}-2)\big]$, für die Genealogien betrachtet ist c_N die Paar-, d_N die Tripelverschmelzungsw'keit, vgl. (1.5), (1.6)) ist

$$\mathbb{E}\big[|Z_{t+c_N}^{(N)} - z|^4 \,\big|\, Z_t^{(N)} = z\big] = o(c_N)$$

(siehe Lemma 1.27 unten), somit ist für $f \in C^2([0,1])$ und $z_N \in \frac{1}{N}\mathbb{Z} \cap [0,1]$

$$\begin{aligned}
&\frac{1}{c_N}\mathbb{E}\big[f\big(Z^{(N)}_{t+c_N}\big) - f(z_N)\,\big|\, Z^{(N)}_t = z_N\big] \\
&= \frac{1}{c_N} \sum_{w \in \frac{1}{N}\mathbb{Z}\cap[0,1]} \mathbb{P}\big(Z^{(N)}_{t+c_N} = w \,\big|\, Z^{(N)}_t = z_N\big)\big(f(w) - f(z_N)\big) \\
&= \frac{1}{c_N} \sum_{w \in \frac{1}{N}\mathbb{Z}\cap[0,1]} \mathbb{P}\big(Z^{(N)}_{t+c_N} = w \,\big|\, Z^{(N)}_t = z_N\big) \\
&\qquad\qquad \times \Big((w - z_N)f'(z_N) + \frac{1}{2}(w - z_N)^2 f''(\zeta_{z_N,w})\Big) \\
&= \frac{1}{2} z_N(1 - z_N) f''(z_N) + R_N(z_N)
\end{aligned}$$

mit

$$\begin{aligned}
R_N(z_N) &= \frac{1}{2} \sum_{w \in \frac{1}{N}\mathbb{Z}\cap[0,1]} \mathbb{P}\big(Z^{(N)}_{t+c_N} = w \,\big|\, Z^{(N)}_t = z_N\big)(w - z_N)^2\big(f''(\zeta_{z_N,w}) - f''(z_N)\big) \\
&= \mathbb{E}\Big[(Z^{(N)}_{t+c_N} - z_N)^2\big(f''(\zeta_{z_N, Z^{(N)}_{t+c_N}}) - f''(z_N)\Big|\, Z^{(N)}_t = z_N\Big]
\end{aligned}$$

und einem $\zeta_{z_N,w}$ zwischen z_N und w (wir verwenden Taylor-Entwicklung von f um z_N mit Restglied in Lagrange-Form).

Sei $\varepsilon > 0$. Da $f'' \in C([0,1])$ gibt es $\delta > 0$ mit

$$\sup\{|f''(x) - f''(y)| : x, y \in [0,1], |x - y| < \delta\} \le \varepsilon,$$

somit ist

$$\begin{aligned}
|R_N(z_N)| &\le \varepsilon \mathbb{E}\big[(Z^{(N)}_{t+c_N} - z_N)^2 \,\big|\, Z^{(N)}_t = z_N\big] + \frac{2\|f''\|_\infty}{\delta^2}\mathbb{E}\big[(Z^{(N)}_{t+c_N} - z_N)^4 \,\big|\, Z^{(N)}_t = z_N\big] \\
&= \varepsilon c_N z_N(1 - z_N) + \frac{2\|f''\|_\infty}{\delta^2} o(c_N),
\end{aligned}$$

also $\limsup_N |R_N(z_N)|/c_N \le \varepsilon$, d.h. $R_N(z_N) = o(c_N)$.

Somit gilt, sofern $z_N \to z \in [0,1]$ konvergiert,

$$\lim_{N\to\infty} \frac{1}{c_N}\mathbb{E}\big[f\big(Z^{(N)}_{t+c_N}\big) - f(z_N)\,\big|\, Z^{(N)}_{t+c_N} = z_N\big] = \frac{1}{2} z(1 - z) f''(z).$$

Wir nehmen die Bedingungen $c_N \to 0$, $d_N/c_N \to 0$ implizit für den Rest dieses Kapitels an.

Bericht (Diffusionsprozesse auf $\mathbb{R}$)**.** Ein starker Markovprozess $(X_t)_{t\ge 0}$ auf $\mathbb{R}$ mit stetigen Pfaden heißt ein Diffusionsprozess. Die Dynamik eines solchen Prozesses kann man durch seine „infinitesimalen Carakterististiken“ beschreiben:

$$\mathbb{E}\big[X_{t+h} - X_t \,\big|\, \mathscr{F}_t\big] = h\mu(X_t) + o(h), \quad \operatorname{Var}\big[X_{t+h} - X_t \,\big|\, \mathscr{F}_t\big] = h\sigma^2(X_t) + o(h)$$

für $h \downarrow 0$ mit gewissen Funktionen μ und σ^2 ($\mathscr{F}_t = \sigma(X_s, s \le t)$ beschreibt die durch Beobachtung des Pfads bis t verfügbare Information). Dann ist (für eine genügend glatte Funktion f auf $\mathbb{R}$)

$$Lf(x) = \lim_{h\downarrow 0} \frac{1}{h}\mathbb{E}\big[f(X_{t+h}) - f(X_t) \,\big|\, X_t = x\big] = \mu(x)f'(x) + \frac{1}{2}\sigma^2(x)f''(x)$$

der sogenannte Generator von X.

Das „kanonischste“ Objekt dieser Klasse, die Brownsche Bewegung $(B_t)_{t\ge 0}$, erfüllt dies mit $\mu(\cdot) \equiv 0$, $\sigma^2(\cdot) \equiv 1$. Man kann zeigen, dass man ein allgemeines X (unter geeigneten Bedingungen an μ und σ^2) als Lösung einer sogenannten stochastischen Differentialgleichung

$$dX_t = \mu(X_t)\,dt + \sigma(X_t)\,dB_t$$

gewinnen kann.

Lesenswerte Einführungen zu diesem Thema finden sich beispielsweise bei Breiman [B, Ch. 16] und bei Kersting und Wakolbinger [KW] [siehe Kap. 3 für die Brownsche Bewegung, Kap. 5 für Markovprozesse und Generatoren].

Definition 1.24. Der starke Markovprozess $Z = (Z_t)_{t\ge 0}$ mit Werten in $[0,1]$ und stetigen Pfaden, dessen Generator für $f \in C^2([0,1])$ gegeben ist durch

$$Lf(z) := \lim_{h\downarrow 0} \frac{1}{h}\mathbb{E}_z\big[f(Z_h) - f(Z_0)\big] = \frac{1}{2}z(1-z)f''(z), \quad z \in [0,1]$$

heißt die (neutrale 2-Typ) *Wright-Fisher-Diffusion.*

Bemerkung 1.25. 1. Die Wright-Fisher-Diffusion Z ist die eindeutige Lösung des folgenden wohlgestellten Martingalproblems: Für jedes $f \in C^2([0,1])$ ist der Prozess

$$M_t(f) := f(Z_t) - f(Z_0) - \int_0^t \frac{1}{2}Z_s(1-Z_s)f''(Z_s)\,ds \tag{1.20}$$

ein stetiges Martingal (bzgl. seiner kanonischen Filtration), wobei $M_0(f) = 0$. (Die Wohl-Gestelltheit zeigen wir unten in Satz 1.26, man könnte sie auch aus allgemeineren Argumenten ableiten, vgl. [St3] und [EK].)

2. Äquivalent könnten wir die Eigenschaft

$$Z \text{ stetiges Martingal mit Werten in } [0,1] \text{ und } \langle Z\rangle_t = \int_0^t Z_s(1-Z_s)\,ds$$

betrachten.

Die Itō-Formel zeigt, dass Z dann das Martingalproblem mit Generator

$$Lf(z) = \frac{1}{2}z(1-z)f''(z) \quad \text{für } f \in C^2([0,1])$$

löst, denn

$$\begin{aligned} f(Z_t) - f(Z_0) &= \int_0^t f'(Z_s)\,dZ_s + \frac{1}{2}\int_0^t f''(Z_s)\,d\langle Z\rangle_s \\ &= \int_0^t f'(Z_s)\,dZ_s + \frac{1}{2}\int_0^t Z_s(1-Z_s)f''(Z_s)\,ds, \end{aligned}$$

d.h. $f(Z_t) - f(Z_0) - \int_0^t \frac{1}{2} Z_s(1-Z_s) f''(Z_s)\, ds = \int_0^t f'(Z_s)\, dZ_s$ *ist* ein Martingal.

3. Ein weiterer Blickpunkt: Z ist die (eindeutige, starke) Lösung der stochastischen Differentialgleichung

$$dZ_t = \sqrt{Z_t(1-Z_t)}\, dB_t,$$

wo (B_t) standard-Brownsche Bewegung.

Bemerkung. Um die Konvergenz von $Z^{(N)}$ gegen Z in Verteilung zu beschreiben, müssen wir die Objekte auf einem gemeinsamen Wertebereich formulieren. Dazu dient „kanonischerweise“ der Skorohod-Raum der càdlàg-Funktionen (vgl. z.B. [EK, Ch. 3.5] und Übung 1.4* a).

Alternativ können wir $Z^{(N)}$ via Polygonzug-Approximation zum stetigen Pfad machen (vgl. etwa den Beweis des Invarianzprinzips in [St3, Kap. 1.3]), dann können wir als gemeinsamen Wertebereich die Menge der stetigen Funktionen $C([0,\infty),[0,1])$ verwenden. Wir metrisieren ihn z.B. via

$$d(f,g) := \sum_{T=1}^{\infty} 2^{-T} \Big(\|(f-g)\mathbf{1}_{[0,T]}\|_\infty \wedge 1 \Big)$$

(dies metrisiert die lokal-gleichmäßige Konvergenz), damit wird $C([0,\infty),[0,1])$ polnisch.

Wir betrachten dann im Folgenden (ggfs. implizit) die stetige Polygonzug-Version von $Z^{(N)}$,

$$Z_t^{(N)} = \big(c_N(\lfloor t/c_N \rfloor + 1) - t\big) \frac{1}{N} X^{(N)}_{\lfloor t/c_N \rfloor} + \big(t - c_N \lfloor t/c_N \rfloor\big) \frac{1}{N} X^{(N)}_{\lfloor t/c_N \rfloor + c_N}, \quad t \geq 0$$

und fassen diese als ZV mit Werten in $C([0,\infty),[0,1])$ auf.

Satz 1.26. *Es gelte $c_N \to 0$, $d_N/c_N \to 0$ und $Z_0^{(N)} = z_N \to z \in (0,1)$. Dann konvergiert der zufällige Pfad $(Z_t^{(N)})_{t\geq 0}$ in Verteilung gegen die Wright-Fisher-Diffusion (Z_t) mit Startpunkt $Z_0 = z$.*

Diskussion. Satz 1.26 besagt, dass für jedes stetige, beschränkte Funktional $\varphi : C([0,\infty),[0,1]) \to \mathbb{R}$ gilt

$$\mathbb{E}\Big[\varphi\big((Z_t^{(N)})_{t\geq 0}\big)\Big] \underset{N\to\infty}{\longrightarrow} \mathbb{E}\Big[\varphi\big((Z_t)_{t\geq 0}\big)\Big].$$

Beispiele solcher Funktionale sind $\varphi(f) = f(t)$,

$$\varphi(f) = \psi(f(t_1), f(t_2), \ldots, f(t_k)) \quad \text{für } \psi : [0,1]^k \to \mathbb{R} \text{ stetig},$$

(d.h. Satz 1.26 impliziert insbesondere die Konvergenz der endlich-dimensionalen Verteilungen), $\varphi(f) = \sup_{s\leq t} f(s)$, $\varphi(f) = \int_0^t \psi(f(s))\, ds$ für $\psi : [0,1] \to \mathbb{R}$ stetig.

Allerdings $\varphi : f \mapsto \inf\{t \geq 0 : f(t) = 0\}$ ist nicht stetig , d.h. Satz 1.26 impliziert nicht direkt Satz 1.22.

Zum Beweis von Satz 1.26 benötigen wir folgendes Lemma:

Lemma 1.27. *Für $y \in [0,1]$ sei $B^{(N,y)} := \frac{1}{N}\sum_{i=1}^{\lfloor Ny\rfloor} \nu_i^{(N)}$. (d.h. $B^{(N,y)} =^d \mathscr{L}(Z_{c_N}^{(N)} | Z_0 = \lfloor Ny\rfloor)/N$, vgl. Beob. 1.23 für 1. und 2. Moment).*

Falls $c_N \to 0$, $d_N/c_N \to 0$, so gilt für $y \in \frac{1}{N}\mathbb{Z} \cap [0,1]$

$$\sup_{y\in\frac{1}{N}\mathbb{Z}\cap[0,1]} \left|\mathbb{E}\Big[\big(B^{(N,y)} - y\big)^3\Big]\right| \le c_N y(1-y) \times r_{N,3} \tag{1.21}$$

$$\mathbb{E}\Big[\big(B^{(N,y)} - y\big)^4\Big] \le c_N y(1-y) \times r_{N,4} \tag{1.22}$$

mit gewissen Folgen $r_{N,3} \to_{N\to\infty} 0$, $r_{N,4} \to_{N\to\infty} 0$, insbesondere sind $\mathbb{E}\Big[\big(B^{(N,y)} - y\big)^3\Big]$, $\mathbb{E}\Big[\big(B^{(N,y)} - y\big)^3\Big] = o(c_N)$ für $N \to \infty$ gleichmäßig in y.

Falls zusätzlich

$$\sigma^2 := \lim_{N\to\infty} \mathrm{Var}[\nu_1^{(N)}] \in (0,\infty) \quad \text{existiert und} \quad K_4 := \limsup_{N\to\infty} \mathbb{E}[(\nu_1^{(N)})^4] < \infty, \tag{1.23}$$

so gilt für ein $C < \infty$

$$\left|\mathbb{E}\Big[\big(B^{(N,y)} - y\big)^3\Big]\right| \le \frac{C}{N^2} y(1-y), \quad \mathbb{E}\Big[\big(B^{(N,y)} - y\big)^4\Big] \le \frac{C}{N^2} y^2(1-y)^2 \tag{1.24}$$

gleichmäßig in $y \in [0,1] \cap \frac{1}{N}\mathbb{Z}$ und $N \ge N_0$ (d.h. $r_{N,3} \vee r_{N,4} = O(1/N)$, wie im Wright-Fisher-Modell).

Beweis von Lemma 1.27 . Zu (1.21) [Wir lassen im Folgenden den oberen Index von $\nu_i^{(N)}$ weg und schreiben kürzer ν_i] : Aus Symmetriegründen genügt es, $0 < y \le 1/2$ zu betrachten (sonst gehe zu $1-y$ über, beachte $1 - B^{(N,1-y)} =^d B^{(N,y)}$).

Es ist für $y \in \{0, 1/N, 2/N, \dots, 1/2\}$

$$\begin{aligned}\left|\mathbb{E}\Big[\big(B^{(N,y)} - y\big)^3\Big]\right| &= \left|\frac{1}{N^3}\mathbb{E}\Big[\big(\sum_{i=1}^{Ny}(v_i - 1)\big)^3\Big]\right| \\ &= \Big|\frac{y}{N^2}\mathbb{E}[(\nu_1-1)^3] + 3\frac{y(Ny-1)}{N^2}\mathbb{E}[(\nu_1-1)^2(\nu_2-1)] \\ &\qquad + \frac{y(Ny-1)(Ny-2)}{N^2}\mathbb{E}[(\nu_1^{(N)}-1)(\nu_2^{(N)}-1)(\nu_3^{(N)}-1)]\Big| \end{aligned} \tag{1.25}$$

Analog zu (1.22):

$$\begin{aligned}&\mathbb{E}\Big[\big(B^{(N,y)} - y\big)^4\Big] = \frac{1}{N^4}\mathbb{E}\Big[\big(\sum_{i=1}^{Ny}(v_i - 1)\big)^4\Big] \\ &= \frac{y}{N^3}\mathbb{E}[(\nu_1-1)^4] + 4\frac{y(Ny-1)}{N^3}\mathbb{E}\big[(\nu_1-1)^3(\nu_2-1)\big] \\ &\quad + 3\frac{y(Ny-1)}{N^3}\mathbb{E}\big[(\nu_1-1)^2(\nu_2-1)^2\big] + 3\frac{y(Ny-1)(Ny-2)}{N^3}\mathbb{E}\big[(\nu_1-1)^2(\nu_2-1)(\nu_3-1)\big] \\ &\quad + \frac{y(Ny-1)(Ny-2)(Ny-3)}{N^3}\mathbb{E}\big[(\nu_1-1)(\nu_2-1)(\nu_3-1)(\nu_4-1)\big] \end{aligned} \tag{1.26}$$

Kombinatorische Vorbetrachtungen, Folgerungen aus der Austauschbarkeit: Es ist

$$\begin{aligned}\mathbb{E}\big[(\nu_1-1)^2(\nu_2-1)\big] &= \frac{1}{N-1}\mathbb{E}\Big[(\nu_1-1)^2\Big(\underbrace{(\nu_2-1)+(\nu_3-1)+\cdots+(\nu_N-1)}_{=-(\nu_1-1)}\Big)\Big]\\ &= -\frac{1}{N-1}\mathbb{E}\big[(\nu_1-1)^3\big],\end{aligned}$$

$$\begin{aligned}\mathbb{E}\big[(\nu_1-1)^3(\nu_2-1)\big] &= \frac{1}{N-1}\mathbb{E}\Big[(\nu_1-1)^3\Big(\underbrace{(\nu_2-1)+(\nu_3-1)+\cdots+(\nu_N-1)}_{=-(\nu_1-1)}\Big)\Big]\\ &= -\frac{1}{N-1}\mathbb{E}\big[(\nu_1-1)^4\big], \qquad (1.27)\end{aligned}$$

analog

$$\begin{aligned}\mathbb{E}\big[(\nu_1-1)^2(\nu_2-1)(\nu_3-1)\big] &= \frac{1}{N-2}\mathbb{E}\Big[(\nu_1-1)^2(\nu_2-1)\Big(\underbrace{(\nu_3-1)+(\nu_4-1)+\cdots+(\nu_N-1)}_{=-(\nu_1-1)-(\nu_2-1)}\Big)\Big]\\ &= -\frac{1}{N-2}\mathbb{E}\big[(\nu_1-1)^3(\nu_2-1)\big]-\frac{1}{N-2}\mathbb{E}\big[(\nu_1-1)^2(\nu_2-1)^2\big]\\ &= \frac{1}{(N-1)(N-2)}\mathbb{E}\big[(\nu_1-1)^4\big]-\frac{1}{N-2}\mathbb{E}\big[(\nu_1-1)^2(\nu_2-1)^2\big]\\ &\qquad (1.28)\end{aligned}$$

(wobei wir in der letzten Zeile (1.27) verwenden),

$$\begin{aligned}&\mathbb{E}\big[(\nu_1-1)(\nu_2-1)(\nu_3-1)(\nu_4-1)\big]\\ &= -\frac{1}{N-3}\mathbb{E}\big[(\nu_1-1)(\nu_2-1)(\nu_3-1)\big((\nu_1-1)+(\nu_2-1)+(\nu_3-1)\big)\big]\\ &= -\frac{3}{N-3}\mathbb{E}\big[(\nu_1-1)^2(\nu_2-1)(\nu_3-1)\big] = -\frac{3}{(N-1)_{3\downarrow}}\mathbb{E}\big[(\nu_1-1)^4\big]+\frac{3}{(N-2)_{2\downarrow}}\mathbb{E}\big[(\nu_1-1)^2(\nu_2-1)^2\big]\end{aligned}$$

(wobei wir in der letzten Zeile (1.28) verwenden) und

$$\begin{aligned}\mathbb{E}\big[(\nu_1-1)(\nu_2-1)(\nu_3-1)\big] &= -\frac{1}{N-2}\mathbb{E}\big[(\nu_1-1)(\nu_2-1)\big((\nu_1-1)+(\nu_2-1)\big)\big]\\ &= -\frac{2}{N-2}\mathbb{E}\big[(\nu_1-1)^2(\nu_2-1)\big] = -\frac{2}{(N-1)_{2\downarrow}}\mathbb{E}\big[(\nu_1-1)^3\big]\end{aligned}$$

(wobei wir in der letzten Zeile (1.28) verwenden).

Weiter ist

$$\begin{aligned}(\nu_1-1)^3 &= (\nu_1)_{3\downarrow}+\nu_1-1,\\ (\nu_1-1)^2(\nu_2-1)^2 &= (\nu_1)_{2\downarrow}(\nu_2)_{2\downarrow}-(\nu_1)_{2\downarrow}\nu_2-\nu_1(\nu_2)_{2\downarrow}+(\nu_1)_{2\downarrow}+(\nu_2)_{2\downarrow}+\nu_1\nu_2-\nu_1-\nu_2+1,\\ (\nu_1-1)^4 &= (\nu_1)_{4\downarrow}+2(\nu_1)_{3\downarrow}+(\nu_1)_{2\downarrow}-\nu_1+1,\end{aligned}$$

also

$$\begin{aligned}
\mathbb{E}\big[(\nu_1-1)^3\big] &= \mathbb{E}\big[(\nu_1)_{3\downarrow}\big],\\
\mathbb{E}\big[(\nu_1-1)^2(\nu_2-1)^2\big] &= \mathbb{E}\big[(\nu_1)_{2\downarrow}(\nu_2)_{2\downarrow}\big] + 2\mathbb{E}\big[(\nu_1)_{2\downarrow}\nu_2\big] + 2\mathbb{E}\big[(\nu_1)_{2\downarrow}\big] + \mathbb{E}\big[\nu_1\nu_2\big] - 1,\\
\mathbb{E}\big[(\nu_1-1)^4\big] &= \mathbb{E}\big[(\nu_1)_{4\downarrow}\big] + 2\mathbb{E}\big[(\nu_1)_{3\downarrow}\big] + \mathbb{E}\big[(\nu_1)_{2\downarrow}\big],
\end{aligned}$$

etc. Nach Voraussetzung ist

$$\lim_N \frac{1}{c_N}\frac{\mathbb{E}\big[(\nu_1)_{3\downarrow}\big]}{N^2} = \lim_N \frac{\mathbb{E}\big[(\nu_1)_{3\downarrow}\big]}{N\mathbb{E}\big[(\nu_1)_{2\downarrow}\big]} = \lim_N \frac{d_N}{c_N} = 0,$$

Argumente wie im Beweis von Satz 1.8 zeigen, dass dann auch

$$\frac{1}{c_N}\frac{\mathbb{E}\big[(\nu_1)_{4\downarrow}\big]}{N^2},\ \frac{1}{c_N}\frac{\mathbb{E}\big[(\nu_1)_{2\downarrow}(\nu_2)_{2\downarrow}\big]}{N^2} \xrightarrow[N\to\infty]{} 0 \quad \text{gilt.}$$

Demnach erfüllt

$$\begin{aligned}
e_N := \max\Big\{ &\frac{\mathbb{E}\big[(\nu_1-1)^4\big]}{N^3}, \frac{\big|\mathbb{E}\big[(\nu_1-1)^3\big]\big|}{N^2}, \frac{\big|\mathbb{E}\big[(\nu_1-1)^2(\nu_2-1)^2\big]\big|}{N^2}, \frac{\big|\mathbb{E}\big[(\nu_1-1)^3(\nu_2-1)\big]\big|}{N^2},\\
&\frac{\big|\mathbb{E}\big[(\nu_1-1)^2(\nu_2-1)(\nu_3-1)\big]\big|}{N}, \frac{\big|\mathbb{E}\big[(\nu_1-1)^2(\nu_2-1)\big]\big|}{N},\\
&\big|\mathbb{E}\big[(\nu_1-1)(\nu_2-1)(\nu_3-1)(\nu_4-1)\big]\big|, \big|\mathbb{E}\big[(\nu_1-1)(\nu_2-1)(\nu_3-1)\big]\big|\Big\}
\end{aligned}$$

$\lim_{N\to\infty} e_N/c_N = 0$.

Zurück zu (1.21): Aus Symmetriegründen genügt es, $0 < y \le 1/2$ zu betrachten. Wir sehen aus (1.25) und obigem: Für $y \in \{0, 1/N, 2/N, \dots, 1/2\}$

$$\Big|\mathbb{E}\Big[\big(B^{(N,y)} - y\big)^3\Big]\Big| \le \big(y + 3y^2 + y^3\big)e_N.$$

Analog für (1.22), aus (1.26) und obigem:

$$\mathbb{E}\Big[\big(B^{(N,y)} - y\big)^4\Big] \le \big(y + 7y^2 + 3y^3 + y^4\big)e_N \tag{1.29}$$

Wenn weiter (1.23) gilt, können wir per Inspektion (1.24) verifizieren. Wir beachten, dass

$$\begin{aligned}
\mathbb{E}[(\nu_1-1)^4] &\le 1 + \mathbb{E}[\nu_1^4] \le K_4 + 1,\\
\mathbb{E}[|\nu_1-1|^3] &= \mathbb{E}\big[((\nu_1-1)^4)^{3/4}\big] \le \big(\mathbb{E}[\nu_1^4]\big)^{3/4} \le K_4^{3/4},\\
\mathbb{E}[\nu_1(\nu_1-1)\nu_2(\nu_2-1)] &\le (\mathbb{E}[\nu_1^4]\mathbb{E}[\nu_2^4])^{1/2} \le K_4
\end{aligned}$$

gilt, d.h. es ist $e_N = O(1/N)$.

Weiterhin können wir

$$\frac{y}{N^3}\mathbb{E}[(\nu_1-1)^4] \le \frac{y^2}{N^2}(1+K_4)$$

abschätzen (denn $y \ge 1/N$), d.h. wir können die rechte Seite von (1.29) ersetzen durch $(8y^2+3y^3+y^4)e_N$.

□

Beweis von Satz 1.26. 1. Schritt (Jeder Häufungspunkt löst das Martingalproblem, vgl. Bem. 1.25):

Sei $\mathscr{F}_t^{(N)} := \sigma(Z_s^{(N)}, s \le t)$, es ist $\mathbb{E}\big[Z_{t+c_N}^{(N)} \,\big|\, \mathscr{F}_t^{(N)}\big] = Z_t^{(N)}$ und somit $\mathbb{E}\big[Z_{t+h}^{(N)} \,\big|\, \mathscr{F}_t^{(N)}\big] = Z_t^{(N)}$ für alle $t \ge 0, h > 0$; wie oben ist (siehe Beob. 1.23)

$$\mathbb{E}\big[Z_{t+c_N}^{(N)}(1-Z_{t+1/c_N}^{(N)}) \,\big|\, \mathscr{F}_t^{(N)}\big] = (1-c_N)Z_t^{(N)}(1-Z_t^{(N)}),$$

also

$$\mathbb{E}\big[(Z_{t+c_N}^{(N)})^2 \,\big|\, \mathscr{F}_t^{(N)}\big] = (Z_t^{(N)})^2 + c_N Z_t^{(N)}(1-Z_t^{(N)})$$

und somit ist

$$(Z_t^{(N)})^2 - c_N \sum_{j=0}^{\lfloor t/c_N\rfloor -1} Z_{c_N j}^{(N)}(1-Z_{c_N j}^{(N)})$$

ein Martingal. Demnach löst jeder Häufungspunkt der Familie $\mathscr{L}\big((Z_t^{(N)})_{t\ge 0}\big)$, $N \in \mathbb{N}$ das Martingalproblem der Wright-Fisher-Diffusion:

Die Summe oben kann bis auf einen kleinen „Randkorrekturterm“ als $\int_0^t Z_s^{(N)}(1-Z_s^{(N)})\,ds$ geschrieben werden, für jeden Limespunkt Z gilt

$$\mathbb{E}\Big[\big(Z_t - Z_s\big)g_0(Z_{s_0})\cdots g_m(Z_{s_m})\Big] = 0 \quad \text{und}$$

$$\mathbb{E}\Big[\big(Z_t^2 - Z_s^2 - \int_s^t Z_s(1-Z_s)\,ds\big)g_0(Z_{s_0})\cdots g_m(Z_{s_m})\Big] = 0$$

für $0 \le s_0 < s_1 < \cdots < s_m \le s < t$ und $g_0, g_1, \ldots, g_m \in C([0,1])$.

2. Schritt (Straffheit auf dem Pfadraum):

Wir zeigen für $T > 0$, $\varepsilon > 0$

$$\lim_{\delta \searrow 0} \limsup_{N\to\infty} \mathbb{P}\big(\exists\, 0 \le s < t \le T : t \le s+\delta \text{ und } |Z_t^{(N)} - Z_s^{(N)}| \ge \varepsilon\big) = 0. \tag{1.30}$$

Sei $s = c_N i < c_N j = t$.

$$\begin{aligned}
\mathbb{E}\Big[\big(Z_t^{(N)} - Z_s^{(N)}\big)^4\Big] &= \mathbb{E}\Big[\big(\sum_{k=i}^{j-1} Z_{c_N(k+1)}^{(N)} - Z_{c_N k}^{(N)}\big)^4\Big] \\
&= \sum_{k=i}^{j-1} \mathbb{E}\Big[\big(Z_{c_N(k+1)}^{(N)} - Z_{c_N k}^{(N)}\big)^4\Big] + 6 \sum_{i\le k_1<k_2<j} \mathbb{E}\Big[\big(Z_{c_N(k_1+1)}^{(N)} - Z_{c_N k_1}^{(N)}\big)^2\big(Z_{c_N(k_2+1)}^{(N)} - Z_{c_N k_2}^{(N)}\big)^2\Big] \\
&\quad + 4 \sum_{i\le k_1<k_2<j} \mathbb{E}\Big[\big(Z_{c_N(k_1+1)}^{(N)} - Z_{c_N k_1}^{(N)}\big)\big(Z_{c_N(k_2+1)}^{(N)} - Z_{c_N k_2}^{(N)}\big)^3\Big] \\
&\quad + 6 \sum_{i\le k_1<k_2<k_3<j} \mathbb{E}\Big[\big(Z_{c_N(k_1+1)}^{(N)} - Z_{c_N k_1}^{(N)}\big)\big(Z_{c_N(k_2+1)}^{(N)} - Z_{c_N k_2}^{(N)}\big)\big(Z_{c_N(k_3+1)}^{(N)} - Z_{c_N k_3}^{(N)}\big)^2\Big]
\end{aligned}$$

(beachte: Da $Z^{(N)}$ ein Martingal ist, gilt für $k_1 \le k_2 \le k_3 < k_4$ stets

$$\mathbb{E}\Big[\prod_{i=1}^{4}\big(Z^{(N)}_{c_N(k_i+1)} - Z^{(N)}_{c_N k_i}\big)\Big] = 0$$

— berechne zunächst den Erwartungswert bedingt auf $\mathscr{F}^{(N)}_{c_N k_4}$)

Die beiden ersten Terme auf der rechten Seite sind beschränkt durch

$$\sum_{k=i}^{j-1} c_N r_{N,4} + 6 \sum_{i\le k_1<k_2<j} c_N^2 \le (j-i)c_N r_{N,4} + (c_N(j-i))^2 \le (t-s)(r_{N,4} + (t-s)),$$

der dritten Term

$$= 4 \sum_{i+1\le k_2<j} \mathbb{E}\Big[\big(Z^{(N)}_{c_N k_2} - Z^{(N)}_{c_N i}\big)\big(Z^{(N)}_{c_N(k_2+1)} - Z^{(N)}_{c_N k_2}\big)^3\Big] \le 4(j-i)c_N r_{N,3} = 4r_{N,3}(t-s)$$

(beachte

$\mathbb{E}\Big[\big(Z^{(N)}_{c_N k_2} - Z^{(N)}_{c_N i}\big)\big(Z^{(N)}_{c_N(k_2+1)} - Z^{(N)}_{c_N k_2}\big)^3\Big] = \mathbb{E}\Big[\big(Z^{(N)}_{c_N k_2} - Z^{(N)}_{c_N i}\big)\mathbb{E}\Big[\big(Z^{(N)}_{c_N(k_2+1)} - Z^{(N)}_{c_N k_2}\big)^3 \,\Big|\, \mathscr{F}^{(N)}_{c_N k_2}\Big]\Big] \le$
$r_{N,3}$),

der vierte Term

$$= 3 \sum_{k_3=i+2}^{j-1} \mathbb{E}\Big[\big(Z^{(N)}_{c_N k_3} - Z^{(N)}_{c_N i}\big)^2\big(Z^{(N)}_{c_N(k_3+1)} - Z^{(N)}_{c_N k_3}\big)^2\Big] \le 3(j-i)c_N^2 = 3c_N(t-s)$$

Insgesamt:

$$\mathbb{E}\Big[\big(Z^{(N)}_t - Z^{(N)}_s\big)^4\Big] \le (t-s)(r_{N,4} + (t-s) + 4r_{N,3} + 3c_N)$$

gleichmäßig in $s < t \le T$, somit:

$$\begin{aligned}\sup_{s\le T}\mathbb{P}\Big(\sup_{s\le t\le s+\delta}\big|Z^{(N)}_t - Z^{(N)}_s\big| \ge \epsilon/4\Big) &\le \sup_{s\le T}\Big(\frac{4}{\epsilon}\Big)^4 \mathbb{E}\Big[\sup_{s\le t\le s+\delta}\big|Z^{(N)}_t - Z^{(N)}_s\big|^4\Big] \\ &\le \frac{C}{\epsilon^4}\sup_{s\le T}\mathbb{E}\Big[\big|Z^{(N)}_{s+\delta} - Z^{(N)}_s\big|^4\Big] \le \frac{C}{\epsilon^4}\delta(r_{N,4} + \delta + 4r_{N,3} + 3c_N)\end{aligned}$$

(mit Markov-Ungleichung und Doobs L^4-Ungleichung).

Wir zerlegen nun $[0,T]$ in Intervalle der Länge δ:

$$\begin{aligned}&\mathbb{P}\big(\exists\, 0 \le s < t \le T : t \le s+\delta, |Z^{(N)}_t - Z^{(N)}_s| \ge \epsilon\big) \\ &\le \sum_{i=0}^{\lceil T/\delta\rceil+1}\mathbb{P}\Big(\sup_{i\delta\le t\le(i+1)\delta}\big|Z^{(N)}_t - Z^{(N)}_{i\delta}\big| \ge \epsilon/4\Big) + \sum_{i=0}^{\lceil T/\delta\rceil+1}\mathbb{P}\Big(\big|Z^{(N)}_{(i+1)\delta} - Z^{(N)}_{i\delta}\big| \ge \epsilon/4\Big) \\ &\le \frac{C}{\epsilon^4}\big(\lceil T/\delta\rceil + 1\big)\delta(r_{N,4} + \delta + 4r_{N,3} + 3c_N)\end{aligned}$$

und für jedes $\epsilon > 0, \delta > 0$ ist der $\limsup_{N\to\infty}$ der rechten Seite $= \frac{C}{\epsilon^4}\delta(T+\delta)$, mit $\delta \downarrow 0$ konvergiert dies gegen 0, d.h. (1.30) gilt.

(1.30) impliziert Straffheit auf $C([0,T],[0,1])$, dazu verwenden wir das Kriterium (siehe z.B. P. Billingsley, Convergence of probability measures, Ch. 2, Thm. 8.2)

$$\forall\, T>0, \epsilon>0\ \exists\, \delta>0 \ \text{ mit } \ \sup_N \mathbb{P}\big(w(Z_\cdot^{(N)},\delta,T)\geq\epsilon\big)\leq\epsilon, \tag{1.31}$$

wobei für $g\in C([0,T],[0,1])$ der Stetigkeitsmodul definiert ist durch

$$w(g,\delta,T) := \sup_{u,v\in[0,T],|u-v|<\delta} |g(u)-g(v)|.$$

(Für ein Argument „von Hand“: Wähle δ_n so, dass

$$\sup_N \mathbb{P}\big(w(Z_\cdot^{(N)},\delta_n,T)\geq\varepsilon/2^n\big)\leq\varepsilon/2^n,$$

setze

$$K_\varepsilon := \bigcap_{n\in\mathbb{N}}\{f : w(f,\delta_n,T)<\varepsilon/2^n\}.$$

Die Menge $K_\varepsilon\subset C([0,T],[0,1])$ ist relativkompakt (nach dem Satz von Arzelà-Ascoli) und erfüllt $\inf_N \mathbb{P}(Z^{(N)}\in K_\varepsilon)>1-\varepsilon$.)

(Wenn wir mit $Z^{(N)}$ als $D([0,\infty),[0,1])$-wertiger ZV arbeiten, d.h. die „Stufenprozessversion“ von $Z^{(N)}$ verwenden, müssen wir wörtlich ein Kriterium für Straffheit auf dem Skorohod-Raum $D([0,\infty),[0,1])$ verwenden (siehe z.B. [EK, Thm. 3.7.2]): Straffheit der 1-dim. Randverteilungen (ist trivialerweise erfüllt) und (1.31), wobei für $g\in D([0,\infty),[0,1])$ der Stetigkeitsmodul definiert ist durch

$$w(g,\delta,T) := \inf_{(t_i)}\max_i \sup_{u,v\in[t_{i-1},t_i)} |g(u)-g(v)|$$

(das Infinimum erstreckt sich über alle Zerlegungen $0=t_0<t_1<\cdots<t_{n-1}\leq T<t_n$ $(n\in\mathbb{N})$ mit $t_i-t_{i-1}>\delta$; große Sprünge sind erlaubt, solange sie sich nicht häufen). Aus (1.30) folgt auch dies und zudem auch, dass jeder Häufungspunkt auf $C([0,\infty),[0,1])$ konzentriert ist, vgl. [EK].)

3. Schritt (Das Martingalproblem ist eindeutig (via Dualität mit dem Klassenzählprozess des Kingman-Koaleszenten)):

Wir zeigen: Sei $Z=(Z_t)_{t\geq0}$ stetiges Martingal mit Werten in $[0,1]$, $Z_0=z$ und $\langle Z\rangle_t=\int_0^t Z_s(1-Z_s)\,ds$ [d.h. $Z_t^2-Z_0^2-\int_0^t Z_s(1-Z_s)\,ds$ ist Martingal], so ist Z die Wright-Fisher-Diffusion.

Mit Itō-Formel ist für $n\in\mathbb{N}$

$$Z_t^n = Z_0^n + \int_0^t nZ_s^{n-1}\,dZ_s + \frac{1}{2}\int_0^t n(n-1)Z_s^{n-2}d\langle Z\rangle_s,$$

also ist

$$Z_t^n - Z_0^n - \frac{1}{2}\int_0^t n(n-1)\big[Z_s^{n-1}-Z_s^n\big]\,ds = \int_0^t nZ_s^{n-1}\,dZ_s$$

ein Martingal und somit

$$\begin{aligned} f(n,z,t) := \mathbb{E}_z[Z_t^n] &= z^n + \binom{n}{2}\int_0^t \mathbb{E}_z[Z_t^{n-1}]-\mathbb{E}_z[Z_t^n]\,ds \\ &= z^n + \binom{n}{2}\int_0^t f(n-1,z,t)-f(n,z,t)\,ds \end{aligned}$$

(und $f(1,z,t) = z$ für alle $t \geq 0$, $z \in [0,1]$), d.h. die Familie von Funktionen $f(n,z,t)$ löst ein (lineares) System von Differentialgleichungen

$$\frac{\partial}{\partial t} f(n,z,t) = \binom{n}{2}\big(f(n-1,z,t) - f(n,z,t)\big),\ t \geq 0, n \in \mathbb{N} \quad \text{und} \quad f(n,z,0) = z^n. \tag{1.32}$$

Dieses System ist eindeutig lösbar (z.B. rekursiv in n).

Sei $(N_t)_{t\geq 0}$ ($N_t = |R_t^{(n)}|$, wenn Startpunkt n Klassen hat) der Blockzählprozess (oder Klassenzählprozess) des Kingman-Koaleszenten (Def. 1.7), wir lesen dort ab, dass $(N_t)_{t\geq 0}$ eine Markovkette auf $\mathbb{N}$ mit Sprungratenmatrix

$$q_{m,m-1} = \binom{m}{2}, \quad q_{m,m} = -\binom{m}{2}, \quad q_{m,\ell} = 0 \text{ falls } \ell \notin \{m-1, m\}$$

ist. Setze

$$g(n,z,t) := \mathbb{E}_n[z_t^N].$$

Gemäß Kolmogorovs Rückwärtsgleichung (vgl. Blatt 1, Aufg. 1.3) löst $g(n,z,t)$ ebenfalls (1.32). Da dieses System eindeutig lösbar ist, folgt

$$\mathbb{E}_z[Z_t^n] = \mathbb{E}_n[z^{N_t}], \quad t \geq 0, z \in [0,1], n \in \mathbb{N}, \tag{1.33}$$

d.h. alle Momente von Z_t, und damit (da Z_t beschränkt ist) auch die Verteilung von Z_t ist festgelegt.

Für die Lösung eines Martingalproblems impliziert Eindeutigkeit der eindimensionalen Verteilungen die Eindeutigkeit der Verteilung des gesamten Prozesses (vgl. z.B. [EK, Thm. 4.4.2] oder [St3, Prop. 4.3]).

Zusammensetzen: Wegen 2. gibt es Häufungspunkte (Straffheit impliziert Relativkompaktheit nach dem Satz von Prohorov), wegen 1. löst jeder Häufungspunkt das MP, wegen 3. sind alle Häufungspunkte gleich $\Rightarrow$ Konvergenz □

Schritt 3 des Beweises von Satz 1.26 ist eine Beispiel-Anwendung eines allgemeineren Prinzips:

Bemerkung/Erinnerung (Dualität, allgemeiner)**.** In [St3, Def. 4.6; in Kap. 4.1] hatten wir folgendes betrachtet: Seien $X^{(x)} = (X_t^{(x)})_{t\geq 0}$, $x \in E$ und $Y^{(y)} = (Y_t^{(y)})_{t\geq 0}$, $y \in E'$ Familien von stochastischen Prozessen mit Werten in E bzw. in E' (E, E' seien polnische Räume, sagen wir), es gelte $X_0^{(x)} = x$, $Y_0^{(y)} = y$ f.s. X und Y heißen *dual* mit *Dualitätsfunktion* $H : E \times E' \to \mathbb{C}$, wenn gilt

$$\mathbb{E}\big[H(X_t^{(x)}, y)\big] = \mathbb{E}\big[H(x, Y_t^{(y)})\big] \quad \forall\, t \geq 0,\ x \in E,\ y \in E'.$$

(Wir nehmen an, dass H geeignete Messbarkeits- und Beschränktheits/-Wachstumsannahmen erfüllt, so dass die betrachteten Erwartungswerte existieren.)

Wir hatten dort (insbesondere) gesehen [St3, Satz 4.7; in Kap. 4.1]: Es gebe eine Familie $Y^{(y)}$, $y \in E'$ von Markovprozessen mit Werten in E' und $Y_0^{(y)} = y$ (wie oben) und $H : [0,1] \times E' \to \mathbb{C}$ m.b. so dass $\mathbb{E}\big[|H(x, Y_t^{(y)})|\big] < \infty$ für $x \in \mathbb{R}, y \in E'$, $t \geq 0$ und die Funktionenmenge

$$\big\{H(\cdot, y), y \in E'\big\} \quad \text{sei trennend für } \mathcal{M}_1([0,1]),$$

(d.h. $\int_{[0,1]} H(x,y)\,\mu(dx) = \int_{[0,1]} H(x,y)\,\nu(dx)$ für alle $y \Rightarrow \mu = \nu$). Weiter gebe es für jedes $x \in [0,1]$ eine Lösung $X^{(x)}$ des Martingalproblems mit Generator L und Startwert $X_0^{(x)} = x$, die

$$\mathbb{E}\big[H(X_t^{(x)}, y)\big] = \mathbb{E}\big[H(x, Y_t^{(y)})\big] \quad \forall\, t \geq 0,\ y \in E' \tag{1.34}$$

erfüllt. Dann ist das Martingalproblems zum Generator L wohlgestellt, d.h. die Lösung ist eindeutig.

In Schritt 3 des Beweises von Satz 1.26 oben haben wir diesen Zusammenhang zwischen der (neutralen 2 Typ-)Wright-Fisher-Diffusion und dem (Klassenzählprozess des) Kingman-Koaleszent(en), via ein explizites Argument mit dem Momentenproblem, mit Dualitätsfunktion $H(z,n) = z^n$ ausgenutzt. (Man spricht daher auch von einem „Momentendual.“)

Man könnte die Eindeutigkeit des Martingalproblems der 2 Typ-Wright-Fisher-Diffusion auch auf andere Art beweisen. Es ist bekannt, dass die Lösung des Martingalproblems (aus Bem. 1.25, 1) und die Lösung der SDgl (aus Bem. 1.25, 3) äquivalent sind (vgl. z.B. [St3, Satz 3.22]), das Yamada-Watanabe-Kriterium (vgl. z.B. [St3, Satz 3.17]) garantiert Eindeutigkeit der SDgl. aus Bem. 1.25, 3, daher ist auch das Martingalproblem eindeutig lösbar.

Seine „volle“ Macht zeigt der Dualitätsansatz erst in der mehrdimensionalen Situation, beispielsweise wenn mehr als 2 Typen betrachtet werden oder die Population zusätzlich noch räumlich strukturiert ist. Zudem gestattet die Dualität recht explizite Berechnungen der Momente, die wir im nächsten Unterkapitel anschauen werden.

Ein alternativer Zugang zur Wright-Fisher-Diffusion und Dualität mit dem Kingman-Koaleszenten*

Ein alternativer Zugang zur Konvergenz von $(Z^{(N)})_{t\geq 0}$ — der zumindest die Konvergenz der endlich-dimensionalen Verteilungen beweist und der mit weniger „technischem Apparat“ als der Beweis von Satz 1.26 auskommt — liegt darin, die Momente des Grenzprozesses via Stichproben-Interpretation und die Konvergenz der Genealogien (aus Satz 1.8) zu berechnen.

Beobachtung 1.28. Sei $x \in (0,1)$, im Modell mit Populationsgröße N vergeben wir an die N Individuen der Startgeneration 0 rein zufällig $\lfloor Nx \rfloor$-mal den Typ 1 und $(N - \lfloor Nx \rfloor)$-mal den Typ 0, in der Notation von Beob. 1.3:

$$\big((t_{0,1}^{(N)}, \ldots, t_{0,N}^{(N)})\big) \quad \text{ist eine zufällige Permutation von } \big(\underbrace{1,1,\ldots,1}_{\lfloor Nx \rfloor}, \underbrace{0,\ldots,0}_{N-\lfloor Nx \rfloor}\big).$$

Sei $t > 0$ und $n \in \mathbb{N}$ $(n < N)$, wir ziehen zum Zeitpunkt $\lfloor t/c_N \rfloor$ eine Stichprobe von n

Individuen aus der Population. Die Wahrscheinlichkeit, dass alle n den Typ 1 haben ist,

$$\begin{aligned}
&\mathbb{E}_{\frac{\lfloor Nx\rfloor}{N}}\big[Z^{(N)}_{\lfloor t/c_N\rfloor}(Z^{(N)}_{\lfloor t/c_N\rfloor}-\tfrac{1}{N})\cdots(Z^{(N)}_{\lfloor t/c_N\rfloor}-\tfrac{n-1}{N})\big]\\
&=\mathbb{P}_{\frac{\lfloor Nx\rfloor}{N}}\big(t^{(N)}_{\lfloor t/c_N\rfloor,J_1}=1,\dots,t^{(N)}_{\lfloor t/c_N\rfloor,J_n}=1\big)\\
&=\mathbb{P}_{\frac{\lfloor Nx\rfloor}{N}}\big(t^{(N)}_{0,A^{(N)}_{\lfloor t/c_N\rfloor,J_1}[\lfloor t/c_N\rfloor]}=1,\dots,t^{(N)}_{0,A^{(N)}_{\lfloor t/c_N\rfloor,J_n}[\lfloor t/c_N\rfloor]}=1\big)\\
&=\mathbb{E}\Big[\prod_{i=0}^{|R^{(N,n)}_{\lfloor t/c_N\rfloor}|-1}\frac{\lfloor Nx\rfloor-i}{N-i}\Big]
\end{aligned}$$

mit $J_1,\dots,J_n$ eine rein zufällige Stichprobe vom Umfang n aus $1,\dots,N$ (verwende Beob. 1.3 zum Typenzählprozess im N-ten Modell).

Mit Satz 1.8 (Konvergenz von $R^{(N,n)}_{\lfloor\cdot/c_N\rfloor}$ gegen den Kingman-Koaleszenten) folgt

$$\lim_{N\to\infty}\mathbb{E}_{\frac{\lfloor Nx\rfloor}{N}}\big[(Z^{(N)}_{\lfloor t/c_N\rfloor})^n\big]=\mathbb{E}\big[x^{|R^{(n)}_t|}\big]$$

Für $x\in[0,1]$, $t>0$ wird daher ein W'maß $\kappa_t(x,dy)$ eindeutig (das Momentenproblem auf dem kompakten Intervall $[0,1]$ ist eindeutig lösbar, vgl. z.B. [Kl, Kor. 15.32] oder [St2, Kor. 3.31]) durch die Forderung

$$\int_{[0,1]} y^n\,\kappa_t(x,dy)=\mathbb{E}\big[x^{|R^{(n)}_t|}\big]\quad\text{für } n\in\mathbb{N} \tag{1.35}$$

festgelegt und es gilt $\mathscr{L}(Z^{(N)}_{\lfloor t/c_N\rfloor}\,|\,Z^{(N)}_0=\lfloor Nx\rfloor)\to\kappa_t(x,\cdot)$. (Wir wissen bereits aus Satz 1.26, dass $\kappa_t(x,dy)=\mathbb{P}_x(Z_t\in dy)$, d.h. κ_t ist der Übergangskern der Wright-Fisher-Diffusion.)

Sei $(N_t)_{t\ge0}$ der Blockzählprozess (oder Klassenzählprozess) des Kingman-Koaleszenten, d.h. die zeitkontinuierliche Markovkette auf $\mathbb{N}$ mit Sprungraten

$$q_{m,m-1}=\binom{m}{2},\quad q_{m,m}=-\binom{m}{2},\quad q_{m,\ell}=0\text{ falls }\ell\notin\{m-1,m\}$$

(in der Notation von Def. 1.7 ist $N_t=|R^{(n)}_t|$, wenn $N_0=n$).

Für $1\le m<n$ ist mit Lemma 1.14 (und S_j u.a., $S_j\sim\mathrm{Exp}(\binom{j}{2})$)

$$\begin{aligned}
\mathbb{P}_n(N_t\le m)&=\mathbb{P}(S_n+S_{n-1}+\cdots+S_{m+1}\le t)\\
&=\int_0^t\sum_{j=m+1}^n\binom{j}{2}\exp\big(-\tbinom{j}{2}s\big)\prod_{\substack{k=m+1,\\ k\ne j}}^n\frac{\binom{k}{2}}{\binom{k}{2}-\binom{j}{2}}\,ds\\
&=\sum_{j=m+1}^n\Big(1-\exp\big(-\tbinom{j}{2}s\big)\Big)\prod_{\substack{k=m+1,\\ k\ne j}}^n\frac{\binom{k}{2}}{\binom{k}{2}-\binom{j}{2}}
\end{aligned}$$

(und $\mathbb{P}_n(N_t \le n) = 1$), somit für $1 < m < n$

$$\begin{aligned}
\mathbb{P}_n(N_t = m) &= \mathbb{P}_n(N_t \le m) - \mathbb{P}_n(N_t \le m-1) \\
&= \sum_{j=m+1}^{n} \Big(1 - \exp\big(-\tbinom{j}{2}s\big)\Big)\Big(1 - \frac{\binom{m}{2}}{\binom{m}{2} - \binom{j}{2}}\Big) \prod_{\substack{k=m+1,\\ k\neq j}}^{n} \frac{\binom{k}{2}}{\binom{k}{2} - \binom{j}{2}} \\
&\quad - \Big(1 - \exp\big(-\tbinom{m}{2}s\big) \prod_{\substack{k=m+1,\\ k\neq j}}^{n} \frac{\binom{k}{2}}{\binom{k}{2} - \binom{j}{2}} \\
&= -\sum_{j=m}^{n} \Big(1 - \exp\big(-\tbinom{j}{2}s\big)\Big) \frac{\binom{j}{2}}{\binom{m}{2}} \prod_{\substack{k=m,\\ k\neq j}}^{n} \frac{\binom{k}{2}}{\binom{k}{2} - \binom{j}{2}} \\
&= -\frac{\prod_{\ell=m}^{n}\binom{\ell}{2}}{\binom{m}{2}} \sum_{j=m}^{n} \Big(1 - \exp\big(-\tbinom{j}{2}s\big)\Big) \prod_{\substack{k=m,\\ k\neq j}}^{n} \frac{1}{\binom{k}{2} - \binom{j}{2}} \\
&= \frac{\prod_{\ell=m}^{n}\binom{\ell}{2}}{\binom{m}{2}} \sum_{j=m}^{n} \exp\big(-\tbinom{j}{2}s\big) \prod_{\substack{k=m,\\ k\neq j}}^{n} \frac{1}{\binom{k}{2} - \binom{j}{2}}
\end{aligned}$$

(in der letzten Zeile nutzen wir aus, dass $\sum_{j=m}^{n} \prod_{k=m,k\neq j}^{n} \frac{1}{\binom{k}{2}-\binom{j}{2}} = 0$ gilt, vgl. (1.14) in Lemma 1.14).

Weiter ist

$$\prod_{\ell=m}^{n} \binom{\ell}{2} = \prod_{\ell=m}^{n} \frac{\ell(\ell-1)}{2} = 2^{m-n} \frac{n!(n-1)!}{(m-1)!(m-2)!}$$

und

$$\begin{aligned}
\prod_{\substack{k=m,\\ k\neq j}}^{n} \Big(\tbinom{k}{2} - \tbinom{j}{2}\Big) &= \prod_{\substack{k=m,\\ k\neq j}}^{n} \frac{k(k-1) - j(j-1)}{2} = \prod_{\substack{k=m,\\ k\neq j}}^{n} \frac{(k+j-1)(k-j)}{2} \\
&= 2^{-(n-m)} \frac{(n+j-1)!}{(2j-1)(m+j-2)!} (-1)^{j-m} (j-m)!(n-j)!.
\end{aligned}$$

Insgesamt folgt

$$\begin{aligned}
\mathbb{P}_n(N_t = m) &= \sum_{j=m}^{n} \exp\big(-\tbinom{j}{2}t\big) \frac{(2j-1)(-1)^{j-m}(m)_{(j-1)\uparrow}(n)_{j\downarrow}}{m!(j-m)!(n)_{j\uparrow}}, \quad 2 \le m \le n, \\
\mathbb{P}_n(N_t = 1) &= 1 - \sum_{j=2}^{n} \exp\big(-\tbinom{j}{2}t\big) \frac{(2j-1)(-1)^{j}(n)_{j\downarrow}}{(n)_{j\uparrow}}
\end{aligned}$$

(mit $(x)_{j\downarrow} = x(x-1)\cdots(x-j+1)$, $(x)_{j\uparrow} = x(x+1)\dots(x+j-1)$ fallende bzw. wachsende Faktorielle).

