T-79.300 Stochastic Algorithms

Ergodicity and convergence in Markov chains

Anne Patrikainen
Laboratory of Computer and Information Science
20.10.2003

Outline of the presentation

- Part 1: Review of Markov chains and linear algebra
  - Irreducibility, ergodicity, reversibility....
  - Eigenvectors, eigenvalues...
- Part 2: Estimates for the convergence speed of Markov Chains
  - We will look at the well-known Perron-Frobenius theorem on the speed of convergence
  - The second largest eigenvalue modulus of the transition matrix turns out to be extremely important
  - But often it cannot be calculated explicitly. We will therefore derive various upper and lower bounds for it.

Material

- The main reference: Chapter 6 of P. Brémaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer-Verlag, New York, 1999.
- The basic concepts are nicely explained in O. Häggström, *Finite Markov Chains and Algorithmic Applications*. Cambridge University Press, 2002. We will cover chapters 1–6 in the introductory part of the presentation.
- As a linear algebra reference, I warmly recommend R. A. Horn, C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

Part 1:
Review of Markov chains and linear algebra

Markov chains

- Let $P = (P_{ij})$ be a kxk matrix. A random process $(X_0, X_1, \ldots)$ with finite state space $S = \{s_1, \ldots, s_k\}$ is said to be a homogeneous first-order **Markov chain** with transition matrix P, if for all n, all $i, j \in \{1, \ldots, k\}$, and all $i_0, \ldots, i_{n-1} \in \{1, \ldots, k\}$ we have

$$\begin{aligned} &\mathrm{P}(X_{n+1} = s_j | \, X_0 = s_{i_0}, X_1 = s_{i_1}, \ldots, X_{n-1} = s_{i_{n-1}}, X_n = s_i) \\ &= \mathrm{P}(X_{n+1} = s_j | X_n = s_i) \\ &= P_{ij} \end{aligned}$$

- Every transition matrix P satisfies $P_{ij} \geq 0$ for all $i, j \in \{1, \ldots, k\}$ and $\sum_{j=1}^{k} P_{ij} = 1$ for every $i \in \{1, \ldots, k\}$. This kind of a matrix is referred to as a **stochastic matrix**.

Irreducible Markov chain

- State s_i **communicates** with another state s_j, written as $s_i \to s_j$, if the chain has positive probability of ever reaching s_j when started from s_i. In other words, there exists n such that $(P^n)_{ij} > 0$.
- If $s_i \to s_j$ and $s_j \to s_i$, we say that the states **intercommunicate** and write $s_i \leftrightarrow s_j$.
- A Markov chain with state space S and transition matrix P is said to be **irreducible** if for all $s_i, s_j \in S$ we have $s_i \leftrightarrow s_j$. Otherwise the chain is **reducible**.

Aperiodic Markov chain

- The **period** $\mathrm{d}(s_i)$ of a state s_i is the greatest common divisor of the set of times after which the chain can return to s_i, given that we start with s_i.
- If $\mathrm{d}(s_i) = 1$, we say that the state s_i is aperiodic.
- A Markov chain is said to be **aperiodic** if all its states are aperiodic. Otherwise the chain is said to be **periodic**.

Markov chains and distributions

- We consider a probability distribution $\mu(0)$ on the state space $S = \{s_1, \ldots, s_k\}$. That is, $\mu(0) = (\mu_1(0), \mu_2(0), \ldots, \mu_k(0))^T = (\mathrm{P}(X_0 = s_1), \mathrm{P}(X_0 = s_2), \ldots, \mathrm{P}(X_0 = s_k))^T$.
- After one time step, the distribution becomes $\mu(1)^T = \mu(0)^T P$.
- After n time steps, we have $\mu(n)^T = \mu(n-1)^T P = \mu(0)^T P^n$.

Stationary distribution of a Markov chain

- Consider a distribution π that does not change in time: $\pi^T = \pi^T P$.
- This kind of a distribution is referred to as a stationary distribution of the Markov chain.
- Any irreducible and aperiodic Markov chain has exactly one stationary distribution.
- In the case of undirected transition graph, the i:th element of the stationary distribution is proportional to the degree of the i:th vertex of the graph (corresponding to the i:th state).
- But in the general directed case, it is more difficult to get an intuition on the form of the stationary distribution without calculations.

