J. Indian Inst. Sci., Jan.–Feb. 1995, 75, 5–24.

Recent trends in Markov decision processes

VIVEK S. BORKAR[1] AND MRINAL K. GHOSH[2]
[1]Department of Electrical Engineering and [2]Department of Mathematics, Indian Institute of Science, Bangalore 560 012

Received on September 10, 1993.

Abstract

Markov decision processes provide a rigorous mathematical framework for sequential decision making under uncertainty. In recent years, the field has seen explosive activity because of new application areas thrown up by advances in technology. These have not only stretched the limits of the existing theory but have also brought about novel methodologies to handle problems that do not fit the existing theoretical constructs. The present survey gives a short tutorial introduction to Markov decision processes and briefly outlines the thrust areas in this field.

Keywords: Markov decision processes, optimal control, dynamic programming, control under partial information, applications of MDPs.

1. Introduction

Markov decision processes (MDPs for short) are a popular paradigm for sequential decision making under uncertainty. An offspring of the operations research boom of the post-World War years, it quickly blossomed into a major subdiscipline not only of operations research, but also of control engineering and mathematical statistics. By seventies, it already accounted for a vast number of articles, texts and surveys. But it did not get fossilized like some of its siblings from the boom years because of the continuous input of new problems thrown up by the emerging application areas. In recent years, such an impetus has come from the technological advances in communication networks and flexible manufacturing systems. In this survey we hope to give a flavour of some of the recent developments in MDPs, beginning with a brief tutorial introduction to the subject. At this juncture, we must warn the reader that this survey is by no means exhaustive, but aims to serve mainly as a pointer to this field. We must also admit to the unavoidable bias in favour of topics that we ourselves have been involved with in recent years.

The paper is organized as follows: Section 2 gives a brief account of the classical theory, most notably of dynamic programming and its consequences, and of computational techniques for MDPs. Even within the classical framework, many issues remain open and much of the ongoing activity remains firmly within the classical fold. The first subsection of Section 3 gives a brief overview of some such issues. The second subsection surveys some 'nonclassical' problems that have attracted a lot of attention lately. These include problems with nonclassical costs and multiobjective problems. Problems with partial information about the state of the process or with model uncertainty merit a separate section. Section 4 surveys these and related issues. Section 5 briefly mentions some applications areas and the closely related areas of continuous time stochastic control and stochastic games.

2. Classical theory

2.1. *Preliminaries*

An MDP is a random process $\{X_n, n = 0, 1, 2, \dots\}$ (where n is the discrete time index) taking values in a discrete (finite or countably infinite) state space S, with an evolution law we shall presently describe. Without any loss of generality, we label S as $\{0, 1, 2, \dots\}$. If the process is at state $i \in S$ at time n, it moves to $j \in S$ at time $(n+1)$ with probability $p(i, j, u)$, where u is the 'action' or 'control' parameter chosen by a controller in the background at time n. This usually takes values in a finite set or a closed bounded subset of an euclidean space (more generally, a compact metric space) denoted by U. The 'transition probability function' p is taken to be continuous and clearly satisfies

$$p(i, j, u) \in [0, 1], \quad \sum_k p(i, k, u) = 1, \quad i, j \in S, \quad u \in U.$$

Obviously, the controller is constrained to choose a control based only upon his observations up to that time, possibly involving some independent randomization (*e.g.*, he may choose to toss a coin to decide between two alternatives), but never anticipating the future trajectory of the process. At each time he receives a reward or pays a cost that depends on the current state and his choice of control. The problem then is to maximize the overall reward or minimize the overall cost. The control problems are classified according to how the word 'overall' is interpreted. We return to this classification following some illustrative examples. These are oversimplified caricatures of real-life situations, but should suffice to convey the spirit of the matter. We shall use letters $f, g, h, \dots$ to denote 'some function of ...'.

Example 1 (*Inventory control*)

A storage facility has a stock of X_i units of a certain good at time i, acquiring r_i additional units thereof and then supplying $\min(d_i, X_i + r_i)$ to the customers when confronted with a demand for d_i units. Assuming that $\{d_i\}$ are independent and identically distributed (i.i.d.) nonnegative integer-valued random variables, $\{X_i\}$ is an MDP obeying the equation

$$X_{i+1} = X_i + r_i - \min(d_i, X_i + r_i), \quad i \geq 0.$$

The cost at time i is the sum of the acquisition cost $f(r_i)$, the storage cost $g(X_i + r_i)$ and the penalty for any shortfall, given by $h((X_i + r_i - d_i)^-)$. The last-mentioned is usually much larger than the rest when $d_i > X_i + r_i$ and zero when not.

Example 2 (*Control of competing queues*)

A communication channel receives packets from two sources at different rates, which are either transmitted or queued up in distinct queues. The channel can transmit only one packet at a time and the 'control' variable is the decision as to which queue to serve. The 'state' now is the pair of queue lengths and the cost a function of the weighted sum thereof, dictated by the relative priority given to the two sources.

Example 3 (*Machine scheduling*)

A factory has M machines $m_1, \dots, m_M$ of different ratings to manufacture a common perishable good. Machine m_i, $1 \le i \le M$, has three possible states (a_i, b_i, c_i), where a_i = functional but inoperative, b_i = operative, c_i = malfunctioning. When in a_i, the decision is whether to switch it on incurring a 'start-up cost' of C_i units and moving thereafter to b_i with probability 1, or to remain inoperative, *i.e.*, in a_i with probability 1 at zero cost. When in b_i, the decision is whether to switch it off and move to a_i at zero cost, or to remain operative incurring an 'operational cost' of D_i units and then moving to c_i with probability $p_i \in (0, 1)$ or remaining in b_i with probability $1 - p_i$. When in c_i, the decision is to either try to repair the machine at 'repairing cost' R_i and then move to a_i with probability $q_i \in (0, 1)$ or remain in c_i with probability $1 - q_i$, or to not repair, incurring zero cost and remaining in c_i with probability 1. When in b_i, m_i produces r_i units of the manufactured goods, nil in either a_i or c_i. There is a demand for d_n units at time n, where $\{d_i\}$ are i.i.d. nonnegative integer-valued random variables. Letting X_{in} = the output of machine i at time n ($= r_i$ if it is in b_i, zero otherwise), one pays a wastage cost of $f(\sum_{i=1}^{M} X_{in} - d_n)$ when $\sum_{i=1}^{M} X_{in} > d_n$ and a shortfall cost of $g(d_n - \sum_{i=1}^{M} X_{in})$ when $d_n > \sum_{i=1}^{M} X_{in}$, This problem can be formulated as an MDP. The important observation to make is that though each machine functions independently, the decision variables may depend on the current states of all M machines at any given time.

