10 Schnell mischende Markov-Ketten

Allgemeines zu schnell mischenden Markov-Ketten findet man zum Beispiel in dem Buch „*Introduction to Markov Chains*“ von Behrends (2000). Außerdem haben wir von einem Teil eines Vorlesungsskriptes von Steger (2001) über schnell mischende Markov-Ketten profitiert. Außerdem ist der Überblick von Randall (2006) empfehlenswert.

10.1 Anmerkungen zu Eigenwerten

10.1 Lemma. Es sei P eine zeilenstochastische Matrix. Dann ist 1 ein Eigenwert von P und für jeden Eigenwert λ von P gilt: $|\lambda| \leqslant 1$.

10.2 Beweis. Der erste Teil der Aussage ist wegen $P(1,\dots,1)^\top = (1,\dots,1)^\top$ klar.

Sei nun $\lambda \in \mathbb{C}$ Eigenwert und $P(x_1,\dots,x_n)^\top = \lambda(x_1,\dots,x_n)^\top$. Es sei i_0 ein Index mit $|x_{i_0}| = \max_j |x_j|$. Es liegen also alle x_j im Kreis um den Ursprung mit Radius $|x_{i_0}|$. Es gilt:

$$|\lambda| \cdot |x_{i_0}| = |\lambda \cdot x_{i_0}| = \left|\sum_j P_{i_0 j} x_j\right| \leqslant \sum_j P_{i_0 j}|x_j| \leqslant \sum_j P_{i_0 j}|x_{i_0}| = |x_{i_0}| \sum_j P_{i_0 j} = |x_{i_0}| \,.$$

Also ist $|\lambda| \leqslant 1$. ∎

Für reversible Markov-Ketten kann man die Aussage von Lemma 10.1 verschärfen.

10.3 Lemma. Ist P die stochastische Matrix einer reversiblen Markov-Kette, dann hat P nur reelle Eigenwerte.

Für einen Beweis konsultiere man z. B. das Vorlesungsskript von Steger (2001) oder Kapitel 3 im Manuskript von Aldous und Fill (1999)

10.4 Im folgenden benutzen wir die Abkürzung $\lambda_{max} = \max\{\, |\lambda| \mid \lambda \text{ ist Eigenwert von } P \text{ und } \lambda \neq 1\}$. Auf Grund des Vorangegangenen ist klar, dass $\lambda_{max} \leqslant 1$ ist.

10.5 Falls $\lambda_N < 0$ ist geht man auf bewährte Weise von $\mathbf{P}$ zur Matrix $\mathbf{P}' = \frac{1}{2}(\mathbf{P} + \mathbf{I})$ über. Für deren Eigenwerte gilt dann $\lambda_i' = \frac{1}{2}(\lambda_i + 1)$, so dass alle Eigenwerte echt größer 0 sind. In diesem Fall ist dann also $\lambda_{max} = \lambda_2$.

10.2 Konvergenzverhalten ergodischer Markov-Ketten

Wir interessieren uns nun für die Frage, wie lange es dauert, bis man bei einer Markov-Kette, die man mit einer Verteilung oder in einem bestimmten Zustand begonnen hat, davon ausgehen

kann, dass die Wahrscheinlichkeiten, in bestimmten Zuständen zu sein, denen der stationären Verteilung „sehr nahe“ sind.

Dazu definieren wir als erstes eine Art Abstandsbegriff für diskrete Wahrscheinlichkeitsverteilungen.

10.6 Definition Für zwei Verteilungen $\mathbf{p}$ und $\mathbf{q}$ ist die *totale Variationsdistanz*

$$\|p - q\|_{tv} = \frac{1}{2} \sum_{j \in S} |\mathbf{p}_j - \mathbf{q}_j| \,.$$

◇

Man kann sich überlegen, dass $\|p - q\|_{tv} = \max_{T \subseteq S} |\mathbf{p}(T) - \mathbf{q}(T)|$ ist, wobei $\mathbf{p}(T)$ zu verstehen ist als $\sum_{j \in T} \mathbf{p}_j$ (und analog $\mathbf{q}(T)$).

