Stochastic Target Hitting Time and the Problem of Early Retirement

Kang Boda, Jerzy A. Filar, Yuanlie Lin, and Lieneke Spanjers

***Abstract*—We consider a problem of optimal control of a "retirement investment fund" over a finite time horizon with a target hitting time criteria. That is, we wish to decide, at each stage, what percentage of the current retirement fund to allocate into the limited number of investment options so that a decision maker can maximize the probability that his or her wealth exceeds a target x prior to his or her retirement. We use Markov decision processes with probability criteria to model this problem and give an example based on data from certain options available in an Australian retirement fund.**

***Index Terms*—Markov decision processes, probability criterion, retirement fund, target hitting time.**

I. Introduction

In THIS paper, we study a problem of optimal control of a "retirement investment fund" with, loosely speaking, the goal of ensuring that an adequate capital accumulates sufficiently quickly with sufficiently high probability. The objective is to develop a tool that could be used to advise nonprofessional investors who place their retirement benefits in a fund that permits only a limited number of options and offers only limited opportunity to reallocate the money among these options; say, once a year. We assume that such an investor is primarily interested in maximizing the probability of being to afford early retirement by certain age, and that the word "afford" means that the fund will equal or exceed a certain specified target amount at that terminal time. As such, we believe that the problem is a realistic one.

Since the mathematical framework in which we model this problem is that of *Markov decision processes* (MDPs) (e.g., see [11]) and since a vast majority of MDPs have objective criteria that depend on one of a number of "expected utility" criteria, it follows immediately that our problem is essentially different from these classical MDP models. Instead, the problem belongs to a class of models that are sometimes called "risk-sensitive MDPs." The latter can, perhaps, be traced back to [6] and constitutes an area where there has been a fair bit of research activity in recent years (e.g., see [2], [3], [8], [9], [12], and [15]–[18]). Some of these contributions tried to capture risk in terms of tradeoffs between mean and variance of suitable random variables, some have followed [20] in considering the expected value of a suitable exponential utility criterion and some have focussed on the so-called "percentile optimality" (e.g, [2], [4], [9], and [17]). The present paper is, perhaps, best classified as a continuation of this last line of research. Of course, Markowitz [5] pioneered the notion of mean-variance tradeoffs in finance literature and many more sophisticated, dynamic and stochastic, financial models involving closely related issues have been studied in recent years (e.g., see [1], [13], and [19]).

More precisely, we consider a finite-horizon discounted MDP model in which the decision-maker, at each stage, needs to decide what percentage of the current retirement fund to allocate into the limited (small) number of investment options. We assume that both the initial investment s_0 and the target retirement capital x are known and that the number of stages is n. Now, the first target hitting time $\tau(x)$ is a random variable whose distribution is specified by the choice of a policy. As mentioned above the decision-maker's goal is to find a policy π which maximizes $P_\pi(\tau(x) \leq n)$.

While at first sight, this might appear to be a very difficult problem it turns out a version of optimality principle can be shown to hold under mild conditions when we work in an "extended" state space (see Theorem 1). However, even in the extended state space the new process $\{e_t = (i_t, x_t), t \geq 1\}$ is not a Markov process under a general policy. Hence the existence and characterization of optimal policies cannot be obtained by standard techniques. Instead, the techniques used here are similar to those developed in [2] which dealt with a related problem of minimizing the probability that the total discounted wealth is less than a specified traget level. From the preceding optimality principle, structural results about optimal policies can be easily derived (Theorems 2 and 3) which, in turn, lead to a dynamic-programming type algorithm that is discussed in Section II-C and an enhanced dynamic-programming in Section II-D.

The above theoretical results are illustrated with an example based on real data from certain options available in an Australian retirement fund (Section III). Under a number of simplifying, but reasonable, assumptions the problem becomes computationally tractable. The results of these calculations are discussed in terms of their meaning for the decision-maker's original problem.

Manuscript received October 15, 2002; revised April 15, 2003. Recommended by Guest Editor B. Pasik-Duncan. This work was supported in part by the Australian Research Council under Grant A49532206 and by the National Natural Science Foundation of China under Grant 19871046 and Grant 79970120. This work was completed while L. Spanjers was an Exchange Student at the University of South Australia.

K. Boda and Y. Lin are with the Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China.

J. A. Filar is with the Center for the Industrial and Applicable Mathematics, School of Mathematics, University of South Australia, Mawson Lakes, SA 5095, Australia (e-mail: j.filar@unisa.edu.au).

L. Spanjers is with the University of Twente, Twente 7500, The Netherlands.

Digital Object Identifier 10.1109/TAC.2004.824469

II. Background and Preliminary Results

A. Model Description

We consider the following discrete-time and stationary *Markov decision process*:

$$\Gamma = (S, A, W, P, \beta) \tag{1}$$

where the *state space* S is countable, the *action space* $A(i)$ in each state i is finite and the *overall action space* $A = \cup_{i\in S} A(i)$ is countable. The *reward set* W is a bounded countable subset of $\mathcal{R} = (-\infty, +\infty)$. For each t from $1, \ldots$, let i_t, a_t and r_t denote the state of the system, the action taken by the decision maker, and the reward received at stage t, respectively. The stationary, single-stage, conditional *transition probabilities* are defined by

$$p^a_{ijr} := P(i_{t+1} = j, r_t = r \,|\, i_t = i, a_t = a) \quad i, j \in S, \quad a \in A(i) \quad r \in W \quad n \geq 1. \tag{2}$$

$$\sum_{j\in S, r\in W} p^a_{ijr} = 1, \qquad i \in S \quad a \in A(i). \tag{3}$$

We shall also assume that future rewards are discounted by the *discount factor* $\beta \in (0, 1]$.

In our formulation, when making a decision and taking an action at each stage, the decision maker considers not only the state of the original system but also his *target*. Effectively, this means that a new *hybrid state* $(i, x) \in S \times \mathcal{R}$ is introduced. Hence, we expand MDP Γ by enlarging the state space. We refer to (i, x) as the *hybrid state* of the decision maker to distinguish it from the system's state i, where x is the target value. Note that if the initial state of the decision maker is (i, x) and an action a is taken according to (2), the decision-maker's new hybrid state transits from (i, x) to $(j, (x - r)/\beta)$ with probability p^a_{ijr}.

Thus, if we denote E as the extended (hybrid) state–space, then the extended MDP $\tilde{\Gamma}$ has the following structure:

$$\tilde{\Gamma} = (E, A, W, P, \beta) \tag{4}$$

where the state–space $E = S \times \mathcal{R}$, the action space $A = \cup_{(i,x)\in E} A(i, x) = \cup_{i\in S} A(i)$. Note that $A(i, x) = A(i), (i, x) \in E$, the extended transition probabilities are simply $P : P(e_{t+1} = (j, (x - r)/\beta) \,|\, e_t = (i, x), a_t = a) = p^a_{ijr}, i, j \in S, a \in A(i), r \in W, x \in \mathcal{R}$. The reward set W and the discount factor β are the same as in MDP Γ.

Since in the model (4), the target is important when making decisions we must define policies which depend both on the system's state and the target, that is on the hybrid state.

Let the vector $h_t = (i_1, x_1, a_1, \ldots, i_{t-1}, x_{t-1}, a_{t-1}, i_t, x_t)$ denote the admissible history up to stage t, where $(i_k, x_k) \in E$, $a_k \in A(i_k, x_k), k = 1, \ldots, t - 1, (i_t, x_t) \in E$. If we denote the k-th hybrid state by $e_k = (i_k, x_k), k = 1, \ldots, t$, then $h_t = (e_1, a_1, \ldots, e_{t-1}, a_{t-1}, e_t)$.

Define the set $K := \{(e, a) \,|\, e \in E, a \in A(e)\}$, then we can denote the set of all (admissible) histories up to stage t by H_t. Now, we see $H_1 = E$, and $H_t = K^{t-1} \times E = K \times H_{t-1}, t > 1$. Observe that, for each t, H_t is a subspace of $\tilde{H}_t = (E \times A)^{t-1} \times E = (E \times A) \times \tilde{H}_{t-1}, t > 1$, and $\tilde{H}_1 := H_1 = E$.

Definition 1: A *decision rule* π_t at time t is a conditional transition probability measure on the control set A given H_t satisfying the constraint

$$\pi_t(A(e_t) \,|\, h_t) = 1 \qquad \forall h_t \in H_t, \qquad t = 1, 2, \ldots.$$

A decision rule π_t is called *deterministic*, if π_t is a measurable mapping from H_t to A, such that $\pi_t(h_t) \in A(e_t)$ for all $h_t \in H_t$.

Definition 2: A *policy* π is a sequence of decision rules. The set of all policies is denoted by Π. A policy $\pi = \{\pi_t, t = 1, 2, \ldots\} \in \Pi$ is said to be the following.

