SURVEY PAPER

Mean, Variance, and Probabilistic Criteria in Finite Markov Decision Processes: A Review[1]

D. J. WHITE[2]

Communicated by P. L. Yu

Abstract. This paper is a survey of papers which make use of nonstandard Markov decision process criteria (i.e., those which do not seek simply to optimize expected returns per unit time or expected discounted return). It covers infinite-horizon nondiscounted formulations, infinite-horizon discounted formulations, and finite-horizon formulations. For problem formulations in terms solely of the probabilities of being in each state and taking each action, policy equivalence results are given which allow policies to be restricted to the class of Markov policies or to the randomizations of deterministic Markov policies. For problems which cannot be stated in such terms, in terms of the primitive state set I, formulations involving a redefinition of the states are examined.

Key Words. Markov decision processes, infinite horizon, finite horizon, mean, variance, probabilistic criteria.

1. Introduction

In this paper, we will discuss particular aspects of criteria for finite Markov decision processes. We will formally define our framework later on, but for the moment concentrate on the *raison d'être* for the paper in general terms, in the context of finding optimal control policies over a finite or infinite time horizon, where random elements enter into the problem in a specific manner.

[1] The author would like to thank two referees for a very thorough and helpful refereeing of the original article and for the extra references (Refs. 47–52) now added to the original reference list.

[2] Professor of Decision Theory, University of Manchester, Manchester, England.

For the vast majority of work in the area of Markov decision processes, standard criteria have been the expected total return over the time horizon, the expected discount total return over the time horizon, or the limiting return per unit over an infinite time horizon where this limit exists, as it will for the class of finite problems with which we deal, providing we restrict ourselves to particular classes of policies.

However, the use of expected total (discounted or not) return or limiting return per unit time may be quite insufficient to characterize the problem from the point of view of the decision maker, and it may be necessary to select a criterion (or criteria, if more than one needed) to reflect the variability-risk features of the problem. For example, perhaps the best known approach stems from the earlier work of Markowitz (see Ref. 1) on portfolio analysis, where mean and variance of return are used to characterize the problem. Alternatively, one might formulate the problem in terms of the probabilities of certain outcomes, and optimize, for example, the expected return subject to constraints on these probabilities (see Charnes and Cooper, Ref. 2, for the earlier work, and Hogan, Morris and Thompson, Ref. 3, for the list of references where these ideas have been applied). These references relate to single-period optimization, whereas the present paper deals with multi-period optimization, and specifically with problem formulations involving means, variances (in different senses, viz., variances of total discounted reward and variances of rewards in each period), probabilistic constraints, and threshold probabilities.

Before doing so, there is one fundamental issue which has to be considered. For those analysts who prefer to work in terms of expected utility theory, it may, in principle, be possible to establish the existence of a utility function over the realizable histories of the process being studied. In such cases it may be possible to establish equivalent results to those which exist for the standard criteria mentioned earlier on. We will not concern ourselves with this approach in this paper, and the reader is referred to work of this kind specifically in the Markov decision process area in papers such as those of Jacquette (see Refs. 4, 5), Porteus (see Ref. 6), White (see Ref. 7), Howard and Matheson (see Ref. 8), Kreps (see Refs, 9, 10), and Rothblum (see Ref. 11).

Central to the standard Markov decision process theory are five concepts, which we will define formally later on, but must be mentioned here (viz., stationarity, uniformity, pureness, history independence, and the principle of optimality), all of which hold for our class of problems and for some of the utility-oriented work referred to earlier on, but some or all of which fail for the approaches that we will discuss in this paper.

Stationarity requires policies to be specified independently of the time of the decision to be made. Uniformity requires policies to be specified independently of the starting state and seeks policies which are uniformly optimal for all initial states. Pureness requires that neither policies, nor actions, be randomized. History independence requires policies to be specified only in terms of current state at the time of the decision (although, of course, past history may be formally incorporated into the state description). The principle of optimality, in effect, requires that any residual subpolicy of an optimal policy is also optimal for the residual duration of the process beginning at any time of the process.

These are all extremely useful features from a computational point of view, which make standard Markov decision process theory attractive, although even then, not always computationally feasible. The failure of these properties to hold, collectively or individually, in the approaches to be considered does raise serious computational issues in some cases and merits serious consideration of the utility point of view. The determination of utility functions in practice, and even the acceptance of utility ideas by decision makers, pose difficulties. In addition, many analysts work with nonutility-based problem formulations and many decision makers, in practice, work with mean variance and probabilistic approaches in their decision making. Of course, it is not necessary to actually obtain an explicit utility function to obtain the characteristic features referred to earlier on and, under suitable conditions, one might be able to restrict one's attention to stationary, pure, nonhistory-dependent, uniformly optimal policies. The use of the principle of optimality would pose problems without a utility function, since somehow the preferences over probability distributions of returns would have to be structured in a computationally useful way (Sobel, Ref. 12, touches upon this briefly, but the ideas remain to be developed). Even without the principle of optimality, the other features would be useful, but the actual search procedure for an optimal policy would be hampered. Of course, if we knew that we could restrict ourselves to a finite number of pure, history-independent, stationary policies, we might just determine the probability distributions of returns for each and just leave the decision maker to choose, but this could be computationally impracticable.

One particular point which has a bearing on the following material is that of pure policies. Given the existence of a utility function (such as a linear or exponential one), we need consider neither randomized actions nor randomized policies. For some problem formulations (e.g., see Kallenberg, Ref. 13), optimal policies have randomized actions. There is, therefore, some conflict with the implications of utility theory. For example, in Section 4.2 a problem formulation of Kallenberg (see Ref. 13) is given in terms of the steady-state probabilities of being in a given state and using a given action. Constraints are placed upon these probabilities, and it is required to maximize the expected return per unit time subject to these probabilities.

It is possible that no pure optimal policies will exist or indeed pure feasible policies. However, for utility functions which are a function of the probabilities, a nonrandomized optimal policy would exist. Thus, the Kallenberg formulation conflicts, in policy selection, with the existence of such utility functions. At the same time, restriction to pure policies will, for some problem formulations, add significantly to the computational work. We will not, in this paper, deal with this quite vital issue, but merely put it on record.

There exist various examples which illustrate the failure of some or all of the attributes referred to. We do not have the space to reproduce them here, and refer the reader to Sobel (see Ref. 14), Miller (see Ref. 15), White (see Ref. 16) (for examples where the principle of optimality fails for specific problems), Kallenberg (see Ref. 13), White (see Ref. 16) (who gave examples where nonuniformity applies, for probabilistically constrained problems in the first reference, and for variance minimization in the second reference), Kallenberg (see Ref. 13) (for an example where only randomized action optimal policies exist for a probabilistically constrained problem), and White (see Ref. 16) (where no stationary optimal policy exists for a variance minization problem).

Let us now specify the framework for the remainder of the paper. Although this is quite a limited one, it is to be expected that some results will carry over to more general classes of problem.

2. The Framework

Decisions will be made at the beginning of each of a sequence of unit time intervals, $t=1, 2, 3, \dots$, where the time horizon may be deterministic, finite, or infinite (although we will make some observations on other cases).

The state at the beginning of each unit time interval t will be a random variable S_t, taking one of a finite number of values $\{i\}=I$, $i=1, 2, \dots, m$.

For each pair (i, t), there will be a finite feasible action set $\{k\}=K(i, t)$.

We let Z_t be the random variable representing the action at time t.

In each unit time interval t there will be a return Y_t, which is a random variable, depending upon i, k, t.

We will now superimpose the Markov requirement. First of all, the history of the process up to unit time interval t takes the form

$$h_t=(i_1, k_1, i_2, k_2, \dots, i_{t-1}, k_{t-1}, i_t).$$

Let H_t be the set of all such realiazable histories, $t=1, 2, \dots$.

The Markov requirement is that the probability distributions of (S_{t+1}, Y_t) depend only on $S_t(=i_t)$ and on Z_t $(=k_t)$. For the marginal distributions of S_{t+1} and Y_t, we will require

$$P(S_{t+1}=j\,|\,S_t=i, Z_t=k)=p_{ij}^k(t),$$
$$P(Y_t\leq y\,|\,S_t=i, Z_t=k)=F_i^k(y, t).$$

A comment on the state-history framework is required here. The state i is a primitive state, and the process is assumed to be Markov in terms of i. The history h is the standard history description used in the theories of Derman (see Ref. 17) and Van der Wal (see Ref. 18) used in this paper. It must not be assumed that, because the process is Markov in i, the history can be ignored. The main purpose of Section 3 is to show that, where problems may be formulated in terms of the probabilities $\{x_i^k(t)\}$, we may ignore this history, but not necessarily otherwise. We do not formally incorporate the individual returns up to time t in the history. This may be done, in effect, by redefining the primitive state to be a new state $s=(i, r)$, where r is the realized return in the previous period. To fit within our framework, r must take a finite set of values, although this is clearly not necessary to get some results. This approach is used in some of the papers cited later on. In addition, we may wish to include the total reward or the total discounted reward to date in the state description, again giving a new state $s=(i, r)$, where r is defined appropriately. We will make reference to such approaches later on. Again, strictly speaking, to fit within our framework, r must take a finite number of values, although, for infinite-horizon problems, this will not be the case.

