Commun. Optim. Theory 2025 (2025) 5 https://doi.org/10.23952/cot.2025.5

Communications in Optimization Theory
Available online at http://cot.mathres.org

ON A VERIFICATION THEOREM FOR OPTIMALITY OF STRATEGIES IN MARKOV DECISION PROCESSES WITH TOTAL COST

ALEXEY PIUNOVSKIY[1,*], YI ZHANG[2]

[1]Department of Mathematical Sciences, University of Liverpool, Liverpool, UK
[2]School of Mathematics, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

Dedicated to the memory of Professor Uriel Rothblum

Abstract. The goal of this brief article is to present a verification theorem, which can be used to show the optimality of a given strategy, without computing the Bellman function in advance, for a Markov decision process (MDP) with a general-signed cost function. An example is elaborated on, which shows that one of the minor conditions in the verification theorem is really important.

Keywords. Markov decision process; Total cost; Optimal strategy.

2020 Mathematics Subject Classification. 90C40, 49L20.

1. INTRODUCTION

Markov decision processes (MDPs) is a well developed branch of applied probability, statistics and operational research with numerous applications to real-life problems. The current paper is devoted to the MDPs with a total expected cost. One can say that there are two main approaches, in some sense dual: convex analytic approach and dynamic programming (DP). As for the first one, a substantial progress was made in the recent years by Professor U. Rothblum [4].

Professor U. Rothblum was also active in developing the dynamic programming approach; see, e.g., [5]. One can find many useful statements on DP in the famous monographs by Bertsekas and Shreve [1] and by Puterman [10]. Usually, one of the main steps is calculating the Bellman function, given by the minimal possible expected cost with any given initial state. After that, a control strategy is optimal if it is so called conserving and equalizing; see, e.g. [1, Prop. 9.12] or [10, Thm. 7.1.7]. Both conditions are stated in terms of the Bellman function. There are fewer theorems which help to check the optimality of a strategy without calculating the Bellman function *a priori*. For example, Proposition 9.13 of [1] is about only negative models and discounted ones with the bounded cost function; [6, Thm. 4.2.3(c)] is about discounted MDPs with an unbounded cost function. The goal of the current paper is to prove a similar statement in the more general case: no restrictions on the sign of the cost function, and the MDP is not necessarily absorbing.

*Corresponding author.
E-mail address: piunov@liverpool.ac.uk (A. Piunovskiyr), y.zhang.29@bham.ac.uk (Y. Zhang).
Received November 9, 2023; Accepted March 14, 2024.

In Sections 2 and 3, we describe the MDP under investigation, formulate and prove our main result, Theorem 3.1. In Section 4, we elaborate on a counter-example showing that one of the conditions in that theorem cannot be omitted. All over the text, when calculating expectations, we accept the following conventions:

$$\mathsf{E}[Y] := \mathsf{E}[Y^+] + \mathsf{E}[Y^-], \text{ where } \infty - \infty := \infty \text{ and } Y^+ := \max\{Y,0\},\ Y^- := \min\{Y,0\}.$$

2. Description of the model

The primitives of an MDP are the following.

- The state and action spaces $\mathbf{X}$ and $\mathbf{A}$ are nonempty Borel spaces, endowed with their σ-algebras $\mathscr{B}(\mathbf{X})$ and $\mathscr{B}(\mathbf{A})$.
- The transition probability $p(dy|x,a)$ is a stochastic kernel from $\mathbf{X}\times\mathbf{A}$ to $\mathscr{B}(\mathbf{X})$.
- The $[-\infty,+\infty]$-valued one-step cost function $c(\cdot,\cdot)$ on $\mathbf{X}\times\mathbf{A}$.

