Endliche Markovketten
Konvergenzsatz, alternativer Abstand und Mischzeiten

Vortrag III im Bachelorseminar zur Wahrscheinlichkeitstheorie bei Prof. Marcel Ortgiese und Fabian Buckmann

Daniel Goseberg

6. Mai 2014

Der erste Teil des Vortrags wird sich mit dem Konvergenzsatz für Markovketten mit endlichem Zustandsraum beschäftigen, dessen hier aufgeführte Form auf D. Aldous und P. Diaconis (1986) zurückgeht. Im weiteren Verlauf untersuchen wir das Verhalten des Abstands einer Markovkette zu ihrer stationären Verteilung noch genauer und führen den wichtigen Begriff der Mischzeit ein.

Dabei orientiert sich dieser Vortrag in der Wahl der Notation, sowie auch des Inhaltes an dem exzellenten Werk [1] von Levin, Peres und Wilmer. Wird von dort ein Resultat zitiert, so findet es der Leser der Vollständigkeit halber ebenfalls im Anhang. Im gesamten Vortrag hat jede Markovkette $X = (X_0, X_1, \dots)$ einen endlichen Zustandsraum Ω.

Um überhaupt über die Konvergenz von Markovketten gegen ihre stationäre Verteilung sprechen zu können, benötigt man einen geeigneten Abstandsbegriff. Zur Erinnerung: Für zwei Wahrscheinlichkeitsmaße μ und ν auf einer endlichen Menge Ω wird der Abstand in Totalvariation definiert durch

$$\|\mu - \nu\|_{TV} := \max_{A \subset \Omega} |\mu(A) - \nu(A)| .$$

Wie bereits im zweiten Vortrag festgehalten wurde, wird damit die Menge aller Wahrscheinlichkeitsmaße auf Ω zu einem metrischen Raum mit entsprechendem Konvergenzbegriff.

1. Konvergenzsatz

Variieren wir die Startverteilung eine irreduzible Markovkette mit Übergangsmatrix P, gelte also beispielsweise $P\{X_0 = \cdot\} = \delta_x$ für ein $x \in \Omega$ und schauen uns die Verteilung

nach t Schritten $P\{X_t = \cdot\}$ an., so sind wir an der Frage interessiert, wie weit diese schlimmstenfalls von der stationären Verteilung entfernt ist. Dies führt uns zu folgender Definition:

Definition 1.1 (Abstand zur stationären Verteilung)**.** *Ist $X = (X_0, X_1, \dots)$ eine irreduzible Markovkette mit Übergangsmatrix P und stationärer Verteilung π, so wird der Abstand zur stationären Verteilung, im Folgenden auch Abstandsfunktion genannt, definiert durch:*

$$d(t) := \max_{x\in\Omega} \left\| P^t(x,\cdot) - \pi \right\|_{TV}. \tag{1}$$

Wann und in welchem Sinne konvergiert eine Markovkette gegen ihre stationäre Verteilung? Die wesentliche Aussage des weiter unten folgenden Konvergenzsatzes besagt $d(t) \to 0$ für $t \to \infty$, falls X irreduzibel und aperiodisch ist. Dass beide Vorraussetzungen wirklich notwendig sind kann man sich schnell klarmachen. Betrachtet man eine irreduzible periodische Markovkette mit Periode d (jeder Zustand hat dann Periode d) so bemerkt man, dass diese bei Start in einer Dirac-Verteilung immer genau nach d Schritten zu dieser zurückkehrt und die Markovkette somit nicht konvergieren kann.

Zunächst brauchen wir aber noch ein vorbereitendes Resultat:

Satz 1.2. *Sei X eine eine Markovkette mit Übergangsmatrix P. Falls X irreduzibel und aperiodisch ist, existiert ein $r \in \mathbb{N}$ mit $P^r(x,y) > 0$ für alle $x, y \in \Omega$.*

Beweis. Wir benutzen folgendes Ergebnis der Zahlentheorie: Eine Menge M von natürlichen Zahlen, die abgeschlossen ist bezüglich der Addition und mit $ggT(M) = 1$, enthält fast alle, d.h. alle bis auf endlich viele natürlichen Zahlen.

