Randomisierte Algorithmen

9. Schnell mischende Markov-Ketten

Thomas Worsch

Fakultät für Informatik
Karlsruher Institut für Technologie

Wintersemester 2016/2017

Überblick

Der Metropolis-Algorithmus

Anmerkungen zu Eigenwerten

Schnell mischende Markov-Ketten

Ein schönes Buch

David A. Levin, Yuval Peres, Elizabeth L. Wilmer:
Markov Chains and Mixing Times
AMS (2008)
http://darkwing.uoregon.edu/~dlevin/MARKOV/markovmixing.pdf

Überblick

Der Metropolis-Algorithmus

Anmerkungen zu Eigenwerten

Schnell mischende Markov-Ketten

9.1 Definition

Eine ergodische Markov-Kette ist
(zeit-)reversibel (*in detailed balance*),
wenn für stationäre Verteilung $\mathbf{w}$ und alle $i, j \in S$ gilt:

$$\mathbf{w}_i P_{ij} = \mathbf{w}_j P_{ji} \ .$$

9.2 Definition

- $G = (V, E)$ ein zusammenhängender ungerichteter Graph ohne Schlingen
- $0 < \beta \leq 1$ eine reelle Zahl
- $d(i)$ Grad von Knoten i und $d = \max_{i \in V} d(i)$
- definiere Markov-Kette $M_{G,\beta}$:

$$P_{ij} = \begin{cases} \beta/d & \text{falls } i \neq j \text{ und } (i,j) \in E \\ 0 & \text{falls } i \neq j \text{ und } (i,j) \notin E \\ 1 - d(i)\beta/d & \text{falls } i = j \end{cases}$$

9.3 Beobachtung/Übung

- Kette irreduzibel, da Graph zusammenhängend
- für $\beta < 1$ Kette aperiodisch
- stationäre Verteilung $M_{G,\beta}$: Gleichverteilung
- für $\beta < 1$ Kette reversibel

9.4 Verallgemeinerung

- $\mathbf{p}$ eine Wahrscheinlichkeitsverteilung auf V, nirgends 0
- $G = (V, E)$ ein zusammenhängender ungerichteter Graph
- $0 < \beta \leq 1$ eine reelle Zahl
- $d(i)$ Grad von Knoten i und $d = \max_{i \in V} d(i)$
- definiere Markov-Kette $M_{G,\beta}$:

$$P_{ij} = \begin{cases} \min(1, \frac{\mathbf{p}_j}{\mathbf{p}_i}) \cdot \beta/d & \text{falls } i \neq j \text{ und } (i,j) \in E \\ 0 & \text{falls } i \neq j \text{ und } (i,j) \notin E \\ 1 - \sum_{i \neq k} P_{ik} & \text{falls } i = j \end{cases}$$

9.4 Verallgemeinerung (2)

Sei $i \neq j$ und $(i,j) \in E$ (alles andere trivial):

$$\begin{aligned}
\frac{\mathbf{p}_i P_{ij}}{\mathbf{p}_j P_{ji}} &= \frac{\mathbf{p}_i \min(1, \frac{\mathbf{p}_j}{\mathbf{p}_i}) \cdot \beta/d}{\mathbf{p}_j \min(1, \frac{\mathbf{p}_i}{\mathbf{p}_j}) \cdot \beta/d} \\
&= \begin{cases} \frac{\mathbf{p}_i}{\mathbf{p}_j} / \frac{\mathbf{p}_i}{\mathbf{p}_j} & \text{falls } \mathbf{p}_i \leq \mathbf{p}_j \\ \frac{\mathbf{p}_i}{\mathbf{p}_j} \cdot \frac{\mathbf{p}_j}{\mathbf{p}_i} & \text{falls } \mathbf{p}_i > \mathbf{p}_j \end{cases} \\
&= 1
\end{aligned}$$

Die Markov-Kette ist reversibel mit stationärer Verteilung $\mathbf{p}$, denn …

9.5 Lemma

- Sei M Markov-Kette und
- $\mathbf{q}$ Verteilung mit
$$\mathbf{q}_i P_{ij} = \mathbf{q}_j P_{ji}$$
für alle Zustände i und j.
- Dann ist $\mathbf{q}$ eine stationäre Verteilung von M.

