JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS **225,** 193–223 (1998)
ARTICLE NO. AY986015

Optimization Models for the First Arrival Target Distribution Function in Discrete Time

Stella X. Yu*

Department of Automation, Tsinghua University

Yuanlie Lin

Department of Applied Mathematics, Tsinghua University, Beijing 100084, People's Republic of China

and

Pingfan Yan

Department of Automation, Tsinghua University

Submitted by Augustine O. Esogbue

Received July 15, 1996

1. INTRODUCTION

The study of expectation optimality criteria (standard criteria) has constituted most previous work in the area of Markov decision processes (MDPs). However, the optimal policies obtained from such models are not reliable when considering a single or a few decision processes, since only the average performance over many trials is guaranteed to be optimal. In fact, the expectation optimality criteria are insufficient to characterize the variability–risk features of practical problems [1–5]. A typical requirement for a long-term application, for example, unmanned space flight and satellites, is to have a 0.95 or greater probability of being operational at the end of a 10-year period, whereas a typical requirement for space shuttle, aircraft flight control, and military systems is to have a reliability of 0.97 at the end of a 3-h time period. Likewise, chemical reactions must

* To whom correspondence should be addressed, at Center for the Neural Basis of Cognition, Carnegie Mellon University, 115 Mellon Institute, 4400 Fifth Ave., Pittsburgh, PA 15213-2683. E-mail: stella.yu@cmu.edu.

0022-247X/98 $25.00

be precisely controlled to prevent explosions or other unwanted effects. In some controllable stochastic dynamic systems [2, 3, 6, 7], it is desirable to maximize the reliability of normal operation. For the optimal regulation of a hydropower station reservoir considered in [8, 9], the probability of generating electric power at more than some given level should be no less than 0.95 for whatever working state the plant is in. In insurance services, the risk of total capital being less than some lower limit should generally be avoided as much as possible. In dynamic portfolio selection, investors are interested in strategies that can help them reach a given profit with maximal probability [10].

In all of these applications, which demand high reliability, system performance is controlled on a single trial basis, and thus the task requirements are formulated as probabilities rather than expectations. There have been some papers devoted to the probability criteria for various rewards. References [11, 12] studied the percentile performance criteria for the limiting average return. References [8, 13] considered the threshold probability criteria for discounted MDPs and focused on the properties of the optimality equations without discussion of the existence and properties of the optimal policies. We are further motivated to investigate the stochastic order [3–5, 10, 13–14] optimization problems, mainly on the distribution function criteria for nondiscounted first arrival target total reward [9, 15–22].

In this paper, the target level problems are recast into the total reward and optimal stopping setting. The target is a prescribed set of system states, corresponding to the failure set in reliability applications. Once the system is in one of these states, the decision process is terminated. Different terminal states may have different exit rewards. For a policy π, the first arrival target total return $W(\pi)$ is the sum of single stage rewards plus the exit reward upon system's first visit to the target. The objective function of this model, $V_i(\pi, l)$, is defined as the probability that the total reward exceeds a certain reward level l when the initial state is i. For example, the optimal regulation of a hydropower station reservoir should be to maximize the probability that electric power generation is more than some given value under normal water levels. The general optimization model is to find a policy π that maximizes $V_i(\pi, x)$ for every initial state i and some return levels of interest. Three classes of the set of these levels, namely, the infinite interval, a finite interval, and a single point, are studied in this paper.

We begin by describing these models in Section 2. The basic recursive properties of the objective functions are shown in Section 3. The convex combination and various cut-and-paste properties (in the spirit of [23]) of the optimal policies are presented in Section 4. The value functions and the optimal action sets are introduced in Section 5, and the optimality

equation and the optimality conditions for all three classes of return level sets are established. These results are refined in Section 6 for finite state space and action space. It is shown that nonstationary deterministic optimal policies must exist for the single point optimization problem. If a finite/countable intersection of the optimal action sets is not empty, then the finite interval/infinite interval optimal policies must exist. An algorithm is developed for computing the value functions and the optimal action sets, from which any optimal policy can be derived. In Section 7, numerical examples and computational experiments are used to illustrate the existence and structure of the optimal policies for these models. The optimality constraints on system parameters are probed as well.

2. MODEL DESCRIPTION

Let S be the *state space* with countable system states, denoted by $S = \{0, 1, 2, \dots\}$. For each $i \in S$, $A(i)$ is the set of all possible actions when the system is in state i. $A(i)$ is also countable. Let A be the *action space* or *control space*, where $A = \bigtimes_{i \in S} A(i)$. The transition law of the homogeneous controlled Markov chain is $q(j_{n+1} \mid j_0, a_0, \dots, j_n, a_n) = q(j_{n+1} \mid j_n, a_n)$, $n \geq 0$, $j_k \in S$, $a_k \in A(j_k)$, $0 \leq k \leq n$, $j_{n+1} \in S$. Let h_n denote the *history of the process up to stage n*, $h_n = (j_0, a_0, \dots, j_n, a_n)$, $j_k \in S$, $a_k \in A(j_k)$, $0 \leq k \leq n$. Throughout the paper, h_{-1} is assumed to be empty. $H_n = \{h_n\}$ is the set of all possible n-stage histories. A *policy* is a series of decision rules, denoted by $\pi = (\pi_0, \pi_1, \dots, \pi_n, \dots)$, where $\pi_n(a_n \mid h_{n-1}, j_n)$ is a probability measure on $A(j_n)$. Given an initial distribution and a control policy π, an MDP denoted by $(Y, \Delta) = \{(Y_n(\pi), \Delta_n(\pi)), n \geq 0\}$ can be uniquely determined with probability 1, where $Y_n(\pi)$ and $\Delta_n(\pi)$ represent the state and the action at stage n, respectively. Let the *policy space* $\Pi = \{\pi\}$ be the set of all policies. In particular, $f^\infty = \{f, f, \dots\}$ is said to be a *deterministic stationary policy* where f is a *decision function* mapping the state space S into the action space A. Let $F = \{f\}$ and $\Pi_s^d = \{f^\infty\}$ be the set of all of the decision functions and all of the deterministic stationary policies, respectively. For $n \geq 0$, θ_n is called a *history-dependent decision function* if for any given $h_{n-1} \in H_{n-1}$, $\theta_n(\cdot \mid h_{n-1}) \in F$. A policy that is made up entirely of history-dependent decision functions is deterministic but not stationary.

The general optimization models for the first arrival target distribution function in discrete time are specified by the nine-tuple $\{S, A, q, S_0, T, r, e, W, L\}$, where $S_0 \subset S$ is the *target set*; T is the *first arrival time at the target set* S_0; r is the *one stage running reward function* in the nontarget set $S_1 = S - S_0$; e is the *exit reward function* or *terminal reward function* in the target set S_0; W is the *total return until the first arrival time* T. The objective

function $V_i(\pi, x)$ is the remaining distribution function of W. For any policy $\pi \in \Pi$, let

$$T = \inf\{n : Y_k(\pi) \in S_1, 0 < k \le n-1, Y_n(\pi) \in S_0, n \ge 0\}; \quad (1)$$

$$W(\pi) = \sum_{0 \le n \le T-1} r(Y_n(\pi), \Delta_n(\pi)) + e(Y_T(\pi)); \quad (2)$$

$$V_i(\pi, x) = P\{W(\pi) > x \mid Y_0(\pi) = i\}, i \in S, x \in R, \quad (3)$$

where R is the set of all real numbers.[1] Obviously, $T = 0$ if $Y_0(\pi) \in S_0$; $T = +\infty$ if the set in (1) is empty.[2] The states in S_0 are assumed to be absorbing, i.e., there exists an *exit decision function* χ, such that the action $\chi(i)$ will be used whenever the system reaches $i \in S_0$ and $q(i \mid i, \chi(i)) = 1$. L is the *return level set* of interest in optimization. A policy π^* is *optimal for the first arrival target distribution function* in L if

$$V_i(\pi^*, x) \ge V_i(\pi, x), \qquad \forall \pi \in \Pi, \quad i \in S, \quad x \in L. \quad (4)$$

Let $\Pi^*(L)$ be the set of all optimal policies,

$$\Pi^*(L) = \{\pi^* : V_i(\pi^*, x) \ge V_i(\pi, x), \qquad \forall \pi \in \Pi, \quad i \in S, \quad x \in L\}. \quad (5)$$

$\Pi^*(L)$ is a nonincreasing set function of L:

$$\Pi^*(L_1) \subseteq \Pi^*(L_2), \qquad L_2 \subseteq L_1. \quad (6)$$

Moreover, for any index set K,

$$\Pi^*\left(\bigcup_{k \in K} L_k\right) = \bigcap_{k \in K} \Pi^*(L_k). \quad (7)$$

In particular, three classes of L are considered in this paper:

I. $L = R$ for the *complete stochastic order* optimization model.

II. $L = [0, l]$, $l > 0$ for the *local stochastic order* optimization model.

III. $L = \{l\}$ for the *single point stochastic order* optimization model.

The reward functions are assumed to be nonnegative; thus $L = R$ in case I can be reduced to $L = [0, +\infty)$. From (6) and (7), it is straightforward to

[1] As seen in this definition, the objective function is nonlinear with respect to a constant shift in the running reward functions. If r is increased by a constant c, then the objective function becomes $P\{W(\pi) + Tc > x \mid Y_0(\pi) = i\}$. Determining the effects of a constant shift in running rewards is still an open problem.

[2] In this case, $V_i(\pi, x) = 1$ for all x.

show that

$$\Pi^*([0, +\infty)) \subseteq \Pi^*([0, l]) \subseteq \Pi^*(\{x\}), \qquad \forall 0 \le x \le l < +\infty, \quad (8)$$

and

$$\Pi^*([0, +\infty)) = \bigcap_{x \in [0, +\infty)} \Pi^*(\{x\}). \quad (9)$$

The above models characterize decision and optimization problems in several important application areas. For reliability engineering, let the target set S_0 correspond to the set of all failure states, and S_1 corresponds to the set of all running states. To maximize the objective function $V_i(\pi, x)$ is to find an optimal policy to control the system toward yielding the desired outcome level until it breaks down. If a decision-maker has a profit goal in mind, he might want to use the single point stochastic order optimal policy, which reaches the given level of profit with maximum reliability. If he is not sure whether this level is reachable, for instance, if the chance of earning this much profit is unacceptably small, he might conservatively consider using the local stochastic order optimal policy, which ensures the maximum reliability for any profit below this given level. This approach can be regarded as an extension of multilevel optimization in many applications. For example, the probability of generated electrical power being more than a should be no less than 0.95, while the probability of the power being more than b, $b < a$, should be no less than 0.98, etc. Since these values, 0.95 or 0.98, might be impossible to achieve, a reasonable approach is to maximize both the probabilities of $l > a$ and $l > b$, where l is the generated electrical power. The ideal and dominant optimal policy is the complete stochastic order optimal policy, which consistently provides the maximum reliability for any outcome level.

3. BASIC PROPERTIES OF THE OBJECTIVE FUNCTIONS

In this section, the recursive equations of the objective functions are derived. With the assumptions of positive running rewards, it is shown that the objective function of any policy and any given level is determined by the first finite stages' decision rules.

First we introduce some notation and assumptions[3]:

$$r_m(i) = \inf\{r(i, a): a \in A(i)\}, \quad i \in S_1;$$
$$r_{\min} = \inf\{r_m(i): i \in S_1\}; \; e_{\min} = \inf\{e(j): j \in S_0\}; \quad (10)$$

[3] Throughout the paper, a subscript small m denotes some inferior limit/minimum value, and a subscript capital M denotes some superior limit/maximum value.

$$l_m(i) = (r_m(i) + e_{\min})I(i \in S_1) + e(i)I(i \in S_0);$$
$$l_{\min} = \inf\{l_m(i): i \in S_1\} = r_{\min} + e_{\min}. \tag{11}$$

In the above, I is the indicator function: $I(\text{True}) = 1$, $I(\text{False}) = 0$. The reward functions are assumed to be bounded: $r_{\min} > 0$, $e_{\min} \geq 0$, $\sup\{r(i,a): i \in S_1, a \in A(i)\} \leq B$; $\sup\{e(i): i \in S_0\} \leq B$. To simplify notation, the definition of the running reward function is extended to the target set, i.e., $r(i,a) = 0$, $\forall i \in S_0$, $a \in A(i)$. Let $n \in N = \{0,1,2,\dots\}$. For $\pi = (\pi_0, \pi_1, \dots, \pi_n, \dots) \in \Pi$, ${}^{[n]}\pi = (\pi_0, \pi_1, \dots, \pi_{n-1})$ denotes the *truncation of* π *to n stages*. Given history h_{n-1}, $\pi^{h_{n-1}} = (\pi_0^{h_{n-1}}, \pi_1^{n_{n-1}}, \dots)$ denotes the *n-remainder policy*, where

$$\pi_k^{h_{n-1}}(\cdot \mid h_{k-1}, j_k) = \pi_{n+k}(\cdot \mid h_{n-1}, h_{k-1}, j_k),$$
$$h_{n-1} \in H_{n-1}, \qquad j_k \in S, \qquad h_{k-1} \in H_{k-1}, \qquad k \geq 0. \tag{12}$$

Let ${}^{[n]}\Pi = \{{}^{[n]}\pi: \pi \in \Pi\}$ and $\pi^{[n]} = (\pi_n, \pi_{n+1}, \dots)$ for simplicity.