Returning to the mathematical formalism, let S and U be the 'state' and 'control' spaces as above, with $p = S \times S \times U \to [0, 1]$ the transition function. More generally (as suggested by the above example), one may consider a different control space U_i for each $i \in S$, with $p(i, ., .) : S \times U_i \to [0, 1]$. In other words, the nature of the decision variables depends on the current state of the process. This can, however, be reduced to the former set-up by replacing each U_i by $U = \Pi_i U_i$ and the corresponding $p(i, j, .)$, $j \in S$, by their composition with the projection map $U \to U_i$. Introduce the following notation: for a metric space X, $P(X)$ is the space of probability measures on X with the topology of weak convergence[1]. A (control) policy $\{\pi = \pi_0, \pi_1, \dots\}$ is a sequence of (measurable) maps $\pi_n : (S \times U)^n \times S \to \mathcal{P}(U)$, $n \ge 0$. Thus, π_n takes as its argument the state sequence till $n + 1$ and the control sequence till $n - 1$ these together constituting the 'history' h_n at time n and yields a probability measure on U according to which one picks the control Z_n (say) at time n. That is, the conditional distribution of Z_n given $h_n = [X_0, Z_0, \dots, X_{n-1}, Z_{n-1}, X_n]$ is $\pi_n(h_n)$. This allows the controller full use of observations up to n as well as the use of an additional randomization device (such as tossing a coin) for picking the control. Summarizing,

$$P(X_{n+1} = j/h_n, Z_n) = p(X_n, j, Z_n), \quad j \in S,$$

$$P(X_{n+1} = j/h_n) = \int p(X_n, j, u)\, \pi_n(h_n)\, (du), \quad j \in S.$$

One often expects or seeks an optimal policy within certain subclasses of polices, such as those which depend only on the current state and time or only the former and/or do not require any randomization. Call a policy a Markov-randomized policy if, for $n \geq 0$, $\pi_n(h_n) = v(X_n, n)$ for some $v : S \times \{0, 1, \dots\} \to \mathcal{P}(U)$ and a Markov deterministic policy if in addition $v(i, m)$ is concentrated at a single point for each i, m. Call it a stationary randomized policy if $\pi_i(h_n) = v(X_n)$, $n \geq 0$, for some $v : S \to \mathcal{P}(U)$ and a stationary policy if in addition $v(i)$ is concentrated at a single point in U for each $i \in S$. By abuse of terminology, the foregoing are sometimes identified with the function v in question. The implications of these definitions should be clear: all four classes do not require the knowledge of the history up to time $n-1$. The first two require an explicit time count, the rest do not. The first and the third require extraneous randomization, the others do not. Let Π denote the set of all policies.

Note that maximizing a reward is the same as minimizing a cost set equal to its negative. Thus, we shall consider only the minimization problems henceforth. Let $c : S \times U \to R$ be a bounded continuous 'running cost' function, *i.e.*, $c(X_n, Z_n)$ is the cost paid at time n, $n \geq 0$. Let $\mu \in \mathcal{P}(S)$ be the distribution of X_0 and π the policy in use. Let $N \geq 1$, $\beta \in (0, 1)$. Some standard ways of defining the 'overall' cost are:

Finite horizon cost

$$J_N(\mu, \pi, c) = E\left[\sum_{n=0}^{N-1} c(X_n, Z_n)\right].$$

Discounted cost

$$J_\beta(\mu, \pi, c) = E\left[\sum_{n=0}^{\infty} \beta^n c(X_n, X_n)\right].$$

Total (undiscounted) cost

$$J(\mu, \pi, c) = E\left[\sum_{n=0}^{\infty} c(X_n, Z_n)\right] \quad \text{(possibly} \pm\infty \text{ or undefined).}$$

Average (or 'ergodic') cost

$$J(\mu, \pi, c) = \limsup_{N \to \infty} \frac{1}{N} E\left[\sum_{n=0}^{N-1} c(X_n, Z_n)\right].$$

In each case, $\pi^* \in \Pi$ is said to be optimal if it attains the minimum of the cost over Π for a prescribed μ, and ε-optimal if it is within ε thereof for a prescribed $\varepsilon > 0$. The main aims of Markov decision theory are to establish the existence of an optimal (or failing that, an ε-optimal) control in a prescribed class, to characterize it via accessible necessary and sufficient conditions and to develop computational schemes for computing it. In the next subsection, we study the powerful dynamic programming heuristic initiated by Bellman and others, which is the principal tool in accomplishing this programme.

3. Dynamic programming

Dynamic programming is an ideal tool for sequential decision problems that are made up of several stages (which is always so, by definition) with the total cost being a composite of per stage 'running cost'. Crudely put, the dynamic programming principle says that the minimum cost to go from a stage on is the minimum of the sum of the cost at that stage and the minimum cost to go from the next stage on. The important point to note is the 'backward recursion' implicit in this statement. We illustrate its use in case of the finite horizon problem. For $0 \le n < N$ and $i \in S$, define

$$V(i,n) = \inf E\left[\sum_{m=n}^{N-1} c(X_m, Z_m) / X_n = i\right],$$

where the infimum is over all admissible choices of $\{Z_n, \ldots, Z_{N-1}\}$. Thus, $V(i, n)$ is the 'minimum cost to go' at time n if you are at state i. Clearly, $V(i, N) = 0$ for all i. The dynamic programming principle now leads to

$$\begin{aligned} V(i,n) &= \min E[c(X_n, Z_n) + V(X_{n+1}, n+1)/X_n = i] \\ &= \min_u \left[c(i,u) + \sum_j p(i,j,u) V(j, n+1)\right], \quad n < N, \qquad (\dagger) \\ V(i, N) &= 0. \end{aligned}$$

It is not difficult to prove this rigorously. One can solve this system of equations backwards in time to find its unique solution V (called the 'value function'). What is more, for a chain $\{X_n\}$ controlled by $\{Z_n\}$, we have

$$V(X_n, n) \le c(X_n, Z_n) + E[V(X_{n+1}, n+1)/X_n, Z_n], \quad n \ge 0.$$

Iterating, taking expectations and using the definition of V one sees that $\{Z_n\}$ is optimal if and only if the above equality is an equality with probability 1 for each n. It follows that if $u = v(i, n)$ attains the infimum on the right-hand side of ($\dagger$), then the Markov deterministic policy $Z_n = v(X_n, n)$, $n \ge 0$, is optimal regardless of the initial law. For the discounted problem, one similarly has

$$V(i) = \inf_{\{Z_n\}} E\left[\sum_{n=0}^{\infty} \beta^n c(X_n, Z_n) / X_0 = i\right], \quad i \in S,$$

satisfying

$$V(i) = \inf_u [c(i,u) + \beta \sum_j p(i,j,u) V(j)], \quad i \in S.$$

Furthermore, if $u = v(i)$ attains the infimum on the right, the stationary deterministic policy $Z_n = v(X_n)$, $n > 0$, is optimal for any initial law. The solution $V(.)$ of this system of equations is unique for bounded $c(.,.)$ (This follows easily from the 'contraction mapping principle'.)