Finally, we could redefine a state to include all relevant historical information, and then the state becomes the history. We will not do this, because we wish to separate out the primitive state i_t (which is all that is required in some problem formulations) and because infinite-horizon problems would then require the state space to be infinite.

Let us now turn to decision rules and policies.

A decision rule is a function δ_t from H_t to the set of all probability distributions over $K(i_t, t)$ [noting that $K(i_t, t)$ depends only on the present state i_t, part of the history, with a redefinition of "state" if dependence on earlier history is required]. Let Δ_t be the set of all such decision rules, $t=1, 2, \ldots$.

A policy π is a sequence of decision rules $\pi=(\delta_1, \delta_2, \ldots, \delta_t \ldots)$ covering the time horizon of the process, which effectively gives us the probability distribution of actions to be taken at epoch t in terms of the history to date.

For problems with unbounded time horizon (even though termination may arise at a finite epoch), we let C be the set of all such policies. For problems when the time horizon is a deterministic number n, we let $C(n)$ be the set of all such policies.

Within these universal classes of policy, we will wish to consider particular subsets. The first subset is the set of Markov policies, where the decision rules at time t are functions of the current state i_t only. We label these C_M, $C_M(n)$ respectively, noting that the decision rules may be time dependent.

Within the set C_M, we will wish to consider policies made up of infinite repetitions of a single decision rule δ, independently of the time. This class is labelled C_S, the class of stationary policies.

We will wish to consider policies made up of decision rules which assign a specific action to a specific state and not use randomization of actions. The appropriate subset of C_S will be labelled C_D, and the appropriate subset of $C_M(n)$ will be labelled $C_D(n)$. These are the deterministic policies.

We also mention the class C_{MD} of all Markov deterministic, but not necessarily stationary, policies. This set does not appear to be mentioned in the literature. We use $C_{MD}(n)$ for the finite-horizon case.

A key concept in the development will be the probability of being in a certain state i and taking action k at epoch t, which will run throughout the paper. Formally, let

$$x_i^k(t) \text{ be the probability that } (S_t = i, Z_t = k),$$

for a given initial state i_1 and a given policy π.

We will suppress (i_1, π) for notational convenience, but stress that, in view of the nonexistence of the uniformity property in general, the dependence on the initial state must always be borne in mind.

For infinite-horizon problems, we will require the following limit, as a representation of the long-run proportion of times state i and action k are realised. In such cases, we restrict ourselves to the case when $\{p_{ij}^k(t), F_i^k(y, t), K(i, t)\}$ are stationary (i.e., independent of time). Thus, we define

$$x_i^k = \operatorname*{limit}_{n\to\infty} \left[\left(\sum_{t=1}^{n} x_i^k(t) \right) \Big/ n \right],$$

for a given initial state i_1 and policy π (both suppressed), and where this limit exists (as it will, for example, always exist for $\pi \in C_S$).

Let

$$x(t) \text{ be the vector } (x_i^k(t)) \in \mathbb{R}^{q_t}, \qquad q_t = \sum_{i\in I} \# K(i, t).$$

For the stationary, infinite-horizon case, let X be the set of all limit points, taken over all policies (again for a given initial state i_1) of the vector sequence $\{\sum_{t=1}^{n} x(t)/n\}$, noting that, for some policies, a single sequence may have more than one limit point.

For some policies, such a sequence may have only one limit point. Let $\tilde{C}$ be the set of all such policies, and let $\tilde{X}$ be the corresponding set of (unique) limit points. Finally, let X_M, X_S, X_D, X_{MD} be the corresponding sets of (multiple if necessary) limit points for the policy sets C_M, C_S, C_D, C_{MD}, respectively.

For the infinite-horizon case we will also discuss discounted problems. We will assume the random variables $\{Y_t\}$ to be bounded uniformly in t, but will retain time dependence. The discount factor will be $\rho < 1$. We will, instead of studying the problem in terms of the limiting vectors discussed above, discuss the problem in terms of the infinite sets of vectors $(x(1), x(2), \ldots, x(t) \ldots)$.

We define $X(\infty)$, $X_M(\infty)$, $X_S(\infty)$, $X_D(\infty)$, $X_{MD}(\infty)$ to be the sets of all such infinite vectors for the discounted processes beginning in a given state and using any of the policies in sets C, C_M, C_S, C_D, C_{MD}, respectively.

For the finite-horizon case, discounted or otherwise, we use corresponding sets $X(n)$, $X_M(n)$, $X_D(n)$, $X_{MD}(n)$.

Note that, by convention, the suffix D on its own refers to Markov, stationary, deterministic and the suffices MD refer to Markov, deterministic, but not necessarily stationary.

Before we proceed to look at some theoretical results, we introduce the notion of randomized vectors, analogously to the notion of randomized actions. We define X^*, $\tilde{X}^*$, X_M^*, X_S^*, X_D^*, X_{MD}^* to be the set of all randomizations over countable subsets of X, $\tilde{X}$, X_M, X_S, X_D, X_{MD}, respectively. Similar definitions are used for policy randomization. The use of countable randomizations removes the need to include randomizations over impure policies and their corresponding probability vectors.

We likewise use a asterisk (*) to denote appropriate randomization for the infinite-horizon discounted case and the finite-horizon discounted or nondiscounted cases.

We also use $\bar{A}$ to denote the closure of a set A.

There are obvious inclusion relationships between the various sets defined. These are, for the infinite-horizon cases,

$$C_D \subseteq C_S \subseteq C_M \subseteq C,$$

$$C_{MD} \subseteq C_M, \qquad C_S \subseteq \tilde{C} \subseteq C,$$

$$X_D \subseteq X_S \subseteq X_M \subseteq X,$$

$$X_{MD} \subseteq X_M, \qquad X_S \subseteq \tilde{X} \subseteq X,$$

with corresponding inclusions for the finite-horizon cases, with the exception that $\tilde{C}$, $\tilde{X}$ are irrelevant, except in a trivial sense.

For the infinite-horizon case, the constructs with which we will work are those of the policy space C (and its special subsets) and the action state frequency space X (and its special subsets), for the nondiscounted and discounted cases. For the finite-horizon case, we deal with $C(n)$, $X(n)$ and their special subsets. We also deal with randomizations over the policy sets.

There are relationships between policies and action-state frequencies. It is possible to do this in general for $\pi \in C$, but this is rather cumbersome. Since we are essentially interested in C_M and its special subsets, we confine ourselves to this case. If $\pi_i^k(t)$ is the probability, for policy π, that if we are in state i at time t we will then take action $k \in K(i, t)$, we have

$$x_i^k(t) = \sum_{j \in I} \sum_{l \in K(j,t-1)} \pi_j^k(t) p_{ji}^l(t-1) x_j^l(t-1), \qquad t \geq 2,$$

$$\text{with } x_i^k(1) = 0, \qquad \text{if } i \neq i_1.$$

For $\pi \in C_S$, both $\{x_i^k(t)\}$ and $(\sum_{t=1}^n x_i^k(t))/n$ will have limits as $t \to \infty$ (possibly, but not always, dependent upon the initial state $i = i_1$).

The formulations with which we will deal involve the X spaces rather than the C spaces. Hence, we need the corresponding inverse results relating X to C. Again, we confine ourselves to C_M. Indeed, from Van der Wal (see Ref. 18), if we are given $\{x_i^k(t)\}$ for any policy $\pi \in C$, then we can find a policy $\pi \in C_M$ with the same $\{x_i^k(t)\}$ values, given by

$$\pi_i^k(t) = x_i^k(t) / \sum_{l \in K(i,t)} x_i^l(t),$$

providing the denominator is positive, with $\pi_i^k(i)$ arbitrarily chosen otherwise.

We are now in a position to consider theoretical results for these problems: (a) infinite-horizon nondiscounted problems; (b) infinite-horizon discounted problems; and (c) finite-horizon discounted and nondiscounted problems.

3. Policy Equivalence Results

3.1. Infinite-Horizon Nondiscounted Stationary Problems. The conditions under which we work have been stated earlier on. The fundamental result is that (again noting that everything is relative to a suppressed initial state i_1)

$$X = \tilde{X} = X_M = \bar{X}_S^* = X_D^*.$$

This result is to be found in Hordijk and Kallenberg (see Ref. 23) and is based on earlier work of Derman (see Refs. 17, 19) and Derman and Strauch (see Ref. 20). White (see Ref. 16) also demonstrates that, for the unichain case, $\tilde{X} \subseteq X_D^*$, and $X^* = X_D^*$, a weaker result than above, but using a different approach based on a multi-objective result of Hartley (see Ref. 21).