Für $x \in \Omega$ sei $\mathcal{T}(x) = \{t \geq 1 | P^t(x,x) > 0\}$ die Menge der Zeitprunkte an denen man bei Startpunkt x mit positiver Wahrscheinlichkeit zu x zurückkehren kann. Da P aperiodisch ist, gilt $ggT(\mathcal{T}(x)) = 1$.

Durch eine simple Anwendung des Matrixproduktes notieren wir, dass $\mathcal{T}(x)$ abgeschlossen unter Addition ist: $P^{s+t}(x,x) \geq P^s(x,x)P^t(x,x) > 0$ für $s, t \in \mathcal{T}(x)$. Also existiert ein $t(x)$, sodass für alle $t \geq t(x)$ gilt: $t \in \mathcal{T}(x)$.

Da X irreduzibel, existiert für $y \in \Omega$ ein $r(x,y)$ mit $P^r(x,y) > 0$, also folgt für alle $t \geq t(x) + r(x,y)$:

$$P^t(x,y) \geq P^{t-r}(x,x)P^r(x,y) > 0.$$

Für $t \geq \hat{t}(x) := t(x) + \max_{y\in\Omega} r(x,y)$ gilt $P^t(x,y) > 0$ für alle $y \in \Omega$.

Sind $x, y \in \Omega$ beliebig, so folgt $P^t(x,y) > 0$ für alle $t \geq \max_{x\in\Omega} \hat{t}(x)$. □

Vorab noch ein Lemma, dessen Aussagen uns im Beweis des Konvergenzsatzes von Nutzen sind:

Lemma 1.3. *Sei π eine Verteilung auf Ω und Π die $|\Omega| \times |\Omega|$ Matrix mit identischen Zeilen π. Dann gilt für $|\Omega| \times |\Omega|$ Matrizen M, M':*

(i) Ist M stochastisch, d. h. : $\sum_y M(x,y) = 1$ für alle $x \in \Omega$, so ist $M\Pi = \Pi$.
(ii) Für jede Matrix M mit $\pi M = \pi$ gilt $\Pi M = \Pi$.

(iii) Sind M, M' stochastisch, so auch MM'.

Beweis. Nachrechnen. □

Satz 1.4 (Konvergenzsatz)**.** *Sei P die Übergangsmatrix einer irreduziblen und aperiodischen Markovkette $X = (X_0, X_1, \dots)$ mit stationärer Verteilung π. Dann existieren $\alpha \in (0,1)$ und $C > 0$, sodass:*

$$\max_{x \in \Omega} ||P^t(x, \cdot) - \pi||_{TV} \leq C\alpha^t. \tag{2}$$

Beweis. Da P irreduzibel und aperiodisch ist, existiert nach [1, Satz 1.7] ein $r > 0$, sodass $P^r(x,y) > 0$ für alle $x, y \in \Omega$.
Sei Π diejenige $|\Omega| \times |\Omega|$ Matrix mit identischen Zeilen π. Wir wählen ein $\theta \in (0,1)$ hinreichend klein mit

$$P^r(x,y) \geq \delta\pi(y). \tag{3}$$

für alle $x, y \in \Omega$. Wir setzten $\theta := 1 - \delta$. Dann wird durch $Q := \theta^{-1}P^r + (1 - \theta^{-1})\,\Pi$ wieder eine stochastische Matrix definiert mit:

$$P^r = (1-\theta)\,\Pi + \theta Q. \tag{IA}$$

Warum ist Q stochastisch? Wegen (3) sind alle Einträge positiv und die Normiertheit folgt sofort aus der Definition.

Mittels Induktion können wir nun diese Gleichung für $k \in \mathbb{N}$ verallgemeinern:

$$P^{rk} = \left(1 - \theta^k\right)\Pi + \theta^k Q^k. \tag{$\star$}$$

Der Induktionsanfang gilt nach (IA). Angenommen ($\star$) gelte für $k = n$ so gilt:

$$P^{r(n+1)} = P^{rn}P^r = \left[(1-\theta^n)\,\Pi + \theta^n Q^n\right] P^r$$

Einsetzen von (IA) im zweiten Summanden liefert nun:

$$\begin{aligned} P^{r(n+1)} &= (1-\theta^n)\,\Pi P^r + (1-\theta)\,\theta^n Q^n \Pi + \theta^{n+1}Q^{n+1} \\ &= \left(1 - \theta^{n+1}\right)\Pi + \theta^{n+1}Q^{n+1}. \end{aligned}$$

Für die letzte Gleichheit benötigten wir Lemma 1.3 (i) und (ii): Damit ist ($\star$) gezeigt.