The initial distribution P_0 on $\mathscr{B}(\mathbf{X})$ is fixed. $\Omega := (\mathbf{X}\times\mathbf{A})^\infty$ is the sample space, endowed with the product σ-algebra $\mathscr{F}$. The component-wise projections

$$\Omega \ni (x_0,a_1,x_1,\ldots) \to x_t, t=0,1,2,\ldots$$

and

$$\Omega \ni (x_0,a_1,x_1,\ldots) \to a_t, t=1,2,\ldots$$

are denoted by X_t and A_t, and $\{X_t\}_{t=0}^\infty$, $\{A_t\}_{t=1}^\infty$ are the controlled and the controlling random processes.

Definition 2.1 (Strategy)**.** A control strategy $\pi = \{\pi_n\}_{n=1}^\infty$ is a sequence of stochastic kernels such that for each $n=1,2,\ldots$, $\pi_n(da|x_0,a_1,\ldots,x_{n-1})$ is a stochastic kernel from $(\mathbf{X}\times\mathbf{A})^{n-1}\times\mathbf{X}$ to $\mathscr{B}(\mathbf{A})$, where $(\mathbf{X}\times\mathbf{A})^0\times\mathbf{X} := \mathbf{X}$. A strategy is called stationary if, for all $n=1,2,\ldots$, $\pi_n(da|x_0,a_1,\ldots,x_{n-1}) = \pi(da|x_{n-1})$ is the same stochastic kernel from $\mathbf{X}$ to $\mathscr{B}(\mathbf{A})$. If for a stationary strategy π, there is a measurable mapping φ from $\mathbf{X}$ to $\mathbf{A}$ such that $\pi(da|x) = \delta_{\varphi(x)}(da)$ for all $x\in\mathbf{X}$, where $\delta_{\varphi(x)}(da)$ is the Dirac measure concentrated on the singleton $\{\varphi(x)\}$, then π is called a deterministic stationary strategy, identified with the (measurable) selector φ.

As is well known, for each control strategy π and initial distribution P_0, there is a unique strategic measure on the sample space $(\Omega,\mathscr{F})$, denoted as $\mathsf{P}^\pi_{\mathsf{P}_0}$, which is specified by the following conditions:

$$\mathsf{P}^\pi_{\mathsf{P}_0}(X_0\in dy) = \mathsf{P}_0(dy);$$

and for each $n=1,2,\ldots$, $\Gamma^{\mathbf{X}}_i\in\mathscr{B}(\mathbf{X})$ $(i=0,1,\ldots,n)$ and $\Gamma^{\mathbf{A}}_i\in\mathscr{B}(\mathbf{A})$ $(i=1,2,\ldots,n)$,

$$\begin{aligned}&\mathsf{P}^\pi_{\mathsf{P}_0}(X_0\in\Gamma^{\mathbf{X}}_0,\ A_1\in\Gamma^{\mathbf{A}}_1,\ \ldots,\ X_{n-1}\in\Gamma^{\mathbf{X}}_{n-1},\ A_n\in\Gamma^{\mathbf{A}}_n)\\ =&\int\limits_{\Gamma^{\mathbf{X}}_0\times\Gamma^{\mathbf{A}}_1\times\cdots\times\Gamma^{\mathbf{X}}_{n-1}} \pi_n(\Gamma^{\mathbf{A}}_n|x_0,a_1,\ldots,x_{n-1})\mathsf{P}^\pi_{\mathsf{P}_0}(X_0\in dx_0,A_1\in da_1,\ldots,X_{n-1}\in dx_{n-1});\end{aligned}$$

and

$$\begin{aligned}&\mathsf{P}^\pi_{\mathsf{P}_0}(X_0\in\Gamma^{\mathbf{X}}_0,\ A_1\in\Gamma^{\mathbf{A}}_1,\ldots,\ X_n\in\Gamma^{\mathbf{X}}_n)\\ =&\int\limits_{\Gamma^{\mathbf{X}}_0\times\Gamma^{\mathbf{A}}_1\times\cdots\times\Gamma^{\mathbf{X}}_{n-1}\times\Gamma^{\mathbf{A}}_n} p(\Gamma^{\mathbf{X}}_n|x_{n-1},a_n)\\ &\times\mathsf{P}^\pi_{\mathsf{P}_0}(X_0\in dx_0,A_1\in da_1,\ldots,X_{n-1}\in dx_{n-1},A_n\in da_n).\end{aligned}$$

