APPLICATIONES MATHEMATICAE
44,2 (2017), pp. 149–161

Hubert Asienkiewicz (Zielona Góra)
Anna Jaśkiewicz (Wrocław)

# A NOTE ON A NEW CLASS OF RECURSIVE UTILITIES IN MARKOV DECISION PROCESSES

*Abstract.* This paper deals with Markov decision processes on a general state space under standard compactness-continuity assumptions. The purpose is to obtain a new class of so-called recursive utilities with the aid of the entropic risk measure. Within this framework we show that there exists a stationary policy for a discounted payoff problem in the infinite time horizon. Our result is illustrated by examples.

**1. Introduction.** Recursive utility models, suggested by Kreps and Porteus [13] and introduced into the asset pricing literature by Epstein and Zin [8], represent preferences as the solution to a non-linear difference equation with a terminal condition. This is now the standard approach in dynamic problems. To list a few such problems let us mention applications to business cycle models [19], to macroeconomics [14, 20], or to asset pricing models [14]. The popularity of recursive utilities stems from the fact that they allow partial separation between risk aversion and elasticity of intertemporal substitution (see [14, Chapter 20]).

Let us assume that uncertainty is modelled by a probability space $(\Omega, \mathcal{F}, P)$ and that information is modelled by an increasing filtration $(\mathcal{F}_t)$. Within such a framework most applications are based on ranking of stochastic payoff streams, say $(r_t)$, adapted to a given filtration. This ranking is fixed by a stream of preference functionals $(V_t)$ also adapted to $(\mathcal{F}_t)$. The

2010 *Mathematics Subject Classification*: Primary 90C40; Secondary 90C39, 91B70.
*Key words and phrases*: Borel space, dynamic programming, optimal policy, entropic risk measure, certainty equivalent.
Received 15 September 2016; revised 27 November 2016.
Published online 24 May 2017.

DOI: 10.4064/am2317-1-2017  

standard time additive case has the form

$$V_t = r_t + \beta E_t[V_{t+1}], \quad t = 1, 2, \ldots,$$

where $\beta \in [0, 1)$ is a time discount factor and $E_t$ stands for the expectation operator with respect to period $t$ information. We propose to replace $E_t[V_{t+1}]$ by a quasi-arithmetic certainty equivalent operator

$$M_t(V_{t+1}) = \phi^{-1}(E_t(\phi \circ V_{t+1})),$$

where $\phi$ is a monotone function. In particular, we deal with the exponential function $\phi(x) = e^{-\gamma x}$, where $\gamma > 0$ is a risk coefficient of the decision maker. Moreover, $\gamma$ affects decision maker's attitude towards risk in future utility. Thus, the preference functional is of the form

$$V_t = r_t - \frac{\beta}{\gamma} \ln E_t[e^{-\gamma V_{t+1}}]. \tag{1.1}$$

The preferences (1.1) were first studied by Weil [21] and then by Hansen and Sargent [10], who used them to deal with a linear quadratic Gaussian control model. Furthermore, it is worth emphasising that risk sensitive preferences of the form (1.1) have found several applications, for instance, in problems of Pareto optimal allocations [1] or in finance [9, 15], where the certainty equivalent $M_t(V_{t+1}) = -\frac{\beta}{\gamma} \ln E_t[e^{-\gamma V_{t+1}}]$ is known as the entropic risk measure.

In this paper we apply the preference form introduced in (1.1) to define a non-expected payoff in the infinite time horizon for Markov decision processes on a general state space. Under relatively mild conditions we prove that the corresponding Bellman equation has a solution and there exists an optimal stationary strategy. In this sense our paper generalises the approach of Blackwell [4], who considered expected discounted payoff in the infinite time horizon for Markov decision processes with the time additive functional.

The paper is organised as follows. Section 2 describes the model and risk sensitive preferences of the decision maker, and provides essential assumptions. In Section 3, we use the dynamic programming approach to show that the agent has an optimal stationary policy and the lifetime utility is a solution to the optimality (Bellman) equation. In Section 4, we solve some examples and compute both the value function and the optimal policy.

**2. The model.** Let $\mathbb{R} := (-\infty, \infty)$ and $\mathbb{N}$ be the set of all positive integers. We consider a discrete-time controlled Markov process described by the following objects.

(i) The *state space* $X$ is a standard Borel space (that is, a non-empty Borel subset of some Polish space).
(ii) The *action space* $A$ is also a standard Borel space.

(iii) $D$ is a non-empty Borel subset of $X \times A$. We assume that for each $x \in X$, the non-empty $x$-section

$$A(x) = \{a \in A : (x, a) \in D\}$$

of $D$ is compact and represents the *set of actions available* in state $x$.