- *Markov policy*, if each π_t only depends on the current state at time t, that is, $\pi_t(\cdot \,|\, h_t) = \pi_t(\cdot \,|\, i_t, x_t) = \pi_t(\cdot \,|\, e_t), \forall h_t \in H_t, t \geq 1$.
- *Stationary policy*, if the policy π is a Markov policy, and the decision rules of π are all identical, that is,$\pi_t = \pi_1, \forall t > 1$ which is denoted by $\pi = \pi_1^\infty$.
- *Deterministic policy*, is any policy π such that all of its decision rules are deterministic.
- *TI-policy*, a policy which are independent of all targets $x_t (t \geq 1)$.

Let $\Pi_m, \Pi^d_m, \Pi_s, \Pi^d_s, \Pi_0$ denote the set of all Markov policies, all deterministic Markov policies, all stationary policies, all deterministic stationary policies, and all *TI-policies* respectively.

Now, given any $\pi = (\pi_t, t \geq 1) \in \Pi$, and a state-action triple (i, x, a), we construct the *"cut-head policy"* from π and (i, x, a) which is defined by $\bar{\pi}^{(i,x,a)} = (\pi_t^{(i,x,a)}, t \geq 1)$, where $\forall h_t \in H_t$, $t \geq 1, \pi_t^{(i,x,a)}(\cdot \,|\, h_t) = \pi_{t+1}(\cdot \,|\, i, x, a, h_t) = \pi_{t+1}(\cdot \,|\, i_1 = i, x_1 = x, a_1 = a, \ldots)$.

For any two policies $\pi = (\pi_t, t \geq 1), \sigma = (\sigma_t, t \geq 1) \in \Pi$, let $\pi(n) = (\pi_1, \pi_2, \ldots, \pi_n)$ denote the truncation of π to the first n stages and $(\pi(n), \sigma) = (\pi_1, \pi_2, \ldots, \pi_n, \sigma_1, \sigma_2, \ldots,)$ denote the policy in which π is implemented during the first n stages, and σ is implemented from the $(n+1)$st stage onwards. Hence, $(\pi(n), \sigma)$ is called an n stage *switching policy* from π to σ. Observe that, if $\pi = \delta^\infty \in \Pi_s$, then $\pi(n) = \delta^n$.

Note that a transition law P and a policy π determine the conditional probability measure P_π on the space of all possible histories of the process. Let W^π_n denote the random variable that is the sum of discounted rewards generated by policy π for the n-stage finite horizon problem. That is, $W^\pi_n = \sum_{t=1}^n \beta^{t-1} r_t$, for $n \geq 1$. To simplify the notation, we will use W_n instead of W^π_n when the choice of the policy is clear in the context.

Definition 3: Let $\tau(x)$ denote the first time at which the random total discounted reward exceeds the target value x. Note that $\tau(x)$ is also a random variable which we name the *target hitting time*, that is

$$\tau(x) = \inf\{k \,|\, W_k \geq x, k \geq 1\}.$$

Note that for any $\pi \in \Pi$, the functions

$$G^\pi_n(i, x) = P_\pi(\tau(x) > n \,|\, e_1 = (i, x)) = P_\pi\left(\cap^n_{k=1}(W_k < x) \,|\, e_1 = (i, x)\right) \quad (i, x) \in E \quad n \geq 1$$

are the objective functions that the decision maker wishes to minimize if he or she is interested in achieving the target as soon as possible. Consequently, $G^\pi_n(i, x)$ is called the *objective function* generated by π. It is clear that: $(\tau(x) \leq n) = \bigcup^n_{k=1}(W_k > x)$.

Definition 4: The following functions $G_n^*(i,x) = \inf_{\pi\in\Pi}\{G_n^\pi(i,x)\}, (i,x) \in E, n \geq 1$ are called the *optimal value functions*.

Definition 5: If the policy $\pi^* \in \Pi$ is such that $G_n^{\pi^*}(i,x) = G_n^*(i,x), \forall (i,x) \in E, n \geq 1$, then π^* is called an *n-stage optimal policy*. Equivalently

$$P_{\pi^*}(\tau(x) \leq n \,|\, e_1 = (i,x)) = \sup_{\pi\in\Pi} P_\pi(\tau(x) \leq n \,|\, e_1 = (i,x)).$$

Remark: It can be checked that with the above definitions, π^* is an n-stage optimal policy if and only if π^* is the policy that minimizes the probability that the cumulative discounted reward over the first n stages does not exceed x.

If we now define the lower and upper bounds on the rewards by $M_1 := \inf\{r \,|\, r \in W\}, M_2 := \sup\{r \,|\, r \in W\}$, then the objective functions G_n^π and G_n^* have the following properties.

1) If $M_2 \leq 0$, then obviously the discounted reward over n stages can be only lower or equal to the discounted reward after the first stage, hence trivially
$$G_n^\pi = G_1^\pi \quad G_n^* = G_1^*, \qquad n > 1.$$
2) If $M_2 > 0$, then $\cap_{k=1}^{n+1}(W_k < x) \subset \cap_{k=1}^{n}(W_k < x)$ and hence the probability of the target hitting time being greater than n is monotone nonincreasing, that is
$$G_{n+1}^\pi \leq G_n^\pi \quad G_{n+1}^* \leq G_n^*, \qquad n \geq 1.$$
Also, in this case
$$G_n^\pi(i,x) = G_n^*(i,x) = \begin{cases} 0, & \text{when } x \leq M_1 \\ 1, & \text{when } x > d_n \end{cases}$$
where $d_n = nM_2$ if $\beta = 1, d_n = M_2(1-\beta^n)/(1-\beta)$ if $\beta < 1$.

Now, let us introduce the space $\mathcal{D} := \{u \,|\, u : E \to [0,1], \text{measurable}\}$ of measurable functions on the extended state space E. For $\delta^\infty \in \Pi_s$, and $u \in \mathcal{D}$ define the operators K, T^δ and T by

$$Ku(i,x,a) := \sum_{j\in S, r\in W} p_{ijr}^a u(j,(x-r)/\beta) I_{(0,+\infty)}(x-r) \qquad (i,x) \in E \quad a \in A(i) \quad (5)$$

$$T^\delta u(i,x) := \sum_{a\in A(i)} \delta(a \,|\, i,x) Ku(i,x,a) \quad (i,x) \in E \quad (6)$$

$$Tu(i,x) := \min_{a\in A(i)} \{Ku(i,x,a)\} \quad (i,x) \in E$$

$$(T^\delta)^n u = T^\delta((T^\delta)^{n-1}u) \quad T^n u = T(T^{n-1}u)$$

where $I_{(0,+\infty)}(x)$ is the indicator function of the set $(0,+\infty)$. Obviously, when $f^\infty \in \Pi_s^d$, then $T^f u(i,x) = Ku(i,x,f(i,x))$.

In addition, we define

$$G_0^\pi(i,x) = G_0^*(i,x) = I_{(0,+\infty)}(x) \qquad \forall (i,x) \in E, \qquad \pi \in \Pi. \quad (7)$$

It can be easily checked that the operators K, T^δ and T defined above possess the usual monotonicity properties of dynamic programming (e.g., see [11]). These are stated, without proof, in the following Lemma.

Lemma 1: Let $u, v \in \mathcal{D}$. i) If $u \leq v$, then $Ku \leq Kv, T^\delta u \leq T^\delta v, Tu \leq Tv$; ii) If $u(i,x)$ is a nondecreasing and a left continuous function of x for any $i \in S$, then $Tu(i,x)$ is also a nondecreasing and a left continuous function of x for any $i \in S$; and iii) There exists $f \in F$ such that $T^f u = Tu$.

Proof: The proof is analogous to the classical results in [11, p. 163]. □

B. Finite Horizon Model

This subsection studies the finite horizon model. The objective is to prove the existence of a policy which minimizes the probability (risk) that the total discounted reward does not exceed the target value in the preceding finite number of stages.

Lemma 2: Let $\pi = (\pi_k, k \geq 1) \in \Pi$. Then, for each $(i,x) \in E, n \geq 1$

$$G_n^\pi(i,x) = T^{\pi_1} G_{n-1}^{\bar\pi}(i,x), \qquad n \geq 1. \quad (8)$$

and $G_n^\pi \in \mathcal{D}$, is determined by the truncated policy $\pi(n)$.