THEOREM 1. *For any* $\pi = (\pi_0, \pi_1, \dots, \pi_n, \dots) \in \Pi$,

$$V_i(\pi, x) = I(e(i) > x), \qquad \forall i \in S_0, \quad x \in (-\infty, +\infty); \tag{13}$$
$$V_i(\pi, x) = 1, \qquad \forall i \in S_1, \quad x < l_m(i); \tag{14}$$
$$V_i(\pi, x) = \sum_{a \in A(i)} \pi_0(a \mid i) \sum_{j \in S} q(j \mid i, a) V_j(\pi^{(i,a)}, x - r(i,a)),$$
$$\forall i \in S_1, \quad x \geq l_m(i), \tag{15}$$

where $l_m(i)$ *is the least total reward for initial state* i *and* $\pi^{(i,a)} = \pi^{h_0}$ *is the remainder policy after the zero-stage history* $h_0 = (i,a)$, $i \in S_1$, $a \in A(i)$.

Proof. (13) and (14) follow from the definition of $V_i(\pi, x)$ and the nonnegativity of $W(\pi)$. Here is the proof for (15). For $i \in S_1$, $x \geq l_m(i)$, from the homogeneity and the Markov property of q, we have

$$V_i(\pi, x) = P\{W(\pi) > x \mid Y_0(\pi) = i\}$$
$$= P\left\{W(\pi) > x, \bigcup_{a \in A(i)} (\Delta_0(\pi) = a) \,\middle|\, Y_0(\pi) = i\right\}$$
$$= \sum_{a \in A(i)} \pi_0(a \mid i) P\left\{W(\pi) > x, \right.$$
$$\left. \bigcup_{j \in S} (Y_1(\pi) = j) \,\middle|\, Y_0(\pi) = I, \Delta_0(\pi) = a\right\}$$

$$= \sum_{a \in A(i)} \pi_0(a \mid i) \times \sum_{j \in S} q(j \mid i, a) P\{W(\pi^{(i,a)}) > x - r(i,a) \mid Y_0(\pi) = j\}$$

$$= \sum_{a \in A(i)} \pi_0(a \mid i) \sum_{j \in S} q(j \mid i, a) V_j(\pi^{(i,a)}, x - r(i,a)).$$

∎

COROLLARY 1.1. *For any* $\pi = (\pi_0, \pi_1, \ldots, \pi_n, \ldots) \in \Pi$, $i \in S_1$, *and* $x \geq l_m(i)$,

$$V_i(\pi, x) = \sum_{a \in A(i)} \pi_0(a \mid i) \Big\{ \sum_{j \in S_0} q(j \mid i, a) I(e(j) > x - r(i,a)) + \sum_{j \in S_1} q(j \mid i, a) V_j(\pi^{(i,a)}, x - r(i,a)) \Big\}. \quad (16)$$

Since the states in the target set S_0 are assumed to be absorbing, (15) holds for $i \in S_0$. In other words, as $\pi_0(\chi(i) \mid i) = 1$, $q(i \mid i, \chi(i)) = 1$, and $r(i, \chi(i)) = 0$, $V_i(\pi, x) = \sum_{a \in A(i)} \pi_0(a \mid i) \sum_{j \in S} q(j \mid i, a) V_j(\pi^{(i,a)}, x - r(i,a)) = V_i(\pi^{(i,a)}, x) = I(e(i) > x)$. Thus Corollary 1.2 follows.

COROLLARY 1.2. *For any* $\pi = (\pi_0, \pi_1, \ldots, \pi_n, \ldots) \in \Pi$, $j_0 \in S_1$, *and* $x \geq l_m(j_0)$,

$$V_{j_0}(\pi, x) = \sum_{h_{n-1} \in H_{n-1, j_n} \in S} P_\pi\{h_{n-1}, j_n \mid j_0\} V_{j_n}(\pi^{h_{n-1}}, x - \varphi(h_{n-1})), \qquad n \geq 0, \quad (17)$$

where $h_{n-1} = (j_0, a_0, \ldots, j_{n-1}, a_{n-1}) \in H_{n-1}$, $h_{-1} = \varnothing$; $P_\pi\{h_{n-1}, j_n \mid j_0\}$ *is the transition probability from initial state* j_0 *to state* j_n *via history* h_{n-1} *under policy* π, *and* $\varphi(h_{n-1})$ *is the total reward for history* h_{n-1}, *i.e.*,

$$P_\pi\{h_{n-1}, j_n \mid j_0\} = \bigtimes_{0 \leq k \leq n-1} \pi_k(a_k \mid h_{k-1}, j_k) q(j_{k+1} \mid j_k, a_k), \; P_\pi\{h_{-1}, j_0 \mid j_0\} = 1; \quad (18)$$

$$\varphi(h_{n-1}) = \sum_{0 \leq k \leq n-1} r(j_k, a_k), \qquad \varphi(h_{-1}) = 0. \quad (19)$$

Therefore, the objective functions of any two given policies should only be compared for $i \in S_1$ and $x \geq l_m(i)$. The optimization problems for the complete stochastic order, the local stochastic order, and the single point stochastic order can be simplified to (20), (21), and (22), respectively:

$$V_i(\pi^*, x) \geq V_i(\pi, x), \qquad \forall \pi \in \Pi, \quad i \in S_1, \quad \text{and} \quad x \in [l_m(i), +\infty); \tag{20}$$

$$V_i(\pi^*, x) \geq V_i(\pi, x), \qquad \forall \pi \in \Pi, \quad i \in S_1, \quad x \in [l_m(i), l]; \tag{21}$$

$$V_i(\pi^*, l) \geq V_i(\pi, l), \qquad \forall \pi \in \Pi, \quad i \in S_1, \quad l \geq l_m(i). \tag{22}$$

THEOREM 2. *Given* $i \in S_1$ *and level* l, $\exists \underline{n}(i, l) \geq 0$, *s.t.* $\forall \pi = (\pi_0, \pi_1, \ldots, \pi_n, \ldots) \in \Pi$, $V_i(\pi, l)$ *only depends on* ${}^{[\underline{n}(i,l)]}\pi$, *where*[4]

$$\underline{n}(i, l) = \lceil (l - e_{\min} - r_m(i))/r_{\min} \rceil. \tag{23}$$

The function $\lceil x \rceil$ *is the smallest integer greater than or equal to* x.

Proof. Given $i \in S_1$ and level l, $\forall \pi = (\pi_0, \pi_1, \ldots, \pi_n, \ldots) \in \Pi$ and $n \geq 0$, $V_i(\pi, x) = \sum_{h_{n-1} \in H_{n-1}, j_n \in S} P_\pi\{h_{n-1}, j_n \mid j_0\} V_{j_n}(\pi^{h_{n-1}}, x - \varphi(h_{n-1}))$, $h_{n-1} = (j_0, a_0, \ldots, j_{n-1}) \in H_{n-1}$, $j_0 = i$. Given q, $P_\pi\{h_{n-1}, j_n \mid i\}$ is determined by ${}^{[n]}\pi$. If $n = \underline{n}(i, l) = \lceil (l - e_{\min} - r_m(i))/r_{\min} \rceil = \inf\{k : l - [r_m(i) + (k-1)r_{\min}] < l_{\min}\}$, $V_{j_n}(\pi^{h_{n-1}}, l - \varphi(h_{n-1})) = 1$ for any $\pi \in \Pi$, $h_{n-1} \in H_{n-1}$, $j_n \in S_1$. Thus, $V_i(\pi, l)$ is determined by ${}^{[\underline{n}(i,l)]}\pi$.

COROLLARY 2.1. *Given* $i \in S_1$ *and* $n \in N$, $\exists \underline{l}(i, n)$, *s.t.* $\forall \pi = (\pi_0, \pi_1, \ldots, \pi_n, \ldots) \in \Pi$, $x < \underline{l}(i, n)$, $V_i(\pi, x)$ *only depends on* ${}^{[n]}\pi$, *where*

$$\underline{l}(i, n) = \sup\{x : \underline{n}(i, x) = n\} = e_{\min} + n \cdot r_{\min} + r_m(i). \tag{24}$$

Defining $\underline{l}(i, n)$ in (24), an alternative definition for $\underline{n}(i, l)$ is

$$\underline{n}(i, l) = \inf\{n : \underline{l}(i, n) > l, n \geq 0\}. \tag{25}$$

Obviously, for any $i \in S_1$, $\underline{n}(i, l)$ is a nondecreasing function in l, while $\underline{l}(i, n)$ is a linearly increasing function in n. Furthermore,

$$\underline{n}(i, \underline{l}(i, n)) = n, \qquad \underline{l}(i, \underline{n}(i, l)) \geq l. \tag{26}$$

We assume $\underline{n}(i, l) = 0$ and $\underline{l}(i, n) = +\infty$ for any $i \in S_0$, any real number l, and $n \in N$.

[4] The property that the objective function of any finite level is determined by the decision rules of a finite number of stages is guaranteed from the assumption $r_{\min} > 0$. There is much more work to do when $r_{\min} = 0$, since it changes some properties of the model, for instance, the objective functions are no longer recursive at some $r(i, a) = 0$.

Now, for any $i \in S$, $\underline{n}(i, l)$ is called the *truncation stage number* for the initial state i and the level l; $\underline{l}(i, n)$ is called the *reachable return level* for the initial state i and the stage number n. Therefore, only the decision rules of a finite number of stages starting from the initial stage need to be considered in the local and the single point stochastic order optimization models.

4. CONVEX COMBINATION AND CUT-AND-PASTE PROPERTIES OF THE OPTIMAL POLICIES

To simplify notation, for any $f \in F$, $\pi \in \Pi$, $i \in S_1$, and $j \in S$, let

$$r(i, f) = r(i, f(i)), \qquad q(j \mid i, f) = q(j \mid i, f(i)), \qquad \pi^{(i,f)} = \pi^{(i, f(i))},$$
$$V_i(f, x - r(i, f)) = V_i(f^\infty, x - r(i, f(i))). \tag{27}$$

THEOREM 3. *If $\pi = (\pi_0, \pi_1, \ldots, \pi_n, \ldots) \in \Pi^*(L)$ and a decision function f satisfies $f(i) \in A_{\pi_0}(i)$ for any $i \in S_1$, where*

$$A_{\pi_0}(i) = \{a: a \in A(i), \pi_0(a \mid i) > 0\}, \tag{28}$$

then $f\pi^{[1]} = (f, \pi_1, \ldots, \pi_n, \ldots) \in \Pi^*(L)$.

