Markov Decision Chains

A. Hordijk

1. Introduction

Many real-life phenomena have a stochastic dynamic behaviour. Mathematical models for analyzing these phenomena are stochastic processes. For example, in order to study the queue-length at a counter, the mathematical model supposes an arrival process of customers and a distribution of service times. The most simple process already studied early this century by A.K. Erlang, the pioneer in queueing theory (see figure 1), assumes that the probability of an arrival in an interval is linear in the length of this interval with a rest term that is of smaller order than the length of the interval. A similar assumption is made for the service process. This model with Poisson arrivals and exponential service times is denoted by $M/M/1$. Erlang used this mathematical model to compute the long run blocking probability of a telephone-exchange. His goal was to study the quality of service provided by the Danish telephone company he was working for.

In modern telecommunication technology high-speed networks are designed to carry different types of traffic, like audio, video, and data.

One of the challenging problems is to derive the optimal admission control. Given a certain load in the network, should a new arrival, which generates traffic of a certain type, be accepted to the network or should it be blocked? The mathematical problem now becomes: what is the optimal control of the underlying stochastic process?

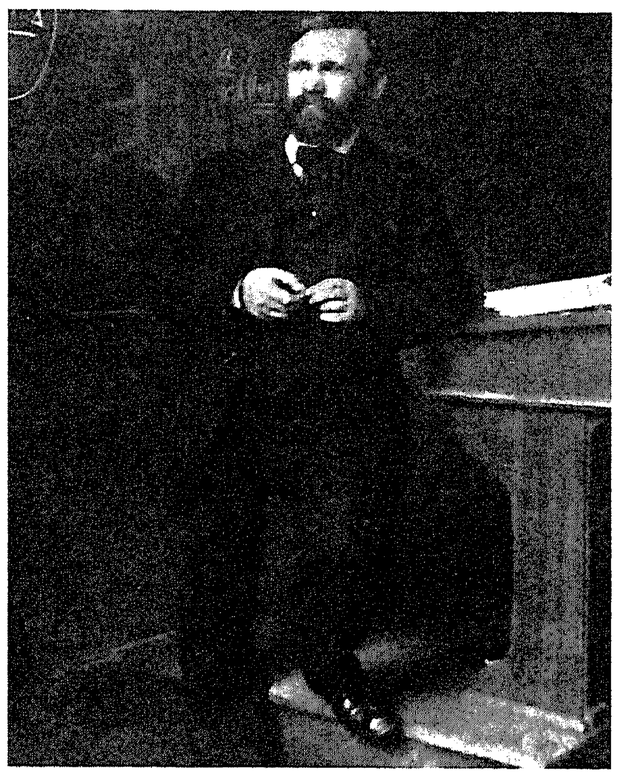

Figure 1. A.K. Erlang, the pioneer in queueing theory.

2. MARKOV DECISION CHAINS

There are several variants of this mathematical model. Let us describe in more detail the discrete-time Markov decision chain (MDC) with a discrete state space. In this model at discrete-time points, which may be stochastic, a decision or control has to be taken. In the admission control model these time points are the arrival times (epochs) of customers. The state of the controlled stochastic process is an element of a subset of the points with integer coordinates in a space of finite dimension. In the admission control model this is the number of customers of the various types at the various nodes in the network.

Each transition from a state at a decision epoch to the state at the next decision epoch has a certain probability. These transition probabilities depend on the chosen control. The control also influences the stochastic rewards and costs until the next decision epoch. For example, the acceptance or rejection of an arrival induces different costs and/or rewards.

The controlled stochastic process may be considered over a finite or an infinite time horizon. In the earlier case the total expected cost is relevant, in the latter, infinite case the total (expected) discounted cost is often considered: when discounted, the cost and/or reward at time instant t is multiplied by α^t with α the discount factor, thus yielding a finite total expected cost. For discount factors close to one, the Laurent expansion of the

total discounted cost in the variable $(1-\alpha)$ is important:

$$\frac{g}{1-\alpha} + u_0 + (1-\alpha)u_1 + \dots$$

In this expansion, g in the first term is the average cost per time unit, and the second term u_0 gives the bias which is the limit of the difference of the total cost over a finite time horizon t minus t times the average cost, as t tends to infinity. All higher order terms have similar interpretations. The Laurent expansion and all its terms depend on the chosen control or policy. One considers several optimality criteria in the nondiscounted case. Average optimality means optimizing the first term of the Laurent expansion. Bias optimality corresponds to lexicographic optimization of the first two terms. (In comparing two policies, this means that if the average cost of policy 1 is lower than that of policy 2, or if, the average costs being equal, the

bias of policy 1 is lower than that of policy 2, then policy 1 is preferred to policy 2.) And in more sensitive optimality criteria more terms of the Laurent expansion are taken into account. The most sensitive criterion, Blackwell optimality, is lexicographic optimization over all terms of the Laurent expansion.

The history of Markov decision chains goes back to the fifties. The first papers were on optimal inventory control. The pioneer R. Bellman wrote his book on Dynamic programming (1957) including a chapter on MDCs. This book also has a chapter on Markov games, a closely connected mathematical model. In Markov games, which were introduced by L. Shapley (1953), there are two or more controllers, called players. The players have different objective functions and mostly play against each other. R.A. Howard, in his book Dynamic Programming and Markov processes (1960), focusses on algorithms and applications.

G. de Leve introduced MDCs in The Netherlands with his Ph.D. thesis: Generalized Markovian decision processes (1964). This started a school of researchers in this field, first at CWI, and later also at universities.

In the late sixties a rather complete theory was available for MDCs with a finite number of states. This theory contains theorems on the existence of optimal policies for the discounted, average, bias and more sensitive optimality criteria. Moreover, optimality equations and methods to solve them were obtained. For a denumerable state space only isolated results were available.

Although a dozen of papers on denumerable MDCs are still appearing each year, we can safely say that a rather complete theory for the denumerable case has been established now, almost twenty years later.

3. DENUMERABLE MARKOV DECISION CHAINS

One can claim that all problems in practice have a finite state space, and one can question the importance of a theory for nonfinite models. However, in many applications the size of the state space is large but unknown and then often the denumerable state space is the natural model. Also the simple structure of the optimal control is often lost, when the denumerable state space is truncated to a finite one. For example, in the admission control problem the optimal control in the case of linear holding costs is of control-limit type, i.e., a customer of specific type is accepted as long as the total number of customers of that type is below a certain number, the control-limit. For computing the optimal control and even more for implementing it in practice, the simple structure is crucial. By truncation of the state space this simple structure is lost.