10.7 Definition Für eine ergodische Markov-Kette mit Matrix $\mathbf{P}$ und stationärer Verteilung $\mathbf{w}$ und für alle Verteilungen $\mathbf{p}$ sei

$$\delta_{\mathbf{p}}(t) = \|\mathbf{p}\mathbf{P}^t - \mathbf{w}\|_{tv}$$

die totale Variationsdistanz zwischen $\mathbf{w}$ und der Verteilung, die man nach t Schritten ausgehend von $\mathbf{p}$ erreicht hat.

Als *Variationsdistanz zum Zeitpunkt* t bezeichnen wir das Maximum $\Delta(t) = \max_{\mathbf{p}} \delta_{\mathbf{p}}(t)$. ◇

Man kann zeigen, dass das Maximum für einen Einheitsvektor $\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ angenommen wird. Also gilt:

$$\Delta(t) = \max_i \|P_i^t - \mathbf{w}\|_{tv}$$

wobei $\mathbf{P}_i^t$ die i-te Zeile von $\mathbf{P}^t$ sei.

10.8 Satz. *Für eine ergodische Markov-Kette mit Matrix* $\mathbf{P}$ *und stationärer Verteilung* $\mathbf{w}$ *existieren Konstanten* C *und* $\alpha < 1$ *so, dass gilt:*

$$\Delta(t) = \max_i \|\mathbf{P}_i^t - \mathbf{w}\|_{tv} \leqslant C\alpha^t$$

Der nachfolgende Beweis stammt aus dem Buch von Levin, Peres und Wilmer (2009).

10.9 Beweis. Wegen der Ergodizität existiert ein t_0 so, dass in $\mathbf{P}^{t_0}$ nur echt positive Einträge vorkommen. Bezeichnet $\mathbf{w}$ die stationäre Verteilung (und $\mathbf{W}$ die Matrix, deren Zeilen alle gleich $\mathbf{w}$ sind), dann gibt es ein δ mit $0 < \delta < 1$ so, dass für alle $i, j \in S$ gilt:

$$\mathbf{P}_{ij}^{t_0} \geqslant \delta \mathbf{w}_j$$

Es sei $\vartheta = 1 - \delta$, also auch $0 < \vartheta < 1$. Durch

$$\mathbf{P}^{t_0} = (1 - \vartheta)\mathbf{W} + \vartheta \mathbf{Q}$$

wird eine Matrix $\mathbf{Q}$ festgelegt. Q ist eine zeilenstochastische Matrix.

Die Überlegungen in Abschnitt 8.2 zeigen, dass für jede stochastische Matrix $\mathbf{M}$ gilt, dass $\mathbf{MW} = \mathbf{W}$ ist. Zum Beispiel gilt für jedes $k \in \mathbb{N}_0$: (i) $\mathbf{Q}^k\mathbf{W} = \mathbf{W}$. Außerdem ist (ii) $\mathbf{W}\mathbf{P}^{t_0} = \mathbf{W}$.

Durch Induktion zeigt man nun: Für jedes $k \in \mathbb{N}_+$ ist

$$\mathbf{P}^{t_0 k} = (1 - \vartheta^k)\mathbf{W} + \vartheta^k \mathbf{Q}^k$$

Der Induktionsanfang ist klar. Für den Induktionsschritt rechnet man

$$\begin{aligned}
\mathbf{P}^{t_0(k+1)} &= \mathbf{P}^{t_0 k}\mathbf{P}^{t_0} \\
&= \left((1-\vartheta^k)\mathbf{W} + \vartheta^k\mathbf{Q}^k\right)\mathbf{P}^{t_0} && \text{nach Induktionsvoraussetzung} \\
&= (1-\vartheta^k)\mathbf{W}\mathbf{P}^{t_0} + \vartheta^k\mathbf{Q}^k\mathbf{P}^{t_0} \\
&= (1-\vartheta^k)\mathbf{W} + \vartheta^k\mathbf{Q}^k\mathbf{P}^{t_0} && \text{nach Bemerkung (ii) oben} \\
&= (1-\vartheta^k)\mathbf{W} + \vartheta^k\mathbf{Q}^k\left((1-\vartheta)\mathbf{W} + \vartheta\mathbf{Q}\right) \\
&= (1-\vartheta^k)\mathbf{W} + \vartheta^k(1-\vartheta)\mathbf{Q}^k\mathbf{W} + \vartheta^{k+1}\mathbf{Q}^{k+1} \\
&= (1-\vartheta^k)\mathbf{W} + \vartheta^k(1-\vartheta)\mathbf{W} + \vartheta^{k+1}\mathbf{Q}^{k+1} && \text{nach Bemerkung (i) oben} \\
&= (1-\vartheta^{k+1})\mathbf{W} + \vartheta^{k+1}\mathbf{Q}^{k+1}
\end{aligned}$$

