Vidale M.L. and Wolfe H.B. (1957). An Operations Research Study of Sales Response to Advertising. *Operations Research* 5, 370–381.

von Weizsäcker C.C. (1965). Existence of Optimal Programs of Accumulation for an Infinite Horizon. *Review of Economic Studies* 32, 85–104.

Wickwire K. (1977). Mathematical Models for the Control of Pests and Infectious Diseases: A Survey. *Theoretical Population Biology* 11, 182–238.

Wu C.B. (1997). Continuous Time Markov Decision Processes with Unbounded Reward and Non-Uniformly Bounded Transition Rate Under Discounted Criterion. *Acta Mathematicae Applicandae Sinica* 20, 196–208.

Ye L., Guo X.P. and Hernández-Lerma O. (2006). Existence and Regularity of Nonhomogeneous $Q(t)$-Processes under Measurability Conditions. Preprint.

Yosida K. (1980). *Functional Analysis, Sixth Edition.* Springer.

Yushkevich A.A. (1973). On a Class of Strategies in General Markov Decision Models. *Theory of Probability and Its Applications* 18, 777–779.

Yushkevich A.A. (1977). Controlled Markov Models with Countable State and Continuous Time. *Theory of Probability and its Applications* 22, 215–235.

Yushkevich A.A. (1994). Blackwell Optimal Policies in a Markov Decision Process with a Borel State Space. *Mathematical Methods of Operations Research* 40, 253–288.

Yushkevich A.A. (1997). Blackwell Optimality in Continuous in Action Markov Decision Processes. *SIAM Journal on Control and Optimization* 35, 2157–2182.

Yushkevich A.A. and Feinberg E. A. (1979). On Homogeneous Markov Model with Continuous Time and Finite or Countable State Space. *Theory of Probability and its Applications* 24, 156–161.

DISCUSSION

Xi-Ren Cao and Junyu Zhang

Hong Kong University of Science and Technology, Hong Kong

Markov decision processes (MDPs), also known as Markov controlled processes, stochastic problems, Markov decision problems, or stochastic dynamic programming, with applications in engineering, economics, operations, statistics, resource management, queueing systems and control of epidemics, etc, are a class of stochastic control problems, which consist of five elements: decision epochs, states, actions, transition probabilities, and rewards. Choosing an action in a state at a decision epoch generates a reward and determines the state in the future through a transition probability function. Policies are prescriptions of which action to choose at every decision epoch. Decision-makers seek policies which are optimal under some criterion. A main analysis of MDPs includes:

- providing conditions under which there exists an optimal policy;
- developing algorithms for computing optimal policies;
- applying MDPs to practical problems.

These analysis depend on the decision epochs and the criterion used to compare policies. When the set of decision epochs is discrete (or continuous), the corresponding MDP is called a discrete-time (or continuous-time) MDP. Most of the existing literature is concentrated on discrete-time MDPs. However, in many real-world situations, for instance, communication engineering, queueing systems, population processes, and control of epidemics, the state processes evolve in continuous time; and so it is very natural and suitable to use continuous-time MDPs for some optimality problems of such systems. Therefore, continuous-time MDPs become one of the topics that deserve some attention. The most common and basic optimality criteria in continuous-time MDPs are the long-run expected average and discounted reward criteria. When the sets of states and actions are both finite, as shown in this paper, the existence of discounted- and average- optimal stationary policies is indeed guaranteed, and the policy iteration and linear programming algorithms for computing such optimal policies have been given. However, as shown in Dynkin and Yushkevich (1979), Puterman (1994), and Sennott (1999), an average-optimal stationary policy may not exist for the case of continuous-time MDPs with infinitely many states (i.e., denumerable continuous-time MDPs). Thus, the research on denumerable continuous-time MDPs becomes more complex. First of all, since the transition functions in continuous-time MDPs are determined by given transition rates which may be unbounded, some underlying (possibly nonhomogeneous) continuous-time Markov process may have more than one transition functions under one policy. Moreover, the expected discounted and long-run average reward values of a policy may be infinite when rewards are unbounded. These lead to two important questions:

- What conditions can guarantee the regularity of a possibly nonhomogeneous continuous-time Markov process with respect to any given policy?
- What conditions can be used to verify the finiteness of the expected discounted or average reward values of a policy?

The above two questions are at the outset of the research on continuous-time MDPs. The answers to these questions depend heavily on the structure theory of transition functions of a continuous-time Markov process, and this paper by Guo, Hernández-Lerma and Prieto-Rumeau makes a significant contribution to this research area since it gives very mild conditions that ensure the desirable results. In particular, Lemma 3.1 in this paper is fundamental since it can be used not only to answer the above questions but also give more generalized results. We now turn to the existence of discounted- or average- optimal stationary policies. The question is:

- What conditions can guarantee the existence of discounted- or average-optimal stationary policies?

Obviously, such an existence problem of optimal policies is most essential and should be first answered in MDPs. Indeed, many authors have worked on this problem and obtained many interesting results. However, most of them are restricted to the case that either transition rates or rewards are bounded. As mentioned in Bather (1976), the challenging and most difficult problem is to study the case when the transition rates and rewards are both unbounded. Recently, important development has been achieved by the authors of this paper. This survey paper by Guo, Hernández-Lerma and Prieto-Rumeau gives a deep insight and excellent overview over different approaches and variant conditions for the existence of optimal stationary policies, as well as the relationship between the approaches and conditions. In particular, many of these conditions are imposed on the preliminary data of the model of continuous-time MDPs, and so they are desired and easy to be verified.

