Stochastische Modellierung im Gymnasialunterricht

Hansruedi Künsch
Seminar für Statistik
ETH Zürich

2. Februar 2020

1 Einleitung

Im Mathematikunterricht im Gymnasium wird seit einiger Zeit auch die mathematische Modellierung anhand von Differentialgleichungen besprochen. In Anwendungen ausserhalb der Physik, insbesondere in der Biologie, ist jedoch eine rein deterministische Modellierung oft fragwürdig, da sie der inhärenten Unsicherheit der betrachteten Prozesse nicht gerecht wird und die natürliche Variabilität nicht wiedergeben kann. Ferner gibt es viele Situationen, wo die betrachteten Variablen Anzahlen sind, die sich nur sprungweise verändern können.

Als Alternative bieten sich stochastische Modelle an, insbesondere Markovprozesse, bei denen die weitere Entwicklung nur vom jetzigen Zustand und nicht von der ganzen Vergangenheit abhängt. Sind die möglichen Zustände diskret, also z.B. Anzahlen, dann spricht man von Markovketten. Ich stelle hier ein paar Markovkettenmodelle in stetiger Zeit vor, die stochastische Erweiterungen von einfachen Differentialgleichungsmodellen sind, und illustriere, wie sich das Verhalten der Lösungen dabei ändert.

Es ist mir bewusst, dass eine umfassende Behandlung von stochastischen dynamischen Modellen im Grundlagenfach wohl kaum möglich ist. Ich finde es aber wichtig, im Sinne einer kritischen Behandlung von deterministischen Modellen auch auf deren Schwachpunkte hinzuweisen und einen kurzen Ausblick auf die stochastischen Modelle zu geben. Computersimulationen dieser Modelle sind ein gutes Hilfsmittel zum anschaulichen Verständnis. Ich erkläre daher auch die dazu verwendeten Algorithmen und stelle deren Implementation im statistischen Softwaresystem R in einem separaten File zur Verfügung.

Im Ergänzungsfach werden Markovketten manchmal besprochen, allerdings meist in diskreter Zeit. Man findet dazu auch Unterlagen auf der Gymnasialstufe, z.B. in den Büchern von Lambacher-Schweizer. Dies ist insbesondere dann naheliegend, wenn auch eine Einführung in Matrizen und lineare Algebra gegeben wird, weil es viele Querverbindungen gibt. Im Kapitel 4 diskutiere ich daher noch allgemeine Markovketten in stetiger Zeit und zeige insbesondere einige dieser Querverbindungen. Man kann sich bei der Lektüre dieses Textes aber auch auf die Kapitel 2 und 3 beschränken, die weniger anspruchsvoll sind. Auch die Beweise der drei Lemmata in Kapitel 2 kann man ohne weiteres überspringen.

Systematische Behandlungen von Markovketten und deren Anwendungen findet man in vielen Büchern. Ich erwähne nur vier davon: Karlin and Taylor (1975) ist ein Klassiker, der zu meiner Studienzeit mein Interesse am Thema geweckt hat. Modernere Darstellungen, die nicht allzu hohe mathematische Ansprüche stellen, sind Resnick (1992) oder Durrett (2016). Allen (2011) behandelt Anwendungen in der Biologie und ist ein guter Einstieg, wenn man mehr über die Modellierung von Epidemien wissen möchte als was in Kapitel 3 steht. In diesem Buch findet man auch MATLAB Programme zur Simulation von stochastischen Prozessen.

2 Stochastische Modelle für das Wachstum von Populationen

Die erste Differentialgleichung, denen SchülerInnen begegnen, ist wohl das exponentielle Wachstum

$$\dot{X}(t) = \lambda X(t) \tag{1}$$

mit der Lösung $X(t) = X(0)\exp(\lambda t)$. Als Modell für das Wachstum einer biologischen Population hat diese Differentialgleichung jedoch die in der Einleitung erwähnten Schwachstellen: $X(t)$ ist notwendigerweise ganzzahlig, und die Zeit, bis sich ein Individuum reproduziert oder stirbt, kann nicht exakt vorhergesagt werden, sondern ist auch vom Zufall bestimmt.

In einem stochastischen Wachstumsmodell einer Population ist die Grösse der Population $X(t)$ zur Zeit t eine Zufallsvariable mit Werten in $\{0, 1, 2, ...\}$. Veränderungen in $X(t)$ geschehen an zufälligen Zeitpunkten, an denen $X(t)$ um 1 nach oben (im Fall einer Geburt) oder nach unten (bei einem Todesfall) springt. Im stochastischen Modell ist der Verlauf von $X(t)$ somit eine zufällige Sprungfunktion.

Sowohl das deterministische als auch das stochastische Modell überlegen, wie sich $X(t)$ in einem kurzen Zeitschritt verändet, und lassen dann diesen Zeitschritt gegen Null konvergieren. Während eine Differentialgleichung eine Veränderung postuliert, welche proportional zum betrachteten Zeitschritt ist, postuliert ein stochastisches Modell, dass die Wahrscheinlichkeiten für mögliche Sprünge proportional sind zum betrachteten Zeitschritt. Beide Modelltypen nehmen an, dass der weitere Verlauf des Prozesses, d.h. die Werte $X(u)$ für $u > t$, nur vom aktuellen Zustand $X(t)$ und nicht von der Vorgeschichte, d.h. den Werten $X(s)$ für $s < t$, abhängt. Dies wird bei stochastischen Modellen die **Markoveigenschaft** genannt.

2.1 Der reine Geburtsprozess

Das einfachste Modell betrachtet eine Population, in der nur Geburten, das heisst nur Sprünge der Höhe +1 auftreten. Wenn die Population zur Zeit t aus n Individuen besteht, $X(t) = n$, dann ist die Wahrscheinlichkeit für eine Geburt, d.h. einen Sprung von n nach $n + 1$, im Zeitraum $(t, t + h]$ in erster Näherung proportional zur Schrittlänge h multipliziert mit der Anzahl Individuen n zur Zeit t. Ferner ist die Wahrscheinlichkeit für mehr als einen Sprung im Zeitraum $(t, t + h]$ von kleinerer Grössenordnung als h. In Formeln besagt dieses Modell also

$$P(X(t + h) = n + 1 \mid X(t) = n) \approx \lambda \cdot n \cdot h, \tag{2}$$

$$P(X(t + h) = n \mid X(t) = n) \approx 1 - \lambda \cdot n \cdot h. \tag{3}$$

wobei $\approx$ bedeutet, dass der Fehler rascher als h gegen Null konvergiert.

Wie lange bleibt die Grösse der Population konstant gleich n? Wenn man die Zeit diskretisiert mit einer kleinen Schrittlänge h und annimmt, dass (2) und (3) exakt gelten, dann kann man die Verteilung dieser Zufallsvariablen leicht angeben: Die Wahrscheinlichkeit, dass sie den Wert h annimmt, ist offensichtlich gleich λnh. Sie nimmt den Wert $2h$ an, falls im ersten Intervall kein Sprung und im zweiten Intervall ein Sprung erfolgt. Die Wahrscheinlichkeit dafür ist also $(1 - \lambda nh)\lambda nh$ (wegen der Markoveigenschaft werden die Wahrscheinlichkeiten multipliziert). Allgemein nimmt diese Zufallsvariable den Wert kh an, wenn in den ersten $k-1$ Intervallen kein Sprung und im k-ten Intervall ein Sprung erfolgt, und die Wahrscheinlichkeit dafür ist gleich $(1 - \lambda nh)^{k-1}\lambda nh$. Die Zeit, während der die Grösse der Population gleich n ist, hat somit eine geometrische Verteilung mit Parameter λnh.

Da Wahrscheinlichkeiten nicht grösser als 1 sein können, macht diese Approximation jedoch nur dann Sinn, wenn $h \leq \frac{1}{\lambda n}$ ist. Es gibt daher keinen festen Zeitschritt h, für den man diese Approximation zu beliebigen Zeiten t verwenden kann. Für $h \to 0$ konvergiert die geometrische Verteilung mit Parameter λnh gegen die Exponentialverteilung mit Parameter λn. Das folgende

Lemma gibt eine direkte Begründung dafür, dass die Zeit, während der die Grösse der Population gleich n ist, exponentialverteilt ist.

Lemma 1. *Wenn die Markoveigenschaft gilt und wenn $P(X(t+h) = n \mid X(t) = n)$ nicht von t, sondern nur von h und n abhängt, dann ist (3) äquivalent dazu, dass für alle $h > 0$ und alle t*

$$P(X(t+h) = n \mid X(t) = n) = \exp(-\lambda \cdot n \cdot h). \tag{4}$$

Gegeben $X(t) = n$ ist also die Wartezeit auf den Sprung von n nach $n+1$ exponentialverteilt mit Parameter λn. Wenn die Markoveigenschaft gilt, sind diese Wartezeiten unabhängig für verschiedene Werte von n und dann ist auch (2) erfüllt.

Beweis: Weil $\exp(-x) \approx 1 - x$ für $x \to 0$ mit einem Fehler von kleinerer Ordnung als x, folgt aus (4) sofort (3). Für die Umkehrung benutzt man, dass für beliebige Zeitschritte $h_1 > 0$ und $h_2 > 0$ gilt

$$\begin{aligned} & P(X(t+h_1+h_2) = n \mid X(t) = n) = P(X(t+h_1+h_2) = n, X(t+h_1) = n \mid X(t) = n) \\ = \;& P(X(t+h_1+h_2) = n \mid X(t+h_1) = n) \cdot P(X(t+h_1) = n \mid X(t) = n). \end{aligned}$$

Die erste Gleichheit gilt, weil X nur Sprünge von $+1$ hat. Die zweite Gleichheit folgt aus der allgemeinen Formel

$$P(A \cap B \mid C) = P(A \mid B \cap C) \cdot P(B \mid C)$$

und daraus, dass wegen der Markoveigenschaft

$$P(X(t+h_1+h_2) = n \mid X(t+h_1) = n, X(t) = n) = P(X(t+h_1+h_2) = n \mid X(t+h_1) = n).$$

Die Funktion $h \mapsto f_n(h) := P(X(t+h) = n \mid X(t) = n)$ erfüllt also die Funktionalgleichung der Exponentialfunktion und wegen (3) konvergiert $f_n(h)$ gegen 1 für $h \to 0$. Somit ist $f_n(h) = \exp(ah)$ und aus (3) folgt $a = -\lambda n$.