Sei $j \in \mathbb{N}$. Nach einer Umsortierung erhalten wir durch Multiplikation mit P^j von rechts:

$$P^{rk+j} - \Pi = \theta^k\left(Q^kP^j - \Pi\right) \qquad (j \in \mathbb{N}). \tag{4}$$

Nun sind wir fast fertig. Sei $x \in \Omega$ fixiert. Schauen wir uns den Abtand in Totalvariation von $P^{rk+j}(x,\cdot)$ und π an, können wir dazu die Charakterisierung aus [1, Satz 4.2] verwenden. Mit Lemma 1.3 (iii) erkennen wir $P^kQ^j(x,\cdot)$ zudem wieder als eine Wahrscheinlichkeitsverteilung auf Ω:

$$\begin{aligned} \left\|P^{rk+j}(x,\cdot) - \pi\right\|_{TV} &= \frac{1}{2}\sum_y \left|P^{rk+j}(x,y) - \pi(y)\right| \\ &\overset{(4)}{=} \theta^k\left(\frac{1}{2}\sum_y \left|Q^kP^j(x,y) - \pi(y)\right|\right) \\ &= \theta^k \left\|Q^kP^j(x,\cdot) - \pi\right\|_{TV} \leq \theta^k, \end{aligned} \tag{$\star$}$$

wobei letztere Abschätzung daraus folgt, dass der Totalvariationsabstand nach oben durch 1 beschränkt ist.

Wie folgt daraus die Behauptung? Sei $t \in \mathbb{N}$ und o.E. $t > r$. Wir verwenden Division mit Rest durch r und schätzen dann mittels $(\star)$ ab. Es existieren $k \in \mathbb{N}, l \in \{0, 1, \ldots, r-1\}$, sodass $t = rk + l$. Dann

$$\left\|P^t(x,\cdot) - \pi\right\|_{TV} = \left\|P^{rk+m}(x,\cdot) - \pi\right\|_{TV} \overset{(\star)}{\leq} \theta^k = \left(\theta^{\frac{1}{r}}\right)^t \left(\frac{1}{\theta}\right)^{\frac{m}{r}} \overset{(m<r)}{<} \alpha^t C.$$

Wobei wir α als $\theta^{(r^{-1})} \in (0,1)$ und C als θ^{-1} setzen. □

Als kleine Anmerkung sei noch gesagt, dass man den Konvergenzsatz auch mittels der Kopplungsmethode beweisen kann [1, Aufg. 5.1].

2. Abstand zur stationären Verteilung

Nachdem wir nun den zentralen Konvergenzsatz für Markovketten mit endlichem Zustandsraum bewiesen haben, liegt es natürlich nahe, das Konvergenzverhalten von Markovketten genauer zu betrachten. Dazu schauen wir uns die Abstandsfunktion (1) zur stationären Verteilung genauer an. Sei P die Übergangsmatrix einer irreduziblen Markovkette und π ihre stationäre Verteilung. Wir setzen:

$$\bar{d}(t) := \max_{x,y\in\Omega} \left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{TV}. \tag{5}$$

Unser erstes Anliegen wird es sein, gewisse Eigenschaften der Abstände $d(t)$ und $\bar{d}(t)$ herzuleiten. Dabei wird uns das folgende Lemma helfen:

Lemma 2.1. *Sei P die Übergangsmatrix einer Markovkette mit Zustandsraum Ω. Seien μ und ν Verteilungen auf Ω. Dann gilt:*

$$\|\mu P - \nu P\|_{TV} \leq \|\mu - \nu\|_{TV}, \tag{6}$$

also induktiv auch $\|\mu P^{t+1} - \nu P^{t+1}\|_{TV} \leq \|\mu P^t - \nu P^t\|_{TV}$,
sowie $\|\mu P^{t+1} - \pi\|_{TV} \leq \|\mu P^t - \pi\|_{TV}$, falls eine stationäre Verteilung existiert.

Beweis. Nachrechnen. □

Der folgende Satz fasst nun einige der wichtigsten Eigenschaften der Abstandsfunktionen zusammen.

Satz 2.2. *Sei P die Übergangsmatrix einer irreduziblen Markovkette mit stationärer Verteilung π. Dann gilt:*
(i) $d(t)$ und $\bar{d}(t)$ sind monoton fallend.
(ii) $d(t) \leq \bar{d}(t) \leq 2d(t)$.