Proof: i) By the properties of P_π and the definition of G_n^π, we have

$$\begin{aligned} &G_1^\pi(i,x) \\ &= P_\pi(\tau(x) > 1 \,|\, (i,x)) \\ &= \sum_{a\in A(i)} \pi_1(a \,|\, i,x) \\ &\quad \times \sum_{j\in S, r\in W} p_{ijr}^a P_\pi(\tau(x) > 1 \,|\, (i,x), a, (j,(x-r)/\beta)) \\ &= \sum_{a\in A(i)} \pi_1(a \,|\, i,x) \sum_{j\in S, r\in W} p_{ijr}^a I_{(0,+\infty)}(x-r) \\ &= T^{\pi_1} G_0^{\bar\pi}(i,x). \end{aligned}$$

So, from (7), the lemma holds in the case $n = 1$. Now, for general $n > 1$, we can argue similarly that

$$\begin{aligned} G_n^\pi(i,x) &= P_\pi(\tau(x) > n \,|\, (i,x)) \\ &= \sum_{a\in A(i)} \pi_1(a \,|\, i,x) \sum_{j\in S, r\in W} p_{ijr}^a P_\pi \\ &\quad \times (\tau(x) > n \,|\, (i,x), a, (j,(x-r)/\beta)) \\ &= \sum_{a\in A(i)} \pi_1(a \,|\, i,x) \sum_{j\in S, r\in W} p_{ijr}^a P_{\bar\pi^{(i,x,a)}} \\ &\quad \times (\tau((x-r)/\beta) > n-1 \,|\, (j,(x-r)/\beta)) \\ &\quad \times I_{(0,+\infty)}(x-r) \\ &= \sum_{a\in A(i)} \pi_1(a \,|\, i,x) \sum_{j\in S, r\in W} p_{ijr}^a G_{n-1}^{\bar\pi^{(i,x,a)}} \\ &\quad \times (j,(x-r)/\beta) I_{(0,+\infty)}(x-r) = T^{\pi_1} G_{n-1}^{\bar\pi}(i,x). \end{aligned}$$

This completes the proof of (8) for all $n \geq 1$. Using (8) repeatedly we immediately obtain the last part of the lemma. □

In Theorem 1, we establish the "optimality principle" for the target hitting time criterion studied in this paper.

Theorem 1: (i) The optimal value function $\{G_n^*, n \geq 0\}$ satisfies the following optimality equations:

$$G_0^* = I_{[0,\infty)} \quad G_n^* = TG_{n-1}^*, \qquad n \geq 1$$

ii) For all $n \geq 0, i \in S, G_n^*(i,x)$ is a distribution function of some random variable X taking on values x;

iii) For all $n \geq 0$, there exists a policy $\pi \in \Pi_m^d$ such that $G_n^\pi = G_n^*$.

Proof: We prove this theorem by induction. When $n = 0$, by (7), the theorem holds. By inductive hypothesis, assume the

theorem also holds for $n = k$. Hence, $G_k^*(i,x)$ has the properties of a probability distribution function. Thus from part iii) of Lemma 1, $\exists \delta^\infty \in \Pi_s^d$ such that $T^\delta G_k^* = TG_k^*$. Similarly, by induction assumption, there exists a policy $\sigma \in \Pi_m^d$ such that $G_k^\sigma = G_k^*$. Let $\pi = (\delta, \sigma)$. Clearly, $\pi \in \Pi_m^d$. By Lemma 2, we have

$$\begin{aligned} G_{k+1}^*(i,x) &\le G_{k+1}^\pi(i,x) = T^\delta G_k^\sigma(i,x) \\ &= T^\delta G_k^*(i,x) = TG_k^*(i,x). \end{aligned}$$

Hence, $G_{k+1}^*(i,x) \le TG_k^*(i,x)$.

On the other hand, $\forall \eta \in \Pi$, again by Lemma 2, we have

$$G_{k+1}^\eta(i,x) = T^{\eta_1} G_k^{\bar\eta}(i,x) \ge T^{\eta_1} G_k^*(i,x) \ge TG_k^*(i,x).$$

Hence, the opposite inequality, $G_{k+1}^*(i,x) \ge TG_k^*(i,x)$ also holds.

From these inequalities, we now have the desired equality: $G_{k+1}^* = G_{k+1}^\pi = TG_k^*$. Thus by Lemma 1, $\forall i \in S, G_{k+1}^*(i,x)$ is a distribution function of some random variable X taking values x. This completes the Proof of Theorem 1. □

Remark: A Markov decision problem with a probability criterion such as the one we use might be regarded as quite difficult. However, a significant simplification can be achieved by extending the definition of the decision-maker's state space as was done here. With this extension under similar conditions to those normally used in the case of an expectation criterion we obtained results analogous to those known to hold in the classical model: there exists a deterministic Markov policy which minimizes the probability (risk) of the target hitting time exceeding a specified value n.

Corollary 1: There is no loss of generality in restricting consideration to deterministic Markov policies only, that is

$$\begin{aligned} G_n^*(i,x) &= \inf_{\pi \in \Pi} \{G_n^\pi(i,x)\} \\ &= \inf_{\pi \in \Pi_m^d} \{G_n^\pi(i,x)\}, \qquad (i,x) \in E \quad n \ge 1. \end{aligned}$$

Definition 6: We define *optimal action sets* by

$$\begin{aligned} A_n^*(i,x) &:= \{a \,|\, a \in A(i) \text{ and} \\ &\qquad KG_{n-1}^*(i,x,a) = G_n^*(i,x)\}, \quad (i,x) \in E \quad n \ge 1 \\ A_n^*(i) &:= \bigcap_{x \in \mathcal{R}} A_n^*(i,x), \qquad i \in S, \quad n \ge 1. \end{aligned} \tag{9}$$

Note that by the finiteness of $A(i)$ and Theorem 1, it follows that $A_n^*(i,x) \ne \emptyset, \forall e = (i,x) \in E, n \ge 1$.

Lemma 3: Let δ_k be a measurable mapping from E to A which satisfies $\delta_k(e) \in A_k^*(e), \forall e \in E, 1 \le k \le n$. Then, any policy $\pi \in \Pi$ which satisfies $\pi(n) = (\delta_n, \delta_{n-1}, \ldots, \delta_1)$ is n-stages optimal.

Proof: By induction. By the definition of $A_n^*(i,x)$, we note that: $T^{\delta_k} G_{k-1}^* = G_k^*, 1 \le k \le n$. When $n = 1$, by Lemma 2 and (7) we have that $G_1^\pi = T^{\pi_1} G_0^{\bar\pi} = T^{\delta_1} G_0^* = G_1^*$.

Assume that the lemma holds when $n = l$. Now, let $n = l + 1$, then because $\bar\pi^{(i,x,a)}(l) = (\delta_l, \ldots, \delta_1)$, and by inductive hypothesis $G_l^{\bar\pi^{(i,x,a)}} = G_l^*$, then by Lemma 2 we have: $G_{l+1}^\pi = T^{\pi_1} G_l^{\bar\pi} = T^{\delta_{l+1}} G_l^* = G_{l+1}^*$. So, the lemma holds when $n = l + 1$ and, hence, for all $n \ge 1$. □

With respect to the structure of n-stages optimal policies we have the following result.

Theorem 2: Let $\pi = (\pi_k, k \ge 1) \in \Pi$, for a given $(i,x) \in E$, then $G_n^\pi(i,x) = G_n^*(i,x)$ if and only if $\pi_1(A_n^*(i,x) \,|\, i,x) = 1$ and

$$\begin{aligned} &G_{n-1}^{\bar\pi^{(i,x,a)}}(j,(x-r)/\beta) I_{(0,+\infty)}(x-r) \\ &\qquad = G_{n-1}^*(j,(x-r)/\beta) I_{(0,+\infty)}(x-r) \end{aligned} \tag{10}$$

whenever $\pi_1(a \,|\, i,x) p_{ijr}^a > 0$.

Proof: Assume that $G_n^\pi(i,x) = G_n^*(i,x)$. By Theorem 1 iii) applied with $n - 1$ in place of n, there exists a policy $\sigma \in \Pi_m^d$ such that $G_{n-1}^\sigma = G_{n-1}^*$ and, hence, $T^{\pi_1} G_{n-1}^*(i,x) = T^{\pi_1} G_{n-1}^\sigma(i,x)$. Now, we have

$$\begin{aligned} G_n^*(i,x) = G_n^\pi(i,x) &= T^{\pi_1} G_{n-1}^{\bar\pi}(i,x) \ge T^{\pi_1} G_{n-1}^*(i,x) \\ &= T^{\pi_1} G_{n-1}^\sigma(i,x) = G_n^{(\pi_1,\sigma)}(i,x) \ge G_n^*(i,x) \end{aligned}$$

where the second and the forth equalities follow from Lemma 2 and the first inequality follows from Lemma 1. Thereby

$$\begin{aligned} G_n^*(i,x) &= T^{\pi_1} G_{n-1}^*(i,x) \quad \text{and} \\ T^{\pi_1} G_{n-1}^{\bar\pi}(i,x) &= T^{\pi_1} G_{n-1}^*(i,x). \end{aligned}$$

These equations can be converted with the help of the operators (5) and (6) to

$$\sum_{a \in A(i)} \pi_1(a \,|\, i,x) \{KG_{n-1}^*(i,x,a) - G_n^*(i,x)\} = 0 \tag{11}$$

$$\begin{aligned} &\sum_{a \in A(i)} \sum_{j \in S, r \in W} \pi_1(a \,|\, i,x) p_{ijr}^a \\ &\quad \times \left\{ G_{n-1}^{\bar\pi^{(i,x,a)}}(j,(x-r)/\beta) - G_{n-1}^*(j,(x-r)/\beta) \right\} \\ &\quad \times I_{(0,+\infty)}(x-r) = 0. \end{aligned} \tag{12}$$

Thus, by Theorem 1 and (11), we have $\pi_1(A_n^*(i,x) \,|\, i,x) = 1$. Similarly, from (12), we obtain (10) whenever $\pi_1(a \,|\, i,x) p_{ijr}^a > 0$.