Wenn $X(t) = n$, dann ist das Ereignis $X(t+h) = n$ äquivalent dazu, dass die Wartezeit von t bis zum Zeitpunkt des nächsten Sprungs grösser ist als h. Daher ist diese Wartezeit exponentialverteilt mit Parameter λn. Wenn $X(t) = n$ und $X(t+h) > n+1$, dann muss sowohl die Wartezeit auf den Sprung von n nach $n+1$ als auch die Wartezeit auf den Sprung von $n+1$ nach $n+2$ kleiner oder gleich h sein. Daher gilt

$$\begin{aligned} P(X(t+h) > n+1 \mid X(t) = n) &\leq P(X(t+h) = n+1 \mid X(t) = n) \cdot \\ P(X(t+h) = n+2 \mid X(t) = n+1) &\leq (1 - \exp(-\lambda n h))(1 - \exp(-\lambda(n+1)h)) = O(h^2). \end{aligned}$$

Somit ist (2) erfüllt. □

Dieses Lemma gibt eine explizite Konstruktion der zufälligen Sprungfunktion X. Die Wartezeit T_n auf den Sprung von n nach $n+1$ kann man mit Hilfe einer auf $(0,1)$ gleichverteilten Zufallsvariablen U_n konstruieren, indem man $T_n = -\log(U)/(\lambda n)$ setzt. Dann gilt nämlich

$$P(T_n > h) = P(-\log(U) > h\lambda n) = P(U \leq \exp(-h\lambda n)).$$

Es ist also nicht nötig, Zufallsvariablen mit allgemeiner stetiger Verteilung zu besprechen, weil das Konzept einer auf $(0,1)$ gleichverteilten Zufallsvariable unmittelbar einleuchtet.

Der folgende Algorithmus simuliert Sprungfunktionen gemäss dem stochastischen Modell für das exponentielle Wachstum (Die Variable S gibt an, wie weit die Simulation fortgeschritten ist):

1. Wähle den Anfangswert $X(0)$ und setze $S = 0$.
2. Für $k = 1, 2, \ldots$ wähle auf $(0,1)$ gleichverteilte und unabhängige Zufallsvariablen U_k und setze

- $T_k = -\log(U_k)/(\lambda \cdot X(S))$,
- $X(t) = X(S)$ für $S < t < S + T_k$,
- $X(S + T_k) = X(S) + 1$,
- $S = S + T_k$.

Um den Zusammenhang mit der deterministischen Differentialgleichung herzustellen, berechnen wir den Erwartungswert von $X(t + h)$, wenn $X(t) = n$ gegeben ist. Wenn $X(t) = n$, dann kann $X(t + h)$ in erster Näherung nur die beiden Werte n und $n + 1$ annehmen. Wegen (2) und (3) gilt also

$$E(X(t + h) \mid X(t) = n) \approx n(1 - \lambda nh) + (n + 1)n\lambda h = n + n\lambda h$$

bzw. ohne Festlegung eines spezifischen Werts von $X(t)$,

$$E(X(t + h) - X(t) \mid X(t)) \approx \lambda h X(t).$$

Bildet man auf beiden Seiten der letzten Gleichung den Erwartungswert über $X(t)$, dann folgt daraus

$$E(X(t + h)) - E(X(t)) \approx \lambda h E(X(t)).$$

Die Approximationen (2) und (3) ergeben also im Erwartungswert die Euler-Approximation der deterministischen Gleichung.

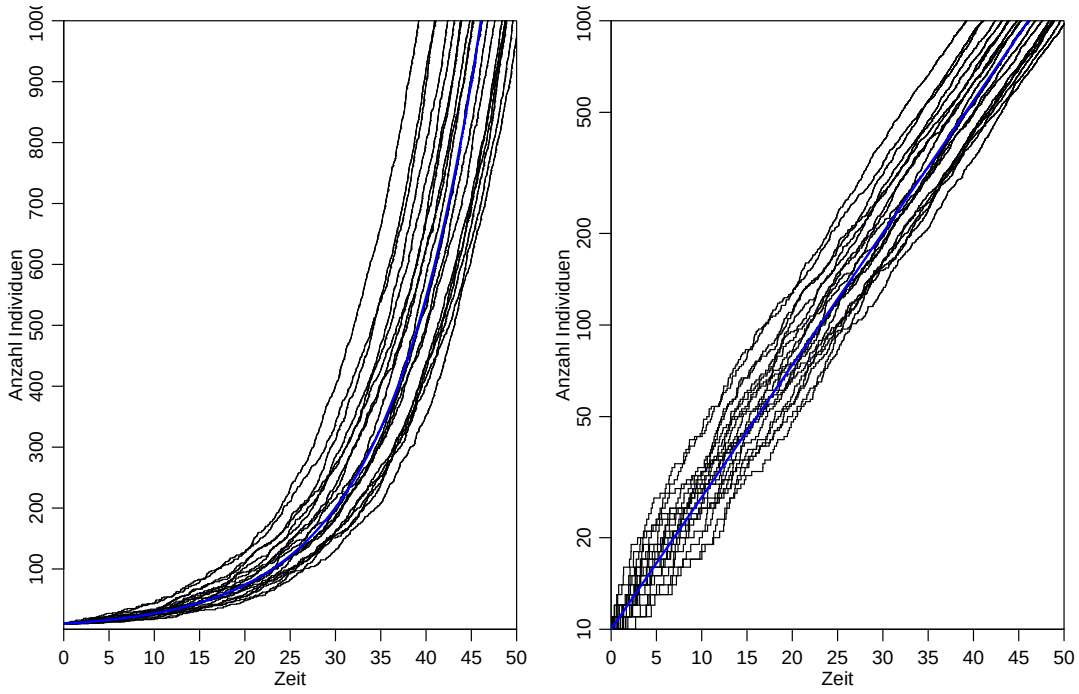

Abbildung 1: 20 simulierte Verläufe des Geburtsprozesses mit $\lambda = 0.1$ und $X(0) = 10$ in ursprünglicher (links) und logarithmischer (rechts) Skala. In Blau ist die Lösung des zugehörigen deterministischen Systems dargestellt.

Um zu sehen, wie stark sich Lösungen des stochastischen Modells von der Lösung des deterministischen Modells unterscheiden, sind in Abbildung 1 zwanzig unabhängige Sprungfunktionen gezeichnet, welche mit obigem Algorithmus generiert wurden. Man sieht, dass zwar die Sprünge immer rascher aufeinander folgen, dass aber die Funktionen für das Auge trotzdem nicht glatt aussehen. Die Variabilität zwischen den einzelnen Funktionen ist beträchtlich. Die rechte Figur mit der logarithmischen Skala auf der y-Achse legt die Vermutung nahe, dass sich die relativen Abweichungen von $X(t)$ vom deterministischen Verlauf $X(0)e^{\lambda t}$ für $t \to \infty$ bei einem zufälligen Wert stabilisieren.

Da die Sprünge immer rascher aufeinander folgen, ist es naheliegend zu fragen, ob die im obigen Algorithmus konstruierte Sprungfunktion für beliebige Zeiten t definiert ist, oder ob die Sprungzeiten gegen einen endlichen Grenzwert konvergieren (analog wie im Problem von Achilles und der Schildkröte). Das folgende Lemma zeigt, dass die Zeit S in obigem Algorithmus mit Wahrscheinlichkeit 1 gegen unendlich konvergiert.

Lemma 2. *Für unabhängige Zufallsvariablen $T_k, k = 1, 2, \ldots$, welche exponentialverteilt sind mit Parameter λk konvergiert $\sum_{k=1}^{n} T_k$ mit Wahrscheinlichkeit 1 gegen unendlich.*

Beweis: Der Erwartungswert von T_k ist gleich $\frac{1}{\lambda k}$ und die Varianz von T_k ist gleich $\frac{1}{\lambda^2 k^2}$. Der Erwartungswert μ_n von $\sum_{k=1}^{n} T_k$ divergiert somit, während die Varianz σ_n^2 konvergiert. Aus der Chebyshev-Ungleichung folgt daher für beliebiges festes s und alle n mit $\mu_n \geq s$

$$P(\sum_{k=1}^{n} T_k < s) \leq P(|\sum_{k=1}^{n} T_k - \mu_n| > |\mu_n - s|) \leq \frac{\sigma_n^2}{(\mu_n - s)^2} \to 0.$$

Aus der Stetigkeit der Wahrscheinlichkeit für absteigende Folgen von Ereignissen folgt damit $P(\sum_{k=1}^{n} T_k < s \text{ für alle } n) = 0$. □

2.2 Der Geburts- und Todesprozess

Wir betrachten als nächstes eine Population, in der sowohl Geburten als auch Todesfälle, auftreten. Es sind dann also Sprünge der Höhe ± 1 möglich. Die Gleichungen (2) und (3) werden dann ersetzt durch

$$\begin{aligned} P(X(t+h) = n+1 \mid X(t) = n) &\approx \nu \cdot n \cdot h, & (5)\\ P(X(t+h) = n-1 \mid X(t) = n) &\approx \mu \cdot n \cdot h, & (6)\\ P(X(t+h) = n \mid X(t) = n) &\approx 1 - (\nu + \mu) \cdot n \cdot h. & (7) \end{aligned}$$

Der Parameterν ist also die Geburtenrate und μ die Sterberate. (7) ist äquivalent dazu, dass die Wahrscheinlichkeit für mehr als einen Sprung in einem Intervall $(t, t+h]$ von kleinerer Ordnung ist als h. Analog zu (4) verschärfen wir dies und verlangen, dass für beliebiges $h > 0$ und beliebiges t gelten soll

$$P(\text{Keine Veränderung in } (t, t+h] \mid X(t) = n) = \exp(-h(\nu + \mu)n). \quad (8)$$

Wenn $X(t) = n$ ist, dann hat die Wartezeiten auf den nächsten Sprung wieder eine Exponentialverteilung, und zwar mit Parameter $(\nu + \mu)n$. Für $n = 0$ bleibt $X(s) = 0$ für alle $s > t$, d.h. die Population ist ausgestorben. Die verschiedenen Wartezeiten sind wegen der Markoveigenschaft wieder unabhängig. Für jeden Sprung muss noch entschieden werden, ob er nach oben oder nach unten gehen soll. Wenn man die Wahrscheinlichkeiten dafür wie folgt festlegt

$$P(\text{Geburt}) = \frac{\nu}{\nu + \mu}, \quad P(\text{Tod}) = \frac{\mu}{\nu + \mu},$$

dann folgt für kleine Zeitschritte

$$\begin{aligned} P(X(t+h) = n+1 \mid X(t) = n) &\approx (1 - \exp(-h(\nu + \mu)n))\frac{\nu}{\nu + \mu} \approx \nu \cdot n \cdot h\\ P(X(t+h) = n-1 \mid X(t) = n) &\approx (1 - \exp(-h(\nu + \mu)n))\frac{\mu}{\nu + \mu} \approx \mu \cdot n \cdot h \end{aligned}$$

d. h. (5) und (6) sind erfüllt. Es wird also eine Münze mit Erfolgsparameter $\frac{\nu}{\nu+\mu}$ geworfen, was man mit einer zweiten auf $(0, 1)$ gleichverteilten Zufallsvariablen bewerkstelligen kann.