Mit $n \to \infty$ ergibt sich (vgl. auch Beob. 1.10 und den anschließenden Bericht zum Start in $N_0 = \infty$)

$$\begin{aligned}
\mathbb{P}_\infty(N_t = m) &= \sum_{k=m}^{\infty} \exp\big(-\tbinom{k}{2}t\big) \frac{(2k-1)(-1)^{k-m}(m)_{(k-1)\uparrow}}{m!(k-m)!}, \quad 2 \le m \le n, \\
\mathbb{P}_\infty(N_t = 1) &= 1 - \sum_{k=2}^{\infty} \exp\big(-\tbinom{k}{2}t\big)(2k-1)(-1)^{k}.
\end{aligned}$$

Somit ist

$$\mathbb{P}_z(Z_t = 1) = \lim_{n\to\infty} \mathbb{E}_z[Z_t^n] = \lim_{n\to\infty} \sum_{j=1}^{n} \mathbb{P}_n(N_t = j) z^j = \sum_{j=1}^{\infty} \mathbb{P}_\infty(N_t = j) z^j \tag{1.36}$$

und analog

$$\mathbb{P}_z(Z_t = 0) = \sum_{j=1}^{\infty} \mathbb{P}_\infty(N_t = j)(1-z)^j. \tag{1.37}$$

Um aus (1.35) auch die Dichte im Inneren $(0,1)$ des Einheitsintervalls zu bestimmen, benötigen wir noch einen weiteren Gedanken:

Für $a, b > 0$ sei

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}, \quad y \in (0,1)$$

die Beta-(a,b)-Dichte, für $Y_{a,b} \sim \mathrm{Beta}(a,b)$ ist

$$\mathbb{E}[Y_{a,b}] = a/(a+b) \quad \text{und} \quad \mathrm{Var}[Y_{a,b}] = ab/((a+b)^2(a+b+1)).$$

Für stetiges $f : [0,1] \to \mathbb{R}$ gilt somit

$$\mathbb{E}[f(Y_{a,b})] \to f(y) \quad \text{wenn } a, b \to \infty \text{ mit } \frac{a}{a+b} \to x \in (0,1)$$

(verwende z.B. die Chebychev-Ungleichung um zu sehen, dass $Y_{a,b}$ dann sehr eng um seinen Erwartungswert konzentriert ist).

Für $n_1, n_2 \in \mathbb{N}$ gilt

$$\begin{aligned} &\mathbb{E}_z\Big[\binom{n_1+n_2}{n_1} Z_t^{n_1}(1-Z_t)^{n_2}\Big] \\ &\qquad = \sum_{j=2}^{n_1+n_2} \mathbb{P}_{n_1+n_2}(N_t = j) \sum_{k=1}^{j-1} \binom{j}{k} z^k (1-z)^{j-k} \frac{\binom{n_1-1}{k-1}\binom{n_2-1}{j-k-1}}{\binom{n_1+n_2-1}{j-1}}. \end{aligned} \tag{1.38}$$

Der Erwartungswert auf der linken Seite ist die Wahrscheinlichkeit, in einer zufälligen Stichprobe der Größe $n_1 + n_2$ aus der Population zur Zeit t den Typ 1 n_1-mal und den Typ 0 n_2-mal zu sehen.

Für die Gleichung betrachten wir die folgende alternative Berechnung dieser Wahrscheinlichkeit, die auf die rechte Seite führt:

- Auf $\{N_t^{(n_1+n_2)} = j\}$ $(2 \le j \le n_1 + n_2)$ denken wir uns die j Linien, die die Ahnen der Stichprobe zur Zeit 0 bilden, (zufällig) nummeriert.
- Eine gegebene Teilmenge K_1 der Größe k (mit $1 \le k < j$) dieser j Linien erhält Typ 1, die übrigen Typ 2: dies hat W'keit $z^k(1-z)^{j-k}$

  (und: es gibt $\binom{j}{k}$ verschiedene solche Teilmengen).

- Die W'keit, dass die k Linien aus K_1 genau n_1 Stichproben repräsentieren (und demnach repräsentieren die $j-k$ Linien aus $[j]\setminus K_1$ genau n_2 Stichproben), ist

$$\frac{\big|\{(m_1,\dots,m_j)\in\mathbb{N}^j : \sum_{i\in K_1} m_i = n_1, m_1+\cdots+m_j = n_1+n_2\}\big|}{|\{(m_1,\dots,m_j)\in\mathbb{N}^j : m_1+\cdots+m_j = n_1+n_2\}|} = \frac{\binom{n_1-1}{k-1}\binom{n_2-1}{j-k-1}}{\binom{n_1+n_2-1}{j-1}},$$

  vergleiche Kor. 1.12.

Nehmen wir an, $\mathbb{P}_x(Z_t\in\cdot)$ besitzt Dichte $f_t(x,y)$ im Inneren $(0,1)$.

Sei nun $y\in(0,1)$, wir setzen $n_1 := \lfloor ny\rfloor$, $n_2 := n-n_1$, dann ist (setze $a=n_1+1$, $b=n_2+1$ in (1.38) oben)

$$\begin{aligned}
&\int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} z^{a-1}(1-z)^{b-1} f_t(x,z)\,dz = \mathbb{E}_x\Big[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} Z_t^{a-1}(1-Z_t)^{b-1}\Big]\\
&= (n+1)\mathbb{E}_x\Big[\binom{n_1+n_2}{n_1} Z_t^{n_1}(1-Z_t)^{n_2}\Big]\\
&= \sum_{j=2}^{n_1+n_2} \mathbb{P}_{n_1+n_2}(N_t=j) \sum_{k=1}^{j-1}\binom{j}{k} x^k(1-x)^{j-k}(n+1)\frac{\binom{n_1-1}{k-1}\binom{n_2-1}{j-k-1}}{\binom{n_1+n_2-1}{j-1}} \qquad (1.39)
\end{aligned}$$

[beachte $\Gamma(a+b)=\Gamma(n_1+n_2+2)=(n+1)!$]

Es ist

$$\begin{aligned}
(n+1)\frac{\binom{n_1-1}{k-1}\binom{n_2-1}{j-k-1}}{\binom{n_1+n_2}{j-1}} &= \frac{(j-1)!}{(k-1)!(j-k-1)!}(n+1)\frac{(\lfloor ny\rfloor-1)_{(k-1)\downarrow}(n-\lfloor ny\rfloor-1)_{(j-k-1)\downarrow}}{(n-1)_{(j-1)\downarrow}}\\
&\xrightarrow[n\to\infty]{} \frac{(j-1)!}{(k-1)!(j-k-1)!} y^{k-1}(1-y)^{j-k-1} \qquad (1.40)
\end{aligned}$$

Demnach finden wir insgesamt

Beobachtung 1.29. $\kappa_t(x,\cdot)$ besitzt auf $(0,1)$ die Dichte

$$f_t(x,y) = \sum_{j=2}^{\infty}\mathbb{P}(N_t=j)\sum_{k=1}^{j-1}\binom{j}{k}x^k(1-x)^{j-k}\frac{(j-1)!}{(k-1)!(j-k-1)!}y^{k-1}(1-y)^{j-k-1},\quad x,y\in(0,1),\ t>0$$

[also eine Mischung aus Beta$(k,j-k)$-Dichten] und

$$\kappa(x,dy) = \mathbb{P}_x(Z_t=1)\delta_1(dy) + \mathbb{P}_x(Z_t=0)\delta_0(dy) + f_t(x,y)\mathbf{1}_{(0,1)}(y)dy. \qquad (1.41)$$

(Die Annahme, dass $\mathbb{P}_x(Z_t\in\cdot)$ tatsächlich eine stetige Dichte im Inneren $(0,1)$ besitzt, können wir nachträglich durch Berechnung der Laplace-Transformierten rechtfertigen, vgl. [T84].)

Beweis von Satz 1.22*

Wir gehen von Typenhäufigkeitsprozess (X_t) zum Typenanteilsprozess $Y_t^{(N)} := \frac{1}{N} X_t$ (mit Werten in $\{0, 1/N, 2/N, \dots, 1\} \subset [0,1]$) über und notieren $\mathbb{P}_p(\cdot)$, $\mathbb{E}_p[\cdot]$ für die Situation, dass $Y_0^{(N)} = p$, $p \in [0,1]$, d.h. dass $X_0 = pN$.

Sei $H(p) = -p\log(p) - (1-p)\log(1-p)$ für $p \in [0,1]$ wie oben, nach Voraussetzung gilt $\sigma_N^2 := \mathrm{Var}[\nu_1^{(N)}] \to \sigma^2 \in (0,\infty)$, $K_4 := \sup_N \mathbb{E}[(\nu_1^{(N)})^4] < \infty$.

(Beachte $c_N = \frac{\sigma_N^2}{N-1} \sim \sigma^2/N$.)

Wir können $\delta > 0$, $K_1, K_2 < \infty$ finden, so dass für alle $N \geq N_0$ gilt

$$\inf_{p \in \{\frac{1}{N}, \frac{2}{N}, \dots, \frac{N-1}{N}\}} \Big\{ \mathbb{P}_p\Big(|Y_1^{(N)} - p| \geq \delta \sqrt{p(1-p)/N} \Big) \Big\} \geq \frac{1}{K_1} > 0 \tag{1.42}$$

und

$$\max_{p \in \{\frac{1}{N}, \frac{2}{N}, \dots, \frac{N-1}{N}\}} \frac{1}{p \wedge (1-p)} \Big| \mathbb{E}_p\Big[Y_1^{(N)} - p \,\Big|\, |Y_1^{(N)} - p| \geq \delta \sqrt{p(1-p)/N} \Big] \Big| \leq K_2. \tag{1.43}$$

Für (1.42) beachte

$$\mathbb{P}_p\Big(|Y_1^{(N)} - p| \geq \delta \sqrt{p(1-p)/N} \Big) = \mathbb{P}_p\Big((Y_1^{(N)} - p)^2 \geq \delta^2 \tfrac{N-1}{N} \sigma_N^2 \mathbb{E}_p\big[(Y_1^{(N)} - p)^2 \big] \Big)$$
$$\geq \Big(1 - \delta^2 \tfrac{N-1}{N} \sigma_N^2 \Big)^2 \frac{\big(\mathbb{E}_p\big[(Y_1^{(N)} - p)^2 \big] \big)^2}{\mathbb{E}_p\big[(Y_1^{(N)} - p)^4 \big]} \geq \Big(1 - \delta^2 \tfrac{N-1}{N} \sigma_N^2 \Big)^2 \frac{\sigma_N^4 N^{-2} p^2 (1-p)^2}{C N^{-2} p^2 (1-p)^2}$$

gemäß Lemma 1.27 (siehe (1.24)).

Für (1.43): Sei $p = m/N$, es ist

$$\Big| \mathbb{E}_p\big[(Y_1^{(N)} - p) \mathbf{1}_{\{ |Y_1^{(N)} - p| < \delta \sqrt{p(1-p)/N} \}} \big] \Big| \leq \frac{\mathbb{E}_p\big[(Y_1^{(N)} - p)^2 \big]}{\delta \sqrt{p(1-p)/N}} = \frac{\frac{\sigma_N^2}{N-1} p(1-p)}{\delta \sqrt{p(1-p)/N}}$$
$$= \frac{\sigma_N^2}{\delta} \Big(\frac{N}{(N-1)^2} p(1-p) \Big)^{1/2} \leq \frac{\sigma_N^2}{\delta} \frac{\sqrt{m} \wedge \sqrt{N-m}}{N-1} \leq \frac{2\sigma_N^2}{\delta} \big(p \wedge (1-p) \big).$$

Folglich (beachte $\mathbb{E}_p[Y_1^{(N)} - p] = 0$ und (1.42))

$$\Big| \mathbb{E}_p\Big[Y_1^{(N)} - p \,\Big|\, |Y_1^{(N)} - p| \geq \delta \sqrt{p(1-p)/N} \Big] \Big| = \frac{\Big| \mathbb{E}_p\big[(Y_1^{(N)} - p) \mathbf{1}_{\{ |Y_1^{(N)} - p| \geq \delta \sqrt{p(1-p)/N} \}} \big] \Big|}{\mathbb{P}_p\Big(|Y_1^{(N)} - p| \geq \delta \sqrt{p(1-p)/N} \Big)}$$
$$\leq K_1 \Big| \mathbb{E}_p\big[(Y_1^{(N)} - p) \mathbf{1}_{\{ |Y_1^{(N)} - p| < \delta \sqrt{p(1-p)/N} \}} \big] \Big| \leq K_1 \frac{2\sigma_N^2}{\delta} \big(p \wedge (1-p) \big).$$

Wir zeigen zunächst, dass es $\alpha < \infty$ gibt mit

$$\mathbb{E}_{pN}[T_{\mathrm{fix}}] \leq \alpha \frac{N-1}{\sigma_N^2} H(p) \quad \text{für alle } p \in \{0, \tfrac{1}{N}, \tfrac{2}{N}, \dots, 1\} \quad \text{gleichmäßig in } N. \tag{1.44}$$

Für $p \in \{0,1\}$ ist (1.44) offensichtlich erfüllt.

Gemäß Taylor-Formel (Entwicklung von H um p bis zur 2. Ordnung) gilt

$$H(q) = (q-p)H'(p) + (q-p)^2 \int_0^1 (1-t)H''\big(p+t(q-p)\big)\,dt$$

(für $0 < p < 1$, $0 \le q \le 1$).

Es ist

$$\sum_{q\in\{0,\frac{1}{N},\dots,1\}} \mathbb{P}_p(Y_1^{(N)} = q)\big[H(q) - H(p)\big] = H'(p)\underbrace{\mathbb{E}_p[Y_1^{(N)} - p]}_{=0} + R_N(p)$$

mit

$$\begin{aligned} R_N(p) &= -\mathbb{E}_p\Big[(Y_1^{(N)} - p)^2 \int_0^1 \frac{1-t}{(p+t(Y_1^{(N)}-p))(1-p-t(Y_1^{(N)}-p))}\,dt\Big] \\ &= -\mathbb{E}_p\Big[(Y_1^{(N)} - p)^2 \int_0^1 f_{p,t}(Y_1^{(N)})\,dt\Big] = -\int_0^t \mathbb{E}_p\Big[(Y_1^{(N)} - p)^2 f_{p,t}(Y_1^{(N)})\Big]\,dt \end{aligned}$$

[es ist $H''(p) = -1/(p(1-p))$ $(\le 0$, H ist konkav)] mit

$$f_{p,t}(y) = \frac{1-t}{(p+t(y-p))(1-p-t(y-p))}, \quad y \in [0,1].$$

Für jedes $p,t \in (0,1)$ ist die Funktion $f_{p,t}$ nicht-negativ und konvex, also

$$\begin{aligned} R_N(p) &\le -\int_0^{1/(2K_2)} \frac{\mathbb{E}_p\Big[(Y_1^{(N)} - p)^2 f_{p,t}(Y_1^{(N)}) \,\Big|\, |Y_1^{(N)} - p| \ge \delta\sqrt{p(1-p)/N}\Big]}{\mathbb{P}_p\Big(|Y_1^{(N)} - p| \ge \delta\sqrt{p(1-p)/N}\Big)}\,dt \\ &\le -K_1 \int_0^{1/(2K_2)} \delta^2 \frac{p(1-p)}{N} \mathbb{E}_p\Big[f_{p,t}(Y_1^{(N)}) \,\Big|\, |Y_1^{(N)} - p| \ge \delta\sqrt{p(1-p)/N}\Big]\,dt. \end{aligned}$$

Mit der Jensenschen Ungleichung ist gleichmäßig in N und $p \in (0,1) \cap \frac{1}{N}\mathbb{Z}$ (für $0 \le t \le K_2/2$)

$$\begin{aligned} &\mathbb{E}_p\Big[f_{p,t}(Y_1^{(N)}) \,\Big|\, |Y_1^{(N)} - p| \ge \delta\sqrt{p(1-p)/N}\Big] \\ &\ge f_{p,t}\Big(\mathbb{E}_p\Big[Y_1^{(N)} \,\Big|\, |Y_1^{(N)} - p| \ge \delta\sqrt{p(1-p)/N}\Big]\Big) \\ &\ge \frac{1-t}{(p+tK_2p)(1-p+tK_2(1-p))} = \frac{1}{p(1-p)}\frac{1-t}{(1+tK_2)^2} \ge \frac{K_3}{p(1-p)} \end{aligned}$$

für ein $K_3 > 0$ (wir verwenden in der zweiten Ungleichung (1.43)) und insgesamt

$$R_N(p) \le -\frac{K_1K_3\delta^2}{2K_2}\frac{1}{N}.$$

Um die Asymptotik von $\mathbb{E}_p[T_{\text{fix}}]$ wie in (1.17) angegeben präzise zu fassen, betrachten wir die Taylor-Entwicklung von H bis zur vierten Ordnung.

Sei $\varepsilon > 0$, $p \in (\varepsilon, 1-\varepsilon) \cap \frac{1}{N}\mathbb{Z}$:

$$\begin{aligned}
&\sum_{q\in\{0,\frac{1}{N},\dots,1\}} \mathbb{P}_p(Y_1^{(N)} = q)\big[H(q) - H(p)\big] \\
&= \sum_{q\in\{0,\frac{1}{N},\dots,1\}} \mathbb{P}_p(Y_1^{(N)} = q)\Big[(q-p)H'(p) + \frac{1}{2}(q-p)^2 H''(p) + \frac{1}{6}(q-p)^3 H^{(3)}(p) \\
&\qquad\qquad + \frac{1}{24}(q-p)^4(1-\vartheta_{p,q})^3 H^{(4)}\big(p + \vartheta_{p,q}(q-p)\big)\Big] \\
&= -\frac{1}{2}\frac{\sigma_N^2}{N-1} + \widetilde{R}_N(p)
\end{aligned}$$

mit einem $\vartheta_{p,q}$, das $p \wedge q < \vartheta_{p,q} < p \vee q$ erfüllt, und

$$\widetilde{R}_N(p) = \frac{1}{6}H^{(3)}(p)\mathbb{E}_p\big[(Y_1^{(N)} - p)^3\big] \tag{1.45}$$

$$+ \frac{1}{24}\mathbb{E}_p\Big[(Y_1^{(N)} - p)^4(1-\vartheta_{Y_1^{(N)},q})^3 H^{(4)}\big(p + \vartheta_{Y_1^{(N)},q}(Y_1^{(N)} - p)\big)\Big]. \tag{1.46}$$

Für $p \in (\varepsilon, 1-\varepsilon)$, $q \in (0,1)$, $\vartheta \in (0,1)$ ist

$$\big|H^{(3)}(p)\big| = \frac{|1-2p|}{p^2(1-p)^2} \le \frac{1}{p^2(1-p)^2} \le \frac{1}{\varepsilon^2(1-\varepsilon)^2}$$

und

$$\big|(1-\vartheta)^3 H^{(4)}\big(p+\vartheta(q-p)\big)\big| \le \frac{2(1-\vartheta)^3}{\big(p+\vartheta(q-p)\big)^3\big(1-p-\vartheta(q-p)\big)^3} \le \frac{2}{p^3(1-p)^3} \le \frac{2}{\varepsilon^3(1-\varepsilon)^3}.$$

Mit Lemma 1.27 (siehe (1.24)) folgt

$$\max_{p\in\{0,\frac{1}{N},\dots,1\}\cap(\varepsilon,1-\varepsilon)} |R_N(p)| = O\Big(\frac{1}{N^2}\Big). \tag{1.47}$$

Zerlegung nach dem ersten Schritt zusammen mit obigem zeigt, dass

$$f_{N,\varepsilon}(p) := 2\frac{N-1}{\sigma_N^2}H(p) - \mathbb{E}_p[T_\varepsilon]$$

erfüllt

$$\sum_{q\in\{0,\frac{1}{N},\dots,1\}} \mathbb{P}_p(Y_1^{(N)} = q)\big[f_{N,\varepsilon}(q) - f_{N,\varepsilon}(p)\big] = 2\frac{N-1}{\sigma_N^2}\widetilde{R}_N(p), \quad p \in \{0,\tfrac{1}{N},\dots,1\} \cap (\varepsilon, 1-\varepsilon)$$

und

$$0 \le f_{N,\varepsilon}(p) \le 2\frac{N-1}{\sigma_N^2}H(\varepsilon), \quad p \in \{0,\tfrac{1}{N},\dots,1\} \cap \big([0,\varepsilon] \cap [1-\varepsilon,1]\big),$$

somit ist

$$\begin{aligned}
\Big|\frac{\sigma_N^2}{2(N-1)} f_{N,\varepsilon}(p)\Big| &= \Big|\frac{\sigma_N^2}{2(N-1)}\mathbb{E}_p\Big[f_{N,\varepsilon}(Y_{T_\varepsilon}^{(N)}) + \sum_{j=0}^{T_\varepsilon - 1}\widetilde{R}_N(Y_j^{(N)})\Big]\Big| \\
&\le H(\varepsilon) + \mathbb{E}_p\big[T_\varepsilon\big] \max_{p\in\{0,\frac{1}{N},\dots,1\}\cap(\varepsilon,1-\varepsilon)} |\widetilde{R}_N(p)|.
\end{aligned}$$

Da $\mathbb{E}_p\big[T_\varepsilon\big] \le \mathbb{E}_p\big[T_{\text{fix}}\big] \le \alpha \frac{N-1}{\sigma_N^2} H(p)$ gilt, ist zusammen mit (1.47) somit

$$\limsup_{N\to\infty} \max_{p\in\{0,\frac{1}{N},\dots,1\}} \Big|\frac{\sigma_N^2}{2(N-1)} f_{N,\varepsilon}(p)\Big| \le H(\varepsilon)$$

(und es ist $H(\varepsilon) \to 0$ für $\varepsilon \downarrow 0$).

Offenbar ist $T_{\text{fix}} - T_\varepsilon \ge 0$ stets. Die starke Markov-Eigenschaft (sowie Spiegelungssymmetrie um $p = 1/2$) zeigt, dass

$$\mathbb{E}_p\big[T_{\text{fix}} - T_\varepsilon\big] \le \sup_{0<p'\le\varepsilon} \mathbb{E}_{p'}\big[T_{\text{fix}}\big] \le \alpha \frac{N-1}{\sigma_N^2} H(\varepsilon).$$

Insgesamt ergibt sich

$$\begin{aligned}\limsup_{N\to\infty} \Big|\frac{\sigma_N^2 \mathbb{E}_p\big[T_{\text{fix}}\big]}{2(N-1)} - H(p)\Big| &\le \limsup_{N\to\infty} \Big|\frac{\sigma_N^2 \mathbb{E}_p\big[T_\varepsilon\big]}{2(N-1)} - H(p)\Big| \\ &\quad + \limsup_{N\to\infty} \frac{\sigma_N^2}{2(N-1)} \mathbb{E}_p\big[T_{\text{fix}} - T_\varepsilon\big] \le \big(1 + \frac{\alpha}{2}\big) H(\varepsilon),\end{aligned}$$

mit $\varepsilon \downarrow 0$ folgt die Behauptung. □

1.2.2 Modelle für den diploiden Fall

Viele Spezies sind *diploid*, besitzen also zwei Kopien jedes (autosomalen) Chromosoms, und typischerweise hat jedes Individuum zwei (verschiedene) Eltern. Gemäß den Mendelschen Regeln erbt ein Kind von jedem Elter jeweils eine Kopie eines der beiden Chromosomen dieses Elters (welche, wird im Idealfall rein zufällig ausgewählt), siehe Abb. 1.2 für ein schematisches Diagramm. Die bisher betrachteten Populationsmodelle beschrei-

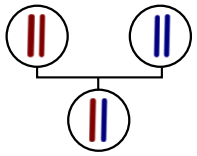


Abbildung 1.2: Schematisches Bild zur Vermehrung im diploiden Fall

ben den *haploiden* Fall, in dem jedes Individuum nur ein Elter besitzt. Wir diskutieren hier kurz, wie die Modelle und die Ergebnisse an den diploiden Fall angepasst werden können.

Wir betrachten eine Population von N diploiden Individuuen pro Generation und interessieren uns für einen gewissen Ort im Genom, auf einem gewissen Chromosom. Wir nehmen der Einfachheit halber an, dass es nur ein Geschlecht gibt (die Individuuen sind Hermaphroditen) und dass Selbstbefruchtung prinzipiell möglich ist. Ein einfaches stochastisches Modell, das in dieser Situation Zufälligkeit in der Nachkommensverteilung beschreibt, ist das diploide Wright-Fisher-Modell:

Definition 1.30 (diploides Wright-Fisher-Modell)**.** Die Population besteht aus $2N$ Chromosomen pro Generation (die in N diploide Individuen aufgeteilt sind, z.B. Chrom. 1 und 2 in Ind. 1, Chrom. 3 und 4 in Ind. 2, etc.). Es findet Zufallspaarung in folgendem Sinn statt: Für jedes der N Kinder unabhängig werden zwei Individuen der Elterngeneration rein zufällig (sagen wir, mit Zurücklegen) als Eltern ausgewählt. Für die beiden Chromosomen in einem Kind bedeutet dies, dass jedes eine Kopie eines rein zufällig aus den $2N$ Chromosomen der Elterngeneration gezogenen Chromosoms ist.

Wir ersetzen also N durch $2N$ im „haploiden Wright-Fisher-Modell“ (Bsp. 1.1) und ändere die Interpretation des Begriffs „Individuum“ entsprechend.

Demnach (nach Satz 1.8) konvergiert die Genealogie einer zufälligen Stichprobe von n Chromosomen im diploiden Wright-Fisher-Modell (z.B. aus $n/2$ diploiden Individuen), wenn man die Zeit mit $2N$ reskaliert, gegen den Kingman-Koaleszenten (eine Koaleszenten-Zeiteinheit entspricht im Modell mit N Individuen nun etwa $2N$ Generationen).

Analog gilt in der neutralen 2 Typ-Situation mit

$$X_r^{(N)} = \text{Anz. Typ 1-Chromosomen in Generation } r,$$

dass der reskalierte Anteilsprozess $(\frac{1}{2N} X^{(N)}_{\lfloor t/2N \rfloor})_{t\geq 0}$ gegen die Wright-Fisher-Diffusion konvergiert (vgl. Satz 1.26).

Bemerkung 1.31 (Hardy-Weinberg-Gleichgewicht[8])**.** Zwei Typen von Chromosomen (im Jargon der Genetik: zwei „Allele“) bedeuten, dass es drei mögliche (diploide) Genotypen der Individuen gibt

$$11, \quad 01, \quad 00$$

(Typ 1-Homozygot, Heterozygot, Typ 0-Homozygot).

Anhand der Information $X_r^{(N)} = k$ allein ist die Aufteilung in diploide Individuen nicht festgelegt. Seien

$$\big(Y_r^{(N)}(11), Y_r^{(N)}(01), Y_r^{(N)}(00)\big)$$

die Anzahlen diploider Ind. (der drei möglichen Genotypen) in Generation r. Gegeben $X_r^{(N)} = k = \lfloor 2Np \rfloor$ (mit $p \in [0,1]$) ist

$$\big(Y_{r+1}^{(N)}(11), Y_{r+1}^{(N)}(01), Y_{r+1}^{(N)}(00)\big) \sim \text{Multinom}\big(N, p^2, 2p(1-p), (1-p)^2\big).$$

Mit dem Gesetz der großen Zahlen und dem multivariaten zentraler Grenzwertsatz folgt: Gegeben $\frac{1}{2N} X_r^{(N)} = p$ ist

$$\Big(\frac{1}{N} Y_{r+1}^{(N)}(11), \frac{1}{N} Y_{r+1}^{(N)}(01), \frac{1}{N} Y_{r+1}^{(N)}(00)\Big) = (p^2, 2p(1-p), (1-p)^2) + O_{\mathbb{P}}\Big(\frac{1}{\sqrt{N}}\Big)$$

(wobei $O_{\mathbb{P}}(1/\sqrt{N})$ einen (zufälligen) Korrekturterm beschreibt, dessen „typische“ Größe höchstens $C/\sqrt{N}$ für eine Konstante $C < \infty$ ist).

[8]Nach Godfrey Harold Hardy (1877–1947) und Wilhelm Weinberg (1862–1937) benannt. G.H. Hardy, Mendelian proportions in a mixed population, *Science* 28 (706), 49–50, (1908); W. Weinberg, Über den Nachweis der Vererbung beim Menschen, *Jahreshefte des Vereins für Vaterländische Naturkunde in Württemberg*, 64, 369–382, (1908).

Die Aufteilung in Genotypen $(11, 01, 00)$ gemäß $p^2 : 2p(1-p) : (1-p)^2$ heißt *Hardy-Weinberg-Gleichgewicht*. Wir sehen: In einer sehr großen Population mit „Zufallspaarungen“ und Startanteil p von Allel 1 stellt sich nach nur einer Generation für die Verteilung der diploiden Genotypen nahezu das Hardy-Weinberg-Gleichgewicht ein (in der mathematischen Idealisierung im Grenzwert $N \to \infty$ stellt es sich exakt ein). Andererseits braucht es nach Satz 1.26 Zeit $\approx \Theta(2N)$, bis eine spürbare Änderung des Allelanteils auftritt, im Vergleich dazu stellt sich HW-Glgw. also „instantan“ ein. Dies ist eine Begründung, warum man sich in Populationsgenetik-Modellierung oft auf den haploiden Fall beschränkt (und beschränken darf).

Bericht 1.32 (Allgemeinere diploide Cannings-Modelle)**.** Das haploide Wright-Fisher-Modell aus Bsp. 1.1 ist ein (wichtiger) Spezialfall der allgemeineren Klasse der (haploiden) Cannings-Modelle aus Def. 1.2. Analog kann das diploide Wright-Fisher-Modell aus Def. 1.30 in eine wesentlich allgemeinere Klasse von diploiden Cannings-Modellen eingebettet werden, und es gelten zu Satz 1.8 analoge Aussagen bezüglich der Konvergenz der Genealogie einer zufälligen Stichprobe gegen den Kingman-Koaleszenten, siehe z.B. M. Möhle und S. Sagitov, Coalescent patterns in diploid exchangeable population models, *J. Math. Biol.*, 337–352, (2003).

1.2.3 Beispiel: Das Experiment von P. Buri

Peter Buri, Gene frequency in small populations of mutant Drosophila, *Evolution* 10, 367–402 (1956) berichtet ein Experiment in „künstlicher Evolution“:

- 105 Populationen von jeweils konstant[9] 16 Taufliegen (8 weibl., 8 männl.) wurden für 19 Generationen (1 Gen. $\approx$ 14d) unter konstanten Bedingungen gehalten.
- 2 Allele: bw und bw^{75}, die 3 Genotypen bw/bw, bw/bw^{75}, bw^{75}/bw^{75} sind anhand der Augenfarbe unterscheidbar
- Vorexperimente legten nahe, dass diese Genotypen keinen Einfluss auf den erwarteten Reproduktionserfolg haben.
- Die Anzahl bw^{75}-Chromosomen in jeder Population und Generation wurde beobachtet, s.a. Abb. 1.3.

In dieser Laborsituation kann man die Wirkung der Gendrift direkt beobachten. Beispielsweise passt der beobachtete Abfall der (empirischen) Heterozygotie, gemittelt über die 105 Populationen, recht gut zum mittels Lemma 1.19 theoretisch vorhergesagten geometrischen Abfallen der erwartete Stichprobenheterozygotie, allerdings muss der „reale“ Populationsgrößenparameter $2N = 32$ durch die „effektive“ Populationsgröße $2N = 23$ ersetzt werden, siehe Abb. 1.4.

[9] Die konstante Populationsgröße wurde jeweils „von Hand“ beim Umsetzen der nächsten Generation in ein neues Glas erzwungen, zwischenzeitlich waren die Populationen natürlich angewachsen.

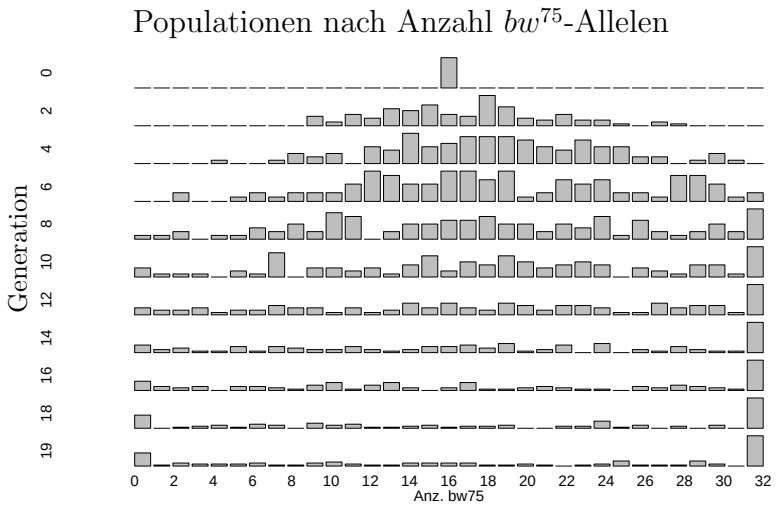


Abbildung 1.3: Beobachtungen aus 105 *Drosophila melanogaster*-Populationsexperimenten (aus P. Buri, *loc. cit.*, Table 14, S. 387)

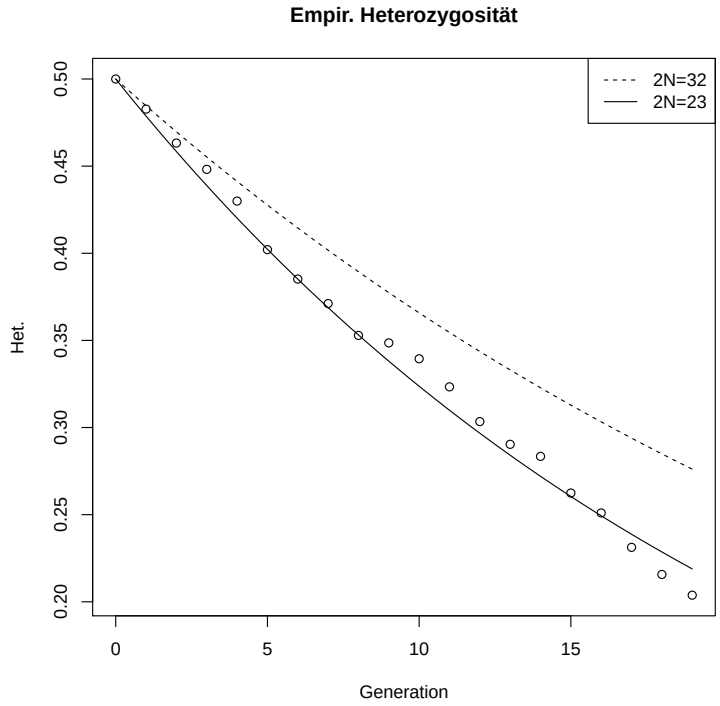


Abbildung 1.4: Beobachtete empirische Heterozygotie und nach Lemma 1.19 angepasste Kurven für die (theoretische) erwartete Stichprobenheterozygotie für zwei Parameterwahlen ($2N = 32$ und $2N = 23$)

Kapitel 2

Mutationen und der markierte Koaleszent

2.1 Zwei (neutrale) Typen

Wir betrachten zunächst wieder Populationsmodelle mit zwei verschiedenen Typen (sagen wir $E = \{0,1\}$) und schreiben

$$X_t = \text{Anz. Typ-1-Individuen in Generation } t.$$

Wie in den Cannings-Modellen aus Def. 1.16 nehmen wir an, dass im Modell mit Populationsgröße N die Nachkommensverteilungen durch einen austauschbaren Vektor $\nu^{(N)}$ beschrieben werden (u.i.v. für verschiedene Generationen).

Wir erlauben nun zusätzlich die Möglichkeit der Mutation:

- Das Kind eines Typ 1-Elters habe mit W'keit $\mu_{10}^{(N)}$ den Typ 0 (und mit W'keit $1-\mu_{10}^{(N)}$ den Typ des Elters),
- das Kind eines Typ 0-Elters habe mit W'keit $\mu_{01}^{(N)}$ den Typ 1 (und mit W'keit $1-\mu_{01}^{(N)}$ den Typ des Elters),
- Mutationen entstehen u.a. für verschiedene Kinder.

Definition 2.1 (Typenanzahlprozess eines 2-Typ Cannings-Modells mit Mutationen)**.** Die Markovkette auf $\{0,1,\dots,N\}$ mit Übergangsmatrix

$$\begin{aligned}&\mathbb{P}\big(X_{t+1} = y \,\big|\, X_t = x\big)\\ &= \sum_{k=0}^{N} \mathbb{P}\Big(\sum_{i=1}^{x} \nu_i^{(N)} = k\Big) \sum_{\ell=0}^{k\wedge y} \binom{k}{\ell} (1-\mu_{10}^{(N)})^{\ell} (\mu_{10}^{(N)})^{k-\ell} \binom{N-k}{y-\ell} (\mu_{01}^{(N)})^{y-\ell} (1-\mu_{01}^{(N)})^{N-k-y+\ell}\end{aligned}$$

d.h. gegeben $X_t = x$ ist

$$X_{t+1} \sim B(X'_{t+1}, 1-\mu_{10}^{(N)}) + \widetilde{B}(N - X'_{t+1}, \mu_{01}^{(N)}) \tag{2.1}$$

mit $B(x,p), \widetilde{B}(x,p) \sim \mathrm{Bin}(x,p)$, u.a. ($x \in \{0,\dots,N\}$, $p \in [0,1]$) und $X'_{t+1} =^d \sum_{i=1}^{x} \nu_i^{(N)}$ heißt *Typenanzahlprozess eines 2-Typ Cannings-Modells mit Mutationen* (mit Nachkommensvektor(verteilung) $\mathscr{L}(\nu^{(N)})$ und Mutationswahrscheinlichkeiten $\mu_{10}^{(N)}$, $\mu_{01}^{(N)}$).

Beobachtung 2.2. 1. Falls $\mu_{10}^{(N)}, \mu_{01}^{(N)} > 0$, so besitzt (X_t) stets ein eindeutiges Gleichgewicht (es handelt sich um eine irreduzible Markovkette auf der endlichen Menge $\{0, 1, \dots, N\}$).

2. Aus der Darstellung (2.1) lassen sich leicht bedingter Erwartungswert und Varianz bestimmen:

$$\mathbb{E}\big[X_{t+1} \,\big|\, X_t = x\big] = x(1-\mu_{10}^{(N)}) + (N-x)\mu_{01}^{(N)},$$
$$\mathrm{Var}\big[X_{t+1} \,\big|\, X_t = x\big] = (1-\mu_{10}^{(N)}-\mu_{01}^{(N)})^2 c_N x(N-x) + x(1-\mu_{10}^{(N)})\mu_{10}^{(N)} + (N-x)\mu_{01}^{(N)}(1-\mu_{01}^{(N)})$$

(zur Berechnung der Varianz ist die allgemeine Formel $\mathrm{Var}[Y] = \mathrm{Var}[\mathbb{E}[Y|X']] + \mathbb{E}[\mathrm{Var}[Y|X']]$ angenehm); falls $\mu_{10}^{(N)} = \mu_{01}^{(N)} = 0$, so ergibt sich natürlich wieder die Formel aus Beob. 1.18.

Insbesondere gilt (analog zu Beob. 1.23) für $z \in [0,1] \cap \frac{1}{N}\mathbb{Z}$

$$\mathbb{E}\big[\frac{1}{N}X_{t+1} - z \,\big|\, X_t = Nz\big] = -z\mu_{10}^{(N)} + (1-z)\mu_{01}^{(N)}$$
$$\mathbb{E}\Big[\Big(\frac{1}{N}X_{t+1} - z\Big)^2 \,\Big|\, X_t = Nz\Big] = \mathrm{Var}\Big[\frac{1}{N}X_{t+1} \,\Big|\, X_t = Nz\Big] + \big(-z\mu_{10}^{(N)} + (1-z)\mu_{01}^{(N)}\big)^2$$
$$= (1-\mu_{10}^{(N)}-\mu_{01}^{(N)})^2 c_N z(1-z) + \frac{1}{N}z(1-\mu_{10}^{(N)})\mu_{10}^{(N)} + \frac{1}{N}(1-z)\mu_{01}^{(N)}(1-\mu_{01}^{(N)}) + \big(-z\mu_{10}^{(N)} + (1-z)\mu_{01}^{(N)}\big)^2$$

(für das 2. Moment verwende man z.B. die Formel $\mathbb{E}[(Y-z)^2] = \mathrm{Var}[Y] + (\mathbb{E}[Y]-z)^2$).

Falls $c_N \to 0$ und

$$\mu_{01}^{(N)} \sim c_N \frac{\theta_1}{2}, \quad \mu_{10}^{(N)} \sim c_N \frac{\theta_0}{2} \quad \text{für } N \to \infty \tag{2.2}$$

mit $\theta_0, \theta_1 \in (0, \infty)$ gilt, so gilt für den reskalierten Anteilsprozess

$$Z_t^{(N)} := \frac{1}{N} X_{\lfloor t/c_N \rfloor}^{(N)}, \quad t \geq 0$$

analog zur Argumentation in Beob. 1.23

$$\lim_{N\to\infty} \frac{1}{c_N}\mathbb{E}\big[f\big(Z_{t+c_N}^{(N)}\big) - f(z_N) \,\big|\, Z_t^{(N)} = z_N\big] = \Big(-\frac{\theta_0}{2}z + \frac{\theta_1}{2}(1-z)\Big)f'(z) + \frac{1}{2}z(1-z)f''(z)$$

für $f \in C^2([0,1])$ und $z_N \in [0,1] \cap \frac{1}{N}\mathbb{Z}$ mit $z_N \to z \in [0,1]$.

Bemerkung. Die Annahme (2.2), die Populationsgröße und Mutationswahrscheinlichkeiten aneinander koppelt, mag auf den ersten Blick unnatürlich erscheinen: Warum sollte die Mutationswahrscheinlichkeit, die sich aus der Effektivität der biochemischen Kopier- und Reparaturmechanismen innerhalb der Zellen eines Individuums, ggfs. im Zusammenspiel mit diversen Umwelteinflüssen, ergibt, irgend etwas mit der Populationsgröße zu tun haben?

Annahme (2.2) bedeutet, dass Mutation und Gendrift „auf derselben Zeitskala“ wirken. Wenn $\mu_{01}^{(N)}, \mu_{10}^{(N)} \ll c_N$, so wirkt die Gendrift (wie wir wissen, über Zeiten der

Größenordnung $O(1/c_N)$), bevor Mutationen irgendeinen merklichen Einfluss auf die Zusammensetzung der Population haben, wenn $\mu_{01}^{(N)}, \mu_{10}^{(N)} \gg c_N$, so stellt sich „reines Mutationsgleichgewicht“ ein, an dem Gendrift nichts ändert.

Anders gewendet: Für eine gegebene Population (mit endlichen, aber sehr großem N) sind

$$\frac{2}{c_N}\mu_{01}^{(N)} \approx \theta_1 \quad \text{und} \quad \frac{2}{c_N}\mu_{10}^{(N)} \approx \theta_0$$

die „entscheidenden“ Parameter, um die Zeitentwicklung des Typenanteils zu beschreiben.

Bericht 2.3. Falls $c_N \to 0$, $d_N/c_N \to 0$ sowie (2.2) und $Z_0^{(N)} = z_N \to z \in [0,1]$ gelten, so konvergiert der reskalierte Anteilsprozess $(Z_t^{(N)})_{t\geq 0}$ gegen den starken Markovprozess $Z = (Z_t)_{t\geq 0}$ auf $[0,1]$ mit Generator

$$Lf(z) = \Big(-\frac{\theta_0}{2}z + \frac{\theta_1}{2}(1-z)\Big)f'(z) + \frac{1}{2}z(1-z)f''(z), \quad z \in [0,1], \ \ f \in C^2([0,1]). \tag{2.3}$$

Z heißt die (neutrale 2 Typ-)Wright-Fisher-Diffusion mit Mutation.

Analog zu Bem. 1.25 kann man Z auch als Lösung des Martingalproblems mit Generator L aus (2.3) oder durch die Forderung

$$\begin{aligned} &Z \text{ ist stetiges Semimartingal mit Werten in } [0,1] \text{ und } \langle Z\rangle_t = \int_0^t Z_s(1-Z_s)\,ds, \\ &M_t := Z_t - \int_0^t \Big(-\frac{\theta_0}{2}Z_s + \frac{\theta_1}{2}(1-Z_s)\Big)\,ds, \ t \geq 0 \quad \text{ist ein Martingal} \end{aligned}$$

charakterisieren. Z ist ebenfalls die (eindeutige, starke) Lösung der stochastischen Differentialgleichung

$$dZ_t = \Big(-\frac{\theta_0}{2}Z_t + \frac{\theta_1}{2}(1-Z_t)\Big)\,dt + \sqrt{Z_t(1-Z_t)}\,dB_t$$

(mit (B_t) standard-Brownbewegung).

Man kann dies analog zum Beweis von Satz 1.26 beweisen.

Die Gleichgewichtsverteilung $\pi^{(N)} \in \mathcal{M}_1(\{0,1,\dots,N\})$ von $X^{(N)}$ ist zwar prinzipiell durch das Gleichungssystem

$$\pi^{(N)}(x) = \sum_{y\in\{0,1,\dots,N\}} \pi^{(N)}(y)\mathbb{P}\big(X_1^{(N)} = x \,\big|\, X_0^{(N)} = y\big), \quad x \in \{0,1,\dots,N\}$$

(zusammen mit der Bedingung $\sum_{x=0}^N \pi^{(N)}(x) = 1$) eindeutig festgelegt, aber diese Gleichungssystem ist (zumal für große N) i.A. nicht explizit lösbar.

Bericht 2.4. Für jedes $z \in [0,1]$ gilt

$$\mathscr{L}(Z_t \,|\, Z_0 = z) \underset{N\to\infty}{\longrightarrow} \mathrm{Beta}(\theta_1,\theta_0)$$

und $\mathrm{Beta}(\theta_1,\theta_0)$ ist das eindeutige Gleichgewicht von Z. (Die Dichte von $\mathrm{Beta}(\theta_1,\theta_0)$ ist $\Gamma(\theta_0+\theta_1)/(\Gamma(\theta_0)\Gamma(\theta_1))z^{\theta_1-1}(1-z)^{\theta_0-1}$.)

Wir betrachten dazu (hier nur) eine Heuristik: Z löst

$$\begin{aligned} dZ_t &= \Big(-\frac{\theta_0}{2} Z_t + \frac{\theta_1}{2}(1 - Z_t)\Big) dt + \sqrt{Z_t(1-Z_t)}\, dB_t \\ &= -\frac{\theta}{2}\Big(Z_t - \frac{\theta_1}{\theta}\Big) dt + \sqrt{Z_t(1-Z_t)}\, dB_t = \mu(Z_t) + \sigma(Z_t)\, dB_t \end{aligned}$$

(wir schreiben $\theta := \theta_0 + \theta_1$).

Betrachte allgemeiner die Lösung X der stochastischen Differentialgleichung

$$dX_t = \mu(X_t)\, dt + \sigma(X_t)\, dB_t$$

(mit $\mu(\cdot)$ "Driftkoeffizient", $\sigma(\cdot)$ Diffusionskoeffizient). Sei

$$\phi(x) := \exp\Big(-\int_{1/2}^{x} \frac{2\mu(z)}{\sigma^2(z)}\, dz\Big), \quad x \in [0,1].$$

Die Funktion

$$s(x) := \int^x \phi(y)\, dy$$

ist eine sog. Skalenfunktion für X: Itōs Formel zeigt, dass $(s(X_t))_{t\geq 0}$ ein Martingal ist.

Die Gleichgewichtsdichte von X ist dann gegeben durch

$$\pi(x) = C/(\phi(x)\sigma^2(x)) \tag{2.4}$$

(mit Normierung $1/C = \int_0^1 1/(\phi(x)\sigma^2(x))\, dx$), siehe z.B. Rogers & Williams [RW], Vol. 2, Ch. V.54 und speziell Thm. (54.5).

Heuristisch kann man (2.4) folgendermaßen einsehen: Nehmen wir an, dass es eine Gleichgewichtsdichte $\pi(x)$ gibt, so gilt für genügend glatte testfunktionen f

$$\begin{aligned} 0 = \mathbb{E}_\pi[Lf(X_0)] &= \int \pi(x)\big[\mu(x) f'(x) + \frac{1}{2}\sigma^2(x) f''(x)\big]\, dx \\ &= \int \big[-(\pi\mu)'(x) + \frac{1}{2}(\pi\sigma^2)''(x)\big] f(x)\, dx \end{aligned}$$

wobei wir in der zweiten Zeile partiell integrieren (und annehmen, dass f und seine Ableitungen am Rand verschwinden, so dass es beim partiellen Integrieren keine Randterme gibt).

Da obiges Integral für beliebige (genügend glatte) f verschwindet, muss der Integrand $= 0$ sein, d.h.

$$(\pi\mu)' = \frac{1}{2}(\pi\sigma^2)''.$$

Dafür ist $\pi\mu = \frac{1}{2}(\pi\sigma^2)'$, oder äquivalent

$$\big(\log(\pi\sigma^2)\big)' = (\pi\sigma^2)'/(\pi\sigma^2) = 2\mu/\sigma^2,$$

hinreichend, d.h. $\pi = \text{const}/(\sigma^2\phi)$ ist eine Lösung.

Für Z ergibt sich mit $\mu(z) = \theta(\frac{\theta_1}{\theta} - z)/2$, $\sigma^2(z) = z(1-z)$

$$\int_{1/2}^{x} \frac{2\mu(z)}{\sigma^2(z)}\,dz = \int_{1/2}^{x} -\frac{\theta(z-\frac{\theta_1}{\theta})}{z(1-z)}\,dz = \int_{1/2}^{x} \frac{\theta_1-\theta}{1-z} + \frac{\theta_1}{z}\,dz = \theta_0\log(1-x) + \theta_1\log(x) + \text{const.},$$

somit

$$\pi(x) = C\exp\Big(\theta_0\log(1-x) + \theta_1\log(x)\Big)\frac{1}{x(1-x)} = C\,x^{\theta_1-1}(1-x)^{\theta_0-1}.$$

Beobachtung 2.5 (Stichprobenformel im 2-Typ-Fall)**.** Die W'keit, dass in einer Stichprobe der Größe $n = n_0 + n_1$ aus der Population im Gleichgewicht genau n_1 Individuen Typ 1 haben, ist

$$\begin{aligned} p(n_1,n_0) &= \binom{n}{n_1}\mathbb{E}_\pi\big[Z^{n_1}(1-Z)^{n_0}\big] = \binom{n}{n_1}\frac{\Gamma(\theta_0+\theta_1)}{\Gamma(\theta_0)\Gamma(\theta_1)}\int_0^t z^{n_1}(1-z)^{n_0}z^{\theta_1-1}(1-z)^{\theta_0-1}\,dz \\ &= \frac{(n_1+n_0)!}{n_1!\,n_0!}\,\frac{\Gamma(\theta_0+\theta_1)}{\Gamma(\theta_0)\Gamma(\theta_1)}\,\frac{\Gamma(n_1+\theta_1)\Gamma(n_0+\theta_0)}{\Gamma(n_1+n_0+\theta_1+\theta_0)} = \frac{(n_1+n_0)!}{n_1!\,n_0!}\,\frac{(\theta_1)_{n_1\uparrow}(\theta_0)_{n_0\uparrow}}{(\theta_1+\theta_0)_{(n_1+n_0)\uparrow}} \end{aligned}$$

(beachte $x\Gamma(x) = \Gamma(x+1)$ und somit $\Gamma(\theta+n)/\Gamma(\theta) = (\theta+n-1)(\theta+n-2)\cdots(\theta+1)\theta = (\theta)_{n\uparrow}$). Die Anzahl Typ 1-Individuen in einer n-Stichprobe aus dem Gleichgewicht ist also Beta-Binomial(n,θ_0,θ_1)-verteilt.

Alternativ können wir diese Stichprobenformel folgendermaßen gewinnen: Betrachte einen Kingman-n-Koaleszent, längs dessen Ästen mit Rate $\theta/2$ Mutationen fallen ($\theta := \theta_1 + \theta_0$),

- mit W'keit θ_1/θ ist es eine "$\to$ 1-Mutation",
- mit W'keit θ_0/θ ist es eine "$\to$ 0-Mutation".
- Markiere Typ der Wurzel zufällig (Typ 1 mit W'keit θ_1/θ, Typ 0 mit W'keit θ_0/θ) und
- propagiere Typen längs den Ästen des Baums von der Wurzel zu den Blättern (der Typ ändert sich auf 1, wenn man eine "$\to$ 1-Mutation" trifft und analog zu 0, wenn man eine "$\to$ 0-Mutation" trifft),
- lese Typen an den Blättern ab.

`[Bild an der Tafel]`

Sei

$$p(n_1,n_0) = \text{ W'keit, dass genau } n_1 \text{ Blätter Typ 1, } n_0 \text{ Blätter Typ 0 erhalten}$$

und

$$\widetilde{p}(n_1,n_0) = \begin{array}{l}\text{W'keit, dass } n_1 \text{ vorgegebene Blätter Typ 1,}\\ \text{die übrigen } n_0 \text{ Blätter Typ 0 erhalten}\end{array}.$$

Es ist $p(n_1,n_0) = \binom{n_1+n_0}{n_1}\widetilde{p}(n_1,n_0)$ wegen Symmetrie, in der Definition von $\widetilde{p}(n_1,n_0)$ denken wir uns die Blätter mit $1,\dots,n$ nummeriert und verlangen, dass eine vorgebene Teilmenge $A \subset [n]$ von Blättern mit $|A| = n_1$ Typ 1 erhält.

$\widetilde{p}$ löst (zerlege nach dem "jüngsten Ereignis")

$$\begin{aligned}\widetilde{p}(n_1,n_0) &= \frac{n_1\theta_1/2}{\binom{n_1+n_0}{2}+(n_0+n_1)\theta/2}\widetilde{p}(n_1-1,n_0) + \frac{n_0\theta_0/2}{\binom{n_1+n_0}{2}+(n_0+n_1)\theta/2}\widetilde{p}(n_1,n_0-1)\\ &\quad + \frac{\binom{n_1}{2}}{\binom{n_1+n_0}{2}+(n_0+n_1)\theta/2}\widetilde{p}(n_1-1,n_0) + \frac{\binom{n_0}{2}}{\binom{n_1+n_0}{2}+(n_0+n_1)\theta/2}\widetilde{p}(n_1,n_0-1)\\ &\quad + \frac{n_1\theta_0/2+n_0\theta_1/2+n_1n_0}{\binom{n_1+n_0}{2}+(n_0+n_1)\theta/2}\times 0\\ &= \frac{n_1(n_1-1+\theta_1)}{(n_1+n_0)(n_1+n_0-1+\theta)}\widetilde{p}(n_1-1,n_0) + \frac{n_0(n_0-1+\theta_0)}{(n_1+n_0)(n_1+n_0-1+\theta)}\widetilde{p}(n_1,n_0-1)\end{aligned}$$

mit Randwerten $\widetilde{p}(1,0) = \theta_1/\theta$, $\widetilde{p}(0,1) = \theta_0/\theta$.