Convergence of Markov chains

- We wish to consider the asymptotic behavior of the distribution $\mu(n)^T = \mu(0)^T P^n$, when the initial distribution $\mu(0)$ is arbitrary.
- We need to define what it means for a sequence of probability distributions $\mu(0), \mu(1), \mu(2), \ldots$ to converge to a limiting probability distribution π.
- There are several possible metrics in the space of probability distributions; the one usually considered with Markov chains is the so-called **total variation distance**.

Convergence of Markov chains

- Let $\mu = (\mu_1, \ldots, \mu_k)^T$ and $\nu = (\nu_1, \ldots, \nu_k)^T$ be probability distributions on state space $S = \{s_1, \ldots, s_k\}$. We now define the total variation distance between μ and ν as

$$\mathrm{d}_{\mathrm{TV}}(\mu, \nu) = \frac{1}{2} \sum_{i=1}^{k} |\mu_i - \nu_i| = \frac{1}{2} ||\mu - \nu||_1.$$

- We say that $\mu(n)$ **converges to** μ **in total variation as** $n \to \infty$, writing $\mu(n) \stackrel{\mathrm{TV}}{\to} \mu$, if $\lim_{n\to\infty} \mathrm{d}_{\mathrm{TV}}(\mu(n), \mu) = 0$.

- The constant $\frac{1}{2}$ is designed to make the total variation distance take values between 0 and 1.

The Markov chain convergence theorem

- Let $(X_0, X_1, \ldots)$ be an irreducible aperiodic Markov chain with state space $S = \{s_1, \ldots, s_k\}$, transition matrix P, and arbitrary initial distribution $\mu(0)$. Then, for the stationary distribution π, we have $\mu(n) \stackrel{\mathrm{TV}}{\rightarrow} \pi$.

- In other words, regardless of the initial distribution, we always end up with the stationary distribution.

Reversible Markov chains

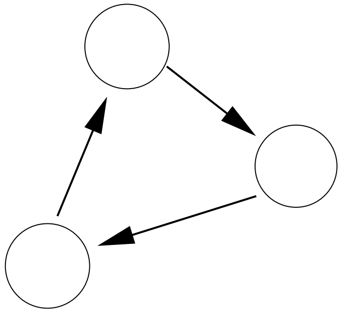
- Consider a Markov chain with state space S and transition matrix P. A probability distribution π on S is said to be reversible for the chain if for all $i, j \in \{1, \ldots, k\}$ we have

$$\pi_i P_{ij} = \pi_j P_{ji}.$$

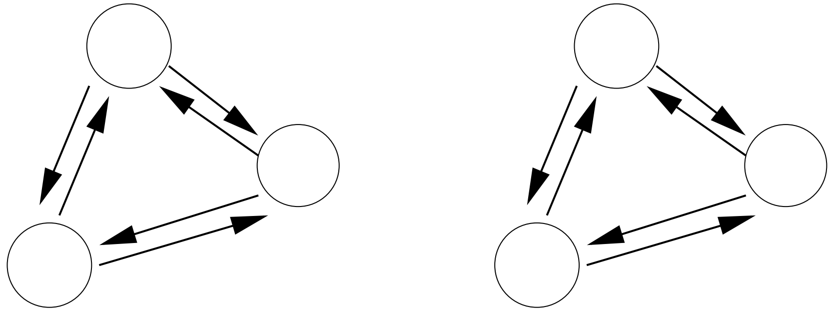
  A Markov chain is said to be **reversible** if there exists a reversible distribution for it.

- The amount of probability mass flowing from state s_i to state s_j equals to the mass flowing from s_j to s_i.
- Any reversible distribution is also a stationary distribution.
- But a stationary distribution might not be a reversible distribution.

