1 ON MODELING RISK IN MARKOV DECISION PROCESSES

Steve Levitt
Taro Pharmaceuticals, Hawthorne NY 10532, USA

Adi Ben-Israel
RUTCOR–Rutgers Center for Operations Research, Rutgers University
640 Bartholomew Rd, Piscataway, NJ 08854-8003, USA

Abstract: Markov decision processes are solved recursively, using the Bellman optimality principle,

$$V(s,t) := \max_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \alpha \sum_{j \in \mathcal{S}} p_{s,j}(a)\, V(j,t+1) \right\} \tag{A}$$

where $V(s,t)$ is the optimal value of state s at stage t, $r(s,a)$ is the instantaneous profit from action a at state s, $\mathcal{S}$ is the state space, $\mathcal{A}(s)$ the set of feasible actions at state s and $p_{i,j}(a)$ the transition probabilities from i to j. This solution maximizes the expected value of the discounted sum of future profits (the right side of (A)), and assumes risk neutrality, i.e. the decision maker is indifferent between a random variable and its expected value.

We propose an alternative solution, with explicit modeling of risk, using the recursion

$$V(s,t) := \max_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \alpha\, \mathrm{S}_\beta(V(\mathbf{Z}(s,a), t+1)) \right\} \tag{B}$$

where $Z(s,a)$ is the next state, S_β is the quadratic certainty equivalent

$$\mathrm{S}_\beta(\mathbf{X}) := \mathrm{E}\,\mathbf{X} - \frac{\beta}{2} \mathrm{Var}\,\mathbf{X}$$

and β is a parameter modeling the attitude of the decision maker towards risk: $\beta > 0$ if risk-averse, $\beta < 0$ if risk seeking and $\beta = 0$ if risk-neutral (in which case (B) reduces to (A)).

We apply our model to solve two problems of maintenance and inventory and compare with the classical solution.
Key words: Decision-making under uncertainty. Certainty equivalents. Risk aversion. Dynamic programming. Markov decision process.
Mathematics Subject Classification (2000) 90C39, 91B30, 91B06

1 INTRODUCTION

We use the following notation for a Markov Decision Process (MDP)

T	the number of **stages** (assumed finite)
$\mathcal{S}$	**state space** (discrete)
s_t	the **state** at the beginning of stage $t = 1, \cdots, T$
s_1	the **initial state**, given
$\mathcal{A}(s)$	**action set** (finite) for each state $s \in \mathcal{S}$
a_t	the **action** taken at stage $t = 1, \cdots, T$
$r(s, a)$	the **stage return** from state s and action a
$p_{ij}(a)$	the **transition probabilities** (from i to j, depending on the action a)
α	the **discount factor**
$V(s, t)$	the **optimal value** (**OV**) **function** in stage t with state $s \in \mathcal{S}$.

The MDP is solved recursively, using Bellman's optimality principle, as follows

$$V(s_t, t) := \max_{a \in \mathcal{A}(s_t)} \left\{ r(s_t, a) + \alpha \sum_{j \in \mathcal{S}} p_{s_t, j}(a)\, V(j, t+1) \right\} \quad (1.1a)$$

$$s_t \in \mathcal{S}\ ,\ t = 1, \cdots, T$$

$$V(s_{T+1}, T+1) := \text{the } \textbf{salvage value} \text{ of the } \textbf{terminal state}\ , \quad (1.1b)$$

and the maximizing a_t^* give the **optimal policy** $\{a_t^* : t = 1, \cdots, T\}$.

The recursion (1.1a) can be written as

$$V(s_t, t) := \max_{a \in \mathcal{A}(s_t)} \left\{ r(s_t, a) + \alpha \, \mathrm{E}\, V(\mathbf{Z}(s_t, a), t+1) \right\}\ , \quad t = 1, \cdots, T \quad (1.2)$$

where $\mathbf{Z}(s_t, a)$, the **next state**, is a random variable (RV). The OV function $V(\mathbf{Z}(s_t, a), t+1)$, in the RHS of (1.2), is random through its argument. It is replaced, in this computation, by its expected value $\mathrm{E}\, V(\mathbf{Z}(s_t, a), t+1)$. An optimal policy obtained from (1.1) is therefore risk–neutral (indifferent between a RV and its expected value), unless some other risk–attitude is implicit in the return functions $r(s, a)$.

We propose here an alternative formulation of the MDP, with explicit modeling of risk–attitude. The classical model (1.1) is a special case of our model, corresponding to risk–neutrality.

We replace (1.2) by

$$V(s_t, t) := \max_{a \in \mathcal{A}(s_t)} \left\{ r(s_t, a) + \alpha \, \mathrm{S}(V(\mathbf{Z}(s_t, a), t+1)) \right\}\ , \quad t = 1, \cdots, T \quad (1.3)$$

where $\mathrm{S}(\mathbf{X})$ is a **certainty equivalent** of the RV $\mathbf{X}$ in question, see Appendix A for explanation and justification of (1.3). We use here the **quadratic certainty equivalent**

$$\mathrm{S}_\beta(\mathbf{X}) := \mathrm{E}\, \mathbf{X} - \frac{\beta}{2} \mathrm{Var}\, \mathbf{X} \quad (1.4)$$

with β a **risk parameter**, increasing with risk aversion. The case $\beta = 0$ corresponds to the customary (risk-neutral) recursion (1.3), and $\beta > 0$ [$\beta < 0$] gives **risk averse** [**risk seeking**] behavior. The parameter β is assumed sufficiently small so that $U(x) = x - \frac{\beta}{2}\, x^2$ is increasing throughout the support of $\mathbf{X}$.