Hierbei entsprechen die Terme

- in der ersten Zeile aus einem Mutationsereignis (eine der n_1 Linien, die Typ 1 haben sollen, wir von einer "$\to$ 1-Mutation" getroffen bzw. eine der n_0 Linien, die Typ 0 haben sollen, von einer "$\to$ 0-Mutation"),
- in der zweiten Zeile aus einem Verschmelzungsereignis (eine Verschmelzung innerhalb der Gruppe der n_1 Linien, die Typ 1 haben sollen, oder innerhalb der Gruppe der n_0 Linien, die Typ 0 haben sollen) und
- in der dritten Zeile Ereignissen, die nicht mit dem geforderten Ausgang kompatibel sind (eine der n_1 Linien, die Typ 1 haben sollen, wir von einer "$\to$ 0-Mutation" getroffen oder eine der n_0 Linien, die Typ 0 haben sollen, von einer "$\to$ 1-Mutation" oder eine Linie, die Typ 1 haben soll, verschmilzt mit einer, die Typ 0 haben soll).

Die eindeutige Lösung ist tatsächlich

$$\widetilde{p}(n_1,n_0) = \frac{(\theta_1)_{n_1\uparrow}(\theta_0)_{n_0\uparrow}}{(\theta_1+\theta_0)_{(n_1+n_0)\uparrow}} = \mathbb{E}_\pi\big[Z^{n_1}(1-Z)^{n_0}\big],$$

denn

$$\begin{aligned}&\frac{n_1(n_1-1+\theta_1)}{(n_1+n_0)(n_1+n_0-1+\theta)}\frac{(\theta_1)_{(n_1-1)\uparrow}(\theta_0)_{n_0\uparrow}}{(\theta_1+\theta_0)_{(n_1+n_0-1)\uparrow}} + \frac{n_0(n_0-1+\theta_0)}{(n_1+n_0)(n_1+n_0-1+\theta)}\frac{(\theta_1)_{n_1\uparrow}(\theta_0)_{(n_0-1)\uparrow}}{(\theta_1+\theta_0)_{(n_1+n_0-1)\uparrow}}\\ &= \frac{n_1}{n_1+n_0}\frac{(\theta_1)_{n_1\uparrow}(\theta_0)_{n_0\uparrow}}{(\theta_1+\theta_0)_{(n_1+n_0)\uparrow}} + \frac{n_0}{n_1+n_0}\frac{(\theta_1)_{n_1\uparrow}(\theta_0)_{n_0\uparrow}}{(\theta_1+\theta_0)_{(n_1+n_0)\uparrow}}.\end{aligned}$$

Bemerkung 2.6 (Dualität mit Feynman-Kac-Korrektur zwischen WF-Diffusion mit Mutation und Blockzählprozess eines Koaleszenten „mit Tötung"). Sei $(N_t)_{t\ge 0}$ zeitkont. Markovkette auf $\mathbb{N}_0$ mit Sprungratenmatrix

$$q_{n,n-1} = \binom{n}{2} + \frac{\theta_1}{2}n, \quad q_{n,n} = -q_{n,n-1}, \qquad (\text{sonst} = 0),$$

(d.h. (N_t) ist der Klassenzählprozess eines Kingman-Koaleszenten, in dem jede Klasse zusätzlich mit Rate $\frac{\theta_1}{2}$ „stirbt“).

Es gilt

$$\mathbb{E}_z[Z_t^n] = \mathbb{E}_n\Big[z^{N_t}\exp\big(-\frac{\theta_0}{2}\int_0^t N_s\,ds\big)\Big] \tag{2.5}$$

(im Fall $\theta_0 = \theta_1 = 0$ entspricht dies (1.33) aus dem Beweis von Satz 1.26).

Beweis. Mit Itō-Formel ist

$$\begin{aligned}
Z_t^n &= Z_0^n + \int_0^t nZ_s^{n-1}\,dZ_s + \frac{1}{2}\int_0^t n(n-1)Z_s^{n-2}d\langle Z\rangle_s \\
&= Z_0^n + \int_0^t nZ_s^{n-1}Z_s(1-Z_s)\,dB_s + \int_0^t nZ_s^{n-1}\big(-\frac{\theta_0}{2}Z_s + \frac{\theta_1}{2}(1-Z_s)\big)\,ds \\
&\quad + \frac{1}{2}\int_0^t n(n-1)Z_s^{n-2}Z_s(1-Z_s)\,ds \\
&= Z_0^n + \int_0^t \big(\frac{n(n-1)}{2} + n\frac{\theta_1}{2}\big)\big(Z_s^{n-1} - Z_s^n\big)\,ds - \frac{\theta_0}{2}\int_0^t Z_s\,ds + \text{Martingal}_t,
\end{aligned}$$

d.h. $f(n,z,t) := \mathbb{E}_z[Z_t^n]$ löst

$$\frac{\partial}{\partial t}f(n,z,t) = \big(\frac{n(n-1)}{2} + n\frac{\theta_1}{2}\big)\big(f(n-1,z,t) - f(n,z,t)\big) - \frac{\theta_0}{2}f(n,z,t),\ t\ge 0, n\in\mathbb{N}$$

(mit Startwert $f(n,z,0) = z^n$).

Ebenso löst

$$g(n,z,t) := \mathbb{E}_n\Big[z^{N_t}\exp\big(-\frac{\theta_0}{2}\int_0^t N_s\,ds\big)\Big]$$

dieses (eindeutig lösbare) System linearer Differentialgleichungen. Diese Aussage ist ein Spezialfall einer sogenannten Feynman-Kac-Formel, siehe z.B. Rogers & Williams [RW], Vol. 1, Ch. III.19 und Vol. 2, Ch. IV.22, Ex. (22.11). Für einen direkten Beweis — in diesem Kontext eine kleine Variation über die Kolmogorov-Rückwärtsgleichung — beachte

$$\begin{aligned}
&\frac{1}{h}\Big(g(n,z,t+h) - g(n,z,t)\Big) = \frac{1}{h}\Big(\mathbb{E}_n\big[\mathbb{E}_n\big[z^{N_{t+h}}e^{-\frac{\theta_0}{2}\int_0^{t+h}N_s\,ds}\big)\,|\,N_h\big]\big] - \mathbb{E}_n\big[z_t^N e^{-\frac{\theta_0}{2}\int_0^t N_s\,ds}\big)\big]\Big) \\
&= \frac{1}{h}\Big(\big(1 - h\tfrac{n(n-1+\theta_1)}{2}\big)e^{-\frac{\theta_0}{2}nh}\mathbb{E}_n\big[z^{N_t}e^{-\frac{\theta_0}{2}\int_0^t N_s\,ds}\big] + h\tfrac{n(n-1+\theta_1)}{2}\mathbb{E}_{n-1}\big[z^{N_t}e^{-\frac{\theta_0}{2}\int_0^t N_s\,ds}\big] + O(h^2) \\
&\qquad - \mathbb{E}_n\big[z^{N_t}e^{-\frac{\theta_0}{2}\int_0^t N_s\,ds}\big]\Big) \\
&= \frac{n(n-1+\theta_1)}{2}\big(g(n-1,z,t) - g(n,z,t)\big) - \frac{\theta_0}{2}ng(n,z,t) + O(h),
\end{aligned}$$

mit $h\downarrow 0$ folgt die Behauptung. □

Bemerkung. Man kann (2.5) auch anhand des mit Mutationen markierten Koaleszenten-Baums aus Beob. 2.5 interpretieren: $\int_0^t N_s\,ds$ ist die Gesamtlänge der Zweige, die noch keine “→ 1-Mutation” getroffen haben. Wenn auf einen solchen Teil eines Asts eine “→ 0-Mutation” fällt, tritt das gewünschte Ereignis (alle n Stichproben sind vom Typ 1) nicht ein. Da die “→ 0-Mutation” gemäß einem Poissonprozess mit Rate $\theta_0/2$ auftreten, ist $\exp\big(-\frac{\theta_0}{2}\int_0^t N_s\,ds\big)$ die bedingte Wahrscheinlichkeit gegeben (N_t), dass in diesem Teil der Genealogie keine “→ 0-Mutation” auftritt.

2.1.1 Allgemeine endliche Typenmenge

Wir berichten hier kurz die Situation, wenn es anstelle von zwei eine beliebige endliche Anzahl möglicher Typen gibt.

Sei E endliche Typenmenge (sagen wir, $E = \{1, \dots, d\}$), p irred. stoch. Matrix auf E. Betrachte Cannings-Modelle mit Nachkommensvektoren $\nu^{(N)}$, ein Kind eines Typ k-Elters habe mit W'keit $\mu_{k,\ell}^{(N)}$ den Typ ℓ und es gelte

$$\lim_{N\to\infty} \frac{1}{c_N} \mu_{k,\ell}^{(N)} = \frac{\theta}{2} p_{k,\ell} \quad \text{für } \ell \neq k$$

mit einem $\theta > 0$. Sei

$$Z_{t,k}^{(N)} = \frac{1}{N}\big(\text{Anz. Typ } k\text{-Individuen in Generation } \lfloor t/c_N \rfloor\big).$$

Bericht 2.7. Falls $Z_0^{(N)} \to z \in [0,1]^d$ (mit $z_1 + \cdots + z_d = 1$), so gilt in dieser Situation $Z^{(N)} \to^d Z$, die d-dim. Wright-Fisher-Diffusion mit Mutation. Z ist ein starker Markovprozess auf dem $(d-1)$-dimensionalen Simplex $\{(x_1, \dots, x_d) \in [0,1]^d : x_1 + \cdots + x_d = 1\}$ mit Generator

$$Lf(x) = \frac{1}{2} \sum_{i,j=1}^{d} x_i(\delta_{ij} - x_j) \frac{\partial^2}{\partial x_i x_j} f(x) + \frac{\theta}{2} \sum_{i=1}^{d} \Big(\sum_{j\neq i}^{d} (p_{ji} x_j - p_{ij} x_i) \Big) \frac{\partial}{\partial x_i} f(x).$$

für $x = (x_1, \dots, x_d)$. Z besitzt ein eindeutiges Gleichgewicht.

Definition 2.8 („markierter n-Koaleszent“)**.**

- Betrachte n-Koaleszent als (zufälligen) gewurzelten Baum „mit Astlängen,“
- längs den Kanten fallen Mutationsereignisse als unabhängige Poissonprozesse mit Rate $\theta/2$,
- markiere die Wurzel mit einem gemäß π, dem Gleichgewicht von p, zufällig ausgewählten Typ,
- propagiere Typen von der Wurzel zu den Blättern, bei jedem Mutationsereignis führe u.a. einen Schritt gemäß p aus.

(Der Typenprozess längs den Ästen ist eine Baum-indizierte Markovkette.)

Bericht 2.9. Die Verteilung einer n-Stichprobe aus dem Gleichgewicht der d-Typ Wright-Fisher-Diffusion mit Mutation (aus Bericht 2.7) entspricht der Verteilung der Typen an den Blättern eines entsprechend markierten n-Koaleszenten aus Def. 2.8. Man kann gemischte Momente der Gleichgewichtsverteilung von Z auf diese Weise charakterisieren (ähnlich wie in Beob. 2.5 im Fall von 2 Typen).

2.2 Infinitely-many-alleles-Modell (IMA)

Definition 2.10 (Infinitely-many-alleles-Mutationsmechanismus)**.** Wir treffen die Modellannahme, dass jede Mutation einen völlig neuen Typ erzeugt (und die Mutationen sind „neutral“, d.h. sie beeinflussen den Fortpflanzungserfolg nicht). In der Literatur ist auch der Name “infinite alleles model” üblich.

Mathematisch realisieren wir dies durch die Wahl $E = [0,1]$ als Typenmenge, bei jedem Mutationsereignis wird (unabhängig) der neue Typ uniform($[0,1]$)-verteilt gewählt.

Wenn Mutationen sich mit Rate $\theta/2 > 0$ ereignen, ist die Vorwärtsentwicklung des Typs längs einer Abstammungslinie demnach beschrieben durch den Markov-Prozess mit Generator

$$Bf(x) = \frac{\theta}{2} \int_0^1 f(u) - f(x)\, du, \quad x \in [0,1]$$

für $f : [0,1] \to \mathbb{R}$ beschränkt und messbar.

Betrachte Stichprobe der Größe n: Beobachtete genetische Variabilität modelliert durch n-Koaleszent, längs dessen Kanten sich mit Rate $\frac{\theta}{2}$ Mutationen gem. IMA-Modell ereignen.

`[Bild an der Tafel]`

Offenbar ist nur der Teil der Genealogie jeweils „oberhalb“ der jüngsten Mutation relevant, für die Beobachtungen an den Blättern können wir also folgende äquivalente Dynamik betrachten:

Definition 2.11 („Getöteter n-Koaleszent“)**.**

- Beginne mit $\{\{1\}, \{2\}, \ldots, \{n\}\}$ (alle aktiv).
- Jedes Paar von aktiven Klassen verschmilzt mit Rate 1.
- Jede Klasse wird mit Rate $\frac{\theta}{2}$ getötet/inaktiviert: allen Elementen wird derselbe, unif($[0,1]$)-verteilte Typ zugeordnet und die Klasse wird inaktiviert

  (sie hat ihre „definierende Mutation“ getroffen).
- Ende, wenn keine aktiven Klassen mehr übrig.

Analog zu Def. 1.7 (Kingman-Koaleszent) kann man dies formal als zeitkontinuierliche Markovkette auf

$$\widetilde{\mathcal{E}}_n = \big\{\text{Äquivalenzrelationen auf } [n], \text{ deren Klassen als aktiv/inaktiv markiert sind}\big\}$$

ausformulieren (Übung: man stelle die Sprungratenmatrix auf).

Bemerkung 2.12. Angesichts der Symmetrien des Koaleszenten ist die eigentlich relevante Information das *Typenhäufigkeitsspektrum* $(B_1, \ldots, B_n)$, wobei

$$B_i = \#\text{Typen, die } i\text{-mal in der Stichprobe vorkommen}, \quad i = 1, \ldots, n$$

(offenbar ist $\sum_{i=1}^n i\, B_i = n$).

Gegeben $(B_1, \dots, B_n) = (b_1, \dots, b_n)$ mit $\sum_{i=1}^n i b_i = n$ und

$$\sum_{i=1}^n b_i = k$$

sind die beobachteten Typen in der Stichprobe folgendermaßen verteilt: Zerlege $\{1, \dots, n\}$ uniform in k Teilmengen der Größen

$$\underbrace{1, \dots, 1}_{b_1}, \underbrace{2, \dots, 2}_{b_2}, \dots, \underbrace{n}_{b_n},$$

ordne jeder Teilmenge u.a. einen unif($[0,1]$)-verteilten Typ zu.

Seien $E_n, \dots, E_1$ ZVn mit Werten in $\{\text{koal}, \text{mut}\}$,

E_k = Typ des Ereignisses, das die Anz. aktiver Klassen von k auf $k-1$ reduziert.

Es gilt

$$\mathbb{P}(E_k = \text{coal}) = \frac{\binom{k}{2}}{\binom{k}{2} + k\frac{\theta}{2}} = \frac{k-1}{k-1+\theta},$$
$$\mathbb{P}(E_k = \text{mut}) = \frac{k\frac{\theta}{2}}{\binom{k}{2} + k\frac{\theta}{2}} = \frac{\theta}{k-1+\theta}$$

und $E_n, \dots, E_1$ sind unabängig (verwende „konkurrierende-Raten“-Argument und Symmetrien der Sprungraten). Also gilt für $e_n, e_{n-1}, \dots, e_1 \in \{\text{coal}, \text{mut}\}$

$$\mathbb{P}(E_n = e_n, \dots, E_1 = e_1) = \frac{\prod_{k=1}^n \left(\theta \mathbf{1}(e_k = \texttt{mut}) + (k-1)\mathbf{1}(e_k = \texttt{coal})\right)}{\prod_{k=1}^n (k-1+\theta)}. \tag{2.6}$$

Definition 2.13 (Hoppe[1]-Urne)**.** Urne enthält eine schwarze Kugel (“Mutationskugel”) mit Masse θ, farbige Kugeln jew. mit Masse 1.

- Zu Beginn: Urne enthält nur die schwarze Kugel .
- In jedem Schritt: Ziehe eine Kugel mit W’keit proportional zu ihrer Masse.
- Falls farbige Kugel gezogen: Lege zurück zusammen mit einer weiteren Kugel derselben Farbe.
- Falls schwarze Kugel gezogen: Lege zurück zusammen mit einer weiteren Kugel einer völlig neuen Farbe.

[1]Fred M. Hoppe, Pólya-like urns and the Ewens’ sampling formula, *J. Math. Biol.* 20, no. 1, 91–94, (1984).

Lemma 2.14. *Die von den n nicht-schwarzen Kugeln erzeugte Verteilung der Familiengrößen (Typenhäufigkeitsspektrum) nach n Zügen der Hoppe-Urne entspricht der des getöteten n-Koaleszenten (aus Def. 2.11).*

Beweisskizze. Lese (2.6) „rückwärts." □

Sei

$$K_n = \#\text{verschiedene Typen in } n\text{-Stichprobe.}$$

Satz 2.15. *Im IMA-Modell ($\theta > 0$ fest) gilt für $n \to \infty$*

$$\mathbb{E}[K_n] = \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1} \sim \theta \log n, \quad \mathrm{Var}[K_n] = \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1} \cdot \frac{i-1}{\theta + i - 1} \sim \theta \log n,$$

$$\textit{und} \quad \frac{K_n - \mathbb{E}[K_n]}{\sqrt{\mathrm{Var}(K_n)}} \xrightarrow{d} \mathcal{N}(0,1).$$

Beweis. Verwende Hoppe-Urne, schreibe

$$K_n = A_1 + \cdots + A_n,$$

mit

$$A_i = \mathbf{1}(\text{im } i\text{-ten Zug wurde die schwarze Kugel gezogen}).$$

Nach Konstruktion sind $A_1, \ldots, A_n$ u.a. mit

$$\mathbb{P}(A_i = 1) = \theta/(\theta + i - 1).$$

Für die Asymptotik vergleiche mit Riemann-Integral, für asymptotische Normalität bilde ein (unabängiges, zentriertes, normiertes) Dreiecksschema

$$X_{ni} = \frac{A_i - \frac{\theta}{\theta+i-1}}{\sqrt{\mathrm{Var}[K_n]}},$$

dies erfüllt (trivialerweise) die Lindeberg-Bedingung: Es gilt für

$$S_n = X_{n1} + \cdots + X_{nn}$$

$\mathrm{Var}[S_n] = 1$, für jedes $\varepsilon > 0$ gilt wegen $\mathrm{Var}[K_n] \to \infty$, dass $\mathbb{E}[X_{ni}^2 \mathbf{1}(X_{ni}^2 > \varepsilon)] = 0$ für n genügend groß, insbesondere

$$L_n(\varepsilon) := \sum_{i=1}^{n} \mathbb{E}[X_{ni}^2 \mathbf{1}(X_{ni}^2 > \varepsilon)] \to 0.$$

□

Bemerkung 2.16 (Hoppes Urne und zufällige Permutationen[2])**.** n Züge aus der Hoppe-Urne generieren sukzessive eine zufällige Permutation Π_n von $\{1, \ldots, n\}$ (in Zyklen-Darstellung) folgendermaßen:
Nummeriere die farbigen Kugeln in der Reihenfolge des Erscheinens, wenn im k-ten Zug

[2]Eine Beobachtung aus Paul Joyce, Simon Tavaré, Cycles, permutations and the structure of the Yule process with immigration, *Stochastic Process. Appl.* 25, no. 2, 309–314, (1987).

- schwarze Kugel gezogen : füge neuen Zyklus (k) hinzu,
  (insbesondere: nach dem ersten Zug entsteht die Identität (1))
- farbige Kugel j_k gezogen : füge im Zyklus, der j_k enthält, links von j_k ein.

Für jede Permutation π mit k Zyklen gilt dann

$$\mathbb{P}(\Pi_n = \pi) = \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)},$$

denn π legt die Reihenfolge der Ereignisse i.d. Urne fest, wenn im k-ten Zug schwarze Kugel gezogen: Faktor $\frac{\theta}{\theta+k-1}$, wenn farbige gezogen: Faktor $\frac{1}{\theta+k-1}$.

Man kann dies als eine Version des sogenannten China-Restaurant-Prozesses auffassen, siehe z.B. [Kl], Kap. 24.3 (und speziell S. 523f dort).

Satz 2.17 (Ewens'sche Stichprobenformel[3])**.** *Seien $b_1, \ldots, b_n \in \mathbb{N}_0$ mit*

$$\sum_{j=1}^{n} b_j = k \leq n \quad \textit{und} \quad \sum_{j=1}^{n} j b_j = n$$

gegeben. Die Wahrscheinlichkeit, in einer n-Stichprobe (im IMA-Modell) jeweils b_j Typen mit genau j Repräsentanten (für $j = 1, \ldots, n$) zu beobachten, ist

$$\frac{n!}{1^{b_1} 2^{b_2} \cdots n^{b_n}} \cdot \frac{1}{b_1! b_2! \cdots b_n!} \cdot \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)}. \tag{2.7}$$

Man kann (2.7) auch schreiben als

$$C(n,\theta) \times \prod_{j=1}^{n} e^{-\theta/j} \frac{(\theta/j)^{b_j}}{b_j!} \tag{2.8}$$

mit

$$C(n,\theta) = n! \exp\big(\theta \sum_{j=1}^{n} 1/j\big) / (\theta(\theta+1)\cdots(\theta+n-1)),$$

d.h. die Verteilung des Typenhäufigkeitsspektrums $(B_1, \ldots, B_n)$ in einer n-Stichprobe ist $\otimes_{j=1}^{n} \mathrm{Poi}(\theta/j)$, bedingt auf $\sum_{j=1}^{n} j B_j = n$.

Beweis. Man kann dies per Induktion beweisen, indem man (analog zum Argument in Beob. 2.5) nach dem „jüngsten" Ereignis im markierten Koaleszenten zerlegt. Wir betrachten hier ein direktes, kombinatorisches Argument aus dem Artikel Robert C. Griffiths, Sabin Lessard, Ewens' sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles, *Theoretical Population Biology* 68, no. 3, 167–177, (2005).

Seien $E_n, E_{n-1}, \ldots E_1$ die „Elementarübergänge" in der Historie des getöteten Koaleszenten aus Def. 2.11 (vgl. (2.6) und Diskussion vor der Hoppe-Urne, Def. 2.13),

E_m beschreibt das Ereignis, das die Anzahl aktiver Linien von m auf $m-1$ reduziert.

[3] Warren J. Ewens, The sampling theory of selectively neutral alleles, *Theoretical Population Biology* 3, 87–112, (1972) und S. Karlin, J. McGregor, Addendum to a paper of W. Ewens, *Theoretical Population Biology* 3, 113–116, (1972).

Wir nummerieren die Linien mit $1, \dots, n$ und führen (genauer) Buch, welche Linie von einer Mutation getroffen wird bzw. welche Linie mit welcher Linie verschmilzt, mögliche Werte der Elementarübergänge sind also

$$\texttt{mut}(i), \text{ Linie } i \text{ trifft eine Mutation, und}$$

$$\texttt{koal}(i \to j), \text{ Linie } i \text{ verschmilzt in Linie } j\ (\neq i).$$

Wir denken bei den Verschmelzungereignissen an „gerichtete“ Verschmelzungen, d.h. für jedes aktuell noch aktive Paar von Linien i und j

$$\text{verschmilzt Linie } i \text{ in Linie } j \text{ mit Rate } \frac{1}{2}.$$

Eine Liste $e_n, e_{n-1}, \dots, e_1$ solcher möglicher Elementarübergänge nennen wir ein *Feinprotokoll*.

Sei für $m \geq 2$

$$p_m(e_m) = \begin{cases} \dfrac{1/2}{\frac{1}{2}m(m-1)+\frac{\theta}{2}m} = \dfrac{1}{m(m-1+\theta)}, & \text{wenn } e_m \text{ eine Verschmelzung,} \\ \dfrac{\theta/2}{\frac{1}{2}m(m-1)+\frac{\theta}{2}m} = \dfrac{\theta}{m(m-1+\theta)}, & \text{wenn } e_m \text{ eine Mutation,} \end{cases}$$

$p_1(e_1) = 1$, wenn e_1 eine Mutation ist, und $p_1(e_1) = 0$ für eine (dann unmögliche) Verschmelzung.

Für ein gegebenes mögliches Feinprotokoll $e_n, e_{n-1}, \dots, e_2, e_1$ mit k Mutationsereignissen (und somit k Typen) ist

$$\mathbb{P}(E_n = e_n, \dots, E_1 = e_1) = p_n(e_n)p_{n-1}(e_{n-1})\cdots p_1(e_1) = \frac{\theta^k}{\prod_{m=1}^{n} m(m-1+\theta)} = \frac{\theta^k}{n!(\theta)_{n\uparrow}} \quad (2.9)$$

(Produkt der Übergangswahrscheinlichkeiten der Skelettkette des getöteten Kingman-n-Koaleszenten).

Für $b_1, \dots, b_n \in \mathbb{N}_0$ mit $\sum_{j=1}^{n} b_j = k$ (und $\sum_{j=1}^{n} j b_j = n$) gibt es

$$\frac{(n!)^2}{\prod_{j=1}^{n}(b_j! j^{b_j})} \quad (2.10)$$

mögliche Feinprotokolle, die auf dieses Typenhäufigkeitsspektrum $(b_1, \dots, b_n)$ führen. Das Produkt von (2.9) und (2.10) liefert (2.7).

Zu (2.10): Stellen wir uns für den Moment die k Typen („künstlich“) nummeriert vor, mit

$$\text{Typenhäufigkeitsvektor } (n_1, n_2, \dots, n_k),$$

d.h. n_ℓ Stichproben sind vom ℓ-ten Typ, und es gilt

$$\big|\{\ell : n_\ell = j\}\big| = b_j, \quad j = 1, 2, \dots, n.$$

Es gibt

$$n! \times \binom{n}{n_1 \dots n_k} \times (n_1 - 1)!\cdots(n_k - 1)! = \frac{(n!)^2}{\prod_{\ell=1}^{k} n_\ell} = \frac{(n!)^2}{\prod_{j=1}^{n} j^{b_j}} \quad (2.11)$$

verschiedene Feinprotokolle mit k nummerierten Typen, die auf diesen Typenhäufigkeitsvektor $(n_1, n_2, \dots, n_k)$ führen:

1. $n!$ Möglichkeiten für die Reihenfolge, in der die Linien inaktiv werden,
2. $\binom{n}{n_1 \dots n_k}$ Möglichkeiten, die n Linien auf die k Typen aufzuteilen
   (n nummerierte Kugeln in k Schachteln legen),
3. für $\ell = 1, \dots, k$ gibt es $(n_\ell - 1)!$ viele Arten, innerhalb von Typ ℓ die „Verschmelzungsziele“ festzulegen.

   (Nehmen wir an, wir haben in Schritt 1 und 2 festgelegt, dass Linien $i_1, i_2, \dots, i_{n_\ell}$ vom Typ ℓ sind und dass diese in der Reihenfolge $i_1, i_2, \dots$ inaktiviert werden. Dann gibt es $n_\ell - 1$ viele Wahlen für das Verschmelzungsziel von i_1, $n_\ell - 2$ viele Wahlen für das Verschmelzungsziel von i_2, u.s.w.)

Schließlich führen

$$b_1! \cdot b_2! \cdot \dots \cdot b_n! \tag{2.12}$$

verschiedene Feinprotokolle mit nummerierten Typen auf dasselbe Feinprotokoll ohne Typennummerierung:

$$\text{Typen } \ell \text{ und } \ell' \text{ können vertauscht werden, sofern } n_\ell = n_{\ell'}. \tag{2.13}$$

Der Quotient von (2.11) und (2.12) liefert (2.10). □

Bemerkung 2.18. Nach Satz 2.15 ist

$$\widehat{\theta}_{\text{naiv}} := \frac{K_n}{\log n}$$

ein (schwach) konsistenter Schätzer für die Mutationsrate θ, d.h. für jedes $\theta \in (0, \infty)$ gilt

$$\widehat{\theta}_{\text{naiv}} \underset{n\to\infty}{\to} \theta \quad \text{stochastisch bzw. in Verteilung.}$$

Satz 2.15 liefert auch asymptotische Normalität von $\widehat{\theta}_{\text{naiv}}$. Allerdings ist

$$\operatorname{Var}_\theta[\widehat{\theta}_{\text{naiv}}] \sim \frac{\theta}{\log n}.$$

(Dies ist allerdings „deprimierend langsam“: z.B. müsste $n = e^{100} \approx 2.7 \cdot 10^{43}$ sein, damit die Streuung $\approx 0.1\sqrt{\theta}$ ist.)

Beobachtung 2.19. Im IMA-Modell ist K_n suffizient für θ, d.h. gegeben $K_n = k$ hängt die Verteilung der beobachteten Typen nicht von θ ab.

Beweis. Sei

$$C_{n,k} := \sum\nolimits'_{(b_1,\dots,b_n)} \frac{n!}{\prod_{i=1}^n i^{b_i}\, b_i!},$$

($\sum'$ bezeichnet die Summe über $(b_1, \dots, b_n) \in \mathbb{N}_0^n$ mit $\sum b_i = k$, $\sum i b_i = n$).

Satz 2.17 (Ewens-Formel) liefert

$$\mathbb{P}_\theta(K_n = k) = C_{k,n}\frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)},$$

für $b_1, \ldots, b_n$ mit $b_1 + \cdots + b_n = k$ (und $\sum ib_i = n$) also

$$\mathbb{P}_\theta(B_1 = b_1, \ldots, B_n = b_n | K_n = k) = \frac{1}{C_{k,n}}\frac{n!}{1^{b_1}2^{b_2}\cdots n^{b_n}} \cdot \frac{1}{b_1!b_2!\cdots b_n!}.$$

□

Sei $K_n = k$ beobachtet. Der Maximum-Likelihood-Schätzer $\widehat{\theta}_{\mathrm{ML}}$ ist dasjenige $\theta \geq 0$, das die Likelihood

$$L_n(\theta, k) = \mathbb{P}_\theta(K_n = k) = C_{k,n}\frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)}$$

(als Funktion von θ) maximiert.

Es ist

$$\frac{\partial}{\partial\theta}\log L_n(\theta, k) = \frac{\partial}{\partial\theta}\left(k\log\theta - \sum_{i=0}^{n-1}\log(\theta+i)\right) = \frac{k}{\theta} - \sum_{i=0}^{n-1}\frac{1}{\theta+i} = \frac{1}{\theta}\left(k - \sum_{i=0}^{n-1}\frac{\theta}{\theta+i}\right),$$

also ist $\widehat{\theta}_{\mathrm{ML}}$ Lösung von

$$k = \sum_{i=0}^{n-1}\frac{\widehat{\theta}_{\mathrm{ML}}}{\widehat{\theta}_{\mathrm{ML}}+i} \quad \left(\text{„} = \mathbb{E}_{\widehat{\theta}_{\mathrm{ML}}}[K_n]\text{“}\right)$$

(d.h. derjenige θ-Wert, unter dem erwartet=beobachtet).

Die Fisher-Information ist

$$I(\theta) \;=\; \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log L_n(\theta, K_n)\right)^2\right] = \frac{1}{\theta^2}\mathbb{E}_\theta\left[\left(K_n - \sum_{i=0}^{n-1}\frac{\theta}{\theta+i}\right)^2\right] = \frac{1}{\theta^2}\mathrm{Var}_\theta(K_n)$$

($\sim \frac{1}{\theta}\log n$ nach Satz 2.15).

2.2.1 Die GEM-Verteilung

Betrachte die Hoppe-Urne (für festes n alternativ: den Koaleszent mit Mutationen gemäß IMA-Modell), die Typen/Familien seien in Reihenfolge des Erscheinens nummeriert (“age order”)

$$X_k(n) := \text{Größe der } k\text{-ten Familie nach dem } n\text{-ten Zug aus der Hoppe-Urne,}$$

(offenbar $X_1(1) = 1$, $X_k(n) = 0$ für $k > n$).

Frage:

$$\left(\frac{1}{n}X_1(n), \frac{1}{n}X_2(n), \frac{1}{n}X_3(n), \ldots\right) \underset{n\to\infty}{\longrightarrow} ? \tag{2.14}$$

Beobachtung: $(n+\theta)^{-1}X_1(n)$, $n = 2,3,\dots$, ist ein (beschränktes) Martingal:

$$\mathbb{E}\Big[\frac{1}{n+1+\theta}X_1(n+1)\,\Big|\,\mathcal{F}_n\Big] = \frac{X_1(n)+1}{n+1+\theta}\frac{X_1(n)}{n+\theta} + \frac{X_1(n)}{n+1+\theta}\frac{\theta+n-X_1(n)}{n+\theta} = \frac{X_1(n)}{n+\theta}$$

(mit $\mathcal{F}_n = \sigma$(Beobachtungen bis einschließlich n-tem Zug)), konvergiert also f.s. Analog ist für $k \geq 2$ mit $\alpha_k =$ Zeitpunkt des ersten Auftretens von Typ k

$$(n+\alpha_k+\theta)^{-1}X_k(n+\alpha_k), \quad n = 1,2,\dots$$

ein (beschr.) Martingal.

Demnach: (2.14) konvergiert f.s. (zumindest koordinaten-weise).

Satz 2.20 (GEM-Verteilung[4])**.** *Seien $B_1, B_2, \dots$ u.i.v.* Beta$(1,\theta)$*, d.h. sie besitzen die Dichte $\theta(1-b)^{\theta-1}$ auf $[0,1]$. Die Verteilung des Grenzwerts in (2.14) ist gegeben durch*

$$\Big(B_1, (1-B_1)B_2, (1-B_1)(1-B_2)B_3, (1-B_1)(1-B_2)(1-B_3)B_4, \dots\Big).$$

Definition 2.21 (Yule[5]-Prozess)**.** Ein zeitkontinuierlicher reiner Geburtsprozess (jedes Individuum erzeugt u.a. mit Rate 1 ein neues Individuum) heißt ein Yule-Prozess.

Es handelt sich also um eine zeitkontinuierliche Markovkette auf $\mathbb{N}$ mit Sprungraten

$$q_{n,n+1} = n = -q_{n,n},\ n \in \mathbb{N} \quad (q_{n,m} = 0 \text{ für } m \neq n, n+1).$$

Es ist [auch] ein Spezialfall eines zeitkont. (Galton-Watson-)Verzweigungsprozesses.

Lemma 2.22. *Sei $(Y_t)_{t\geq 0}$ ein Yule-Prozess (mit Geburtsrate 1 pro Individuum) und Startwert $Y_0 = 1$.*

Es gilt $\mathscr{L}(Y_t) =$ geom(e^{-t}) *und $(e^{-t}Y_t)_{t\geq 0}$ ist ein L^2-beschränktes Martingal mit*

$$\lim_{t\to\infty} e^{-t}Y_t \overset{d}{=} \text{Exp}(1).$$

Beweis. Sei

$$T_i := |\{t : Y_t = i\}| \text{ die Länge des Zeitintervalls, in dem } i \text{ Ind. leben.}$$

Die Form der Raten zeigt:

$$T_1, T_2, \dots \text{ sind u.a.}, T_i \sim \text{Exp}(i).$$

Somit

$$\mathbb{P}(Y_t > n) \;=\; \mathbb{P}(T_1 + \cdots + T_n < t) = \mathbb{P}\Big(\max_{i=1,\dots,n} \tau_i < t\Big) = (1-e^{-t})^n, \quad n = 0,1,2,\dots$$

mit τ_i u.i.v. Exp(1), d.h. $\mathscr{L}(Y_t) =$ Geom(e^{-t}).

[4]Nach Bob Griffiths, Steinar Engen und John William Thomas Mccloskey benannt

[5]nach George Udny Yule, 1871–1951

(Alternativ beachte, dass die Lösung der Vorwärtsgleichung

$$\frac{\partial}{\partial t}\mathbb{P}_1(Y_t = n) = (n-1)\mathbb{P}_1(Y_t = n-1) - n\mathbb{P}_1(Y_t = n),\ \ \mathbb{P}_1(Y_0 = n) = \delta_{1n}$$

gegeben ist durch $\mathbb{P}_1(Y_t = n) = e^{-t}(1-e^{-t})^{n-1}$)

Zusammen mit der Verzweigungseigenschaft

$$\mathscr{L}(Y_t|Y_0 = k+j) = \mathscr{L}(Y_t|Y_0 = k) * \mathscr{L}(Y_t|Y_0 = j)$$

folgt: $\mathbb{E}[Y_{t+h}|Y_t] = e^h Y_t$, d.h. $(e^{-t}Y_t)_{t\geq 0}$ ist Martingal.

Weiter folgt leicht: $\sup_{t\geq 0}\mathbb{E}[(e^{-t}Y_t)^2] < \infty$ und $e^{-t}Y_t \to^d \mathrm{Exp}(1)$. □

Lemma 2.23. *Seien G_1 und G_2 u.a., $G_i \sim \mathrm{Gamma}(\theta_i)$ (d.h. Dichte $(\Gamma(\theta_i))^{-1}g^{\theta_i-1}e^{-g}$ auf $\mathbb{R}_+$). Dann ist*

$$\mathscr{L}\Big(G_1+G_2, \frac{G_1}{G_1+G_2}\Big) = \mathrm{Gamma}(\theta_1+\theta_2)\otimes\mathrm{Beta}(\theta_1,\theta_2).$$

Beweis. $G := G_1 + G_2$ ($G \sim \mathrm{Gamma}(\theta_1+\theta_2)$). Die gemeinsame Dichte von (G_1, G) ist

$$f_{(G_1,G)}(g_1,g) = c\mathbf{1}(0\leq g_1\leq g)g_1^{\theta_1-1}e^{-g_1}(g-g_1)^{\theta_2-1}e^{-(g-g_1)} = ce^{-g}\mathbf{1}(0\leq g_1\leq g)g_1^{\theta_1-1}(g-g_1)^{\theta_2-1},$$

demnach ist die bedingte Dichte von G_1, gegeben $G = g$

$$f_{G_1|G=g}(g_1) = c(g)\mathbf{1}(0\leq g_1\leq g)g_1^{\theta_1-1}(g-g_1)^{\theta_2-1},$$

und somit die bedingte Dichte von G_1/G, gegeben $G = g$

$$f_{(G_1/G)|G=g}(b) = \tilde{c}(g)\mathbf{1}(0\leq b\leq 1)b^{\theta_1-1}(1-b)^{\theta_2-1}.$$

Da $\int_0^1 f_{(G_1/G)|G=g}(b)\,db = 1$ gilt, muss

$$\tilde{c}(g) = \Gamma(\theta_1+\theta_2)/(\Gamma(\theta_1)\Gamma(\theta_2))$$

für jedes $g > 0$ gelten. □

Beweis von Satz 2.20. Darstellung via Yule-Prozess mit Immigration:

Seien $0 < T_1 < T_2 < \cdots$ die Sprungzeitpunkte eines Poissonprozesses auf $[0,\infty)$ mit Rate θ.

Der i-te Immigrant erscheint zum Zeitpunkt T_i, gründet i-te Familie, diese wächst ab dann als Yule-Prozess (vgl. Def. 2.21) unabhängig von allen anderen.

`[Bild an der Tafel]`

Seien

$Z_i(t) :=$ Größe der i-ten Familie zur Zeit t (wir setzen $Z_i(t) = 0$ für $t < T_i$, $Z_i(T_i) = 1$),

$S(t) := \sum_{i=1}^{\infty} Z_i(t)$ die Gesamtgröße der Population zur Zeit t,

$\tau_n := \min\{t : S(t) = n\}$ der Zeitpunkt, zu dem die Population auf n anwächst.

Es gilt

$$\Big(\frac{1}{n}Z_1(\tau_n), \frac{1}{n}Z_2(\tau_n), \frac{1}{n}Z_3(\tau_n), \dots\Big)_{n=1,2,\dots} \overset{d}{=} \Big(\frac{1}{n}X_1(n), \frac{1}{n}X_2(n), \frac{1}{n}X_3(n), \dots\Big)_{n=1,2,\dots} \tag{2.15}$$

Dazu Vergleich der Sprungraten: Es gebe aktuell

$$k \text{ Familien d. Größen } j_1, j_2, \dots, j_k \text{ mit } j_1 + \dots + j_k = n.$$

- $S(t)$ springt nach $n+1$ mit Rate $n+\theta$,
- der Zuwachs
  - betrifft i-te Familie mit W'keit $j_i/(n+\theta)$,
  - erzeugt neue Familie mit W'keit $\theta/(n+\theta)$.

Also: Skelettkette des Yule-Prozesses mit Immigration $\widehat{=}$ Hoppe-Urne.

Lemma 2.22 zeigt:

$$\Big(e^{-t}Z_1(t), e^{-t}Z_2(t), e^{-t}Z_3(t), \dots\Big) \to \Big(e^{-T_1}A_1, e^{-T_2}A_2, e^{-T_3}A_3, \dots\Big) \quad \text{f.s.}, \tag{2.16}$$

(koordinaten-weise) mit $A_1, A_2, \dots$ u.i.v. Exp(1).

Daraus folgt

$$e^{-t}S(t) \to \sum_{n=1}^{\infty} e^{-T_n}A_n \quad \text{f.s.} \tag{2.17}$$

(Zur Rechtfertigung der Grenzwertvertauschung beachte, dass $M_i := \sup_{t\geq 0} e^{-t}Z_i(T_i+t)$ u.i.v. sind mit $\mathbb{E}M_1 < \infty$, also $\limsup M_n/n = 0$ gemäß Borel-Cantelli-Lemma ($\sum_{n=1}^{\infty}\mathbb{P}(M_i > \epsilon n) < \infty$ für jedes $\epsilon > 0$).

Weiterhin gilt $T_n/n \to \theta^{-1}$ f.s. gem. dem starken Gesetz der großen Zahlen, somit ist für $m \geq N_0$

$$\sup_{t\geq 0}\sum_{n=m}^{\infty} e^{-T_n}e^{-(t-T_n)}Z_n(t) \leq \sum_{n=m}^{\infty} e^{-T_n}M_n \leq \sum_{n=m}^{\infty} ne^{-2n/\theta}. \quad)$$

Weiterhin ist

$$\mathscr{L}\Big(\sum_{n=1}^{\infty} e^{-T_n}A_n\Big) = \text{Gamma}(\theta), \tag{2.18}$$

denn $\sum_i \delta_{(A_i,T_i)}$ ist ein Poissonscher Punktprozess auf $\mathbb{R}_+ \times \mathbb{R}_+$ mit Intensitätsmaß $\theta dt \otimes e^{-x}dx$ (und dies ist das Lévy-Maß des Gamma-Prozess/der Gamma-Verteilung, siehe z.B. Klenke [Kl], Bsp. 16.15)

Für die Form des Intensitätsmaßes: sei $h : (0,\infty) \to \mathbb{R}_+$, sagen wir, stetig mit kompaktem Träger, so ist

$$\int_0^\infty \int_0^\infty h(e^{-t}a)\,\theta dt\, e^{-a}da = \int_0^\infty \int_0^a h(r)\theta\frac{dr}{r}e^{-a}da = \int_0^\infty h(r)\int_r^\infty e^{-a}\,da\,\theta\frac{dr}{r} = \int_0^\infty h(r)\theta e^{-r}\frac{dr}{r}.$$

(Die allgemeine Beobachtung dahinter ist folgende: Wenn $\Pi = \sum\delta_{a_i}$ ein PPP auf E mit Intensitätsmaß ν ist und $f : E \to E'$, dann ist $\tilde{\Pi} = \sum\delta_{f(a_i)}$ ein PPP auf E' mit Intensitätsmaß $\tilde{\nu} = \nu \circ f^{-1}$.)

Das Argument für (2.18) zeigt auch

$$\mathscr{L}(G,T) = \text{Gamma}(1+\theta) \otimes \text{Exp}(1) \;\Rightarrow\; \mathscr{L}(e^{-T/\theta}G) = \text{Gamma}(\theta), \tag{2.19}$$

denn $\sum_{n=1}^{\infty} e^{-T_n} A_n = e^{-T_1}\left(A_1 + \sum_{n=2}^{\infty} e^{-(T_n - T_1)} A_n\right)$.

(Alternativ beachte man, dass $e^{-T/\theta} \sim \text{Beta}(\theta, 1)$ gilt und verwende Lemma 2.23.)

Somit gilt

$$\frac{Z_1(t)}{S(t)} = \frac{e^{T_1 - t} Z_1(t)}{e^{T_1 - t} Z_1(t) + \sum_{i=2}^{\infty} e^{T_1 - t} Z_i(t)} \to \frac{A_1}{A_1 + \sum_{i=2}^{\infty} e^{-(T_i - T_1)} A_i} =: B_1 \quad \text{f.s.},$$

und $\mathscr{L}(B_1) = \text{Beta}(1, \theta)$, wobei B_1 und $A_1 + \sum_{i=2}^{\infty} e^{-(T_i - T_1)} A_i$ u.a. (Lemma 2.23).

Sei

$$C_n := A_n + \sum_{i=n+1}^{\infty} e^{-(T_i - T_n)} A_i, \quad B_n := \frac{A_n}{C_n}.$$

Zeige induktiv:

$$\mathscr{L}\Big(C_1, B_1, B_2, \ldots, B_n\Big) = \text{Gamma}(1+\theta) \otimes \text{Beta}(1,\theta)^{\otimes n} \quad \text{für } n \in \mathbb{N}. \tag{2.20}$$

Der Fall $n = 1$ stimmt nach obigem.

Für den Schluss von $n \to n+1$: I.V. und Stationarität sowie Unabhängigkeit der Poisson-Zuwächse liefern

$$\mathscr{L}(C_2, B_2, B_3, \ldots, B_{n+1}) = \text{Gamma}(1+\theta) \otimes \text{Beta}(1,\theta)^{\otimes n};$$

zudem sind $(C_2, B_2, B_3, \ldots, B_{n+1})$ und $(A_1, T_2 - T_1)$ unabhängig.

Nach Def. ist

$$C_1 = A_1 + e^{-(T_2 - T_1)} C_2, \; B_1 = \frac{A_1}{C_1} = \frac{A_1}{A_1 + e^{-(T_2 - T_1)} C_2}.$$

Es ist $e^{-(T_2 - T_1)} C_2 \sim \text{Gamma}(\theta)$ nach (2.19) und $A_1 \sim \text{Exp}(1)$ u.a. von $e^{-(T_2 - T_1)} C_2$, d.h.

$$(C_1, B_1) \sim \text{Gamma}(\theta) \otimes \text{Beta}(1, \theta)$$

nach Lemma 2.23, dies liefert den Induktionsssschluss.

Schließlich gilt

$$\begin{aligned}
\frac{e^{-t} Z_n(t)}{e^{-t} S(t)} \quad \to \quad & \frac{e^{-T_n} A_n}{\sum_{i=1}^{\infty} e^{-T_i} A_i} = \frac{\sum_{i=2}^{\infty} e^{-T_i} A_i}{\sum_{i=1}^{\infty} e^{-T_i} A_i} \times \cdots \times \frac{\sum_{i=n}^{\infty} e^{-T_i} A_i}{\sum_{i=n-1}^{\infty} e^{-T_i} A_i} \times \frac{e^{-T_n} A_n}{\sum_{i=n}^{\infty} e^{-T_i} A_i} \\
= \quad & (1 - B_1) \times \cdots \times (1 - B_n) \times \frac{A_n}{A_n + \sum_{i=n+1}^{\infty} e^{-(T_i - T_n)} A_i} \\
= \quad & (1 - B_1) \times \cdots \times (1 - B_{n-1}) B_n.
\end{aligned}$$

□

Beobachtung 2.24 (Poisson-Dirichlet-Verteilung)**.** Seien $1 \geq V_1 > V_2 > \cdots$ die (der Größe) nach sortierten Einträge (=Typenhäufigkeiten) aus GEM-verteiltem

$$\Big(B_1, (1-B_1)B_2, (1-B_1)(1-B_2)B_3, (1-B_1)(1-B_2)(1-B_3)B_4, \dots\Big).$$

Sei $\Pi = \sum_i \delta_{X_i}$ PPP auf $\mathbb{R}_+$ mit Intensitätsmaß $(\theta/x)e^{-x}dx$ (Π beschreibt die Sprünge eines Standard-Gamma-Subordinators bis zur Zeit θ) und $S := \sum X_i$, seien $X_{[1]} > X_{[2]} > \cdots$ die Ordnungsstatistik der X_is. Dann ist

$$\big(X_{[1]}/S, X_{[2]}/S, \dots\big) \stackrel{d}{=} \big(V_1, V_2, \dots\big).$$

Dies folgt aus dem Beweis von Satz 2.20. Diese Verteilung (ein W'maß auf $\{(x_1, x_2, \dots) \in [0,1]^{\mathbb{N}} : x_1 + x_2 + \cdots = 1\}$), heißt die Poisson-Dirichlet-Verteilung (mit Parameter θ).

Bericht 2.25 (Endlich viele Typen mit elternunabhängiger Mutation)**.** Betrachte Mutationsmodell mit d neutralen Typen (Typenmenge $E = \{1, \dots, d\}$), jede Linie mutiert mit Rate $\theta/2$, Typ nach Mutation ist j mit W'keit $\pi_j\,(>0)$ ($(\pi_1, \dots, \pi_d)$ Ws-Gewichte auf $\{1, \dots, d\}$), u.a. vom vorigen Typ.

Die Typenverteilung in einer unendlichen Population im Gleichgewicht ist dann

$$\mathscr{L}\Big(Z_1(\infty), \dots, Z_d(\infty)\Big) \;=\; \text{Dirichlet}(\theta\pi_1, \dots, \theta\pi_d), \tag{2.21}$$

d.h. die Dichte ist

$$\frac{\Gamma(\theta)}{\Gamma(\theta\pi_1)\cdots\Gamma(\theta\pi_d)} x_1^{\theta\pi_1 - 1} \cdots x_k^{\theta\pi_d - 1}$$

bezüglich dem Lebesgue-Maß auf $\{(x_1, \dots, x_d) : 0 \leq x_i \leq 1, x_1 + \cdots + x_d = 1\}$.

Im Fall von $d = 2$ Typen hatten wir dies bereits in Beob. 2.5 gesehen. Man kann den allgemeinen Fall aus Beob. 2.24 herleiten: Wenn man jeden Sprung des Gamma-Subordinators unabhängig gemäß π mit einer „Farbe“ aus $\{1, \dots, d\}$ einfärbt, so bilden die Sprünge jeder Farbe für sich jeweils unabhängige Gamma-Subordinatoren (mit entsprechend verkleinerter Intensität), somit: $Y_i \sim \text{Gamma}(\theta\pi_i)$ und $Y_1, \dots, Y_d$ unabhängig, so ist

$$\Big(Z_1(\infty), \dots, Z_d(\infty)\Big) \stackrel{d}{=} \left(\frac{Y_1}{Y_1 + \cdots + Y_k}, \dots, \frac{Y_k}{Y_1 + \cdots + Y_k}\right)$$

und die rechte Seite ist Dirichlet$(\theta\pi_1, \dots, \theta\pi_d)$-verteilt (dies ist eine multivariate Verallgemeinerung von Lemma 2.23, siehe z.B. Ch. 40.5 in Norman L. Johnson, Samuel Kotz, *Distributions in statistics: continuous multivariate distributions*, Wiley, 1972).

2.3 Infinitely-many-sites-Modell (IMS)

Definition 2.26 (Infinitely-many-sites-Modell[6])**.** Man nimmt an, dass jede Mutation eine neue, bisher noch nie mutierte Position am betrachteten Lokus betrifft.

[6]Eingeführt in G. A. Watterson, On the number of segregating sites in genetical models without recombination, Theoretical Population Biology 7 (2), 256–276, (1975).

Mathematisch realisiert man dies z.B. folgendermaßen: Die betrachtete Stelle im Genom (eine gewisse Abfolge von Nukleotiden im DNS-Doppelstrang eines Chromosoms) entspricht $[0,1]$, jede Mutation erhält eine neue, uniform aus $[0,1]$ gewählte „Position“, der Typ eines Individuums ist ein (einfaches) Zählmaß auf $[0,1]$ (bzw. alternativ eine Teilmenge von $[0,1]$), der Typ eines Individuums gibt an, wo dieses relativ zu einen „Referenztyp“ (oder „Wildtyp“) mutiert ist.

Das infinitely-many-sites-Modell (IMS-Modell, in der Literatur auch infinite-sites-Modell genannt) ist für viele praktische Zwecke eine angemessene Approximation für die Beschreibung von Mutationen auf dem Niveau der DNS-Sequenz : Wenn die Mutationsrate pro Basenpaar sehr klein und die betrachtete Stelle im Genom (der sog. Lokus) nicht „zu lang“ ist, ist es plausibel, die Möglichkeit der Mehrfachmutation einer Stelle (und andere Effekte, die im IMS-Modell nicht berücksichtigt werden, etwa Rekombination oder Insertionen/Deletionen längerer Stücke im Genom) zu vernachlässigen. (Man denke an eine Abfolge von $L \gg 1$ Basenpaaren, bei Mutation wird eine der zufällig gewählte der L Positionen zufällig modifiziert.)

Beispiel 2.27. John Parsch, Colin D. Meiklejohn, and Daniel L. Hartl, Patterns of DNA sequence variation suggest the recent action of positive selection in the janus-ocnus region of drosophila simulans, *Genetics* 159:647–657, (2001) berichten (u.a.) genetische Variabilität in einem ca. 1.700 Basenpaare langen Stück des Chromosoms 3 in einer (weltweiten) Stichprobe von 8 *Drosophila simulans* und einer Stichprobe von *Drosophila melanogaster*, zwei verwandten Arten von Taufliegen. An insgesamt 31 Stellen sind Unterschiede zwischen den Individuen sichtbar (siehe Tabelle 2.1, es sind nur die sogenannten variablen oder segregierenden Positionen aufgeführt).

Die Sequenzinformation von *Drosophila melanogaster* — diese Stichprobe bildet bezüglich der 8 Stichproben von *Drosophila simulans* eine „outgroup“ — gestattet (zusammen mit den IMS-Modellannahmen) an jeder Position zu entscheiden, welche Base die anzestrale und welche die mutierte ist.

Zur Beschreibung von beobachteten Sequenzdaten in einer Stichprobe der Größe n im Kontext des IMS-Modells betrachten wir folgende Modellvorstellung: Die n-Stichprobe entsteht aus einem n-Koaleszent, längs dessen Ästen sich mit Rate $\frac{\theta}{2}$ Mutationen ereignen (und jede trifft eine völlig neue Position).