Reversibility - examples



Irreversible chain
Unique stationary distribution



Reversible chain that is not irreducible
No unique stationary distribution

Ergodicity

- We are almost done with the review of Markov chains — but how about ergodicity mentioned in the title of the presentation?
- Ergodicity is an important concept in the general theory of Markov chains: The **ergodicity theorem** tells us that an ergodic chain has a unique stationary distribution.
- But in this course, we are dealing with chains on finite state spaces only. Therefore the only conditions needed for uniqueness of the stationary distribution are irreducibility and aperiodicity.

Ergodicity

- In general, a Markov chain is **ergodic** if it is irreducible, aperiodic, and positive recurrent.
- A chain is **positive recurrent** if all its states are. State s_i is positive recurrent if it can be returned to in a finite number of steps with probability 1, and if the expected return time to s_i is finite.
- A given state is transient if it cannot be returned to in a finite number of steps with probability 1. If a state is not transient nor positive recurrent, it is null recurrent.
- If a chain is finite and irreducible, it is also positive recurrent. Therefore a finite, irreducible, and aperiodic chain is also ergodic.

A prelude to the Perron-Frobenius theorem

- In case of a finite state space, a Markov Chain is wholly defined by a transition matrix P.
- The asymptotic behavior of the chain depends on the behavior of P^n, when the number of steps n approaches infinity.
- The behavior of P^n depends in turn on the eigenstructure of P.
- The Perron-Frobenius theorem relates the speed of convergence of the chain to the eigenstructure of the transition matrix.
- We will therefore go on to review some basics concepts of linear algebra.

Eigenvectors and eigenvalues - a review

- The **right eigenvectors** v of a matrix P are given by $Pv = \lambda v$. Here λ is the corresponding eigenvalue.
- The **left eigenvectors** u are given by $u^T P = \mu u^T$. Here μ is an eigenvalue and u^T stands for the transpose of u.
- The set of eigenvalues is the same for the left and the right eigenvectors.
- The **algebraic multiplicity** of an eigenvalue tells how many times the eigenvalue appears as a root of the characteristic polynomial. The **geometric multiplicity** is the dimension of the corresponding eigenspace.

Eigenvectors and eigenvalues - a review

- If the matrix P has eigenvalues $\{\lambda_i\}$, the matrix P^n has eigenvalues $\{\lambda_i^n\}$ (the eigenvectors are the same).
- If the kxk matrix P has distinct eigenvalues, we have the **spectral decomposition** $P = \sum_{i=1}^{k} \lambda_i v_i u_i^T$.
- Furthermore, $P^n = \sum_{i=1}^{k} \lambda_i^n v_i u_i^T$.

The eigenvalues and eigenvectors of the transition matrix P

- Recall that the stationary distribution is defined as $\pi^T = \pi^T P$. Thus the left eigenvector corresponding to eigenvalue 1 is $u_1 = \pi$.
- Associated with an eigenvalue 1 we also have a right eigenvector $v_1 = \mathbf{1}$, the vector of all ones.

Part 2:
Estimates for the convergence speed of Markov chains

The Perron-Frobenius theorem

Let P be stochastic, irreducible, aperiodic kxk matrix.

Then there exists a real eigenvalue $\lambda_1 = 1$ with algebraic as well as geometric multiplicity one. For any other eigenvalue λ_j (might be complex-valued), $\lambda_1 > |\lambda_j|$. We order the eigenvalues by modulus, i.e. $\lambda_1 > |\lambda_2| \geq \ldots \geq |\lambda_k|$. Let us denote the algebraic multiplicity of the eigenvalue λ_i by m_i. Now

$$\begin{aligned} P^n &= \lambda_1^n v_1 u_1^T + \Theta(n^{m_2-1}|\lambda_2|^n) \\ &= \mathbf{1}\pi^T + \Theta(n^{m_2-1}|\lambda_2|^n) \end{aligned}$$

Here $\Theta(f(n))$ represents a function of n such that there exist constants α, β, n_0, $0 < \alpha \leq \beta < \infty$, such that $\alpha f(n) \leq \Theta(f(n)) \leq \beta f(n)$ for all $n > n_0$.