For details, see [3, 6, 8]. The expectation taken with respect to $\mathsf{P}^\sigma_{\mathsf{P}_0}$ is denoted by $\mathsf{E}^\sigma_{\mathsf{P}_0}$. In case $\mathsf{P}_0(dx) = \delta_{x_0}(dx)$ is a Dirac measure, we use notations $\mathsf{P}^\pi_{x_0}$ and $\mathsf{E}^\pi_{x_0}$.

The problem is about minimizing the total expected cost:

$$v^{\pi}_{\mathsf{P}_0} := \limsup_{T\to\infty} \mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\sum_{t=1}^{T} c(X_{t-1},A_t)\right] \to \inf_{\pi}. \tag{2.1}$$

Here and below,

$$\mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\sum_{t=1}^{T} c(X_{t-1},A_t)\right] := \mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\sum_{t=1}^{T} c^{+}(X_{t-1},A_t)\right] + \mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\sum_{t=1}^{T} c^{-}(X_{t-1},A_t)\right]. \tag{2.2}$$

Recall that convention $\infty-\infty:=\infty$ is in use here.

Usually, it is assumed that the expression $\infty-\infty$ never appears in (2.2), or, more specifically, when considering only Dirac measures P_0, it is assumed that

$$\inf_{\pi} \mathsf{E}^{\pi}_{x_0}\left[\sum_{t=1}^{\infty} c^{-}(X_{t-1},A_t)\right] > -\infty,\ \forall x_0\in\mathbf{X}, \tag{2.3}$$

see [7, Ass. 9.3.2]. Under this and other mild conditions, a very strong theorem characterizing the optimal strategies was proved in [7]; see Theorem 9.5.5 therein. This and similar statements in [1, 10] assume that the Bellman function

$$v^{*}_{x} := \inf_{\pi} v^{\pi}_{x},\ x\in\mathbf{X}$$

was preliminary calculated. In the next section, we formulate and prove our main result which allows to check the optimality of a strategy without calculating function v^{*}_{x}. We also compare our result with the similar previously published ones.

3. Main result

Theorem 3.1. *Suppose that $\hat{\pi}$ is a strategy such that $v^{\hat{\pi}}_{\mathsf{P}_0}$ and $v^{\hat{\pi}}_{x}$ for all $x\in\mathbf{X}$ are finite, and equation*

$$v^{\hat{\pi}}_{x} = \inf_{a\in\mathbf{A}}\left\{c(x,a)+\int_{\mathbf{X}} v^{\hat{\pi}}_{y}\,p(dy|x,a)\right\},\ x\in\mathbf{X}. \tag{3.1}$$

is satisfied. Assume that, for each strategy π, $\mathsf{E}^{\pi}_{\mathsf{P}_0}[v^{\hat{\pi}}_{X_t}]$ is finite for all $t=0,1,\ldots$ and

$$\liminf_{T\to\infty} \mathsf{E}^{\pi}_{\mathsf{P}_0}[v^{\hat{\pi}}_{X_T}] \le 0. \tag{3.2}$$

Then $\hat{\pi}$ is optimal for the initial distribution P_0.

Proof. The target is to show that, for any strategy π, $v^{\pi}_{\mathsf{P}_0}\ge v^{\hat{\pi}}_{\mathsf{P}_0}$. This inequality is obvious in the case $v^{\pi}_{\mathsf{P}_0}=+\infty$, so below we assume that $v^{\pi}_{\mathsf{P}_0}<\infty$.