Proof. For a decision function f with $f(i) \in A_{\pi_0}(i)$ for any $i \in S_1$, because $\pi \in \Pi^*(L)$, $V_i(\pi, x) \geq V_i(f\pi^{[1]}, x) = \sum_{j \in S} q(j \mid i, f) V_j(\pi^{(i,f)}, x - r(i, f))$ for any $i \in S_1$ and $x \in L$, which implies

$$V_i(\pi, x) \geq \sup\left\{ \sum_{j \in S} q(j \mid i, a) V_j(\pi^{(i,a)}, x - r(i, a)): a \in A_{\pi_0}(i) \right\}. \tag{29}$$

On the other hand,

$$V_i(\pi, x) = \sum_{a \in A_{\pi_0}(i)} \pi_0(a \mid i) \sum_{j \in S} q(j \mid i, a) V_j(\pi^{(i,a)}, x - r(i, a))$$
$$\leq \sup\left\{ \sum_{j \in S} q(j \mid i, a) V_j(\pi^{(i,a)}, x - r(i, a)): a \in A_{\pi_0}(i) \right\}. \tag{30}$$

Combining (29) and (30) yields

$$V_i(\pi, x) = \sup\left\{ \sum_{j \in S} q(j \mid i, a) V_j(\pi^{(i,a)}, x - r(i, a)): a \in A_{\pi_0}(i) \right\},$$
$$\forall \pi \in \Pi^*(L), \quad i \in S_1, \quad x \in L. \tag{31}$$

Now we will show $V_i(\pi, x) = V_i(f\pi^{[1]}, x)$ for any $i \in S_1$ and $x \in L$. If this does not hold, then there must exist some $i \in S_1$ and some $x \in L$ such that $V_i(f\pi^{[1]}, x) < V_i(\pi, x)$, i.e.,

$$\sum_{j \in S} q(j \mid i, f) V_j\big(\pi^{(i,f)}, x - r(i,f)\big) < \sup\left\{ \sum_{j \in S} q(j \mid i, a) V_j\big(\pi^{(i,a)}, x - r(i,a)\big) \colon a \in A_{\pi_0}(i) \right\}. \quad (32)$$

Thus,

$$\begin{aligned} V_i(\pi, x) = & \sum_{a \in A_{\pi_0}(i) - \{f(i)\}} \pi_0(a \mid i) \sum_{j \in S} q(j \mid i, a) V_j\big(\pi^{(i,a)}, x - r(i,a)\big) \\ & + \pi_0\big(f(i) \mid i\big) \sum_{j \in S} q(j \mid i, f) V_j\big(\pi^{(i,f)}, x - r(i,f)\big) \\ & < \sup\left\{ \sum_{j \in S} q(j \mid i, a) V_j\big(\pi^{(i,a)}, x - r(i,a)\big) \colon a \in A_{\pi_0}(i) \right\}, \end{aligned}$$

contradicting (31). Hence $V_i(\pi, x) = V_i(f\pi^{[1]}, x)$, $f\pi^{[1]} \in \Pi^*(L)$. ∎

Let $\pi = (\pi_0, \pi_1, \ldots, \pi_n, \ldots) \in \Pi$ and K be some finite or countable index set. If there is a list of numbers $\{\alpha_n(k) \colon \alpha_n(k) \geq 0, \sum_{k \in K} \alpha_n(k) = 1, n \geq 0, k \in K\}$ such that $\pi_n = \sum_{k \in K} \alpha_n(k) \zeta_{k,n}$, $n \geq 0$, where $\zeta_{k,n}$ is the decision rule at stage n in policy $\zeta_k = (\zeta_{k,0}, \zeta_{k,1}, \ldots, \zeta_{k,n}, \ldots)$, $k \in K$, then π is called the *convex combination* of the policies $\{\zeta_k, k \in K\}$. In particular, if $\sum_{k \in K} \pi_0(f_k(i) \mid i) = 1$, then π is called the convex combination of the policies $\{f_k \pi^{[1]} = (f_k, \pi_1, \ldots, \pi_n, \ldots), k \in K\}$.

COROLLARY 3.1. *If $\pi = (\pi_0, \pi_1, \ldots, \pi_n, \ldots) \in \Pi^*(L)$, then π can be decomposed into a convex combination of optimal policies such as $f\pi^{[1]} = (f, \pi_1, \ldots, \pi_n, \ldots)$ in Theorem* 3.

Theorem 3 is valid for any return level set L. It shows that the first decision rule of any optimal policy need not be randomized. In other words, actions at the first stage with selection probability greater than zero can be assumed to be equal in producing an optimal policy. For complete and local stochastic order optimization models, if an optimal policy exists, then a deterministic stationary optimal policy must exist. This further result will be established in Theorem 4 and its corollaries after the introduction of the concept of concatenated policies below.

A concatenated policy is a new policy made up of decision rules from a set of known policies. For example, Theorem 3 shows a concatenated policy of decision rule f and the 1-remainder policy of π. This concatena-

tion can continue. Let $f \in F$ and $\pi = (f, \pi_1, \ldots, \pi_n, \ldots) \in \Pi$. $f\pi$ is the *concatenated policy* of f and π: $f\pi = (f, f, \pi_1^{(1)}, \ldots, \pi_n^{(1)}, \ldots)$, where $\pi_n^{(1)}(\cdot \mid i, f(i), h_{n-1}, j_n) = \pi_n(\cdot \mid h_{n-1}, j_n)$, $\forall i, j_n \in S, h_{n-1} \in H_{n-1}$. Without any confusion, $f\pi$ can be written as $(f, f, \pi_1, \ldots, \pi_n, \ldots)$ for short. Throughout the rest of the paper, detailed definitions for each stage of a concatenated policy will be omitted, and the policy is given by linking each composite policy's stages together.

THEOREM 4. *For* $L = [0, +\infty)$ *or* $[0, l]$, *if* $\pi = (f, \pi_1, \ldots, \pi_n, \ldots) \in \Pi^*(L)$, *then* $f\pi = (f, f, \pi_1, \ldots, \pi_n, \ldots) \in \Pi^*(L)$.

Proof. For $L = [0, +\infty)$ or $[0, l]$, because $\pi = (f, \pi_1, \ldots, \pi_n, \ldots) \in \Pi^*(L)$, for any $i \in S_1$ and $x \in L$, $V_i(\pi, x) \geq V_i(f\pi, x)$, $V_i(\pi, x) \geq V_i(\pi^{[1]}, x)$. Therefore, $V_i(f\pi, x) = \sum_{j \in S} q(j \mid i, f) V_j(\pi, x - r(i, f)) \geq \sum_{j \in S} q(j \mid i, f) V_j(\pi^{[1]}, x - r(i, f)) = V_i(\pi, x)$. Thus $V_i(f\pi, x) = V_i(\pi, x)$, implying $f\pi \in \Pi^*(L)$. ∎

As seen in the proof, if $L = \{l\}$, $V_i(f\pi, l) = V_i(\pi, l)$ cannot be obtained from the recursive equation, since $V_i(\pi, x) \geq V_i(\pi^{[1]}, x)$ does not hold for $x < l$. It will be shown in the next section that single point stochastic order optimal policies are generally not stationary.

COROLLARY 4.1. *For* $L = [0, +\infty)$ *or* $[0, l]$, *if* $\pi = (f, \pi_1, \ldots, \pi_n, \ldots) \in \Pi^*(L)$, *then* $f^\infty \in \Pi^*(L)$.

COROLLARY 4.2. *For* $L = [0, +\infty)$ *or* $[0, l]$, *if* $\Pi^*(L) \neq \varnothing$, *then* $\exists f \in F$, *s.t.* $f^\infty \in \Pi^*(L)$.

COROLLARY 4.3. *For* $L = [0, +\infty)$ *or* $[0, l]$, *if* $\pi \in \Pi^*(L)$, *then* π *can be decomposed into a convex combination of some deterministic stationary optimal policies*.

Therefore, for both complete and local stochastic order optimization problems, if the optimal policies do exist, then at least one deterministic stationary policy can be derived from the initial decision rule of an optimal policy. The existence ensures that the optimal policies can be discussed in the deterministic stationary policy set Π_s^d instead of the whole policy set Π.

Another result of studying the concatenated policies is the *cut-and-paste* properties of the optimal policies. For $\pi = (\pi_0, \pi_1, \ldots, \pi_n, \ldots) \in \Pi$ and $\pi' = (\pi_0', \pi_1', \ldots, \pi_n', \ldots) \in \Pi$, let ${}^{[n]}\pi\pi'$ be the concatenated policy of ${}^{[n]}\pi$ and π': ${}^{[n]}\pi\pi' = (\pi_0, \ldots, \pi_{n-1}, \pi_0', \pi_1', \ldots, \pi_n', \ldots)$.

THEOREM 5. *For* $L = [0, +\infty)$ *or* $[0, l]$, *if* $\pi, \pi' \in \Pi^*(L)$, *then* ${}^{[n]}\pi\pi' \in \Pi^*(L)$, $n = 1, 2, \ldots$.

Proof. Given $L = [0, +\infty)$ or $[0, l]$, since π, $\pi' \in \Pi^*(L)$, for any $j_0 \in S_1$ and $x \geq l_m(j_0)$, $V_{j_0}({}^{[n]}\pi\pi', x) = \sum_{h_{n-1} \in H_{n-1}, j_n \in S} P_\pi\{h_{n-1}, j_n \mid j_0\}$ $V_{j_n}(\pi', x - \varphi(h_{n-1})) \geq \sum_{h_{n-1} \in H_{n-1}, j_n \in S} P_\pi\{h_{n-1}, j_n \mid j_0\}$ $V_{j_n}(\pi^{h_{n-1}}, x - \varphi(h_{n-1})) = V_{j_0}(\pi, x)$. Thus $V_{j_0}({}^{[n]}\pi\pi', x) = V_{j_0}(\pi, x)$, ${}^{[n]}\pi\pi' \in \Pi^*(L)$. ∎

COROLLARY 5.1. *For $L = [0, +\infty)$ or $[0, l]$, if $\pi \in \Pi^*(L)$, then ${}^{[n]}\pi\pi \in \Pi^*(L)$; ${}^{[n_1]}\pi^{[n_2]}\pi \cdots {}^{[n_k]}\pi\pi \in \Pi^*(L)$; $n, k, n_1, n_2, \ldots, n_k = 1, 2, \ldots$.*

COROLLARY 5.2. *For $L = [0, +\infty)$ or $[0, l]$, if $\pi = (\pi_0, \pi_1, \ldots, \pi_n, \ldots) \in \Pi^*(L)$, then $\pi_0^n\pi = (\pi_0, \ldots, \pi_0, \pi_0, \pi_1, \ldots, \pi_n, \ldots) \in \Pi^*(L)$, $n = 1, 2, \ldots$, $\pi_0^\infty \in \Pi^*(L)$.*

For the complete and the local stochastic order optimal policies, we can cut off an arbitrary number of stages from the initial stage. The concatenated policy of these pieces of the optimal policies is still an optimal policy. In particular, the initial stage decision rule, which may be random, can constitute a stationary optimal policy. Now, can the cut operation be relaxed from any stage instead of the initial stage? Theorem 6 gives a sufficient condition to shift the cut properties of the optimal policies.

Given $j \in S_1$, $n > 0$, and policy π, if there exists some state $i \in S_1$, s.t. $P\{Y_n(\pi) = j \mid Y_0(\pi) = i\} > 0$, then the state j is said to be *n-step arrivable under policy π*.

THEOREM 6. *For $L = [0, +\infty)$ or $[0, l]$, if $\pi \in \Pi^*(L)$, each $j \in S_1$ is n-step arrivable under π, then the n-remainder policy $\pi^{[n]} \in \Pi^*(L - n \cdot r_M)$, where $r_M = \sup\{r(i, a)\colon i \in S_1, a \in A(i)\}$, $L - y = \{x - y\colon x \in L\}$.*

Proof. If the statement is not true, then $\exists j \in S_1$, $x \in L - n \cdot r_M$, s.t. $V_j(\pi, x) > V_j(\pi^{[n]}, x)$. Since j is n-step arrivable under π, there exists some state $j_0 \in S_1$, s.t. $P\{Y_n(\pi) = j \mid Y_0(\pi) = j_0\} > 0$, implying: $\exists h_{n-1} \in H_{n-1}$, s.t. $P_\pi\{h_{n-1}, j \mid j_0\} > 0$. Then $V_{j_0}({}^{[n]}\pi\pi, x + \varphi(h_{n-1})) > V_{j_0}(\pi, x + \varphi(h_{n-1}))$, contradicting $\pi \in \Pi^*(L)$, since $\varphi(h_{n-1}) < n \cdot r_M$. Therefore $\pi^{[n]} \in \Pi^*(L - n \cdot r_M)$. ∎

COROLLARY 6.1. *For $L = [0, +\infty)$ or $[0, l]$, if $\pi \in \Pi^*(L)$, and each $j \in S_1$ is n-step arrivable under π, then $\pi_n^\infty \in \Pi^*(L - n \cdot r_M)$.*

Therefore, for a complete stochastic order optimal policy, if all running states are arrivable at some stage, then the decision rule at this stage can also constitute a stationary optimal policy, while the remainder of the policy from this stage is still optimal.

