Projektaufgaben Block 1

Carlo Michaelis, <Matr.-Nr.>; Lukas Ruff, 572521

15 November 2016

1 Infinite-Monkey-Theorem

1.1 Formulierung und Beweis des Infinite-Monkey-Theorems

Laut dem Infinite-Monkey-Theorem wird ein Affe, der unendlich lange zufällig auf einer Schreibmaschine tippt, fast sicher jede beliebige Zeichenkette unendlich oft schreiben. Diese bildhafte Interpretation des mathematischen Satzes soll der gedanklichen Einordnung von Unendlichkeit dienen. Das Infinite-Monkey-Theorem ist ein Beispiel für die Anwendung des Lemmas von Borel-Cantelli.

Satz 1 (Lemma von Borel-Cantelli)**.** *Sei $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum.*

a) Ist $(A_n)_{n\in\mathbb{N}}$ eine Folge von Ereignissen mit $\sum_{n\in\mathbb{N}} P(A_n) < \infty$, so gilt

$$P\left(\limsup_{n\to\infty} A_n\right) = 0.$$

b) Ist $(A_n)_{n\in\mathbb{N}}$ eine Folge von unabhängigen Ereignissen mit $\sum_{n\in\mathbb{N}} P(A_n) = +\infty$, so gilt

$$P\left(\limsup_{n\to\infty} A_n\right) = 1.$$

Um das Infinite-Monkey-Theorem formulieren zu können, modellieren wir zunächst einen geeigneten Wahrscheinlichkeitsraum. Definiere dazu ein endliches Alphabet $\Omega := \{\text{a, b, c, } \ldots\}$, d.h. Ω ist eine Menge von Zeichen (Buchstaben, Satzzeichen, Zahlen, etc.) mit $|\Omega| = n$ für ein $n \in \mathbb{N}$. Setze weiter die zugehörige σ-Algebra als Potenzmenge $\mathcal{F} := \mathcal{P}(\Omega)$ und definiere das Wahrscheinlichkeitsmaß P als diskrete Gleichverteilung, d.h. $P(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{n}$ für $A \in \mathcal{F}$. Somit modelliert $(\Omega, \mathcal{F}, P)$ das Ziehen von Zeichen aus einem Alphabet mit gleicher Wahtscheinlichkeit. Den Übergang zu Zeichenketten (Folgen von Zeichen aus dem Alphabet Ω) können wir nun mittels des (stets existierenden und eindeutigen) Produktraumes $(\Omega^{\mathbb{N}}, \mathcal{F}^{\mathbb{N}}, \mathbb{P})$ mit $\mathbb{P} := P^{\mathbb{N}}$ bewerkstelligen. Hiermit können wir das Infinite-Monkey-Theorem wie folgt formulieren und beweisen:

Satz 2 (Infinite-Monkey-Theorem)**.** *Betrachte den oben definierten Produktraum $(\Omega^{\mathbb{N}}, \mathcal{F}^{\mathbb{N}}, \mathbb{P})$ von (unendlich langen) Zeichenketten und sei $s = (s_1, \ldots, s_m)^\top \in \Omega^m$ eine beliebige Zeichenkette der Länge $m \in \mathbb{N}$ (String). Definiere zu $k \in \mathbb{N}$ das Ereignis*

$$A_k := \left\{\omega = (\omega_j)_{j\in\mathbb{N}} \in \Omega^{\mathbb{N}} : \omega_k = s_1, \ldots, \omega_{k+m-1} = s_m\right\},$$

d.h. A_k ist das Ereignis, dass String s der Länge m an der Stelle k in einer Zeichenkette beginnt. Dann ist $(A_{jm+1})_{j\in\mathbb{N}_0}$ eine Folge unabhängiger Ereignisse und es gilt

$$\mathbb{P}\left(\limsup_{j\to\infty} A_{jm+1}\right) = 1.$$

Beweis. Es sei $\pi_i : \Omega^{\mathbb{N}} \to \Omega, (\omega_j)_{j\in\mathbb{N}} \mapsto \omega_i$ die i-te Koordinatenprojektion auf dem kartesischen Produkt $\Omega^{\mathbb{N}}$. Weiter sei $(I_j)_{j\in\mathbb{N}_0}$ eine Zerlegung von $\mathbb{N}$ in disjunkte Blöcke der Länge m, d.h.

$$\mathbb{N} = \bigcup_{j\in\mathbb{N}_0}^{\cdot} I_j \qquad \text{mit} \qquad I_j := \bigcup_{k=1}^{m} \{jm + k\},\ j \in \mathbb{N}_0.$$