A corresponding certainty equivalent of a random stream $\mathbf{X} = (\mathbf{X}_1\,, \cdots, \mathbf{X}_T)$ is

$$\begin{aligned} \mathrm{S}_{\{\beta_1,\cdots,\beta_T\}}(\mathbf{X}) \;&:=\; \sum_{t=1}^{T} \alpha^{t-1}\, \mathrm{S}_{\beta_t}(\mathbf{X}_t) & (1.5a)\\ &=\; \sum_{t=1}^{T} \alpha^{t-1} \left\{ \mathrm{E}\,\mathbf{X}_t - \frac{\beta_t}{2}\,\mathrm{Var}\,\mathbf{X}_t \right\} & (1.5b) \end{aligned}$$

where the β_t allow modeling different risk attitudes in different stages. If all $\beta_t = \beta$ we denote (1.5a) by

$$\mathrm{S}_{\boldsymbol{\beta}}(\mathbf{X}) \;:=\; \sum_{t=1}^{T} \alpha^{t-1}\, \mathrm{S}_{\beta}(\mathbf{X}_t) \qquad (1.6)$$

Using the certainty equivalent (1.4), the recursion (1.3) is about as easy to compute as (1.1a). However, the optimal policy obtained by (1.3) reflects the risk–attitude of the certainty equivalent $\mathrm{S}(\cdot)$, and is in general different than the optimal policy of (1.1).

We illustrate this for a class of MDP's where the optimal policies are myopic, see § 2, and the maintenance example in § 3, and for an inventory problem, § 5, where there are optimal order-to levels.

2 MYOPIC OPTIMA IN MDP'S

Following Sobel (1981) we show here that certain MDP's, solved by (1.3), have myopic optimal solutions. As there we assume that

- the set

$$\begin{aligned} \mathcal{W} \;&:=\; \{(s,a) : a \in \mathcal{A}(s), s \in \mathcal{S}\} \text{ is finite, and denote} & (2.1a)\\ \mathcal{S}(a) \;&:=\; \{s \in \mathcal{S} : a \in \mathcal{A}(s)\}\;, & (2.1b)\\ &\qquad \text{the set of states in which action } a \text{ is feasible}\,. \end{aligned}$$

- the returns $r(s,a)$ have the form

$$r(s,a) \;=\; K(a) + L(s)\;, \quad (s,a) \in \mathcal{W}\;, \qquad (2.2)$$

where $L(s)$ is the salvage value function as in (1.1b), and
- the transition probabilities $p_{ij}(a)$ do not depend on i,

$$p_{ij}(a) \;:=\; q_j(a)\;, \quad \forall\, i,j \in \mathcal{S},\, a \in \mathcal{A} \qquad (2.3)$$

The last assumption implies that the next state $\mathbf{Z}(s,a)$ depends only on $a \in \mathcal{A}$, i.e. there is a RV $\boldsymbol{\zeta}(a)$ with the same distribution, a fact denoted by

$$\mathbf{Z}(s,a) \;\sim\; \boldsymbol{\zeta}(a) \;, \quad \forall\, (s,a) \in \mathcal{W} \tag{2.4}$$

Theorem 2.1 (After Sobel, Sobel (1981), Theorem 1). Let (2.2)–(2.3) hold, and use the certainty equivalents S_β and $\mathrm{S}_{\boldsymbol{\beta}}$ of (1.4) and (1.6). Denote

$$G_\beta(a) \;:=\; K(a) + \alpha\, \mathrm{S}_\beta\left(L(\boldsymbol{\zeta}(a))\right) \;, \quad a \in \mathcal{A} \tag{2.5}$$

Let $a^*(\beta)$ maximize $G_\beta(a)$ on $\mathcal{A}$

$$K(a^*(\beta)) + \alpha\, \mathrm{S}_\beta\left(L(\boldsymbol{\zeta}(a^*(\beta)))\right) \;\geq\; K(a) + \alpha\, \mathrm{S}_\beta\left(L(\boldsymbol{\zeta}(a))\right) \;, \quad \forall\, a \in \mathcal{A} \tag{2.6}$$

and suppose

$$a^*(\beta) \;\in\; \mathcal{A}(s_1) \tag{2.7a}$$

$$\sum_{j \in \mathcal{S}(a^*(\beta))} q_j(a^*(\beta)) \;=\; 1 \tag{2.7b}$$

Then the policy $a_t \;:=\; a^*(\beta)$, $t = 1, 2, \cdots$, is optimal.