Die Anzahl segregierender Stellen ist

$$S_n = \#\text{ verschiedene Mutationen, die in } n\text{-Stichprobe vorkommen}$$

(im Sinne von: Positionen, an denen sich mindestens zwei Stichproben unterscheiden). Wenn $S_n = s$, so entsprechen die Beobachtungen einer $n \times s$-Datenmatrix $(D_{ik})_{i=1,\dots,n;k=1,\dots,s}$

$$D_{ik} = \mathbf{1}(\text{Stichprobe } i \text{ ist an } k\text{-ter segregierender Stelle mutiert}).$$

Beispielsweise sehen wir in Abbildung 2.1 eine Realisierung mit $S_4 = 3$.

Bemerkung 2.28 (Unbekannter Wildtyp)**.** Wenn man „nur“ die Stichprobe sieht und keine externen Zusatzinformationen (z.B. eine „outgroup“ durch inter-Spezies-Vergleich wie in Bsp. 2.27) besitzt, kann man an den segregierenden Stellen nicht entscheiden,

	Position																													
	3	8	9	2 4	2 9	2 9	3 6	5 3	5 4	5 8	6 4	6 5	6 5	6 5	6 5	6 6	6 6	7 0	8 4	1 3	1 3	1 2	1 3	1 3	1 3	1 4	1 6	1 6	1 6	1 6
	5	3	3	6	1	4	2	8	9	5	4	1	2	7	8	5	7	7	3	2	9	1	0	2	8	2	4	3	1	2	4
s1	c	g	a	t	c	c	a	a	t	a	t	a	a	a	g	c	t	c	g	a	t	a	a	g	c	c	g	a	t	t	c
s2	.	.	c	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	g	.	.	.	.	.
s3	a	c	.	c	a	t	g	c	c	c	g	g	g	g	a	t	c	t	a	t	c	c	t	c	t	g	t	t	g	c	a
s4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
s5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	g	.	.	.	.	.
s6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
s7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	g	.	.	.	.	.
s8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	g	.	.	.	.	.
m1	.	.	.	c	a	t	g	c	c	c	a	g	.	.	.	.	.	t	.	.	c	c	.	.	t	g	.	t	g	c	a

Tabelle 2.1: Beobachtete genetische Variabilität in einer Region in Chromosom 3 aus einer Stichprobe von 8 *Drosophila simulans* (Zeilen s1–s8) und einer Stichprobe von *Drosophila melanogaster* (Zeile m1) aus Parsch et al, *Genetics* 159:647–657, (2001). Siehe Figure 2 dort, wir betrachten hier nur den Teil der Sequenz, der die Gene *janA* und *janB* umfasst.

welcher Typ der Wildtyp und welcher die Mutante ist (im Genetik-Jargon: die Mutationen sind „unpolarisiert“). In dieser Situation ist obige Datenmatrix nur bis auf „Umklappen“ von Spalten definiert, d.h. die eigentliche Information ist S_n und

$$\Delta_{i,j}(k) = \begin{cases} 1, & \text{Stichproben } i \text{ und } j \text{ an } k\text{-ter segregierender Stelle verschieden,} \\ 0, & \text{sonst} \end{cases}$$

für $k = 1, \dots, S_n$.

Beobachtung 2.29. Es gilt

$$\mathbb{E}_\theta[S_n] = \theta\, h_n, \quad \operatorname{Var}_\theta[S_n] = \theta\, h_n + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2}$$

mit $h_n = \sum_{i=1}^{n-1} \frac{1}{i}$.

$$\widehat{\theta}_W := \frac{S_n}{h_n}$$

ist ein erwartungstreuer Schätzer für θ (der sogenannte Watterson-Schätzer) mit

$$\operatorname{Var}_\theta\left[\widehat{\theta}_W\right] \sim \frac{\theta}{\log n} \quad \text{und} \quad \frac{\widehat{\theta}_W - \theta}{\sqrt{\operatorname{Var}_\theta\left[\widehat{\theta}_W\right]}} \xrightarrow{d} \mathcal{N}(0,1) \qquad \text{für } n \to \infty.$$

Beweis. Schreibe

$$S_n = S_{n,n} + S_{n,n-1} + \cdots + S_{n,2}$$

mit

$$S_{n,j} = \#\text{ Mutationen, während die Genealogie aus } j \text{ Linien besteht.}$$

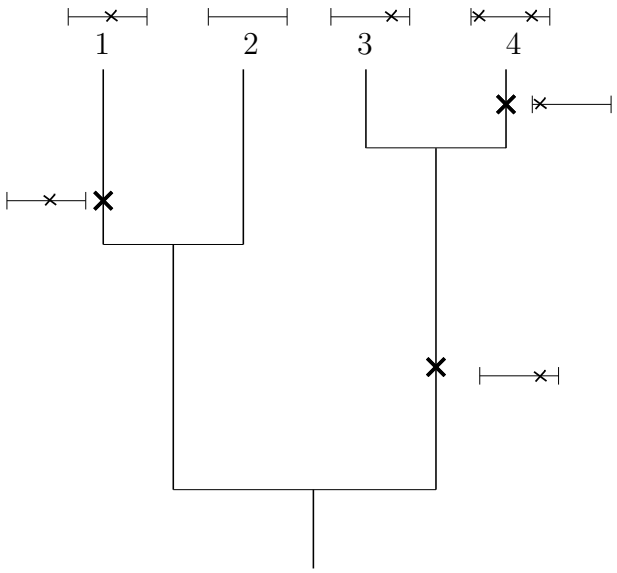


Abbildung 2.1: Ein 4-Koaleszent, längs dessen Kanten Mutationen gemäß IMS-Modell auftreten. Wir registrieren für jede Mutation die mutierte Position (in $[0,1]$) und an den Blättern (den Stichproben) sämtliche Mutationen, die sich auf dem kürzesten Weg vom jeweiligen Blatt zur Wurzel befinden.

Wir hatten in der Diskussion in Kapitel 1.1.1 gesehen, dass $S_{n,j} \sim \mathrm{geom}(\frac{j-1}{\theta+j-1})$ (denn gegeben T_j, die Zeit, während der j Linien in der Genealogie existieren, ist $S_{n,j} \sim \mathrm{Pois}(\frac{\theta}{2} j T_j)$) und $S_{n,n}, \dots, S_{n,2}$ sind unabhängig. Somit

$$\mathbb{E}_\theta[S_{n,j}] = \tfrac{\theta+j-1}{j-1} - 1 = \theta/(j-1), \quad \mathrm{Var}_\theta[S_{n,j}] = \big(\tfrac{\theta}{\theta+j-1}\big)/\big(\tfrac{j-1}{\theta+j-1}\big)^2 = \frac{\theta(\theta+j-1)}{(j-1)^2} = \frac{\theta}{j-1} + \frac{\theta^2}{(j-1)^2},$$

was die Formeln für Erwartungswert und Varianz von S_n beweist.

Zur asymptotischen Normalität von $\widehat{\theta}_W$: Schreibe

$$X_{n,j} := \frac{S_{n,j} - \theta/(j-1)}{\sqrt{\mathrm{Var}_\theta\big[\widehat{\theta}_W\big]}}, \quad j = 2, 3, \dots, n.$$

Die $X_{n,j}$ bilden ein unabhängiges, zentriertes und normiertes Dreiecksschema (für festes n sind $X_{n,2}, \dots, X_{n,n}$ unabhängig mit $\mathbb{E}_\theta[X_{n,j}] = 0$ und $\sum_{j=2}^n \mathrm{Var}_\theta[X_{n,j}] = 1$), für $\varepsilon > 0$ und n so groß, dass $\varepsilon \mathrm{Var}_\theta\big[\widehat{\theta}_W\big] > \theta^2$ gilt, ist

$$\begin{aligned}\mathbb{E}\big[X_{n,j}^2 \mathbf{1}(X_{n,j}^2 > \varepsilon)\big] &= \frac{1}{\mathrm{Var}_\theta\big[\widehat{\theta}_W\big]} \mathbb{E}\Big[\big(S_{n,j} - \theta/(j-1)\big)^2 \mathbf{1}\big(S_{n,j} > \theta/(j-1) + \sqrt{\varepsilon \mathrm{Var}_\theta\big[\widehat{\theta}_W\big]}\big)\Big] \\ &\le \frac{1}{\mathrm{Var}_\theta\big[\widehat{\theta}_W\big]} \mathbb{E}\Big[S_{n,j}^2 \mathbf{1}\big(S_{n,j} > \theta/(j-1) + \sqrt{\varepsilon \mathrm{Var}_\theta\big[\widehat{\theta}_W\big]}\big)\Big].\end{aligned}$$

Demnach erfüllt das Schema die Lindeberg-Bedingung

$$\lim_{n\to\infty} \mathbb{E}\big[X_{n,j}^2 \mathbf{1}(X_{n,j}^2 > \varepsilon)\big] = 0 \tag{2.22}$$

und daher ist das renormierte $\widehat{\theta}_W$ asymptotisch normalverteilt (siehe z.B. [Kl, Satz 15.43] oder [St2, Satz 4.2]).

(beachte: Für $X \sim \mathrm{geom}(p)$ gilt

$$\begin{aligned}
\mathbb{E}\big[X^2\mathbf{1}(X \geq m)\big] &= \sum_{k=m}^{\infty} k^2 p(1-p)^{k-1} \leq \sum_{k=m}^{\infty} (k+1)kp(1-p)^{k-1} = p\sum_{k=m}^{\infty}\Big[\frac{d^2}{dy^2}(1-y)^{k+1}\Big]_{y=p} \\
&= p\Big[\frac{d^2}{dy^2}\sum_{k=m}^{\infty}(1-y)^{k+1}\Big]_{y=p} = p\Big[\frac{d^2}{dy^2}\frac{(1-y)^{m+1}}{y}\Big]_{y=p} \\
&= p\,\frac{(1-p)^{m-1}}{p}\Big((m+1)m + 2(m+1)\frac{1-p}{p} + 2\frac{(1-p)^2}{p^2}\Big) \\
&\leq \big((m+1)(m+2)+2\big)\frac{(1-p)^{m-1}}{p^2},
\end{aligned}$$

demnach für $X = S_{n,j}$ mit $p = p_j = \frac{j-1}{\theta+j-1}$ und z.B. $m = \sqrt{\frac{1}{2}\varepsilon\theta\log n}$ ist

$$\frac{1}{\mathrm{Var}_\theta\big[\widehat{\theta}_W\big]}\mathbb{E}\Big[S_{n,j}^2\mathbf{1}\big(S_{n,j} > \sqrt{\tfrac{1}{2}\varepsilon\theta\log n}\big)\Big] \leq C_\theta\big(\tfrac{\theta}{\theta+j-1}\big)^{\sqrt{\frac{1}{2}\varepsilon\theta\log n}}$$

für ein $C_\theta < \infty$, d.h (2.22) gilt.) □

Bemerkung 2.30 (Alternativer Zugang zu Beob. 2.29)**.** Wir könnten Erwartungswert und Varianz von S_n auch folgendermaßen berechnen: Gegeben die Gesamtlänge L_{ges} des Koaleszenten ist S_n $\mathrm{Poi}\big((\theta/2)L_{\mathrm{ges}}\big)$-verteilt.

$$L_{\mathrm{ges}} \stackrel{d}{=} \sum_{j=2}^{n} jT_j,$$

mit $T_n, T_{n-1}, \ldots, T_2$ u.a., $\mathscr{L}(T_j) = \mathrm{Exp}\big(\binom{j}{2}\big)$, somit

$$\begin{aligned}
\mathbb{E}_\theta[S_n] &= \frac{\theta}{2}\sum_{j=2}^{n} j\binom{j}{2} = \theta\sum_{i=1}^{n-1}\frac{1}{i}, &(2.23)\\
\mathrm{Var}_\theta[S_n] &= \mathbb{E}_\theta\big[\mathrm{Var}_\theta[S_n \,|\, L_{\mathrm{ges}}]\big] + \mathrm{Var}_\theta\big[\mathbb{E}_\theta[S_n \,|\, L_{\mathrm{ges}}]\big] \\
&= \mathbb{E}_\theta\Big[\frac{\theta}{2}L_{\mathrm{ges}}\Big] + \mathrm{Var}_\theta\Big[\frac{\theta}{2}L_{\mathrm{ges}}\Big] = \theta\sum_{i=1}^{n-1}\frac{1}{i} + \theta^2\sum_{i=1}^{n-1}\frac{1}{i^2}. &(2.24)
\end{aligned}$$

Die Konvergenzordnung $O(1/\log(n))$ der Varianz des Watterson-Schätzers ist zwar von Standpunkt der Statistik gesehen „frustrierend langsam“ (zumal im Vergleich zur klassischen Situation, in der man einen Parameter basierend auf n *unabhängigen* Beobachtungen schätzt, dort hat man typischerweise Abfall der Varianz $O(1/n)$ für plausible Schätzer). Andererseits haben Y. X. Fu and W. H. Li, Maximum Likelihood Estimation of Population Parameters, *Genetics* 134 (4), 1261–1270, (1993) gezeigt, dass es (zumindest asymptotisch) auch nicht besser möglich ist:

Satz 2.31. *Jeder erwartungstreue Schätzer für θ im IMS-Modell*

$$\textit{hat unter } \mathbb{P}_\theta \textit{ mindestens Varianz } \theta\Big/\sum_{k=1}^{n-1}\frac{1}{\theta+k} \quad (\sim \theta/\log n \textit{ für } n\to\infty).$$

Beweis. Nehmen wir (zunächst) an, wir könnten $S_{n,2} = s_{n,2}, S_{n,3} = s_{n,3}, \ldots, S_{n,n} = s_{n,n}$ beobachten (was anhand von Sequenzdaten an den Blättern des Koaleszenten nicht immer möglich ist):

Die Likelihoodfunktion (die Verteilungsgewichte von $(S_{n,n}, \ldots, S_{n,2})$, aufgefasst als Funktion des Parameters θ) ist

$$\begin{aligned} L_n(s_{n,2}, \ldots, s_{n,n}; \theta) &= \prod_{j=2}^{n} \frac{j-1}{\theta+j-1} \left(\frac{\theta}{\theta+j-1} \right)^{s_{n,j}} \\ &= (n-1)! \theta^{s_n} \prod_{j=2}^{n} \big(\theta+j-1\big)^{-(s_{n,j}+1)} \end{aligned}$$

mit $s_n = s_{n,2} + \cdots + s_{n,n}$, also

$$\frac{\partial}{\partial\theta} \log L_n(s_{n,2}, \ldots, s_{n,n}; \theta) = \frac{s_n}{\theta} - \sum_{j=2}^{n} \frac{s_{n,j}+1}{\theta+j-1}$$

d.h. $\widehat{\theta}_{\mathrm{ML,hyp}}$, der Maximum-Likelihood-Schätzer für θ basierend auf $(S_{n,n}, \ldots, S_{n,2})$, ist die Lösung (in θ) von $s_n = \theta \sum_{j=2}^{n} \frac{s_{n,j}+1}{\theta+j-1}$.

Weiter ist

$$\frac{\partial^2}{\partial\theta^2} \log L_n(s_{n,2}, \ldots, s_{n,n}; \theta) = -\frac{s_n}{\theta^2} + \sum_{j=2}^{n} \frac{s_{n,j}+1}{(\theta+j-1)^2},$$

die Fisher-Information ist somit

$$\begin{aligned} I(\theta) &= -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial\theta^2} \log L_n(s_{n,2}, \ldots, s_{n,n}; \theta) \right] \\ &= \mathbb{E}_\theta \left[\frac{S_n}{\theta^2} \right] - \sum_{j=2}^{n} \mathbb{E}_\theta \left[\frac{S_{n,j}+1}{(\theta+j-1)^2} \right] = \frac{\sum_{k=1}^{n-1} 1/k}{\theta} - \sum_{j=2}^{n} \frac{\theta+j-1}{(j-1)\,(\theta+j-1)^2} \\ &= \frac{\sum_{k=1}^{n-1} 1/k}{\theta} - \sum_{j=2}^{n} \frac{1}{(j-1)(\theta+j-1)} = \frac{1}{\theta} \sum_{k=1}^{n-1} \left(\frac{1}{k} - \frac{\theta}{k(\theta+k)} \right) = \frac{1}{\theta} \sum_{k=1}^{n-1} \frac{1}{\theta+k}. \end{aligned}$$

Gemäß der Cramér-Rao-Ungleichung (siehe z.B. John A. Rice, *Mathematical statistics and data analysis*, Duxbury Press, 1995, Ch. 8.6 oder die knappe Diskussion unten) gilt für jeden erwartungstreuen Schätzer

$$T = T(S_{n,n}, \ldots, S_{n,2})$$

für θ (d.h. der Schätzer wird durch eine Funktion $T : \mathbb{N}_0^{n-1} \to (0, \infty)$ mit $\mathbb{E}_\theta\big[T(S_{n,n}, \ldots, S_{n,2})\big] = \theta$ für alle $\theta > 0$ dargestellt)

$$\mathrm{Var}_\theta[T(S_{n,n}, \ldots, S_{n,2})] \geq \frac{1}{I(\theta)} = \frac{\theta}{\sum_{k=1}^{n-1} \frac{1}{\theta+k}},$$

d.h. die Behauptung.

Nachträge:

1. Heuristik zur Cramér-Rao-Ungleichung:

   Betrachten wir die allgemeine Situation, dass die Beobachtungen X ein Zufallsvektor (mit Werten in einer geeigneten Teilmenge von $\mathbb{R}^d$ für ein d) sind, die Dichte-/Likelihood-Funktion $f(x;\theta)$ (im diskreten Fall: die Gewichte) sei genügend glatt, so dass die folgenden Vertauschungen von Ableitung und Integral gerechtfertigt sind.

   Sei $V(X) := \frac{\partial}{\partial\theta} \log f(X;\theta) = \frac{1}{f(X;\theta)} \frac{\partial}{\partial\theta} f(X;\theta)$ die „Score-Funktion“, es ist $\mathbb{E}_\theta[V(X)] = 0$ stets, denn

   $$\int f(x;\theta) \frac{1}{f(x;\theta)} \frac{\partial}{\partial\theta} f(x;\theta)\, dx = \int \frac{\partial}{\partial\theta} f(x;\theta)\, dx = \frac{\partial}{\partial\theta} \int f(x;\theta)\, dx = \frac{\partial}{\partial\theta} 1 = 0.$$

   $$I(\theta) := \mathrm{Var}_\theta[V(X)] = \mathbb{E}_\theta[V(X)^2]$$

   heißt die Fisher-Information, beachte auch

   $$I(\theta) = -\mathbb{E}_\theta[\frac{\partial^2}{\partial\theta^2} \log f(X;\theta)]$$

   (d.h. wir können die Fisher-Information als (erwartete) Krümmung der Likelihood-Funktion an den beobachteten Daten interpretieren), denn

   $$\frac{\partial^2}{\partial\theta^2} \log f(x;\theta) = \frac{\frac{\partial^2}{\partial\theta^2} f(x;\theta)}{f(x;\theta)} - \Big(\frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)}\Big)^2 = \frac{\frac{\partial^2}{\partial\theta^2} f(x;\theta)}{f(x;\theta)} - \Big(\frac{\partial}{\partial\theta} \log f(x;\theta)\Big)^2$$

   und

   $$\mathbb{E}_\theta[\frac{\frac{\partial^2}{\partial\theta^2} f(X;\theta)}{f(X;\theta)}] = \int \frac{\frac{\partial^2}{\partial\theta^2} f(x;\theta)}{f(x;\theta)} f(x;\theta)\, dx = \int \frac{\partial^2}{\partial\theta^2} f(x;\theta)\, dx = \frac{\partial^2}{\partial\theta^2} \int f(x;\theta)\, dx = \frac{\partial^2}{\partial\theta^2} 1 = 0.$$

   Sei nun $T(X)$ irgendein Schätzer für θ (d.h. eine Funktion der Beobachtungen X mit Werten in $[0,\infty)$ (und $\mathbb{E}_\theta[(T(X))^2] < \infty$), dann ist

   $$\begin{aligned}\mathrm{Cov}_\theta[T(X), V(X)] = \mathbb{E}_\theta[T(X)V(X)] &= \int T(x) \frac{1}{f(x;\theta)} \frac{\partial}{\partial\theta} f(x;\theta)\, f(x;\theta)\, dx \\ &= \int T(x) \frac{\partial}{\partial\theta} f(x;\theta)\, dx = \frac{\partial}{\partial\theta} \mathbb{E}_\theta[T(X)].\end{aligned}$$

   Die Cauchy-Schwarz-Ungleichung liefert

   $$\sqrt{\mathrm{Var}_\theta[T(X)]\mathrm{Var}_\theta[V(X)]} \geq |\mathrm{Cov}_\theta[T(X), V(X)]| = \Big|\frac{\partial}{\partial\theta} \mathbb{E}_\theta[T(X)]\Big|$$

   und somit gilt

   $$\mathrm{Var}_\theta[T(X)] \geq \frac{\Big|\frac{\partial}{\partial\theta} \mathbb{E}_\theta[T(X)]\Big|^2}{\mathrm{Var}_\theta[V(X)]} = \frac{\Big|\frac{\partial}{\partial\theta} \mathbb{E}_\theta[T(X)]\Big|^2}{I(\theta)}.$$

   falls $T(X)$ ein unverzerrter Schätzer für θ ist, d.h. $\mathbb{E}_\theta[T(X)] = \theta$ für alle θ, so ist natürlich $\frac{\partial}{\partial\theta} \mathbb{E}_\theta[T(X)] = 1$ (und dies ist die Form der Cramér-Rao-Ungleichung, die wir oben verwendet haben).

2. Wir hatten die Schranke unter der Annahme hergeleitet, dass wir $(S_{n,n}, \dots, S_{n,2})$ tatsächlich beobachten könnten, was anhand der Daten i.A. nicht möglich ist. Intuitiv erscheint es zumindest sehr plausibel, dass jeder „reale“ erwartungstreue Schätzer für θ (d.h. jedes $\widetilde{T} = \widetilde{T}(D)$, das eine Funktion der Datenmatrix D ist, das also weniger Informationen verwenden darf, als wir oben angenommen hatten), ebenfalls mindestens Varianz $1/I(\theta)$ hat.

   Diese Intuition kann man durch den Begriff der (statistischen) Suffizienz folgendermaßen formalisieren: Sei X die „volle“ Information, die die Genealogie der n-Stichprobe und die darauf vorkommenden Mutationen beschreibt, d.h. X enthält die „topologische“ Information, in welcher Reihenfolge die Verschmelzungen der Linien stattfinden, und für jede Kante im Baum die Information, welche Mutationen auf dieser liegen (wir verzichten hier darauf, dies in Formeln zu fassen). Offenbar kann man aus X die Datenmatrix D ablesen und somit kann $\widetilde{T} = \widetilde{T}(D(X))$ als eine Funktion von X interpretiert werden.

   Die entscheidende Beobachtung ist, dass $Y := (S_{n,2}, S_{n,3}, \dots, S_{n,n})$ *suffizient* für θ ist, d.h. die bedingte Verteilung $\mathscr{L}_\theta(X \,|\, Y)$ hängt nicht von θ ab — gegeben $S_{n,2} = s_{n,2}, \dots, S_{n,n} = s_{n,2}$ entsteht X, indem man für $j = n, n-1, \dots, 2$ auf den j Kanten in „Niveau“ j des Koaleszenten $s_{n,j}$ Mutationen uniform verteilt und unter allen aktuell möglichen Verschmelzungen uniform eine auswählt; demnach enthalten die Gewichte von $\mathscr{L}_\theta(X \,|\, Y)$ nur kombinatorische Terme, aber keine θ-Abhängigkeit.

   Nun ist (beachte, dass wir den bedingten Erwartungswert bilden können, ohne θ zu kennen)

$$\widehat{T} := \widehat{T}(Y) := \mathbb{E}[\widetilde{T} \,|\, Y]$$

   ebenfalls ein erwartungstreuer Schätzer für θ und nach Konstruktion ist $\widehat{T}$ eine gewisse Funktion von $Y = (S_{n,2}, S_{n,3}, \dots, S_{n,n})$, d.h. nach obigem ist $\mathrm{Var}_\theta[\widehat{T}] \geq 1/I(\theta)$ und folglich auch

$$\mathrm{Var}_\theta[\widetilde{T}] = \mathbb{E}_\theta[\underbrace{\mathrm{Var}_\theta[\widetilde{T} \,|\, Y]}_{\geq 0}] + \mathrm{Var}_\theta[\underbrace{\mathbb{E}_\theta[\widetilde{T} \,|\, Y]}_{=\widehat{T}}] \geq \mathrm{Var}_\theta[\widehat{T}] \geq \frac{1}{I(\theta)}.$$

□

Definition 2.32 (Frequenzspektrum)**.** Sei

$$\xi_i^{(n)} := \#\text{ Mutationen, die in genau } i \text{ der } n \text{ Stichproben vorkommen}, \quad i = 1, \dots, n-1.$$

(Wir nehmen dabei an, dass an jeder Position der anzestrale oder „Wildtyp“ bekannt ist, z.B. durch Interspezies-Vergleich.) Der Vektor

$$\xi^{(n)} = \left(\xi_1^{(n)}, \xi_2^{(n)}, \dots, \xi_{n-1}^{(n)}\right)$$

heißt das Frequenzspektrum (der segregrierenden Stellen).

Wenn der anzestrale Typ nicht bekannt ist, betrachtet man stattdessen das gefaltete Frequenzspektrum $\left(\eta_1^{(n)}, \eta_2^{(n)}, \dots, \eta_{\lfloor n/2 \rfloor}^{(n)}\right)$ mit

$$\eta_i^{(n)} := \xi_i^{(n)} + \xi_{n-i}^{(n)} \mathbf{1}_{i \neq n/2}, \quad 1 \leq i \leq \lfloor n/2 \rfloor.$$

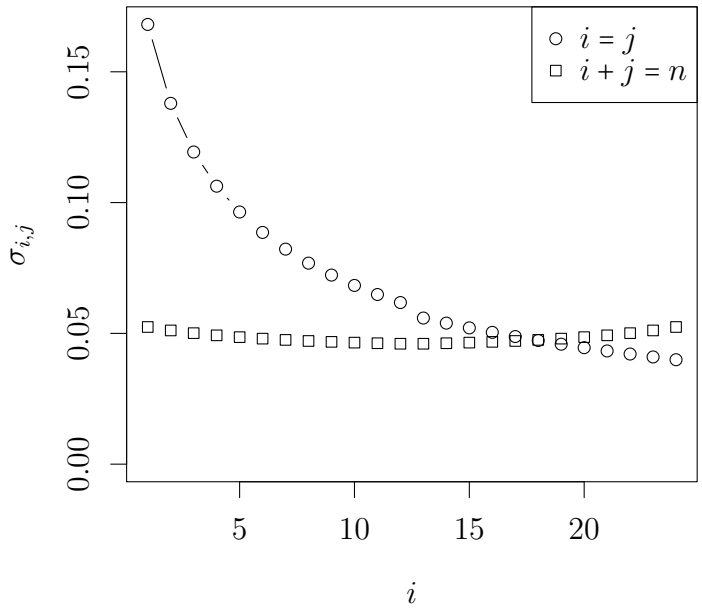


Abbildung 2.2: Diagonaleinträge $\sigma_{i,i}$ und Antidiagonaleinträge $\sigma_{i,n-i}$ der Kovarianzmatrix von $\xi^{(n)}$ für $n = 25$, $\theta = 1$

Satz 2.33. *Es gilt*

$$\mathbb{E}_\theta\left[\xi_i^{(n)}\right] = \frac{\theta}{i}, \quad \mathrm{Cov}_\theta\left[\xi_i^{(n)}, \xi_j^{(n)}\right] = \mathbf{1}_{i=j}\frac{\theta}{i} + \theta^2\sigma_{ij}, \quad 1 \le i \le j \le n$$

mit $h_n := \sum_{i=1}^{n-1} \frac{1}{i}$, $\beta_n(i) := \frac{2n}{(n-i+1)(n-i)}\left(h_{n+1} - h_i\right) - \frac{2}{n-i}$

$$\sigma_{ii} = \begin{cases} \beta_n(i+1), & i < \frac{n}{2}, \\ 2\frac{h_n - h_i}{n-i} - \frac{1}{i^2}, & i = \frac{n}{2}, \\ \beta_n(i) - \frac{1}{i^2}, & i > \frac{n}{2}, \end{cases} \quad \textit{für } i > j \textit{ ist} \quad \sigma_{ij} = \begin{cases} \frac{\beta_n(i+1) - \beta_n(i)}{2}, & i + j < n, \\ \frac{h_n - h_i}{n-i} + \frac{h_n - h_j}{n-j} & \\ \quad - \frac{\beta_n(i) + \beta_n(j+1)}{2} - \frac{1}{ij}, & i + j = n, \\ \frac{\beta_n(j) - \beta_n(j+1)}{2} - \frac{1}{ij}, & i + j > n \end{cases}$$

(und $\sigma_{ij} = \sigma_{ji}$).

Die Diagonaleinträge $\sigma_{i,i}$ (d.h. $\mathrm{Var}_\theta\left[\xi_i^{(n)}\right]$) dominieren die Kovarianzmatrix $(\sigma_{i,j})$: Abb. 2.2 zeigt die Diagonaleinträge $\sigma_{i,i}$ und die Antidiagonaleinträge $\sigma_{i,n-i}$ für $n = 25$ und $\theta = 1$, Abb. 2.3 zeigt eine dreidimensionale Darstellung von $(\sigma_{i,j})$ für $n = 25$ und $\theta = 1$, Abb. 2.3 zeigt $(-\sigma_{i,j})$, wobei der besseren Sichtbarkeit wegen die (auch betragsmäßig) deutlich größeren Diagonal- und Antidiagonaleinträge auf 0 gesetzt wurden.

Beweis (der Formel für den Erwartungswert). Wir denken uns die Kanten des n-Koaleszenten auf jedem Niveau (u.a. zufällig) nummeriert.

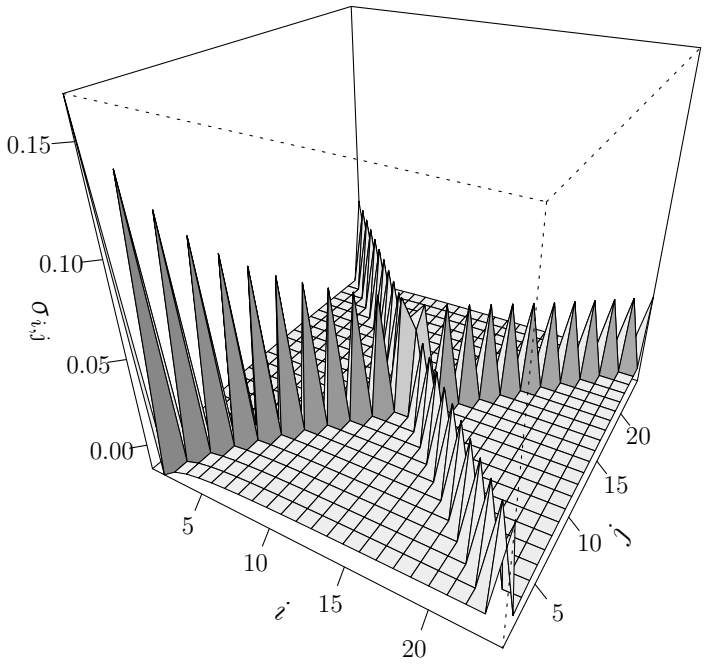


Abbildung 2.3: Kovarianzmatrix von $\xi^{(n)}$ für $n = 25$, $\theta = 1$

Sei

$$\nu_{k,\ell} := \#\,\text{Mutationen auf } \ell\text{-ter Kante auf Niveau } k,$$
$$J_{k,\ell} := \#\,\text{Blätter oberhalb } \ell\text{-ter Kante auf Niveau } k,$$

damit ist

$$\xi_i^{(n)} = \sum_{k=2}^{n-i+1} \sum_{\ell=1}^{k} \nu_{k,\ell} \mathbf{1}(J_{k,\ell} = i). \tag{2.25}$$

Somit gilt

$$\begin{aligned}
\mathbb{E}_\theta\big[\xi_i^{(n)}\big] &= \sum_{k=2}^{n-i+1} \sum_{\ell=1}^{k} \mathbb{E}_\theta\big[\nu_{k,\ell}\mathbf{1}(J_{k,\ell} = i)\big] = \sum_{k=2}^{n-i+1} \sum_{\ell=1}^{k} \mathbb{E}_\theta\big[\nu_{k,\ell}\big]\mathbb{P}_\theta(J_{k,\ell} = i) \\
&= \sum_{k=2}^{n-i+1} \sum_{\ell=1}^{k} \frac{\theta}{k(k-1)} \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} = \sum_{k=2}^{n-i+1} k \frac{\theta}{k(k-1)} \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \\
&= \theta \sum_{k=2}^{n-i+1} \frac{1}{k-1} \frac{(n-i-1)!}{(k-2)!(n-i-k+1)!} \frac{(k-1)!(n-k)!}{(n-1)!} \times \frac{i!}{i(i-1)!} \\
&= \frac{\theta}{i} \frac{1}{\binom{n-1}{i}} \sum_{k=2}^{n-i+1} \binom{n-k}{i-1} = \frac{\theta}{i}
\end{aligned}$$

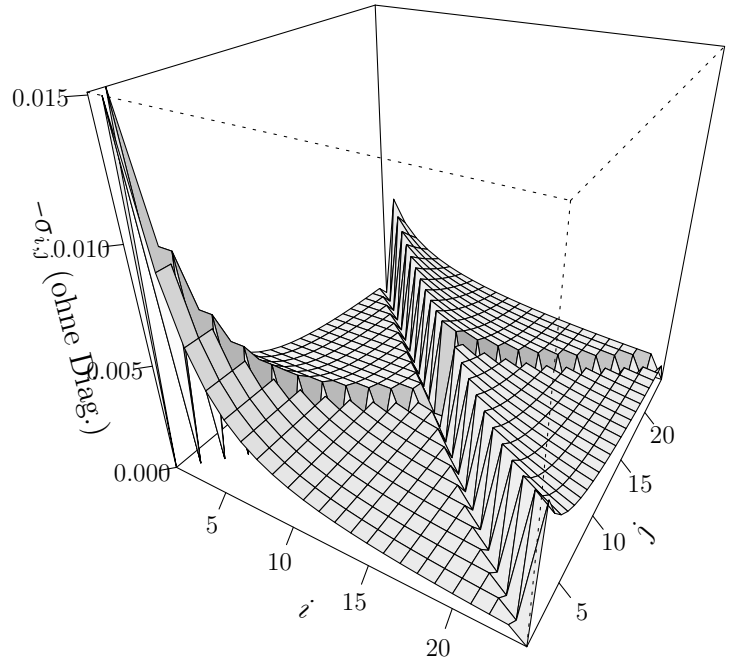


Abbildung 2.4: $(-1)\times$Kovarianzmatrix von $\xi^{(n)}$ für $n = 25$, $\theta = 1$, wobei Diagonal- und Antidiagonaleinträge auf 0 gesetzt wurden

Wir verwenden hierbei in der ersten Zeile, dass

$$\mathbb{E}_\theta\big[\nu_{k,\ell}\big] = \mathbb{E}_\theta\big[\mathbb{E}_\theta\big[\nu_{k,\ell}\,\big|\,T_k\big]\big] = \mathbb{E}_\theta\big[\frac{\theta}{2}T_k\big] = \frac{\theta}{2}\frac{2}{k(k-1)} = \frac{\theta}{k(k-1)}$$

gilt und dass $\nu_{k,\ell}$ (das nur vom Poissonprozess der Mutationen abhängt) und $J_{k,\ell}$ (das nur die Kombinatorik der Abstammungsverhältnisse widerspiegelt) unabhängig sind.

In der zweiten Zeile ersetzen wir

$$\mathbb{P}_\theta(J_{k,\ell} = i) = \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}},$$

denn mit Korollar 1.12 ist die Aufteilung in k (zufällig nummerierte) Familiengrößen uniform auf allen $\big\{(m_1,\dots,m_i)\in\mathbb{N}^i : m_1+\dots+m_i = n\big\}$: Es gibt $\binom{n-1}{k-1}$ viele Möglichkeiten, bei $\binom{n-i-1}{k-2}$ davon ist $J_{k,\ell} = i$.

Schließlich beachte in der letzten Zeile $\sum_{k=2}^{n-i+1}\binom{n-k}{i-1} = \binom{n-1}{i}$, denn es gibt $\binom{n-1-(k-1)}{i-1} = \binom{n-k}{i-1}$ viele Teilmengen von $\{1,\dots,n-1\}$ der Größe i, deren kleinstes Element $k-1$ ist, und insgesamt $\binom{n-1}{i}$ Teilmengen von $\{1,\dots,n-1\}$ der Größe i.

Um $\mathbb{E}_\theta[\xi_i^{(n)}\xi_j^{(n)}]$ zu bestimmen kann man die Darstellungen (2.25) für i und für j miteinander multiplizieren und erhält analog zu oben eine Darstellung via eine Doppelsumme über Paare von Kanten im Koaleszenten-Baum. Mittels einer Verfeinerung von Korollar 1.12 kann man den kombinatorischen Ausdruck $\mathbb{P}_\theta(J_{k,\ell} = i, J_{k',\ell'} = j)$ bestimmen (man unterscheidet verschiedene Fälle, je nachdem ob die betrachtete Kante (k',ℓ') ein Nachfahre der Kante (k,ℓ) im Baum ist oder nicht) und erhält nach recht umfangreichen Umformungen die oben angegebenen Ausdrücke für $\mathrm{Cov}_\theta\big[\xi_i^{(n)}, \xi_j^{(n)}\big]$, für Details siehe den Artikel von Yun-Xin Fu, Statistical Properties of Segregating Sites, *Theor. Pop. Biol.* 48, 172–197 (1995), in dem dieser Satz bewiesen wurde. □

Tajimas[7] Test

Betrachte eine n-Stichprobe (im IMS-Modell), für $1 \le i < j \le n$ sei

$\Delta_{i,j}$:= Anzahl segregierende Stellen, an denen sich Stichproben i und j unterscheiden.

Die mittlere Anzahl paarweiser Unterschiede,

$$\widehat{\theta}_\pi := \frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n}\Delta_{ij}$$

(„Tajimas $\widehat{\theta}_\pi$“), ist ein (auf den beobachteten Sequenzen basierender) Schätzer für die Mutationsrate θ.

Beobachtung 2.34. Es gebe s segregierende Stellen.

$$\widehat{\theta}_\pi = \frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n}\sum_{m=1}^{s}\mathbf{1}(\text{Stichpr. } i \text{ und } j \text{ unterschiedl. an } m\text{-ter segr. Stelle}) = \frac{1}{\binom{n}{2}}\sum_{k=1}^{n-1}\xi_k^{(n)}\,k(n-k)$$

[7] Fumio Tajima, Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism, *Genetics* 123, 585–595, (1989)

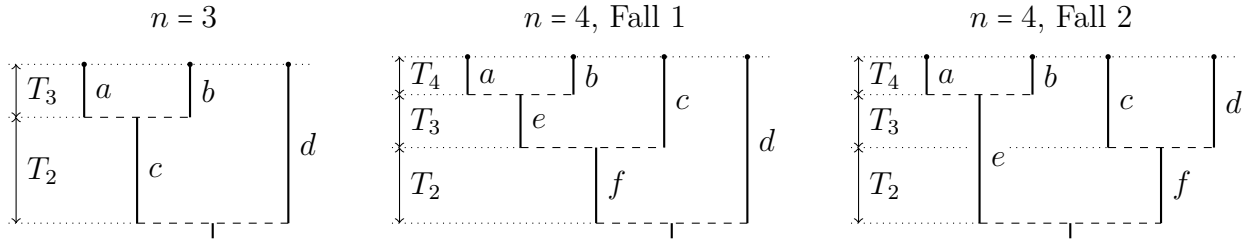


Abbildung 2.5: Formen eines n-Koaleszenten für $n = 3$ (bis auf Umnummerierung der Bläter nur eine Möglichkeit) und $n = 4$ (zwei Möglichkeiten)

(mit $\xi_k^{(n)} = \#$ Mut., die in k Stichpr. vorkommen, aus Def. 2.32), d.h. $\widehat{\theta}_\pi$ ist eine (lineare) Funktion des Frequenzspektrums.

(Darüberhinaus kann $\widehat{\theta}_\pi$ ebenso wie S_n und $\widehat{\theta}_W$ als Funktion des gefalteten Frequenzspektrums aufgefasst werden, d.h. wir können dies auch dann bestimmen, wenn wir die anzestralen Typen nicht kennen.)

Proposition 2.35. *Es gilt*

$$\mathbb{E}_\theta\left[\widehat{\theta}_\pi\right] = \theta, \quad \mathrm{Var}_\theta\left(\widehat{\theta}_\pi\right) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2.$$

Insbesondere: $\widehat{\theta}_\pi$ ist erwartungstreuer Schätzer für θ, allerdings ist es nicht konsistent:

$$\lim_{n\to\infty} \mathrm{Var}_\theta\left(\widehat{\theta}_\pi\right) = \tfrac{1}{3}\theta + \tfrac{2}{9}\theta^2 > 0.$$

Bemerkung. $\widehat{\theta}_\pi = \frac{1}{\binom{n}{2}} \sum_{1\le i<j\le n} \Delta_{ij}$ wird in der Literatur auch mit π bezeichnet und die (empirische) „Nukleotid-Diversität“ (engl. “nucleotide diversity”) genannt.

($\mathbb{E}_\theta[\pi] = \theta$ ist einer der Gründe für die Parametrisierung, dass Mutationen mit Rate $\theta/2$ längs der Genealogie erscheinen.)

Beweis. Betrachte zunächst eine Stichprobe der Größe $n = 2$: Es ist

$$\mathbb{E}_\theta[\Delta_{1,2}] = \mathbb{E}_\theta\big[\mathbb{E}_\theta[\Delta_{1,2}\,|\,T_2]\big] = \mathbb{E}_\theta[\theta T_2] = \theta\mathbb{E}_\theta[T_2] = \theta$$

(mit $T_2 =$ Zeit, währenddessen die Genealogie aus 2 Linien besteht $=$ Zeit bis zum jgV der beiden Stichproben, $T_2 \sim \mathrm{Exp}(1)$) und

$$\begin{aligned}\mathrm{Var}_\theta[\Delta_{1,2}] &= \mathrm{Var}_\theta\big[\mathbb{E}_\theta[\Delta_{1,2}\,|\,T_2]\big] + \mathbb{E}_\theta\big[\mathrm{Var}_\theta[\Delta_{1,2}\,|\,T_2]\big] \\ &= \mathrm{Var}_\theta[\theta T_2] + \mathbb{E}_\theta[\theta T_2] = \theta^2\mathrm{Var}_\theta[T_2] + \theta\mathbb{E}_\theta[T_2] = \theta^2 + \theta.\end{aligned}$$

Betrachte nun eine Stichprobe der Größe $n = 3$, es bezeichne η_a die Anzahl Mutationen auf Kante a, etc., siehe Abbildung 2.5. Jede Kante kommt in 2 von 3 paarweisen Vergleichen vor, also ist

$$\widehat{\theta}_{\pi,n=3} = \frac{1}{3}\big(\Delta_{1,2} + \Delta_{1,3} + \Delta_{2,3}\big) = \frac{2}{3}\big(\eta_a + \eta_b + \eta_c + \eta_d\big).$$

Sei T_j die Länge der Zeitspanne, währenddessen der Koaleszent aus j Linien besteht. Nach Definition sind $\eta_a, \eta_b, \eta_c, \eta_d$ unabhängig, gegeben T_3 und T_2, und $\eta_a, \eta_b \sim \mathrm{Poi}(\frac{\theta}{2}T_3), \eta_c \sim \mathrm{Poi}(\frac{\theta}{2}T_2), \eta_d \sim \mathrm{Poi}(\frac{\theta}{2}(T_3+T_2))$, d.h. $\eta_a+\eta_b+\eta_c+\eta_d \sim \mathrm{Pois}(\theta L_3/2)$, wobei $L_3 = 3T_3+2T_2$ die Gesamtlänge des Baums ist. Daher ist

$$\begin{aligned}
\mathrm{Var}_\theta\big[\widehat{\theta}_{\pi,n=3}\big] &= \frac{4}{9}\mathrm{Var}_\theta\big[\eta_a+\eta_b+\eta_c+\eta_d\big] \\
&= \frac{4}{9}\mathrm{Var}_\theta\big[\mathbb{E}_\theta[\eta_a+\eta_b+\eta_c+\eta_d \,|\, L_3]\big] + \frac{4}{9}\mathbb{E}_\theta\big[\mathrm{Var}_\theta[\eta_a+\eta_b+\eta_c+\eta_d \,|\, L_3]\big] \\
&= \frac{4}{9}\mathrm{Var}_\theta\big[\frac{\theta}{2}L_3\big] + \frac{4}{9}\mathbb{E}_\theta\big[\frac{\theta}{2}L_3\big] = \frac{4}{9}\frac{\theta^2}{4}\big(9\cdot\frac{1}{3^2}+4\cdot 1\big) + \frac{4}{9}\frac{\theta}{2}\big(3\cdot\frac{1}{3}+2\cdot 1\big) = \frac{5}{9}\theta^2+\frac{2}{3}\theta.
\end{aligned}$$

Andererseits ist wegen der Symmetrien der Verteilung des Koaleszenten $\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{1,3}] = \mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{2,3}]$, etc. und somit

$$\mathrm{Var}_\theta\big[\widehat{\theta}_{\pi,n=3}\big] = \frac{1}{9}\cdot 3\cdot\mathrm{Var}_\theta[\Delta_{1,2}] + \frac{1}{9}\cdot 6\cdot\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{1,3}] = \frac{1}{3}(\theta^2+\theta) + \frac{2}{3}\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{1,3}],$$

folglich

$$\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{1,3}] = \frac{3}{2}\mathrm{Var}_\theta\big[\widehat{\theta}_{\pi,n=3}\big] - \frac{1}{2}(\theta^2+\theta) = \frac{1}{3}\theta^2+\frac{1}{2}\theta. \tag{2.26}$$

Betrachte nun eine Stichprobe der Größe $n=4$: Es gibt 2 mögliche Baumtopologien (siehe Abb. 2.5).

Wir untersuchen zunächst Fall 1 (das mittlere Diagramm in Abb. 2.5). Gegeben T_4, T_3, T_2 sind hier $\eta_a \sim \mathrm{Poi}(\frac{\theta}{2}T_4), \eta_b \sim \mathrm{Poi}(\frac{\theta}{2}T_4), \eta_c \sim \mathrm{Poi}(\frac{\theta}{2}(T_4+T_3)), \eta_d \sim \mathrm{Poi}(\frac{\theta}{2}(T_4+T_3+T_2)), \eta_e \sim \mathrm{Poi}(\frac{\theta}{2}T_3), \eta_f \sim \mathrm{Poi}(\frac{\theta}{2}T_2)$ und unabhängig. Weiter ist in diesem Fall

$$\widehat{\theta}_{\pi,n=4} = \Delta(1) = \frac{1}{\binom{4}{2}}\big(3\eta_a+3\eta_b+3\eta_c+3\eta_d+4\eta_e+3\eta_f\big) =: \frac{1}{6}X_1$$

(beachte: wenn oberhalb einer Kante ℓ Blätter liegen, so tritt sie in $\ell\cdot(n-\ell)$ paarweisen Vergleichen auf), somit ist

$$\begin{aligned}
\mathbb{E}_\theta\big[\widehat{\theta}_{\pi,n=4}\,\big|\,\text{Top.}=1\big] &= \frac{1}{6}\frac{\theta}{2}\mathbb{E}\big[3T_4+3T_4+3(T_4+T_3)+3(T_4+T_3+T_2)+4T_3+3T_2\big] \\
&= \frac{\theta}{12}\Big(\frac{3}{\binom{4}{2}}+\frac{3}{\binom{4}{2}}+3\Big(\frac{1}{\binom{4}{2}}+\frac{1}{\binom{3}{2}}\Big)+\frac{4}{\binom{3}{2}}+3\Big(\frac{1}{\binom{4}{2}}+\frac{1}{\binom{3}{2}}+1\Big)+3\cdot 1\Big) = \frac{17}{18}\theta, \\
\mathrm{Var}_\theta\big[\widehat{\theta}_{\pi,n=4}\,\big|\,\text{Top.}=1\big] &= \frac{1}{36}\mathbb{E}_\theta\big[\mathrm{Var}_\theta[X_1\,|\,T_4,T_3,T_2]\big] + \frac{1}{36}\mathrm{Var}_\theta\big[\mathbb{E}_\theta[X_1\,|\,T_4,T_3,T_2]\big] \\
&= \frac{1}{36}\mathbb{E}_\theta\big[\tfrac{\theta}{2}(9T_4+9T_4+9(T_3+T_4)+9(T_4+T_3+T_2)+16T_3+9T_2)\big] \\
&\quad + \frac{1}{36}\mathrm{Var}_\theta\big[\frac{\theta}{2}(3T_4+3T_4+3(T_4+T_3)+3(T_4+T_3+T_2)+4T_3+3T_2)\big] \\
&= \frac{\theta}{72}\big(\tfrac{9}{6}+\tfrac{9}{6}+9(\tfrac{1}{6}+\tfrac{1}{3})+9(\tfrac{1}{6}+\tfrac{1}{3}+1)+\tfrac{16}{3}+9\cdot 1\big) \\
&\quad + \frac{\theta^2}{144}\mathrm{Var}_\theta\big[12T_4+10T_3+6T_2\big] \\
&= \frac{53}{108}\theta + \frac{\theta^2}{144}\big(\tfrac{12^2}{6^2}+\tfrac{10^2}{3^2}+\tfrac{6^2}{1^2}\big) = \frac{53}{108}\theta+\frac{115}{324}\theta^2.
\end{aligned}$$

Untersuchen wir nun Fall 2 (das rechte Diagramm in Abb. 2.5). Gegeben T_4, T_3, T_2 sind hier $\eta_a \sim \mathrm{Poi}(\frac{\theta}{2}T_4), \eta_b \sim \mathrm{Poi}(\frac{\theta}{2}T_4), \eta_c \sim \mathrm{Poi}(\frac{\theta}{2}(T_4+T_3)), \eta_d \sim \mathrm{Poi}(\frac{\theta}{2}(T_4+T_3)), \eta_e \sim \mathrm{Poi}(\frac{\theta}{2}(T_3+T_2)), \eta_f \sim \mathrm{Poi}(\frac{\theta}{2}T_2)$ und unabhängig, weiter ist in diesem Fall (mit Argumentation analog zu Fall 1)

$$\widehat{\theta}_{\pi,n=4} = \Delta(2) = \frac{1}{\binom{4}{2}}\big(3\eta_a + 3\eta_b + 3\eta_c + 3\eta_d + 4\eta_e + 4\eta_f\big) =: \frac{1}{6}X_2,$$

somit ergibt sich

$$\begin{aligned}
\mathbb{E}_\theta\big[\widehat{\theta}_{\pi,n=4}\,\big|\,\text{Top.}=2\big] &= \frac{1}{6}\frac{\theta}{2}\mathbb{E}\big[3T_4 + 3T_4 + 3(T_4+T_3) + 3(T_4+T_3) + 4(T_3+T_2) + 4T_2\big] \\
&= \frac{\theta}{12}\mathbb{E}\big[12T_4 + 10T_3 + 8T_2\big] = \frac{\theta}{12}\big(\tfrac{12}{6} + \tfrac{10}{3} + 8\big) = \frac{10}{9}\theta, \\
\mathrm{Var}_\theta\big[\widehat{\theta}_{\pi,n=4}\,\big|\,\text{Top.}=2\big] &= \frac{1}{36}\mathbb{E}_\theta\big[\mathrm{Var}_\theta[X_2\,|\,T_4,T_3,T_2]\big] + \frac{1}{36}\mathrm{Var}_\theta\big[\mathbb{E}_\theta[X_2\,|\,T_4,T_3,T_2]\big] \\
&= \frac{1}{36}\mathbb{E}_\theta\big[\frac{\theta}{2}(9T_4 + 9T_4 + 9(T_4+T_3) + 9(T_4+T_3) + 16(T_3+T_2) + 16T_2)\big] \\
&\quad + \frac{1}{36}\mathrm{Var}_\theta\big[\frac{\theta}{2}(3T_4 + 3T_4 + 3(T_4+T_3) + 3(T_4+T_3) + 4(T_3+T_2) + 4T_2)\big] \\
&= \frac{\theta}{72}\mathbb{E}\big[36T_4 + 34T_3 + 32T_2\big] + \frac{\theta^2}{144}\mathrm{Var}_\theta\big[12T_4 + 10T_3 + 8T_2\big] \\
&= \frac{\theta}{72}\big(\tfrac{36}{6} + \tfrac{34}{3} + 32\cdot 1\big) + \frac{\theta^2}{144}\big(\tfrac{12^2}{6^2} + \tfrac{10^2}{3^2} + \tfrac{8^2}{1^2}\big) = \frac{37}{54}\theta + \frac{89}{162}\theta^2.
\end{aligned}$$

Insgesamt ist mit $\mathbb{P}(\text{Top.}=1) = 2/3 = 1 - \mathbb{P}(\text{Top.}=2)$ (denn damit der 2. Fall für die Baumtopologie eintritt, muss die zweitjüngste Verschmelzung das Paar von Linien betreffen, das bis dahin noch an keiner Verschmelzung teilgenommen hat, dies ist dann 1 von 3 Möglichkeiten)

$$\mathbb{E}_\theta\big[\widehat{\theta}_{\pi,n=4}\big] = \frac{2}{3}\cdot\frac{17}{18}\theta + \frac{1}{3}\cdot\frac{10}{9}\theta = \theta$$

und

$$\begin{aligned}
\mathrm{Var}_\theta\big[\widehat{\theta}_{\pi,n=4}\big] &= \mathbb{E}_\theta\big[\mathrm{Var}_\theta[\widehat{\theta}_{\pi,n=4}\,|\,\text{Top.}]\big] \;+\; \mathrm{Var}_\theta\big[\mathbb{E}_\theta[\widehat{\theta}_{\pi,n=4}\,|\,\text{Top.}]\big] \\
&= \frac{2}{3}\big(\frac{53}{108}\theta + \frac{115}{324}\theta^2\big) + \frac{1}{3}\big(\frac{37}{54}\theta + \frac{89}{162}\theta^2\big) \;+\; \frac{2}{3}\big(\frac{17}{18}\theta - \theta\big)^2 + \frac{1}{3}\big(\frac{10}{9}\theta - \theta\big)^2 \\
&= \frac{23}{54}\theta^2 + \frac{5}{9}\theta.
\end{aligned}$$

Andererseits ist wie oben wegen der Symmetrien der Verteilung des Koaleszenten

$$\begin{aligned}
\mathrm{Var}_\theta\big[\widehat{\theta}_{\pi,n=4}\big] &= \frac{1}{36}\mathrm{Cov}_\theta\Big[\sum_{1\le i<j\le 4}\Delta_{i,j}\,,\,\sum_{1\le k<\ell\le 4}\Delta_{k,\ell}\Big] \\
&= \frac{1}{36}\Big(6\mathrm{Var}_\theta[\Delta_{1,2}] + 6\cdot 2\cdot 2\,\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{1,3}] + 6\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{3,4}]\Big) \\
&= \frac{1}{6}(\theta^2+\theta) + \frac{2}{3}\big(\frac{1}{3}\theta^2 + \frac{1}{2}\theta\big) + \frac{1}{6}\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{3,4}]
\end{aligned}$$

und somit

$$\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{3,4}] = 6\mathrm{Var}_\theta\big[\widehat{\theta}_{\pi,n=4}\big] - (\theta^2+\theta) - 4\Big(\frac{1}{3}\theta^2+\frac{1}{2}\theta\Big) = \frac{2}{9}\theta^2+\frac{1}{3}\theta. \tag{2.27}$$

Schließlich betrachten wir den allgemeinen Fall einer Stichprobe der Größe n:

$$\begin{aligned}
\mathbb{E}_\theta\big[\widehat{\theta}_\pi\big] &= \frac{1}{\binom{n}{2}} \sum_{1\le i<j\le n} \mathbb{E}_\theta[\Delta_{i,j}] = \frac{1}{\binom{n}{2}}\binom{n}{2}\mathbb{E}_\theta[\Delta_{1,2}] = \theta,\\
\mathrm{Var}_\theta\big[\widehat{\theta}_\pi\big] &= \frac{1}{\binom{\binom{n}{2}}{2}}\mathrm{Cov}_\theta\Big[\sum_{1\le i<j\le n}\Delta_{i,j}\,,\sum_{1\le k<\ell\le n}\Delta_{k,\ell}\Big]\\
&= \frac{1}{\binom{\binom{n}{2}}{2}}\Big(\tbinom{n}{2}\mathrm{Var}_\theta[\Delta_{1,2}] + \tbinom{n}{2}2(n-2)\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{1,3}]\\
&\qquad\qquad + \tbinom{n}{2}\tbinom{n-2}{2}\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{3,4}]\Big)\\
&= \frac{1}{\binom{n}{2}}\Big(\theta^2+\theta+2(n-2)\Big(\frac{1}{3}\theta^2+\frac{1}{2}\theta\Big) + \tbinom{n-2}{2}\Big(\frac{2}{9}\theta^2+\frac{1}{3}\theta\Big)\Big)\\
&= \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2.
\end{aligned}$$

□

Passen beobachtete Sequenzdaten zum Modell? Unser wahrscheinlichkeitstheoretisches Modell beschreibt die Verteilung von n beobachteten Sequenzen, die wir an den Blättern eines Kingman-n-Koaleszenten ablesen, auf dessen Kanten gemäß einem Poissonprozess mit einer gewissen Rate $\theta/2$ Mutationen liegen, die den Typ jeweils gemäß dem IMS-Modell ändern. Angesichts Satz 1.8 ist die biologische Interpretation, dass wir n Stichproben aus einer „panmiktischen" Population konstanter Größe sehen und dass die genetische Variabilität (am betrachteten Ort im Genom) „neutral" ist (und dass die Annahmen des IMS-Modells wenigstens approximativ zutreffen).