(iv) $q$ is a regular *conditional distribution* from $D$ to $X$.

(v) The one-step *reward function* $r : D \to \mathbb{R}$ is a Borel measurable non-negative function such that $r(x, a) \leq d$ for $(x, a) \in D$ and some constant $d > 0$.

Then, $H_k = D^{k-1} \times X$ for $k \geq 2$, and $H_1 = X$, denotes the set of all feasible histories of the process up to stage $k$. Clearly, $H_\infty := D \times D \times \cdots$ (infinitely many times) is the set of all infinite feasible histories. As usual, a *policy* $\pi = (\pi_k)_{k \in \mathbb{N}}$ is a sequence of Borel measurable mappings (decision rules) from $H_k$ to $A$ such that $\pi_k(h_k) \in A(x_k)$, where $h_k = (x_1, a_1, x_2, a_2, \ldots, x_k) \in H_k$. The set of all policies is denoted by $\Pi$. The class of *stationary policies* is identified with the class $F$ of measurable functions $f$ from $X$ to $A$ such that $f(x) \in A(x)$. It is well-known that $F$ is non-empty if $A(x)$ is compact for each $x \in X$ (see [5, Theorem 1]).

In this section, we shall consider the *non-expected utility* of the decision maker in the infinite time horizon derived with the aid of the so-called entropic risk measure. In order to define this measure, let $(\Omega, \mathcal{F}, P)$ be a probability space and let $X \in L^\infty(\Omega, \mathcal{F}, P)$ be a random payoff. The *entropic risk measure* of $X$ is

$$\rho(X) = -\frac{1}{\gamma} \ln \int_\Omega e^{-\gamma X(\omega)} \, P(d\omega),$$

where $\gamma > 0$ is the *risk sensitive coefficient.* Let $X$, $Y$ be random variables from $L^\infty(\Omega, \mathcal{F}, P)$. The following properties of $\rho$ are of importance (see [9, p. 184]):

(P1) monotonicity: $X \leq Y \Rightarrow \rho(X) \leq \rho(Y)$,
(P2) translation invariance: $\rho(X + x) = \rho(X) + x$ for all $x \in \mathbb{R}$,
(P3) the Jensen inequality: $\rho(X) \leq \mathbb{E}X$.

However, $\rho$ is not positive homogeneous. Here, we only mention that making use of the Taylor expansions for the exponential and logarithmic functions, we can approximate $\rho(X)$ as follows:

$$\rho(X) \approx \mathbb{E}X - \frac{\gamma}{2} \operatorname{Var} X, \tag{2.1}$$

if $\gamma > 0$ is sufficiently close to 0. Therefore, if $X$ is a random payoff, then the decision maker who evaluates his expected payoff with the aid of the entropic risk measure, thinks not only of the expected value $\mathbb{E}X$ of the

random payoff $X$, but also of its variance. Further comments on the entropic risk measure can be found in [9] and the references cited therein.

For $k \in \mathbb{N}$, let $B(H_k)$ be the set of all Borel measurable bounded non-negative real-valued functions defined on $H_k$, equipped with the supremum norm $\|\cdot\|$. Let $\pi = (\pi_k)_{k\in\mathbb{N}} \in \Pi$ be any policy. For $v_{k+1} \in B(H_{k+1})$ and $h_k \in H_k$ we set

$$\rho_{\pi_k,h_k}(v_{k+1}) := -\frac{1}{\gamma} \ln \int_X e^{-\gamma v_{k+1}(h_k,\pi_k(h_k),y)}\, q(dy \,|\, x_k, \pi_k(h_k)). \tag{2.2}$$

Observe that by (P3) we have

$$\begin{aligned} 0 &\le \rho_{\pi_k,h_k}(v_{k+1}) \\ &\le \int_X v_{k+1}(h_k,\pi_k(h_k),y)\, q(dy \,|\, x_k,\pi_k(h_k)) \le \|v_{k+1}\| \end{aligned} \tag{2.3}$$

for $h_k \in H_k$. Next, we define an operator $L_{\pi_k}$ as follows:

$$(L_{\pi_k} v_{k+1})(h_k) = L_{\pi_k} v_{k+1}(h_k) := r(x_k, \pi_k(h_k)) + \beta \rho_{\pi_k,h_k}(v_{k+1}),$$

where $\beta \in (0,1)$ is a discount factor. By (P1), $L_{\pi_k}$ is monotone, i.e., $L_{\pi_k} v_{k+1}(h_k) \le L_{\pi_k} \hat{v}_{k+1}(h_k)$ for $h_k \in H_k$ and $v_{k+1} \le \hat{v}_{k+1}$. Moreover, by (2.3),

$$0 \le L_{\pi_k} v_{k+1}(h_k) \le d + \beta \|v_{k+1}\| \tag{2.4}$$

for every $h_k \in H_k$.

