Prof. Dr. Matthias Birkner JGU Mainz

Stochastische Modelle der Populationsbiologie: Übungen und Ergänzungen **WS 2022/23**

Blatt 7

Aufgabe 7.1 (Verteilung der Anzahl Mutationen in einer n-Stichprobe im IMS-Modell) Sei S_n die Anzahl Mutationen in einer n-Stichprobe im Infinite-sites-Modell mit Mutationsratenparameter θ (d.h. Mutationen erscheinen mit Rate $\theta/2$ längs den Ästen des n-Koaleszenten). Zeigen Sie

$$\mathbb{P}_\theta\left(S_n = m\right) = \frac{n-1}{\theta}\sum_{k=1}^{n-1}(-1)^{k-1}\binom{n-2}{k-1}\Big(\frac{\theta}{\theta+k}\Big)^{m+1}, \quad m \in \mathbb{N}_0$$

$$\mathbb{P}_\theta\left(S_n \leqslant s\right) = 1 - \sum_{k=1}^{n-1}(-1)^{k-1}\binom{n-1}{k}\Big(\frac{\theta}{\theta+k}\Big)^{s+1}, \quad s \in \mathbb{N}_0$$

Hinweis. Sie können folgendermaßen vorgehen: Stellen Sie $S_n = S_{n,n} + S_{n,n-1} + \cdots + S_{n,2}$ dar, wobei $S_{n,j}$ die Anzahl Mutationen ist, die in der Genealogie auftreten, während es genau j Äste gibt. Verwenden Sie dies, um einen Ausdruck für die erzeugende Funktion $\mathbb{E}_\theta[u^{S_n}]$ anzugeben, lesen Sie dann deren Koeffizienten ab.

Aufgabe 7.2 ("Variation is the spice of life") a) Wieviele Mutationen, die zwischen 1% und 99% der Population betreffen (sogenannte SNPs), erwarten Sie in einer sehr großen Population, wenn Mutationen mit Rate $\theta/2$ gemäß dem Infinite-sites-Modell auftreten?

b) Nehmen wir an, für das menschliche Genom, dessen Gesamtlänge ca. 3×10^9 Basenpaare beträgt, gilt $\theta \approx 1/1331$ pro Basenpaar (auf der evolutionären Zeitskala). Wieviele SNPs würden wir dann in der Population ungefähr erwarten?

(Die Zahlenwerte stammen aus L. Kruglyak, D. A. Nickerson, Variation is the spice of life, *Nature Genetics* 156 (2001), 234–236, zitiert nach R. Durrett, *Probability Models for DNA sequence evolution*, 2nd ed., Springer 2008, S. 51.)

Hinweis. Erinnern Sie sich an die Formel $\mathbb{E}_\theta[\xi_i^{(n)}] = \theta/i$ für die Erwartungswerte des Frequenzspektrums.

Aufgabe 7.3* (Eine Metrik für càdlàg-Pfade und Konvergenz im diskreten Fall) a) Sei E ein polnischer Raum, $D([0,\infty), E)$ die Menge aller Funktionen $f : [0,\infty) \to E$, die rechtsstetig sind und an jeder Stelle einen linken Limes besitzen (oft mit dem französischen Akronym càdlàg bezeichnet).

Sei $\Lambda := \{\lambda : [0,\infty) \to [0,\infty) : \lambda \text{ bijektiv und stetig}\}$, die Lipschitz-Konstante von $\lambda(\cdot)$ bezeichnen wir mit $\gamma(\lambda) := \sup_{0\leqslant s<t} \big|\log \frac{\lambda(t)-\lambda(s)}{t-s}\big|$ ($\leqslant \infty$). Für $f, g \in D([0,\infty), E)$ definieren wir die Skorokhod-Metrik als

$$d(f,g) := \inf_{\lambda\in\Lambda}\Big\{\gamma(\lambda) \vee \int_0^\infty e^{-t}\sup_{s\geqslant 0} d_E\big(f(s\wedge t), g(\lambda(s)\wedge t)\big)\,dt\Big\}$$

Zeigen Sie: Dies *ist* eine Metrik [d.h. $d(\cdot,\cdot)$ ist symmetrisch, $d(f,g) = 0 \iff f = g$, die Dreiecksungleichung gilt], damit ausgestattet ist $D([0,\infty), E)$ ein vollständiger und separabler metrischer Raum.