However, whereas the earlier results are dependent on the initial state (except in special circumstances), White's results are independent of the initial state; i.e., for each $x_{i_1} \in \tilde{X}$ achievable for some starting state $i_1 \in I$, there exists a randomized policy in X_D^* for which the corresponding x_i, for all $i \in I$, is equal to x_{i_1}.

Derman (see Ref. 19) also shows that, if we now explicitly make X and the other sets dependent on an initial state i_1, then, providing all $\pi \in C_D$ are irreducible,

$$X(i_1) = X_D^*(i_1) = \bar{X}_S^*(i_1) = \bigcup_{j_1 \in I} X(j_1),$$

in which case all sets in the initial result are identical, independently of the starting state i_1.

Derman (see Refs. 19, 22) gives examples to show that, in general, $X \neq X_S$, where, in the latter case, even an initial recurrent state i_1 may not produce this result. However, if every $\pi \in C_D$ has a single chain (see Derman, Ref. 17) or if $x[x = \lim_{n \to \infty}(\sum_{t=1}^{n} x^{\pi}(t)/n)$ for a given policy $\pi \in C_S$, where this limit always exists in $C_S]$ is continuous on C_S (see Hordijk and Kallenberg, Ref. 23), then

$$X = X_S.$$

Hordijk and Kallenberg (see Ref. 23) also give an example to show that the continuity property is not necessary to obtain this result.

A weaker result is given in Derman (see Ref. 22), viz., for the case when each $\pi \in C_D$ is irreducible; then, again introducing dependence on the initial state i_1,

$$\bigcup_{i_1 \in I} X(i_1) = \bigcup_{i_1 \in I} X_S(i_1).$$

It is to be noted that we do not refer to C_{MD}, X_{MD} in the infinite horizon nondiscounted case; but, if desired, since $C_D \subseteq C_{MD} \subseteq C_M$, we may add X_{MD}^* to our identities in our initial main result.

The significance of the sets of limiting x vectors lies to some extent in the relationships with linear programming as a computational tool. Hordijk and Kallenberg (see Ref. 23) introduce the set X^0 defined as follows:

$$\begin{aligned} X^0 = \{x \in \mathbb{R}^q \mid \exists y \in \mathbb{R}^q \text{ with } (x, y) \text{ feasible for the constraints} \\ \sum_{i \in I} \sum_{k \in K(i)} (\sigma_{ij} - p_{ij}^k) x_i^k = 0, \forall j \in I, \\ \sum_{k \in K(j)} x_j^k + \sum_{i \in I} \sum_{k \in K(i)} (\sigma_{ij} - p_{ij}^k) y_i^k = \sigma_{i_1 j}, \forall j \in I, \\ x_i^k \geq 0, y_i^k \geq 0, \forall i \in I, k \in K(i)\}, \end{aligned}$$

where $\{\sigma_{ij}\}$ are the Kronecker delta numbers and i_1 is the initial state.

For cases where each $\pi \in C_S$ is unichain, these constraints may be simplified (see Kallenberg, Ref. 13).

Hordijk and Kallenberg show that

$$X_D^* = X^0,$$

which may be added to the initial main results to reduce X to X^0.

If, for a given problem formulation, we have obtained an appropriate limiting x solution, there remains the question of obtaining an appropriate policy.

If x is obtained, together with an appropriate y, from the Hordijk-Kallenberg constraints, then, if $X = X_S$, a policy $\pi \in C_S$ may be obtained by

$$\pi \equiv \{\pi_i^k\}.$$

π_i^k is the probability that, if we are in state i, we take action k and

$$\begin{aligned}\pi_i^k &= x_i^k \Big/ \Big(\sum_{k \in K(i)} x_i^k \Big), \quad \text{if } \sum_{k \in K(i)} x_i^k > 0\\ &= y_i^k \Big/ \Big(\sum_{k \in K(i)} y_i^k \Big), \quad \text{if } \sum_{k \in K(i)} x_i^k = 0, \ \sum_{k \in K(i)} y_i^k > 0\\ &= \text{arbitrary value}, \quad \text{otherwise}.\end{aligned}$$

It is to be noted that it is not necessary that, if $x \in X^0$, then $x \in X_S$. If $x \in X^0$, it will be possible to find $\pi \in C_M \cap \tilde{C}$ such that $x = x^\pi$ by first of all translating it to X_D^* and then back to $C_M \cap \tilde{C}$.

If x is obtained directly or indirectly via the use of C_D^*, Derman (see Ref. 17), Derman and Veinott (see Ref. 24), and Strauch and Veinott (see Ref. 25) identify an equivalent $\pi \in C_M \cap \tilde{C}$ (remembering that π need not be stationary) as follows. Let

$$x = \sum_{u=1}^{U} \alpha_u x_u,$$

where $\{x_u\}$ are the limiting vectors for the policies $\{\delta_u^\infty\} = \{\pi_u\}$ making up C_D, and let

$$\sum_{u=1}^{U} \alpha_u = 1, \qquad \alpha_u \geq 0, \quad \forall u.$$

Let $\{x_{iu}^k(t)\}$ be the probabilities of being in state i and taking action k at epoch t using policy π_u, $i \in I$, $1 \leq u \leq U$, $k \in K(i)$, $t \geq 1$. Then, a policy

$$\pi = \{\pi_i^k(t)\} \in C_M$$

is defined by

$$\pi_i^k(t) = \sum_{u=1}^{U} \alpha_u x_{iu}^k(t) \Big/ \left(\sum_{u=1}^{U} \sum_{k \in K(i)} \alpha_u x_{iu}^k(t) \right),$$

if the denominator is positive, and is arbitrary otherwise.

This transformation applies also for any $x \in X$ (see also Van der Wal, Ref. 18). It is also be noted that, for such a π, we will have $\pi \in \tilde{C}$.

Although the long-run frequencies implicit in our definition of $\{x_i^k\}$ are identical with the expected frequencies for $\pi \in C_S$, this is not true in general. Derman (see Refs. 17, 19) produces the probabilistic result

$$P\left(L \subseteq \bigcup_{i_1 \in I} X_D^*(i_1) \right) = 1,$$

where L is the set of all limit points of a sample sequence. It is difficult to say how this result might be used.

3.2. Infinite-Horizon Discounted Problems. For this class of problems, we deal with the total discounted return

$$R = \sum_{t=1}^{\infty} \rho^{t-1} Y_t,$$

which will be a random variable.

Using the transformation of Van der Wal (see Ref. 18), we immediately obtain

$$X(\infty) = X_M(\infty),$$
$$X_{MD}^*(\infty) \subseteq X_M(\infty).$$

Let us define a metric d on the set of all vectors $x \in X(\infty)$ as follows:

$$d_\rho(x) = \sum_{t=1}^{\infty} \rho^{t-1} |x(t)|,$$

where $|x(t)|$, $x \in \mathbb{R}^{q_t}$, is any metric (e.g., supremum metric). Then, using the topology induced by this metric, it is easy to see, using the finite-time horizon results of the next section, that

$$\overline{X_{MD}^*}(\infty) = X_M(\infty).$$

These results allow us to restrict ourselves to the latter two sets, providing our problems may be expressed solely in terms of $x^\pi \in X(\infty)$. For example (see Section 4.2), we may wish to optimize the expected discounted return subject to constraints on the $\{x_i^k(t)\}$, and the problem will be expressible completely in these terms. Restriction to $X_{MD}^*(\infty)$ will give ε optimal solutions, where ε is arbitrarily small. Unfortunately, this equivalence does not carry over to equivalence in terms of total discounted returns. Thus, for example (see Section 4.2), if we wish to minimize the variance of the total discounted return subject to a constraint on the total discounted return, it would not be possible, in general, to express this problem solely in terms of the $\{x_i^k(t)\}$, without including the individual returns in the state variable, and to do so might result in suboptimal policies. A redefinition of the state to include the total discounted return to date (see Ref. 26) would, however, allow this problem to be expressed in terms of the new $\{x_s^k(t)\}$ variables, although computationally more demanding to solve.

3.3. Finite-Horizon Problems.

For this class of problems, we deal with the total discounted return

$$R = \sum_{t=1}^{n} \rho^{t-1} Y_t,$$

which will be a random variable, and we allow ρ to be any real number.