Mit folgendem Algorithmus kann man Sprungfunktionen gemäss diesem Modell erzeugen:

1. Wähle den Anfangswert $X(0)$ und setze $S = 0$.

2. Für $k = 1, 2, \ldots$ wähle unabhängige auf $(0, 1)$ gleichverteilte Zufallsvariablen U_k und V_k und setze

   - $T_k = -\log(U_k)/((\nu + \mu)X(S))$,
   - $X(t) = X(S)$ für $S < t < S + T_k$,
   - $X(S + T_k) = X(S) + 1$, falls $V_k \leq \frac{\nu}{\nu+\mu}$, $X(S + T_k) = X(S) - 1$ sonst,
   - $S = S + T_k$.

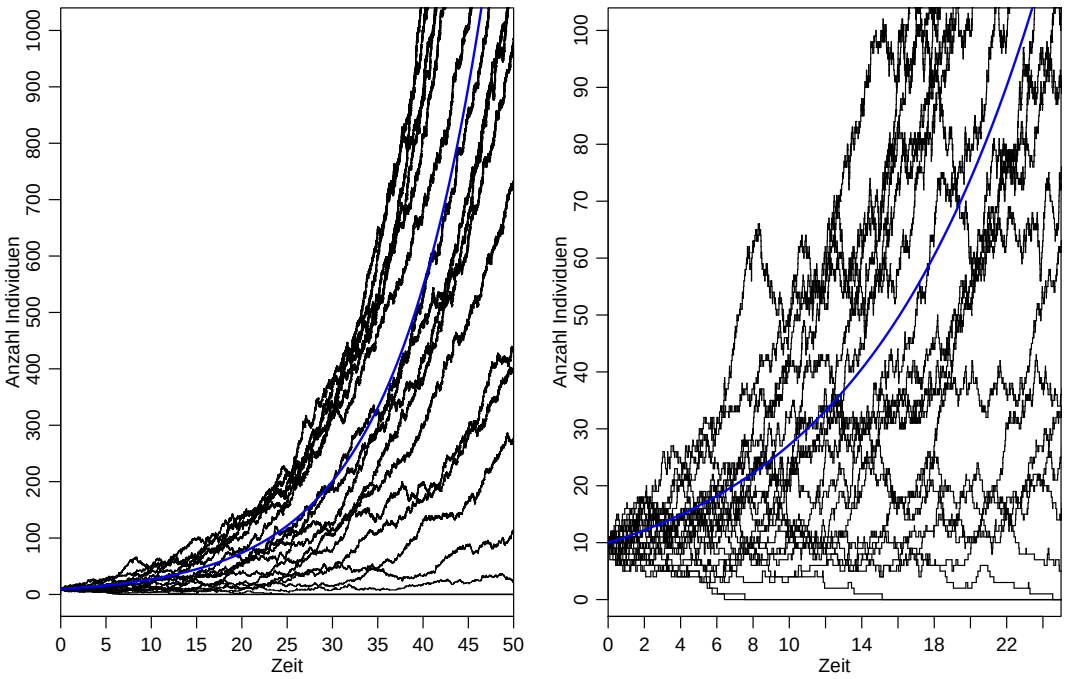

Abbildung 2: 20 simulierte Verläufe des Geburts- und Todesprozesses mit $\nu = 0.55$, $\mu = 0.45$ und $X(0) = 10$. Die Figur rechts ist ein Zoom der Figur links, um das Verhalten zu Beginn deutlicher zu machen.

Mit dem gleichen Argument wie zuvor folgt aus der Approximation (5) – (7) für den Erwartungswert die Euler-Approximation des deterministischen Modells (1):

$$E(X(t + h)) - E(X(t)) \approx (\nu - \mu)E(X(t))h.$$

Das deterministische und das stochastische Modell unterscheiden sich jetzt jedoch grundlegend, wie die Abbildungen 2 und 3 zeigen. Wenn $\nu > \mu$, dann kann die Population im stochastischen Modell entweder exponentiell wachsen oder aussterben. In Abb. 2, wo $\nu > \mu$ ist, sind im Zeitraum $[0, 25]$ vier der zwanzig Realisierungen ausgestorben, bei einer ist der weitere Verlauf noch unsicher, während die restlichen 15 mit an Sicherheit grenzender Wahrscheinlichkeit exponentiell wachsen. In Abb. 3, wo $\nu < \mu$ ist, sind 19 von 20 Realisierungen nach einer Zeit von 50 ausgestorben, und bei der letzten wird das vermutlich auch bald geschehen.

Das folgende Lemma liefert einen mathematischen Beweis, dass diese empirischen Befunde tatsächlich das Verhalten des stochastischen Modells darstellen.

Lemma 3. *Wenn $\nu \leq \mu$, dann stirbt die Population für beliebigen Startwert mit Wahrscheinlichkeit 1 in endlicher Zeit aus. Für $\nu = \mu$ hat die Zeit bis zum Aussterben unendlichen Erwartungswert. Für $\nu > \mu$ und beliebigen Startwert stirbt die Population entweder in endlicher Zeit aus oder sie konvergiert gegen unendlich. Beide Möglichkeiten haben strikt positive Wahrscheinlichkeit.*

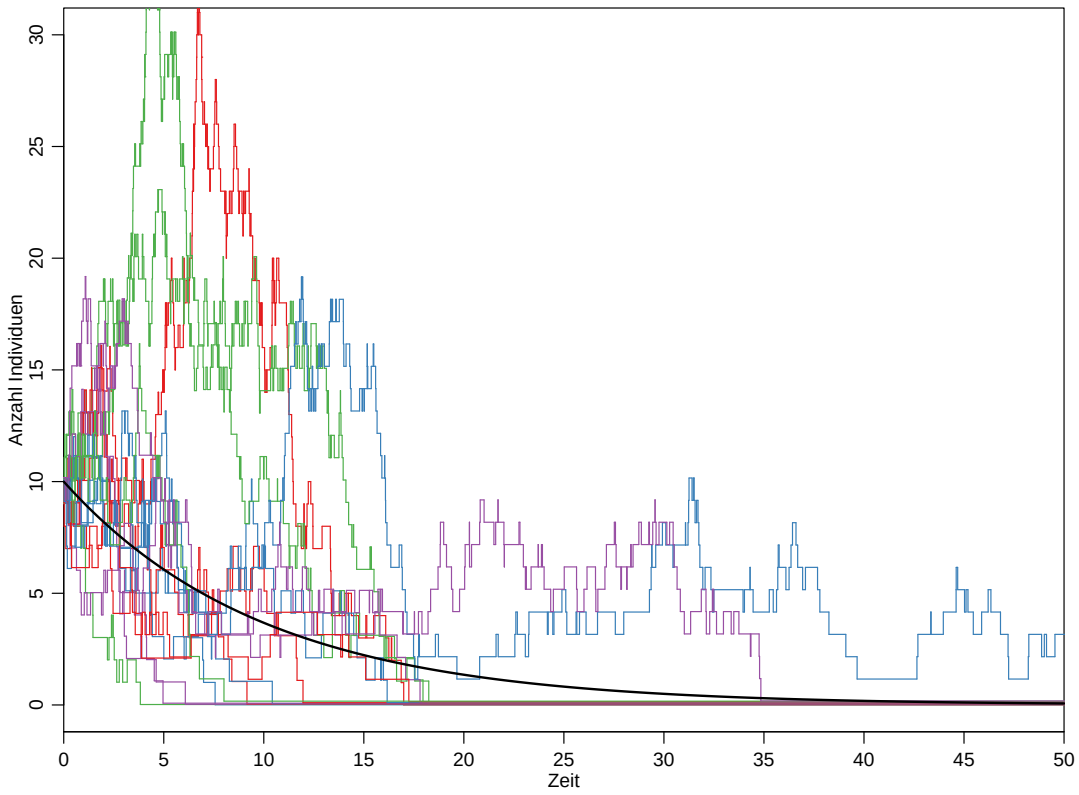

Abbildung 3: 20 simulierte Verläufe des Geburts- und Todesprozesses mit $\nu = 0.45$, $\mu = 0.55$ und $X(0) = 10$. Um die einzelnen Funktionen besser auseinanderhalten zu können, wurden Farben verwendet und die Graphen leicht vertikal versetzt.

