Westfälische Wilhelms-Universität Münster
Institut für Mathematische Statistik

Mathematische Modellierung und Analyse der Polymerase-Kettenreaktion

Diplomarbeit

vorgelegt von
Carsten Magnus

Thema gestellt von
Professor Dr. G. Alsmeyer

29. September 2006

Inhaltsverzeichnis

Einleitung

Am 17. Juli 2006 befand sich eine Meldung folgenden Wortlautes in der Frankfurter Allgemeinen Zeitung:

> COSWIG, 16. Juli (dpa). In Sachsen hat am Wochenende ein Massengentest zur Ermittlung eines Sexualstraftäters begonnen. Das Landeskriminalamt bezeichnete den Beginn der Aktion als Erfolg. Mehr als 1000 Männer zwischen 25 und 45 Jahren hatten in Coswig bei Dresden freiwillig eine Speichelprobe abgegeben. Das entspricht etwa einem Drittel der dort für den Test in Frage kommenden Männer. Die Untersuchung wird am nächsten Wochenende fortgesetzt werden. Die Behörden haben sich auf bis zu 100 000 Teilnehmer im Raum Dresden eingerichtet. Das wäre der bislang größte derartige Test in Deutschland. Mit der Massenuntersuchung soll der Vergewaltiger von zwei Mädchen ausfindig gemacht werden. An beiden Tatorten wurde eine übereinstimmende DNA-Spur gefunden. Bei den bisherigen Ermittlungen wurden die Alibis von rund 3500 Personen überprüft und mehr als 1000 Männer durch einen DNA-Vergleich als Täter ausgeschlossen.

Bei der Untersuchung der DNA-Proben der Testpersonen kommt eine biochemische Methode zum Einsatz, die seit ihrer Entdeckung in den 1980ern aus biochemischen Laboren nicht mehr wegzudenken ist: die Polymerase-Kettenreaktion. Diese Reaktion ermöglicht es, eine kleine Menge an DNA-Sequenzen so zu vervielfältigen, dass eine analysierbare Menge entsteht. Zur Erstellung eines genetischen Fingerabdrucks als Mittel zur Aufklärung der oben beschriebenen Straftat werden nur eine Speichelprobe der in Frage kommenden potentiellen Täter und ein paar Haare, Fingernägel oder sonstige Körperzellen des Täters benötigt. Mit Hilfe chemischer Verfahren wird die sich in diesen Zellen befindende DNA extrahiert, mit der Polymerase-Kettenreaktion vervielfältigt und die entstandenen Produkte analysiert. Durch einen Vergleich der so gewonnen Daten kann dann der Täter einer Straftat überführt werden, oder wenn dies nicht möglich ist, zumindest potentielle Täter ausgeschlossen werden - wie dies im oben beschriebenen Delikt der Fall war.
Doch nicht nur in der Gerichtsmedizin findet die Polymerase-Kettenreaktion Anwendung. Mit ihrer Hilfe können Krankheiten noch vor der Bildung von Antikörpern entdeckt werden. Geschichtswissenschaftler können Erbfolgen und Verwandtschaftsverhältnisse klären. Paläontologen und Paläozoologen erforschen mit ihrer Hilfe die urzeitliche Pflanzen- und Tierwelt. Besonders zu erwähnen sind natürlich auch die Biochemiker, die mit Hilfe der Polymerase-Kettenreaktion den Aufbau und die Funktion der DNA untersuchen.

Die nachfolgende Arbeit widmet sich dieser essentiellen Reaktion. Im Mittelpunkt steht dabei ihre mathematische Modellierung, mit deren Hilfe die Reaktion besser verstanden und die Genauigkeit der Analysemethoden verbessert werden kann. Kapitel 1 führt kurz in die biochemischen Grundlagen der Reaktion ein. Für den darüber hinaus interessierten Leser stellt Anhang B weitere Informationen zur Verfügung, die für das Verständnis des mathematischen Teils aber nicht vonnöten sind. Mittels Modellierung der Reaktion als Galton-Watson-Prozess kann die Verteilung der Anzahl an Mutationen eines Reaktionsproduktes bestimmt und die Mutationsrate, d.h. die Wahrscheinlichkeit mit der eine Base fehlerhaft übertragen wird, geschätzt werden (☞ Kapitel 2).

Der zweite Teil der Arbeit führt kurz in die Theorie der größenabhängigen Verzweigungsprozesse ein. Wir studieren das asymptotische Verhalten des Prozesses und schätzen die für den Prozess wichtigen Parameter (☞ Kapitel 3).
Mit Hilfe dieser Theorie wird dann das Modell der Polymerase-Kettenreaktion verfeinert, die Asymptotik des Prozesses betrachtet und die Effizienz, d.h. die Wahrscheinlichkeit, mit der eine Sequenz in einem Reaktionsschritt vervielfältigt wird, geschätzt (☞ Kapitel 4).
Somit erlaubt die mathematische Modellierung alle die Reaktion charakterisierenden Größen anhand von Messdaten zu schätzen. Eine Anwendung dieser Theorie bestünde darin, die Sicherheit von DNA-Tests zu untersuchen. Die obige Meldung betreffend, könnte also die Frage geklärt werden, wie wahrscheinlich die irrtümliche Verurteilung eines unschuldigen Mannes, bei Verwendung des genetischen Fingerabdrucks und der zur Analyse der DNA-Proben eingesetzten Polymerase-Kettenreaktion als Beweismittel, ist.

Zum Abschluss dieser Einleitung möchte ich mich bei allen Menschen bedanken, auf deren Hilfe und Unterstützung ich immer zählen konnte. Ganz besonderer Dank gilt meinen Eltern, die mich in all den zurückliegenden Jahren liebevoll und liebend begleitet haben. Ohne sie wäre diese Arbeit niemals zu Stande gekommen.
Herrn Prof. Dr. Alsmeyer danke ich für die Vergabe dieses interessanten Themas und für die teils durch einige Kilometer Land und Wasser erschwerte Betreuung während der Anfertigung der Diplomarbeit.

1. Die Polymerase-Kettenreaktion aus biochemischer Sicht

Die Polymerase-Kettenreaktion (engl. *polymerase chain reaction*, kurz: *PCR*) dient zur Vervielfältigung von *DNA*-Abschnitten (DNA von engl. *deoxyribonucleic acid*, im deutschen Sprachraum auch *DNS* von *Desoxyribonukleinsäure*).
Die DNA besteht aus drei verschiedenen Bausteinen: Zucker, Phosphatgruppen und vier verschiedenen Basen. Die Zuckermoleküle, genauer die Desoxyribosemoleküle, sind über Phosphatgruppen miteinander verknüpft und bilden das Grundgerüst der DNA. An jedem Zuckermolekül hängt eine der Basen Thymin, Cytosin, Adenin oder Guanin, die in der Biologie der Einfachheit halber mit ihren Anfangsbuchstaben abgekürzt werden. Alle in der DNA kodierten Informationen werden allein durch die Abfolge dieser Basen bestimmt. Je zwei Stränge sind im Inneren des DNA-Moleküls über die Basen miteinander verknüpft. Dabei liegen sich nur die Basen Guanin und Cytosin sowie Adenin und Thymin gegenüber, da die Verknüpfung über Wasserstoffbrücken unter diesen Basen zu Stande kommt. Das heißt, dass die Einzelstränge verschieden sind, aber über die Adenin/Thymin- bzw. Guanin/Cytosin-Äquivalenz miteinander identifiziert werden können. Man spricht hier auch von der *komplementären* Anordnung der Basen.
Bei der PCR wird der allen Zellen eigene Mechanismus zur Duplizierung von DNA mit Hilfe eines Enzyms, der Polymerase, ausgenutzt. Neben dem Begriff *Duplizierung* verwenden die Biologen häufig die Begriffe *Amplifikation* und *Replikation* für die komplementäre, exakte Vervielfältigung von DNA-Sequenzen. Zu Anfang der Reaktion müssen die doppelsträngig vorliegenden DNA-Moleküle getrennt werden (*Denaturierung*). Die gesuchte Sequenz (auch *Target* genannt) wird mit Startmolekülen (den sogenannten *Primern*) markiert, von denen dann in einer dritten Phase die Polymerase die ursprüngliche Sequenz beginnend bei den Primern komplementär vervielfältigt. Das Maximum der Reaktionsgeschwindigkeit der drei Phasen liegt bei verschiedenen Temperaturen. Die Denaturierung erfolgt bei einer Temperatur von ca. 90°C, die Primeranlagerung bei ca. 50°C und die für die PCR verwendete Polymerase arbeitet bei ca. 70°C optimal. Abbildung 1.1 zeigt die Phasen eines kompletten Zyklus.
Durch die verschiedenen Reaktionstemperaturen lässt sich ein PCR-Zyklus genau abgrenzen. Ist die Phase der Amplifikation abgeschlossen und wird das Gemisch wieder auf 90°C erhitzt, so beginnt ein neuer Zyklus. Jede neu gebildete Sequenz dient im nächsten Zyklus wieder als Matrize. Pro Zyklus können die vorhandenen Sequenzen also maximal verdoppelt werden. Die Reaktion findet in wässrigem Milieu statt und die Reaktanden sind im Überschuss hinzugefügt; das bedeutet, dass sich eine weit größere Anzahl an Nukleotiden und Polymerasen, als für eine komplette Umsetzung theoretisch benötigt wird, in der Reaktionslösung befindet. Doch auch bei diesen optimalen Reaktionsbedingungen dient nicht jedes DNA-Molekül als Replikationsmatrize. Es kann passieren, dass nicht alle DNA-Stränge im ersten Schritt denaturiert werden. Die Anlagerung der Primer kann fehlerhaft sein oder gar nicht erfolgen, so dass eine Vervielfältigung dieser Stränge nicht möglich ist. Im folgenden Schritt muss die Polymerase am Primer andocken, um dann mit der Amplifikation beginnen zu können. Diese beiden Schritte können durch nicht näher bestimmbare Faktoren gestört werden, so dass auch hier eine Vervielfältigung nicht erfolg-

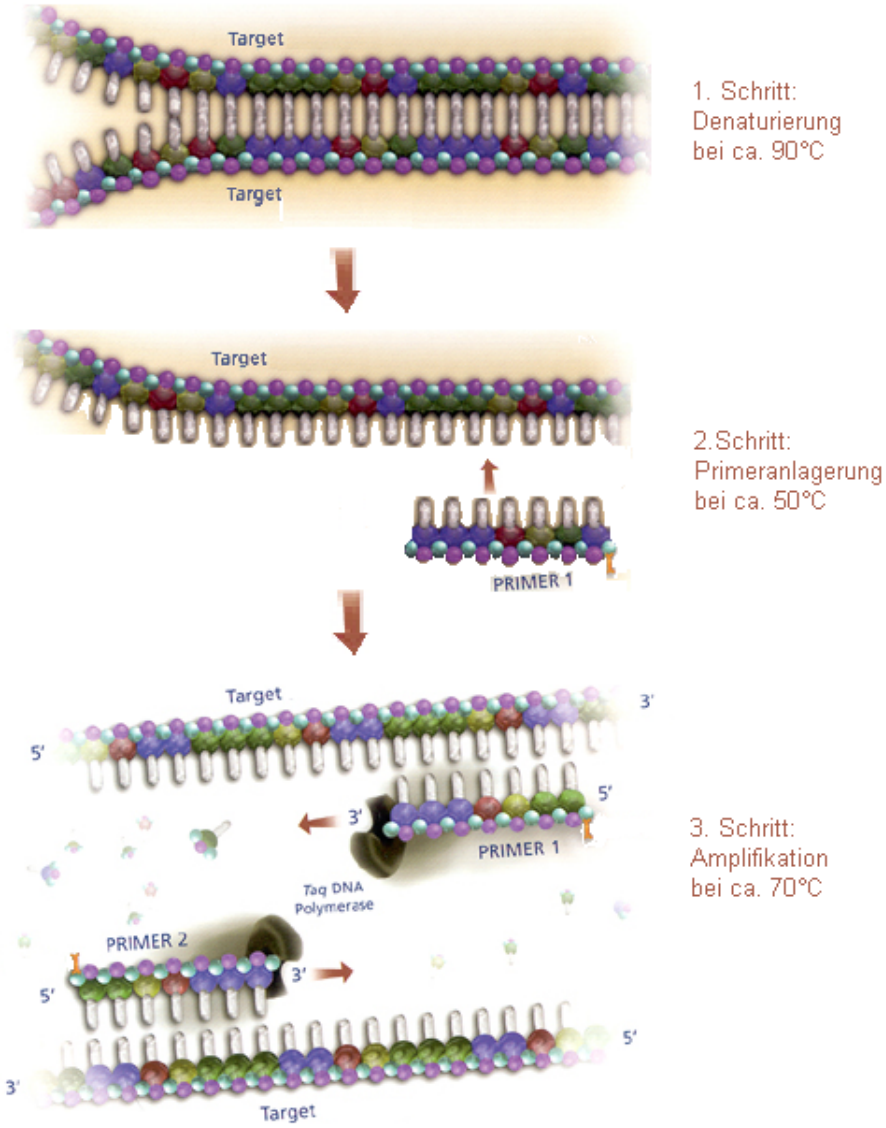

Abbildung 1.1.: Die Schritte eines PCR-Zyklus

reich ist. Die Reaktion ist stark vom Zufall abhängig. Die Wahrscheinlichkeit, mit der sich eine Sequenz vervielfältigt, wird *Effizienz* der Reaktion genannt und mit λ bezeichnet. Doch nicht nur bei der Vervielfältigung als solche können Probleme auftreten. Während des Duplizierungsvorganges kann es passieren, dass die Polymerase an einer Stelle eine falsche, keine oder eine zusätzliche Base einbaut. Dieser Vorgang wird *Mutation* genannt und die Wahrscheinlichkeit mit der eine Base mutiert heißt *Mutationsrate* μ.

In den nachfolgenden Kapiteln wird diese Reaktion in ein mathematisches Modell gebettet, das hilft, die PCR besser zu verstehen. Eine ausführlichere Beschreibung des Aufbaus der DNA, der Polymerase-Kettenreaktion und weitere biochemische Hintergründe zur PCR sind im Anhang Teil B zu finden.

2. Die Polymerase-Kettenreaktion als Galton-Watson-Prozess

Dieses Kapitel beschreibt ein erstes einfaches mathematisches Modell der Polymerase-Kettenreaktion. Es orientiert sich an einer Arbeit von Fengzhu Sun [27]. Die ausführliche Beschreibung des Modells in Kapitel 2.1 ist jedoch an die des in [4] vorgestellten Standardmodells der Galton-Watson-Prozess-Theorie angelehnt. Durch die Einbettung der PCR in einen Verzweigungsprozess lässt sich die Mutationsverteilung genauer studieren (☞ Kapitel 2.2), ein geeigneter Schätzer für die Mutationsrate finden (☞ Kapitel 2.3) und die Verteilung der Anzahl unterschiedlicher Basen zweier Sequenzen, des sogenannten Hamming-Abstandes, bestimmen (☞ Kapitel 2.4).

2.1. Ein erstes mathematisches Modell

1. Der Prozess
Zu Beginn der Polymerase-Kettenreaktion befindet sich eine gewisse Anzahl an DNA-Strängen, die das Target enthalten, in der Reaktionslösung. Die Denaturierung trennt die Stränge in zwei Einzelstränge auf, die sich nur durch die komplementäre Anordnung der Basen unterscheiden, aber auf Grund der Adenin/Thymin- bzw. Guanin/Cytosin-Äquivalenz miteinander identifiziert werden können. Im mathematischen Modell seien daher alle DNA-Einzelstränge identisch und die *Anzahl der Startsequenzen* sei mit Z_0 bezeichnet. Diese Sequenzen werden auch *Urahnen* genannt. Nach dem Ablauf der Phasen Denaturierung, Primeranlagerung und Replikation ist ein Reaktionszyklus vollendet. Die *Anzahl aller Sequenzen nach n PCR-Zyklen* sei Z_n.
Zur mathematischen Präzisierung sei ein messbarer Raum $(\Omega, \mathcal{A})$ und darauf Wahrscheinlichkeitsmaße P_j, für $j \in \mathbb{N}_0$ gegeben, unter denen $Z_0 = j$ f.s. gelte. Während des Reaktionszyklus kann sich jede Sequenz mit der Wahrscheinlichkeit $\lambda \in [0,1]$, *der Effizienz*, vervielfältigen und verbleibt anschließend in der Lösung. Die Anzahl der Sequenzen nach n Zyklen kann für alle $n \geq 1$ durch folgende rekursive Formel beschrieben werden:

$$Z_n = Z_{n-1} + \sum_{j=1}^{Z_{n-1}} I_{n,j} = \sum_{j=1}^{Z_{n-1}} (I_{n,j} + 1) \ , \tag{2.1}$$

wobei die $I_{n,j}$ für alle $j, n \in \mathbb{N}$ stochastisch unabhängige, identisch $\mathcal{B}(1,\lambda)$-verteilte Zufallsgrößen seien. Diese Zufallsgrößen geben an, ob sich das j-te Molekül im n-ten Schritt erfolgreich dupliziert oder nicht. In diesem Kapitel sei die Effizienz konstant, der Prozess $(Z_n)_{n\geq 0}$ entpuppt sich dann unter jedem P_j als Galton-Watson-Prozess (GWP), dessen Definition lautet (näheres zur Theorie der Verzweigungsprozesse siehe z.B. [4] und [5]):

Definition 2.1 Ein *Galton-Watson-Prozess* mit *Reproduktionsverteilung* $(p_k)_{k\geq 0}$ ist eine diskrete Markov-Kette $(Z_n)_{n\geq 0}$ mit Zustandsraum $\mathbb{N}_0$ und Übergangswahrscheinlichkeiten der Form:

$$p_{ij} = P\left(Z_{n+1} = j \mid Z_n = i\right) = \begin{cases} p_j^{*(i)} & \text{für} \quad i \geq 1,\ j \geq 0 \\ \delta_{0j} & \text{für} \quad i = 0,\ j \geq 0 \end{cases} \tag{2.2}$$

wobei $p_j^{*(i)}$ für $i \geq 1$ die i-fache Faltung von $(p_j)_{j\geq 0}$ und δ_{0j} das Kronecker-Delta bezeichne, d.h. $\delta_{0j} = 1$ für $j = 0$ und sonst $\delta_{0j} = 0$.

Satz 2.2 *Der durch Gleichung (2.1) definierte Prozess $(Z_n)_{n\geq 0}$ ist unter jedem P_j, $j \in \mathbb{N}_0$ ein Galton-Watson-Prozess.*

Beweis: Die Markov-Eigenschaft ergibt sich sofort aus der rekursiven Darstellung der Z_n in (2.1). Die Reproduktionsverteilung $(0, (1-\lambda), \lambda, 0, 0, \ldots)$ liefert die genaue Gestalt der Übergangswahrscheinlichkeiten

$$p_{ij} = P\left(Z_{n+1} = j \mid Z_n = i\right) = \begin{cases} \binom{i}{j-i}\lambda^{j-i}(1-\lambda)^{2i-j} & \text{für } i \leq j \leq 2i \\ 0 & \text{sonst} \end{cases} \tag{2.3}$$

Diese erfüllen (2.2). □

Für Galton-Watson-Prozesse gilt folgendes

Lemma 2.3 *Jeder GWP $(Z_n)_{n\geq 0}$ mit Reproduktionsverteilung $(p_k)_{k\geq 0}$ und j Urahnen ist die Summe von j unabhängigen GWP $(Z_n(i))_{n\geq 0}$, $1 \leq i \leq j$, mit derselben Reproduktionsverteilung und einem Urahnen. Ist $(Z_n)_{n\geq 0}$ in einem Standardmodell gegeben, folgt insbesondere*

$$P_j\left((Z_n)_{n\geq 0} \in \cdot\right) = P_1\left((Z_n)_{n\geq 0} \in \cdot\right)^{*(j)}$$

für alle $j \in \mathbb{N}_0$.

Beweis: Dieses Lemma ist [4] entnommen; dort ist auch der Beweis zu finden. □

Dieses Lemma besagt, dass die Reaktion mit j Startsequenzen als j unabhängige Reaktionen mit einer Startsequenz aufgefasst werden kann.

2. Mutationen und Markierungen

Es sei G die *Länge des Targets*, also der zu amplifizierenden Sequenz. Während des Replikationsvorgangs können fehlerhafte Basen eingebaut werden, eine Mutation tritt auf. Die Wahrscheinlichkeit, dass eine Base fehlerhaft durch die Polymerase übertragen wird, heiße Mutationsrate μ. Für jede Base entscheidet eine $\mathcal{B}(1,\mu)$-verteilte Zufallsgröße $M_{n,j}$, $1 \leq j \leq G, n \geq 1$, ob die j-te Base der Sequenz im n-ten PCR-Zyklus mutiert oder nicht. Diese Zufallsgrößen sind stochastisch unabhängig, die Anzahl der Mutationen nach einem PCR-Zyklus einer Sequenz ist somit $\mathcal{B}(G,\mu)$-verteilt. Da die Mutationsrate sehr klein und die Länge des Targets sehr groß ist, ersetzt man diese Verteilung durch eine Poisson-Verteilung mit dem Parameter μG. Der Poissonsche Grenzwertsatz (siehe [3] Satz 29.4) rechtfertigt dieses Vorgehen. Bei diesem Übergang wird streng genommen die Sequenz bestehend aus einer endlichen Anzahl an Basen durch eine Sequenz unendlicher Länge ersetzt. Mir positiver Wahrscheinlichkeit können nämlich unter der Annahme einer Poisson-verteilten Anzahl an Mutationen einer Sequenz mehr Mutationen auftreten, als

Basen in der Sequenz vorhanden sind. Da diese Wahrscheinlichkeit aber äußerst klein ist, gehen wir im Folgenden nicht weiter darauf ein.
Je häufiger eine Sequenz im Laufe der PCR repliziert wurde, desto wahrscheinlicher tritt eine hohe Anzahl an Mutationen auf. Es ist wichtig zu wissen, welche Sequenzen wie oft repliziert wurden, und wie viele Sequenzen es von einer Generation gibt. Die folgende Definition stellt wichtige Begriffe zur Beschreibung dieser Tatsachen bereit.

Definition 2.4 Z_0 bezeichne die Anzahl der Startmoleküle, die auch *Urahnen* oder *Sequenzen der 0-ten Generation* genannt werden. Die in einer PCR daraus hervorgehenden Sequenzen heißen *Sequenzen der ersten Generation* und diejenigen, die im $(k+1)$-ten Schritt aus der k-ten Generation hervorgehen, *Sequenzen der $(k+1)$-ten Generation.*
X_k^n sei die *Gesamtanzahl aller Sequenzen der k-ten Generation nach n PCR-Zyklen.*
$X_k^n(i)$ sei die *Anzahl der vom i-ten Startmolekül*, $1 \leq i \leq Z_0$, *erzeugten Sequenzen der k-ten Generation nach n PCR-Zyklen*, also $X_k^n = \sum_{i=1}^{Z_0} X_k^n(i)$.
Für alle $1 \leq i \leq Z_0$ sei $Z_n(i)$ die *Gesamtanzahl aller vom i-ten Startmolekül erzeugten Sequenzen nach n PCR-Zyklen.*

Anders als im Standardmodell der GWP-Theorie stirbt ein Individuum nach einem Schritt nicht aus, d.h. jede Sequenz verbleibt in der Lösung und kann zusätzlich als Matrize zur Herstellung einer Kopie dienen. Das Markierungssystem der GWP-Theorie kann zur Beschreibung der PCR nicht übernommen werden. Vielmehr wird ein System benötigt, in dem deutlich wird, welcher Generation eine Sequenz entstammt, und wie viele Sequenzen einer Generation existieren. Die folgende Definition basiert auf diesen Forderungen.

Definition 2.5 Jeder in einer PCR entstandenen Sequenz wird genau ein Vektor

$$\alpha := (i, j) \in \mathbb{N}_0 \times \mathbb{N}$$

zugeordnet. Dabei bedeutet $\alpha = (i, j)$, dass es sich bei der Sequenz α um die *j-te Sequenz der i-ten Generation* handelt. Die Menge aller nach dem n-ten PCR Zyklus vorhandenen Sequenzen sei mit $\mathcal{Z}_n$ bezeichnet.

Gibt es $Z_0 \in \mathbb{N}$ Startsequenzen, so tragen diese in Übereinstimmung mit Definition 2.5 die Markierungen $(0,1), (0,2), \ldots, (0, Z_0)$. Die Abbildung 2.1 zeigt ein Beispiel für eine 3-stufige PCR mit einem Startmolekül. Man erkennt, dass ein Molekül der ersten Generation nicht unbedingt im ersten PCR-Zyklus entstanden sein muss, ein Beispiel hierfür ist die im dritten Zyklus gebildete Sequenz $(1,3)$.

Welche Sequenzen sind in der Menge aller nach dem n-ten PCR-Zyklus vorhandenen Sequenzen enthalten? Zu Beginn der Reaktion liegen nur die Anfangssequenzen vor, d.h.

$$\mathcal{Z}_0 = \{(0,1), \ldots, (0, Z_0)\} .$$

Für $n \geq 1$ besteht $\mathcal{Z}_n$ aus zwei disjunkten Teilmengen, nämlich der Menge $\mathcal{Z}_{n-1}$ aller schon im $(n-1)$-ten Schritt vorhandenen Sequenzen und der Menge aller im n-ten Schritt entstandenen Sequenzen, die mit $\mathcal{Z}'_n$ bezeichnet sei.
Nach dem $(n-1)$-ten Zyklus gibt es X_1^{n-1} Sequenzen der ersten Generation. Also entstehen im n-ten Zyklus die Sequenzen

$$\left(1, X_1^{n-1} + 1\right), \left(1, X_1^{n-1} + 2\right), \ldots, \left(1, X_1^{n-1} + (X_1^n - X_1^{n-1})\right) .$$

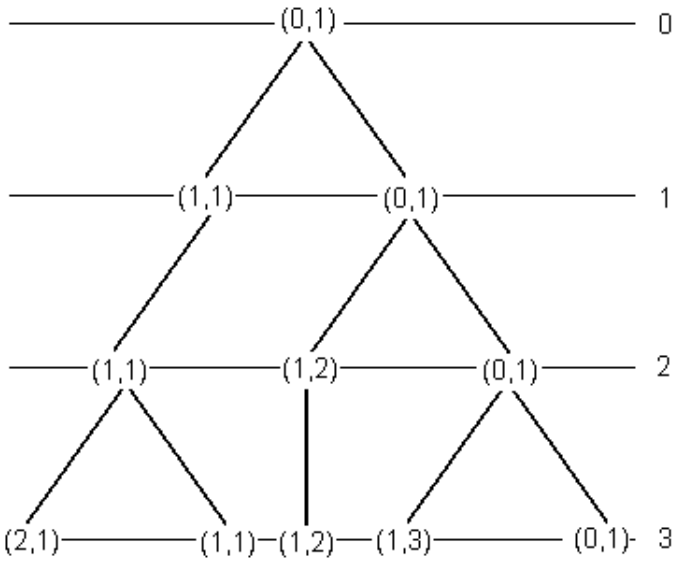

Abbildung 2.1.: Beispiel einer 3-stufigen PCR mit einem Urahn

Genauso werden im n-ten Zyklus folgende Sequenzen der zweiten Generation gebildet

$$\left(2, X_2^{n-1}+1\right), \left(2, X_2^{n-1}+2\right), \ldots, \left(2, X_2^{n-1}+(X_2^n - X_2^{n-1})\right) .$$

Da im n-ten Schritt nur Sequenzen entstehen können, deren Generationsnummer maximal n beträgt, hat $\mathcal{Z}_n'$ die Gestalt

$$\mathcal{Z}_n' = \bigcup_{k=1}^{n} \left\{ \left(k, X_k^{n-1}+1\right), \ldots, \left(k, X_k^{n-1}+(X_k^n - X_k^{n-1})\right) \right\} - \bigcup_{k=1}^{n} \{(k,0)\} .$$

Die Menge $\bigcup_{k=1}^{n}\{(k,0)\}$ muss abgezogen werden, da die Zufallsgrößen X_k^{n-1} und X_k^n beide gleichzeitig den Wert 0 annehmen können. Dies ist der Fall, wenn im $(n-1)$-ten Schritt noch keine Sequenz der k-ten Generation gebildet wurde und im n-ten Schritt auch keine solche entsteht. Dann wäre aber der Vektor $(k,0)$ in der Menge enthalten, der keine sinnvolle Markierung darstellt. Insgesamt erhalten wir für $n \geq 1$

$$\mathcal{Z}_n = \mathcal{Z}_{n-1} \cup \bigcup_{k=1}^{n} \left\{ \left(k, X_k^{n-1}+1\right), \ldots, \left(k, X_k^{n-1}+(X_k^n - X_k^{n-1})\right) \right\} - \bigcup_{k=1}^{n} \{(k,0)\} \quad (2.4)$$

Desweiteren lassen sich die Zufallsgrößen X_k^n für $k, n \geq 0$ über die Mengen $\mathcal{Z}_n$ definieren als:

$$X_k^n := \max_{(k,l)\in\mathcal{Z}_n} \{0, l\} \text{ für alle } k \geq 0. \quad (2.5)$$

In Abbildung 2.1 stammt die Sequenz $(2,1)$ von der Sequenz $(1,1)$ ab, diese ist also ein Vorfahr von $(2,1)$. Eine exakte Definition dieses Begriffes liefert:

Definition 2.6 Gegeben sei eine Sequenz $\alpha = (i_\alpha, j_\alpha)$, die aus einer n-stufige PCR mit Z_0 Urahnen hervorgeht.
Falls eine Kette von Sequenzen $\alpha = \alpha_t, \alpha_{t-1}, \ldots, \alpha_0 = (0, \iota)$ mit $\iota \in \{1, \ldots, Z_0\}$ und $t \in \mathbb{N}$ existiert, so dass für alle $1 \leq i \leq t$ die α_i durch Amplifikation aus α_{i-1} hervorgehen, heißen die Sequenzen α_j mit $1 \leq j \leq t-1$ *Vorfahren von* α.

Zur Beschreibung von Mutationen werden zusätzlich folgende Begriffe benötigt (zur genauen Bedeutung des Begriffes „zufällig ziehen“ siehe die Erläuterungen im Abschnitt 3):

Definition 2.7 Gegeben sei eine n-stufige PCR.
(i) $M(\alpha)$ sei die *Anzahl der Mutationen* der Sequenz $\alpha \in \mathcal{Z}_n$.
(ii) M sei die *Anzahl an Mutationen einer zufällig gezogenen Sequenz* nach n PCR-Zyklen.
(iii) Die *Generationsnummer* der Sequenz $\alpha = (i_\alpha, j_\alpha) \in \mathcal{Z}_n$ wird definiert als $g(\alpha) := i_\alpha$.
(iv) K sei die (zufällige) *Generationsnummer einer zufällig gezogenen Sequenz.*

Jetzt lässt sich auch die eingangs formulierte Äußerung präzisieren: Je mehr Vorfahren eine Sequenz $\alpha \in \mathcal{Z}_n$ besitzt, desto größer ist die Anzahl der Mutationen $M(\alpha)$. Genauer:

Lemma 2.8 *Es sei α eine in einer PCR erzeugte Sequenz. Dann gilt für die Verteilung der Anzahl der Mutationen*

$$P^{M(\alpha)} = Poi\left(g(\alpha)\mu G\right). \tag{2.6}$$

Beweis: Diese Aussage ergibt sich aus der Tatsache, dass die Anzahl der Mutationen nach einem PCR-Zyklus $Poi(\mu G)$-verteilt sind und der Faltungsformel für Poisson-Verteilungen $Poi(\theta) * Poi(\eta) = Poi(\theta + \eta)$ (siehe [3] Satz 29.3). □

Damit kann $M(\alpha)$ für $\alpha \in \mathcal{Z}_n$ als Summe stochastisch unabhängiger, identisch $Poi(\mu G)$-verteilter Zufallsgrößen $X_1, \ldots, X_{g(\alpha)}$ dargestellt werden.
Ganz ähnlich lässt sich auch die Anzahl der Mutationen einer zufällig gezogenen Sequenz als Summe

$$M = \sum_{i=1}^{K} X_i$$

von stochastisch unabhängigen, identisch $Poi(\mu G)$-verteilten Zufallsgrößen X_i schreiben.

3. Die zufällige Auswahl einer Sequenz und die n-stufige PCR

Am Ende einer n-stufigen PCR liegt ein Gemisch aus Z_n Sequenzen der verschiedensten Generationen vor. Die Extraktion einer Sequenz kann man sich als Urnenziehung mit Z_n Kugeln vorstellen. Dabei ist jede Kugeln mit einem anderen $\alpha \in \mathcal{Z}_n$ beschriftet. Bei der Extraktion handelt es sich also um ein Laplace-Experiment. Im folgenden werden zufällig gezogene Sequenzen mit großen griechischen Buchstaben bezeichnet; $A, B, \Gamma, \ldots$ sind demnach Laplace-verteilte Zufallsgrößen auf $\mathcal{Z}_n$. Die Anzahl an Mutationen M einer zufällig gezogenen Sequenz nach n PCR Zyklen ist in dieser Schreibweise also $M = M(A)$ und die Generationsnummer K einer zufällig gezogenen Sequenz $K = g(A)$.
Da es für alle $0 \leq k \leq n$ genau X_k^n Kugeln mit der Generationsnummer k gibt, gilt für die unter $X_0^n, X_1^n, \ldots, X_n^n$ bedingte Verteilung von K

$$P(K = k \mid X_0^n, X_1^n, \ldots, X_n^n) = \frac{X_k^n}{Z_n} \tag{2.7}$$

und damit

$$P(K = k) = EP(K = k \mid X_0^n, X_1^n, \ldots, X_n^n) = E\frac{X_k^n}{Z_n}. \tag{2.8}$$

Zur Beschreibung des Zugmechanismus ist die Anzahl der Sequenzen der k-ten Generation nach n PCR-Zyklen X_k^n ausreichend. Deshalb enthält das in der nachfolgenden Definition eingeführte mathematische Modell alle für die Beschreibung und Untersuchung einer Polymerase-Kettenreaktion erforderlichen Größen.

Definition 2.9 Gegeben sei ein messbarer Raum $(\Omega, \mathcal{A})$, auf dem eine Familie von Wahrscheinlichkeitsmaßen $(P_j)_{j\geq 0}$, eine Zufallsgröße Z_0, für die unter P_j f.s. $Z_0 = j$ gilt, und stochastisch unabhängige, identisch $\mathcal{B}(1, \lambda)$-verteilte Zufallsgrößen $I_{n,j},\ n, j \geq 1$ existieren. Für alle $k \geq 0,\ n \geq 1$ seien $Z_n, \mathcal{Z}_n$ und X_k^n durch die Gleichungen (2.1), (2.4) bzw. (2.5) definiert. Dann heißt das Modell

$$\Big((\Omega, \mathcal{A}), (P_j)_{j\geq 0}, (I_{n,j})_{j,n\geq 1}, (Z_n)_{n\geq 0}, (\mathcal{Z}_n)_{n\geq 0}, (X_k^n)_{k,n\geq 0}\Big)$$

Polymerase-Kettenreaktion, kurz *PCR*. Werden nur die ersten n Schritte betrachtet, heißt das Modell *n-stufige PCR* oder *PCR mit n Zyklen*. Ist Z_0 konstant vorgegeben, so wird das Modell *(n-stufige) PCR mit Z_0 Startsequenzen* genannt.

Wie Lemma 2.3 beweist, ist es in vielen Fällen ausreichend, den Prozess unter der Verteilung P_1 zu betrachten. Für den Erwartungswert, die Varianz und die Kovarianz unter der Verteilung P_j schreiben wir E_j, Var_j bzw. Cov_j, falls die Angabe der Anfangsverteilung notwendig erscheint. Ist die Spezifizierung nicht nötig, notieren wir diese Größen weiterhin mit E, Var sowie Cov, insbesondere dann, wenn der Prozess nur einen Urahn besitzt.

2.2. Die Mutationsverteilung

Wie schon in Lemma 2.8 festgestellt, ist die Anzahl an Mutationen einer Sequenz α einer n-stufigen PCR $Poi(g(\alpha)\mu G)$-verteilt. Anhand dieser Verteilung lässt sich aber nicht die Wahrscheinlichkeit dafür, dass eine zufällig gezogene Sequenz genau $m \in \mathbb{N}_0$ Mutationen aufweist, berechnen. Unter einer speziellen Voraussetzung kann diese explizit angegeben werden und ist das Hauptresultat dieses Kapitels. Doch vorerst müssen die erwartete Anzahl von Sequenzen der k-ten Generation nach n PCR-Zyklen X_k^n und die erwartete Gesamtzahl aller Sequenzen Z_n berechnet werden. In diesem Kapitel wird der Übersicht halber auf die Trennung zwischen den Anfangsverteilungen P_j verzichtet.

Lemma 2.10 *Gegeben sei eine n-stufige PCR mit der Effizienz $\lambda \in [0,1]$ und Z_0 Startsequenzen. Dann gilt*

$$EZ_n = Z_0(1+\lambda)^n. \tag{2.9}$$

Beweis: Da Z_n als Summe $\sum_{k=1}^{Z_0} Z_n(k)$ mit stochastisch unabhängigen, identisch verteilten $Z_n(k)$, $1 \le k \le Z_0$, darstellbar ist, gilt

$$EZ_n = E\sum_{k=1}^{Z_0} Z_n(k) = Z_0 EZ_n(1).$$

Aus $EZ_n(1) = (1+\lambda)^n$ (siehe [5] Kapitel 1, Abschnitt 2) folgt die Behauptung. □

Lemma 2.11 *Es seien $\lambda \in (0,1)$ die Effizienz einer PCR mit einer Startsequenz, X_k^n die Anzahl der Sequenzen der k-ten Generation nach n PCR-Zyklen. Dann gilt*

$$EX_k^n = \binom{n}{k}\lambda^k \text{ für alle } k \ge 0 \text{ und } n \ge 1. \tag{2.10}$$

Beweis: Diese Aussage wird durch Induktion bewiesen.
Zuerst sei $n = 1$. Dann gibt es drei Möglichkeiten:

1. Für $k = 0$ ist $EX_0^1 = 1 = \binom{1}{0}\lambda^0$.
2. Für $k = 1$ ist $EX_1^1 = \lambda = \binom{1}{1}\lambda^1$.
3. Für $k > 1$ ist $EX_k^1 = 0 = \binom{1}{k}\lambda^k$, wegen der Definition der Binomialkoeffizienten.

Die Behauptung gelte nun für ein festes n und beliebiges k. Auf Grund des PCR-Mechanismus lässt sich die Anzahl der Sequenzen der k-ten Generation nach $(n+1)$ PCR-Zyklen X_k^{n+1} schreiben als

$$X_k^{n+1} = X_k^n + \sum_{j=1}^{X_{k-1}^n} I_j,$$

mit von X_{k-1}^n und untereinander stochastisch unabhängigen, identisch $\mathcal{B}(1,\lambda)$-verteilten Zufallsgrößen I_j, $j \ge 1$, die angeben, ob sich das j-te Individum im n-ten PCR-Schritt vervielfältigt ($I_j = 1$) oder nicht ($I_j = 0$). Nach der ersten Waldschen Gleichung (siehe [2] Lemma 10.3 (a)) gilt

$$E\sum_{j=1}^{X_{k-1}^n} I_j = EI_1 EX_{k-1}^n = \lambda EX_{k-1}^n\,.$$

Dies impliziert:

$$\begin{aligned} EX_k^{n+1} &= EX_k^n + E\sum_{j=1}^{X_{k-1}^n} I_j \\ &\overset{\text{I.V.}}{=} \binom{n}{k}\lambda^k + \lambda\binom{n}{k-1}\lambda^{k-1} \\ &= \binom{n+1}{k}\lambda^k \end{aligned}$$

Insgesamt folgt die Behauptung. □

Gleichung (2.8) liefert eine Formel zur Berechnung der Verteilung der Generationsnummer K einer zufällig gezogenen Sequenz nach n PCR-Zyklen, nämlich $P(K=k) = E\frac{X_k^n}{Z_n}$. Der Erwartungswert lässt sich nicht ad hoc errechnen, er lässt sich aber für eine große Anzahl an Startsequenzen aus dem starken Gesetz der großen Zahlen ableiten:
Die Gesamtanzahl aller Sequenzen nach n PCR-Zyklen Z_n ist gleich der Summe aller Gesamtanzahlen der vom i-ten Startmolekül erzeugten Moleküle, d.h. $Z_n = \sum_{i=1}^{Z_0} Z_n(i)$, mit stochastisch unabhängigen, identisch verteilten $Z_n(i)$. Zusätzlich folgt aus Lemma 2.10, dass $EZ_n(1) = (1+\lambda)^n$. Vermöge des Satzes von Etemadi (siehe [3] Satz 35.4) gilt dann

$$\lim_{Z_0\to\infty} \frac{\sum_{i=1}^{Z_0} Z_n(i)}{Z_0} = EZ_n(1) = (1+\lambda)^n \quad \text{f.s.}$$

Auch die Anzahlen der Sequenzen der k-ten Generation nach n PCR-Zyklen, die vom i-ten Startmolekül abstammen, $(X_k^n(i))_{i\geq 1}$ sind stochastisch unabhängig, identisch verteilt mit $EX_k^n(1) = \binom{n}{k}\lambda^k$ (siehe Lemma 2.11). Dies impliziert

$$\lim_{Z_0\to\infty} \frac{\sum_{i=1}^{Z_0} X_k^n(i)}{Z_0} = EX_k^n(1) = \binom{n}{k}\lambda^k \quad \text{f.s.}$$

Insgesamt folgt

$$\lim_{Z_0\to\infty} \frac{X_k^n}{Z_n} = \lim_{Z_0\to\infty} \frac{\sum_{i=1}^{Z_0} X_k^n(i)}{\sum_{i=1}^{Z_0} Z_n(i)} = \lim_{Z_0\to\infty} \frac{Z_0}{\sum_{i=1}^{Z_0} Z_n(i)} \cdot \frac{\sum_{i=1}^{Z_0} X_k^n(i)}{Z_0} = \frac{\binom{n}{k}\lambda^k}{(1+\lambda)^n} \quad \text{f.s.} \quad (2.11)$$

Dies bedeutet mit dem Verweis auf (2.8), dass bei einer hinreichend großen Anzahl von Startmolekülen Z_0 die Verteilung der Generationsnummer K einer zufällig gezogenen Sequenz nach n PCR-Zyklen durch eine Binomialverteilung mit Parametern n und $\frac{\lambda}{1+\lambda}$ approximiert werden kann. Dies rechtfertigt die folgende

Voraussetzung (V.1) Die Verteilung der Generationsnummer einer zufällig gezogenen Sequenz nach n PCR-Zyklen sei binomialverteilt, genauer

$$K \sim \mathcal{B}\left(n, \frac{\lambda}{1+\lambda}\right).$$

Bemerkung 2.12 Nach F. Sun [27] haben Simulationen ergeben, dass diese Approximation auch dann für jede beliebige Anzahl von Startmolekülen Z_0 gut ist, wenn $\lambda > 0,85$ gilt.
Im PCR-Alltag ist die Anzahl der Startsequenzen im Allgemeinen sehr hoch. Selbst bei kriminalistischen Methoden, wie der eingangs vorgestellten DNA-Analyse in der Gerichtsmedizin, stehen ausreichend viele Sequenzen zur Verfügung, um Voraussetzung V.1 anwenden

zu können. Zum Beispiel können aus einem Haar von 1cm Länge mehrere tausend DNA-Sequenzen extrahiert werden. Für den Biochemiker reicht diese Menge für eine Analyse nicht aus, der Mathematiker sieht jedoch eine Rechtfertigung für Voraussetzung V.1.

Der nachfolgende Satz gibt nun unter V.1 die Mutationsverteilung, den Erwartungswert sowie die Varianz für die Anzahl an Mutationen einer zufällig gezogenen Sequenz nach n PCR-Zyklen an und beschreibt, wie die Mutationsverteilung durch eine Standardnormalverteilung bzw. eine Poisson-Verteilung angenähert werden kann. Dieses Ergebnis ist wegen V.1 unabhängig von der Anzahl der Startsequenzen.

Satz 2.13 *Unter V.1 gilt für die Anzahl an Mutationen M einer zufällig gezogenen Sequenz nach n PCR-Schritten:*
(i)

$$P(M=m)=\frac{(\mu G)^m\left(1+\lambda e^{-\mu G}\right)^n}{m!(1+\lambda)^n}ES^m \text{ für alle } m\in\mathbb{N}_0$$

wobei S eine $\mathcal{B}\left(n,\frac{\lambda e^{-\mu G}}{1+\lambda e^{-\mu G}}\right)$-verteilte Zufallsgröße sei.
(ii) Die erzeugende Funktion von M ist

$$f_M(s)=\left(\frac{1+\lambda\exp\left(\mu G(s-1)\right)}{1+\lambda}\right)^n.$$

Der Erwartungswert sowie die Varianz von M existieren und sind

$$EM=\mu G\frac{n\lambda}{1+\lambda}\qquad VarM=\frac{n\lambda\mu G}{(1+\lambda)^2}(\mu G+1+\lambda).$$

(iii) Für alle $x\in\mathbb{R}$ gilt:

$$\lim_{n\to\infty}P\left(\frac{(1+\lambda)M-n\lambda\mu G}{\sqrt{n\lambda\mu G(\mu G+1+\lambda)}}\le x\right)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{s^2}{2}}ds$$

(iv) Die Mutationsrate und die Länge des Targets seien von n abhängig, bezeichnet mit μ_n bzw. G_n, und zwar derart, dass $\lim_{n\to\infty}n\mu_nG_n=\nu$ für ein $\nu\in\mathbb{R}$. Dann gilt

$$\lim_{n\to\infty}P(M=m)=Poi\left(\frac{\lambda\nu}{1+\lambda}\right)(\{m\})\quad \text{für alle } m\in\mathbb{N}_0.$$

Beweis:
(i) Da die Anzahl der Mutationen M von Sequenzen der k-ten Generation $Poi(k\mu G)$-

verteilt und K nach Voraussetzung $\mathcal{B}\left(n, \frac{\lambda}{1+\lambda}\right)$-verteilt ist, gilt für jedes $m \in \mathbb{N}_0$

$$\begin{aligned}
P(M=m) &= \sum_{k=0}^{n} P(M=m, K=k) \\
&= \sum_{k=0}^{n} P(M=m \mid K=k) P(K=k) \\
&= \sum_{k=0}^{n} e^{-k\mu G} \frac{(k\mu G)^m}{m!} \binom{n}{k} \frac{\lambda^k}{(1+\lambda)^n} \\
&= \frac{(\mu G)^m}{m!(1+\lambda)^n} \sum_{k=0}^{n} \binom{n}{k} \left(\lambda e^{-\mu G}\right)^k k^m \frac{\left(1+\lambda e^{-\mu G}\right)^n}{\left(1+\lambda e^{-\mu G}\right)^n} \\
&= \frac{(\mu G)^m}{m!(1+\lambda)^n} \left(1+\lambda e^{-\mu G}\right)^n \sum_{k=0}^{n} k^m \binom{n}{k} \left(\frac{\lambda e^{-\mu G}}{1+\lambda e^{-\mu G}}\right)^k \underbrace{\left(\frac{1}{1+\lambda e^{-\mu G}}\right)}_{=1-\frac{\lambda e^{-\mu G}}{1+\lambda e^{-\mu G}}}^{n-k} \\
&= \frac{(\mu G)^m}{m!(1+\lambda)^n} \left(1+\lambda e^{-\mu G}\right)^n \sum_{k=0}^{n} k^m P(S=k) \\
&= \frac{(\mu G)^m}{m!(1+\lambda)^n} \left(1+\lambda e^{-\mu G}\right)^n ES^m
\end{aligned}$$

wobei $S \sim \mathcal{B}\left(n, \frac{\lambda e^{-\mu G}}{1+\lambda e^{-\mu G}}\right)$ sei. Dies ist die Behauptung.