Derman and Klein (see Ref. 27), by embedding the finite-horizon problem in an infinite-horizon problem and using the results of Section 3.1 with a new, finite-state structure $s = (i, t)$, $i \in I$, $i \leq t \leq n$, effectively produce the following equivalence:

$$X(n) = X_M(n) = X_{MD}^*(n).$$

These results allow us to restrict ourselves to the latter two sets, providing our problems, may be expressed solely in terms of $x^\pi \in X(n)$. Unfortunately, as in the infinite-horizon case, this equivalence does not carry over to equivalence in terms of probability distributions of total discounted returns. Thus, for example, when we turn to problems involving variance, should we restrict ourselves to $X_{MD}^*(n)$, this may involve a loss of optimality in some sense.

Let us now look at the manner in which these theoretical results may be brought to bear on the various Markov decision processes using nonstandard criteria as introduced in Section 1.

4. Mean, Variance, and Probabilistic Criteria

4.1. Infinite-Horizon, Nondiscounted, Stationary Problems.

Hordijk and Kallenberg (see Ref. 23) and White (see Refs. 16, 28) consider problems of maximizing the long-run return per unit time, subject to constraints on the long-run proportion of times in states i and actions k are realized.

In Ref. 23, the feasible set of x vectors X^0 is specified. Additional constraints are as below, which determine a region $\hat{X}$:

$$\sum_{i \in I} \sum_{k \in K(i)} a_{is}^k x_i^k \leq b_s, \qquad s = 1, 2, \ldots, S.$$

If

$$r_i^k = E(Y | S = i, Z = k),$$

the problem is

$$\underset{x \in X^0 \cap \hat{X}}{\text{maximize}} \left[\sum_{i \in I} \sum_{k \in K(i)} r_i^k x_i^k \right].$$

As an illustration of this type of problem, consider a simple inventory control problem, with no backlogs, in which i represents the current stock level and k represents the order quantity. The constraints might take the form

$$\sum_{k \in K(i)} x_0^k \leq 0.05$$

(i.e., the probability of running out of stock is less than or equal to 0.05) and

$$\sum_{i \in I} \sum_{k \in K(i): k > 0} x_i^k \leq 0.2$$

(i.e., the probability of placing a new order is less than or equal to 0.2).

Using the Van der Wal (see Ref. 18) approach one may then find $\pi \in C_M \cap \tilde{C}$ to give the requisite $x = x^\pi$.

Hordijk and Kallenberg also give an example (from Derman, Ref. 29) where no constrained optimal solution in C_S exists. However, $x = x^\pi$ for some $\pi \in \overline{C_S^*}$. If $X = X_S$, then the Hordijk-Kallenberg approach gives $\pi \in C_S$ directly. Even when $X \neq X_S$, it is possible that, for the particular x obtained, $x = x^\pi$ for some $\pi \in C_S$, and Hordijk-Kallenberg give a method for testing this possibility.

White (Ref. 16) suggests a linear programming approach in terms of X_D^*, using equalities in place of inequalities in $\hat{X}$, to give a set of constraints $\hat{X}$, but this makes no essential difference to the formulation. Only the unichain C_D case is considered, in which case everything is state independent. If $\{\pi_u\}$, $1 \leq u \leq U$, are the policies of C_D, then policies of the form

$$\pi \equiv \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_u \ldots \alpha_U)$$

are considered, where α_u is the probability of using policy $\pi_u \in C_D$. The problem is reduced to

$$\underset{\alpha}{\text{maximize}} \left[\sum_{u=1}^{U} \alpha_u r_u \right],$$

subject to

$$\sum_{u=1}^{U} a_{us}\alpha_u = b_s, \qquad s = 1, 2, \ldots, S+1,$$

where r_u is the long-run return per unit time using policy π_u, the $(S+1)$th constraint is

$$\sum_{u=1}^{U} \alpha_u = 1, \qquad \alpha \geq 0,$$

and

$$\alpha_{us} = \sum_{i\in I} \sum_{k\in K(i)} a_{is}^{k} x_{iu}^{k}, \qquad 1 \leq u < U,$$

where x_{iu}^k is the probability of being in state i and taking action $k \in K(i)$ for policy $\pi_u \in C_D$.

The column generation method of Dantzig and Wolfe (see Ref. 30) may be used; and, if $\{\lambda_s\}$ are the current simplex multipliers, the new basic variable α_u to be brought in is obtained as follows:

$$\underset{u}{\text{maximize}} \left[\sum_{i\in I} \left(r_i^{\delta_u(i)} - \sum_{s=1}^{S+1} \lambda_s a_{is}^{\delta_u(i)} \right) x_{iu} \right],$$

where $\pi_u = (\delta_u^\infty)$, $\{x_{iu}\}$ are the limiting long-run proportion of times in state i using policy π_u.

This subproblem may be solved by the usual policy iteration method of Howard (see Ref. 31), with the returns $\{r_i^k\}$ modified by the simplex multipliers.

Again, when an optimal α is obtained, a corresponding policy in $C_M \cap \tilde{C}$ may be obtained, or in C_S, since in this case $X = X_S$. The method of Hordijk and Kallenberg (see Ref. 23) will achieve this, using

$$x_i^k = \sum_{u=1}^{U} \alpha_u x_{iu}^k.$$

White (see Ref. 28) is the author of an earlier paper introducing the ideas of Lagrange multipliers directly, resulting in the determination of optimal policies in C_D for the range of Lagrange multipliers, and then determining all the optimal expected returns and levels $\{b_s\}$ in the constraints determining $\hat{X}$, which are achievable by taking probability mixtures at the threshold Lagrange parameter values.

An alternative approach to constrained optimization is to introduce the ideas of mean-variance analysis. In the cases where long-run frequencies and expected frequencies are the same (e.g., for the policy space C_S), the variance of long-run return per unit time is zero. However, the variation from one time unit to the next time unit may be important.

Filar and Lee (see Ref. 32) first of all determine the maximal expected return per unit time (again, note that everything is conditional on an initial state i_1). Let this be r^0. Then, they consider the random variation in unit time interval, viz.,

$$Y_t^0 = |Y_t - r^0|.$$

There are other equally acceptable definitions of Y_t^0, e.g.,

$$Y_t^0 = \max[0, c - Y_t].$$

A new problem is created in which $\{Y_t, -Y_t^0\}$ are taken as two returns, and multi-objective methods are applied to obtain the efficient solution set, with respect to the long-run values per unit time of $\{Y_t\}$ and $\{-Y_t^0\}$ for a specified prior probability distribution over the initial states.

If $\hat{Y}^\pi$, $-\hat{Y}^{0\pi}$ are the expected values per unit time of $\{Y_t\}$, $\{-Y_t^0\}$, respectively, using policy π, then π is said to be efficient if there is no other policy τ such that

$$\hat{Y}^\tau \geq \hat{Y}^\pi,$$
$$-\hat{Y}^{0\tau} \geq -\hat{Y}^{0\pi},$$

with at least one strict inequality.

The results developed in this paper are all relative to a given initial starting state. Filar and Lee's results relate to a given probability distribution over the starting states. However, if we introduce an artificial state i^0, with a fixed action which moves the system from state i^0 to the other state $i \in I$ with the specified probabilities, then their work fits into the framework of this paper, and we may apply the X_S^* approach using the Hordijk-Kallenberg results (see Ref. 23) or the X_D^* approach using the White results (see Ref. 16), since the problem may be stated in terms of $\{x_i^k\}$.

Given the linear programming equivalence, the set of efficient solutions is obtained by using a weighted reward

$$Y_t(\alpha) = (1-\alpha)Y_t - \alpha Y_t^0$$

varying α over $(0, 1)$ (see Ref. 33).

The authors also suggest an alternative approach. Let r be the long-run return per unit time for a given policy π. The return $Y^0(t)$ is replaced by $Y_t - r$, and a weighted reward $Y_t(\alpha)$ is used, where

$$Y_t(\alpha) = (1-\alpha)Y_t - \alpha(Y_r - r)^2.$$

Using the approaches of Hordijk and Kallenberg (see Ref. 23) or of White (see Ref. 16), the first problem reduces to a linear program and the second to a quadratic program.

Mendelssohn (see Ref. 34) considers a variance type problem in which it is the variation between successive returns which is important. To fit within our formulation, the state i would have two components, i.e., $s=(i, r)$, where r is the return in the last unit time interval, but decisions are still based only on i. His new return takes the form

$$\tilde{Y}_t = Y_t - \lambda |Y_t - Y_{t-1}|, \qquad t \geq 2,$$

for some $\lambda > 0$. For the problems to remain finite state, $\{Y_t\}$ may only take a finite set of values to fit in with our framework.

Mendelssohn's state description involves two components, viz., our primitive state i and the return level r in the previous time interval. In order to fit within our framework, our new state becomes $s=(i, r)$. With the new state description, the results of this paper apply and the problem would be solved using the new formulation. Mendelssohn restricts his decision rules to be functions of i only, i.e., $K(s)$ becomes $K(i)$.