The Perron-Frobenius theorem — intuition

- Consider having a transition matrix $A = \mathbf{1}\pi^T$ and an initial distribution $\mu(0)$.
- After one time step, we have $\mu(1)^T = \mu(0)^T A = \mu(0)^T \mathbf{1}\pi^T = \pi^T$, the stationary distribution.

The Perron-Frobenius theorem — an example

Consider the doubly stochastic matrix

$$P = \frac{1}{12} \begin{bmatrix} 0 & 6 & 6 \\ 4 & 3 & 5 \\ 8 & 3 & 1 \end{bmatrix}.$$

The eigenvalues are $\lambda_1 = 1$, $\lambda_2 = -\frac{1}{2}$, $\lambda_3 = -\frac{1}{3}$.
The right and the left eigenvectors are $u_1 = (1,1,1)^T$, $v_1 = \frac{1}{3}(1,1,1)^T$, $u_2 = \frac{1}{12}(2,-1,-1)^T$, $v_2 = (4,1,-5)^T$, $u_3 = \frac{1}{4}(-2,3,-1)^T$, and $v_3 = (0,1,-1)^T$.

The Perron-Frobenius theorem — an example

Now

$$
\begin{aligned}
P^n &= \sum_{i=1}^{3} \lambda_i^n v_i u_i^T = \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \\
&\quad + (-\tfrac{1}{2})^n \frac{1}{12}\begin{bmatrix} 8 & -4 & -4 \\ 2 & -1 & -1 \\ -10 & 5 & 5 \end{bmatrix} + (-\tfrac{1}{6})^n \frac{1}{4}\begin{bmatrix} 0 & 0 & 0 \\ -2 & 3 & -1 \\ 2 & -3 & 1 \end{bmatrix}
\end{aligned}
$$

The convergence is geometric with relative speed $\frac{1}{2}$.

The Perron-Frobenius theorem in practice

- We are able to estimate the speed of convergence of a Markov chain based on the second eigenvalue modulus of the transition matrix.
- But in practice it may be impossible to calculate the eigenvalues.
- For instance, in a MCMC simulation, we do not have the means to calculate them.
- But we would like to know how long to run our simulation — how long does it take to get close to the stationary distribution.
- Good upper bounds for the second eigenvalue modulus would be useful.

Bounds for the second eigenvalue modulus

- We will assume that our Markov chain is reversible in addition to being finite, aperiodic, and irreducible. This makes the analysis easier.
- In order to proceed, we will need some new definitions.
- If π is a strictly positive probability distribution on the state space S with k states, let $l^2(\pi)$ be the real vector space R^k endowed with the inner product $< x, y >_\pi := \sum_i x(i)y(i)\pi(i)$.
- It follows that the norm is $||x||_\pi^2 := \sum_i x^2(i)\pi(i)$.
- A convenient definition for the expectation follows: $\mathrm{E}_\pi(x) :=< x, 1 >_\pi$.
- Similarly for the variance: $\mathrm{Var}_\pi(x) := ||x||_\pi^2 - \mathrm{E}_\pi^2(x)$.

Bounds for the second eigenvalue modulus

- The **Dirichlet form** $\mathcal{E}_\pi(x,x)$ associated with a reversible pair (P,π) is defined by

$$\mathcal{E}_\pi(x,x) =< (I-P)x, x >_\pi .$$

- We change the notation and order the eigenvalues of P as $\lambda_1 > \lambda_2 \geq \lambda_3 \geq \ldots$ (by value, not by modulus).

- We are able to calculate an upper bound for λ_2. If $A > 0$ is such that for all $x \in R^k$, $\mathrm{Var}_\pi(x) \leq A\mathcal{E}_\pi(x,x)$, then $\lambda_2 \leq 1 - \frac{1}{A}$.

- We also need a lower bound for the smallest eigenvalue λ_k. If $B > 0$ is such that for all $x \in R^k$, $< Px, x >_\pi + ||x||_\pi^2 \geq B||x||_\pi^2$, then $\lambda_k \geq B - 1$.

- The second largest eigenvalue modulus $\rho = \max(\lambda_2, |\lambda_k|)$.