The total and ergodic cost criteria are much more difficult to handle. In the former case, the cost can be infinite or undefined. When the minimum cost is finite, the existence of an optimal stationary deterministic policy was proved by Ornstein[2]. The dynamicprogramming equation, when justified, corresponds to $\beta = 1$ in the above. In the caseof the ergodic control problem, the difficulty arises because the cost suppresses all effects of finite time behaviour and depends only on long-run averages. Hence, the dynamic programming heuristic cannot be directly applied. In a major breakthrough, Howard[3] derived the dynamic programming equations for this problem by treating it as a 'vanishing discount' (*i.e.*, $\beta \to 1$) limit of the discounted cost problem in a suitable sense. These are:

$$\rho + V(i) = \inf_{n} [c(i,u) + \sum_{j} p(i,j,u)\, V(j)], \quad j \in S.$$

They are solved for the pair $(\rho, V(\,.\,))$, where ρ turns out to be the minimum cost, attained by the stationary deterministic control $v(.)$, for which $v(i)$ attains the infimum on the right. All this, however, presupposes the well-posedness of this system of equations, which does not come by easily. In fact, the early work[3,4] on this problem uses very stringent conditions, such as finite S or 'strong uniform rescurrence' condition. More on this later.

Dynamic programming equations form a basis for most computational schemes for MDPs. We sketch below the three archetypical schemes in the case of the discounted problem.

(i) *Value iteration*

In this scheme, one starts with a guess $V_0(\,.\,)$ for $V(\,.\,)$ and improves it through successive iterations

$$V_{n+1}(i) = \inf_{u} [c(i,u) + \beta \sum_{j} p(i,j,u) V_n(j)], \quad i \in S.$$

Under suitable conditions, V_n's converge to $V_\infty = V$ and provide an approximation thereof for large n. An optimal or near-optimal control can be constructed by performing the minimization above for $n = \infty$, n large, respectively.

(ii) *Policy iteration*

Start with a guess $v_0 : S \to U$ for the optimal stationary deterministic policy $v : S \to U$ and improve it successively as follows: At step n, find $V_n(\,.\,)$ by solving

$$V_n(i) = c(i, v_n(i)) + \beta \sum_{j} p(i,j,v_n(i)) V_n(j), \quad j \in S,$$

and find $v_{n+1} : S \to U$ such that for $i \in S$, $v_{n+1}(i)$ minimizes

$$u \to c(i,u) + \beta \sum_{j} p(i,j,u)\, V_n(j).$$

Under suitable conditions, v_n is near-optimal for large n.

(iii) *Linear programming*

Let U be finite, If $W : S \to R$ satisfies

$$W(i) \le \inf_u [c(i,u) + \beta \sum_j p(i,j,u) W(j)], \quad i \in S,$$

it is easy to see that $W \le V$ termwise. Thus, V solves the linear program

$$\text{maximize} \sum_j a_j W(j) \text{ s.t.}$$

$$W(i) \le c(i,u) + \beta \sum_j p(i,j,u)\, W(j), \quad i \in S, \ u \in U,$$

where $a_i \in (0, 1)\ \forall_i$ and $\sum_i a_i = 1$. The dual linear program is

$$\text{minimize} \sum_{i,u} x(i,u) c(i,u) \text{ s.t.}$$

$$\sum_{u \in U} x(i,u) - \beta \sum_{j \in S} \sum_{u \in U} p(j,i,u) x(j,u) = a_i, \quad i \in S,$$

$$x(i,u) \ge 0 \quad \forall\, i, u.$$

If $x(.\,,.)$ solves this problem, the stationary randomized strategy that picks in state i control u with probability $x(i,u)/(\sum_b x(i,b))$ is optimal.

This concludes our survey of the classical theory. It should be remarked that these results have not been presented at the greatest level of generality and some generalizations are immediately possible. To mention just one, U can be allowed to be unbounded by ensuring that the running cost c penalizes 'large' u. For further reading, some excellent texts are those by Dynkin and Yuskevich[5], Kumar and Varaiya[6], Ross[7], Tijms[8] and Whittle[9, 10]. See also Puterman's survey[11] for an excellent account of the algorithmic aspects.

4. Recent developments

4.1. *Extensions of classical theory*

Much of the ongoing work in MDPs remains firmly within the folds of the classical framework described above. Here we briefly list some of the dominant strands therein.

(i) *Generalizations*

A considerable effort in MDPs continues to be directed towards extending known results to more general situations. This is particularly true for the difficult problems of total cost[12,13] and ergodic cost[14]. The latter, in particular, attracts much attention due to its popularity with the communication networks community. A major advance in ergodic control has been a new convex analytic approach based on a characterization of limit points of empirical processes associated with the state and control sequences[15-17]. This approach completely circumvents the vanishing discount argument. The latter in turn has been extended to more general situations[18]. It should be added that the 'convex analytic' approach, which treats the 'dynamic' control problem as a 'static' optimization problem on a set of suitably defined 'occupation measures', can also be applied fruitfully to other cost criteria to gain useful insight[17, 19].

Another direction for generalization has been towards more general state spaces, notably Borel spaces, *i.e.*, Borel subsets of complete separable metric spaces[20, 21]. These problems lead to difficult measurability issues and have had a fruitful relationship with descriptive set theory[22].

(ii) *Algorithms*

The computational schemes described above continue to be refined, modified and tuned for special classes of problems and their convergence properties analysed[23,24]. Also, special algorithms are developed for specific problems[25]. Two important developments in this context are the development of parallel algorithms[26] and a computational-complexity-based study of MDPs[27,28].

(iii) *Special structures*

Several specific classes of MDPs have an additional structure such as the convexity of the value function, which allows one to say something more about the structure of the optimal policy. There have been quite a few success stories of this sort, the most prominent being the discovery of various index rules. These date back to the discovery of the Gittins index[29] for multiarmed bandit problems. This class of problems can be briefly described as follows: One has a finite family of Markov chains called 'bandits'. If the ith bandit is at some state x and is selected to be played, a reward of $R(x)$ is received and the bandit remains active over a time period of $T(x)$, ending up in a random state y. At this point one selects a new (possibly the same) bandit to be played. Under the usual cost criteria, the optimal policy for such problems was shown to be based on simple comparison of certain indices (the Gittins indices) associated with the states. Specifically, one picks the bandit whose state has the highest index. The original work has undergone many simplifications and refinements[30,31], including extentions to arm-acquiring bandits[32], restless bandits[33] and so on. There is also work on computation of these indices[34]. A major development in this domain is the work of Klimov[35] on a class of controlled networks of queues.

Another important type of special structure often sought is a switching policy wherein the state space splits into two or more (but not too many) connected sets such that the optimal policy is to switch between corresponding finitely many choices of controls whenever the process crosses the boundaries between these sets. Once again there are important instances of this from controlled queues[36].