Beweis: Die Folge der Werte von $X(t)$ an den Zeitpunkten S_n, an denen ein Sprung erfolgt, bildet eine sogenannte Irrfahrt mit Absorption in Null: Solange $S_n \neq 0$, gilt $X(S_n) = X(0) + \sum_{k=1}^{n} Z_k$, wobei die Z_k unabhängig sind mit $P(Z_k = 1) = \frac{\nu}{\nu+\mu}$, $P(Z_k = -1) = \frac{\mu}{\nu+\mu}$. Das Gesetz der grossen Zahlen besagt, dass $\frac{1}{n}\sum_{k=1}^{n} Z_k$ mit Wahrscheinlichkeit 1 gegen $\frac{\nu-\mu}{\nu+\mu}$ konvergiert. Daraus folgt sofort, dass die Population mit Wahrscheinlichkeit 1 ausstirbt, wenn $\nu < \mu$. Wenn $\nu > \mu$, dann ist die Wahrscheinlichkeit auszusterben grösser als $\left(\frac{\mu}{\nu+\mu}\right)^{X(0)} > 0$, und wenn sie nicht ausstirbt, dann ist $X(S_n) = X(0) + \sum_{k=1}^{n} Z_k$ für alle n, also geht sie wegen des Gesetzes der grossen Zahlen gegen unendlich. Für $\nu = \mu$ ist die Irrfahrt symmetrisch, und mit Wahrscheinlichkeit 1 erreicht diese Irrfahrt für beliebigen Startpunkt irgendwann den Ursprung, allerdings hat die Wartezeit darauf unendlichen Erwartungswert, siehe z.B. den Abschnitt 1.6 in Resnick (1992) oder den Abschnitt 2.3.2 in Föllmer and Künsch (2013). □

2.3 Der Geburts- und Todesprozess mit Immigration

Um ein Modell zu erhalten, beim dem die Population nicht nur aussterben oder unbegrenzt wachsen kann, führe ich beim Geburts- und Todesprozess mit $\mu > \nu$ noch Immigration mit konstanter Rate γ ein, welche das Aussterben verhindert. Das deterministische Modell lautet dann

$$\dot{X}(t) = \gamma - (\mu - \nu)X(t). \tag{9}$$

Die Lösung von (9) ist

$$X(t) = (X(0) - \frac{\gamma}{\mu - \nu})e^{-(\mu-\nu)t} + \frac{\gamma}{\mu - \nu}.$$

Die Population konvergiert somit exponentiell gegen den Fixpunkt $\frac{\gamma}{\mu-\nu}$.

Das stochastische Modell lautet

$$\begin{aligned} P(X(t+h) = n+1 \mid X(t) = n) &\approx (\gamma + \nu \cdot n) \cdot h, && (10)\\ P(X(t+h) = n-1 \mid X(t) = n) &\approx \mu \cdot n \cdot h, && (11)\\ P(X(t+h) = n \mid X(t) = n) &\approx 1 - (\gamma + (\nu+\mu) \cdot n) \cdot h. && (12) \end{aligned}$$

Es kann mit dem folgenden Algorithmus simuliert werden

1. Wähle den Anfangswert $X(0)$ und setze $S = 0$.
2. Für $k = 1, 2, \ldots$ wähle unabhängige auf $(0, 1)$ gleichverteilte Zufallsvariablen U_k und V_k und setze
   - $T_k = -\log(U_k)/(\gamma + (\nu+\mu)X(S))$,
   - $X(t) = X(S)$ für $S < t < S + T_k$,
   - $X(S + T_k) = X(S) + 1$, falls $V_k \leq \frac{\gamma+\nu n}{\gamma+(\nu+\mu)n}$, $X(S + T_k) = X(S) - 1$ sonst,
   - $S = S + T_k$.

In Abb. 4 sind fünf mit diesem Algorithmus simulierte Sprungfunktionen für $\nu = 0.45$, $\mu = 0.55$ und $\gamma = 1$ gezeichnet. Man sieht, dass im Unterschied zum deterministischen Modell die Grösse der Populationen nicht zu einem festen Wert tendiert, sondern stets fluktuiert. Es scheint sich aber ein Gleichgewicht in dem Sinne einzustellen, dass sehr bald keine Trends oder andere systematische Veränderungen (z.B. in der Grösse der Fluktuationen) mehr sichtbar sind. Dies wird noch deutlicher sichtbar in Abb. 5, wo der Anteil der Zeit zwischen 0 und 1000, während der die Population eine feste Grösse n hat, für verschiedene Werte von n dargestellt ist. Die drei schwarzen Stäbe gehören zu je einer Simulation mit den Startwerten $X(0) = 0, 10, 20$. Man sieht, dass diese Anteile trotz unterschiedlicher Startwerte ähnlich sind.

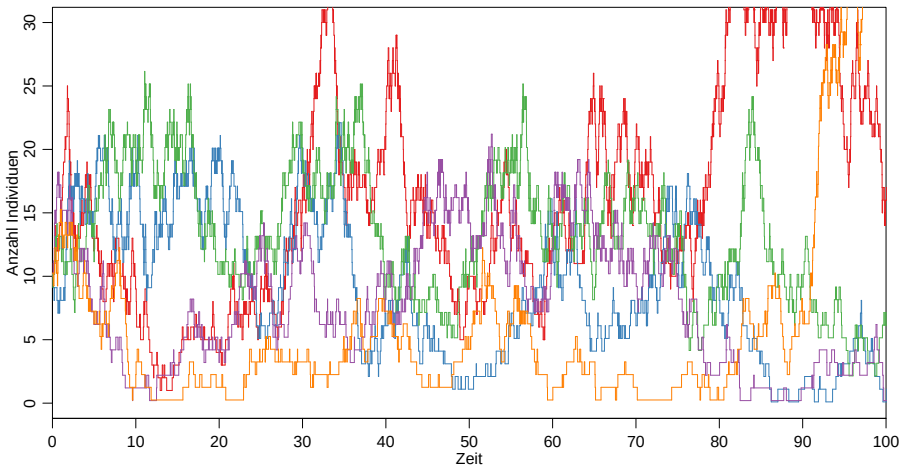

Abbildung 4: 5 simulierte Verläufe des Geburts- und Todesprozesses mit Immigration für $\nu = 0.55$, $\mu = 0.45$, $\gamma = 1$ und $X(0) = 10$. Um die einzelnen Funktionen besser auseinanderhalten zu können, wurden Farben verwendet und die Graphen leicht vertikal versetzt.

Dieser empirische Befund kann auch mathematisch streng formuliert und bewiesen werden. Wenn $\nu < \mu$, dann gilt für beliebigen Startwert $X(0) = k$ und beliebiges n

$$\lim_{t\to\infty} P(X(t) = n \mid X(0) = k) = \lim_{t\to\infty} \frac{1}{t} \int_0^t 1_n(X(s))ds = \frac{\Gamma(\frac{\gamma}{\nu} + n)}{n!\,\Gamma(\frac{\gamma}{\nu})} \left(\frac{\nu}{\mu}\right)^n \left(1 - \frac{\nu}{\mu}\right)^{\gamma/\nu}. \quad (13)$$

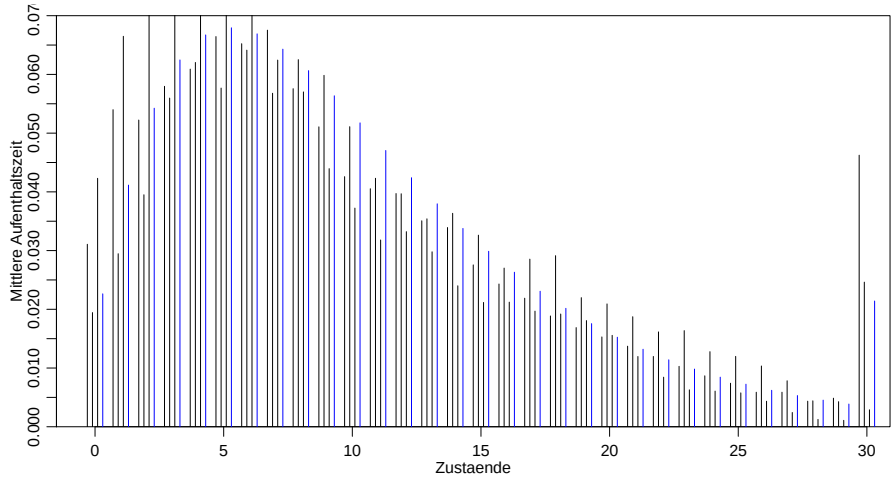

Abbildung 5: Die Anteile der Zeit zwischen 0 und 1000, während der sich 3 Realisierungen des Prozesses mit den Startwerten $X(0) = 0$ (linke schwarze Stäbe), bzw. $X(0) = 10$ (mittlere schwarze Stäbe), bzw. $X(0) = 20$ (rechte schwarze Stäbe) in einem Zustand n, bzw. im Intervall $[30, \infty)$, aufgehalten haben. Die blauen Stäbe zeigen die Grenzverteilung (13) dieser Aufenthaltszeiten.

Dabei ist $1_n(.)$ die Indikatorfunktion für den Wert n, d.h. $I_n(x) = 1$ falls $x = n$ und $I_n(x) = 0$ sonst, und Γ ist die Gamma-Funktion $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$. Die blauen Stäbe in Abb. 5 zeigen die Grenzverteilung. Der Abschnitt 4.3 unten enthält etwas mehr intuitive Erklärung, worauf diese Resultate beruhen.

3 Das SIR-Modell in der Epidemiologie

3.1 Definition und Simulation des Modells

Ich diskutiere hier das sogenannte SIR-Modell (susceptible – infected – recovered), siehe z.B. Allen (2011). In der einfachsten Form wird dabei eine geschlossene Population der Grösse N modelliert. Die Anzahl $X(t)$ der infizierbaren (susceptible) und die Anzahl $Y(t)$ der infizierten (infected) Individuen genügen in der deterministischen Version der Differentialgleichung

$$\dot{X}(t) = -\alpha X(t)Y(t), \quad \dot{Y}(t) = \alpha X(t)Y(t) - \mu Y(t). \tag{14}$$

Die Anzahl der Immunen (recovered) berechnet sich dann als $N - X(t) - Y(t)$. In obiger Differentialgleichung ist $\alpha X(t)Y(t)$ die Änderungsrate durch eine Neuinfizierung, welche proportional ist zur Anzahl möglicher Kontakte zwischen Infizierbaren und Infizierten, und $\mu Y(t)$ ist die Änderungsrate durch Gesundung eines Infizierten, der nachher immun ist. Da die Immunen für die Dynamik keine Rolle spielen, kann man ohne Beschränkung der Allgemeinheit $Y(0) + X(0) = N$ annehmen.

