Stationäre Verteilungen bei Mehrstufigen Prozessen

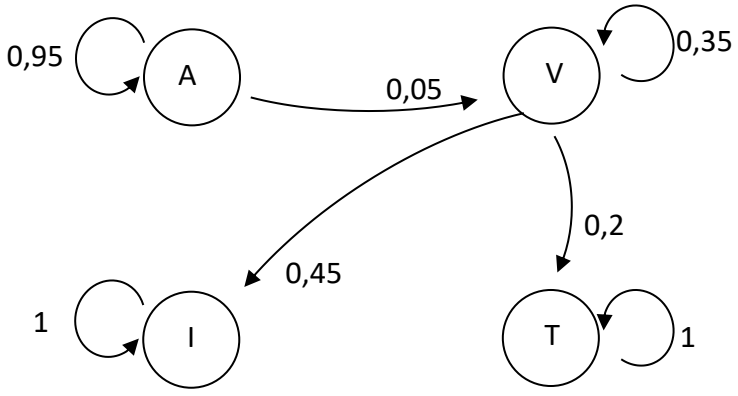
1. Ziele der Lerneinheit

In der folgenden Lerneinheit lernen Sie,

- was ein Fixvektor und was eine stationäre Verteilung ist.
- dass es zu jeder vorgegebenen Komponentensumme eine stationäre Verteilung gibt.
- unter welcher Bedingung es zu jeder vorgegebenen Komponentensumme nur eine stationäre Verteilung gibt.
- wie man diese eindeutige stationäre Verteilung durch Lösen eines linearen Gleichungssystems berechnen kann.
- dass unter der Doeblin-Bedingung die Spaltenvektoren der Matrizen M^n gegen die eindeutige stationäre Verteilung mit Komponentensumme 1 konvergieren.
- dass sich bei jedem mehrstufigen Prozess, der die Doeblin-Bedingung erfüllt, aus jeder Startverteilung schließlich die stationäre Verteilung wird.
- wie Google Internetseiten gewichtet und was dies mit stationären Verteilungen bei mehrstufigen Prozessen zu tun hat.

2. Ein letztes Mal die Pandemie-Simulation

Durch den folgenden Gozintographen simulieren Mathematiker eines Think Tanks eine durch einen Virus hervorgerufene Pandemie in einer Population von Menschen,



wobei die Zustände die Ansteckbaren (A), die mit dem Virus infizierten (V), die Immunen (I) und die Toten (T) kennzeichnen. Die Übergangswahrscheinlichkeiten beziehen sich dabei auf den Zeitraum von einem Tag.

Die Wissenschaftler fragen sich abschließend, ob es eine Verteilung einer Population von 1.000.000 Menschen auf die drei Zustände gibt, die sich – zumindest den Zahlen nach – nicht ändert: Gibt es einen Verteilungsvektor $\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$ mit $\begin{pmatrix} 0{,}95 & 0 & 0 & 0 \\ 0{,}05 & 0{,}35 & 0 & 0 \\ 0 & 0{,}45 & 1 & 0 \\ 0 & 0{,}2 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$?

Eine (mathematische) Lösung ist natürlich der Nullvektor $\vec{x} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$, der aber im Sachproblem nicht interessiert, da insgesamt 1.000.000 Individuen auf die Zustände verteilt sein sollen, sodass $x_1 + x_2 + x_3 + x_4 = 1.000.000$ sein muss. (Man spricht bei dieser Summe auch von der **Komponentensumme**.)

Definition. Jeder *vom Nullvektor verschiedene* Vektor $\vec{x}$, für den $M \cdot \vec{x} = \vec{x}$ gilt, heißt **Fixvektor** der Matrix M.

Wenn M eine stochastische Matrix und der Fixvektor $\vec{x}$ ein Verteilungsvektor (*keine* der Komponenten von $\vec{x}$ ist negativ) ist, dann heißt $\vec{x}$ auch **stationäre Verteilung.**

Um eine stationäre Verteilung zu finden, schreiben die Mathematiker die zu lösende Gleichung als vier lineare Gleichungen:

$$\begin{array}{lrcrcrcrcl} \text{I} & 0{,}95x_1 & & & & & & & = & x_1 \\ \text{II} & 0{,}05x_1 & + & 0{,}35x_2 & & & & & = & x_2 \\ \text{III} & & & 0{,}45x_2 & + & x_3 & & & = & x_3 \\ \text{IV} & & & 0{,}2x_2 & & & + & x_4 & = & x_4 \end{array}$$

Indem die Unbekannten auf die linke Seite gebracht werden, entsteht ein lineares Gleichungssystem.

$$\begin{array}{lrcrcrcl} \text{I} & -0{,}05x_1 & & & & & = & 0 \\ \text{II} & 0{,}05x_1 & - & 0{,}65x_2 & & & = & 0 \\ \text{III} & & & 0{,}45x_2 & & & = & 0 \\ \text{IV} & & & 0{,}2x_2 & & & = & 0 \end{array}$$

Aus Gleichung I folgt $x_1 = 0$, aus Gleichung III und Gleichung IV folgt $x_2 = 0$. Über x_3 und x_4 macht das LGS überhaupt keine Aussage. Dies bedeutet: Für jede Wahl von x_3 und x_4 ist

$$\vec{x} = \begin{pmatrix} 0 \\ 0 \\ x_3 \\ x_4 \end{pmatrix}$$

eine stationäre Verteilung. Man könnte für die Komponentensumme 1.000.000 zum Beispiel $x_3 = 600.000$ und $x_4 = 400.000$ wählen.