The necessity part of the theorem is now proved. Note, however, that the preceding proof is reversible. Hence, the sufficiency part of the theorem also holds. □

Remark: i) Theorem 2 shows that a policy π is optimal for a finite horizon model if and only if the action taken by π at each realizable state is an optimal action and before the total discounted reward exceeds the target value the corresponding cut-head policy is also optimal at each stage.

ii) From Lemma 2 and Theorem 1, we can further see that π is n stages optimal if and only if the actions taken by π in the preceding n stages are optimal.

The next result gives a sufficient and a necessary condition for the existence of a finite horizon optimal TI-policy, namely, one that does not depend on the targets.

Theorem 3: i) If there exists a policy $\pi \in \Pi_0$ such that $G_n^\pi = G_n^*$, then $A_n^*(i) \ne \emptyset$ and $\pi_1(A_n^*(i) \,|\, i) = 1, \forall i \in S$; ii) If $A_k^*(i) \ne \emptyset, \forall i \in S, 1 \le k \le n$, then there exists a policy $\pi \in \Pi_0$ such that $G_n^\pi = G_n^*$.

Proof: i) Let $\pi \in \Pi_0$ and $G_n^\pi = G_n^*$. Then by Theorem 2, $\pi_1(A_n^*(i,x) \,|\, i) = 1$ for all $x \in \mathcal{R}$ and $i \in S$, it follows that $\pi_1(A_n^*(i) \,|\, i) = 1, \forall i \in S$. Hence $A_n^*(i) \ne \emptyset$ for each $i \in S$.

ii) Select $\delta_k : S \to A$ such that $\delta_k(i) \in A_k^*(i)$ for each $i \in S$ and $1 \le k \le n$. Then, by Lemma 2, for the policy $\pi \in \Pi_0$ which satisfies $\pi(n) = (\delta_n, \delta_{n-1}, \ldots, \delta_1), G_n^\pi = G_n^*$ holds. □

C. DP-Algorithm

Since we have now demonstrated that our target hitting time criterion possesses many of the properties of classical dynamic programming problems, it is not surprising that the backward recursion algorithm of dynamic programming can be adapted to apply to our problem.

Later, we present such an adaptation that computes optimal value functions, optimal action sets, and optimal policies in a finite horizon model with the target hitting time criterion.

Henceforth, we assume that S and W are both finite sets and that $W = \{r_1, r_2, \ldots, r_m\}$, with $r_k \le r_{k+1}, k = 1, \ldots m-1$. By Theorem 1,we have

$$\begin{aligned} G_0^*(i,x) &= I_{(0,+\infty)}(x) \\ G_n^*(i,x) &= \min_{a\in A(i)} \left\{ \sum_{j\in S, r\in W} p_{ijr}^a G_{n-1}^*(j,(x-r)/\beta) \right. \\ &\qquad \left. \times\, I_{(0,+\infty)}(x-r) \right\} \\ &\qquad i \in S, \qquad x \in \mathcal{R} \quad n \ge 1. \quad (13) \end{aligned}$$

Then, for notational convenience, define

$$\begin{aligned} b_n(i,x,a) &:= \sum_{j\in S, r\in W} p_{ijr}^a G_{n-1}^*(j,(x-r)/\beta) \\ &\qquad \times I_{(0,+\infty)}(x-r) \\ M_n(i,x) &:= \min_{a\in A(i)} \{b_n(i,x,a)\}. \end{aligned}$$

With the help of Theorem 1, Lemma 2, and Definition 4, we obtain the following algorithm.

Step 1) Calculate

$$\begin{aligned} b_1(i,r_k,a) &= \sum_{j\in S} \sum_{r\in W, r\le r_k} p_{ijr}^a, \qquad i \in S \quad a \in A(i) \\ M_1(i,r_k) &= \min_{a\in A(i)} \{b_1(i,r_k,a)\}, \qquad i \in S \\ A_1^*(i,r_k) &= \{a \,|\, a \in A(i) \\ b_1(i,r_k,a) &= M_1(i,r_k)\} \qquad i \in S \end{aligned}$$

and select an action $g_1(i,r_k) \in A_1^*(i,r_k), k = 1, \ldots, m-1$, and an arbitrary action $g_1(i,r_m) \in A(i)$. Then, by (13) and (9)

$$G_1^*(i,x) = \begin{cases} 0, & x \le r_1 \\ M_1(i,r_k), & r_k < x \le r_{k+1} \quad k = 1, \ldots, m-1 \\ 1, & x > r_m \end{cases}$$

$$A_1^*(i,x) = \begin{cases} A(i), & x \le r_1 \quad \text{or} \quad x > r_m \\ A_1^*(i,r_k), & r_k < x \le r_{k+1} \quad k = 1, \ldots, m-1. \end{cases}$$

Let

$$g_1(i,x) = \begin{cases} g_1(i,r_m), & x \le r_1 \quad \text{or} \quad x > r_m \\ g_1(i,r_k), & r_k < x \le r_{k+1}, \quad k = 1, \ldots, m-1. \end{cases}$$

Step 2) Assume that G_l^*, A_l^* and g_l have already been calculated and all the jump points of $G_l^*(i,x)(\forall i \in S) x_1 < x_2 < \cdots < x_\rho$ are known. Calculate the elements of the set $\{\beta x_k + r_h \,|\, k = 1, 2, \ldots, \rho, h = 1, 2, \ldots, m\}$ and denote them by $u_1 < u_2 < \cdots < u_L$ $(L \le m\rho)$, in an ascending order. Then, for any $j \in S$ and $r \in W$, we have

$$G_l^*(j,(x-r)/\beta) = \begin{cases} 0, & x \le u_1 \\ G_l^*(j,(u_k-r)/\beta), & u_k < x \le u_{k+1}, 1 \le k < L \\ 1, & x > u_L. \end{cases} \quad (14)$$

If $r_1 > u_L$, then $I_{[0,+\infty)}(u_k - r) = 0, k = 1, \ldots, L$ and, hence, from (13) $G_{l+1}^*(i,u_k) = 0, \forall k$. Or, there exists some N such that $u_{N-1} \le r_1 < u_N$ (note that if $r_1 < u_1$ we can simply define $u_0 = r_1$ and take $N = 1$).

Calculate

$$\begin{aligned} b_{l+1}(i,r_1,a) &= \sum_{j\in S} p_{ijr_1}^a G_l^*(j,0) \quad i \in S \quad a \in A(i) \\ b_{l+1}(i,u_k,a) &= \sum_{j\in S, r\in W, r\le u_k} p_{ijr}^a G_l^*(j,(u_k-r)/\beta) \\ &\qquad i \in S, \qquad a \in A(i) \quad k \ge N \\ M_{l+1}(i,r_1) &= \min_{a\in A(i)} \{b_{l+1}(i,r_1,a)\}, \qquad i \in S \\ M_{l+1}(i,u_k) &= \min_{a\in A(i)} \{b_{l+1}(i,u_k,a)\} \quad i \in S, \qquad k \ge N \\ A_{l+1}^*(i,r_1) &= \{a \,|\, a \in A(i) \\ b_{l+1}(i,r_1,a) &= M_{l+1}(i,r_1)\} \quad i \in S \\ A_{l+1}^*(i,u_k) &= \{a \,|\, a \in A(i) \\ b_{l+1}(i,u_k,a) &= M_{l+1}(i,u_k)\}, \qquad i \in S \quad k \ge N. \end{aligned}$$

Next, select actions $g_{l+1}(i,r_1) \in A_{l+1}^*(i,r_1)$, $g_{l+1}(i,u_k) \in A_{l+1}^*(i,u_k), k = N, \ldots, L-1$, and an arbitrary action $g_{l+1}(i,u_L) \in A(i)$. Then, by (13), (14), and (9)

$$G_{l+1}^*(i,x) = \begin{cases} 0, & x \le r_1 \\ M_{l+1}(i,r_1), & r_1 < x \le u_N \\ M_{l+1}(i,u_k), & u_k < x \le u_{k+1} \quad k = N, \ldots, L-1 \\ 1, & x > u_L \end{cases}$$

$$A_{l+1}^*(i,x) = \begin{cases} A_{l+1}^*(i,r_1), & r_1 < x \le u_N \\ A_{l+1}^*(i,u_k), & u_k < x \le u_{k+1} \quad k = N, \ldots, L-1 \\ A(i), & x \le r_1 \text{ or } x > u_L. \end{cases}$$

Let the decision rule at the next stage be defined by

$$g_{l+1}(i,x) = \begin{cases} g_{l+1}(i,r_1), & r_1 < x \le u_N \\ g_{l+1}(i,u_k), & u_k < x \le u_{k+1} \quad k = N, \ldots, L-1 \\ g_{l+1}(i,u_L), & x \le r_1 \quad \text{or} \quad x > u_L. \end{cases}$$

Step 3) Repeat Step 2 until $l + 1 = n$.

