Convergence of Finite Markov Chains

Philippe Robert
INRIA-Rocquencourt
February 2, 1998

[summary by Jean-Marc Lasgouttes]

1. Framework

Let $X(t)$ be an aperiodic and irreducible Markov chain on a finite set $S$, with transition probabilities $P = (p(x, y))_{x, y \in S}$ and equilibrium distribution $(\pi(x))_{x \in S}$. It is often desirable to know how far $\Pr(X(t) = x)$ is from $\pi(x)$, in particular when $\pi(x)$ has a nice closed form, but the transient distribution is difficult to express. It is known (Doeblin) that there exists $\alpha \in [0, 1[$, such that

$$|P_y(t)(x) - \pi(x)| \leq C\alpha^t, \quad \forall x, y \in S,$$

where $P_y(t)$ is the law of $X(t)$ when $X(0) = y$.

This talk is concerned with methods allowing to get more accurate estimates on this difference. The distance that will be used in the following is the total variation distance between two probability distributions $P$ and $Q$ on $S$, defined by

$$d_{tv}(P, Q) = \sup_{A \subseteq S} \{|P(A) - Q(A)|\} = \frac{1}{2} \sum_{i \in S} |P(\{i\}) - Q(\{i\})|,$$

or, more precisely

$$d(t) = \sup_{x \in S} d_{tv}(P_x(t), \pi).$$

One particularly interesting property is the existence of a cutoff (see Diaconis [2]):

**Definition 1.** There is a cutoff if there exist $a_n, b_n \to \infty$, $b_n / a_n \to 0$, such that

$$\lim_{n \to \infty} d(a_n + tb_n) = H(t),$$

with $\lim_{t \to -\infty} H(t) = 1$ and $\lim_{t \to +\infty} H(t) = 0$.

Two main methods can be used to evaluate $d(t)$:

- **Geometric** (see Diaconis and Stroock [3]): when $X$ is a reversible Markov chain, then

$$d_{tv}(P_x(t), \pi) \leq \frac{1}{2} \sqrt{\frac{1 - \pi(x)}{\pi(x)}} |\max\{\beta_1, |\beta_{m-1}|\}|^t,$$

where $-1 < \beta_{m-1} \leq \beta_{m-2} \leq \cdots \leq \beta_1 < \beta_0 = 1$ are the eigenvalues of $P$. The values of $\beta_1$ and $\beta_{m-1}$ can be obtained from the Rayleigh-Ritz principle.

- **Coupling** (see Aldous [1]): let $X$ and $\tilde{X}$ be 2 versions of the Markov chain with transition matrix $P$, such that $X(0) = x$ and $\tilde{X}(0) \sim \pi$. A coupling time is a finite random variable $T$ such that $X(t) = \tilde{X}(t)$, for all $t \geq T$. The following inequality holds for such $T$:

$$d_{tv}(P_x(t), \pi) \leq \Pr(T > t).$$
Moreover, there exists a coupling $T^*$ such that, for all $x \in S$, $d_{tv}(P_x(t), \pi) = \Pr(T^* > t)$.

2. Application to Erlang’s Model

As an example of these techniques, let $X_N(t), t \in \mathbb{R}_+$ be the Markov process associated with a M/M/N/N queue with arrival rate $\lambda N$ and service rate 1. This process, known as an Erlang loss system with $N$ slots, has the transition rates

$$x \rightarrow x + 1 : \lambda N 1_{\{x \leq N\}}$$
$$x \rightarrow x - 1 : x$$

and its equilibrium distribution is

$$\pi(x) = C_N \frac{(\lambda N)^x}{x!}, \quad x \leq N.$$ 

This process has three different regimes:

- $\lambda > 1$. The queue becomes full after a finite time and $N - X_N(t)/N$ is a Markov process whose generator tends as $N \to \infty$ to the generator of a birth and death process.
- $\lambda < 1$. The queue is never full and the process $(X_N(t) - \lambda N)/\sqrt{N}$ has a generator which tends to the generator of an Ornstein–Uhlenbeck process with parameter $\lambda$.
- $\lambda = 1$. The queue becomes full at infinity and the process $(N - X_N(t))/\sqrt{N}$ has a generator which tends to the generator of the reflected Ornstein–Uhlenbeck process on $\mathbb{R}_+$.

The main result obtained in Fricker, Robert and Tibi [4] is the existence of a cutoff in the possible regimes of Erlang’s model:

**Proposition 1.** In the case $\lambda > 1$,

$$\lim_{N \to \infty} \frac{d_N(t)}{N} = \begin{cases} 1 & \text{if } t < \log \frac{\lambda}{\lambda - 1}, \\ 0 & \text{if } t > \log \frac{\lambda}{\lambda - 1}. \end{cases}$$

In the case $\lambda \geq 1$ the behaviour of $d_N$ is such that, for any sequence $\phi(N)$ and for any $a \in \mathbb{R}^*$,

$$\lim_{N \to \infty} d_N \left[ \frac{\log N}{2} + a\phi(N) \right] = \begin{cases} 1 & \text{if } a < 0, \\ 0 & \text{if } a > 0. \end{cases}$$

These results are obtained by coupling techniques and use as an auxiliary process the M/M/1 queue $Y_N(t)$ with input rate $\lambda N$, which is the unbounded version of $X_N(t)$. A central tool in the proof is the process

$$(E_c(t))_{t \geq 0} = \left( (1 + ce^t)Y_N(t)e^{-\lambda N c e^t} \right)_{t \geq 0},$$

which turns out to be a martingale for any $c \geq 0$.

Bibliography