Finally, one sometimes detects other structural aspects like hysteresis[37] or 'monotonicity' in a suitable sense[38].

(iv) *Miscellaneous*

In addition to the foregoing, there is also work on sensitivity analysis[39], perturbation analysis[40,41], comparison of policies[42], easily computed bounds on performance[43] and so on. Approximation of complex MDPs is a major issue and in this direction one should mention the work on state space reduction[44] and singular perturbation analysis of chains whose transition probabilities exhibit different 'scales'[45]. More recently, Koehler[46] has worked on general optimization problems with formal structural similarity to MDPs and Dutta[47] has studied the asymptotics of discounted cost problems in the vanishing discount limit.

4.2. *Nonclassical problems*

In this subsection we consider some problems that have attracted much attention lately and are distinguished by the fact that they do not completely fit into the classical framework described above.

(i) *Multiobjective MDPs*

Suppose we wish to minimize simultaneously n distinct cost functionals (comprising a 'vector cost') of the same type (*e.g.*, all ergodic or all discounted with the same discount factor β). This is not in general possible and one has to extend the concept of a 'solution'. The minimal natural requirement then is that it be a policy such that no other policy gives a cost vector which is at least as good in all the components and strictly better in at least one. Policies satisfying this are said to be Pareto-optimal. Clearly, a policy that minimizes a strict convex combination of the costs will be Pareto-optimal. Conversely, each Pareto-optimal policy is obtainable as an optimal policy for a convex combination, not necessarily strict (except in the finite state/action case), of given costs[48]. Parametric linear programming can be used for this problem[49]. Another approach[50] is to cast this problem as a specially structured partially observed problem (see the next section) and treat it as a special case thereof.

A general scheme for converting a vector cost to a scalar cost is to take as the cost a scalar-valued function of the original costs that is monotone-increasing in each argument. Convex combinations mentioned above yield one such 'utility function'. Another is the distance in $\mathcal{R}^m$ of the cost vector from the 'utopian point' $[u_1, \ldots, u_m]$, where $u_i =$ the minimum of the ith cost functional over all admissible policies. Minimizing this gives a unique Pareto-optimal point which in finite state/action case can be found through a combined linear–quadratic program[48].

(ii) *Constrained problems*

Another way of handling multiple costs is to minimize one of them while keeping the rest within the prescribed bounds. Under reasonable conditions, such problems admit a 'Lagrange multiplier' formulation; moreover, one can show[51–53] that with m independent constraints, the optimal stationary randomized policy requires at most m randomizations (*i.e.*, randomization between m_i controls at state i subject to $\sum_i (m_i - 1) \le m$). In ergodic case, 'pathwise' constraints of the type

$$P\left(\limsup_{n\to\infty}\frac{1}{n}\sum_{m=0}^{n-1}k(X_m, Z_m)\le a\right)=1$$

have also been considered[54]. Constrained problems are of great interest in controlled queueing networks, where further structure can sometimes be found[55].

(iii) *Weighted cost criteria*

One may wish to combine the advantages of different cost criteria by combining them. For example, one may want to account for both short- and long-term costs by considering a weighted sum of the discounted and ergodic costs. Such problems have attracted a lot of attention in recent times[56,57], culminating in the following result[58]: Let $c_1, \ldots, c_m, c$ be the running cost functions as before and $\alpha_1, \beta_1, \ldots, \beta_m \in (0, 1)$. Consider the cost

$$(1-\alpha)\sum_{i=1}^{m}(1-\beta_i)E\left[\sum_{n=0}^{\infty}\beta_i^n c_i(X_n, Z_n)\right]+\alpha\limsup_{n\to\infty}\frac{1}{n}E\left[\sum_{m=0}^{n-1}c(X_m, Z_m)\right].$$

Such problems need not always have an optimal policy nor need a policy optimal among Markov deterministic policies be optimal overall. One can, however, find for each $\varepsilon>0$ an ε-optimal policy with the following structure: use a policy π' up to a prescribed time N (dependent on ε) and another policy π'' thereafter, where π'' is optimal for the ergodic problem with running cost c and π' is optimal for the discounted problem with discount factor β_1 and the time-dependent running cost

$$\sum_{k=1}^{m}\left(\frac{1-\beta_k}{1-\beta_1}\right)\left(\frac{\beta_k}{\beta_1}\right)^n c_k(i, u).$$

(iv) *Overtaking criterion*

Introduced first in economics literature[59,60], this criterion may be considered a refinement of the ergodic criterion. Here one requires the policy to be not only ergodic-optimal but also finite-horizon-optimal for all sufficiently long finite horizons. Under suitable hypotheses, the overtaking optimal policies can be shown to be those ergodic-optimal stationary deterministic policies that further maximize $\lim_{n\to\infty} E[V(X_n)]=\sum_i \pi(i)V(i)$, where V is the ergodic value function and π, the stationary distribution under the given stationary deterministic policy[58,61].

(v) *Variance-sensitive control*

The standard cost functionals aim at minimizing some averages (or limits thereof) of the type $E[F(\psi)]$ for $\psi=[X_0, Z_0, X_1, Z_1, \ldots]$, $F:(S\times U)^\infty\to R$. These do not account for the variability of the actual sample pathwise cost around this average. This motivates variance-sensitive control where we add to the cost a 'variance' term

$$aE[(F(\psi)-E[F(\psi)])^2]$$

for some $a > 0$. More generally, one may consider the cost $E[h(F(\psi), E[F(\psi)])]$, where h is the 'variability function' ($h(x, y) = x + a(x - y)^2$ in the above instance). A variant for the ergodic case is

$$\limsup_{N\to\infty} \frac{1}{N} E\left[\sum_{n=b}^{N-1} h\left(c(X_n, Z_n), N^{-1}E\left[\sum_{n=0}^{N-1} c(X_n, Z_n)\right]\right)\right],$$

which is the most extensively studied case in the literature[62-64]. The convex analytic framework mentioned in Section 3.1 allows one to establish the existence of an optimal stationary deterministic policy in some cases. This class of problems is important in manufacturing[65].

5. Problems with partial information

Problems with partial information are mainly of two types: those involving partial (or 'noisy') observations of the state of the process and those involving model uncertainty, *i.e.*, ignorance about the transition probability function p. (More complicated situations can arise and will be briefly mentioned later.)