Proof: This proof is an adaptation of the proof of Sobel (1981), Theorem 1. We use the shift additivity property of the RCE $\mathrm{S}_\beta(\cdot)$, see (A.5),

$$\mathrm{S}_\beta(\mathbf{X} + c) \;=\; \mathrm{S}_\beta(\mathbf{X}) + c \;, \quad \text{for all RV } \mathbf{X} \text{ and constant } c \;. \tag{2.8}$$

Denote the $(T+1)$–dimensional random vector of (undiscounted) rewards by

$$\begin{aligned} \mathbf{X} \;&=\; \left(r(s_1,a_1)\,,\, r(s_2,a_2)\,,\, \cdots\,,\, r(s_T,a_T)\,,\, L(s_{T+1})\right) \\ &=\; \left(r(s_1,a_1)\,,\, r(\boldsymbol{\zeta}(a_1),a_2)\,,\, \cdots\,,\, r(\boldsymbol{\zeta}(a_{T-1}),a_T)\,,\, L(\boldsymbol{\zeta}(a_T)\right) \;, \quad \text{by (2.4)} \;. \end{aligned}$$

Therefore, by (1.5a),

$$\begin{aligned} \mathrm{S}_{\{\beta,\cdots,\beta\}}(\mathbf{X}) \;&=\; r(s_1,a_1) + \sum_{t=2}^{T} \alpha^{t-1} \left\{\mathrm{S}_\beta\left(r(\boldsymbol{\zeta}(a_{t-1}),a_t)\right)\right\} + \alpha^T\, \mathrm{S}_\beta\left(L(\boldsymbol{\zeta}(a_T))\right) \\ &= K(a_1) + L(s_1) + \sum_{t=2}^{T} \alpha^{t-1}\, \mathrm{S}_\beta\left(K(a_t) + L(\boldsymbol{\zeta}(a_{t-1}))\right) + \alpha^T\, \mathrm{S}_\beta\left(L(\boldsymbol{\zeta}(a_T)\right) \;, \\ &\qquad \text{by (2.2)} \;, \\ &= K(a_1) + L(s_1) + \sum_{t=2}^{T} \alpha^{t-1} \left\{K(a_t) + \mathrm{S}_\beta\left(L(\boldsymbol{\zeta}(a_{t-1}))\right)\right\} + \alpha^T\, \mathrm{S}_\beta\left(L(\boldsymbol{\zeta}(a_T)\right) \;, \\ &\qquad \text{by (2.8)} \;, \\ &=\; L(s_1) + \sum_{t=1}^{T} \alpha^{t-1} \left\{K(a_t) + \alpha\, \mathrm{S}_\beta\left(L(\boldsymbol{\zeta}(a_t))\right)\right\} \;. \end{aligned}$$

From (2.6) it follows that the policy $\{a^*(\beta)\}$ is optimal, if it is feasible, i.e. if

$$a^*(\beta) \in \mathcal{A}(s_t) \;, \quad t = 1, 2, \cdots$$

This is guaranteed by (2.7a)–(2.7b). △

3 A MAINTENANCE PROBLEM

We solve the maintenance problem in Sobel (1981), § 5. A system has N identical and independent units, each is in one of two states, **functioning** or **broken**.

The **state** s of the system is the number of functioning units, $s = 0, 1, \cdots, N$.

Before each period, the state s is observed, and a decision is made how many units to repair. The number of units to be repaired is denoted by $a - s$, so that after repair there are a functioning units. The **cost of repair** is C_r per unit.

Each functioning unit may break, during the period, with **probability** p. The probability that exactly j units are still functioning at the end of the period (out of the a units functioning at the beginning of the period) is

$$q_j(a) \;=\; \binom{a}{j} p^{a-j} (1-p)^j \;, \quad j = 0, \cdots, a \;.$$

and is independent of the beginning state, i.e. the transition probabilities satisfy (2.3).

If all units break during the period, i.e. if the state becomes $s = 0$, a **penalty** of C_p is paid. Otherwise (i.e. if $s \geq 1$ at the end of the period) a **revenue** of R is collected.

For convenience we assume that the **salvage value** per functioning unit (at the end of the last period), is equal to the **cost of repair** C_r.

The return $r(s,a)$ is therefore

$$\begin{aligned} r(s,a) &= (1 - q_0(a))\, R - C_p\, q_0(a) - C_r\,(a - s) && \text{(3.1a)} \\ &= K(a) + L(s)\;, \quad \text{as in (2.2)}\;, \\ \text{where}\quad K(a) &= (1 - q_0(a))\, R - C_p\, q_0(a) - C_r\, a && \text{(3.1b)} \\ L(s) &= C_r\, s\;. && \text{(3.1c)} \end{aligned}$$

The optimal policy, see Sobel (1981), is:

repair $\max\{a^* - s, 0\}$ **units if the state is** s

where the optimal level a^* is the maximizer of

$$\begin{aligned} G(a) &:= K(a) + \alpha\, \mathrm{E}\, L(\boldsymbol{\zeta}(a)) \\ &= R\,(1 - q_0(a)) - C_p\, q_0(a) - C_r\, a + \alpha\, C_r \sum_{j=0}^{a} j\, q_j(a) && \text{(3.2a)} \\ &= R - (R + C_p)\; p^a - C_r(1 - \alpha\, p)\, a && \text{(3.2b)} \end{aligned}$$