In diesem Fall gibt es also unendlich viele stationäre Verteilungen für die vorgegebene Komponentensumme, die hier die Anzahl der Individuen in der Population darstellt. Ist das immer so?

3. Existenz und Eindeutigkeit von stationären Verteilungen

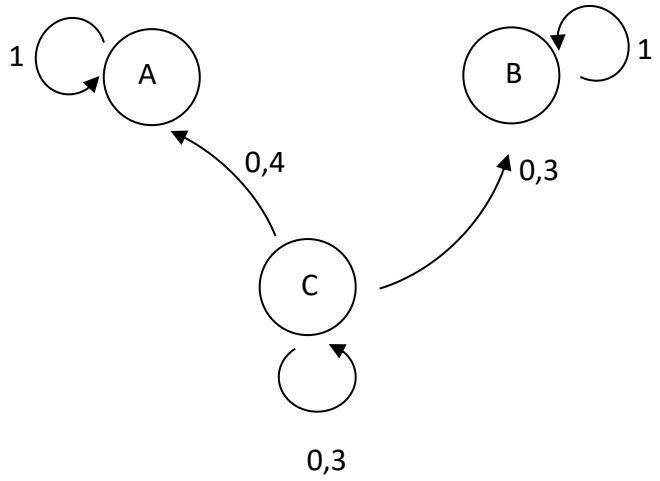
Satz. (Existenz von stationären Verteilungen)
Wenn M eine stochastische Matrix und p eine beliebige positive Zahl ist, so gibt es *mindestens* eine stationäre Verteilung $\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ mit Komponentensumme $x_1 + x_2 + \ldots + x_n = p$.

Wie wir im Pandemie-Beispiel gesehen haben, kann es mehrere (unendlich viele!) solche Verteilungen geben. Unter welchen Voraussetzungen es genau eine stationäre Verteilung gibt, sehen wir jetzt:

Satz. (Eindeutige Existenz der stationären Verteilung)

Wenn M eine stochastische Matrix und p eine beliebige positive Zahl ist, so gibt es genau dann **nur eine stationäre Verteilung** mit Komponentensumme $x_1 + x_2 + \ldots + x_n = p$, wenn es mindestens einen Zustand gibt, der von allen anderen Zuständen aus erreichbar ist.

Was soll „erreichbar“ bedeuten? In dem mehrstufigen Prozess mit dem Gozintographen



kann A nicht von B aus, B nicht von A aus und C kann sogar von keinem anderen Zustand aus über Kanten mit positiven Wahrscheinlichkeiten erreicht werden. Dies bedeutet: Es gibt *keinen* Zustand, der von jedem anderen aus erreichbar ist. Und in der Tat hat die zu diesem Prozess gehörende stochastische Matrix

$$M = \begin{pmatrix} 1 & 0 & 0{,}4 \\ 0 & 1 & 0{,}3 \\ 0 & 0 & 0{,}3 \end{pmatrix}$$

für die Zahl $p = 1$ die stationären Verteilungen

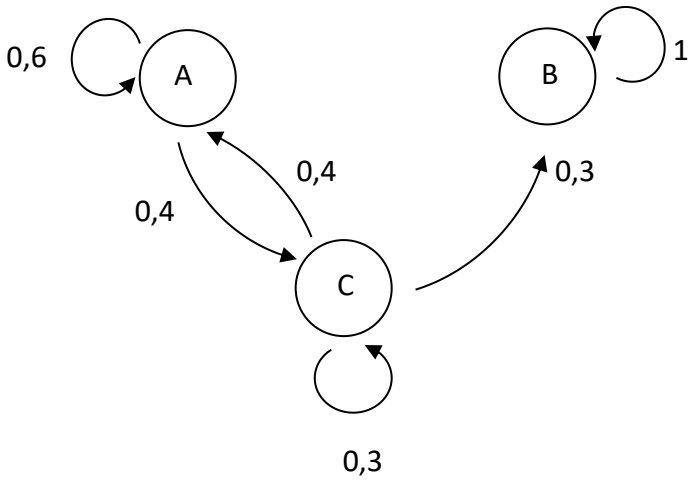
$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \text{ bzw. } \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \text{ bzw. } \begin{pmatrix} 0{,}5 \\ 0{,}5 \\ 0 \end{pmatrix}.$$

Es ist sogar jeder der Vektoren

$$\begin{pmatrix} x \\ 1-x \\ 0 \end{pmatrix} \text{ mit } 0 \le x \le 1$$

eine stationäre Verteilung von M, deren Komponentensumme $p = 1$ ist.

Wir modifizieren den Prozess, indem wir einen Übergang von Zustand A nach C zulassen:



Nun kann der Zustand B von allen anderen Zuständen aus erreicht werden. Die zu diesem Prozess gehörende stochastische Matrix

$$M = \begin{pmatrix} 0{,}6 & 0 & 0{,}4 \\ 0 & 1 & 0{,}3 \\ 0{,}4 & 0 & 0{,}3 \end{pmatrix}$$

hat nur eine stationäre Verteilung mit Komponentensumme $p = 1$, nämlich $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$.