We follow the approach of Hansen and Sargent [10] and model the preferences of the decision maker recursively. For any state $x_1 = x$ and $N \in \mathbb{N}$ we define an $N$-stage total discounted utility by

$$J_N(x,\pi) := (L_{\pi_1} \circ \cdots \circ L_{\pi_N})\mathbf{0}(x), \tag{2.5}$$

where $\mathbf{0}$ is a function such that $\mathbf{0}(h_k) \equiv 0$ for every $h_k \in H_k$ and $k \in \mathbb{N}$. For instance, if $N = 2$, definition (2.5) reads

$$\begin{aligned} J_2(x,\pi) &= (L_{\pi_1} \circ L_{\pi_2})\mathbf{0}(x) = L_{\pi_1}(L_{\pi_2}\mathbf{0})(x) \\ &= r(x,\pi_1(x)) - \frac{\beta}{\gamma} \ln \int_X e^{-\gamma L_{\pi_2}\mathbf{0}(x,\pi_1(x),y)}\, q(dy \,|\, x, \pi_1(x)) \\ &= r(x,\pi_1(x)) - \frac{\beta}{\gamma} \ln \int_X e^{-\gamma r(y,\pi_2(x,\pi_1(x)),y))}\, q(dy \,|\, x, \pi_1(x)). \end{aligned}$$

Observe that (P1) implies that the sequence $(J_N(x,\pi))_{N\in\mathbb{N}}$ is non-decreasing and bounded from below by 0 for all $x \in X$ and $\pi \in \Pi$. Moreover,

$$J_N(x,\pi) \le \frac{d}{1-\beta} \quad \text{for } x \in X,\ \pi \in \Pi,\ N \in \mathbb{N}.$$

This follows easily from (P1) and (P3). Hence, $\lim_{N\to\infty} J_N(x,\pi)$ exists for all $x \in X$ and $\pi \in \Pi$.

PROBLEM 2.1. *For an initial income $x \in X$ and a policy $\pi \in \Pi$ we define the non-expected discounted payoff in the infinite time horizon as follows:*

$$J(x,\pi) := \lim_{N\to\infty} J_N(x,\pi). \tag{2.6}$$

*The aim of the decision maker is to find the value function of this payoff, i.e.,*

$$J^*(x) = \sup_{\pi\in\Pi} J(x,\pi), \qquad x \in X,$$

*and a policy $\pi^* \in \Pi$ such that*

$$J(x,\pi^*) = J^*(x) \quad \text{for all } x \in X.$$

REMARK 2.2. When $\gamma \to 0^+$, the non-expected payoff (utility) in (2.6) tends to the von Neumann–Morgenstern expected payoff that was first studied by Blackwell [4] and, for instance, by Stockey et al. [18] in economic models. The greater $\gamma > 0$, the more risk averse is the agent.

**3. The optimality equation.** In order to solve the aforementioned problem we shall use the dynamic programming approach.

We start by formulating our compactness and semicontinuity assumptions, which will be used alternatively.

*Condition* (**S**):

(i) The set $A(x)$ is compact.
(ii) For each $x \in X$ and every Borel set $C \subset X$, the function $q(C \,|\, x, \cdot)$ is continuous on $A(x)$.
(iii) The reward function $r(x,\cdot)$ is upper semicontinuous on $A(x)$ for each $x \in X$.

*Condition* (**W**):

(i) The set $A(x)$ is compact and the set-valued mapping $x \mapsto A(x)$ is upper semicontinuous, that is, $\{x \in X : A(x) \cap B \neq \emptyset\}$ is closed for every closed set $B$ in $A$.
(ii) The transition law $q$ is weakly continuous on $D$, that is,

$$(x,a) \mapsto \int_X u(y)\, q(dy \,|\, x,a)$$

is continuous for each continuous bounded function $u$.
(iii) The reward function $r$ is upper semicontinuous on $D$.

We denote by $U(X)$ $[B(X)]$ the set of all bounded non-negative upper semicontinuous [bounded non-negative] Borel measurable functions on $X$.