Mendelssohn's paper is for discounted problems, but the ideas clearly carry over to the nondiscounted case, so we mention them here.

Filar (see Ref. 35) introduces a percentile approach. A critical return λ is chosen. A new return function $Y_t(\lambda)$ is defined by

$$\begin{aligned} &Y_t(\lambda) = 1, \qquad &&\text{if } Y_t \geq \lambda, \\ &Y_t(\lambda) = 0, \qquad &&\text{otherwise.} \end{aligned}$$

A critical long-run return per unit time c is chosen.

Equivalently (although not stated), for each policy π, a number $\lambda^\pi(c)$ is determined by

$$\lambda^\pi(c) = \{\max[\lambda]: r^\pi(\lambda) \geq c\},$$

where $r^\pi(\lambda)$ is the long-run return per unit time using policy π and random returns $\{Y_t(\lambda)\}$, and where

$$\lambda^\pi(c) = -\infty, \qquad \text{if } r^\pi(\lambda) < c, \forall \lambda.$$

Then, $\lambda(c)$ is defined by

$$\lambda(c) = \sup_\pi [\lambda^\pi(c)],$$

and finally a policy π is chosen to

$$\text{maximize}[r^\pi(\lambda(c))].$$

Filar's approach is a risk-oriented approach, but, instead of risk being evaluated in terms of the total return, it is evaluated in terms of the proportion of times the return in each period reaches a critical level λ. Clearly, the problem can be formulated in terms of $\{x_i^k\}$, and we can use the earlier equivalence results and restrict ourselves to X_M or to X_D^*, if we wish. The final problem, once $\lambda(c)$ has been determined, reduces to a standard Markov decision process problem, for which a solution in X_S will exist.

The linear programming approach of Kallenberg (see Ref. 13) is used to solve the problem for a long-run limiting vector, which may be transformed into an appropriate policy. Alternatively, we may use X_D^* and use White (see Ref. 16).

In Section 1, we made reference to the use of expected utilities, using utility functions whose domain was the realizable histories of the process. For some problems (e.g., those formulated in terms of probabilistic constraints on the state-action probabilities, discussed earlier on in this section), a unique randomized optimal solution can arise (e.g., see Kallenberg, Ref. 13). If an approach which seeks to optimize expected utility over the realizable histories had been used, optimal nonrandomized policies would exist. There is therefore an incompatibility between the two approaches in this case.

The other problems discussed, if formulated in terms of $\{x_i^k\}$ or $\{\alpha_u\}$ (i.e., in X_S or X_D^*), produce overall objective functions which are linear or convex quadratic in the variables used and will produce optimal nonrandomized solutions (maximizing a convex function, over a convex set), and thus will produce results which are compatible with utility theory in that nonrandomized optima are produced. However, this is not the same as saying that the specified objective functions are compatible with utility theory axioms. For example, consider the alternative approach of Filar and Lee (see Ref. 32), and let us assume that, in each period, the returns $\{Y_t\}$ are independent and identically distributed for any policy. The equivalent problem is then to maximize

$$[(1-\alpha)E(Y)-\alpha(Y-E(Y))^2].$$

The expression does not satisfy the axiom that a probability mixture of two indifferent policies is indifferent to both [see White, Ref. 36, page 146, Example 12(iii)].

Thus, the objective functions used do induce an order relationship over C, but some will not satisfy some axioms of expected utility theory.

4.2. Infinite-Horizon Discounted Problems. Our earlier results indicate that, if we can formulate our problems in terms of $x^\pi \in X(\infty)$, then we may restrict ourselves to $X_M(\infty)$ or to $X_{MD}^*(\infty)$ without loss (or rather to within an ε-loss in this case).

For C_M, we may specify the feasible set in terms of $\{x_i^k(t)\}$ (the probabilities of being in state i and taking action k at time t) and $\{\pi_i^k(t)\}$

(the probabilities of taking action k at time t if we are in state i). We have a set of equations of the form

$$\sum_{k\in K(i,t+1)} x_i^k(t+1) = \sum_{j\in I, k\in K(j,t)} x_j^k(t)p_{ji}^k(t), \qquad \forall i\in I,\ t\geq 1,$$

$$\sum_{k\in K(i_1,1)} x_{i_1}^k(1)=1,$$

$$x_i^k(1)=0, \qquad \forall i\neq 1_1, \qquad k\in K(i,1),$$

$$x_i^k(t)\geq 0, \qquad \forall i\in I, \qquad k\in K(i,t),\ t\geq 1.$$

Given $\{x_i^k(t)\}$, we may calculate the policy probabilities $\{\pi_i^k(t)\}$ by

$$\pi_i^k(t)=x_i^k(t)\Big/\Big(\sum_{k\in K(i,t)} x_i^k(t)\Big),$$

providing the denominator is not zero, and arbitrary if the denominator is zero.

For C_{MD}^*, we may specify the feasible set in terms of $\alpha=(\alpha_1, \alpha_2, \ldots, \alpha_u, \ldots)$, the countably infinite-dimensional probability vector over the policies $\{\pi^u\}$ in C_{MD}.

In the former approach, one might wish to get optimal or ε-optimal policies with respect the the expected total discounted return, subject to various constraints on the $\{x_i^k(t)\}$ or some function of these, e.g.,

$$\sum_{i\in I}\sum_{k\in K(i,t)} a_{is}^k \sum_{t=1}^{\infty} \rho^{t-1}x_i^k(t)\leq b_s, \qquad s=1,2,\ldots,S.$$

Ideally, one would wish to work in terms of the probability distributions of the total discounted returns, but this appears to be quite a difficult problem. Slightly extending a result of Sobel (see Ref. 14), we may express the cumulative probability distribution of the total discounted return as follows, for any policy π:

$$P_i(r,t)=\sum_{k\in K(i,t)}\sum_{j\in I}\int_y \pi_i^k(t)p_{ij}^k P_j((r-y)/\rho, t+1)dF_i(y,t),$$

$$\forall i\in I,\ t\geq 1,$$

where, for a given policy π, $P_i(.,t)$ is the cumulative probability distribution function for the total discounted returns from time t onward beginning in state i and, as given in Section 2, $F_i(.,t)$ is the cumulative probability distribution function for the immediate return Y_t, given (i,t). We may impose constraints on $\{P_i(.,t)\}$.

It should be pointed out that the above formulation is not strictly within the framework of this paper for infinite-horizon problems, since we may need to cater for an infinite number of values of r. Also, there is no prior reason why the policies should be restricted to C_M in terms of the primitive state I. White (see Ref. 37) extends this problem to an infinite-horizon optimization problem with appropriate computational algorithms.

Instead of working with probability constraints, we may work in terms of a finite set of the moment vectors. If we accept the underlying utility theory of Markowitz (see Ref. 1), there will exist a polynomial utility function of the discounted return, and a nonrandomized optimal policy will exist. However, this need not be in C_{MD} [see, for example, White (Ref. 16), where a variance minimization problem requires a history-remembering policy as its solution]. It is also possible that one might wish to work within C_M.

For such solutions, Sobel (see Ref. 14) gives recurrence relations for moments, for stationary policies, which may again be adapted to the C_M class. Of particular interest are the first two moments, mean v, and variance V (again recall that everything is conditional on the initial state i_1). It is possible to extend Sobel's recurrence relations for the stationary case to the C_M case using the variables $\{x_i^k(t)\}$, $\{\pi_i^k(t)\}$. In Sobel's case, the random return Y_t is determined by i, j, k, i.e., $Y_t = r_{ij}^k(t)$, a deterministic quantity.

The alternative procedure is to consider C_{MD}^*. In this case, each $x_i^k(t)$ may be expressed in the form

$$x_i^k(t) = \sum_{u=1}^{\infty} \alpha_u x_{iu}^k(t),$$

where

$$\begin{aligned} x_{iu}^k(t) &= 1, \qquad \text{if } \delta^u(i, t) = k, \\ x_{iu}^k(t) &= 0, \qquad \text{otherwise,} \end{aligned}$$

noting that the decision rules are also functions of time in this case. This means that all linear constraints in $\{x_i^k(t)\}$ become linear constraints in $\{\alpha_u\}$; and, if we wish to optimize (or ε-optimize) the expected total discounted return subject to these constraints, we have a semi-infinite linear program, viz.,

$$\text{maximize}\left[\sum_{u=1}^{\infty} \alpha_u \sum_{t=1}^{\infty} \rho^{t-1} \sum_{i\in I} \sum_{i\in I} \sum_{k\in K(i,t)} x_{iu}^k(t) r_i^k(t)\right],$$

subject to, for example,

$$\sum_{u=1}^{\infty} \alpha_u \sum_{i\in I} \sum_{k\in K(i,t)} a_{is}^k \sum_{t=1}^{\infty} \rho^{t-1} x_{iu}^k(t) \leq b_s, \qquad s = 1, 2, \ldots, S,$$

$$\sum_{u=1}^{\infty} \alpha_u = 1, \qquad \alpha_u \geq 0, \qquad 1 \leq u < \infty.$$

Again with a little bit of manipulation, using column-generating methods (see Dantzig and Wolfe, Ref. 30), each subproblem is reducible to solving a standard optimal infinite-horizon discounted problem in C_{MD}, with $r_i^k(t)$ replaced by $r_i^k(t) - \sum_{s=1}^{s} \lambda_s a_{is}^k$, where $\{\lambda_i\}$ are the current simplex multipliers. Of course, more complicated constraints may be used and, if we keep to C_D, the problem is much easier.