Using our approach, the $G_\beta(a)$ of (2.5) is

$$\begin{aligned} G_\beta(a) &:= G(a) - \alpha\, \frac{\beta}{2}\, \mathrm{Var}\{C_r \boldsymbol{\zeta}(a)\} \\ &= G(a) - \alpha\, \frac{\beta}{2}\, C_r^2\, \mathrm{Var}\{\boldsymbol{\zeta}(a)\} \\ &= G(a) - \alpha\, \frac{\beta}{2}\, C_r^2\, a\, p\,(1 - p) && \text{(3.3)} \end{aligned}$$

Corollary 3.1 The maximizer $a^*(\beta)$ of $G_\beta(a)$ is a non-increasing function of β. In particular,

$$a^*(\beta) \geq a^* \quad \text{if} \quad \beta < 0 \tag{3.4a}$$
$$a^*(\beta) \leq a^* \quad \text{if} \quad \beta > 0 \tag{3.4b}$$

Proof: The maximizer $a^*(\beta)$ of $G_\beta(a)$ satisfies

$$\begin{aligned} G_\beta(a) - G_\beta(a-1) &\geq 0 \\ G_\beta(a+1) - G_\beta(a) &\leq 0 \end{aligned}$$

and the proof follows since, by (3.3),

$$G_\beta(a+1) - G_\beta(a) = G(a+1) - G(a) - \alpha \frac{\beta}{2} C_r^2 \, p \, (1-p) \tag{3.5}$$

a decreasing function of β. △

It follows from (3.4a) that the risk seeking manager, with $\beta < 0$, will never repair less units than the risk-neutral manager.

4 A MAINTENANCE EXAMPLE

This example is based on the maintenance example in Sobel (1981), § 5. There are $N = 4$ identical units, which break independently with a probability of $p = 0.3$. If any of the units are working at the end of a stage, the system generates $R = 1000$, otherwise, a penalty of $C_p = 1500$ is incurred. Before each stage, the number of functioning units, s, is observed and a decision, a, is made to decide how many units will be operational for the stage (i.e. $a - s$ units are repaired). The cost to repair a machine is $C_r = 500$.

So for this specific example:

$$r(s,a) = 1000\,(1 - q_0(a)) - 1500\,q_0(a) - 500\,(a - s) \tag{4.1}$$

is of the form $r(s,a) = K(a) + L(s)$, see (3.1), where

$$\begin{aligned} K(a) &= 1000\,(1 - q_0(a)) - 1500\,q_0(a) - 500\,a \\ L(s) &= 500\,s \end{aligned}$$

and $G_\beta(a)$ takes the form (3.3),

$$\begin{aligned} G_\beta(a) &= K(a) + \alpha S_\beta(L(\zeta(a)))) \\ &= K(a) + 500\,\alpha \sum_{j=1}^{a} j\,q_j(a) - \alpha \frac{\beta}{2} C_r^2 \, a \, p \, (1-p) \; . \end{aligned} \tag{4.2}$$

We seek the maximizer of $G_\beta(a)$. The table below gives values of $G_\beta(a)$ for three typical values of β,

$\beta = 0.006$, **risk averse**,
$\beta = 0$, **risk neutral** , and
$\beta = -.01$, **risk seeking** .

For each β we underline the maximum value of $G_\beta(a)$.

β	a				
	0	1	2	3	4
.006	−1500	−75	<u>141</u>	116	−248
0	−1500	83	440	<u>565</u>	350
−.01	−1500	345	939	1313	<u>1347</u>

The maximzing $a^*(\beta)$ are

$$\begin{aligned} a^*(.006) &= 2 \\ a^*(0) &= 3 \\ a^*(-.01) &= 4 \end{aligned}$$

showing, in agreement with Corollary 3.1, that the risk averse manager will invest less in repair.

5 AN INVENTORY PROBLEM

The model in this section is based on Denardo (1982), pp. 117–125. It concerns inventory of a single (discrete) commodity with random demand. We denote

T the number of **stages** (possibly infinite)
$\mathbf{D}_t$ the **demand** in stage t
$p_t(j)$ the probability that $\mathbf{D}_t = j$, $j = 0, 1, 2, \cdots$
W the **wholesale price** [$/unit]
R the **retail price** [$/unit]
S the **salvage value** [$/unit] at the end of the horizon (stage $T + 1$)
r the **interest rate** per stage
α $= 1/(1 + r)$, the **discount factor**
M the maximum capacity of the warehouse

and assume

$$S < W < R . \tag{5.1}$$

We further denote

s_t the **inventory level** just before stage t (**state variable**)
a_t the **inventory level** at the beginning of stage t (**decision variable**)