THEOREM 3.1. *Assume* (**W**) [(**S**)]. *Then:*

(a) *There exist a unique function* $V \in U(X)$ [$V \in B(X)$] *and a decision rule* $f^* \in F$ *such that*

$$V(x) = \sup_{a\in A(x)} \left( r(x,a) - \frac{\beta}{\gamma} \ln \int_X e^{-\gamma V(y)}\, q(dy \,|\, x,a) \right) \tag{3.1}$$

$$= r(x, f^*(x)) - \frac{\beta}{\gamma} \ln \int_X e^{-\gamma V(y)}\, q(dy \,|\, x, f^*(x)) \tag{3.2}$$

*for all* $x \in X$.

(b) *Moreover,* $V(x) = J^*(x) = J(x, f^*)$ *for all* $x \in X$, *i.e.,* $f^*$ *is an optimal stationary policy.*

*Proof.* Let us first assume (**W**) . For any $v \in U(X)$, define an operator $L$ as follows:

$$Lv(x) := \sup_{x\in A(x)} \left( r(x,a) - \frac{\beta}{\gamma} \ln \int_X e^{-\gamma v(y)}\, q(dy \,|\, x,a) \right) \tag{3.3}$$

for all $x \in X$.

Part (a) follows from the Banach fixed point theorem applied to $L$. Note that $L : U(X) \to U(X)$. Indeed, this follows from [11, Theorem 3.3.5] and our assumption (**W**). Thus, it only remains to prove that $L$ is contractive when we endow $U(X)$ with the supremum norm $\|\cdot\|$. Assume that $v_1, v_2 \in U(X)$. Then

$$\begin{aligned}
&Lv_1(x) - Lv_2(x) \\
&\le \sup_{a\in A(x)} \left( -\frac{\beta}{\gamma} \ln \int_X e^{-\gamma v_1(y)}\, q(dy \,|\, x,a) + \frac{\beta}{\gamma} \ln \int_X e^{-\gamma v_2(y)}\, q(dy \,|\, x,a) \right) \\
&\le \frac{\beta}{\gamma} \sup_{a\in A(x)} \left( -\ln \int_X e^{-\gamma\|v_1 - v_2\| - \gamma v_2(y)}\, q(dy \,|\, x,a) + \ln \int_X e^{-\gamma v_2(y)}\, q(dy \,|\, x,a) \right) \\
&\le \frac{\beta}{\gamma} \gamma \|v_1 - v_2\| = \beta \|v_1 - v_2\|,
\end{aligned}$$

and consequently, by changing the roles of $v_1$ and $v_2$,

$$\|Lv_1 - Lv_2\| \le \beta \|v_1 - v_2\|. \tag{3.4}$$

Since $U(X)$ equipped with the supremum norm is complete, by the Banach fixed point theorem there exists a unique $V \in U(X)$ satisfying (3.1). The existence of $f^* \in F$ in (3.2) follows from the measurable selection theorem (see [5, Corollary 1] or [11, Proposition D.5]).

Now we prove part (b). From (3.1) we obtain

$$V(x) \ge r(x,a) - \frac{\beta}{\gamma} \ln \int_X e^{-\gamma V(y)}\, q(dy \,|\, x,a)$$

for every $(x,a) \in D$. Let $\pi = (\pi_k)_{k\in\mathbb{N}} \in \Pi$ be any policy. Then, for every history $h_k \in H_k$, $k \in \mathbb{N}$, the above display implies that

$$V(x_k) \geq L_{\pi_k} V(h_k). \tag{3.5}$$

Fix any $T \in \mathbb{N}$. Set $k := T$ in (3.5), i.e., $V(x_T) \geq L_{\pi_T} V(h_T)$. Then, applying (3.5) consecutively for $k = T-1, \ldots, 1$ we obtain

$$V(x) \geq (L_{\pi_1} \circ \cdots \circ L_{\pi_T}) V(x).$$

Since $V \geq 0$ and $L_{\pi_k}$ is monotone for every $\pi_k$, $k \in \mathbb{N}$, we obtain

$$V(x) \geq (L_{\pi_1} \circ \cdots \circ L_{\pi_T}) V(x) \geq (L_{\pi_1} \circ \cdots \circ L_{\pi_T}) \mathbf{0}(x) = J_T(x,\pi) \tag{3.6}$$

for any $\pi \in \Pi$ and $x \in X$. Letting $T \to \infty$ in (3.6), we conclude that $V(x) \geq J(x,\pi)$ for any $\pi \in \Pi$ and $x \in X$. Hence,

$$V(x) \geq \sup_{\pi\in\Pi} J(x,\pi), \qquad x \in X. \tag{3.7}$$

Let $f^* \in F$ be as in (3.2). We define

$$r_{f^*}(x) = r(x, f^*(x)), \qquad \rho_{f^*,x}(V) = -\frac{1}{\gamma} \ln \int_X e^{-\gamma V(y)}\, q(dy \,|\, x, f^*(x))$$