9.6 Beweis

Für alle i ist

$$(\mathbf{q}P)_i = \sum_j \mathbf{q}_j P_{ji} = \sum_j \mathbf{q}_i P_{ij} = \mathbf{q}_i \sum_j P_{ij} = \mathbf{q}_i \ .$$

9.7 Vorbereitung Metropolis-Algorithmus

- Q zeilenstochastische Matrix, symmetrisch
  - *proposal matrix*
  - Voraussetzung Symmetrie von Hastings beseitigt
- Ziel reversible Markov-Kette
  - mit einer Übergangsmatrix P
  - stationäre Verteilung $\mathbf{w}$
- $\mathbf{w}$ „indirekt“ gegeben durch $H\colon S \to \mathbb{R}$
  - *energy function*
  - $\mathbf{w}_i$ proportional zu $e^{-\beta H(i)}$
    - für ein $\beta \geq 0$
    - «inverse Temperatur»
    - nachfolgend $\beta = 1$ angenommen
- Proportionalitätsfaktor $Z(\beta) = \sum_{i \in S} e^{-\beta H(i)}$
  - *partition function*

9.8 Metropolis-Algorithmus

- starte in beliebigem $i_0 \in S$
- Schritt von Zustand $i_t \in S$ zu i_{t+1} in zwei Phasen:
  - gemäß der Verteilung $Q(i_t, \cdot)$ zufällig $j \in S$ wählen
  - dann
    - falls $H(j) \leq H(i_t)$, also $w_j \geq w_{i_t}$, wähle $i_{t+1} = j$.
    - falls $H(j) > H(i_t)$, also $w_j < w_{i_t}$, wähle

$$i_{t+1} = \begin{cases} j & \text{mit Wahrscheinlichkeit } w_j / w_{i_t} = e^{H(i_t) - H(j)} \\ i_t & \text{sonst} \end{cases}$$

9.9 Satz

Metropolis-Algorithmus, Markov-Kette auf S:

- reversibel
- stationäre Verteilung: i hat Wahrscheinlichkeit $e^{-H(i)}/Z$.

9.10 Beweis (Übung)

P durch Algorithmus festgelegte Übergangsmatrix
genügt zu zeigen: reversibel

$$e^{-H(i)}/Z \cdot P_{ij} = e^{-H(j)}/Z \cdot P_{ji}$$

einfache Rechnung analog zu der in Punkt 9.4

Überblick

Der Metropolis-Algorithmus

Anmerkungen zu Eigenwerten

Schnell mischende Markov-Ketten

9.11 Lemma

Es sei P eine zeilenstochastische Matrix.

- Dann ist 1 ein Eigenwert von P.
- Für jeden Eigenwert λ von P gilt: $|\lambda| \leq 1$.

9.12 Beweis

- Erster Teil klar wegen $P(1,\ldots,1)^T = (1,\ldots,1)^T$.
- sei $\lambda \in \mathbb{C}$ Eigenwert mit $P(x_1,\ldots,x_n)^T = \lambda(x_1,\ldots,x_n)^T$
- Sei i_0 ein Index mit $|x_{i_0}| = \max_j |x_j|$.

9.12 Beweis

- Erster Teil klar wegen $P(1,\ldots,1)^T = (1,\ldots,1)^T$.
- sei $\lambda \in \mathbb{C}$ Eigenwert mit $P(x_1,\ldots,x_n)^T = \lambda(x_1,\ldots,x_n)^T$
- Sei i_0 ein Index mit $|x_{i_0}| = \max_j |x_j|$.
- Es gilt:

$$\begin{aligned} |\lambda| \cdot |x_{i_0}| &= |\lambda \cdot x_{i_0}| = \left|\sum_j P_{i_0 j} x_j\right| \\ &\leq \sum_j P_{i_0 j}|x_j| \leq \sum_j P_{i_0 j}|x_{i_0}| \\ &= |x_{i_0}| \sum_j P_{i_0 j} = |x_{i_0}| \,. \end{aligned}$$

Also ist $|\lambda| \leq 1$.

9.13 Lemma

sei P stochastische Matrix

- einer reversiblen Markov-Kette
- mit $|S| = N$ Zuständen

dann besitzt P

- nur reelle Eigenwerte
- mit $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_N > -1$.