Im stochastischen SIR-Modell sind zwei Typen von Sprüngen möglich: Entweder $(X(t), Y(t)) \to (X(t) - 1, Y(t) + 1)$, entsprechend einer Ansteckung, oder $(X(t), Y(t)) \to (X(t), Y(t) - 1)$, entsprechend einer Gesundung. Analog zum deterministischen Modell werden die folgenden Wahrscheinlichkeiten für diese zwei möglichen Veränderungen postuliert:

$$P(X(t+h) = n-1, Y(t+h) = m+1 \mid X(t) = n, Y(t) = m) \approx \alpha \cdot n \cdot m \cdot h, \tag{15}$$

$$P(X(t+h) = n, Y(t+h) = m-1 \mid X(t) = n, Y(t) = m) \approx \mu \cdot m \cdot h, \tag{16}$$

$$P(X(t+h) = n, Y(t+h) = m \mid X(t) = n, Y(t) = m) \approx 1 - \lambda(n, m) \cdot h, \tag{17}$$

wobei wir zur Vereinfachung der Notation setzen

$$\lambda(n,m) = \alpha nm + \mu m.$$

Wir ersetzen wieder (17) durch die Forderung, dass für jedes $h > 0$ und jedes t gelten soll:

$$P(\text{Keine Veränderung in } (t, t+h] \mid X(t) = n, Y(t) = m) = \exp(-h\lambda(n,m)). \tag{18}$$

Wenn $X(t) = n$ und $Y(t) = m$, dann ist die Zeit bis zur nächsten Veränderung exponentialveteilt mit Parameter $\lambda(n,m)$. Nachdem der Zeitpunkt der nächsten Veränderung bestimmt ist, muss noch entschieden werden, welcher der beiden möglichen Sprünge an diesem Zeitpunkt auftreten soll. Die Wahrscheinlichkeiten dafür sind die folgenden

$$P(\text{Ansteckung}) = \frac{\alpha nm}{\lambda(n,m)}, \quad P(\text{Genesung}) = 1 - P(\text{Ansteckung}) = \frac{\mu m}{\lambda(n,m)}.$$

Wie zuvor gelten dann (15) und (16).

Der Algorithmus zur Simulation des stochastischen SIR-Modells lautet

1. Wähle die Anfangswerte $X(0)$ und $Y(0)$ und setze $S = 0$.
2. Für $k = 1, 2, \ldots$ wähle unabhängige, auf $(0,1)$ gleichverteilte Zufallsvariablen U_k und V_k und setze
   - $T_k = -\log(U_k)/\lambda(X(S), Y(S))$, $X(t) = X(S)$, $Y(t) = Y(S)$ für $S < t < S + T_k$,
   - $X(S+T_k) = X(S) - 1, Y(S+T_k) = Y(S) + 1$, falls $V_k \leq \alpha X(S)Y(S)/\lambda(X(S), Y(S))$,
   - $X(S + T_k) = X(S), Y(S + T_k) = Y(S) - 1$ sonst,
   - $S = S + T_k$.

3.2 Vergleich des deterministischen und des stochastischen Modells

Im deterministischen Modell nimmt die Anzahl der Infizierbaren $X(t)$ monoton ab mit der Zeit, weil die Ableitung stets negativ ist. Das Vorzeichen von $\dot{Y}(t)$ hängt davon ab, ob $X(t)$ grösser oder kleiner ist als μ/α. Wenn $X(0) \leq \mu/\alpha$, nimmt $Y(t)$ ebenfalls monoton ab mit der Zeit. Wenn hingegen $X(0) > \mu/\alpha$, dann wächst $Y(t)$ so lange, bis $X(t) = \mu/\alpha$, und nimmt danach monoton ab. Die Epidemie breitet sich also zunächst aus und geht dann zurück, wenn die Anzahl von Infizierbaren genügend klein geworden ist. Weil

$$\dot{Y}(t) + \left(1 - \frac{\mu}{\alpha X(t)}\right) \dot{X}(t) = 0,$$

bleibt im deterministischen Modell die Grösse

$$Y(t) + X(t) - \frac{\mu}{\alpha} \log X(t)$$

in der Zeit unverändert. Wegen $Y(0) + X(0) = N$ gilt somit

$$Y(t) = N - X(t) + \frac{\mu}{\alpha}(\log X(t) - \log X(0)).$$

Weil $x \mapsto \frac{\mu}{\alpha}\log(x) - x$ maximal ist für $x = \frac{\mu}{\alpha}$ folgt in Übereinstimmung mit obigem, dass $\max_t Y(t) = Y(0) = N - X(0)$ wenn $X(0) \leq \frac{\mu}{\alpha}$, während für $X(0) > \frac{\mu}{\alpha}$

$$\max_t Y(t) = N - \frac{\mu}{\alpha}\left(1 + \log X(0) - \log(\mu/\alpha)\right).$$

Im deterministischen Modell dauert die Epidemie unendlich lange: $Y(t)$ konvergiert zwar gegen null, ist jedoch für alle endlichen Zeiten strikt positiv. Der Limes $\lim_{t\to\infty} X(t)$ ist gleich der kleineren der beiden Lösungen von $x - \frac{\mu}{\alpha}\log x = N - \frac{\mu}{\alpha}\log X(0)$ und somit stets grösser als 0. Es wird also auch im Limes $t \to \infty$ nur ein Teil der Population infiziert.

Das stochastische Modell verhält sich in mancher Hinsicht anders: Die Epidemie endet nach einer zufälligen, aber endlichen Zeit, nämlich sobald $Y(t) = 0$. $X(t)$ ist zwar immer noch monoton abnehmend, aber nicht mehr strikt, und bei $Y(t)$ wechseln sich Sprungphasen nach oben und nach unten im Allgemeinen mehrmals ab. Die Wahrscheinlichkeit, dass $Y(t)$ über die ganze Zeit monoton abnehmend ist (und somit niemand angesteckt wird), ist jedoch für beliebige Parameter und beliebige Startwerte grösser als Null. Wenn $Y(0) = 1$, dann ist die Wahrscheinlichkeit, dass sich diese Person erholt ohne jemanden anzustecken, gleich $\mu/(\mu + \alpha(N-1))$. Ebenso ist die Wahrscheinlichkeit, dass alle angesteckt werden, dass also $X(t) = 0$ eintritt, grösser als 0.

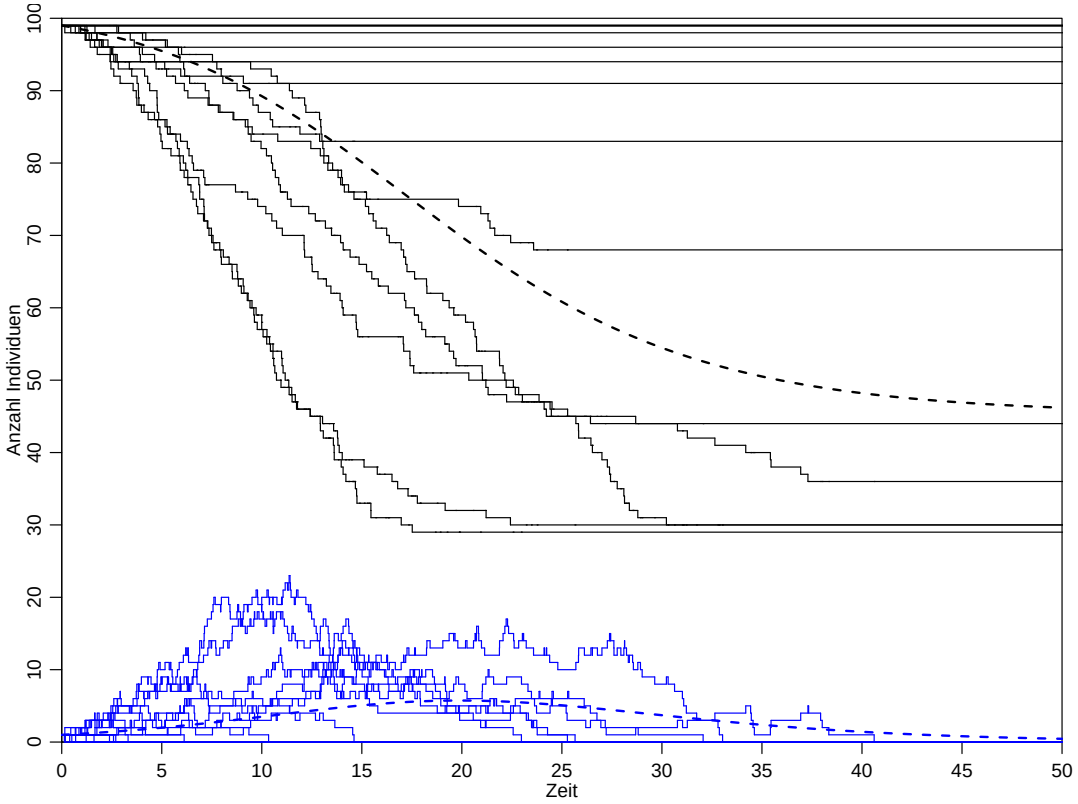

Abbildung 6: 20 simulierte Verläufe des SIR-Modells mit $\alpha = 0.005$, $\mu = 0.35$, $N = 100$ und $X(0) = 99, Y(0) = 1$. Schwarz: $X(t)$, blau: $Y(t)$, gestrichelt: Lösung des deterministischen Systems.

Die Abb. 6 zeigt zwanzig simulierte Verläufe des Modells mit $N = 100$, $\alpha = 0.005$, $\mu = 0.35$, und dem Startwert $Y(0) = 1$, $X(0) = 99 > \mu/\alpha = 70$. Der maximale Wert von $Y(t)$ im deterministischen Modell ist gleich 5.74 und $X(t)$ konvergiert für $t \to \infty$ gegen 45.35, d.h. etwas mehr als die Hälfte wird angesteckt. Die Dauer der Epidemie in den zwanzig Realisierungen des stochastischen Modells variiert von fast 0 bis etwas mehr als 40, das Maximum von $Y(t)$ variiert von 1 bis 23 und der Wert von $X(t)$ am Ende der Epidemie von 99 bis 29. Das stochastische Modell verhält sich also sehr variabel und unterscheidet sich stark vom deterministischen Modell.