Die Tatsache, dass sowohl $\widehat{\theta}_W$ als auch $\widehat{\theta}_\pi$ in diesem Modell erwartungstreue Schätzer für (das unbekannte) θ sind, gestattet es, für die Nullhypothese „das Modell beschreibt die Daten zutreffend" einen statistischen Test zu formulieren. Wenn das Modell zutrifft, sollte nämlich

$$\widehat{\theta}_\pi - \widehat{\theta}_W \approx 0$$

bis auf „zufällige Fluktuationen" gelten. Diese Idee geht auf F. Tajima zurück, siehe den in Fußnote 7 auf S. 83 zitierten Artikel.

Um einzuschätzen, wie groß die „typischen" Fluktuationen sind, sollten wir (zumindest) $\mathrm{Var}_\theta\big[\widehat{\theta}_\pi - \widehat{\theta}_W\big]$ bestimmen können.

Bericht 2.36. Es gilt $\mathrm{Cov}_\theta\big[S_n,\widehat{\theta}_\pi\big] = \theta + \big(\frac{1}{2}+\frac{1}{n}\big)\theta^2$, also $\mathrm{Cov}_\theta\big[\widehat{\theta}_W,\widehat{\theta}_\pi\big] = \frac{\theta}{h_n} + \big(\frac{1}{2}+\frac{1}{n}\big)\frac{\theta^2}{h_n}$ und somit

$$\mathrm{Var}_\theta\big[\widehat{\theta}_\pi - \widehat{\theta}_W\big] = \Big(\tfrac{n+1}{3(n-1)} - \tfrac{1}{h_n}\Big)\theta + \Big(\tfrac{2(n^2+n+3)}{9n(n-1)} - \tfrac{n+2}{nh_n} + \tfrac{g_n}{h_n^2}\Big)\theta^2$$

(mit $h_n = \sum_{i=1}^{n-1}\frac{1}{i}$, $g_n = \sum_{i=1}^{n-1}\frac{1}{i^2}$).

Weiterhin ist

$$\widehat{V} := \alpha_1 S_n + \alpha_2 S_n(S_n - 1) \tag{2.28}$$

mit

$$\alpha_1 = \left(\tfrac{n+1}{3(n-1)} - \tfrac{1}{h_n}\right)/h_n, \quad \alpha_2 = \left(\tfrac{2(n^2+n+3)}{9n(n-1)} - \tfrac{n+2}{nh_n} + \tfrac{g_n}{h_n^2}\right)\Big/(h_n^2 + g_n) \tag{2.29}$$

ein erwartungstreuer Schätzer für $\mathrm{Var}_\theta\left[\widehat{\theta}_\pi - \widehat{\theta}_W\right]$.

Die Formel für $\mathrm{Cov}_\theta[S_n, \widehat{\theta}_\pi]$ kann man mittels einer ähnlichen Zerlegung wie im Beweis von Proposition 2.35 beweisen, siehe F. Tajima, a.a.O. Zusammen mit Beobachtung 2.29 und Proposition 2.35 ergibt sich daraus die Formel für $\mathrm{Var}_\theta[\widehat{\theta}_\pi - \widehat{\theta}_W]$.

Aus Beobachtung 2.29 folgt auch

$$\mathbb{E}_\theta[S_n] = \theta h_n \quad \text{und} \quad \mathbb{E}_\theta[S_n(S_n-1)] = \mathrm{Var}_\theta[S_n] + \big(\mathbb{E}_\theta[S_n]\big)^2 - \mathbb{E}_\theta[S_n] = \theta^2 h_n + \theta^2 g_n,$$

d.h. $\mathbb{E}_\theta[\widehat{V}] = \mathrm{Var}_\theta[\widehat{\theta}_\pi - \widehat{\theta}_W]$.

Definition 2.37 (Tajimas D)**.** $D := \dfrac{\widehat{\theta}_\pi - \widehat{\theta}_W}{\sqrt{\widehat{V}}}$ mit $\widehat{V}$ aus (2.28) heißt Tajimas D.

Die Teststatistik D erfüllt $\mathbb{E}_\theta[D] \approx 0$, $\mathrm{Var}_\theta(D) \approx 1$ (die Erwartung ist nicht exakt = 0, da Zähler und Nenner nicht unabhängig sind, die Varianz ist nicht exakt = 1, da $\widehat{V}$ nur ein Schätzer für die Varianz des Zählers ist). Die Formulierung ist (beispielsweise) durch den klassischen t-Test inspiriert: Dort normiert man einen empirischen Mittelwert von n Beobachtungswerten mit dem Standardfehler, einem Schätzer für die Streuung.

Um anhand von D einen statistischen Test zu formulieren, benötigen wir (für ein vorgegebenes Signifikanzniveau α) sogenannte kritische Werte, d.h. geeignete Quantile von D unter der Nullhypothese.

Auf dem Ereignis $\{S_n = s\}$ gilt

$$\widehat{\theta}_W = s/h_n, \quad \widehat{V} = \alpha_1 s + \alpha_2 s(s-1),$$

der kleinste möglicher Wert von $\widehat{\theta}_\pi$ ist dann

$$\frac{1}{\binom{n}{2}} s(n-1) = \frac{2s}{n}.$$

Dies geschieht, wenn $\xi_1^{(n)} + \xi_{n-1}^{(n)} = n$, $\xi_i^{(n)} = 0$ für $2 \le i \le n-2$ gilt (insbesondere, wenn alle Mutationen auf sogenannten externen Kanten – die direkt zu einem Blatt führen – liegen). Der kleinste mögliche Wert von D ist dann somit

$$\frac{2s/n - s/h_n}{\sqrt{\alpha_1 s + \alpha_2 s(s-1)}} \xrightarrow[s\to\infty]{} \frac{2/n - 1/h_n}{\sqrt{\alpha_2}} =: d_{\min} \ \big(= d_{\min}(n)\big).$$

Während ein so kleiner Wert von D unter dem Modell, in dem Mutationen auf den Kingman-Koaleszenten fallen, eher untypisch ist, wäre dies ist in einer „sternförmigen“

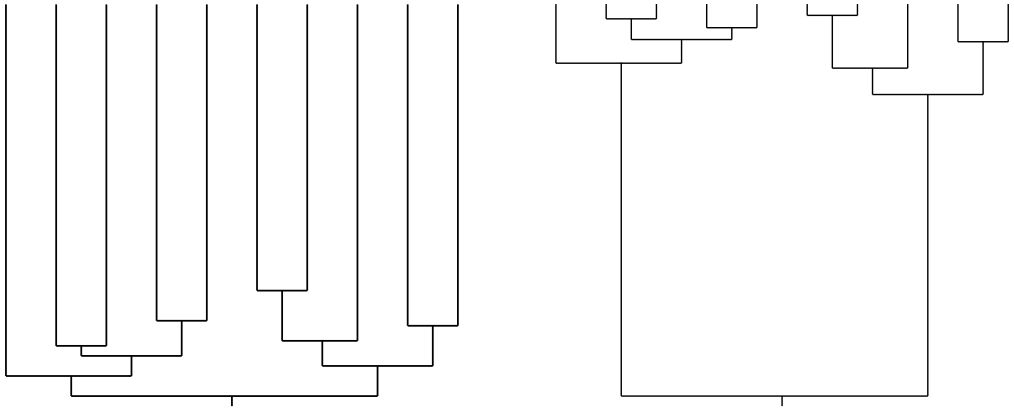


Abbildung 2.6: Ein „sternförmiger“ (links) und ein „Hühnerbein“- (rechts) Koaleszentenbaum

Genealogie, in der die externen Äste die Gesamtlänge des Baumes dominieren (siehe Abbildung 2.6), typisch.

Andererseits ist auf $\{S_n = s\}$ der größte mögliche Wert von $\widehat{\theta}_\pi$

$$\frac{1}{\binom{n}{2}} s\lceil n/2\rceil\lfloor n/2\rfloor = 2s\frac{\lceil n/2\rceil\lfloor n/2\rfloor}{n(n-1)}.$$

Dies geschieht, wenn $\xi^{(n)}_{\lceil n/2\rceil} = n$, $\xi^{(n)}_i = 0$ für $i \neq \lceil n/2\rceil$ (d.h. wenn alle Mutationen auf sehr „balanzierten“ Kanten liegen, die die Blätter in genau zwei Hälften teilen). Der größte mögliche Wert von D ist dann

$$\frac{\frac{2s\lceil n/2\rceil\lfloor n/2\rfloor}{n(n-1)} - s/h_n}{\sqrt{\alpha_1 s + \alpha_2 s(s-1)}} \xrightarrow[s\to\infty]{} \frac{\frac{2\lceil n/2\rceil\lfloor n/2\rfloor}{n(n-1)} - 1/h_n}{\sqrt{\alpha_2}} =: d_{\max} \ \big(= d_{\max}(n)\big)$$

Während ein so kleiner Wert von D unter dem Kingman-Koaleszenten untypisch wäre, wäre dies ist in einer (sehr balanzierten) “Hühnerbein-artigen”-Genealogie, in der zwei innere Äste die Gesamtlänge des Baums dominieren (siehe Abbildung 2.6), typisch.

Im Gegensatz etwa zum klassischen t-Test ist die Verteilung von D unter der Nullhypothese

„die Beobachtungen entstehen durch die Typen an den Blättern eines n-Koaleszenten, längs dessen Kanten sich mit Rate $\theta/2$ Mutationen gemäß IMS-Modell ereignen“ (2.30)

nicht explizit bekannt und hängt von dem unbekannten θ ab.

Tajimas pragmatisch-heuristische Lösung: Approximiere die Verteilung von D durch eine skalierte Beta-Verteilung, so dass der Träger $= [d_{\min}, d_{\max}]$, EW$= 0$ und Var$= 1$ gilt (was recht plausibel passt, siehe Abbildung 2.7): Verwende die approximative Dichte

$$f_{\text{appr}}(d) = \frac{\Gamma(u+v)(d-d_{\min})^{u-1}(d_{\max}-d)^{v-1}}{\Gamma(u)\Gamma(v)(d_{\max}-d_{\min})^{u+v-1}}, \quad d_{\min} < d < d_{\max} \tag{2.31}$$

mit

$$u = \frac{(1 + d_{\max} d_{\min}) d_{\min}}{d_{\max} - d_{\min}}, \quad v = -\frac{(1 + d_{\max} d_{\min}) d_{\max}}{d_{\max} - d_{\min}}. \tag{2.32}$$

(beachte $d_{\min} < 0 < d_{\max}$).

Diese Formeln entspringen dem Ansatz

$$D \approx (d_{\max} - d_{\min}) B + d_{\min} \quad \text{mit} \quad B \sim \text{Beta}(u, v).$$

$\text{Beta}(u, v)$ hat EW $\frac{u}{u+v}$ und Var $\frac{uv}{(u+v)^2(u+v+1)}$, aus dem Ansatz und den geforderten Normierungen ergibt sich

$$\begin{aligned}
\frac{u}{u+v} &= \frac{-d_{\min}}{d_{\max} - d_{\min}} \qquad \Longrightarrow \quad v = u\frac{d_{\max} - d_{\min}}{-d_{\min}} - u = u\frac{d_{\max}}{-d_{\min}}, \\
\frac{uv}{(u+v)^2(u+v+1)} &= \frac{u^2 \frac{d_{\max}}{-d_{\min}}}{u^2\big(1 + \frac{d_{\max}}{-d_{\min}}\big)^2\big(u(1 + \frac{d_{\max}}{-d_{\min}}) + 1\big)} = \frac{-d_{\max} d_{\min}}{(d_{\max} - d_{\min})^2\big(u(1 + \frac{d_{\max}}{-d_{\min}}) + 1\big)} \\
&= \frac{1}{(d_{\max} - d_{\min})^2} \\
\Longrightarrow \quad u &= \frac{-d_{\max} d_{\min} - 1}{1 - \frac{d_{\max}}{d_{\min}}} = \frac{d_{\min}(1 + d_{\max} d_{\min})}{d_{\max} - d_{\min}}, \quad v = -\frac{d_{\max}(1 + d_{\max} d_{\min})}{d_{\max} - d_{\min}},
\end{aligned}$$

woraus sich (2.32) ergibt.

Definition 2.38 (Tajimas Test)**.** Sei $\alpha \in (0, 1)$, $q_{\text{Beta}(u,v)}(\alpha/2)$, $q_{\text{Beta}(u,v)}(1 - \alpha/2)$ das $\alpha/2$- bzw. $(1 - \alpha/2)$-Quantil der $\text{Beta}(u, v)$-Verteilung mit angepassten Parametern u, v aus (2.32).

Lehne H_0 : (2.30) ab, wenn

$$\begin{aligned}
&D < (d_{\max} - d_{\min}) q_{\text{Beta}(u,v)}(\alpha/2) + d_{\min} \qquad \text{oder} \\
&D > (d_{\max} - d_{\min}) q_{\text{Beta}(u,v)}(1 - \alpha/2) + d_{\min}.
\end{aligned}$$

Dieser Test hält (zumindest approximativ) das Signifikanzniveau α ein.

Beispiel. Für die Daten aus Bsp. 2.27 ergibt sich $n = 8$, $s = 31$, $\xi_1^{(8)} = 13$, $\xi_2^{(8)} = 1$, $\xi_7^{(8)} = 17$, somit $\widehat{\theta}_\pi \doteq 7.93$, $\widehat{\theta}_W \doteq 11.96$, $D \doteq -1.79$

Tajimas Approximation liefert ein 95%-Konfidenzintervall für D unter dem Standard-Kingman-Koaleszenten von $[-1.663, 1.975]$, d.h. die Abweichung von 0 ist auf dem 5%-Niveau signifikant (s. Tajima, a.a.O., Table 2, S 592).

Diskussion. In der biologischen Interpretation nennt man Tajimas Test gelegentlich etwas salopp einen „Test auf Neutralität“, da die Nullhypothese (2.30) aus einem Modell ohne Selektion stammt.

Signifikante Abweichungen von $D \approx 0$ legen Alternativhypothesen nahe, unter denen der Baum, der die Stichproben verbindet, eher nicht wie ein „typischer“ Koaleszent aussieht.

Ein signifikant negatives $D < 0$ passt eher zu einem Baum, in dem externe Äste dominieren (Abbildung 2.6, links). Biologische Szenarien, in denen solche Genealogien

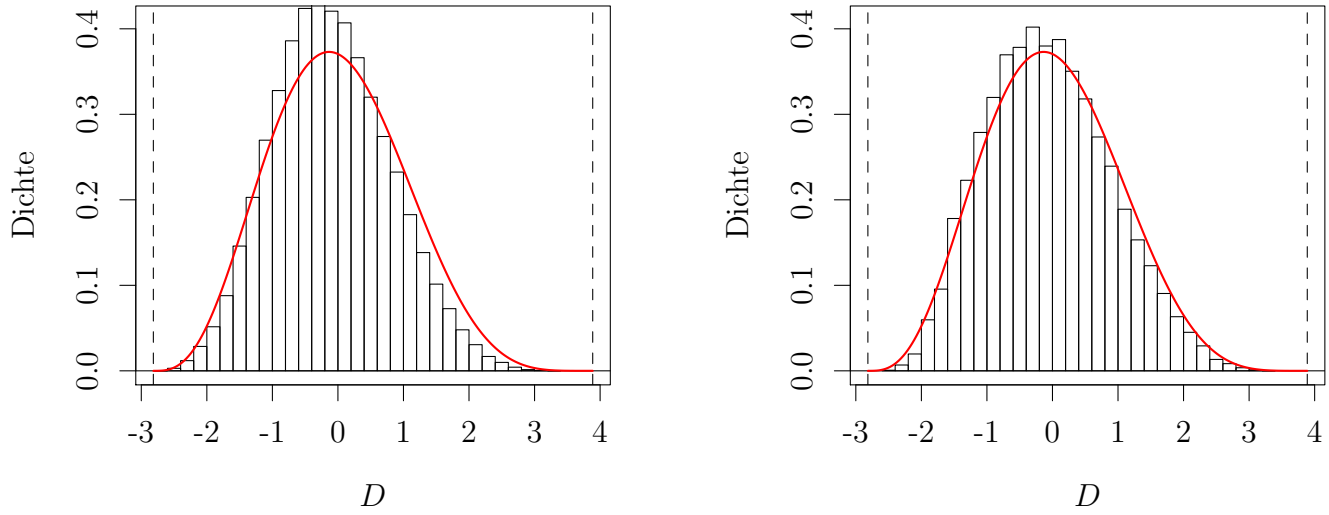


Abbildung 2.7: Simulation der Verteilung von D für $n = 25$ und $\theta = 10$ (links) bzw. $\theta = 2$ (rechts) unter dem Kingman-Koaleszenten mit IMS-Mutationen sowie angepasste skalierte Beta-Dichte aus (2.31). Histogramm jeweils basierend auf 100.000 simulierten Datensätzen.

typisch sind, wären beispielsweise gerichtete Selektion am betrachteten Ort im Genom (oder in dessen „Nähe“, ein sogenannter selektiver “sweep”) oder eine stark wachsende Population.

Ein signifikant positives $D > 0$ passt eher zu einem Baum, in dem wenige interne Äste dominieren (Abbildung 2.6, rechts). Populationsszenarien, in denen solche Genealogien typisch sind, sind beispielsweise (räumlich stark) strukturierte Populationen oder sogenannte balanzierende Selektion (bei der selektive Kräfte gewissermaßen eine genetische Substruktur in der Population aufrecht erhalten).

Eine „exakte“ Version von Tajimas Test

K. L. Simonsen, G. A. Churchill und C. F. Aquadro haben in dem Artikel Properties of statistical tests of neutrality for DNA polymorphism data, *Genetics* 141:413–429, (1995) eine Version von Tajimas Test vorgeschlagen, die ohne die (nicht wörtlich gerechtfertigte) Approximation von D durch eine Beta-Verteilung auskommt[8].

Das unbekannte θ wird dabei als „Störparameter“ (engl. “nuisance parameter”) aufgefasst. Wir wählen $\beta > 0$ (und typischerweise klein) und konstruieren zunächst ein Konfidenzintervall für θ zum Irrtumsniveau β:

Sei für $s \in \mathbb{N}_0$

$$\widehat{\theta}_L(s) = \min\big\{\theta > 0 : \mathbb{P}_\theta(S_n \geq s) > \beta/2\big\}, \quad \widehat{\theta}_R(s) = \max\big\{\theta > 0 : \mathbb{P}_\theta(S_n \leq s) > \beta/2\big\}.$$

Dies ist mittels der Verteilungsfunktion von S_n unter $\mathbb{P}_\theta$ aus Lemma 2.39 unten zumindest numerisch möglich; da diese als Funktion von θ stetig ist, gilt tatsächlich $\mathbb{P}_{\widehat{\theta}_L(s)}(S_n \geq s) = \beta/2$ und $\mathbb{P}_{\widehat{\theta}_R(s)}(S_n \leq s) = \beta/2$ für $s \in \mathbb{N}_0$.

Da $\theta \mapsto \mathbb{P}_\theta(S_n \geq s)$ monoton wachsend in θ ist, gilt

$$\widehat{\theta}_L(s) > \theta \iff \mathbb{P}_\theta(S_n \geq s) \leq \beta/2 \quad \text{und} \quad \widehat{\theta}_R(s) < \theta \iff \mathbb{P}_\theta(S_n \leq s) \leq \beta/2.$$

Damit gilt

$$\forall\, \theta > 0 \,:\, \mathbb{P}_\theta\big([\widehat{\theta}_L(S_n), \widehat{\theta}_R(S_n)] \ni \theta\big) \geq 1 - \beta,$$

denn für $\theta > 0$ ist

$$\begin{aligned}
&\mathbb{P}_\theta\big([\widehat{\theta}_L(S_n), \widehat{\theta}_R(S_n)] \not\ni \theta\big) = \mathbb{P}_\theta\big(\theta < \widehat{\theta}_L(S_n)\big) + \mathbb{P}_\theta\big(\theta > \widehat{\theta}_R(S_n)\big) \\
&= \mathbb{P}_\theta\big(S_n \in \{s : \theta < \widehat{\theta}_L(s)\}\big) + \mathbb{P}_\theta\big(S_n \in \{s : \theta > \widehat{\theta}_R(s)\}\big) \\
&= \sum_{s\,:\,\mathbb{P}_\theta(S_n \geq s) \leq \beta/2} \mathbb{P}_\theta(S_n = s) + \sum_{s\,:\,\mathbb{P}_\theta(S_n \leq s) \leq \beta/2} \mathbb{P}_\theta(S_n = s) \leq \frac{\beta}{2} + \frac{\beta}{2}.
\end{aligned}$$

Dann bestimmt man bei beobachtetem Wert von S_n (mittels Simulation, für θ-Werte aus einem geeignet feinen Gitter in $[\widehat{\theta}_L(S_n), \widehat{\theta}_R(S_n)]$)

$$\begin{aligned}
D_L^* &= \min\Big\{\frac{\alpha}{2}\text{-Quantil von } \mathscr{L}_\theta(D) \,:\, \theta \in [\widehat{\theta}_L(S_n), \widehat{\theta}_R(S_n)]\Big\}, \\
D_L^* &= \max\Big\{\Big(1 - \frac{\alpha}{2}\Big)\text{-Quantil von } \mathscr{L}_\theta(D) \,:\, \theta \in [\widehat{\theta}_L(S_n), \widehat{\theta}_R(S_n)]\Big\}.
\end{aligned}$$

[8] Die Konstruktion verwendet ein allgemeines statistisches Prinzip, siehe R. L. Berger und D. D. Boos, *P* values maximized over a confidence set for the nuisance parameter, *Journal of the American Statistical Association* 89, No. 427, 1012–1016, (1994).

Somit gilt

$$\forall\, \theta > 0 \,:\, \mathbb{P}_\theta\big(D \notin [D_L^*, D_R^*]\big) \le \alpha + \beta,$$

d.h. der Test

$$\text{lehne } H_0 : (2.30) \text{ ab, wenn } D < D_L^* \text{ oder } D > D_R^*$$

hält Niveau $\alpha + \beta$ ein (zumindest theoretisch, wenn man die Quantile im 2. Schritt exakt bestimmen könnte).

Beispiel. Für die Daten aus Bsp. 2.27 ($n = 8$, $D \doteq -1.79$) berichten Simonsen, Churchill und Aquadro, a.a.O., Table 3 gemäß diesem Ansatz ein 95%-Konfidenzintervall für D unter dem Standard-Kingman-Koaleszenten von $[-1.80, 1.83]$ (für $n = 10$, $S_n \in [27, 41]$), d.h. die Abweichung ist „gerade so“ nicht signifikant auf dem 5%-Niveau.

Lemma 2.39 (Explizite Verteilung von S_n)**.** *Es gilt*

$$\mathbb{P}_\theta\,(S_n = m) = \frac{n-1}{\theta} \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-2}{k-1} \Big(\frac{\theta}{\theta+k}\Big)^{m+1}, \quad m \in \mathbb{N}_0,$$

$$\mathbb{P}_\theta\,(S_n \le s) = 1 - \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-1}{k} \Big(\frac{\theta}{\theta+k}\Big)^{s+1}, \quad s \in \mathbb{N}_0.$$

Bemerkung. Dies ist eine Version von Lemma 1.14 (Dichte der Faltung exponentieller ZVn) für den diskreten Fall (Faltung geometrischer ZVn).

Beweis. $S_n = S_{n,n} + S_{n,n-1} + \cdots + S_{n,2}$ mit $S_{n,j} \sim \mathrm{geom}(\frac{i-1}{\theta+i-1})$ u.a.

Sei $u \in [0,1]$: Es ist

$$\mathbb{E}\big[u^{S_{n,j}}\big] = \sum_{\ell=0}^{\infty} u^\ell \tfrac{j-1}{\theta+j-1} \Big(\tfrac{\theta}{\theta+j-1}\Big)^\ell = \tfrac{j-1}{\theta+j-1} \frac{1}{1 - u\frac{\theta}{\theta+j-1}} = \frac{j-1}{j-1+\theta(1-u)},$$

somit

$$\mathbb{E}\big[u^{S_n}\big] = \prod_{j=2}^{n} \mathbb{E}\big[u^{S_{n,j}}\big] = \prod_{k=1}^{n-1} \frac{k}{k+\theta(1-u)}.$$

Weiter ist

$$\prod_{k=1}^{n-1} \frac{k}{k+z} = \sum_{k=1}^{n-1} \frac{a_{n,k}}{k+z} \quad (z \in \mathbb{C} \setminus -\mathbb{N}) \qquad \text{mit } a_{n,k} = \tfrac{(n-1)!}{\prod_{j\ne k}^{n-1}(j-k)} = (-1)^k (n-1)\tbinom{n-2}{k-1},$$

also

$$\mathbb{E}_\theta\big[u^{S_n}\big] = \sum_{m=0}^{\infty} u^m \mathbb{P}_\theta\,(S_n = m) = \sum_{k=1}^{n-1} a_{n,k} \sum_{m=0}^{\infty} \Big(\frac{\theta}{\theta+k}\Big)^m u^m = \sum_{m=0}^{\infty} u^m \sum_{k=1}^{n-1} a_{n,k} \Big(\frac{\theta}{\theta+k}\Big)^m$$

und

$$\begin{aligned}
\mathbb{P}_\theta(S_n \le s) &= \sum_{m=0}^{s} \mathbb{P}_\theta(S_n = m) = \sum_{m=0}^{s} \frac{n-1}{\theta} \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-2}{k-1} \Big(\frac{\theta}{\theta+k}\Big)^{m+1} \\
&= \frac{n-1}{\theta} \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-2}{k-1} \sum_{m=0}^{s} \Big(\frac{\theta}{\theta+k}\Big)^{m+1} \\
&= \frac{n-1}{\theta} \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-2}{k-1} \frac{\theta}{\theta+k} \frac{1-\big(\frac{\theta}{\theta+k}\big)^{s+1}}{1-\big(\frac{\theta}{\theta+k}\big)} \\
&= \frac{n-1}{\theta} \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-2}{k-1} \frac{\theta}{k} \left(1 - \Big(\frac{\theta}{\theta+k}\Big)^{s+1}\right) \\
&= \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-1}{k} \left(1 - \Big(\frac{\theta}{\theta+k}\Big)^{s+1}\right) \\
&= 1 - \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-1}{k} \Big(\frac{\theta}{\theta+k}\Big)^{s+1}
\end{aligned}$$

(denn $-\sum_{k=1}^{n-1}(-1)^k\binom{n-1}{k} = 1 - (1-1)^{n-1} = 1$). □

2.4 Kombinatorik, Likelihoods und anzestrale Inferenz im IMS

Angesichts der Modellannahmen des infinitely-many-sites-Modells stellt sich folgende

Frage: Sind beobachtete Sequenz-Daten $D = (D_{ij})$ mit dem IMS-Modell verträglich, d.h. gibt es einen (gewurzelten) Baum mit n Blättern und s eindeutigen Mutationen (wir können uns die Mutationen als Markierungen auf den Kanten vorstellen oder sie als „innere Knoten“ auffassen), so dass

$$\mathcal{O}_j := \{1 \le i \le n : D_{ij} = 1\} = \{\text{Blätter oberhalb Mutation } j\} \quad \text{für } j = 1, \dots, s \text{ gilt ?} \tag{2.33}$$

Bemerkung. Der Baum muss nicht notwendigerweise binär („vollständig aufgelöst“) sein. Insbesondere kann es mehrere / viele mit Mutationen markierte Koaleszenten-Bäume geben, die zu den Daten passen.

Beispiel 2.40. Folgende Datenmatrix ist mit mehreren Bäumen verträglich

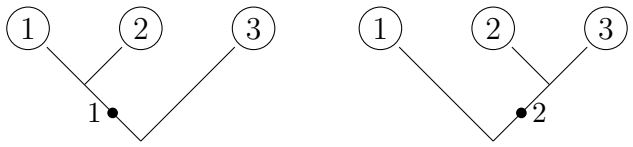
$$\begin{pmatrix} 1 & 1 & . & . \\ 1 & 1 & . & 1 \\ 1 & 1 & . & . \\ 1 & 1 & 1 & . \\ . & . & . & . \end{pmatrix}$$

```
[Skizze an der Tafel,
vgl. auch Abb. 2.8]
```

Die folgende Datenmatrix ist nicht mit einem Baum verträglich:

$$\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix} \tag{2.34}$$

Die erste Spalte zusammen mit der IMS-Annahme erzwingt, dass ein solcher Baum die Struktur des Baums unten links haben müsste, während die zweite Spalte erzwingt, dass er die Struktur des Baums unten rechts haben müsste.



Bedingung 2.41.

$$\forall\, 1 \le j,k \le s \;:\; \mathcal{O}_j \cap \mathcal{O}_k \neq \varnothing \;\Rightarrow\; \Big(\mathcal{O}_j \subset \mathcal{O}_k \text{ oder } \mathcal{O}_k \subset \mathcal{O}_j\Big) \tag{2.35}$$

d.h. D enthält keine Teilmatrix der Form (2.34) und auch keine Zeilenpermutation dieser Form.

Satz 2.42. *D ist mit dem IMS-Modell verträglich, d.h. Frage* (2.33) *ist zu bejahen, g.d.w.* (2.35) *gilt. In diesem Fall kann man den*[9] *„minimalen“ Baum mittels Gusfields Algorithmus konstruieren.*

Offensichtlich gilt (2.35), wenn es einen entsprechenden Baum gibt.

Algorithmus 2.43 (Gusfields Algorithmus[10])**.** Gegeben sei eine $n \times s$-Datenmatrix D mit 0-1-Einträgen.

1. Streiche identische Zeilen und Spalten aus D (für Zeilen: vervielfältige später entspr. Blätter, für Spalten: lege später mehrere Mutationen – in beliebiger Reihenfolge – auf dieselbe Kante)

   (wir nehmen für die weitere Notation an, dass D keine identische Zeilen und Spalten enthält)

2. Sortiere Spalten lexikographisch

   (äquivalent: absteigend als Binärzahlen mit erste Zeile $\widehat{=}$ führendes Bit, d.h. Spalte j entspricht $z_j = \sum_{i=1}^{n} D_{ij} 2^{n-i}$)

3. Für (i,j) mit $D_{ij} = 1$ sei $L_{ij} = \max\big(\{k < j : D_{ik} = 1\} \cup \{0\}\big)$;
   setze $L_j := \max\{L_{ij} : 1 \le i \le n, D_{ij} = 1\}$ $(1 \le j \le s)$

4. Falls $L_{ij} < L_j$ für ein (i,j) mit $D_{ij} = 1$: breche ab

   (die Daten sind nicht mit einem Baum verträglich)

[9] Wir werden sehen, dass es gewisse Uneindeutigkeiten gibt: Für manche Paare von Mutationen legt die Datenmatrix die Reihenfolge nicht per se fest, siehe Schritt 1 in Gusfields Algorithmus.

[10] Aus Dan Gusfield, Efficient algorithms for inferring evolutionary trees, *Networks* 21, 19–28 (1991). Der Kontext dort ist das „Problem der perfekten Phylogenie“ mit binären Charakteren (die sich nicht notwendigerweise auf DNS-Sequenzen beziehen müssen).

5. Erzeuge Knoten $b_1, \dots, b_n$ (für die Blätter), m_0 (für Wurzel), $m_1, \dots, m_s$ (für Mutationen);
   für $j = s, s-1, \dots, 1$: Verbinde m_j mit m_{L_j},
   für $i = 1, \dots, n$: Verbinde b_i mit $m_{c(i)}$, wo $c(i) := \max\big(\{j : D_{ij} = 1\} \cup \{0\}\big)$.

6. Mache ggfs. die Streichungen aus 1. rückgängig, d.h. vervielfältige entsprechende Blätter bzw. Mutationen.

`[Skizze an der Tafel für ein Beispiel]`

Beweisskizze für Satz 2.42. Sei $z_j = \sum_{i=1}^{n} D_{ij} 2^{n-i}$ = Binärzahl, die Spalte j darstellt (nach Sortierung in Schritt 2.), beachte $\mathcal{O}_j \subset \mathcal{O}_k \Rightarrow z_j \le z_k$, d.h. $k \le j$.

Wenn in 4.) abgebrochen wird: Es gibt j, i, i' mit $(j >)\ L_j = L_{ij} =: k > k' := L_{i'j}$, d.h. $\mathcal{O}_j \cap \mathcal{O}_k \neq \varnothing$ (denn $D_{ij} = D_{ik} = 1$), aber $\mathcal{O}_j \not\subset \mathcal{O}_k$ (denn $D_{i'j} = 1$, $D_{i'k} = 0$), wegen $j > k$ kann $\mathcal{O}_k \subset \mathcal{O}_j$ nicht gelten. Demnach ist (2.35) verletzt.

Falls nicht abgebrochen wird, zeige: In 5.) wird ein Baum mit Wurzel m_0 erzeugt.

Da nur Verbindungen $m_j \to m_{L_j}$ mit $L_j < j$ erzeugt werden, gibt es keine Schleifen.

Es ist

$$m_j \to m_k \text{ (direkt)} \quad \Longleftrightarrow \quad \mathcal{O}_j \subsetneq \mathcal{O}_k \text{ und es gibt kein } \ell \text{ mit } \mathcal{O}_j \subsetneq \mathcal{O}_\ell \subsetneq \mathcal{O}_k$$

Stichprobe i hat „größte“ Mutation $c(i)$, demnach: auf dem Pfad von b_i zur Wurzel m_0 kommen alle in Stichprobe i sichtbaren Mutationen vor. Der erzeugte Graph ist somit auch zusammenhängend. □

Bericht 2.44 (Diskussion des ungewurzelten Falls)**.** Wenn die anzestralen Typen nicht bekannt sind und somit die Datenmatrix nur bis auf „Umklappen“ von Spalten bestimmt ist (Vertauschung von 0en und 1en in einer Spalte bedeutet eine Uminterpretation des anzestralen Typs an dieser segregierenden Stelle), dann entsprechen die Daten einem ungewurzelten Baum g.d.w. keine Teilmatrix

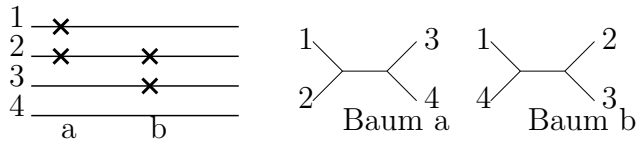
$$\begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

(und auch keine Zeilenpermutation dieser Form) in D vorkommt (im Genetik-Jargon heißt dies die „4-Gameten-Regel“).

In diesem Fall ergibt die Wahl „häufigster Typ“ := 0 in jeder Spalte einen erlaubten gewurzelten Baum. Man kann leicht zeigen, dass mit dieser Wahl dann eine Datenmatrix entsteht, die Bedingung 2.41 erfüllt (siehe auch das Lemma auf S. 135 unten in F.R. McMorris, On the compatibility of binary qualitative taxonomic characters, Bull. Math. Biology 39, no. 2, 133–138, (1977)).

Ein Verschieben der Wurzel im gewurzelten Baum entspricht dem „Umklappen“ gewisser Spalten in D.

Eine Teilmatrix wie oben ist mit keinem Baum zusammen mit dem IMS-Modell verträglich, wie das Bild unten zeigt: Mutation a erzwänge Baum a, während Mutation b Baum b erzwänge.



Wir betrachten im Folgenden nur den Fall bekannter anzestraler Typen. Wenn diese nicht bekannt sind, kann man sich prinzipiell wie in Bericht 2.44 zunächst „künstlich“ anzestrale Typen beschaffen, daraus einen gewurzelten Baum konstruieren und dann beispielsweise in den unten folgenden Rechnungen für die Wahrscheinlichkeiten gewisser Sequenzbeobachtungen über alle Möglichkeiten aussummieren, wie die Position der Wurzel im Baum „verschoben“ werden kann.

Aus Gusfields Algorithmus ergibt sich:

Beobachtung 2.45 (Alternative Darstellung von D)**.** Sei D eine $n \times s$-Datenmatrix, die mit IMS verträglich ist, d.h., die (2.35) aus Bedingung 2.41 erfüllt. Dann ist

$$D \mathrel{\widehat{=}} (\mathbf{x}_1, \ldots, \mathbf{x}_n)$$

$$\text{mit } \mathbf{x}_i = (x_{i,0}, x_{i,1}, \ldots, x_{i,j(i)}) \text{ geordnete Liste mit Einträgen aus } \mathbb{N}_0 \text{ (bzw. aus } \{0, 1, \ldots, s\},$$

$\mathbf{x}_i$ ist die geordnete Liste von Mutationen, die auf dem Pfad zwischen Blatt i und der Wurzel liegen, mit Def. $x_{i,j(i)} = 0$ „Wurzelmutation“), wobei gilt:

1. die Einträge in $\mathbf{x}_i$ sind paarw. versch.,
2. $x_{i,j} = x_{i',j'} \Rightarrow x_{i,j+k} = x_{i',j'+k}$ für $k \geq 0$,
3. jedes Paar $\mathbf{x}_i$, $\mathbf{x}_{i'}$, besitzt mind. einen gemeinsamen Eintrag.

Bedingungen 1. und 2. besagen, dass man die $\mathbf{x}_i$ als schleifenfreie Pfade interpretieren kann, die „zusammenlaufen“, sobald sie sich treffen, d.h. die Setzung „verbinde Mutationen s und s', wenn $s = x_{i,k}$, $s' = x_{i,k+1}$ für ein i und k“ definiert einen Wald, dessen Knoten die Mutationen bilden. Bedingung 3. garantiert, dass dieser zusammenhängend ist und somit ein (gewurzelter) Baum.

Umgekehrt kann man aus $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, das diesen Bedingungen genügt, auch stets eine eindeutige Datenmatrix rekonstruieren, die (2.35) aus Bedingung 2.41 erfüllt.

`[Skizze an der Tafel für ein Beispiel]`

Definition 2.46. Sei $n \in \mathbb{N}$, $\mathcal{T}_n$ die Menge der n-Stichproben (bzw. zug. Datenmatrizen) $\mathbf{x}$ aus Beob. 2.45, die den dort genannten Bedingungen genügen.

Für $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n), \mathbf{x}' = (\mathbf{x}'_1, \ldots, \mathbf{x}'_n) \in \mathcal{T}_n$ schreiben wir

$$\mathbf{x} \sim \mathbf{x}' :\Longleftrightarrow \exists \text{ Permutation } \xi : \mathbb{N}_0 \to \mathbb{N}_0 \text{ so dass } x_{i,j} = \xi(x'_{i,j}) \text{ für alle } i, j$$

$\sim$ beschreibt Äquivalenz von Daten unter Umnummerierung von Mutationen (beachte, dass die Mutationsnummern „künstlich“ sind und alle W'keiten invariant unter Permutationen derselben).

Wir bezeichnen mit

$$(\mathcal{T}_n/\sim) \text{ die Äquivalenzklassen unter } \sim$$

und mit

$$(\mathcal{T}_n/\sim)_0 \subset (\mathcal{T}_n/\sim)$$

die Teilmenge, bei der alle n Stichproben einen verschiedenen Typ haben (beachte, dass dies nicht von der Wahl des Repräsentanten bgzl. $\sim$ abhängt).

Für $\mathbf{x} \in \mathcal{T}_n$ bezeichnet

$$\begin{aligned} s &= s(\mathbf{x}) \text{ die Anzahl Mutationen und} \\ d &= d(\mathbf{x}) = \left|\{\mathbf{x}_i : i = 1, \dots, n\}\right| \text{ die Anzahl verschiedener Typen.} \end{aligned}$$

Mit etwas Missbrauch der Notation definieren wir dies auch für $\mathbf{x} \in \mathcal{T}_n/\sim$ ($s(\mathbf{x})$ und $d(\mathbf{x})$ hängen nicht von der Wahl Repräsentanten bezüglich $\sim$ ab).

Eine Stichprobe der Größe n mit s Mutationen hat *Komplexität* $n + s - 1$.

$\mathbf{x} \in (\mathcal{T}_n/\sim)$ mit d verschiedenen (angeordneten) Typen ist äquivalent beschrieben / parametrisiert als ein Paar

$$\mathbf{t} \in (\mathcal{T}_d/\sim)_0,\ \mathbf{a} = \big(A_1, \dots, A_d\big)$$

mit $A_i = \{j : \mathbf{t}_i = \mathbf{x}_j\}$ für $i = 1, \dots, d$. $(A_1, \dots, A_d)$ ist somit eine geordnete Partition von $\{1, \dots, n\}$, d.h. $A_i \cap A_j = \varnothing\ \forall i \neq j$ and $\bigcup_{i=1}^{d} A_i = \{1, \dots, n\}$.

Wir wählen / fixieren damit, ggfs. implizit, eine gewisse Reihenfolge / Nummerierung der Typen. Wir können dabei an eine zufällige Wahl der Typennummerierung oder an die Wahl „Typ 1=Typ von Stichprobe 1, Typ 2=Typ derjenigen Stichprobe mit der kleinsten Nummer, die nicht Typ 1 hat, etc.“ oder an irgend ein anderes Schema denken.

Wir gehen dann von $(\mathbf{t}, \mathbf{a})$ über zu $(\mathbf{t}, \mathbf{n})$ mit

$$\mathbf{t} = (\mathbf{t}_1, \dots, \mathbf{t}_d),\ \mathbf{n} = (n_1, \dots, n_d)$$

mit $n_i = |A_i|$. (Dies liegt nahe, da auch die Nummerierung der Stichproben „künstlich“ ist und alle W'keiten invariant sind unter Permutationen derselben).

Offenbar gibt es $\binom{n}{n_1, \dots, n_d} = n!/(n_1! \cdots n_d!)$ viele Wahlen von $\mathbf{a} = (A_1, \dots, A_d)$ zu gegebenem $\mathbf{n} = (n_1, \dots, n_d)$ mit $n_1 + \cdots + n_d = n$.

$$\mathcal{T}^* := \bigcup_{d=1}^{\infty} \big((\mathcal{T}_d/\sim)_0 \times \mathbb{N}^d\big)$$

nennen wir die Menge aller (unnummerierten, geordneten) „Gen-Bäume“.

`[Siehe Tafel für Beispiele]`

Bemerkung 2.47. Man kann prinzipiell auch Mutations- und Stichprobennummern „in einem Schritt herausfaktorisieren“. Dazu betrachtet man für $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n), \mathbf{x}' = (\mathbf{x}'_1, \dots, \mathbf{x}'_n) \in \mathcal{T}_n$

$$\mathbf{x} \approx \mathbf{x}' \ :\Longleftrightarrow \ \begin{array}{l}\text{wenn es eine Permutation } \sigma \in S_n \text{ und} \\ \text{eine Permutation } \zeta\colon \mathbb{N}_0 \to \mathbb{N}_0 \text{ gibt mit } x'_{ij} = \zeta(x_{\sigma(i)j}).\end{array}$$

Wir notieren die Äquivalenzklasse von $\mathbf{x} \in \mathcal{T}_n$ modulo $\approx$ als $[\mathbf{x}]_\approx$ Wenn in $\mathbf{x} \in \mathcal{T}_n$ d verschiedene Typen vorkommen und somit $\mathbf{x}$ in der Parametrisierung aus Def. 2.46 einem $(\mathbf{t}, \mathbf{n}) \in (\mathcal{T}_d/\sim)_0 \times \mathbb{N}^d$ (mit einer gewissen Anordnung der d Typen und $n_1 + \dots + n_d = n$) entspricht, notieren wir die Äquivalenzklasse auch als $[\mathbf{t}, \mathbf{n}]_\approx$. Demnach ist für $(\mathbf{t}, \mathbf{n}), (\mathbf{t}', \mathbf{n}') \in (\mathcal{T}_d/\sim)_0 \times \mathbb{N}^d$

$$[\mathbf{t}, \mathbf{n}]_\approx = [\mathbf{t}', \mathbf{n}']_\approx \iff \begin{array}{l}\text{wenn es eine Permutation } \sigma \in S_d \text{ und eine} \\ \text{Permutation } \zeta\colon \mathbb{N}_0 \to \mathbb{N}_0 \text{ gibt mit } t'_{ij} = \zeta(t_{\sigma(i)j}) \text{ und } n'_i = n_{\sigma(i)}.\end{array}$$

Beachte, dass in $[\mathbf{t}, \mathbf{n}]_\approx$ die Reihenfolge der Typen keine Rolle spielt.

Für $(\mathbf{t}, \mathbf{n}) \in (\mathcal{T}_d/\sim)_0 \times \mathbb{N}^d$ sei

$$c(\mathbf{t}, \mathbf{n}) := \left|\left\{\sigma \in S_d : \mathbf{t} \sim \mathbf{t}_\sigma \text{ und } \mathbf{n} = \mathbf{n}_\sigma\right\}\right| \tag{2.36}$$

mit $\mathbf{n}_\sigma = (n_{\sigma(1)}, \dots, n_{\sigma(d)})$, $\mathbf{t}_\sigma = (t_{\sigma(1)}, \dots, t_{\sigma(d)})$ die Anzahl Umordnungen der Typen, die im Sinne von $\sim$ auf dieselbe Stichprobe führen (dies hängt nur von der Äquivalenzklasse modulo $\approx$ ab). Dann gibt es $d!/c(\mathbf{t}, \mathbf{n})$ Elemente von $[\mathbf{t}, \mathbf{n}]_\approx$, die als geordnete Gen-Bäume, d.h. bezüglich Äquivalenz modulo $\sim$, unterschiedlich sind.

`[Siehe Tafel für ein Beispiel]`

Wir bleiben für das Weitere bei angeordneten Typen, da diese für die folgendenden Argumente und Rechnungen angenehmer sind und uns insbesondere sonst auftretende kombinatorische Korrekturfaktoren ersparen.

2.4.1 Wahrscheinlichkeiten von Beobachtungen

Sei $\theta > 0$, $n \in \mathbb{N}$ und D_n Sequenzbeobachtungen an den Blättern eines n-Koaleszenten, längs dessen Kanten sich mit Rate $\theta/2$ Mutationen gemäß dem IMS-Modell ereignen (wir notieren das zugrundeliegende Wahrscheinlichkeitsmaß als $\mathbb{P}_{n,\theta}$, wenn die Parameter nicht aus dem Kontext klar sind).

In diesem Abschnitt geht es darum,

$$p_\theta(\mathbf{t}, \mathbf{n}) := \mathbb{P}_{n,\theta}\big(D_n \sim (\mathbf{t}, \mathbf{n})\big)$$

zu bestimmen (diese Aufgabe entspricht gewissermaßen der Suche nach einem Ersatz für die Ewens'sche Stichprobenformel, Satz 2.17, für das IMS-Modell). Wir nummerieren hierbei die verschiedenen Typen in D_n zufällig und parametrisieren D_n gemäß Beob. 2.45. Beachte: $\mathbf{n} = (n_1, \dots, n_d)$ fixiert implizit die Stichprobengröße als $n = n_1 + \dots + n_d$, ansonsten ist die rechte Seite = 0.

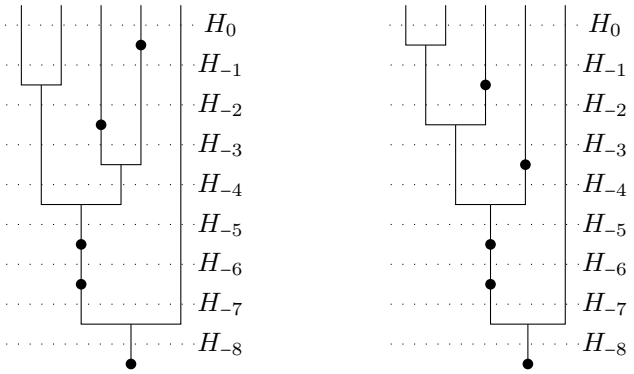


Abbildung 2.8: Zwei mögliche Historien, die beide zu den beobachteten Daten (modulo $\sim$) $(\mathbf{t}, \mathbf{n}) = \big(\big(\,(2,1,0),(3,2,1,0),(4,2,1,0),(0)\,\big),\,(2,1,1,1)\big)$ aus Beispiel 2.40 führen

Definition 2.48. Wir interpretieren die Genealogie mit IMS-Mutationen darauf als „Historie“, d.h. als Abfolge der Zustände in der Geschichte der Stichprobe / im markierten Koaleszenten-Baum zwischen jüngstem gemeinsamem Vorfahren und Gegenwart

$$\mathcal{H} = \big(H_{-\tau+1} = ((1),(0)), H_{-\tau+2}, \ldots, H_{-1}, H_0\big)$$

(wir notieren also die zeitliche Abfolge der Zustände, nicht aber die tatsächlichen reellwertigen Astlängen im Koaleszenten-Baum).

In dieser Parametrisierung gibt es τ Zustände in der Historie; wenn wir $H_0 = (\mathbf{t}, \mathbf{n})$ kennen, so ist τ ist durch die Komplexität von $(\mathbf{t}, \mathbf{n})$ festgelegt (denn in jedem Schritt erhöht sich die Komplexität um 1 – es wird entweder eine weitere Stichprobe oder eine weitere Mutation hinzugefügt).

Bemerkung. Es können (u.U. sehr viele) verschiedene Historien zur selben Stichprobenkonfiguration $(\mathbf{t}, \mathbf{n})$ führen. Siehe Abbildung 2.8 für ein Beispiel.

Die Wahrscheinlichkeiten von Sequenzbeobachtungen im IMS-Modell, $p_\theta(\mathbf{t}, \mathbf{n})$, erfüllen die folgende Rekursion:

Proposition 2.49. *Es gilt für* $(\mathbf{t}, \mathbf{n}) = \big((\mathbf{t}_1, \ldots, \mathbf{t}_d),\,(n_1, \ldots, n_d)\big) \in \mathcal{T}^*$

$$\begin{aligned} p_\theta(\mathbf{t}, \mathbf{n}) = {} & \frac{\binom{n}{2}}{n\theta/2 + \binom{n}{2}} \sum_{i\,:\,n_i \geq 2} \frac{n_i - 1}{n-1} p_\theta\big(\mathbf{t}, \mathbf{n} - \mathbf{e}_i\big) \\ & + \frac{\theta/2}{n\theta/2 + \binom{n}{2}} \sum_{\substack{i\,:\,n_i = 1,\, t_{i,0} \text{ einzig}, \\ \mathfrak{s}(\mathbf{t}_i) \neq \mathbf{t}_j\, \forall\, j \neq i}} p_\theta\big(\mathfrak{s}_i(\mathbf{t}), \mathbf{n}\big) \\ & + \frac{\theta/2}{n\theta/2 + \binom{n}{2}} \frac{1}{d} \sum_{\substack{i\,:\,n_i = 1, \\ t_{i,0} \text{ einzig}}} \sum_{j\,:\,\mathfrak{s}(\mathbf{t}_i) = \mathbf{t}_j} (n_j + 1) p_\theta\big(\mathfrak{r}_i(\mathbf{t}), \mathfrak{r}_i(\mathbf{n} + \mathbf{e}_j)\big) \end{aligned} \tag{2.37}$$

mit Randbedingung $p_\theta\big(\,(\,(0)\,),\,(1)\big) = 1$.

Hierbei bedeutet „$t_{i,0}$ einzig“, dass die „vorderste“ (d.h. die jüngste) Mutation $t_{i,0}$ von Typ

$$\mathbf{t}_i = (t_{i,0}, t_{i,1}, \ldots, t_{i,j(i)})$$

nur einmal in der gesamten Stichprobe vorkommt (d.h. in keinem der anderen Typen $\mathbf{t}_j$, $j \neq i$ enthalten ist – diese Information ist [auch] unter $\sim$ wohldefiniert);

$\mathfrak{s}(\mathbf{t}_i)$ *entfernt die vorderste Mutation aus Typ i, d.h.*

$$\mathfrak{s}(\mathbf{t}_i) = (t_{i,1}, t_{i,2}, \ldots, t_{i,j(i)}),$$

und $\mathfrak{s}_i(\mathbf{t})$ entfernt im Typenvektor $\mathbf{t}$ im i-ten Typ die vorderste Mutation.

$$\mathfrak{s}_i(\mathbf{t}) = \big(\mathbf{t}_1, \ldots, \mathbf{t}_{i-1}, (t_{i,1}, t_{i,2}, \ldots, t_{i,j(i)}), \mathbf{t}_{i+1}, \ldots, \mathbf{t}_d\big);$$

$\mathfrak{r}_i(\cdot)$ *entfernt den i-ten Typ, d.h.*

$$\mathfrak{r}_i(\mathbf{t}) = (\mathbf{t}_1, \ldots, \mathbf{t}_{i-1}, \mathbf{t}_{i+1}, \ldots, \mathbf{t}_d), \quad \mathfrak{r}_i(\mathbf{n} + \mathbf{e}_j) = (n_1, \ldots, n_{i-1}, n_{i+1}, \ldots, n_j + 1, \ldots, n_d)$$

(bzw. wörtlich $\mathfrak{r}_i(\mathbf{n} + \mathbf{e}_j) = (n_1, \ldots, n_j + 1, \ldots, n_{i-1}, n_{i+1}, \ldots, n_d)$ falls $j < i$).