Since

$$\mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\int_{\mathbf{X}} (v^{\hat{\pi}}_{y})^{-}p(dy|X_t,A_{t+1})\right] = \mathsf{E}^{\pi}_{\mathsf{P}_0}[(v^{\hat{\pi}}_{X_{t+1}})^{-}] > -\infty,\ t=0,2,\ldots,$$

$$\mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\int_{\mathbf{X}} v^{\hat{\pi}}_{y}\,p(dy|X_t,A_{t+1})\right] = \mathsf{E}^{\pi}_{\mathsf{P}_0}[v^{\hat{\pi}}_{X_{t+1}}] \in \mathbb{R}$$

for each $t=0,1,\ldots,T-1$.

Now equation (3.1) implies that, for each $T=1,2,\ldots$,

$$\mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\sum_{t=1}^{T}\left\{c(X_{t-1},A_t)-v^{\hat{\pi}}_{X_{t-1}}+\int_{\mathbf{X}} v^{\hat{\pi}}_{y}\,p(dy|X_{t-1},A_t)\right\}\right]\ge 0. \tag{3.3}$$

The left-hand side of (3.3) takes the form of the sum of expectations:

$$\begin{aligned}
&\mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\sum_{t=1}^{T} c(X_{t-1},A_t)\right] - \mathsf{E}^{\pi}_{\mathsf{P}_0}[v^{\hat\pi}_{X_0}] + \mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\int_{\mathbf{X}} v^{\hat\pi}_y p(dy|X_0,A_1)\right] \\
&-\mathsf{E}^{\pi}_{\mathsf{P}_0}[v^{\hat\pi}_{X_1}] + \mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\int_{\mathbf{X}} v^{\hat\pi}_y p(dy|X_1,A_2)\right] - \ldots + \mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\int_{\mathbf{X}} v^{\hat\pi}_y p(dy|X_{T-1},A_T)\right] \\
= \;& \mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\sum_{t=1}^{T} c(X_{t-1},A_t)\right] - v^{\hat\pi}_{\mathsf{P}_0} + \mathsf{E}^{\pi}_{\mathsf{P}_0}[v^{\hat\pi}_{X_T}] \ge 0. \qquad (3.4)
\end{aligned}$$

For each $T \ge 1$, $\mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\sum_{t=1}^{T} c(X_{t-1},A_t)\right]$ cannot equal $-\infty$ in view of the last inequality, where $v^{\hat\pi}_{\mathsf{P}_0}$ and $\mathsf{E}^{\pi}_{\mathsf{P}_0}[v^{\hat\pi}_{X_T}]$ are finite. Hence, in view of (2.2), for each $T \ge 1$, $\mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\sum_{t=1}^{T} c^-(X_{t-1},A_t)\right] > -\infty$. Also, for each $T \ge 1$, $\mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\sum_{t=1}^{T} c(X_{t-1},A_t)\right]$ is different from ∞ because we assumed that $v^{\pi}_{\mathsf{P}_0} < \infty$. In greater detail, if, for some $N \ge 1$, $\mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\sum_{t=1}^{N} c(X_{t-1},A_t)\right] = \infty$, then, in view of (2.2), $\mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\sum_{t=1}^{N} c^+(X_{t-1},A_t)\right] = \infty$, so that for all $T \ge N$, $\mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\sum_{t=1}^{T} c^+(X_{t-1},A_t)\right] = \infty$ and, again by (2.2) and the previous observation, $\mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\sum_{t=1}^{T} c(X_{t-1},A_t)\right] = \infty$. This would contradict the assumption of $v^{\pi}_{\mathsf{P}_0} < \infty$. Thus, all terms in (3.4) are finite. After passing to the limit, we conclude that

$$\begin{aligned}
v^{\pi}_{\mathsf{P}_0} = \limsup_{T\to\infty} \mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\sum_{t=1}^{T} c(X_{t-1},A_t)\right] \;&\ge\; v^{\hat\pi}_{\mathsf{P}_0} + \limsup_{T\to\infty}\{-\mathsf{E}^{\pi}_{\mathsf{P}_0}[v^{\hat\pi}_{X_T}]\} \\
&=\; v^{\hat\pi}_{\mathsf{P}_0} - \liminf_{T\to\infty} \mathsf{E}^{\pi}_{\mathsf{P}_0}[v^{\hat\pi}_{X_T}] \ge v^{\hat\pi}_{\mathsf{P}_0}.
\end{aligned}$$

□

Before we compare Theorem 3.1 with other existing results and provide a corollary, let us recall several definitions.