Die Tatsache, dass $\mathbf{P}^{t_0 k} = (1-\vartheta^k)\mathbf{W} + \vartheta^k\mathbf{Q}^k$ ist, nutzt man wie folgt. Multiplikation mit $\mathbf{P}^j$ und Umordnen liefert

$$\mathbf{P}^{t_0 k+j} - \mathbf{W} = \vartheta^k\left(\mathbf{Q}^k\mathbf{P}^j - \mathbf{W}\right)$$

Man betrachtet nun eine beliebige Zeile i dieser Matrizengleichung, summiert die Beträge der Einträge in dieser Zeile und dividiert durch 2. Auf der linken Seite ergibt sich $\|\mathbf{P}_i^{t_0 k+j} - \mathbf{w}\|_{tv}$. Auf der rechten Seite ergibt sich $\vartheta^k\|(\mathbf{Q}^k\mathbf{P}^j)_i - \mathbf{w}\|_{tv}$, was man durch ϑ^k nach oben abschätzen kann.

Für beliebiges $t \in \mathbb{N}_+$ sei nun $k = t \operatorname{div} t_0$ und $j = t \bmod t_0$ (also $t = t_0 k + j$). Dann ist

$$\begin{aligned}
\|\mathbf{P}_i^t - \mathbf{w}\|_{tv} &= \|\mathbf{P}_i^{t_0 k+j} - \mathbf{w}\|_{tv} \\
&\leqslant \vartheta^k \\
&= \frac{1}{\vartheta}\left(\vartheta^{1/t_0}\right)^{t_0}\left(\vartheta^{1/t_0}\right)^{t_0 k} \\
&\leqslant \frac{1}{\vartheta}\left(\vartheta^{1/t_0}\right)^{j+t_0 k} && \text{da } j < t_0 \text{ und } \vartheta < 1 \\
&= \frac{1}{\vartheta}\left(\vartheta^{1/t_0}\right)^{t}
\end{aligned}$$

■

10.10 Im folgenden benutzen wir die Abkürzung

$$w_{min} = \min_j \mathbf{w}_j$$

10.11 Satz. *Für jede reversible Markov-Kette mit stationärer Verteilung* $\mathbf{w}$ *gilt:*

$$\Delta(t) \leqslant \frac{\lambda_{max}^t}{w_{min}} .$$

Wenn also $\lambda_{max} < 1$ ist (was bei reversiblen Markov-Ketten der Fall ist), dann nähert sich die Markov-Kette in einem gewissen Sinne „schnell“ der stationären Verteilung. Wir präzisieren das noch wie folgt:

10.3 Schnell mischende Markov-Ketten

10.12 Häufig ist man an einer ganzen Familie von Markov-Ketten $M(I)$ interessiert. Dabei ergibt sich jedes $M(I)$ auf Grund einer Instanz I des eigentlich zu lösenden Problems. Zum Beispiel könnte jedes I ein Graph sein, und die Zustände von $M(I)$ sind gerade die Matchings von I.