(ii) Wie in Kapitel 2.1 bemerkt, lässt sich M als Summe $\sum_{i=1}^{K} X_i$ mit untereinander und von K stochastisch unabhängigen, identisch $Poi(\mu G)$-verteilten Zufallsgrößen X_i für alle $i = 1, 2, \ldots$ darstellen. Seien f_M, f_K und f_{X_1} die zu M, K bzw. X_1 gehörenden erzeugenden Funktionen, dann gilt

$$\begin{aligned}
f_M(s) &= Es^M = Es^{\sum_{i=1}^{K} X_i} = E \sum_{\kappa=0}^{n} \left(\mathbb{1}_{\{K=\kappa\}}\right) s^{\sum_{i=1}^{\kappa} X_i} \\
&= \sum_{\kappa=0}^{n} E\mathbb{1}_{\{K=\kappa\}} \underbrace{Es^{\sum_{i=1}^{\kappa} X_i}}_{=Es^{X_1} Es^{X_2} \cdots Es^{X_\kappa} = \left(Es^{X_1}\right)^\kappa \text{ da } X_i \text{ iid}} \\
&= f_K\left(f_{X_1}(s)\right)
\end{aligned}$$

Da $K \sim \mathcal{B}\left(n, \frac{\lambda}{1+\lambda}\right)$ und $X_1 \sim Poi(\mu G)$, folgen (nach Satz 26.3 und Satz 29.2 in [3])

$$f_K(s) = \left(\frac{1+\lambda s}{1+\lambda}\right)^n \quad \text{und} \quad f_{X_1}(s) = \exp\left(\mu G(s-1)\right).$$

Daraus ergibt sich

$$f_M(s) = \left(\frac{1+\lambda \exp\left(\mu G(s-1)\right)}{1+\lambda}\right)^n.$$

Nach dem Eindeutigkeitssatz und dem Multiplikationssatz für erzeugende Funktionen (z.B. zu finden in [12], Chapter 4, Theorem 7.1 und Theorem 7.2) ist M genauso verteilt wie die Summe stochastisch unabhängiger, identisch verteilter Zufallsgrößen $Y_1, Y_2, \ldots, Y_n$. Diese Y_i, $i = 1, \ldots, n$, besitzen die erzeugende Funktion $f_{Y_i}(s) = \frac{1+\lambda \exp(\mu G(s-1))}{1+\lambda}$ und damit die gemischte Verteilung $P^{Y_i} = \frac{1}{1+\lambda}\delta_0 + \frac{\lambda}{1+\lambda} Poi(\mu G)$. Da

$$EY_1 = f'_{Y_1}(1) = \frac{\lambda \mu G}{1+\lambda}$$

und

$$VarY_1 = f''_{Y_1}(1) + f'_{Y_1}(1) - \left(f'_{Y_1}(1)\right)^2 = \frac{\lambda\mu G(\mu G + 1 + \lambda)}{(1+\lambda)^2}\,,$$

ergeben sich mit der Unabhängigkeit, also insbesondere der Unkorreliertheit, sowie der identischen Verteilung der $Y_1, Y_2, \ldots, Y_n$

$$EM = \mu G \frac{n\lambda}{1+\lambda}$$

und

$$VarM = \frac{n\lambda\mu G}{(1+\lambda)^2}(\mu G + 1 + \lambda)\,.$$

Damit ist (ii) bewiesen .

(iii) M lässt sich, wie im Beweis zu (ii) beschrieben, als Summe stochastisch unabhängiger, identisch verteilter Zufallsgrößen $Y_1, Y_2, \ldots, Y_n$ schreiben, deren Erwartungswert und Varianz, wie oben errechnet, existieren. Dies impliziert nach dem Satz von Lindeberg (siehe [3] Satz 37.5)

$$\frac{M - EM}{\sqrt{VarM}} = \frac{\sum_{i=0}^{n}(Y_i - EY_i)}{\sqrt{Var\left(\sum_{i=0}^{n}(Y_i - EY_i)\right)}} \xrightarrow{d} \mathcal{N}(0,1)\,, \tag{2.12}$$

da eine unabhängige Folge identisch verteilter Zufallsgrößen mit Erwartungswert 0 und positiver, endlicher Varianz die Lindebergbedingung erfüllt und somit dem zentralen Grenzwertsatz genügt. Die Aussage (2.12) entspricht der Behauptung.
(iv) Auf Grund des Stetigkeitssatzes für erzeugende Funktionen (siehe [3] Kapitel 45) ist zum Beweis dieser Aussage der Grenzwert der erzeugenden Funktion $f_M(s)$ zu bestimmen.

$$\begin{aligned} f_M(s) &= \left(\frac{1 + \lambda\exp\left(\mu_n G_n(s-1)\right)}{1+\lambda}\right)^n \\ &= \left(1 + \frac{\frac{\lambda}{1+\lambda} n(\exp\left(\mu_n G_n(s-1)\right) - 1)}{n}\right)^n \\ &\xrightarrow[n\to\infty]{} \exp\left(\frac{\lambda}{1+\lambda}\nu(s-1)\right) \end{aligned} \tag{2.13}$$

Der Grenzübergang ist richtig, denn wegen der Voraussetzung $\lim_{n\to\infty} n\mu_n G_n = \nu$ folgt:

$$\begin{aligned} n\left(\exp\left(\mu_n G_n(s-1)\right) - 1\right) &= n\left(\sum_{k=0}^{\infty}\frac{\left(\mu_n G_n(s-1)\right)^k}{k!} - 1\right) \\ &= n\mu_n G_n(s-1) + n\mu_n G_n(s-1)\underbrace{\sum_{k=2}^{\infty}\frac{\left(\mu_n G_n(s-1)\right)^{k-1}}{k!}}_{\xrightarrow[n\to\infty]{} 0} \\ &\xrightarrow[n\to\infty]{} \nu(s-1) \end{aligned}$$

Die Grenzfunktion in Gleichung (2.13) ist die erzeugende Funktion einer $Poi\left(\frac{\lambda}{1+\lambda}\nu\right)$-Verteilung und daraus folgt die Behauptung. □

Bemerkung 2.14
(a) Die erwartete Anzahl an Mutationen M einer zufällig gezogenen Sequenz entspricht der intuitiven Vermutung. EM ist nämlich das Produkt des Erwartungswertes einer $Poi(\mu G)$-verteilten Zufallsgröße und des Erwartungswertes einer $\mathcal{B}\left(n, \frac{\lambda}{1+\lambda}\right)$-verteilten Zufallsgröße.
(b) Obgleich die Voraussetzungen an μ und G in Satz 2.13 (iv) chemisch nicht sinnvoll sind, kann diese Aussage zur approximativen Berechnung der Wahrscheinlichkeit von m Mutationen in einer zufällig gezogenen Sequenz herangezogen werden. Wie bei der Approximation einer Binomialverteilung durch eine Normalverteilung bzw. durch eine Poisson-Verteilung, liefert die Verwendung der Normalapproximation in Satz 2.13 (iii) bessere Resultate, wenn $n\mu G$ groß ist und die Verwendung der Poisson-Approximation in (iv), wenn $n\mu G$ klein ist. Sun zeigt in einem Vergleich der exakten und der angenäherten Wahrscheinlichkeiten für das Vorkommen von 0,1,2 sowie mehr als 2 Mutationen in einer zufällig ausgewählten Sequenz nach 20 bzw. 50 Zyklen, dass in diesen Fällen die Poisson-Approximation deutlich bessere Werte liefert als die Normalapproximation (siehe [27]).
(c) Unter Zuhilfenahme der Verteilungen P_j können die Aussagen in Satz 2.13 ohne Voraussetzung V.1 als Grenzwertsätze formuliert werden. Dann lautet z.B. Aussage (i):

$$\lim_{j\to\infty} P_j(M = m) = \frac{(\mu G)^m \left(1 + \lambda e^{-\mu G}\right)^n}{m!(1+\lambda)^n} ES^m \text{ für alle } m \in \mathbb{N}_0$$

Dies gilt, da die Generationsnummer K einer zufällig gezogenen Sequenz asymptotisch $\mathcal{B}\left(n, \frac{\lambda}{1+\lambda}\right)$-verteilt ist.

2.3. Ein Schätzer der Mutationsrate

Wird eine n-stufige Polymerase-Kettenreaktion durchgeführt, ist vor dieser Reaktion im Allgemeinen nicht bekannt, wie hoch die Mutationsrate ist. Es ist aber von außerordentlichem Interesse, diese annähernd genau zu bestimmen. In der Aufklärung der in der Einleitung beschriebenen Straftat sollte kein Unschuldiger verurteilt werden, nur weil zufällig während der PCR Mutationen aufgetreten sind. Das Schätzen der Mutationsrate wird jedoch dadurch erschwert, dass bei einer zufällig ausgewählten Sequenz nicht bekannt ist, welcher Generation sie entstammt, d.h. wie oft die Ursprungssequenz amplifiziert wurde, um zu der vorliegenden Sequenz zu gelangen. Mit Blick auf Lemma 2.8 ist die Wahrscheinlichkeit einer hohen Anzahl an Mutationen in einer Sequenz umso größer, je mehr Vorfahren sie hat. Das nachfolgende Kapitel entwickelt einen geeigneten Schätzer für die Mutationsrate μ und analysiert dessen Varianz, um ein Gütekriterium für den Schätzer zu erhalten.
In diesem Kapitel wird streng zwischen Erwartungswert, Varianz und Kovarianz unter den verschiedenen Verteilungen P_j unterschieden und diese werden mit E_j, Var_j bzw. Cov_j notiert.

Nach n PCR-Zyklen werden zufällig s Sequenzen $A_1, \ldots, A_s \in \mathcal{Z}_n$ mit Zurücklegen gezogen. Es handelt sich also um $\mathcal{Z}_n$-wertige, Laplace-verteilte Zufallsvektoren. Dann sind die Zufallsgrößen $M(A_1), \ldots, M(A_s)$, die die Anzahl an Mutationen der Sequenzen $A_1, \ldots, A_s$ angeben, identisch verteilt. Die erwartete Anzahl aller Mutationen ist unter Voraussetzung V.1 mit Satz 2.13 (ii)

$$E\left(\sum_{i=1}^{s} M(A_i)\right) = sEM(A_1) = s\mu G\frac{n\lambda}{1+\lambda}.$$

Wie man leicht nachrechnet, ist dann

$$\hat{\mu} : \mathbb{N}_0^s \to [0,\infty), \quad \hat{\mu}\left((m_1,\ldots,m_s)\right) := \frac{(1+\lambda)\sum_{i=1}^{s} m_i}{n\lambda G s} \tag{2.14}$$

ein erwartungstreuer Schätzer für μ. Es handelt sich um den Momentenmethode-Schätzer (siehe dazu auch [1] Definition 6.1).

Für die Berechnung der Varianz von μ betrachtet man zuerst die Varianz der Summe der $M(A_i)$ und erhält auf Grund ihrer identischen Verteilung für jede Anfangsverteilung P_j:

$$\begin{aligned} Var_j\left(\sum_{i=1}^{s} M(A_i)\right) &= \sum_{i=1}^{s} Var_j M(A_i) + 2\sum_{1\le i<k\le s} Cov_j\left(M(A_i), M(A_k)\right) \\ &= sVar_j M(A_1) + 2\binom{s}{2} Cov_j\left(M(A_1), M(A_2)\right) \end{aligned} \tag{2.15}$$

Die in (2.15) vorkommenden Terme $Var_j M(A_1)$ und $Cov_j\left(M(A_1), M(A_2)\right)$ werden nun getrennt voneinander untersucht. Den Anfang macht die Kovarianz von $M(A_1)$ und $M(A_2)$. Doch zuvor muss noch ein wichtiges Hilfsmittel bereitgestellt werden. Definition 2.6 klärt den Begriff der Vorfahren einer Sequenz. Im Laufe der PCR dient jede Sequenz als Matrize für weitere Kopien, so dass sich am Ende der Reaktion Sequenzen in der Reaktionslösung befinden, die von einer Sequenz höherer Generationsnummer abstammen. Dies führt zu

Definition 2.15 Die Sequenzen α und β entstammen einer n-stufigen PCR mit Z_0 Startmolekülen, d.h. $\alpha, \beta \in \mathcal{Z}_n$. Ist $\gamma \in \mathcal{Z}_n$ eine weitere Sequenz, die sowohl Vorfahr von α als auch von β ist, heißt γ *gemeinsamer Vorfahr von α und β*.
Gibt es keinen weiteren gemeinsamen Vorfahren vor γ, so heißt dieser *letzter gemeinsamer Vorfahr von α und β*, kurz *MRCA* vom englischen *„most recent common ancestor“*. Man schreibt dann

$$\mathrm{MRCA}(\alpha, \beta) = \gamma.$$

Stammen α und β von verschiedenen Urahnen ab, so wird ihr MRCA definiert als

$$\mathrm{MRCA}(\alpha, \beta) := (0, 0).$$

Nun kann die Varianz zweier zufällig gezogener Sequenzen berechnet werden:

Lemma 2.16 *Gegeben seien eine n-stufige PCR, zwei $\mathcal{Z}_n$-wertige, Laplace-verteilte Zufallsvektoren A, B und $\Gamma := \mathrm{MRCA}(A, B)$. Dann gilt für jede Anfangsverteilung P_j:*

$$Cov_j(M(A), M(B)) = \mu G E_j g(\Gamma) + (\mu G)^2 Cov_j(g(A), g(B))$$

Beweis: Betrachtet wird nur der Fall eines Urahns, also $j = 1$ (siehe Lemma 2.3). Nach A.1 (ii) gilt:

$$\begin{aligned} Cov(M(A), M(B)) &= E(Cov(M(A), M(B)|A, B, \Gamma)) \\ &\quad + Cov(E[M(A)|A, B, \Gamma], E[M(B)|A, B, \Gamma]) \end{aligned}$$

Wir berechnen nun die Terme auf der rechten Seite. Dazu seien $M(\Gamma A)$, $M(\Gamma B)$ die Anzahl von Mutationen, die sich zwischen Γ und A bzw. B ereignet haben. Nach den zur PCR gemachten Voraussetzungen sind $M(\Gamma)$, $M(\Gamma A)$, $M(\Gamma B)$ bedingt unter A, B, Γ stochastisch unabhängig und es gilt für $X \in \{A, B, \Gamma\}$

$$E\left(M(X) \mid A, B, \Gamma\right) = g(X)\mu G. \tag{2.16}$$

Daraus folgt

$$\begin{aligned} Cov\big(E[M(A)|A, B, \Gamma], E[M(B)|A, B, \Gamma]\big) &= Cov(g(A)\mu G, g(B)\mu G) \\ &= (\mu G)^2 Cov(g(A), g(B))\,. \end{aligned} \tag{2.17}$$

Wegen der bedingten stochastischen Unabhängigkeit von $M(\Gamma A)$, $M(\Gamma B)$ und $M(\Gamma)$ sowie Gleichung (2.16) gilt weiter:

$$\begin{aligned} ECov(M(A), M(B)|A, B, \Gamma) &= ECov[M(\Gamma) + M(\Gamma A), M(\Gamma) + M(\Gamma B)|A, B, \Gamma] \\ &= ECov[M(\Gamma), M(\Gamma)|A, B, \Gamma] \\ &= EVar[M(\Gamma)|A, B, \Gamma] \\ &= VarM(\Gamma) - Var(E[M(\Gamma]|A, B, \Gamma]) \\ &= VarM(\Gamma) - (\mu G)^2 Var\, g(\Gamma) \end{aligned} \tag{2.18}$$

Die vorletzte Gleichheit gilt wegen der bedingten Varianz-Formel in A.1 (i). Um den letzten Term weiter umformen zu können, muss die Varianz von $M(\Gamma)$ untersucht werden.

Wegen Lemma 2.8 und der sich daran anschließenden Bemerkung ist $M(\Gamma) \sim \sum_{i=1}^{g(\Gamma)} X_i$ mit stochastisch unabhängigen, identisch $Poi(\mu G)$-verteilten Zufallsgrößen X_i. Daraus folgt:

$$\begin{aligned}
VarM(\Gamma) &= Var\left(\sum_{i=1}^{g(\Gamma)} X_i\right) \\
&= E\left(\sum_{i=1}^{g(\Gamma)} X_i\right)^2 - \left(E\sum_{i=1}^{g(\Gamma)} X_i\right)^2 \\
&= E\left(\sum_{i=1}^{g(\Gamma)} X_i\right)^2 - \big(\underbrace{Eg(\Gamma)EX_1}_{=:\zeta}\big)^2 \\
&= E\sum_{\kappa=1}^{n} \mathbb{1}_{\{g(\Gamma)=\kappa\}}\left(\sum_{i=1}^{\kappa} X_i\right)^2 - \zeta^2 \\
&= \sum_{\kappa=1}^{n} E\left[\mathbb{1}_{\{g(\Gamma)=\kappa\}}\left(\sum_{i=1}^{\kappa} X_i\right)^2\right] - \zeta^2 \\
&= \sum_{\kappa=1}^{n} E\mathbb{1}_{\{g(\Gamma)=\kappa\}} E\left(\sum_{i=1}^{\kappa} X_i\right)^2 - \zeta^2 \\
&= \sum_{\kappa=1}^{n} E\mathbb{1}_{\{g(\Gamma)=\kappa\}} E\left(\sum_{i=1}^{\kappa} X_i^2 + 2\sum_{1\le i<j\le\kappa} X_iX_j\right) - \zeta^2 \\
&= \sum_{\kappa=1}^{n} E\mathbb{1}_{\{g(\Gamma)=\kappa\}}\left(\kappa EX_1^2 + 2\binom{\kappa}{2} EX_1X_2\right) - \zeta^2 \\
&= EX_1^2\sum_{\kappa=1}^{n}\kappa P(g(\Gamma)=\kappa) + (EX_1)^2\sum_{\kappa=1}^{n}\kappa(\kappa-1)P(g(\Gamma)=\kappa) - \zeta^2 \\
&= EX_1^2 Eg(\Gamma) + (EX_1)^2 Eg(\Gamma)^2 - (EX_1)^2 Eg(\Gamma) - (Eg(\Gamma)EX_1)^2 \\
&= Eg(\Gamma)\left(EX_1^2 - (EX_1)^2\right) + (EX_1)^2\left(Eg(\Gamma)^2 - (Eg(\Gamma))^2\right) \\
&= Eg(\Gamma)VarX_1 + (EX_1)^2 Var\, g(\Gamma) \\
&= \mu G Eg(\Gamma) + (\mu G)^2 Var\, g(\Gamma) \qquad (2.19)
\end{aligned}$$

Aus den Gleichungen (2.17), (2.18) und (2.19) ergibt sich mit der Formel für die bedingte Kovarianz:

$$\begin{aligned}
Cov(M(A),M(B)) &= ECov[M(A),M(B)|A,B,\Gamma] \\
&\quad + Cov\big(E[M(A)|A,B,\Gamma], E[M(B)|A,B,\Gamma]\big) \\
&= \mu G Eg(\Gamma) + (\mu G)^2 Cov(g(A),g(B))
\end{aligned}$$

Dies ist die Behauptung. □

Der Weg zur Kovarianz von $M(A)$ und $M(B)$ führt also über den Erwartungswert von $g(\Gamma)$ und die Kovarianz von $g(A)$ und $g(B)$. Diese werden in den nachfolgenden Lemmata bestimmt. Doch vorerst muss geklärt werden, wie groß die erwartete Anzahl von Paaren mit einem MRCA der Generation k nach n PCR-Zyklen ist.

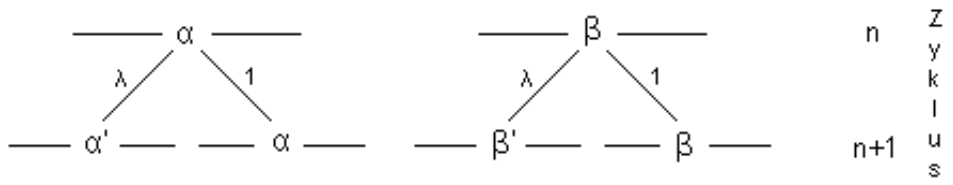

Abbildung 2.2.: Mögliche Nachkommen der Sequenzen α und β

Lemma 2.17 *Gegeben sei eine n-stufige PCR, $n \geq 1$, mit einem Urahn und für $k \in \mathbb{N}_0$ sei*

$$C_n(k) := E|\{\{\alpha, \beta\} : \ \alpha, \beta \in \mathcal{Z}_n,\ \alpha \neq \beta,\ g(\mathrm{MRCA}(\alpha, \beta)) = k\}| \,.$$

$C_n(k)$ ist also die erwartete Anzahl von Paaren mit einem MRCA der Generation k nach n PCR-Zyklen, wobei die Sequenzen des Paares unterschiedlich sind und die Zugreihenfolge nicht berücksichtigt wird. Dann gilt:
(i)

$$C_{n+1}(k) = (1+\lambda)^2 C_n(k) + \binom{n}{k} \lambda^{k+1}$$

wobei $C_n(k) = 0$ für alle $k \geq n$.
(ii) Die erzeugende Funktion von $(C_n(k))_{k\geq 0}$, definiert durch $\varphi_{C_n}(s) := \sum_{k=0}^{n-1} C_n(k) s^k$ ist

$$\varphi_{C_n}(s) = \lambda \frac{(1+\lambda s)^n - (1+\lambda)^{2n}}{(1+\lambda s) - (1+\lambda)^2} \,.$$

Beweis:
(i) Für ein Sequenzpaar mit einem MRCA der k-ten Generation nach $(n+1)$ PCR-Zyklen gibt es zwei Möglichkeiten, wie dieses Paar im $(n+1)$-ten Schritt entstanden ist:
1. Fall: Bei dem Paar handelt es sich um eine Sequenz der k-ten Generation und ihre im (n+1)-ten Schritt entstandene Kopie. Es gibt X_k^n Sequenzen aus denen ein solches Paar entstehen kann. Die (zufällige) Anzahl solcher Paare ist also eine Summe $\sum_{j=1}^{X_k^n} I_j$ mit (auch von X_k^n) stochastisch unabhängigen, identisch $\mathcal{B}(1,\lambda)$-verteilten I_j, $j \geq 1$. Die erwartete Anzahl dieser Paare ist dann

$$E \sum_{j=1}^{X_k^n} I_j = E X_k^n E I_1 \overset{(2.1)}{=} \binom{n}{k} \lambda^k \, \lambda = \binom{n}{k} \lambda^{k+1}$$

2. Fall: Die beiden Sequenzen haben im n-ten PCR-Zyklus verschiedene Vorfahren α und β. Abbildung 2.2 zeigt diese Situation, wobei α' und β' die Nachkommen von α bzw. β seien, die jeweils mit der Wahrscheinlichkeit λ entstehen. Es gibt dann vier mögliche Paare im $(n+1)$-ten Zyklus.

- Bei dem Paar handelt es sich um α, β. Die erwartete Anzahl solcher Paare ist $C_n(k)$.
- Das Paar ist α', β, die erwartete Anzahl dieser Paare ist $\lambda C_n(k)$.
- Das Paar besteht aus α, β', für diese Kombination werden $\lambda C_n(k)$ Paare erwartet.
- Es liegt das Paar α', β' vor; die erwartete Anzahl hiervon ist $\lambda^2 C_n(k)$.

Zählt man alle Möglichkeiten zusammen, erhält man

$$C_{n+1}(k) = \binom{n}{k}\lambda^{k+1} + (1+2\lambda+\lambda^2)C_n(k) = \binom{n}{k}\lambda^{k+1} + (1+\lambda)^2 C_n(k)$$

also (i).

(ii) Es sei $\varphi_{C_n}(s)$ die wie oben definierte erzeugende Funktion von $C_n(k)$, $k \in \mathbb{N}_0$. Da im n-ten PCR-Schritt höchstens eine Sequenz der Generation n entstehen kann, gibt es nach n Zyklen kein Paar mit einem MRCA der Generation n, d.h. $C_n(n) = 0$. Dann gilt mit (i):

$$\begin{aligned}
\varphi_{C_{n+1}}(s) &= \sum_{k=0}^{n} C_{n+1}(k)s^k \\
&= \sum_{k=0}^{n} \left((1+\lambda)^2 C_n(k) + \binom{n}{k}\lambda^{k+1} \right) s^k \\
&= (1+\lambda)^2 \sum_{k=0}^{n-1} C_n(k)s^k + \lambda \sum_{k=0}^{n} \binom{n}{k}(\lambda s)^k \\
&= (1+\lambda)^2 \varphi_{C_n}(s) + \lambda(1+\lambda s)^n
\end{aligned}$$

In die letzte Gleichung gehen die Definition der erzeugenden Funktion und der binomische Lehrsatz ein. Nun folgt eine Induktion über n. Der Fall $n = 1$ ist schnell erledigt. Die erwartete Anzahl an Paaren mit einem MRCA der 0-ten Generation ist dann gleich λ, d.h. $\varphi_{C_1}(s) = C_1(0) = \lambda$. Gleichzeitig ist $\lambda\frac{(1+\lambda s)^1-(1+\lambda)^2}{(1+\lambda s)-(1+\lambda)^2} = \lambda$. Gelte nun die Behauptung für beliebiges aber festes n. Dann folgt:

$$\begin{aligned}
\varphi_{C_{n+1}}(s) &= (1+\lambda)^2\varphi_{C_n}(s) + \lambda(1+\lambda s)^n \\
&\overset{\text{I.V.}}{=} (1+\lambda)^2\lambda\frac{(1+\lambda s)^n - (1+\lambda)^{2n}}{(1+\lambda s)-(1+\lambda)^2} + \lambda(1+\lambda s)^n \\
&= \lambda\frac{(1+\lambda)^2(1+\lambda s)^n - (1+\lambda)^{2(n+1)} + (1+\lambda s)^n\left((1+\lambda s)-(1+\lambda)^2\right)}{(1+\lambda s)-(1+\lambda)^2} \\
&= \lambda\frac{(1+\lambda s)^{n+1} - (1+\lambda)^{2(n+1)}}{(1+\lambda s)-(1+\lambda)^2}
\end{aligned}$$

Dies impliziert die Behauptung (ii). □

Das nächste Lemma erläutert das Verhalten der erwarteten Generation des MRCA eines zufällig gezogenen Paares für eine große Anzahl an Startmolekülen $j = Z_0$. Danach stammt der MRCA zweier zufällig gezogener PCR-Produkte für $j \to \infty$ aus der 0-ten Generation.

Lemma 2.18 *Gegeben sei eine PCR und für $n \geq 1$ sei $\bar{A}_n := g\big(\mathrm{MRCA}(A,B)\big)$, wobei A, B zwei $\mathcal{Z}_n$-wertige, Laplace-verteilte Zufallsgrößen seien. $\bar{A}_n$ bezeichnet also die Generationsnummer des MRCA eines zufällig mit Zurücklegen gezogenen Paares nach n PCR-Zyklen. Dann gilt für alle $n \in \mathbb{N}$:*

$$\lim_{j\to\infty} jE_j\bar{A}_n = \frac{2}{(1+\lambda)^2} - \frac{2+n\lambda(1-\lambda)}{(1+\lambda)^{n+2}}$$

Beweis: Zuerst betrachten wir die PCR mit einem Urahn. $C_n^*(k)$ sei die Anzahl der erwarteten Paare mit einem MRCA der k-ten Generation, wobei die Zugreihenfolge berücksichtigt und mit Zurücklegen gezogen wird, also

$$C_n^*(k) := E|\{(\alpha,\beta) \in \mathcal{Z}_n \times \mathcal{Z}_n : \ g(\mathrm{MRCA}(\alpha,\beta)) = k\}| \,.$$

Nach Lemma 2.11 gibt es $\binom{n}{k}\lambda^k$ erwartete gleiche Paare mit einem MRCA der Generation k, daher folgt

$$C_n^*(k) = 2C_n(k) + \binom{n}{k}\lambda^k \tag{2.20}$$

Nun sei die Anzahl an Startmolekülen $Z_0 = j \in \mathbb{N}$ beliebig. $Y_k^n(i),\ 1 \le i \le j$, bezeichne die Anzahl aller Paare aus $\mathcal{Z}_n \times \mathcal{Z}_n$ mit einem MRCA der k-ten Generation und demselben Urahn $(0,i)$. Wie immer sei $Z_n(i)$ die Anzahl der Sequenzen mit dem Urahn $(0,i)$ nach n PCR-Schritten. Für die bedingte Verteilung der Generationsnummer $\bar{A}_n$ gegeben $(Y_k^n(i))_{1\le i\le j}, (Z_n(i))_{1\le i\le j}$ erhält man für alle $1 \le k \le n$:

$$P_j\left(\bar{A}_n = k \mid (Y_k^n(i))_{1\le i\le j}, (Z_n(i))_{1\le i\le j}\right) = \frac{\sum_{i=1}^j Y_k^n(i)}{\left(\sum_{i=1}^j Z_n(i)\right)^2} \text{ f.s.}$$

Also gilt für alle $1 \le k \le n$:

$$P_j\left(\bar{A}_n = k\right) = E_j \frac{\sum_{i=1}^j Y_k^n(i)}{\left(\sum_{i=1}^j Z_n(i)\right)^2} \text{ f.s.} \tag{2.21}$$

Da die $Y_k^n(i),\ 1 \le i \le j$ stochastisch unabhängig und identisch verteilt sind mit $EY_k^n(1) = C_n^*(k)$ (siehe oben), gilt das starke Gesetz der großen Zahlen für die Summenfolge der $Y_k^n(i)$ vermöge des Satzes von Etemadi (siehe [3] Satz 35.4). Somit erhält man:

$$\begin{aligned}
\lim_{j\to\infty} jP_j(\bar{A}_n = k) &\overset{(2.21)}{=} \lim_{j\to\infty} E_j\left(\frac{\sum_{i=1}^j Y_k^n(i)}{j} \cdot \frac{j^2}{\left(\sum_{i=1}^j Z_n(i)\right)^2}\right) \\
&= \frac{EY_k^n(1)}{(EZ_n(1))^2} \\
&= \frac{C_n^*(k)}{(1+\lambda)^{2n}}
\end{aligned} \tag{2.22}$$

Da $\sum_{k=1}^n k\binom{n}{k}\lambda^k = n\lambda\sum_{k=1}^{n-1}\binom{n-1}{k}\lambda^{k-1}$ und die erste Ableitung der in Lemma 2.17 (ii) eingeführten erzeugenden Funktion die Gestalt $\varphi'_{C_n}(s) = \sum_{k=1}^{n-1} kC_n(k)s^{k-1}$ für $n > 1$

sowie $\varphi'_{C_1}(s) = 0$ besitzt, gilt:

$$\begin{aligned}
\lim_{j\to\infty} jE_j\bar{A}_n &= \lim_{j\to\infty} j\sum_{k=1}^{n} kP_j(\bar{A}_n = k) \\
&= \sum_{k=1}^{n} k \lim_{j\to\infty} jP_j(\bar{A}_n = k) \\
&\overset{(2.22)}{=} \sum_{k=1}^{n} k\frac{C_n^*(k)}{(1+\lambda)^{2n}} \\
&\overset{(2.20)}{=} \frac{1}{(1+\lambda)^{2n}} \sum_{k=1}^{n} k\left(2C_n(k) + \binom{n}{k}\lambda^k\right) \\
&= \frac{1}{(1+\lambda)^{2n}}\left(2\varphi'_{C_n}(1) + n\lambda(1+\lambda)^{n-1}\right) \\
&= \frac{2}{(1+\lambda)^2} - \frac{2+n\lambda(1-\lambda)}{(1+\lambda)^{n+2}}
\end{aligned}$$

Die letzte Gleichheit ergibt sich durch

$$\varphi'_{C_n}(1) = (n-1)(1+\lambda)^{n-2} - n(1+\lambda)^{n-1} + (1+\lambda)^{2n-2}.$$

Einsetzen und Auflösen liefert die gewünschte Gleichung. □

Die gemeinsame Verteilung der Generationsnummer $g(A)$ und $g(B)$ der $\mathcal{Z}_n$-wertigen und Laplace-verteilten Zufallsvektoren A, B lässt sich ähnlich wie Gleichung (2.21) über die bedingte Verteilung bestimmen und man erhält:

$$P\big(g(A) = k,\ g(B) = l\big) = E\frac{X_k^n X_l^n}{Z_n^2}$$

Da $g(A)$ und $g(B)$ identisch verteilt sind, ergibt sich für die Kovarianz von $g(A)$ und $g(B)$ unter jeder Anfangsverteilung P_j:

$$\begin{aligned}
Cov_j\left(g(A),\ g(B)\right) &= E_j\sum_{k=0}^{n}\sum_{l=0}^{n}\frac{klX_k^nX_l^n}{Z_n^2} - E_j\sum_{k=0}^{n}\frac{kX_k^n}{Z_n}E_j\sum_{l=0}^{n}\frac{lX_l^n}{Z_n} \\
&= E_j\sum_{k=0}^{n}\sum_{l=0}^{n}\frac{klX_k^nX_l^n}{Z_n^2} - \left(E_j\sum_{k=0}^{n}\frac{kX_k^n}{Z_n}\right)^2 \\
&= E_j\left(\sum_{k=0}^{n}\frac{kX_k^n}{Z_n}\right)^2 - \left(E_j\sum_{k=0}^{n}\frac{kX_k^n}{Z_n}\right)^2 \\
&= Var_j\left(\sum_{k=0}^{n}\frac{kX_k^n}{Z_n}\right) \qquad (2.23)
\end{aligned}$$

Weiter seien für $1 \le i \le j$

$$T_n := \sum_{k=0}^{n} kX_k^n \quad \text{und} \quad T_n(i) := \sum_{k=0}^{n} kX_k^n(i)\,,$$

also $T_n = \sum_{i=1}^{j} T_n(i)$. Somit lässt sich (2.23) schreiben als

$$Cov_j\left(g(A),\ g(B)\right) = Var_j\left(\frac{T_n}{Z_n}\right) = Var_j\left(\frac{\bar{T}_n}{\bar{Z}_n}\right) \qquad (2.24)$$

mit $\bar{T}_n := \frac{1}{j}\sum_{i=1}^{j} T_n(i)$ und $\bar{Z}_n := \frac{1}{j}\sum_{i=1}^{j} Z_n(i)$. Also lässt sich das Grenzverhalten von $Cov_j\left(g(A),\ g(B)\right)$ auf dasjenige von $Var_j\left(\frac{\bar{T}_n}{\bar{Z}_n}\right)$ zurückführen. Zu dessen Bestimmung ist folgendes technische Hilfslemma vonnöten.

Lemma 2.19 *Es sei $\left((X_i, Y_i)^t\right)_{i\geq 1}$ eine Folge von stochastisch unabhängigen, identisch verteilten Zufallsvektoren mit Erwartungswert $(\xi, \upsilon)^t$ und Kovarianzmatrix $\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. Ferner seien für alle $i \geq 1$ die $Y_i \geq c > 0$, damit $\upsilon \neq 0$, und die X_i beschränkt durch $a \in \mathbb{R}$. Mit $\bar{X} := \frac{1}{m}\sum_{i=1}^{m} X_i$ und $\bar{Y} := \frac{1}{m}\sum_{i=1}^{m} Y_i$ gilt dann*
(i)

$$\lim_{m\to\infty} m\left(E\frac{\bar{X}}{\bar{Y}} - \frac{\xi}{\upsilon}\right) = \frac{\xi}{\upsilon^3}\sigma_2^2 - \frac{1}{\upsilon^2}\rho\sigma_1\sigma_2$$

(ii)

$$\lim_{m\to\infty} mVar\frac{\bar{X}}{\bar{Y}} = \frac{1}{\upsilon^2}\sigma_1^2 - \frac{2\xi}{\upsilon^3}\rho\sigma_1\sigma_2 + \frac{\xi^2}{\upsilon^4}\sigma_2^2\,.$$

Beweis:
(i) Mit dem Satz von Taylor (siehe [14] Satz 61.1) ergibt sich:

$$E\frac{\bar{X}}{\bar{Y}} = E\bar{X}\left(\frac{1}{E\bar{Y}} - \frac{1}{(E\bar{Y})^2}(\bar{Y} - E\bar{Y}) + \frac{2}{2!(E\bar{Y})^3}(\bar{Y} - E\bar{Y})^2 - \frac{3}{3!\Upsilon^4}(\bar{Y} - E\bar{Y})^3\right)$$

wobei Υ zwischen $\bar{Y}$ und $E\bar{Y}$ liegt. Da $E\bar{X} = \frac{1}{m}\sum_{i=1}^{m} EX_i = \xi$ und $E\bar{Y} = \frac{1}{m}\sum_{i=1}^{m} EY_i = \upsilon$, gilt weiter

$$m\left(E\frac{\bar{X}}{\bar{Y}} - \frac{\xi}{\upsilon}\right) = -m\frac{E\bar{X}(\bar{Y} - E\bar{Y})}{\upsilon^2} + m\frac{E\bar{X}(\bar{Y} - E\bar{Y})^2}{\upsilon^3} - m\frac{E\bar{X}(\bar{Y} - E\bar{Y})^3}{2\Upsilon^4}$$

Die in dieser Gleichung auftauchenden Terme werden nun näher betrachtet:

$$\begin{aligned}
mE\bar{X}(\bar{Y} - E\bar{Y}) &= m\left(E\bar{X}\bar{Y} - E\bar{X}E\bar{Y}\right) \\
&= mCov\left(\bar{X}, \bar{Y}\right) \\
&= mCov\left(\frac{1}{m}\sum_{i=1}^{m} X_i, \frac{1}{m}\sum_{j=1}^{m} Y_j\right) \\
&= \frac{1}{m}\left(\sum_{i=1}^{m}\sum_{j=1}^{m} Cov\left(X_i, Y_j\right)\right) \\
&= \frac{1}{m}\left(\sum_{i=1}^{m} Cov\left(X_i, Y_i\right)\right) \\
&= Cov\left(X_1, Y_1\right) \\
&= \rho\sigma_1\sigma_2
\end{aligned}$$

wobei die Unabhängigkeit und somit die Unkorreliertheit der $X_i,\ Y_j$ für $i \neq j$ ausgenutzt wurde. Für die Untersuchung des nächsten Terms ist zu bemerken, dass durch die Beschränktheit der X_i für alle $i \geq 1$ auch $\bar{X}$ durch a beschränkt ist. Außerdem gilt $E(\bar{Y} - E\bar{Y})^2 = Var\bar{Y} = \frac{1}{m}VarY_1$. Damit lässt sich der Satz von der majorisierten Konvergenz (siehe [3] Satz 9.9 (c)) anwenden. Desweiteren impliziert der Satz von Etemadi

(siehe [3] Satz 35.4), dass $\lim_{m\to\infty}\bar{X} = EX_1$ f.s. Also folgt

$$\begin{aligned}\lim_{m\to\infty} mE\bar{X}(\bar{Y}-E\bar{Y})^2 &= E\lim_{m\to\infty}\bar{X}m(\bar{Y}-E\bar{Y})^2\\ &= EX_1E\lim_{m\to\infty}m(\bar{Y}-E\bar{Y})^2\\ &= EX_1\lim_{m\to\infty}mE(\bar{Y}-E\bar{Y})^2\\ &= EX_1VarY_1\\ &= \xi\sigma_2^2\end{aligned}$$

Die gleiche Vorgehensweise liefert $\lim_{m\to\infty} mE\bar{Y}(\bar{Y}-E\bar{Y})^2 = \upsilon VarY_1$. Da Y_i für alle $i \geq 1$ nach unten durch c beschränkt ist, gilt dies auch für Υ. Man erhält also

$$\begin{aligned}m\frac{E\bar{X}(\bar{Y}-E\bar{Y})^3}{2\Upsilon^4} &\leq \frac{a}{2c^4}mE(\bar{Y}-E\bar{Y})^3\\ &= \frac{a}{2c^4}mE(\bar{Y}-E\bar{Y})(\bar{Y}-E\bar{Y})^2\\ &= \frac{a}{2c^4}m\left(E\bar{Y}(\bar{Y}-E\bar{Y})^2 - E\left(E\bar{Y}(\bar{Y}-E\bar{Y})^2\right)\right)\\ &= \frac{a}{2c^4}\left(mE\bar{Y}(\bar{Y}-E\bar{Y})^2 - \upsilon VarY_1\right)\\ &\xrightarrow[m\to\infty]{} 0.\end{aligned}$$

Insgesamt folgt

$$m\left(E\frac{\bar{X}}{\bar{Y}} - \frac{\xi}{\upsilon}\right) \xrightarrow[m\to\infty]{} \frac{\xi}{\upsilon^3}\sigma_2^2 - \frac{1}{\upsilon^2}\rho\sigma_1\sigma_2$$

(ii) Auch diesen Teil erledigt eine Taylor-Entwicklung. Es gilt

$$Var\frac{\bar{X}}{\bar{Y}} = Var\bar{X}\left(\frac{1}{E\bar{Y}} - \frac{1}{(E\bar{Y})^2}(\bar{Y}-E\bar{Y}) + \frac{2}{2!(E\bar{Y})^3}(\bar{Y}-E\bar{Y})^2 - \frac{3}{3!\Upsilon^4}(\bar{Y}-E\bar{Y})^3\right),$$

mit einer Zwischenstelle $\Upsilon \in (\bar{Y}\wedge E\bar{Y}, \bar{Y}\vee E\bar{Y})$. Man erhält also

$$\begin{aligned}mVar\frac{\bar{X}}{\bar{Y}} &= mVar\frac{\bar{X}}{E\bar{Y}} + mVar\frac{\bar{X}(\bar{Y}-E\bar{Y})}{(E\bar{Y})^2} + mVar\frac{\bar{X}(\bar{Y}-E\bar{Y})^2}{(E\bar{Y})^3}\\ &\quad + mVar\frac{\bar{X}(\bar{Y}-E\bar{Y})^3}{2\Upsilon^4} + 2m\sum_{i=1}^{6}\Gamma_i,\end{aligned}$$

mit

$$\begin{aligned}\Gamma_1 &= Cov\left(\frac{\bar{X}}{E\bar{Y}}, -\frac{\bar{X}(\bar{Y}-E\bar{Y})}{(E\bar{Y})^2}\right)\\ \Gamma_2 &= Cov\left(\frac{\bar{X}}{E\bar{Y}}, \frac{\bar{X}(\bar{Y}-E\bar{Y})^2}{(E\bar{Y})^3}\right)\\ \Gamma_3 &= Cov\left(\frac{\bar{X}}{E\bar{Y}}, -\frac{\bar{X}(\bar{Y}-E\bar{Y})^3}{2\Upsilon^4}\right)\\ \Gamma_4 &= Cov\left(-\frac{\bar{X}(\bar{Y}-E\bar{Y})}{(E\bar{Y})^2}, \frac{\bar{X}(\bar{Y}-E\bar{Y})^2}{(E\bar{Y})^3}\right)\\ \Gamma_5 &= Cov\left(-\frac{\bar{X}(\bar{Y}-E\bar{Y})}{(E\bar{Y})^2}, -\frac{\bar{X}(\bar{Y}-E\bar{Y})^3}{2\Upsilon^4}\right)\\ \Gamma_6 &= Cov\left(\frac{\bar{X}(\bar{Y}-E\bar{Y})^2}{(E\bar{Y})^3}, -\frac{\bar{X}(\bar{Y}-E\bar{Y})^3}{2\Upsilon^4}\right).\end{aligned}$$

Das Grenzverhalten der auftretenden Terme wird nun getrennt voneinander untersucht.

$$mVar\frac{\bar{X}}{E\bar{Y}} = \frac{m}{\upsilon^2}\frac{1}{m}VarX_1 = \frac{1}{\upsilon^2}\sigma_1^2$$

Für die Berechnung des Grenzwertes des nächsten Terms benötigt man ähnlich wie in (i) den Satz von der monotonen Konvergenz.

$$\begin{aligned}\lim_{m\to\infty} mVar\frac{\bar{X}(\bar{Y}-E\bar{Y})}{(E\bar{Y})^2} &= \frac{1}{\upsilon^4}\lim_{m\to\infty} m\left(E\bar{X}^2(\bar{Y}-E\bar{Y})^2 - \left(E\bar{X}(\bar{Y}-E\bar{Y})\right)^2\right)\\ &= \frac{1}{\upsilon^4}\Big(\xi^2\underbrace{\lim_{m\to\infty} mVar\bar{Y}}_{=VarY_1} - \lim_{m\to\infty} m\big(\underbrace{Cov(\bar{X},\bar{Y})}_{=\frac{1}{m}Cov(X_1,Y_1)}\big)^2\Big)\\ &= \frac{\xi^2}{\upsilon^4}\sigma_2^2\end{aligned}$$

Da $\bar{X}$ nach oben durch a beschränkt ist, verschwindet auch der dritte Term, wie folgende Abschätzung zeigt.

$$\begin{aligned}mVar\frac{\bar{X}(\bar{Y}-E\bar{Y})^2}{(E\bar{Y})^3} &\leq \frac{a^2}{\upsilon^6}m\left(E(\bar{Y}-E\bar{Y})^4 - \left(E(\bar{Y}-E\bar{Y})^2\right)^2\right)\\ &= \frac{a^2}{\upsilon^6}\Big(mE(\bar{Y}-E\bar{Y})^2(\bar{Y}-E\bar{Y})^2 - \frac{1}{m}\big(\underbrace{mVar\bar{Y}}_{=VarY_1}\big)^2\Big)\\ &\xrightarrow[m\to\infty]{} 0\end{aligned}$$

Wobei für den letzten Grenzübergang der Satz von der monotonen Konvergenz und die Tatsache, dass $mE(\bar{Y}-E\bar{Y})^2 = \frac{1}{m}VarY_1$ gilt, benutzt werden. Ganz ähnlich lässt sich auch der vierte Term abschätzen:

$$\begin{aligned}mVar\frac{\bar{X}(\bar{Y}-E\bar{Y})^3}{2\Upsilon^4} &\leq \frac{a^2}{4c^8}mVar(\bar{Y}-E\bar{Y})^3\\ &= \frac{a^2}{4c^8}m\left(E(\bar{Y}-E\bar{Y})^6 - \left(E(\bar{Y}-E\bar{Y})^3\right)^2\right)\\ &= \frac{a^2}{4c^8}\Big(mE(\bar{Y}-E\bar{Y})^2(\bar{Y}-E\bar{Y})^4 - \frac{1}{m}\big(\underbrace{mE(\bar{Y}-E\bar{Y})^3}_{\xrightarrow[m\to\infty]{}0}\big)^2\Big)\\ &\xrightarrow[m\to\infty]{} 0\end{aligned}$$

Nun muss nur noch $m\sum_{i=1}^6\Gamma_i$ genauer betrachtet werden. In den Kovarianztermen Γ_3, Γ_5 und Γ_6 taucht die Zwischenstelle Υ auf. Sie lassen sich ähnlich wie die Varianz abschätzen und verschwinden durch den Grenzübergang $m\to\infty$. Für die anderen drei Terme ergibt sich mit dem Satz von der monotonen Konvergenz:

$$\begin{aligned}m\left(\Gamma_1+\Gamma_2+\Gamma_4\right) &= \frac{3m}{\upsilon^2}Var\bar{X} - \frac{7m}{\upsilon^3}Cov\left(\bar{X},\bar{X}\bar{Y}\right) + \frac{2m}{\upsilon^4}Cov\left(\bar{X},\bar{X}\bar{Y}^2\right)\\ &\quad + \frac{2m}{\upsilon^4}Cov\left(\bar{X}\bar{Y},\bar{X}\bar{Y}\right) - \frac{m}{\upsilon^5}Cov\left(\bar{X}\bar{Y},\bar{X}\bar{Y}^2\right)\\ &\xrightarrow[m\to\infty]{} VarX_1\left(\frac{3}{\upsilon^2} - \frac{7}{\upsilon^2} + \frac{2}{\upsilon^2} + \frac{2}{\upsilon^2}\right) - \frac{1}{\upsilon^3}EX_1Cov(X_1,Y_1)\\ &= -\frac{\xi}{\upsilon^3}\rho\sigma_1\sigma_2\end{aligned}$$

Insgesamt folgt die Behauptung. □

Um dieses Lemma anwenden zu können, müssen nun die geforderten Größen berechnet werden. Dies erledigt das folgende Lemma.