Definition 3.2. (a) An MDP is called absorbing at a specific (cemetery) state $\Delta \in \mathbf{X}$ under the initial distribution P_0 if $p(\{\Delta\}|\Delta,a) \equiv 1$, $c(\Delta,a) \equiv 0$ and $\mathsf{E}^{\pi}_{\mathsf{P}_0}[T_0] < \infty$ for all strategies π, where $T_0 := \inf\{t \ge 0 :\ X_t = \Delta\}$ is the time to absorption, having accepted that the infimum over the empty set equals ∞.

(b) An MDP, which is absorbing at Δ under each initial distribution P_0, is called discounted if $p(\{\Delta\}|x,a) \equiv 1-\beta$ for all $x \ne \Delta$, where $\beta \in (0,1)$ is the "discount factor". (Note that the requirement $\mathsf{E}^{\pi}_{\mathsf{P}_0}[T_0] < \infty$ is satisfied automatically for all P_0 as soon as $p(\{\Delta\}|x,a) \equiv 1-\beta$ with $\beta \in (0,1)$.)

If the model is negative (i.e., $c(\cdot,\cdot) \le 0$) or discounted with a bounded cost function $c(\cdot,\cdot)$, then the following version of Theorem 3.1 is known; see e.g., [1, Prop. 9.13]: a stationary strategy $\hat\pi$ is optimal (for all initial states $x \in \mathbf{X}$) if and only if equation (3.1) holds.

For the discounted model with an unbounded cost function $c(\cdot,\cdot)$, under some additional conditions (e.g., $c(x,\cdot) \ge 0$ is inf-compact on $\mathbf{A}$ for each $x \in \mathbf{X}$, etc) the following version of Theorem 3.1 is known in e.g., [6, Thm. 4.2.3(c)]: if $\lim_{T\to\infty} \mathsf{E}^{\pi}_{x}[v^{\hat\pi}_{X_T}] = 0$ for all $x \in \mathbf{X}$ and all strategies π such that $v^{\pi}_{x} < \infty$, then $\hat\pi$ is optimal (for all initial states $x \in \mathbf{X}$) if and only if equation (3.1) holds.

In the general model, under some additional conditions (e.g., inequality (2.3) etc), if $\limsup_{T\to\infty} \mathsf{E}^{\pi}_{x}[v^{\hat\pi}_{X_T}] = 0$ for all $x \in \mathbf{X}$ and all strategies π, then $\hat\pi$ is optimal (for all initial states $x \in \mathbf{X}$) if and only if equation (3.1) holds: see [7, Thm. 9.5.13].

Corollary 3.3 (from Theorem 3.1)**.** *Suppose* $\sup_{(x,a)\in\mathbf{X}\times\mathbf{A}} |c(x,a)| \le K < \infty$ *and the model is absorbing at* Δ *under each initial state* $x \in \mathbf{X}$.

(a) *If the model is absorbing at* Δ *also under the initial distribution* P_0 *and equations (3.1) and (3.2) are satisfied, then the strategy* $\hat{\pi}$ *is optimal for* P_0.
(b) *If equation (3.1) is satisfied and (3.2) holds for* $\mathsf{P}_0(dx) = \delta_{x_0}(dx)$ *for all* $x_0 \in \mathbf{X}$*, then the strategy* $\hat{\pi}$ *is optimal for all initial states* $x \in \mathbf{X}$.