3.3 Verhalten in grossen Populationen

In welchem Sinn ist das deterministische Modell eine Approximation des stochastischen? Man könnte vermuten, dass wie beim Wachstumsmodell der Erwartungswert $(E(X(t)), E(Y(t)))$ dem

deterministischen Modell genügt, doch das ist nicht der Fall. Es gilt zwar immer noch, dass

$$\begin{aligned} E(X(t+h) - X(t) \mid X(t), Y(t)) &\approx -\alpha \cdot X(t) \cdot Y(t) \cdot h, \\ E(Y(t+h) - Y(t) \mid X(t), Y(t)) &\approx (\alpha \cdot X(t) \cdot Y(t) - \mu \cdot Y(t)) \cdot h. \end{aligned}$$

Weil $E(X(t)Y(t)) \neq E(X(t))E(Y(t))$, ergibt die Bildung des Erwartungswerts auf beiden Seiten dieser Gleichungen jedoch nicht die Differentialgleichung (14).

Das deterministische System ist eigentlich gedacht als Modell für die Anteile

$$x(t) = \frac{X(t)}{N}, \quad y(t) = \frac{Y(t)}{N}$$

in grossen Populationen, weil dann die Sprünge klein sind und im Limes $N \to \infty$ sogar verschwinden. Dividiert man die Gleichung (14) durch N, so erhält man die Differentialgleichung

$$\dot{x}(t) = -\alpha' x(t) y(t), \quad \dot{y}(t) = \alpha' x(t) y(t) - \mu y(t) \tag{19}$$

wobei $\alpha' = N\alpha$. Dass der Parameter α in (14) umgekehrt proportional zu N sein sollte, sieht man wie folgt: Die Anzahl näherer Kontakte einer infizierten Person mit anderen Personen in einem kurzen Intervall h ist ungefähr proportional zu h, und ein Anteil $x(t)$ dieser Kontakte ist mit infizierbaren Personen, die dann infiziert werden. Also ist die Anzahl Neuansteckungen durch alle Infizierten in einem kurzen Intervall h ungefähr proportional zu $hY(t)x(t) = h\frac{1}{N}Y(t)X(t)$.

In grossen Populationen ist es plausibel, dass für das stochastische Modell ein Gesetz der grossen Zahlen gilt. Das heisst, die Anteile $x(t)$ und $y(t)$ für $N \to \infty$ konvergieren gleichmässig auf jedem Intervall $[0, u]$ gegen einen Grenzwert, falls dies für die Anfangsbedingungen $x(0)$ und $y(0)$ gilt. Weiter ist die Dynamik des Grenzwertes nicht mehr zufällig, sondern durch die Differentialgleichung (19) gegeben. Dies wird in Theorem 8.1 in Kurtz (1981) gezeigt. Die Herleitung davon sprengt jedoch den Rahmen der Mittelschulmathematik.

4 Allgemeine Markovketten in stetiger Zeit

Ich hoffe, dass es klar ist, wie man die hier besprochenen Modelle verallgemeinern kann, z.B. für alters- oder räumlich strukturierte Populationen, für Räuber- und Beute-Modelle, oder für andere Formen der Änderungsraten durch Geburten oder Tod. Ich gebe im Folgenden noch eine kurze Einführung in die Theorie der Markovketten und deren asymptotisches Verhalten. Insbesondere will ich die Konvergenz gegen eine Grenzverteilung besprechen, die wir im Beispiel des Geburts- und Todesprozesses mit Immigration angetroffen haben.

4.1 Infinitesimale Übergänge

Markovketten beschreiben eine stochastische Dynamik eines Systems mit einem endlichen oder abzählbaren Zustandsraum E. Elemente von E bezeichnen wir im Folgenden mit x, y, z. Die Dynamik wird festgelegt durch die Angabe der sogenannten Übergangsraten $\alpha(x, z)$ für $x, z \in E$, $x \neq z$. Dies bedeutet, dass für $h \to 0$:

$$P(X(t+h) = z \mid X(t) = x) \approx \alpha(x, z) \cdot h \tag{20}$$

$$P(X(t+h) = x \mid X(t) = x) \approx 1 - \lambda(x) \cdot h, \tag{21}$$

wobei verlangt wird, dass für alle x

$$\lambda(x) = \sum_{z \neq x} \alpha(x, z) < \infty.$$

Wenn $X(t) = x$, dann ist die Wartezeit T auf den nächsten Sprung exponentialverteilt mit Parameter $\lambda(x)$. Wenn T bestimmt ist, dann wird der nächste Zustand mit den folgenden Wahrscheinlichkeit ausgewürfelt:

$$P(X(t+T) = z \mid X(t) = x) = \frac{\alpha(x,z)}{\lambda(x)}.$$

Damit lässt sich die Markovkette mit einem analogen Algorithmus wie zuvor simulieren.

4.2 Die Berechnung der Übergangswahrscheinlichkeiten

Die bedingten Wahrscheinlichkeiten

$$P(X(t+h) = z \mid X(t) = x)$$

werden die Übergangswahrscheinlichkeiten des Modells genannt und mit $P_h(x,z)$ bezeichnet. Wenn wir die Zustände durchnummerieren, können wir P_h als endliche, bzw. unendliche Matrix auffassen. Die Elemente von P_h sind stets nicht-negativ und alle Zeilensummen sind gleich 1.

Die Zeile x von P_t stellt die Verteilung von $X(t)$ bei Start in x dar. Wenn der Startzustand nicht fest ist, sondern gemäss einer Verteilung $p = (p(x))$ gewählt wird, dann erhält man mit dem Satz der totalen Wahrscheinlichkeit

$$P(X(t) = z) = \sum_{x \in E} P(X(t) = z \mid X(0) = x) \cdot P(X(0) = x) = \sum_{x \in E} P(x,z)p(x). \tag{22}$$

In Matrixform ist die Verteilung von $X(t)$ somit gleich $P_t^T p$.

Offensichtlich ist $P_0(x,z) = 1$, falls $x = z$, und $P_0(x,z) = 0$ sonst, d.h. P_0 ist die Einheitsmatrix I. Die Bedingungen (20) und (21) legen die Ableitungen von $P_h(x,z)$ nach h an der Stelle 0 fest:

$$\left.\frac{d}{dh}P_h(x,z)\right|_{h=0} = A(x,z) = \begin{cases} -\lambda(x) & (z = x) \\ \alpha(x,z) & (z \neq x) \end{cases} \tag{23}$$

Aus der Markoveigenschaft folgt sogar eine Differentialgleichung für die Übergangswahrscheinlichkeiten. Wenn wir wissen wollen, wie wahrscheinlich die verschiedenen Übergänge im Zeitraum $(0, s+t]$ sind, können wir die Übergänge in den beiden Intervallen $(0, s]$ und $(s, s+t]$ betrachten. Für das zweite Intervall ist nur der Zustand zum Zeitpunkt s relevant. Daher können wir alle möglichen Zustände zum Zeitpunkt s betrachten und die entsprechenden Wahrscheinlichkeiten aufsummieren:

$$P(X(s+t) = z \mid X(0) = x) = \sum_{y \in E} P(X(s) = y \mid X(0) = x) P(X(s+t) = z \mid X(s) = y),$$

bzw. kürzer geschrieben

$$P_{s+t}(x,z) = \sum_y P_s(x,y) P_t(y,z).$$

Dies ist die sogenannte Chapman-Kolmogorov-Gleichung. Fasst man P_t als Matrix auf, dann lautet die Chapman-Kolmogorov-Gleichung $P_{s+t} = P_s P_t$.

Aus der Chapman-Kolmogorov-Gleichung folgt

$$\frac{P_{t+h} - P_t}{h} = \frac{P_h P_t - P_t}{h} = \frac{P_h - I}{h} P_t.$$

Wenn E endlich ist, dann folgt zusammen mit (23) sofort, dass

$$\frac{d}{dt}P_t = AP_t,$$

bzw. ausgeschrieben

$$\frac{d}{dt}P_t(x,z) = \sum_y A(x,y)P_t(y,z).$$

Dies ist die sogenannte Rückwärtsgleichung von Kolmogorov, ein lineares Differentialgleichungssystem mit konstanten Koeffizienten. Weil P_0 die Einheitsmatrix I ist, lässt sich die Lösung mit Hilfe der Matrixexponentialfunktion schreiben als

$$P_t = \exp(tA) = \sum_{k=0}^{\infty} \frac{t^k A^k}{k!}.$$

Wenn E abzählbar unendlich ist, dann gilt die Rückwärtsgleichung immer noch, aber die Herleitung ist komplizierter, siehe z.B. Proposition 5.4.3 in Resnick (1992). Wenn E endlich ist, erfüllt P_t auch die Vorwärtsgleichung $\frac{d}{dt}P_t = P_t A$. Diese erhält man, direkt aus $P_t = \exp(tA)$, oder indem man einen Zeitschritt der Länge t gefolgt von einem Zeitschritt der Länge h betrachtet

$$\frac{P_{t+h} - P_t}{h} = \frac{P_t P_h - P_t}{h} = P_t \frac{P_h - I}{h},$$

und dann h gegen Null gehen lässt. Wenn E abzählbar unendlich ist, dann kann der Fall eintreten, dass die Rückwärtsgleichung mehr als eine Lösung hat und die Vorwärtsgleichung nicht für alle diese Lösungen erfüllt ist.