In this fashion, we construct the optimal function G_n^* and an optimal policy $\pi^* = (g_n, g_{n-1}, \ldots, g_1)^\infty$. In the process, the corresponding optimal action sets $A_1^*(i,x), A_2^*(i,x), \ldots, A_n^*(i,x)$ are constructed as well. By Theorem 2, these sets characterize all n stages optimal policies.

D. Enhanced DP-Algorithm

As we know, DP-algorithm can calculate the optimal value functions, optimal policies and optimal action sets accurately, however, it can quickly become computationally prohibitive. At each iteration more and more points (u_k) need to be considered. For a large state space, a large action space and a large reward set this will have drastic consequences. The number of points that need to be considered and thereby the time to do this will grow exponentially.

To overcome this problem a new algorithm is presented below. This algorithm approximates the solution found by the DP-algorithm by calculating a fixed number of points at each iteration. However, by taking this number large enough, the approximation will be quite good and the computational time will decrease significantly. We will assume that all rewards in the problem are positive.

The idea is that—irrespective of the iteration index l—a bounded monotone decreasing function such as $G_l^*(i,x)$ on an interval $[0, v_m]$ can be well approximated by an array of values $\{(v_1, G_l^*(i,v_1)), (v_2, G_l^*(i,v_2)), \ldots, (v_m, G_l^*(i,v_m))\}$ provided that $|v_{i+1} - v_i|$ is "sufficiently" small. The interpolation between the values $G_l^*(i,v_i)$ and $G_l^*(i,v_{i+1})$ at v_i and v_{i+1} can be carried out in a number of ways. In the implementation below the upper end is used. That is,

$$G_l^*(i,v) = G_l^*(i,v_{i+1}) \quad \forall\, v \in (v_i, v_{i+1}].$$

The following enhanced dynamic programming algorithm can now be used. For notational convenience, assume $\beta = 1$ and define

$$b_n(i,x,a) := \sum_{j\in S, r\in W} p_{ijr}^a G_{n-1}^*(j, x-r)$$
$$M_n(i,x) := \min_{a\in A(i)} \{b_n(i,x,a)\}.$$

Step 1) Initialize:
Choose m points $v_1 < v_2 < \cdots < v_k < \cdots < v_m$ that will represent the target values. The value of v_1 needs to be 0. The value of v_m is the largest target value that will be computed. The larger the m, the more accurate the approximation of the optimal value functions will be. Taking equi-spaced v_k's will have computational advantages.
Now by Theorem 1:

$$G_0^*(i,x) = \begin{cases} 0, & x \le v_1 \\ 1, & v_{k-1} < x \le v_k \quad k = 2, \ldots, m. \end{cases}$$

Step 2) Assume that G_l^* has already been calculated. Now calculate

$$b_{l+1}(i,v_k,a) = \sum_{j\in S, r\in W} p_{ijr}^a G_l^*(j, v_k - r) \qquad i \in S, \quad a \in A(i) \quad k \ge 1$$
$$M_{l+1}(i,v_k) = \min_{a\in A(i)} \{b_{l+1}(i,v_k,a)\} \qquad i \in S, \quad k \ge 1$$
$$A_{l+1}^*(i,v_k) = \{a \in A(i) \,|\, b_{l+1}(i,v_k,a) = M_{l+1}(i,v_k)\} \qquad i \in S \quad k \ge 1.$$

Next, select actions $g_{l+1}(i,v_k) \in A_{l+1}^*(i,v_k), k = 1, \ldots, m$. Then

$$G_{l+1}^*(i,x) = \begin{cases} 0, & x \le v_1 \\ M_{l+1}(i,v_k), & v_{k-1} < x \le v_k \quad k = 2, \ldots, m \end{cases}$$
$$A_{l+1}^*(i,x) = \begin{cases} A(i), & x \le v_1 \\ A_{l+1}^*(i,v_k), & v_{k-1} < x \le v_k \quad k = 2, \ldots, m. \end{cases}$$

Let the decision rule at the next stage be defined by

$$g_{l+1}(i,x) = \begin{cases} g_{l+1}(i,v_1), & x \le v_1 \\ g_{l+1}(i,v_k), & v_{k-1} < x \le v_k \quad k = 2, \ldots, m. \end{cases}$$

Step 3) Repeat Step 2) until $l + 1 = n$.
The approximate optimal function G_n^* and an optimal policy $\pi^* = (g_n, g_{n-1}, \ldots, g_1)^\infty$ have now been constructed. The corresponding approximate optimal action sets $A_1^*(i,x), A_2^*(i,x), \ldots, A_n^*(i,x)$ have been constructed as well.

III. Application

In this section, we will apply the above theory to the problem of allocating a fixed amount of funds in a number of investment options with the goal of attaining enough money for "early retirement." This is an important problem facing many people who are not professional investors. Most retirement funds in developed and even some developing countries offer its members the flexibility of choosing between a, typically small, number of investment options. Generally, the more "risky" options are associated with higher short term interest payments.

The real-life problem is complicated by the fact that the above "risks" and amounts of interest are not known precisely, or remain constant throughout the rather long planning horizons (e.g., twenty plus years) that many people are interested in. For the purpose of illustrating the theory and the algorithm derived above we shall not address these difficulties. Instead, we shall assume that the historical data on the performance of the various investment options that are available at the beginning of the planning horizon, accurately capture their future performance[1].

A. The Model

For the ease of intuitive understanding we present the results of the corresponding problem where the decision maker wishes to maximize, $P_\pi(\tau(x) \le n)$ namely, the probability that the total wealth exceeds the target x prior to his or her retirement which is assumed to occur at n years in the future. For instance, if $n = 15$, the target $x = 130\,000$ and π^* is such that $P_{\pi^*}(\tau(x) \le 15) = 0.8$, then the decision maker will believe that by implementing the policy π^* he or she will ensure, with probability 0.8, that the retirement fund will exceed 130 000 within 15 years. We can now exploit the theory and the algorithm presented earlier to solve this problem under the following set of simplifying, but reasonable, assumptions.

[1]There are a number of simple modifications to our problem that can be implemented to address these difficulties. We do not discuss these modifications in detail because that would necessitate the use of even more complex notation and further computational effort.

Assumptions:

- The decision maker invests a fixed amount only once at the beginning of the n-stage planning horizon. However, once a year, he or she can allocate the current amount in the fund among three different investment options.
- Following one Australian example the three given options are: Perpetual's Inv Choice Pension—Industrial Share, BT Lifetime Super Pers—Australian Share, Zurich FIP-Equity.
- Each year prior to n, the full amount in the fund is reinvested in these three options.
- The historical data available at the initial stage can be used to predict future behavior.
- Only options with positive rewards are considered.

Based on the previous assumptions, we can consider the model

$$\Gamma_r = (E, A, \{A(e), e = (s_i, x) \in E\}, W)$$

where, $E = S \times \mathcal{R}$ is the decision maker's state–space and $\mathcal{R} = (-\infty, +\infty)$. The underlying system's state–space S is composed of disjoint nonoverlapping intervals which represent the current wealth. More precisely, we define the system's state–space as

$$S = \{s_0, s_1, s_2, \ldots, s_\eta\} \quad s_k \geq 0 \qquad \forall \quad 0 \leq k \leq \eta$$

where, without loss of generality, we assume: $s_0 \leq s_1 \leq \cdots \leq s_\eta$. Since, the wealth is a continuous rather than a discrete variable, we interpret the statement "the wealth is $(s_{i-1} + s_i)/2$" to mean that the actual amount of wealth lies in the interval $(s_{i-1}, s_i]$ for $i = 1, \ldots, \eta$[2]. Depending on the real situation, it may be possible for $s_0 = -\infty$ and $s_n = +\infty$.