[*Hinweis.* Vgl. auch Kap. 3.5 in S.N. Ethier, T.G. Kurtz, *Markov processes: characterization and convergence*, Wiley, 1986]

b) Sei nun $|E| < \infty$ und E mit der diskreten Metrik $d_E(x,y) = \mathbf{1}(x \neq y)$ ausgestattet. Für $f \in D([0,\infty), E)$ sei $\tau_0^{(f)} := 0$, sofern $\tau_{i-1}^{(f)} < \infty$ setzen wir

$$\tau_i^{(f)} := \inf\left\{t > \tau_{i-1}^{(f)} : f(t) \neq f(\tau_{i-1}^{(f)})\right\} \quad i \in \mathbb{N}$$

(ansonsten sei auch $\tau_i^{(f)} = \infty$) und

$$\xi_j^{(f)} := f(\tau_j^{(f)}) \quad \text{für } j \in \mathbb{N}_0 \text{ mit } j \leqslant a^{(f)} := \inf\left\{k : \tau_k^{(f)} = \infty\right\} - 1$$

Seien $f_n, f \in D([0,\infty), E)$. Zeigen Sie: Die Folge $\tau_{i-1}^{(f)}$, $i \in \mathbb{N}$ besitzt keine Häufungspunkte im Endlichen. Es gilt $d(f_n, f) \to 0$ für $n \to \infty$ genau dann, wenn gilt

1. $a^{(f_n)} \to a^{(f)}$ für $n \to \infty$
2. für jedes $k \in \mathbb{N}$ ist $\xi_{k \wedge a^{(f)}}^{(f_n)} = \xi_{k \wedge a^{(f)}}^{(f)}$ für alle genügend großen n und
3. für jedes $k \in \mathbb{N}$ gilt $\tau_{k \wedge a^{(f)}}^{(f_n)} \underset{n \to \infty}{\longrightarrow} \tau_{k \wedge a^{(f)}}^{(f)}$.

c) In der Vorlesung hatten wir folgende Situation betrachtet: Sei E endliche Menge, $Q = (q_{xy})_{x,y \in E}$ eine Sprungratenmatrix, $X^{(N)}$, $N \in \mathbb{N}$ zeitdiskrete E-wertige Markovketten mit zugehörigen Übergangsmatrizen

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

wo $c_N \to 0$ für $N \to \infty$ und $X_0^{(N)} = x_0 \in E$. Wir hatten argumentiert, dass dann die (zeitlich reskalierten) Prozesse $(X_{\lfloor t/c_N \rfloor}^{(N)})_{t \geqslant 0}$ für $N \to \infty$ gegen die zeitkontinuierliche Markovkette X mit Sprungratenmatrix Q im Sinne der endlich-dimensionalen Verteilungen konvergieren, d.h. für $0 < t_1 < t_2 < \cdots < t_k$, $x_1, \ldots, x_k \in E$, $k \in \mathbb{N}$ gilt

$$\mathbb{P}_{x_0}(X_{t_1}^{(N)} = x_1, \ldots, X_{t_k}^{(N)} = x_k) \underset{N \to \infty}{\longrightarrow} \mathbb{P}_{x_0}(X_{t_1} = x_1, \ldots, X_{t_k} = x_k)$$

Zeigen Sie: Man kann $X^{(N)}$ und X so koppeln, dass auch gilt

$$\limsup_{N \to \infty} \mathbb{P}_{x_0}\big(d(X^{(N)}, X) > \varepsilon\big) = 0 \quad \text{für jedes } \varepsilon > 0$$

und folgern Sie, dass die $X^{(N)}$, aufgefasst als $D([0,\infty), E)$-wertige Zufallsvariablen, in Verteilung gegen X konvergieren.