Beweis: lineare Algebra …

9.14 Abkürzung

- $\lambda_{max} = \max\{|\lambda| \mid \lambda \text{ ist Eigenwert von } P \text{ und } \lambda \neq 1\}$.
- Eben bewiesen:
  - für ergodische Markov-Ketten: $\lambda_{max} \leq 1$.
  - für reversible Markov-Ketten: $\lambda_{max} = \max\{|\lambda_2|, |\lambda_N|\} < 1$

9.15 Bemerkung

Falls $\lambda_N < 0$:

- Gehe von $\mathbf{P}$ zu $\mathbf{P}' = \frac{1}{2}(\mathbf{P} + \mathbf{I})$ über.
- Eigenwerte $\lambda_i' = \frac{1}{2}(\lambda_i + 1) > 0$, also
- $\lambda_{max} = \lambda_2$

Überblick

Der Metropolis-Algorithmus

Anmerkungen zu Eigenwerten

Schnell mischende Markov-Ketten

9.16 Definition

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

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

- Übung:

$$\|\mathbf{p} - \mathbf{q}\|_{tv} = \max_{T \subseteq S} |\mathbf{p}(T) - \mathbf{q}(T)|$$

wobei $\mathbf{p}(T) = \sum_{j \in T} \mathbf{p}_j$

9.17 Definition

- ergodische Markov-Kette $\mathbf{P}$ stationäre Verteilung $\mathbf{w}$; für alle Verteilungen $\mathbf{p}$ sei

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

- Die *Variationsdistanz zum Zeitpunkt t* ist

$$\Delta(t) = \max_{\mathbf{p}} \delta_{\mathbf{p}}(t) \ .$$

- Übung: Maximum für einen Einheitsvektor

$$\Delta(t) = \max_{i} \|\mathbf{P}_i^t - \mathbf{w}\|_{tv}$$

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

$$\lim_{t \to \infty} \Delta(t) = 0$$

9.18 Satz

ergodische Markov-Kette, Matrix $\mathbf{P}$, stationäre Verteilung $\mathbf{w}$
Es existieren Konstanten C und $\alpha < 1$ mit

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

9.19 Beweis (1)

- $\mathbf{P}^{t_0}$ enthalte nur positive Werte
- $\mathbf{W}$ wie bisher
- dann exisitiert $0 < \delta < 1$ so, dass für alle i und j gilt: $\mathbf{P}^{t_0}_{ij} \geq \delta \mathbf{w}_j$
- $\vartheta = 1 - \delta$
- stochastische Matrix $\mathbf{Q}$ mit

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

- es gilt:
  - $\mathbf{Q}^k \mathbf{W} = \mathbf{W}$
  - $\mathbf{W}\mathbf{P}^{t_0} = \mathbf{W}$
- Induktion: für jedes $k \in \mathbb{N}_+$ ist

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

9.19 Beweis (2)

- $\mathbf{P}^{t_0 k} = (1 - \vartheta^k)\mathbf{W} + \vartheta^k \mathbf{Q}^k$
- $\mathbf{P}^{t_0 k + j} - \mathbf{W} = \vartheta^k \left(\mathbf{Q}^k \mathbf{P}^j - \mathbf{W}\right)$
- Zeile i: $\|\mathbf{P}_i^{t_0 k + j} - \mathbf{w}\|_{tv} \leq \vartheta^k$
- für beliebiges $t = t_0 k + j$:

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

9.20 Abkürzung

reversible Markov-Kette mit stationärer Verteilung $\mathbf{w}$:

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

9.21 Satz

Für reversible Markov-Kette mit stationärer Verteilung $\mathbf{w}$ gilt:

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

9.22

Mitunter folgende Situation:

- Problem mit Instanzen I
- Familie von Markov-Ketten $M(I)$
- Beispiel:
  - I: ein Graph
  - Zustände von $M(I)$: die Matchings von I

9.23 Definition

ergodische Markov-Kette mit stationärer Verteilung $\mathbf{w}$

für $\varepsilon > 0$ definiere ihre *ε-Konvergenzzeit*:

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

Von Probleminstanzen zu Markovketten

mitunter vorliegende Situation

- gegeben: unendliche Menge von «Probleminstanzen» I
- zu jedem I wird eine Markovkette $M(I)$ konstruiert

Beispiel:

- I: Graph (V, E)
- $M(I)$: Markovkette mit
  - Zustände: Matchings in I
  - Übergangswahrscheinlichkeiten für Matching m:
    - mit W.keit $1/2$ kein Übergang
    - mit W.keit $1/2$ Übergang zu

$$m' = \begin{cases} m - e, & \text{falls } e \in m \\ m + e, & \text{falls } e \notin m \text{ und } m + e \text{ Matching} \\ m, & \text{sonst} \end{cases}$$

für zufällig gleichverteilt gewählte Kante $e \in E$

9.24 Definition

Familie $M(I)$ heißt *schnell mischend*, falls die ε-Konvergenzzeit polynomiell in $|I|$ und $\ln 1/\varepsilon$ ist.