5.1. *State uncertainty*

This concerns the case where there is another countably valued ($\bar{S}$-valued, say) observation process $\{Y_n\}$ with the joint evolution of $\{X_n\}$, $\{Y_n\}$ governed by

$$P(X_{n+1} = i, Y_{n+1} = j/h_n, Y_0, \ldots, Y_n, Z_n) = p(X_n, i, j, Z_n), \quad n \geq 0,$$

for a suitable transition function : $S \times S \times \bar{S} \times U \to [0, 1]$. The problem is to control $\{X_n\}$ with one of the standard cost criteria, but with Z_n constrained to depend only on $\{Y_i, i \leq n\}$ at time n (plus, possibly, some independent randomization). This problem can be converted to a problem with complete observations by moving over to a new state space, *viz.*, the space $\mathcal{P}(S)$ of probability measures on S. The 'state' at time n now is the conditional law η_n of X_n given $\{Y_i, Z_i, i \leq n\}$. This is recursively computed by the discrete nonlinear filter

$$\eta_{n+1} = F(\eta_n P_n), \quad n \geq 1,$$

with $\eta_0 =$ the law of X_0, $P_n =$ the matrix $[[p(i, j, Y_{n+1}, Z_n)]]_{i,j \in S}$, and $F =$ the map $[x_1, x_2, \ldots] \to [x_1/a, x_2/a, \ldots]$, $a = \sum_i |x_i|$. (Here η_n is being written as a row vector for each n.) The 'running cost' correspondingly becomes $\bar{c}(\eta_n, Z_n) = \sum_i \eta_n(i)c(i, Z_n)$. This is then a special case of MDPs on a general (Borel) state space and can be handled accordingly[6,66]. Of special interest is the ergodic problem, which continues to elude a satisfactory treatment. The existence of optimal stationary randomized policies can be derived by analysis of pathwise empirical measures as in the completely observed case[17], but their characterization through suitable 'dynamic programming' equations is hard to come by. Problems arise because the process can have a complicated control-dependent ergodic decomposition under stationary randomized controls. Platzman made some progress on this problem under very restrictive 'reachability' conditions[67]. More recently, seemingly more general but intuitively unappealing conditions have been used[68] to justify the dynamic programming equations and have been verified for an important special case.

An alternative state process sometimes used is the 'unnormalized conditional law' $\{v_n\}$ given by[6]

$$v_{n+1} = v_n P_n, \quad n \geq 0, \quad v_0 = \pi_0 .$$

For $n \geq 0$, v_n is a finite nonnegative measure on S which yields η_n on normalization to a probability measure. The advantage here is the linear dynamics, also leading to some simplification in the dynamic programming equations. A variant of this with a slightly different linear dynamics coupled with a 'measure transformation' leads to a linear dynamics 'driven' by $\{Y_n\}$ which become i.i.d. under the new measure[17]. This has some analytic advantages.

A recent related development[69,70] is to view the nonlinear filter (without control) as an iteration of random maps and use the theory of the latter to analyse its attractors. It will be interesting to extend these results to the controlled case and to explore their implications for the ergodic control problem.

Finally, computational aspects of control under partial observations have been investigated[71].

5.2. *Model uncertainty*

This refers to the situation when the system model is unknown and has to be inferred from the observed state while simultaneously controlling the process. Thus, the control process has to play the dual role of optimizing the system while probing it so as to reveal its structure.

There are two broad philosophies for handling such problems. The first is that of adaptive control, wherein one explicitly estimates the model 'on-line' using a suitable statistical scheme and uses at each time that control which would be the optimal choice were the current estimate the true model. This is the 'self-tuning' or 'certainty equivalence' control. For the sake of completeness, we mention another standard paradigm for adaptive control, quite popular in linear control systems literature, but for some reason unexplored for MDPs. This is the 'model reference adaptive control', wherein one feeds the control input to the system and to a putative model based on which the control is derived and whose parameters are updated based on the error signal given by the difference between the system output and the model output.

The second approach is that of 'learning control'. The broad philosophy here is to make probing moves in the control or the parameter space and, depending upon whether the performance is improved or degraded, either reinforce or discourage future moves in that direction.

In the literature, what we call adaptive and learning control are sometimes referred to as indirect and direct adaptive control.

Self-tuning control for MDPs was pioneered by Mandl[72] for the finite case. For a class of estimation schemes that include maximum likelihood, he showed that for a parameterized model set containing the true model, the parameter estimates converge to the true parameter and the ergodic cost to the optimal, with probability one. Similar results for the discounted case followed[73]. The main problem with these was a strong 'identifiability condition' which ensures complete model discrimination under any arbitrary policy. In the absence of this, one may end up in a trap where one uses a nonoptimal policy that consistently leads to a wrong choice of the parameter estimate (by virtue of not distinguishing it from the true parameter), which in turn leads to the choice of the said policy[74]. (More complicated scenarios are possible.) Subsequent works[75,76] relaxed this condition by taking recourse to randomization of parameter estimates or controls. A significant development to follow (for the finite case) was the introduction of an explicit, asymptotically negligible cost bias in the estimation scheme which favours parameters with lower optimal costs[77,78]. This leads to the optimal cost even when the parameter estimates do not converge. These works have been recently extended to a broad class of MDPs[79,80].

The foregoing used maximum-likelihood estimates. Other estimation schemes have also been used, such as Bayesian[81,82] or nonparametric[83]. Furthermore, algorithmic aspects have been investigated, involving a stochastic approximation algorithm for parameter estimation[84] or value iteration for the control update[85]. A more recent development of interest is the work on asymptotically efficient adaptive control schemes[86]. Extending the earlier work in this vein on bandit problems[87], this work derives a lower bound for the 'loss', *i.e.*, the difference between the actual cost and the ideal optimal, uniform with respect to whatever value the true parameter may take. The aim then is to find a policy whose loss equals this bound for every possible value of the true parameter.

In learning control, an important recent contribution is that of Wheeler and Narendra[88], who propose a decentralized learning scheme using a team of learning automata each of which uses a very simple estimation scheme to improve its policy. An alternative approach is provided by Santharam and Sastry[89], who use a stochastic neural network to implement learning in policy space. This work is in the spirit of 'Q-learning' introduced by Watkins[90,91], which can briefly be described as follows. The agent tries all state–action combinations repeatedly and evaluates which are the best overall by looking at the costs incurred. The Watkins algorithm is similar in structure to stochastic approximation and this fact was exploited by Tsitsiklis[92] to simplify its analysis and give a parallel asynchronous version.

Comparing different adaptive control and/or learning schemes is not easy and one expects different comparative merits for different problem classes. Learning schemes are cruder and therefore simpler to implement, but appear less desirable for large MDPs. Also, parametric self-tuning versus nonparametric self-tuning or learning may be expected to exhibit the 'bias-variance' dilemma[93]: Observing that the inclusion of true model in the model class under consideration is a theoretical convenience not often met in practice, one expects parametric methods to have a built-in bias because of modelling limitations, but low fluctuations around this bias. Nonparametric schemes assume less structure and should exhibit lower bias, but the variance may be high.

Finally, Araposthasis *et al* have considered joint state–parameter estimation, *i.e.*, adaptive filtering[94] and, subsequently, adaptive control under partial observations[95].