(2.37) *ist strikt rekursiv bezüglich der Komplexität der Stichprobe: Jeder der Terme auf der rechten Seite enthält entweder eine Stichprobe oder eine Mutation weniger als $(\mathbf{t}, \mathbf{n})$.*

Beweisskizze. Wir zerlegen nach dem „jüngsten“ Ereignis (das am nächsten an den Blättern liegt) im markierten Koaleszenten, das den Übergang von Zustand H_{-1} zu Zustand H_0 in der Historie bescheibt (vergleiche auch den Gedankengang rund um Beob. 2.5, dort hatten wir die analoge Argumentation im 2 Typ-Fall betrachtet; formell nutzen wir die Markoveigenschaft des Kingman-Koaleszenten zusammen mit der Markoveigenschaft der Mutations-Poissonprozesse auf den Kanten aus):

- Mit W'keit $\frac{\binom{n}{2}}{n\theta/2+\binom{n}{2}}$ ist das jüngste Ereignis eine Verschmelzung, damit $D_n \sim (\mathbf{t}, \mathbf{n})$ gilt, muss es notwendigerweise eine Verschmelzung innerhalb eines der d Typen sein und dieser Typ muss notwendigerweise Vielfacheit ≥ 2 haben. Vom Zustand H_{-1} vor dem Ereignis aus gesehen bedeutet dies, dass einer der d Typen verzweigt, dies betrifft Typ i mit W'keit $\frac{n_i-1}{n-1}$ (es gab vor dem Ereignis $n-1$ Linien, $n_i - 1$ davon vom Typ i) und dann ist die Konfiguration vor dem Ereignis $\big(\mathbf{t}, \mathbf{n} - \mathbf{e}_i\big)$.

- Mit W'keit $\frac{n\theta/2}{n\theta/2+\binom{n}{2}}$ ist das jüngste Ereignis eine Mutation, diese muss notwendigerweise die äußerste Mutation auf einem Typ mit Vielfachheit 1 in $(\mathbf{t}, \mathbf{n})$ sein und diese Mutation darf nur einmal in der gesamten Stichprobe vorkommen.

  - Wenn diese jüngste Mutation Typ i betraf und Typ i auch nach Entfernen dieser Mutation immer noch eindeutig ist (d.h. $\mathfrak{s}(\mathbf{t}_i) \neq \mathbf{t}_j \forall j \neq i$), so muss die Mutation genau die eine Linie vom Typ i betroffen haben – dies hat W'keit $\frac{1}{n}$ – und die Konfiguration vor dem Ereignis ist dann $\big(\mathfrak{s}_i(\mathbf{t}), \mathbf{n}\big)$.
  - Wenn diese jüngste Mutation (den aktuellen) Typ i betraf und Typ i nach Entfernen dieser Mutation mit (dem aktuellen) Typ j (in $\mathbf{t}$) zusammenfällt, so muss die Mutation Typ j (bzw. den Typ, der vor dem Mutationsereignis Typ

j heißt) betroffen haben, dies hat W'keit $\frac{n_j+1}{n}$ (vor dem Mutationsereignis gab es n_j+1 Linien dieses Typs) und der neu erzeugte Typ muss bei der zufälligen Anordnung der Typen die Position i erhalten haben, dies liefert einen weiteren Faktor $\frac{1}{d}$.

Die Randbedingung ergibt sich aus der Tatsache, dass es im Fall $n = 1$ nur eine mögliche Stichprobenkonfiguration gibt. □

Wörtlich entstehen in unserem Modell die Sequenzbeobachtungen in einem zwei-Schritt-Verfahren: Es wird zuerst ein n-Koaleszent erzeugt (die Genealogie, die die Stichproben verbindet, unabhängig von deren Typen), dann werden, gegeben diesen Baum, Mutationen (im IMS-Modell) längs den Ästen gemäß Poisson-Prozessen mit Rate $\theta/2$ verteilt und die sich an den Blättern ergebenden (Sequenz-)Typen abgelesen. Analog zur Hoppe-Urne (Def. 2.13) im Fall des IMA-Modells gibt es auch hier ein Verfahren, das nur „einen Durchlauf" benötigt und gewissermaßen den zufälligen Baum samt Mutationen „von der Wurzel her" konstruiert:

Definition 2.50 („Ethier-Griffiths[11]-Urne")**.** Das folgende Verfahren generiert eine zufällige n-Stichprobe im IMS-Modell:

- Beginne mit 2 Blättern.
- Wenn aktuell ℓ Blätter:
  - mit W'keit $\dfrac{\theta}{\ell-1+\theta}$: Füge einem Blatt eine neue Mutation hinzu.
  - mit W'keit $\dfrac{\ell-1}{\ell-1+\theta}$: Verdopple ein Blatt.

  (Das Blatt wird jeweils uniform aus den k Blättern gewählt.)
- Stoppe sobald $n+1$ Blätter erreicht, entferne das zuletzt erzeugte Blatt.

  Gebe den so erhaltenen Zustand als Ausgabe zurück. (Wenn wir die Ausgabe als ein Element von $\mathcal{T}^*$, d.h. als Gen-Baum mit angeordneten Typen, auffassen möchten, so werden die Typen in rein zufälliger Reihenfolge ausgegeben.)

`[Skizze an der Tafel für ein Beispiel]`

Für eine vorgegebene „Ziel"-Stichprobengröße $n \geq 2$ können wir die Ethier-Griffiths-Urne als eine Markovkette $(U_k)_{k=0,1,\ldots}$ auf $\mathcal{T}^* \cup \{\partial\}$ mit Startzustand $U_0 = \big((0),(2)\big) \in \mathcal{T}^*$ auffassen. In dieser Form betrachten wir die Typen in den Zuständen U_k als *geordnet* und wir nehmen an, dass, wenn im Verlauf des Verfahrens bei aktuell d Typen eine neue Mutation hinzugefügt wird – und somit ein neuer Typ entsteht –, die Position des neuen Typs in der Anordnung jeweils unabhängig und uniform aus allen $d+1$ Möglichkeiten gewählt wird. ∂ stellt einen zusätzlichen „Friedhofszustand" dar, den wir verwenden, um die Abbruchbedingung zu formalisieren. Mit $\widehat{\tau} := \min\{k : U_k = \partial\}$ ist $U_{\widehat{\tau}-1}$ der Ausgabezustand des Verfahrens.

[11] Aus S.N. Ethier, R.C. Griffiths, The infinitely-many-sites model as a measure-valued diffusion, Ann. Probab. 15 (1987), no. 2, 515–545

Für $(\mathbf{t}', \mathbf{n}') = \big((\mathbf{t}'_1, \dots, \mathbf{t}'_d), (n_1, \dots, n_d)\big) \in (\mathcal{T}_d/\!\sim) \times \mathbb{N}^d$ mit Stichprobengröße $n' = n'_1 + \dots + n'_d$ sind die Übergangswahrscheinlichkeiten demnach

$$\mathbb{P}_{n,\theta}\big(U_{k+1} = u \,\big|\, U_k = (\mathbf{t}', \mathbf{n}')\big)$$
$$= \begin{cases} \dfrac{n'-1}{n'-1+\theta}\dfrac{n'_i}{n'}, & \text{falls } n' < n \text{ und } u = (\mathbf{t}', \mathbf{n}' + \mathbf{e}_i) \text{ für } i = 1, \dots, d, \\ \dfrac{n'-1}{n'-1+\theta}, & \text{falls } n' = n \text{ und } u = \partial, \\ \dfrac{\theta}{n'-1+\theta}\dfrac{1}{n'}, & \text{falls } n'_i = 1 \text{ und } u = \big(\mathfrak{a}_i(\mathbf{t}'), \mathbf{n}'\big) \text{ für } i = 1, \dots, d, \\ \dfrac{\theta}{n'-1+\theta}\dfrac{n'_i}{n'}\dfrac{1}{d+1}, & \text{falls } n'_i > 1 \text{ und } u = \big(\mathfrak{e}_{i,j}(\mathbf{t}'), \mathfrak{e}_j(\mathbf{n}' - \mathbf{e}_i)\big) \\ & \text{für } i = 1, \dots, d \text{ und} \\ & j \in \{1, \dots, d+1\} \setminus \{k, k+1 : k \in \mathsf{dirNachf}_i(\mathbf{t}')\}, \\ \dfrac{\theta}{n'-1+\theta}\dfrac{n'_i}{n'}\dfrac{r_{w,i}(\mathbf{t}') - \ell_{w,i}(\mathbf{t}') + 1}{d+1}, & \text{falls } n'_i > 1 \text{ und } u = \big(\mathfrak{e}_{i,j}(\mathbf{t}'), \mathfrak{e}_j(\mathbf{n}' - \mathbf{e}_i)\big) \text{ für} \\ & i = 1, \dots, d \text{ und } j = \ell_{w,i}(\mathbf{t}'),\ w = 1, \dots, m(\mathbf{t}', i), \\ 0, & \text{sonst.} \end{cases} \tag{2.38}$$

Hierbei fügt $\mathfrak{a}_i(\mathbf{t}')$ eine neue Mutation zu Typ i hinzu, d.h. wenn es s Mutationen in $\mathbf{t}'$ gibt, ist

$$\mathfrak{a}_i(\mathbf{t}') = \big(\mathbf{t}'_1, \dots, \mathbf{t}'_{i-1}, (s+1, t'_{i,0}, t'_{i,1}, \dots, t'_{i,j(i)}), \mathbf{t}'_{i+1}, \dots, \mathbf{t}'_d\big);$$

$\mathfrak{e}_{i,j}(\mathbf{t}')$ kopiert Typ i, fügt eine neue Mutation hinzu und setzt den resultierenden Typ an Position j [vor dem aktuell j-ten Typ, falls $j \le d$ bzw. ans Ende der Liste, falls $j = d+1$] in $\mathbf{t}$ ein, d.h.

$$\mathfrak{e}_{i,j}(\mathbf{t}') = \big(\mathbf{t}'_1, \dots, \mathbf{t}'_{j-1}, (s+1, t'_{i,0}, t'_{i,1}, \dots, t'_{i,j(i)}), \mathbf{t}'_j, \mathbf{t}'_{j+1}, \dots, \mathbf{t}'_d\big)$$

wenn es s Mutationen in $\mathbf{t}'$ gibt; der Typenhäufigkeitsvektor

$$\mathfrak{e}_j(\mathbf{n}') = \mathfrak{e}_j(n'_1, \dots, n'_d) := (n'_1, \dots, n'_{j-1}, 1, n'_j, \dots, n'_d)$$

entsteht, indem ein neuer Typ mit Häufigkeit 1 an Position j eingefügt wird.

$$\mathsf{dirNachf}_i(\mathbf{t}') := \big\{1 \le k \le d : n_k = 1 \text{ und } \mathfrak{s}(\mathbf{t}'_k) = \mathbf{t}'_i\big\}$$

(mit $\mathfrak{s}(\cdot)$ wie in Proposition 2.49) sind diejenigen Typen, die „direkte Nachfahren“ des i-ten Typs mit Vielfachheit 1 sind. Wir nennen Typ k einen direkten Nachfahren von Typ i (in $\mathbf{t}'$), wenn Typ k genau eine (notwendigerweise die äußerste) Mutation mehr besitzt als i. Falls $\mathsf{dirNachf}_i(\mathbf{t}') \neq \varnothing$, sei

$$\{k, k+1 : k \in \mathsf{dirNachf}_i(\mathbf{t}')\}$$
$$= \big([\ell_{1,i}(\mathbf{t}'), r_{1,i}(\mathbf{t}')] \cup [\ell_{2,i}(\mathbf{t}'), r_{2,i}(\mathbf{t}')] \cup \dots \cup [\ell_{m(\mathbf{t}',i),i}(\mathbf{t}'), r_{m(\mathbf{t}',i),i}(\mathbf{t}')]\big) \cap \mathbb{N}$$

mit $m(\mathbf{t}', i) \in \mathbb{N}$ und

$$1 \le \ell_{1,i}(\mathbf{t}') \le r_{1,i}(\mathbf{t}'),\ r_{1,i}(\mathbf{t}') + 1 < \ell_{2,i}(\mathbf{t}') \le r_{2,i}(\mathbf{t}'),\ \dots$$
$$\dots,\ r_{m(\mathbf{t}',i)-1,i}(\mathbf{t}') + 1 < \ell_{m(\mathbf{t}',i),i}(\mathbf{t}') \le r_{m(\mathbf{t}',i),i}(\mathbf{t}') \le d+1$$

die disjunkte Zerlegung von $\{k, k+1 : k \in \mathsf{dirNachf}_i(\mathbf{t}')\}$ in $m(\mathbf{t}', i)$ maximale (diskrete) Intervalle.

Für die Übergangswahrscheinlichkeit in der fünften und vierten Zeile der rechten Seite von (2.38) beachten wir: Mit W'keit $\frac{\theta}{n'-1+\theta}$ wird eine neue Mutation hinzugefügt, mit W'keit $\frac{n_i'}{n'}$ betrifft dies ein Typ i-Blatt. Wenn $n_i' > 1$, so entsteht damit ein neuer Typ, der an prinzipiell $d+1$ Positionen unter den aktuell d Typen eingefügt werden kann. Es ist (da wir Typen unter der Äquivalenzrelation $\sim$ betrachten)

$$\big(\mathfrak{e}_{i,k}(\mathbf{t}'), \mathfrak{e}_k(\mathbf{n}' - \mathbf{e}_i)\big) = \big(\mathfrak{e}_{i,k+1}(\mathbf{t}'), \mathfrak{e}_{k+1}(\mathbf{n}' - \mathbf{e}_i)\big) \quad \text{für jedes } k \in \mathsf{dirNachf}_i(\mathbf{t}).$$

Da der neue Typ nach Definition an eine zufällig ausgewählte Position gesetzt wird, gibt es $r_{w,i}(\mathbf{t}') - \ell_{w,i}(\mathbf{t}') + 1$ Wahlen für diese Position, die alle auf den neuen Zustand

$$\big(\mathfrak{e}_{i,\ell_{w,i}(\mathbf{t}')}(\mathbf{t}'), \mathfrak{e}_{\ell_{w,i}(\mathbf{t}')}(\mathbf{n}' - \mathbf{e}_i)\big)$$

führen. Für Positionen $j \in \{1, \dots, d+1\} \setminus \{k, k+1 : k \in \mathsf{dirNachf}_i(\mathbf{t}')\}$ gibt es solche Uneindeutigkeiten nicht, in diesem Fall legt der „Zielzustand" eindeutig fest, welche Stelle für den neuen Typ zufällig ausgewählt wurde.

Die Ausdrücke und Argumente für die erste bis dritte Zeile von (2.38) sind übersichtlicher:

Um von $(\mathbf{t}', \mathbf{n}')$ nach $\big(\mathbf{t}', \mathbf{n}' + \mathbf{e}_i\big)$ zu gelangen muss ein „Blattverdopplungsereignis" stattfinden – dies hat W'keit $\frac{n'-1}{n'-1+\theta}$ – und das gewählte Blatt muss vom Typ i sein – dies hat W'keit $\frac{n_i'}{n'}$; wenn bereits $n' = n$ gilt, so führt ein Blattverdopplungsereignis definitionsgemäß zum Ende des Verfahrens; wenn ein Mutationsereignis stattfindet – dies hat W'keit $\frac{\theta}{n'-1+\theta}$ – und Typ i, der Vielfachheit $n_i' = 1$ hat, gewählt wird – dies hat W'keit $\frac{1}{n'}$ – so erhält Typ i eine weitere Mutation, aber es entsteht kein neuer Typ.

Lemma 2.51. *Die Verteilung der Abfolge der Zustände der Ethier-Griffiths-Urne entspricht der der Historien unter $\mathbb{P}_{n,\theta}$ (bis auf eine Verschiebung des Startzustands), d.h.*

$$\big(U_0, U_1, \dots, U_{\widehat{\tau}-1}\big) \stackrel{d}{=} \big(H_{-\tau+2}, H_{-\tau+3}, \dots, H_{-1}, H_0\big)$$

Beweisgedanke. Per Inspektion, betrachte Produkte der Übergangsw'keiten aus (2.38) und vergleiche mit der Argumentation aus Prop. 2.49 und ihrem Beweis. □

Nach Definition ist

$$p_\theta(\mathbf{t}, \mathbf{n}) = \sum_{\mathcal{H}\,:\, H_0 = (\mathbf{t}, \mathbf{n})} \mathbb{P}_{n,\theta}\{\mathcal{H}\}, \tag{2.39}$$

wobei wir über alle Historien mit Endzustand $H_0 = (\mathbf{t}, \mathbf{n})$ summieren (und die Rekursion für $p(\mathbf{t}, \mathbf{n})$ aus Prop. 2.49 kann als eine Art der „Buchführung" über alle möglichen Historien mit $H_0 = (\mathbf{t}, \mathbf{n})$ interpretiert werden).

Beobachtung 2.52 („Naiver Simulationsansatz")**.** Ein offensichtliche Art, $p_\theta(\mathbf{t}, \mathbf{n})$ approximativ zu bestimmen besteht darin, R unabhängige n-Stichproben $D^{(1)}, \dots, D^{(R)}$ (mit Mutationsparameter θ) zu simulieren (z.B. via Ethier-Griffiths-Urne), dann ist

$$\frac{1}{R} \sum_{r=1}^{R} \mathbf{1}\big(D^{(r)} \sim (\mathbf{t}, \mathbf{n})\big) \approx p_\theta(\mathbf{t}, \mathbf{n})$$

für genügend großes R gemäß dem Gesetz der großen Zahlen (die linke Seite konvergiert für $R \to \infty$ f.s. gegen $p_\theta(\mathbf{t}, \mathbf{n})$).

Ein Problem ist, dass man typischerweise $R \gg 1/p_\theta(\mathbf{t}, \mathbf{n})$ benötigt, um eine plausible Schätzgenauigkeit zu erreichen (was leicht $> 10^{20}$ sein kann, siehe das Beispiel aus Abschnitt 2.4.2, so dass dieses Vorgehen nur für sehr kleine Stichproben praktikabel ist).

Importance sampling

Lemma 2.53. *Für gegebenes* $(\mathbf{t}, \mathbf{n})$ *und jede Wahrscheinlichkeitsverteilung* $\mathcal{Q}$ *auf Historien, die* $\mathbb{P}_{n,\theta}\big|_{\{H_0=(\mathbf{t},\mathbf{n})\}} \ll \mathcal{Q}$ *erfüllt, ist*

$$\begin{aligned} p_\theta(\mathbf{t}, \mathbf{n}) = \mathbb{P}_{n,\theta}\big(H_0 = (\mathbf{t}, \mathbf{n})\big) &= \sum_{\mathcal{H}\,:\,H_0=(\mathbf{t},\mathbf{n})} \mathbb{P}_{n,\theta}\big(\mathcal{H}\big) \\ &= \sum_{\mathcal{H}\,:\,H_0=(\mathbf{t},\mathbf{n})} w_{\mathcal{Q}}(\mathcal{H})\mathcal{Q}(\mathcal{H}), \end{aligned}$$

wobei

$$w_{\mathcal{Q}}(\mathcal{H}) := \frac{\mathbb{P}_{n,\theta}\big(\mathcal{H}\big)}{\mathcal{Q}\big(\mathcal{H}\big)}$$

das sogenannte importance weight *von* $\mathcal{H}$ *ist.*

Demnach gilt für jedes solche $\mathcal{Q}$

$$p_\theta(\mathbf{t}, \mathbf{n}) \approx \widehat{p}_{\theta,\mathcal{Q},R}(\mathbf{t}, \mathbf{n}) := \frac{1}{R}\sum_{i=1}^{R} w_{\mathcal{Q}}(\mathcal{H}^{(i)})\mathbf{1}_{\{(\mathcal{H}^{(i)})_0=(\mathbf{t},\mathbf{n})\}},$$

mit $\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(R)}$ *u.a.,* $\sim \mathcal{Q}$*, genauer gilt*

$$\begin{aligned} \mathbb{E}_{\mathcal{Q}}[\widehat{p}_{\theta,\mathcal{Q},R}(\mathbf{t}, \mathbf{n})] &= p_\theta(\mathbf{t}, \mathbf{n}), \\ \mathrm{Var}_{\mathcal{Q}}\,[\widehat{p}_{\theta,\mathcal{Q},R}(\mathbf{t}, \mathbf{n})] &= \frac{1}{R}\mathrm{Var}_{\mathcal{Q}}\left[w_{\mathcal{Q}}(\mathcal{H})\mathbf{1}_{\{(\mathcal{H})_0=(\mathbf{t},\mathbf{n})\}}\right] \end{aligned}$$

und $\widehat{p}_{\theta,\mathcal{Q},R}(\mathbf{t}, \mathbf{n}) \underset{R\to\infty}{\longrightarrow} p_\theta(\mathbf{t}, \mathbf{n})$ *f.s. für* $R \to \infty$*.*

Beweis. Es gilt

$$\mathbb{E}_{\mathcal{Q}}[\widehat{p}_{\theta,\mathcal{Q},R}(\mathbf{t}, \mathbf{n})] = \mathbb{E}_{\mathcal{Q}}\left[w_{\mathcal{Q}}(\mathcal{H}^{(1)})\mathbf{1}_{\{(\mathcal{H}^{(1)})_0=(\mathbf{t},\mathbf{n})\}}\right] = \sum_{\mathcal{H}\,:\,H_0=(\mathbf{t},\mathbf{n})} \mathcal{Q}\big(\mathcal{H}\big)\frac{\mathbb{P}_{n,\theta}\big(\mathcal{H}\big)}{\mathcal{Q}\big(\mathcal{H}\big)} = p_\theta(\mathbf{t}, \mathbf{n}).$$

Die Formel für die Varianz ergibt sich aus der Unabhängigkeit der $\mathcal{H}^{(i)}$, die Konvergenzaussage aus dem Gesetz der großen Zahlen. □

Für Praktikabilität wünschenswert (für $\mathcal{Q}$ aus Lemma 2.53):

- $\mathcal{Q}$ sollte konzentriert sein auf $\{\mathcal{H} : H_0 = (\mathbf{t}, \mathbf{n})\}$.
- $H_0, H_{-1}, H_{-2}, \ldots$ sollte unter $\mathcal{Q}$ eine Markovkette sein.

Beispiel 2.54 (Der Vorschlag von Griffiths & Tavaré[12]). $\mathcal{Q}_{\mathrm{GT}}$ (Verteilung der) Markovkette $(H_0, H_{-1}, H_{-2}, \dots)$ auf $\mathcal{T}^*$ startend von $H_0 = (\mathbf{t}, \mathbf{n})$ mit

$$\mathcal{Q}_{\mathrm{GT}}\big(H_{i-1} = (\mathbf{t}'', \mathbf{n}'') \,\big|\, H_i = (\mathbf{t}', \mathbf{n}')\big) \propto \mathbb{P}_{n,\theta}\big(H_i = (\mathbf{t}', \mathbf{n}') \big| H_{i-1} = (\mathbf{t}'', \mathbf{n}'')\big)$$

(vgl. Rekursion in Prop. 2.49 bzw. die Übergangsw'keiten der Ethier-Griffiths-Urne; dies kann als allgemeiner Ansatz zur Lösung linearer Rekursionsgleichungen aufgefasst werden).

Beobachtung 2.55 (Optimale Vorschlagsverteilung). $\mathcal{Q}_{\mathrm{opt}}(\cdot) := \mathbb{P}_{n,\theta}\big(\,\cdot\,\big|H_0 = (\mathbf{t}, \mathbf{n})\big)$ ist optimal:

- Varianz des Schätzers $\widehat{p}_{\theta,\mathcal{Q}_{\mathrm{opt}},R}(\mathbf{t}, \mathbf{n}) = 0$, denn $w_{\mathrm{opt}}(\mathcal{H}^{(i)}) \equiv p_\theta(\mathbf{t}, \mathbf{n})$.
- Leider i.A. nur hypothetische Lösung: $\mathcal{Q}_{\mathrm{opt}}$ zu konstruieren ist genauso schwer wie die Berechnung von $p_\theta(\mathbf{t}, \mathbf{n})$.
- $H_0, H_{-1}, \dots$ ist Markovkette unter $\mathcal{Q}_{\mathrm{opt}}$.

  (Dies ist eine allgemeine Eigenschaft der Zeitumkehr von Markovketten, siehe Lemma 2.56 unten.)

Lemma 2.56 (Zeitumkehr gestoppter Markovketten[13]). *Sei $|E| < \infty$, $x_0 \in E$, (p_{xy}) substochastische Matrix auf E, $p_{x\partial} := 1 - \sum_{y\in E} p_{xy}$, ergänze p zu stochastischer Matrix auf $E \dot\cup \partial$, (X_n) p-Markovkette auf $E \dot\cup \partial$, es gelte $\inf_{x\in E} \mathbb{P}_x(\tau < \infty) = 1$ mit $\tau := \min\{n \geq 0 : X_n = \partial\}$, sei*

$$g(x, y) := \mathbb{E}_x\Big[\sum_{i=0}^{\tau} \mathbf{1}(X_i = y)\Big], \quad x, y \in E$$

die Greensche Funktion,

$$Y = \big(X_\tau, X_{\tau-1}, \dots, X_1, X_0\big)$$

(mit $\widetilde{\partial}$ "unendlich fortgesetzt", sagen wir; $Y_k = X_{\tau-k}$ für $k \leq \tau$, $Y_k = \widetilde{\partial}$ für $k > \tau$).

Unter $\mathbb{P}_{x_0}$ ist Y $\widetilde{p}$-Markovkette auf $E \cup \{\partial, \widetilde{\partial}\}$ mit

$$\begin{aligned}
\widetilde{p}_{xy} &= \frac{g(x_0, y)}{g(x_0, x)} p_{yx}, \quad x, y \in E, \\
\widetilde{p}_{\partial x} &= g(x_0, x) p_{x\partial}, \quad x \in E, \\
\widetilde{p}_{x_0\widetilde{\partial}} &= 1 - \sum_{y\in E} \widetilde{p}_{x_0 y} = \mathbb{P}_{x_0}(X_i \neq x_0 \text{ für alle } i \geq 1) \quad \text{und} \quad \widetilde{p}_{\widetilde{\partial}\widetilde{\partial}} = 1
\end{aligned}$$

und Startwert $Y_0 = \partial$.

Beweis. Aus den Voraussetzungen folgt, dass $g(x, y) < \infty$ für alle $x, y \in E$ gilt, insbesondere sind die $\widetilde{p}_{xy}$ definiert.

[12] R.C. Griffiths, S. Tavaré, Ancestral inference in population genetics. Statist. Sci. 9:307–319, (1994).

[13] Dies ist ein diskreter Spezialfall der sogenannten Nagasawa-Formel, vgl. z.B. [RW], Ch. III.42 und III.46.

Für die Identität in der Definition von $\widetilde{p}_{x_0\widetilde{\partial}}$ beachte

$$\sum_{y\in E} g(x_0,y)p_{yx_0} = \sum_{y\in E}\mathbb{E}_{x_0}\Big[\sum_{i=0}^{\tau}\mathbf{1}(X_i=y)p_{yx_0}\Big] = \sum_{i=0}^{\infty}\sum_{y\in E}\mathbb{P}_{x_0}(X_i=y, i<\tau)p_{yx_0}$$
$$= \sum_{i=0}^{\infty}\mathbb{P}_{x_0}(X_{i+1}=x_0, i\le\tau) = \mathbb{E}_{x_0}\Big[\sum_{i=1}^{\tau}\mathbf{1}(X_i=x_0)\Big] = g(x_0,x_0)-1,$$

mit der starken Markov-Eigenschaft ist $\sum_{i=1}^{\tau}\mathbf{1}(X_i=x_0) = \#\{i\ge 0 : X_i = x_0\}$ unter $\mathbb{P}_{x_0}$ geometrisch verteilt mit Erfolgsparameter $a = \mathbb{P}_{x_0}(X_i\neq x_0 \text{ für alle } i\ge 1)$, also $g(x_0,x_0) = 1/a$ und $\sum_{y\in E}\widetilde{p}_{x_0y} = \sum_{y\in E} g(x_0,y)p_{yx_0}/g(x_0,x_0) = \frac{a^{-1}-1}{a^{-1}} = 1-a$.

Seien $x_0, x_1, \dots, x_{\ell-1}\in E$, $x_\ell = \partial$, es ist

$$P_{x_0}(Y_0=x_\ell, Y_1=x_{\ell-1},\dots,Y_\ell=0, Y_{\ell+1}=\widetilde{\partial})$$
$$= \underbrace{g(x_0,x_{\ell-1})p_{x_{\ell-1}\partial}}_{=\widetilde{p}_{\partial x_{\ell-1}}} \times \prod_{i=1}^{\ell-1}\underbrace{\widetilde{p}_{x_{\ell-i}x_{\ell-i-1}}}_{=\frac{g(x_0,x_{\ell-i-1})}{g(x_0,x_{\ell-i})}p_{x_{\ell-i-1}x_{\ell-i}}} \times \widetilde{p}_{x_0\widetilde{\partial}} = p_{x_0x_1}p_{x_1x_2}\cdots p_{x_{\ell-1}x_\ell}\underbrace{\mathbb{P}_{x_0}(X_j\neq x_0\ \forall\, j\ge 1)g(x_0,x_0)}_{=1}$$
$$= \prod_{i=0}^{\ell-1}p_{x_ix_{i+1}} = \mathbb{P}_{x_0}(X_i=x_i, i=0,1,\dots,\ell).$$

□

Beispiel 2.57 (Der Vorschlag von Stephens & Donnelly[14])**.** $\mathcal{Q}_{\mathrm{SD}}$ (Verteilung der) Markovkette startend von $H_0 = (\mathbf{t},\mathbf{n})$ mit folgenden Übergängen:

- Wähle einen „erlaubten"Typ mit W'keit proportional zu seiner Häufigkeit in der aktuellen Stichprobe(nkonfiguration),

  dabei sind Typen erlaubt, die Vielfachheit ≥ 2 haben oder deren „äußerste" Mutation einzig ist. ,

- wenn gezogener Typ und dessen äußerste Mutation nur einmal vorhanden: entferne diese Mutation

- wenn mindestens zwei Linien diesen Typ haben: verschmelze ein solches Paar

(Dies wäre tatsächlich optimal im Fall des IMA-Modells, vgl. Ewens'sche Stichprobenformel, Satz 2.17.)

Der Vorschlag von Hobolt, Uyenoyama & Wiuf

Lemma 2.58. *Für Stichproben, die nur eine Mutation enthalten,*

$$(\mathbf{t},\mathbf{n}) = \big(((1,0),(0)),(m,n-m)\big)$$

[14] M. Stephens, P. Donnelly, Inference in Molecular Population Genetics, Journal of the Royal Statistical Society, Series B, 62, 605–655, (2000)

kann man die Dynamik unter $\mathcal{Q}_{\text{opt}} = \mathbb{P}_{n,\theta}\big(\,\cdot\,\big|\,H_0 = (\mathbf{t},\mathbf{n})\big)$ *bestimmen:*

$$\begin{aligned} p_\theta^{(1)}(n,m) &:= \mathbb{P}_{n,\theta}\Big(H_{-1} = \big(((1,0),(0)),(m-1,n-m)\big) \\ &\qquad\qquad \Big|\, H_0 = \big(((1,0),(0)),(m,n-m)\big)\Big) \\ &= \frac{\sum_{k=2}^{n-m+1} \frac{m-1}{n-k}\frac{1}{k-1+\theta}\binom{n-m-1}{k-2}\Big/\binom{n-1}{k-1}}{\sum_{k'=2}^{n-m+1} \frac{1}{k'-1+\theta}\binom{n-m-1}{k'-2}\Big/\binom{n-1}{k'-1}} \end{aligned} \tag{2.40}$$

(für $m \geq 2$*), für* $m = 1$ *ist*

$$\begin{aligned} p_\theta^{(1)}(n,1) &= \mathbb{P}_{n,\theta}\Big(H_{-1} = \big(((0)),(n)\big)\,\Big|\, H_0 = \big(((1,0),(0)),(1,n-1)\big)\Big) \\ &= \frac{\frac{1}{n-1+\theta}}{\sum_{k'=2}^{n} \frac{1}{k'-1+\theta}\frac{k'-1}{n-1}} \end{aligned} \tag{2.41}$$

Beweis. Sei

$$\mathcal{M}_m^n := \Big\{ \begin{array}{l} \text{es gibt genau eine Mutation im } n\text{-Koaleszenten,} \\ \text{diese ist in } m \text{ Blättern sichtbar} \end{array} \Big\}$$

(es ist $\mathbb{P}_{n,\theta}\big(H_0 = \big(((1,0),(0)),(m,n-m)\big)\big) = \mathbb{P}_\theta(\mathcal{M}_m^n)$ n. Def.)

$$\begin{aligned} \mathcal{I}_n^k &:= \Big\{ \begin{array}{l} \text{es gibt genau eine Mutation im } n\text{-Koaleszenten,} \\ \text{diese tritt auf, während es } k \text{ Kanten im Baum gibt} \end{array} \Big\}, \\ \mathcal{J}_n^k &:= \Big\{ \begin{array}{l} \text{es gibt im } n\text{-Koaleszenten auf dem Niveau } k \text{ genau eine} \\ \text{Mutation und keine Mutation auf den Niveaus } \ell = 2,3,\ldots,k-1 \end{array} \Big\} \end{aligned}$$

(Wir stellen uns vor, dass der n-Koaleszent samt Mutationen mittels der Ethier-Griffiths-Urne erzeugt wird).

Nach Kor. 1.12 (zur Skelettkette des Kingman-Koaleszenten) ist

$$\mathbb{P}_{n,\theta}\big(\mathcal{M}_m^n \,\big|\, \mathcal{I}_n^k\big) = \frac{\binom{n-m-1}{k-2}}{\binom{n-1}{k-1}}$$

Es ist

$$\mathbb{P}_{n,\theta}\big(\mathcal{I}_n^k \,\big|\, \mathcal{J}_n^k\big) = \frac{k}{k+\theta}\cdot\frac{k+1}{k+1+\theta}\cdot\ldots\cdot\frac{n-1}{n-1+\theta}$$

(vgl. Ethier-Griffiths-Urne oder beachte: die Anz. Mutationen M_ℓ auf Niveau ℓ ist geometrisch verteilt mit Erfolgsparameter $\frac{\ell-1}{\ell-1+\theta}$, wie wir in Kap. 1.1.1 gesehen hatten), somit

$$\mathbb{P}_{n,\theta}\big(\mathcal{M}_m^n \,\big|\, \mathcal{J}_n^k\big) = \mathbb{P}_{n,\theta}\big(\mathcal{I}_n^k \,\big|\, \mathcal{J}_n^k\big)\mathbb{P}_{n,\theta}\big(\mathcal{M}_m^n \,\big|\, \mathcal{I}_n^k\big) = \frac{\binom{n-m-1}{k-2}}{\binom{n-1}{k-1}}\prod_{\ell=k}^{n-1}\frac{\ell}{\ell+\theta}.$$

Weiter ist für $2 \le k \le n-m+1$ [wir stellen uns einen markierten $n-1$-Koaleszent eingebettet vor in den markierten n-Koaleszent, der durch die Ethier-Griffiths-Urne simuliert wird: der letzte Zustand mit $n-1$ Blättern hat die Verteilung des markierten $n-1$-Koaleszenten, und auf diesen bezieht sich das Ereignis $\mathcal{M}_{m-1}^{n-1}$ unten]

$$\begin{aligned}\mathbb{P}_{n,\theta}\big(\mathcal{M}_{m-1}^{n-1} \,\big|\, \mathcal{M}_m^n \cap \mathcal{J}_n^k\big) &= \mathbb{P}_{n,\theta}\big(\mathcal{M}_m^n \,\big|\, \mathcal{M}_{m-1}^{n-1} \cap \mathcal{J}_n^k\big) \frac{\mathbb{P}_{n,\theta}\big(\mathcal{M}_{m-1}^{n-1} \,\big|\, \mathcal{J}_n^k\big)}{\mathbb{P}_{n,\theta}\big(\mathcal{M}_m^n \,\big|\, \mathcal{J}_n^k\big)} \\ &= \frac{n-1}{n-1+\theta}\frac{d-1}{n-1}\frac{\frac{\binom{(n-1)-(m-1)-1}{k-2}}{\binom{(n-1)-1}{k-1}}\prod_{\ell=k}^{n-2}\frac{\ell}{\ell+\theta}}{\frac{\binom{n-m-1}{k-2}}{\binom{n-1}{k-1}}\prod_{\ell=k}^{n-1}\frac{\ell}{\ell+\theta}} = \frac{m-1}{n-k}\end{aligned}$$

(beachte

$$\mathbb{P}_{n,\theta}\big(\mathcal{M}_m^n \,\big|\, \mathcal{M}_{m-1}^{n-1} \cap \mathcal{J}_n^k\big) = \frac{n-1}{n-1+\theta}\frac{m-1}{n-1},$$

denn das geforderte Ereignis verlangt, dass das jüngste Ereignis (währenddessen es $n-1$ Linien gab) ein Verzweigungs-/Verschmelzungsereignis war und dass eine der (dann) $m-1$ "mutierten" Linien verdoppelt wurde).

Schließlich gilt [beachte $\mathcal{J}_n^k \cap \mathcal{M}_m^n = \mathcal{I}_n^k \cap \mathcal{M}_m^n$]

$$\mathbb{P}_{n,\theta}\big(\mathcal{J}_n^k \,\big|\, \mathcal{M}_m^n\big) = \mathbb{P}_{n,\theta}\big(\mathcal{I}_n^k \,\big|\, \mathcal{M}_m^n\big) = \frac{\frac{1}{k-1+\theta}\binom{n-m-1}{k-2}\Big/\binom{n-1}{k-1}}{\sum_{\ell=2}^{n-m-1}\frac{1}{\ell-1+\theta}\binom{n-m-1}{\ell-2}\Big/\binom{n-1}{\ell-1}},$$

denn für festes n, m, θ gilt

$$\mathbb{P}_{n,\theta}\big(\mathcal{I}_n^k \,\big|\, \mathcal{M}_m^n\big) \propto \mathbb{P}_{n,\theta}\big(\mathcal{I}_n^k \cap \mathcal{M}_m^n\big) = \mathbb{P}_{n,\theta}\big(\mathcal{I}_n^k\big)\mathbb{P}_{n,\theta}\big(\mathcal{M}_m^n \,\big|\, \mathcal{I}_n^k\big) \propto \frac{1}{k-1+\theta}\binom{n-m-1}{k-2}\Big/\binom{n-1}{k-1}$$

(beachte $\mathbb{P}_{n,\theta}\big(\mathcal{I}_n^k\big) = \frac{\theta}{k-1+\theta}\frac{k-1}{k-1+\theta} \times \prod_{\ell=2,\,\ell\neq k}^{n}\frac{\ell-1}{\ell-1+\theta} = \frac{1}{k-1+\theta} \times \theta\prod_{\ell=2}^{n}\frac{\ell-1}{\ell-1+\theta}$), andererseits ist $\sum_{k=2}^{n-m+1}\mathbb{P}_{n,\theta}\big(\mathcal{I}_n^k \,\big|\, \mathcal{M}_m^n\big) = 1$.

Insgesamt ist

$$\begin{aligned}\mathbb{P}_{n,\theta}\big(\mathcal{M}_{m-1}^{n-1} \,\big|\, \mathcal{M}_m^n\big) &= \sum_{k=2}^{n-m+1}\mathbb{P}_{n,\theta}\big(\mathcal{M}_{m-1}^{n-1} \cap \mathcal{J}_n^k \,\big|\, \mathcal{M}_m^n\big) \\ &= \sum_{k=2}^{n-m+1}\mathbb{P}_{n,\theta}\big(\mathcal{M}_{m-1}^{n-1} \,\big|\, \mathcal{M}_m^n \cap \mathcal{J}_n^k\big)\mathbb{P}_{n,\theta}\big(\mathcal{J}_n^k \,\big|\, \mathcal{M}_m^n\big),\end{aligned}$$

zusammensetzen ergibt (2.40).

Für den Fall $m = 1$: $\mathbb{P}_{n,\theta}\big(\mathcal{M}_0^n \,\big|\, \mathcal{M}_1^n\big)$ ist die Wahrscheinlichkeit, dass das jüngste Ereignis im n-Koaleszenten eine Mutation war, gegeben dass nur genau eine mutierte Linie vorkommt. Nach Def. ist

$$\mathbb{P}_{n,\theta}\big(\mathcal{M}_0^{n-1} \,\big|\, \mathcal{M}_1^n \cap \mathcal{J}_n^k\big) = \mathbf{1}(k = n)$$

und

$$\mathbb{P}_{n,\theta}\big(\mathcal{M}_1^n \,\big|\, \mathcal{J}_n^n\big) = \mathbb{P}_{n,\theta}\big(\mathcal{M}_0^{n-1} \,\big|\, \mathcal{J}_n^n\big) = 1,$$

zusammensetzen wie oben ergibt dann (2.41). □

Beispiel 2.59 (Der Vorschlag von Hobolt, Uyenoyama & Wiuf[15]). Sei $(\mathbf{t}, \mathbf{n}) \in \mathcal{T}^*$,

$$u_{i,m} = \begin{cases} p_\theta^{(1)}(n, d_m) \cdot \frac{n_i}{d_m} & \text{falls Typ } i \text{ Mutation } m \text{ hat,} \\ \big(1 - p_\theta^{(1)}(n, d_m)\big) \cdot \frac{n_i}{n - d_m} & \text{sonst,} \end{cases}$$

wobei d_m = Anz. Stichproben in der aktuellen Konfiguration $(\mathbf{t}, \mathbf{n})$, die Mutation m tragen.

$\mathcal{Q}_{\mathrm{HUW}}$ schlägt Typ i vor mit W'keit

$$q_{\mathrm{HUW}}\big(i \,|\, (\mathbf{t}, \mathbf{n})\big) \propto \begin{cases} \sum_m u_{i,m} & \text{wenn Typ } i \text{ erlaubt} \\ 0 & \text{sonst.} \end{cases}$$

(Ein Typ ist nicht erlaubt, wenn er Vielfachheit 1 hat, aber seine äusserste Mutation in einem weiteren Typ vorkommt.)

Übergang, wenn Typ i vorgeschlagen und dieser

- Vielfachheit 1 hat : entferne äusserste Mutation
- Vielfachheit ≥ 2 hat : Verschmelzung innerhalb Typ i

Bedingte Verteilung der Genealogie, gegeben die Daten

Man kann die Verfahren aus dem vorigen Abschnitt auch verwenden, um beispielsweise die bedingte Verteilung von T_{jgV}, der Zeit bis zum jüngsten gemeinsamen Vorfahren der Stichprobe (in Koaleszenten-Zeiteinheiten), gegeben die beobachteten Sequenzdaten $D_n \sim (\mathbf{t}, \mathbf{n})$ zu bestimmen bzw. zu approximieren (hierauf bezieht sich die „anzestrale Inferenz" aus dem Titel des Abschnitts).

Beobachtung 2.60. Gegeben $\mathcal{H} = (H_0, H_1, \ldots, H_\tau)$ ist $T_{\mathrm{jgV}} \sim \sum_{i=1}^{\tau} W_i$ unter $\mathbb{P}_{n,\theta}$ mit W_i u.a., $W_i \sim \mathrm{Exp}(\ell_i \frac{\theta}{2} + \binom{\ell_i}{2})$, wenn Zustand H_i ℓ_i Linien enthält.

Sei $(\mathbf{t}, \mathbf{n})$ ein geordneter Gen-Baum, der Sequenzbeobachtungen in einer n-Stichprobe beschreibt. Für gegebenes $\theta > 0$ kann man die bedingte Verteilung von T_{jgV} unter $\mathbb{P}_{n,\theta}$, gegeben $D_n \sim (\mathbf{t}, \mathbf{n})$, folgendermaßen approximieren:

Simuliere $\mathcal{H}^{(r)}$, $r = 1, \ldots, R$ u.a. gemäß einem $\mathcal{Q}$, das den Bedingungen von Lemma 2.53 genügt und $\mathcal{Q}(H_0 = (\mathbf{t}, \mathbf{n})) = 1$ erfüllt, dazu sei jeweils $T^{(r)} = \sum_{i=1}^{\tau} W_i^{(r)}$ wie oben (u.a. für verschiedene r). Dann gilt für $t \geq 0$

$$\frac{\frac{1}{R} \sum_{r=1}^{R} \frac{\mathbb{P}_{n,\theta}(\mathcal{H}^{(r)})}{\mathcal{Q}(\mathcal{H}^{(r)})} \mathbf{1}(T^{(r)} \leq t)}{\widehat{p}_{\theta,\mathcal{Q},R}(\mathbf{t}, \mathbf{n})} \approx \mathbb{P}_{n,\theta}\Big(T_{\mathrm{jgV}} \leq t \,\Big|\, D_n \sim (\mathbf{t}, \mathbf{n})\Big)$$

(in dem Sinne, dass die linke Seite für $R \to \infty$ gegen die Wahrscheinlichkeit auf der rechten Seite konvergiert) gemäß dem Gesetz der großen Zahlen, denn

$$\begin{aligned} \mathbb{E}_{\mathcal{Q}}\Big[\frac{\mathbb{P}_{n,\theta}(\mathcal{H})}{\mathcal{Q}(\mathcal{H})} \mathbf{1}(T_{\mathrm{jgV}} \leq t)\Big] &= \mathbb{E}_{\mathcal{Q}}\Big[\frac{\mathbb{P}_{n,\theta}(\mathcal{H})}{\mathcal{Q}(\mathcal{H})} \mathbf{1}(T_{\mathrm{jgV}} \leq t, H_0 = (\mathbf{t}, \mathbf{n}))\Big] \\ &= \mathbb{E}_{n,\theta}\Big[\mathbf{1}(T_{\mathrm{jgV}} \leq t, H_0 = (\mathbf{t}, \mathbf{n}))\Big] = \mathbb{P}_{n,\theta}\Big(T_{\mathrm{jgV}} \leq t, D_n \sim (\mathbf{t}, \mathbf{n})\Big). \end{aligned}$$

[15] A. Hobolt, M.K. Uyenoyama, C. Wiuf, Importance sampling for the infinite sites model, Stat. Appl. Genet. Mol. Biol. 7 (2008)

Bemerkung. Eine Frage nach der bedingten Verteilung von T_{jgV} gegeben gewisse Beobachtungen an den Blättern hatten wir bereits in dem Beispiel in Kapitel 1.1.1 betrachtet, dort war die Antwort angesichts der – sehr übersichtlichen – Datenlage deutlich einfacher. Die Beobachtungen dort, 38 Stichproben und 0 Mutationen, entsprechen einem Gen-Baum $(\mathbf{t}, \mathbf{n}) = \big(((0)), (38)\big)$ und in einem solchen Fall bestimmt $H_0 = \big(((0)), (n)\big)$ die Historie eindeutig.

Bericht. Das auf Lemma 2.53 basierende Monte Carlo-Verfahren zur approximativen Berechnung der Wahrscheinlichkeiten von Beobachtungen aus Beispiel 2.54 ist in `genetree` von Robert Griffiths, `http://www.stats.ox.ac.uk/~griff/software.html` implementiert. `genetree` implementiert auch das auf Beobachtung 2.60 zusammen mit Beispiel 2.54 basierende Monte Carlo-Verfahren zur approximativen Berechnung der bedingten Verteilung von T_{jgV}; zudem enthält `genetree` auch das Programm `seq2tre`, das Gusfields Algorithmus 2.43 implementiert.

Monte Carlo-Verfahren, die auf Lemma 2.53 und Beobachtung 2.60 zusammen mit den Beispielen 2.57 und 2.59 (und auch 2.54) basieren, sind in dem Programm `metagenetree` von Matthias Steinrücken, `http://sourceforge.net/projects/metagenetree/` implementiert.

Aus technischen Gründen berechnen `genetree` und `metagenetree` für $(\mathbf{t}, \mathbf{n}) \in (\mathcal{T}_d/\sim)_0 \times \mathbb{N}^d$ nicht wörtlich $p_\theta(\mathbf{t}, \mathbf{n})$ wie vor Proposition 2.49 definiert, sondern $c(\mathbf{t}, \mathbf{n}) p_\theta\big([\mathbf{t}, \mathbf{n}]_\approx\big) = d! p_\theta(\mathbf{t}, \mathbf{n})$ mit kombinatorischen Faktoren wie in Bemerkung 2.47 beschrieben, siehe M. Birkner, J. Blath, M. Steinrücken, Importance sampling for Lambda-coalescents in the infinitely many sites model, Theor. Pop. Biology 79, 155–173, (2011) und speziell Remark 1.2 dort. `metagenetree` implementiert noch weitere, auf Lemma 2.53 basierende Monte Carlo-Verfahren, die in dem genannten Artikel beschrieben sind.

2.4.2 Beispiel: Ward et als Nuu-Chah-Nulth-Daten

R.H. Ward, B.L. Frazier, K. Dew-Jager, and S. Pääbo, Extensive Mitochondrial Diversity Within a Single Amerindian Tribe, Proc. Nat. Acad. Sci. USA 88, 8720–8724, (1991) berichten beobachtete genetische Variabilität im mitochondrialen Genom in einer Stichprobe von 63 weiblichen Nuu-Chah-Nulth. Die Nuu-Chah-Nulth sind eine indigene Population ("first nation"), die hauptsächlich auf Vancouver Island, Kanada leben; es wurde jeweils ein 360 Basenpaare langes Stück der sogenannten mitochondrialen Kontrollregion sequenziert (Mitochondrien sind Zellorganellen, die auch über etwas eigene Erbinformation verfügen).

Die Daten wurden erneut analysiert im Kontext des IMS-Modells in dem Artikel R.C. Griffiths, S. Tavaré, Ancestral inference in population genetics, Statist. Sci. 9, 307–319, (1994). Wir folgen hier (gewissen Aspekten) dieser Analyse, die Daten wie von Griffiths und Tavaré editiert sind in Tabelle 2.2 wiedergegeben. In den Originaldaten von Ward *et al* (63 Stichproben) sind Verletzungen der IMS-Annahme sichtbar, d.h. die Bedingung aus Bericht 2.44 ist verletzt. Griffiths und Tavaré wählen eine Teilmenge der Daten, die mit den IMS-Annahmen verträglich ist. Es handelt sich hierbei um eine Stichprobe vom Umfang $n = 55$, in der $s = 18$ Mutationen sichtbar sind, es gibt $d = 14$ verschiedene Typen (Tabelle 2.2). Wir treffen für die weitere Diskussion die Annahme, dass an jeder segregierenden Stelle die jeweils häufigere Base anzestral ist, dann entsprechen

2	:	a	g	g	a	a	t	c	c	t	c	t	t	c	t	c	t	t	c
2	:	a	g	g	a	a	t	c	c	t	t	t	t	c	t	c	t	t	c
1	:	g	a	g	g	a	c	c	c	t	c	t	t	c	c	c	t	t	t
3	:	g	g	a	g	a	c	c	c	c	c	t	t	c	c	c	t	t	c
19	:	g	g	g	a	a	t	c	c	t	c	t	t	c	t	c	t	t	c
1	:	g	g	g	a	g	t	c	c	t	c	t	t	c	t	c	t	t	c
1	:	g	g	g	g	a	c	c	c	t	c	c	c	c	c	c	t	t	t
1	:	g	g	g	g	a	c	c	c	t	c	c	c	t	c	c	t	t	t
4	:	g	g	g	g	a	c	c	c	t	c	t	t	c	c	c	c	c	t
8	:	g	g	g	g	a	c	c	c	t	c	t	t	c	c	c	c	t	t
5	:	g	g	g	g	a	c	c	c	t	c	t	t	c	c	c	t	t	c
4	:	g	g	g	g	a	c	c	c	t	c	t	t	c	c	c	t	t	t
3	:	g	g	g	g	a	c	c	t	t	c	t	t	c	c	c	t	t	c
1	:	g	g	g	g	a	c	t	c	t	c	t	t	c	c	t	t	t	c

Tabelle 2.2: Die Beobachtungen von Ward *et al*, wie von Griffiths und Tavaré editiert: Die erste Spalte gibt jeweils an, wie oft die Sequenz in dieser Zeile beobachtet wurde, jede der Spalten rechts beschreibt eine segregierende Position

die Beobachtungen aus Tabelle 2.2 dem Gen-Baum aus Abbildung 2.9 (für die Analyse in der – realistischeren – Situation mit unbekannten Ahnentypen siehe den zitierten Artikel von Griffiths und Tavaré).

Man findet für den Watterson-Schätzer $\widehat{\theta}_W \approx 3.93$ (vgl. Beob. 2.29), die mittlere Anzahl paarweiser Unterschiede ist $\widehat{\theta}_\pi \approx 3.30$, Tajimas $D \approx -0.27$ (vgl. Def. 2.37).

Ein 95%-Konfidenzintervall für Tajimas D nach dem Ansatz von Simonsen, Churchill und Aquadro (vgl. Seite 93) ist $[-1.77, 2.11]$, insoweit gibt Tajimas D keinen Anlass, angesichts der Daten an der Annahme des Kingman-Koaleszenten für die Genealogie zu zweifeln.

Eine Approximation der Likelihood-Funktion ist in Abbildung 2.10 dargestellt, man findet $\widehat{\theta}_{ML} \approx 4.8$.

Tabelle 2.3 zeigt einen Vergleich der Schätzgüte in Abhängigkeit der Anzahl unabhängiger Simulationen gemäß verschiedener Vorschlagsverteilungen $\mathcal{Q}$ für den Schätzer $\widehat{p}_{\theta,\mathcal{Q},R}(\mathbf{t},\mathbf{n})$ aus Lemma 2.53. Die Tabelle zeigt jeweils den (geschätzten) relativen Fehler, d.h. Standardfehler/geschätzer Wert. (In Formeln: Sei $x_i = w_{\mathcal{Q}}(\mathcal{H}^{(i)})$ das importance weight des i-ten Replikats, so ist $\overline{x} := \frac{1}{R}\sum_{i=1}^{R} x_i = \widehat{p}_{\theta,\mathcal{Q},R}(\mathbf{t},\mathbf{n})$ der Schätzwert, $s^2 = \frac{1}{R-1}\sum_{i=1}^{R}(x_i - \overline{x})^2$ die geschätze Varianz der importance weights und $s/\sqrt{R}$ der Standardfehler; die Tabelleneinträge zeigen $(s/\sqrt{R})/\overline{x}$.)