5. THE VALUE FUNCTIONS AND THE OPTIMALITY CONDITIONS

The preceding section discussed the various properties of optimal polices based on the existence of the optimal policies. This section examines

the existence conditions of the optimal policies. To this end, the properties of the value functions need to be studied first. For any state $i \in S$ and $x \in R$, let $V_i^*(x)$ be the *value function*, where

$$V_i^*(x) = \sup\{V_i(\pi, x): \pi \in \Pi\}. \tag{33}$$

Property 1. For any $i \in S$, $V_i^*(x)$ is nonincreasing and right continuous; and,

$$V_i^*(x) = I(e(i) > x), \qquad \forall i \in S_0, x \in (-\infty, +\infty); \tag{34}$$

$$V_i^*(x) = 1, \qquad \forall i \in S_1, x < l_m(i); \tag{35}$$

$$V_i^*(x) \le \sup\left\{ \sum_{j \in S} q(j \mid i, f) V_j^*(x - r(i, f)): f \in F \right\},$$

$$\forall i \in S_1, \quad x \ge l_m(i). \tag{36}$$

Property 2. If $\Pi^*(L) \neq \emptyset$, then $\forall \pi \in \Pi^*(L)$, $V_i(\pi, x) = V_i^*(x)$ for all $i \in S_1$ and $x \in L$.

Property 3. If $\Pi^*(L) \neq \emptyset$ and $P\{T < +\infty \mid Y_0(\pi) = i\} = 1$ for any $i \in S_1$ and $\pi \in \Pi$, then $\lim_{x \to +\infty} V_i^*(x) = 0$ for any $i \in S_1$.

Proof. Since $\sup\{r(i, a): i \in S_1, a \in A(i)\} \le B$, $\sup\{e(i): i \in S_0\} \le B$ and $\pi \in \Pi^*(L) \neq \emptyset$, for any $i \in S_1$, $x \in L$, $V_i(\pi, x) = P\{W(\pi) > x \mid Y_0(\pi) = i\} \le P\{TB + B > x \mid Y_0(\pi) = i\} = P\{T > x/(B + 1) \mid Y_0(\pi) = i\}$. Because $P\{T < +\infty \mid Y_0(\pi) = i\} = 1$ for any $i \in S_1$ and $\pi \in \Pi$, the right-hand side of the above inequality will approach 0 as x approaches $+\infty$. By Property 2 it follows immediately that $\lim_{x \to +\infty} V_i^*(x) = 0$. ∎

Properties 1 and 3 show that the value functions are the optimal remaining distribution functions under very weak conditions.

Property 4. For $L = [0, +\infty)$ or $[0, l]$, if $\Pi^*(L) \neq \emptyset$, then $\exists g \in F$, s.t. $V_i(g, x) = V_i^*(x) = \sup\{V_i(f, x): f \in F\}$ for all $i \in S_1$ and $x \in L$.

Proof. For any $i \in S_1$ and $x \in L$, $V_i^*(x) = \sup\{V_i(\pi, x): \pi \in \Pi\} \ge \sup\{V_i(f, x): f \in F\}$. If $\Pi^*(L) \neq \emptyset$, where $L = [0, +\infty)$ or $[0, l]$, then $\exists g \in \Pi^*(L)$, s.t. $V_i^*(x) = V_i(g, x) \le \sup\{V_i(f, x): f \in F\}$, completing the proof. ∎

Next, the optimality equations will be derived by casting the optimization problem into a more general model, namely, the ε-optimization model. Given a small positive number ε, π^ε is called an *ε-optimal policy for the first arrival target distribution function* if

$$V_i^*(x) - \varepsilon \le V_i(\pi^\varepsilon, x) \le V_i^*(x), \qquad \forall i \in S_1, \quad x \in L. \tag{37}$$

Let $\Pi^*(L, \varepsilon)$ be the set of all ε-optimal policies:

$$\Pi^*(L, \varepsilon) = \{\pi^\varepsilon : V_i^*(x) - \varepsilon \le V_i(\pi^\varepsilon, x) \le V_i^*(x), \quad \forall i \in S_1, \quad x \in L\}. \tag{38}$$

$\Pi^*(L, \varepsilon)$ is nonincreasing in L while nondecreasing in ε:

$$\Pi^*(L_1, \varepsilon) \subseteq \Pi^*(L_2, \varepsilon), \quad \forall L_2 \subseteq L_1;$$
$$\Pi^*(L, \varepsilon_1) \subseteq \Pi^*(L, \varepsilon_2), \quad \forall \varepsilon_1 \le \varepsilon_2. \tag{39}$$

Similar to (7), for any ε and index set K,

$$\Pi^*\Big(\bigcup_{k\in K} L_k, \varepsilon\Big) = \bigcap_{k\in K} \Pi^*(L_k, \varepsilon). \tag{40}$$

In particular,

$$\Pi^*(L) = \Pi^*(L, 0) \subseteq \Pi^*(L, \varepsilon) \subseteq \Pi^*(L, \infty) = \Pi; \tag{41}$$
$$\Pi^*(L) \subseteq \lim_{\varepsilon \to 0} \Pi^*(L, \varepsilon). \tag{42}$$

THEOREM 7. *For any given l and $\varepsilon > 0$, $\Pi^*(\{l\}, \varepsilon) \neq \emptyset$.*

Proof. Given l and ε, for each $i \in S_1$, there must exist some policy $\pi[i] \in \Pi$, s.t. $V_i^*(l) - \varepsilon \le V_i(\pi[i], l) \le V_i^*(l)$. Let π^ε be a composite policy of these policies: $\{\pi[i], i \in S_1\}$, such that for any history beginning from state i, policy $\pi[i]$ is used, i.e., $\pi_n(\cdot \mid h_{n-1}, j_n) = \pi[i]_n(\cdot \mid h_{n-1}, j_n)$ for any $h_{n-1} = (j_0, a_0, \ldots, j_{n-1}, a_{n-1}) \in H_{n-1}$, $j_0 = i, n \ge 0$. Thus $\pi^\varepsilon \in \Pi^*(\{l\}, \varepsilon) \neq \emptyset$. ∎

Therefore, there always exists an ε-optimal policy for any single point stochastic order ε-optimization models.

THEOREM 8. *For any $i \in S_1$ and $x \ge l_m(i)$, $V_i^*(x)$ is the unique solution that satisfies the initial conditions of* (34) *and* (35) *and the following optimality functional equations*:

$$V_i^*(x) = \sup\Big\{\sum_{j\in S} q(j \mid i, f) V_j^*(x - r(i, f)) : f \in F\Big\}. \tag{43}$$

Proof. Suppose there exists $f \in F$, $i \in S_1$, $x \ge l_m(i)$, such that $V_i^*(x) < \sum_{j\in S} q(j \mid i, f) V_j^*(x - r(i, f))$. Let α be such that $0 < \alpha < \sum_{j\in S} q(j \mid i, f) V_j^*(x - r(i, f)) - V_i^*(x)$. By Theorem 7, $\exists \pi \in \Pi^*(\{x - r(i, f)\}, \alpha)$, s.t. $V_i(f\pi, x) = \sum_{j\in S} q(j \mid i, f)\ V_j(\pi, x - r(i, f)) > \sum_{j\in S} q(j \mid i, f)\ V_j^*(x$

$- r(i, f)) - \alpha > V_i^*(x)$, a contradiction. Therefore, $V_i^*(x) \geq \sup\{\sum_{j \in S} q(j \mid i, f)V_j^*(x - r(i, f)): f \in F\}$. Combining this result with (36) proves the optimality equation. The uniqueness of the solution comes from the initial conditions in Property 1. ▌

THEOREM 9. $\Pi^*(L) \neq \varnothing \Leftrightarrow \exists \pi \in \Pi$, *s.t.* $\forall i \in S_1$, $x \in L$, $V_i(\pi, x) = \sup\{\sum_{j \in S} q(j \mid i, f)V_j^*(x - r(i, f)): f \in F\}$.

COROLLARY 9.1. *For* $L = [0, +\infty)$ *or* $[0, l]$, $\Pi^*(L) \neq \varnothing \Leftrightarrow \exists g \in F$, *s.t.* $\forall i \in S_1$, $x \in L$, $V_i(g, x) = \sup\{\sum_{j \in S} q(j \mid i, f)V_j^*(x - r(i, f)): f \in F\}$.

To link the optimality equations with action selection in decision, we introduce the concept of the optimal action set. Given $i \in S_1$ and $x \in (-\infty, +\infty)$, let $A_i^*(x)$ denote the *optimal action set*:

$$A_i^*(x) = \left\{a: V_i^*(x) = \sum_{j \in S} q(j \mid i, a)V_j^*(x - r(i, a)), a \in A(i)\right\}. \quad (44)$$

Obviously, $A_i^*(x) = A(i)$ when $x < l_m(i)$. We assume $A_i^*(x) = \{x(i)\}$, $i \in S_0$, $x \in R$. Given $l > 0$, $H_n^*(l)$, $n \geq 0$ is called the *optimal history up to stage n toward level l*, where

$$H_0^*(l) = \{h_0 = (j_0, a_0): j_0 \in S, a_0 \in A_{j_0}^*(l - \varphi(h_{-1})) = A_{j_0}^*(l)\}, \quad (45)$$

$$H_n^*(l) = H_{n-1}^*(l)\{(j_n, a_n): h_{n-1} \in H_{n-1}^*(l), j_n \in S, a_n \in A_{j_n}^*(l - \varphi(h_{n-1}))\}. \quad (46)$$

For $h_{n-1} = (j_0, a_0, \ldots, j_{n-1}, a_{n-1}) \in H_{n-1}$, let

$$P\{h_{n-1}, j_n \mid j_0\} = \max\{P_\pi\{h_{n-1}, j_n \mid j_0\}: \pi \in \Pi\} = \bigtimes_{0 \leq k \leq n-1} q(j_{k+1} \mid j_k, a_k). \quad (47)$$

By induction, the optimality equation can also be written as

$$V_{j_0}^*(l) = \sum_{j_1, j_2, \ldots, j_n \in S} P\{h_{n-1}, j_n \mid j_0\}V_{j_n}^*(l - \varphi(h_{n-1})), \quad h_{n-1} \in H_{n-1}^*(l), n = 1, 2, \ldots. \quad (48)$$

THEOREM 10. *For* $L = [0, +\infty)$ *or* $[0, l]$,

I. $\Pi^*(L) \neq \varnothing \Leftrightarrow \bigcap_{x \in L} A_i^*(x) \neq \varnothing, i \in S_1$.

II. *If* $\bigcap_{x \in L} A_i^*(x) \neq \varnothing$, *then* $f^\infty \in \Pi^*(L)$, *where* $f(i) \in \bigcap_{x \in L} A_i^*(x)$, $i \in S_1$.

Proof. "$\Leftarrow$": If $\forall i \in S_1$, $\bigcap_{x \in L} A_i^*(x) \neq \varnothing$, let f be a decision function such that $f(i) \in \bigcap_{x \in L} A_i^*(x)$, $i \in S_1$. From the recursive equation $V_i(f, x) = \sum_{j \in S} q(j \mid i, f) V_j(f, x - r(i, f))$, $V_i(f, x) = V_i^*(x)$ for all $x \in L$. Therefore $f^\infty \in \Pi^*(L)$, $\Pi^*(L) \neq \varnothing$.

"$\Rightarrow$": Given l, for $L = [0, +\infty)$ or $[0, l]$, if $\Pi^*(L) \neq \varnothing$, by Property 4, $\exists f \in F$, s.t. $\forall i \in S_1$, $x \in L$, $V_i(f, x) = V_i^*(x)$. Thus $V_i(f, x) = \sum_{j \in S} q(j \mid i, f) V_j(f, x - r(i, f)) = \sum_{j \in S} q(j \mid i, f) V_j^*(x - r(i, f)) = V_i^*(x)$, implying $f(i) \in A_i^*(x)$, $i \in S_1$, $x \in L$. Therefore $f(i) \in \bigcap_{x \in L} A_i^*(x) \neq \varnothing$, $\forall i \in S_1$. ∎

In addition to the existence of a deterministic stationary policy provided in Corollary 4.2, Theorem 10 also gives the form of that policy. If the complete or the local stochastic order optimal policies do exist, then there must exist deterministic stationary optimal policies and they can be derived from the intersection of the optimal action sets over L. For the single point stochastic order optimization model, the optimal policies may not be stationary.

THEOREM 11. *Given l,*

I. $\Pi^*(\{l\}) \neq \varnothing \Leftrightarrow A_{j_n}^*(l - \varphi(h_{n-1})) \neq \varnothing$, $h_{n-1} = (j_0, a_0, \ldots, j_{n-1}, a_{n-1}) \in H_{n-1}^*(l)$, $j_n \in S_1$, $0 \leq n \leq \underline{n}(j_0, l) - 1$.