(iii) Die Funktion $\bar{d}$ ist submultiplikativ:

$$\bar{d}(s+t) \leq \bar{d}(s)\bar{d}(t), \tag{7}$$

für alle $s, t \in \mathbb{N}$.

Beweis. (i) Dies folgt unmittelbar aus Lemma 2.1, wenn man dortige Aussage mit $\mu = \delta_x$ und $\nu = \delta_y$ für $x, y \in \Omega$ liest.

(ii) Aus der Dreiecksungleichung für den Totalvariationsabstand folgt $\bar{d}(t) \leq 2d(t)$. Um $d(t) \leq \bar{d}(t)$ zu zeigen, bemerken wir zunächst, dass $\pi(A) = \sum_y \pi(y)P^t(y, A)$. Mithilfe dieser Eigenschaft und der Dreiecksungleichung erhalten wir:

$$\begin{aligned}
\left\|P^t(x,\cdot) - \pi\right\|_{TV} &= \max_{A\subset\Omega} \left|P^t(x,A) - \pi(A)\right| \\
&= \max_{A\subset\Omega} \left|\sum_y \pi(y)\left[P^t(x,A) - P^t(y,A)\right]\right| \\
&\leq \sum_y \pi(y) \max_{A\subset\Omega} \left|P^t(x,A) - P^t(y,A)\right| \\
&= \sum_y \pi(y) \left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{TV}.
\end{aligned}$$

Zunächst können wir den hinteren Faktor innerhalb der Summe nach oben durch $\max_{y\in\Omega} \|P^t(x,\cdot) - P^t(y,\cdot)\|_{TV}$ abschätzen. Da $\sum_y \pi(y) = 1$, folgt Teil *(ii)* durch maximieren über x.

(iii) Seien also $s, t \in \mathbb{N}$.. Wir fixieren $x, y \in \Omega$ und nehmen eine optimale Kopplung (X_s, Y_s) von $P^s(x,\cdot)$ und $P^s(y,\cdot)$ her (siehe [1, Satz 4.7]). Also

$$\|P^s(x,\cdot) - P^s(y,\cdot)\|_{TV} = \mathbf{P}\left\{X_s \neq Y_s\right\}. \tag{8}$$

Beachten wir nun, dass P^{s+t} das Matrixprodukt von P^t und P^s ist und $P^s(x,\cdot)$ die Verteilung von X_s, so erhalten wir:

$$\begin{aligned}
P^{s+t}(x,w) &= \sum_z P^s(x,z)P^t(z,w) \\
&= \sum_z \mathbf{P}\left\{X_s = z\right\} P^t(z,w) = \mathbf{E}\left[P^t(X_s,w)\right].
\end{aligned}$$

Mit einer analogen Rechnung erhält man $P^{s+t}(y,w) = \mathbf{E}\left[P^t(Y_s,w)\right]$. Also, da X_s und Y_s auf demselben Wahrscheinlichkeitsraum definiert sind:

$$\begin{aligned}
P^{s+t}(x,w) - P^{s+t}(y,w) &= \mathbf{E}\left[P^t(X_s,w)\right] - \mathbf{E}\left[P^t(Y_s,w)\right] \\
&= \mathbf{E}\left[P^t(X_s,w) - P^t(Y_s,w)\right]
\end{aligned} \tag{9}$$

Betrachten wir den Totalvariationsabstand der Verteilungen $P^{s+t}(x,\cdot)$ und $P^{s+t}(y,\cdot)$, so können wir uns nun (9) zu Nutzen machen indem wir dort über alle $w \in \Omega$ summieren und Charakterisierung [1, Satz 4.2] verwenden:

$$\left\|P^{s+t}(x,\cdot) - P^{s+t}(y,\cdot)\right\|_{TV} = \frac{1}{2}\sum_w \left|\mathbf{E}\left[P^t(X_s,w) - P^t(Y_s,w)\right]\right|.$$