Remark 3.4. The requirement that the model is absorbing at Δ also under the initial distribution P_0 is independent, because there are examples when absorbing under each initial state $x \in \mathbf{X}$ model is not absorbing under another initial distribution P_0. See, e.g., Example 2.2.2 or 2.2.21 in [9].

Proof of Corollary 3.3. (a) It is known that, for the absorbing model, $\mathsf{E}^{\pi}_{\mathsf{P}_0}[T_0] \le M < \infty$ for all strategies π, where the constant M can depend on P_0: see [4, p.132]. Therefore, for all strategies π, $\left|\mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\sum_{t=1}^{\infty} c^{\pm}(X_{t-1},A_t)\right]\right| \le KM$ and $\mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\sum_{t=1}^{\infty} c(X_{t-1},A_t)\right]$ is finite and coincides with $v^{\pi}_{\mathsf{P}_0}$ (see (2.1)) for all strategies π.

The same reasoning holds for $\mathsf{P}_0(dx) = \delta_{x_0}(dx)$, for each $x_0 \in \mathbf{X}$. Thus, $v^{\hat{\pi}}_{\mathsf{P}_0}$ and $v^{\hat{\pi}}_x$ for all $x \in \mathbf{X}$ are finite.

Now, for each strategy π,

$$\mathsf{E}^{\pi}_{\mathsf{P}_0}\left[v^{\hat{\pi}}_{X_t}\right] = \mathsf{E}^{\pi}_{\mathsf{P}_0}\left[\mathsf{E}^{\hat{\pi}}_{X_t}\left[\sum_{n=1}^{\infty} c(X_{n-1},A_n)\right]\right] = \mathsf{E}^{\tilde{\pi}}_{\mathsf{P}_0}\left[\sum_{\tau=t+1}^{\infty} c(X_{\tau-1},A_{\tau})\right],$$

where $\tilde{\pi}$ is the following combination of the strategies π and $\hat{\pi}$:

$$\tilde{\pi}_n(da|x_0,a_1,\dots,x_{n-1}) = \begin{cases} \pi_n(da|x_0,a_1,\dots,x_{n-1}), & \text{if } n \le t; \\ \hat{\pi}_{n-t}(da|x_t,a_{t+1},\dots,x_{n-1}), & \text{if } n > t. \end{cases}$$

Thus, again $\left|\mathsf{E}^{\pi}_{\mathsf{P}_0}\left[v^{\hat{\pi}}_{X_t}\right]\right| \le KM$ is finite, and assertion (a) follows from Theorem 3.1.

(b) After we put $\mathsf{P}_0(dx) = \delta_{x_0}(dx)$ for an arbitrarily fixed $x_0 \in \mathbf{X}$, the strategy $\hat{\pi}$ is optimal for the initial state x_0 according to (a). □

4. Counter-example

Example 3.1.4 of [2] presents a non-optimal strategy $\hat{\pi}$ in a discounted positive model (when $c(\cdot,\cdot) \ge 0$) satisfying equation (3.1), for which $v^{\hat{\pi}}_x = \infty$ and $\lim_{T\to\infty} \mathsf{E}^{\pi}_x\left[v^{\hat{\pi}}_{X_T}\right] = \infty$ for some states x and some strategies π. Below, we elaborate on another illustrative example, based on [9, Subsec. 3.2.3], where all functions are finite, which confirms that requirement (3.2) in Theorem 3.1 is important. It also confirms that one cannot omit the condition $\sup_{(x,a)\in\mathbf{X}\times\mathbf{A}} |c(x,a)| < \infty$ in Proposition 9.13 of [1].

The MDP we plan to consider is discounted, with discrete state and actions spaces. In this case, the model is absorbing at Δ under each initial distribution and, traditionally, the notations are slightly modified. First of all, with some abuse of notation, we write $p(y|x,a)$ for stochastic kernels, rather than $p(\{y\}|x,a)$. Secondly, the cemetery Δ is omitted and the transition probability on $\mathbf{X} \not\ni \Delta$ is normalized:

$$p(y|x,a) = \beta q(y|x,a) \ \text{ for } y \in \mathbf{X};$$

The initial distribution P_0 is on $\mathscr{B}(\mathbf{X})$.