Concerning the algorithms for computing discounted- or average- optimal stationary policies, Theorem 3.2 in this paper gives a value iteration algorithm to compute, or at least to approximate, both the optimal discounted value and a discounted- optimal stationary policy. In particular, the choice of initial value (i.e., u_0 in (3.8)) for the value iteration algorithm is very interesting since it plays a key role in this algorithm and it is also rather different from any traditional choice. Moreover, Theorem 4.2 in this paper also presents a policy iteration algorithm for computing or at least approximating both the optimal average value and an average-optimal stationary policy. Of course, convergence rates and numerable examples about the two algorithms are also worth to be further studied.

For the advanced optimality criteria such as the bias and sensitive discounted optimality criteria, the authors of this paper first illustrate a strong motivation to study these criteria, and then present many results, which include the existence and properties of optimal policies for the so-called advanced criteria (e.g., Theorems 5.1, 5.2, 5.4, 5.5 and 5.7). These results are very interesting and important contributions to the development of MDPs.

It would be interesting to note the recent works by Cao and his colleagues (2003a-2003c, 2004) on discrete-time MDPs with finite states and actions. In their works, a stochastic system is viewed as a discrete-event dynamic system, and the optimization problem is explored by using the dynamic features of a system. This view is different from the traditional approaches to MDPs, and it, therefore, leads to different insights and methods. With this view, a sensitivity-based approach is proposed to finding the optimal solutions to the discounted and average optimization problems. This approach provides a new perspective to the optimization problems and establishes a relationship among MDPs, perturbation analysis, and reinforcement learning. In particular, the paper Cao and Zhang (2006) proposes the n-bias optimality criteria, which is closely related to, and essentially equivalent to, the sensitive discounted optimality criteria; but the problem is directly defined on long-run averages without discounting, and the solution approach is intuitive clear and simple. The extension of these results to continuous MDPs is in a recent paper to be finalized, and to extend these results to Markov processes with general state spaces is certainly a worthwhile research topic.

Another point to mention is that in the above continuous-time MDPs, any choice of actions is independent to each other. We call such continuous-time MDPs standard ones. On the other hand, there are some systems in which choice of action may depend. For example, in Jackson network Dijk

(1993) each decision/action is taken only when a customer arrives at the network. A customer arrival is described by an event. Optimality problems for such Jackson network cannot be described as standard continuous-time MDPs, because when an event occurs (a customer arrives), the system can be in many different states, and therefore, an action may affect the transition probabilities of many states. The optimality problems for such systems are called event-based optimization of Markov systems; see for instance, Cao (2005) for discrete-time Markov chains. Based on the framework of continuous-time MDPs and the idea of event-based optimality, we will add the following question which may be interesting for the future.

- To extend the event-based optimality problems to continuous-time Markov processes with general state spaces, and then find optimality conditions and algorithms for so-called event-based optimal policies.

In summary, this paper by Guo, Hernández-Lerma and Prieto-Rumeau gives an excellent overview over the most interesting results with respect to the discounted, average and advanced criteria in continuous-time MDPs. This is an important source for this challenging research area on continuous-time MDPs, and may open up many possibilities for further study on other related important problems such as stochastic games, minimax control, and event-based optimization. The works by Cao et al. provide an alternative way to the optimization of stochastic systems.

References

Bather J. (1976). Optimal Stationary Policies for Denumerable Markov Chains in Continuous Time. *Advances in Applied Probability* 8, 148-155.

Cao X.-R. (2003a). Semi-Markov Decision Problems and Performance Sensitivity Analysis. *IEEE Transactions on Automatic Control* 48, 758–769.

Cao X.-R. (2003b). A Sensitivity View of Markov Decision Processes and Reinforcement Learning. In: Gong W. and Shi L. (eds.), *Modeling, Control and Optimization of Complex systems.* Kluwer, 261–283.

Cao X.-R. (2003c). From Perturbation Analysis to Markov Decision Processes and Reinforcement Learning. *Discrete Event Dynamic Systems* 13, 9-39.

Cao X.-R. (2004). The Potential Structure of Sample Paths and Performance Sensitivities of Markov Systems. *IEEE Transactions on Automatic Control* 49, 2129–2142.

Cao X.-R. (2005). Basic Ideas for Event-Based Optimality of Markov Systems. *Discrete Event Dynamic Systems: Theory and Applications* 15, 169–197.

Cao X.-R. and Zhang J.Y. (2007). The nth-Order Bias Optimality for Multichain Markov Decision Processes. *IEEE Transactions on Automatic Control* (to appear).

Dijk N.V. (1993). *Queueing Networks and Product Forms: A System Approach.* Wiley.

Dynkin E.B. and Yushkevich A.A. (1979). *Controlled Markov Processes.* Springer.

Puterman M.L. (1994). *Markov Decision Processes.* Wiley.

Sennott L.I. (1999). *Stochastic Dynamic Programming and the Control of Queueing Systems.* Wiley.

Qiying Hu
Shanghai University, China

This paper gives a good survey of recent results on continuous-time Markov decision processes (CTMDPs) with countable state space. The paper focuses on unbounded transition rates and unbounded reward rates. The authors analyzed first the two main optimality criteria for infinite-horizon CTMDPs, i.e., the discounted reward and the average reward optimality criteria. Then they paid their attention to some “advanced” optimality criteria, including the bias, sensitive discount, and Blackwell optimality criteria.

The main steps to analyze the first two optimality criteria in the survey are as follows. First, some assumptions are presented. Then, the standard results are shown under the assumptions. Here, the standard results include that (a) the optimality equations are true, and (b) any stationary policy achieving the supremum or ϵ-supremum of the optimality equation is an optimal or ϵ'-optimal policy. Here, ϵ' depends on ϵ and tends to zero whenever ϵ tends to zero.

Discounted Reward Optimality

The authors showed first that the expected discounted total reward $V_\alpha(i, \varphi)$ is well defined with finite norm with respect to $w(i)$ under Assumptions A