Dann folgt für $j \in \mathbb{N}_0$

$$\begin{aligned}\mathbb{P}(A_{jm+1}) &= \mathbb{P}\left(\left\{\omega : \omega_{jm+1} = s_1, \ldots, \omega_{(j+1)m} = s_m\right\}\right) \\ &= \mathbb{P}\left(\bigcap_{k=1}^{m} \pi_{jm+k}^{-1}(\{s_k\})\right) \\ &= \prod_{k=1}^{m} P(\{s_k\}) = \left(\frac{1}{n}\right)^m,\end{aligned}$$

da $\mathbb{P} = P^{\mathbb{N}}$ Produktmaß ist. Weiter folgt für $j_1, j_2 \in \mathbb{N}_0, j_1 \neq j_2$:

$$\begin{aligned}\mathbb{P}\left(A_{j_1 m+1} \cap A_{j_2 m+1}\right) &= \mathbb{P}\left(\left\{\omega : \omega_{j_1 m+1} = s_1, \ldots, \omega_{(j_1+1)m} = s_m\right\} \cap \left\{\omega : \omega_{j_2 m+1} = s_1, \ldots, \omega_{(j_2+1)m} = s_m\right\}\right) \\ &= \mathbb{P}\left(\left(\bigcap_{k=1}^{m} \pi_{j_1 m+k}^{-1}(\{s_k\})\right) \cap \left(\bigcap_{k=1}^{m} \pi_{j_2 m+k}^{-1}(\{s_k\})\right)\right) \\ &= \left(\prod_{k=1}^{m} P(\{s_k\})\right)^2 = \left(\frac{1}{n}\right)^m \left(\frac{1}{n}\right)^m = \mathbb{P}(A_{j_1 m+1}) \cdot \mathbb{P}(A_{j_2 m+1}),\end{aligned}$$

da $I_{j_1} \cap I_{j_2} = \varnothing$ für $j_1 \neq j_2$. Somit ist $(A_{jm+1})_{j \in \mathbb{N}_0}$ eine Folge unabhängiger Ereignisse mit

$$\sum_{j \in \mathbb{N}_0} P(A_{jm+1}) = \sum_{j \in \mathbb{N}_0} \left(\frac{1}{n}\right)^m = +\infty.$$

D.h. mit Teil b) des Lemma von Borel-Cantelli folgt die Behauptung. □

Es gilt also

$$\begin{aligned}\mathbb{P}\left(\limsup_{j \to \infty} A_{jm+1}\right) &= \mathbb{P}\left(\bigcap_{j \in \mathbb{N}_0} \bigcup_{k \geq j} A_{km+1}\right) \\ &= \mathbb{P}\left(\left\{\omega \in \Omega^{\mathbb{N}} : \forall j \in \mathbb{N}_0\, \exists k \geq j : \omega \in A_{km+1}\right\}\right) \\ &= \mathbb{P}\left(\left\{\omega \in \Omega^{\mathbb{N}} : \omega \in A_{jm+1} \text{ für unendlich viele } j \in \mathbb{N}_0\right\}\right) = 1,\end{aligned}$$

d.h. jeder beliebige String s der Länge m erscheint fast sicher unendlich oft.

1.2 Simulation eines Infinite-Monkey Experiments

In diesem Abschnitt wollen wir das Infinite-Monkey-Theorem experimentell simulieren. Wir schreiben hierzu eine R-Funktion die solange zufällig Zeichen aus dem Alphabet $\Omega = \{a, b, \ldots, y, z\}$ mit gleicher Wahrscheinlichkeit (d.h. $\frac{1}{26}$) auswählt, bis eine vorgegebene Zeichenkette $s = (s_1, \ldots, s_m)^\top \in \Omega^m, m \in \mathbb{N}$, vollständig erschienen ist. Die R-Funktion gibt dann die Anzahl der bis zum Erscheinen der Zeichenkette s gezogenen Zeichen zurück. Damit simulieren wir die Zufallsvariable $X : \Omega^{\mathbb{N}} \to \mathbb{N}$ mit

$$X(\omega) = \min\{k + m - 1 : \omega_k = s_1, \ldots, \omega_{k+m-1} = s_m \text{ für } k \in \mathbb{N}\}$$

Wir haben die Funktion wie folgt in R implementiert:[1]

[1] Eine effizientere Implementierung wäre je Schleifendurchlauf eine größere Anzahl an Zeichen zu generieren und den resultierenden Zeichen-Block mit einem Fenster nach `strTarget` zu durchsuchen. Dabei muss beachtet werden, dass `strTarget` auch in der Überlappung zweier Blöcke erscheinen könnte. Für unsere Untersuchung ist eine vereinfachte (ineffiziente) Implementierung jedoch ausreichend. Für große Samples oder einen langen String `strTarget` sollte jedoch auf jeden Fall eine effizientere Implementierung herangezogen werden.

```r
fnInfiniteMonkey <- function(strTarget) {
  # This function is a simulation of the Infinite-Monkey-Theorem. It generates a
  # random sequence of letters until a given target string appears.
  #
  # Args:
  #   strTarget: The target string which should be matched
  #
  # Returns:
  #   The number of generated letters until the target string appeared

  # Split target string to vector of chars
  vecCharTarget <- strsplit(strTarget, "")[[1]]
  # Get the number of letters in target string
  nTarget <- length(vecCharTarget)

  # Set counting variable (at least nTarget letters needed)
  nCounter <- nTarget

  # Switch on the monkey (i.e. sample the first nTarget letters)
  vecLetterSeqTail <- sample(letters, nTarget, replace = TRUE)

  # Let the monkey type until target string was written
  while(!identical(vecCharTarget, vecLetterSeqTail)) {

    # Let the monkey hit another key (sample next letter) and store in
    # vecLetterSeqTail (first in, first out)
    if (nTarget == 1) {
      vecLetterSeqTail <- sample(letters, 1)
    } else {
      vecLetterSeqTail <- c(vecLetterSeqTail[2:nTarget],
                            sample(letters, 1, replace = TRUE))
    }
    # Count
    nCounter <- nCounter + 1
  }

  # Return the length of the generated letter sequence
  return(nCounter)
}
```