Now equations (3.1) and (3.2) take the form

$$v^{\hat{\pi}}_x = \inf_{a\in\mathbf{A}} \left\{ c(x,a) + \beta \int_{\mathbf{X}} v^{\hat{\pi}}_y q(dy|x,a) \right\}, \ x \in \mathbf{X}. \tag{4.1}$$

and

$$\liminf_{T\to\infty} \beta^T \mathsf{E}^{\pi}_{\mathsf{P}_0}[v^{\hat{\pi}}_{X_T}] \le 0: \tag{4.2}$$

in the original notations, $(1-\beta^T)$ is the probability that $X_T=\Delta$ leading to $v^{\hat\pi}_{X_T}=0$.

Let $\mathbf{X}=\{0,1,2\ldots\}$, $\mathbf{A}=\{1,2\}$, $q(0|0,a)\equiv 1$, $c(0,a)\equiv 0$. For $x>0$ we put $q(x+1|x,1)=q(0|x,1)\equiv 1/2$, $q(x+1|x,2)\equiv 1$, other transition probabilities $q(\cdot)$ being zero; $c(x,1)=2^x$, $c(x,2)\equiv 1$. The discount factor is $\beta=1/2$. (See Fig. 1, where the probabilities q are shown.)

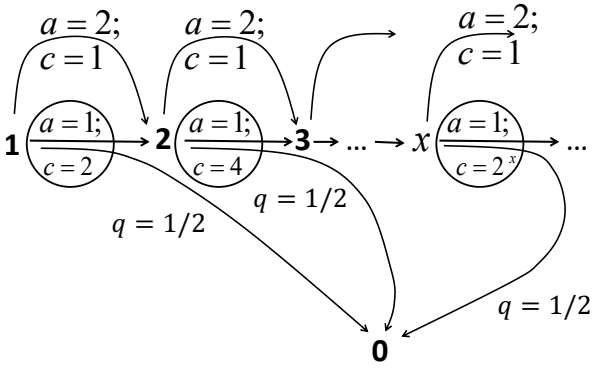

FIGURE 1. The selector $\hat\varphi(x)\equiv 1$ is not optimal.

The optimality equation

$$v(x)=\inf_{a\in\mathbf{A}}\left\{c(x,a)+\beta\int_{\mathbf{X}}v(y)q(dy|x,a)\right\},\ x\in\mathbf{X} \tag{4.3}$$

takes the form

$$v(0)=\frac{1}{2}v(0);$$

$$\text{for } x>0,\qquad v(x)=\min\{2^x+\frac{1}{4}v(x+1)+\frac{1}{4}v(0);\ 1+\frac{1}{2}v(x+1)\},$$

and has the minimal non-negative solution

$$v(x)=v^*_x\equiv 2\ \text{ for } x>0;\ v(0)=0,$$

coincident with the Bellman function by [10, Thm. 7.3.2]. The stationary selector $\varphi^*(x)\equiv 2$ is conserving and equalizing and hence optimal for each initial state $x\in\mathbf{X}$: see [1, Prop. 9.12].

Now consider the stationary selector $\hat\varphi(x)\equiv 1$. The performance functional $v^{\hat\varphi}_x$ is given by $v^{\hat\varphi}_x=2^{x+1}$ (for $x>0$) and satisfies equation (4.1):

$$1+\frac{1}{2}v^{\hat\varphi}_{x+1}=1+2^{x+1}>2^{x+1}=v^{\hat\varphi}_x=2^x+\frac{1}{4}v^{\hat\varphi}_{x+1}.$$