and

$$L_{f^*} V(x) = r(x, f^*(x)) + \beta \rho_{f^*,x}(V)$$

for $x \in X$. Hence, (3.2) can be written as $L_{f^*} V = V$. By iterating this equality $T-1$ times we get

$$V(x) = L_{f^*}^{(T)} V(x), \qquad x \in X, \tag{3.8}$$

where $L_{f^*}^{(T)}$ denotes the $T$-fold composition of the operator $L_{f^*}$ with itself. From (3.8), making use of properties (P1) and (P3) we obtain

$$\begin{aligned}
V(x) &= L_{f^*}^{(T-1)}(r_{f^*} + \beta \rho_{f^*,.}(V))(x) \\
&\leq L_{f^*}^{(T-1)}(r_{f^*} + \beta \|V\|)(x) \\
&= L_{f^*}^{(T-2)}\big(r_{f^*} + \beta \rho_{f^*,.}(r_{f^*} + \beta \|V\|)\big)(x) \\
&= L_{f^*}^{(T-2)}\big(r_{f^*} + \beta \rho_{f^*,.}(r_{f^*}) + \beta^2 \|V\|\big)(x) \qquad \text{by (P2)} \\
&= L_{f^*}^{(T-3)}\big(r_{f^*} + \beta \rho_{f^*,.}(r_{f^*} + \beta \rho_{f^*,.}(r_{f^*}) + \beta^2 \|V\|)\big)(x) \\
&= L_{f^*}^{(T-3)}\big(r_{f^*} + \beta \rho_{f^*,.}(r_{f^*} + \beta \rho_{f^*,.}(r_{f^*})) + \beta^3 \|V\|\big)(x)
\end{aligned} \tag{3.9}$$

for $x \in X$; the last equality follows again from (P2). Continuing, we deduce

$$V(x) \leq J_T(x, f^*) + \beta^T \|V\|, \qquad x \in X. \tag{3.10}$$

Letting $T \to \infty$ in (3.10) we obtain

$$V(x) \le J(x, f^*), \quad x \in X. \tag{3.11}$$

Now, (3.7) and (3.11) combined yield (b).

The proof under assumption (**S**) proceeds along similar lines. We have to show that $L : B(X) \to B(X)$, which follows from (**S**) and [11, Theorem 3.3.5]. Further, we prove that $L$ is contractive when $B(X)$ is endowed with the supremum norm (see (3.4)). Since $B(X)$ is complete, we can apply the Banach fixed point theorem to $L$. The existence of a Borel measurable selection follows from [5, Corollary 1]. Part (b) is the same as in the previous case. ∎

Remark 3.2. For simplicity we have assumed that the reward function is positive. Clearly, Theorem 3.1 holds true for the model with bounded rewards (not necessarily positive), say $\tilde{r}$. Then the solution to equation (3.1) is

$$\tilde{V}(x) := V(x) + \frac{c}{1-\beta}, \quad x \in X,$$

where $c > 0$ is a constant such that $\tilde{r}(x, a) + c \ge 0$ for all $(x, a) \in D$.

Remark 3.3. It is worth mentioning that Markov decision processes with risk measures that help to evaluate the discounted payoff (utility) in the infinite time horizon have already been studied in the literature. There are two ways of dealing with this problem. The first is similar to the one studied above, i.e., at every stage the decision maker uses some measure or certainty equivalent to evaluate the payoff or utility in the next step. Such an approach leads to a stationary optimal policy and was considered for *coherent* risk measures in [16, 17]. However, our entropic risk measure is not positive homogeneous, and thus not coherent. In the second approach, the decision maker applies a risk measure to the aggregated discounted reward in the infinite time horizon, i.e., he/she is interested in the value $E_x^\pi \exp\{-\gamma \sum_{t=1}^{\infty} r(x_t, a_t)\}$, where $E_x^\pi$ is the expectation with respect to the unique measure $P_x^\pi$, defined on the space of all trajectories of a Markov process starting at $x \in X$ and governed by policy $\pi \in \Pi$ and transition probability $q$. However, in this context the decision maker may not have an optimal stationary policy [3, 6, 7].

Remark 3.4. Note that if $\gamma > 0$ is close to zero, then we may view the Bellman equation (3.2) through formula (2.1). Namely, the maximised payoff, in the infinite time horizon is the sum of the today payoff and the expected value and the variance, multiplied by $-\gamma/2$, of the infinite time horizon payoff calculated with respect to an uncertain tomorrow state.