10.13 Definition Es sei M eine ergodische Markov-Kette mit stationärer Verteilung $\mathbf{w}$. Für $\varepsilon > 0$ sei

$$\tau(\varepsilon) = \min\{t \mid \forall t' \geqslant t : \Delta(t') \leqslant \varepsilon\}$$

die *ε-Konvergenzzeit* der Markov-Kette M. ◇

10.14 Definition Eine Familie von Markov-Ketten $M(I)$ heißt *schnell mischend*, falls die ε-Konvergenzzeit polynomiell in $|I|$ und $\ln 1/\varepsilon$ ist. ◇

10.15 Wegen Satz 10.11 ist eine reversible Markov-Kette jedenfalls dann schnell mischend, wenn für ein t, das polynomiell in $|I|$ und $\log 1/\varepsilon$ ist, gilt:

$$\frac{\lambda_{max}^t}{w_{min}} \leqslant \varepsilon\ .$$

Äquivalente Umformungen ergeben

$$\begin{aligned} \lambda_{max}^t &\leqslant \varepsilon w_{min} \\ \left(\frac{1}{\lambda_{max}}\right)^t &\geqslant \frac{1}{\varepsilon w_{min}} \\ t &\geqslant \frac{\ln \varepsilon^{-1} + \ln w_{min}^{-1}}{\ln \lambda_{max}^{-1}} \end{aligned}$$

Wegen $1 - x \leqslant \ln x^{-1}$ für $0 < x < 1$ ist das jedenfalls dann der Fall, wenn

$$t \geqslant \frac{\ln \varepsilon^{-1} + \ln w_{min}^{-1}}{1 - \lambda_{max}}$$

Schnelles Mischen liegt also jedenfalls dann vor, wenn $\ln w_{min}^{-1}$ und $1/(1-\lambda_{max})$ polynomiell in $|I|$ sind.

Damit haben wir zumindest eine Hälfte des folgenden Satzes bewiesen, der obere und untere Schranken für $\tau(\varepsilon)$ angibt:

10.16 Satz.

$$\begin{aligned} \tau(\varepsilon) &\leqslant \frac{1}{1-\lambda_{max}} \log \frac{1}{w_{min}\varepsilon} \\ \tau(\varepsilon) &\geqslant \frac{1}{2(1-\lambda_{max})} \log \frac{1}{2\varepsilon} \end{aligned}$$

Es stellt sich die Frage, woher man (zumindest näherungsweise) Kenntnis von λ_{max} bzw. im Falle von reversiblen Markov-Ketten von λ_2 bekommen kann. Eine Möglichkeit ist der sogenannte Leitwert einer Markov-Kette. Wir werden ihn mit Hilfe gewichteter Graphen einführen, die später selbst noch nützlich sein werden.

10.17 Definition Für eine reversible Markov-Kette $M = (S, \mathbf{P})$ mit stationärer Verteilung $\mathbf{w}$ sei F_M der gerichtete gewichtete Graph mit Knotenmenge S und Kantenmenge $E_F = \{(i,j) \mid i \neq j \wedge P_{ij} > 0\}$. Jede Kante (i,j) ist gewichtet mit der reellen Zahl $c(i,j) = \mathbf{w}_i P_{ij}$. ◇

10.18 Definition Für eine reversible Markov-Kette mit Zustandsmenge S und stationärer Verteilung $\mathbf{w}$ definieren wir für jede Teilmenge $T \subseteq S$

$$\begin{aligned} \text{die Kapazität} \qquad C(T) &= \sum_{i \in T} \mathbf{w}_i \\ \text{den Fluß} \qquad F(T) &= \sum_{i \in T, j \notin T} \mathbf{w}_i \mathbf{P}_{ij} \\ \text{und} \qquad \Phi(T) &= F(T)/C(T) \end{aligned}$$

Der *Leitwert* Φ der Markov-Kette ist dann

$$\Phi = \min_{T \subseteq S} \max\{\Phi(T), \Phi(S \setminus T)\}\,.$$

◇

Eine kurze Überlegung zeigt, dass $\Phi(T)$ die bedingte Wahrscheinlichkeit ist, dass man bei der Markov-Kette mit stationärer Verteilung einen Übergang von innerhalb von T nach außerhalb von T beobachtet. Wenn $\Phi(T)$ klein ist, dann ist T sozusagen eine Art „Falle“, aus der die Markov-Kette schlecht heraus kommt.