9.25 Rechnung

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

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

9.25 Rechnung (2)

- $$\begin{aligned} \left(\frac{1}{\lambda_{max}}\right)^t &\geq \frac{1}{\varepsilon w_{min}} \\ t &\geq \frac{\ln \varepsilon^{-1} + \ln w_{min}^{-1}}{\ln \lambda_{max}^{-1}} \end{aligned}$$
- Wegen $1 - x \leq \ln x^{-1}$ für $0 < x < 1$ ist dafür hinreichend:
$$t \geq \frac{\ln \varepsilon^{-1} + \ln w_{min}^{-1}}{1 - \lambda_{max}}$$
- schnelles Mischen, falls $\ln w_{min}^{-1}$ und $1/(1 - \lambda_{max})$ polynomiell in $|I|$

9.26 Satz

Für reversible Markov-Ketten gilt:

$$\tau(\varepsilon) \leq \frac{1}{1-\lambda_{max}} \log \frac{1}{w_{min}\varepsilon}$$
$$\tau(\varepsilon) \geq \frac{1}{2(1-\lambda_{max})} \log \frac{1}{2\varepsilon}$$

Woher bekommt man λ_{max}?

9.27 Definition

- reversible Markov-Kette $M = (S, \mathbf{P})$,
  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) gewichtet mit Zahl $c(i, j) = \mathbf{w}_i P_{ij}$.

9.28 Definition

- Für reversible Markov-Kette mit Zustandsmenge S und stationärer Verteilung $\mathbf{w}$
- definiere für jede Teilmenge $T \subseteq S$

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

- Der *Leitwert* Φ der Markov-Kette ist dann

$$\Phi = \min_{T \subseteq S} \max\{\Phi(T), \Phi(S \setminus T)\} = \min_{T \subseteq S, 0 < C(T) \leq 1/2} \Phi(T) \ .$$

9.29 Satz

Für jede reversible Markov-Kette gilt:

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

9.30 Korollar

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

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

9.32 Beobachtung

Für die Markov-Ketten $M_{G,\beta}$ aus Definition 9.2 gilt:

$$\Phi = \beta\mu/d \ .$$

Woher bekommt man im allgemeinen Φ?

9.33 Definition

- für 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.
- gesucht: Flüsse $f_{ij} : E_F \longrightarrow \mathbb{R}_+$, so dass gilt:

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

9.33 Definition (2)

- Gesamtfluss durch eine Kante e sei

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

und

- die *relative Kantenauslastung*

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

mit $c((i,j)) = \mathbf{w}_i P_{ij}$

9.34 Lemma

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

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

9.35 Beispiel: Hyperwürfel

- reversible Markov-Kette M gemäß Def. 9.2 aus Hyperwürfel H_n der Dimension n mit $\beta = 1/2$.
- $P_{ii} = 1/2$ und $P_{ij} = 1/2n$ für $i \neq j$.
- stationäre Verteilung
  - Gleichverteilung mit $w_{min} = 1/2^n$
  - also $\log 1/w_{min} = n$ polynomiell in n
- Schnelles Mischen: Suche Flüsse f_{ij}
  - die $1/2^{2n}$ transportieren und
  - kleine Kantenauslastung erzeugen.

9.35 Beispiel: Hyperwürfel (2)

- Wähle: gleichmäßige Verteilung auf alle $d!$ kürzeste Pfade zwischen i und j (d = Hammingdistanz)
- Aus Symmetriegründen gleicher Gesamtfluss auf jeder Kante.
- 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. Also:

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

9.35 Beispiel (3)

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

9.35 Beispiel (4)

- eben: $f(e) = 1/(2 \cdot 2^n)$.
- andererseits $c(e) = \mathbf{w}_i P_{ij} = 1/(2n \cdot 2^n)$
- also $\rho(f) = \frac{1/(2 \cdot 2^n)}{1/(2n \cdot 2^n)} = n$
- wegen 9.32 daher $\Phi \geq \frac{1}{2\rho(f)} = \frac{1}{2n}$
- wegen 9.27 daher $1 - \frac{1}{2n^2} \leq \lambda_2 \leq 1 - \frac{1}{8n^2}$
- also $1/(1 - \lambda_2) \in \Theta(n^2)$
- zusammen mit $\ln w_{min}^{-1} = n$
- Korollar 9.28: Markov-Kette schnell mischend

Zusammenfassung

- oft ergodische Markov-Ketten, die sogar reversibel sind
- Metropolis-Algorithmus
- Kriterien, um festzustellen, ob sie schnell mischend sind