4. Berechnen einer eindeutigen stationären Verteilung

Um in der Rent-a-Bike-Situation mit der Übergangsmatrix

$$M = \begin{pmatrix} 0{,}8 & 0{,}5 & 0{,}5 \\ 0{,}12 & 0{,}05 & 0{,}1 \\ 0{,}08 & 0{,}45 & 0{,}4 \end{pmatrix}$$

eine stationäre Verteilung der vorhandenen 100 Fahrräder auf die Stationen A, B, C zu zu finden, muss das Gleichungssystem

$$\begin{pmatrix} 0{,}8 & 0{,}5 & 0{,}5 \\ 0{,}12 & 0{,}05 & 0{,}1 \\ 0{,}08 & 0{,}45 & 0{,}4 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

mit der Zusatzbedingung $x_1 + x_2 + x_3 = 100$ gelöst werden. Wir schreiben das Gleichungssystem mit drei Gleichungen auf und bringen die Unbekannten direkt auf die linke Seite:

$$\begin{array}{lrcrcrcl} \text{I} & -0{,}2x_1 & + & 0{,}5x_2 & + & 0{,}5x_3 & = & 0 \\ \text{II} & 0{,}12x_1 & - & 0{,}95x_2 & + & 0{,}1x_3 & = & 0 \\ \text{III} & 0{,}08x_1 & + & 0{,}45x_2 & - & 0{,}6x_3 & = & 0 \end{array}$$

Dummerweise ist *dieses LGS nicht eindeutig lösbar*: Zuerst addieren wir die dritte Zeile zu ersten:

$$\begin{array}{l} \text{I} \to \text{I} + \text{III} \\ \begin{array}{lrcrcrcl} \text{I} & -0{,}12x_1 & + & 0{,}95x_2 & + & -0{,}1x_3 & = & 0 \\ \text{II} & 0{,}12x_1 & - & 0{,}95x_2 & + & 0{,}1x_3 & = & 0 \\ \text{III} & 0{,}08x_1 & + & 0{,}45x_2 & - & 0{,}6x_3 & = & 0 \end{array} \end{array}$$

und dann addieren wir die zweite Zeile zu der ersten:

$$\begin{array}{l} \text{I} \to \text{I} + \text{II} \\ \begin{array}{lrcrcrcl} \text{I} & & & & & 0 & = & 0 \\ \text{II} & 0{,}12x_1 & - & 0{,}95x_2 & + & 0{,}1x_3 & = & 0 \\ \text{III} & 0{,}08x_1 & + & 0{,}45x_2 & - & 0{,}6x_3 & = & 0 \end{array} \end{array}$$

Aus der Lösungstheorie für LGS ergibt sich, dass dieses Gleichungssystem unendlich viele Lösungen hat. Dies gilt auch ganz allgemein: Bei jeder stochastischen Matrix kann man im Gleichungssystem zur Bestimmung der stationären Verteilung durch Addition von Vielfachen der anderen Zeilen stets die erste Zeile (oder auch jede der anderen Zeilen!) zu einer Nullzeile machen, bei der auch die rechte Seite Null ist. Das LGS hat also immer unendlich viele Lösungen.

Im vorliegenden Kontext interessieren uns aber nur eine spezielle stationäre Verteilung, nämlich eine, bei der die Komponentensumme 100 ist: $x_1 + x_2 + x_3 = 100$.

Um eine stationäre Verteilung mit vorgegebene Komponentensumme $x_1 + x_2 + \ldots + x_n = p$ zu bestimmen, ersetzt man in dem zu lösenden Gleichungssystem $M \cdot \vec{x} = \vec{x}$ *eine* der Zeilen durch die Gleichung $x_1 + x_2 + \ldots + x_n = p$.

Wir ersetzen im Gleichungssystem

$$\begin{array}{lrcrcrcl} \text{I} & -0{,}2x_1 & + & 0{,}5x_2 & + & 0{,}5x_3 & = & 0 \\ \text{II} & 0{,}12x_1 & - & 0{,}95x_2 & + & 0{,}1x_3 & = & 0 \\ \text{III} & 0{,}08x_1 & + & 0{,}45x_2 & - & 0{,}6x_3 & = & 0 \end{array}$$

die erste Gleichung durch $x_1 + x_2 + x_3 = 100$

$$\begin{array}{lrcrcrcr} \text{I} & x_1 & + & x_2 & + & x_3 & = & 100 \\ \text{II} & 0{,}12x_1 & - & 0{,}95x_2 & + & 0{,}1x_3 & = & 0 \\ \text{III} & 0{,}08x_1 & + & 0{,}45x_2 & - & 0{,}6x_3 & = & 0 \end{array}$$

und bringen dieses LGS auf Zeilenstufenform:

$$\begin{array}{lrcrcrcr} \text{I} & x_1 & + & x_2 & + & x_3 & = & 100 \\ \text{II} & & & x_2 & + & \frac{2}{107}x_3 & = & \frac{1200}{107} \\ \text{III} & & & & & -\frac{147}{214}x_3 & = & -\frac{1300}{107} \end{array}$$

Hieraus ergeben sich die Lösungen

- $x_3 = \frac{1300}{107} : \frac{147}{214} = \frac{2600}{147}$
- $x_2 = \frac{1200}{107} - \frac{2}{107} \cdot \frac{2600}{147} = \frac{1600}{147}$
- $x_1 = 100 - \frac{2600}{147} - \frac{1600}{147} = \frac{500}{7}$.