The action space is defined as

$$A = \left\{ a \,|\, a = (\theta_1, \theta_2, \theta_3), 0 \leq \theta_j \leq 1, \sum_{j=1}^{3} \theta_j = 1 \right\}$$

where θ_i is the fraction of the total wealth invested in the option i. We also assume that $A(e) = A, \forall e \in E$. For example $a = (0.25, 0.25, 0.5)$ is an action which means that the decision maker allocates 25 percent of the current wealth to the first option (Perpetual's Inv Choice Pension—Industrial Share), 25 percent to the second one and 50 percent to the last one.

Next, we will derive the exact form of the immediate rewards r and the reward sets $W(s_i, a)$. For ease of understanding, we shall divide this derivation into a number of separate steps. First, assume that the current wealth is denoted by $\bar{W}$ and the interest rates on 1 invested in each of the three options[3]: r_1, r_2, r_3 are all random variables.

If the decision maker takes an action $a = (\theta_1, \theta_2, \theta_3)$, then his or her expected wealth next year will be $\sum_{i=1}^{3} \theta_i \bar{W}(1 + r_i)$. Hence, the *portfolio interest rate* $r(a)$ will satisfy $\bar{W}(1 + r(a)) = \sum_{i=1}^{3} \theta_i \bar{W}(1 + r_i)$, and from $\sum_{j=1}^{3} \theta_j = 1$ we know that

$$r(a) = \sum_{i=1}^{3} \theta_i r_i.$$

Hence, we will name the $r(a)$ as the *immediate percentage rewards*, and note that it is also a random variable. We will present the second step after the definition of the total reward.

By Corollary 1, we know that we can find an optimal deterministic Markov policy. So, when we calculate the optimal value function, we only need to find an optimal policy in the set Π_m^d. That means, we are required to take deterministic actions that depend only on the current state $e = (s_i, x)$.

For a given policy $\pi = \{a_1, a_2, \ldots, a_n, \ldots\} \in \Pi_m^d$, we define the n-year *total rewards* from the initial state s_{i_1} as

$$W_n^\pi = s_{i_1} \Pi_{k=1}^{n} (1 + r(a_k)).$$

We will change this multiplicative formula to an additive one. If we start from s_{i_1} for some i_1, a random additive reward $r_1 = s_{i_1} r(a_1)$ will be received when the action a_1 is taken, following, the state will transit to s_{i_2} with probability $p_{s_{i_1} s_{i_2} r_1}^{a_1}$; similarly, in the next step, another random additive reward $r_2 = s_{i_2} r(a_2)$ will be received when the action a_2 is taken, and so on. So, we can rewrite the above total rewards from s_{i_1} as

$$\begin{aligned} W_n^\pi &= s_{i_1} + s_{i_1} r(a_1) + s_{i_2} r(a_2) + \cdots + s_{i_n} r(a_n) \\ &= s_{i_1} + \sum_{j=1}^{n} s_{i_j} r(a_j). \end{aligned}$$

Thus, the reward sets, for each state-action pair (s_i, a), are now given by

$$W(s_i, a) = \{w \,|\, w = s_i r(a)\} \cup \{s_i\} \quad \forall s_i \in S \quad a \in A.$$

So, the *aggregate reward set* can be written as: $W = \cup_{s_i \in S, a \in A} W(s_i, a)$. This completes the second step of the definition of the reward r.

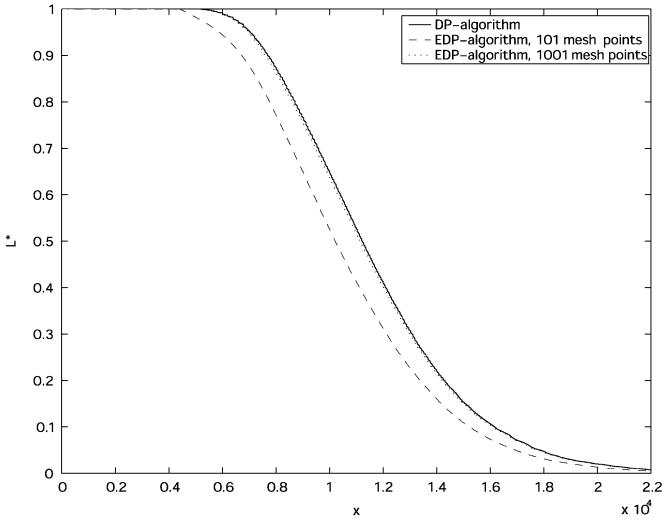


Fig. 1. Probabilities $L_{10}^*(10000, x)$ of reaching target x after ten years using different algorithms, with initial capital of \$10 000, when an optimal policy is followed.

[2]Instead of the mid-point of the interval $(s_{i-1}, s_i]$ we could also have used the right end-point, or some other point. If these intervals are narrow and the time horizon is long the results will not be significantly different.

[3]For example, if the current wealth is \$7000 this year, and the decision maker allocates his or her total wealth to the option-j, then his or her expected wealth next year will be $\$7000(1 + r_j), j = 1, 2, 3$.

TABLE I
HISTORICAL DATA

Three Options				The	eight	years			
		1993	1994	1995	1996	1997	1998	1999	2000
Option1	Capital value($)	1000	1075	1125	1375	1425	1475	1925	2350
	Percentage(%)		7.5	4.65	22.22	3.64	3.51	30.51	22.08
Option2	Capital value($)	1000	1025	1050	1400	1425	1475	1825	2125
	Percentage(%)		2.5	2.44	33.33	1.79	3.51	23.73	20.55
Option3	Capital value($)	1000	1125	1225	1425	1550	1625	1925	2225
	Percentage(%)		12.5	8.89	16.33	8.77	4.84	18.46	15.58

Now, the target hitting time, namely the first random time at which the total reward W_n^π exceeds the target value x, for a fixed $\pi \in \Pi_m^d$, is given by

$$\tau(x) = \inf\{k \mid W_k^\pi \geq x\,, k \geq 1\}.$$

As explained at the beginning of this section, we define the objective functions by

$$\begin{aligned} L_n^\pi(s_i, x) &= P_\pi(\tau(x) \leq n \mid e_1 \\ &= (s_i, x)) \qquad \forall (s_i, x) \in E \end{aligned}$$

and the optimal value functions by

$$L_n^*(s_i, x) = \sup_{\pi \in \Pi_m^d} \{L_n^\pi(s_i, x)\} \qquad \forall (s_i, x) \in E.$$

A policy π^* is called an n stages optimal policy, if it satisfies: $L_n^{\pi^*}(s_i, x) = L_n^*(s_i, x), \forall (s_i, x) \in E$.

Obviously, the functions $G_n^\pi(s_i, x)$ discussed in Section II and $L_n^\pi(s_i, x)$ are connected via the simple complementary relationships

$$L_n^\pi(s_i, x) + G_n^\pi(s_i, x) = 1 \quad L_n^*(s_i, x) + G_n^*(s_i, x) = 1.$$

Therefore, we can use the theory and algorithm of Section II to calculate the optimal value function and the optimal action set for the function $G_n^\pi(s_i, x)$ and then apply the above relation to compute the corresponding $L_n^\pi(s_i, x)$ function.

Now, the exact values of the transition probabilities $p^a_{s_i s_j r}$ and the reward sets $W(s_i, a)$ can be calculated from the historical data of the performance of the three investment options. The details of these calculations are supplied in the Appendix.

B. Example

As an example, take $S = \{11000, 13000, 15000, 17000, 19000, 21000, 23000, 25000, 27000, 29000, 31\,000\}$, $A = \{(1,0,0), (0,1,0), (0,0,1)\}$ and the historical data[4]

$$H = \begin{bmatrix} 0.04 & 0.08 & 0.12 \\ 0.01 & 0.01 & 0.16 \\ 0.05 & 0.06 & 0.07 \end{bmatrix}.$$

Then for $n = 10$ the DP-algorithm takes 6655.5 s, EDP-algorithm takes 21.86 s for 101 mesh points and 214.43 s for 1001 mesh points.

The results can be used to compare the algorithms. In Fig. 1, both figures using DP-algorithm and EDP-algorithms are drawn. It can easily be seen that the enhanced algorithm approximates the DP-algorithm extremely well for 1001 mesh points (it is hard to see the difference between the two functions). Furthermore, it takes the EDP-algorithm 12.4 h to compute the solution for the stochastic target hitting time problem with dimensions $|S| = 80, |A| = 6, |W(i)| = 7, \forall i \in S$ and time horizon $n = 20$ years (see the illustration in Section IV where 501 mesh points are used). The DP-algorithm however can not even compute two years in this time.

[4]The details of H can be found in the Appendix.

IV. RESULTS AND INTERPRETATIONS

In this section, we solve an illustrative example constructed from historical data of three (out of five) top performing pension funds listed in the Australian Financial Services Directory (http://www.client.afsd.com.au/). These three funds will correspond to the investment options in the theoretical model discussed earlier. They are: Option1-Perpetual's Inv. Choice Pension—Industrial Share, Option2-BT Lifetime Super Pers—Australian Share, Option3-Zurich FIP-Equity.