Lemma 2.20 *Gegeben sei eine n-stufige PCR, $n \geq 1$, mit einem Startmolekül ($Z_0 = 1$). Unter den üblichen Bezeichnungen und $R_n := \sum_{k=0}^{n} k^2 X_k^n$ gelten:*

$$(i) \qquad VarZ_n = (1-\lambda)(1+\lambda)^{n-1}\left((1+\lambda)^n - 1\right)$$

$$(ii) \qquad Cov(Z_n, T_n) = (1-\lambda)(1+\lambda)^{n-2}(n\lambda+1)\left((1+\lambda)^n - 1\right)$$

$$(iii) \qquad VarT_n = (1-\lambda)(1+\lambda)^{n-3}\left[(1+\lambda)^n\left((n\lambda)^2 + 2n\lambda + 2\right) - n\lambda(n\lambda+3) - 2\right]$$

$$(iv) \qquad Cov(R_n, Z_n) = (1-\lambda)(1+\lambda)^{n-3}\left((n\lambda)^2 + (3n-1)\lambda + 1\right)\left((1+\lambda)^n - 1\right)$$

Beweis:
(i) Diese Formel ist ein Standardergebnis der Verzweigungsprozesstheorie und anwendbar, da $(Z_n)_{n\geq 0}$ ein Galton-Watson-Prozess mit $EZ_1 = 1+\lambda$ und $VarZ_1 = \lambda(1-\lambda)$ ist (siehe [5] Kapitel 1, Abschnitt 2).

(ii) Zum Beweis dieses Ergebnisses müssen wir zuerst den Erwartungswert $EZ_{n+1}T_{n+1}$ errechnen, indem wir unter dem Verhalten des Startmoleküls bedingen. Dabei gibt es zwei Fälle:
1. Fall: Mit der Wahrscheinlichkeit $1-\lambda$ wird das Startmolekül nicht reproduziert. Dann gibt es nach einem PCR-Zyklus nur das eine Startmolekül, so dass man den zweiten Zyklus als Startpunkt ansehen kann. Die diesem Zyklus korrespondierenden Zufallsgrößen für die Gesamtanzahl aller Sequenzen und T_n nach n Zyklen - gezählt vom neuen Startpunkt - bezeichne man mit Z_n^* und T_n^*. Offensichtlich sind (Z_n, T_n) und (Z_n^*, T_n^*) identisch verteilt.
2. Fall: Das Startmolekül wird im ersten Zyklus dupliziert. Dies geschieht mit einer Wahrscheinlichkeit von λ. Nach dem ersten Zyklus existieren dann eine Sequenz der ersten und eine der nullten Generation. Die Sequenz der nullten Generation lässt sich wie im ersten Fall als Ausgangspunkt einer neuen PCR ansehen, deren korrespondierende Zufallsgrößen mit $Z_n^{(0)}$ bzw. $T_n^{(0)}$ gekennzeichnet werden. Auch die Sequenz der ersten Generation bildet den Ausgangspunkt einer neuen PCR. Es sei $\tilde{X}_k^n$ die Anzahl der Sequenzen der k-ten Generation nach n PCR-Zyklen, mit der im ersten Schritt entstandene Sequenz als „Urahnen“, sowie

$$T_n^{(1)} := \sum_{k=0}^{n} k\tilde{X}_k^n \quad \text{und} \quad Z_n^{(1)} := \sum_{k=0}^{n} \tilde{X}_k^n.$$

Um die tatsächliche Generationsnummer zu erhalten, muss in diesem Fall aber noch 1 hinzuaddiert werden und man erhält für die korrespondierende Zufallsgröße

$$\tilde{T}_n^{(1)} = \sum_{k=0}^{n}(k+1)\tilde{X}_k^n = \sum_{k=0}^{n} k\tilde{X}_k^n + \sum_{k=0}^{n}\tilde{X}_k^n = T_n^{(1)} + Z_n^{(1)}.$$

Also gilt für den zweiten Fall:

$$T_{n+1} = T_n^{(0)} + T_n^{(1)} + Z_n^{(1)} \quad \text{und} \quad Z_{n+1} = Z_n^{(0)} + Z_n^{(1)}$$

In dem beschriebenen Modell sind $(Z_n^{(0)}, T_n^{(0)})$ und $(Z_n^{(1)}, T_n^{(1)})$ stochastisch unabhängig und genauso verteilt wie (Z_n, T_n).

Der Satz von der totalen Wahrscheinlichkeit (siehe [25] Satz 4.4) liefert unter Berücksichtigung der vorangegangenen Überlegungen:

$$\begin{aligned} EZ_{n+1}T_{n+1} &= \underbrace{(1-\lambda)EZ_n^*T_n^*}_{1.\ Fall} + \underbrace{\lambda E\left(Z_n^{(0)}+Z_n^{(1)}\right)\left(T_n^{(0)}+T_n^{(1)}+Z_n^{(1)}\right)}_{2.\ Fall} \\ &= (1-\lambda)EZ_nT_n + \lambda\Big(EZ_n^{(0)}T_n^{(0)} + EZ_n^{(0)}T_n^{(1)} + EZ_n^{(0)}Z_n^{(1)} \\ &\qquad + EZ_n^{(1)}T_n^{(0)} + EZ_n^{(1)}T_n^{(1)} + EZ_n^{(1)}Z_n^{(1)}\Big) \\ &= (1+\lambda)EZ_nT_n + 2\lambda EZ_nET_n + \lambda\left(EZ_n\right)^2 + \lambda EZ_n^2 \end{aligned} \tag{2.25}$$

Nach Lemma 2.10 gilt $EZ_n = (1+\lambda)^n$. Weiter errechnet man:

$$\begin{aligned} ET_n &= E\sum_{k=0}^{n} kX_k^n \\ &= \sum_{k=0}^{n} EkX_k^n \\ &\overset{(2.10)}{=} \sum_{k=0}^{n} k\binom{n}{k}\lambda^k \\ &= n\lambda\sum_{k=1}^{n}\binom{n-1}{k-1}\lambda^{k-1} \\ &= n\lambda(1+\lambda)^{n-1} \end{aligned} \tag{2.26}$$

Und für den Erwartungswert des Quadrates der Gesamtzahl nach n PCR-Zyklen ergibt sich:

$$\begin{aligned} EZ_n^2 &= VarZ_n + \left(EZ_n\right)^2 \\ &\overset{(2.9)}{=} (1-\lambda)(1+\lambda)^{n-1}\left((1+\lambda)^n-1\right) + (1+\lambda)^{2n} \\ &= (1+\lambda)^{n-1}\left((1-\lambda)(1+\lambda)^n - (1-\lambda) + (1+\lambda)^{n+1}\right) \\ &= (1+\lambda)^{n-1}\left(2(1+\lambda)^n - (1-\lambda)\right) \end{aligned}$$

So erhält man dann in Gleichung (2.25):

$$\begin{aligned} EZ_{n+1}T_{n+1} &= (1+\lambda)EZ_nT_n + 2\lambda EZ_nET_n + \lambda\left(EZ_n\right)^2 + \lambda EZ_n^2 \\ &= (1+\lambda)EZ_nT_n \\ &\quad + 2n\lambda^2(1+\lambda)^{2n-1} + \lambda(1+\lambda)^{2n} + \lambda(1+\lambda)^{n-1}\left(2(1+\lambda)^n - (1-\lambda)\right) \\ &= (1+\lambda)EZ_nT_n \\ &\quad + \lambda(1+\lambda)^{n-1}\left(2n\lambda(1+\lambda)^n + (1+\lambda)^{n+1} + 2(1+\lambda)^n - (1-\lambda)\right) \\ &= (1+\lambda)EZ_nT_n + \lambda(1+\lambda)^{n-1}\left[(\lambda(2n+1)+3)(1+\lambda)^n - (1-\lambda)\right] \end{aligned}$$

Damit lässt sich eine Rekursionsformel für die Kovarianz von Z_{n+1} und T_{n+1} errechnen:

$$\begin{aligned}
Cov(Z_{n+1},T_{n+1}) &= EZ_{n+1}T_{n+1} - EZ_{n+1}ET_{n+1}\\
&= (1+\lambda)EZ_nT_n + \lambda(1+\lambda)^{n-1}\left[(\lambda(2n+1)+3)(1+\lambda)^n - (1-\lambda)\right]\\
&\qquad -(1+\lambda)^{n+1}(n+1)\lambda(1+\lambda)^n\\
&= (1+\lambda)EZ_nT_n + \lambda^2(1+\lambda)^{2n-1}(2n+1) + 3\lambda(1+\lambda)^{2n-1}\\
&\qquad -\lambda(1+\lambda)^{n-1}(1-\lambda) - (1+\lambda)^{2n+1}(n+1)\lambda\\
&= (1+\lambda)EZ_nT_n + \lambda^2(1+\lambda)^{2n-1}(2n+1) - \lambda(1+\lambda)^{n-1}(1-\lambda)\\
&\qquad +\lambda(1+\lambda)^{2n-1}\left(3-(1+\lambda)^2(n+1)\right)\\
&= (1+\lambda)EZ_nT_n + \lambda^2(1+\lambda)^{2n-1}(2n+1) - \lambda(1+\lambda)^{n-1}(1-\lambda)\\
&\qquad +\lambda(1+\lambda)^{2n-1}\left(2-n-2n\lambda-n\lambda^2-2\lambda-\lambda^2\right)\\
&= (1+\lambda)EZ_nT_n + \lambda^2(1+\lambda)^{2n-1}(2n+1) - \lambda(1+\lambda)^{n-1}(1-\lambda)\\
&\qquad +\lambda(1+\lambda)^{2n-1}\left[\left(2-n\lambda-n\lambda^2-2\lambda-\lambda^2\right)-n(1+\lambda)\right]\\
&= (1+\lambda)EZ_nT_n - \underbrace{n(1+\lambda)\lambda(1+\lambda)^{2n-1}}_{=(1+\lambda)EZ_nET_n} - \lambda(1+\lambda)^{n-1}(1-\lambda)\\
&\qquad +\lambda(1+\lambda)^{2n-1}\Big(\underbrace{\lambda(2n+1)+2-n\lambda-n\lambda^2-2\lambda-\lambda^2}_{=(1-\lambda)((n+1)\lambda+2)}\Big)\\
&= (1+\lambda)Cov(Z_n,T_n) + (1-\lambda)\lambda(1+\lambda)^{n-1}\\
&\qquad \cdot\left[((n+1)\lambda+2)(1+\lambda)^n - 1\right] \qquad (2.27)
\end{aligned}$$

Den Rest erledigt man induktiv. Sei dazu zuerst $n=1$. Da $T_1 = X_1^1$, also nur die Werte 0 und 1 mit den Wahrscheinlichkeiten $P(T_1=1)=\lambda=1-P(T_1=0)$ annehmen kann, und $P(Z_1=2)=\lambda=1-P(Z_1=1)$ gilt, folgt für die Kovarianz von Z_1 und T_1

$$Cov\,(Z_1,T_1) = EZ_1T_1 - EZ_1ET_1 = 2\lambda - (1+\lambda)\lambda = \lambda-\lambda^2.$$

Die in (ii) auftauchende Formel für die Kovarianz liefert

$$(1-\lambda)(1+\lambda)^{1-2}(\lambda+1)(1+\lambda-1) = (1-\lambda)\lambda = \lambda-\lambda^2$$

also Übereinstimmung. Gegeben sei nun die Behauptung für beliebiges aber festes n, dann gilt:

$$\begin{aligned}
Cov(Z_{n+1},T_{n+1}) &\overset{(2.27)}{=} (1+\lambda)Cov(Z_n,T_n) + (1-\lambda)\lambda(1+\lambda)^{n-1}\\
&\qquad \cdot\left[((n+1)\lambda+2)(1+\lambda)^n - 1\right]\\
&\overset{\text{I.V.}}{=} (1-\lambda)(1+\lambda)^{(n+1)-2}(n\lambda+1)\left((1+\lambda)^n-1\right)\\
&\qquad +(1-\lambda)\lambda(1+\lambda)^{n-1}\left[((n+1)+\lambda+2)(1+\lambda)^n-1\right]\\
&= (1-\lambda)(1+\lambda)^{(n+1)-2}\big[(n\lambda+1)\left((1+\lambda)^n-1\right)\\
&\qquad +\lambda\left(n\lambda+\lambda+2\right)(1+\lambda)^n-\lambda\big]\\
&= (1-\lambda)(1+\lambda)^{(n+1)-2}\big[(n\lambda+1)(1+\lambda)^n-(n\lambda+1)\\
&\qquad +\lambda(n\lambda+1)(1+\lambda)^n+\lambda(1+\lambda)^{n+1}-\lambda\big]\\
&= (1-\lambda)(1+\lambda)^{(n+1)-2}\big[(n\lambda+1)\left((1+\lambda)^n(1+\lambda)-1\right)\\
&\qquad +\lambda\left((1+\lambda)^{n+1}-1\right)\big]\\
&= (1-\lambda)(1+\lambda)^{(n+1)-2}\big((n+1)\lambda+1\big)\left((1+\lambda)^{n+1}-1\right)
\end{aligned}$$

Daraus folgt die Behauptung.

(iii) Mit derselben Idee wie in (i) lässt sich für die Varianz der Zufallsgröße T_n eine Rekursionsformel angeben:

$$\begin{aligned} VarT_{n+1} &= (1+\lambda)VarT_n + (1-\lambda)\lambda(1+\lambda)^{n-2} \\ &\quad \cdot \left[\left((n+1)^2\lambda^2 + (4n+3)\lambda + 4\right)(1+\lambda)^n - \left((2n+1)\lambda + 3\right)\right] \end{aligned}$$

Über eine Induktion beweist man die angegebene Formel.

(iv) Das Bedingen unter dem Anfangsverhalten des Startmoleküls liefert die Rekursionsformel:

$$\begin{aligned} Cov\left(R_{n+1}, Z_{n+1}\right) &= (1+\lambda)Cov\left(R_n, Z_n\right) + (1-\lambda)\lambda(1+\lambda)^{n-2} \\ &\quad \cdot \left[\left((n+1)^2\lambda^2 + (5n+3)\lambda + 4\right)(1+\lambda)^n - \left((2n+1)\lambda + 3\right)\right] \end{aligned}$$

Eine Induktion liefert die Behauptung. □

Jetzt lässt sich das Grenzverhalten der Kovarianz der Generationsnummern zweier zufällig gezogener Sequenzen für eine gegen ∞ strebende Anzahl von Startmolekülen bestimmen.

Lemma 2.21 *Gegeben seien eine n-stufige PCR und zwei $\mathcal{Z}_n$-wertige, Laplace-verteilte Zufallsvektoren A, B. Dann gilt:*

$$\lim_{j\to\infty} jCov_j\left(g(A), g(B)\right) = \frac{1-\lambda}{(1+\lambda)^3}\left(2 - \frac{n\lambda+2}{(1+\lambda)^n}\right)$$

Beweis: Die Folge $\left((T_n(i), Z_n(i))^t\right)_{1\le i\le j}$ ist eine Folge stochastisch unabhängiger, identisch verteilter Zufallsvektoren mit Erwartungswert $(\xi, \upsilon)^t = \left(n\lambda(1+\lambda)^{n-1}, (1+\lambda)^n\right)^t$ nach (2.9) und (2.26). Die Einträge der Kovarianzmatrix $\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ sind in Lemma 2.20 (i), (ii), (iii) zu finden. Des Weiteren sind für alle $i \ge 1$ die $Z_n(i) \ge 1$, die $T_n(i)$ beschränkt und $EZ_n(1) = (1+\lambda)^n > 0$. Das Hilfslemma 2.19 (ii) lässt sich demnach anwenden und man erhält:

$$\begin{aligned} \lim_{j\to\infty} jVar_j\frac{\bar{T}_n}{\bar{Z}_n} &= \frac{\sigma_1^2}{\upsilon^2} - \frac{2\xi}{\upsilon^3}\rho\sigma_1\sigma_2 + \sigma_2^2\frac{\xi^2}{\upsilon^4} \\ &= \frac{(1-\lambda)(1+\lambda)^{n-3}\left[(1+\lambda)^n((n\lambda)^2 + 2n\lambda + 2) - n\lambda(n\lambda+3) - 2\right]}{(1+\lambda)^{2n}} \\ &\quad -\frac{2n\lambda(1+\lambda)^{n-1}}{(1+\lambda)^{3n}}(1-\lambda)(1+\lambda)^{n-2}(n\lambda+1)\left((1+\lambda)^n - 1\right) \\ &\quad +(1-\lambda)(1+\lambda)^{n-1}\left((1+\lambda)^n - 1\right)\frac{\left(n\lambda(1+\lambda)^{n-1}\right)^2}{(1+\lambda)^{4n}} \\ &= \frac{1-\lambda}{(1+\lambda)^{n+3}}\big[\left((1+\lambda)^n(n^2\lambda^2 + 2n\lambda + 2) - n\lambda(n\lambda+3) - 2\right) \\ &\qquad -2n\lambda(n\lambda+1)\left((1+\lambda)^n - 1\right) \\ &\qquad +\left((1+\lambda)^n - 1\right)n^2\lambda^2\big] \\ &= \frac{1-\lambda}{(1+\lambda)^{n+3}}\big[(1+\lambda)^n\left(n^2\lambda^2 + 2n\lambda + 2 - 2n^2\lambda^2 - 2n\lambda + n^2\lambda^2\right) \\ &\qquad -n^2\lambda^2 - 3n\lambda - 2 + 2n^2\lambda^2 + 2n\lambda - n^2\lambda^2\big] \\ &= \frac{1-\lambda}{(1+\lambda)^3}\left(2 - \frac{n\lambda+2}{(1+\lambda)^n}\right) \end{aligned}$$

Nach Gleichung (2.24) ist $Cov_j\left(g(A), g(B)\right)$ gleich $Var_j \frac{\bar{T}_n}{\bar{Z}_n}$, also gilt die Behauptung. □

Dieses Ergebnis liefert zusammen mit den Lemmata 2.16 und 2.18 das Grenzverhalten der Kovarianz der Mutationen zweier zufällig gezogener Sequenzen nach n PCR-Zyklen.

Um in Gleichung (2.15) die Varianz der Summe der Mutationen von s zufällig gezogenen Sequenzen abschätzen zu können, muss nun die Varianz der Mutation einer Sequenz betrachtet werden. Da $M \sim \sum_{i=1}^{K} X_i$ mit - auch von K - stochastisch unabhängigen, identisch $Poi(\mu G)$-verteilten Zufallsgrößen X_i, folgt für die Varianz unter jeder Anfangsverteilung P_j:

$$\begin{aligned} Var_j M &= Var_j \sum_{i=1}^{K} X_i = E_j K Var_j X_1 + Var_j K (E_j X_1)^2 \\ &= \mu G E_j K + (\mu G)^2 Var_j K \end{aligned} \tag{2.28}$$

Das nächste Lemma gibt an, wie stark der Erwartungswert von K bzw. dessen Varianz von der in der Voraussetzung V.1 geforderten $\mathcal{B}\left(n, \frac{\lambda}{1+\lambda}\right)$-Verteilung für eine gegen ∞ strebende Anzahl an Urahnen abweicht.

Lemma 2.22 *Gegeben sei eine n-stufige PCR. K bezeichne die Generationsnummer einer zufällig gezogenen Sequenz. Dann gilt:*
(i)

$$\lim_{j\to\infty} j\left(E_j K - \frac{n\lambda}{1+\lambda}\right) = -\frac{1-\lambda}{(1+\lambda)^2}\left(1-(1+\lambda)^{-n}\right)$$

(ii)

$$\lim_{j\to\infty} j\left(Var_j K - \frac{n\lambda}{(1+\lambda)^2}\right) = -\frac{(1-\lambda)^2}{(1+\lambda)^3}\left(1-(1+\lambda)^{-n}\right)$$

Beweis:
(i) Es sei $Z_0 = j$ f.s. und $j \in \mathbb{N}$ beliebig. Gleichung (2.8) liefert:

$$E_j K = \sum_{k=0}^{n} k E_j \frac{X_k^n}{Z_n} = E_j \frac{\sum_{k=0}^{n} k X_k^n}{Z_n} = E_j \frac{T_n}{Z_n} = E_j \frac{\sum_{i=1}^{j} T_n(i)}{j} \frac{j}{\sum_{i=0}^{j} Z_n(i)} = E_j \frac{\bar{T}_n}{\bar{Z}_n}$$

Die $T_n(i)$ sowie die $Z_n(i)$, $1 \le i \le j$, sind jeweils stochastisch unabhängig und identisch verteilt. Außerdem sind für alle $i \ge 1$ die $Z_n(i) \ge 1$, die $T_n(i)$ beschränkt und $EZ_n(1) = (1+\lambda)^n > 0$. Mit $\xi = ET_n(1)$, $\upsilon = EZ_n(1)$, $\sigma_2^2 = VarZ_n(1)$ und $\rho\sigma_1\sigma_2 = Cov\left(T_n(1), Z_n(1)\right)$, liefert das Lemma 2.19 (i) unter Verwendung der in Lemma 2.20 (i) bis (iii) bewiesenen Formeln und Gleichung (2.26):

$$\begin{aligned} \lim_{j\to\infty} j\left(E_j \frac{\bar{T}_n}{\bar{Z}_n} - \frac{\xi}{\upsilon}\right) &= \lim_{j\to\infty} j\left(E_j \frac{\bar{T}_n}{\bar{Z}_n} - \frac{n\lambda}{1+\lambda}\right) \\ &= \frac{\xi}{\upsilon^3}\sigma_2^2 - \frac{1}{\upsilon^2}\rho\sigma_1\sigma_2 \\ &= \frac{n\lambda(1+\lambda)^{n-1}}{(1+\lambda)^{3n}}(1-\lambda)(1+\lambda)^{n-1}\left((1+\lambda)^n - 1\right) \\ &\quad - \frac{1}{(1+\lambda)^{2n}}(1-\lambda)(1+\lambda)^{n-2}(n\lambda+1)\left((1+\lambda)^n - 1\right) \\ &= \frac{1-\lambda}{(1+\lambda)^2}\left[n\lambda\left(1-(1+\lambda)^{-n}\right) - (n\lambda+1)\left(1-(1+\lambda)^{-n}\right)\right] \\ &= -\frac{1-\lambda}{(1+\lambda)^2}\left(1-(1+\lambda)^{-n}\right) \end{aligned}$$

(ii) Wieder sei $Z_0 = j$ f.s. und $j \in \mathbb{N}$ beliebig. Da $Var_j K = E_j K^2 - (E_j K)^2$ muss der Erwartungswert von K^2 berechnet werden. Dabei ist $R_n(i) := \sum_{k=0}^{n} k^2 X_k^n(i)$ für alle $1 \leq i \leq j$, und $\bar{R}_n := \frac{1}{j} \sum_{i=1}^{j} R_n(i)$:

$$E_j K^2 = \sum_{k=0}^{n} k^2 \underbrace{P_j(K=k)}_{\stackrel{(2.8)}{=} E_j \frac{X_k^n}{Z_n}} = E_j \frac{\sum_{i=1}^{j} \sum_{k=0}^{n} k^2 X_k^n(i)}{j} \frac{j}{\sum_{i=1}^{j} Z_n(i)} = E_j \frac{\bar{R}_n}{\bar{Z}_n}$$

Wieder soll zur Abschätzung des Grenzverhaltens des Erwartungswertes Lemma 2.19 (i) benutzt werden. Für alle $1 \leq i \leq j$ sind die $Z_n(i) \geq 1$, $EZ_n(1) = (1+\lambda)^n > 0$ sowie die $R_n(i)$ stochastisch unabhängig, identisch verteilt und beschränkt. $\upsilon = EZ_n$, $\rho\sigma_1\sigma_2 = Cov\,(R_n, Z_n)$ und $\sigma_2^2 = VarZ_n$ sind in den Lemmata 2.10 und 2.20 zu finden. Nur der Erwartungswert von R_n muss an dieser Stelle noch berechnet werden. Dies geschieht mit Hilfe von Lemma 2.11:

$$\begin{aligned}
\xi &= ER_n \\
&= E \sum_{k=0}^{n} k^2 X_k^n \\
&= \sum_{k=0}^{n} k^2 \binom{n}{k} \lambda^k \\
&= \sum_{k=0}^{n} k(k-1) \binom{n}{k} \lambda^k + \sum_{k=0}^{n} k \binom{n}{k} \lambda^k \\
&= n(n-1)\lambda^2 (1+\lambda)^{n-2} + n\lambda(1+\lambda)^{n-1} \\
&= n\lambda(n\lambda+1)(1+\lambda)^{n-2}
\end{aligned}$$

Nun lässt sich Lemma 2.19 (i) anwenden:

$$\begin{aligned}
\lim_{j\to\infty} \left(E_j K^2 - \frac{\xi}{\upsilon} \right) &= \lim_{j\to\infty} \left(E_j \frac{\bar{R}_n}{\bar{Z}_n} - \frac{\xi}{\upsilon} \right) \\
&= \lim_{j\to\infty} \left(E_j \frac{\bar{R}_n}{\bar{Z}_n} - \frac{n\lambda(n\lambda+1)}{(1+\lambda)^2} \right) \\
&= \frac{\xi}{\upsilon^3} \sigma_2^2 - \frac{1}{\upsilon^2} \rho\sigma_1\sigma_2 \\
&= \frac{n\lambda(n\lambda+1)(1+\lambda)^{n-2}}{(1+\lambda)^{3n}} (1-\lambda)(1+\lambda)^{n-1} \left((1+\lambda)^n - 1\right) \\
&\quad - \frac{1}{(1+\lambda)^{2n}} (1-\lambda)(1+\lambda)^{n-3} \left((n\lambda)^2 + (3n-1)\lambda + 1\right) \left((1+\lambda)^n - 1\right) \\
&= \frac{1-\lambda}{(1+\lambda)^{n+3}} \left((1+\lambda)^n - 1\right) \left[n\lambda(n\lambda+1) - \left((n\lambda)^2 + (3n-1)\lambda + 1\right) \right] \\
&= \frac{1-\lambda}{(1+\lambda)^3} \left(1 - (1+\lambda)^{-n}\right) \left((1-2n)\lambda - 1\right)
\end{aligned}$$

Da K asymptotisch $\mathcal{B}\left(n, \frac{\lambda}{1+\lambda}\right)$-verteilt ist, gilt außerdem:

$$\begin{aligned}
j \left((E_j K)^2 - \left(\frac{n\lambda}{1+\lambda} \right)^2 \right) &= j \underbrace{\left(E_j K - \frac{n\lambda}{1+\lambda} \right)}_{\text{siehe (i)}} \underbrace{\left(E_j K + \frac{n\lambda}{1+\lambda} \right)}_{\xrightarrow[j\to\infty]{} \frac{2n\lambda}{1+\lambda}} \\
&\xrightarrow[j\to\infty]{} -\frac{2n\lambda(1-\lambda)}{(1+\lambda)^3} \left(1 - (1+\lambda)^{-n}\right)
\end{aligned}$$

Dies impliziert:

$$\begin{aligned}\lim_{j\to\infty} j\left(Var_j K - \frac{n\lambda}{(1+\lambda)^2}\right) &= \lim_{j\to\infty} j\left(E_j K^2 - \frac{n\lambda(n\lambda+1)}{(1+\lambda)^2}\right) \\ &\quad - \lim_{j\to\infty} j\left((E_j K)^2 - \left(\frac{n\lambda}{1+\lambda}\right)^2\right) \\ &= \frac{1-\lambda}{(1+\lambda)^3}\left(1-(1+\lambda)^{-n}\right)((1-2n)\lambda - 1) \\ &\quad + \frac{2n\lambda(1-\lambda)}{(1+\lambda)^3}\left(1-(1+\lambda)^{-n}\right) \\ &= -\frac{(1-\lambda)^2}{(1+\lambda)^3}\left(1-(1+\lambda)^{-n}\right)\end{aligned}$$

Damit ist das Lemma bewiesen. □

Jetzt stehen alle Hilfsmittel bereit, um eine asymptotische Aussage für die Varianz des Schätzers der Mutationsrate zu bestimmen.

Satz 2.23 *Gegeben seien eine n-stufige PCR und Zufallsvektoren $A_1, \ldots, A_s$, $s \in \mathbb{N}$, derart, dass A_1 $\mathcal{Z}_n$-wertig, Laplace-verteilt ist, A_2 $\mathcal{Z}_n - \{A_1\}$-wertig, Laplace-verteilt, usw. Diese stehen also für s zufällig gezogene Sequenzen ohne Zurücklegen nach eine n-stufigen PCR. Für $\lambda = 1$ und jede beliebige Anfangsverteilung P_j gilt*

$$Var_j\left(\sum_{i=1}^{s} M(A_i)\right) = \frac{sn\mu G}{4}(2+\mu G) + \binom{s}{2}\frac{\mu G}{j}\left(1-\frac{1}{2^n}\right). \tag{2.29}$$

Liegt die Effizienz echt zwischen 0 und 1, d.h. $\lambda \in (0,1)$, gilt

$$\lim_{j\to\infty} j\left(Var_j\left(\sum_{i=1}^{s} M(A_i)\right) - \frac{sn\lambda\mu G}{(1+\lambda)^2}(\mu G + 1 + \lambda)\right) = sC_1 + 2\binom{s}{2}C_2, \tag{2.30}$$

wobei

$$\begin{aligned}C_1 &= -\mu G\frac{1-\lambda}{(1+\lambda)^2}\left(1-(1+\lambda)^{-n}\right)\left(1+\mu G\frac{1-\lambda}{1+\lambda}\right) \\ C_2 &= \mu G\left(\frac{2}{(1+\lambda)^2} - \frac{2+n\lambda(1-\lambda)}{(1+\lambda)^{n+2}} + \frac{\mu G(1-\lambda)}{(1+\lambda)^3}\left(2-\frac{n\lambda+2}{(1+\lambda)^n}\right)\right).\end{aligned}$$

Beweis: Wir beweisen zuerst Gleichung (2.30). Dabei seien die $A_1, \ldots, A_s$ vorerst alle $\mathcal{Z}_n$-wertig und Laplace-verteilt. Das Ziehen erfolgt also mit Zurücklegen. Mit Gleichung (2.28) und Lemma 2.16 gilt dann für jede Anfangsverteilung P_j und $\lambda \in (0,1)$:

$$j\left(Var\left(\sum_{i=1}^{s} M(A_i)\right) - \frac{sn\lambda\mu G}{(1+\lambda)^2}(\mu G + 1 + \lambda)\right)$$

$$
\begin{aligned}
&\overset{(2.15)}{=} && j\left(sVar_jM(A_1)+2\binom{s}{2}Cov_j\left(M(A_1),M(A_2)\right)-\frac{sn\lambda\mu G}{(1+\lambda)^2}(\mu G+1+\lambda)\right)\\
&= && j\Bigg(s\left(\mu GE_jK+(\mu G)^2Var_jK\right)\\
&&&\quad +2\binom{s}{2}\left(\mu GE_j\bar{A}_n+(\mu G)^2Cov_j\left(g(A_1),g(A_2)\right)\right)\\
&&&\quad -\frac{sn\lambda\mu G}{(1+\lambda)^2}(\mu G+1+\lambda)\Bigg)\\
&= && s\mu G\underbrace{j\left(E_jK-\frac{n\lambda}{1+\lambda}\right)}_{\text{siehe Lemma 2.22 (i)}}+s(\mu G)^2\underbrace{j\left(Var_jK-\frac{n\lambda}{(1+\lambda)^2}\right)}_{\text{siehe Lemma 2.22 (ii)}}\\
&&&\quad +2\binom{s}{2}\mu G\underbrace{jE_jA_n}_{\text{siehe Lemma 2.18}}+2\binom{s}{2}(\mu G)^2\underbrace{jCov_j\left(g(A_1),g(A_2)\right)}_{\text{siehe Lemma 2.21}}\\
&\xrightarrow[j\to\infty]{} && -s\mu G\frac{1-\lambda}{(1+\lambda)^2}\left(1-(1+\lambda)^{-n}\right)-s(\mu G)^2\frac{(1-\lambda)^2}{(1+\lambda)^3}\left(1-(1+\lambda)^{-n}\right)\\
&&&\quad +2\binom{s}{2}\mu G\left(\frac{2}{(1+\lambda)^2}-\frac{2+n\lambda(1-\lambda)}{(1+\lambda)^{n+2}}\right)\\
&&&\quad +2\binom{s}{2}(\mu G)^2\frac{1-\lambda}{(1+\lambda)^3}\left(2-\frac{n\lambda+2}{(1+\lambda)^n}\right)\\
&= && s\Bigg(\underbrace{-\mu G\frac{1-\lambda}{(1+\lambda)^2}\left(1-(1+\lambda)^{-n}\right)\left(1+\mu G\frac{1-\lambda}{1+\lambda}\right)}_{=:C_1}\Bigg)\\
&&&\quad +2\binom{s}{2}\underbrace{\mu G\left(\frac{2}{(1+\lambda)^2}-\frac{2+n\lambda(1-\lambda)}{(1+\lambda)^{n+2}}+\frac{\mu G(1-\lambda)}{(1+\lambda)^3}\left(2-\frac{n\lambda+2}{(1+\lambda)^n}\right)\right)}_{=:C_2}
\end{aligned}
$$

Das Ziehen ohne Zurücklegen wird beim Grenzübergang $j\to\infty$ durch das Ziehen mit Zurücklegen approximiert. Deshalb ist Gleichung (2.30) auch im Falle der Auswahl von s Sequenzen ohne Zurücklegen, wie es in den Voraussetzungen des Satzes gefordert wird, richtig.
Ist $\lambda=1$, so sind Z_n, T_n und alle X_k^n für $k\geq 0$ konstant. Die Gleichung (2.24) liefert in diesem Fall $Cov_j\left(g(A_1),g(A_2)\right)=0$. Des Weiteren ist die Generationsnummer einer zufällig gezogenen Sequenz K dann $\mathcal{B}\left(n,\frac{1}{2}\right)$-verteilt und das Lemma 2.18 ist ohne Grenzwertbetrachtung richtig. Damit gilt für jede Anfangsverteilung P_j

$$
\begin{aligned}
Var_j\left(\sum_{i=1}^{s}M(A_i)\right) &= sVar_jM(A_1)+2\binom{s}{2}Cov_j\left(M(A_1),M(A_2)\right)\\
&= s\big(\mu G\underbrace{E_jK}_{=\frac{n}{2}}+(\mu G)^2\underbrace{Var_jK}_{=\frac{n}{4}}\big)\\
&\quad +2\binom{s}{2}\big(\mu G\underbrace{E_j\bar{A}_n}_{=\frac{1}{j}\left(\frac{2}{4}-\frac{2}{2^{n+2}}\right)}+(\mu G)^2\underbrace{Cov_j\left(g(A_1),g(A_2)\right)}_{=0}\big)\\
&= \frac{sn\mu G}{4}(2+\mu G)+\binom{s}{2}\frac{\mu G}{j}\left(1-\frac{1}{2^n}\right)
\end{aligned}
$$

Dies ist Gleichung (2.29) □

Bemerkung 2.24 Die in Satz 2.23 auftauchenden Konstanten C_1 und C_2 sind durch eine nur von der Anzahl der PCR-Zyklen n abhängige Konstante beschränkt. Unter den in diesem Satz geltenden Voraussetzungen gilt dann:

$$\lim_{j\to\infty} Var_j\left(\sum_{i=1}^{s} M(A_i)\right) = \frac{sn\lambda\mu G}{(1+\lambda)^2}(\mu G + 1 + \lambda).$$

Damit gilt für unseren Schätzer $\hat{\mu}$ genauer:

Korollar 2.25 *Gegeben seien eine PCR mit n Zyklen und $\lambda \in (0,1)$, Zufallsvektoren $A_1, \ldots, A_s$, $s \in \mathbb{N}$, derart, dass A_1 $\mathcal{Z}_n$-wertig, Laplace-verteilt ist, A_2 $\mathcal{Z}_n - \{A_1\}$-wertig, Laplace-verteilt, usw. Für den durch Gleichung (2.14) definierten Schätzer der Mutationsrate gilt dann:*

$$\lim_{j\to\infty} Var_j\, \hat{\mu}\,(M(A_1), \ldots, M(A_s)) = \frac{\mu(\mu G + 1 + \lambda)}{n\lambda G s}$$

Beweis: Da $Var_j\, \hat{\mu}\,(M(A_1), \ldots, M(A_s)) = \left(\frac{1+\lambda}{n\lambda G s}\right)^2 Var_j\,(\sum_{i=1}^{s} M(A_i))$, folgt die Behauptung direkt aus Satz 2.23 und Bemerkung 2.24. □

Beispiel 2.26 In einem von Saiki et al. [26] erhobenen Datensatz traten bei 28 Sequenzen, die nach einer 30-stufigen PCR gewonnen wurden, insgesamt 17 falsch eingebaute Basen auf. Die Länge einer jeden Sequenz betrug 239 Basen. Die Effizienz dieser PCR wird mit $\lambda = 0,85$ angegeben. Die Mutationsrate kann nun mit Hilfe des Schätzers $\hat{\mu}$ errechnet werden:

$$\hat{\mu}\,((m_1, \ldots, m_s)) = \frac{(1+\lambda)\sum_{i=1}^{s} m_i}{n\lambda G s} = \frac{(1+0,85)\cdot 17}{30 \cdot 0,85 \cdot 239 \cdot 28} \approx 1,84 \cdot 10^{-4}$$

Die Standardabweichung des Schätzers σ kann durch

$$\sigma = \sqrt{Var\hat{\mu}} = \sqrt{\frac{\mu(\mu G + 1 + \lambda)}{n\lambda G s}} = \sqrt{\frac{1,84 \cdot 10^{-4}(1,84 \cdot 10^{-4} \cdot 239 + 1,85)}{30 \cdot 0,85 \cdot 239 \cdot 28}} \approx 4,5 \cdot 10^{-5}$$

approximiert werden.

2.4. Die paarweise Distanz und der Hamming-Abstand

Bis jetzt sind wir davon ausgegangen, dass die genaue Abfolge der Basen im Target bekannt ist und so die Anzahl der Mutationen durch den Vergleich mit der ursprünglichen Sequenz bestimmt werden kann. In vielen Fällen sind aber vom Target nur der Anfang und das Ende, an die die Primer angelagert werden, bekannt, aber die Basenabfolge im dazwischen liegenden Stück nicht. Die Anzahl der Mutationen einer zufällig gezogenen Sequenz nach n PCR-Schritten lässt sich nicht wie im vorigen Kapitel beschrieben ermitteln und die Mutationsrate kann nicht mit dem Schätzer $\hat{\mu}$ geschätzt werden. Sind aber zwei Sequenzen gegeben, können die Unterschiede in der Basenabfolge durch Vergleich der Sequenzen bestimmt werden. Der nachfolgende Abschnitt beschreibt die Verteilung dieser Unterschiede. Es ist offensichtlich, dass die Wahrscheinlichkeit von Unterschieden in der Basenabfolge mit der Anzahl der Replikationsschritte, die zwischen den beiden Molekülen liegen, größer wird.
Wir verzichten der Übersicht halber wieder auf die strikte Trennung zwischen den Erwartungswerten, usw. unter den verschiedenen Verteilungen P_j. Es sei noch einmal an die mathematisch korrekte Interpretation der Sprechweise „zwei zufällig gezogene Sequenzen“ erinnert. Es handelt sich hierbei um zwei auf $\mathcal{Z}_n$ Laplace-verteilte Zufallsvektoren A und B. Der Zusatz „ohne Zurücklegen“ bedeutet, dass A auf $\mathcal{Z}_n$ und B auf $\mathcal{Z}_n - \{A\}$ Laplace-verteilt ist.
Beginnen wir mit der Definition der Anzahl der Zwischenschritte zwischen zwei Sequenzen.

Definition 2.27 Gegeben seien zwei Sequenzen $\alpha, \beta \in \mathcal{Z}_n$, die einer n-stufigen PCR entstammen, γ sei ihr MRCA. Dann heißt

$$d(\alpha, \beta) := (g(\alpha) - g(\gamma)) + (g(\beta) - g(\gamma))$$

der *Abstand zwischen α und β* oder auch die *Distanz zwischen α und β*.
Die Zufallsgröße D sei der *Abstand zweier zufällig ohne Zurücklegen gezogener Sequenzen* kurz die *paarweise Distanz* einer PCR mit n Zyklen, d.h. für A, B wie oben gilt: $D = d(A, B)$.

Der Abstand zwischen zwei Sequenzen gibt also die Anzahl der Sequenzen an, in denen Mutationen auftreten können. Als Beispiel sei auf die Abbildung 2.1 verwiesen. Dort haben die Sequenzen (2,1) und (1,3) den Abstand $d((2,1),(1,3)) = 3$. Stammen die Sequenzen von unterschiedlichen Urahnen ab, so liefert die Definition 2.27 in Übereinstimmung mir der Definition 2.15 die Summe der Anzahl der Vorfahren der einzelnen Sequenzen.

Im nachfolgenden Lemma wird eine rekursive Formel zur Ermittlung der erwarteten Anzahl an Paaren mit vorgegebenem Abstand bereitgestellt.

Lemma 2.28 *$P_n(k)$ sei die erwartete Anzahl von Paaren mit einem Abstand $k \in \mathbb{N}$ (d.h. Ziehen ohne Zurücklegen) nach einer PCR mit $n \geq 1$ Zyklen und einem Startmolekül, also $Z_0 = 1$. Dann gelten folgende Aussagen:*
(i)

$$P_n(1) = (1 + \lambda)^n - 1 \tag{2.31}$$

(ii)

$$P_{n+1}(k) = P_n(k) + 2\lambda P_n(k-1) + \lambda^2 P_n(k-2) \quad \textit{für alle } 2 \leq k \leq 2n+1$$

wobei $P_n(k) = 0$ *falls* $k = 0$ *oder* $k \geq 2n$.

(iii) Die erzeugende Funktion von $(P_n(k))_{k\geq 0}$ *ist*

$$\varphi_{P_n}(s) = \lambda s \frac{(1+\lambda s)^{2n} - (1+\lambda)^n}{(1+\lambda s)^2 - (1+\lambda)} \,. \tag{2.32}$$

Beweis: Die (zufällige) Anzahl der Paare mit Abstand k nach n PCR-Zyklen sei mit $N_n(k)$ bezeichnet.
(i) Dieses Ergebnis wird über eine Induktion nach n bewiesen. Da nur bei erfolgreicher Duplizierung des Startmoleküls im ersten Schritt ein Paar mit Abstand 1 entsteht, gilt für die erwartete Anzahl an Paaren mit dem Abstand 1 nach einem PCR-Schritt $P_1(1) = \lambda = 1 + \lambda - 1$.
Gelte nun die Gleichung (2.31) für beliebiges aber festes n. Für ein Paar mit Distanz 1 nach dem $(n+1)$-ten Schritt gibt es dann zwei Möglichkeiten:
1. Fall: Das Paar hat schon im n-ten Zyklus den Abstand 1, dafür gibt es $N_n(1)$ Möglichkeiten.
2. Fall: Eine der beiden Sequenzen entsteht aus der anderen im $(n+1)$-ten Schritt. Da es insgesamt Z_n Sequenzen nach dem n-ten Zyklus gibt lässt sich die Anzahl dieser Paare durch eine Summe von stochastisch unabhängigen, identisch $\mathcal{B}(1,\lambda)$-verteilten und von Z_n unabhängigen I_j, $j \geq 1$ beschreiben, also durch $\sum_{j=1}^{Z_n} I_j$.
Insgesamt ist die Anzahl aller Paare mit dem Abstand 1 nach $(n+1)$ PCR-Zyklen gleich

$$N_{n+1}(1) = N_n(1) + \sum_{j=1}^{Z_n} I_j \,.$$

Dies impliziert:

$$\begin{aligned}
P_{n+1}(1) &= EN_{n+1}(1) \\
&= EN_n(1) + E\sum_{j=1}^{Z_n} I_j \\
&\overset{\text{I.V.}}{=} (1+\lambda)^n - 1 + E\sum_{j=1}^{Z_n} I_j \\
&= (1+\lambda)^n - 1 + EZ_nEI_1 \\
&= (1+\lambda)^n - 1 + (1+\lambda)^n\lambda \\
&= (1+\lambda)(1+\lambda)^n - 1 \\
&= (1+\lambda)^{n+1} - 1
\end{aligned}$$

Dies ist die Behauptung.

(ii) Ein Paar mit einem Abstand $k > 1$ nach $(n+1)$ PCR-Zyklen muss schon nach n Zyklen verschiedene Vorfahren haben. Es seien also α und β die Vorfahren im n-ten Schritt eines solchen Paares. Es gibt 4 Möglichkeiten, so dass das Paar im nächsten Zyklus den Abstand k besitzt (siehe dazu auch Abbildung 2.2):

- Bei dem Paar handelt es sich um α, β. Dann muss für ihren Abstand $d(\alpha,\beta) = k$ gelten. Es werden $P_n(k)$ solcher Paare erwartet.
- Das Paar ist α', β mit $d(\alpha,\beta) = k-1$. Davon gibt es $\lambda P_n(k-1)$ erwartete Paare.

- Das gleiche gilt für das Paar α, β'.
- Es liegt das Paar α', β' vor. Der Abstand von α und β muss dann $d(\alpha,\beta) = k-2$ betragen. Davon gibt es $\lambda^2 P_n(k-2)$ erwartete Paare.

Dies impliziert:

$$P_{n+1}(k) = P_n(k) + 2\lambda P_n(k-1) + \lambda^2 P_n(k-2).$$

(iii) Da die erwartete Anzahl an Paaren mit einer Distanz von 0 bzw. mit einer Distanz, die mehr als doppelt so groß wie die Anzahl an PCR-Schritten ist, verschwindet, gilt für die erzeugende Funktion von $P_{n+1}(k)$ mit den schon bewiesenen Teilen (i) und (ii):

$$\begin{aligned}
\varphi_{P_{n+1}}(s) &= \sum_{k=1}^{2n+1} P_{n+1}(k)s^k \\
&= P_{n+1}(1)s + \sum_{k=2}^{2n+1} P_{n+1}(k)s^k \\
&= \left((1+\lambda)^{n+1}-1\right)s + \sum_{k=2}^{2n+1}\left(P_n(k) + 2\lambda P_n(k-1) + \lambda^2 P_n(k-2)\right)s^k \\
&= \left((1+\lambda)^{n+1}-1\right)s + \underbrace{\sum_{k=2}^{2n-1} P_n(k)s^k}_{=\varphi_{P_n}(s)-((1+\lambda)^n-1)s} + 2\lambda s \underbrace{\sum_{k=1}^{2n-1} P_n(k)s^k}_{=\varphi_{P_n}(s)} \\
&\qquad + \lambda^2 s^2 \underbrace{\sum_{k=1}^{2n-1} P_n(k)s^k}_{=\varphi_{P_n}(s)} \\
&= \left((1+\lambda)^{n+1}-1\right)s - \left((1+\lambda)^n - 1\right)s + \varphi_{P_n}(s)\left(1 + 2\lambda s + \lambda^2 s^2\right) \\
&= \lambda(1+\lambda)^n s + (1+\lambda s)^2 \varphi_{P_n}(s)
\end{aligned}$$

Gleichung (2.32) wird nun mit Induktion bewiesen. Für $n=1$ ergibt sich $\varphi_{P_1}(s) = \lambda s = \lambda s \frac{(1+\lambda s)^{2*1}-(1+\lambda)}{(1+\lambda s)^2-(1+\lambda)}$. Die Behauptung gelte für beliebiges aber festes n. Dann folgt:

$$\begin{aligned}
\varphi_{P_{n+1}}(s) &= \lambda(1+\lambda)^n s + (1+\lambda s)^2 \varphi_{P_n}(s) \\
&= \lambda(1+\lambda)^n s + (1+\lambda s)^2 \lambda s \frac{(1+\lambda s)^{2n} - (1+\lambda)^n}{(1+\lambda s)^2 - (1+\lambda)} \\
&= \lambda s \frac{(1+\lambda)^n\left((1+\lambda s)^2 - (1+\lambda)\right) + (1+\lambda s)^{2(n+1)} - (1+\lambda s)^2(1+\lambda)^n}{(1+\lambda s)^2 - (1+\lambda)} \\
&= \lambda s \frac{(1+\lambda s)^{2(n+1)} - (1+\lambda)^{(n+1)}}{(1+\lambda s)^2 - (1+\lambda)}
\end{aligned}$$

Damit sind alle Aussagen des Lemmas bewiesen. □

Ähnlich wie in Kapitel 2.2 konvergiert die Verteilung des Abstandes D zweier zufällig ohne Zurücklegen gezogener Sequenzen für eine große Anzahl an Startmolekülen gegen eine einfacher handhabbare Verteilung. Dazu sei eine n-stufige PCR gegeben. Die erwartete Anzahl an Paaren mit einem Abstand k, von denen beide vom selben Urahnen abstammen, ist nach Lemma 2.28 gleich $Z_0 P_n(k)$. Um ein Paar zu erhalten, bei dem die Urahnen unterschiedlich sind, muss zuerst eine Sequenz mit dem Urahn i, $1 \le i \le Z_0$, und dann

eine Sequenz mit dem Urahn $j \neq i,\ 1 \leq j \leq Z_0$, gezogen werden. Stammen diese aus der Generation k_1 bzw. k_2, so muss die Summe $k_1 + k_2 = k$ sein. Damit erhält man für die erwartete Anzahl an Paaren mit dem Abstand k, die nicht vom selben Urahnen abstammen

$$\begin{aligned} E\left(\sum_{1\leq i<j\leq Z_0}\sum_{k_1+k_2=k} X^n_{k_1}(i)X^n_{k_2}(j)\right) &= \sum_{1\leq i<j\leq Z_0}\sum_{k_1+k_2=k} EX^n_{k_1}(i)X^n_{k_2}(j) \\ &= \sum_{1\leq i<j\leq Z_0}\sum_{k_1+k_2=k} EX^n_{k_1}(i)EX^n_{k_2}(j) \\ &= \binom{Z_0}{2}\sum_{k_1+k_2=k}\binom{n}{k_1}\lambda^{k_1}\binom{n}{k_2}\lambda^{k_2} \\ &= \binom{Z_0}{2}\lambda^k\binom{2n}{k}. \end{aligned}$$

Bei dieser Rechnung wird ausgenutzt, dass die Summen endlich und die $X^n_{k_1}(i),\ X^n_{k_2}(j)$ für alle $1 \leq i,j \leq Z_0,\ i \neq j$ stochastisch unabhängig sind. Die letzte Gleichheit ergibt sich aus der Gleichung $\sum_{k_1+k_2=k}\binom{n}{k_1}\binom{n}{k_2} = \binom{2n}{k}$, die in [11] Satz 7.7 zu finden ist.
Die Gesamtzahl der erwarteten Paare mit Distanz k ist somit

$$Z_0P_n(k) + \binom{Z_0}{2}\lambda^k\binom{2n}{k}. \tag{2.33}$$

Die Gesamtzahl aller erwarteten Paare lässt sich berechnen zu

$$\begin{aligned} E\binom{Z_n}{2} &= \sum_{k=0}^{2n}\left(Z_0P_n(k) + \binom{Z_0}{2}\lambda^k\binom{2n}{k}\right) \\ &= Z_0(1+\lambda)^{n-1}\left((1+\lambda)^n - 1\right) + \binom{Z_0}{2}(1+\lambda)^{2n}, \end{aligned} \tag{2.34}$$

wobei der erste Term aus Lemma 2.28 sowie einer Induktion und der zweite Term aus dem binomischen Lehrsatz folgt.