The states evolve according to

$$s_{t+1} := (a_t - \mathbf{D}_t)^+ , \quad t = 1, 2, \cdots, T \tag{5.2a}$$

$$\text{where} \quad s_1 := \text{the } \textbf{initial state} \text{ (given)} . \tag{5.2b}$$

In stage $t = 1, \cdots, T$,
the **sales** are $\min\{\mathbf{D}_t, a_t\}$, and accordingly the **revenue** is $R \min\{\mathbf{D}_t, a_t\}$,
the **amount ordered** is $(a_t - s_t)$, so the **ordering cost** is $W(a_t - s_t)$, and finally

the **interest on inventory** is $\alpha\, r\, a_t\, W$.
We assume, following Denardo (1982), p. 119, that **revenue** and **interest on inventory** occur at the end of the stage. Consequently, the **profit** in stage t, with state s and action a, is

$$\mathbf{\Pi}_t(s,a) = \alpha R \min\{\mathbf{D}_t, a\} - W(a-s) - \alpha r a W \ , \ t = 1, \cdots, T \quad (5.3a)$$

$$\mathbf{\Pi}_{T+1}(s) = s\, S \quad (5.3b)$$

where end-of-stage money is multiplied by α. We denote

$$g_t(a) := \alpha R\, \mathrm{E}\, \{\min\{\mathbf{D}_t, a\}\} - \alpha\, r\, a\, W \quad (5.4)$$

so the expected profit in stage t is $\mathrm{E}\,\mathbf{\Pi}_t(s,a) = g_t(a) - W\,(a-s)$, $t = 1, \cdots, T$.

5.1 *The classical solution*

Let $V(s_t, t)$ denote the maximal profit resulting from beginning stage t with state s_t. The Bellman optimality principle (1.1) then gives

$$V\,(s_t, t) := \max_{s_t \le a \le M} \left\{ g_t(a) - W\,(a - s_t) + \alpha\, \mathrm{E}\, \left\{ V\left((a-\mathbf{D}_t)^+, t+1\right)\right\}\right\}$$
$$\text{for } t = 1, \cdots, T\ , \quad (5.5a)$$

$$V\,(s_{T+1}, T+1) := s_{T+1}\, S\ . \quad (5.5b)$$

It is convenient to change from $V(s,t)$ to

$$\overline{V}(s,t) := V(s,t) - sW\ . \quad (5.6)$$

Then, using the facts:

$$\alpha\, r = 1 - \alpha\ , \quad \text{and}$$
$$(a - \mathbf{D}_t)^+ = a - \min\{\mathbf{D}_t, a\} \quad (5.7)$$

we can rewrite (5.5a) as

$$\overline{V}\,(s_t, t) := \max_{s_t \le a \le M} \left\{ G_t(a) + \alpha\, \mathrm{E}\, \left\{ \overline{V}\left((a-\mathbf{D}_t)^+, t+1\right)\right\}\right\} \quad (5.8a)$$

$$\text{where } G_t(a) := \alpha(R-W)\, \mathrm{E}\, \{\min\{\mathbf{D}_t, a\}\} - 2(1-\alpha)aW\ . \quad (5.8b)$$

The maximand in (5.8a) is independent of s_t. We denote it by

$$L_t(a) := G_t(a) + \alpha\, \mathrm{E}\, \left\{ \overline{V}\left((a-\mathbf{D}_t)^+, t+1\right)\right\}\ . \quad (5.9)$$

We assume now $M = \infty$ (i.e. unlimited storage capacity) and finite $\mathrm{E}\,\{\mathbf{D}_t\}$ for all t. For $t = 1, \cdots, T$ it follows then that the maximand $L_t(\cdot)$ is concave on $\{0, 1, 2, \cdots\}$ and $\lim_{a\to\infty} L_t(a) = -\infty$. Consequently there is a nonnegative integer S_t such that

$$L_t(S_t) = \max_{a \ge 0}\{L_t(a)\} \quad (5.10a)$$

$$\text{and } \overline{V}\,(s,t) = \begin{cases} L_t(S_t) & , \text{ if } s \le S_t \\ L_t(s) & , \text{ if } s > S_t \end{cases} \quad (5.10b)$$

see Denardo (1982), Theorem 6.2. This means, for $t = 1, \cdots, T$, and beginning stock level s_t, that the optimal order is $(S_t - s_t)^+$.

5.2 Solution based on the quadratic certainty equivalent

Consider now the alternative approach, of applying the certainty equivalent $\mathrm{S}_{\boldsymbol{\beta}}$ of (1.6) to evaluate the stream of profits (5.3). The corresponding optimal value functions

$$V(s_t,t) = \max_{a_t,a_{t+1},\cdots,a_T} \mathrm{S}_{\boldsymbol{\beta}}\left(\boldsymbol{\Pi}_t(s_t,a_t),\boldsymbol{\Pi}_{t+1}(s_{t+1},a_{t+1}),\cdots,\boldsymbol{\Pi}_T(s_T,a_T),\boldsymbol{\Pi}_{T+1}(s_{T+1})\right)$$

then satisfy

$$\begin{aligned} V(s,t) &= \max_{s\le a\le M}\{-W(a-s)-\alpha\, r\, a\, W\, + \qquad (5.11a)\\ &+ \alpha \mathrm{S}_\beta\left(R\,\min\{\mathbf{D},a\}+V((a-\mathbf{D})^+,t+1)\right)\}\ ,\ t=1,\cdots,T\\ V(s,T+1) &= s\,S \qquad (5.11b)\end{aligned}$$