II. *If $A_{j_n}^*(l - \varphi(h_{n-1})) \neq \varnothing$ for any $h_{n-1} \in H_{n-1}^*(l)$, $j_n \in S_1$, $0 \leq n \leq \underline{n}(j_0, l) - 1$, then ${}^{[\underline{n}_M(l)]}\pi = (\theta_0, \theta_1, \ldots, \theta_{\underline{n}_M(l)-1}) \in \Pi^*(\{l\})$, where $\underline{n}_M(l) = \sup\{\underline{n}(j_0, l)\colon j_0 \in S_1\}$,*

$$\theta_n(j_n \mid h_{n-1}) \in A_{j_n}^*(l - \varphi(h_{n-1})), \quad h_{n-1} \in H_{n-1}^*(l), \quad 0 \leq n \leq \underline{n}(j_0, l) - 1, \quad j_n \in S_1. \tag{49}$$

In addition, there must be

$$\pi^{h_{n-1}} \in \Pi^*\big(\{l - \varphi(h_{n-1})\}\big), h_{n-1} \in H_{n-1}^*(l), \qquad n \geq 0. \tag{50}$$

Proof. By Theorem 2, $\forall \pi \in \Pi$, $V_i(\pi, x)$ only depends on ${}^{[\underline{n}(i, x)]}\pi$; hence the optimization for a given level l can be restricted to ${}^{[\underline{n}_M(l)]}\Pi$, where $n = \underline{n}_M(l) = \sup\{\underline{n}(j_0, l)\colon j_0 \in S_1\}$.

"$\Leftarrow$": Let $m = \underline{n}(j_0, l)$, $j_0 \in S_1$. For any $h_{m-1} = (j_0, a_0, \ldots, j_{m-1}, a_{m-1}) \in H_{m-1}$, since $l - \varphi(h_{m-1}) < l_{\min}$, $V_{j_m}(\pi^{h_{m-1}}, l - \varphi(h_{m-1})) = 1 = V_{j_m}^*(l - \varphi(h_{m-1}))$, $j_m \in S_1$, $\pi \in \Pi$. Hence, $\pi^{j_{n-1}} \in \Pi^*(\{l - \varphi(h_{n-1})\})$ for $n \geq m$. Combining this condition with (48) and (49), we have $V_{j_n}(\pi^{h_{n-1}}, l - \varphi(h_{n-1})) = V_{j_n}^*(l - \varphi(h_{n-1}))$, $h_{n-1} \in H_{n-1}^*(l)$, $n = m - 1, \ldots, 0$, $j_n \in S_1$, i.e., $\pi^{h_{n-1}} \in \Pi^*(\{l - \varphi(h_{n-1})\})$, $n = m - 1, \ldots, 0$. As for $n = 0$, $\pi^{h_{-1}} = \pi \in \Pi^*(\{l\})$. Thus ${}^{[\underline{n}_M(l)]}\pi \in \Pi^*(\{l\})$, $\Pi^*(\{l\}) \neq \varnothing$.

"$\Rightarrow$": Induction. First we show that (49) and (50) hold for $n = 0$. By Theorem 3, $\exists f \in F$, $\underline{\pi}^{h_0} \in \Pi$, s.t. $f\underline{\pi}^{h_0} \in \Pi^*(\{l\})$. For any $j_0 \in S_1$, because

$$V_{j_0}(f\underline{\pi}^{h_0}, l) = \sum_{j_1 \in S} q(j_1 \mid j_0, f) V_{j_1}(\underline{\pi}^{h_0}, l - \varphi(h_0))$$

$$= V_{j_0}^*(l) = \sup\left\{ \sum_{j_1 \in S} q(j_1 \mid j_0, f) V_{j_1}^*(l - \varphi(h_0)) : f \in F \right\}, \tag{51}$$

$$V_{j_1}(\underline{\pi}^{h_0}, l - \varphi(h_0)) = V_{j_1}^*(l - \varphi(h_0)) \qquad \text{for all } j_1 \in S. \tag{52}$$

Otherwise, $V_{j_0}(f\underline{\pi}^{h_0}, l) < \sum_{j_1 \in S} q(j_1 \mid j_0, f) V_{j_1}^*(l - \varphi(h_0)) \le V_{j_0}^*(l)$, a contradiction. From (52) and (51), $\underline{\pi}^{h_0} \in \Pi^*(\{l - \varphi(h_0)\})$, $f(j_0) \in A_{j_0}^*(l)$, $h_0 = (j_0, f(j_0)) \in H_0^*(l)$, $j_0 \in S_1$. Next, suppose (49) and (50) hold for $0 \le k \le n$, i.e., $\exists \pi = (\theta_0, \theta_1, \ldots, \theta_n, \pi_{n+1}, \ldots)$, $\theta_k(j_k \mid h_{k-1}) \in A_{j_k}^*(l - \varphi(h_{k-1}))$, $j_k \in S_1$, $\pi^{h_k} \in \Pi^*(\{l - \varphi(h_k)\})$, $h_k \in H_k^*(l)$. For each given $h_n \in H_n^*(l)$, apply Theorem 3 to $\pi^{h_n} \in \Pi^*(\{l - \varphi(h_n)\})$. Then, similar to (51) and (52), $\exists \theta_{n+1}(\cdot \mid h_n) \in F$, $\underline{\pi}^{h_{n+1}} \in \Pi$, s.t. $\theta_{n+1}\underline{\pi}^{h_{n+1}} \in \Pi^*(\{l - \varphi(h_n)\})$. Hence for any $j_{n+1} \in S_1$, we derive $\theta_{n+1}(j_{n+1} \mid h_n) \in A_{j_{n+1}}^*(l - \varphi(h_n))$ and $\underline{\pi}^{h_{n+1}} \in \Pi^*[\{l - \varphi(h_{n+1})\}]$, $h_{n+1} = (h_n, j_{n+1}, \theta_{n+1}(j_{n+1} \mid h_n)) \in H_{n+1}^*(l)$. Let $\pi^{h_n} = \theta_{n+1}\underline{\pi}^{h_{n+1}}$. Then (49) and (50) also hold for $k = n + 1$, completing the induction. The existence of such optimal policy implies $A_{j_n}^*(l - \varphi(h_{n-1})) \ne \varnothing$, $h_{n-1} \in H_{n-1}^*(l)$, $0 \le n \le \underline{n}(j_0, l) - 1$, $j_n \in S_1$. ∎

COROLLARY 11.1. *Given* l, $\Pi^*(\{l\}) \ne \varnothing \Rightarrow \Pi^*(\{l - \varphi(h_{n-1})\}) \ne \varnothing$, $h_{n-1} \in H_{n-1}^*(l)$, $n \ge 0$.

Therefore, if there exist the single point stochastic order optimal policies for some level l, there must exist a finite-stage nonstationary optimal policy that is made up of history-dependent decision functions. Moreover, since each of its n-remainder policy π^{h_n} is optimal for level $l - \varphi(h_n)$ along the optimal history h_n, there exist optimal policies for many single levels below l. Consequently, the existence of the optimal policy for $L = \{l\}$ is not only determined by the optimal action set on L, which is the case for the complete and local stochastic order optimization models, but is also dependent on some optimal action sets outside of L.

6. OPTIMIZATION FOR FINITE STATE SPACE AND ACTION SPACE

The various properties and the existence conditions of optimal policies presented above lead to the following questions. Are there any sufficient

optimality conditions? Are these conditions computationally verifiable? These questions are explored by examining the model in finite state space and action space. For finite state space and action space, $W(\pi)$ is a discrete random variable for any given policy π. This is because each stage has the same number of finitely many possible rewards, either $r(i, a)$, $i \in S_1$, $a \in A(i)$, or $e(j)$, $j \in S_0$, with $r_{\min} > 0$ and $e_{\min} \geq 0$. Therefore, when the initial state is $i \in S_1$, there are only finitely many possible values for the first arrival target total return $W(\pi)$ in any return level interval $[l_m(i), l]$. These possible values can be denoted by an ordered list w_i:

$$w_i = \{w_i[k]\colon w_i[k] < w_i[k+1], k = 0, 1, 2, \dots\}, \tag{53}$$

$$\sup\{k\colon w_i[k] < x, k = 0, 1, \dots\} < +\infty \qquad \text{for any } x < +\infty. \tag{54}$$

It is obvious that $w_i = \{e(i), +\infty\}$ for any $i \in S_0$. The objective function $V_i(\pi, x)$, which is the remaining distribution function of $W(\pi)$ for initial state i, is thus determined by a countable set of values $\{V_i(\pi, w_i(k))\colon k = 0, 1, 2, \dots\}$, where

$$V_i(\pi, x) = 1, \quad x \in (-\infty, w_i[0]), \quad w_i[0] = l_m(i), \qquad i \in S, \quad \pi \in \Pi; \tag{55}$$

$$V_i(\pi, x) = V_i(\pi, w_i[k]), \quad x \in [w_i[k], \quad w_i[k+1]);$$
$$k = 0, 1, 2, \dots, \quad i \in S, \quad \pi \in \Pi. \tag{56}$$

From (56) and (15): $V_i(\pi, x) = \sum_{a \in A(i)} \pi_0(a \mid i) \sum_{j \in S} q(j \mid i, a) V_j(\pi^{(i,a)}, x - r(i, a))$, there must be

$$V_j(\pi^{(i,a)}, x - r(i, a)) = V_j(\pi^{(i,a)}, w_i[k] - r(i, a)),$$
$$\forall a \in A(i), \quad j \in S, \quad x \in [w_i[k], w_i[k+1]). \tag{57}$$

Otherwise, $\exists a \in A(i)$, $j \in S$, $x \in [w_i[k], w_i[k+1])$, $\pi \in \Pi$, s.t. $V_j(\pi^{(i,a)}, x - r(i, a)) \neq V_j(\pi^{(i,a)}, w_i[k] - r(i, a))$, which results in $V_i(\pi, x) \neq V_i(\pi, w_i[k])$, contradicting (56). The constraints in (57) lead to a way to compute w_i. Suppose $\{w_j[k]\colon k = k_j, k_j - 1, \dots, 0, j \in S\}$ are all known currently. The goal is to get $w_i[k_i + 1]$, $i \in S_1$ from these known values. Let

$$\underline{k}_j(i, a) = \min\{k\colon w_j[k] > w_i[k_i] - r(i, a), \quad k = k_j, k_j - 1, \dots, 0\},$$
$$i \in S_1, \quad a \in A(i), \quad j \in S. \tag{58}$$

$\underline{k}_j(i, a) = k_j + 1$ if the above set is empty. Let

$$\delta_j(i, a) = \big(w_j[\underline{k}_j(i, a)] - w_i[k_i] + r(i, a)\big) I\big(\underline{k}_j(i, a) \leq k_j\big). \tag{59}$$

Now, for $x \in [w_i[k_i] - r(i,a), w_i[k_i] - r(i,a) + \delta_j(i,a))$, $V_j(\pi^{(i,a)}, x - r(i,a)) = V_j(\pi^{(i,a)}, w_i[k_i] - r(i,a))$. Let

$$\delta(i) = \min\{\delta_j(i,a): j \in S, a \in A(i)\}, \qquad i \in S_1. \tag{60}$$

Then, by Eq (15), $V_i(\pi, x) = V_j(\pi, w_i[k_i])$ for $x \in [w_i[k_i], w_i[k_i] + \delta(i))$. Therefore,

$$w_i[k_i + 1] := w_i[k_i] + \delta(i). \tag{61}$$

Let $k_i := k_i + 1$ to complete a loop from (58) to (61). This loop can iterate until $\delta(i) = 0$. Since k_i is updated, there must be some k_j, $j \neq i$, that can be updated in the same way. Otherwise, $\delta(i) = 0$ for all $i \in S_1$. Then, $\forall i \in S_1$, $\exists a(i) \in A(i)$, $j(i) \in S_1$, s.t. $\delta_{j(i)}(i, a(i)) = 0$, or equivalently, $w_i[k_i] - r(i, a(i)) \geq w_{j(i)}[k_{j(i)}]$. Notice that $j(i) \neq i$ for all $i \in S_1$. As S_1 is finite, $\{n: j^n(i) = i, n \geq 1, i \in S_1\} \neq \emptyset$. Hence $w_i[k_i] \geq w_{j(i)}[k_{j(i)}] + r(i, a(i)) \geq \cdots \geq w_i[k_i] + r(i, a(i))$, contradicting the assumption of a strictly positive running reward. Together with (54), it follows that any return level can be reached in finite steps by using this asynchronous update scheme for computing $\{w_i, i \in S_1\}$.