Dies bedeutet:

- $\frac{2600}{147} \approx 18$ aller Fahrräder sind in Station C zu platzieren;
- $\frac{1600}{147} \approx 11$ aller Fahrräder sind in Station B aufzustellen;
- $\frac{500}{7} \approx 71$ aller Fahrräder sind in Station A zu deponieren.

5. Konvergenz gegen die stationäre Verteilung

Für die Rent-a-Bike-Situation hatten wir für 100 Fahrräder die stationäre Verteilung

$$\vec{x} = \begin{pmatrix} 500/7 \\ 1600/147 \\ 2600/147 \end{pmatrix} \approx \begin{pmatrix} 71{,}4286 \\ 10{,}8844 \\ 17{,}6871 \end{pmatrix}$$

ausgerechnet. Sind die Fahrräder auf diese Weise morgens auf die drei Stationen verteilt, so stellt sich abends auch genau diese Verteilung wieder ein.

Was ist aber, wenn man mit einer anderen Verteilung der Fahrräder auf die Stationen beginnt? Wir wissen bereits, dass wir die Verteilung nach einem, zwei, drei, allgemein nach n Tagen erhalten, indem wir die n-te M^n Potenz der Übergangsmatrix M mit der Startverteilung multiplizieren. Wir untersuchen zunächst, was bei der als Startverteilung $\vec{y} = \begin{pmatrix} 100 \\ 0 \\ 0 \end{pmatrix}$ passiert, wenn sich also alle Fahrräder zu Beginn in Station A befinden. Personen, die am ersten Morgen von einer der anderen Stationen ein Fahrrad mieten wollen, müssten also warten, bis ein bei Station A gemietetes Fahrrad dort abgegeben wird. Mithilfe des GTR ergeben sich – auf vier Stellen hinter dem Komma gerundet – für die folgenden Tage die folgenden Verteilungen:

n	1	2	5	10	30
$M^n \cdot \vec{y}$	$\begin{pmatrix} 80 \\ 12 \\ 8 \end{pmatrix}$	$\begin{pmatrix} 74 \\ 11 \\ 15 \end{pmatrix}$	$\begin{pmatrix} 71{,}498 \\ 10{,}8883 \\ 17{,}6137 \end{pmatrix}$	$\begin{pmatrix} 71{,}4287 \\ 10{,}8843 \\ 17{,}6869 \end{pmatrix}$	$\begin{pmatrix} 71{,}4286 \\ 10{,}8844 \\ 17{,}6871 \end{pmatrix}$

Die Verteilungen nähern sich der stationären Verteilung immer weiter an. Nach einem Monat ist kein Unterschied zur stationären Verteilung mehr zu sehen. Verteilt man die Fahrräder zu Beginn gleichmäßiger auf die Stationen, indem man als Startverteilung zum Beispiel

$$\vec{y} = \begin{pmatrix} 40 \\ 30 \\ 30 \end{pmatrix}$$

wählt, ergeben sich die folgenden Werte:

n	1	2	5	10	30
$M^n \cdot \vec{y}$	$\begin{pmatrix} 62 \\ 9{,}3 \\ 28{,}7 \end{pmatrix}$	$\begin{pmatrix} 68{,}6 \\ 10{,}775 \\ 20{,}625 \end{pmatrix}$	$\begin{pmatrix} 71{,}3522 \\ 10{,}88 \\ 17{,}7678 \end{pmatrix}$	$\begin{pmatrix} 71{,}4284 \\ 10{,}8843 \\ 17{,}6873 \end{pmatrix}$	$\begin{pmatrix} 71{,}4286 \\ 10{,}8844 \\ 17{,}6871 \end{pmatrix}$

Wiederum hat sich nach (spätestens) einem Monat automatisch die Gleichgewichtsverteilung eingestellt. Das ist kein Zufall:

Satz. (Konvergenz gegen die Gleichgewichtsverteilung)

Angenommen, M ist eine stochastische Matrix, für die die **Doeblin-Bedingung** gilt, d.h., mindestens eine der Matrizen $M, M^2, M^3, M^4, \ldots$ hat eine *Zeile*, in der *alle* Einträge *strikt*

positiv sind. Dann gibt es für jede positive Komponentensumme p genau eine stationäre Verteilung $\vec{x}$ mit Komponentensumme p. Außerdem gilt:

- Alle Komponenten von $\vec{x}$ sind positiv.
- Für jede Verteilung $\vec{y}$ mit Komponentensumme p konvergiert die Folge $M\cdot\vec{y}, M^2\cdot\vec{y}, M^3\cdot\vec{y}, \ldots$ gegen $\vec{x}$: Für alle hinreichend großen n ist $M^n\cdot\vec{y}\approx\vec{x}$, und der Unterschied von $M^n\cdot\vec{y}$ zu $\vec{x}$ wird immer kleiner, wenn n größer wird.
- Für genügend großes n sind die Spaltenvektoren von M^n alle gleich und stimmen mit der stationären Verteilung mit Komponentensumme 1 überein.