Als Beispiel betrachten wir den reinen Geburtsprozess. Die Rückwärtsgleichung lautet

$$\frac{d}{dt}P_t(k,n) = -k\lambda P_t(k,n) + k\lambda P_t(k+1,n)$$

und die Vorwärtsgleichung lautet

$$\frac{d}{dt}P_t(k,n) = -n\lambda P_t(k,n) + (n-1)\lambda P_t(k,n-1).$$

Weil $P_t(k,n) \equiv 0$ ist für $k > n$ und $P_0(n,n) = 1$, folgt in Übereinstimmung mit (4), dass $P_t(n,n) = \exp(-n\lambda t)$. Die Funktionen $P_t(k,n)$ für $k < n$ kann man dann rekursiv aus der Rückwärts- oder der Vorwärtsgleichung bestimmen: Der zweite Term auf der rechten Seite ist dann jeweils bereits bekannt und man hat eine inhomogene lineare Differentialgleichung mit konstanten Koeffizienten. Die Rechnung ist allerdings mühsam. Man erhält

$$P_t(k,n) = \binom{n-1}{k-1} \exp(-\lambda t)^k (1 - \exp(-\lambda t))^{n-k}.$$

Beim Startwert $X(0) = k$ hat der Zuwachs $X(t) - X(0)$ somit eine sogenannte Negativ-Binomialverteilung, die Verteilung der Anzahl Misserfolge bis zum k-ten Erfolg beim Münzwurf mit Erfolgsparameter $\exp(-\lambda t)$.

4.3 Asymptotisches Verhalten der Übergangswahrscheinlichkeiten

In den bisherigen Beispielen haben wir unterschiedliches Verhalten von Markovketten über lange Zeitspannen gesehen: Beim reinen Geburtsprozess verlässt die Kette asymptotisch jede endliche Menge von Zuständen. Beim SIR-Modell gibt es irgendwann einmal keine infizierten Individuen mehr und damit endet der Prozess in einem der Zustände $(x, 0)$. Diese Zustände heissen daher auch absorbierend. In welchem dieser Zustände der Prozess endet, ist zufällig, und die Wahrscheinlichkeiten dafür hängen vom Startwert ab. Der Geburts- und Todesprozess hat einen einzigen absorbierenden Zustand, nämlich 0. Für $\nu \leq \mu$ wird dieser Zustand sicher erreicht,

während für $\nu > \mu$ mit positiver Wahrscheinlichkeit nicht absorbiert wird, sondern jede endliche Menge von Zuständen verlässt. Der Geburts- und Todesprozess mit Immigration hat keine absorbierenden Zustände, und für $\mu \geq \nu$ stellt sich unabhängig vom Startwert ein stochastisches Gleichgewicht ein, bei dem sich der Prozess in jedem Zustand während eines bestimmten Zeitanteils aufhält.

In diesem Abschnitt will ich noch etwas ausführlicher diskutieren, wie man das Verhalten des Prozesses für grosse Zeiten bereits aus den Übergangsraten ableiten kann. Absorbierende Zustände kann man sofort erkennen: x ist genau dann absorbierend, wenn alle Raten $\alpha(x, z)$ gleich Null sind. Interessanter und schwieriger ist die Frage, wann es ein stochastisches Gleichgewicht wie beim Geburts- und Todesprozess mit Immigration für $\mu \geq \nu$ gibt.

Zunächst betrachte ich den Fall, wo E endlich ist.

Lemma 4. *Wenn E endlich ist und wenn $P_t(x, y)$ für alle $x, y, \in E$ gegen einen Grenzwert $\pi(y)$ konvergiert, der nicht vom Startwert x abhängt, dann ist π eine Wahrscheinlichkeitsverteilung. Wenn $P(X(0) = y) = \pi(y)$ ist, dann ist $P(X(t) = y) = \pi(y)$ für alle t, d.h. π ist eine stationäre Verteilung. Ferner ist 0 $A^T\pi = 0$, d.h. π ist ein Eigenvektor von A^T zum Eigenwert 0, die geometrische Vielfachheit von 0 ist gleich eins und π ist die einzige stationäre Verteilung.*

Beweis: Offensichtlich ist $\pi(y) \geq 0$, und wegen

$$\sum_{y\in E} \pi(y) = \sum_{y\in E} \lim_t P_t(x, y) = \lim_t \sum_{y\in E} P_t(x, y) = 1$$

ist π eine Wahrscheinlichkeit. Für die Stationarität müssen wir wegen (22) $P_t^T\pi = \pi$ zeigen:

$$\sum_{x\in E} \pi(x)P_t(x, y) = \sum_{x\in E} \lim_s P_s(z, x)P_t(x, y) = \lim_s \sum_{x\in E} P_s(z, x)P_t(x, y) = \lim_s P_{t+s}(z, y) = \pi(y).$$

Die Gleichung $A^T\pi = 0$ folgt sofort durch Ableiten von $P_t^T\pi = \pi$ nach t. Wenn q ein Eigenvektor zum Eigenwert 0 von A^T ist, dann ist q auch ein Eigenvektor zum Eigenwert 1 von P_t^T. Weil

$$q(y) = \sum_{x\in E} q(x)P_t(x, y) \stackrel{t\to\infty}{\longrightarrow} \sum_{x\in E} q(x)\pi(y),$$

ist q ein Vielfaches von π. Insbesondere ist π also die einzige stationäre Verteilung. □

Weil die Zeilensummen von A gleich Null sind, ist 0 immer ein Eigenwert von A und damit auch von A^T. Daher prüft man als erstes, ob es einen zugehörigen Eigenvektor mit positiven Elementen gibt. Anhand der anderen Eigenwerte von A kann man meist feststellen, ob die Übergangswahrscheinlichkeiten konvergieren.

Als einfaches Beispiel betrachten wir $E = \{1, 2\}$ mit $\alpha(1, 2) = \nu$ und $\alpha(2, 1) = \mu$. Dann ist

$$A = \begin{pmatrix} -\nu & \nu \\ \mu & -\mu \end{pmatrix},$$

und die Eigenwerte von A sind 0 und $-(\nu + \mu)$ mit den Eigenvektoren $(1, 1)^T$, bzw. $(-\nu, \mu)^T$. Der Vektor $(\mu, \nu)^T$ ist ein Eigenvektor von A^T zum Eigenwert 0. Durch Diagonalisieren von A erhält man

$$P_t = \begin{pmatrix} 1 & -\nu \\ 1 & \mu \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & \exp(-(\mu+\nu)t) \end{pmatrix} \cdot \frac{1}{\nu+\mu} \begin{pmatrix} \mu & \nu \\ -1 & 1 \end{pmatrix} \to \frac{1}{\nu+\mu} \begin{pmatrix} \mu & \nu \\ -1 & 1 \end{pmatrix}.$$

Es gibt ein anschauliches Kriterium für die Konvergenz der Übergangswahrscheinlichkeiten gegen eine stationäre Verteilung, das ohne die Berechnung der Eigenwerte und ohne die Annahme der Diagonalisierbarkeit von A auskommt: Es muss möglich sein, von jedem Zustand x jeden anderen Zustand y zu erreichen, d.h. für beliebiges x und $y \neq x$ muss entweder $\alpha(x, y) > 0$ sein oder es muss Zustände $z_1, \ldots, z_k$ geben, so dass $\alpha(x, z_1) \cdots \alpha(z_k, y) > 0$.

Lemma 5. *Wenn E endlich ist und wenn man von jedem Zustand x jeden anderen Zustand y erreichen kann, dann sind die Übergangswahrscheinlichkeiten $P_t(x,y) > 0$ für alle $t > 0$ und alle $x, y \in E$. Ferner hat dann die Markovkette genau eine stationäre Verteilung und die Verteilung von X_t konvergiert für beliebigen Startwert gegen diese stationäre Verteilung:*

$$\lim_{t\to\infty} P_t(x,y) = \pi(y).$$

Beweis: Wenn $\alpha(x,y) > 0$, dann geht $X(t)$ in der Zeit t von x nach y, wenn im Intervall $(0,t)$ genau ein Sprung stattfindet und dieser von x nach y führt (es gibt natürlich noch andere Möglichkeiten mit mehr als einem Sprung). Die Wahrscheinlichkeit dafür ist grösser oder gleich

$$(1-\exp(-\lambda(x)t)\frac{\alpha(x,y)}{\lambda(x)}\exp(-\lambda(y)t) > 0.$$

Der erste Faktor in dieser Formel ist die Wahrscheinlichkeit, dass bei Start in x der erste Sprung vor der Zeit t stattfindet, der zweite Faktor ist die Wahrscheinlichkeit, dass dieser Sprung nach y führt, und der dritte Faktor die Wahrscheinlichkeit, dass der Prozess mindestens während der Zeit t in y bleibt.

Wenn $\alpha(x,z_1)\alpha(z_1,z_2)\cdots\alpha(z_{m-1},y) > 0$ ist, dann gilt wegen der Chapman-Kolmogorov-Gleichung

$$P_t(x,y) \geq P_{t/m}(x,z_1)P_{t/m}(z_1,z_2)\cdots P_{t/m}(z_{m-1},y) > 0.$$

Als nächstes fixieren wir ein $t > 0$ und wählen einen Eigenvektor π von P_t^T zum Eigenwert 1 (die Zeilensummen von P_t sind gleich 1, also ist 1 ein Eigenwert von P_t, und P_t^T hat das gleiche charakteristische Polynom wie P_t). Wir wollen zeigen, dass alle Elemente von π das gleiche Vorzeichen haben. Da alle $P_t(x,y) > 0$, wäre anderenfalls für alle x

$$|\pi(x)| = |\sum_{y\in E} P_t(y,x)\pi(y)| < \sum_{y\in E} P_t(y,x)|\pi(y)|.$$

Es gibt sogar ein $\varepsilon > 0$, so dass für alle x $|\pi(x)| < (1-\varepsilon)\sum_y P_t(y,x)|\pi(y)|$. Mit Hilfe der Chapman-Kolmogorov-Gleichung folgt dann durch Iteration

$$|\pi(x)| < (1-\varepsilon)^k \sum_{y\in E} P_{kt}(y,x)|\pi(y)|.$$

Weil alle $P_{kt}(y,x) < 1$, muss π der Nullvektor sein, was einen Widerspruch darstellt.