Using (5.7) we can rewrite (5.11a) as

$$\begin{aligned} V(s,t) &= \max_{s\le a\le M}\{-W(a-s)-\alpha\, r\, a\, W\, +\\ &\qquad + \alpha \mathrm{S}_\beta\left(R(a-(a-\mathbf{D})^+)+V((a-\mathbf{D})^+,t+1)\right)\}\\ &= \max_{s\le a\le M}\{-W(a-s)-\alpha\, r\, a\, W+\alpha\, a\, R\, +\\ &\qquad + \alpha \mathrm{S}_\beta\left(V((a-\mathbf{D})^+,t+1)-R(a-\mathbf{D})^+\right)\} \qquad (5.12)\end{aligned}$$

where we used the shift–additivity property (2.8) to take the deterministic quantity aR outside S_β. In analogy with (5.6) we define

$$\widehat{V}(s,t) := V(s,t)-sR\,. \qquad (5.13)$$

The recursions (5.11) then become,

$$\begin{aligned} \widehat{V}(s,t) &= -(R-W)\,s+\max_{s\le a\le M}\{(\alpha(R-W)-2(1-\alpha)W)\,a\, + \qquad (5.14a)\\ &+ \alpha\,\mathrm{S}_\beta\left(\widehat{V}((a-\mathbf{D}_t)^+,t+1)\right)\}\ ,\ t=1,\cdots,T\\ &= -(R-W)\,s+\max_{s\le a\le M}\{(\alpha(R-W)-2(1-\alpha)W)\,a\, + \qquad (5.14b)\\ &+ \alpha\,\mathrm{E}\{\widehat{V}((a-\mathbf{D}_t)^+,t+1)\}-\alpha\frac{\beta}{2}\mathrm{Var}\{\widehat{V}((a-\mathbf{D}_t)^+,t+1)\}\\ \widehat{V}(s,T+1) &= s\,(S-R) \qquad (5.14c)\end{aligned}$$

The maximand in (5.14a) is independent of s. We denote it

$$\begin{aligned} \widehat{L}_t(a) &:= G(a)\ +\ \alpha\,\mathrm{S}_\beta\left(\widehat{V}((a-\mathbf{D}_t)^+,t+1)\right) =\\ &= G(a)+\alpha\,\mathrm{E}\,\widehat{V}((a-\mathbf{D}_t)^+,t+1)-\alpha\frac{\beta}{2}\mathrm{Var}\,\widehat{V}((a-\mathbf{D}_t)^+,t+1) \qquad (5.15)\\ \text{where } G(a) &:= (\alpha(R-W)-2(1-\alpha)W)\,a \qquad (5.16)\end{aligned}$$

Note that (5.14b) reduces to (5.8a) if $\beta = 0$.

The following theorem, establishing optimal order-to-levels, is analogous to Denardo (1982), Theorem 6.2,.

Theorem 5.1 Let $M = \infty$, let the random variables $\mathbf{D}_t$ have bounded supports for all t, and let the risk-parameter β be positive. Then for $t = 1, \cdots, T$ the function $\widehat{L}_t(\cdot)$ of (5.15) is concave on $a = 0, 1, 2, \cdots$, and there exists a nonnegative integer S_t such that

$$\widehat{L}_t(S_t) = \max_{a \geq 0} \{\widehat{L}_t(a)\} \tag{5.17a}$$

$$\text{and} \quad \widehat{V}(s,t) = -(R-W)\,s + \begin{cases} \widehat{L}_t(S_t) & , \quad \text{if } s \leq S_t \\ \widehat{L}_t(s) & , \quad \text{if } s > S_t \end{cases} \tag{5.17b}$$

Proof: The theorem follows from the following

claim: for $t = 1, \cdots, T$ the functions $\widehat{L}_t(a)$ are concave, and $\lim_{a\to\infty} \widehat{L}_t(a) = -\infty$

that we prove by induction on t.

(i) The function $\widehat{L}_T$ is concave and $\lim_{a\to\infty} \widehat{L}_T(a) = -\infty$:

$$\begin{aligned} \widehat{L}_T(a) &= G(a) + \alpha\,\mathrm{S}_\beta\left(\widehat{V}((a-\mathbf{D}_T)^+, T+1\right) \\ &= G(a) + \alpha\,\mathrm{S}_\beta\left((S-R)\,(a-\mathbf{D}_T)^+\right) \ , \quad \text{by (5.14c)} \ , \\ &= G(a) + \alpha\,(S-R)\,\mathrm{E}\left\{(a-\mathbf{D}_T)^+\right\} - \alpha\,\frac{\beta}{2}\,(S-R)^2\,\mathrm{Var}\left\{(a-\mathbf{D}_T)^+\right\} \end{aligned}$$