Man kann nun mit einigem technischen Aufwand zeigen:

10.19 Satz. *Für jede reversible Markov-Kette gilt:*

$$1 - 2\Phi^2 \leqslant \lambda_2 \leqslant 1 - \frac{\Phi^2}{2}\,.$$

Zusammen mit Punkt 10.15 ergibt sich:

10.20 Korollar. Für reversible Markov-Ketten (mit $\lambda_2 = \lambda_{max}$) ist

$$\tau(\varepsilon) \leqslant \frac{2}{\Phi^2}(\ln \varepsilon^{-1} + \ln w_{min}^{-1})$$

Man kennt mehrere Methoden, um den Leitwert jedenfalls mancher Markov-Ketten zu berechnen.

Die folgende Definition ist eine Art graphentheoretische Version des Leitwertes:

10.21 Definition Für einen ungerichteten Graphen (V, E) ist die *Kantenvervielfachung* μ das Minimum der Zahlen

$$\frac{|\{(i,j) \mid i \in T \wedge j \notin T \wedge (i,j) \in E\}|}{|T|}$$

wobei über alle Teilmengen $T \subseteq V$ minimiert werde mit $|T| \leqslant |V|/2$. ◇

10.22 Man kann sich überlegen, dass für die Markov-Ketten $M_{G,\beta}$ aus Definition 9.2 gilt: $\Phi = \beta\mu/d$.

Damit ist man bei der Aufgabe gelandet, die Kantenvervielfachung von Graphen zu bestimmen. Das kann man zum Beispiel mit Hilfe der Methode der sogenannten kanonischen Pfade von Sinclair machen. Die Verallgemeinerung für beliebige reversible Markov-Ketten betrachtet Mehrgüterflüsse.

10.23 Definition Für jedes Paar (i,j) von Knoten in F_M soll von einem „Gut“ g_{ij} die Menge $\mathbf{w}_i\mathbf{w}_j$ von i nach j transportiert werden. Dazu werden Flüsse $f_{ij} : E_F \to \mathbb{R}_+$ gesucht, so dass die folgenden naheliegenden Forderungen erfüllt sind:

$$\begin{aligned} \sum_k f_{ij}(i,k) &= \mathbf{w}_i\mathbf{w}_j \\ \text{für alle } l \neq i,j: \sum_k f_{ij}(k,l) &= \sum_m f_{ij}(l,m) \\ \sum_k f_{ij}(k,j) &= \mathbf{w}_i\mathbf{w}_j \end{aligned}$$

Der Gesamtfluss durch eine Kante e sei

$$f(e) = \sum_{i\neq j} f_{ij}(e)$$

und die *relative Kantenauslastung*

$$\rho(f) = \max_{e\in E_F} f(e)/c(e) \; .$$ ◇

Dann gilt die folgende Aussage, die wir hier nicht beweisen:

10.24 Lemma. Für jede Markov-Kette mit Flüssen f_{ij} gilt:

$$\Phi \geqslant \frac{1}{2\rho(f)} \; .$$

Um auf einen großen Leitwert schließen zu können, muss man daher versuchen, Flüsse mit kleiner (i. e. polynomieller) relativer Kantenauslastung zu finden.

Wir wollen dies nun auf Random Walks im Hyperwürfel anwenden.

10.25 Beispiel. Dazu sei eine Dimensionalität n beliebig aber fest gewählt und M die Markov-Kette, die sich gemäß Definition 9.2 für $\beta = 1/2$ aus dem n-dimensionalen Hyperwürfel H_n als zu Grunde liegenden Graphen ergibt. Das n sei per definitionem die „Größe“ der Probleminstanz.

M ist reversibel gemäß Punkt 9.3. Da in H_n jeder Knoten Grad n hat, sind die Übergangswahrscheinlichkeiten also $P_{ii} = 1/2$ und $P_{ij} = 1/2n$ für $i \neq j$. Aus Symmetriegründen ist klar, dass die stationäre Verteilung die Gleichverteilung ist mit $\mathbf{w}_i = 1/2^n$; damit ist natürlich auch $w_{min} = 1/2^n$.

Offensichtlich ist $\ln 1/w_{min} \in \Theta(n)$ polynomiell in n. Um einzusehen, dass M schnell mischend ist, genügt es folglich wegen Lemma 10.24, Flüsse f_{ij} zu finden, so dass $\rho(f)$ polynomiell (in n) ist.