**4. Examples.** In this section, we present two examples of Markov decision processes with transition probabilities independent of a state. Such models are known in the literature under the name of "invariant models" [2]. According to Remark 3.2 we may consider bounded rewards. Below, we describe how to solve the optimality equation (3.1) for invariant dynamic programming problems with a *finite* action set (cf. [2] for a discounted cost criterion in a risk-neutral setting). A similar algorithm, however, does not work for Markov control processes with a finite action set and general state space.

Let $A(x) = A := \{a_0, a_1, \dots, a_n\}$ and set $q_i(\cdot) := q(\cdot \,|\, a_i)$ and

$$w(x) = e^{-\gamma V(x)} \quad \text{for } x \in X,$$
$$g(x,a) = e^{-\gamma r(x,a)} \quad \text{for } (x,a) \in D := X \times A.$$

Then (3.1) can be written in the form

$$w(x) = \min_{a \in A} \left[ g(x,a) \left( \int_X w(y)\, q(dy \,|\, a) \right)^{\beta} \right], \qquad x \in X. \tag{4.1}$$

Furthermore, we define

$$g_i(x) := g(x, a_i), \qquad W_i = \int_X w(y)\, q_i(dy)$$

for $i = 0, \dots, n$ and $x \in X$, and we set

$$X_i := \{x \in X : g_i(x) W_i^{\beta} < g_j(x) W_j^{\beta} \text{ for } j = 1, \dots, i-1, \text{ and } g_i(x) W_i^{\beta} \le g_j(x) W_j^{\beta} \text{ for } j = i+1, \dots, n\}$$

for $i = 0, \dots, n$. In other words, $X_i$ is the set of states for which $i$ is the least index such that the action $a_i$ minimises the right-hand side of (4.1), or equivalently maximises the right-hand side of (3.1). Therefore, (4.1) reduces to

$$w(x) = \sum_{i=0}^{n} g_i(x) W_i^{\beta} 1_{X_i}(x),$$

where $1_{X_i}$ is the indicator of $X_i$. Integrating this inequality with respect to $q_j$ we obtain

$$W_j = \sum_{i=0}^{n} G_{ij} W_i^{\beta}, \qquad j = 0, \dots, n, \tag{4.2}$$

where

$$G_{ij} := \int_{X_i} g_i(x)\, q_j(dx).$$

Hence, (4.2) represents a system of $n+1$ non-linear equations with unknowns $W_j$, $j = 1, \dots, n$. Observe that in this setting the assumptions (**W**)(i, ii) or

(**S**)(i, ii) automatically hold. Therefore, if additionally (**W**)(iii) or (**S**)(iii) is satisfied, then by Theorem 3.1 there exists a solution to (4.2). Abusing the notation a little, let us denote it by the same symbols, i.e., $W_j$, $j = 1, \dots, n$. Then

$$V(x) = -\frac{1}{\gamma} \ln\Big(\sum_{i=0}^{n} g_i(x) W_i^{\beta} 1_{X_i}(x)\Big),$$

and the optimal policy is

$$f^*(x) = \begin{cases} a_0, & x \in X_0, \\ \vdots & \\ a_n, & x \in X_n. \end{cases}$$

Example 4.1 (Maintenance problem; see also [2, Example 4.2] or [12, Example 2]). We consider a complex system composed of a central unit and a large number of small units. The system is checked regularly each year, and its current situation is recorded. Assume that the central unit is half of the system and can be either in working condition, recorded as 0, or in non-working condition, recorded as 0.5. The collection of small units is treated as a continuum and their situation is recorded as a number between 0 (perfect condition) and 0.5 (all broken down). The recorded state of the system is the sum of these two numbers, and is thus a number between 0 and 1. The following actions may be chosen.

- *Minor repair* $a_0$. The repair results in a probability of 0.5 that the central unit is in working condition, and a uniform distribution on the state of small units. The state of the system after such a repair is thus uniformly distributed over $[0, 1]$.
- *Overall repair* $a_1$. The repair results in a probability of 0.5 that the central unit is in working condition and that the small units are in perfect condition.
- *Replacement* $a_2$. The system is replaced by a new one. The trade-in value of the used system affects the price of a new one.