Voraussetzung (V.2) Die Verteilung des paarweisen Abstandes D zweier zufällig ohne Zurücklegen gezogener Sequenzen einer n-stufigen PCR mit Z_0 Urahnen genüge der Gleichung

$$P(D=k) = \frac{Z_0P_n(k) + \binom{Z_0}{2}\lambda^k\binom{2n}{k}}{E\binom{Z_n}{2}} \quad \text{für alle } k \geq 1,$$

und $P(D=0) = \frac{\binom{Z_0}{2}}{E\binom{Z_n}{2}}$.

Diese vereinfachende Voraussetzung ermöglicht es nun, die erzeugende Funktion und damit den Erwartungswert sowie die Varianz von D zu bestimmen.

Satz 2.29 *Gegeben eine PCR mit n Zyklen, gilt unter Voraussetzung V.2:*
(i) Die erzeugende Funktion von D ist

$$f_D(s) = \frac{1}{E\binom{Z_n}{2}}\left(Z_0\varphi_{P_n}(s) + \binom{Z_0}{2}(1+\lambda s)^{2n}\right).$$

(ii)

$$ED = \frac{2n\lambda}{1+\lambda} - \frac{2}{(1+\lambda)Z_0 + 1 - \lambda} + \mathcal{O}\left(\frac{1}{Z_0(1+\lambda)^n}\right)$$

(iii)

$$VarD = \frac{2n\lambda}{(1+\lambda)^2} - \frac{2(3+\lambda)}{(1+\lambda)^2 Z_0 + 1 - \lambda^2} - \frac{2}{\left((1+\lambda)Z_0 + 1 - \lambda\right)^2} + \mathcal{O}\left(\frac{1}{Z_0(1+\lambda)^n}\right)$$

Beweis:
(i) Die Wahrscheinlichkeit dafür, dass der Abstand zweier zufällig ohne Zurücklegen gezogener Sequenzen einer n-stufigen PCR mehr als $2n$ beträgt, ist 0. Für die erzeugende Funktion folgt deswegen

$$\begin{aligned} f_D(s) &= \sum_{k=0}^{2n} P(D=k)s^k \\ &\overset{\text{V.2}}{=} \frac{1}{E\binom{Z_n}{2}} \sum_{k=0}^{2n} \left(Z_0 s^k P_n(k) + \binom{Z_0}{2}\binom{2n}{k}(\lambda s)^k \right) \\ &= \frac{1}{E\binom{Z_n}{2}} \left(Z_0 \varphi_{P_n}(s) + \binom{Z_0}{2}(1+\lambda s)^{2n} \right) . \end{aligned}$$

Die letzte Gleichheit ergibt sich durch Anwendung des binomischen Lehrsatzes und der Definition der erzeugenden Funktion einer Zahlenfolge.
(ii) Wie schon im Beweis des Satzes 2.13 (ii) wird die Theorie der erzeugenden Funktionen zur Berechnung des Erwartungswertes verwendet. Zuerst ist die Ableitung der erzeugenden Funktion $\varphi_{P_n}(s)$ zu bestimmen:

$$\begin{aligned} \varphi'_{P_n}(s) &= \lambda \frac{(1+\lambda s)^{2n} - (1+\lambda)^n}{(1+\lambda s)^2 - (1+\lambda)} + \lambda s \left(\frac{2n\lambda(1+\lambda s)^{2n-1}\left((1+\lambda s)^2 - (1+\lambda)\right)}{\left((1+\lambda s)^2 - (1+\lambda)\right)^2} \right. \\ &\quad \left. - \frac{2\lambda\left((1+\lambda s)^{2n} - (1+\lambda)^n\right)(1+\lambda s)}{\left((1+\lambda s)^2 - (1+\lambda)\right)^2} \right) \end{aligned}$$

Mit $(1+\lambda)^2 - (1+\lambda) = \lambda(1+\lambda)$ folgt:

$$\begin{aligned} \varphi'_{P_n}(1) &= (1+\lambda)^{2n-1} - (1+\lambda)^{n-1} + \lambda \frac{2n\lambda^2(1+\lambda)^{2n} - 2\lambda\left((1+\lambda)^{2n} - (1+\lambda)^n\right)(1+\lambda)}{\lambda^2(1+\lambda)^2} \\ &= (1+\lambda)^{n-1} - (1+\lambda)^{2n-1} + 2n\lambda(1+\lambda)^{2n-2} \end{aligned}$$

Da $f'_D(s) = \frac{Z_0}{E\binom{Z_n}{2}}\varphi'_{P_n}(s) + \frac{\binom{Z_0}{2}}{E\binom{Z_n}{2}} 2n\lambda(1+\lambda s)^{2n-1}$, ergibt sich:

$$\begin{aligned} ED &= f'_D(1) \\ &= \frac{1}{E\binom{Z_n}{2}} \Big[Z_0(1+\lambda)^{n-1} - Z_0(1+\lambda)^{2n-1} \\ &\qquad + Z_0 2n\lambda(1+\lambda)^{2n-2} + \binom{Z_0}{2} 2n\lambda(1+\lambda s)^{2n-1} \Big] \\ &= \frac{1}{E\binom{Z_n}{2}} \Big[\frac{2n\lambda}{1+\lambda} \Big(\underbrace{Z_0(1+\lambda)^{2n-1} - Z_0(1+\lambda)^{n-1}}_{=Z_0(1+\lambda)^{n-1}((1+\lambda)^n - 1)} + \binom{Z_0}{2}(1+\lambda s)^{2n} \Big) \\ &\qquad + 2n\lambda Z_0(1+\lambda)^{n-2} + Z_0(1+\lambda)^{n-1} - Z_0(1+\lambda)^{2n-1} \Big] \\ &\overset{(2.34)}{=} \frac{2n\lambda}{1+\lambda} + \frac{2n\lambda Z_0(1+\lambda)^{n-2} + Z_0(1+\lambda)^{n-1}}{Z_0(1+\lambda)^{n-1}\left((1+\lambda)^n - 1\right) + \binom{Z_0}{2}(1+\lambda)^{2n}} \\ &\qquad - \frac{Z_0(1+\lambda)^{2n-1}}{Z_0(1+\lambda)^{n-1}\left((1+\lambda)^n - 1\right) + \binom{Z_0}{2}(1+\lambda)^{2n}} \end{aligned}$$

Nun gilt

$$\begin{aligned}\frac{2n\lambda Z_0(1+\lambda)^{n-2}+Z_0(1+\lambda)^{n-1}}{Z_0(1+\lambda)^{n-1}\left((1+\lambda)^n-1\right)+\binom{Z_0}{2}(1+\lambda)^{2n}} &= \frac{\frac{2n\lambda}{1+\lambda}+1}{((1+\lambda)^n-1)+\frac{Z_0-1}{2}(1+\lambda)^{n+1}} \\ &= \mathcal{O}\left(\frac{1}{Z_0(1+\lambda)^n}\right)\end{aligned}$$

sowie

$$\begin{aligned}\frac{Z_0(1+\lambda)^{2n-1}}{Z_0(1+\lambda)^{n-1}\left((1+\lambda)^n-1\right)+\binom{Z_0}{2}(1+\lambda)^{2n}} &= \frac{1}{1-\frac{1}{(1+\lambda)^n}+\frac{Z_0-1}{2}(1+\lambda)} \\ &= \frac{2}{2+(Z_0-1)(1+\lambda)}+\mathcal{O}\left(\frac{1}{(1+\lambda)^n}\right) \\ &= \frac{2}{Z_0(1+\lambda)+1-\lambda}+\mathcal{O}\left(\frac{1}{(1+\lambda)^n}\right).\end{aligned}$$

Zusammen folgt

$$ED=\frac{2n\lambda}{1+\lambda}-\frac{2}{Z_0(1+\lambda)+1-\lambda}+\mathcal{O}\left(\frac{1}{Z_0(1+\lambda)^n}\right),$$

also die Behauptung.
(iii) Die Identität $VarD = f_D''(1)+f_D'(1)-(f_D'(1))^2$ und eine zu Teil (ii) analoge Rechnung liefern die gewünschte Gleichung. □

Die Zufallsgröße D erlaubt nun, die Anzahl unterschiedlicher Basen zweier zufällig ohne zurücklegen gezogener Sequenzen zu beschreiben. D gibt an, wie viele Zyklen zwischen den beiden Sequenzen liegen. In jedem Zyklus kann eine zufällige Anzahl an Mutationen entstehen, die gemäß des Modells $Poi(\mu G)$-verteilt ist. Die Gesamtanzahl an Mutationen erhält einen besonderen Namen:

Definition 2.30 Gegeben seien eine n-stufige PCR und eine Folge $(X_i)_{i\geq 1}$ stochastisch unabhängiger, identisch $Poi(\mu G)$-verteilter sowie von D stochastisch unabhängiger Zufallsgrößen. Die Zufallsgröße

$$H := \sum_{i=1}^{D} X_i$$

heißt *Hamming-Abstand* zweier zufällig ohne Zurücklegen gezogener Sequenzen.

Ähnlich wie für die Anzahl an Mutationen einer zufällig gezogenen Sequenz in Satz 2.13, lässt sich nun ein Satz formulieren, der für den Hamming-Abstand den Erwartungswert und die Varianz, sowie eine Normal- und eine Poisson-Approximation bereitstellt.

Satz 2.31 *Gegeben sei eine PCR mit n Zyklen. Unter Voraussetzung V.2 gilt:*
(i) Die erzeugende Funktion des Hamming-Abstandes H ist

$$f_H(s) = f_D\left(\exp(\mu G(s-1))\right).$$

(ii) Der Erwartungswert und die Varianz von H genügen den Gleichungen

$$EH = \mu GED \qquad VarH = \mu GED + (\mu G)^2 VarD\,.$$

(iii) Die Zufallsgröße $\hat{H} := \frac{(1+\lambda)H - 2n\lambda\mu G}{\sqrt{2n\lambda\mu G(\mu G+1+\lambda)}}$ *konvergiert in Verteilung gegen die Standardnormalverteilung, es gilt also*

$$\lim_{n\to\infty} P\left(\frac{(1+\lambda)H - 2n\lambda\mu G}{\sqrt{2n\lambda\mu G(\mu G+1+\lambda)}} \leq x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds \quad \text{für alle } x \in \mathbb{R}.$$

(iv) Hängen μ *und* G *von* n *ab, mit* μ_n, G_n *bezeichnet, und zwar so, dass* $\lim_{n\to\infty} n\mu_n G_n = \nu \in \mathbb{R}$, *gilt*

$$\lim_{n\to\infty} P(H = h) = Poi\left(\frac{2\lambda\nu}{1+\lambda}\right)(\{h\}) \quad \textit{für alle } h \in \mathbb{N}_0.$$

Beweis:
(i) Da $H = \sum_{i=1}^{D} X_i$ mit stochastisch unabhängigen, identisch $Poi(\mu G)$-verteilten Zufallsgrößen X_i, gilt unter Verwendung des Satzes von der monotonen Konvergenz:

$$\begin{aligned}
f_H(s) &= Es^{\sum_{i=1}^{D} X_i} \\
&= E\sum_{k\geq 0} \mathbb{1}_{\{D=k\}} \prod_{i=1}^{k} s^{X_i} \\
&= \sum_{k\geq 0} E\mathbb{1}_{\{D=k\}} \prod_{i=1}^{k} Es^{X_i} \\
&= \sum_{k\geq 0} P(D=k)\left(Es^{X_1}\right)^k \\
&= E\exp(\mu G(s-1))^D \\
&= f_D\left(\exp(\mu G(s-1))\right)
\end{aligned}$$

Die vorletzte Gleichheit gilt, da die erzeugende Funktion einer $Poi(\mu G)$-verteilten Zufallsgröße die Gestalt $\exp(\mu G(s-1))$ besitzt.
(ii) Die erste Waldsche Gleichung liefert für den Erwartungswert des Hamming-Abstandes

$$EH = E\sum_{i=1}^{D} X_i = EDEX_1 = \mu GED\,.$$

Da X_1 $Poi(\mu G)$-verteilt ist, ergibt eine ähnliche Rechnung wie im Beweis des Lemmas 2.16

$$VarH = EDVarX_1 + (EX_1)^2\,VarD = \mu GED + (\mu G)^2 VarD\,.$$

(iii) Zum Beweis der Normalapproximation wird der Hamming-Abstand in zwei Teile aufgeteilt. Entweder besitzt ein zufällig ohne Zurücklegen gezogenes Paar denselben Urahn, oder die Sequenzen des Paares stammen von unterschiedlichen Urahnen ab. Der Hamming-Abstand im ersten Fall sei mit H_1 und im zweiten mit H_2 bezeichnet. Wie im ersten Teil bewiesen, lässt sich die erzeugende Funktion des Hamming-Abstandes durch die erzeugende Funktion des Abstandes ausdrücken. Deswegen wird zuerst dessen Verteilung untersucht. D_1 sei der Abstand zweier zufällig ohne Zurücklegen gezogener Sequenzen mit demselben Urahn und D_2 der Abstand zweier Sequenzen mit unterschiedlichen Startsequenzen. Mit dem Satz von der totalen Wahrscheinlichkeit gilt dann:

$$P(D=k) = c_n P(D_1 = k) + (1-c_n)P(D_2 = k)$$

wobei $c_n := \frac{Z_0(1+\lambda)^{n-1}((1+\lambda)^n-1))}{E\binom{Z_n}{2}}$ die Wahrscheinlichkeit, ein Paar mit demselben Urahnen zu erwischen, ist. Weiter gilt:

$$P(D_1 = k) = \frac{P_n(k)}{(1+\lambda)^n\left((1+\lambda)^n - 1\right)} \quad \text{für alle } 1 \le k \le 2n-1$$

und

$$P(D_2 = k) = \frac{\binom{Z_0}{2}\binom{2n}{k}\lambda^k}{\binom{Z_0}{2}(1+\lambda)^{2n}} = \frac{\binom{2n}{k}\lambda^k}{(1+\lambda)^{2n}} \quad \text{für alle } 0 \le k \le 2n.$$

Die erzeugenden Funktionen von D_1 bzw. D_2 sind demnach

$$\begin{aligned} f_{D_1}(s) &= \sum_{k=1}^{2n-1} s^k P(D_1 = k) = \frac{\varphi_{P_n}(s)}{(1+\lambda)^n\left((1+\lambda)^n - 1\right)} \\ f_{D_2}(s) &= \frac{1}{(1+\lambda)^{2n}} \sum_{k=0}^{2n} (\lambda s)^k \binom{2n}{k} \lambda^k = \left(\frac{1+\lambda s}{1+\lambda}\right)^{2n}. \end{aligned}$$

Nach (ii) ergibt sich somit für die erzeugenden Funktionen des aufgeteilten Hamming-Abstandes

$$f_{H_1}(s) = \frac{\varphi_{P_n}(\exp(\mu G(s-1)))}{(1+\lambda)^{n-1}\left((1+\lambda)^n - 1\right)} \tag{2.35}$$

$$f_{H_2}(s) = \left(\frac{1+\lambda\exp(\mu G(s-1))}{1+\lambda}\right)^{2n}. \tag{2.36}$$

Wie im Beweis des Satzes 2.13 (ii) stimmt dann die Verteilung von H_2 mit der Verteilung der Summe von stochastisch unabhängigen, identisch verteilten Zufallsgrößen $Y_1, \ldots, Y_{2n}$ überein, wobei Y_1 die gemischte Verteilung $\frac{1}{1+\lambda}\delta_0 + \frac{\lambda}{1+\lambda} Poi(\mu G)$ besitzt. Da $EY_1 = \frac{\lambda\mu G}{1+\lambda}$ und $VarY_1 = \frac{\lambda\mu G(\mu G+1+\lambda)}{(1+\lambda)^2}$, gilt vermöge des Satzes von Lindeberg

$$\hat{H}_2 := \frac{(1+\lambda)H_2 - 2n\lambda\mu G}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}} \sim \frac{(1+\lambda)\sum_{i=1}^{2n} Y_i - 2n\lambda\mu G}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}} \xrightarrow{d} \mathcal{N}(0,1)\,.$$

Da für die Fourier-Transformierte einer $\mathbb{N}_0$-wertigen Zufallsgröße X die Gleichheit $\phi_X(s) = f_X(\exp(is))$ gilt, kann nun auf die Ebene der Fourier-Transformierten gewechselt werden. Für die F.T. von $\hat{H}_2$ liefert der Stetigkeitssatz von Lévy (siehe [3] Satz 45.2)

$$\begin{aligned} \lim_{n\to\infty} \phi_{\hat{H}_2}(t) &= \lim_{n\to\infty} \phi_{H_2}\left(\frac{(1+\lambda)t}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}\right) \\ &\qquad \cdot \exp\left(-it\frac{2n\lambda\mu G}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}\right) \\ &= \phi_{\mathcal{N}(0,1)}(t)\,. \end{aligned} \tag{2.37}$$

Mit $\varepsilon := \exp\left(-it\frac{2n\lambda\mu G}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}\right)$, $\tau(t) := \exp\left(\mu G\left(\exp\left(i\frac{(1+\lambda)t}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}\right)-1\right)\right)$ gilt für die Fouriertransformierte von $\hat{H}_1 := \frac{(1+\lambda)H_1 - 2n\lambda\mu G}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}$:

$$\begin{aligned}
\phi_{\hat{H}_1}(t) &= \phi_{H_1}\left(\frac{(1+\lambda)t}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}\right)\varepsilon \\
&= f_{H_1}\left(\exp\left(i\frac{(1+\lambda)t}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}\right)\right)\varepsilon \\
&\overset{(2.35)}{=} \frac{\varphi_{P_n}(\tau(t))}{(1+\lambda)^{n-1}\left((1+\lambda)^n-1\right)}\varepsilon \\
&\overset{(2.32)}{=} \lambda\tau(t)\frac{(1+\lambda\tau(t))^{2n}-(1+\lambda)^n}{((1+\lambda\tau(t))^2-(1+\lambda))\,(1+\lambda)^{n-1}\left((1+\lambda)^n-1\right)}\varepsilon \\
&= \lambda\tau(t)\varepsilon\left(\frac{1+\lambda\tau(t)}{1+\lambda}\right)^{2n}\frac{1+\lambda}{((1+\lambda\tau(t))^2-(1+\lambda))\left(1-\frac{1}{(1+\lambda)^n}\right)} \\
&\quad -\lambda\tau(t)\varepsilon\frac{(1+\lambda)^n}{((1+\lambda\tau(t))^2-(1+\lambda))\,(1+\lambda)^{n-1}\left((1+\lambda)^n-1\right)}
\end{aligned}$$

Die Grenzwerte für $n \to \infty$ der in dieser Gleichung auftretenden Terme werden nun getrennt voneinander betrachtet:

$$\begin{aligned}
&\tau(t) = \exp\left(\mu G\left(\exp\left(\frac{it(1+\lambda)}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}\right)-1\right)\right) \xrightarrow[n\to\infty]{} 1 \\
&\varepsilon = \exp\left(-it\frac{2n\lambda\mu G}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}\right) \xrightarrow[n\to\infty]{} 0 \\
&\frac{1+\lambda}{((1+\lambda\tau(t))^2-(1+\lambda))\left(1-\frac{1}{(1+\lambda)^n}\right)} \xrightarrow[n\to\infty]{} \frac{1+\lambda}{(1+\lambda)^2-(1+\lambda)} = \frac{1}{\lambda} \\
&\frac{(1+\lambda)^n}{((1+\lambda\underbrace{\tau(t)}_{\xrightarrow[n\to\infty]{}1})^2-(1+\lambda))(1+\lambda)^{n-1}\left((1+\lambda)^n-1\right)} \xrightarrow[n\to\infty]{} 0
\end{aligned}$$

Zusätzlich gilt:

$$\begin{aligned}
\left(\frac{1+\lambda\tau(t)}{1+\lambda}\right)^{2n} &= \left(\frac{1+\lambda\exp\left(\mu G\left(\exp\left(i\frac{(1+\lambda)t}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}\right)-1\right)\right)}{1+\lambda}\right)^{2n} \\
&\overset{(2.36)}{=} f_{H_2}\left(\exp\left(i\frac{(1+\lambda)t}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}\right)\right) \\
&= \phi_{H_2}\left(\frac{(1+\lambda)t}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}\right)
\end{aligned}$$

Insgesamt ergibt sich für den Grenzwert der F.T. von $\hat{H}_1$

$$\begin{aligned}\lim_{n\to\infty}\phi_{\hat{H}_1}(t) &= \lim_{n\to\infty}\phi_{H_2}\left(\frac{(1+\lambda)t}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}\right)\\ &\quad\cdot\exp\left(-it\frac{2n\lambda\mu G}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}\right)\\ &\overset{(2.37)}{=} \phi_{\mathcal{N}(0,1)}(t)\,.\end{aligned}$$

Da

$$\begin{aligned}c_n &= \frac{Z_0(1+\lambda)^n\left((1+\lambda)^n-1\right)\right)}{E\binom{Z_n}{2}}\\ &\overset{(2.34)}{=} \frac{Z_0(1+\lambda)(1+\lambda)^{n-1}\left((1+\lambda)^n-1\right)\right)}{Z_0(1+\lambda)^{n-1}\left((1+\lambda)^n-1\right)+\binom{Z_0}{2}(1+\lambda)^{2n}}\\ &= \frac{(1+\lambda)}{1+\frac{Z_0(Z_0-1)(1+\lambda)^{2n}}{2Z_0(1+\lambda)^{n-1}((1+\lambda^n-1)}}\\ &= \frac{2(1+\lambda)}{2+(Z_0-1)(1+\lambda)\underbrace{\frac{(1+\lambda)^n}{(1+\lambda)^n-1}}_{\xrightarrow[n\to\infty]{}1}}\\ &\xrightarrow[n\to\infty]{} \frac{2(1+\lambda)}{Z_0(1+\lambda)+1-\lambda}\,,\end{aligned}$$

gilt mit $c_\infty := \lim_{n\to\infty} c_n$ für alle $x\in\mathbb{R}$:

$$\begin{aligned}\lim_{n\to\infty}P\left(\frac{(1+\lambda)H-2\lambda n\mu G}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}\le x\right) &= c_\infty\lim_{n\to\infty}P\left(\frac{(1+\lambda)H_1-2\lambda n\mu G}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}\le x\right)\\ &\quad+(1-c_\infty)\lim_{n\to\infty}P\left(\frac{(1+\lambda)H_2-2\lambda n\mu G}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}\le x\right)\\ &= c_\infty F_{\mathcal{N}(0,1)}(x)+(1-c_\infty)F_{\mathcal{N}(0,1)}(x)\\ &= F_{\mathcal{N}(0,1)}(x)\,,\end{aligned}$$

wobei $F_{\mathcal{N}(0,1)}(x)$ die Verteilungsfunktion einer $\mathcal{N}(0,1)$-verteilten Zufallsgröße sei. Mit der Charakterisierung der Verteilungskonvergenz mittels Verteilungsfunktionen (siehe [3] Satz 36.5) konvergiert die Zufallsgröße $\frac{(1+\lambda)H-2\lambda n\mu G}{\sqrt{2n\lambda\mu G(1+\lambda+\mu G)}}$ in Verteilung gegen eine standardnormalverteilte Zufallsgröße.

(iv) Der Beweis dieser Aussage folgt der Idee des Beweises zu (iii). Wieder sei der Hamming-Abstand H unterteilt in H_1 und H_2. Wie im Beweis zu Satz 2.13 (iv) berechnet, gilt $\lim_{n\to\infty} n\left(\exp(\mu_n G_n(s-1))-1\right)=\nu(s-1)$ und damit folgt für die erzeugende Funktion von H_2

$$\begin{aligned}f_{H_2}(s) &= \left(\frac{1+\lambda\exp(\mu_n G_n(s-1))}{1+\lambda}\right)^{2n}\\ &= \left(1+\frac{\frac{\lambda}{1+\lambda}2n\left(\exp\left(\mu_n G_n(s-1)\right)-1\right)}{2n}\right)^{2n}\\ &\xrightarrow[n\to\infty]{} \exp\left(\frac{2\lambda\nu}{1+\lambda}(s-1)\right)\,.\end{aligned}$$

Dies ist die erzeugende Funktion der $Poi\left(\frac{2\lambda\nu}{1+\lambda}\right)$-Verteilung. Mit dem Stetigkeitssatz für erzeugende Funktionen folgt, dass H_2 in Verteilung gegen die $Poi\left(\frac{2\lambda\nu}{1+\lambda}\right)$-Verteilung konvergiert. Aus der Voraussetzung an die Mutationsrate μ_n und die Targetlänge G_n ergibt sich

$$\lim_{n\to\infty} \exp\left(\mu_n G_n(s-1)\right) = 1\,.$$

Dies bedeutet für die erzeugende Funktion von H_1:

$$\begin{aligned} f_{H_1}(s) &= \frac{\varphi_{P_n}\left(\exp\left(\mu_n G_n(s-1)\right)\right)}{(1+\lambda)^{n-1}\left((1+\lambda)^n-1\right)} \\ &= \frac{\lambda\exp\left(\mu_n G_n(s-1)\right)\left(1+\lambda\exp\left(\mu_n G_n(s-1)\right)\right)^{2n}-(1+\lambda)^n}{(1+\lambda)^{n-1}\left((1+\lambda)^n-1\right)\left(1+\lambda\exp\left(\mu_n G_n(s-1)\right)\right)^2-(1+\lambda)} \\ &= \exp\left(\frac{2\lambda\nu}{1+\lambda}(s-1)\right) \end{aligned}$$

Eine ähnliche Rechnung wie in (iii) liefert die Konvergenz der Verteilungsfunktion gegen die Verteilungsfunktion einer Poisson-verteilten Zufallsgröße. Aus der Charakterisierung der Verteilungskonvergenz mittels Verteilungsfunktionen folgt die Behauptung. □

Bemerkung 2.32 Ähnlich wie bei der Normal- und Poisson-Approximation der Anzahl an Mutationen einer zufällig gezogenen Sequenz M in Satz 2.13, können die Grenzwertsätze in Satz 2.31 (iii) und (iv) zur approximativen Berechnung des Hamming-Abstandes verwendet werden. Dabei liefert die Normalapproximation für großes $n\mu G$ bessere Werte und für kleines $n\mu G$ die Poisson-Approximation.
Desweiteren kann die Mutationsrate μ auch mit Hilfe des Hamming-Abstandes geschätzt werden. Dazu sei $H(A_1, A_2)$ der Hamming-Abstand zweier zufällig ohne Zurücklegen gezogener Sequenzen A_1, A_2 nach n PCR-Zyklen. Gegeben sei eine Stichprobe $h_{i,j},\ 1 \le i < j \le s$ von Hamming-Abständen von s zufällig gezogenen Sequenzen einer n-stufigen PCR. Die Momentenmethode (siehe [1] Kapitel II.6) liefert dann den Schätzer

$$\hat{\mu}\left((h_{i,j})_{1\le i<j\le s}\right) := \frac{\sum_{1\le i<j\le s} h_{i,j}}{\binom{s}{2} GED}\,.$$

Auf die Berechnung der Varianz dieses Schätzers verzichten wir an dieser Stelle.
Es wird aber an den sehr praktisch orientierten Artikel von Weiss und Haesseler [28] verwiesen, in dem einige der in diesem Kapitel theoretisch entwickelten Ergebnisse an realen und simulierten Datensätzen erläutert werden.

3. Eigenschaften allgemeiner größenabhängiger Verzweigungsprozesse

In Kapitel 2 wurde die Effizienz der PCR als konstant vorausgesetzt. Dies ist eine Vereinfachung der Reaktion, die angesichts der guten Ergebnisse für die Schätzung der Mutationsrate gerechtfertigt erscheint, doch der chemischen Realität nicht entspricht (siehe dazu auch B.4). Vielmehr gibt es im Laufe der Reaktion zwei Phasen: Zu Beginn tritt exponentielles Wachstum auf, nach einiger Zeit hingegen verlangsamt sich das Wachstum, bis die Nachkommen annähernd linear mit der Zeit zunehmen, dies wird als *gesättigte Phase* bezeichnet. Dieser Sättigungseffekt hat verschiedene Ursachen: die Primerkonzentration nimmt im Laufe der Reaktion ab, die Enzymaktivität sinkt möglicherweise und die bei den vorangegangenen Schritten entstandenen Moleküle behindern die Reaktion sterisch.

Abbildung 3.1 zeigt den Übergang der exponentiellen Phase zur gesättigten am Beispiel

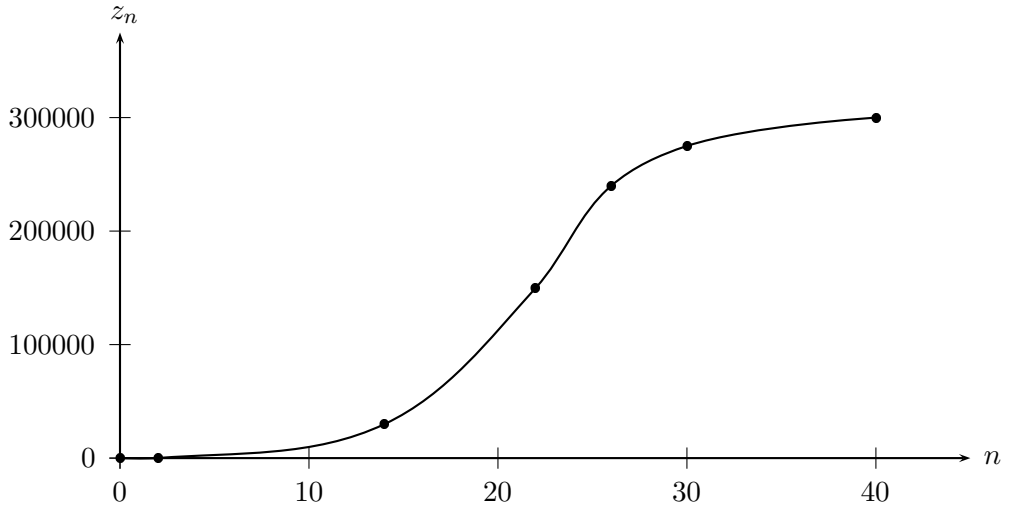

Abbildung 3.1.: Experimentell bestimmte Amplifikationsdaten z_n, siehe [22]

real gemessener PCR-Daten. In Kapitel 4.1 wird das in Kapitel 2.1 beschrieben Modell erweitert, indem die konstante Effizienz durch eine größenabhängige Effizienzfunktion ersetzt wird. Dadurch rückt das Modell in die Klasse der größenabhängigen Verzweigungsprozesse. Um dann mit Hilfe dieser Klasse weitere Aussagen über die PCR formulieren zu können, wird diese Klasse im folgenden Kapitel eingeführt und näher untersucht. Dieses Kapitel basiert weitgehend auf einem Artikel von Lalam und Jacob [20].

3.1. Modellbeschreibung

Den Begriff des größenabhängigen Verzweigungsprozesses klärt

Definition 3.1 Gegeben sei ein messbarer Raum $(\Omega, \mathcal{A})$, darauf eine Familie von Wahrscheinlichkeitsmaßen $(P_j)_{j\geq 0}$ und eine Zufallsgröße Z_0, für die unter P_j f.s. $Z_0 = j$ gilt. Desweiteren seien Zufallsgrößen Z_n für alle $n \geq 1$ definiert durch

$$Z_n := \sum_{j=1}^{Z_{n-1}} I_{n,j} \tag{3.1}$$

mit bedingt unter $\mathcal{F}_{n-1} := \sigma(Z_0, \ldots, Z_{n-1})$ stochastisch unabhängigen, identisch verteilten Zufallsgrößen $(I_{n,j})_{j\geq 1}$. Deren bedingter Erwartungswert wird mit $E(I_{n,j} \mid \mathcal{F}_{n-1}) =: m(Z_{n-1})$ bezeichnet und *Nachkommenmittel* oder *Reproduktionsmittel* genannt; die bedingte Varianz $Var(I_{n,j} \mid \mathcal{F}_{n-1}) =: \sigma^2(Z_{n-1})$ heißt *Reproduktionsvarianz*. Dann ist

$$((\Omega, \mathcal{A}), (P_j)_{j\geq 0}, (I_{n,j})_{j,n\geq 1}, (Z_n)_{n\geq 0})$$

das *Standardmodell eines größenabhängigen Verzweigungsprozesses*. $(Z_n)_{n\geq 0}$ heißt *größenabhängiger Verzweigungsprozess*.

Im Folgenden untersuchen wir nur Prozesse mit einem Urahn.
Ähnlich wie in der Theorie gewöhnlicher Galton-Watson-Prozesse kann man den größenabhängigen Verzweigungsprozess anhand der Eigenschaften des Nachkommenmittels eines Individuums klassifizieren:

Definition 3.2 Gegeben sei ein größenabhängiger Verzweigungsprozess $(Z_n)_{n\geq 0}$, für dessen Nachkommenmittel $\lim_{N\to\infty} m(N) = m$ gelte. Dieser Prozess heißt
(i) *superkritisch* (englisch *supercritical*) falls $m > 1$ und
(ii) *fast-kritisch* (engl. *near-critical*) falls $m = 1$.

Gleichung (3.1) lässt sich auch in einer anderen Weise schreiben:

$$Z_n = m(Z_{n-1})Z_{n-1} + \eta_n, \tag{3.2}$$

wobei $\eta_n := \sum_{j=1}^{Z_{n-1}} (I_{n,j} - m(Z_{n-1}))$ eine Martingaldifferenz ist. Auf Grund der $\mathcal{F}_{n-1}$-Messbarkeit der Summe und der ersten Waldschen Gleichung gilt nämlich:

$$\begin{aligned} E(\eta_n \mid \mathcal{F}_{n-1}) &= E\left(\sum_{j=1}^{Z_{n-1}} I_{n,j} \mid \mathcal{F}_{n-1}\right) - m(Z_{n-1})EZ_{n-1} \\ &= EZ_{n-1}E\left(I_{n,1} \mid \mathcal{F}_{n-1}\right) - m(Z_{n-1})EZ_{n-1} = 0 \text{ f.s.} \end{aligned}$$

Das Ziel ist nun, für das Nachkommenmittel $m(Z_n)$ einen geeigneten Schätzer zu finden. Dabei sind die Stichprobenwerte die experimentell bestimmten Gesamtanzahlen der Nachkommen in den verschiedenen Zyklen. Geeignet ist der Schätzer dann, wenn die Genauigkeit des geschätzten Wertes mit zunehmender Datenmenge steigt, genauer (siehe [29]):

Definition 3.3 Gegeben sei ein statistisches Experiment, in dem der Parameter $\theta_0 \in \Theta$ zu schätzen ist. Die Folge von Schätzern $\left(\hat{\theta}_n\right)_{n\geq 1}$ heißt *stark konsistent*, wenn gilt:

$$\hat{\theta}_n \xrightarrow[n\to\infty]{} \theta_0 \text{ f.s.}$$

Im weiteren Verlauf betrachten wir nur noch speziellen Voraussetzungen genügende Modelle. Das Modell für die PCR erfüllt diese Voraussetzungen, wie sich in Kapitel 4 herausstellt. Die ersten beiden Voraussetzungen schränken die in Frage kommenden Funktionen für das Reproduktionsmittel und die Reproduktionsvarianz ein. Voraussetzung 5 ist eine technische Bedingung, die bei der Untersuchung des asymptotischen Verhaltens eine Rolle spielt. Voraussetzung 6 garantiert, dass triviale Modelle ausgeschlossen sind.

Vorausetzung 3 (V.3) Für das Nachkommenmittel $m(\cdot) : \mathbb{N} \to \mathbb{R}$, $N \mapsto m(N)$ gelte für alle $N \geq 1$

$$m(N) = m + f(N),$$

wobei $m \geq 1$ und $f : \mathbb{N} \to \mathbb{R}$ eine Funktion ist, die die Bedingung

$$|f(N)| \leq \frac{K}{N^\alpha}$$

für ein $K < \infty$ und ein $\alpha > 0$ erfüllt.

Vorausetzung 4 (V.4) Für die Varianz als Funktion in N aufgefasst gelte

$$\sigma^2(N) \leq \sigma^2 N^\beta$$

im superkritischen Fall (d.h. $m > 1$) für alle $N \geq 1$ und im fast-kritischen Fall (also $m = 1$) nur für alle $N \geq N_0$ für ein $N_0 \in \mathbb{N}$ groß genug. Dabei seien $\sigma^2 \in (0, \infty)$ und $\beta \in [-1, 1)$.

Vorausetzung 5 (V.5) Im fast-kritischen Fall sei $m(\cdot) > 0$ und monoton fallend. Die Funktion $\frac{\sigma^2(x)}{x^2(m(x)-1)}$ sei streng monoton fallend auf $[1, \infty)$ und genüge folgender Integralbedingung:

$$\int_1^\infty \frac{\sigma^2(x)}{x^2(m(x)-1)} dx < \infty$$

Vorausetzung 6 (V.6) Die Wahrscheinlichkeit dafür, dass zu jedem Zeitpunkt ein Individum 0 oder 1 Nachkommen erzeugt, liege echt zwischen 0 und 1, d.h.

$$0 < p_{N0} + p_{N1} < 1$$

wobei $(p_{Ni})_{i \in \mathbb{N}}$ die Nachkommenverteilung für die Populationsgröße $N \in \mathbb{N}$ bezeichne.

3.2. Asymptotisches Verhalten

In diesem Abschnitt wird eine Wachstumsrate a_n für den superkritischen und den fast-kritischen größenabhängigen Verzweigungsprozess angegeben. Teil (a) des nachfolgenden Satzes entstammt [18], Teil (b) ist eine Zusammenstellung mehrerer Aussagen aus [16], [17] und [19].

Satz 3.4 (*a*) *Gegeben sei ein den Voraussetzungen V.3 und V.4 genügender, superkritischer, größenabhängiger Verzweigungsprozess* $(Z_n)_{n\geq 0}$*, für den also*

$$(i)\ m(N) = m + f(N) \quad mit \quad |f(N)| \leq \frac{K_1}{N^\alpha}$$
$$(ii)\ \sigma^2(N) \leq K_2 N^\beta,$$

mit $K_1, K_2, \alpha > 0$ *und* $0 \leq \beta < 1$ *gelten. Weiter seien* $a_n := m^n$ *und* $W_n := \frac{Z_n}{a_n}$ *für alle* $n \geq 0$*. Dann existiert eine integrierbare Zufallsvariable* W*, mit* $0 \leq W < \infty$ *f.s., derart dass*

$$W_n \xrightarrow[n\to\infty]{} W \quad f.s.\ und\ in\ \mathcal{L}_2 \quad sowie\ P(W > 0) > 0.$$

(*b*) *Gegeben seien ein fast-kritischer, größenabhängiger Verzweigungsprozess* $(Z_n)_{n\geq 0}$*, der neben V.3 und V.4 der Voraussetzung V.5 genüge,* $a_n := a_{n-1} m(a_{n-1})$*,* $W_n := \frac{Z_n}{a_n}$ *und* $E_\infty := \{\lim_{n\to\infty} Z_n = \infty\}$*. Dann gilt* $P(E_\infty) > 0$*,* $\lim_{n\to\infty} a_n = \infty$ *und*

$$\lim_{n\to\infty} W_n = 1 \quad f.s.\ auf\ E_\infty\,.$$

Beweis: (a) Der Beweis wird in zwei Fälle aufgeteilt, $0 < \alpha < 1$ und $\alpha \geq 1$. Der Kniff im Beweis des ersten Falles ist die Anwendung des im Anhang bewiesenen Lemmas A.2 auf die Folgen $(EW_n)_{n\in\mathbb{N}}$ und $\left(EW_n^2\right)_{n\in\mathbb{N}}$.
Sei also zuerst $0 < \alpha < 1$. Aus der Definition des Prozesses (Gleichung (3.1)) folgt mit der ersten Waldschen Gleichung

$$E\left(Z_{n+1} \mid \mathcal{F}_n\right) = Z_n m(Z_n) = mZ_n + Z_n f(Z_n) \tag{3.3}$$

und daraus ergibt sich

$$\left|E\left(Z_{n+1} \mid \mathcal{F}_n\right) - mZ_n\right| = \left|Z_n f(Z_n)\right| \leq K_1 Z_n^{1-\alpha},$$

wobei Voraussetzung (i) die letzte Ungleichung liefert. Folglich gilt

$$\begin{aligned} \left|E\left(W_{n+1} \mid \mathcal{F}_n\right) - W_n\right| &= \frac{1}{m^{n+1}} \left|E\left(Z_{n+1} \mid \mathcal{F}_n\right) - mZ_n\right| \\ &\leq \frac{1}{m^{n+1}} K_1 \frac{Z_n^{1-\alpha}}{m^{n-n\alpha}} m^{n-n\alpha} = K_1 W_n^{1-\alpha} m^{-n\alpha-1} \end{aligned} \tag{3.4}$$

und durch die Bildung der Erwartungswerte sowie der Anwendung der Standardungleichung $|EX| \leq E|X|$ im ersten Schritt und der Jensenschen Ungleichung auf die konkave Funktion $x \mapsto x^{1-\alpha}$ im letzen Schritt (die beiden Ungleichung sind z.B. in [3] Satz 17.4 zu finden)

$$\begin{aligned} \left|EW_{n+1} - EW_n\right| &\leq E\left|E\left(W_{n+1} \mid \mathcal{F}_n\right) - W_n\right| \\ &\leq K_1 EW_n^{1-\alpha} m^{-n\alpha-1} \\ &\leq K_1 (EW_n)^{1-\alpha} m^{-n\alpha-1}. \end{aligned} \tag{3.5}$$

Da $m(N) \xrightarrow[N\to\infty]{} m$ und $m(N) > 0$ für alle $n \geq 0$, ist $\tilde{m} := \inf_{N\in\mathbb{N}_0} m(N) > 0$. Mit (3.3) folgt dann

$$EZ_{n+1} \geq \tilde{m} EZ_n \geq \tilde{m}^n Z_0 > 0$$

und somit ist auch $EW_n > 0$ für alle $n > 0$.
Damit sind die Voraussetzungen des Lemmas A.2 für die Folge $(EW_n)_{n\geq 0}$ erfüllt. Also existiert der Grenzwert $\lim_{n\to\infty} EW_n$ und ist endlich. Da $Z_0, a_0 > 0$, gilt weiter

$$0 < \lim_{n\to\infty} EW_n < \infty. \tag{3.6}$$

Nun seien Y_n und S_n definiert durch

$$Y_n := W_n + \underbrace{K_1 \sum_{k=0}^{n-1} W_k^{1-\alpha} m^{-k\alpha-1}}_{=:S_n} = W_n + S_n. \tag{3.7}$$

Mit Gleichung (3.4) und der $\mathcal{F}_n$-Messbarkeit der W_n folgt

$$\begin{aligned} E\left(Y_{n+1} \mid \mathcal{F}_n\right) &= E\left(W_{n+1} \mid \mathcal{F}_n\right) + K_1 \sum_{k=0}^{n} W_k^{1-\alpha} m^{-k\alpha-1} \\ &\geq W_n - K_1 W_n^{1-\alpha} m^{-n\alpha-1} + K_1 \sum_{k=0}^{n} W_k^{1-\alpha} m^{-k\alpha-1} \\ &= Y_n, \end{aligned}$$

d.h. $(Y_n)_{n\geq 1}$ ist ein Submartingal bezüglich der Filtration $(\mathcal{F}_n)_{n\geq 1}$. Zusätzlich sind die Y_n nicht-negativ, da die $W_n > 0$ sind. Desweiteren gilt

$$\sup_{n\in\mathbb{N}} EY_n \leq \sup_{n\in\mathbb{N}} EW_n + \sup_{n\in\mathbb{N}} ES_n$$

sowie mit der Jensenschen Ungleichung

$$ES_n \leq K_1 \sum_{k=0}^{n-1} \left(EW_k\right)^{1-\alpha} m^{-k\alpha-1}.$$

Da der Grenzwert der EW_n für $n \to \infty$, wie oben festgestellt, existiert und endlich ist, folgt aus diesen beiden Gleichungen, dass $\sup_{n\in\mathbb{N}} EY_n < \infty$. Nach dem Martingalkonvergenzsatz (siehe [2] Satz 21.2) existiert also eine Zufallsgröße Y mit endlichen Erwartungswert, so dass

$$Y_n \xrightarrow[n\to\infty]{} Y \quad \text{f.s.} \tag{3.8}$$

Die Folge $(S_n)_{n\geq 1}$ ist beschränkt und monoton wachsend, also existiert der Grenzwert $\lim_{n\to\infty} S_n =: S$ f.s. Außerdem gilt, da sowohl EY als auch $\lim_{n\to\infty} EW_n$ endlich sind, die Gleichung

$$ES = E\left(K_1 \sum_{k=0}^{\infty} W_k^{1-\alpha} m^{-k\alpha-1}\right) < \infty\,.$$

Folglich ist S f.s. endlich und für $W := Y - S$ ergibt sich dann aus den Gleichungen (3.7) und (3.8)

$$W_n \xrightarrow[n\to\infty]{} W \quad \text{f.s.}$$

sowie $0 \leq W < \infty$ f.s. und $EW = EY - ES < \infty$.