In Figure 1 haben wir ein Histogramm der simulierten Anzahlen von Zeichen und die durchschnittliche Anzahl von Zeichen, bis die Zeichenkette `'ab'` vollständig erschienen ist, für 10^4 Samples geplottet. Wir können sehen, dass die Anzahl von Zeichen X einer rechtsschiefen Verteilung folgt. Die durch das starke Gesetz der großen Zahlen motivierte Monte-Carlo-Approximation von $\mathbb{E}[X]$ liegt bei 686.9719.

Zur Simulation von Infinite-Monkey Experimenten wollen wir abschließend noch folgende Überlegung durchführen: würden wir anstatt einer fortlaufenden Zeichenfolge, bei welcher je Schleife nur ein weiterer Letter generiert wird, je Schleifendurchlauf einen Zeichen-Block der Länge m generieren und mit s vergleichen, so entspräche jeder Schleifendurchlauf einem Bernoulli-Experiment mit Erfolgswahrscheinlichkeit $p = \left(\frac{1}{n}\right)^m$. In diesem Fall wäre die Anzahl der Blöcke, die notwendig sind um eine Übereinstimmung mit s zu erhalten, gerade geometrisch-verteilt mit Erfolgswahrscheinlichkeit $p = \left(\frac{1}{n}\right)^m$. Diese Zerlegung in Blöcke der Länge m entspricht genau der Folge unabhängiger Ereignisse aus dem Beweis des Theorems.

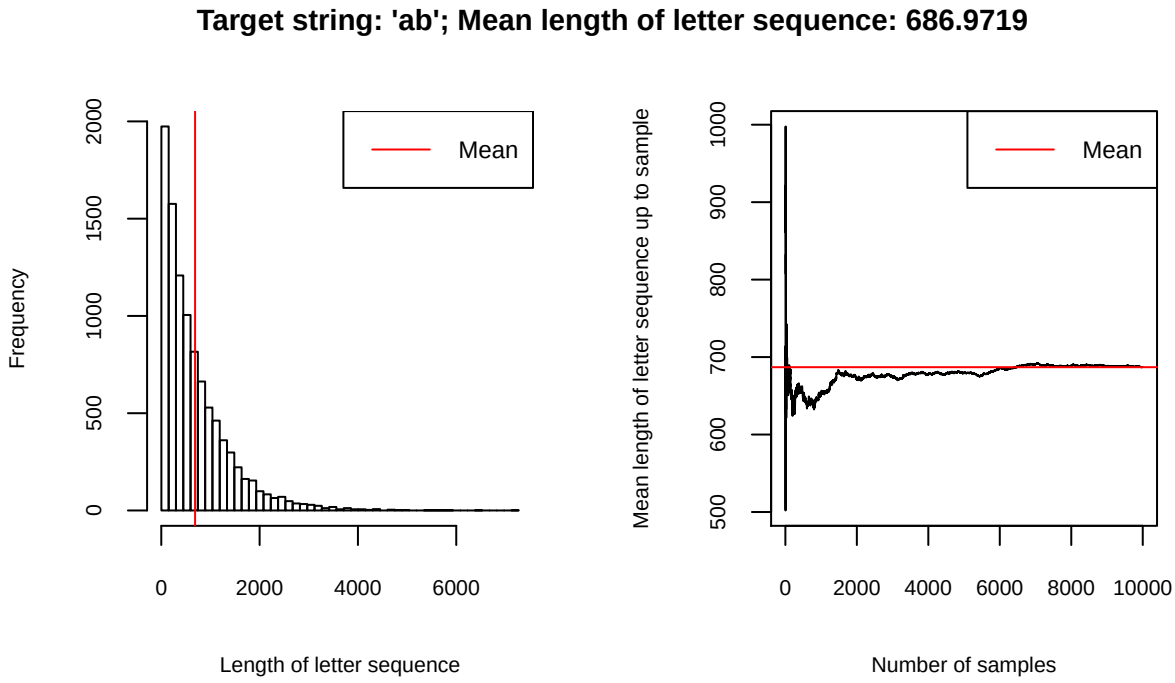

Figure 1: Histogram of letter sequence lengths and mean letter sequence length by number of samples.

2 Monte-Carlo-Approximationen

3 Eine (naive) Datenanalyse

4 Normalverteilung