Wenn die Doeblin-Bedingung gilt, kann man also die stationäre Verteilung mit Komponentensumme p wie folgt berechnen:

- Berechne M^k für ein sehr großes k.
- Nimm einen der Spaltenvektoren und multipliziere ihn mit p.

→ *Übung 1*

Die **Doeblin-Bedingung** bedeutet, dass es einen Zustand Z gibt, sodass man im Gozintographen von jedem Zustand aus über genau k Knoten (diese müssen nicht verschieden sein) entlang von Kanten mit positiven Wahrscheinlichkeiten nach Z gelangen kann. Mit anderen Worten: Ein Objekt, das sich zu Beginn in irgendeinem Zustand befindet, kann mit positiver Wahrscheinlichkeit nach genau k Übergängen in Z landen.

Die Doeblin-Bedingung ist benannt nach dem Mathematiker Wolfgang Doeblin (1915- 1940), einem Sohn des Schriftstellers Alfred Döblin. Wolfgang Doeblin emigrierte in den dreißiger Jahren nach Frankreich und erhielt dort die französische Staatsbürgerschaft. Im Krieg von Frankreich eingezogen wurde sein Bataillon im Juni 1940 von deutschen Truppen eingenommen. Auf der Flucht wählte Doeblin den Freitod. Wenige Monate vor seinem Tod – im Februar 1940 – hinterlegte Doeblin seine neuesten mathematischen Resultate in einem versiegelten Brief bei der Académie des Sciences in Paris. Der Brief wurde erst im Jahr 2000 geöffnet. Die Aufzeichnungen Doeblins nehmen bedeutende Resultate vorweg, die von anderen erst in der zweiten Hälfte des zwanzigsten Jahrhunderts (wieder-)gefunden wurden.


Wolfgang Doeblin

6. Eine Anwendung: Google's PageRank-Algorithmus

Seit September 1998 ist die Suchmaschine Google online. Sie wird betrieben von der Firma Google Inc., die von Larry Page und Sergei Brin, zwei ehemaligen Informatik-Studenten an der Stanford University, im Jahr 1998 gegründet wurde.

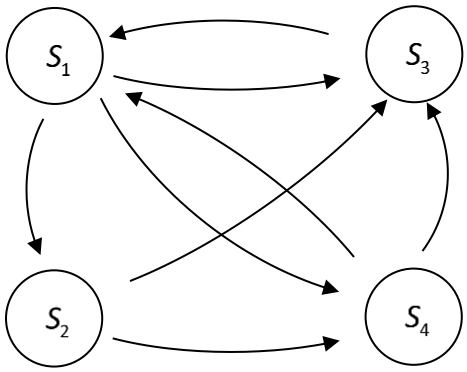
Besonders in den ersten Jahren - mittlerweile haben die Konkurrenten sich verbessert – zeichnete sich Google dadurch aus, dass die zu einem Suchbegriff gefundenen Internetseiten nicht in mehr oder weniger zufälliger Reihenfolge aufgelistet werden. Die Suchergebnisse werden vielmehr in einer Reihenfolge präsentiert, die der Wichtigkeit der Seiten entspricht. Google verwendet hierzu den von Page und Brin noch an der Stanford University entwickelten Algorithmus **PageRank**.

PageRank basiert im Wesentlichen darauf, dass jeder Seite S des www ein Gewicht x_S, dessen Wert zwischen 0 und 1 liegt, zugemessen wird. Google präsentiert dann seine Suchergebnisse nach dem Gewicht absteigend sortiert.

Für die Berechnung des Gewichtes wird davon ausgegangen, dass ein Surfer auf die Seite S gerät, indem er sich zuvor auf einer anderen Seite befindet, die einen Link *auf* S hat. Auf dieser Seite hat er den Link auf die Seite S ausgewählt und ist so auf S gelangt.

Für jede Seite S des www wird nun die Anzahl n_S der Seiten bestimmt, auf die sie verlinkt ist[1]. Auf jede Seite, auf die S verlinkt ist, vererbt sie dann den Wert x_S/n_S. Das Gewicht einer Seite wird also gleichmäßig auf die Seiten übertragen, auf die diese Seite verlinkt ist.

Die Frage ist: Wie können die Gewichte berechnet werden? Wir betrachten zur Veranschaulichung zunächst ein einfaches, kleines www, das aus nur vier Seiten S_1, S_2, S_3, S_4 besteht. Im folgenden Graphen bedeutet ein Pfeil zwischen zwei Knoten, dass ein Link von der ersten auf die zweite Internetseite zeigt.



Für die Gewichte der Seiten schreiben wir kurz x_1 statt x_{S_1} usw.

Das Gewicht der Seite S_1 ergibt sich aus den Gewichten der Seiten, die auf sie verlinken. Dies sind S_3 und S_4. Da von S_3 nur ein Link ausgeht, erhält S_1 das gesamte Gewicht dieser Seite; da von S_4 zwei Links ausgehen, erhält S_1 die Hälfte des Gewichtes von S_4: $\underline{\underline{x_1 = x_3 + 1/2\,x_4}}$.