Nun wird die $\mathcal{L}_2$-Konvergenz bewiesen. Dazu sei $\|\cdot\|_2$ die $\mathcal{L}_2$-Norm. Aus der Definition des Prozesses und der zweiten Waldschen Gleichung folgt

$$\begin{aligned} E\left(Z_{n+1}^2 \mid \mathcal{F}_n\right) &= Var\left(Z_{n+1} \mid \mathcal{F}_n\right) + \left(E\left(Z_{n+1} \mid \mathcal{F}_n\right)\right)^2 \\ &= \sigma^2(Z_n)Z_n + m^2(Z_n)Z_n^2 \\ &= \sigma^2(Z_n)Z_n + \left(m + f(Z_n)\right)^2 Z_n^2 \\ &= \sigma^2(Z_n)Z_n + m^2 Z_n^2 + 2mf(Z_n)Z_n^2 + f(Z_n)^2 Z_n^2. \end{aligned}$$

Damit ergibt sich

$$EW_{n+1}^2 = E\frac{Z_{n+1}^2}{m^{2n+2}} = E\frac{\sigma^2(Z_n)Z_n}{m^{2n+2}} + EW_n^2 + 2E\frac{f(Z_n)Z_n^2}{m^{2n+1}} + E\frac{f(Z_n)^2 Z_n^2}{m^{2n+2}},$$

und folglich ist

$$\left|EW_{n+1}^2 - EW_n^2\right| \le E\frac{\sigma^2(Z_n)Z_n}{m^{2n+2}} + 2E\left|\frac{f(Z_n)Z_n^2}{m^{2n+1}}\right| + E\frac{f(Z_n)^2 Z_n^2}{m^{2n+2}}. \tag{3.9}$$

Die rechte Seite der Ungleichung (3.9) wird nun termweise betrachtet. Mit $\gamma := \alpha \wedge (1-\beta)$ liefern die Voraussetzung (ii), die Jensensche Ungleichung für konkave Funktionen und $m > 1$ die Abschätzung

$$\begin{aligned} E\frac{\sigma^2(Z_n)Z_n}{m^{2n+2}} &\le K_2 E\frac{Z_n^{1+\beta}}{m^{2n+2}} \le K_2 E\frac{Z_n^{2-\gamma}}{m^{2n+2}} = K_2 m^{-n\gamma-2} EW_n^{2-\gamma} \\ &< K_2 m^{-n\gamma-1}\left(EW_n^2\right)^{1-\frac{\gamma}{2}}. \end{aligned}$$

Weiter folgt mit Voraussetzung (i)

$$\begin{aligned} E\left|\frac{f(Z_n)Z_n^2}{m^{2n+1}}\right| &\le K_1 E\frac{Z_n^{2-\alpha}}{m^{2n+1}} \le K_1 E\frac{Z_n^{2-\gamma}}{m^{2n+1}} = K_1 m^{-n\gamma-1} EW_n^{2-\gamma} \\ &\le K_1 m^{-n\gamma-1}\left(EW_n^2\right)^{1-\frac{\gamma}{2}} \end{aligned}$$

Für die nächste Abschätzung wird $|f(Z_n)|$ einmal großzügig durch K_1 nach oben abgeschätzt, und man erhält mit der eben bewiesenen Ungleichung

$$\begin{aligned} E\frac{f(Z_n)^2 Z_n^2}{m^{2n+2}} &\le \frac{K_1}{m} E\left|\frac{f(Z_n)Z_n^2}{m^{2n+1}}\right| \le K_1^2 m^{-n\gamma-2}\left(EW_n^2\right)^{1-\frac{\gamma}{2}} \\ &< K_1^2 m^{-n\gamma-1}\left(EW_n^2\right)^{1-\frac{\gamma}{2}} \end{aligned}$$

Die Ungleichung (3.9) lässt sich also unter Verwendung von $C := 4\max\{K_1, K_1^2, K_2\}$ zusammenfassen zu

$$\left|EW_{n+1}^2 - EW_n^2\right| < Cm^{-n\gamma-1}\left(EW_n^2\right)^{1-\frac{\gamma}{2}} \tag{3.10}$$

Die Folge $\left(EW_n^2\right)_{n\ge 0}$ genügt somit den Voraussetzungen des Lemmas A.2 und folglich besitzt diese Folge einen endlichen Grenzwert, d.h.

$$\lim_{n\to\infty} \|W_n\|_2 < \infty. \tag{3.11}$$

Die Minkowski-Ungleichung (siehe [3] Satz 17.4 (e)) und die Definition der Zufallsgröße Y_n in Gleichung (3.7) liefern

$$\|Y_n\|_2 \le \|W_n\|_2 + \|S_n\|_2. \tag{3.12}$$

Da für jede Zufallsgröße $X \in \mathcal{L}_2$ und für alle $r,\ 0 < r < 1$ aus der Jensenschen Ungleichung (für den konkaven Fall)

$$\|X^r\|_2 = \left(EX^{2r}\right)^{\frac{1}{2}} \leq \left(EX^2\right)^{\frac{r}{2}} = \|X\|_2^r$$

folgt, liefert Gleichung (3.11) mit Blick auf die Definition der Zufallsgrößen S_n in (3.7) und $S = \lim_{n\to\infty} S_n$:

$$\lim_{n\to\infty} \|S - S_n\|_2 = 0$$

Das heißt: $S_n \xrightarrow{\mathcal{L}_2} S$ und es gilt $\sup_{n\in\mathbb{N}} \|S_n\|_2 < \infty$. Mit (3.11) und (3.12) folgt nun $\sup_{n\in\mathbb{N}} \|Y_n\|_2 < \infty$. Der Submartingal-Einschachtelungs-Satz („submartingale closure theorem“, siehe [23] S. 59) liefert

$$Y_n \xrightarrow[n\to\infty]{} Y \quad \text{f.s. und in } \mathcal{L}_2.$$

Damit gilt $\|W - W_n\|_2 \leq \|Y - Y_n\|_2 + \|S - S_n\|_2 \xrightarrow[n\to\infty]{} 0$, was

$$W_n \xrightarrow[n\to\infty]{} W \quad \text{f.s. und in } \mathcal{L}_2.$$

impliziert. Das bedeutet auch, dass EW und $\lim_{n\to\infty} EW_n$ übereinstimmen. Aus Gleichung (3.6) folgt dann für alle $Z_0 > 0$, dass

$$P(W > 0) > 0.$$

Zweiter Fall: $\alpha \geq 1$.
Der Beweis für diesen Fall läuft weitestgehend analog zum ersten Fall. Da in der Abschätzung (3.5) die Jensensche Ungleichung für den konkaven Fall verwendet wird, der nur bei $\alpha < 1$ erfüllt ist, muss nun auf andere Weise abgeschätzt werden:

$$\begin{aligned} |EW_{n+1} - EW_n| &\leq E\,|E(W_{n+1}|\mathcal{F}_n) - W_n| \\ &\leq K_1 \underbrace{EZ_n^{1-\alpha}}_{\leq 1} m^{-(n-1)\alpha-2} \\ &\leq K_1 m^{\alpha-2} m^{-n} \xrightarrow[n\to\infty]{} 0 \end{aligned}$$

Die Folge $(EW_n)_{n\in\mathbb{N}}$ ist also eine Cauchy-Folge, sie besitzt einen endlichen Grenzwert. Die Anwendung des Lemmas A.2 entfällt hier. Y_n sei nun definiert als

$$Y_n := W_n + \hat{K}_1 \sum_{k=0}^{n-1} m^{-k\alpha-1}$$

mit $\hat{K}_1 := K_1 m^{\alpha-2}$. Mit diesen Änderungen lässt sich der Beweis des ersten Falles übernehmen.

(b) Auf einen Beweis des Ergebnisses muss an dieser Stelle verzichtet werden, da die Aussage auf mehreren Lemmata und Sätzen aus [16], [17] und [19] beruht. Die Darstellung dieses Beweises würde den Rahmen dieser Arbeit sprengen. □

3.3. Identifizierbarkeit und starke Konsistenz

Das Nachkommenmittel $m(N)$ zur Populationsgröße $N \in \mathbb{N}_0$ genüge Voraussetzung V.3, also der Gleichung

$$m(N) = m + f(N),$$

und hänge von einem zu schätzenden Parameter $(\theta_0, \nu_0) \in \overset{\circ}{\Theta} \times \mathcal{N}$ ab. Dabei stamme θ_0 aus dem Inneren einer kompakten Menge $\Theta \subset \mathbb{R}^{d_1}$, $d_1 \in \mathbb{N}$ und der Parameter ν_0 aus einer kompakten Teilmenge $\mathcal{N} \subset \mathbb{R}^{d_2}$. Er kann endlich dimensional sein, dann ist $d_2 \in \mathbb{N}$, oder unendlich dimensonial, dann gilt $d_2 = \mathbb{N}$. Hierbei stelle man sich vor, dass der zu schätzende Parameter aus zwei Teilen besteht. Ein endlich dimensionaler Teil ist von großem Interesse und wird mit θ_0 bezeichnet. Ein anderer, möglicherweise unendlich dimensionaler Teil ν_0, ist nur ein Störparameter, auf dessen Bestimmung unter gewissen Umständen (☞ Definition 3.7) verzichtet werden kann.
Um die Abhängigkeit des Nachkommenmittels von (θ_0, ν_0) zum Ausdruck zu bringen, wird im folgenden $m_{\theta_0,\nu_0}(N)$ anstelle von $m(N)$ geschrieben. Weiter sei für alle $\delta > 0$ und $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,d})$

$$B_\delta := \left\{ \theta = (\theta_1, \ldots, \theta_d) \in \Theta : \sum_{i=1}^{d} \mid \theta_i - \theta_{0,i} \mid < \delta \right\}.$$

Das Ziel im Auge behaltend, einen konsistenten Schätzer für das Nachkommenmittel, genauer für (θ_0, ν_0) zu finden, wird an dieser Stelle der Begriff der *Identifizierbarkeit* eingeführt. Dazu sei mit $a \in \mathbb{N}$ die Prozessgröße bezeichnet, bis zu der der Prozess betrachtet wird. Wir nehmen an, dass für jedes a eine Halbnorm zur Verfügung steht, mit der der Abstand zwischen realem Nachkommenmittel und von dem Parameter θ abhängenden Nachkommenmittel gemessen werden kann, ohne diese Folge zuerst genauer zu spezifizieren. Grob gesagt, heißt ein zu schätzender Parameter θ_0 *identifizierbar*, wenn außerhalb jeder δ-Umgebung um den wahren Parameter θ_0 f.s. kein weiterer Parameter existiert, in dem das Nachkommenmittel den „realen“ Wert annimmt. Der Abstand wird mit der Halbnorm gemessen.
Wird mit zunehmender maximaler Prozessgröße a außerhalb der δ-Umgebung der Abstand zwischen dem Nachkommenmittel und dem „realen“ Nachkommenmittel nicht 0, so heißt der Parameter θ_0 *asymptotisch identifizierbar*.
Gilt dies auch für alle ν, so wird θ_0 *gleichgradig asymptotisch identifizierbar* genannt. Hier die genauen mathematischen Definitionen dieser Begriffe:

Definition 3.5 Gegeben seien ein größenabhängiger Verzweigungsprozess $(Z_n)_{n\geq 0}$ mit der Nachkommenmittelfunktion $m_{\theta_0,\nu_0}(\cdot)$, die der Voraussetzung V.3 genüge und von einem zu schätzenden Parameter $(\theta_0, \nu_0) \in \overset{\circ}{\Theta} \times \mathcal{N}$ abhänge. Weiter sei für alle $a \in \mathbb{N}$ $\|\cdot\|^a$ eine Halbnorm auf der Menge $\{f(N) : \ N \leq a\}$.
(i) Dann heißt θ_0 *identifizierbar* auf $\{m_{\theta_0,\nu}(N) : \ N \leq a\}$ für die Halbnorm $\|\cdot\|^a$, falls es eine Funktion $v(\cdot)$ gibt, so dass für alle $\delta > 0$

$$\inf_{\theta \in B_\delta^c} \|(m_{\theta_0,\nu}(\cdot) - m_{\theta,\nu}(\cdot))\, v(\cdot)\|^a \neq 0 \text{ f.s.}$$

gilt.
(ii) θ_0 heißt *asymptotisch identifizierbar* in $m_{\theta_0,\nu}(\cdot)$ für $(\|\cdot\|^a)_{a\in\mathbb{N}}$, wenn es eine Funktion $v(\cdot)$ gibt, derart dass für alle $\delta > 0$

$$\liminf_{a\to\infty} \inf_{\theta \in B_\delta^c} \|(m_{\theta_0,\nu}(\cdot) - m_{\theta,\nu}(\cdot))\, v(\cdot)\|^a \neq 0 \text{ f.s.}$$

(iii) θ_0 heißt *gleichgradig asymptotisch identifizierbar* in $m_{\theta_0,\cdot}(\cdot)$ für $(\|\cdot\|^a)_{a\in\mathbb{N}}$, wenn es eine Funktion $v(\cdot)$ gibt, so dass für alle $\delta > 0$

$$\liminf_{a\to\infty} \inf_{\theta\in B_\delta^c, \nu\in\mathcal{N}} \|(m_{\theta_0,\nu}(\cdot) - m_{\theta,\nu}(\cdot))\, v(\cdot)\|^a \neq 0 \text{ f.s.}$$

Beispiel 3.6 Das Nachkommenmittel eines größenabhängigen Verzweigungsprozesses genüge der Gleichung

$$m(N) = \theta_0 + \mathcal{O}(N^{-\alpha})$$

mit $\alpha > 0$. Es handelt sich hier um einen Spezialfall der in V.3 geforderten Klasse. Die unbekannten Parameter sind $\theta_0 \in [0,\infty)$ und $\alpha \in (0,\infty)$. Dann ist θ_0 asymptotisch identifizierbar in $m_{\theta_0,\alpha}$ für $|\cdot|$ mit der Rate 1. Für beliebiges $N \in \mathbb{N}$ und $\alpha > 0$ gilt nämlich

$$\begin{aligned} \inf_{\theta\in B_\delta^c} |m_{\theta_0,\alpha}(N) - m_{\theta,\alpha}(N)| &= \inf_{m\in B_\delta^c} |\theta_0 - \theta + \mathcal{O}(N^{-\alpha}) - \mathcal{O}(N^{-\alpha})| \\ &\geq \delta\,. \end{aligned}$$

Da $\alpha > 0$ beliebig gewählt war, ist θ_0 sogar gleichgradig asymptotisch identifizierbar.

Wie gerade gesehen, gibt es Fälle, in denen das Nachkommenmittel in zwei Teile zerlegt werden kann, von denen einer nur vom uns interessierenden Parameter θ abhängt, und der andere (möglicherweise von diesem und) von ν abhängt. Verschwindet letzterer bezüglich der Halbnorm mit zunehmender Prozessgröße a, so wird er *asymptotisch vernachlässigbar* genannt und ist somit nichts anderes als ein Störparameter. Genauer:

Definition 3.7 Die vom Parameter ν abhängende Funktion $g_\nu(\cdot)$ heißt *gleichgradig asymptotisch vernachlässigbar*, falls

$$\limsup_{a\to\infty} \sup_{\nu\in\mathcal{N}} \|g_\nu(\cdot)\|^a = 0 \text{ f.s.}$$

Um ein im weiteren Verlauf benötigtes Ergebnis formulieren zu können, notieren wir

Definition 3.8 Die Folge $\left(\hat\theta_a\right)_{a\in\mathbb{N}}$ heißt *fast sicher $m_{\theta_0,\nu}(\cdot)$-konsistent* für $(\|\cdot\|^a)_{a\in\mathbb{N}}$ falls es eine Funtkion $v(\cdot)$ gibt, so dass

$$\lim_{a\to\infty} \left\| \left(m_{\theta_0,\nu(\cdot)} - m_{\hat\theta_a,\nu}(\cdot) \right) v(\cdot) \right\|^a = 0 \text{ f.s.}$$

Hier das angekündigte Ergebnis.

Lemma 3.9 *Es sei*

$$\mathcal{M}_{v,\nu} := \left\{ \left(\hat\theta_a\right)_{a\in\mathbb{N}} :\ \left(\hat\theta_a\right)_{a\in\mathbb{N}} \textit{ ist fast sicher } m_{\theta_0,\nu}(\cdot)\textit{-konsistent mit der Rate } v(\cdot) \right\}.$$

Ist dann θ_0 asymptotisch identifizierbar in $v(\cdot)$ für $(\|\cdot\|^a)_{a\in\mathbb{N}}$ folgt, dass für alle $\left(\hat\theta_a\right)_{a\in\mathbb{N}} \in \mathcal{M}_{v,\nu}$ die Folge $\left(\hat\theta_a\right)_{a\in\mathbb{N}}$ stark konsistent ist.
Die Umkehrung gilt, falls zusätzlich angenommen wird, dass für alle abgeschlossenen $F \subset \Theta$ das Infimum $\inf_{\theta\in F} \|(m_{\theta_0,\nu}(\cdot) - m_{\theta,\nu}(\cdot))\, v(\cdot)\|^a$ für ein θ_a angenommen wird.

Beweis: Angenommen, es gäbe unter den oben angegebenen Voraussetzungen eine Folge $\left(\hat{\theta}_a\right)_{a\in\mathbb{N}}$ aus $\mathcal{M}_{v,\nu}$, die nicht f.s. konsistent ist. Dann existiert für alle $\delta > 0$ ein a_δ und eine Teilfolge $\left(\hat{\theta}_{a'}\right)_{a'>a_\delta}$ derart, dass $\hat{\theta}_{a'} \in B_\delta^c$ für alle $a' > a_\delta$. Da $\left(\hat{\theta}_a\right)_{a\in\mathbb{N}}$ aus $\mathcal{M}_{v,\nu}$ stammt, ist der Grenzwert

$$\lim_{a\to\infty} \left\|\left(m_{\theta_0,\nu}(\cdot) - m_{\theta_a,\nu}(\cdot)\right) v(\cdot)\right\|^a = 0 \text{ f.s.}$$

und damit gilt auch

$$\liminf_{a\to\infty} \inf_{\theta\in B_\delta^c} \left\|\left(m_{\theta_0,\nu}(\cdot) - m_{\theta,\nu}(\cdot)\right) v(\cdot)\right\|^a = 0 \text{ f.s.}$$

Folglich ist θ_0 nicht asymptotisch identifizierbar mit der Rate $v(\cdot)$. Das ist ein Widerspruch zur Voraussetzung.
Zum Beweis der umgekehrten Implikation gelte die Zusatzbedingung. Angenommen, θ_0 sei nicht asymptotisch identifizierbar in $m_{\theta_0,\nu}(\cdot)$ mit der Rate $v(\cdot)$. Dann existiert ein $\delta > 0$, so dass $\liminf_{a\to\infty} \inf_{\theta\in B_\delta^c} \left\|\left(m_{\theta_0,\nu}(\cdot) - m_{\theta,\nu}(\cdot)\right) v(\cdot)\right\|^a = 0$ f.s. Mit

$$\theta_a := \arg\min_{\theta\in B_\delta^c} \left\|\left(m_{\theta_0,\nu}(\cdot) - m_{\theta,\nu}(\cdot)\right) v(\cdot)\right\|^a$$

Die Folge gehört zu $\mathcal{M}_{v,\nu}$, ist aber nicht konsistent, da sie zu B_δ^c gehört. Dies ist ein Widerspruch, woraus die Behauptung folgt. □

Der Schätzer $\hat{\theta}$ für den Parameter θ_0 soll nicht nur die Eigenschaft der starken Konsistenz besitzen, sondern auch gute Schätzergebnisse der Anzahl der Nachkommen nach n Schritten liefern. Mit Blick auf die zu Anfang dieses Kapitels eingeführte Zerlegung von Z_n in Gleichung (3.2) $Z_n = m(Z_{n-1})Z_{n-1} + \eta_n$, soll das von θ abhängende Nachkommenmittel möglichst genau durch $m_{\hat{\theta}}$ geschätzt werden. Dies ist der Fall, wenn folgende Kleinste-Quadrate-Gleichung minimiert wird:

$$\tilde{S}_{h,n,\nu,\gamma}(\theta) := \sum_{k=h+1}^{n} \left(Z_k - m_{\theta,\nu}(Z_{k-1})Z_{k-1}\right)^2 Z_{k-1}^{-\gamma}, \tag{3.13}$$

wobei $\gamma \in \mathbb{R}$ und entweder $h \in \mathbb{N}$ fest gewählt oder die Differenz $n - h \in \mathbb{N}$ konstant ist − wir messen also entweder ab einem bestimmten Zyklus (h konstant) oder nur eine bestimmte Anzahl an Zyklen ($n - h$ konstant). Die so definierte Größe $\tilde{S}_{h,n,\nu,\gamma}$ wird auch *Kontrast* genannt. Wird $\psi \in \mathbb{R}$ so gewählt, dass die Funktion $v(N) = N^\psi$ die Identifizierbarkeitsrate von θ_0 ist, lässt sich Gleichung (3.13) umschreiben:

$$\tilde{S}_{h,n,\nu,\gamma}(\theta) := \sum_{k=h+1}^{n} \left(\frac{X_k}{Z_{k-1}^{\frac{1-\beta-2\psi}{2}}} + \delta_{k,\theta,\nu}\right)^2 W_{k-1}^{2(1-\psi)-\gamma} a_{k-1}^{2(1-\psi)-\gamma}, \tag{3.14}$$

mit

$$\begin{aligned} X_k &:= \left(Z_k - m_{\theta,\nu}(Z_{k-1})Z_{k-1}\right) Z_{k-1}^{-\frac{1+\beta}{2}} \\ \delta_{k,\theta,\nu} &:= \left(m_{\theta_0,\nu_0}(Z_{k-1}) - m_{\theta,\nu}(Z_{k-1})\right) Z_{k-1}^{\psi}. \end{aligned}$$

Der *normalisierte Kontrast* wird mit

$$D_n := \sum_{k=h+1}^{n} a_{k-1}^{2(1-\psi)-\gamma}$$

definiert als

$$S_{h,n,\nu,\gamma}(\theta) := \frac{\tilde{S}_{h,n,\nu,\gamma}(\theta)}{D_n}. \tag{3.15}$$

Die Forderung, dass der von h, n, ν und γ abhängende Schätzer, im Folgenden mit $\hat{\theta}_{h,n,\nu,\gamma}$ bezeichnet, Gleichung (3.13) erfüllen soll, ist äquivalent dazu, dass er $S_{h,n,\nu,\gamma}(\theta)$ minimiert. Der Schätzer wird deshalb definiert als:

$$\hat{\theta}_{h,n,\nu,\gamma} := \arg\min_{\theta\in\Theta} S_{h,n,\nu\,\gamma}(\theta)\,. \tag{3.16}$$

In Anlehnung an den Kontrast kann die zur „Abstandsmessung“ der Nachkommenmittel benötigte Folge von Halbnormen $(\|\cdot\|^a)_{a\in\mathbb{N}}$ nun näher spezifiziert werden. Für eine Funktion $u(\cdot)$ sei

$$\|u(\cdot)\|_n^2 := \frac{\sum_{k=h+1}^{n} u^2(Z_{k-1}) a_{k-1}^{2(1-\psi)-\gamma}}{\sum_{k=h+1}^{n} a_{k-1}^{2(1-\psi)-\gamma}}. \tag{3.17}$$

Durch das Überprüfen der definierenden Eigenschaften einer Halbnorm (Nichtnegativität, Homogenität, Dreiecksungleichung), ist leicht zu sehen, dass für alle $n \in \mathbb{N}$ durch (3.17) eine Halbnorm $\|\cdot\|_n$ auf dem Funktionenraum $\{f(N) : N \leq n\}$ definiert wird. Diese wird im Folgenden zur Überprüfung des Abstandes zwischen den Nachkommenmittelfunktionen verwendet.
Ist θ_0 gleichgradig asymptotisch identifizierbar in $m(\cdot)$ für $(\|\cdot\|_n)_{n\in\mathbb{N}}$ mit der Rate N^ψ und ist der Störparameter von $m(N)$ gleichgradig asymptotisch vernachlässigbar, so kann unter weiteren Voraussetzungen die Konsistenz der Folge $\left(\hat{\theta}_{h,n,\nu,\gamma}\right)_{n\in\mathbb{N}}$ gezeigt werden. Dazu wird folgendes Lemma benötigt (siehe [30]):

Lemma 3.10 (Lemma von Wu) *Es seien $S_{h,n,\nu,\gamma}$ und $\hat{\theta}_{h,n,\nu,\gamma}$ wie in (3.15) und (3.16) definiert. Gilt zusätzlich für jedes $\delta > 0$ die Bedingung*

$$\liminf_{n\to\infty} \inf_{|\theta-\theta_0|\geq\delta} (S_{h,n,\nu,\gamma}(\theta) - S_{h,n,\nu,\gamma}(\theta_0)) > 0 \text{ f.s.},$$

folgt

$$\hat{\theta}_{h,n,\nu,\gamma} \xrightarrow[n\to\infty]{} \theta_0 \text{ f.s.},$$

also die starke Konsistenz der Folge $\left(\hat{\theta}_{h,n,\nu,\gamma}\right)_{n\in\mathbb{N}}$.

Beweis: Angenommen, $\hat{\theta}_{h,n,\nu,\gamma} \xrightarrow[n\to\infty]{} \theta_0$ f.s. gelte nicht. Dann existiert ein $\delta > 0$, so dass $P\left(\limsup_{n\to\infty} \left|\hat{\theta}_{h,n,\nu,\gamma}(\cdot) - \theta_0\right| \geq \delta\right) > 0$. Mit der Definition der $\hat{\theta}_{h,n,\nu,\gamma}$ als $\arg\min_{\theta\in\Theta} S_{h,n,\nu,\gamma}(\theta)$ impliziert dies

$$P\left(\liminf_{n\to\infty} \inf_{|\theta-\theta_0|\geq\delta} (S_{h,n,\nu,\gamma}(\theta) - S_{h,n,\nu,\gamma}(\theta_0)) \leq 0\right) > 0$$

und das ist ein Widerspruch zur Voraussetzung, womit die Behauptung gezeigt ist. □

Nun kann die versprochene Konsistenzaussage formuliert und bewiesen werden:

Satz 3.11 *Die Folge der Halbnormen $(\|\cdot\|_n)_{n\in\mathbb{N}}$ sei durch Gleichung (3.17) definiert, $v(N) := N^\psi$. Es seien θ_0 gleichgradig asymptotisch identifizierbar in $m_{\theta_0,\cdot}(\cdot)$ für $(\|\cdot\|_n)_{n\in\mathbb{N}}$*

und die Funktion $(m_{\theta_0,\nu_0}(\cdot) - m_{\theta_0,\nu}(\cdot))\, v(\cdot)$ *gleichgradig asymptotisch vernachlässigbar. Desweiteren gelten noch die Voraussetzungen:*

(*i*) $\limsup_{n\to\infty} \sup_{\theta\in B_\delta^c,\nu\in\mathcal{N}} \|(m_{\theta_0,\nu}(\cdot) - m_{\theta,\nu}(\cdot))\, v(\cdot)\|_{n,\infty} < \infty$ *f.s.*

(*ii*) $D_n = \sum_{k=h+1}^{n} a_{k-1}^{2(1-\psi)-\gamma} \xrightarrow[n\to\infty]{} \infty$

(*iii*) *es gibt ein* h_0 *so dass,* $\sum_{k=h_0+1}^{\infty} a_{k-1}^{4(1-\psi)-2\gamma} a_{k-1}^{\beta+2\psi-1} \left(\sum_{l=h+1}^{k} a_{l-1}^{2(1-\psi)-\gamma} \right)^{-2} < \infty$

(*iv*) *für alle* $\delta > 0$ *und für alle* N *werde das Supremum und das Infimum von* $(m_{\theta_0,\nu}(N) - m_{\theta,\nu}(N))$ *bezüglich* (θ,ν) *angenommen.*

Dann ist $\left(\hat{\theta}_{h,n,\nu,\gamma}\right)_{n\in\mathbb{N}}$ *stark konsistent.*

Beweis: Die Behauptung wird mit Hilfe des Lemmas von Wu (Lemma 3.10) bewiesen. Dazu ist für alle $\delta > 0$ die Bedingung

$$\liminf_{n\to\infty} \inf_{|\theta-\theta_0|\geq\delta} (S_{h,n,\nu,\gamma}(\theta) - S_{h,n,\nu,\gamma}(\theta_0)) > 0 \text{ f.s.}$$

zu prüfen.
Es sei also $\delta > 0$ beliebig. Aus der Definition des normalisierten Kontrastes $S_{h,n,\nu,\gamma}$ in Gleichung (3.14) und (3.15) folgt mit $\xi_k := W_{k-1}^{2(1-\psi)-\gamma} a_{k-1}^{2(1-\psi)-\gamma}$ sowie $\rho := \frac{1-\beta-2\psi}{2}$:

$$\begin{aligned}
S_{h,n,\nu,\gamma}(\theta) - S_{h,n,\nu,\gamma}(\theta_0) &= \frac{1}{D_n} \sum_{k=h+1}^{n} \xi_k \left(\left(\frac{X_k}{Z_{k-1}^\rho} + \delta_{k,\theta,\nu} \right)^2 - \left(\frac{X_k}{Z_{k-1}^\rho} + \delta_{k,\theta_0,\nu} \right)^2 \right) \\
&= \frac{1}{D_n} \sum_{k=h+1}^{n} \xi_k \left(2\frac{X_k}{Z_{k-1}^\rho} (\delta_{k,\theta,\nu} - \delta_{k,\theta_0,\nu}) + \delta_{k,\theta,\nu}^2 - \delta_{k,\theta_0,\nu}^2 \right) \\
&\overset{(*)}{=} \underbrace{\frac{2}{D_n} \sum_{k=h+1}^{n} \xi_k \frac{X_k}{Z_{k-1}^\rho} \big(m_{\theta_0,\nu}(Z_{k-1}) - m_{\theta,\nu}(Z_{k-1})\big) Z_{k-1}^\psi}_{:=S_{3,n}(\theta)} \\
&\quad + \underbrace{\frac{2}{D_n} \sum_{k=h+1}^{n} \xi_k \delta_{k,\theta_0,\nu} \big(m_{\theta_0,\nu}(Z_{k-1}) - m_{\theta,\nu}(Z_{k-1})\big) Z_{k-1}^\psi}_{:=S_{2,n}(\theta)} \\
&\quad + \underbrace{\frac{1}{D_n} \sum_{k=h+1}^{n} \xi_k \big(m_{\theta_0,\nu}(Z_{k-1}) - m_{\theta,\nu}(Z_{k-1})\big)^2 Z_{k-1}^{2\psi}}_{:=S_{1,n}(\theta)} \\
&= S_{1,n}(\theta) + S_{2,n}(\theta) + S_{3,n}(\theta)
\end{aligned}$$

Die Gleichheit $(*)$ ergibt sich durch die Definition der $\delta_{k,\theta,\nu} = \big(m_{\theta_0,\nu_0}(Z_{k-1}) - m_{\theta,\nu}(Z_{k-1})\big) Z_{k-1}^\psi$ und das Auflösen der quadratischen Terme. Im weiteren Verlauf werden die drei Terme $S_{1,n}(\theta)$, $S_{2,n}(\theta)$ und $S_{3,n}(\theta)$ getrennt voneinander betrachtet.

Für $S_{1,n}(\theta)$ ergibt sich mit $v(N) = N^{\psi}$

$$\begin{aligned}
S_{1,n}(\theta) &= \frac{1}{D_n} \sum_{k=h+1}^{n} \xi_k \big(m_{\theta_0,\nu}(Z_{k-1}) - m_{\theta,\nu}(Z_{k-1})\big)^2 Z_{k-1}^{2\psi} \\
&= \frac{1}{D_n} \sum_{k=h+1}^{n} W_{k-1}^{2(1-\psi)-\gamma} a_{k-1}^{2(1-\psi)-\gamma} \big(m_{\theta_0,\nu}(Z_{k-1}) - m_{\theta,\nu}(Z_{k-1})\big)^2 Z_{k-1}^{2\psi} \\
&\geq \left(\inf_{h \leq k < n} W_k^{2(1-\psi)-\gamma} \right) \frac{\sum_{k=h+1}^{n} \big(m_{\theta_0,\nu}(Z_{k-1}) - m_{\theta,\nu}(Z_{k-1})\big)^2 Z_{k-1}^{2\psi} a_{k-1}^{2(1-\psi)-\gamma}}{\sum_{k=h+1}^{n} a_{k-1}^{2(1-\psi)-\gamma}} \\
&\overset{(3.17)}{\geq} \inf_{h \leq k < n} W_k^{2(1-\psi)-\gamma} \inf_{\nu \in \mathcal{N}} \big\| \big(m_{\theta_0,\nu}(\cdot) - m_{\theta,\nu}(\cdot)\big) v(\cdot) \big\|_n^2 .
\end{aligned}$$

Dies impliziert

$$\liminf_{n\to\infty} \inf_{\theta \in B_\delta^c} S_{1,n}(\theta) \geq \inf_{h \leq k < n} W_k^{2(1-\psi)-\gamma} \liminf_{n\to\infty} \inf_{\theta \in B_\delta^c, \nu \in \mathcal{N}} \big\| \big(m_{\theta_0,\nu}(\cdot) - m_{\theta,\nu}(\cdot)\big) v(\cdot) \big\|_n^2 .$$

Nach Satz 3.4 konvergiert die Folge $(W_n)_{n\geq 0}$ f.s. gegen eine Zufallsgröße W, die mit positiver Wahrscheinlichkeit größer 0 ist. Aus dieser Tatsache und der gleichgradig asymptotischen Identifizierbarkeit von θ_0 folgt

$$\liminf_{n\to\infty} \inf_{\theta \in B_\delta^c} S_{1,n}(\theta) > 0 \text{ f.s.} \tag{3.18}$$

Der zweite Term lässt sich mit der Hölder-Ungleichung (siehe [10] Kapitel VI Satz 1.5) abschätzen:

$$\begin{aligned}
|S_{2,n}(\theta)| &\leq 2 \sup_{h \leq k < n} W_{k-1}^{2(1-\psi)-\gamma} \\
&\quad \cdot \left| \frac{\sum_{k=h+1}^{n} \delta_{k,\theta_0,\nu} \big(m_{\theta_0,\nu}(Z_{k-1}) - m_{\theta,\nu}(Z_{k-1})\big) Z_{k-1}^{\psi} a_{k-1}^{2(1-\psi)-\gamma}}{\sum_{k=h+1}^{n} a_{k-1}^{2(1-\psi)-\gamma}} \right| \\
&\leq 2 \sup_{h \leq k < n} W_{k-1}^{2(1-\psi)-\gamma} \big\| \big(m_{\theta_0,\nu_0}(\cdot) - m_{\theta_0,\nu}(\cdot)\big) v(\cdot) \big\|_n \big\| \big(m_{\theta_0,\nu}(\cdot) - m_{\theta,\nu}(\cdot)\big) v(\cdot) \big\|_n
\end{aligned}$$

Da $\left|\liminf_{n\to\infty} \inf_{\theta \in B_\theta^c} S_{2n}(\theta)\right| \leq \limsup_{n\to\infty} \sup_{\theta \in B_\theta^c} |S_{2n}(\theta)|$ ergibt sich

$$\begin{aligned}
\left| \liminf_{n\to\infty} \inf_{\theta \in B_\theta^c} S_{2n}(\theta) \right| &\leq \limsup_{n\to\infty} \sup_{\theta \in B_\theta^c} |S_{2n}(\theta)| \\
&\leq 2 \sup_{h \leq k < n} W_{k-1}^{2(1-\psi)-\gamma} \limsup_{n\to\infty} \sup_{\theta \in B_\delta^c, \nu \in \mathcal{N}} \big\| \big(m_{\theta_0,\nu_0}(\cdot) - m_{\theta_0,\nu}(\cdot)\big) v(\cdot) \big\|_n \\
&\qquad \cdot \big\| \big(m_{\theta_0,\nu}(\cdot) - m_{\theta,\nu}(\cdot)\big) v(\cdot) \big\|_n \\
&= 0 \text{ f.s.}
\end{aligned}$$

Der Grenzübergang ist richtig, denn nach Satz 3.4 gilt $\lim_{n\to\infty} W_n = W$ f.s., nach Voraussetzung ist $\big(m_{\theta_0,\nu_0}(\cdot) - m_{\theta_0,\nu}(\cdot)\big) v(\cdot)$ gleichgradig asymptotisch vernachlässigbar und Voraussetzung (i) sichert, dass der letzte Term nicht über alle Schranken wächst. Es folgt

$$\liminf_{n\to\infty} \inf_{\theta \in B_\theta^c} S_{2n}(\theta) = 0 \text{ f.s.} \tag{3.19}$$

Nun wird der letzte Summand betrachtet. Durch Einsetzen der Definition von ξ_k ergibt sich

$$\begin{aligned} S_{3,n}(\theta) &= \frac{2}{D_n}\sum_{k=h+1}^{n}\xi_k\frac{X_k}{Z_{k-1}^{\rho}}\big(m_{\theta_0,\nu}(Z_{k-1})-m_{\theta,\nu}(Z_{k-1})\big)Z_{k-1}^{\psi} \\ &= \frac{2}{D_n}\sum_{k=h+1}^{n}\frac{X_k}{Z_{k-1}^{\rho}}\big(m_{\theta_0,\nu}(Z_{k-1})-m_{\theta,\nu}(Z_{k-1})\big)Z_{k-1}^{\psi}Z_{k-1}^{2(1-\psi)-\gamma} \\ &= \frac{2}{D_n}\sum_{k=h+1}^{n}\zeta_k(\theta,\nu) \end{aligned}$$

mit

$$\zeta_k(\theta,\nu) := \frac{X_k}{Z_{k-1}^{\rho}}\big(m_{\theta_0,\nu}(Z_{k-1})-m_{\theta,\nu}(Z_{k-1})\big)Z_{k-1}^{\psi}Z_{k-1}^{2(1-\psi)-\gamma}\,.$$

Es seien $(\theta_k^{min},\nu_k^{min}) := \arg\min_{\theta\in B_\delta^c,\nu\in\mathcal{N}}\zeta_k(\theta,\nu)$ und $(\theta_k^{max},\nu_k^{max}) := \arg\max_{\theta\in B_\delta^c,\nu\in\mathcal{N}}\zeta_k(\theta,\nu)$. Dann impliziert Voraussetzung (iv)

$$2\frac{\sum_{k=h+1}^{n}\zeta_k(\theta_k^{min},\nu_k^{min})}{D_n} \le \inf_{\theta\in B_\delta^c} S_{3,n}(\theta) \le 2\frac{\sum_{k=h+1}^{n}\zeta_k(\theta_k^{max},\nu_k^{max})}{D_n} \tag{3.20}$$

Im Folgenden werden zwei Fälle unterschieden:
1. Fall: Für konstantes h, wird gezeigt, dass die Folgen $\big(\sum_{k=h+1}^{n}\zeta_k(\theta_k^{min},\nu_k^{min})\big)_{n\ge1}$ und $\big(\sum_{k=h+1}^{n}\zeta_k(\theta_k^{max},\nu_k^{max})\big)_{n\ge1}$ Martingale bezüglich der Filtration $(\mathcal{F}_n)_{n\ge1}$, $\mathcal{F}_n := \sigma(Z_0,\dots,Z_n)$ sind. Dann liefert das starke Gesetz der großen Zahlen für Martingale A.3, dass die rechte und linke Seite der Ungleichung (3.20) f.s. verschwinden. Es sei $L_n := \sum_{k=h+1}^{n}\zeta_k(\theta_k^{max},\nu_k^{max})$. Da für alle $k \ge 1$ $E\left(|\zeta_k(\theta_k^{max},\nu_k^{max})|\,|\mathcal{F}_{k-1}\right) < \infty$ ist, gilt

$$\begin{aligned} E\left(L_n|\mathcal{F}_{n-1}\right) &= \underbrace{E\left(L_{n-1}|\mathcal{F}_{n-1}\right)}_{=L_{n-1}\,f.s.} + E\left(\zeta_n(\theta_n^{max},\nu_n^{max})|\mathcal{F}_{n-1}\right) \\ &= L_{n-1} + E\left[E\left(\zeta_n(\theta_n^{max},\nu_n^{max})|\mathcal{F}_{n-1},\theta_n^{max},\nu_n^{max}\right)|\mathcal{F}_{n-1}\right] \\ &= L_{n-1} + Z_{n-1}^{\frac{3+\beta-2\gamma}{2}}E\Big[\big(m_{\theta_0,\nu_n^{max}}(Z_{n-1})-m_{\theta_n^{max},\nu_n^{max}}(Z_{n-1})\big) \\ &\qquad\cdot E\left(X_n|\mathcal{F}_{n-1},\theta_n^{max},\nu_n^{max}\right)|\mathcal{F}_{n-1}\Big]\ \text{f.s.} \end{aligned}$$

Da X_n weder von θ_n^{max} noch von ν_n^{max} abhängt und $E(X_n|\mathcal{F}_{n-1}) = 0$ ist, folgt

$$E\left(L_n|\mathcal{F}_{n-1}\right) = L_{n-1}\ \text{f.s.},$$

d.h. $(L_n)_{n\ge1}$ ist ein Martingal bezüglich der Filtration $(\mathcal{F}_n)_{n\ge1}$.
Sei nun weiter $s_k := \frac{1}{D_k^2}E\left(\zeta_k^2(\theta_k^{max},\nu_k^{max})|\mathcal{F}_{k-1}\right)$. Nach Voraussetzung (ii) wächst die Folge $(D_n)_{n\ge1}$ über alle Grenzen und nach der dem Modell zu Grunde liegenden Voraussetzung V.4 gilt

$$\begin{aligned} \sum_{k=h+1}^{n}s_k &\le \sigma^2\sup_{\theta\in B_\delta^c,\nu\in\mathcal{N}}\sup_{h+1\le k\le n}\big(m_{\theta_0,\nu}(Z_{k-1})-m_{\theta,\nu}(Z_{k-1})\big)^2Z_{k-1}^{2\psi} \\ &\qquad\cdot\sup_{h+1\le k\le n}W_{k-1}^{2(2(1-\psi)-\gamma)-(1-\beta-2\psi)}\sum_{k=h+1}^{n}\frac{a_{k-1}^{2(2(1-\psi)-\gamma)-(1-\beta-2\psi)}}{D_k^2} \\ &\xrightarrow[n\to\infty]{} \infty\ \text{f.s.} \end{aligned}$$

Damit sind die Voraussetzungen des Satzes A.3 erfüllt und dieser liefert:

$$\lim_{n\to\infty} \frac{\sum_{k=h+1}^{n} \zeta_k(\theta_k^{max}, \nu_k^{max})}{D_n} = \lim_{n\to\infty} \frac{L_n}{D_n} = 0 \text{ f.s.}$$

Mit demselben Vorgehen erhalten wir

$$\lim_{n\to\infty} \frac{\sum_{k=h+1}^{n} \zeta_k(\theta_k^{min}, \nu_k^{min})}{D_n} = 0 \text{ f.s.}$$

Mit Blick auf Gleichung (3.20) impliziert dies:

$$\lim_{n\to\infty} \inf_{\theta\in B_\delta^c} S_{3,n}(\theta) = 0 \text{ f.s.}$$

2. Fall: $n-h$ ist konstant. Da h von n abhängt, ist die oben definierte Folge $(L_n)_{n\geq 1}$ kein Martingal. Mit $c := n-h$ ergibt sich aber eine Aufspaltung von L_n, nämlich

$$L_n = \sum_{k=c}^{n} \zeta_k\left(\theta_k^{max}, \nu_k^{max}\right)) - \sum_{k=c}^{n-c} \zeta_k\left(\theta_k^{max}, \nu_k^{max}\right)),$$

in denen die beiden Summen Martingale bezüglich der Filtration $(\mathcal{F}_n)_{n\geq 1}$ bilden. Folglich gilt mit Satz A.3 und Voraussetzung (ii)

$$\lim_{n\to\infty} \frac{\sum_{k=c}^{n} \zeta_k(\theta_k^{min}, \nu_k^{min})}{D_n} = 0 \text{ f.s.}$$

und da $D_n > D_{n-c}$ auch

$$\lim_{n\to\infty} \frac{\sum_{k=c}^{n-c} \zeta_k(\theta_k^{min}, \nu_k^{min})}{D_n} = 0 \text{ f.s.}$$

Mit $\zeta_k\left(\theta_k^{min}, \nu_k^{min}\right)$ wird analog verfahren und aus den beiden Fällen folgt:

$$\lim_{n\to\infty} \inf_{\theta\in B_\delta^c} S_{3,n}(\theta) = 0 \text{ f.s.} \tag{3.21}$$

Die Gleichungen (3.18), (3.19) und (3.21) liefern zusammen:

$$\liminf_{n\to\infty} \inf_{\theta\in B_\delta^c} (S_{h,n,\nu,\gamma}(\theta) - S_{h,n,\nu,\gamma}(\theta_0)) > 0 \text{ f.s.}$$

Da δ beliebig gewählt wurde, ist die Bedingung des Lemmas von Wu erfüllt und damit ist die Folge $\left(\hat{\theta}_{h,n,\nu,\gamma}\right)_{n\in\mathbb{N}}$ stark konsistent. □

Bei Einschränkung der Klasse der betrachteten Nachkommenmittelfunktionen kann nicht nur eine Konsistenzaussage bewiesen, sondern auch die Konvergenzgeschwindigkeit des Schätzers gegen den wahren Wert ermittelt werden (☞ Kapitel 3.4). Deswegen betrachten wir im Folgenden nur noch Modelle des Typs:

$$m_{\theta,\nu}(N) = m_1 \mathbb{1}_{\{\alpha_*>0\}} + m_2 N^{-\alpha_*} + r_{\theta,\nu}(N), \tag{3.22}$$

wobei $r_{\theta,\nu}(N) := \mathcal{O}\left(N^{-\alpha_{**}}\right)$ und $\alpha_{**} > \alpha_* \geq 0$. Zwei Schätzprobleme können betrachtet werden. Im ersten ist der zu schätzende Parameter $\theta_0 = m_1$. Dazu muss $\alpha_* = 0$ und m_2 ein Störparameter sein. Dies ist der Fall, wenn ν darstellbar ist als $\nu = (m_2, \nu')$. Im zweiten Fall sind m_1 und $\alpha_* > 0$ bekannt und $\theta_0 = m_2$ ist zu schätzen. In beiden Fällen ist die Funktion

$$(m_{\theta_0,\nu_0}(N) - m_{\theta_0,\nu}(N))N^{\alpha_*} = (r_{\theta_0,\nu_0}(N) - r_{\theta,\nu}(N))N^{\alpha_*}$$

offensichtlich gleichgradig asymptotisch vernachlässigbar, falls $\alpha_{**} > \alpha_*$ gilt. Die starke Konsistenz für den Schätzer $\left(\hat{\theta}_{h,n,\nu,\gamma}\right)_{n\in\mathbb{N}}$ liefert das folgende Korollar.