Furthermore, the set of recursive equations implies a backward propagation from values of some n-remainder policy to those for the original policy. However, to even compute the objective function value at a single point, x, is not trivial. Roughly speaking, if there are N_S states, each with N_A actions, and $n = \underline{n}(i, x)$ stages are involved in computing $V_i(\pi, x)$, then $V_j(\pi^{h_n}, l - \varphi(h_n))$, for all $h_n \in H_n$, are obtained from the initial conditions (55). $O[(N_S N_A)^{\underline{n}(i,x)}]$ multiplication and additions are needed to propagate these $(N_S N_A)^{\underline{n}(i,x)}$ initial values to finally get $V_i(\pi, x)$. An example of $N_S = N_A = \underline{n}(i, l) = 2$ is shown in Fig. 1. Although $V_j(\pi^{h_n}, x - \varphi(h_n))$ for all $h_n \in H_n$ must be computed, since $\pi \neq \pi^{h_n}$ in general, all of these computations only give the value of $V_i(\pi, x)$ at one point x. Hence, the computation is impractical for large state spaces and action spaces, unless π is stationary.

The optimality equations, as well as Eq. (15) for stationary policies, are *self*-recursive (thus the objective functions for a given stationary policy may be evaluated in an algorithm similar to the optimization algorithm in Fig. 2). Therefore, the optimal function values may be recursively computed from their initial conditions. The way in which the value functions are computed is quite different from that in conventional optimization models, for example, value iteration or policy iteration. Here the goal is to compute function $V_i^*(x)$, which is specified by a countable list of values, not a single value V_i^*; the optimality equations are recursive simultaneous equations with initial conditions, not purely simultaneous equations. On the other hand, each of the simultaneous optimality equations is a highly

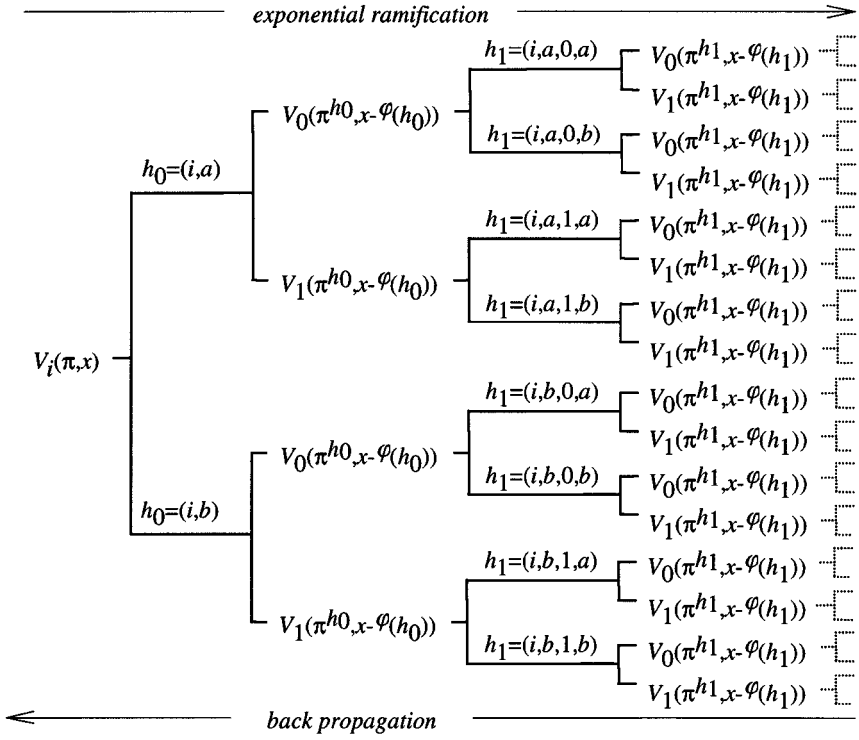

FIG. 1. Back propagation and exponential ramification in the computation of the objective functions. In this example, $S = \{0, 1\}$, $A(0) = A(1) = \{a, b\}$, $\underline{n}(i, x) = 2$. To compute $V_i(\pi, x)$, an exponentially increasing number of the objective functions for π's n-remainder policies at $x - \varphi(h_{n-1})$ are needed. Until some n, which is 2 in this case, $V_j(\pi^{h_{n-1}}, x - \varphi(h_{n-1}))$, $j \in S_1$, are first known as initial values; then they are propagated back through the recursive equations to get $V_i(\pi, x)$. If π is not stationary, all of these computations only give $V_i(\pi, x)$ at a single point x, although 20 $V_j(\pi^{h_{n-1}}, x - \varphi(h_{n-1}))$, $j \in S_1$, must be computed.

nonlinear equation involving all of the value functions for different states and levels. Because the reward r appears inside the objective function on the right-hand side of the recursive equations, it leads to the asynchronous update scheme for computing nonlinear calculation steps. Using these steps, the optimality equations are turned into a set of discrete recursive equations.

For any $i \in S$, since $V_i(\pi, x)$ is given by $\{V_i(\pi, w_i[k])$: $k = 0, 1, 2, \ldots\}$, $\forall \pi \in \Pi$, $V_i^*(x)$ is also determined by its values at w_i. Let $V_i^* = \{V_i^*[k], k = 0, 1, 2, \ldots\}$, where for $i \in S_0$, $V_i^* = \{1, 0\}$; for $i \in S_1$, $V_i^*[0] = 1$, $V_i^*[k] = V_i^*(w_i[k-1])$, $k > 0$. Then $V_i^*(x) = V_i^*[0]$, $x \in (-\infty, w_i(0))$; $V_i^*(x) = V_i^*[k]$, $x \in [w_i[k-1], w_i[k])$, $k > 0$. From the way w_i is constructed in

(58)–(61), the optimality equation becomes

$$V_i^*[k_i + 1] = \max\left\{ \sum_{j \in S} q(j \mid i, a) V_j^*\left[\underline{k}_j(i, a)\right], a \in A(i) \right\}, \qquad i \in S_1. \tag{62}$$

Let $A_i^*[0] = A(i)$, $A_i^*[k] = A_i^*(w_i[k-1])$, $i \in S_1$. According to the definition of $\underline{k}_j(i, a)$ in (58), there must be $\arg V_i^*(x) = \arg V_i^*[k_i]$, $x \in [w_i[k_i - 1], w_i[k_i])$, $k > 0$. Thus, $A_i^*(x) = A_i^*[0]$, $x \in (-\infty, w_i[0])$; $A_i^*(x) = A_i^*[k]$, $x \in [w_i[k-1], w_i[k])$, $k > 0$.

In Fig. 2, we are computing a *function*, not a *number*, for each initial state. Because of self-recursion, the computational complexity of computing $V_i^*(x)$ is greatly reduced to $O[N_S^2 N_A \underline{n}(i, x)]$. To compute $V_i^*(x)$, all of the $V_i^*(y)$ for $y < x$ are obtained, together with some related values for $V_j^*(x)$, $j \in S_1$. Once the optimal action sets are obtained, the optimal policies can be derived.

THEOREM 12. *For finite state space and action space,*

I. $\Pi^*([0, +\infty)) \neq \varnothing \Leftrightarrow \bigcap_{k \geq 0} A_i^*[k] \neq \varnothing$, $i \in S_1$. *Furthermore,* $f^\infty \in \Pi^*([0, +\infty))$, $f(i) \in \bigcap_{k \geq 0} A_i^*[k]$.

II. $\Pi^*([0, l]) \neq \varnothing \Leftrightarrow \bigcap_{0 \leq k \leq \underline{n}(i,l)-1} A_i^*[k] \neq \varnothing$, $i \in S_1$. *Furthermore,* $f^\infty \in \Pi^*([0, l])$, $f(i) \in \bigcap_{0 \leq k \leq \underline{n}(i,l)-1} A_i^*[k]$.

III. $\Pi^*(\{l\}) \neq \varnothing$. *Furthermore,* $\pi = (\theta_0, \theta_1, \ldots, \theta_{\underline{n}_M(l)-1}, \ldots) \in \Pi^*(\{l\})$, $\theta_n(j_n \mid h_{n-1}) \in A_{j_n}^*(l - \varphi(h_{n-1}))$, $h_{n-1} \in H_{n-1}^*(l)$, $j_n \in S_1$, $0 \leq n \leq \underline{n}(j_0, l) - 1$.

Proof. I and II are restatements of Theorem 10 for finite state space and action space. For finite state space and action space, $A_i^*(x) = \arg V_i^*(x) \neq \varnothing$, $x \in R$, $i \in S_1$. Thus by Theorem 11, $\Pi^*(\{l\}) \neq \varnothing$. The rest of III comes from Theorem 11. ▌

Hence, for finite state space and action space, the complete/local stochastic order optimality condition given in Theorem 10 is a countable/finite intersection of optimal action sets. If the intersection of these optimal action sets is empty, the optimal policies do not exist; otherwise a deterministic stationary optimal policy f^∞ with $f(i) \in \bigcap_{x \in L} A_i^*(x)$, $i \in S_1$, can be obtained. The single point stochastic order optimal policies must exist and can be derived from the optimal action sets on some levels related to the optimal histories.

(* Initialization *)

For all $i \in S_0$, do:

$w_i=\{e(i),+\infty\}$, $V_i=\{1,0\}$, $k_i=1$

End For

$e_{\min}=\min\{e(i): i \in S_0\}$

For all $i \in S_1$, do:

$l_m(i)=\min\{r(i,a): a \in A(i)\}+e_{\min}$

$w_i[0]=l_m(i)$, $V_i[0]=1$, $A_i^*[0]=A(i)$, $k_i=0$

End For

(* Iterations *)

$S_1=S-S_0=\{i_1, i_2, \cdots, i_U\}$

For $u=1$ to U, let $i=i_u$, do:

Repeat

For all $a \in A(i)$, $j \in S$, do:

$k=k_j$; While $k \geq 0$ and $w_j[k]>w_i[k_i]-r(i,a)$, $k:=k-1$

$\underline{k}_j(i,a)=k+1$

$\delta_j(i,a)=\{w_j[\underline{k}_j(i,a)]-w_i[k_i]+r(i,a)\}I(\underline{k}_j(i,a) \leq k_j)$

End For

$\delta(i)=\min\{\delta_j(i,a): j \in S, a \in A(i)\}$

If $\delta(i)>0$, then

$V_i^*[k_i+1]=\max\{\Sigma_{j \in S}\, q(j|i,a)V_j^*[\underline{k}_j(i,a)], a \in A(i)\}$

$A_i^*[k_i+1]=\arg V_i^*[k_i+1]$

$w_i[k_i+1]:=w_i[k_i]+\delta(i)$

$k_i:=k_i+1$

End If

Until $\delta(i)=0$

End For

FIG. 2. Algorithm for the value functions and the optimal action sets, given S, A, q, S_0, r, e.

7. OPTIMALITY CONSTRAINTS ON SYSTEM PARAMETERS AND NUMERICAL EXAMPLES

All of the optimality conditions in the preceding section are given in the optimal action sets, which are obtained from the computation of the value functions. Can the optimality conditions be given in terms of system parameters, namely, transition probabilities and reward functions? Not only is this question theoretically important, it might also lead to computationally verifiable optimality conditions for the complete stochastic order optimal policies. However, the problem is not trivial, since the optimization models are highly nonlinear with respect to the transition probabilities and reward functions. In this section, Example 1 is first used to illustrate the computation in the optimization algorithm and some intuitive ideas of the optimal action selection. The optimality constraint on system parameters is generalized to a special class of systems. Example 2 is then given to show that this constraint does not hold and that it becomes harder to figure out the optimal policies with the increasing complexity of systems. Finally, to see how likely it is for a complete stochastic order optimal policy to exist, computational experiments using random system parameters are carried out and the number of trials, in which sufficiently large local stochastic order optimal policies exist, is reported. These results may shed some light on further research on these optimization models.

EXAMPLE 1. $S = \{0, 1\}$; $S_0 = \{0\}$, $e(0) = 0$; $S_1 = \{1\}$, $A(1) = \{a, b\}$. r and q are shown in Table I. The transition probabilities from the states in S_0 are omitted, as they are all absorbing.