Dazu gehen wir wie folgt vor. Jeder Fluss f_{ij} muss gerade die „Menge“ $1/2^{2n}$ transportieren. Sie wird wie folgt verteilt: Zwischen i und j gibt es $d!$ kürzeste Pfade, wobei d die Hammingdistanz zwischen i und j ist. Auf jedem dieser Pfade transportieren wir den gleichen Anteil der Gesamtmenge.

Die Bestimmung der relativen Kantenauslastung wird dadurch erleichtert, dass aus Symmetriegründen auf jeder Kante der gleiche Gesamtfluss vorliegt.

Für jedes d gibt es $2^n \cdot \binom{n}{d}$ Paare (i,j) mit Abstand d. Für ein *festes* Paar (i,j) haben alle für den Fluss f_{ij} verwendeten Pfade Länge d und es ist folglich

$$\sum_{e\in E_F} f_{ij}(e) = d\mathbf{w}_i\mathbf{w}_j \; .$$

Also ist

$$\begin{aligned} f(e) &= \frac{1}{|E_F|} \sum_{d=1}^{n} 2^n \binom{n}{d} \cdot d \cdot w_i w_j \\ &= \frac{1}{n2^n} \sum_{d=1}^{n} 2^n \binom{n}{d} \cdot d \cdot \frac{1}{2^{2n}} \\ &= \frac{1}{2^{2n}} \sum_{d=1}^{n} \binom{n}{d} \cdot \frac{d}{n} \\ &= \frac{1}{2^{2n}} \sum_{d=1}^{n} \binom{n-1}{d-1} \\ &= \frac{1}{2^{2n}} \sum_{d=0}^{n-1} \binom{n-1}{d} \\ &= \frac{2^{n-1}}{2^{2n}} = \frac{1}{2 \cdot 2^n} \end{aligned}$$

Andererseits ist für alle Kanten $c(e) = w_i P_{ij} = 1/(2n \cdot 2^n)$ und somit

$$\rho(f) = \frac{1/(2 \cdot 2^n)}{1/(2n \cdot 2^n)} = n \,.$$

Wegen Korollar 10.20 sind diese Markov-Ketten also schnell mischend.

Man mache sich noch einmal klar, was das bedeutet (siehe Definition 10.14): Es sei $\mathbf{p} = \mathbf{e}_0$ die Anfangs-„Verteilung“ bei der man sicher im Knoten $(0, 0, \ldots, 0)$ des n-dimensionalen Hyperwürfels startet. Es sei $\mathbf{p}_t$ die nach t Schritten erreichte Wahrscheinlichkeitsverteilung. Für jedes $\varepsilon = 2^{-k}$ ist dann die ε-Konvergenzzeit, also die Anzahl Schritte nach der $\|p_t - w\|_{tv} < \varepsilon$ immer gilt, polynomiell in $\ln 1/\varepsilon = k$ und n. Und das, obwohl der Hyperwürfel 2^n Knoten hat!

Zusammenfassung

Oft hat man es mit ergodischen Markov-Ketten zu tun, die sogar reversibel sind. In diesem Fall gibt es Kriterien, um festzustellen, ob sie schnell mischend sind.

Literatur

Aldous, David und James Allen Fill (1999). „Reversible Markov Chains and Random Walks on Graphs“. In: URL: http://www.stat.berkeley.edu/~aldous/RWG/book.html (siehe S. 81).

Behrends, Ehrhard (2000). *Introduction to Markov Chains*. Advanced Lectures in Mathematics. Vieweg (siehe S. 81).

Levin, Davind A., Yuval Peres und Elizabeth L. Wilmer (2009). *Markov Chains and Mixing Times*. AMS (siehe S. 82).

Randall, Dana (2006). „Rapidly Mixing Markov Chains with Applications in Computer Science and Physics“. In: *Computing in Science and Engineering* 8.2, S. 30–41 (siehe S. 81).

Steger, Angelika (2001). „Schnell mischende Markov-Ketten“. In: URL: http://wwwmayr.informatik.tu-muenchen.de/lehre/2001SS/ra/slides/markov-skript.ps (siehe S. 81).