Using a similar procedure, White (see Ref. 16) considers mean-variance analysis. If $\{(v_u, V_u)\}$ are the means and variances of policies in C_{MD}, and $\{(v(\alpha), V(\alpha))\}$ are the means and variances for policies in C_{MD}^* for the various probability vectors α, then

$$v(\alpha) = \sum_{u=1}^{\infty} \alpha_u v_u,$$

$$V(\alpha) = \sum_{u=1}^{\infty} \alpha_u (V_u + v_u^2) - \left(\sum_{u=1}^{\infty} \alpha_u v_u \right)^2.$$

In this form, it is possible to analyze the problem in terms of, for example, minimizing $V(\alpha)$ subject to $v(\alpha) = c$, or in terms of finding efficient sets in terms of $(v, -V)$ or (v, V) (see White, Ref. 33, where some of the weighting factor results carry over from finite-dimensional problems to infinite-dimensional problems). Column-generation methods may be used and result in subproblem optimization of the form

$$\underset{u}{\text{minimize}}[V_u + v_u^2 - \lambda v_u],$$

where λ is the current simplex multiplier for the constraint $v(\alpha) = c$.

Once a solution in C_{MD}^* is found, an equivalent solution in C_M may be found using Van der Wal's transformation (see Ref. 18), which will give the same x^π vector.

The work of Filar and Lee (see Ref. 32) and of Mendelssohn (see Ref. 34), discussed in Section 4.1 for nondiscounted problems, is also studied by Filar and Lee (see Ref. 32), with different immediate return functions, for the infinite-horizon discounted problem. The new reward in interval t is, for a given policy π,

$$\tilde{Y}_t = Y_t - \lambda f(Y_t - r),$$

where r is the expected value of Y_t for the given policy and $f(\cdot)$ is a penalty function which may be approximated by the first two terms of its Taylor series.

Finally, we refer to Mendelssohn (see Ref. 34), referred to in Section 4.1 for nondiscounted problems, whose actual paper deals with the discounted version of his variance approach.

4.3. Finite-Horizon Problems. As we have seen in Section 3.3, we may reproduce all the x^π vectors by confining ourselves to $X_M(n)$ or to $X^*_{MD}(n)$ or equivalently to $C_M(n)$ or to $C^*_{MD}(n)$, the policy sets.

We may approach this problem using the $\{(x_i^k(t), \pi_i^k(t)\}$ approach for $C_M(n)$ or the $\{\alpha_u\}$ approach for $C^*_{MD}(n)$, but when $1 \le t \le n$ in the case of $C^*_M(n)$ and $1 \le u \le U < \infty$ in the case of $C^*_{MD}(n)$. The former approach is given in White (see Ref. 28). In the latter case, we may find an equivalent solution in C_M by using Van der Wal's transformation (see Ref. 18).

There are several approaches to mean-variance analysis.

The mean-variance analysis of Sobel (see Ref. 14) and White (see Ref. 16) may be adapted for the finite-horizon cases. An alternative approach is to convert the problem into a final-value problem with state $s = (i, r)$, where r is the return (discounted or nondiscounted) up to time t. If we then want to solve the problem of minimizing $V(\alpha)$, subject to $v(\alpha) = c$, as in the discounted problem, the column generation method will reduce to solving the subproblem

$$\min[E_u(r_0^2 - \lambda r_0)],$$

where E_u is the expectation for policy π_u and r_0 is the cumulative discounted or nondiscounted return at the end of the time horizon. To fit in with our finiteness of state set I, r_0 would have to take a finite number of values. It may be possible to extend the earlier theory for finite state sets I to infinite-state sets, replacing, for example, $\{x_i^k(t)\}$ by distribution functions over $\{i, k\}$ for each value of t, in which case the fact that r_0' may have an infinite number of values could be handled. Alternatively, some computational approximation scheme might be developed.

It is easy to solve the subproblem by the usual dynamic programming approach for final-value systems, and this is given in White (see Ref. 28).

As with the infinite-horizon discounted problem, efficiency analysis in terms of $(v, -V)$ or (v, V) may be considered. Since $V(\alpha)$ is concave in α, for (v, V) we may use the parametrized single objective function $V + \lambda v$ for $\lambda \ge 0$ and $\lambda = \infty$, to generate the efficient solutions, noting that, for $\lambda = 0$ or $\lambda = \infty$, formally we have to eliminate the nonefficient solutions generated. For $(v, -V)$, since (see White, Ref. 33) any efficient solution is an optimal solution of a subproblem of the form "minimize $V(\alpha)$, subject to $v(\alpha) = c$," and since this is now a linear programming problem (as we have seen in the infinite-horizon case), the weighting factor result goes through again, with the same provisos for $\lambda = 0$, $\lambda = \infty$. In this case, the subproblem is to maximize $[\lambda v - V]$.

Let us now turn to a few problems related, to some degree, to our main framework, although not, in general, wholly matching it.

Henig (see Refs. 38, 39) introduces the criteria given below, for a specified policy:

(TC) target criterion: $1-P(c)$,

(PC) percentile criterion: $\inf[r: P(r) \geq p]$.

Henig's problem is a routing problem with a finite number of states, an initial state $i=1$, and a target state $i=m$, with random returns at each step dependent upon (i, k) as with our main framework. The transitions from i to j are, however, taken to be deterministic and, in effect, $k=j$. This is trivially Markov, but, since we may introduce randomization, it may be made nontrivially Markov. It is assumed that, for all policies, the target state is reached in a finite number of moves because Henig keeps to nonrandomized policies.

This problem does not fit exactly into any of our problem classes as they stand. However, by using a similar approach to that of White (see Ref. 16) for the infinite-horizon discounted problem, with conditions that ensure finiteness of termination with probability one, similar results to those of White given in the infinite-horizon discounted case will follow, providing we allow policies to be time dependent (in this case, we label the moves $1, 2, \ldots, t, \ldots$). For such cases, we may wish to restrict ourselves then to C_M or to C^*_{MD}, remembering that these induce nonstationary policies. It is then possible to adapt similar methods for C_M or C^*_{MD} using the $\{(x^k_i(t), \pi^k_i(t))\}$ or α approaches mentioned in the infinite-horizon discounted cases.

Henig (see Refs. 38, 39) restricts himself to the situation where the random returns Y_t (in our stage-dependent framework) are independently, normally distributed at each step. The distribution of the total return is then normal for policies $\pi \in C_{MD}$, but not in C_M. Henig actually restricts himself to $\pi \in C_D$ (our deterministic stationary set). For normally distributed total returns, the maximization of the (TC) and (PC) criteria are equivalent, respectively to maximizing (see Charnes and Cooper, Ref. 40) $(v-c)/\sqrt{V}$ and $v+r(p)\sqrt{V}$, where $r(p)$ is the level r of the standardized normal variate for which $\Phi(r)=p$. Henig relates such optimization problems to finding efficient sets, using weighting factor approaches, with respect to (v, V) and $(v, -V)$.

It is to be stressed that, if $\pi \notin C_D$, then the equivalence of the (TC) and (PC) criteria and the stipulated mean-variance optimization problems are no longer valid.

If we use the C^*_{MD} policy approach, as referred to earlier on, for $C^*_{MD}(n)$, then efficient solutions are obtainable by using $\lambda v + V$ or $\lambda v - V$, as may be the case, and maximizing over $\alpha = (\alpha_u)$. In the latter case, since $V = V(\alpha)$ is concave in α, an optimal solution in C_{MD} will exist, although not necessarily uniquely so. Since Henig's transformations are deterministic and the returns are independently distributed, $\lambda v - V$, for policies in C_{MD}, may be optimized using the usual dynamic programming approach, as Henig does for his restriction to C_D (see Ref. 38).

Henig's argument for keeping to C_D is the utility one discussed in Section 1.