5.3. *Decentralized control*

Consider the 'team' theoretic problem of several agents trying to control a common process, but with access to different sets of observations. This is an important situation in practice, where the control is required to perform yet another function in addition to optimization and probing, *viz.*, that of signalling. The agents can use controls to signal to each other a part of their information[96]. This is a difficult problem to analyse and only a few special instances have been studied[97].

6. Conclusions

In conclusion, we briefly mention some application areas and allied disciplines.

While the traditional application areas of MDPs, like inventory control, continue to draw inputs[98,99], the area really bursting forth with activity is the area of control of queuing networks, notably in the specific application areas of flexible manufacturing systems[100] and communication networks[101, 102]. These are vast fields in themselves that merit separate full-length surveys, so we shall confine ourselves to mentioning a few salient features thereof. An important aspect of this class of problems is the frequent use of very novel techniques, distinct from dynamic programming, for solving specific problems. These include interchange arguments, forward induction and so on[102]. These (or traditional dynamic programming, for that matter) can often be combined with the special features of the problem to deduce additional structure of the policy, say a switching structure or an index rule. One also encounters here multiagent control problems with each agent seeking to optimize his own cost criterion, with a notion of overall performance in the background. Thus, considerations such as 'individual *versus* social optimality' arise[102]. Sometimes these problems are fruitfully analysed as stochastic games[103]. Finally, effort is also directed towards evaluating and comparing simple and intuitively appealing policies[104] (such as 'first come first served').

Some of the recent work in controlled queues concerns optimal scheduling of processors executing a communication protocol stack[105], admission control to queues with delayed queue length information[106] and admission control subject to a fairness criterion to compare services allotted to different queues[107].

Finally, MDPs occasionally find unexpected applications in novel problems, such as stochastic shortest path problems[108], to mention but one instance.

In this survey we have not touched the related areas of stochastic games[109] and control of continuous-time Markov processes[110]. The former entails several agents seeking to optimize their own costs or rewards, with or without cooperation and with or without additional information constraints. The field has many novel features not present in single-agent MDPs and is of great interest to economists trying to model group behaviour in economic phenomena. On the other hand, control of Markov processes in $\mathcal{R}^n$ in continuous time has formal similarity to MDPs, but has a far richer mathematical structure, a high point of which is its link with a class of nonlinear partial differential equations.

References

1. BILLINGSLEY, P. *Concergence of probability measures*, 1968, Wiley.

2. ORNSTEIN, D. On the existence of stationary optimal sratetegies, *Proc. Am. Math. Soc.*, 1969, **20**, 563–569.

3. HOWARD, R. *Dynamic programming and Markov processes*, 1960, MIT Press.

4. DERMAN, C. Denumerable state Markov decision processes-average cost criterion, *Ann. Math. Stat.*, 1966, **37**, 1545–1554.

5. DYNKIN, E AND YUSHKEVICH, Y. Y. *Controlled Markov processes*, 1979, Springer-Verlag.

6. KUMAR, P. R. AND VARAIYA, P. *Stochastic systems–estimation, identification and adaptive control*, 1986, Prentice-Hall.

7. ROSS, S. *Introduction to stochastic dynamic programming*, 1983, Academic Press.

8. TIJMS, H. C. *Stochastic modelling and analysis: a computational approach*, 1986, Wiley.

9. WHITTLE, P. *Optimization over time: dynamic programming and stochastic control*, Vol. 1, 1983, Wiley.

10. WHITTLE, P. *Optimization over time: dynamic programming and stochastic control*, Vol. 2, 1983, Wiley.

11. PUTERMAN, M. Markov decision processes. In '*Handbooks in OR and MS*', Vol. 2 (D. P. Heyman and M. J. Sobel, eds), 1990, pp 331–434, Elsevier.

12. VAN DAWEN, R. Pointwise and uniformly good strategies for dynamic programming models, *Math. OR*, 1986, **11**, 521–535.

13. FASSBENDER, M. Optimal stationary strategeties in leavable Markov decision processes *J. Appl. Prob.*, 1990, **27**, 134–145.

14. ARAPOSTHASIS, A., BORKAR, V. S., FERNANDEZ–GAUCHERNAND, E., GHOSH, M. K. AND MARCUS S. I. Discrete-time controlled Markov processes with average cost criterion – a survey, *SIAM J. Control Opt.*, 1992, **31**, 282–344.

15. BORKAR, V. S. Control of Markov chains with long-run average cost criterion. In *Stochastic differential systems, stochastic control theory and applications*. (W. Fleming and P. L. Lions, eds), IMA, Vol. 10, pp. 57–77, 1988, Springer-Verlag.

16. BORKAR, V. S. Control of Markov chains with long-run average cost criterion: the dynamic programming equations, *SIAM J. Control Opt.*, 1989, **27**, 642–657.

17. BORKAR, V. S. *Topics in controlled Markov chains*, Pitman Research Notes in Maths. No. 240, 1991, Longman Scientific and Technical.

18. SENNOTT, L. I. Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs, *Op. Res.*, 1989, **37**, 626–633.

19. BORKAR, V. S. A convex analytic approach to Markov decision processes, *Prob. Theory Related Fields*, 1988, **78**, 583–602.

20. BHATTACHARYA, R. N. AND MAJUMDAR, M. Controlled semi-Markov model under long-run average rewards, *J. Stat. Plann. Inference*, 1939, **22**, 223–242.

21. FEINBERG, E. A. On stationary strategies in Borel dynmaic programming, *Math. OR*, 1992, **17**, 392–397.

22. BERTSEKAS, D. AND SHREVE, S. *Stochastic optimal control–the discrete time case*, 1978, Academic Press.

23. HORDIJK, A. On the convergence of policy iteration in finite state undiscounted Markov decision processes: the unichain case, *Math.OR*, 1987, **12**, 163–176.

24. PORTEUS, E. Survey of numerical methods for finite Markov and semi-Markov chains, *XII Conf. on Stochastic Process Applications*, Ithaca, NY, 1983.

25. VARAIYA, P. Optimal and suboptimal stationary controls for Markov chains, *IEEE Trans. Automat. Control*, 1978, **AC-23**, 388–394.

26. BERTSEKAS, D. AND TSITSIKLIS, J. *Parallel and distributed computation–numerical methods*, Ch. 4, pp. 323–324, 1989, Prentice-Hall.

27. PAPADIMITRIOU, V. AND TSITSIKLIS, J. The complexity of Markov decision processes, *Math. OR*, 1987, **12**, 441–450.

28. TSENG, P. *Polynomial time algorithms for finite horizon, stationary Markov decision processes*, Technical Report CICS-P-65, Center for Intelligent Control Systems, MIT, Massachusetts, USA, 1988.

29. GITTINS, J. C. AND JONES, D. M. A dynamic allocation index for design of experiments. In *Progress in statistics*, Vol. 1, (J. Gani, K. Sarkadi and J. Vince, eds), 1974, pp. 161–173, North-Holland.