Although it is natural in this context to think about costs rather than payoffs, we consider the equivalent maximum problem. The payoff functions equal $r_0(x) = 4 - x$, $r_1(x) = 5 - 3x$, $r_2(x) = 5 - 5x$ for $x \in X := [0, 1]$, and the transition probabilities are $q_0 \sim U(X)$, $q_1(0) = q_1(0.5) = 0.5$ and $q_2(0) = 1$. Let $\gamma = 1$ and $K_0 = W_0/W_1$, $K_2 = W_0/W_2$. If we assume that $\beta = 0.5$, then it can be proved that $0 \in X_2$, $0.5 \in X_1$ and $1 \in X_0$, where

$$X_0 = \big\{x \geq \ln \sqrt{K_0^{\beta} e}\big\}, \qquad X_2 = \big\{x < \ln \sqrt{K_2^{\beta}}\big\},$$
$$X_1 = \big\{\ln \sqrt{K_2^{\beta}} \leq x < \ln \sqrt{K_0^{\beta} e}\big\}.$$

Furthermore, we have

$$G_{00} = e^{-4}(e - (K_0^\beta e)^{1/2}), \quad G_{10} = \tfrac{1}{3}e^{-5}((K_0^\beta e)^{3/2} - K_2^{3\beta/2}),$$
$$G_{20} = \tfrac{1}{5}e^{-5}(K_2^{5\beta/2} - 1), \quad G_{01} = 0, \quad G_{11} = \tfrac{1}{2}e^{-7/2},$$
$$G_{21} = \tfrac{1}{2}e^{-5}, \quad G_{02} = G_{12} = 0, \, G_{22} = e^{-5}.$$

Making use of (4.2), we obtain

$$W_0 = 0.00053384, \quad W_1 = 0.0002715, \quad W_2 = 0.0000454.$$

Thus, the optimal policy is

$$f^*(x) = \begin{cases} a_0, & x \in [0.669034; 1], \\ a_1, & x \in [0.447113; 0.669034), \\ a_2, & x \in [0; 0.447113), \end{cases}$$

and the solution to the Bellman equation is

$$V(x) = \begin{cases} -7.767707 + x, & x \in X_0, \\ -9.105774 + 3x, & x \in X_1, \\ -10 + 5x, & x \in X_2. \end{cases}$$

If $\beta = 0.75$, then $0.5 \in X_2$ and therefore

$$G_{00} = e^{-4}(e - (K_0^\beta e)^{1/2}), \quad G_{10} = \tfrac{1}{3}e^{-5}((K_0^\beta e)^{3/2} - K_2^{3\beta/2}),$$
$$G_{20} = \tfrac{1}{5}e^{-5}(K_2^{5\beta/2} - 1), \quad G_{01} = G_{11} = 0,$$
$$G_{21} = \tfrac{1}{2}e^{-5} + \tfrac{1}{2}e^{-5/2}, \quad G_{02} = G_{12} = 0, \quad G_{22} = e^{-5}.$$

The system (4.2) yields

$$W_0 = 4.67043 * 10^{-8}, \quad W_1 = 1.358557 * 10^{-8},$$
$$W_2 = 2.061154 * 10^{-9}.$$

Hence, the optimal policy is

$$f_{0.75} = \begin{cases} a_0, & x \in [0.963061; 1], \\ a_1, & x \in [0.7071533; 0.963061), \\ a_2, & x \in [0; 0.7071533), \end{cases}$$

and the solution to the Bellman equation is

$$V(x) = \begin{cases} -16.6596 + x, & x \in X_0, \\ -18.5857 + 3x, & x \in X_1, \\ -20 + 5x, & x \in X_2. \end{cases}$$

Example 4.2 (see [12, Example 3]). We consider a system that is checked regularly each month, and its current situation is recorded as a number from the interval $X := [0, 1]$. The number 0 means that the system is completely broken down, whereas 1 means that it is in perfect condition.

Each month a controller decides to call one of the two available mechanics to service the system. Let $A := \{a_0, a_1\}$, where $a_0$ means calling the first mechanic, and $a_1$ calling the other one. Mechanic I wishes to be paid $2(1-x)$ and after his service the distribution function of the next state is $q_0(dx) := \frac{e^x}{e-1}dx$, which results in reaching 'worst states' with greater probability. The cost of mechanic II's service is independent of the current state and equals 1. The distribution function of the next state after his service is $q_1(dx) := \frac{e^{1-x}}{e-1}dx$, which means that greater probability is assigned to the 'better states'. We can again view this problem as an equivalent maximum problem with the payoffs obtained from costs by the change of sign, i.e., $r_0(x) = 2(x-1)$ and $r_1(x) = -1$. We obtain

$$X_0 = \left\{x \in X : x \geq \frac{1+\beta\ln(W_0/W_1)}{2} =: m\right\}, \quad X_1 := X \setminus X_0.$$