Beyond the Perron-Frobenius theorem

- Perron-Frobenius theorem is not the only way to estimate the speed of convergence. However, the second largest eigenvalue modulus keeps showing up.
- We again consider reversible, irreducible, aperiodic Markov chains with state space $S = \{s_1, \ldots, s_k\}$, transition matrix P and stationary distribution π.
- For all n and all $i \in \{1, \ldots, k\}$ we have

$$\mathrm{d}_{TV}(\delta_i^T P^n, \pi)^2 \leq \frac{P_{ii}(2)}{\pi(i)} \rho^{2n-2},$$

where ρ is the second largest eigenvalue modulus of P, and δ_i is the Dirac's delta vector.

Beyond the Perron-Frobenius theorem

- For any probability distribution μ, and for all $n \geq 1$,

$$||\mu^T P^n - \pi^T||_{\frac{1}{\pi}} \leq \rho^n ||\mu - \pi||_{\frac{1}{\pi}}.$$

- It also holds that

$$\mathrm{d}_{TV}(\delta_i^T P^n, \pi)^2 \leq \frac{1 - \pi(i)}{4\pi(i)} \rho^{2n}.$$

- In both bounds, we have the familiar second largest eigenvalue modulus ρ. Again, we need to derive bounds for it.

Eigenvalue bounds with weighted paths

- We will continue considering reversible, finite, irreducible, aperiodic Markov chains.
- We will consider oriented edges e of the transition graph associated with P.
- For each oriented edge e, define $Q(e) = \pi(i)P_{ij}$.
- For each ordered pair of distinct states (s_i, s_j), select arbitrarily one path from s_i to s_j. That is, a sequence $i, i_1, \ldots, i_m, j$ which does not use the same edge twice.
- Let Γ be the collection of paths so selected. For a path $\gamma_{ij} \in \Gamma$, define

$$|\gamma_{ij}|_Q := \sum_{e \in \gamma_{ij}} \frac{1}{Q(e)} = \frac{1}{\pi(i)P_{ii_1}} + \frac{1}{\pi(i_1)P_{i_1 i_2}} + \ldots + \frac{1}{\pi(i_m)P_{i_m j}}.$$

Eigenvalue bounds with weighted paths

- Define the **Poincaré coefficient**

$$\kappa = \kappa(\Gamma) = \max_e \sum_{\gamma_{ij} \ni e} |\gamma_{ij}|_Q \pi(i)\pi(j).$$

- An upper bound for the second largest eigenvalue of P is given by

$$\lambda_2 \leq 1 - \frac{1}{\kappa}.$$

- But again, in order to derive an upper bound for the second largest eigenvalue modulus, we need a lower bound for the smallest eigenvalue λ_k.

Eigenvalue bounds with weighted paths

- For each state s_i, select exactly one closed path σ_i from s_i to s_i such that it does not pass twice through the same edge, and with an odd number of edges.
- Let Σ be the collection of paths so selected. For a path $\sigma_i \in \Sigma$, let

$$|\sigma_i|_Q = \sum_{e \in \sigma_i} \frac{1}{Q(e)}.$$

- Define

$$\alpha = \alpha(\Sigma) = \max_e \sum_{\sigma_i \ni e} |\sigma_i|_Q \pi(i).$$

- Then we get the lower bound

$$\lambda_k \geq \frac{2}{\alpha} - 1.$$

An aside: The other adventures of the second eigenvalue

- The magical second eigenvalue comes up also in contexts that are not directly related to Markov chains.
- The second eigenvalue of the so-called Laplacian matrix of a graph can be utilized in partitioning the graph.
- Spectral clustering is based on calculating the second (or related) eigenvalue of various matrices derived from a data set.
- Spectral clustering is observed to be a valuable technique, but sound theoretical results are rare.
- More theory on the second eigenvalue is needed.

Summary

- The speed of convergence of a Markov chain depends greatly on the second largest eigenvalue modulus of the transition matrix.
- The Perron-Frobenius theorem is the most famous theorem related to this.
- Often in practice, for instance in MCMC applications, it is impossible to calculate the second largest eigenvalue modulus explicitly.
- Bounds are therefore needed. There are various approaches to deriving them. Some were presented, many others can be found in Brémaud's book.