The capital values and the percentage returns of these three options over a period of 8 years are listed in Table I. Let h_{jt} be the percentage return on \$1 invested in Option j in year t. For instance, $h_{25} = 3.51\%$, the underlined number in Table I. Under the assumption that the data from these eight years are representative of the future performance of these three pension funds it is now possible to construct the reward sets $W(s_i, a)$ for each $s_i \in S$ and $a \in A$ as well as the transition probabilities $p^a_{s_i s_j r}$. The details of these constructions are given in the Appendix.

The system's state–space needs to be finite and to consist of nonoverlapping adjacent intervals $(s_{i-1}, s_i]$. The more intervals, the larger the dimensionality of the problem and thus the greater the computational time. However, with a small number of intervals, there will be a lot of rounding that renders results unreliable. In the case where the state represents accumulated wealth over a long time horizon an argument can be made that when the state of the wealth is small, greater accuracy is required. On the other hand, when the state is large, say of the order of \$300 000 discrepancies of one or two thousands are no longer important. This leads to the following construction of S:

$$\begin{aligned} s_k &= 10\,000 + 50 \times k(k+1) \qquad \forall \quad k = 1, \ldots, 79 \\ s_{80} &= +\infty. \end{aligned}$$

This leads to $S = \{10100, 10300, 10600, 11000, \ldots, 318100, 326000, +\infty\}$, where $s_3 = 10\,600$ means that the current wealth is in the interval $(s_2, s_3] = (10300, 10\,600]$. For computing the transition probabilities and rewards the value $(s_2 + s_3)/2 = 10\,450$ is used. The other intervals are similar, however the endpoints require an adjustment. Here $s_1 = 10\,100$ means that the current wealth is in the interval [10000, 10 100] but for the purpose of computation the value of 10 000 is used.

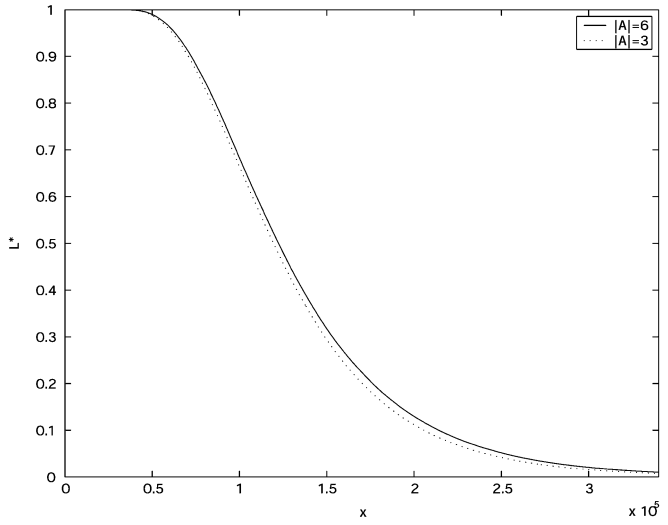


Fig. 2. Probability $L^*_{20}(10000, x)$ of achieving target x after 20 years with initial capital of \$10 000. Comparison between the optimal policy with $|A| = 3$ and the optimal policy with $|A| = 6$.

Further, $s_{80} = +\infty$ means the current wealth is greater than 326 000 but for computation the value of 330 000 is used.

The action space is taken to be

$$A = A(e) = \left\{ a = (\theta_1, \theta_2, \theta_3) \,|\, \theta_j \in \{0, 0.5, 1\} \text{ for } j = 1, 2, 3; \text{ s.t. } \sum_{j=1}^{3} \theta_j = 1 \right\}.$$

So the cardinality of the action space $|A| = 6$. The decision maker has the choice to put all the money in one fund or to divide the money equally among two funds. This will be seen to be less restrictive than might appear at first by comparing with the action space

$$\tilde{A} = \tilde{A}(e) = \left\{ a = (\theta_1, \theta_2, \theta_3) \,|\, \theta_j \in \{0, 1\} \text{ for } j = 1, 2, 3; \text{ s.t. } \sum_{j=1}^{3} \theta_j = 1 \right\}.$$

The latter means all the money has to be put in one option, so $|\tilde{A}| = 3$. The result of this halving of the action space is shown in Fig. 2. The maximum difference is 0.0238. This means that by using a simpler policy that comes from the restricted $\tilde{A}$, one will have at worst 2.38% less chance of getting amount x after 20 years than with the policy described in Section IV. However the computation time halved. For these data it can be said that this simplified strategy is very good. There is also the added "intangible" benefit of easier decision making for the user.

Using the EDP-algorithm 2.4 to compute the optimal policy to maximize the probability L to achieve target x after 20 years with initial capital of \$10 000, leads to the optimal value function $L^*_{20}(10000, x)$. Representative points are presented in Table II. As a comparison also the results by using a superstationary policy are given. A superstationary policy means that all money is placed in one fund and never reallocated. Here, $L^j_{20}(i, x)$ means that all money is put in fund j. In Fig. 3, the complete optimal value function is drawn.

TABLE II
PROBABILITY $L^*_{20}(10000, x)$ OF ACHIEVING TARGET x AFTER 20 YEARS WITH INITIAL CAPITAL OF \$10 000, WHEN AN OPTIMAL POLICY IS FOLLOWED. COMPARED WITH THE NAIVE POLICIES THAT PLACE ALL THE MONEY IN ONE FUND, PERMANENTLY

Target x	$L^*_{20}(10000, x)$	$L^1_{20}(10000, x)$	$L^2_{20}(10000, x)$	$L^3_{20}(10000, x)$
\$ 25,000	1	1	0.9500	0.9985
\$ 50,000	0.9894	0.9778	0.7211	0.9503
\$ 75,000	0.8795	0.6129	0.4618	0.7904
\$ 100,000	0.6774	0.1566	0.2774	0.5747
\$ 125,000	0.4787	0.0209	0.1652	0.3865
\$ 150,000	0.3157	0.0016	0.0971	0.2417
\$ 175,000	0.2031	0	0.0573	0.1457
\$ 200,000	0.1283	0	0.0341	0.0868
\$ 225,000	0.0818	0	0.0209	0.0513
\$ 250,000	0.0514	0	0.0128	0.0301
\$ 275,000	0.0322	0	0.0078	0.0177
\$ 300,000	0.0204	0	0.0049	0.0103
\$ 325,000	0.0132	0	0.0032	0.0062

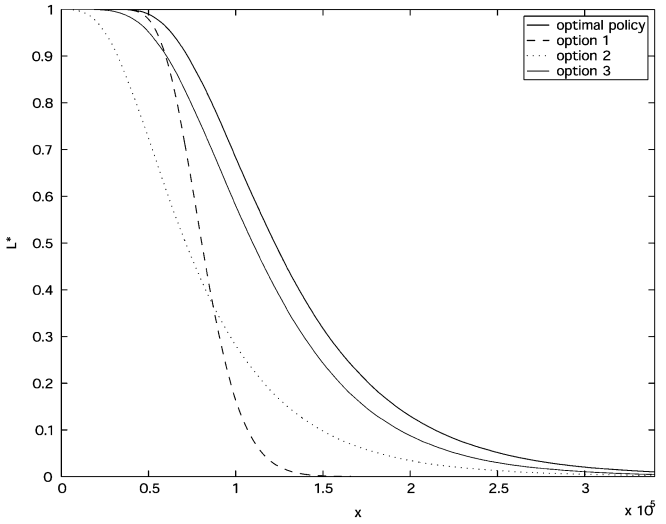


Fig. 3. Probability $L^*_{20}(10000, x)$ of achieving target x after 20 years with initial capital of \$10000, when an optimal policy is followed. Compared with the naive policies that place all the money in one fund, permanently.

V. CONCLUSION

It should be mentioned that the preceding theory and implementation are simpler than the "real-life" problem of investment for retirement in a number of aspects that were already mentioned in Section II. Some of these problems could be easily incorporated into our method. For instance, each year new data become available about the performance of the investment funds. This means that we could, in principle, update our rewards and transition probabilities every year prior to making our next decision on the allocations.

In further research it would be interesting to implement better predictions for future rewards in the model. In this model the historical data are used to predict future performance. It is assumed that the yield of a given fund in every year in the future is best modeled as a random variable that takes on the past observed yields from that fund with equal probability. This is a rather simplistic assumption that may not correspond to reality.[5]

To try to alleviate this problem one could consider a model with *rolling horizon policies*. The idea of this approach is that an optimal policy is found and the first decision rule is implemented. Then, if new data are available, a problem with updated parameters and a new time horizon is solved. The first decision rule from an optimal policy of the latter is then implemented and so on. Rolling horizons have been used by many researchers (e.g., see [14]).