Auf die Seite S_2 verlinkt nur die Seite S_1. Da S_1 drei Outlinks hat, erhält S_2 ein Drittel des Gewichts von S_1: $\underline{\underline{x_2 = 1/3\,x_1}}$.

Die Seite S_3 erhält Links von allen drei anderen Seiten. Durch Zählen der Outlinks jeder Seite erhält an für das Gewicht von Seite S_3 die Gleichung $\underline{\underline{x_3 = 1/3\,x_1 + 1/2\,x_2 + 1/2\,x_4}}$.

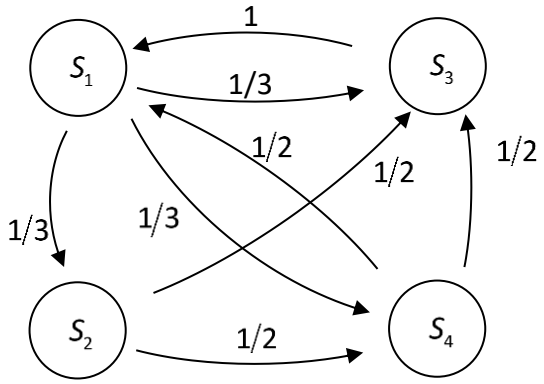
Ebenso ergibt sich für das Gewicht der Seite S_4: $\underline{\underline{x_4 = 1/3\,x_1 + 1/2\,x_2}}$.

Um die Gewichte der Seiten zu bestimmen, ist somit die Gleichung

[1] Links, die von einer Seite ausgehen, werden im Folgenden auch als **Outlinks** bezeichnet.

$$\begin{pmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

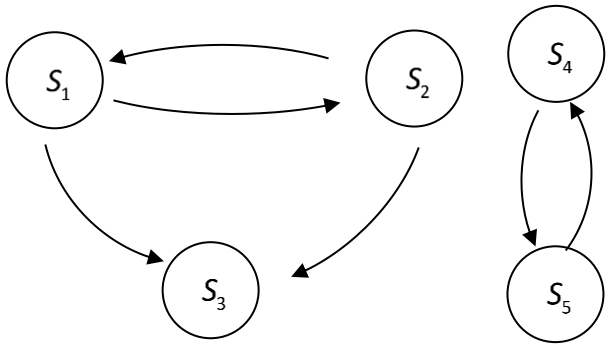
zu lösen, und damit eine stationäre Verteilung für eine stochastische Matrix zu bestimmen! In der graphischen Darstellung des Webs entsprechen dann den Knoten und Kanten denen des Gozintographen dieser Matrix:



Da z.B. S_1 von jedem anderen Knoten aus erreichbar ist, existiert in dieser Situation genau eine stationäre Verteilung mit Komponentensumme 1. Diese ist

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 12/31 \\ 4/31 \\ 9/31 \\ 6/31 \end{pmatrix}.$$

Die Situation ist nicht immer so offensichtlich wie in diesem einfachen Fall, da nicht immer eine stochastische Matrix entstehen muss. Hierzu betrachten wir ein etwas größeres www mit fünf Seiten $S_1, S_2, \ldots, S_5$. Die Links zwischen den Seiten werden durch den folgenden Graphen wiedergegeben:



Als problematisch werden sich zwei Aspekte herausstellen:

- Das Netz zerfällt in Teilnetze, die nicht miteinander verbunden sind. Das eine Teilnetz bilden die Seiten S_1, S_2, S_3, das zweite Teilnetz bilden die Seiten S_4, S_5.

- Die Seite s_3 hat keinen Outlink.

Aus dem ersten Aspekt folgt, dass es keinen Zustand gibt, der von allen anderen aus erreichbar ist, sodass es keine eindeutige stationäre Verteilung geben kann!

Bilden wir wie eben wieder das Lineare Gleichungssystem zur Bestimmung der Gewichte, ergibt sich die Matrix

$$M = \begin{pmatrix} 0 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Die Nullspalte rührt daher, dass von Seite S_3 kein Link ausgeht. Die Matrix M ist also nicht stochastisch.

Eine Möglichkeit, dies für die weitere Berechnung zu beheben, besteht darin, der Seite S_3 einen Link auf jede der vier anderen Seiten künstlich hinzuzufügen. Inhaltlich entspricht dem, dass ein Surfer, der auf Seite S_3 gelangt, zufällig mit jeweils gleicher Wahrscheinlichkeit eine der anderen Seiten zum Weitersurfen auswählt. Die Matrix M wird damit zu

$$M' = \begin{pmatrix} 0 & 1/2 & 1/4 & 0 & 0 \\ 1/2 & 0 & 1/4 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/4 & 0 & 1 \\ 0 & 0 & 1/4 & 1 & 0 \end{pmatrix}.$$

Diese Modifikation führt nun auch dazu, dass hier zum Beispiel die Seite S_5 von jeder anderen Seite aus erreichbar ist, sodass genau eine stochastische Gleichgewichtsverteilung existiert. Wegen

$$\begin{pmatrix} 0 & 1/2 & 1/4 & 0 & 0 \\ 1/2 & 0 & 1/4 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/4 & 0 & 1 \\ 0 & 0 & 1/4 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1/2 \\ 1/2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1/2 \\ 1/2 \end{pmatrix}$$

muss

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1/2 \\ 1/2 \end{pmatrix}$$

diese Gleichgewichtsverteilung sein. Dies ist jedoch für die vorliegende Situation höchst unbefriedigend, da mehr als der Hälfte der Seiten des www das Gewicht Null zugewiesen würde.