Korollar 3.12 *Gegeben sei ein größenabhängiger Verzweigungsprozess mit Nachkommenmittel der Form (3.22). Folgende Voraussetzungen seien zusätzlich erfüllt:*

(*i*) *für alle* $\delta > 0$ *und für alle* N *werde das Supremum und das Infimum von* $(m_{\theta_0,\nu}(N) - m_{\theta,\nu}(N))$ *bezüglich* (θ, ν) *angenommen*

(*ii*) *es existiere ein* $C < \infty$, *so dass für alle* $N \in \mathbb{N}$ *gilt* $\sup_{\theta,\nu} |r_{\theta,\nu}(N)|\, N^{\alpha_{**}} \leq C$

(*iii*) $\sum_{k=h+1}^{n} a_{k-1}^{2(1-\alpha_*)-\gamma} \xrightarrow[n\to\infty]{} \infty$

(*iv*) *es gibt ein* h_0 *so dass,* $\sum_{k=h_0+1}^{\infty} a_{k-1}^{4(1-\alpha_*)-2\gamma} a_{k-1}^{\beta+2\alpha_*-1} \left(\sum_{l=h+1}^{k} a_{l-1}^{2(1-\alpha_*)-\gamma} \right)^{-2} < \infty$

Dann ist $\left(\hat{\theta}_{h,n,\nu,\gamma}\right)_{n\in\mathbb{N}}$ *stark konsistent.*

Beweis: Sind die Voraussetzungen des Satzes 3.11 erfüllt, so liefert dieser die starke Konsistenz. Die gleichgradig asymptotische Vernachlässigbarkeit wurde oben gezeigt. Die gleichgradig asymptotische Identifizierbarkeit von θ_0 in $m_{\theta_0,\cdot}(\cdot)$ für $(\|\cdot\|_n)_{n\in\mathbb{N}}$ mit der Rate $v(N) := N^{\alpha_*}$, d.h. $\psi = \alpha_*$, folgt leicht aus

$$(m_{\theta_0,\nu}(N) - m_{\theta,\nu}(N))\, v(N) = (r_{\theta_0,\nu}(N) - r_{\theta,\nu}(N))\, N^{\alpha_*}\,.$$

Es bleibt nur noch Voraussetzung (i) des Satzes 3.11 zu überprüfen. Dazu betrachte

$$\begin{aligned}
\|(m_{\theta_0,\nu}(\cdot) &- m_{\theta,\nu}(\cdot))v(\cdot)\|_n^2 \\
&= \frac{1}{D_n} \sum_{k=h+1}^{n} \left((\theta_0-\theta) Z_{k-1}^{-\alpha_*} + r_{\theta_0,\nu}(Z_{k-1}) - r_{\theta,\nu}(Z_{k-1})\right)^2 Z_{k-1}^{2\alpha_*} a_{k-1}^{2(1-\alpha_*)-\gamma} \\
&= (\theta_0-\theta)^2 \frac{\sum_{k=h+1}^{n} a_{k-1}^{2(1-\alpha_*)-\gamma}}{D_n} + \|(r_{\theta_0,\nu}(\cdot) - r_{\theta,\nu}(\cdot))\, v(\cdot)\|_n^2 \\
&\qquad \underbrace{+ \frac{2(\theta_0-\theta)}{D_n} \sum_{k=h+1}^{n} (r_{\theta_0,\nu}(Z_{k-1}) - r_{\theta_0,\nu}(Z_{k-1}))\, Z_{k-1}^{\alpha_*} a_{k-1}^{2(1-\alpha_*)-\gamma}}_{=:R_{\theta,\nu,n}} \\
&= (\theta_0-\theta)^2 + \|(r_{\theta_0,\nu}(\cdot) - r_{\theta,\nu}(\cdot))\, v(\cdot)\|_n^2 + R_{\theta,\nu,n}
\end{aligned}$$

Mit Θ ist auch $\theta_0 - \theta$ beschränkt. Aus Voraussetzung (ii) folgt die Beschränktheit von $\|(r_{\theta_0,\nu}(\cdot) - r_{\theta,\nu}(\cdot))\, v(\cdot)\|_n^2$. Wegen der Voraussetzungen (iii) und (iv) ist auch $R_{\theta,\nu,n}$ beschränkt und dies liefert

$$\limsup_{n\to\infty} \sup_{\theta\in B_\delta^c, \nu\in\mathcal{N}} \|(m_{\theta_0,\nu}(\cdot) - m_{\theta,\nu}(\cdot))v(\cdot)\|_{n,\infty} < \infty.$$

Insgesamt folgt die Behauptung. □

Unter einer weiteren Zusatzbedingung lässt sich im Modell (3.22) die starke Konsistenz auf eine weitere Art zeigen. Der Beweis basiert auf der Taylor-Entwicklung der ersten Ableitung des Kontrastes.

Satz 3.13 *Zusätzlich zu den Voraussetzungen (i)-(iv) des Korollars 3.12 gelten die Voraussetzungen:*

(v) *für alle N sei die Funktion $r_{\theta,\nu}(N)$ zweimal in einer Umgebung von θ_0 in θ stetig differenzierbar und es existieren $C', C'' < \infty$, so dass für alle $N \in \mathbb{N}$ die Suprema* $\sup_{\theta\in\Theta,\nu\in\mathcal{N}} \left|r'_{\theta,\nu}(N)\right| N^{\alpha_{**}} \leq C'$ *und* $\sup_{\theta\in\Theta,\nu\in\mathcal{N}} \left|r''_{\theta,\nu}(N)\right| N^{\alpha_{**}} \leq C''$ *sind*

(vi) *für alle $N \in \mathbb{N}$ werden das Supremum und das Infimum der zweiten Ableitung $r''_{\theta,\nu}(N)$ in einem $\left(\theta_N^{sup}, \nu_N^{sup}\right)$ bzw. $\left(\theta_N^{inf}, \nu_N^{inf}\right)$ angenommen*

Dann ist $\left(\hat{\theta}_{h,n,\nu,\gamma}\right)_{n\in\mathbb{N}}$ stark konsistent.

Beweis: Es sei mit $\tilde{S}'_{h,n,\nu,\gamma}$ die erste Ableitung bezüglich θ bezeichnet, d.h.

$$\tilde{S}'_{h,n,\nu,\gamma} := \frac{\mathrm{d}\tilde{S}_{h,n,\nu,\gamma}}{\mathrm{d}\theta}.$$

Nach dem Satz von Taylor (siehe [14] Satz 61.1) angewendet auf $\tilde{S}'_{h,n,\nu,\gamma}$ in θ_0 gilt

$$\tilde{S}'_{h,n,\nu,\gamma}(\theta) = \tilde{S}'_{h,n,\nu,\gamma}(\theta_0) + \tilde{S}''_{h,n,\nu,\gamma}(\theta_n) \cdot (\theta - \theta_0)$$

für ein $\theta_n \in (\theta \wedge \theta_0, \theta \vee \theta_0)$. Nach der Definition des Schätzers $\hat{\theta}_{h,n,\nu,\gamma}$ in Gleichung (3.16) als der Wert, in dem $\tilde{S}_{h,n,\nu,\gamma}$ ein Minimum annimmt, also die erste Ableitung in diesem Wert gleich 0 ist, folgt daraus unter der Annahme, dass in einer Umgebung von θ_0 die zweite Ableitung $\tilde{S}''_{h,n,\nu,\gamma}$ nicht verschwindet,

$$\hat{\theta}_{h,n,\nu,\gamma} - \theta_0 = -\frac{\tilde{S}'_{h,n,\nu,\gamma}(\theta_0)}{\tilde{S}''_{h,n,\nu,\gamma}(\theta_n)} = \frac{1}{Q_{h,n,\nu,\gamma}} \frac{L_n}{D_n} \tag{3.23}$$

für ein $\theta_n \in \left(\hat{\theta}_{h,n,\nu,\gamma} \wedge \theta_0 \, , \, \hat{\theta}_{h,n,\nu,\gamma} \vee \theta_0\right)$. Dabei sind

$$L_n := \sum_{k=h+1}^{n} \left(Z_k - m_{\theta_0,\nu}(Z_{k-1})Z_{k-1}\right) m'_{\theta_0,\nu}(Z_{k-1}) Z_{k-1}^{1-\gamma}$$

und

$$Q_{h,n,\nu,\gamma} := \frac{1}{D_n} \sum_{k=h+1}^{n} \left(\sum_{i=1}^{3} U_{i,k}(\theta_n)\right) W_{k-1}^{2(1-\alpha_*)-\gamma} a_{k-1}^{2(1-\alpha_*)-\gamma} \tag{3.24}$$

mit

$$\begin{aligned}
U_{1,k}(\theta_n) &:= m'^{\,2}_{\theta_n,\nu}(Z_{k-1}) Z_{k-1}^{2\alpha_*}\,, \\
U_{2,k}(\theta_n) &:= -\left(m_{\theta_0,\nu_0}(Z_{k-1}) - m_{\theta_n,\nu}(Z_{k-1})\right) m''_{\theta_n,\nu}(Z_{k-1}) Z_{k-1}^{2\alpha_*}\,, \\
U_{3,k}(\theta_n) &:= -X_k Z_{k-1}^{\frac{\beta-1}{2}} m''_{\theta_n,\nu}(Z_{k-1})\,.
\end{aligned}$$

Nun folgen zwei Schritte. Im ersten wird gezeigt, dass

$$\lim_{n\to\infty} \frac{L_n}{D_n} = 0 \text{ f.s.} \tag{3.25}$$

gilt und im zweiten

$$\lim_{n\to\infty} Q_{h,n,\nu,\gamma} = W^{2(1-\alpha_*)-\gamma} \text{ f.s.} \tag{3.26}$$

Die Gleichungen (3.23), (3.25) und (3.26) implizieren dann, dass $\hat{\theta}_{h,n,\nu,\gamma} \xrightarrow[n\to\infty]{} \theta_0$ f.s.

1. Schritt: Es seien

$$\zeta_k := (Z_k - m_{\theta_0,\nu_0}(Z_{k-1})Z_{k-1})\, m'_{\theta_0,\nu}(Z_{k-1})Z_{k-1}^{1-\gamma}$$

und

$$\bar{r}_{\theta_0,\nu} := r_{\theta_0,\nu_0} - r_{\theta_0,\nu} = m_{\theta_0,\nu_0} - m_{\theta_0,\nu}.$$

Dann ist

$$\begin{aligned}
\frac{L_n}{D_n} &= \frac{1}{D_n}\sum_{k=h+1}^{n} [Z_k - m_{\theta_0,\nu}(Z_{k-1})Z_{k-1}]\, m'_{\theta_0,\nu}(Z_{k-1})Z_{k-1}^{1-\gamma} \\
&= \frac{1}{D_n}\sum_{k=h+1}^{n} [Z_k - m_{\theta_0,\nu}(Z_{k-1})Z_{k-1} + m_{\theta_0,\nu_0}(Z_{k-1})Z_{k-1} - m_{\theta_0,\nu_0}(Z_{k-1})Z_{k-1}] \\
&\qquad\qquad \cdot m'_{\theta_0,\nu}(Z_{k-1})Z_{k-1}^{1-\gamma} \\
&= \frac{1}{D_n}\sum_{k=h+1}^{n} [Z_k - m_{\theta_0,\nu_0}(Z_{k-1})Z_{k-1} + \bar{r}_{\theta_0,\nu}(Z_{k-1})Z_{k-1}]\, m'_{\theta_0,\nu}(Z_{k-1})Z_{k-1}^{1-\gamma} \\
&= \frac{\sum_{k=h+1}^{n}\zeta_k}{D_n} + \frac{\sum_{k=h+1}^{n}\bar{r}_{\theta_0,\nu}(Z_{k-1})m'_{\theta_0,\nu}(Z_{k-1})Z_{k-1}^{2-\gamma}}{D_n} \qquad (3.27)
\end{aligned}$$

Die beiden Terme dieser Gleichung werden nun einzeln betrachtet. Da für $n > h+1$ die Gleichung

$$\begin{aligned}
E\left(\sum_{k=h+1}^{n}\zeta_k|\mathcal{F}_{n-1}\right) &= E(\zeta_n|\mathcal{F}_n) + \sum_{k=h+1}^{n-1}\zeta_k \\
&= E\left((Z_n - m_{\theta_0,\nu_0}(Z_{n-1})Z_{n-1})\, m'_{\theta_0,\nu}(Z_{n-1})Z_{n-1}^{1-\gamma}|\mathcal{F}_n\right) + \sum_{k=h+1}^{n-1}\zeta_k \\
&= \underbrace{(E(Z_n|\mathcal{F}_n) - m_{\theta_0,\nu_0}(Z_{n-1})Z_{n-1})}_{=0 \text{ f.s.}}\, m'_{\theta_0,\nu}(Z_{n-1})Z_{n-1}^{1-\gamma} + \sum_{k=h+1}^{n-1}\zeta_k \\
&= \sum_{k=h+1}^{n-1}\zeta_k \text{ f.s.}
\end{aligned}$$

gilt, handelt es sich bei der Summenfolge $\left(\sum_{k=h+1}^{n}\zeta_k\right)_{n>h+1}$ um ein Martingal. Das starke Gesetz der großen Zahlen für Martingale, Satz A.3, liefert mit Voraussetzung (iii), dass $\frac{\sum_{k=h+1}^{n}\zeta_k}{D_n}$ für $n\to\infty$ fast sicher verschwindet.
Der zweite Term der rechten Seite der Gleichung (3.27) konvergiert auch für $n\to\infty$ fast sicher gegen 0. Dazu werden die einzelnen Summanden mit Hilfe der Voraussetzungen (ii) und (v) wie folgt abgeschätzt

$$\begin{aligned}
\left|\bar{r}_{\theta_0,\nu}(Z_{k-1})m'_{\theta_0,\nu}(Z_{k-1})\right| Z_{k-1}^{2-\gamma} &\le 2CZ_{k-1}^{-\alpha_{**}}\left|m'_{\theta_0,\nu}(Z_{k-1})\right| Z_{k-1}^{2-\gamma} \\
&\le 2CZ_{k-1}^{-\alpha_{**}}\left(Z_{k-1}^{-\alpha_*} + C'Z_{k-1}^{-\alpha_{**}}\right) Z_{k-1}^{2-\gamma} \\
&= 2CZ_{k-1}^{2(1-\alpha_*)-\gamma}\left(Z_{k-1}^{\alpha_*-\alpha_{**}} + C'Z_{k-1}^{2(\alpha_*-\alpha_{**})}\right) \\
&= a_{k-1}^{2(1-\alpha_*)-\gamma}2CW_{k-1}^{2(1-\alpha_*)-\gamma}\left(Z_{k-1}^{\alpha_*-\alpha_{**}} + C'Z_{k-1}^{2(\alpha_*-\alpha_{**})}\right)
\end{aligned}$$

Satz 3.4 garantiert, dass $W_{k-1}^{2(1-\alpha_*)-\gamma}$ für $k\to\infty$ fast sicher gegen $W^{2(1-\alpha_*)-\gamma}$ konvergiert. Da $\alpha_{**} > \alpha_*$, konvergiert $\left(Z_{k-1}^{\alpha_*-\alpha_{**}} + C'Z_{k-1}^{2(\alpha_*-\alpha_{**})}\right)$ fast sicher gegen 0 und Voraussetzung

(iii) sichert, dass $\sum_{k=h+1}^{n} a_{k-1}^{2(1-\alpha_*)-\gamma}$ über alle Schranken wächst. Damit liefert das Lemma von Toeplitz (siehe A.4 (iii)):

$$\frac{1}{D_n}\sum_{k=h+1}^{n} a_{k-1}^{2(1-\alpha_*)-\gamma} 2C W_{k-1}^{2(1-\alpha_*)-\gamma}\left(Z_{k-1}^{\alpha_*-\alpha_{**}} + C' Z_{k-1}^{2(\alpha_*-\alpha_{**})}\right) \xrightarrow[n\to\infty]{} 0 \text{ f.s.}$$

Folglich konvergiert auch $\frac{1}{D_n}\sum_{k=h+1}^{n} \bar{r}_{\theta_0,\nu}(Z_{k-1}) m'_{\theta_0,\nu}(Z_{k-1}) Z_{k-1}^{2-\gamma}$ fast sicher gegen 0 und dasselbe gilt für $\frac{L_n}{D_n}$ in Gleichung (3.27).
2. Schritt: $Q_{h,n,\nu,\gamma}$ wird in zwei Teile aufgespaltet und diese werden getrennt voneinander untersucht. Der erste Teil besteht aus dem Term mit den Summanden $U_{1,k}(\theta_n)$ und $U_{2,k}(\theta_n)$. Dazu betrachte:

$$\begin{aligned}
&\Big|\frac{1}{D_n}\sum_{k=h+1}^{n}\big(U_{1,k}(\theta_n)+U_{2,k}(\theta_n)\big)W_{k-1}^{2(1-\alpha_*)-\gamma}a_{k-1}^{2(1-\alpha_*)-\gamma} - W^{2(1-\alpha_*)-\gamma}\Big| \\
&\quad\le \frac{1}{D_n}\sum_{k=h+1}^{n}\sup_{\theta\in\Theta,\nu\in\mathcal{N}}\big(|U_{1,k}(\theta)|+|U_{2,k}(\theta)|\big)\left|W_{k-1}^{2(1-\alpha_*)-\gamma} - W^{2(1-\alpha_*)-\gamma}\right| a_{k-1}^{2(1-\alpha_*)-\gamma} \\
&\qquad + \frac{1}{D_n}\sum_{k=h+1}^{n}\sup_{\theta\in\Theta,\nu\in\mathcal{N}}\big(|U_{1,k}(\theta)-1|+|U_{2,k}(\theta)|\big)W^{2(1-\alpha_*)-\gamma}a_{k-1}^{2(1-\alpha_*)-\gamma} \qquad (3.28)
\end{aligned}$$

Nach der Definiton von $U_{1,k}$ und $U_{2,k}$ gilt, dass $\sup_{\theta\in\Theta,\nu\in\mathcal{N}}\big(|U_{1,k}(\theta)|+|U_{2,k}(\theta)|\big)$ beschränkt ist. Nach Satz 3.4 konvergiert die Folge der $\left|W_n^{2(1-\alpha_*)-\gamma} - W^{2(1-\alpha_*)-\gamma}\right|$ fast sicher gegen 0. Damit sind wie in Schritt 1 die Voraussetzungen für das Lemma von Toeplitz (A.4 (iii)) erfüllt und der erste Term auf der rechten Seite von Gleichung (3.28) konvergiert fast sicher gegen 0. Weiter gilt wegen Voraussetzung (v) und $m'_{\theta,\nu}(N) = N^{-\alpha_*} + r'_{\theta,\nu}(Z_{k-1})$

$$\begin{aligned}
\sup_{\theta\in\Theta,\nu\in\mathcal{N}}|U_{1,k}(\theta)-1| &= \sup_{\theta\in\Theta,\nu\in\mathcal{N}}\left|\big(Z_{k-1}^{-\alpha_*}+r'_{\theta,\nu}(Z_{k-1})\big)^2 Z_{k-1}^{2\alpha_*}-1\right| \\
&= \sup_{\theta\in\Theta,\nu\in\mathcal{N}}\left|2Z_{k-1}^{\alpha_*}r'_{\theta,\nu}(Z_{k-1}) + r'^{\,2}_{\theta,\nu}(Z_{k-1})Z_{k-1}^{2\alpha_*}\right| \\
&\le 2C' Z_{k-1}^{\alpha_*-\alpha_{**}} + C'^2 Z_{k-1}^{2(\alpha_*-\alpha_{**})}\,.
\end{aligned}$$

Da $m_{\theta_0,\nu_0}(N) - m_{\theta,\nu}(N) = (\theta_0-\theta)Z_{k-1}^{-\alpha_*} + \bar{r}_{\theta,\nu}(Z_{k-1})$, folgt mit den Voraussetzungen (i) und (v)

$$\begin{aligned}
\sup_{\theta\in\Theta,\nu\in\mathcal{N}}|U_{2,k}(\theta)| &= \sup_{\theta\in\Theta,\nu\in\mathcal{N}}\left|(\theta_0-\theta)Z_{k-1}^{-\alpha_*}+\bar{r}_{\theta,\nu}(Z_{k-1})\right|\left|m''_{\theta,\nu}(Z_{k-1})\right| Z_{k-1}^{2\alpha_*} \\
&\le \big(|\theta_0-\theta|Z_{k-1}^{-\alpha_*}+|\bar{r}_{\theta,\nu}(Z_{k-1})|\big)\,C'' Z_{k-1}^{2\alpha_*-\alpha_{**}} \\
&\le \Big(\underbrace{|\theta_0-\theta|}_{<\infty \text{ da } \Theta \text{ beschränkt}} + 2C Z_{k-1}^{\alpha_*-\alpha_{**}}\Big)C'' Z_{k-1}^{\alpha_*-\alpha_{**}}\,.
\end{aligned}$$

Wegen $\alpha_{**} > \alpha_*$ sind auch hier die Bedingungen des Lemmas von Toeplitz erfüllt und damit geht der zweite Term in Gleichung (3.28) fast sicher gegen 0 für $n\to\infty$. Dies impliziert, dass

$$\lim_{n\to\infty}\frac{1}{D_n}\sum_{k=h+1}^{n}\left(\sum_{i=1}^{2}U_{i,k}(\theta_n)\right)W_{k-1}^{2(1-\alpha_*)-\gamma}a_{k-1}^{2(1-\alpha_*)-\gamma} = W^{2(1-\alpha_*)-\gamma} \text{ f.s.}$$

Der zweite Teil von $Q_{h,n,\nu,\gamma}$, also $\frac{1}{D_n}\sum_{k=h+1}^{n}U_{3,k}(\theta_n)W_{k-1}^{2(1-\alpha_*)-\gamma}a_{k-1}^{2(1-\alpha_*)-\gamma}$, wird ähnlich wie im Beweis zu Satz 3.11 abgeschätzt. Dazu werden die Voraussetzungen (iv) und (vi)

benötigt. Diese Argumentation, auf deren ausführliche Behandlung an dieser Stelle verzichtet wird, liefert

$$\lim_{n\to\infty} \frac{1}{D_n} \sum_{k=h+1}^{n} U_{3,k}(\theta_n) W_{k-1}^{2(1-\alpha_*)-\gamma} a_{k-1}^{2(1-\alpha_*)-\gamma} = 0 \text{ f.s.}$$

Daraus folgt Gleichung (3.26).
Die Ergebnisse des 1. und 2. Schrittes ergeben zusammen die Behauptung. □

3.4. Die Konvergenzrate des Schätzers

In diesem Abschnitt wird die Frage nach der Konvergenzgeschwindigkeit des im vorigen Abschnitt ausführlich auf Konsistenz geprüften Schätzers $\hat{\theta}_{h,n,\nu,\gamma}$ gegen den wahren Wert θ_0 beantwortet. Dazu wird die Klasse der Modelle weiter eingeschränkt und zwar zu solchen, deren Varianz strengeren Bedingungen genügt, als das bisher der Fall war. Genauer gelte für die Varianz für alle $N \geq 1$ im superkritischen Fall und für alle $N \geq \tilde{N}$ für ein geeignetes $\tilde{N}$ im fast-kritischen Fall die Gleichung

$$\sigma^2(N) = \sigma^2 N^\beta - r_+(N) \tag{3.29}$$

mit $r_+(N) = \mathcal{O}\left(N^{\bar{\beta}}\right)$ wobei $\bar{\beta} < \beta$, $r_+(N) \geq 0$ und unabhängig von θ und ν seien. Benötigt werden noch folgende Größen:

$$\begin{aligned}
B_n &:= \sqrt{\sum_{k=h+1}^{n} a_{k-1}^{1+\beta+2(1-\alpha_*)-2\gamma}} \\
B_{k,n} &:= B_n a_{k-1}^{\frac{-\beta-2(1-\alpha_*)+2\gamma}{2}} \\
R_k &:= (I_{k,1} + 1 - m_{\theta_0,\nu_0}(Z_{k-1}))\, m'_{\theta_0,\nu}(Z_{k-1}) W_{k-1}^{1-\gamma} a_{k-1}^{-\frac{\beta}{2}+\alpha_*} \\
U_{h,n,\nu,\gamma} &:= \frac{1}{B_n} \sum_{k=h+1}^{n} \sum_{i=1}^{Z_{k-1}} (I_{k,i} + 1 - m_{\theta_0,\nu}(Z_{k-1}))\, m'_{\theta_0,\nu}(Z_{k-1}) Z_{k-1}^{1-\gamma} \\
\Phi_{h,n,\gamma} &:= \frac{B_n}{D_n}
\end{aligned}$$

Mit diesen Größen kann die im Beweis zu Satz 3.13 durch Ableiten des Kontrastes erhaltene Gleichung (3.23) umgeschrieben werden zu

$$\hat{\theta}_{h,n,\nu,\gamma} - \theta_0 = \frac{U_{h,n,\nu,\gamma}\Phi_{h,n,\gamma}}{Q_{h,n,\nu,\gamma}}\,. \tag{3.30}$$

Nun kann der angekündigte Grenzwertsatz formuliert werden.

Satz 3.14 *Zusätzlich zu den Voraussetzungen des Satzes 3.13 gelte noch:*

(*vii*) *Für festes h sei*

$$\lim_{n\to\infty} B_n = \infty\,.$$

(*viii*) *Für $\nu \neq \nu_0$ sei*

$$\lim_{n\to\infty} \frac{1}{B_n} \sum_{k=h+1}^{n} a_{k-1}^{2-\gamma-\alpha_{**}-\alpha_*} = 0\,.$$

(*ix*) *Für alle $x \in \mathbb{R}$ sei*

$$\lim_{h\to\infty} \lim_{n\to\infty} \sup_{h+1\leq k\leq n} E\left(R_k^2 \mathbb{1}_{\{R_k^2 \geq B_{k,n}^2 x^2\}} | \mathcal{F}_{k-1}\right) = 0 \text{ f.s.}$$

Dann existiert eine Zufallsgröße U, so dass

$$\lim_{n\to\infty} U_{h,n,\nu,\gamma} \stackrel{D}{=} U. \tag{3.31}$$

Die Fourier-Transformierte dieser Zufallsgröße ist

$$\phi_U(s) = Ee^{isU} = E\left(\exp\left(-\frac{s^2}{2}\sigma^2 W^{1+\beta+2(1-\alpha_*)-2\gamma}\right)\right) . \tag{3.32}$$

Außerdem gilt

$$\lim_{n\to\infty} Q_{h,n,\nu,\gamma} = W^{2(1-\alpha_*)-\gamma}. \tag{3.33}$$

Desweiteren wird die beste Konvergenzrate erreicht, wenn $\gamma = 1+\beta$ ist. Zwei Fälle sind möglich:

(α) *Falls $\beta + 2\alpha_* < 1$ ist, gilt im superkritischen Fall*

$$\Phi_{h,n,\gamma} = \mathcal{O}\left(a_n^{\frac{2\alpha_*+\beta-1}{2}}\right)$$

und im fast-kritischen Fall

$$\Phi_{h,n,\gamma} \le a_h^{\frac{2\alpha_*+\beta-1}{2}} \frac{1}{\sqrt{n-h}} .$$

(β) *Falls $\beta + 2\alpha_* = 1$ ist, gilt*

$$\Phi_{h,n,\gamma} = \frac{1}{\sqrt{n-h}} .$$

Beweis: Gleichung (3.33) ist ein Zwischenergebnis des Beweises zu Satz 3.13 (☞ Gleichung (3.26)).

Die asymptotische Verteilung in Gleichung (3.31) und die Fourier-Transformierte in Gleichung (3.32) folgen aus dem Satz A.5 nach Rahimov, dessen Voraussetzungen im Folgenden zu prüfen sind. Dazu seien folgende Größen definiert:

$$\begin{aligned}
\xi_{k,i} &:= (I_{k,i} - m_{\theta_0,\nu}(Z_{k-1}))\, m'_{\theta_0,\nu}(Z_{k-1}) Z_{k-1}^{1-\gamma} \mathbb{1}_{\{k\ge h+1\}} \\
\vartheta_{k,i} &:= E\left(\xi_{k,i}|\mathcal{F}_{k-1}\right) = \bar{r}_{\theta_0,\nu}(Z_{k-1}) m'_{\theta_0,\nu}(Z_{k-1}) Z_{k-1}^{1-\gamma} \mathbb{1}_{\{k\ge h+1\}} \\
\sigma^2_{k,i} &:= E\left[(\xi_{k,i} - \vartheta_{k,i})^2 |\mathcal{F}_{k-1}\right] = \sigma^2(Z_{k-1}) m'^{\,2}_{\theta_0,\nu}(Z_{k-1}) Z_{k-1}^{2(1-\gamma)} \mathbb{1}_{\{k\ge h+1\}}
\end{aligned}$$

Voraussetzung (A.9): Betrachte

$$\begin{aligned}
\frac{1}{B_n^2}\sum_{k=1}^{n}\sum_{i=1}^{Z_{k-1}} \sigma^2_{k,i} &= \frac{1}{B_n^2}\sum_{k=1}^{n}\sum_{i=1}^{Z_{k-1}} \sigma^2(Z_{k-1}) m'^{\,2}_{\theta_0,\nu}(Z_{k-1}) Z_{k-1}^{2(1-\gamma)} \mathbb{1}_{\{k\ge h+1\}} \\
&= \frac{1}{B_n^2}\sum_{k=h+1}^{n} Z_{k-1}\sigma^2(Z_{k-1}) m'^{\,2}_{\theta_0,\nu}(Z_{k-1}) Z_{k-1}^{2(1-\gamma)} \\
&= \frac{1}{B_n^2}\sum_{k=h+1}^{n} \sigma^2(Z_{k-1}) Z_{k-1}^{-\beta} m'^{\,2}_{\theta_0,\nu}(Z_{k-1}) Z_{k-1}^{2\alpha_*} \\
&\qquad \cdot W_{k-1}^{2(1-\gamma)-2\alpha_*+\beta+1} a_{k-1}^{2(1-\gamma)-2\alpha_*+\beta+1}
\end{aligned}$$

Aus Voraussetzung (v) und $\alpha_{**} > \alpha_*$ folgt

$$m'_{\theta_0,\nu}(N)N^{\alpha_*} = N^{-\alpha_*} + r'_{\theta_0,\nu}(N)N^{\alpha_*} = 1 + C'N^{\alpha_*-\alpha_{**}} \xrightarrow[N\to\infty]{} 1 , \tag{3.34}$$

Da $\bar{\beta} < \beta$, liefert Voraussetzung (3.29)

$$\sigma^2(N)N^\beta = \sigma^2 - \mathcal{O}(N^{\bar{\beta}})N^{-\beta} \uparrow \sigma^2 \quad \text{für } N \to \infty\,. \tag{3.35}$$

Mit Satz 3.4 und Voraussetzung (vii) sind alle Voraussetzungen des Lemmas von Toeplitz (A.4 (iii)) erfüllt. Dann folgt:

$$\lim_{n\to\infty} \frac{1}{B_n^2} \sum_{k=1}^{n} \sum_{i=1}^{Z_{k-1}} \sigma_{k,i}^2 = \sigma^2 W^{2(1-\gamma)-2\alpha_*+\beta+1} \text{ f.s.}$$

Es sei nun T eine von $\sigma^2 W^{2(1-\gamma)-2\alpha_*+\beta+1}$ nach unten begrenzte Zufallsgröße, dann gilt

$$\lim_{n\to\infty} P\left(\frac{1}{B_n^2} \sum_{k=1}^{n} \sum_{i=1}^{Z_{k-1}} \sigma_{k,i}^2 > T \right) = 0\,.$$

Dies war zu prüfen.
Voraussetzung (A.10): Nach Gleichung (3.35) gilt:

$$\begin{aligned}\max_{h+1\le k\le n} \frac{\sigma^2(Z_{k-1}) Z_{k-1}^{2(1-\gamma)} m'^{\,2}_{\theta_0,\nu}(Z_{k-1})}{B_n^2} \quad &\le \quad \sigma^2 \max_{h+1\le k\le n} W_{k-1}^{\beta+2(1-\alpha_*)-2\gamma} m'^{\,2}_{\theta_0,\nu}(Z_{k-1}) Z_{k-1}^{2\alpha_*} \\ &\quad \cdot \frac{1}{B_n^2} \max_{h+1\le k\le n} a_{k-1}^{\beta+2(1-\alpha_*)-2\gamma}\end{aligned}$$

Nach Satz 3.4 und Gleichung (3.34) ist der Grenzwert des ersten Faktors endlich, d.h.

$$\lim_{n\to\infty} \max_{h+1\le k\le n} W_{k-1}^{\beta+2(1-\alpha_*)-2\gamma} m'^{\,2}_{\theta_0,\nu}(Z_{k-1}) Z_{k-1}^{2\alpha_*} < \infty$$

Der Grenzwert des zweiten Faktors ist 0, was wie folgt gesehen werden kann.
Ist $\beta + 2(1-\alpha_*) - 2\gamma \le 0$, so wird das Maximum in h angenommen, genauer

$$\frac{1}{B_n^2} \max_{h+1\le k\le n} a_{k-1}^{\beta+2(1-\alpha_*)-2\gamma} = \frac{a_h^{\beta+2(1-\alpha_*)-2\gamma+1}}{B_n^2} a_h^{-1}\,.$$

Bei festem h folgt aus Voraussetzung (vii), dass

$$\lim_{n\to\infty} \frac{1}{B_n^2} \max_{h+1\le k\le n} a_{k-1}^{\beta+2(1-\alpha_*)-2\gamma} = 0 \tag{3.36}$$

gilt. Bei festem $n-h$ ist $a_h^{\beta+2(1-\alpha_*)-2\gamma+1} \le B_n^2$ und $\lim_{n\to\infty} a_{h-1}^{-1} = 0$, was (3.36) impliziert.
Ist $\beta + 2(1-\alpha_*) - 2\gamma > 0$, so gilt

$$\frac{1}{B_n^2} \max_{h+1\le k\le n} a_{k-1}^{\beta+2(1-\alpha_*)-2\gamma} \le a_{n-1}^{-1}\,,$$

woraus (3.36) folgt. In beiden Fällen gilt also

$$\lim_{n\to\infty} \max_{h+1\le k\le n} \frac{\sigma_{k,i}^2}{B_n^2} = 0 \text{ f.s.}$$

und da aus der fast sicheren Konvergenz die stochastische folgt, gilt Voraussetzung (A.10).
Voraussetzung (A.11): Auch hier wird die stärke Aussage der fast sicheren Konvergenz gezeigt. Es gilt:

$$\begin{aligned}\frac{1}{B_n} \sum_{k=1}^{n} \sum_{i=1}^{Z_{k-1}} \theta_{k,i} \quad &= \quad \frac{1}{B_n} \sum_{k=h+1}^{n} Z_{k-1} \bar{r}_{\theta_0,\nu}(Z_{k-1}) m'_{\theta_0,\nu}(Z_{k-1}) Z_{k-1}^{1-\gamma} \\ &= \quad \frac{1}{B_n} \sum_{k=h+1}^{n} \bar{r}_{\theta_0,\nu}(Z_{k-1}) \left(Z_{k-1}^{-\alpha_*} + r'_{\theta_0,\nu}(Z_{k-1}) \right) (Z_{k-1}) Z_{k-1}^{2-\gamma} \qquad (3.37)\end{aligned}$$

Für $\nu = \nu_0$ ist $\bar{r}_{\theta_0,\nu}(N) = r_{\theta_0,\nu_0}(N) - r_{\theta_0,\nu}(N) = 0$, und somit ist auch Gleichung (3.37) gleich 0.
Für $\nu \neq \nu_0$ ist dies nicht so schnell zu sehen. Voraussetzungen (ii) und (v) implizieren aber:

$$\begin{aligned}
\left| \frac{1}{B_n} \sum_{k=}^{n} \sum_{i=1}^{Z_{k-1}} \theta_{k,i} \right| &\leq \frac{1}{B_n} \sum_{k=h+1}^{n} 2C Z_{k-1}^{-\alpha_{**}} \left(Z_{k-1}^{\alpha_*} + C' Z_{k-1}^{-\alpha_{**}} \right) Z_{k-1}^{2-\gamma} \\
&= \frac{1}{B_n} \sum_{k=h+1}^{n} 2C \big(1 + C' \underbrace{Z_{k-1}^{\alpha_*-\alpha_{**}}}_{\leq 1 \text{ da } \alpha_* < \alpha_{**}} \big) Z_{k-1}^{2-\gamma-\alpha_*-\alpha_{**}} \\
&\leq 2C(1+C') \sup_{h+1 \leq k \leq n} W_{k-1}^{2-\gamma-\alpha_*-\alpha_{**}} \frac{\sum_{k=h+1}^{n} a_{k-1}^{2-\gamma-\alpha_*-\alpha_{**}}}{B_n} \\
&\xrightarrow[n\to\infty]{} 0 \text{ f.s.}
\end{aligned}$$

Der Grenzübergang ist richtig, da nach Satz 3.4 W_n fast sicher gegen eine Zufallsgröße W konvergiert und Voraussetzung (viii) garantiert, dass der Bruch für $n \to \infty$ verschwindet. Somit gilt Voraussetzung (A.11).
Voraussetzung (A.12): Der zu Beginn dieses Kapitels definierte Ausdruck R_k lässt sich unter Verwendung der oben definierten $\xi_{k,i}$ und $\vartheta_{k,i}$ umschreiben zu

$$R_k = (\xi_{k,1} - \vartheta_{k,1})\, a_{k-1}^{\frac{-\beta-2(1-\alpha_*)+2\gamma}{2}} .$$

Weiter sei

$$\hat{R}_n := \sum_{k=1}^{n} \sum_{i=1}^{Z_{k-1}} \frac{1}{B_n^2} E\left[(\xi_{k,i} - \vartheta_{k,i})^2 \mathbb{1}_{\{\frac{1}{B_n}(\xi_{k,i}-\vartheta_{k,i})<x\}} | \mathcal{F}_{k-1} \right] .$$

Da die Zufallsgrößen $I_{k,i}$ für festes k identisch verteilt sind, lässt sich dies weiter umformen:

$$\hat{R}_n = \sum_{k=1}^{n} \sum_{i=1}^{Z_{k-1}} \frac{1}{B_{k,n}^2} E\left[R_k^2 \mathbb{1}_{\{R_k < x B_{k,n}\}} | \mathcal{F}_{k-1} \right]$$

Gezeigt wird nun, dass die Folge $\left(\hat{R}_n\right)_{n\leq 1}$ für $n \to \infty$ für $x > 0$ fast sicher gegen $\sigma^2 W^{1+\beta+2(1-\alpha_*)-2\gamma}$ und für $x \leq 0$ fast sicher gegen 0 konvergiert.
Es seien also $x > 0$, $l \in \mathbb{N}$ mit $h \leq l \leq n$ fest gewählt und $\Delta_n := \left| \hat{R}_n - \sigma^2 W^{1+\beta+2(1-\alpha_*)-2\gamma} \right|$.
Mit $y := 1 + \beta + 2(1-\alpha_*) - 2\gamma$ liefern dann Einsetzen der Definitionen und die Dreiecks-

ungleichung die Abschätzung

$$\begin{aligned}
\Delta_n &= \left| \sum_{k=1}^{n} \sum_{i=1}^{Z_{k-1}} \frac{1}{B_{k,n}^2} E\left[R_k^2 \mathbb{1}_{\{R_k < xB_{k,n}\}} | \mathcal{F}_{k-1}\right] - \sigma^2 W^y \right| \\
&= \left| \sum_{k=1}^{n} Z_{k-1} \frac{a_{k-1}^{\beta+2(1-\alpha_*)-2\gamma}}{B_n^2} E\left[R_k^2 \mathbb{1}_{\{R_k < xB_{k,n}\}} | \mathcal{F}_{k-1}\right] - \sigma^2 W^y \right| \\
&\leq \left| \sum_{k=1}^{l} W_{k-1} \frac{a_{k-1}^y}{B_n^2} E\Big[R_k^2 \underbrace{\mathbb{1}_{\{R_k < xB_{k,n}\}}}_{\leq 1} | \mathcal{F}_{k-1}\Big] \right| \\
&\qquad + \left| \sup_{l+1 \leq k \leq n} W_{k-1} E\left[R_k^2 \mathbb{1}_{\{R_k < xB_{k,n}\}} | \mathcal{F}_{k-1}\right] \underbrace{\frac{\sum_{k=l+1}^{n} a_{k-1}^y}{\sum_{k=h+1}^{n} a_{k-1}^y}}_{\leq 1} - \sigma^2 W^y \right| \\
&\leq \left(\sup_{1 \leq k \leq l} W_{k-1} E\left[R_k^2 | \mathcal{F}_{k-1}\right] \right) \frac{\sum_{k=1}^{l} a_{k-1}^y}{\sum_{k=h+1}^{n} a_{k-1}^y} \\
&\qquad + \sup_{l+1 \leq k \leq n} \left| W_{k-1} E\left[R_k^2 \mathbb{1}_{\{R_k < xB_{k,n}\}} | \mathcal{F}_{k-1}\right] - \sigma^2 W^y \right| \qquad (3.38)
\end{aligned}$$

Da

$$\begin{aligned}
W_{k-1} E\left(R_k^2 | \mathcal{F}_{k-1}\right) &= \sigma^2(Z_{k-1}) m'^2_{\theta_0,\nu}(Z_{k-1}) W_{k-1}^{1+2(1-\gamma)} a_{k-1}^{-\beta+2\alpha_*} \\
&= \sigma^2(Z_{k-1}) Z_{k-1}^{-\beta} m'^2_{\theta_0,\nu}(Z_{k-1}) Z_{k-1}^{2\alpha_*} W_{k-1}^{1+\beta+2(1-\alpha_*)-2\gamma} \qquad (3.39)
\end{aligned}$$

gilt, ist nach Satz 3.4 sowie Voraussetzungen (v) und (3.29) das Supremum

$$\sup_{1 \leq k \leq l} W_{k-1} E\left[R_k^2 | \mathcal{F}_{k-1}\right] < \infty \text{ f.s.}$$

Voraussetzung (vii) impliziert dann, dass der erste Term der rechten Seite der Gleichung (3.38) für $n \to \infty$ fast sicher verschwindet, da l fest gewählt ist.
Der zweite Term der rechten Seite der Gleichung (3.38) kann weiter aufgespaltet werden:

$$\begin{aligned}
&\sup_{l+1 \leq k \leq n} \left| W_{k-1} E\left[R_k^2 \mathbb{1}_{\{R_k < xB_{k,n}\}} | \mathcal{F}_{k-1}\right] - \sigma^2 W^y \right| \\
&\quad \leq \sup_{l+1 \leq k \leq n} W_{k-1} \sup_{l+1 \leq k \leq n} \left| E\left[R_k^2 \mathbb{1}_{\{R_k < xB_{k,n}\}} | \mathcal{F}_{k-1}\right] - \sigma^2 W^y \right| \\
&\qquad + \sigma^2 W^{\beta+2(1-\alpha_*)-2\gamma} \sup_{l+1 \leq k \leq n} |W_{k-1} - W|
\end{aligned}$$

Nach Satz 3.4 ist der Grenzwert $\lim_{l\to\infty} \lim_{n\to\infty} \sup_{l+1 \leq k \leq n} W_{k-1}$ fast sicher endlich und $\lim_{l\to\infty} \lim_{n\to\infty} \sup_{l+1 \leq k \leq n} |W_{k-1} - W| = 0$ fast sicher. Das heißt, es bleibt nur noch ein Term zu betrachten:

$$\begin{aligned}
&\lim_{l\to\infty} \lim_{n\to\infty} \sup_{l+1 \leq k \leq n} \left| E\left[R_k^2 \mathbb{1}_{\{R_k < xB_{k,n}\}} | \mathcal{F}_{k-1}\right] - \sigma^2 W^y \right| \\
&\quad \leq \lim_{l\to\infty} \lim_{n\to\infty} \sup_{l+1 \leq k \leq n} \left| E\left[R_k^2 | \mathcal{F}_{k-1}\right] - \sigma^2 W^y \right| \\
&\qquad + \lim_{l\to\infty} \lim_{n\to\infty} \sup_{l+1 \leq k \leq n} E\left[R_k^2 \mathbb{1}_{\{R_k \geq xB_{k,n}\}} | \mathcal{F}_{k-1}\right]
\end{aligned}$$

Aus Gleichung (3.39) folgt mit Satz 3.4 sowie Voraussetzungen (v) und (3.29), dass der erste Term beim Grenzübergang $n \to \infty$ gefolgt vom Grenzübergang $l \to \infty$ fast sicher

gegen 0 konvergiert. Der zweite Term konvergiert nach Voraussetzung (ix) gegen 0 fast sicher.
Für $x \leq 0$ konvergiert nach Voraussetzung (ix) $\hat{R}_n$ fast sicher gegen 0. Damit konvergiert diese Folge für $n \to \infty$ für $x > 0$ fast sicher gegen $\sigma^2 W^{1+\beta+2(1-\alpha_*)-2\gamma}$ und für $x \leq 0$ fast sicher gegen 0. Der Beweis für festes $n - h$ verläuft ähnlich.
Nun sind alle Voraussetzungen des Satzes A.5 erfüllt und daraus folgen (3.31) und (3.32).

Zum Beweis der ergänzenden Konvergenzaussagen (α) und (β) sei noch einmal an die Definition von $\Phi_{h,n,\gamma}$ erinnert:

$$\Phi_{h,n,\gamma} = \frac{B_n}{D_n} = \frac{\sqrt{\sum_{k=h+1}^{n} a_{k-1}^{1+\beta+2(1-\alpha_*)-2\gamma}}}{\sum_{k=h+1}^{n} a_{k-1}^{2(1-\alpha_*)-\gamma}}$$

Da entweder h oder $n - h$ fest ist, hängt $\Phi_{h,n,\gamma}$ nur von γ ab. In Hinblick auf Gleichung (3.30) ist also die Konvergenzrate minimal und damit am besten, wenn $\Phi_{h,n,\gamma}$ als Funktion von γ ein Minimum aufweist. Dazu wird die erste Ableitung betrachtet:

$$\begin{aligned}\frac{\mathrm{d}\Phi_{h,n,\gamma}}{\mathrm{d}\gamma} = & \left[\frac{1}{2}\left(\sum_{k=h+1}^{n} a_{k-1}^{y}\right)^{-\frac{1}{2}}\left(\sum_{k=h+1}^{n}\left(-2\log(a_{k-1})\right)a_{k-1}^{y}\right)\left(\sum_{k=h+1}^{n} a_{k-1}^{2(1-\alpha_*)-\gamma}\right)\right. \\ & \left.-\sqrt{\sum_{k=h+1}^{n} a_{k-1}^{y}}\left(\sum_{k=h+1}^{n}\left(-\log(a_{k-1})\right)a_{k-1}^{2(1-\alpha_*)-\gamma}\right)\right]\left(\sum_{k=h+1}^{n} a_{k-1}^{2(1-\alpha_*)-\gamma}\right)^{-2}\end{aligned}$$

Folglich ist die erste Ableitung genau dann gleich 0, wenn

$$\begin{aligned}&\left(\sum_{k=h+1}^{n}\log(a_{k-1})a_{k-1}^{y}\right)\left(\sum_{k=h+1}^{n} a_{k-1}^{2(1-\alpha_*)-\gamma}\right) \\ &\qquad = \left(\sum_{k=h+1}^{n} a_{k-1}^{y}\right)\left(\sum_{k=h+1}^{n}\log(a_{k-1})a_{k-1}^{2(1-\alpha_*)-\gamma}\right) \qquad (3.40)\end{aligned}$$

Eine Lösung der Gleichung (3.40) liegt dann vor, wenn $y = 2(1-\alpha_*) - \gamma$ gilt. Da $y = 1 + \beta + 2(1-\alpha_*) - 2\gamma$, ist das der Fall, wenn $\gamma = 1 + \beta$ ist. Dass dies die einzige Lösung ist, wird nun bewiesen. Dazu betrachten wir die Funktion:

$$\omega(x) := \frac{\sum_{k=h+1}^{n}\log(a_{k-1})a_{k-1}^{x}}{\sum_{k=h+1}^{n} a_{k-1}^{x}}$$

Gleichung (3.40) gilt genau dann, wenn $\omega(1+\beta+2(1-\alpha_*)-2\gamma) = \omega(2(1-\alpha_*)-\gamma)$, also genau dann wenn $\gamma = 1 + \beta$ gilt. Ableiten von ω ergibt:

$$\frac{\mathrm{d}\omega(x)}{\mathrm{d}x} = \frac{\sum_{k=h+1}^{n}\left(\log(a_{k-1})\right)^2 a_{k-1}^{x}}{\sum_{k=h+1}^{n} a_{k-1}^{x}} - \left(\frac{\sum_{k=h+1}^{n}\log(a_{k-1})a_{k-1}^{x}}{\sum_{k=h+1}^{n} a_{k-1}^{x}}\right)^2$$

Die Cauchy-Schwarzsche Ungleichung (siehe [10] Kapitel VI.1) liefert

$$\left|\sum_{k=h+1}^{n}\left(\log(a_{k-1})a_{k-1}^{\frac{x}{2}}\right)a_{k-1}^{\frac{x}{2}}\right| < \sqrt{\sum_{k=h+1}^{n}\left(\log(a_{k-1})\right)^2 a_{k-1}^{x}}\sqrt{\sum_{k=h+1}^{n} a_{k-1}^{x}},$$

folglich gilt

$$\left(\frac{\sum_{k=h+1}^{n}\log(a_{k-1})a_{k-1}^{x}}{\sum_{k=h+1}^{n}a_{k-1}^{x}}\right)^{2}<\frac{\sum_{k=h+1}^{n}\left(\log(a_{k-1})\right)^{2}a_{k-1}^{x}}{\sum_{k=h+1}^{n}a_{k-1}^{x}}\,,$$

und damit ist $\frac{\mathrm{d}\omega(x)}{\mathrm{d}x}>0$. Das heißt die Funktion ω ist streng monoton wachsend und somit gilt $1+\beta+2(1-\alpha_*)-2\gamma=2(1-\alpha_*)-\gamma$. Mit dem gleichen Kniff ist zu sehen, dass $\frac{\mathrm{d}^2\Phi_{h,n,\gamma}}{\mathrm{d}\gamma^2}\geq 0$ ist. In $\gamma=1+\beta$ liegt also ein Minimum vor.
In diesem Fall gilt für $\beta+2\alpha_*=1$, dass

$$\Phi_{h,n,\gamma}=\frac{\sqrt{\sum_{k=h+1}^{n}a_{k-1}^{2(1-\alpha_*)-(1+\beta)}}}{\sum_{k=h+1}^{n}a_{k-1}^{2(1-\alpha_*)-(1+\beta)}}=\frac{1}{\sqrt{\sum_{k=h+1}^{n}a_{k-1}^{2(1-\alpha_*)-(1+\beta)}}}=\frac{1}{\sqrt{n-h}}\,.$$

Für $\beta+2\alpha_*<1$ gilt im superkritischen Fall

$$\Phi_{h,n,\gamma}=\mathcal{O}\left(a_n^{\frac{2\alpha_*+\beta-1}{2}}\right)$$

und im fast-kritischen Fall

$$\Phi_{h,n,\gamma}\leq a_h^{\frac{2\alpha_*+\beta-1}{2}}\frac{1}{\sqrt{n-h}}$$

Damit sind alle Behauptungen dieses Satzes bewiesen. □

4. Die PCR als größenabhängiger Verzweigungsprozess

Nun steht ein breites Spektrum an Aussagen über größenabhängige Verzeigungsprozesse zur Verfügung, um zur PCR zurückkehren zu können. Wie schon zu Beginn des Kapitels 3 erläutert, muss das Modell erweitert werden, um den gesamten Prozess chemisch realitätsnah beschreiben zu können. Dies geschieht in Kapitel 4.1. Dadurch kann die Asymptotik des Prozesses untersucht (☞ Kapitel 4.2) und ein Schätzer für die Effizienz gefunden werden (☞ Kapitel 4.3).