Es zeigt sich in diesem Fall, dass die Vorschlagsverteilung $\mathcal{Q}_{GT}$ deutlich größere Streuung des Schätzwerts erzeugt als die anderen beiden und dass $\mathcal{Q}_{HUW}$ etwas besser ist als $\mathcal{Q}_{SD}$.

Schließlich zeigt Abbildung 2.11 eine Approximation der Verteilungsfunktion der Zeit T_{jgV} bis zum jüngsten gemeinsamen Vorfahren der Stichprobe, gegeben die Daten, basierend auf dem Verfahren aus Beobachtung 2.60. Bei $\theta = \widehat{\theta}_{ML} = 4.8$ ist der Erwartungswert von T_{jgV}, gegeben die Daten ≈ 1.20 (Koaleszenten-Zeiteinheiten) $\widehat{=}14.400$ Jahre mit der Wahl $N = 600$ und 20 Jahre/Generation (vgl. Griffiths & Tavaré, Table 3, S. 316). Der

2	:	1	14	6	4	0
2	:	10	1	14	6	4	0
1	:	2	18	0		
3	:	9	3	0		
19	:	14	6	4	0	
1	:	5	14	6	4	0
1	:	12	11	18	0	
1	:	13	12	11	18	0
4	:	17	16	18	0	
8	:	16	18	0		
5	:	0				
4	:	18	0			
3	:	8	0			
1	:	15	7	0		

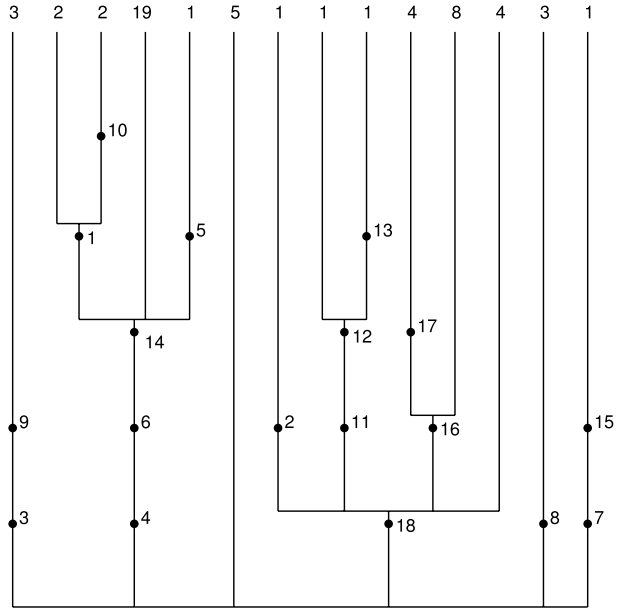


Abbildung 2.9: Die Daten aus Tabelle 2.2 als Gen-Baum (wobei wir an jeder segregierenden Stelle die häufigere Base als anzestral annehmen). Links: Jede Zeile entspricht einem Typ (kodiert wie in Beob. 2.45 als Abfolge der Mutationen), die Zahl vor dem Doppelpunkt gibt die Vielfachheit an. Rechts: Grafische Darstellung der Anordnung der Typen und Mutationen als Gen-Baum

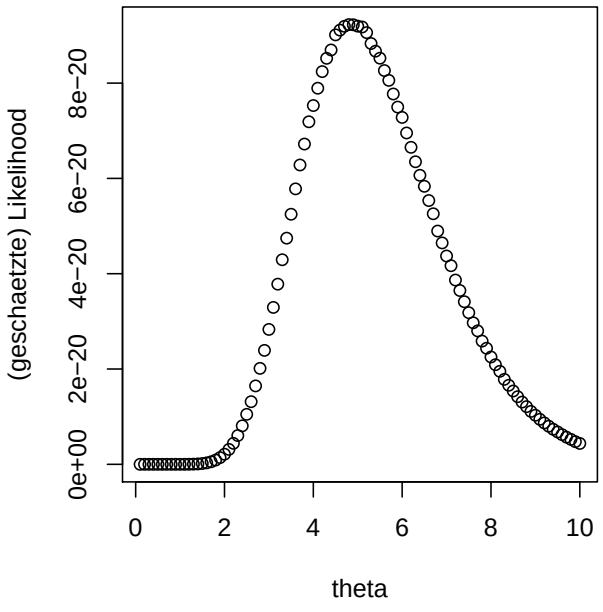


Abbildung 2.10: Likelihood-Funktion $\theta \mapsto p_\theta(\mathbf{t}, \mathbf{n})$ für die Daten aus Abbildung 2.9/Tabelle 2.2 (basierend auf dem Schätzer aus Lemma 2.53 mit $\mathcal{Q}_{\mathrm{HUW}}$ aus Beispiel 2.59 und $R = 1.2 \cdot 10^7$ Replikaten pro Punkt)

R	10^5	$5 \cdot 10^5$	10^6
rel. Fehler ($\mathcal{Q}_{GT}$)	0.53	0.11	0.08
rel. Fehler ($\mathcal{Q}_{SD}$)	0.049	0.037	0.027
rel. Fehler ($\mathcal{Q}_{HUW}$)	0.072	0.024	0.015

Tabelle 2.3: Eine Mini-Studie zur Konvergenzgeschwindigkeit der verschiedenen Importance Sampling-Verfahren: Für die Daten aus Abbildung 2.9/Tabelle 2.2 mit $\theta = \widehat{\theta}_{ML} = 4.8$ ist für verschiedene Anzahlen R der Replikate und verschiedene Vorschlagsverteilungen jeweils der (geschätzte) relative Fehler vermerkt

a apriori-Erwartungswert von T_{jgV} ist dagegen $\approx 1.96 \,\widehat{=}\, 23.500$ Jahre. Griffiths und Tavaré bemerken, dass ein Schätzwert für T_{jgV} von ca. 14.000 Jahren recht gut zu anderen diesbezüglichen Werten in der wissenschaftlichen Diskussion passt (siehe S. 316 in dem zitierten Artikel).

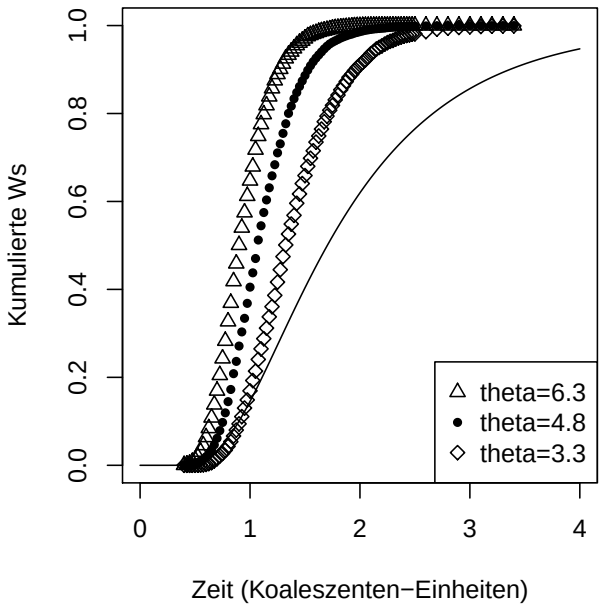


Abbildung 2.11: Verteilungsfunktion der Zeit T_{jgV} bis zum jüngsten gemeinsamen Vorfahren der Stichprobe, gegeben die Daten, für verschiedene Wahlen von θ, in Koaleszenten-Zeiteinheiten. Die durchgezogene Linie ist die Verteilungsfunktion von T_{jgV} unter dem Koaleszenten, d.h. ohne Bedingen auf die Beobachtungen.

Kapitel 3

Selektion

3.1 Vorbemerkung: Deterministische Dynamik

Wir betrachten eine sehr große Population aus diploiden, (der mathematischen Einfachheit halber) hermaphroditischen Individuen und nehmen an, dass es bezüglich eines gewissen Orts im Genom zwei verschiedene Typen, sogenannte Allele, sagen wir A und a, gibt. Jedes Individuum besitzt also zwei Chromosomenkopien, von denen es jeweils eine von jedem seiner beiden Eltern geerbt hat (gemäß den Mendelschen Regeln, d.h. die vererbte Chromosomenkopie wird rein zufällig ausgewählt); es gibt aber keine expliziten Geschlechter, jedes Individuum könnte sich prinzipiell mit jedem anderen paaren. Die zeitliche Entwicklung laufe in diskreten Generationen ab. Sei die Populationsgröße konstant N und

$$\big(Y_r^{(N)}(AA), Y_r^{(N)}(Aa), Y_r^{(N)}(aa)\big)$$

seien die Anzahlen diploider Individuen (der drei möglichen diploiden Genotypen AA, Aa, aa) in Generation r,

$$X_r^{(N)} = 2Y_r^{(N)}(AA) + Y_r^{(N)}(Aa)$$

die Anzahl A-Chromosomen (unter den insgesamt $2N$ Chromosomen) in Generation r.

Erinnerung. In Abschnitt 1.2.2 hatten wir diese Situation im Kontext des diploiden Wright-Fisher-Modells (Def. 1.30) betrachtet. Dort hatten wir vorausgesetzt, dass die genetischen Typen für den Fortpflanzungserfolg irrelevant sind und im Modell angenommen, dass die N Kinder der Nachfolgegeneration aus unabhängigen Zügen aus der Elterngeneration (wörtlich: mit Zurücklegen entstehen). Demnach ist gegeben $X_r^{(N)} = k = 2Np$ (mit einem $p \in [0,1] \cap \frac{1}{2N}\mathbb{Z}$)

$$\big(Y_{r+1}^{(N)}(AA), Y_{r+1}^{(N)}(Aa), Y_{r+1}^{(N)}(aa)\big) \sim \mathrm{Multinom}\big(N, p^2, 2p(1-p), (1-p)^2\big)$$

und $\mathscr{L}(X_{r+1}^{(N)} \mid X_r^{(N)} = k) = \mathrm{Bin}(2N, \frac{k}{2N})$.

Wenn $\frac{1}{2N} X_0^{(N)} =: p_N \to p$ für $N \to \infty$ gilt, so folgt mit dem Gesetz der großen Zahlen

$$\big(\tfrac{1}{N} Y_1^{(N)}(AA), \tfrac{1}{N} Y_1^{(N)}(Aa), \tfrac{1}{N} Y_1^{(N)}(aa)\big) \to \big(p^2, 2p(1-p), (1-p)^2\big)$$

(in Wahrscheinlichkeit) und iterativ ergibt dann sich für jedes $T \in \mathbb{N}$, $\varepsilon > 0$ (beachte $p = p^2 + \frac{1}{2}2p(1-p)$)

$$\mathbb{P}\Big(\big|\tfrac{1}{2N}X_r^{(N)} - p\big| < \varepsilon \text{ für } r = 1,2,\dots,T \,\Big|\, \tfrac{1}{2N}X_0^{(N)} = p_N\Big) \underset{N\to\infty}{\longrightarrow} 1 \quad \text{und}$$

$$\mathbb{P}\Big(\Big\|\big(\tfrac{1}{N}Y_r^{(N)}(AA), \tfrac{1}{N}Y_r^{(N)}(Aa), \tfrac{1}{N}Y_r^{(N)}(aa)\big) - \big(p^2, 2p(1-p), (1-p)^2\big)\Big\| < \varepsilon \text{ für } r = 1,2,\dots,T \,\Big|\, \tfrac{1}{2N}X_0^{(N)} = p_N\Big) \underset{N\to\infty}{\longrightarrow} 1,$$

d.h. es stellt sich Hardy-Weinberg-Gleichgewicht ein (vgl. Bem. 1.31) und die Allelanteile ändern sich (auf dieser Zeitskala) überhaupt nicht.

(Diploide) Selektion Wir betrachten nun die allgemeine Situation, in der der diploide Genotyp eines Individuums seinen Überlebens- und Fortpflanzungserfolg beeinflusst. Dazu seien $w_{AA}, w_{Aa}, w_{aa} \geq 0$ gegeben und wir nehmen an, dass die Chance eines Typ AA-Kinds, bis zum Reproduktionsalter zu überleben und somit als potentielles Elter in Frage zu kommen, proportional zu w_{AA} ist, etc. Der Vektor (w_{AA}, w_{Aa}, w_{aa}) gibt die relative Fitness der drei Genotypen an.

Wir betrachten wieder eine feste Populationgröße N und bezeichnen mit

$$\big(Y_r^{(N)}(AA), Y_r^{(N)}(Aa), Y_r^{(N)}(aa)\big)$$

die Anzahlen diploider Individuen sowie mit $X_r^{(N)} = 2Y_r^{(N)}(AA) + Y_r^{(N)}(Aa)$ die Anzahl A-Chromosomen in Generation r. Für das *diploide Wright-Fisher-Modell mit Selektion* ist gegeben $X_r^{(N)} = k = 2Np$ (mit einem $p \in [0,1] \cap \frac{1}{2N}\mathbb{Z}$)

$$\big(Y_{r+1}^{(N)}(AA), Y_{r+1}^{(N)}(Aa), Y_{r+1}^{(N)}(aa)\big) \sim \mathrm{Multinom}\Big(N, \frac{p^2 w_{AA}}{w_{\mathrm{ges}}(p)}, \frac{2p(1-p)w_{Aa}}{w_{\mathrm{ges}}(p)}, \frac{(1-p)^2 w_{aa}}{w_{\mathrm{ges}}(p)}\Big) \tag{3.1}$$

mit

$$w_{\mathrm{ges}}(p) := p^2 w_{AA} + 2p(1-p)w_{Aa} + (1-p)^2 w_{aa}$$

der „Gesamtfitness“ der Population (bei A-Anteil p). Offenbar bildet

$$\big(Y_{r+1}^{(N)}(AA), Y_{r+1}^{(N)}(Aa), Y_{r+1}^{(N)}(aa)\big)_{r\in\mathbb{N}_0}$$

mit Setzung (3.1) eine Markovkette.

Wir können (3.1) folgendermaßen interpretieren (siehe auch Abbildung 3.2): Zunächst wird ein großes Reservoir von $\gg N$ „Juvenilen“ gemäß Zufallspaarung wie im neutralen diploiden Wright-Fisher-Modell gebildet, d.h. wenn der Anteil A-Allele in der Elterngeneration p ist, verhalten sich die Anteile der Genotypen AA, Aa, aa in diesem Reservoir wie $p^2 : 2p(1-p) : (1-p)^2$. Dann werden aus diesem Reservoir N Juvenile herausgezogen, wobei ein Typ AA-Juveniles mit Wahrscheinlichkeit proportional zu w_{AA} gewählt wird, etc., und diese N Herausgezogenen reifen zu den N Erwachsenenindividuen der Folgegeneration heran. (Eine denkbare Assoziation sind Pflanzen, die viele Früchte produzieren, von denen aber wegen Ressourcenbeschränkungen nur wenige tatsächlich keimen.)

Alternativ können wir (3.1) folgendermaßen generieren: o. E. dürfen wir $w_{AA}, w_{Aa}, w_{aa} \in [0,1]$ annehmen (es kommt nur auf die Verhältnisse an). Wenn im Modell ein Kind erzeugt werden soll, werden zwei Eltern rein zufällig gewählt und in diesen jeweils rein zufällig ein Chromosom, dessen Typ kopiert wird; wenn dabei ein AA-Kind entsteht, überlebt es mit Wahrscheinlichkeit w_{AA}, etc. Die Wahrscheinlichkeit, dass ein Kind überlebt, ist dann $p^2 w_{AA} + 2p(1-p)w_{Aa} + (1-p)^2 w_{aa} = w_{\text{ges}}(p)$ und der Typ eines überlebenden Kinds ist somit

$$AA \text{ mit W'keit } \frac{p^2 w_{AA}}{w_{\text{ges}}(p)},\ Aa \text{ mit W'keit } \frac{2p(1-p)w_{Aa}}{w_{\text{ges}}(p)},\ aa \text{ mit W'keit } \frac{(1-p)^2 w_{aa}}{w_{\text{ges}}(p)},$$

wenn der Anteil A-Allele in der Elterngeneration p ist.

Wir wiederholen diesen Mechanismus unabhängig so lange, bis N überlebende Kinder erzeugt wurden und registrieren deren Typen, dies ergibt (3.1).

Analog zum neutralen Fall folgt aus (3.1) mit dem Gesetz der großen Zahlen, dass, sofern $\frac{1}{2N} X_0^{(N)} =: p_N \to p$ für $N \to \infty$ gilt,

$$\left(\tfrac{1}{N} Y_1^{(N)}(AA), \tfrac{1}{N} Y_1^{(N)}(Aa), \tfrac{1}{N} Y_1^{(N)}(aa)\right) \underset{N\to\infty}{\longrightarrow} \left(\frac{p^2 w_{AA}}{w_{\text{ges}}(p)}, \frac{2p(1-p)w_{Aa}}{w_{\text{ges}}(p)}, \frac{(1-p)^2 w_{aa}}{w_{\text{ges}}(p)}\right)$$

und daher auch

$$\frac{1}{2N} X_1^{(N)} \underset{N\to\infty}{\longrightarrow} \frac{1}{2} \frac{2p^2 w_{AA} + 2p(1-p)w_{Aa}}{w_{\text{ges}}(p)} = \frac{p^2 w_{AA} + p(1-p)w_{Aa}}{w_{\text{ges}}(p)}$$

(in Wahrscheinlichkeit).

Sei

$$f(p) := \frac{p^2 w_{AA} + p(1-p)w_{Aa}}{w_{\text{ges}}(p)}, \tag{3.2}$$

für $x_0 \in [0,1]$ definiert die Funktionsiteration

$$x_{n+1} := f(x_n), \quad n \in \mathbb{N}_0 \tag{3.3}$$

eine Folge in $[0,1]$.

Durch Iteration der Argumente oben ergibt sich (analog zum neutralen Fall), sofern $\frac{1}{2N} X_0^{(N)} \to x_0$ gilt, für jedes $T \in \mathbb{N}$, $\varepsilon > 0$

$$\mathbb{P}\left(\left|\tfrac{1}{2N} X_r^{(N)} - x_r\right| < \varepsilon \text{ für } r \le T \,\middle|\, \tfrac{1}{2N} X_0^{(N)} = p_N\right) \underset{N\to\infty}{\longrightarrow} 1 \quad \text{und} \tag{3.4}$$

$$\mathbb{P}\Big(\Big\|\big(\tfrac{1}{N} Y_r^{(N)}(AA), \tfrac{1}{N} Y_r^{(N)}(Aa), \tfrac{1}{N} Y_r^{(N)}(aa)\big) - \big(\tfrac{x_r^2 w_{AA}}{w_{\text{ges}}(x_r)}, \tfrac{2x_r(1-x_r)w_{Aa}}{w_{\text{ges}}(x_r)}, \tfrac{(1-x_r)^2 w_{aa}}{w_{\text{ges}}(x_r)}\big)\Big\| < \varepsilon \text{ für } r \le T \,\Big|\, \tfrac{1}{2N} X_0^{(N)} = p_N\Big) \underset{N\to\infty}{\longrightarrow} 1, \tag{3.5}$$

d.h. der Prozess des A-Anteils konvergiert gegen die deterministische, durch (3.3) beschriebene Folge. Auf dieser (unskalierten) Zeitskala spielen somit für eine große Population Fluktuationen, die aus der Zufälligkeit des Reproduktionsprozesses stammen, nahezu keine Rolle.

Da nur Quotienten der relativen Fitnesswerte w_{AA}, w_{Aa}, w_{aa} die Dynamik bestimmen, parametrisiert man (3.3) zur leichteren Interpretation o.E. folgendermaßen um:

$$w_{AA} = 1,\ w_{Aa} = 1 - hs,\ w_{aa} = 1 - s \tag{3.6}$$

mit $s \in [0,1]$ und $h \in \mathbb{R}$. s misst den selektiven Nachteil des Homozygoten aa im Vergleich zu AA (wir nehmen o.E. an, dass AA mindestens so fit ist wie aa, sonst vertausche die Rollen von A und a), h heißt der (Koeffizient des) „Heterozygoteneffekt(s)".

Die biologische Interpretation von (3.2), (3.3) variiert je nach dem Wert von h:

$h = 1$	:	Allel A ist rezessiv
$h = 0$	:	Allel A ist dominant
$0 < h < 1$	:	unvollständige Dominanz
$h < 0$	:	Überdominanz (Aa „am fittesten")
$h > 1$	:	Unterdominanz

In der Parametrisierung (3.6) lautet (3.2)

$$f(p) = \frac{p^2 + p(1-p)(1-hs)}{p^2 + 2p(1-p)(1-hs) + (1-p)^2(1-s)}$$

und die Änderung über eine Generation ist

$$d(p) = f(p) - p = \frac{sp(1-p)\big(ph + (1-p)(1-h)\big)}{p^2 + 2p(1-p)(1-hs) + (1-p)^2(1-s)}.$$

Langzeitverhalten Das Langzeitverhalten von $(x_r)_{r\in\mathbb{N}_0}$ aus (3.3) wird (hauptsächlich) von h bestimmt (wir nehmen $s > 0$ an, sonst ist $x_r \equiv x_0$). Offensichtlich gilt $d(0) = d(1) = 0$ und, sofern $h \neq 1/2$, auch $d(p_*) = 0$ mit $p_* = \frac{1-h}{1-2h}$.

Falls $0 \leq h \leq 1$: Es gilt $d(p) > 0$ für $0 < p < 1$, somit $x_r \nearrow 1$ für $r \to \infty$ sobald $x_0 > 0$. Man spricht von „gerichteter Selektion": Das „fittere" Allel A verdrängt a.

Falls $h < 0$: Es gilt $0 < p_* < 1$ und $d(p) > 0$ für $0 < p < p_*$, $d(p) < 0$ für $p_* < p < 1$. Daher gilt $x_r \nearrow p_*$ für $r \to \infty$ falls $0 < x_0 \leq p_*$ und $x_r \searrow p_*$ falls $p_* < x_0 < 1$. Man spricht von „balancierender Selektion": Beide Allele bleiben langfristig in der Population erhalten, das genaue Verhältnis hängt von h ab (das hier die Stärke der Überdominanz misst).

Falls $h > 1$: Es gilt $0 < p_* < 1$ und $d(p) < 0$ für $0 < p < p_*$, $d(p) > 0$ für $p_* < p < 1$. Daher gilt $x_r \searrow 0$ für $r \to \infty$ falls $0 < x_0 < p_*$ und $x_r \nearrow 1$ falls $p_* < x_0 < 1$. Man spricht von „disruptiver Selektion": Langfristig setzt sich einer der beiden Typen durch (es sei denn, man beginnt mit genau $x_0 = p_*$), welcher, hängt von der Startbedingung ab. Zwei Populationen, deren Startanteil an A sehr ähnlich ist, aber einer ober- und einer unterhalb von p_*, werden sich auf lange Sicht auseinander entwickeln.

Beispiel. Das *medionigra*-Allel (a) wurde in einer Population von *Callimorpha dominula* (ein Nachtfalter, deutsch Schönbär) in der Umgebung von Oxford recht intensiv studiert (a verändert im Vergleich zum „Wildtyp" die Flügelfärbung). Abbildung 3.1 zeigt den beobachteten a-Anteil für die Jahre 1939–1972 und eine angepasste deterministische Folge (x_r), erzeugt aus (3.3), (3.6) mit $s = 0{,}1$, $h = 0{,}5$ (d.h. $w_{AA} = 1$, $w_{Aa} = 0{,}95$, $w_{aa} = 0{,}9$). Die Werte sind Kap. 3 des Buches John H. Gillespie, *Population genetics : a concise guide*,

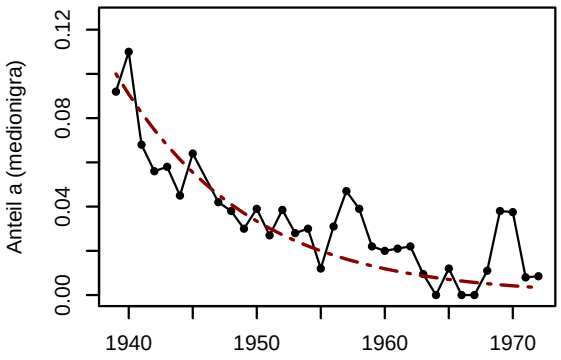


Abbildung 3.1: Beobachteter Anteil des a-Allels 1939–1972 in *Callimorpha dominula* und Modellvorhersage

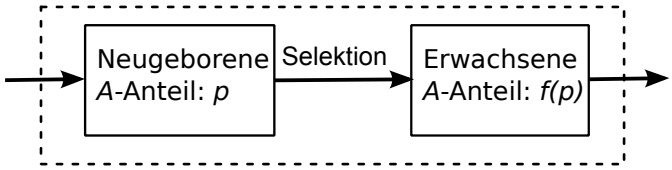


Abbildung 3.2: Schematische Darstellung eines Lebenszyklus im Modell mit Selektion

Johns Hopkins Univ. Press, 1998, entnommen. Die Modellkurve passt – zumindest dem Augenschein nach – recht gut.

(Das beweist allerdings nicht, dass tatsächlich gerichtete Selektion für die beobachteten Änderungen des a-Anteils verantwortlich ist — es könnte andere Effekte geben, die z.T. kontrovers in der Literatur diskutiert wurden, siehe z.B. E.B. Ford, P.M. Sheppard, The medionigra polymorphism of Panaxia dominula. *Heredity* 24, 112—134, 1969.)

Beispiel. Ein „klassisches Lehrbuchbeispiel“ für den Effekt balancierender Selektion ist die Verbreitung der Sichelzellenanämie in Gebieten mit endemischer Malaria. Das a-Allel (in unserer Notation) ruft in der homozygoten Form eine Verformung der roten Blutkörperchen hervor, die zum Krankheitsbild der Sichelzellenanämie führt. Aa-Individuen haben nur eine milde Form der Krankheit und sind zugleich vor gewissen Formen der Malaria geschützt. Sie haben daher in Regionen, in denen Malaria weit verbreitet ist – zumal, wenn keine Therapien zur Verfügung stehen – gegenüber AA-Homozygoten effektiv einen Vorteil. J.H. Gillespie, a.a.O., Kap. 3.3 berichtet, dass die Wahl $s = 1$, $h = -0{,}17$ und somit $p_* \approx 0{,}87$, d.h. ein Anteil a von $1 - p_* \approx 0{,}13$ relativ gut zu Beobachtungswerten in West- und Zentralafrika passt.

3.2 Intermezzo: Zeitkontinuierliches Moran-Modell

Das Moran-Modell ist gewissermaßen das „zeitkontinuierliche Analogon“ zum Wright-Fisher-Modell, es ist (ebenfalls) eines der fundamentalen Modelle der mathematischen

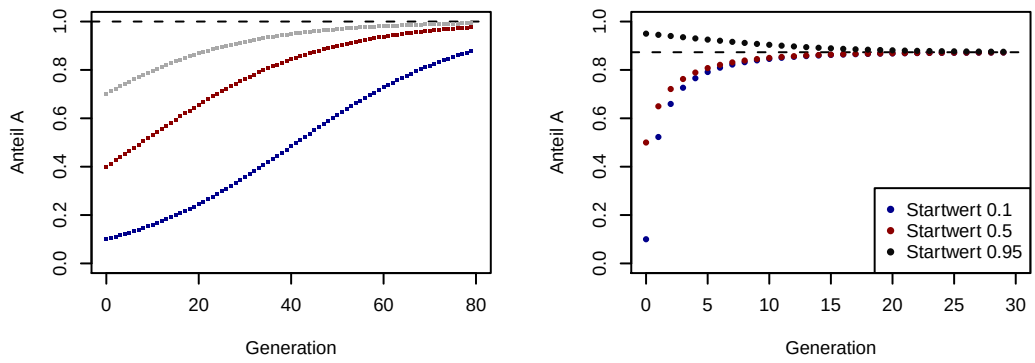


Abbildung 3.3: Zeitliche Entwicklung des A-Anteils x_r bei verschiedenen Startwerten. Links s=0,1, h=0,5 (gerichtete Selektion), rechts: s=1, h=-0,17 (balancierende Selektion)

Populationsgenetik.

Definition 3.1 ((Neutrales 2 Typ-)Moran-Modell[1])**.** Man betrachtet eine Population von konstant N (haploiden) Individuen, jedes Individuum besitzt eine unabhängige, Exp(1)-verteilte Lebenszeit und wird am Ende seiner Lebenszeit durch den Nachkommen eines rein zufällig aus der Population gezogenen Individuums ersetzt (es gibt nur ein Elter und, sagen wir, man kann durch sein eigenes Kind ersetzt werden).

Wir nehmen zusätzlich an, dass es zwei Typen A und a gibt, die ohne Mutation vererbt werden. Sei

$$X_t^{(N)} = \text{Anzahl Typ } A\text{-Ind. zur Zeit } t.$$

Angesichts der Gedächtnislosigkeit der Exponentialverteilung ist $(X_t^{(N)})_{t\geq 0}$ eine zeitkontinuierliche Markovkette mit Werten in $\{0,1,\dots,N\}$ und Sprungraten

$$q_{i,i+1} = i\frac{(N-i)}{N} = (N-i)\frac{i}{N} = q_{i,i-1}, \quad q_{i,i} = -2\frac{i(N-i)}{N}.$$

Insbesondere: 0 und N sind absorbierende Zustände, $(X_t^{(N)})_{t\geq 0}$ ist ein Martingal, für die Fixierungswahrscheinlichkeiten gilt

$$\mathbb{P}\big(\text{Absorption in } N \,\big|\, X_0^{(N)} = x\big) = \frac{x}{N}$$

(mit Argument analog zu Lemma 1.20).

Bemerkung. Eine zeitdiskrete Version hatten wir bereits in Bsp. 1.17 betrachtet, dies war ein (zeitdiskretes) Cannings-Modell, in dem in jeder „Generation“ genau ein Individuum zwei Kinder hat und dafür eines null. Insoweit entspricht eine Generation in Bsp. 1.17 einem Reproduktionsereignis im Moran-Modell, wenn man für die Individuen mit genau einem Kind jeweils Elter und Kind identifiziert.

Bericht. Wenn $\frac{1}{N}X_0^{(N)} \to x_0$ gilt, so konvergiert der reskalierte Anteilsprozess $\big(\frac{1}{N}X_{Nt/2}^{(N)}\big)_{t\geq 0}$ gegen die (neutrale 2 Typ-)Wright-Fisher-Diffusion (vgl. Def. 1.24 und Bem. 1.25).

[1]Nach Patrick Alfred Pierce Moran, 1917–1988 benannt

Man kann dies analog zum Beweis von Satz 1.26 – dort hatten wir den zeitdiskreter Fall betrachtet – beweisen oder via Zeittransformation prinzipiell auch auf diesen zurückführen; für die entscheidende Heuristik beachte

$$\frac{1}{h}\mathbb{E}\Big[\big(\tfrac{1}{N}X^{(N)}_{N(t+h)/2} - \tfrac{1}{N}X^{(N)}_{Nt/2}\big)^2 \,\Big|\, X^{(N)}_{Nt/2} = x\Big] = \frac{1}{N^2}\Big(2\frac{N}{2}h\frac{x(N-x)}{N} + o\big(\frac{N}{2}h\big)\Big) = \frac{x}{N}\big(1-\frac{x}{N}\big) + o(1)$$

für $h \downarrow 0$.

Graphische Konstruktion

Für jedes geordnete Paar (i,j), $i,j \in \{1,\dots,N\}$, $i \neq j$ sei $(N_t^{(i,j)})_{t\geq 0}$ ein Poissonprozess auf $\mathbb{R}_+$ mit Rate $\frac{1}{N}$, u.a. für verschiedene Paare. Zu den Sprungzeiten von $(N_t^{(i,j)})_{t\geq 0}$ stirbt Individuum j und wird durch einen Nachkommen von Individuum i ersetzt (s.a. Abb. 3.4).

[Bild an der Tafel]

Abbildung 3.4: Im Bild: N Kopien der Zeitachse, gerichtete Pfeile zwischen ihnen zu den Sprungzeitpunkten von u.a. Poissonprozessen; das Individuum an der Pfeilspitze stirbt jeweils und wird durch einen Nachkommen des Individuums am Pfeilschaft ersetzt.

Sei

$$X_t(i) = \text{Typ von Individuum } i \text{ zur Zeit } t.$$

Die Dynamik des Prozesses $\big(X_t(1), X_t(2), \dots, X_t(N)\big)_{t\geq 0}$, der über die Typen der Individuen in der Population Buch führt (und nicht nur über die Anzahlen) ist somit folgende:

Ersetze zu jedem Sprungzeitpunkt t von $N^{(i,j)}$ den Typ $X_{t-}(j)$ durch $X_t(j) = X_{t(-)}(i)$.

Dies ist wohldefiniert, da unabhängige Poissonprozesse f.s. keine gemeinsamen Sprungzeitpunkte besitzen. Diese Konstruktion ist ein Spezialfall einer sogenannten Harris-Konstruktion[2], ein in der Theorie der interagierenden Teilchensysteme übliches (und nützliches) Werkzeug.

Bemerkung 3.2 (Ablesen der Genealogie und der Typen aus der graphischen Konstruktion)**.** Für $t > 0$, $i \in [N]$ sei

$$A_s^{(i,t)} = \text{Nr. des Ahnenindividuums zur Zeit } t-s \text{ von Ind. } i \text{ zur Zeit } t$$
$$(\text{für } 0 \leq s \leq t, \text{ Werte in } [N])$$

Zur Konstruktion von $A^{(i,t)} = (A_s^{(i,t)})_{0\leq s\leq t}$ verfolgen wir die derzeitige „Zeitachse“ rückwärts und folgen den Pfeilen jeweils in entgegengesetzter Richtung, vgl. auch Abb. 3.4.

In Formeln können wir den Pfad von $A^{(i,t)}$ beispielsweise folgendermaßen fassen (wir schreiben $N^{(j,i)}([a,b))$ für die Anzahl Sprünge des Poissonprozesses $N^{(j,i)}$ im Zeitintervall $[a,b)$):

[2] nach Theodore Edward Harris, 1919–2005 benannt

Sei $T_0^{(i,t)} := 0$, $\widetilde{A}_0^{(i,t)} := A_0^{(i,t)} := i$, für $k \in \mathbb{N}$ setzen wir

$$T_k^{(i,t)} := \inf\left\{u > T_{k-1}^{(i,t)} : \text{es gibt ein } j \neq i \text{ mit } N^{(j,\widetilde{A}_{k-1}^{(i,t)})}\big([t-u, t-T_{k-1}^{(i,t)}]\big) = 1\right\}$$

bzw. $T_k^{(i,t)} := t$, falls es kein solches u gibt. Falls $T_k^{(i,t)} = t$ gilt, so setzen wir $M^{(i,t)} := k$ und wir brechen die Konstruktion hier ab, andernfalls sei

$$\widetilde{A}_k^{(i,t)} \text{ das (f.s.) eindeutig bestimmte } j \text{ mit } N^{(j,\widetilde{A}_{k-1}^{(i,t)})}\big([t-T_k^{(i,t)}, t-T_{k-1}^{(i,t)}]\big) = 1$$

und wir setzen die Konstruktion fort. Da die endlich vielen Poissonprozesse $N^{(j,i)}$ f.s. keine Häufungspunkte in $[0,t]$ besitzen, bricht die Konstruktion mit Wahrscheinlichkeit 1 nach endlich vielen Schritten ab und wir setzen dann für $0 < s \leq t$

$$A_s^{(i,t)} := \widetilde{A}_\ell^{(i,t)} \text{ falls } T_\ell^{(i,t)} \leq s < T_{\ell+1}^{(i,t)} \text{ für } 0 \leq \ell < M^{(i,t)}$$

bzw. $A_t^{(i,t)} := \widetilde{A}_{M^{(i,t)}-1}^{(i,t)}$.

$(A_s^{(i,t)})_{0\leq s\leq t}$ ist eine zeitkontinuierliche Markovkette mit (vollkommen symmetrischen) Sprungraten

$$q_{jk} = \begin{cases} \frac{1}{N}, & k \neq j, \\ -\frac{N-1}{N}, & k = j, \end{cases} \tag{3.7}$$

man nennt eine solche Kette auch eine (zeitkontinuierliche) „Irrfahrt auf dem vollständigen Graphen V_N der Ordnung N“.

Für $i_1 \neq i_2$ bewegen sich $A^{(i_1,t)}$ und $A^{(i_2,t)}$ unabhängig bis zum „Verschmelzungszeitpunkt“

$$\tau_{i_1,i_2} := \inf\{s \in [0,t] : A_s^{(i_1,t)} = A_s^{(i_2,t)}\},$$

ab dann, d.h. für $u \geq \tau_{i_1,i_2}$, gilt $A_u^{(i_1,t)} = A_u^{(i_2,t)}$.

Für paarweise verschiedene $i_1, i_2, \ldots i_n$ $(\leq N)$ bilden

$$A^{(i_1,t)}, \ldots, A^{(i_n,t)} \text{ ein System verschmelzender Irrfahrten auf } V_n$$

und mit

$$k \sim_{s,N} \ell \iff A_s^{(i_k,t)} = A_s^{(i_\ell,t)}, \quad 1 \leq k, \ell \leq n$$

ist

$$\mathcal{R}_s^{(n,N)} := \text{Äquivalenzklassen bezüglich } \sim_{s,N},\ s \in [0,t]$$

ein (zeittransformierter) Kingman-n-Koaleszent. $((\mathcal{R}_{Ns/2}^{(n,N)})_{s\geq 0}$ wäre wörtlich ein Koaleszent, wenn wir die Zeitachsen in der graphischen Konstruktion „bis $-\infty$ fortsetzten.“)

Aus der Konstruktion ergibt sich folgende (realisierungsweise Form) der „Dualität“:

$$X_t(i) = X_0(A_t^{(i,t)}) \quad \text{für } 1 \leq i \leq N, t > 0. \tag{3.8}$$

(Vergleiche auch Beobachtung 1.28 für den zeitdiskreten Fall.)

Beweisskizze. Die Tatsache, dass $A^{(i,t)}$ eine zeitkontinuierliche Markovkette ist, folgt anschaulich gesehen aus der Unabhängigkeit der Zuwächse der „treibenden“ Poissonprozesse $N^{(j,k)}$, die symmetrische Form der Sprungratenmatrix (3.7) stammt daher, dass alle Poissonprozesse dieselbe Rate $1/N$ haben. Wenn aktuell $A_s^{(i,t)} = j$, so gibt es für $0 < h \ll 1$ und jedes $j' \neq j$ mit Wahrscheinlichkeit $\approx h/N$ einen Sprung von $N^{(j',j)}$ im Zeitintervall $[t-s-h, t-s)$ und dann springt $A^{(i,t)}$ von j nach j'.

Etwas formaler: Sei

$$\mathcal{F}_u^t := \sigma\big(N^{(j,k)}([a,b)) : j \neq k, t-u \leq a < b \leq t\big)$$

die σ-Algebra, die die Informationen über alle Sprünge der $N^{(j,k)}$ zwischen $t-u$ und t enthält. Offenbar kann man $A_s^{(i,t)}$ für $s \leq u$ anhand der Pfade der $N^{(j,k)}$ zwischen $t-u$ und t rekonstruieren (d.h. $A_s^{(i,t)}$ ist $\mathcal{F}_u^t$-messbar für $s \leq u$) und für $s < t$, $j \neq j' \in [N]$ ist auf dem Ereignis $\{A_s^{(i,t)} = j\}$

$$\begin{aligned}
&\frac{1}{h}\mathbb{P}\big(A_{s+h}^{(i,t)} = j' \,|\, \mathcal{F}_s^t\big) \\
&= \frac{1}{h}\mathbb{P}\big(N^{(j',j)}([t-s-h, t-s)) = 1, N^{(j'',j)}([t-s-h, t-s)) = 0 \text{ für } j'' \neq j'\big) + \frac{1}{h}R_h \\
&= \frac{1}{h} \cdot e^{-h/N}\frac{h/N}{1!} \cdot \big(e^{-h/N}\big)^{N-2} + R_h = \frac{1}{N} + o(1)
\end{aligned}$$

für $h \downarrow 0$, wobei der Resterm

$$|R_h| \leq \mathbb{P}\Big(\sum_{k\neq\ell}^N N^{(k,\ell)}([t-s-h, t-s)) \geq 2\Big) = O(h^2)$$

erfüllt. □

3.3 (2-Typ) Moran-Modell mit (gerichteter) Selektion

Wir schränken uns (aus Zeit- und Platzgründen) im Folgenden bei der detaillierteren Diskussion von Modellen, die sowohl Selektion als auch Gendrift enthalten, auf das Moran-Modell (mit gerichteter Selektion) ein, da hier (im Ggs. zum Wright-Fisher-Modell) wegen der einfacheren Struktur der Übergangsdynamik verschiedene explizite Rechnungen möglich sind. Analog hatten wir im neutralen Fall gesehen: Die erwartete Fixierungszeit im Moran-Modell (Beob. 1.21) ließ sich durch eine relativ übersichtliche Rechnung bestimmen, die analoge Aussage für allgemeine Cannings-Modelle, insbesondere das Wright-Fisher-Modell (Satz 1.22) war deutlich aufwändiger.

Definition 3.3 ((2 Typ-)Moran-Modell mit gerichteter Selektion)**.** Sei $s \geq 0$. Man betrachtet eine Population von konstant N (haploiden) Individuen, die zwei mögliche Typen A und a haben. Typ A-Ind. vermehren sich mit Rate $1+s$, Typ a-Ind. vermehren sich mit Rate 1 und bei jedem Vermehrungsereignis wird ein rein zufällig gezogenes Individuum durch den gerade erzeugten Nachkommen ersetzt (der den Typ des Elters erbt und, sagen wir, man kann durch sein eigenes Kind ersetzt werden). Sei

$$X_t^{(N)} = \#\text{Typ } A\text{-Ind. zur Zeit } t,$$

$(X_t^{(N)})_{t\geq 0}$ ist eine zeitkontinuierliche Markovkette mit Werten in $\{0,1,\dots,N\}$ und Sprungratenmatrix

$$q_{ij} = \begin{cases} (1+s)i\frac{N-i}{N}, & j = i+1, \\ (N-i)\frac{i}{N}, & j = i-1, \\ -(2+s)\frac{i(N-i)}{N}, & j = i, \\ 0, & \text{sonst.} \end{cases}$$

Bemerkung. 1. Wörtlich unterscheiden sich in der Formulierung von Def. 3.3 die Typen bezüglich der Vermehrungsgeschwindigkeit (man nennt dies auch "fecundity selection").

Wir könnten dieselben Sprungraten allerdings auch als Unterschiede in der Überlebensfähigkeit ("viability selection") interpretieren: Nehmen wir an, Typ A-Individuen sterben mit Rate 1 (d.h. ihre Lebenszeit ist Exp(1)-verteilt), Typ a-Individuen sterben mit Rate $1+s$ (d.h. ihre Lebenszeit ist $\sim \mathrm{Exp}(1+s)$) und das gestorbene Ind. wird instantan durch den Nachkommen eines rein zufällig gezogenen Ind. ersetzt, so liefert dies dieselbe Dynamik von $(X_t^{(N)})_{t\geq 0}$.

Bemerkung 3.4 (Graphische Konstruktion des Moran-Modells mit Selektion)**.** Man kann die graphische Konstruktion aus Abschnitt 3.2 auf den Fall mit Selektion übertragen:

Für jedes geordnete Paar (i,j), $i,j \in \{1,\dots,N\}$, $i \neq j$ sei $(N_t^{(i,j)})_{t\geq 0}$ ein Poissonprozess auf $\mathbb{R}_+$ mit Rate $\frac{1}{N}$, u.a. für verschiedene Paare. Zusätzlich sei für jedes Paar $(S_t^{(i,j)})_{t\geq 0}$ ein weiterer, unabhängiger Poissonprozess auf $\mathbb{R}_+$ mit Rate s/N.

Wie vorher erzeugt zu den Sprungzeiten von $(N_t^{(i,j)})_{t\geq 0}$ Individuum i einen Nachkommen, der Individuum j ersetzt, wir legen in der Konstruktion zu einem solchen Zeitpunkt einen Pfeil von i nach j. Zu den Sprungzeiten von $(S_t^{(i,j)})_{t\geq 0}$ legen wir „s-Pfeile", zu einem solchen Zeitpunkt erzeugt Individuum i einen Nachkommen, der j verdrängt, aber nur, wenn i aktuell Typ A hat (s-Pfeile sind also nur „wirksam", wenn der Typ am „Pfeilschaft" A ist).

`[Bild an der Tafel]`

Offenbar sind 0 und N (wiederum) absorbierende Zustände von $X^{(N)}$. Die Fixierungswahrscheinlichkeiten sind für $s > 0$ (für $s \downarrow 0$ ergibt sich z.B. mit der Regel von l'Hospital i/N, was zur Diskussion in Abschnitt 3.2 passt) wie folgt:

Satz 3.5. *Sei* $\tau_k = \inf\{t \geq 0 : X_t^{(N)} = k\}$. *Es gilt*

$$\mathbb{P}_i(\tau_N < \tau_0) = \frac{1-(1+s)^{-i}}{1-(1+s)^{-N}}.$$

(Wir schreiben $\mathbb{P}_i(\cdot)$ für $\mathbb{P}(\cdot \,|\, X_0^{(N)} = i)$ und unterdrücken die Abhängigkeit von N in der Notation von τ_k.)

Beobachtung 3.6. Für $s > 0$ (fest), $N \gg 1$ ist demnach

$$h(1) \approx 1 - \frac{1}{1+s} = \frac{s}{1+s} \quad (\approx s \text{ für } s \text{ klein});$$

für $N \gg 1$, $s \ll 1$ mit $\sigma := Ns \in (0,\infty)$, $x \in (0,1)$ ist

$$h(\lfloor Nx \rfloor) = \frac{1-(1+\sigma/N)^{-\lfloor Nx \rfloor}}{1-(1+\sigma/N)^{-N}} \approx \frac{1-e^{-\sigma x}}{1-e^{-\sigma}}$$

Beweis von Satz 3.5. Sei $h(i) := \mathbb{P}_i(\tau_N < \tau_0)$, offenbar gilt $h(0) = 0$, $h(N) = 1$, Zerlegung gemäß dem ersten Sprung zeigt

$$\begin{aligned} h(i) &= \frac{1}{(2+s)\frac{i(N-i)}{N}} \left((1+s)\frac{i(N-i)}{N} h(i+1) + \frac{i(N-i)}{N} h(i-1) \right) \\ &= \frac{1+s}{2+s} h(i+1) + \frac{1}{2+s} h(i-1) \qquad \text{für } 1 \le i < N. \end{aligned}$$

Dies ist ein homogenes, lineares Differenzengleichungssystem 2. Ordnung, demnach gilt

$$h(i) = c_1 u_1^i + c_2 u_2^i$$

mit $u_{1/2}$ Lösungen von $(1+s)u^2 + (2+s)u + 1 = 0$, d.h.

$$u_{1/2} = \frac{(2+s) \pm \sqrt{(2+s)^2 - 4(1+s)\cdot 1}}{2(1+s)} = \frac{(2+s) \pm s}{2(1+s)} = \{1, \tfrac{1}{1+s}\}$$

und Koeffizienten $c_1, c_2 \in \mathbb{R}$, deren Werte $c_1 = 1/(1-(1+s)^{-N})$, $c_2 = -1/(1-(1+s)^{-N})$ sich aus den Randbedingungen ergeben. □

Stellen wir uns vor, in einer bisher homogenen a-Population ist gerade eine vorteilhafte Mutation A aufgetreten (weitere Mutationen sieht unser Modell zunächst nicht vor). Obwohl A selektiv bevorzugt ist, könnte der Typ A aufgrund der zufälligen Fluktuationen im Reproduktionsprozess trotzdem wieder verschwinden. Beobachtung 3.6 gibt die Wahrscheinlichkeit an, dass A tatsächlich fixiert. Der folgende Satz geht der Frage nach, wie lange es dauert, bis eine solche „frisch aufgetretene“ vorteilhafte Mutation fixiert (gegeben, dass dies tatsächlich passiert).

Satz 3.7. *Sei $s > 0$ fest, starte mit $X_0^{(N)} = 1$. Für $N \to \infty$ gilt*

$$\frac{s}{2\log N} \tau_N \to 1 \quad \textit{in Verteilung unter } \mathbb{P}(\cdot \,|\, \tau_N < \tau_0),$$

d.h. $\mathbb{P}(\tau_N \le b \log N \,|\, \tau_N < \tau_0) \xrightarrow[N\to\infty]{} \begin{cases} 0, & \textit{falls } b < 2/s, \\ 1, & \textit{falls } b > 2/s. \end{cases}$

Bew.idee für Satz 3.7: Wir zerlegen den Pfad (auf dem Weg von $X_0^{(N)} = 1$ nach $X_{\tau_N}^{(N)} = N$) in drei Phasen.

1. Solange $X^{(N)}$ klein ist (sagen wir, $\le \varepsilon N$), geschehen die Sprünge

$$\begin{aligned} x \to x+1 \text{ mit Rate } (1+s)i\frac{N-i}{N} &\approx (1+s)i, \\ x \to x-1 \text{ mit Rate } i\frac{N-i}{N} &\approx i, \end{aligned}$$

d.h. $X^{(N)}$ verhält sich „beinahe“ wie ein superkritischer binärer Galton-Watson(-Verzweigungs)-Prozess (Y_t) (vgl. Def. 3.9 unten). Da

$$Y_t \approx \text{eine zufällige Konstante} \times e^{st}$$

(siehe Proposition 3.10 unten), dauert es etwa Zeit $\approx \frac{1}{s}\log N$, bis $X^{(N)}$ auf εN angewachsen ist.

2. Sobald $X^{(N)}$ „mittelgroß“ ist (sagen wir, $\varepsilon N < X_t^{(N)} < (1-\varepsilon)N$) folgt der Pfad des Anteilsprozesses $\frac{1}{N}X_t^{(N)}$ nahezu einer deterministischen, „makroskopischen“ Dynamik (vgl. auch Diskussion in Abschnitt 3.1).

   Diese Phase ist daher „kurz“ im Vergleich zu Phasen 1 und 3 (ihre Länge divergiert nicht mit $N \to \infty$)

3. Sobald $X^{(N)}$ „groß“ ist (sagen wir, $\geq (1-\varepsilon)N$) können wir analog zu Phase 1 mit einem subkritischem Galton-Watson-Prozess vergleichen (oder wir vertauschen Rollen von a und A und kehren Zeit um, um wörtlich auf Beweisschritte für Phase 1 zurückzugreifen).

   Die Phase dauert ebenfalls $\approx \frac{1}{s}\log N$.

`[Skizze an der Tafel]`

Offensichtlich benötigen wir für den Beweis von Satz 3.7 Informationen über das Verhalten von Markovketten, die wir auf das Erreichen gewisser Zustände bedingen. Eine allgemeine Antwort gibt die Doob-Transformation.

Sei X zeitkontinuierliche Markovkette auf E (E endl. oder abz.b. unendliche Menge) mit (konservativer) Ratenmatrix $Q = (q_{ij})$ (und wir nehmen an $\sup_x -q_{xx} < \infty$, d.h. die Gesamtsprungrate ist global beschränkt), sei $E_0 \subset E$ eine endliche Menge von absorbierenden Zuständen, Q sei irreduzibel auf $E \setminus E_0$ und jeder Punkt $z \in E_0$ mit pos. W'keit erreichbar von $E \setminus E_0$ aus, für

$$\tau := \inf\{t \geq 0 : X_t \in E_0\}$$

gelte $\mathbb{P}_x(\tau < \infty) = 1$ für alle $x \in E$.

Sei $z_0 \in E_0$ und $h : E \to [0,\infty)$ (beschränkt) mit $h(z_0) = 1$, $h(z) = 0$ für $z \in E_0 \setminus \{z_0\}$, h (Q-)harmonisch in $E \setminus E_0$, d.h.

$$\sum_y q_{xy} h(y) = 0 \quad \text{für } x \in E \setminus E_0. \tag{3.9}$$

Lemma 3.8 (Doob-Transformation)**.** *Sei $X_0 \in E \setminus E_0$ (f.s). $(X)_{t\geq 0}$ bedingt auf $\{X_\tau = z_0\}$ ist verteilt wie die zeitkontinuierliche Markovkette $\widetilde{X}$ mit Sprungraten $\widetilde{q}_{ij} = \frac{h(j)}{h(i)} q_{ij}$.*

(Beachte: $\sum_{j\neq i} \frac{h(j)}{h(i)} q_{ij} = \frac{1}{h(i)}(-h(i)q_{ii}) = -q_{ii}$ nach Voraussetzung, d.h. $(\widetilde{q}_{ij})$ ist ebenfalls eine (konservative) Ratenmatrix und $\widetilde{X}$ und X haben in jedem Punkt dieselbe Gesamtsprungrate).

Beweis. Sei $p_{ij} := \frac{q_{ij}}{-q_{ii}}$ für $i \neq j \in E$, $p_{ii} := 0$ die Übergangsmatrix der (zeitdiskreten) Skelettkette X' von X (d.h. im Zustand i verbringt X eine $\mathrm{Exp}(-q_{ii})$-verteilte Wartezeit und springt dann wie X' von i aus gemäß p_{ij}, $j \in E$ in einen neuen Zustand j) und analog $\widetilde{p}_{ij} := \frac{\widetilde{q}_{ij}}{-\widetilde{q}_{ii}}$ für $i \neq j \in E$, $\widetilde{p}_{ii} := 0$ die Übergangsmatrix der (zeitdiskreten) Skelettkette $\widetilde{X}'$ von $\widetilde{X}$. Dann gilt auch

$$\widetilde{p}'_{ij} = \frac{h(j)}{h(i)} p'_{ij}.$$

Es ist $\{X_\tau = z_0\} = \{X'_{\tau'} = z_0\}$ mit $\tau' := \min\{k \in \mathbb{N}_0 : X'_k \in E_0\}$ und nach Definition folgt aus (3.9) für $x \in E \setminus E_0$

$$-q_{xx} h(x) = \sum_{y \neq x} q_{xy} h(y), \qquad \text{somit } h(x) = \sum_y p_{xy} h(y),$$

d.h. h ist auch p-harmonisch in $E \setminus E_0$. Das Standard-Argument der Zerlegung nach dem ersten Sprung zeigt, dass auch $i \mapsto \mathbb{P}_i(X'_{\tau'} = z_0)$ p-harmonisch in $E \setminus E_0$ ist, auf E_0 stimmt dies natürlich mit h überein, daher gilt

$$h(i) = \mathbb{P}_i(X'_{\tau'} = z_0) = \mathbb{P}_i(X_\tau = z_0).$$

Siehe z.B. Klenke, [Kl, Satz 19.7] für die Eindeutigkeit des Dirichlet-Problems (wörtlich dort im Fall endlichen Zustandsraums).