and concavity follows from: $G(a)$ is linear, $(S-R) < 0$, $\mathrm{E}\,(a-\mathbf{D}_T)^+$ is convex in a, and $\beta > 0$. Now,

$$\begin{aligned} \lim_{a\to\infty} \widehat{L}_T(a) &= \\ &= \lim_{a\to\infty} \{G(a) + \alpha\,(S-R)\,a\} - \alpha\,(S-R)\,\mathrm{E}\,\mathbf{D}_T - \alpha\,\frac{\beta}{2}\,(S-R)^2\,\mathrm{Var}\,\mathbf{D}_T \ , \\ &\qquad \text{since } \mathbf{D}_T \text{ has bounded support} \ , \\ &= \lim_{a\to\infty} \{-(\alpha(W-S) + 2(1-\alpha)W)\,a\} + \text{a constant} \ , \ \text{by (5.16)} \ , \\ &= -\infty \ , \quad \text{by (5.1)} \ . \end{aligned}$$

(ii) Assume the claim true for $t+1, \cdots, T$. substituting (5.17b) in

$$\widehat{L}_t(a) := G(a) + \alpha\,\mathrm{S}_\beta\left(\widehat{V}((a-\mathbf{D}_t)^+, t+1)\right)$$

we note that $\widehat{L}_t(a)$ is concave for any degenerate RV $\mathbf{D}_t$. The concavity of (5.15) then follows from that of $\mathrm{S}_\beta(\cdot)$, see (A.6). The statement $\lim_{a\to\infty} \widehat{L}_t(a) = -\infty$ is proved similarly. $\triangle$

Acknowledgments

The research of Steve Levitt was supported initially by NSF, *Research Experiences for Undergraduates* program, at Rutgers University. The authors wish to thank the referees for their constructive suggestions.

hukal,whi-88, baross, moso, boso, sob-90, koch, filvr, fikale

References

Baykal-Gürsoy, M. and Ross, K.W. (1992), Variability sensitive Markov decision processes, *Math, Oper. Res.*, Vol. 17, pp. 558-571.

Ben-Israel, A. and Ben-Tal, A. (1997), Duality and equilibrium prices in economics of uncertainty, *Math. Meth. of Oper. Res.*, Vol. 46, pp. 51-85.

Ben-Tal, A. (1985), The entropic penalty approach to stochastic programming, *Math. Oper. Res.*, Vol. 10, pp. 263-279.

Ben-Tal, A. and Ben-Israel, A. (1991), A recourse certainty equivalent for decisions under uncertainty, *Annals of Oper. Res.*, Vol. 30, pp. 3-44.

Bouakiz, M. and Sobel, M.J. (1992), Inventory control with an exponential utility criterion, *Oper. Res.*, Vol. 40, pp. 603-608.

Denardo, E.V. (1982), *Dynamic Programming: Models and Applications*, Prentice-Hall, Englwood Cliffs, New Jersey.

Filar, J., Kallenberg, L.C.M. and Lee, H.M. (1989), Variance-penalized Markov decision processes, *Math. Oper. Res.*, Vol. 14, pp. 147-161

Filar, J. and Vrieze, K. (1997), *Competitive Markov Decision Processes*, Springer-Verlag, New York.

Huang, Y. and Kallenberg, L.C.M. (1994), On finding optimal solutions for Markov decision chains: a unifying framework for mean-variance tradeoffs, *Math. Oper. Res.*, Vol. 19, pp. 434-448

Köchel, P. (1985), A note on "Myopic solutions of Markov decision processes and stochastic games" by M.J. Sobel, *Oper. Res.*, Vol. 33, pp. 1394-1398.

Monahan, G.E. and Sobel, M.J. (1997), Risk-sensitive dynamic market share attraction games, *Games Econ. Behav.*, Vol. 20, pp. 149-160.

Sobel, M.J. (1981), Myopic solutions of Markov decision processes and stochastic games, *Oper. Res.*, Vol. 29, pp. 996-1009

Sobel, M.J. (1990), Higher-order and average reward myopic-affine dynamic nodels, *Math. Oper. Res.*, Vol. 15, pp. 299-310

White, D.J. (1988), Mean, variance, and probabilistic criteria in finite Markov decision processes: a review, *J. Optimiz. Theory Appl.*, Vol. 56, pp. 1-29