Equation (4.2) is violated for selector φ^*:

$$\beta^T\mathsf{E}^{\varphi^*}_x\left[v^{\hat\varphi}_{X_T}\right]=\beta^T\ 2^{x+T+1}=2^{x+1},$$

and the selector $\hat\varphi$ is certainly non-optimal for each initial distribution P_0, except $\mathsf{P}_0(dx)=\delta_0(dx)$. If we put $\mathsf{P}_0(dx)=\delta_{x_0}(dx)$ for an arbitrarily fixed $x_0>0$, we see that all the conditions in Theorem 3.1 are satisfied for $\hat\varphi$ apart from equation (3.2). In particular, for each strategy π, $\mathsf{E}^{\pi}_x\left[v^{\hat\varphi}_{X_t}\right]\le\mathsf{E}^{\varphi^*}_x\left[v^{\hat\varphi}_{X_t}\right]=2^{x+t+1}$ is finite: the probability to reach state $X_T\ne 0$ is the highest for φ^*. By the way,

$$\beta^T\mathsf{E}^{\hat\varphi}_x\left[v^{\hat\varphi}_{X_T}\right]=2^{x+1-T}\to 0\ \text{ as } T\to\infty.$$

We see that the both functions $v_x^* = v_x^{\varphi^*}$ and $v_x^{\hat{\varphi}}$ solve the optimality equation (4.3), but only v_x^* is the minimal non-negative solution. Note that the optimality equation (4.3) has many other solutions, e.g. $v(x) = 2 + k \cdot 2^x$ with $k \in [0, 1/2]$.

5. Conclusion

Theorem 3.1 makes it possible to prove optimality of a strategy without knowing the Bellman function. Similar statements in the existing literature are known under additional restrictive conditions. At the same time, Theorem 3.1 can also be used for calculating the Bellman function. Indeed, if that theorem is applicable for each degenerate initial distribution $\mathsf{P}_0(dy) = \delta_x(dy)$, $x \in \mathbf{X}$ (the Dirac measure), then

$$\mathsf{E}_x^{\hat{\pi}} \left[\sum_{t=1}^{\infty} c(X_{t-1}, A_t) \right] = v_x^{\hat{\pi}} = v_x^*$$

equals the Bellman function. In this connection, let us remind that the Bellman equation

$$v_x^* = \inf_{a \in \mathbf{A}} \left\{ c(x,a) + \int_{\mathbf{X}} v_y^* \, p(dy|x,a) \right\}, \; x \in \mathbf{X}$$

in an infinite-horizon model can have many solutions even in finite models [3, Ch. 4, Section 7], [9, Subsection 2.2.3], [10, Ex. 7.2.3 and 7.3.1], see also the above counter-example. Sometimes it is hard to choose the Bellman function among those solutions.

References

[1] D. Bertsekas, S. Shreve, Stochastic Optimal Control, Academic Press, New York, 1978.

[2] D. Bertsekas, Dynamic Programming and Optimal Control, V.II, Athena Scientific, Belmont, MA, USA, 2001.

[3] E.B. Dynkin, A.A. Yushkevich, Controlled Markov Processes, Springer, New York, 1979.

[4] E.A. Feinberg, U. Rothblum, Splitting randomized stationary policies in total-reward Markov decision processes, Math. Oper. Res. 37 (2012) 129-153.

[5] P.G. Canbolat, U.G. Rothblum, (Approximate) iterated successive approximations algorithm for sequential decision processes, Ann. Oper. Res. 208 (2013) 309-320.

[6] O. Hernández-Lerma, J. Lasserre, Discrete-Time Markov Control Processes, Springer Verlag, New York, 1996.

[7] Hernández-Lerma, J. Lasserre, Further Topics in Discrete-Time Markov Control Processes, Springer Verlag, New York, 1999.

[8] A. Piunovskiy, Optimal Control of Random Sequences in Problems with Constraints, Kluwer, Dordrecht, 1997.

[9] A.B. Piunovskiy, Examples in Markov Decision Processes, Imperial College Press, London, 2013.

[10] M.L. Puterman, Markov Decision Processes, Wiley, New York, 1994.