4.1. Verallgemeinerung des PCR-Modells

Das in Definition 2.9 eingeführte PCR-Modell

$$\Big((\Omega,\mathcal{A}),(P_i)_{i\geq 0},(I_{n,j})_{j,n\geq 1},(Z_n)_{n\geq 0},(\mathcal{Z}_n)_{n\geq 0},(X_k^n)_{k\geq 0,n\geq 1}\Big)$$

beschreibt die PCR unter der Voraussetzung konstanter Effizienz λ. Der Galton-Watson-Prozess $(Z_n)_{n\geq 0}$ ist bekanntlich durch $Z_n := Z_{n-1} + \sum_{j=1}^{Z_{n-1}} I_{n,j}$ mit stochastisch unabhängigen, identisch $\mathcal{B}(1,\lambda)$-verteilten Zufallsgrößen gegeben. Ersetzt man in dieser Definition die konstante Effizienz durch eine größenabhängige Funktion $\lambda(\cdot) : \mathbb{N}_0 \to [0,1]$, so wird auch der Prozess $(Z_n)_{n\geq 0}$ größenabhängig. Im Folgenden interessiert das Verhalten dieses größenabhängigen Prozesses, wohingegen Mutationen nicht untersucht werden, so dass die Menge aller nach dem n-ten Schritt vorhandenen Sequenzen $\mathcal{Z}_n$ und die Anzahl der Sequenzen der k-ten Generation nach n Zyklen X_k^n in der folgenden Definition keine Erwähnung mehr finden.

Definition 4.1 Gegeben seien ein messbarer Raum $(\Omega,\mathcal{A})$ und darauf eine Familie von Wahrscheinlichkeitsmaßen $(P_j)_{j\geq 0}$ unter denen für eine Zufallsgröße $Z_0 : \Omega \to \mathbb{N}_0$ die Gleichung $P_j(Z_0 = j) = 1$ für alle $j \geq 0$ gilt. Für $n \geq 1$ seien Zufallsgrößen Z_n rekursiv definiert durch

$$Z_n := Z_{n-1} + \sum_{j=1}^{Z_{n-1}} I_{n,j}$$

mit unter $\mathcal{F}_{n-1} := \sigma(Z_0, Z_1, \ldots, Z_{n-1})$ stochastisch unabhängigen und identisch Bernoulliverteilten Zufallsgrößen $I_{n,j}$ mit einem von Z_{n-1} abhängigen Parameter $\lambda(Z_{n-1})$, wobei $\lambda(\cdot) : \mathbb{N}_0 \to [0,1]$ *größenabhängige Effizienzfunktion*, kurz auch nur *Effizienz*, genannt wird. Dann heißt das Modell

$$((\Omega,\mathcal{A}),(P_j)_{j\geq 0},(Z_n)_{n\geq 0},\lambda(\cdot),(I_{n,j})_{j,n\geq 1})$$

größenabhängige Polymerase-Kettenreaktion; bei fester Vorgabe von Z_0 wird es *größenabhängige Polymerase-Kettenreaktion mit Z_0 Startsequenzen* genannt.

Auch in diesem Kapitel können wir uns auf die Betrachtung der PCR mit einer Startsequenz beschränken. Für die erwartete Anzahl an Nachkommen einer jeden Sequenz im n-ten PCR-Zyklus gilt folgendes Lemma:

Lemma 4.2 *Gegeben sei eine größenabhängige PCR mit der Standardfiltration* $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$, $\mathcal{F}_n := \sigma(Z_0, \ldots, Z_n)$. *Bezeichnen* $m(Z_n) := E(I_{n+1,j} \mid \mathcal{F}_n)$ *die erwartete Anzahl an Nachkommen einer dem n-ten Zyklus entstammenden Sequenz und* $\sigma^2(Z_n) := Var(I_{n+1,j} \mid \mathcal{F}_n)$ *die Varianz einer solchen Sequenz, dann gilt für alle* $n \geq 0$*:*

$$m(Z_n) = 1 + \lambda(Z_n) \tag{4.1}$$
$$\sigma^2(Z_n) = \lambda(Z_n)(1 - \lambda(Z_n)) \tag{4.2}$$

Beweis: Die beiden Gleichungen ergeben sich sofort aus Definition 4.1, da

$$P(I_{n+1,j} = 2 \mid \mathcal{F}_n) = \lambda(Z_n) = 1 - P(I_{n+1,j} = 1 \mid \mathcal{F}_n)$$

gilt. □

Bemerkung 4.3 Aus Gleichung (4.1) folgt für alle $n \geq 0$

$$E(Z_{n+1} \mid \mathcal{F}_n) = Z_n(1 + \lambda(Z_n)) \quad \text{f.s.}$$

Im Vergleich zu Lemma 2.10, das eine Aussage über die erwartete Anzahl an Nachkommen im $(n+1)$-ten Zyklus macht, ist es im größenabhängigen Fall nur möglich, die erwartete Anzahl an Nachkommen rekursiv anzugeben.

Jagers und Klebaner modellieren in [15] die PCR als größenabhängigen Verzweigungsprozess, indem die Effizienz der Gleichung

$$\lambda(Z_n) = \frac{K_M}{K_M + Z_n} \tag{4.3}$$

genügen soll. Dieser Ansatz berücksichtigt durch das Verwenden der Michaelis-Menten-Konstante K_M, dass es sich bei der PCR um eine enzymkatalysierte Reaktion handelt, für die im Allgemeinen die Michaelis-Menten-Kinetik gilt (siehe dazu auch B.6). Den experimentell ermittelten zwei Wachstumsphasen trägt es allerdings nicht Rechnung.
Wie in Abbildung 3.1 zu erkennen, beginnt die PCR mit einer exponentiellen Wachstumsphase, die mit dem in Kapitel 2 bereitgestellten Modell beschrieben werden kann. In dieser Phase ist also die Effizienz eine konstante Größe. Überschreitet die Gesamtanzahl der in einem Schritt erzeugten Sequenzen Z_n einen bestimmten Schwellenwert S, so nimmt die Effizienz mit zunehmender Gesamtanzahl ab.

Definition 4.4 Gegeben seien eine größenabhängige PCR, Konstanten $K_M, C \in \mathbb{R}_0^+$ und $S \in \mathbb{N}$. Weiter sei $\mathscr{S}(x) := S\mathbb{1}_{\{x<S\}} + x\mathbb{1}_{\{x\geq S\}}$ für alle $x \in \mathbb{R}$. Die Effizienz genüge der Gleichung

$$\lambda(Z_n) := \frac{K_M}{K_M + \mathscr{S}(Z_n)} \frac{1 + \exp\left(-C\left(\frac{\mathscr{S}(Z_n)}{S} - 1\right)\right)}{2}. \tag{4.4}$$

Bemerkung 4.5 Da $P(I_{n,j} = 0 \mid \mathcal{F}_{n-1}) = 0$ f.s. für alle $n, j \geq 0$ gilt, ist der Prozess $(Z_n)_{n\geq 0}$ monoton wachsend und die durch Definition 4.4 gegebene Effizienz ist eine in $\mathscr{S}(Z_n)S^{-1}$ monoton fallende Funktion. Ist n_S definiert als der letzte Zyklus der exponentiellen Phase, d.h.

$$n_S := \sup\{n : Z_{n-1} < S\}\,, \tag{4.5}$$

so nimmt die Effizienz für alle $n \leq n_S - 1$ einen konstanten Wert an. Für $n \geq n_S$ wächst der Prozess langsamer, wofür der Korrekturterm $\frac{1}{2}\left[1 + \exp\left(-C\left(\frac{\mathscr{S}(Z_n)}{S} - 1\right)\right)\right]$ sorgt.
Das durch Gleichung (4.3) definierte Modell ist ein Spezialfall von Definition 4.4, mit $S = Z_0$ und $C = 0$.
Somit erfüllt das definierte Effizienzmodell alle experimentell bestimmten Eigenschaften der PCR.

Mit dem Verweis auf Lemma 4.2 folgt mit einer kurzen Rechnung für alle $n \geq 0$

$$m(Z_n) = 1 + \frac{K_C}{\mathscr{S}(Z_n)} + r\left(\mathscr{S}(Z_n)\right), \tag{4.6}$$

mit

$$K_C := \frac{K_M}{2}(1 + \delta_C), \quad \delta_C := \mathbb{1}_{\{C=0\}} \quad \text{und}$$

$$r\left(\mathscr{S}(Z_n)\right) := \frac{K_M}{\mathscr{S}(Z_n)(K_M + \mathscr{S}(Z_n))}\left(\mathscr{S}(Z_n)\frac{\exp\left(-C\left(\frac{\mathscr{S}(Z_n)}{S} - 1\right)\right) - \delta_C}{2} - K_C\right).$$

Da $\lim_{x\to\infty} x\left(\exp\left(-C\left(x-1\right)\right) - \delta_C\right) = 0$, gilt $r(x) = \mathcal{O}(x^{-2})$.
Für die Nachkommensvarianz gilt mit Lemma 4.2 nach leichter Rechnung

$$\sigma^2(Z_n) = \frac{K_C}{\mathscr{S}(Z_n)} - r_+(\mathscr{S}(Z_n))\,, \tag{4.7}$$

wobei

$$\begin{aligned} r_+(\mathscr{S}(Z_n)) \;:=\; & \frac{K_M}{\mathscr{S}(Z_n)(K_M + \mathscr{S}(Z_n))^2}\left[K_M\mathscr{S}(Z_n)\left(\frac{\exp\left(-C\left(\frac{\mathscr{S}(Z_n)}{S} - 1\right)\right) + 1}{2}\right)^2\right. \\ & \left. + \left(K_M + \mathscr{S}(Z_n)\right)\left(K_C - \mathscr{S}(Z_n)\right)\frac{\exp\left(-C\left(\frac{\mathscr{S}(Z_n)}{S} - 1\right)\right) + \delta_C}{2}\right] \end{aligned}$$

ist.

4.2. Asymptotisches Verhalten der größenabhängigen PCR

Das Nachkommenmittel der größenabhängigen PCR genügt Gleichung (4.6). Da für $N \geq S$

$$1 \leq m(N) = 1 + \frac{K_C}{N} + r\,(N) \leq 1 + \frac{K_M}{N} + \mathcal{O}(N^{-2}) \xrightarrow[N\to\infty]{} 1$$

gilt, liegt ein fast-kritischer, größenabhängiger Verzweigungsprozess vor. Also lässt sich Satz 3.4 (b) anwenden.

Korollar 4.6 *Gegeben sei eine größenabhängige PCR. Mit* $a_n := a_{n-1}m(a_{n-1})$ *und* $E_\infty := \{\lim_{n\to\infty} = \infty\}$ *gilt* $P(E_\infty) > 0$ *und*

$$\lim_{n\to\infty} \frac{Z_n}{a_n} = 1 \quad \textit{f.s. auf } E_\infty\,.$$

Beweis: Wie oben angedeutet, müssen die Voraussetzungen des Satzes 3.4 (b) überprüft werden.
Voraussetzung V.3 ergibt sich sofort aus Gleichung (4.6) mit $m = 1$ und $\alpha = 1$.
Voraussetzung V.4: Zu zeigen ist, dass es ein $N_0 \in \mathbb{N}$, ein $\sigma^2 \in (0,\infty)$ und ein $\beta \in [-1,1)$ gibt, so dass für alle $N \geq N_0$ die Ungleichung $\sigma^2(N) \leq \sigma^2 N^\beta$ gilt. Für $N \geq S$ gilt mit Gleichung (4.2) und (4.4) sowie $0 \leq \exp\left(-C\left(\frac{N}{S}-1\right)\right) \leq 1$:

$$\begin{aligned}
\sigma^2(N) &= \lambda(N)\,(1-\lambda(N)) \\
&= \frac{K_M}{K_M+N}\underbrace{\frac{1+\exp\left(-C\left(\frac{N}{S}-1\right)\right)}{2}}_{\leq 1}\left(1 - \frac{K_M}{K_M+N}\underbrace{\frac{1+\exp\left(-C\left(\frac{N}{S}-1\right)\right)}{2}}_{\geq \frac{1}{2}}\right) \\
&\leq \frac{K_M}{K_M+N}\underbrace{\left(1-\frac{K_M}{2(K_M+N)}\right)}_{\leq 1} \\
&\leq \frac{K_M}{K_M+N} \\
&\leq K_M N^{-1}
\end{aligned}$$

Die Voraussetzung V.4 ist also mit $N_0 = S,\ \sigma^2 = K_M$ und $\beta = -1$ erfüllt.
Voraussetzung V.5: Das Nachkommenmittel $m(\cdot)$ ist monoton fallend und echt größer 0, da entweder ein oder zwei Nachkommen produziert werden, aber niemals keine Nachkommen. Nun wird die Funktion $\frac{\sigma^2(x)}{x^2(m(x)-1)}$ betrachtet. Sie ist offensichtlich streng monoton fallend. Da $0 \leq \exp\left(-C\left(\frac{\mathscr{S}(x)}{S}-1\right)\right) \leq 1$ und $\frac{K_M}{K_M+\mathscr{S}(x)} \in (0,1)$ für alle $x \geq 1$ ist, gilt

$$\int_1^\infty \left|\frac{1}{x^2}\,(1-\lambda(x))\right| \mathrm{d}x \leq \int_1^\infty \frac{1}{x^2}\,|1-\lambda(x)|\,\mathrm{d}x \leq \int_1^\infty \frac{1}{x^2}\mathrm{d}x < \infty\,.$$

Damit ist auch diese Voraussetzung erfüllt und die Behauptung gilt. □

4.3. Ein Schätzer der Effizienz

Die Effizienz der größenabhängigen PCR genügt im vorausgesetzten Modell der Gleichung

$$\lambda(Z_n) := \frac{K_M}{K_M + \mathscr{S}(Z_n)} \frac{1 + \exp\left(-C\left(\frac{\mathscr{S}(Z_n)}{S} - 1\right)\right)}{2},$$

wobei die Größen K_M, C und S in $\mathscr{S}(\cdot)$ unbekannt sind. Die experimentell gewonnen Daten liefern eine Messreihe, die zu jedem Zyklus die Gesamtzahl aller Sequenzen nach diesen Zyklen aufzeigt. Die in den Kapiteln 3.3 und 3.4 entwickelte Schätztheorie für größenabhängige Verzweigungsprozesse bezieht sich auf das Schätzen des Nachkommenmittels. Da aber nach Gleichung (4.6)

$$m(Z_n) = 1 + \lambda(Z_n) = 1 + \frac{K_C}{\mathscr{S}(Z_n)} + r\left(\mathscr{S}(Z_n)\right)$$

gilt, lässt sich mit dieser Theorie auch die Effizienz der größenabhängigen PCR ermitteln. Unbekannt sind hier die Größen $K_C = K_M(\frac{1+\delta_C}{2})$, C und S, welche vorerst das Ziel der Schätzung sein sollen. Wie im Modell des größenabhängigen Verzweigungsprozesses beschrieben (Kapitel 3.1), kann die Anzahl aller Sequenzen nach n PCR-Zyklen durch die Gleichung

$$Z_n = m(Z_{n-1})Z_{n-1} + \eta_n$$

ausgedrückt werden, wobei η_n die Schwankung um den mittleren Wachstumsterm angibt. Bedingt unter $\mathcal{F}_{n-1}$ gilt für diese Schwankung

$$E\left[\eta_n|\mathcal{F}_{n-1}\right] = 0, \quad Var\left[\eta_n|\mathcal{F}_{n-1}\right] = Z_{k-1}\sigma^2(Z_{k-1}), \tag{4.8}$$

wobei die letzte Gleichung mit Gleichung (4.7) umgeschrieben werden kann:

$$Var\left[\eta_n|\mathcal{F}_{n-1}\right] = Z_{k-1}\left(\frac{K_C}{\mathscr{S}(Z_n)} - r_+(\mathscr{S}(Z_n))\right)$$

Die Folge der von der Zykluszahl n abhängigen Schätzer für die Konstanten K_C, S und C sei mit $\left((\hat{K_C})_{h,n}, \hat{S}_{h,n}, \hat{C}_{h,n}\right)_{n\geq h}$ bezeichnet, wobei $h \in \mathbb{N}_0$ im Folgenden fest gewählt sei und angibt, mit welchem Zyklus die Datenanalyse beginnt. Der Schätzer soll nicht nur die Eigenschaft besitzen, konsistent zu sein, sondern auf jeden Fall Werte liefern, für die die geschätzte Anzahl an Sequenzen möglichst wenig von der tatsächlichen abweicht. Diese Bedingung erfüllt der Schätzer, wenn er den folgenden Kontrast minimiert:

$$\bar{S}_{h,n}(K_C, S, C) := \sum_{k=h+1}^{n} \left(Z_k - (1 + \lambda(Z_{k-1}))\, Z_{k-1}\right)^2 Z_{k-1}^{-1} \mathscr{S}(Z_{k-1})^{-\beta} \tag{4.9}$$

mit $\beta = -1$. Die Schwelle S, ab der der Prozess nur noch „linear“ wächst, kann mit Hilfe des in Gleichung (4.5) definierten letzten Zyklus der exponentiellen Wachstumsphase $n_S = \sup\{n : Z_{n-1} < S\}$ durch die Gesamtzahl an Sequenzen nach diesem Zyklus Z_{n_S} geschätzt werden. Dadurch erhält der Kontrast in Gleichung (4.9) die Gestalt

$$\begin{aligned} S_{h,n}(K_C, S, C) \quad := \quad & \sum_{k=h+1}^{n_S} \left(Z_k - (1 + \lambda(Z_{k-1}))\, Z_{k-1}\right)^2 Z_{k-1}^{-1} Z_{n_S} \\ & + \sum_{k=n_S+1}^{n} \left(Z_k - (1 + \lambda(Z_{k-1}))\, Z_{k-1}\right)^2 \end{aligned} \tag{4.10}$$

Der Schätzer muss also die Gleichung

$$\left((\hat{K_C})_{h,n}, \hat{S}_{h,n}, \hat{C}_{h,n}\right) = \arg \min_{K_C,C,S} S_{h,n}(K_C, S, C)$$

erfüllen.
Betrachten wir noch einmal das oben aufgeführte Nachkommenmittel. Die Terme $\frac{K_C}{\mathscr{S}(N)}$ und $r\left(\mathscr{S}(N)\right)$ sind von der Ordnung $\mathcal{O}(N^{-\alpha})$ bzw. $\mathcal{O}(N^{-\bar{\alpha}})$ mit $\alpha = 1$ sowie $\bar{\alpha} = 2$. Demnach gehört die größenabhängige PCR zum Modelltyp (3.22). Die starke Konsistenz des Schätzers erfordert aber die gleichgradig asymptotische Identifizierbarkeit von (K_C, C, S) in $m(\cdot)$ mit der Rate $Z_{n-1}^{\bar{\alpha}}$. Dies ist leider nicht der Fall, da dafür $\beta + 2\bar{\alpha} \leq 1$ erfüllt sein müsste und diese Bedingung durch $\beta = -1$ sowie $\bar{\alpha} = 2$ verletzt wird.
Die gute Nachricht ist aber, dass K_C identifizierbar mit Rate Z_n ist. So kann (S, C) als Störparameter aufgefasst werden, der in der allgemeinen Theorie mit ν bezeichnet wurde, d.h. $\nu = (S, C)$. Im weiteren Verlauf wird demnach ein Schätzer für K_C gesucht. In Übereinstimmung mit den Bezeichnungen in Kapitel 3 sei $\theta_0 := K_C$ und $S_{h,n,\nu}(\theta) := S_{h,n}(K_C, S, C)$. Dann muss der Schätzer $\hat{\theta}_{h,n,\nu}$ für θ_0 der Bedingung

$$\hat{\theta}_{h,n,\nu} = \arg \min_{\theta} S_{h,n,\nu}(\theta) \tag{4.11}$$

genügen. Wie zu Beginn des Kapitels 3 existiere eine kompakte Menge $\Theta \subset \mathbb{R}^+$, derart, dass $\theta_0 \in \overset{\circ}{\Theta}$ gilt. Um zu zeigen, dass der durch (4.11) definierte Schätzer konsistent ist, werden noch einige Größen benötigt. Die Effizienz der exponentiellen Wachstumsphase wird mit λ bezeichnet, also $\lambda := \lambda(1)$. Da es sich bei der PCR um einen fast-kritischen Verzweigungsprozess handelt, ist a_n für alle $n \geq 1$ wie in Satz 3.4 definiert als $a_n = a_{n-1}m(a_{n-1})$ und $a_0 = 1$. Weiter seien

$$\hat{\mathscr{S}}(a_{k-1}) := \begin{cases} (1+\lambda)^{n_S} & \text{für} \quad 1 \leq k \leq n_S \\ a_{k-1} & \text{für} \quad k > n_S \end{cases} \tag{4.12}$$

und mit $\beta = -1$, $\alpha = 1$

$$\Phi_{h,n}^{-1}(n_S) := \sqrt{\sum_{k=h+1}^{n} a_{k-1}\hat{\mathscr{S}}(a_{k-1})^{-\beta-2\alpha}} = \sqrt{\sum_{k=h+1}^{n_S} (1+\lambda)^{-n_S+k-1} + (n - n_S)}$$

Damit sind alle aus Kapitel 3 in diesem Zusammenhang wichtigen Begriffe näher spezifiziert und das versprochene Ergebnis kann formuliert werden:

Korollar 4.7 *Gegeben seien eine größenabhängige PCR und $h \in \mathbb{N}_0$ fest gewählt. Die Schätzer der Folge $\left(\hat{\theta}_{h,n,\nu}\right)_{n \geq h}$ für $\theta_0 = K_C$ genügen der Bedingung (4.11). Dann gilt*

$$\lim_{n\to\infty} \hat{\theta}_{h,n,\nu} = K_C \,,$$

also die starke Konsistenz der Folge $\left(\hat{\theta}_{h,n,\nu}\right)_{n \geq h}$. Für die Konvergenzrate gilt

$$\lim_{n\to\infty} \Phi_{h,n}^{-1}(n_S)\left(\theta_{h,n,\nu} - K_C\right) \overset{D}{=} \mathcal{N}(0, K_C) \,. \tag{4.13}$$

Beweis: Zum Beweis der starken Konsistenz wird Korollar 3.12 angewendet. Für den Grenzwertsatz wird Satz 3.14 benötigt. In beiden Fällen heißt dies, dass die Voraussetzungen überprüft werden müssen.

Da S fest ist und $Z_n \to \infty$ f.s. für $n \to \infty$ gilt, ist n_s fast sicher endlich. Der Kontrast $S_{h,n,\nu}(\theta)$ ist eine Verallgemeinerung des in Gleichung (3.13) eingeführten Kontrastes. Desweiteren werden die dort verwendeten a_{k-1}^{β}, $a_{k-1}^{\alpha_*}$ und $a_{k-1}^{\alpha_{**}}$ ersetzt durch $\hat{\mathscr{S}}(a_{k-1})^{\beta}$, $\hat{\mathscr{S}}(a_{k-1})^{\alpha_*}$ bzw. $\hat{\mathscr{S}}(a_{k-1})^{\alpha_{**}}$, außerdem gilt $\alpha_* = \alpha = 1$ sowie $\alpha_{**} = \bar{\alpha} = 2$. Da der Kontrast in (4.10) asymptotisch gleich dem Kontrast in (4.9) ist, kann Korollar 3.12 angewendet werden, dessen Voraussetzungen gelten, denn:
(i) Da Θ kompakt, $C \in \mathbb{R}_0^+$ und $S \in \mathbb{N}$ ist, wird für alle $\delta > 0$ und $N \in \mathbb{N}$ das Supremum bzw. Infimum von

$$m_{\theta_0,\nu}(N) - m_{\theta,\nu}(N) = \frac{1 + \exp\left(-C\left(\frac{\mathscr{S}(N)}{S}\right)\right)}{2}\left(\frac{\theta_0}{\theta_0 + \frac{(1+\delta_C)\mathscr{S}(N)}{2}} - \frac{\theta}{\theta + \frac{(1+\delta_C)\mathscr{S}(N)}{2}}\right)$$

für je ein $(\theta, \nu) = (K_C, (S, C)) \in (\Theta - (\theta_0 - \delta, \theta_0 + \delta)) \times \mathbb{N} \times \mathbb{R}_0^+$ angenommen.
(ii) Zu zeigen ist, dass ein $C' < \infty$ existiert, so dass für alle $N \in \mathbb{N}$ die Bedingung $\sup_{\theta,\nu} |r(\mathscr{S}(N))|\, N^2 \leq C'$ erfüllt ist. Zur Erinnerung: Die Funktion $r(\mathscr{S}(N))$ hat im zu Grunde liegenden Modell die Gestalt:

$$r\left(\mathscr{S}(Z_n)\right) = \frac{K_M}{\mathscr{S}(Z_n)(K_M + \mathscr{S}(Z_n))}\left(\mathscr{S}(Z_n)\frac{\exp\left(-C\left(\frac{\mathscr{S}(Z_n)}{S} - 1\right)\right) - \delta_C}{2} - K_C\right)$$

Da Θ kompakt also insbesondere beschränkt ist, existiert eine Konstante $c \in \mathbb{R}^+$, so dass $K_C \leq c$. Außerdem ist $C \geq 0$ und S fest gewählt. Für $1 \leq N < S$ gilt dann

$$\begin{aligned} |r(\mathscr{S}(N))|N^2 &= \frac{N^2 K_M}{S(K_M + S))}\left|S\frac{1-\delta_C}{2} - K_C\right| \\ &\leq \frac{S}{1 + \frac{S}{K_M}}\left(S\frac{1}{2} + c\right) < \infty \end{aligned}$$

für alle K_C und (S, C). Für $N \geq S$ gilt

$$\begin{aligned} |r(\mathscr{S}(N))|N^2 &= \frac{NK_M}{K_M + N}\left|N\frac{\exp\left(-C\left(\frac{N}{S} - 1\right)\right) - \delta_C}{2} - K_C\right| \\ &\leq \frac{N}{1 + \frac{N}{K_M}}\left(N\frac{1}{2} + c\right) < \infty \end{aligned}$$

für alle K_C und (S, C). In beiden Fällen gilt also Bedingung (ii) des Korollars 3.12.
(iii) Diese Bedingung ist schon für $\gamma = 1 + \beta = 0$ efüllt. Wie oben beschrieben, werden a_{k-1}^{β}, $a_{k-1}^{\alpha_*}$ und $a_{k-1}^{\alpha_{**}}$ ersetzt durch $\hat{\mathscr{S}}(a_{k-1})^{\beta}$, $\hat{\mathscr{S}}(a_{k-1})^{\alpha_*}$ bzw. $\hat{\mathscr{S}}(a_{k-1})^{\alpha_{**}}$. Da $\alpha_* = 1$ und h fest gewählt ist, ergibt sich dann mit der Definition von $\hat{\mathscr{S}}(a_{k-1})$ in (4.12):

$$\sum_{k=h+1}^{n} a_{k-1}^2 \hat{\mathscr{S}}(a_{k-1})^{-2\alpha_*} \hat{\mathscr{S}}(a_{k-1})^{-\gamma} = \left(\sum_{k=h+1}^{n_S} (1+\lambda)^{2(k-1)-2n_S}\right) + n - n_S \xrightarrow[n\to\infty]{} \infty$$

(iv) Mit $h_0 = n_S$ gilt:

$$\begin{aligned}
&\sum_{k=n_S+1}^{\infty} a_{k-1}^{4(1-\alpha_*)-2\gamma} a_{k-1}^{\beta+2\alpha_*-1} \left(\sum_{l=h+1}^{k} a_{l-1}^2 \hat{\mathscr{S}}(a_{l-1})^{-2\alpha_*} \hat{\mathscr{S}}(a_{l-1})^{-\gamma} \right)^{-2} \\
&\quad = \sum_{k=n_S+1}^{\infty} \left(\left(\sum_{l=h+1}^{n_S} (1+\lambda)^{2(l-1)-2n_S} \right) + k - n_S \right)^{-2} \\
&\quad \leq \sum_{k=n_S+1}^{\infty} (k-n_S)^{-2} \\
&\quad = \sum_{k=1}^{\infty} k^{-2} < \infty
\end{aligned}$$

Das ist Bedingung (iv).
Somit sind die Voraussetzungen des Korollar 3.12 erfüllt, welches die starke Konsistenz der Schätzerfolge $\left(\hat{\theta}_{h,n,\nu}\right)_{n \geq h}$ liefert.

Auch für den Beweis der Gleichung (4.13) wurde in Kapitel 3 gute Vorarbeit geleistet. Gelten die Voraussetzungen des Satzes 3.14, impliziert dieser die Behauptung. Die Voraussetzungen werden nun geprüft:
(v) Offensichtlich ist $r(N)$ zweimal in $\theta = K_C$ stetig differenzierbar und die Supremumsbedingung an die Ableitungen erfüllt.
(vi) Da die zweite Ableitung $r''(N)$ bezüglich θ gleich 0 ist, ist auch diese Bedingung erfüllt.
(vii) Für die zu Anfang des Kapitels 3.4 definierte Größe B_n gilt unter Beachtung der üblichen Annahmen:

$$B_n = \sqrt{\sum_{k=h+1}^{n} a_{k-1}^3 \hat{\mathscr{S}}(a_{k-1})^{\beta-2\alpha_*-2\gamma}} = \sqrt{\left(\sum_{k=h+1}^{n_S} (1+\lambda)^{3(k-1)-3n_S} \right) + n - n_S} \xrightarrow[n\to\infty]{} \infty$$

(viii) Es sei $\nu \neq \nu_0$, dann gilt:

$$\frac{1}{B_n} \sum_{k=h+1}^{n} a_{k-1}^2 \hat{\mathscr{S}}(a_{k-1})^{-\alpha_*-\alpha_{**}} = \frac{1}{B_n} \Bigg(\underbrace{\sum_{k=h+1}^{n_S} (1+\lambda)^{2(k-1)-3n_S}}_{=\text{const.}} + \sum_{k=n_S+1}^{n} \frac{1}{a_{k-1}} \Bigg)$$

Da nach (vii) B_n für $n \to \infty$ gegen ∞ geht, muss nur der zweite in der Klammer stehende Term betrachtet werden. Da $a_n = K_C\, n + o(n)$ gilt:

$$\begin{aligned}
\frac{1}{B_n} \sum_{k=n_S+1}^{n} \frac{1}{a_{k-1}} \quad &= \quad \frac{\sum_{k=n_S}^{n-1} \frac{1}{K_C k + o(k)}}{\sqrt{\left(\sum_{k=h+1}^{n_S} (1+\lambda)^{3(k-1)-3n_S}\right) + n - n_S}} \\
&\leq \quad \frac{1}{\sqrt{n-n_S}} \underbrace{\sup_{n_s \leq k \leq n-1} \left| \frac{1}{K_C + \frac{o(k)}{k}} \right|}_{\leq c < \infty} \sum_{k=n_S}^{n-1} \frac{1}{k} \\
&\overset{(*)}{\leq} \quad c \frac{\log(n-1) - \log(n_S - 1)}{\sqrt{n-n_S}} \\
&\xrightarrow[n\to\infty]{} \quad 0
\end{aligned}$$

Wobei an der Stelle $(*)$ benutzt wurde, dass $\sum_{k=n_S}^{n-1} \frac{1}{k} \leq \int_{n_S-1}^{n-1} \frac{1}{t} dt = \log(n-1) - \log(n_S-1)$ gilt. Dann folgt

$$\lim_{n\to\infty} \frac{1}{B_n} \sum_{k=h+1}^{n} a_{k-1}^2 \hat{\mathscr{S}}(a_{k-1})^{-\alpha_* - \alpha_{**}} = 0 \,.$$

(ix) Zu zeigen ist hier, dass mit $R_k = (I_{k,1} - \lambda(Z_{k-1}))\, \lambda'(Z_{k-1}) Z_{k-1} a_{k-1}^{\frac{1}{2}}$ und $B_{k,n} = B_n a_{k-1}^{\frac{1}{2}}$ für alle $x \in \mathbb{R}$ gilt:

$$\lim_{n\to\infty} \sup_{n_S+1\leq k\leq n} E\left(R_k^2 \mathbb{1}_{\{R_k^2 \geq B_{k,n}^2 x^2\}} | \mathcal{F}_{k-1}\right) = 0 \text{ f.s.}$$

Dazu sei $g_\nu(x) := \frac{1}{2}\exp(1 - C(\frac{x}{S} - 1))$, dann gilt für die Ableitung der in Gleichung (4.4) definierten Effizienz für alle $k \geq n_S$

$$\lambda'(Z_{k-1}) = \frac{2Z_{k-1}(1+\delta_C)}{(2K_C + Z_{k-1}(1+\delta_C))^2} g_\nu(Z_{k-1}) \,.$$

Da $K_C = \frac{K_M(1+\delta_C)}{2}$ folgt nun für alle $k \geq n_S$,

$$\begin{aligned}
\lambda'(Z_{k-1}) Z_{k-1} &= \frac{2Z_{k-1}^2(1+\delta_C)}{4K_C^2 + 4K_C Z_{k-1}(1+\delta_C) + Z_{k-1}^2(1+\delta_C)^2} \underbrace{g_\nu(Z_{k-1})}_{\leq 1} \\
&\leq 2\left(\frac{4K_C^2}{Z_{k-1}^2(1+\delta_C)} + \frac{4K_C}{Z_{k-1}} + (1+\delta_C)\right)^{-1} \\
&= 2(1+\delta_C)^{-1}\underbrace{\left(\underbrace{\frac{K_M}{Z_{k-1}^2} + \frac{2K_M}{Z_{k-1}}}_{\geq 0} + 1\right)^{-1}}_{\geq 1} \\
&\leq 2 \,.
\end{aligned}$$

Damit gilt:

$$R_k^2 = \underbrace{(I_{k,1} - \lambda(Z_{k-1}))^2}_{\leq 1} \left(\lambda'(Z_{k-1}) Z_{k-1}\right)^2 a_{k-1} \leq 4a_{k-1}$$

Für alle $x \in \mathbb{R}$ bedeutet dies, da die Folge der a_n monoton wachsend ist und $B_n \to \infty$ für $n \to \infty$:

$$\sup_{n_S+1\leq k\leq n} E\left(R_k^2 \mathbb{1}_{\{R_k^2 \geq B_{k,n}^2 x^2\}} | \mathcal{F}_{k-1}\right) \leq 4a_{n-1} P(4 \geq B_n^2 x^2) \xrightarrow[n\to\infty]{} 0$$

Das beweist die Gültigkeit der letzten Voraussetzung. Satz 3.14 sagt nun, dass

$$\hat{\theta}_{h,n,\nu} - K_C \xrightarrow{D} \frac{U\Phi_{h,n}(n_S)}{W^{2(1-\alpha_*)}} = U\Phi_{h,n}(n_S) \text{ für } n \to \infty$$

wobei die Fourier-Transformierte von U die Gestalt

$$\Phi_U(s) = E\left(\exp\left(-\frac{s^2}{2}\sigma^2 W^{1+\beta+2(1-\alpha_*)}\right)\right) = \exp\left(-\frac{s^2}{2}\sigma^2\right)$$

besitzt. Da die Varianz der größenabhängigen PCR $\sigma^2(Z_n) = \frac{K_C}{\mathscr{S}(Z_n)} - r_+(\mathscr{S}(Z_n))$, nach Gleichung (3.29) also $\sigma^2 = K_C$, ist, handelt es sich bei $\Phi_U(s)$ um die Fourier-Transormierte der $\mathcal{N}(0, K_C)$-Verteilung. Somit gilt (4.13). □

Bemerkung 4.8 Das soeben bewiesene Korollar ermöglicht es, aus real gemessenen PCR-Daten, genauer den nach n Zyklen vorhandenen Sequenzen z_n (☞ B.5), den Parameter K_C zu schätzen. Bestimmt man den Schwellenwert S durch Auswertung des Graphen n/z_n, so kann die Effizienz geschätzt werden. Wie in Kapitel 2 beschrieben, lässt sich damit die Mutationsrate schätzen.
Das heißt, dass mit der in dieser Arbeit beschriebenen Theorie alle die PCR charakterisierenden Größen aus während der Reaktion gesammelten Daten geschätzt werden können.

Anhang

A. Einige Hilfssätze

In diesem Kapitel werden die Aussagen, die für die Beweise der Ergebnisse der vorangegangenen Kapitel nötig, aber nicht zum allgemeinen Verständnis der mathematisch modellierten PCR und der Verzweigungsprozesse notwendig sind, nachgeliefert.

Den Anfang macht die bedingte Varianz- bzw. Kovarianzformel:

Lemma A.1 *Gegeben seien Zufallsvariablen X, Y, Z auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Dann gilt mit $Var[X|Z] := E[(X - E[X|Z])^2|Z]$ bzw. $Cov(X, Y|Z) := E[(X - E[X|Z])(Y - E[Y|Z])|Z]$ (zur Definition der bedingten Varianz siehe [12] Kapitel 10.1.2):*

$$(i) VarX = EVar[X|Z] + Var(E[X|Z])$$
$$(ii) Cov(X,Y) = E(Cov(X,Y|Z)) + Cov(E[X|Z], E[Y|Z])$$

BEWEIS: Da $VarX = Cov(X, X)$, genügt es, (ii) zu zeigen.

$$\begin{aligned}
Cov(X,Y) &= E(X - EX)(Y - EY) \\
&= E\,E\big[\big(X - E[X|Z] + E[X|Z] - EX\big)\big(Y - E[Y|Z] + E[Y|Z] - EY\big)|Z\big] \\
&= E\,E\big[(X - E[X|Z])(Y - E[Y|Z]) + \underbrace{(X - E[X|Z])(E[Y|Z] - EY)}_{E\,E(\cdot|Z=0)} \\
&\quad + \underbrace{(E[X|Z] - EX)(Y - E[Y|Z])}_{E\,E(\cdot|Z)=0} + (E[X|Z] - EX)(E[Y|Z] - EY)|Z\big] \\
&= E\,E\big[(X - E[X|Z])(Y - E[Y|Z])|Z\big] \\
&\quad + E\,E\big[(E[X|Z] - EX)(E[Y|Z] - EY)|Z\big] \\
&= E(Cov(X,Y|Z)) + Cov(E[X|Z], E[Y|Z])
\end{aligned}$$

□

Das nachfolgende Lemma entstammt [18] und wird im Beweis des asymptotischen Verhaltens eines größenabhängigen, superkritischen Verzweigungsprozess (Satz 3.4) benötigt.

Lemma A.2 *Es sei $(a_n)_{n\geq 0}$ eine Folge positiver, reeller Zahlen, die der Bedingung*

$$|a_{n+1} - a_n| \leq K a_n^{\beta} m^{-n\alpha - 1} \tag{A.1}$$

mit Konstanten $K, \alpha > 0$, $m > 1$ und $0 \leq \beta < 1$ genüge. Dann existiert ein $0 \leq a < \infty$ mit $\lim_{n\to\infty} a_n = a$. Existiert zudem ein $0 < z_0 < \infty$, das von K, m, α und β abhängt, mit $a_0 > z_0$ dann ist auch $a > 0$.

BEWEIS: Zuerst definiere $b_n := a_n K^{-\frac{1}{1-\beta}}$. Für die Folge $(b_n)_{n\geq 0}$ gilt $b_n > 0$ und wegen Voraussetzung (A.1)

$$\begin{aligned}
|b_{n+1} - b_n| &= \left|(a_{n+1} - a_n) K^{-\frac{1}{1-\beta}}\right| \\
&= |a_{n+1} - a_n|\, K^{-\frac{1}{1-\beta}} \\
&\leq K^{1-\frac{1}{1-\beta}} a_n^{\beta} m^{-n\alpha - 1} = b_n^{\beta} m^{-n\alpha - 1}.
\end{aligned} \tag{A.2}$$

Weiter sei $x_0 := \prod_{n=0}^{\infty} \frac{1}{1-m^{-n\alpha-1}}$. Es ist leicht zu sehen, dass $1 < x_0 < \infty$ und für alle $0 \leq n_1 \leq n_2 < \infty$ gilt

$$x_0 \prod_{k=n_1}^{n_2} \left(1 - m^{-k\alpha-1}\right) > 1. \tag{A.3}$$

Da $\beta \in [0,1)$, folgt aus Gleichung (A.2) für alle $b_n \geq 1$, dass $|b_{n+1} - b_n| \leq b_n m^{-n\alpha-1}$ und damit

$$b_n \left(1 - m^{-n\alpha-1}\right) \leq b_{n+1} \leq b_n \left(1 + m^{-n\alpha-1}\right) \tag{A.4}$$

Nun lassen sich zwei Fälle unterscheiden:
1. Fall: Es gibt ein n_0, derart, dass $b_{n_0} > x_0$. Dann wird mit Induktion nach n gezeigt, dass die Behauptung

$$1 < x_0 \prod_{k=n_0}^{n-1} \left(1 - m^{-k\alpha-1}\right) \leq b_n \leq b_{n_0} \prod_{k=n_0}^{n-1} \left(1 + m^{-k\alpha-1}\right) \tag{A.5}$$

für alle $n > n_0$ richtig ist. Da $b_{n_0} > x_0 > 1$ gilt, folgt für $n = n_0 + 1$ aus den Gleichungen (A.3) und (A.4)

$$1 < x_0 \left(1 - m^{-n_0\alpha-1}\right) < b_{n_0} \left(1 - m^{-n_0\alpha-1}\right) \leq b_{n_0+1} \leq b_{n_0} \left(1 + m^{-n_0\alpha-1}\right).$$

Gelte nun die Behauptung (A.5) für ein $n > n_0$. Dann gilt wegen (A.3) und (A.4)

$$\begin{aligned}
1 &< x_0 \prod_{k=n_0}^{n} \left(1 - m^{-k\alpha-1}\right) \\
&= \underbrace{x_0 \left(\prod_{k=n_0}^{n-1} \left(1 - m^{-k\alpha-1}\right)\right)}_{\leq b_n \text{ nach I.V.}} \left(1 - m^{-n\alpha-1}\right) \\
&\leq b_n \left(1 - m^{-n\alpha-1}\right) \\
&\leq b_{n+1} \\
&\leq b_n \left(1 + m^{-n\alpha-1}\right) \\
&\overset{\text{I.V.}}{\leq} b_{n_0} \left(\prod_{k=n_0}^{n-1} \left(1 + m^{-k\alpha-1}\right)\right) \left(1 + m^{-n\alpha-1}\right) \\
&= b_{n_0} \prod_{k=n_0}^{n} \left(1 + m^{-k\alpha-1}\right)
\end{aligned}$$

Damit ist die Behauptung (A.5) für alle $n > n_0$ bewiesen. Daraus folgt sofort

$$1 < b_n < b_{n_0} \prod_{k=n_0}^{\infty} \left(1 + m^{-k\alpha-1}\right) =: M,$$

wobei die Konstante $M < \infty$ ist. Nun ergibt sich

$$|b_{n+1} - b_n| \leq b_n m^{-n\alpha-1} < M m^{-n\alpha-1} \xrightarrow[n\to\infty]{} 0.$$

Es handelt sich bei der Folge $(b_n)_{n\geq 0}$ also um eine Cauchy-Folge mit dem Grenzwert b, für den $1 \leq b \leq M$ erfüllt ist.
2. Fall: Für alle $n \geq 0$ ist $b_n \leq x_0$. Dann folgt aus Gleichung (A.2) für alle $n \geq 0$

$$|b_{n+1} - b_n| \leq x_0^{\beta} m^{-n\alpha-1}$$

Da x_0, α, β und m nicht von n abhängen, ist der Grenzwert der rechten Seite für $n \to \infty$ gleich 0; somit ist $(b_n)_{n\geq 0}$ eine Cauchy-Folge, also besitzt sie einen endlichen Grenzwert b. In beiden Fällen folgt auch für $(a_n)_{n\geq 0}$, dass es sich um eine Cauchy-Folge handelt, also gilt $\lim_{n\to\infty} a_n = a < \infty$.
Sei nun $b_0 > x_0$. Also ist der 1. Fall erfüllt mit $n_0 = 0$ und damit gilt für den Grenzwert $b \geq 1$. Da $b_n = a_n K^{-\frac{1}{1-\beta}}$, ist $a = bK^{\frac{1}{1-\beta}}$. Falls $a_0 > x_0 K^{\frac{1}{1-\beta}} =: z_0$ ist, folgt $a > K^{\frac{1}{1-\beta}} > 0$ und somit die Behauptung. □

Der folgende Satz wird zum Beweis der starken Konsistenz in Satz 3.11 benötigt und stammt aus [13].

Satz A.3 (Starkes Gesetz der großen Zahlen für Martingale)
Es sei $(S_n = \sum_{i=1}^n X_i)_{n\geq 1}$ ein Martingal bezüglich der Filtration $(\mathcal{F}_n)_{n\geq 1}$ und $(U_n)_{n\geq 1}$ eine monoton wachsende Folge positiver Zufallsvariablen, so dass für alle $n \geq 1$ U_n $\mathcal{F}_{n-1}$-messbar ist.
Für $1 \leq p \leq 2$ konvergiert dann

$$\sum_{i=1}^{\infty} \frac{X_i}{U_i} \text{ f.s. auf } \left\{ \sum_{i=1}^{\infty} \frac{E\left(|X_i|^p | \mathcal{F}_{i-1}\right)}{U_i^p} < \infty \right\} \tag{A.6}$$

und

$$\lim_{n\to\infty} \frac{S_n}{U_n} = 0 \text{ f.s. auf } \left\{ \lim_{n\to\infty} U_n = \infty,\ \sum_{i=1}^{\infty} \frac{E\left(|X_i|^p | \mathcal{F}_{i-1}\right)}{U_i^p} < \infty \right\} \tag{A.7}$$

Für $2 < p < \infty$ gelten dann die Gleichungen (A.6) und (A.7) aber auf der Menge

$$\left\{ \sum_{i=1}^{\infty} \frac{1}{U_i} < \infty,\ \sum_{i=1}^{\infty} \frac{E\left(|X_i|^p | \mathcal{F}_{i-1}\right)}{U_i^{1+\frac{p}{2}}} < \infty \right\}.$$

Beweis: Da der Beweis auf mehreren Ergebnissen beruhen, die in dieser Arbeit nicht vonnöten sind, wird auf [13] Theorem 2.18 verwiesen. □

Das folgende Lemma entstammt [13] und wird für die Konsistenzaussagen in Kapitel 3 benötigt.