Appendix A: The recourse certainty equivalent

The formulation (1.3) is suggested by the **recourse certainty equivalents** (RCE's) introduced in Ben-Tal (1985), and developed in Ben-Israel and Ben-Tal (1997), and Ben-Tal and Ben-Israel (1991), as criteria for decision making under uncertainty. The RCE of a RV $\mathbf{X}$ is defined as

$$\mathrm{S}_U(\mathbf{X}) := \sup_x \{x + \mathrm{E}U(\mathbf{X} - x)\} \tag{A.1}$$

where $U(\cdot)$ is the decision-maker's **value-risk function**. It induces a complete order "$\succeq$" on RV's,

$$\mathbf{X} \succeq \mathbf{Y} \iff \mathrm{S}_U(\mathbf{X}) \geq \mathrm{S}_U(\mathbf{Y}) \tag{A.2}$$

in which case $\mathbf{X}$ is preferred over $\mathbf{Y}$ by a decision maker (DM) with a value–risk function U. Such a DM is indifferent between a RV $\mathbf{X}$ and the certain payment $\mathrm{S}_U(\mathbf{X})$, denoted by

$$\mathbf{X} \approx \mathrm{S}_U(\mathbf{X}) \tag{A.3}$$

Example. Consider the quadratic value–risk function

$$U(x) := x - \frac{\beta}{2}x^2 \tag{A.4}$$

where β is a risk parameter. If $\beta > 0$ then (A.1) gives the RCE

$$\mathrm{S}_\beta(\mathbf{X}) = \mathrm{E}\,\mathbf{X} - \frac{\beta}{2}\mathrm{Var}\,\mathbf{X} \tag{1.4}$$

Since $\mathbf{X} \approx \mathrm{S}_\beta(\mathbf{X}) \leq \mathrm{E}\,\mathbf{X}$, by (A.3) and (1.4), it follows that a person maximizing the criterion (1.4) is **risk averse** if $\beta > 0$, i.e prefers $\mathrm{E}\,\mathbf{X}$ to $\mathbf{X}$.

If $\beta < 0$ then (A.1) may be unbounded, but we still use the RCE (1.4), to model **risk seeking** behavior. This case is studied in Ben-Israel and Ben-Tal (1997) in the context of maximum buying price. △

An important property of the RCE, that holds for arbitrary value-risk functions U, is **shift additivity**:

$$\mathrm{S}_U(\mathbf{X} + c) = \mathrm{S}_U(\mathbf{X}) + c\,, \quad \text{for all RV } \mathbf{X} \text{ and constant } c\,. \tag{A.5}$$

Thus the RCE separates deterministic changes in wealth from the random variable that it evaluates. For the quadratic value–risk function (A.4), we already encountered shift additivity in (2.8). Another notable property of the RCE is **concavity**: If U is strictly concave then for any RV's $\mathbf{X}_0$, $\mathbf{X}_1$ and $0 < \alpha < 1$,

$$S_U(\alpha\mathbf{X}_1 + (1-\alpha)\mathbf{X}_0) \geq \alpha S_U(\mathbf{X}_1) + (1-\alpha)S_U(\mathbf{X}_0) \tag{A.6}$$

see Ben-Tal and Ben-Israel (1991), Theorem 2.1(f).

The RCE of a vector RV $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_T)$ is defined, analogously to (A.1), as

$$\mathrm{S}_U(\mathbf{X}_1, \cdots, \mathbf{X}_T) = \sup_{x_1, x_2, \cdots, x_T} \left\{ \sum_{t=1}^{T} \alpha^{t-1} x_t + \mathrm{E}\, U(\mathbf{X}_1 - x_1, \cdots, \mathbf{X}_T - x_T) \right\} \tag{A.7}$$

If the value–risk function $U(x_1, \cdots, x_T)$ is of the form (called **separable**)

$$U(x_1, \cdots, x_T) = \sum_{t=1}^{T} \alpha^{t-1} U_t(x_t) \tag{A.8}$$

then the RCE (A.7) is

$$\mathrm{S}_{\{U_1, \cdots, U_T\}}(\mathbf{X}) = \sum_{t=1}^{T} \alpha^{t-1}\, \mathrm{S}_{U_t}(\mathbf{X}_t) \,. \tag{A.9}$$

The RCE $\mathrm{S}_{\{\beta_1, \cdots, \beta_T\}}$ of (1.5b) is a special case of (A.9), if all U_t are quadratic functions (A.4).

At stage t, the current and future rewards form a random vector

$$\begin{aligned} \mathbf{Y}_t &= (r(s_t, a_t)\,,\, r(s_{t+1}, a_{t+1})\,,\, \cdots\,,\, r(s_T, a_T)\,,\, L(s_{T+1})) \\ &= (r(s_t, a_t)\,,\, \mathbf{Y}_{t+1}) \end{aligned}$$

whose RCE is, by (A.9),

$$\mathrm{S}_{\{U_t, U_{t+1} \cdots, U_T, U_{T+1}\}}(\mathbf{Y}_t) = r(s_t, a_t) + \alpha\, \mathrm{S}_{\{U_{t+1} \cdots, U_T, U_{T+1}\}}(\mathbf{Y}_{t+1}) \tag{A.10}$$

An RCE maximizer uses the OV function

$$V(s_t, t) := \max_{a \in \mathcal{A}(s_t)} \mathrm{S}_{\{U_t, U_{t+1} \cdots, U_T, U_{T+1}\}}(\mathbf{Y}_t) \tag{A.11}$$

$$= \max_{a \in \mathcal{A}(s_t)} \left\{ r(s_t, a) + \alpha\, \mathrm{S}_{U_{t+1}}(V(\mathbf{Z}(s_t, a), t+1)) \right\}\,, \tag{A.12}$$

by (A.10), which explains (1.3) .