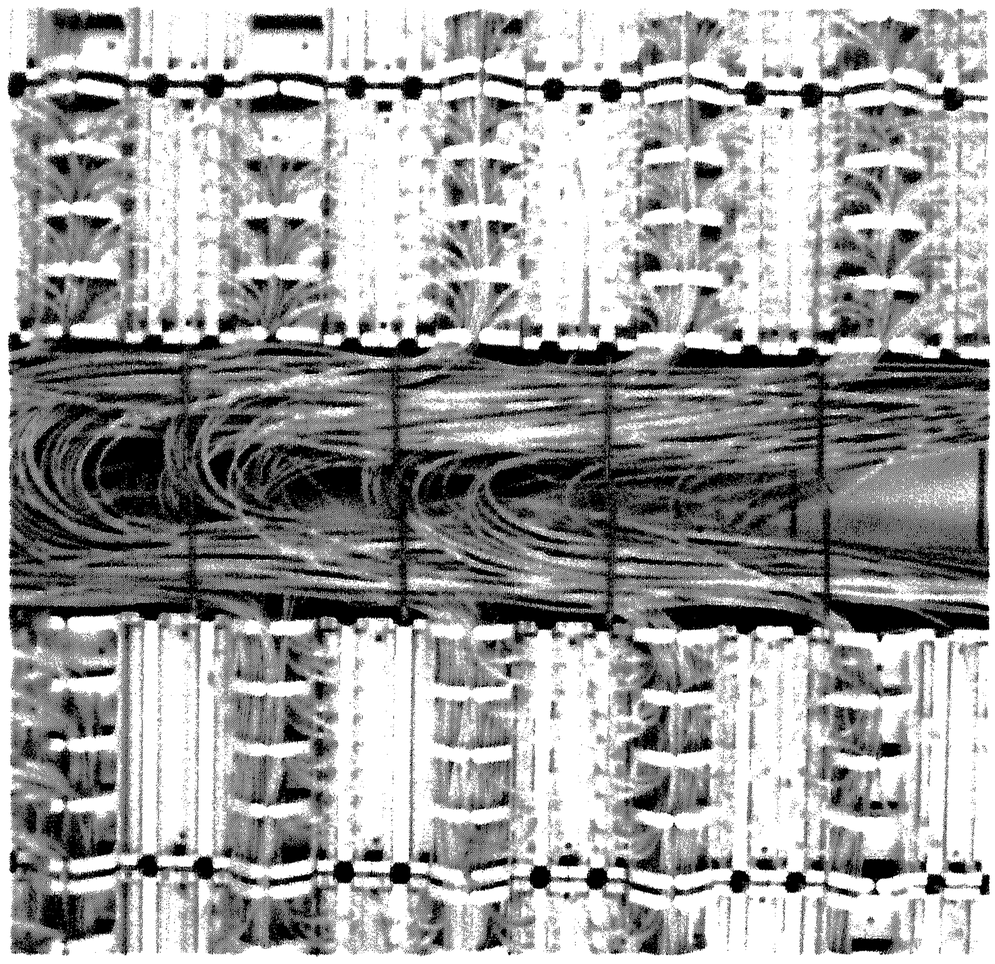

Figure 2. The efficient operation of modern telephone exchanges poses several challenging research problems. (Photo PTT Telecom.)

3.1. Ergodic theory for Markov decision chains

Several of the main steps in developing the theory for denumerable MDC were part of the NWO-SMC projects on Markov decision chains. Let me mention some major results. Whereas the Laurent expansion of the total discounted cost in the finite model always exists, strong recurrence conditions, which guarantee that the stochastic process will return sufficiently 'fast' to a compact set, were necessary for a denumerable number of states. For a satisfactory theory the recurrence conditions should not only be satisfied in the finite state model, they should also be fulfilled in denumerable state applications like the admission control of a telecommunication network.

In classical Markov chain theory there is an extended ergodic theory. In this theory the limiting time average is analyzed, i.e., the behaviour of

$$\frac{1}{T}\sum_{t=1}^{T} X_t$$

is studied as T tends to infinity. The limit, provided it exists, is the long run average reward per time unit, in case X_t is the reward at time t. In

applications discounting is often not appropriate and then the most often used criterion is this long run average reward.

In Markov decision theory not one Markov chain is considered, but a compact set of Markov chains, indexed by the stationary policies, is relevant. For developing a theory for denumerable MDC's it was important to generalize the ergodic theory for one Markov chain to a compact collection.

W. Doeblin was the pioneer in the ergodic theory for Markov chains. His condition for ergodicity, later on called the Doeblin condition, required that the expected recurrence time to a finite set is uniformly bounded in all starting states. Clearly, this condition is too restrictive for almost all queueing models. For example in the simple model with one server, Poisson arrivals and exponential service times (the $M/M/1$ queue), the required number of transitions to the empty state is at least as large as the starting number of customers, so the expectation can never be uniformly bounded for all starting positions. In Markov chain theory Doeblin's work was generalized by T.E. Harris and his recurrence condition is appropriate for queueing models. In Markov decision chains there is the natural requirement that not only the chains are recurrent to a compact set, but that also the expected total cost until this recurrence time is finite. This inspired a condition, which we later on called μ-recurrence. In this assumption the vector μ is a bounding vector of the vector of immediate costs and/or rewards. With this bounding vector weighted supremum norms can be introduced. It is the appropriate extension of the supremum norm, i.e., the μ-norm of vector x is

$$\|x\|_\mu = \sup_i \frac{|x_i|}{\mu_i}, \quad i \in E,$$

where E denotes the state space.

With this vector norm the corresponding norm on the space of matrices is given by

$$\|A\|_\mu = \sup_i \frac{\sum_j |A_{ij}|\mu_j}{\mu_i}.$$

For a given Markov chain let P be the matrix of transition probabilities, and for taboo-set B let ${}_BP$ be the matrix obtained from P by replacing P_{ij} by zero if $j \in B$.