Um diesen Problem abzustellen, haben Page und Brin die folgende Lösung ersonnen: Sie gehen davon aus, dass ein Surfer nur mit einer bestimmten Wahrscheinlichkeit p den Links auf einer Seite folgt und mit Wahrscheinlichkeit $1-p$ im nächsten Schritt eine zufällig ausgewählte Internetseite ansteuert. Dies wird durch die Matrix

$$M_p = p \cdot M' + (1-p) \cdot \begin{pmatrix} 1/5 & \cdots & 1/5 \\ \vdots & \ddots & \vdots \\ 1/5 & \cdots & 1/5 \end{pmatrix}$$

ausgedrückt. Zu lösen ist dann das Gleichungssystem $M_p \cdot \vec{x} = \vec{x}$.

Google arbeitet nach unbestätigten Informationen mit $p = 0{,}85$. Im vorliegenden Fall erhalten wir hiermit

$$M_{0,85} = \begin{pmatrix} 0{,}03 & 0{,}455 & 0{,}2425 & 0{,}03 & 0{,}03 \\ 0{,}455 & 0{,}03 & 0{,}2425 & 0{,}03 & 0{,}03 \\ 0{,}455 & 0{,}455 & 0{,}03 & 0{,}03 & 0{,}03 \\ 0{,}03 & 0{,}03 & 0{,}2425 & 0{,}03 & 0{,}88 \\ 0{,}03 & 0{,}03 & 0{,}2425 & 0{,}88 & 0{,}03 \end{pmatrix}.$$

Mit einem Computer findet man, dass das Gleichungssystem $A_{0,85} \cdot \vec{x} = \vec{x}$ die Lösung

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 291/3155 \\ 291/3155 \\ 342/3155 \\ 2231/6310 \\ 2231/6310 \end{pmatrix} \approx \begin{pmatrix} 0{,}09223 \\ 0{,}09223 \\ 0{,}1084 \\ 0{,}35357 \\ 0{,}35357 \end{pmatrix}$$

hat. Es ist offensichtlich, dass das Lösen eines derartigen LGS sehr zeitaufwändig ist. In Realität müssen nun nicht nur fünf, sondern mehrere Millionen Seiten in die Kalkulation einbezogen werden. Da die – zum Beispiel für $p = 0{,}85$ – entstehende Übergangsmatrix M_p aber die Doeblin-Bedingung erfüllt, kann die stationäre Verteilung auch durch Berechnung von $(M_p)^n$ für großes n gefunden werden: Alle Spaltenvektoren von $(M_p)^n$ sind dann gleich und stimmen mit der stationären Verteilung mit Komponentensumme 1 überein.

Bei typischen in der Praxis vorkommenden Werten ist diese Methode mehrere Millionen mal schneller als die Berechnung der Verteilung durch Lösen des LGS!

Literaturhinweise zum PageRank-Algorithmus

Für Interessierte seien ein Artikel und ein (Workshop-)Skript als Literaturhinweis zum Page-Rank-Algorithmus genannt. Die im Lehrtext mitgeteilten Informationen über den PageRank-Algorithmus entstammen größtenteils diesen Quellen.

KURT BRYAN, TANYA LEISE: *The $25,000,000,000 Eigenvector. The Linear Algebra Behind Google*. Online abrufbar unter http://www.rose-hulman.edu/~bryan/googleFinalVersionFixed.pdf (Zugriff 15.04.2020)

WOLFGANG KONEN: *Google Explained: Eigenwerte, Graphen, Flüsse.* Online abrufbar unter http://www.gm.fh-koeln.de/~konen/Mathe2-SS2007/Workshop-Google/PageRank-Workshop2-ext.pdf (Zugriff 13.07.2012) bzw. https://silo.tips/download/14-google-explained-eigenwerte-graphen-flsse (Zugriff 15.04.2020)

Übungen zur Lerneinheit
Stationäre Verteilungen bei Mehrstufigen Prozessen

Übung 1.

1. Berechnen Sie, wieviel Prozent der Fahrräder an den drei Radstationen A, B und C bei der Übergangsmatrix

$$M = \begin{pmatrix} 0{,}2 & 0{,}3 & 0{,}2 \\ 0{,}4 & 0{,}6 & 0{,}5 \\ 0{,}4 & 0{,}1 & 0{,}3 \end{pmatrix}$$

   verteilt werden müssen, damit die Verteilung der Fahrräder sich nicht ändert. Begründen Sie auch, warum es nur eine stationäre Verteilung geben kann.