30. VARAIYA, P., WALRAND, J. AND BUYUKKOC, C. Extensions of the multi-armed bandit problem, the discounted case, *IEEE Trans.* 1985, **AC-30**, 426–439.

31. TSITSIKLIS, J. *A short proof of the Gittins index theorem*, Technical Report CICS P-363, Centre for Intelligent Control Systems, MIT, Massachusetts, USA, 1993.

32. WHITTLE, P. Arm acquiring bandits, *Ann. Prob*, 1981, **9**, 284–292.

33. WEBER, R. R. AND WEISS, G. On the index policy for restless bandits, *J. Appl. Prob.*, 1990, **27**, 637–648.

34. CHEN, Y. R. AND HATEHAKIS, M. Linear-programming for finite state multi-armed bandit problems, *Math. OR*, 1896, **11**, 180–183.

35. KLIMOV, G. P. Time sharing service systems, I., *Th. Prob. Appl.*, 1974, **19**, 532–551.

36. HAJEK, B. Optimal control of two interacting service stations, *IEEE Trans. Automat. Control*,1984, **AC-29**, 491–499.

37. HIPP, S. K. AND HOLZBAUR, U. D Decision processes with monotone hysteretic policies, *Op. Res.*, 1988, **36**, 585–588.

38. PITTENGER, A. Monotonicity in Markov decision processes, *Math. OR*. 1988, **13**, 65–73.

39. GLAZEBROOK, K. D. Sensitivity analysis for stochastic scheduling problems, *Math. OR*. 1987, **12**, 205–223.

40. VAN DIJK, N. M. AND PUTERMAN, M. Perturbation theory for Markov reward processes with applications to queueing systems, *Adv. Appl. Prob.*, 1988, **20**, 79-98.

41. VAN DIJK, N. M. AND PUTERMAN, M. Perturbation theory for unbounded-Markov reward processess with applications to queueing, *Adv. Appl. Prob.*, 1988, **20**, 99–111.

42. SHWARTZ, A. AND MAKOWSKI, M. Comparing policies in Markov decision processes-Mandl's lemma revisited, *Math. OR*, 1990, **15**, 155–174.

43. LOVEJOY, W. S. Policy bounds for Markov decision processes, *OP. Res.*, 1986, **34**, 630–637.

44. FORESTIER, J. P. AND VARAIYA, P. Multilayer control of large Markov chains, *IEEE Trans. Automat. Control*, 1978,**AC-23**, 298–304.

45. QUADRAT, J. P. Optimal control of perturbed Markov chains. In *Singular perturbations and asymptotic analysis in control systems* (P. Kokotovic, A. Bensoussan and G. Blankenship, eds), Lecture Notes in Control and Information Sciences No. 90, pp. 288–309, Springer-Verlag.

46. KOEHLER, G. Relationships between various Markovian decision problem classes, *SIAM Control Opt.*, 1990, **28**, 1452–1460.

47. DUTTA, P. K. What do discounted optima converge to? A theory of discount rate asymptotics in economic analysis, *J Econ. Theory*, 1991, **55**, 64–94.

48. GHOSH, M. K. Markov decision processes with multiple costs, *OR Lett.*, 1990, **9**, 257–260.

49. VISWANATHAN, B., AGGARWAL, V. V. AND NAIR, K. P. K. Multiple criteria Markov decision processes. In *Multiple criteria decision making*, (M. K. Starr and M. Zeleny, eds) Vol. 6, TIMS Studies in Management Sciences, North-Holland.

50. WHITE, C. C. AND KIM. K. M. Solution procedures for solving vector criterion Markov decision processes, *Large Scale Systems 1*, 1980, 129–140.

51. BEUTLER, F. J. AND ROSS, K. W. Optimal policies for controlled Markov chains with a constraint, *J. Math. Anal. Appl.*, 1985, **122**, 236–252.

52. ROSS, K. W. Randomized and past-dependent policies for Markov decision processes with multiple constraints, *Op. Res.*, 1989, **37**, 474–477.

53. BORKAR, V. S. Ergodic control of Markov chains with constraints – the general case, *SIAM J. Control Opt.*, 1994, to appear.

54. ROSS, K. W. AND VARADARAJAN, R. Markov decision processes with sample path constraints: the communicating case, *Op. Res.*, 1989, **37**, 780–790.

55. HORDIJK, A. AND SPIEKSMA, F. Constrained admission control to a queueing system, *Adv. Appl. Prob.*, 1989, **21**, 409–431.

56. FEINBERG , E. AND SHWARTZ, A. Markov decision models with weighted discounted criteria, *Math. OR*, 1993, to appear.

57. KRASS, D., FILAR, J AND SINHA, S. A weighted Markov decision process, *Op. Res.*, 1992, **40**, 1180–1187.

58. FERNANDEZ-GAUCHERAND, E. GHOSH, M. K. AND MARCUS, S. I. Controlled Markov processes on the infinite planning horizon: weighted and overtaking cost criteria, *Z. Op. Res.*, 1993, to appear.

59. VON WEIZSACKER, C. Existence of optimal programs of accumulation for an infinite time horizon, *Rev. Econ. Stud.*, 1965, **32**, 85–164.

60. GALE, D. On optimal development in a multisector economy, *Rev. Econ. Stud.*, 1967, **34**, 1–19.

61. LEIZAROWITZ, A. Infinite horizon optimization for finite state Markov chains, *SIAM J. Control Opt.*, 1987, **25**, 1601–1618.

62. FILAR, J., KALLENBERG, L. C. M., LEE, H. M. Variance-penalized Markov decision processes, *Math. OR*, 1989, **14**, 147–161.

63. ALTMAN, E. AND SHWARTZ, A Markov decision processes and state-action frequencies, *SIAM J. Control Opt.*, 1991, **29**, 786–809.

64. BAYKAL-GURSOY, M. AND ROSS, K. W. Variability sensitive Markov decision processes, *Math. OR*, 1992, **17**, 558–571.

65. LU, S. C. H., RAMASWAMY, D. AND KUMAR, P. R. *Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants*, preprint, 1993, Coordinated Science Lab., Univ. of Illinois at Urbana-Champaign.

66. MONAHAN, G. E. A survey of partially observable Markov decision processes: theory, models and algorithms, *Mgmt. Sci.*, 1982, **28**, 1–16.

67. PLATZMAN, L. Optimal infinite-horizon undiscounted control of finite probabilistic systems, *SIAM J. Control Opt.*, 1980, **18**, 362–380.

68. FERNANDEZ-GAUCHERAND, E., ARAPOSTHASIS, A. AND MARCUS, S. I. On the average cost optimality equation and the structure of optimal policies for partially observable Markov decision processes, *Ann. OR*, 1991, **29**, 429–470.

69. PICCIONI, M. On the asymptotic behaviour of the predictor of a binary Markov chain, *Boll. Uni. Mat. Ital. Ser. A*, 1990, **7**, 319–329.