Now the μ-recurrence condition is

$$\exists \text{ finite } B \text{ such that } \|\,{}_BP\|_\mu < 1.$$

It generalizes Doeblin's condition, since for the bounding vector $\mu = e$ with $e_i = 1\ \forall i$, e-recurrence is equivalent to Doeblin's condition.

The strong ergodic theorem for Markov chains can be stated as (we assume for the ease of presentation that the Markov chain is aperiodic):

$$\|P^t - \Pi\| \xrightarrow[t\to\infty]{} 0,$$

with Π the matrix of stationary probabilities.

The research in the SMC-NWO-project resulted in a theory for denumerable MDC that uses as basic assumption μ-uniform recurrence for all stationary policies, i.e., $\exists$ finite B such that

$$\sup_f \| {}_B P(f)\|_\mu < 1$$

where $P(f)$ is the matrix of transition probabilities under the stationary policy f. Key results are:

- The continuity of the Laurent expansion as function of the policy.
- The extension of the strong ergodic theorem for Markov chains to μ-norms and Markov decision chains.

The extension to μ-norms is also an original contribution to Markov chain theory. It inspired important research by S.P. Meyn and R.L. Tweedie for Markov chains with a general state space. The generalization of the strong ergodic theorem for Markov decision chains with a general state space is currently in progress.

3.2. Markov decision chains with partial information

With the many applications in telecommunication, models with decentralized control become important. Consider a communication network. In a certain node a controller has to route customers or packets to one of the neighbouring nodes. His control depends of course on the destination of the packet, it may also depend on the number of jobs on the outgoing links of the node. If we model this as a MDC then the control depends on partial information, the controller may not use all information in the complete state description of the network, because his control may not depend on the numbers of customers in links not adjacent to his node.

Within the SMC-NWO project in recent years MDC's with partial information have been investigated. An algorithm has been constructed for computing a memoryless policy that uses partial information and is close to optimal or optimal in that class of policies. The usual approach to solve a MDC with partial information (or partial observation) is to convert this problem to a MDC with a continuous state space via a posteriori probabilities. The drawback of this approach is that the resulting MDC is unsolvable and also that the optimal policy at time t depends on all realized states and actions from time 1 up to time t, so it is far too complicated to implement it in practice. The new approach provides surprisingly good and implementable policies in the various models studied until now.

Applying MDC's in practical problems remains an art, for each problem a special problem oriented method has to be constructed. The main reason is

m	N
1	3
2	11
3	49
4	261
5	1631
6	11743
7	95901
8	876809
9	8877691
10	98641011

Figure 3. The number of states (N), as a function of the number of customers (m).

the curse of dimensionality in real-life applications, since in most cases the number of states increases exponentially fast with the size of the problem. Figure 3 shows the number of states of a recently analyzed closed queueing network with customer routing and only two nodes.

Usually the situation is more favourable, and often we can handle networks with four nodes. However, clearly a lot of research is still waiting in order to overcome this dimensionality problem.

REFERENCES

1. R. DEKKER, A. HORDIJK (1992). Recurrence conditions for average and Blackwell optimality in denumerable state Markov decision chains. *Math. Oper. Res. 17*, 271-289.

2. R. DEKKER, A. HORDIJK, F.M. SPIEKSMA (1994). On the relation between recurrence and ergodicity properties in denumerable Markov decision chains. *Math. Oper. Res. 19*, 539-559.

3. A. HORDIJK, F.M. SPIEKSMA (1992). On ergodicity and recurrence properties of a Markov chain with an application to an open Jackson network. *Adv. Appl. Prob. 24*, 343-376.

4. A. HORDIJK, J.A. LOEVE (1994). Undiscounted Markov decision chains with partial information; an algorithm for computing a locally optimal periodic policy. *ZOR-Math. Meth. Oper. Res. 40*, 163-181.

5. S.P. MEYN, R.L. TWEEDIE (1993). *Markov Chains and Stochastic Stability*, Springer-Verlag, London.

6. H.C. TIJMS (1988). *Stochastic Modelling and Analysis: a Computational Approach*, Wiley, Chichester.