Lemma A.4 (Lemma von Toeplitz) *Es seien $(a_{i,n})_{n\geq 1}^{1\leq i\leq k_n}$ und $(x_n)_{n\geq 1}$ Folgen reeller Zahlen. Für jedes $i \in \mathbb{N}$ gelte $a_{i,n} \xrightarrow[n\to\infty]{} 0$ und für alle $n \in \mathbb{N}$ gelte $\sum_{i=1}^{k_n} |a_{i,n}| \leq C < \infty$. Dann gilt:*

(*i*) *Falls $\lim_{n\to\infty} x_n = 0$, folgt*

$$\sum_{i=1}^{k_n} a_{i,n} x_i \xrightarrow[n\to\infty]{} 0\,.$$

(*ii*) *Falls $\lim_{n\to\infty} \sum_{i=1}^{k_n} a_{i,n} = 1$ und $\lim_{n\to\infty} x_n = x$, folgt*

$$\sum_{i=1}^{k_n} a_{i,n} x_i \xrightarrow[n\to\infty]{} x\,.$$

(*iii*) *Gegeben eine Folge positiver, reeller Zahlen* $(a_n)_{n\geq 1}$ *mit* $b_n := \sum_{i=1}^{n} a_i \xrightarrow[n\to\infty]{} \infty$ *und* $\lim_{n\to\infty} x_n = x$, *gilt insbesondere*

$$\frac{1}{b_n}\sum_{i=1}^{n} a_i x_i \xrightarrow[n\to\infty]{} x\,.$$

Beweis:
(i) Geht $x_n \to 0$ für $n \to \infty$, so existiert zu jedem $\varepsilon > 0$ ein n_ε, so dass für alle $n \geq n_\varepsilon$ die Gleichung $|x_n| \leq \frac{\varepsilon}{C}$ gilt. Damit folgt

$$\left|\sum_{i=1}^{k_n} a_{i,n}x_i\right| \leq \sum_{i=1}^{n_\varepsilon-1} |a_{i,n}x_i| + \sum_{i=n_\varepsilon}^{k_n} |a_{i,n}| \underbrace{|x_i|}_{\leq\frac{\varepsilon}{C}} \leq \sum_{i=1}^{n_\varepsilon-1} |a_{i,n}x_i| + \frac{\varepsilon}{C}\underbrace{\sum_{i=n_\varepsilon}^{k_n} |a_{i,n}|}_{\leq C} \leq \sum_{i=1}^{n_\varepsilon-1} |a_{i,n}x_i| + \varepsilon$$

Der Grenzübergang $n \to \infty$ liefert in Kombination mit dem beliebig gewählten ε die Behauptung.
(ii) Es gilt

$$\sum_{i=1}^{k_n} a_{i,n}x_i = x \underbrace{\sum_{i=1}^{k_n} a_{i,n}}_{\xrightarrow[n\to\infty]{}1 \text{ n.V.}} + \underbrace{\sum_{i=1}^{k_n} a_{i,n}(x_i - x)}_{\xrightarrow[n\to\infty]{}0 \text{ nach (i)}} \xrightarrow[n\to\infty]{} x$$

und damit die Behauptung.
(iii) ist ein Spezialfall von (ii) mit $a_{i,n} := \frac{a_i}{b_n}$. □

Zum Beweis der Konvergenzaussage in Satz 3.14 wird der nachfolgende Satz über zufällige Summen verwendet, der in [24] zu finden ist. Zufällige Summen haben allgemein die Gestalt

$$S_n^{(r)} := \sum_{i_1}^{n} \sum_{i_2}^{\eta_{i_1}(n)} \cdots \sum_{i_r=1}^{\eta_{i_1,\ldots,i_{r-1}}(n)} \xi_{i_1\ldots i_r}(n) \tag{A.8}$$

mit (möglicherweise abhängigen) Zufallsvariablen $\xi_{i_1\ldots i_r}(n)$ für $i_1 \ldots i_r := (i_1, \ldots, i_r) \in \mathbb{N}^r$ auf einem beliebigen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Weiter bezeichne $I := (i_1, \ldots, i_r)$ einen Vektor aus $\mathbb{N}^r$ mit aufsteigenden Einträgen. Die Familie $\big(\xi_I(n)\big)_{I\in\mathbb{N}^r}$ sei eine Folge von Zufallsvektoren auf $(\Omega, \mathcal{A}, P)$ mit zugehöriger Familie von Unter-σ-Algebren $\big(\mathcal{F}_I(n)\big)_{I\in\mathbb{N}^r}$, derart dass $\xi_I(n)$ bezüglich $\mathcal{F}_I(n)$ messbar ist. Es sei

$$\begin{aligned}\mathbf{F}_I(n) \;:=\; & \mathcal{F}_0(n) \otimes \bigotimes_{j_r}^{i_r} \mathcal{F}_{i_1\ldots i_{r-1},j_r}(n) \bigotimes_{l=1}^{r-1} \bigotimes_{j_1}^{i_1-1} \otimes \\ & \bigotimes_{j_{l+1}=1}^{\eta_{i_1\ldots i_{l-1}j_1}(n)} \cdots \bigotimes_{j_r=1}^{\eta_{i_1\ldots i_{l-1}j_l\ldots j_{r-1}(n)}} \mathcal{F}_{i_1\ldots i_{l-1}j_l\ldots j_r}(n)\end{aligned}$$

wobei $\mathcal{F}_0(n)$ irgendeine Unter-σ-Algebra von $\mathcal{A}$ sei und $\bigotimes_{j_1=1}^{0} \mathcal{F}_{j_1}(n) = \{\emptyset, \Omega\}$. Für die unter (A.8) definierte Summe und beliebiges $I \in \mathbb{N}^r$ gelte mit $I' := (i_1, \ldots, i_{r-1}, i_r - 1)$:

$$\big\{i_2 \leq \eta_{i_1}(n), i_3 \leq \eta_{i_1 i_2}(n), \ldots, i_r \leq \eta_{i_1\ldots i_{r-1}(n)}\big\} \in \mathbf{F}_I(n)$$

Folgende Notationen werden noch benötigt:

$$\begin{aligned}\vartheta_I(n) &:= E\left[\xi_I(n)|\mathbf{F}_{I'}(n)\right] \\ \sigma_I^2(n) &:= E\left[\left(\xi_I(n) - \vartheta_I(n)\right)^2 |\mathbf{F}_{I'}(n)\right]\end{aligned}$$

Satz A.5 *Gegeben sei die oben geschilderte Situation im Modell der zufallsabhängigen Summen. Existieren Konstanten B_n, $n \in \mathbb{N}$, $\mathcal{A}$-messbare Zufallsvariablen T und γ und eine $\mathcal{A}$-messbare Funktion K, so dass*

$$\lim_{n\to\infty} P\left(\frac{1}{B_n^2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{\eta_{i_1}(n)}\cdots\sum_{i_r=1}^{\eta_{i_1,\ldots,i_{r-1}}(n)}\sigma_I^2(n)\right) = 0\,, \tag{A.9}$$

$$\frac{1}{B_n^2}\max_{1\le i_1\le n}\max_{1\le i_2\le\eta_{i_1}(n)}\cdots\max_{1\le i_r\le\eta_{i_1\ldots i_{r-1}}(n)}\sigma_I^2(n) \xrightarrow{P} 0 \quad \textit{für } n\to\infty\,, \tag{A.10}$$

$$\frac{1}{B_n}\sum_{i_1=1}^{n}\sum_{i_2=1}^{\eta_{i_1}(n)}\cdots\sum_{i_r=1}^{\eta_{i_1,\ldots,i_{r-1}}(n)}\vartheta_I(n) \xrightarrow{P} \gamma \quad \textit{für } n\to\infty \tag{A.11}$$

und für alle $x \in \mathbb{R}$

$$\sum_{i_1=1}^{n}\sum_{i_2=1}^{\eta_{i_1}(n)}\cdots\sum_{i_r=1}^{\eta_{i_1,\ldots,i_{r-1}}(n)}\frac{1}{B_n^2}E\left[(\xi_I(n)-\vartheta_I(n))^2\,\mathbb{1}_{\{\frac{1}{B_n}(\xi_I(n)-\vartheta_I(n))<x\}}|\boldsymbol{F}_{I'}\right] \xrightarrow{P} K(x) \tag{A.12}$$

gelten, dann folgt:

$$\frac{S_n^{(r)}}{B_n} \xrightarrow{d} \omega \quad \textit{für } n\to\infty\,,$$

wobei ω eine Zufallsgröße ist, deren Fourier-Transformierte die Gestalt

$$\phi_\omega(s) = E\left(\exp\left(is\gamma + \int_{\mathbb{R}} L(s,x)dK(x)\right)\right)\,,$$

mit $L(s,x) := -\frac{s^2}{2}$ für $x = 0$ und $L(s,x) := \frac{e^{isx}-1-isx}{x^2}$ sonst, besitzt.

Beweis: Die einführenden Erläuterungen in die Theorie der zufallsabhängigen Summen zeigt, dass diese Theorie viele technische Voraussetzungen und Resultate benötigt. Dies ist auch im Beweis des Satzes der Fall, weshalb an dieser Stelle auf ihn verzichtet und auf [24] verwiesen werden muss. □

B. Biochemische Hintergründe zur Polymerase-Kettenreaktion

Die in Kapitel 1 nur kurz umrissene Reaktion soll hier nun noch einmal näher betrachtet werden. Zum Verständnis des mathematischen Teils ist das im ersten Kapitel bereitgestellte Material vollkommen ausreichend, doch ist es, was die biochemische Genauigkeit anbelangt, äußerst unzureichend.
Die hier aufgeführten Fakten entstammen [6], [7], [9], [22], [26] sowie [31].

B.1. Der Aufbau der DNA

Die DNA (Desoxyribonucleinsäure) gehört zu den kompliziertesten organischen Molekülen. Könnte man ein DNA-Molekül einer menschlichen Zelle entnehmen und komplett gerade ziehen, so erreichte es eine Länge von 2m. In jeder Zelle befindet sich mindestens ein DNA-Molekül. Jedes DNA-Molekül beherbergt Informationen, die die Proteinsynthese steuern und die genetische Information von einer Zelle zur nächsten weiterleiten können. Die Verschlüsselung der Informationen erfolgt über die Abfolge bestimmter Basen.
Das Grundgerüst der DNA bildet ein modifiziertes Zuckermolekül, die Desoxyribose, die durch Reduktion der Ribose entsteht, und in der Furanoseform, d.h. als Ring, vorliegt. Über eine kovalente Bindung am 1-Kohlenstoffatom der Desoxyribose ist ein aminartiges Molekül, und damit eine Base, mit dem Ring verknüpft. Es gibt nur vier mögliche Basen, die dort angelagert sein können: Adenin, Cytosin, Guanin und Thymin. Die beiden Basen Thymin und Cytosin werden *Pyrimidinbasen* genannt und die Basen Adenin und Guanin *Purinbasen*, da sie sich vom Pyrimidin bzw. vom Purin ableiten. Der durch die Verbindung

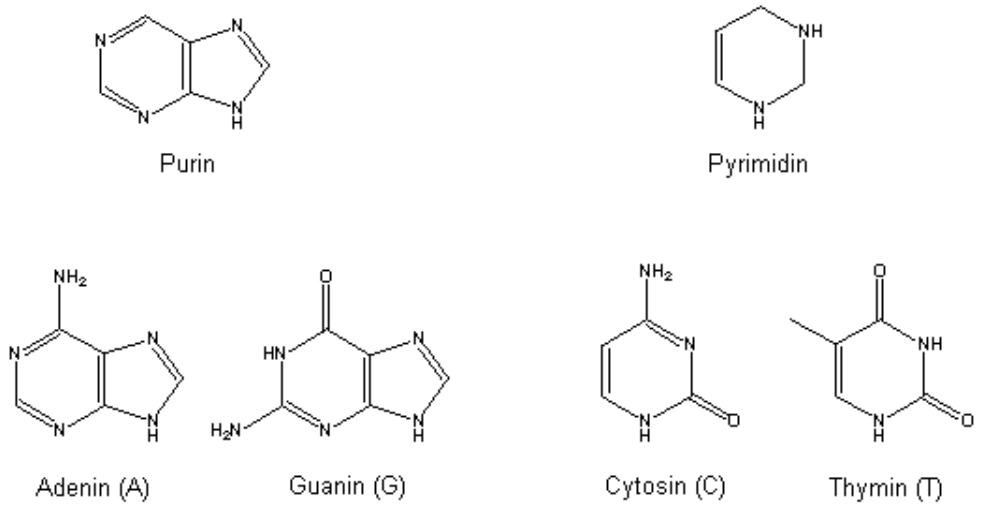

Abbildung B.1.: In der DNA vorkommende Basen und deren Grundgerüst

einer dieser Basen mit einem Desoxyribosemolekül entstandene Baustein wird *Nukleosid*

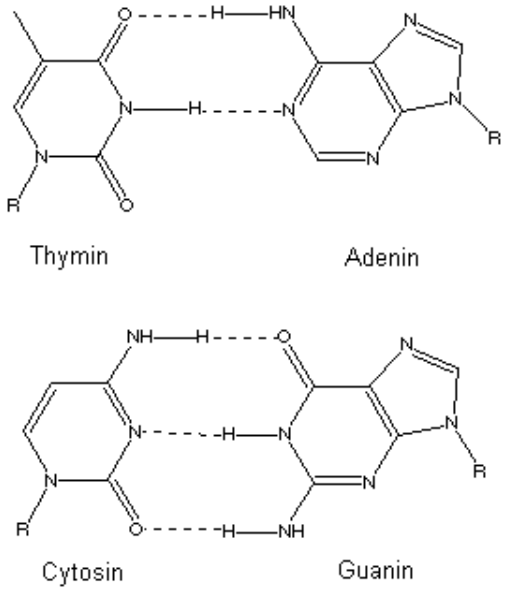

Abbildung B.2.: Wasserstoffbrückenbindungen (gestrichelte Linien) zwischen den Basen, R bezeichnet den Nukleotidrest

genannt. Die Hydroxygruppe am 5er Kohlenstoff wird mit Phosphorsäure H_3PO_4 verestert (die Veresterung ist eine Reaktion zwischen einem Alkohol und einer Säure, bei der ein Wassermolekül abgespalten wird). Dieses aus drei Teilen bestehende Monomer heißt *Nukleotid* und bildet den Baustein der polymeren DNA. Zwei Nukleotide lassen sich in Abbildung B.3 erkennen das obere der beiden Moleküle auf der Reaktandenseite trägt die Base Cytosin, das untere Adenin. Die Phosphatgruppe kann nun mit der Hydroxygruppe am Kohlenstoffatom 3 reagieren, wieder handelt es sich um eine Veresterung, siehe Abbildung B.3. So können immer mehr Nukleotide miteinander reagieren und es entstehen lange Moleküle; das menschliche Genom besteht z.B. aus ca. 3 Milliarden Basenpaaren. Für ein komplettes DNA-Molekül fehlt aber noch ein zweiter Strang. Die DNA ist nämlich eine Doppelhelix bestehend aus zwei mit den Basen einander zugewandten Strängen. Zwischen den Basen der verschiedenen Stränge bilden sich Wasserstoffbrücken aus und sorgen für den Zusammenhalt. Dabei können sich auf Grund der Struktur der Basen nur Adenin und Thymin, bzw. Guanin und Cytosin gegenüberliegen, siehe Abbildung B.2. Zwischen Adenin und Thymin bilden sich nur zwei Wasserstoffbrücken, zwischen Guanin und Cytosin hingegen drei aus.

Der DNA-Doppelstrang wickelt sich auf und die uns bekannte Struktur (Abbildung B.4) entsteht. Durch die Bindung der Basen im Inneren der DNA über Wasserstoffbrücken verlieren die Basen ihre basische Eigenschaft als Protonenakzeptor und die Phosphorsäurebausteine im DNA-Grundgerüst fungieren als Protonendonatoren. Dies führt dazu, dass die DNA im wässrigen Milieu saure Eigenschaften aufweist. So ist die Bezeichnung Desoxyribonuclein*säure* gerechtfertigt.

Abbildung B.3 enthält noch weitere Informationen. Die Kohlenstoffatome des Desoxyriboserings (in der Schreibweise der organischen Chemie werden die Kohlenstoffatome nicht aufgeführt, sondern befinden sich an den Schnittpunkten zweier Strecken, die die Bindung repräsentieren) können nach bestimmten Regeln durchgezählt werden, in der Abbildung zeigen das die blauen Ziffern an. Entsteht durch Polymerisation eine lange Nukleotidkette, so kann man über die Bindung der Desoxyribosemoleküle eine Richtung angeben. Da die Phosphatbindung über den 3'-Kohlenstoff sowie den 5'-Kohlenstoff erfolgt, kann

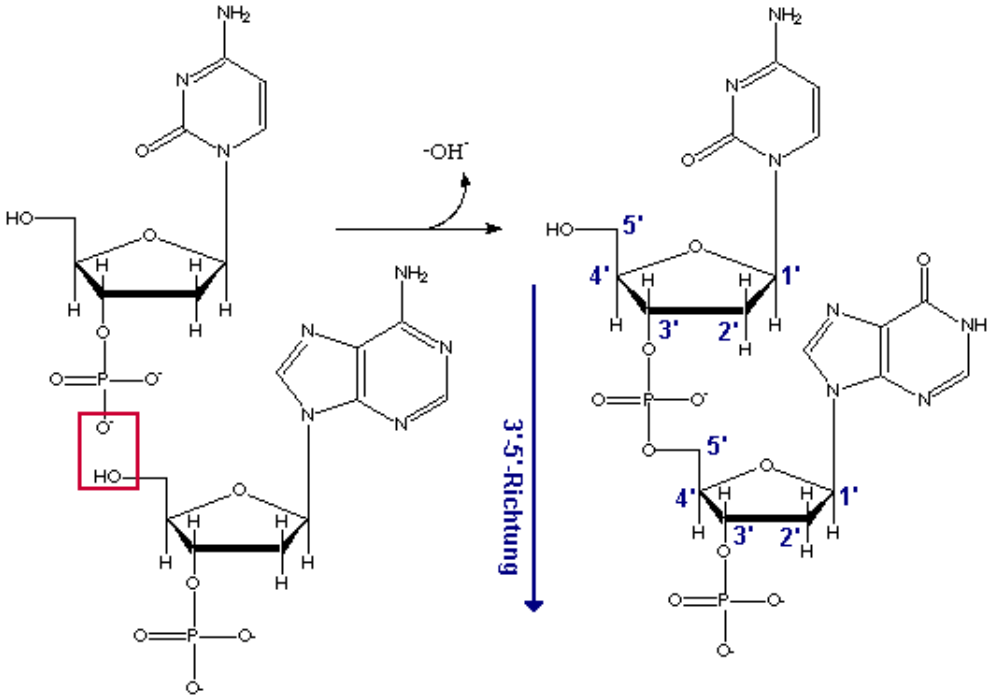

Abbildung B.3.: Polymerisation zweier Nukleotide

ein Strang in 5'-3'-Richtung oder in 3'-5'-Richtung orientiert sein. Die Abkürzung hierfür lauten 3'→5' bzw. 5'→3'.

B.2. Die Reproduktion der DNA

Teilt sich eine Zelle, so muss eine Kopie der DNA angefertigt werden. Dafür setzt das Enzym *Helicase* an der DNA an, trennt an dieser Stelle die Wasserstoffbrücken auf und das Enzym *Polymerase* reproduziert die DNA. Im Detail sieht diese Reaktion folgendermaßen aus.
Die Helicase setzt an einem Punkt in der DNA an und öffnet von dort an die Wasserstoffbrückenbindungen. An diesem Punkt gibt es nun zwei Strangrichtungen. Die Replikation verläuft je nach Richtung des Mutterstranges unterschiedlich ab.
Der *3'→5'-Mutterstrang* bereitet keine Probleme. An der Öffnungsstelle lagert sich eine Startsequenz, ein sogenannter *Pimer* an. Die Polyemarese erkennt diese Sequenz und beginnt dort ihre Arbeit. Sie ergänzt den DNA-Strang durch die Nukleotide mit den komplementären Basen und verknüpft die Nukleotide des neuen Strangs, und zwar in 5'→3'-Richtung.
Weitaus komplizierter ist die Amplifikation des *5'→3'-Mutterstranges*. Die Polymerase kann nämlich nur Basen in 5'→3'-Richtung ergänzen. Deswegen muss sie erst warten, bis die Helicase ein Stück von ca. 1000 Basen geöffnet hat. Dann dockt an der Stelle, bis zu der die Helicase gelangt ist, ein Primer an und die Polymerase ergänzt dort komplementär die Matrize bis sie an die Stelle gelangt, an der die Helicase den DNA-Strang zu öffnen begonnen hat. In der Zwischenzeit hat die Helicase ein weiteres Stück der DNA geöffnet, dort kann sich ein Primer anlagern und eine Polymerase von neuem mit der Amplifikation beginnen. Diese Stücke, bestehend aus Primer und kurzer Sequenz, werden *Okazaki-Fragmente* genannt. Da die Primer aus RNA-Stücken bestehen (siehe Abschnitt

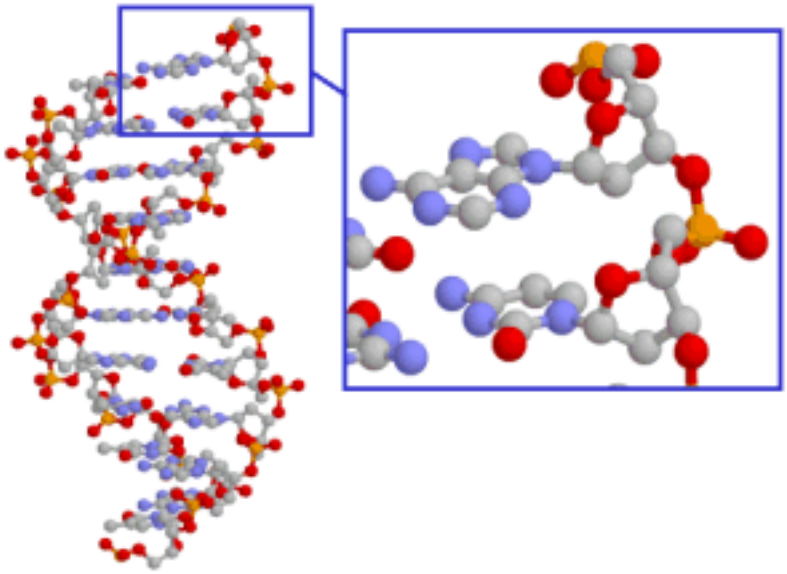

Abbildung B.4.: Modell der DNA-Doppelhelix

B.3), werden die Primer im Anschluss enzymatisch gegen DNA-Sequenzen ausgetauscht. Das Enzym *Ligase* verknüpft zuletzt die von den Primern befreiten Okazaki-Fragmente. So entstehen aus einem DNA-Molekül zwei neue, die jeweils aus einem DNA-Strang der Ursprungssequenz und einem kopierten Strang bestehen.

B.3. Der genetische Code und die RNA

Die Hauptaufgabe der DNA besteht darin, Baupläne zur Erstellung von Proteinen zu liefern. Diese Baupläne, die sich auf bestimmten Abschnitten der DNA, den *Genen*, befinden, bestehen nur aus den oben aufgeführten Basen. Der Baukasten der Proteine besteht aber aus 20 verschiedenen Aminosäuren, die durch die Abfolge der Basen kodiert sind. Deshalb stehen je drei Basen für eine Aminosäure, wobei verschieden Aminosäuren auf mehrere Arten codiert sein können. Eine Übersicht dieser Codes sind z.B. in [6] S. 921 zu finden. Ähnlich wie bei der Replikation der DNA trennt das Enzym Helicase zuerst die beiden Stränge auf und das Enzym RNA-Polymerase setzt zur Vervielfältigung an. Doch gibt es hier zwei Unterschiede zur DNA-Replikation. Damit keine Verwechselungen mit DNA-Replikaten auftreten können, werden andere Nukleotide verwendet. An der Stelle der Desoxyribose sitzt die Ribose (deswegen auch das R in RNA, Ribonucleinsäure) und das Uracil ersetzt das Thymin, siehe Abbildung B.5. Die RNA-Polymerase repliziert nur einen kleinen Teil der DNA, bis ein Stopp-Code auftaucht (z.B. U-A-A). Die so gewonnene „Kopie“, die *messenger-RNA* kurz *mRNA* genannt, wird dann entweder in der Zelle zum Proteinbau genutzt oder in eine andere Zelle geschleust.

B.4. Die PCR

Auch außerhalb einer Zelle kann die DNA vervielfältigt werden. Die von Mullis et al. [26] im Jahre 1988 entwickelte Polymerase-Kettenreaktion (PCR) ist eine solche Methode. Sie basiert weitgehend auf dem natürlichen Mechanismus zur Vervielfältigung der DNA

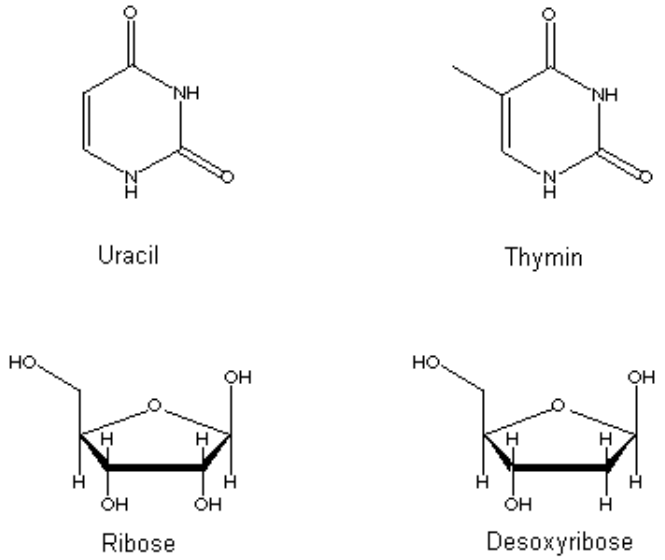

Abbildung B.5.: Gegenüberstellung der unterschiedlichen Moleküle der RNA (links) und der DNA (rechts)

innerhalb der Zellen, kann aber nur Teilsequenzen der DNA von bis zu 2000 Nukleotiden amplifizieren. Eine geringe Menge an DNA, die die zu replizierende Teilsequenz, das Target, enthält, genügt schon, um mit Hilfe der PCR untersuchbare Mengen an Sequenzen zu erzeugen.
Jeder Reaktionszyklus verläuft in drei Phasen. In der ersten Phase, der Denaturierung, wird das Reaktionsgemisch, bestehend aus der Ursprungs-DNA, den Primern, der Polymerase und den Nukleotiden, auf ca. 90°C erhitzt. Dabei trennen sich die beiden Stränge der DNA auf. Das Erhitzen übernimmt also die Arbeit der Helicase. Die Trennung erfolgt vollständig, so dass das Problem der Okazaki-Fragmentierung nicht besteht. An den Anfang des Targets lagern sich im zweiten Schritt die Primer an, dies sind kurzkettige Oligonucleotide mit 15- 30 zum Targetanfang komplementären Basen. Dies geschieht auf beiden Seiten der DNA, es werden also zwei verschiedene Primersorten benötigt. Die Primeranlagerung findet bei ca. 50°C statt. Im letzten Schritt muss die Reaktionslösung auf ca. 70°C erhitzt werden. Dann hat die verwendete Polymerase ihr Reaktionsmaximum und verlängert die DNA komplementär bei den Primern beginnend in 5'-3'-Richtung. Die unterschiedlichen Temperaturen der drei Phasen ermöglicht eine direkte Überwachung und Steuerung der einzelnen Phasen. Ein PCR-Zyklus ist nach Abschluss der dritten Phase beendet und ein neuer Zyklus kann gestartet werden. Zur Verdeutlichung sei auf Abbildung 1.1 verwiesen.
Die Denaturierungstemperatur hängt von der verwendeten DNA ab, je mehr Guanin-Cytosin-Paare im Strang vorkommen, desto höher ist die Schmelztemperatur der DNA. Dies liegt daran, dass diese über drei Wasserstoffbrücken miteinander verbunden sind. Zwischen Adenin und Thymin gibt es nur zwei, siehe Abbildung B.2. Die Denaturierungstemperatur liegt ca. 3°C höher als der Schmelzpunkt der DNA. Diese Temperatur schadet der in den menschlichen Zellen vorkommenden Polymerase, deswegen wird bei der PCR die hitzestabile taq-Polymerase eingesetzt, die aus dem thermophilen Bakterium „Thermus aquaticus“ isoliert wird. Erst mit Erfindung dieser Methode trat die PCR ihren Siegeszug in der Bioanalyse an. Ohne dieses Enzym musste nach jedem Schritt die Polymerase erneuert werden.
Im Laufe der verschiedenen Zyklen entstehen unterschiedliche Produkte. Im ersten Zy-

klus dienen nur die ursprünglichen DNA-Stränge als Matrizen und die Polymerase beginnt den Amplifizierungsprozess an den Primern. Da diese aber auf den Einzelsträngen nur die Anfangspunkte markieren, wird mehr als der eigentlich interessierende Bereich kopiert und so entstehen einseitig terminierte Stränge. Erst im zweiten Zyklus werden zweiseitig terminierte Stränge gebildet. Diese entstehen bei der Replikation einseitig terminierter Stränge. Dort lagern sich die Primer am noch nicht begrenzten Ende des Targets an, und die Amplifikation stoppt mit dem Ende des Ursprungstranges. Die erwartete Anzahl der einseitig terminierten Stränge nach n Zyklen beträgt $Z_0 n(1+\lambda)$, da nur die Ursprungssequenzen diese Stränge erzeugen. Die erwartete Anzahl beidseitig terminierter Stränge nach n Zyklen errechnet sich dann zu $Z_0(1+\lambda)^n - Z_0 n(1+\lambda)$, siehe Lemma 2.10. Da die Unterscheidung zwischen einseitig und beidseitig terminierten Sequenzen mathematisch nicht relevant ist, wurde in der vorangegangenen Arbeit auf diese Unterscheidung verzichtet.
Wie schon zu Beginn des Kapitels 3 erläutert, ist die Effizienz nur in der Anfangsphase der Reaktion konstant. Nach ca. 20 Zyklen sinkt die Wachstumsrate und die exponentielle Wachstumsphase geht in eine lineare Phase, die sogenannte gesättigte Phase über. Abbildung 3.1 zeigt dies an realen PCR-Daten. Die Gründe dafür sind u.a. sterische Hinderung, teilweise Inaktivierung der Polymerase, Verbrauch von Nukleotiden. Eine zu große Anzahl an PCR-Zyklen ist aber nicht nur wegen des Erreichens der gesättigten Phase von Nachteil, sondern auch, weil die Häufigkeit von Mutationen steigt.

B.5. Analyse der PCR-Produkte

In den vorangegangenen Kapiteln sind wir stillschweigend davon ausgegangen, dass sich Größen wie die Gesamtanzahl aller Sequenzen nach n Schritten oder die Anzahl an Mutationen in einem Molekül ohne Probleme bestimmen lassen. Die chemische Realität sieht leider anders aus. Nur über Umwege können diese Größen ermittelt werden.
Die gängigste Methode zur Bestimmung der Anzahl der Sequenzen nach jedem Schritt ist die *Real-Time PCR*. Dabei wird dem Reaktionsgemisch ein fluoreszierender Farbstoff beigesetzt. Von diesem markiert je ein Molekül die Bindung zwischen zwei DNA-Strängen und wird dadurch zur Emission von elektromagnetischer Strahlung in einem bestimmten Wellenlängenbereich angeregt. Direkt nach dem Abschluss eines Zyklus wird die emittierte Strahlung gemessen (daher der Name „Real Time-PCR“) und deren Intensität ist ein Maß für die Anzahl der vorhandenen Sequenzen.
Die Anzahl der Mutationen in einem DNA-Strang wird durch *Sequenzierung* bestimmt. Dabei werden in einer der PCR ähnlichen Reaktion durch Verwendung eines Primertyps ausschließlich einsträngige Sequenzen erzeugt. In verschiedenen Reaktionsansätzen wird jeweils eines der vier Nukleotiden zur Hälfte durch ein verändertes ersetzt. An dieses kann kein weiteres Nukleotid mehr gebunden werden und die Sequenz endet dort. Es entsteht ein Gemisch aus Sequenzen verschiedenster Länge, die durch Gelelektrophorese getrennt werden. Entweder sind die Primer oder die veränderten Nukleotide radioaktiv markiert, so dass anhand der Länge der Sequenzen auf die Stelle geschlossen werden kann, an der das veränderte Nukleotid sitzt. So kann durch sukzessiven Vergleich die gesamte Sequenzabfolge ermittelt werden.

B.6. Die Michaelis-Menten-Kinetik

Um die Polymerase-Kettenreaktion besser zu verstehen, vergegenwärtige man sich, dass es sich um eine von Enzymen katalysierte Reaktion handelt. Damit gilt für ihre Reaktions-

kinetik die klassische, von Leonor Michaelis und Maud Menten im Jahre 1913 entwickelte Enzym-Kinetik. An einer enzymatischen Reaktion sind demnach das Enzym und das Substrat, also der Ausgangsstoff, beteiligt. Über einen real existierenden Zwischenschritt, in dem Enzym und Substrat einen Komplex bilden, reagiert das Substrat mit in der Lösung vorhandenen Stoffen. Das Enzym katalysiert nur die Reaktion des Substrates mit anderen Stoffen und geht unverändert aus der Reaktion hervor. Je höher die Konzentration des Substrates, mit $[S]$ bezeichnet, bei festgehaltener Enzymkonzentration $[E]$ ist, desto höher ist also die Reaktionsgeschwindigkeit. Gilt $[S] > [E]$, ist also die Substratkonzentration größer als die Enzymkonzentration, so hängt die Reaktionsgeschwindigkeit nur von der Geschwindigkeit, mit welcher der Enzym-Substrat-Komplex zerfällt, ab. Diese Maximalgeschwindigkeit wird mit v_{max} bezeichnet. Für jede Enzym-Substrat-Reaktion gibt es dann eine Konstante K_M, die sogenannte Michaelis-Menten-Konstante, so dass für die Reaktionsgeschwindigkeit v die Gleichung

$$v = \frac{v_{max}[S]}{K_M + [S]}$$

gilt. Die Michaelis-Menten-Konstante setzt sich aus verschiedenen Reaktionskonstanten zusammen (zur genaueren Beschreibung der Kinetik und zur Berechnung der Konstante siehe [7] Kapitel 25.2.2) und entspricht derjenigen Substratkonzentration, bei der die Hälfte der maximalen Reaktionsgeschwindigkeit erreicht wird. Ein kleiner K_M-Wert bedeutet, dass eine hohe Affinität zwischen Enzym und Substrat besteht.

Symbolverzeichnis

$\mathbb{1}_A$	Indikatorfunktion der Menge A
$\alpha = (i, j)$	j-te Sequenz der i-ten Generation
$\bar{A}_n$	Generationsnummer des MRCA eines zufällig gezogenen Paares nach n PCR-Zyklen
a_n	$= m^n$ im superkritischen Fall; $= a_{n-1}m(a_{n-1})$ im fast-kritischen Fall wobei $a_0 := 1$
$\mathcal{B}(1, p)$	Bernoulli-Verteilung mit dem Parameter $p \in [0, 1]$
$\mathcal{B}(n, p)$	Binomialverteilung mit den Parametern $n \in \mathbb{N}$ und $p \in [0, 1]$
$C_n(k)$	erwartete Anzahl von Paaren mit einem MRCA der Generation k nach n PCR-Zyklen
$Cov(X, Y)$	Kovarianz der Zufallsvariablen X und Y
$Cov[X, Y\|Z]$	$= E[(X - E[X\|Z])(Y - E[Y\|Z])\|Z]$ für Zufallsvariablen X, Y, Z; bedingte Kovarianz
$d(\alpha, \beta)$	Abstand/Distanz zwischen den Sequenzen α und β
D	Abstand/Distanz zwischen zwei zufällig gezogenen Sequenzen
EX	Erwartungswert der Zufallsvariable X
$E[X\|Z]$	bedingter Erwartungswert der Zufallsvariable X unter Z
$f_X(s)$	$= Es^X$, erzeugende Funktion einer $\mathbb{N}_0$-wertigen Zufallsgröße X
$F_X(x)$	Verteilungsfunktion einer Zufallsvariablen X
G	Länge des Targets
$g(\alpha)$	Generationsnummer der Sequenz α
H	Hamming-Abstand, d.h. Anzahl der verschiedenen Basen zweier zufällig gezogener Sequenzen
K	Generationsnummer einer zufällig gezogenen Sequenz
K_C	$= \frac{K_M}{2}(1 + \delta_C)$ mit $\delta_C = \mathbb{1}_{\{C=0\}}$
K_M	Michaelis-Menten-Konstante
$\mathcal{L}_2(\Omega, \mathcal{A}, P)$	Vektorraum der reellen, 2-fach P-integrierbaren Zufallsgrößen auf $(\Omega, \mathcal{A})$
λ	Effizienz der größenunabhängigen PCR
$\lambda(\cdot)$	Effizienz der größenabhängigen PCR; $\lambda(\cdot) : \mathbb{N} \to [0, 1]$
μ	Mutationsrate
M	Anzahl der Mutationen einer zufällig gezogenen Sequenz
$M(\alpha)$	Anzahl der Mutationen der Sequenz α
$\mathrm{MRCA}(\alpha, \beta)$	letzter gemeinsamer Vorfahr der Sequenzen α und β
$m(Z_n)$	$= E\left(I_{n+1,j} \mid \mathcal{F}_n\right)$, erwartete Anzahl an Nachkommen einer dem n-ten Zyklus entstammenden Sequenz
$m_{\theta,\nu}(\cdot)$	von θ, ν abhängendes Nachkommenmittel
$\mathbb{N}$	Menge der positiven natürlichen Zahlen 1,2,...
$\mathbb{N}_0$	Menge der natürlichen Zahlen 0,1,2,...
$\mathcal{N}(\mu, \sigma^2)$	Normalverteilung mit den Parametern $\mu \in \mathbb{R}$ und $\sigma \in [0, \infty)$
$(\Omega, \mathcal{A})$	messbarer Raum
$(\Omega, \mathcal{A}, P)$	Wahrscheinlichkeitsraum
$\mathcal{O}(g(x)) = f(x)$	für $x \to \xi \in \bar{\mathbb{R}} \ :\Leftrightarrow \ \limsup_{x\to\xi} \frac{\|f(x)\|}{g(x)} < \infty$

$o(g(x)) = f(x)$	für $x \to \xi \in \bar{\mathbb{R}} \ :\Leftrightarrow \ \lim_{x\to\xi} \frac{\lvert f(x)\rvert}{g(x)} = 0$
$P(A)$	Wahrscheinlichkeit der Menge A
P^X	Verteilung der Zufallsvariablen X unter P
$P_n(k)$	erwartete Anzahl von Paaren mit einem Abstand k nach n PCR-Zyklen
$\varphi_C(s)$	$= \sum_{k=0}^{\infty} C(k)s^k$, erzeugende Funktion einer Folge $(C(k))_{k\geq 0}$
$\phi_X(s)$	$= Ee^{isX}$, Fouriertransformierte einer Zufallsvariablen X
$Poi(\lambda)$	Poisson-Verteilung mit Parameter $\lambda \in [0, \infty)$
$\mathbb{R}$	Menge der reellen Zahlen
$\bar{\mathbb{R}}$	$= \mathbb{R} \cup \{-\infty, \infty\}$
$\mathbb{R}^+$	Menge aller reellen Zahlen > 0
$\mathbb{R}_0^+$	Menge aller reellen Zahlen ≥ 0
$\tilde{S}_{h,n,\nu,\gamma}$	Kontrast, von h, n, ν, γ abhängig
$S_{h,n,\nu,\gamma}$	normalisierter Kontrast, von h, n, ν, γ abhängig
$\sigma^2(Z_n)$	$= Var\,(I_{n+1,j} \mid \mathcal{F}_n)$, Varianz einer dem n-ten Zyklus entstammenden Sequenz im größenabhängigen Fall
$\sigma(X)$	von der Zufallsvariablen X erzeugte σ-Algebra
$\mathscr{S}(x)$	$= S\mathbb{1}_{\{x<S\}} + x\mathbb{1}_{\{x\geq S\}}$ für ein $S \in \mathbb{N}$ und alle $x \in \mathbb{R}$
$\hat{\theta}_{h,n,\nu,\gamma}$	Schätzer für θ_0, minimalisiert den Kontrast
$VarX$	Varianz der Zufallsvariablen X
$Var[X\|Z]$	$= E[(X - E[X\|Z])^2\|Z]$ für Zufallsvariablen X, Y; bedingte Varianz
W_n	$= \frac{Z_n}{a_n}$
X_k^n	Gesamtanzahl der Sequenzen der k-ten Generation nach n PCR-Zyklen
$X_k^n(i)$	Anzahl der vom i-ten Startmolekül abstammenden Sequenzen der k-ten Generation nach n PCR-Zyklen
Z_n	Gesamtanzahl der Sequenzen nach n PCR-Zyklen
$Z_n(i)$	Anzahl der vom i-ten Startmolekül abstammenden Sequenzen nach n PCR-Zyklen
$\mathcal{Z}_n$	Menge aller Sequenzen nach dem n-ten PCR-Zyklus
$A \cup B$	Vereinigung der Mengen A und B
$A \cap B$	Schnitt der Mengen A und B
$A - B$	$= A \cap B^c$
$A \times B$	$= \{(a, b) : \ a \in A, b \in B\}$
$\overset{\circ}{A}$	Menge der inneren Punkte der Menge A
$\lvert A\rvert$	Mächtigkeit/Anzahl der Elemente der Menge A
$\xrightarrow{P}$	stochastische Konvergenz
$\xrightarrow{d}$	Verteilungskonvergenz
$\xrightarrow{\mathcal{L}_2}$	Konvergenz im zweiten Mittel
$x \vee y$	$= \max\{x, y\}$
$x \wedge y$	$= \min\{x, y\}$
$\approx$	gerundet
$X \sim Y$	X und Y besitzen dieselbe Verteilung, d.h. $P^X = P^Y$
$X \sim Q$	X besitzt die Verteilung Q, d.h. $P^X = Q$
$\lVert\cdot\rVert_2$	$\mathcal{L}_2$-Norm, d.h. für eine Zufallgröße X auf $(\Omega, \mathcal{A}, P)$ gilt $\lVert X\rVert_2 := \left(\int_\Omega \lvert X\rvert^2\right)^{\frac{1}{2}}$
$:=$	ist definiert als
$\mathcal{A}_1 \otimes \mathcal{A}_2$	Produkt-σ-Algebra

Abkürzungsverzeichnis

DNA	Desoxyribonucleinsäure
f.s.	fast sicher
F.T.	Fourier-Transformierte
i.i.d.	independent, identically distributed; unabhängig identisch verteilt
PCR	polymerase chain reaction; Polymerase-Kettenreaktion
st.u.	stochastisch unabhängig

Literaturverzeichnis

[1] Alsmeyer, G.: *Mathematische Statistik*. Skripten zur Mathematischen Statistik Nr. 36. Universität Münster (2002).

[2] Alsmeyer, G.: *Stochastische Prozesse Teil 1*. Skripten zur Mathematischen Statistik Nr. 33, 2. erweiterte Auflage. Universität Münster (2002).

[3] Alsmeyer, G.: *Wahrscheinlichkeitstheorie*. Skripten zur Mathematischen Statistik Nr. 30, 3. Auflage. Universität Münster (2003).

[4] Alsmeyer, G.: *Verzweigungsprozesse*. Skript. Universität Münster (2005).

[5] Athreya, K.B.; Ney, P.E.: *Branching Processes*. Springer. Berlin, Heidelberg, New York (1972).

[6] Atkins, P.W.; Beran, J.A.: *Chemie einfach alles*. 2., korrigierte Auflage. VCH. Weinheim, New York, Basel, Cambridge, Tokyo (1998).

[7] Atkins, P.W.: *Physikalische Chemie*. 3., korrigierte Auflage. Wiley-VCH. Weinheim, New York, Chinchester (2001).

[8] Bauer, H.: *Wahrscheinlichkeitstheorie*. 4., überarbeitete Auflage. Walter de Gruyter. Berlin, New York (1991).

[9] Campbell, N.; Reece, J.B.: *Biology*. 7. Auflage. Pearson. San Francisco, Boston, New York (2005).

[10] Elstrodt, J.: *Maß- und Integrationstheorie*. 3. erweiterte Auflage. Springer. Berlin, Heidelberg, New York (2002).

[11] Flachsmeyer, J.: *Kombinatorik. Eine Einführung in die mengentheoretische Denkweise*. 3. Auflage. VEB Deutscher Verlag der Wissenschaften. Berlin (1972).

[12] Gut, A.: *Probability: A Graduate Course*. Springer. New York (2005).

[13] Hall, P.; Heyde, C.C.: *Martingale Limit Theory and Its Application*. Academic Press. New York, London, Toronto, Sydney, San Francisco (1980).

[14] Heuser, H.: *Lehrbuch der Analysis Teil 1*. 15. Auflage. Teubner. Stuttgart, Leipzig, Wiesbaden (2003).

[15] Jagers, P.; Klebaner, F.: *Random variation and concentration effects in PCR*. Journal of Theoretical Biology 224, 299-304 (2003).

[16] Keller, G.; Kersting, G.; Rösler, U.: *On the asymptotic behaviour of discrete time stochastic growth processes*. The Annals of Probability, Vol. 15, No. 1, 305-343 (1987).

[17] Kersting, G.: *Some properties of stochastic difference equations*. Stochastic Modelling in Biology, 328-339 (1990).

[18] Klebaner, F.C.: *On population-size-dependent branching processes.* Advances of Applied Pobability 16, 30-55 (1984).

[19] Küster, P.: *Asymptotic growth of controlled Galton-Watson-processes.* The Annals of Probability, Vol. 13, No. 4, 1157-1178 (1985).

[20] Lalam, N.; Jacob, C.: *Estimation of the offspring mean in a supercritical or near-critical size-dependent branching process.* Advances of Applied Probability 36, 582-601 (2004).

[21] Lalam, N.; Jacob, C.; Jagers, P.: *Modelling the PCR amplification process by a size-dependent branching process and estimation of the efficiency.* Advances of Applied Probability 36, 602-615 (2004).

[22] Linz, U.; Degenhardt, H.: *Die Polymerase-Kettenreaktion. Ein Überblick.* Naturwissenschaften 77, 515-530 (1990).

[23] Loève, M: *Probability Theory II.* 4. Auflage. Springer. New York, Heidelberg, Berlin (1978).

[24] Rahimov, I.: *Random Sums and Branching Stochastic Processes.* Lecture Notes in Statistics 96. Springer. New York (1995).

[25] Schmitz, N.: *Stochastik für Lehramtsstudenten.* Münsteraner Einführungen: Mathematik/Informatik, Band 1. Münster (1997).

[26] Saiki, R.K.; Gelfand, D.H.; Stoffel, S. et al.: *Primer-directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase.* Science 239, 487-491 (1988).

[27] Sun, F.: *The Polymerase Chain Reaction and Branching Processes.* Journal of Computational Biology, Volume 2, No. 1, 63-86 (1995).

[28] Weiss, G.; von Haesseler, A.: *Modeling the Polymerase Chain Reaction.* Journal of Computational Biology, Volume 2, No. 1, 49-61 (1995).

[29] Witting, H.; Müller-Funk, U.: *Mathematische Statistik II.* B.G. Teubner. Stuttgart (1995).

[30] Wu, C.-H.: *Asymptotic theory of nonlinear least square estimation.* The Annals of Statistics, Volume 9, No. 3, 501-513 (1981).

[31] http://www.dorak.info/genetics/realtime.html

Hiermit versichere ich, dass ich die vorgelegte Diplomarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen wurden, habe ich unter Angabe der Quellen kenntlich gemacht.

Münster, den _____

_____

Carsten Magnus