Another important aspect is that in practice most salaried employees receive not only a return on the investment from the previous year but also a new contribution (typically a percentage of a salary) from their employers. Once again, our methods and the algorithm can be easily modified to account for this complication by a suitable adaptation of the rewards which will now become stage-dependent. Of course, such a modification could also be used to incorporate the anticipated promotions and jumps in salary.

APPENDIX

In this appendix, we supply details of the derivation of rewards and transition probabilities from historical data on the performance of the three investment funds referred to in Section III.

We assume that the historical data describing the performance of the m investment options[6] over a period of past q years are given by the matrix:

$$H = \begin{bmatrix} h_{11}, & h_{12}, & \cdots, & h_{1q} \\ \vdots & \vdots & \vdots & \vdots \\ h_{m1}, & h_{m2}, & \cdots, & h_{mq} \end{bmatrix}_{m \times q}$$

where each h_{jt} represents the interest rate on \$1 invested in the option-j in the tth year in the past.

One of the assumptions is that the historical data can predict future performances. Consider the fund j. On the basis of the past history the yield from that fund is regarded as a random variable Y_j which takes on values $h_{j1}, h_{j2}, \ldots, h_{jq}$ with probability $1/q$. The assumption made here is that the yield $Y_j^{t^*}$ from the same fund j at year t^* in the future is identically distributed as Y_j. Clearly, alternative assumptions could be made. For instance one could assign higher probabilities to yields that result from historical data from more recent years.

When a decision maker invests s_i and chooses action $a = (\theta_1, \theta_2, \ldots, \theta_m)$ he or she will expect to earn a total reward of $r = s_i \sum_{j=1}^{m} \theta_j r_j$ where r_j is the interest rate of option j. However, this interest rate is a random variable. There are q possibilities for the interest rate, so there will also be q possibilities for the total rewards. The reward sets can now be given by

$$W(s_i, a) = \left\{ r = s_i \sum_{j=1}^{m} \theta_j h_{jt} \,|\, a = (\theta_1, \theta_2, \ldots, \theta_m), \quad t = 1, \ldots, q \right\}.$$

[5]On the other hand the historical data on the funds past performance is usually the best information available.

[6]In our model, we set $m = 3$

The transition probabilities can be constructed in a similar way. Given s_i and a, the next state s_j will be determined by the reward r that is realized. Let $r_1, r_2, \ldots, r_z$ be the distinct values of r in $W(s_i, a)$, where $z = |W(s_i, a)| \leq q$ and $T = \{1, 2, \ldots, q\}$ denotes the columns of H. Let

$$T_k(s_i, a) = \left\{ t \in T \,|\, r_k = s_i \sum_{j=1}^{m} \theta_j h_{jt}, a = (\theta_1, \theta_2, \ldots, \theta_m) \right\}$$

and

$$\delta_k(s_i, a) = |T_k(s_i, a)|.$$

Clearly, $\delta_k(s_i, a) = 1, \forall (s_i, a)$, if $z = q$. The transition probabilities can now be naturally defined by

$$p^a_{s_i s_j r_k} = \frac{\delta_k(s_i, a)}{q}.$$

REFERENCES

[1] A. D. Roy, "Safety first and the holiday of assets," *Econometrica*, vol. 220, pp. 431–449, 1996.

[2] C. Wu and Y. Lin, "Minimizing risk models in Markov decision processes with policies depending on target values," *J. Math. Anal. Appl.*, vol. 231, pp. 47–67, 1999.

[3] D. J. White, "Mean, variance, and probabilistic criteria in finite Markov decision processes: A review," *J. Optim. Theory Appl.*, vol. 56, pp. 1–29, 1988.

[4] ——, "Minimizing a threshold probability in discounted Markov decision processes," *J. Math. Anal. Applicat.*, vol. 173, pp. 634–646, 1993.

[5] H. Markowitz, *Portfolio Selection: Efficient Diversification of Investment*. New York: Wiley, 1959.

[6] R. A. Howard and J. E. Matheson, "Risk-sensitive Markov decision processes," *Manage. Sci.*, vol. 18, pp. 356–369, 1972.

[7] J. B. Krawczyk, "A Markovian approximated solution to a portfolio management problem," *Inform. Technol. Econ. Manage.*, vol. 99, pp. 530–551, 1997.

[8] J. A. Filar, L. C. M. Kallenberg, and H.-M. Lee, "Variance-penalized Markov decision processes," *Math. Oper. Res.*, vol. 14, pp. 147–161, 1989.

[9] J. A. Filar, D. Krass, and K. W. Ross, "Percentile performance criteria for limiting average Markov decision processes," *IEEE Trans. Automat. Contr.*, vol. 40, pp. 2–10, Jan. 1995.

[10] L. Jianyong and S. Huang, "Markov decision processes with distribution function criterion of first-passage time," *Appl. Math. Optim.*, vol. 43, pp. 187–201, 2001.

[11] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY: Wiley, 1994.

[12] M. Bouakiz and Y. Kebir, "Target-level criterion in Markov decision processes," *J. Optim. Theory Appl.*, vol. 86, pp. 1–15, 1995.

[13] R. C. Merton, *Continuous-Time Finance*. Oxford, U.K.: Basic Blackwell, 1990.

[14] R. E. Wildeman, R. Dekker, and A. C. J. M. Smit, "A dynamic policy for grouping maintenance activities," *Eur. J. Oper. Res.*, vol. 99, pp. 530–551, 1997.

[15] R. Cavazos-Cadena and E. Fernández-Gaucherand, "Controlled Markov chains with risk-sensitive criteria: Average cost, optimality equations, and optimal solutions," *Math. Meth. Oper. Res.*, vol. 49, pp. 299–324, 1999.

[16] M. J. Sobel, "The variance of discounted MDPs," *J. Appl. Probab.*, vol. 19, pp. 794–802, 1982.

[17] S. X. Yu, Y. Lin, and P. Yan, "Optimization models for the first arrival target distribution function in discrete time," *J. Math. Anal. Appl.*, vol. 225, pp. 193–223, 1998.

[18] S. C. Jaquette, "A utility criterion for Markov decision processes," *Manage. Sci.*, vol. 23, pp. 43–49, 1976.

[19] S. Roy, "Theory of dynamic portfolio for survival under uncertainty," *Math. Social Sci.*, vol. 30, pp. 171–194, 1995.

[20] P. Whittle, *Risk-Sensitive Optimal Control*. New York: Wiley, 1990.

[21] Y. Lin, B. Kang, and C. Wu, "Finite horizon Markov decision minimizing risk models in Borel state space," working paper, 2004.

[22] Y. Lin and J. Lin, "Models for the first arrival time distribution function optimization and risk minimization," *J. Tsinghua Univ.*, vol. 36, pp. 53–59, 1995.

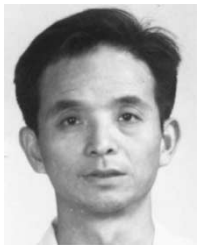
[23] Y. Lin, J. A. Filar, and K. Liu, "Finite horizon portfolio risk models with probability criterion," in *Markov Processes and Controlled Markov Chains*. Norwell, MA: Kluwer, pp. 405–424.

Kang Boda was born in Kunming, China, in 1977. He received the B.S. degree in applied mathematics and the M.S. degree in probability and mathematical statistics, both from Tsinghua University, Beijing, China, in 2000 and 2003, respectively. He is currently working toward the Ph.D. degree at the School of Mathematics and Statistics, the University of South Australia.

His research interests are in Markov decision processes with probability criteria, financial mathematics, and risk analysis in financial markets.

Jerzy A. Filar is a broadly trained applied mathematician with research interests spanning a wide spectrum of both theoretical and applied topics in operations research, optimization, game theory, applied probability and environmental modeling. After a 16-year academic career in the United States, which included appointments at the University of Minnesota, Minneapolis, The Johns Hopkins University, Baltimore, MD, and the University of Maryland, College Park, he returned to Australia in 1992, where he is currently the Foundation Chair of Mathematics and Statistics at the University of South Australia.

Dr. Filar is a Fellow of the Australian Mathematical Society.



Yuanlie Lin was born in Fujian, China, in 1938. He graduated from Tsinghua University, Beijing, China, in 1962.

He is currently a Professor of Mathematics with the Department of Mathematical Sciences, Tsinghua University. His research interests are in the area of Markov chains and their applications and include Markov decision processes with probability criteria, moment optimization, queueing theory, financial risk, and optimal portfolio theory. Recently, he has focused on applying statistics and Markov chains in computational biology and biostatistics.

Lieneke Spanjers is working toward the M.Sc. degree in applied mathematics at the Faculty of Electrical Engineering, Mathematics and Computer Science, the University of Twente, Twente, The Netherlands.