Für $\ell \in \mathbb{N}$, $x_0, x_1, \ldots, x_{\ell-1} \in E \setminus E_0$, $x_\ell := z_0$ ist

$$\mathbb{P}_{x_0}\big(X'_0 = x_0, X'_1 = x_1, \ldots, X'_\ell = x_\ell \,\big|\, X'_{\tau'} = z_0\big) = \frac{1}{h(x_0)} \prod_{i=1}^{\ell} p'_{x_{i-1}x_i} = \frac{1}{h(x_0)} \prod_{i=1}^{\ell} \frac{h(x_{i-1})\widetilde{p}'_{x_{i-1}x_i}}{h(x_i)}$$
$$= \frac{1}{h(x_\ell)} \prod_{i=1}^{\ell} \widetilde{p}'_{x_{i-1}x_i} = \mathbb{P}_{x_0}\big(\widetilde{X}'_0 = x_0, \widetilde{X}'_1 = x_1, \ldots, \widetilde{X}'_\ell = x_\ell\big),$$

denn $h(x_\ell) = h(z_0) = 1$. Demnach gilt die Behauptung für die Skelettketten, da X und $\widetilde{X}$ in jedem Punkt dieselbe Gesamtsprungrate haben, gilt sie auch für X und $\widetilde{X}$. □

Für $X^{(N)}$, das Moran-Modell mit gerichteter Selektion (Def. 3.3) und $E_0 = \{0, N\}$, $z_0 = N$ ergibt sich mit Satz 3.5 aus Lemma 3.8:

$$\widetilde{q}_{i,i+1} = \frac{1-(1+s)^{-(i+1)}}{1-(1+s)^{-i}} \, i \frac{N-i}{N}(1+s) = i\Big(1-\frac{i}{N}\Big)\Big(1+\frac{s}{1-(1+s)^{-i}}\Big),$$
$$\widetilde{q}_{i,i-1} = \frac{1-(1+s)^{-(i-1)}}{1-(1+s)^{-i}} \, (N-i)\frac{i}{N} = i\Big(1-\frac{i}{N}\Big)\Big(1+\frac{s(1+s)^{-i+1}}{1-(1+s)^{-i}}\Big)$$

Für unser Vergleichsargument benötigen wir (gewisse) Verzweigungsprozesse (in Def. 2.21 hatten wir bereits den Yule-Prozess betrachtet, einen Verzweigungsprozess, in dem Individuen niemals sterben).

Definition 3.9 (Binärer (zeitkontinuierlicher) Galton-Watson-Prozess)**.** Wir betrachten eine Population von Individuen, die sich jeweils unabhängig mit Rate $\lambda > 0$ verdoppeln und mit Rate $\mu > 0$ sterben (es gibt keine Restriktionen an die Größe der Population); bezeichne Y_t die Anzahl Individuen zur Zeit t.

$(Y_t)_{t\geq 0}$ ist eine zeitkontinuierliche Markovkette mit Sprungratenmatrix

$$q_{ij}^{\mathrm{GW}} = \begin{cases} \lambda i, & j = i+1, \\ \mu i, & j = i-1, \\ -(\lambda+\mu)i, & j = i, \\ 0, & \text{sonst.} \end{cases}$$

Y heißt ein (binärer zeitkontinuierlicher) Galton-Watson-Prozess.

Man sieht aus der Form der Sprungraten, dass Y die Verzweigungseigenschaft

$$\mathscr{L}(Y_t \,|\, Y_0 = k+j) = \mathscr{L}(Y_t \,|\, Y_0 = k) * \mathscr{L}(Y_t \,|\, Y_0 = j)$$

besitzt (anschaulich: wenn man die Startpopulation in zwei Gruppen aufteilt, dann kann man die beiden Gruppen sich für Zeit t unabhängig getrennt entwickeln lassen und die sich daraus ergebenden Populationen dann wieder zusammenfügen, dies ändert die Verteilung der Gesamtgröße nicht).

Proposition 3.10. *Es gilt*

$$\mathbb{E}_k[Y_t] = e^{(\lambda-\mu)t}k, \quad \mathrm{Var}_k[Y_t] = \frac{\lambda+\mu}{\lambda-\mu}\big(e^{2(\lambda-\mu)t} - e^{(\lambda-\mu)t}\big)k \tag{3.10}$$

(die Formel für die Varianz gilt im Fall $\lambda \neq \mu$, falls $\lambda = \mu$ gilt, ist $\mathrm{Var}_k[Y_t] = 2\lambda k t$). Mit $W_t := e^{-(\lambda-\mu)t}Y_t$ ist $(W_t)_{t\geq 0}$ ein nicht-negatives Martingal, demnach

$$W_t \to W_\infty \quad \text{f.s. für ein } W_\infty \geq 0.$$

Falls $\lambda > \mu$ gilt, ist $(W_t)_{t\geq 0}$ L^2-beschränkt, insbesondere gleichgradig integrierbar, und es gilt $\{W_\infty > 0\} = \{Y_t \neq 0 \ \forall\, t \geq 0\}$ f.s. (d.h. falls der Prozess überlebt, wächst er auch exponentiell mit Rate $\lambda - \mu$) sowie

$$h(k) := \mathbb{P}_k(Y_t = 0 \text{ schließlich}) = (\mu/\lambda)^k, \quad y \in \mathbb{N}.$$

Bemerkung 3.11. Ein binärer zeitkontinuierlicher Galton-Watson-Prozess $Y = (Y_t)_{t\geq 0}$ mit $\lambda > \mu$ heißt *superkritisch*, Proposition 3.10 zeigt, dass Y dann mit positiver Wahrscheinlichkeit überlebt und in diesem Fall unbeschränkt (exponentiell) wächst.

Falls $\lambda < \mu$ gilt, heißt Y *subkritisch*, in diesem Fall gilt $\mathbb{P}_k(Y_t > 0) \leq \mathbb{E}_k[Y_t] = ke^{(\lambda-\mu)t} \to 0$ für $t \to \infty$ und wegen $\{Y_t = 0\} \subset \{Y_u = 0 \text{ für alle } u \geq t\}$ also $\mathbb{P}_k(Y_t = 0 \text{ schließlich}) = 1$.

Im Fall $\lambda = \mu$ heißt Y *kritisch*, man kann mit etwas mehr Aufwand zeigen, dass auch dann gilt $\mathbb{P}_k(Y_t = 0 \text{ schließlich}) = 1$.

Beweisskizze für Proposition 3.10. Für das 1. Moment beachten wir mit $f(y) = y$, dass der Generator $Lf(y) = \lambda y \cdot (y+1) + \mu y \cdot (y-1) - (\lambda + u)y = (\lambda - \mu)y$ erfüllt, also

$$\frac{d}{dt}\mathbb{E}_k[Y_t] = (\lambda - \mu)\mathbb{E}_k[Y_t], \quad \mathbb{E}_k[Y_0] = k$$

gemäß Kolmogorovs Rückwärtsgleichung mit Lösung $\mathbb{E}_k[Y_t] = e^{(\lambda-\mu)t}k$.

Zusammen mit der Markoveigenschaft ergibt sich

$$\mathbb{E}_k\big[Y_{t+h} \,\big|\, \sigma(Y_u, u \le t)\big] = e^{(\lambda-\mu)h}Y_t,$$

d.h. $W_t = e^{-(\lambda-\mu)t}Y_t$ *ist* ein Martingal.

Für das 2. Moment bzw. die Varianz betrachten wir $f(k) = k^2$,

$$Lf(k) = k\big(\lambda(k+1)^2 + \mu(k-1)^2 - (\lambda+\mu)k^2\big) = k\big(2\lambda k + \lambda - 2\mu k + \mu\big) = 2(\lambda-\mu)k^2 + (\lambda+\mu)k$$

Damit ist

$$\frac{d}{dt}\mathbb{E}_1[Y_t^2] = 2(\lambda-\mu)\mathbb{E}_1[Y_t^2] + (\lambda+\mu)\mathbb{E}_1[Y_t] = 2(\lambda-\mu)\mathbb{E}_1[Y_t^2] + (\lambda+\mu)e^{(\lambda-\mu)t}$$

(und $\mathbb{E}_1[Y_0^2] = 1$), Variation der Konstanten liefert

$$\begin{aligned}
\mathbb{E}_1[Y_t^2] &= e^{2(\lambda-\mu)t}\mathbb{E}_1[Y_0^2] + \int_0^t e^{2(\lambda-\mu)(t-u)}(\lambda+\mu)e^{(\lambda-\mu)u}\,du \\
&= e^{2(\lambda-\mu)t} + (\lambda+\mu)e^{2(\lambda-\mu)t}\int_0^t e^{-(\lambda-\mu)u}\,du = e^{2(\lambda-\mu)t} + (\lambda+\mu)e^{2(\lambda-\mu)t}\frac{1}{\lambda-\mu}\big(1 - e^{-(\lambda-\mu)t}\big) \\
&= e^{2(\lambda-\mu)t} + \frac{\lambda+\mu}{\lambda-\mu}\big(e^{2(\lambda-\mu)t} - e^{(\lambda-\mu)t}\big) = \frac{2\lambda}{\lambda-\mu}e^{2(\lambda-\mu)t} - \frac{\lambda+\mu}{\lambda-\mu}e^{(\lambda-\mu)t}.
\end{aligned}$$

Somit

$$\mathrm{Var}_1[Y_t] = \mathbb{E}_1[Y_t^2] - \big(\mathbb{E}_1[Y_t]\big)^2 = \big(\frac{2\lambda}{\lambda-\mu} - 1\big)e^{2(\lambda-\mu)t} - \frac{\lambda+\mu}{\lambda-\mu}e^{(\lambda-\mu)t} = \frac{\lambda+\mu}{\lambda-\mu}\big(e^{2(\lambda-\mu)t} - e^{(\lambda-\mu)t}\big),$$

die Verzweigungseigenschaft liefert $\mathrm{Var}_k[Y_t] = k\mathrm{Var}_1[Y_t]$ (denn $\mathscr{L}(Y_t \,|\, Y_0 = 0) =^d Y_t^{(1)} + \cdots + Y_t^{(k)}$, wobei $Y^{(i)}$ unabhängige Kopien jeweils mit Startwert $Y^{(i)} = 1$ sind).

Die Formel für den Fall $\lambda = \mu$ kann man beispielsweise erhalten, indem man in obigem $\lambda \searrow \mu$ betrachtet (oder die entsprechende Differentialgleichung für $\frac{d}{dt}\mathbb{E}_1[Y_t^2] = 2\lambda\mathbb{E}_1[Y_t]$ direkt löst).

Im Fall $\lambda > \mu$ zeigt die Formel für das 2. Moment, dass $\sup_{t\ge 0}\mathbb{E}\big[(e^{-(\lambda+\mu)t}Y_t)^2\big] < \infty$ gilt.

Zur Aussterbewahrscheinlichkeit: $h(y) := \mathbb{P}_y(Y_t = 0 \text{ schließlich})$ löst (zerlege gemäß dem ersten Sprung, benutze die starke Markov-Eigenschaft)

$$h(y) = \frac{\lambda y}{(\lambda+\mu)y}h(y+1) + \frac{\mu y}{(\lambda+\mu)y}h(y-1) = \frac{\lambda}{\lambda+\mu}h(y+1) + \frac{\mu}{\lambda+\mu}h(y-1), \quad y \in \mathbb{N}$$

mit Randbedingung $h(0) = 1$ (und $\lim_{y\to\infty} h(y) = 0$), die (eind.) Lösung ist $h(y) = (\mu/\lambda)^y$.

Offensichtlich gilt

$$\{Y_t = 0 \text{ schließlich}\} \subset \{W = 0\} \text{ stets},$$

um die Behauptung

$$\{W_\infty > 0\} = \{Y_t \neq 0 \ \forall\, t \geq 0\} \text{ f.s.}$$

zu zeigen, müsste man etwa zeigen, dass $\mathbb{P}_k(W_\infty = 0) = \mathbb{P}_k(Y_t = 0 \text{ schließlich}) = (\mu/\lambda)^k$ gilt.

Siehe dazu beispielsweise Thm. III.2 in K.B. Athreya, P. Ney, Branching processes, Springer, 1972 oder Prop. 5.6 und Cor. 5.7 in Russell Lyons, Yuval Peres, Probability on trees and networks, Cambridge University Press, 2016+ (auch elektronisch unter `http://mypage.iu.edu/~rdlyons/prbtree/prbtree.html`). □

Mit Lemma 3.8 (Doob-Transformation) ist die Sprungratenmatrix von Y (für $\lambda = 1+s$, $\mu = 1$) bedingt auf $\{Y_t \neq 0 \ \forall\, t \geq 0\}$ $(= \{Y_t \to \infty\} = \{W_\infty > 0\})$:

$$\widetilde{q}_{ij}^{\mathrm{GW}} = \frac{1-(1+s)^{-j}}{1-(1+s)^{-i}} q_{ij}^{\mathrm{GW}}$$

(Wörtlich wäre hier noch ein kleines Approximationsargument notwendig, da „∞“ eigentlich kein Punkt des Zustandsraums von Y ist: Bedinge zunächst darauf, $N \gg 1$ vor 0 zu treffen, dann lasse $N \to \infty$.)

Beobachtung 3.12. Sei $\varepsilon > 0$, $\tilde{\tau}_{\varepsilon N} := \inf\{t \geq 0 : Y_t \geq \varepsilon N\}$, es gilt

$$\frac{s}{\log N} \tilde{\tau}_{\varepsilon N} \xrightarrow[N\to\infty]{} 1 \quad \text{f.s. auf } \{W_\infty > 0\}$$

Intuitiv/anschaulich klar: $Y_t \approx e^{st} W_\infty$, also sollte $\tilde{\tau}_{\varepsilon N} \approx \frac{1}{s}(\log N + \log \varepsilon - \log W_\infty)$ sein.

Beweis. Sei $T_\delta := \sup\{t : \left|\frac{Y_t}{e^{st}W_\infty} - 1\right| > \delta\}$ (bzw. $T_\delta = \infty$ auf $\{W_\infty = 0\}$), es gilt $\{T_\delta < \infty\}$ (f.s.) auf $\{W_\infty > 0\}$.

Wähle $\delta < \varepsilon$. Es gilt

$$\{\tilde{\tau}_{\varepsilon N} > \tfrac{1+\delta}{s} \log N\} \subset \{Y_{\frac{1+\delta}{s}\log N} < \varepsilon N\} \subset \{T_\delta \geq \tfrac{1+\delta}{s} \log N\} \cup \{W_\infty \leq N^{-\delta}\},$$

denn auf $\{T_\delta < \frac{1+\delta}{s}\log N\} \cap \{W_\infty > N^{-\delta}\}$ ist $Y_{\frac{1+\delta}{s}\log N} \geq (1-\delta) e^{(1+\delta)\log N} W_\infty \geq (1-\delta) N > \varepsilon N$.

Weiter ist

$$\{\tilde{\tau}_{\varepsilon N} < \tfrac{1-\delta}{s} \log N\} \subset \Big\{ \sup_{t \leq \log\log N} Y_t \geq \varepsilon N \Big\} \cup \{T_\delta \geq \log\log N\} \cup \{W_\infty \geq \tfrac{\varepsilon}{2(1+\delta)} N^\delta\},$$

denn auf $\{T_\delta < \log\log N\} \cap \{W_\infty < \frac{\varepsilon}{2(1+\delta)} N^\delta\}$ ist

$$\sup_{\log\log N \leq t \leq \frac{1-\delta}{s}\log N} Y_t \leq \sup_{\log\log N \leq t \leq \frac{1-\delta}{s}\log N} (1+\delta) e^{st} W_\infty \leq (1+\delta) e^{(1-\delta)\log N} \tfrac{\varepsilon}{2(1+\delta)} N^\delta = \frac{\varepsilon}{2} N.$$

Wegen

$$\limsup_{N\to\infty} \Big\{ \sup_{t \leq \log\log N} Y_t \geq \varepsilon N \Big\} \subset \{W_\infty = \infty\}$$

(und $\mathbb{P}(W_\infty = \infty) = 0$) gilt somit

$$\mathbb{P}\Big(\liminf_{N\to\infty} \big\{ \tfrac{1-\delta}{s} \log N \le \tilde\tau_{\varepsilon N} \le \tfrac{1+\delta}{s} \log N \big\} \Big) = 1.$$

□

Bericht. Es gilt auch $\mathbb{E}_1[\tilde\tau_{\varepsilon N} \,|\, Y_t \neq 0 \;\forall\, t \ge 0] \sim \dfrac{1}{s} \log N$.)

Für den Beweis von Satz 3.7 vergleichen wir verschiedene zeitkontinuierliche Markovketten (nämlich das Moran-Modell und einen Verzweigungsprozess), indem wir sie mittels Zeittransformation ineinander überführen. Die allgemeine Situation beschreibt das folgende Lemma.

Lemma 3.13. *Sei $(X_t)_{t\ge0}$ zeitkontinuierliche Markovkette auf E mit Q-Matrix (q_{ij}), sei $\varphi : E \to (0,\infty)$ (sagen wir, beschr. und glm. positiv), setze*

$$T_u := \int_0^u \varphi(X_v)\, dv, \quad u \ge 0,$$

((T_u) ist ein sogenanntes „additives Funktional“ von X).
$u \mapsto T_u$ hat stetige, strikt wachsende Pfade, die Inverse ist

$$T_t^{-1} := \inf\{u \ge 0 : T_u > t\} \quad \text{für } t \ge 0.$$

Sei $\widehat X_t := X_{T_t^{-1}}$, $t \ge 0$. $\widehat X = (\widehat X_t)_{t\ge0}$ ist zeitkontinuierliche Markovkette auf E mit Sprungratenmatrix $\widehat Q = (\widehat q_{ij})$, wobei

$$\widehat q_{ij} = \frac{q_{ij}}{\varphi(i)}.$$

Beweisskizze. Sei $X_0 = i$,

$$\tau_1 := \text{erster Sprungzeitpkt. von } X \;(\sim \mathrm{Exp}(-q_{ii})),$$

es ist $T_{\tau_1} = \tau_1 \varphi(i)$ und

$$t < T_{\tau_1} \iff T_t^{-1} < T_{T_{\tau_1}}^{-1} = \tau_1$$

d.h. $\widehat\tau_1 =$ erster Sprungzeitpkt. von $\widehat X = T_{\tau_1} = \tau_1\varphi(i)$ $(\sim \mathrm{Exp}(-q_{ii}/\varphi(i))$.

Dann verwendet man die starke Markov-Eigenschaft von X und die Tatsache, dass die Skelettketten von X und von $\widehat X$ nach Konstruktion übereinstimmen. □

Beweis von Satz 3.7. Wähle $\varepsilon > 0$.

1. Phase:

Wende Lemma 3.13 an auf $X^{(N)}$ (startend von $X_0^{(N)} = 1$) mit $\varphi(i) = 1 - \frac{i}{N}$, also

$$T_u = \int_0^u 1 - \frac{X_v^{(N)}}{N}\, dv, \quad T_t^{-1} := \inf\{u \ge 0 : T_u > t\}.$$

Sei $\widetilde Y_t := X^{(N)}_{T_t^{-1}}$ und

$$\widetilde\tau_{\varepsilon N} = \inf\{t \ge 0 : \widetilde Y_t \ge \varepsilon N\}.$$

$(\widetilde{Y}_t)_{t\geq 0}$ ist gemäß Lemma 3.13 verteilt wie ein superkritischer Galton-Watson-Prozess mit $\lambda = 1+s$, $\mu = 1$ (der gestoppt wird, sobald N Ind. erreicht).

Solange $\widetilde{X}_t^{(N)} \leq \varepsilon N$ gilt, ist $(1-\varepsilon)u \leq T_u \leq u$, also

$$t \leq T_t^{-1} \leq \frac{1}{1-\varepsilon} t \quad \text{und daher} \quad \tau_{\varepsilon N} = T^{-1}_{\widetilde{\tau}_{\varepsilon N}} \in \big[\widetilde{\tau}_{\varepsilon N}, \frac{1}{1-\varepsilon}\widetilde{\tau}_{\varepsilon N}\big]$$

Mit Beob. 3.12 gilt

$$\mathbb{P}_1\left(\frac{s}{\log N}\widetilde{\tau}_{\varepsilon N} \in \big(1-\varepsilon, \frac{1}{1-\varepsilon}\big) \,\Big|\, \widetilde{\tau}_N < \widetilde{\tau}_0\right) \to 1 \quad \text{für } N \to \infty.$$

Beachte: $\{\widetilde{Y} \text{ überlebt}\} \subset \{\widetilde{\tau}_N < \widetilde{\tau}_0\}$ mit $\mathbb{P}_1(\widetilde{Y} \text{ überlebt}) > 0$ und wegen

$$\mathbb{P}_1\big(\{\widetilde{\tau}_N < \widetilde{\tau}_0\} \cap \{\widetilde{Y} \text{ überlebt}\}^c\big) = \mathbb{P}_N(\widetilde{Y} \text{stirbt aus}) = 1/(1+s)^N \to 0 \quad \text{für } N \to \infty$$

spielt es keine Rolle, ob wir auf $\{\widetilde{\tau}_N < \widetilde{\tau}_0\}$ oder auf $\{\widetilde{Y} \text{ überlebt}\} = \big\{\lim_{t\to\infty} e^{-st}\widetilde{Y}_t > 0\big\}$ bedingen.

Daher gilt auch

$$\mathbb{P}_1\left(\frac{s}{\log N}\tau_{\varepsilon N} \in \big((1-\varepsilon)^2, \frac{1}{(1-\varepsilon)^2}\big) \,\Big|\, \tau_N < \tau_0\right) \to 1 \quad \text{für } N \to \infty.$$

2. Phase:

Zeige

$$\mathscr{L}(\tau_{(1-\varepsilon)N} - \tau_{\varepsilon N} \,|\, \tau_N < \tau_0),\ N \in \mathbb{N}, \ \text{ ist straff.}$$

Intuitiv ist plausibel (zumindest wenn man ohne die Bedingung $\tau_N < \tau_0$ argumentiert), dass $(\frac{1}{N}X_t^{(N)})_{t\geq 0}$ startend von $X_0^{(N)} = [Nx_0]$ mit $x_0 \in (0,1)$ für $N \to \infty$ gegen die Lösung der logistischen Differentialgleichung

$$\frac{d}{dt}y(t) = sy(t)(1-y(t)), \quad y(0) = x_0$$

konvergiert (beachte: wir transformieren hier nicht die Zeitachse).

Startend von $y(0) = \varepsilon$ erreicht die Kurve $(y(t))_{t\geq 0}$ den Wert $1-\varepsilon$ in endlicher Zeit.

Vgl. auch die Diskussion in Abschnitt 3.1, folgende heuristische Rechnung zeigt, dass die 2. Momente von Inkrementen trivial werden: Für $h \downarrow 0$ ist

$$\begin{aligned}
&\frac{1}{h}\mathbb{E}\big[\tfrac{1}{N}X_{t+h}^{(N)} - \tfrac{1}{N}X_t^{(N)} \,\big|\, \tfrac{1}{N}X_t^{(N)} = x\big] \\
&= \frac{1}{h}\frac{1}{N}\Big(h(1+s)Nx\frac{N-Nx}{N}\cdot(+1) + h(N-Nx)\frac{Nx}{N}\cdot(-1) + o(h)\Big) = sx(1-x) + o(1)
\end{aligned}$$

und

$$\begin{aligned}
&\frac{1}{h}\mathbb{E}\big[\big(\tfrac{1}{N}X_{t+h}^{(N)} - \tfrac{1}{N}X_t^{(N)}\big)^2 \,\big|\, \tfrac{1}{N}X_t^{(N)} = x\big] \\
&= \frac{1}{h}\frac{1}{N^2}\Big(h(1+s)Nx\frac{N-Nx}{N}\cdot(+1)^2 + h(N-Nx)\frac{Nx}{N}\cdot(-1)^2 + o(h)\Big) \\
&= \frac{1}{N}\big((2+s)x(1-x) + o(1)\big),
\end{aligned}$$

was im Licht der Diskussion in Abschnitt 1.2.1 nahelegt, dass $X^{(N)}$ gegen die Lösung von $dX_t = sX_t(1-X_t)\,dt + 0\,dW_t$, d.h. die Lösung einer deterministischen Differentialgleichung konvergiert.

Man kann dieses Argument präzise machen, für unsere Zwecke genügt hier eine gröbere Abschätzung via einem (erneuten) Vergleich mit einem superkritischen Galton-Watson-Prozess:

Für t genügend groß ($t > \frac{1}{s}\log\frac{1-\varepsilon}{\varepsilon}$) ist

$$\lim_{N\to\infty} \mathbb{P}_{[\varepsilon N]}(Y_t \geq (1-\varepsilon)N \,|\, \text{Überleben}) = 1$$

(verwende $\mathscr{L}_{[\varepsilon N]}(Y_t) = \mathscr{L}_1(Y_t)^{*[\varepsilon N]}$ und das Ges.d.gr.Z.)

Solange $X_u^{(N)} \leq (1-\varepsilon)N$ gilt, ist $\varepsilon u \leq T_u \leq u$, also $t \leq T_t^{-1} \leq \frac{1}{\varepsilon}t$, somit

$$\tau_{(1-\varepsilon)N} - \tau_{\varepsilon N} \in \big[\widetilde{\tau}_{(1-\varepsilon)N} - \widetilde{\tau}_{\varepsilon N}, \frac{1}{\varepsilon}(\widetilde{\tau}_{(1-\varepsilon)N} - \widetilde{\tau}_{\varepsilon N})\big]$$

(sofern $\tau_{\varepsilon N} < \infty$) und demnach gilt insbesondere

$$\mathbb{P}\Big(\tau_{(1-\varepsilon)N} - \tau_{\varepsilon N} \leq \frac{1}{s}\frac{1}{\varepsilon}\log\frac{1-\varepsilon}{\varepsilon}\Big) \underset{N\to\infty}{\longrightarrow} 1$$

(und man überlege, dass dies auch unter der Bedingung $\tau_N < \infty$ richtig bleibt: gegeben $\tau_{\varepsilon N} < \infty$ hat $\{\tau_N < \infty\}$ W'keit $\geq 1 - (1/1+s)^{\varepsilon N}$).

3. Phase:

Sobald $X^{(N)}$ den Wert $(1-\varepsilon)N$ erreicht, vertauschen wir die Rollen der Typen A und a :

$\overline{X}_t^{(N)} := N - X_t^{(N)}$ hat Sprungraten

$$q_{ij} = \begin{cases} i\frac{N-i}{N}, & j = i+1, \\ (1+s)i\frac{N-i}{N}, & j = i-1, \\ -(2+s)i\frac{N-i}{N}, & j = i, \\ 0, & \text{sonst.} \end{cases}$$

Wir starten in $\overline{X}_0^{(N)} = \varepsilon N$, die 3. Phase endet, sobald $\overline{X}_t^{(N)} = 0$ gilt (wörtlich müssten wir darauf bedingen, dass $\overline{X}^{(N)}$ niemals den Wert N übersteigt, dies fällt asymptotisch nicht ins Gewicht, da $\mathbb{P}_{\varepsilon N}(\overline{X}^{(N)}$ erreicht $N)$ exponentiell klein in N wird).

Wir vergleichen mit einem subkritischen Galton-Watson-Prozess $(\overline{Y}_t)$ (Individuen sterben mit Rate $1+s$, verdoppeln sich mit Rate 1) mit Sprungraten

$$\overline{q}_{ij} = \begin{cases} i, & j = i+1, \\ (1+s)i, & j = i-1, \\ -(2+s)i, & j = i, \\ 0, & \text{sonst.} \end{cases}$$

Prop. 3.10 liefert

$$\mathbb{E}_k[\overline{Y}_t] = ke^{-st}, \quad \mathrm{Var}_k[\overline{Y}_t] = k\frac{2+s}{s}(e^{-st} - e^{-2st}).$$

Somit gilt für $\delta > 0$:

$$\mathbb{P}_{[\varepsilon N]}(\overline{Y}_{\frac{1+\delta}{s}\log N} > 0) \le \mathbb{E}_{[\varepsilon N]}[\overline{Y}_{\frac{1+\delta}{s}\log N}] = [\varepsilon N]\exp(-(1+\delta)\log N) \sim \varepsilon N^{-\delta} \to 0,$$

andererseits ist (beachte $\mathbb{E}_{[\varepsilon N]}[\overline{Y}_{\frac{1-\delta}{s}\log N}] = [\varepsilon N]N^{\delta-1}$)

$$\begin{aligned}\mathbb{P}_{[\varepsilon N]}\left(\overline{Y}_{\frac{1-\delta}{s}\log N} = 0\right) &\le \mathbb{P}_{[\varepsilon N]}\left(|\overline{Y}_{\frac{1-\delta}{s}\log N} - \mathbb{E}_{[\varepsilon N]}[\overline{Y}_{\frac{1-\delta}{s}\log N}]| \ge \varepsilon N^{\delta}\right)\\ &\le \frac{1}{\varepsilon^2 N^{2\delta}}\mathrm{Var}_{[\varepsilon N]}\left(\overline{Y}_{\frac{1-\delta}{s}\log N}\right)\\ &= \frac{1}{\varepsilon^2 N^{2\delta}}[\varepsilon N]\frac{2+s}{s}\Big(\frac{1}{N^{1+\delta}} - \frac{1}{N^{2(1+\delta)}}\Big) \to 0\end{aligned}$$

Demnach gilt

$$\mathbb{P}_{[\varepsilon N]}\left(\frac{1-\delta}{s}\log N < \inf\{t \ge 0 : \overline{Y}_t = 0\} \le \frac{1+\delta}{s}\log N\right) \underset{N\to\infty}{\longrightarrow} 1$$

und Argumente wie für Phase 1 zeigen, dass Analoges für $\overline{X}^{(N)}$ gilt.

Schließlich folgt die Behauptung mit $\delta \downarrow 0$, dann $\varepsilon \downarrow 0$. □

3.4 (2-Typ) Moran-Modell mit (gerichteter) Selektion und Mutation

Wir nehmen nun zusätzlich zur Dynamik aus Abschnitt 3.3 die Möglichkeit der *Mutation* an:

- Jedes Typ a-Individuum mutiert mit Rate m_A zu Typ A,
- jedes Typ A-Ind. mutiert mit Rate m_a zu Typ a

mit gewissen $m_A, m_a \ge 0$.

Definition 3.14 ((2 Typ-)Moran-Modell mit gerichteter Selektion und Mutation)**.** Die zeitkontinuierliche Markovkette $X^{(N)}$ auf $[N]$ mit Sprungratenmatrix

$$q_{ij} = \begin{cases}(1+s)i\frac{N-i}{N} + (N-i)m_A, & j = i+1,\\ (N-i)\frac{i}{N} + im_a, & j = i-1,\\ -(2+s)\frac{i(N-i)}{N} + (N-i)m_A + im_a, & j = i,\\ 0, & \text{sonst.}\end{cases}$$

interpretieren wir als

$$X_t^{(N)} = \#\text{Typ } A\text{-Individuen zur Zeit } t, \quad t \ge 0$$

in einer Population der Größe N wie oben beschrieben. $X^{(N)}$ heißt (Typenzählprozess des) 2 Typ-Moran-Modell(s) mit gerichteter Selektion und Mutation.

Bemerkung. Für gegebenes $N \in \mathbb{N}$, $s \geq 0$ und $m_a, m_A > 0$ ist $X^{(N)}$ offensichtlich irreduzibel, demnach existiert ein eindeutiges Gleichgewicht.

Beobachtung 3.15 (Gleichgewichte von Geburts- und Todesprozessen)**.** Betrachte eine zeitkontinuierliche Markovkette X auf $\{0, 1, \dots, N\}$ mit Sprungraten

$$q_{i,i+1} = \lambda_i, \quad q_{i,i-1} = \mu_i, \quad q_{i,i} = -(\lambda_i + \mu_i)$$

mit $\lambda_i > 0$ für $0 \leq i < N$, $\mu_i > 0$ für $0 < i \leq N$, $\lambda_N = \mu_0 = 0$ (ein solches X heißt auch ein Geburts- und Todesprozess).

X besitzt eine reversible Gleichgewichtsverteilung $(\pi_k)_{k=0,1,\dots,N}$, d.h. (π_k) erfüllt

$$\pi_{k+1}\mu_{k+1} = \pi_k \lambda_k \quad \text{für } k = 0, \dots, N-1 \tag{3.11}$$

(„detaillierte Balance“). Folglich gilt

$$\pi_k = \frac{\lambda_{k-1}}{\mu_k}\pi_{k-1} = \cdots = \pi_0 \prod_{j=1}^{k} \frac{\lambda_{j-1}}{\mu_j}$$

und somit

$$\pi_k = \frac{\varphi_k}{\sum_{j=0}^{N} \varphi_j} \quad \text{mit} \quad \varphi_k := \prod_{j=1}^{k} \frac{\lambda_{j-1}}{\mu_j} \quad (\text{mit } \varphi_0 = 1).$$

Aus der detaillierten Balancegleichung (3.11) folgt natürlich die Gleichgewichtsbedingung $\pi_k(\lambda_k + \mu_k) = \pi_{k+1}\mu_{k+1} + \pi_{k-1}\lambda_{k-1}$ (für $k = 0, \dots, N$, mit $\pi_{N+1} := \pi_{-1} := 0$), d.h. die (wegen Irreduzibilität eindeutige) Gleichgewichtsverteilung hat tatsächlich diese Gestalt.

Geburts- und Todesprozesse sind also stets reversibel, dies ist eine Spezialität der 1-dim. Situation.

Aus dem Zusammenspiel der verschiedenen evolutionären Kräfte ergibt sich nun ein Mutations-Selektions-Drift-Gleichgewicht:

Satz 3.16 (Mutations-Selektions-Drift-Gleichgewicht im Moran-Modell)**.** *Seien $s \geq 0$, $m_A, m_a > 0$,*

$$\pi^{(N)}(k) = \lim_{t\to\infty} \mathbb{P}(X_t^{(N)} = k), \quad k = 0, \dots, N$$

die Gleichgewichts-Verteilung des 2 Typ-Moran-Modells mit gerichteter Selektion und Mutation (aus Def. 3.14).

a) Es ist

$$\pi^{(N)}(k) = \frac{1}{C_N(m_A, m_a, s)}(1+s)^k \binom{N}{k} \frac{\left(\frac{Nm_A}{1+s}\right)_{k\uparrow} \left(Nm_a\right)_{(N-k)\uparrow}}{\left(N\frac{m_A+(1+s)m_a}{1+s}\right)_{N\uparrow}}$$

mit einer Normierungskonstante $C_N(m_A, m_a, s)$.

Die Normierungskonstante erfüllt

$$C_N(m_A, m_a, s) = \mathbb{E}\left[(1+s\xi)^N\right] \quad \textit{mit } \xi \sim \mathrm{Beta}\left(\frac{Nm_A}{1+s}, Nm_a\right).$$

b) Mit $\theta_A, \theta_a > 0$, $\sigma \geq 0$ gelte

$$s = \sigma/N, m_A = \theta_A/N, m_a = \theta_a/N, \tag{3.12}$$

dann konvergiert $\sum_{k=0}^{N} \pi^{(N)}(k)\delta_{k/N}$ *für* $N \to \infty$ *(schwach als W-Maß auf* $[0,1]$*) gegen das Wahrscheinlichkeitsmaß mit Dichte*

$$\frac{1}{C_{\theta_A,\theta_a,\sigma}} x^{\theta_A-1}(1-x)^{\theta_a-1}e^{\sigma x}, \quad x \in (0,1), \tag{3.13}$$

wobei $C_{\theta_A,\theta_a,\sigma} = \int_0^1 x^{\theta_A-1}(1-x)^{\theta_a-1}e^{\sigma x}\,dx$.

Bemerkung. 1. Im neutralen Fall $\sigma = 0$ handelt es sich bei (3.13) um ein Beta-Integral, $C_{\theta_A,\theta_a,0} = \mathrm{Beta}(\theta_A,\theta_a) = \frac{\Gamma(\theta_A)\Gamma(\theta_a)}{\Gamma(\theta_A+\theta_a)}$, dies hatten wir bereits in Bericht 2.4 gesehen.

2. Annahme (3.12) bewirkt, dass die „evolutionären Kräfte" Selektion, Mutation und Gendrift (im Modell) auf vergleichbarer Zeitskala wirken. Anders gewendet: Das Modell passt zu einer gegebenen Situation, wenn die Population groß und Mutationsraten und Selektionsvorteil (dazu quantitativ passend) klein sind – s.a. die entsprechende Diskussion in Kap. 2.1, speziell die Bemerkung auf Seite 54 zur analogen Annahme (2.2) im neutralen Fall.

Beweis von Satz 3.16. a)

$$\prod_{i=0}^{k-1}\lambda_i = \prod_{i=0}^{k-1}\Big(\underbrace{(1+s)i\frac{N-i}{N} + (N-i)m_A}_{=\frac{1}{N}(1+s)(N-i)\left(i+\frac{Nm_A}{1+s}\right)}\Big) = \frac{1}{N^k}(1+s)^k(N)_{k\downarrow}\prod_{i=0}^{k-1}\Big(\frac{Nm_A}{1+s}+i\Big)$$
$$= \frac{1}{N^k}(1+s)^k(N)_{k\downarrow}\Big(\frac{Nm_A}{1+s}\Big)_{k\uparrow},$$
$$\prod_{i=1}^{k}\mu_i = \prod_{i=1}^{k}\Big(\underbrace{(N-i)\frac{i}{N} + im_a}_{=\frac{i}{N}(N-i+Nm_a)}\Big) = \frac{k!}{N^k}\prod_{i=1}^{k}(N-i+Nm_a) = \frac{k!}{N^k}\frac{(Nm_a)_{N\uparrow}}{(Nm_a)_{(N-k)\uparrow}},$$

also

$$\pi^{(N)}(k) \propto \frac{\prod_{i=0}^{k-1}\lambda_i}{\prod_{i=1}^{k}\mu_i} \propto (1+s)^k\binom{N}{k}\frac{\left(\frac{Nm_A}{1+s}\right)_{k\uparrow}\left(Nm_a\right)_{(N-k)\uparrow}}{\left(N\frac{m_A+(1+s)m_a}{1+s}\right)_{N\uparrow}}$$

Für die Normierung beachte : $\xi \sim \mathrm{Beta}(\alpha_1,\alpha_2)$ erfüllt

$$\mathbb{E}[\xi^k(1-\xi)^\ell] = \frac{\mathrm{Beta}(\alpha_1+k,\alpha_2+\ell)}{\mathrm{Beta}(\alpha_1,\alpha_2)} = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\frac{\Gamma(\alpha_1+k)\Gamma(\alpha_2+\ell)}{\Gamma(\alpha_1+\alpha_2+k+\ell)}$$

und $\Gamma(\alpha+k)/\Gamma(\alpha) = (\alpha)_{k\uparrow}$, (wende an mit $\alpha_1 = \frac{Nm_A}{1+s}$, $\alpha_2 = Nm_a$), erhalte

$$\mathbb{E}\left[(1+s\xi)^N\right] = \mathbb{E}\left[\big((1+s)\xi+(1-\xi)\big)^N\right] = \sum_{k=0}^{\infty}(1+s)^k\binom{N}{k}\mathbb{E}\left[\xi^k(1-\xi)^{N-k}\right].$$

b) Schreibe

$$\lambda_i = N\frac{i}{N}\Big(1-\frac{i}{N}\Big)\Big(1+s+\frac{m_A}{i/N}\Big) \quad (i \neq 0,\ \lambda_0 = Nm_A),$$
$$\mu_i = N\frac{i}{N}\Big(1-\frac{i}{N}\Big)\Big(1+\frac{m_a}{1-i/N}\Big) \quad (i \neq N,\ \mu_N = Nm_a).$$

$$\varphi_k^{(N)} = \frac{\prod_0^{k-1}\lambda_i}{\prod_1^k \mu_i} = \frac{Nm_A}{N\frac{k}{N}\big(1-\frac{k}{N}\big)\big(1+\frac{m_a}{1-k/N}\big)} \prod_{i=1}^{k-1} \frac{1+s+\frac{m_A}{i/N}}{1+\frac{m_a}{1-i/N}}$$
$$= \frac{1}{N}\frac{\theta_A}{\frac{k}{N}\big(1-\frac{k}{N}\big)\big(1+\frac{1}{N}\frac{\theta_a}{1-k/N}\big)} \prod_{i=1}^{k-1} \frac{1+\frac{\sigma}{N}+\frac{\theta_A}{i}}{1+\frac{\theta_a}{N-i}}$$

Sei $k = \lfloor Nx \rfloor$ mit $x \in [\varepsilon, 1-\varepsilon]$ (ε zunächst fest)

$$\log \prod_{i=1}^{k-1} \frac{1+\frac{\sigma}{N}+\frac{\theta_A}{i}}{1+\frac{\theta_a}{N-i}} = \sum_{i=1}^{k-1} \log\left(1+\frac{\sigma}{N}+\frac{\theta_A}{i}\right) - \log\left(1+\frac{\theta_a}{N-i}\right)$$
$$= \sum_{i=1}^{k-1} \frac{\sigma}{N} + \frac{\theta_A}{i} - \frac{\theta_a}{N-i} + R_{N,1}(k)$$

mit $|R_{N,1}(k) - \tilde{R}_\varepsilon| \le C/N^2$ (beachte $\log(1+x) = x + O(x^2)$), dann verwende

$\sum_{i=1}^k \frac{1}{i} = \log k + c_\gamma + o(1)$ mit $c_\gamma \approx 0.57721...$, die Euler-Mascheroni-Konstante

(vgl. auch Fußnote 7 im Beweis von Beob. 1.21).

Somit ist

$$\varphi_k^{(N)} = \frac{1}{N}\frac{\theta_A}{\frac{k}{N}\big(1-\frac{k}{N}\big)\big(1+\frac{1}{N}\frac{\theta_a}{1-k/N}\big)}$$
$$\times \exp\left(\frac{k-1}{N}\sigma + \theta_A \log(k-1) - \theta_a \log(N-1) + \theta_a \log(N-k+1)\right)$$
$$\times \exp\big(\tilde{R}_\varepsilon + O(1/N^2)\big)$$
$$= \frac{C}{N} x^{\theta_A - 1}(1-x)^{\theta_a - 1} e^{\sigma x}\big(1 + O(1/N^2)\big).$$

□

Bericht 3.17 (Konvergenz gegen die Wright-Fisher-Diffusion mit Mutation und Selektion)**.** Unter (3.12) konvergiert

$$\big(\frac{1}{N} X^{(N)}_{tN/2}\big)_{t\ge 0} \to X = (X_t)_{t\ge 0} \quad \text{für } N \to \infty$$

(in Vert. auf $D([0,\infty);[0,1])$), wobei X Lösung von

$$dX_t = \Big(\frac{\sigma}{2} X_t(1-X_t) + \frac{1}{2}(\theta_A - (\theta_A+\theta_a)X_t)\Big)\,dt + \sqrt{X_t(1-X_t)}\,dB_t$$

mit (B_t) Standard-Brownbewegung. X heißt die (2-Typ) Wright-Fisher-Diffusion mit (gerichteter) Selektion und Mutation, sie ist ein Markovprozess auf $[0,1]$ mit Generator

$$Lf(x) = \Big(\frac{\sigma}{2}x(1-x) + \frac{1}{2}(\theta_A - (\theta_A+\theta_a)x)\Big)f'(x) + \frac{1}{2}x(1-x)f''(x) \quad \text{für } f \in C^2([0,1]).$$

Heuristisch können wir hier zumindest beobachten, dass für $h \downarrow 0$ gilt

$$\begin{aligned}
&\mathbb{E}\Big[\frac{1}{N}\big(X^{(N)}_{(t+h)N/2} - X^{(N)}_{tN/2}\big)\Big|X^{(N)}_{tN/2} = k\Big] \\
&= \frac{1}{N}\Bigg(h\frac{N}{2}\Big((1+\tfrac{\sigma}{N})k\frac{N-k}{N} + (N-k)\frac{\theta_A}{N}\Big)(+1) \\
&\qquad + h\frac{N}{2}\Big(k\frac{N-k}{N} + k\frac{\theta_a}{N}\Big)(-1) + O(h^2)\Bigg) \\
&= h\Big(\frac{\sigma}{2}\frac{k}{N}\big(1-\frac{k}{N}\big) + \big(1-\frac{k}{N}\big)\frac{\theta_A}{2} - \frac{k}{N}\frac{\theta_a}{2}\Big) + O(h^2)
\end{aligned}$$

und

$$\begin{aligned}
&\mathbb{E}\Big[\frac{1}{N^2}\big(X^{(N)}_{(t+h)N/2} - X^{(N)}_{tN/2}\big)^2\Big|X^{(N)}_{tN/2} = k\Big] \\
&= \frac{1}{N^2}\Bigg(\frac{N}{2}h\Big((2+\tfrac{\sigma}{N})k\frac{N-k}{N} + \big(1-\frac{k}{N}\big)\theta_A + \frac{k}{N}\theta_a\Big)\Bigg) + O(h^2) \\
&= h\frac{k}{N}\big(1-\frac{k}{N}\big) + O(h/N) + O(h^2),
\end{aligned}$$

was zu obigem Generator führt (vgl. auch Satz 1.26 für die entspr. Konv. des Wright-Fisher-Modells im neutralen Fall).

Die Dichte (3.13)

$$\frac{1}{C_{\theta_A,\theta_a,\sigma}} x^{\theta_A - 1}(1-x)^{\theta_a - 1} e^{\sigma x}$$

aus Satz 3.16 ist die Gleichgewichtsdichte der Wright-Fisher-Diffusion mit Selektion und Mutation, siehe auch Diskussion und Referenzen in Bericht 2.4:

Es ist (in der Notation von Bericht 2.4)

$$\mu(x) = \big(\frac{\sigma}{2}x(1-x) + \frac{1}{2}(\theta_A - (\theta_A+\theta_a)x)\big), \quad \sigma^2(x) = x(1-x),$$

somit

$$\begin{aligned}
\int_{1/2}^{x} \frac{2\mu(z)}{\sigma^2(z)}\,dz &= \int_{1/2}^{x} \sigma - \frac{(\theta_A+\theta_a)z - \theta_A}{z(1-z)}\,dz = \int_{1/2}^{x} \sigma - \frac{\theta_a}{1-z} + \frac{\theta_A}{z}\,dz \\
&= \sigma x + \theta_a \log(1-x) + \theta_A \log(x) + \text{const.}
\end{aligned}$$

und

$$\rho(x) = C\exp\Big(\sigma x + \theta_a\log(1-x) + \theta_A\log(x)\Big)\frac{1}{x(1-x)} = C\,x^{\theta_A-1}(1-x)^{\theta_a-1}e^{\sigma x}.$$

3.4.1 Graphische Konstruktion und anzestraler Selektionsgraph

Für jedes geordnete Paar (i,j), $i \neq j$ sei $(N_t^{(i,j)})_{t\geq 0}$ ein Poissonprozess mit Rate $1/N$, $(S_t^{(i,j)})_{t\geq 0}$ ein Poissonprozess mit Rate s/N, für jedes i sei $(M_t^{(a,i)})_{t\geq 0}$ ein Poissonprozess mit Rate m_a und $(M_t^{(A,i)})_{t\geq 0}$ ein Poissonprozess mit Rate m_A.

Zusätzlich zu den Pfeilen wie in Bem. 3.4 (für den Fall ohne Mutationen) legen wir nun noch „Mutationsereignisse“ auf die „Zeitachsen“: ein Sprung von $M^{(a,i)}$ ändert den Typ von Individuum i auf a, ein Sprung von $M^{(A,i)}$ ändert ihn auf A (wobei „triviale Wechsel“ erlaubt sind, z.B. bleibt ein Typ A-Individuum i nach einem Sprung von $M^{(A,i)}$ vom Typ A).

`[Skizze an der Tafel]`

Beobachtung 3.18 (Ablesen der „Genealogie“ und anzestraler Selektionsgraph (ASG))**.**
1. Eine n-Stichprobe und ihre Genealogie können in einem zweistufigen Prozess gewonnen werden

- zuerst „rückwärts“ (in der Zeit): Linien verschmelzen oder spalten (wenn von der Spitze eines s-Pfeils getroffen), verfolge bis # Linien = 1 (der “ultimate ancestor” ist erreicht)
- dann „vorwärts“: lege Mutationen (die die Typen bestimmen), löse damit potentielle Verzweigungsereignisse auf

  (sobald wir wissen, ob ein gewisses Individuum zum Zeitpunkt eines s-Pfeils Typ A oder a hat, können entscheiden, ob der s-Pfeil benutzt wurde).

2. Alternative Rückwärtsdynamik in „einem Schritt“: Sei $(A_t)_{t\geq 0}$ zeitkontinuierliche Markovkette mit Werten in $2^{\{1,\dots,N\}} \cup \{\partial\}$, wenn aktuell $A_{t-} = B \subset \{1,\dots,N\}$:

$$\begin{aligned}
&\text{für } i \in B\text{: springe von } A_{t-} = B \text{ nach } A_t = B \setminus \{i\} \text{ mit Rate } \frac{|B|-1}{N} + m_a,\\
&\text{für } j \notin B\text{: springe von } A_{t-} = B \text{ nach } A_t = B \cup \{j\} \text{ mit Rate } \frac{s}{N}|B|,\\
&\text{springe nach } \partial \text{ mit Rate } m_A|B| \text{ (der Prozess wird „getötet“)}
\end{aligned}$$

Sei $t > 0$, $n \leq N$, $i_1,\dots,i_n \in \{1,\dots,N\}$, starte in $A_0 = \{i_1,\dots,i_n\}$. Es ist

$$\{X_t(i_1) = a,\dots,X_t(i_n) = a\} = \{A_t \text{ nicht getötet}\} \cap \{X_0(j) = a, \text{ für alle } j \in A_t\}, \quad (3.14)$$

wie wir von der graphischen Konstruktion ablesen.

Der Zählprozess $(Y_t^{(N)})$ mit

$$Y_t^{(N)} = |A_t|, \quad \text{aktuelle Anzahl „potentieller Ahnenlinien“,}$$

hat dann (offenbar) Sprungratenmatrix

$$q_{ij}^{Y,N} = \begin{cases} s\frac{i(N-i)}{N}, & j = i+1,\\ \frac{i(i-1)}{N} + im_a, & j = i-1,\\ im_A, & j = \partial,\\ -(1+s)\frac{i(N-i)}{N} - i(m_a+m_A), & j = i \end{cases}$$

Nehmen wir an $X_0^{(N)}(1), \dots, X_0^{(N)}(N)$ sind austauschbar verteilt (und u.a. von den PPPen der graphischen Konstruktion)

Sei $n \le N$, $t \ge 0$, ziehe n-mal ohne Zurücklegen aus der Population zur Zeit t. Die W'keit, dann n-mal Typ a zu sehen, ist somit

$$\mathbb{E}\left[\frac{(N - X_t^{(N)})(N - X_t^{(N)} - 1)\cdots(N - X_t^{(N)} - n + 1)}{N(N-1)\cdot(N-n+1)}\right] = \frac{1}{(N)_{n\downarrow}}\mathbb{E}\left[\left(N - X_t^{(N)}\right)_{n\downarrow}\right]$$
$$= \frac{1}{(N)_{n\downarrow}}\mathbb{E}\left[\left(N - X_0^{(N)}\right)_{Y_t^{(N)}\downarrow} \mathbf{1}(Y_t^{(N)} \text{ nicht getötet})\right]$$

(durch Ablesen von der graphischen Konstruktion, vgl. (3.14)).

Beobachtung 3.19. $(Y^{(N)}_{tN/2})_{t\ge 0} \to Y = (Y_t)_{t\ge 0}$ (in Vert. auf $D([0,\infty); \mathbb{N}_0 \cup \{\partial\})$), Y zeitk. MK auf $\mathbb{N}_0$ mit Sprungraten

$$q_{ij}^Y = \begin{cases} \frac{\sigma}{2} i, & j = i+1, \\ \binom{i}{2} + i\frac{\theta_a}{2}, & j = i-1, \\ i\frac{\theta_A}{2}, & j = \partial, \end{cases}$$

Die Kette Y ist dual zur Wright-Fisher-Diffusion mit Mutation und Selektion X aus Bericht 3.17 (vgl. auch Bem. 2.6 für den Fall ohne Selektion):

$$\mathbb{E}\left[(1 - X_t)^n \,\middle|\, X_t = x_0\right] = \mathbb{E}\left[(1-x_0)^{Y_t} e^{-\frac{\theta_A}{2}\int_0^t Y_u\, du} \,\middle|\, Y_0 = n\right] \tag{3.15}$$

Betrachte $f(x,n) := (1-x)^n$, es ist

$$\begin{aligned}\left(L^X f(\cdot, n)\right)(x) =& \left(\frac{\sigma}{2}x(1-x) + \frac{1}{2}(\theta_A - (\theta_A + \theta_a)x)\right) n(1-x)^{n-1}(-1) \\ &+ \frac{1}{2}x(1-x)n(n-1)(1-x)^{n-2} \\ =& \frac{\sigma}{2}n\left((1-x)^{n+1} - (1-x)^n\right) + \frac{n(n-1)}{2}\left((1-x)^{n-1} - (1-x)^n\right) \\ &+ \frac{\theta_a}{2}n\left((1-x)^{n-1} - (1-x)^n\right) - \frac{\theta_A}{2}n\,(1-x)^n \\ =& \left(L^Y f(x,\cdot)\right)(n) - \frac{\theta_A}{2}n\, f(x,n),\end{aligned}$$

dann argumentiere analog zu Bem. 2.6.

Interpretation von (3.15) via ancestral selection graph :

- Beginne mit n Linien,
- jede Linie verzweigt mit Rate $\sigma/2$,
- jedes Paar Linien verschmilzt mit Rate 1,
- jede Linie „endet“ mit Rate $\theta_a/2$,

- der gesamte Prozess wird getötet mit Rate aktuelle #Linien $\cdot\, \theta_A/2$.
- Nach Zeit t, sofern der Prozess noch nicht getötet wurde, wählen wir für jede der dann existierenden Linien u.a. einen Typ (A mit W'keit x_0, a mit W'keit $1 - x_0$)

Die rechte Seite von (3.15) ist die Wahrscheinlichkeit, dass Prozess nicht getötet wurde und am Ende alle Linien Typ a erhalten haben, dies ist zugleich die Wahrscheinlichkeit, in einer n-Stichprobe aus einer Population, deren Typenanteils-Evolution von der Wright-Fisher-Diffusion mit Mutation und Selektion beschrieben wird, n-mal den (selektiv benachteiligten) Typ a zu sehen.

Mit $t \to \infty$ kann man auf diese Weise die Momente der Gleichgewichtsverteilung von X beschreiben. Für den Prozess Y genügt es dabei zu warten, bis der sogenannte "ultimate ancestor" erreicht ist (und sein Typ durch eine Mutation festgelegt ist, wörtlich also bis $Y_t = 0$ gilt oder der Prozess getötet wird).

Literaturverzeichnis

[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*, 2. Aufl., Springer, 2008.

[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

[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.