Since there is only one state in S_1 and two actions for this state, there are only two decision functions in the set F, denoted by f and g, where $f(1) = a$ and $g(1) = b$. The initial values are

$$w_0 = \{0, +\infty\}, \qquad V_0^* = \{1, 0\}, \qquad k_0 = 1, \qquad e_{\min} = 0$$

TABLE I
The One Stage Running Reward Function r and the Transition Probability q for Example 1

	r	$i = 1$	$q(j \mid i, action)$	i = 1	
	a	1	$j = 0$	0.2	0.1
$A(i)$	b	2	$j = 1$	0.8	0.9
			action	a	b

and

$$w_1[0] = l_m(1) = \min\{r(1,a), r(1,b)\} + e_{\min} = 1,$$
$$V_1^*[0] = 1, A_1^*[0] = \{a,b\}, \qquad k_1 = 0.$$

For this example, (62) can be written as the inner product of vectors:

$$\begin{aligned} V_1^* &= \max\{(q(0 \mid 1,a), q(1 \mid 1,a)) \\ &\qquad \cdot (V_0^*[\underline{k}_0(1,a)], V_1^*[\underline{k}_1(1,a)]), a \in A(1)\} \\ &= \max\{(0.2, 0.8) \cdot (V_0^*[\underline{k}_0(1,a)], V_1^*[\underline{k}_1(1,a)]), \\ &\qquad (0.1, 0.9) \cdot (V_0^*[\underline{k}_0(1,b)], V_1^*[\underline{k}_1(1,b)])\}. \end{aligned}$$

Step 1: $k_1 = 0$, $w_1 = \{1, \ldots\}$, $V_1^* = \{1, \ldots\}$,

$$w_1[k_1] - r(1,a) = 0, \quad \underline{k}_0(1,a) = 1, \quad \underline{k}_1(1,a) = 0;$$
$$w_1[k_1] - r(1,b) = -1, \quad \underline{k}_0(1,b) = 0, \quad \underline{k}_1(1,b) = 0;$$
$$\delta(1) = \min\{+\infty, 1, 1, 2\} = 1, \qquad w_1[1] = w_1[0] + \delta(1) = 2;$$
$$\begin{aligned} V_1^*[1] &= \max\{(0.2, 0.8) \cdot (0,1), (0.1, 0.9) \cdot (1,1)\} \\ &= \max\{0.8, 1\} = 1, \qquad A_1^*[1] = \{b\}. \end{aligned}$$

Step 2: $k_1 = 1$, $w_1 = \{1, 2, \ldots\}$, $V_1^* = \{1, 1, \ldots\}$,

$$w_1[k_1] - r(1,a) = 1, \quad \underline{k}_0(1,a) = 1, \underline{k}_1(1,a) = 1;$$
$$w_1[k_1] - r(1,b) = 0, \quad \underline{k}_0(1,b) = 1, \quad \underline{k}_1(1,b) = 0;$$
$$\delta(1) = \min\{+\infty, 1, +\infty, 1\} = 1, \qquad w_1[2] = w_1[1] + \delta(1) = 3;$$
$$\begin{aligned} V_1^*[2] &= \max\{(0.2, 0.8) \cdot (0,1), (0.1, 0.9) \cdot (0,1)\} \\ &= \max\{0.8, 0.9\} = 0.9, \qquad A_1^*[2] = \{b\}. \end{aligned}$$

Step 3: $k_1 = 2$, $w_1 = \{1, 2, 3, \ldots\}$, $V_1^* = \{1, 1, 0.9, \ldots\}$,

$$w_1[k_1] - r(1,a) = 2, \qquad \underline{k}_0(1,a) = 1, \qquad \underline{k}_1(1,a) = 2;$$
$$w_1[k_1] - r(1,b) = 1, \qquad \underline{k}_0(1,b) = 1, \qquad \underline{k}_1(1,b) = 1;$$
$$\delta(1) = \min\{+\infty, 1, +\infty, 1\} = 1, \qquad w_1[3] = w_1[2] + \delta(1) = 4;$$
$$\begin{aligned} V_1^*[3] &= \max\{(0.2, 0.8) \cdot (0, 0.9), (0.1, 0.9) \cdot (0,1)\} \\ &= \max\{0.72, 0.9\} = 0.9, \qquad A_1^*[3] = \{b\}. \end{aligned}$$

In fact, since $V_0^*(x - r(1, a)) = V_0^*(x - r(1, b)) = 0$ for $x \geq 2$, $V_1^*(x) = \max\{0.8V_1^*(x - 1), 0.9V_1^*(x - 2)\} = 0.9V_1^*(x - 2)$. Similarly, from $V_1(f, x) = 1$, $x < 1$ and $V_1(g, x) = 1$, $x < 2$, respectively, $V_1(f, x) = \sum_{j \in S} q(j \mid 1, f)V_1(f, x - r(1, f)) = 0.8V_1(f, x - 1)$ and $V_1(g, x) = \sum_{j \in S} q(j \mid 1, g)V_1(g, x - r(1, g)) = 0.9V_1(g, x - 2)$. Thus $V_1(g, x) = V_1^*(x)$. In other words, g^∞ is a complete stochastic order optimal policy. This result is consistent with intuition. From the reward functions and the transition law, action b is better than a because it achieves a better one stage running reward and is less likely to fall into the target set. Hence b is always the more reliable action for any given return level. This rule is generally true for a special class of systems.

THEOREM 13. *If* $S_0 = \{0\}$, $q(0 \mid i, a) > 0$, $i \in S_1$, $a \in A(i)$, *then* $\Pi^*([0, l)) \neq \varnothing$, $l \geq r_M(i) + e(0) + \min\{r_M(j)\colon j \in S_1\} \Rightarrow A_{rM}(i) \cap A_{qm}(i) \neq \varnothing$, $i \in S_1$, *where* $A_{rM}(i) = \{a : r(i, a) = r_M(i), a \in A(i)\}$ *and* $A_{qm}(i) = \{a\colon q(0 \mid i, a) = q_{0m}(i), a \in A(i)\}$, $q_{0m}(i) = \min\{q(0 \mid i, a)\colon a \in A(i)\}$.

Proof. Given $i \in S_1$, $V_i^*(x) = \max\{q(0 \mid i, a)I(e(0) > x - r(i, a)) + \sum_{j \in S_1} q(j \mid i, a)V_j^*(x - r(i, a))\colon a \in A(i)\}$. There must be $V_i^*(x) = 1$, $x < r_M(i) + e(0)$, because when $a \in A_{rM}(i)$, $x - r(i, a) < e(0)$, $V_j^*(x - r(i, a)) = 1$ for each $j \in S$. Thus, $A_{rM}(i) \subset A_i^*(x)$, $x < r_M(i) + e(0)$. On the other hand, since $q(0 \mid i, a) > 0$, $i \in S_1$, $a \in A(i)$, when $a \notin A_{rM}(i)$, there exists some $x \in [r_m(i) + e(0), r_M(i) + e(0)]$, such that $q(0 \mid i, a)I(e(0) > x - r(i, a)) + \sum_{j \in S_1} q(j \mid i, a)V_j^*(x - r(i, a)) < 1$, i.e., $a \notin A_i^*(x)$.

Next, there must be $V_i^*(x) = 1 - q_{0m}(i)$ for $x \in [r_M(i) + e(0), r_M(i) + e(0) + \min\{r_M(j)\colon j \in S_1\})$, because $V_j^*(x - r_M(i)) = 1$ for all $j \in S_1$ and $V_i^*(x) = \max\{\sum_{j \in S_1} q(j \mid i, a)V_j^*(x - r(i, a))\colon a \in A(i)\} = \max\{1 - q(0 \mid i, a)\colon a \in A(i)\}$. Hence, $A_{rM}(i) \cap A_{qm}(i) \neq \varnothing$ if $\Pi^*([0, l)) \neq \varnothing$, $l \geq r_M(i) + e(0) + \min\{r_M(j)\colon j \in S_1\}$. ∎

This simple rule, which selects the action with maximum one stage reward function and minimum exit probability, fails for more complicated systems, as will be shown in the next example. When the system becomes even larger, the interrelationship among all states will be far more complex, and it will become impossible to reach a conclusion qualitatively.

EXAMPLE 2. Let $S = \{0, 1, 2, 3, 4, 5\}$; $S_0 = \{0, 1, 2\}$; $S_1 = \{3, 4, 5\}$, $A(i) = \{a, b, c, d\}$, $i \in S_1$. e, r, and q are shown in Tables II, III, and IV, respectively. The value functions are given in Fig. 3.

TABLE II
Exit Reward Function for Example 2

$i \in S_0$	0	1	2
$e(i)$	0.0	4.5	3.0

TABLE III
One Stage Running Reward Function for Example 2

		$A(i)$			
	r	a	b	c	d
S_1	3	5.0	5.0	3.0	4.5
	4	2.5	6.0	5.0	4.0
	5	4.0	2.5	4.0	2.0

TABLE IV
Transition Probabilities for Example 2

		S						S					
	q	0	1	2	3	4	5	0	1	2	3	4	5
S_1	3	0.15	0.0	0.05	0.15	0.15	0.5	0.1	0.05	0.05	0.15	0.05	0.6
	4	0.0	0.1	0.1	0.2	0.2	0.4	0.05	0.1	0.05	0.3	0.3	0.2
	5	0.1	0.1	0.1	0.1	0.2	0.4	0.1	0.1	0.1	0.0	0.4	0.3
	$A(i)$				a						b		
S_1	3	0.05	0.1	0.05	0.4	0.3	0.1	0.0	0.15	0.05	0.2	0.5	0.1
	4	0.1	0.1	0.1	0.4	0.1	0.2	0.1	0.1	0.1	0.2	0.3	0.2
	5	0.1	0.1	0.1	0.5	0.1	0.1	0.1	0.1	0.1	0.2	0.4	0.1
	$A(i)$				c						d		

The optimization algorithm is implemented in *Mathematica*. On a Pentium 266 PC and Windows NT platform, it takes 42.251 s to do 150 iterations of the algorithm. After 150 iterations, $k_3 = 1118$, $k_4 = 1122$, $k_5 = 1124$; $w_3[k_3] = 562.5$, $w_4[k_4] = 564$, $w_5[k_5] = 564.5$; $V_3^*[k_3] = 2.07 \times 10^{-12}$, $V_4^*[k_4] = 2.01 \times 10^{-12}$, $V_5^*[k_5] = 1.57 \times 10^{-12}$. The value functions are step-like remaining distributions of the first arrival target total return. Part of computed $\{V_i^*, i = 3, 4, 5\}$ are shown in Fig. 3. The optimal action sets are given in Table V.

All optimal policies can be obtained from the optimal action sets. For example, there is one deterministic stationary optimal policy f^∞ for $L = [0, 562]$, where $f(3) = d$, $f(4) = b$, $f(5) = c$. Since the value function values are already very small at level 562, this optimal policy could be considered the complete stochastic order optimal in application. Besides this optimal policy, a nonstationary optimal policy $\pi = (\theta_0, \theta_1, \theta_2, \ldots)$ could also be constructed. Here is an example for $L = \{5\}$. First, as $\theta_0(j_0 \mid h_{-1}) \in A_{j_0}^*(5)$, let $\theta_0(3) = d$, $\theta_0(4) = a$, $\theta_0(5) = a$. Second, as $\theta_1(j_1 \mid h_0) \in A_{j_1}^*(5 - \varphi(h_0))$, where $h_0 \in \{(3, d), (4, a), (5, a)\}$ and $\varphi(h_0) \in \{4.5, 2.5, 4\}$, let $\theta_1(j_1 \mid h_0) =$

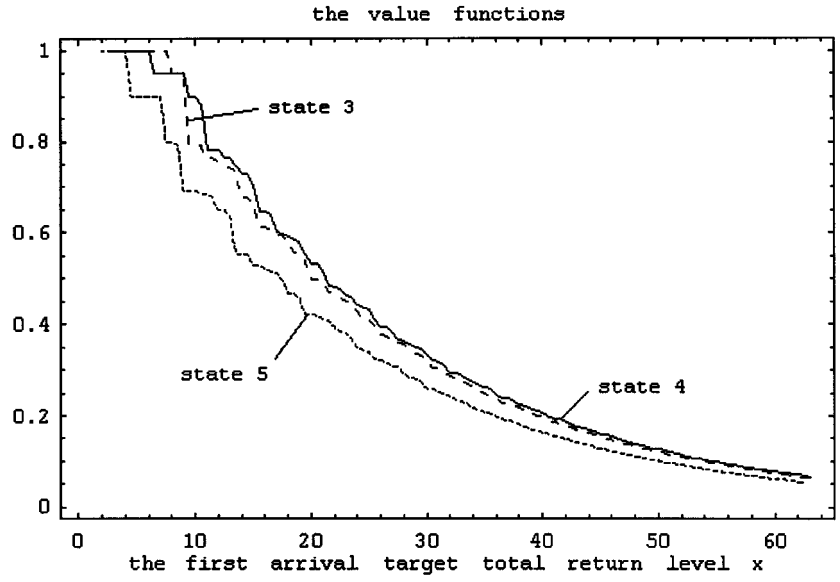

FIG. 3. The value functions for Example 2. In the figure, the dashed line is $V_3^*(x)$, the solid line is $V_4^*(x)$, and the dotted line is $V_5^*(x)$. Each curve is a step-like remaining distribution of the total return.

a, $j_1 \in S_1$. Finally, as $A_{j_2}^*(5 - \varphi(h_1)) = \{a, b, c, d\}$, we simply let $\theta_2(j_2 \mid h_1) = a$, $j_2 \in S_1$.