Computing

$$G_{00} = \frac{e^2}{e-1}(e^{-m} - e^{-1}), \qquad G_{10} = \frac{e}{e-1}(e^m - 1),$$
$$G_{01} = \frac{e^3}{3e-3}(e^{-3m} - e^{-3}), \quad G_{11} = \frac{e^2}{e-1}(1 - e^{-m})$$

and inserting these into (4.2) we obtain the following solution for $\beta = 0.5$:

$$m = 0.459996, \quad W_0 = 4.55251, \quad W_1 = 5.34251.$$

Hence,

$$X_0 = [0.459996; 1], \quad X_1 = [0; 0.459996).$$

Consequently, the function solving (3.1) is

$$V(x) = \begin{cases} 2.757839 - 2x, & x \in X_0, \\ 1.837848, & x \in X_1 \end{cases}$$

and the optimal policy is $f^*(x) = a_0$ for $x \in X_0$ and $f^*(x) = a_1$ for $x \in X_1$. On the other hand, for $\beta = 0.75$ the solutions are

$$m = 0.4361036, \quad W_0 = 22.6295, \quad W_1 = 26.8334,$$

which yields

$$V(x) = \begin{cases} 4.339441 - 2x, & x \in X_0, \\ 3.467236, & x \in X_1, \end{cases}$$

with $X_0 = [0.4361036; 1]$ and $X_1 = [0; 0.4361036)$.

## References

[1] E. W. Anderson, *The dynamics of risk-sensitive allocations*, J. Econom. Theory 125 (2005), 93–150.

[2] D. Assaf, *Invariant problems in discounted dynamic programming*, Adv. Appl. Probab. 10 (1978), 472–490.
[3] N. Bäuerle and U. Rieder, *More risk-sensitive Markov decision processes*, Math. Oper. Res. 39 (2014), 105–120.
[4] D. Blackwell, *Discounted dynamic programming*, Ann. Math. Statist. 36 (1965), 226–235.
[5] L. D. Brown and R. Purves, *Measurable selections of extrema*, Ann. Statist. 1 (1973), 902–912.
[6] K. J. Chung and M. Sobel, *Discounted MDPs: distribution functions and exponential utility maximization*, SIAM J. Control Optim. 25 (1987), 49–62.
[7] G. B. Di Masi and Ł. Stettner, *Infinite horizon risk sensitive control of discrete time Markov processes with small risk*, Systems Control Lett. 40 (2000), 15–20.
[8] L. G. Epstein and S. E. Zin, *Substitution, risk aversion, and the temporal behavior of consumption and asset returns: a theoretical framework*, Econometrica 57 (1989), 937–969.
[9] H. Föllmer and A. Schied, *Stochastic Finance. An Introduction in Discrete Time*, de Gruyter, Berlin, 2004.
[10] L. P. Hansen and T. J. Sargent, *Discounted linear exponential quadratic Gaussian control*, IEEE Trans. Autom. Control 40 (1995), 968–971.
[11] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer, New York, 1996.
[12] A. Jaśkiewicz, *A note on risk-sensitive control of invariant models*, Systems Control Lett. 56 (2007), 663–668.
[13] D. M. Kreps and E. L. Porteus, *Temporal resolution of uncertainty and dynamic choice*, Econometrica 46 (2000), 185–200.
[14] J. Miao, *Economic Dynamics in Discrete Time*, MIT Press, 2014.
[15] M. Pitera and Ł. Stettner, *Long run risk sensitive portfolio with general factors*, Math. Methods Oper. Res. 83 (2016), 265–293.
[16] A. Ruszczyński, *Risk-averse dynamic programming for Markov decision processes*, Math. Program. 125 (2010), 235–261.
[17] Y. Shen, W. Stannat and K. Obermayer, *Risk-sensitive Markov control processes*, SIAM J. Control Optim. 51 (2013), 3652–3672.
[18] N. L. Stockey, R. E. Lucas and E. Prescott, *Recursive Methods in Economic Dynamics*, Harvard Univ. Press, Cambridge, MA, 1989.
[19] T. D. Tallarini, Jr., *Risk-sensitive real business cycles*, J. Monetary Econom. 45 (2000), 507–532.
[20] P. Weil, *Nonexpected utility in macroeconomics*, Quart. J. Econom. 105 (1990), 29–42.
[21] P. Weil, *Precautionary savings and the permanent income hypothesis*, Rev. Econom. Stud. 60 (1993), 367–383.

Hubert Asienkiewicz
Faculty of Mathematics,
Computer Science and Econometrics
University of Zielona Góra
Zielona Góra, Poland
E-mail: h.asienkiewicz@wmie.uz.zgora.pl

Anna Jaśkiewicz
Faculty of Pure and Applied Mathematics
Wrocław University of Science and Technology
Wrocław, Poland
E-mail: anna.jaskiewicz@pwr.edu.pl