70. ELTON, J. AND PICCIONI, M. Iterated function systems arising from recursive estimation problems, *Prob. Theroy Related Fields*, 1992, **91**, 103–114.

71. LOVEJOY, W. Computationally feasible bounds for partially observed Markov decision processes, *Op. Res.* 1991, **39**, 161–175.

72. MANDL, P. Estimation and control in Markov chains, *Adv. Appl. Prov.*, 1974, **6**, 40–60.

73. SCHÄL, M. Estimation and control in discounted dynamic programming, *Stochastics*, 1987, **20**, 51–71.

74. BORKAR, V. S. AND VARAIYA, P. Adaptive control of Markov chains I: finite parameter set, *IEEE Trans. Automat. Control*, 1979, **AC-24**, 953–957.

75. BORKAR, V. S. AND VARAIYA, P. Identification and adaptative control of Markov chains, *SIAM J. Control Opt.*, 1982, **20**, 470–489.

76. DOSHI, B. AND SHREVE, S. Strong consistency of a modified maximum likelihood estimator for controlled Markov chains, *J. Appl. Prob.*, 1980, **17**, 726–734.

77. KUMAR, P. R. AND BECKER, A. A new family of optimal adaptive controllers for Markov chains, *IEEE Trans. Automat. Control*, 1982, **AC-27**, 137–146.

78. Milito, R. and Cruz, J. B. An optimization-oriented approach to adaptive control of Markov chains, *IEEE Trans. Automat. Control*, 1987, **AC-32**, 754–762.

79. Borkar, V. S. The Kumar–Becker–Lin scheme revisited, *J. Opt. Theory Appl.* 1990, **66**, 289–309.

80. Borkar, V. S. On the Milito–Cruz adaptive control scheme for Markov chains, *J. Opt. Appl.*, 1993, **77**, 385–393.

81. Van Hee, K. *Bayesian control of Markov chains*, Math. Centrum Tracts No. 95, 1978, Math. Centrum, Amsterdam.

82. Borkar, V. S. and Mundra, S. Bayesian parameter estimation and adaptive control of Markov processes with time-averaged cost, 1993, Preprint.

83. Hernandez-Lerma, O. and Marcus, S. I. Adaptive policies for discrete-time stochastic control systems with unknown distrubance distribution, *Systems Control Lett.*, 1987, **9**, 307–315.

84. El Fattah, Y. M. Gradient approach for recursive estimation and control in finite Markov chains, *Adv. Appl. Prob.*, 1981, **13**, 778–803.

85. Jalali, A and Ferguson, M. Adaptive control of Markov chains with local updates, *Systems Control Lett.*, **14**, 1990, 209–218.

86. Agarwal, R., Teneketzis, D. and Ananthram, V. Asymptotically efficient adaptive allocation schemes for controlled Markov chains: finite parameter space, *IEEE Trans.Automat. Control*, 1989, AC-34, 1249–1259.

87. Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules, *Adv. Appl. Math.*, 1985, **6**, 4–22.

88. Wheeler, R. and Narendra., K. S. Decentralized learning in finite Markov chains, *IEEE Trans. Automat. Control*, 1986, AC-31, 519–526.

89. Santharam, G. and Sastry, P. S. A reinforcement learning neural network for adaptive control of Markov chains, preprint.

90. Watkins, C. and Dayan, P. Q-learning, *Mach. Learning*, 1992, **8**, 279–292.

91. Barto, A., Bradtke, S. and Singh, S. P. *Real-time learning and control using asynchronous dynamic programming*, Technical Report 91-57, 1991, Dept. of Computer Science, Uni. of Massachusetts.

92. Tsitsiklis, J. Asynchronous stochastic approximation and Q-learning, preprint.

93. Geman, S., Bienenstock, E. and Doursat, R. Neural networks and the bias/variance dilemma. *Neural Computation*, 1992, **4**, 1–58.

94. Araposthasis, A., Fernandez–Gaucherand, E. and Marcus S. I. Analysis of an adaptive control scheme for a partially observed Markov chain, *Proc. 29th Conf. on Decision and Control*, 1990, pp. 1438–1444, IEEE Press.

95. Araposthasis, A. and Marcus, S. I. Analysis of an identification algorithm arising in the adaptive Estimation of Markov chains, *Math. Control. Signals Systems*, 1990, **3**, 1–29.

96. Sandell, N., Varaiya, P., Athans, M. and Safonov, M. Survey of decentralized control methods for large scale systems, *IEEE Trans. Automat. Control*, 1978, AC-23, 108–128.

97. Hsu, K., and Marcus, S. I. Decentralized control of finite state Markov processes, *IEEE Trans. Automat. Control*, 1982, AC-27, 426–431.

98. FEDERGRUEN, A. AND ZIPKIN, P. An inventory model with limited production capacity and uncertain demands I: the average cost criterion, *Math. OR*, 1986, **11**, 193–207.

99. FEDERGRUEN, A. AND ZIPKIN, P. An inventory model with limited production capacity and uncertain demands II: th discounted cost criterion, *Math. OR*, 1986, **11**, 208–215.

100. GUN., L. AND MAKOWSKI, A. *Optimal production strategies for discrete-time machines subject to failures and breakdowns*, Technical Report TR-86-1, 1986, Systems Research Center, Uni. of Maryland, College Park.

101. STIDHAM, S. Optimal control of admission to a queueing system, *IEEE Trans. Automat. Control*, 1985, AC-30, 705–713.

102. WARLAND, J. *Introduction to queueing networks*, Ch. 8–9, pp. 253–314, 1988, Prentice-Hall.

103. HSIAO, M. AND LAZAR, A. A game theoretic approach to decentralized flow control of Markovian queueing networks. In *Performance 87* (P.-J. Courtois, G. Latouche, eds), 1988, pp. 55–73, Elsevier.

104. KUMAR, S. AND KUMAR, P. R. *Performance bounds for queueing networks and scheduling policies*, preprint, 1993, Coordinated Sciences Lab., Univ. of Illinois at Urbana-Champaign.

105. KURAPATI, S. AND KUMAR, A. Optimal scheduling of a processor executing a communication protocol stack, *Proc. NETWRORKS 92*, to appear.

106. KURI, J. AND KUMAR, A. Optimal control of arrivals to queues with delayed queue length information, *Proc. 31st Conf. on Decision and Control*, 1992, IEEE Press.

107. MUKHERJEE, U. AND PILLAI, A. R. Fairness in queue lengths through dynamic access control, in preparation.

108. BERTSEKAS, D. AND TSITSIKLIS, J. An analysis of stochastic shortest path problems, *Math. OR*, 1991, 16, 580–595.

109. FRIEDMAN, J. *Oligopoly and the theory of games*, Ch. 10, pp. 213–233, 1977, North-Holland.

110. BORKAR, V. S. *Optimal control of diffusion processes*, Pitman Research Notes in Maths. No. 203, 1989, Longman Scientific and Technical.