We may induce an order relation over C as follows:

$$\pi \text{ is preferred to } \tau \rightleftarrows (1-P^\pi(c)) \geq (1-P^\tau(c)),$$

where P^π is the distribution function for the total return. The function $(1-P^\pi)$ then satisfies all the axioms of expected utility theory, and we may restrict ourselves to C_{MD} without loss, on the assumption that we are, in the first instance, restricting ourselves to policies defined in terms of the original primitive state space I.

If we wish to introduce the combined state $s=(i, r)$, as we did earlier on in this section, then optimal policies in C_D (with the new state description and r need not be finite) will exist and may be found by combining the cumulative distribution recurrence relation of Sobel (see Ref. 14) with the usual optimality equation approach to give the equation

$$P(i, r) = \min_{j \in I}\left[\int_y P(j, r+y)\, dF_i(y)\right],$$

with

$$P(m, r) = 1, \qquad \text{if } r \leq c,$$

$$P(m, r) = 0, \qquad \text{if } r > c,$$

where $P(i, r)$ = minimal probability that the total return, inclusive of r, to absorption at m, beginning in state (i, r), will be less than or equal to c.

However, the (PC) approach is not compatible with expected utility theory, and hence this argument cannot be used to justify keeping within C_{MD} or C_D using either state description.

A related mean-variance class of problems, again for the case when $\{Y_t\}$ are normally and independently distributed over time, is discussed by Goldwerger (see Ref. 41). Policies are restricted to $C_{MD}(n)$ (finite-stage problems), and it is required to maximize $v/\sqrt{V}$ [slightly different from the Charnes and Cooper problem (see Ref. 40). The authors proceed by successively maximizing the ratio $v(j, t)/\sqrt{V(j, t)}$ over times $t, t+1, \ldots, n$, for any given state j at time t. As Miller (see Ref. 15) points out, this is a misuse of the optimality principle. If the problem is embedded in a class of problems of the kind "minimize V, subject to $v=c$, if $v \geq 0$)," then earlier approaches may be used.

There are other problems which may be cast in the form of finite-time horizon Markov decision processes which are worth mentioning briefly.

One of these is the stochastic knapsack problem discussed by Parks and Steinberg (see Ref. 42). They adapt the (TC) approach of Henig (see Ref. 39), and suggest an approach of Goldwerger (see Ref. 41), which Sneidovitch (see Ref. 43) shows to be erroneous, in that the principle of optimality used is not valid. Sneidovitch also discusses general difficulties in decomposition of preference orders in order to be able to apply the principle of optimality. Sneidovitch (see Ref. 44) considers a stochastic knapsack problem, with two objective functions, in which the problem is to minimize the expectation of one subject to a constraint on the variance of the other. Greenberg (see Ref. 45) also considers the maximization of a function subject to a probabilistic constraint, which may likewise be formulated as deterministic finite time-horizon Markov decision processes.

5. Summary and Comments

The purpose of this paper is to review existing material in the area of finite-state, finite-action, Markov decision processes in the context of nonstandard criteria which, at present, constitutes the vast majority of the work in this area. It is unlikely that single expected total return, expected discount return, or long-run returns per unit time will properly represent the vast majority of decision-making situations. The use of expected utility theory, where it is thought relevant and practicable, would normally be the natural extension, but practical difficulties do arise, and decision making is often undertaken in a manner not always compatible with expected utility theory, e.g., selecting inventory control policies to give a specified probability of run out.

The paper does not seek to validate or to compare criteria, but simply to present some results which have appeared in the literature. It is clearly evident that some criteria suggested may not be sensible. Beja, for example (see Ref. 46), shows how the use of probabilistic criteria may, on occasion, give quite questionable results.

Section 3 studies the possibilities of being able to restrict attention to simpler classes of policies than the set of all possible policies, for infinite-horizon nondiscounted stationary problems, infinite-horizon discounted, and deterministic finite-horizon discounted and nondiscounted problems where, in the two latter cases, time dependence is included. For the first case, equivalent reductions are in terms of long-run proportions of times of being in a given state and taking a given action (or the steady-state probabilities, in special cases). For the second case, equivalent reduction is in terms of infinite vector streams of action-state probabilities. For the last case, equivalent reduction is in terms of the finite vector streams of action-state probabilities. In these terms, the equivalent sets are the sets C_M or C^*_{MD} for the first case, C_M or $\overline{C^*_{MD}}$ in the second case, and $C_M(n)$, $C^*_{MD}(n)$ in the third case. It is important to stress that everything is relative to a given initial state $i = i_1$, since optimal solutions do not always exhibit the state uniformity characteristic of conventional Markov decision processes.

No attempt has been made to look at other classes of problems such as total return problems, nor to go outside the finite requirements, although clearly similar results will be obtainable under appropriate conditions for other situations.

Section 4 looks at approaches to policy optimization, using mean, variance, and probabilistic constraints, which have appeared in the literature. The restriction to the policy classes specified enables two alternative linear programming or quadratic programming approaches to be used, one in terms of $\{(x_i^k(t), \pi_i^k(t)\}$ (or $\{x_i^k, \pi_i^k\}$ in the stationary case), the probabilities of being in state i, and taking action k at time t, and the policy defining probabilities of taking action k at time t if we are in state i, and $\alpha = (\alpha_u)$, α_u being the probability of choosing a policy π_u in C_D or C_{MD} as the case may be.

Finally, for some problems, such as the finite-horizon problem or the optimal routing problems, some advantage both in terms of improved policies and in terms of computations may be obtained by introducing extra state variables to represent the accumulated return up to time t, and then using a final-value approach. Although, to fit within the framework of this paper, such returns should strictly belong to a finite set, it is to be expected that the results will go through in some cases with this restriction removed.

It has not been the intention to identify possible research problems, but clearly the survey results suggest some possibilities such as those listed below.

(i) To what extent can the results of Sections 2 and 3 be extended to more general state-action spaces?

(ii) To what extent can those methods using criteria which have no utility base be said to be acceptable as approximation methods for utility based methods, where the utility models are taken to be the true ones, but where these are seen to be impractible for various reasons?

(iii) If utility models are to be used, what computational procedures might be developed, bearing in mind the possibility that some utility models may involve complex functions of the history of the process at any time?

(iv) Even for the criteria given in the paper, what computational procedures might be developed, e.g., for the Hordijk–Kallenberg problem formulation of Section 4.1, and if we wish to find optimal pure policies, how might this be done? For the mean-variance problem discussed at the end of Section 4.2, what computational schemes may be developed to solve the subproblems of C_{MD} which are generated?

6. Appendix: Column Generating Technique

In Section 4.2 and other sections, the use of column generation techniques is suggested, without verification. Consider the semi-infinite linear program in $\{\alpha_u\}$ given in Section 4.2. This takes the form

$$\text{maximize } z = \sum_{u=1}^{\infty} r_u \alpha_u,$$

subject to

$$\sum_{u=1}^{\infty} b_{us}\alpha_u \leq b_s, \qquad 1 \leq s \leq S,$$

$$\alpha_u \geq 0, \qquad 1 \leq u < \infty,$$

where the coefficients $\{r_u\}$, $\{b_{us}\}$ are given in the text and the constraint

$$\sum_{u=1}^{\infty} \alpha_u = 1$$

is replaced by two inequalities (not strictly necessary if we allow equality constraints).

This is a semi-infinite linear program (a finite set of constraints), with convergence properties of the coefficients which allow the usual simplex method to be used. At any specific stage of the calculations, let NB be the nonbasic set of variables $\{\alpha_u\}$, and let $\{\lambda_s\}$, $1 \leq s \leq S$, be the simplex multipliers at that stage. The canonical form of z is then

$$z = \sum_{u \in NB} \left(r_u - \sum_{s=1}^{S} \lambda_s b_{us} \right),$$

and the next nonbasic variable to be made basic is obtained by solving

$$\underset{u \in NB}{\text{maximize}} \left[r_u - \sum_{s=1}^{S} \lambda_s b_{us} \right].$$

Reinterpreting in the original problem, this is the same as

$$\underset{u \in NB}{\text{maximize}} \left[\sum_{t=1}^{\infty} \rho^{t-1} \sum_{i \in I} \sum_{k \in K(i,t)} x_{iu}^{k}(t) \left(r_i^k(t) - \sum_{s=1}^{S} \lambda_s a_{is}^{k} \right) \right].$$