Der rechte Ausdruck lässt sich nun nach oben durch

$$\mathbf{E}\left[\frac{1}{2}\sum_{w}\left|P^t(X_s,w)-P^t(Y_s,w)\right|\right]=\mathbf{E}\left[\left\|P^t(X_s,\cdot)-P^t(Y_s,\cdot)\right\|_{TV}\right]$$

abschätzen.
Schauen wir uns jetzt einmal die Distanz $\|P^t(X_s,\cdot)-P^t(Y_s,\cdot)\|_{TV}$ genauer an. Zuerst einmal ist diese nach oben durch $\bar{d}(t)$ beschränkt. Da dieser Ausdruck außerdem auf $\{X_s=Y_s\}$ verschwindet erhalten wir insgesamt:

$$\begin{aligned}\left\|P^{s+t}(x,\cdot)-P^{s+t}(y,\cdot)\right\|_{TV} &\leq \mathbf{E}\left[\left\|P^t(X_s,\cdot)-P^t(Y_s,\cdot)\right\|_{TV}\right]\\ &= \mathbf{E}\left[\left\|P^t(X_s,\cdot)-P^t(Y_s,\cdot)\right\|_{TV}\mathbf{1}_{\{X_s\neq Y_s\}}\right]\\ &\leq \bar{d}(t)\mathbf{P}\{X_s\neq Y_s\}\end{aligned}$$

und damit zusammen mit (8) die Behauptung.

□

Mit Hilfe von Satz 2.2 (ii) und (iii) folgt für beliebige positive Zahlen c und t:

$$d(ct)\leq\bar{d}(ct)\leq\bar{d}(t)^c. \tag{10}$$

3. Mischzeiten

An dieser Stelle wird mit der Mischzeit der Zeitpunkt definiert, ab der Abstand einer irreduziblen Markovkette zu ihrer stationären Verteilung kleiner als eine vorgegebene Schranke ε wird.

Dies ist interessant für einige Beispiele von Markovketten, zum Beispiel bei Betrachtung des Kartenmischens als Markovkette mit Zustandsraum $\mathcal{S}_{52}$ und stationärer Verteilung $\pi\equiv\frac{1}{52!}$.

Definition 3.1 (Mischzeit)**.** *Sei P die Übergangsmatrix einer irreduziblen und aperiodischen Markovkette mit stationärer Verteilung π. Dann definieren wir die Mischzeit wie folgt:*

$$t_{mix}(\varepsilon):=\min\{t:d(t)\leq\varepsilon\}$$

und

$$t_{mix}:=t_{mix}\left(\frac{1}{4}\right)$$

Korollar 3.2. *Für alle $l\in\mathbb{N}$ gilt:*

$$d\left(lt_{mix}(\varepsilon)\right)\leq\bar{d}\left(lt_{mix}(\varepsilon)\right)\leq\bar{d}\left(t_{mix}(\varepsilon)\right)^l\leq(2\varepsilon)^l, \tag{11}$$

also insbesondere $d\left(t_{mix}\right)\leq 2^{-l}$. Außerdem gilt:

$$t_{mix}(\varepsilon)\leq\lceil log_2\varepsilon^{-1}\rceil t_{mix}. \tag{12}$$

Beweis. Folgt direkt aus (10), Satz 2.2 (ii) und der Definition der Mischzeit. □

A. Anhang

Mit $\mathbb{N}$ sind im gesamten Vortrag die natürlichen Zahlen ohne 0 gemeint.

Der Vollständigkeit halber sind an dieser Stelle einzelne, für diesen Vortrag wichtige, Resultate zitiert. Der Inhalt der einzelnen Definitionen und Sätze wurde in den vorherigen Vorträgen bereits behandelt. Sie sind der Quelle [1] entnommen und in der Nummerierung selbigem Werke nachempfunden.

A.1. Resultate aus den vorherigen Vorträgen

Satz 4.2 Seien μ und ν zwei Wahrscheinlichkeitsmaße auf Ω. Dann gilt:

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|. \tag{13}$$

für alle $s, t \in \mathbb{N}$.

Satz 4.7 Seien μ und ν zwei Wahrscheinlichkeitsmaße auf Ω. Dann gilt:

$$\|\mu - \nu\|_{TV} = \inf \left\{ \mathbf{P}\left\{X \neq Y\right\} : (X, Y) \text{ ist eine Kopplung von } \mu \text{ und } \nu \right\}. \tag{14}$$

Außerdem existiert eine Kopplung, die die untere Schranke in (A.1) annimmt, wir nennen diese dann optimal.

Literatur

[1] David A. Levin, Yuval Peres, Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2008.