Wir nehmen im Folgenden an, dass π eine Wahrscheinlichkeitsverteilung ist, und schreiben P_t als eine konvexe Kombination von zwei Übergangsmatrizen M und Q:

$$P_t = \varepsilon M + (1-\varepsilon)Q, \quad M = \underline{1}\pi^T, \quad Q = \frac{1}{1-\varepsilon}(P_t - \varepsilon M).$$

Dabei ist $\underline{1} = (1,1,\ldots,1)^T$ und $\varepsilon > 0$ ist so gewählt, dass $P_t(x,y) > \varepsilon M(x,y) = \varepsilon\pi(y)$ für alle $x, y \in E$ (dies ist möglich, da alle $P_t(x,y) > 0$). Weil $\pi^T\underline{1} = \sum_x \pi(x) = 1$ und $\pi^T P_t = \pi^T$, gilt

$$M^2 = \underline{1}(\pi^T\underline{1})\pi^T = M, \quad MQ = \frac{1}{1-\varepsilon}(\underline{1}\pi^T P_t - \varepsilon M^2) = M.$$

Mit Induktion folgt daher, dass

$$P_{kt} = P_t^k = \sum_{j=1}^{k}\binom{k}{j}\varepsilon^j(1-\varepsilon)^{k-j}M + (1-\varepsilon)^k Q^k = (1-(1-\varepsilon)^k)M + (1-\varepsilon)^k Q^k.$$

Weil Q eine Übergangsmatrix ist, sind alle Elemente $Q^k(x,y) \in [0,1]$. Daher konvergiert $P_{kt}(x,y)$ gegen $\pi(y)$ für alle x, y wenn $k \to \infty$. Aus $|P_{kt}(x,y) - \pi(y)| < \varepsilon$ folgt zudem, dass

$$|P_{s+kt}(x,y) - \pi(y)| = |\sum_z P_s(x,z)(P_{kt}(z,y) - \pi(y))| \leq \sum_z P_s(x,z)|P_{kt}(z,y) - \pi(y)| \leq \varepsilon.$$

Deshalb konvergiert $P_t(x,y)$ gegen $\pi(y)$ für $t \to \infty$. Insbesondere hängt π nicht ab vom gewählten Wert t und erfüllt $P_t^T \pi = \pi$ für alle t. Durch Ableiten folgt schliesslich $A^T \pi = 0$ □

Wie das Beispiel einer Irrfahrt auf $\mathbb{Z}$ zeigt, folgt die Existenz einer Grenzverteilung im Fall, wo E unendlich ist, nicht daraus, dass man von jedem Zustand x jeden anderen Zustand y erreichen kann. Man muss zusätzlich voraussetzen, dass die Kette für beliebigen Startpunkt x mit Wahrscheinlichkeit 1 irgendwann nach x zurückkehrt. Diese Eigenschaft wird Rekurrenz genannt (wegen der Markoveigenschaft folgt daraus sogar, dass die Kette unendlich oft zum Startpunkt zurückkehrt). Für eine rekurrente Kette gilt nun das folgende Resultat, siehe z.B. Theorem 5.5.3 und die Korollare 5.5.4 und 5.5.5 von Resnick (1992): Die Gleichung $A^T \pi = 0$ hat eine nichttriviale und bis auf Multiplikation mit einem Faktor eindeutige Lösung mit $\pi(x) \geq 0$ für alle x. Diese Lösung erfüllt $P_t^T \pi = \pi$, d.h. eine stationäre Verteilung existiert genau dann, wenn $\sum_x \pi(x) < \infty$. Ferner konvergieren dann sowohl $P_t(x,y)$ als auch der Anteil der Zeit, während der $X(t)$ im Zustand y ist, gegen $\pi(y)$.

Für das Beispiel des Geburts- und Todesprozesses mit Immigration müssen wir also noch zeigen, dass die Kette rekurrent ist, und wir müssen die Lösung von $A^T \pi = 0$ berechnen. Dies erfolgt im Beweis der folgenden zwei Lemmata.

Lemma 6. *Der Geburts- und Todesprozesses mit Immigration ist für $\mu > \nu$ rekurrent.*

Beweis Es genügt zu zeigen, dass für beliebiges k

$$P(\liminf_t X(t) = 0, \limsup_t X(t) = \infty \mid X(0) = k) = 1.$$

Weil nur Sprünge von ± 1 möglich sind, wird nämlich jeder Wert zwischen dem lim inf und dem lim sup unendlich oft angenommen. Wenn $X(t) = n$ ist mit $n \geq 1$, dann erfolgt der nächste Sprung mit Wahrscheinlichkeit $\frac{n\mu}{\gamma + n(\mu+\nu)} > 0$ nach unten. Wenn $\liminf X(t) = n$, dann müssen alle bis auf endlich viele der Sprünge von n aus nach oben gehen, und die Wahrscheinlichkeit dafür ist wegen der Unabhängigkeit der Sprünge gleich 0. Das gleiche Argument zeigt, dass auch das Ereignis $\limsup_t X(t) = n$ Wahrscheinlichkeit 0 hat für alle n. Also bleibt nur noch zu zeigen, dass das Ereignis $\liminf X(t) = \infty$ Wahrscheinlichkeit 0 hat. Für n gross genug sind die Wahrscheinlichkeiten für einen Sprung nach unten von einem Zustand $> n$ grösser als ein festes $p > \frac{1}{2}$. Durch Vergleich mit der Irrfahrt, die mit Wahrscheinlichkeit p nach unten geht (siehe den Beweis von Lemma 3) sieht man, dass $X(t)$ vom Zustand $n+1$ mit Wahrscheinlichkeit 1 wieder den Zustand n erreicht. □

Lemma 7. *Die einzige Lösung von $A^T \pi = 0$ mit $\pi(0) = 1$ beim Geburts- und Todesprozesses mit Immigration ist*

$$\pi(n) = \left(\frac{\nu}{\mu}\right)^n \frac{\Gamma(\frac{\gamma}{\nu} + n)}{\Gamma(\frac{\gamma}{\nu})\ n!}.$$

Für $\nu < \mu$ ist $\sum_n \pi(n) = (1 - \frac{\nu}{\mu})^{-\gamma/\nu} < \infty$, und die stationäre Verteilung ist die sogenannte Negativ-Binomialverteilung (auch Pascal-Verteilung genannt) mit den Parametern $r = \gamma/\nu$ und $p = 1 - \nu/\mu$.

Bemerkung: Die Negativ-Binomialverteilung ist definiert für $r \in \mathbb{R}_+$ und $p \in (0,1)$. Für ganzzahliges r ist es die Verteilung der Anzahl Misserfolge bis zum r-ten Erfolg beim Werfen einer Münze mit Erfolgsparameter p, für $r = 1$ also die geometrische Verteilung. In R wird der Parameter r als `size` und p als `prob` bezeichnet.

Beweis: Die Gleichung $A^T\pi = 0$ lautet ausgeschrieben

$$-\gamma \cdot \pi(0) + \mu \cdot \pi(1) = 0,$$

sowie für $k = 1, 2, \ldots$

$$(\gamma + (k-1)\nu) \cdot \pi(k-1) - (\gamma + k(\mu+\nu)) \cdot \pi(k) + (k+1)\mu \cdot \pi(k+1) = 0.$$

Summiert man die Gleichungen für $k = 1, 2, \ldots n$, so ergibt sich

$$\sum_{k=0}^{n-1}(\gamma + k\nu)\pi(k) + \sum_{k=2}^{n+1} k\mu\pi(k) = \sum_{k=1}^{n}(\gamma + k\nu)\pi(k) + \sum_{k=1}^{n} k\mu\pi(k),$$

also

$$\gamma\pi(0) + (n+1)\mu\pi(n+1) = (\gamma + n\nu)\pi(n) + \mu\pi(1).$$

Zusammen mit der ersten Gleichung also für $n = 0, 1, 2, \ldots$

$$\pi(n+1) = \frac{\gamma + n\nu}{(n+1)\mu}\pi(n).$$

Wegen $\pi(1) = 1$ folgt durch Iteration

$$\pi(n) = \frac{\prod_{k=0}^{n-1}(\gamma + k\nu)}{\prod_{k=1}^{n}(k\mu)} = \frac{\nu^n \prod_{k=0}^{n-1}(\frac{\gamma}{\nu} + k)}{\mu^n\ n!} = \left(\frac{\nu}{\mu}\right)^n \frac{\Gamma(\frac{\gamma}{\nu} + n)}{\Gamma(\frac{\gamma}{\nu})\ n!}.$$

Man kann auch zeigen, dass $\sum_n \pi(n) = \infty$ falls $\nu \geq \mu$ und anderenfalls $\sum_n \pi(n) = (1 - \frac{\nu}{\mu})^{\gamma/\nu}$. Ich verzichte darauf. □

5 Simulationen mit dem Softwaresystem R

Das Softwaresystem R für die statistische Analyse von Daten ist weit verbreitet an Universitäten und in der Industrie. Es ist ein Open-Source Programm und kann kostenlos von www.R-project.org heruntergeladen werden. Neben statistischen Methoden enthält es auch viele Funktionen für grafische Darstellungen, Erzeugung von Zufallszahlen und auch für numerische lineare Algebra. Für die numerische Lösung von Differentialgleichungen muss noch ein Zusatzpaket geladen werden. Im File markov-sim.R findet man Funktionen, welche die im Text besprochenen Simulationsalgorithmen ausführen und die Resultate grafisch darstellen.

Literatur

Allen, L. S. J. (2011). *An Introduction to Stochastic Processes with Applications to Biology, Second Edition*. CRC Press, Boca Raton.

Durrett, R. (2016). *Essentials of Stochastic Processes, Second Edition*. Springer International, Switzerland.

Föllmer, H. and Künsch, H. (2013). Wahrscheinlichkeitsrechnung und Statistik. Vorlesungsskript ETH Zürich, D-Math, erhältlich auf ftp://ftp.stat.math.ethz.ch/Teaching/kuensch/skript-einf.pdf.

Karlin, S. and Taylor, H. (1975). *A First Course in Stochastic Processes, Second Edition*. Academic Press, New York.

Kurtz, T. G. (1981). *Approximations of Population Processes*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia.

Resnick, S. I. (1992). *Adventures in Stochastic Processes*. Birkhäuser, Boston.