2. Für eine Gruppe von 10.000 Menschen sind die Wahrscheinlichkeiten für jährliche Übergänge zwischen den Nichtrauchern, Gelegenheitsrauchern und täglichen Rauchern durch die Übergangsmatrix

$$M = \begin{pmatrix} 0{,}8 & 0{,}35 & 0{,}05 \\ 0{,}12 & 0{,}33 & 0{,}1 \\ 0{,}08 & 0{,}32 & 0{,}85 \end{pmatrix}$$

   gegeben. Begründen Sie, dass es nur eine stationäre Verteilung geben kann, und berechnen Sie diese. Bestimmen Sie dann, wie viele der 10.000 Menschen der Gruppe nach längerer Zeit Nichtraucher, Gelegenheitsraucher und täglichen Raucher sein werden.

3. In einem sehr einfachen Modell des Wetters in Bonn wird davon ausgegangen, dass ein Tag entweder regnerisch oder trocken ist, wobei die Wahrscheinlichkeiten für die täglichen Übergänge durch die Matrix

$$M = \begin{pmatrix} 0{,}66 & 0{,}25 \\ 0{,}34 & 0{,}75 \end{pmatrix}$$

   gegeben sind. Berechnen Sie hiermit, mit welcher Wahrscheinlichkeit ein Tag in Bonn regnerisch oder trocken ist.
   Hinweis. Wenn x_r die Wahrscheinlichkeit für einen regnerischen Tag und x_t die Wahrscheinlichkeit für einen trockenen Tag ist, so muss $A \cdot \begin{pmatrix} x_r \\ x_t \end{pmatrix} = \begin{pmatrix} x_r \\ x_t \end{pmatrix}$ und $x_r + x_t = 1$ gelten.

4. In einer Population von Insekten wurde die Verteilung zwei verschiedener Merkmale A und B über längere Zeit beobachtet. Dabei hatten Insekten mit Merkmal A zu 70 % Nachkommen mit Merkmal A und Insekten mit Merkmal B zu 20 % Nachkommen mit Merkmal A, was durch die Übergangsmatrix

$$M = \begin{pmatrix} 0{,}7 & 0{,}2 \\ 0{,}3 & 0{,}8 \end{pmatrix},$$

   ausgedrückt wird. Berechnen Sie, auf welche Werte wird sich das Verhältnis von Trägern des Merkmals A zu den Trägern des Merkmals B langfristig einstellen wird.

5. Die Farbe (grün oder gelb) der Schoten einer Erbsenpflanze wird durch zwei Gene bestimmt, wobei das Gen G für grüne Farbe dominant und das Gen g für gelbe Farbe rezessiv ist. Wird ein *mischerbiges Elternteil* mit einer sehr großen Population von Pflanzen beliebigen Genotyps gekreuzt, werden die Wahrscheinlichkeiten für den Übergang „Genotyp 2. Elternteil“ $\rightarrow$ „Genotyp Nachkomme“ durch die Übergangsmatrix

$$\begin{array}{cc} & \text{von} \\ & \begin{array}{ccc} \text{GG} & \text{Gg} & \text{gg} \end{array} \\ \begin{array}{c} \\ \text{nach} \end{array} & \begin{array}{c} \text{GG} \\ \text{Gg} \\ \text{gg} \end{array} \begin{pmatrix} 0{,}5 & 0{,}25 & 0 \\ 0{,}5 & 0{,}5 & 0{,}5 \\ 0 & 0{,}25 & 0{,}5 \end{pmatrix} \end{array}$$

gegeben.

a) Begründen Sie, dass es genau eine stationäre Verteilung mit Komponentensumme 1 gibt, und berechnen Sie diese.

b) Über mehrere Generationen hinweg wird nun jeweils die Population der Nachkommen erneut mit einem mischerbigen Elternteil gekreuzt. Begründen Sie, dass sich die Verteilung der Genotypen in den Nachkommen nach oftmaliger Kreuzung mit einem mischerbigen Elternteil nicht mehr ändert, und geben Sie an, welche wie hoch der Anteile der reinerbig dominanten, der mischerbigen und der reinerbig rezessiven Pflanzen in den Nachkommen schließlich sein werden.

6. Werden in der Situation der vorstehenden Übung als erstes Elternteil nur reinerbig dominante Erbsen verwendet, werden die Wahrscheinlichkeiten für den Übergang „Genotyp 2. Elternteil“ $\rightarrow$ „Genotyp Nachkomme“ durch die Übergangsmatrix

$$\begin{pmatrix} 1 & 0{,}5 & 0 \\ 0 & 0{,}5 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

gegeben. In diesem Fall gibt es unter den Nachkommen keine Erbsen mit Genotyp gg.

a) Begründen Sie, dass es genau eine stationäre Verteilung mit Komponentensumme 1 gibt, und berechnen Sie diese.

b) Auch in der vorliegenden Situation wird über mehrere Generationen hinweg jeweils die Population der Nachkommen erneut mit einem reinerbig dominanten Elternteil gekreuzt. Begründen Sie, dass sich die Verteilung der Genotypen in den Nachkommen nach oftmaliger Kreuzung mit einem reinerbig dominanten Elternteil nicht mehr ändert, und geben Sie an, wie hoch der Anteile der reinerbig dominanten, der mischerbigen und der reinerbig rezessiven Pflanzen in den Nachkommen schließlich sein werden.