The optimal action selections for long-term consideration could be interpreted qualitatively. The probability of state 3 reaching the target is 0.2, no matter which action is selected. This value is smaller than that of state 5, which is 0.3 for all actions. Notice that the state-action pair $(4, b)$ has a much larger one stage running reward and an exit probability of 0.2.

TABLE V
The Optimal Action Sets for Example 2

X	$A_3^*(x)$	x	$A_4^*(x)$	x	$A_5^*(x)$
$(-\infty, 3)$	$\{a, b, c, d\}$	$(-\infty, 4)$	$\{a, b, c, d\}$	$(-\infty, 2)$	$\{a, b, c, d\}$
$[3, 5)$	$\{a, b, d\}$	$[4, 5)$	$\{a, b, c\}$	$[2, 2.5)$	$\{a, b, c\}$
$[5, 9)$	$\{d\}$	$[5, 5.5)$	$\{a, b\}$	$[2.5, 4)$	$\{a, c\}$
$[9, 9.5)$	$\{b, d\}$	$[5.5, 564)$	$\{b\}$	$[4, 5)$	$\{a, b, c, d\}$
$[9.5, 562.5)$	$\{d\}$			$[5, 5.5)$	$\{a, b, c\}$
				$[5.5, 8)$	$\{a, c\}$
				$[8, 564.5)$	$\{c\}$
$\bigcap_{0 \le x < 562.5} A_3^*(x) = \{d\}$		$\bigcap_{0 \le x < 564} A_4^*(x) = \{b\}$		$\bigcap_{0 \le x < 564.5} A_3^*(x) = \{c\}$	

State 5 is the worst state, since it has less running reward and a high probability of reaching the target. A good policy for this system is to try to avoid state 5 and stay in state 3 and 4. Thus, b is the best choice for state 4 and c is the best choice for state 5, both of which obey the rule in Theorem 13. For state 3, although d has a smaller one stage reward than a and b, d turns out to be the best choice for state 3 throughout all levels, since both a and b have a higher probability of causing the system to enter state 5 (the optimality constraint in Theorem 13 does not work here!). To reach target states with different probabilities is also a factor in comparing actions. The analysis here is to show that when the system scale increases, it becomes hard to find out directly from system parameters whether the optimal policy exists and what an optimal policy is. In addition, this example shows that the optimal action sets remain unchanged after some level. This phenomenon is more prominent in the following computational results.

To see how unusual the complete stochastic order optimal policies are, a group of computational experiments are carried out. In each trial, all of the system parameters are randomly generated from some uniform distributions. The optimization algorithm is used to recursively compute the value function and the intersection of the optimal action sets with increasing levels. The exit criterion of all of the trials is that either the optimal policies do not exist ($\cap A_i^*[k] = \varnothing$ for some state i), or the optimal policies exist ($\cap A_i^*[k_i] \neq \varnothing$ for all states) and all of the value functions have $V_i^*[k_i] < 10^{-6}$, which is a good approximation of the complete stochastic order optimization in computation. At the end of each trial, two quantities, v and l, are recorded together with an indication of whether the optimal policies exist or not. $v = \max\{V_i^*[k_i]\colon i \in S_1\}$ and $l = \min\{w_i[k_i]\colon i \in S_1\}$ are, respectively, the minimum value and the maximum return level reached for all states. The results are summarized in Table VI. For each system scale considered, the distribution of number of trials over v is given. The statistics over l are omitted, since they follow a reasonably monotonic pattern: larger l corresponds to smaller v.

In the table, $\|\cdot\|$ is the cardinality of a set. The total number of trials for each system scale setting is 20,000. The first block gives the number of system states and actions. The second block shows the number of trials in which $v = \max\{V_i^*[k_i]\colon i \in S_1\}$ falls into 10 uniform bins in its value range [0, 1] at the end of each trial. The last block shows the number of trials in which the local stochastic order optimal policies exist throughout the computation toward $v < 10^{-6}$ and the mean and the standard deviation of $l = \min\{w_i[k_i]\colon i \in S_1\}$ at the end of each trial. For the left four data columns, the reward function r and e are randomly selected from integers in [0, 10]; while for the two rightmost data columns, they are selected in

TABLE VI
Computational Results for the Existence of Complete Stochastic Order Optimal Policies

$\|\|S_0\|\|$	2	3	2	3	2	3
$\|\|S_1\|\|$	2	3	2	3	2	3
$\|\|A(i)\|\|$	4	4	6	6	4	4
$\|\|\{v \in (0.9, 1.0]\}\|\|$	14,183	18,687	15,882	18,308	15,356	18,292
$\|\|\{v \in (0.8, 0.9]\}\|\|$	1,809	682	1,749	969	1,557	803
$\|\|\{v \in (0.7, 0.8]\}\|\|$	1,383	320	1197	439	1138	422
$\|\|\{v \in (0.6, 0.7]\}\|\|$	934	131	510	148	719	191
$\|\|\{v \in (0.5, 0.6]\}\|\|$	361	40	109	28	300	62
$\|\|\{v \in (0.4, 0.5]\}\|\|$	107	12	25	8	55	25
$\|\|\{v \in (0.3, 0.4]\}\|\|$	37	3	11	2	19	3
$\|\|\{v \in (0.2, 0.3]\}\|\|$	12	2	2	2	13	2
$\|\|\{v \in (0.1, 0.2]\}\|\|$	2	0	0	0	3	1
$\|\|\{v \in (0, 0.1]\}\|\|$	1172	123	515	96	840	199
$\|\|\{\bigcap A_i^*[k_i] \neq \varnothing, v < 10^{-6}\}\|\|$	1170	123	515	96	839	199
mean[l] $\pm$ std[l]	332 ± 147	244 ± 43	411 ± 192	312 ± 80	$3{,}194 \pm 1{,}521$	$2{,}536 \pm 619$

the range [0, 100]. The transition probabilities of q are normalized random numbers in [0, 1].

A comparison of the data in the two leftmost columns with those in the two rightmost columns shows that there is no big difference if the range of rewards increases. This may be due to the fact that scaling the reward functions by a constant does not change the optimization structure. All of the data columns show that complete stochastic order optimal policies do exist for a small percentage of the systems; or more precisely, there exist sufficiently large local stochastic order optimal policies. The percentage decreases with an increasing number of states and actions, as more possible conflicts are introduced. With more states in the system, fewer local stochastic order optimal policies exist for larger intervals. The larger the return level range the optimal policies cover, the fewer the optimal policies. However, it appears that once the optimal policies exist for some finite level interval, they are also optimal for $[0, +\infty)$. In other words, for finite state space and action space, it appears that the countable intersection in the existence condition for complete stochastic order optimal policies might be given as a finite intersection of the first several optimal action sets, yet this reduction is unproved. Finally, it is worth pointing out that the single point stochastic order optimization model and ε-optimization models are most practical and interesting for applications. The ε-optimization models naturally incorporate robustness requirements, and there is a greater likelihood that a complete stochastic order optimal policy exists.

8. SUMMARY

This paper deals with countable state, countable action MDP endowed with a distribution function optimality criterion for the positive first arrival target total return. Based on the basic properties of the objective functions, convex combination, and cut-and-paste properties of the optimal policies, the optimality equations for the value functions and optimality conditions are obtained. If the complete or the local stochastic order optimal policies exist, there must be deterministic stationary optimal policies. If the single point stochastic order optimal policies exist, there must be deterministic nonstationary policies. These results are applied to systems with finite state space and action space. It is shown that the single point stochastic order optimal policies must exist. An algorithm is developed to compute the value functions and the optimal action sets, from which all optimal policies can be constructed. Numerical results are given, and they indicate possible directions of further research on the optimality constraints on system parameters.

ACKNOWLEDGMENTS

This work was supported in part by the Chinese National Natural Science Foundation and Tsinghua University Fundamental Research Foundation. We thank anonymous referees, Arthur Quaid, Lisa Saksida, and Stewart Moorehead for their very valuable comments and suggestions. The first author also thanks her current institutions, the Robotics Institute and the Center for the Neural Basis of Cognition, Carnegie Mellon University, for full support in finishing the revision of this paper.

REFERENCES

1. D. J. White, Mean, Variance and probabilistic criteria in finite Markov decision processes: A review, *J. Optim. Theory Appl.* **56** (1988), 1–29.
2. R. A. Howard and J. E. Matheson, Risk-sensitive Markov decision processes, *Management Sci.* **8** (1972), 356–369.
3. C. Derman and E. Ignall, On the stochastic ordering of Markov chains, *Oper. Res.* **23** (1975), 574–576.
4. S. M. Ross, ''Stochastic Processes,'' Wiley, New York, 1983.
5. D. Stoyan, ''Comparison Methods for Queues and Stochastic Models,'' Wiley, New York, 1983.
6. Kun-Jen Chung and Matthew J. Sobel, Discounted MDPs distribution functions and exponential utility maximization, *SIAM J. Control Optim.* **25** (1987), 49–62.
7. Yuanlie Lin, R. J. Tomkins and Chunglie Wang, Optimal models for the first arrival time distribution function in continuous time—a special case, *Acta Math. Appl. Sinica* **10** (1994), 194–212.

8. Wanqi Zhang, Qiyuan Jiang, and Yuanlie Lin, Reliability-constrained Markov decision programming and penalty factor method, *J. Tsinghua Univ.* **23** (1983), 61–71.
9. Yuanlie Lin and Qiyuan Jiang, Three problems of applying Markov decision programming to the optimal regulation of hydropower station reservoirs, *in* "China–Japan Symposium on Statistics," Beijing, China, 1984, pp. 143–147.
10. M. Bauakiz and Y. Kiber, Target-level criterion in Markov decision processes, *J. Optim. Appl.* **86** (1995), 1–15.
11. J. A. Filar, Percentiles and Markov decision processes, *Oper. Res. Lett.* **2** (1983), 13–15.
12. J. A. Filar, D. Krass, and K. W. Ross, Percentile performance criteria for limiting average Markov decision processes, *IEEE Trans. Automat. Control* **40** (1995), 2–9.
13. D. J. White, Minimizing a threshold probability in discounted Markov decision processes, *J. Math. Anal. Appl.* **173** (1993), 634–646.
14. J. Grandell, "Aspects of Risk Theory," Springer-Verlag, Berlin/New York, 1991.
15. Jianxing Lin, "A Study of Models for Markov Decision Programming in Discrete Time," Master's thesis, Department of Applied Mathematics, Tsinghua Univ., Beijing, 1987.
16. Yuanlie Lin, R. J. Tomkins, and Chunglie Wang, Optimal models for the first arrival time distribution function in continuous time, *Proc. APORS*'91 **14** (1991), 292–299.
17. Yuanlie Lin and Jianxing Lin, Models for the first arrival time distribution function optimization and risk minimization, *J. Tsinghua Univ.* **36** (1995), 53–59.
18. Harold J. Kushner and G. Dupuis, "Numerical Methods for Stochastic Control Problems in Continuous Time," Springer-Verlag, New York, 1992.
19. Yuanlie Lin, Continuous time first arrival target models. I. Discounted moment optimal models, *ACTA Math. Appl. Sinica* **14** (1991), 116–124.
20. Yuanlie Lin, Optimal models for the first arrival target in continuous time. II. L optimal problems, *J. Tsinghua Univ.* **33** (1993), 1–9.
21. Yuanlie Lin, Xingxing Yu, and Jianxing Lin, Models for first arrival target distribution function optimization and risk minimization in discrete time, *in* "Operations Research and Its Applications, First International Symposium, ISORA'95, Beijing, August 1995 Proceedings," pp. 368–375.
22. Xingxing Yu, "Research on Joint Delay-Filter Identification and Reinforcement Learning Control Models," Master's thesis, Department of Automation, Tsinghua Univ., Beijing, 1996.
23. David Blackwell, Discrete dynamic programming, *Ann. Statist.* **33** (1962), 719–726.