This is the same as choosing the policy $\pi_u \in C_S$ to maximize the infinite-horizon discounted expected return with immediate reward structure

$$r_i^k(t) = r_i^k(t) - \sum_{s=1}^{S} \lambda_s a_{is}^{k}.$$

References

1. MARKOWITZ, H., *Portfolio Selection*, Wiley, New York, New York, 1959.
2. CHARNES, A., and COOPER, W. W., *Chance Constrained Programming*, Management Science, Vol. 6, pp. 73-79, 1959.
3. HOGAN, A. J., MORRIS, J. G., and THOMPSON, H. E., *Decision Problems under Risk and Chance Constrained Programming: Dilemmas in the Transition*, Management Science, Vol. 27, pp. 698-716, 1981.
4. JACQUETTE, S. C., *A Utility Criterion for Markov Decision Processes*, Management Science, Vol. 23, pp. 43-49, 1979.
5. JACQUETTE, S. C., *Markov Decision Processes with a New Optimality Criterion, Small Interest Rates*, Annals of Mathematical Statistics, Vol. 1, pp. 1894-1901, 1973.
6. PORTEUS, E. L., *On the Optimality of Structure Policies in Countable Stage Decision Processes*, Management Science, Vol. 22, pp. 148-157, 1975.
7. WHITE, C. C., *The Optimality of Isotone Strategies for Markov Decision Problems with Utility Criterion*, Recent Developments in Markov Decision Processes, Edited by R. Hartley, L. C. Thomas, and D. J. White, Academic Press, New York, New York, 1980.
8. HOWARD, R. A., and MATHESON, J. E., *Risk-Sensitive Markov Decision Processes*, Management Science, Vol. 8, pp. 356-369, 1972.
9. KREPS, D. M., *Decision Problems with Expected Utility Criteria, I: Upper and Lower Convergent Utility*, Mathematics of Operations Research, Vol. 2, pp. 45-53, 1977.
10. KREPS, D. M., *Decision Problems with Expected Utility Criteria, II: Stationarity*, Mathematics of Operations Research, Vol. 2, pp. 266-274, 1977.
11. ROTHBLUM, U. G., *Multiplicative Markov Decision Chains*, Mathematics of Operations Research, Vol. 9, pp. 6-24, 1984.
12. SOBEL, M. J., *Ordinal Dynamic Programming*, Management Science, Vol. 21, pp. 967-975, 1975.
13. KALLENBERG, L. C. M., *Linear Programming and Finite Markovian Control Problems*, Mathematisch Centrum, Amsterdam, Holland, 1983.
14. SOBEL, M. J., *The Variance of Discounted Markov Decision Processes*, Journal of Applied Probability, Vol. 19, pp. 774-802, 1982.
15. MILLER, B., *On Dynamic Programming for a Stochastic Markovian Process with an Application to the Mean Variance Models*, Management Science, Vol. 24, p. 1779, 1978.
16. WHITE, D. J., *Probabilistic Constraints and Variance in Markov Decision Processes*, University of Manchester, Department of Decision Theory, Notes in Decision Theory, No. 149, 1984.
17. DERMAN, C., *Finite State Markovian Decision Processes*, Academic Press, New York, New York, 1970.
18. VAN DER WAL, J., *Stochastic Dynamic Programming*, Mathematisch Centrum, Amsterdam, Holland, 1981.
19. DERMAN, C., *On Sequential Control Procedures*, Annals of Mathematical Statistics, Vol. 35, pp. 341-349, 1964.

20. DERMAN, C., and STRAUCH, R., *A Note on Memoryless Rules for Controlling Sequential Control Processes*, Annals of Mathematical Statistics, Vol. 37, pp. 276–278, 1966.
21. HARTLEY, R., *Finite, Discounted, Vector Markov Decision Processes*, University of Manchester, Department of Decision Theory, Notes in Decision Theory, No. 85, 1979.
22. DERMAN, C., *Stable Sequential Control Rules and Markov Chains*, Journal of Mathematical Analysis and Applications, Vol. 6, pp. 257–265, 1963.
23. HORDJIK, A., and KALLENBERG, L. C. M., *Constrained Stochastic Dynamic Programming*, Mathematics of Operations Research, Vol. 9, pp. 276–289, 1984.
24. DERMAN, C., and VEINOTT, A. F., *Constrained Markov Decision Chains*, Management Science, Vol. 19, pp. 389–390, 1972.
25. STRAUCH, R., and VEINOTT, A., *A Property of Sequential Control Processes*, The Rand Corporation, Santa Monica, California, Research Memorandum No. RM 14772, 1966.
26. WHITE, D. J., *Utility, Probabilistic Constraints, Mean, and Variance in Markov Decision Processes*, University of Manchester, Notes in Decision Theory, No. 163, 1985.
27. DERMAN, C., and KLEIN, M., *Some Remarks on Finite-Horizon Markovian Decision Models*, Operations Research, Vol. 13, pp. 272–278, 1965.
28. WHITE, D. J., *Dynamic Programming with Probabilistic Constraints*, Operations Research, Vol. 22, pp. 654–664, 1972.
29. DERMAN, C., *Optimal Replacement under Markovian Deterioration with Probability Bounds on Failure*, Management Science, Vol. 9, pp. 478–481, 1963.
30. DANTZIG, G. B., and WOLFE, P., *The Decomposition Algorithm for Linear Programming*, Econometrica, Vol. 29, pp. 767–778, 1961.
31. HOWARD, R. A., *Dynamic Programming and Markov Processes*, Massachusetts Institute of Technology, PhD Thesis, 1960.
32. FILAR, J. A., and LEE, H. M., *Gain Variability Tradeoffs in Undiscounted Markov Decision Processes*, Proceedings of the 24th IEEE Conference on Decision and Control, pp. 1106–1112, 1985.
33. WHITE, D. J., *Optimality and Efficiency*, Wiley, Now York, New York, 1982.
34. MENDELSSOHN, R., *A Systematic Approach to Determining Mean Variance Tradeoffs when Managing Randomly Varying Populations*, Mathematical Biosciences, Vol. 50, pp. 75–84, 1980.
35. FILAR, J. A., *Percentiles and Markovian Decision Proceesses*, Operations Research Letters, Vol. 2, pp. 13–15, 1980.
36. WHITE, D. J., *Fundamentals of Decision Theory*, North-Holland, New York, New York, 1976.
37. WHITE, D. J., *Minimizing Threshold Probabilities in Infinite-Horizon Discounted Markov Decision Processes*, University of Manchester, Department of Decision Theory, Notes in Decision Theory, No. 165, 1985.
38. HENIG, M., *Optimality in Dynamic Programming with Deterministic Transitions and Stochastic Rewards*, Tel Aviv University, Faculty of Management, Working Paper No. 721/82, 1982.

39. HENIG, M., *Target and Percentile Criteria in Dynamic Programming with Deterministic Transitions and Stochastic Rewards*, University of Illinois at Urbana-Champaign, Department of Business Administration, 1984.

40. CHARNES, A. and COOPER, W. W., *Chance Constraints and Normal Deviates*, Journal of the American Statistical Association, Vol. 57, pp. 134–148, 1962.

41. GOLDWERGER, J., *Dynamic Programming of a Stochastic Markovian Process with an Application to the Mean Variance Models*, Management Science, Vol. 23, pp. 612–620, 1977.

42. PARKS, M. S., and STEINBERG, E., *A Preference Order Dynamic Program for a Knapsack Problem with Stochastic Rewards*, Journal of the Operational Research Society, Vol. 30, pp. 141–147, 1979.

43. SNEIDOVITCH, M., *Preference Order Stochastic Knapsack Problems: Methodological Issues*, Journal of the Operation Research Society, Vol. 31, pp. 1025–1032, 1980.

44. SNEIDOVITCH, M., *A Class of Variance Constrained Problems*, Operations Research, Vol. 31, pp. 338–353, 1983.

45. GREENBERG, H., *Dynamic Programming with Linear Uncertainty*, Operations Research, Vol. 16, pp. 675–678, 1968.

46. BEJA, A., *Probability Bounds in Replacement Policies for Markov Systems*, Management Science, Vol. 16, pp. 253–264, 1969.

47. BOUAKIZ, M., *Risk Sensitivity in Stochastic Optimization with Applications*, Georgia Institute of Technology, PhD Thesis, 1985.

48. CHUNG, K. J., *Some Topics in Risk-Sensitive Stochastic Dynamic Models*, Georgia Institute of Technology, PhD Thesis, 1985.

49. FILAR, J. A., and LEE, H. M., *Gain Variability Tradeoffs in Discounted Markov Decision Processes*, Johns Hopkins University, Department of Mathematical Sciences, Technical Report No. 408, 1985.

50. LEE, H. M., *Gain Variability Tradeoffs in Markovian Decision Processes and Related Problems*, Johns Hopkins University, Department of Mathematical Sciences, PhD Thesis, 1985.

51. SOBEL, M. J., *Mean-Variance Tradeoffs in an Undiscounted MDP*, Georgia Institute of Technology, Research Memorandum, 1984.

52. SOBEL, M. J., *Maximal Mean/Variance Ratio in an Undiscounted MDP*, Georgia Institute of Technology, Research Memorandum, 1985.