

# When Bellman's Principle Fails

A.B. Piunovskiy\*

Department of Mathematical Sciences, M & O Building, Peach Street, University of Liverpool, Liverpool, L69 3BX, UK

**Abstract:** We present several examples which show that the well known statements about Markov Decision Processes can fail if the loss function is not bounded.

**Keywords:** Markov decision processes, dynamic programming.

## 1. INTRODUCTION

The Dynamic Programming approach elaborated 50 years ago remains the most powerful method for solving optimal control problems. For example, the whole issue N.3 of journal "Control and Cybernetics" was devoted to it in 2006. There exist many excellent monographs on stochastic optimal control with discrete time, e.g. [1-6]. Such models are traditionally called Markov Decision Processes. Dynamic Programming can be adjusted even to constrained optimal control [7]. On the other side, one should be very accurate when dealing with infinities. Several authors formulate and prove their statements without mentioning that their results are valid only if expressions of the type "+∞" and "-∞" do not appear. Even the excellent monograph [3] contains such typos (see e.g. Theorem 3.2.1). Of course, specialists understand those difficulties, but recent article [4] shows that the questions under study are interesting for applied researchers. The goal of the present work is to indicate clearly what can and what cannot happen if the loss functions are unbounded (Lemma 1), and to present several academic counterexamples illustrating that many common statements can fail to hold if "+∞" and "-∞" meet together.

In Sections 2 and 3 we give general ideas about Markov Decision Processes and Dynamic Programming. Although we mainly consider the models with at most countable state and action spaces, Lemmas 1 and 2 and Corollaries 1 and 2 are formulated for general Borel models. In the main Section 4, examples are provided which show

- that Markov and non-randomized strategies are not sufficient for solving optimal control problems;
- that the Bellman's principle can fail, i.e. the final (or the starting) part of an optimal trajectory can be not optimal;
- that a uniformly optimal strategy can also be not optimal;
- that a solution to the optimality equation can provide no boundaries to the performance functional.

Example 4 is the discussion of possible conventions about mathematical expectation and adding together the infinities. It turns out that standard conventions are most natural although they can also lead to inconveniences. The proofs are collected in Appendix.

As far as possible, we use bold letters for spaces and capital letters for random variables.  $I\{\cdot\}$  is the indicator function;  $\mathbb{R} = \{-\infty, +\infty\}$ ,  $\mathbb{R}^* = [-\infty, +\infty]$ .

## 2. MODEL DESCRIPTION

Let us consider the Markov Decision Process  $\{\mathbf{X}, \mathbf{A}, T, p, r, R\}$  with the finite time horizon  $T$  and total expected loss. Here  $\mathbf{X}$  and  $\mathbf{A}$  are (Borel) state and action spaces,  $p_t(dx_t | x_{t-1}, a_t)$  is the transition probability,  $r_t(x_{t-1}, a_t)$  and  $R(x_T)$  are the real-valued loss functions.

As usual, a control strategy (policy)  $\pi$  is a sequence of measurable stochastic kernels  $\pi_t(da_t | h_{t-1})$  on  $\mathbf{A}$ , where  $h_{t-1} = (x_0, a_1, x_1, \dots, a_{t-1}, x_{t-1})$  is a history. A strategy is called Markov if it has the form  $\pi_t(da_t | h_{t-1}) = \pi_t^m(da_t | x_{t-1})$ . In case  $\forall t \pi_t(da | h)$  is concentrated at a single point  $\varphi_t(h)$ , the strategy is called a selector (non-randomized strategy). Markov selector has the form  $\varphi_t(x_{t-1})$ .

Suppose the initial distribution  $P_0(dx)$  is fixed. If a control strategy  $\pi$  is fixed, too, then there exists a unique probability measure  $P_\pi^\pi$  on the space of trajectories

$$\mathbf{H} = \{(x_0, a_1, x_1, \dots, a_T, x_T)\}$$

defined in the usual way:

$$\begin{aligned} P_\pi^\pi \{d(x_0, a_1, x_1, \dots, a_T, x_T)\} \\ = P_0(dx_0) \pi_1(da_1 | x_0) p_1(dx_1 | x_0, a_1) \\ \times \pi_2(da_2 | x_0, a_1, x_1) \dots p_T(dx_T | x_{T-1}, a_T) \end{aligned} \quad (1)$$

The integral wrt the measure  $P_\pi^\pi$  is denoted by  $E_{P_\pi^\pi}^\pi$ . For each  $h \in \mathbf{H}$ , the (realized) total loss equals

$$w(h) = \sum_{t=1}^T r_t(x_{t-1}, a_t) + R(x_T).$$

It is convenient to say that  $\mathbf{H}$  is the sample space and consider the trivial projections  $h \rightarrow x_t$ ,  $h \rightarrow a_t$  and other

\*Address correspondence to this author at the Department of Mathematical Sciences, M & O Building, Peach Street, University of Liverpool, Liverpool, L69 3BX, UK. Tel: +44-(0)151-794-4737; E-mail: piunov@liverpool.ac.uk

functions on  $\mathbf{H}$  as random elements. Thus,  $w(h)$  defines the random total loss  $W$ . In what follows, random variables are denoted with capital letters; small letters are used to show arguments of functions. If all the spaces are countable, no measurability problems arise, and we always have in mind the collections of all subsets of such spaces as  $\sigma$ -algebras.

The performance of control strategy  $\pi$  is given by

$$v^\pi = E_{P_0}^\pi [W] = E_{P_0}^\pi [W^+] + E_{P_0}^\pi [W^-], \quad (2)$$

where  $W^+ = \max\{0, W\}$ ,  $W^- = \min\{0, W\}$  and  
 $" + \infty " + " - \infty " = " + \infty "$ . (3)

The aim is to solve problem

$$v^\pi = E_{P_0}^\pi \left[ \sum_{t=1}^T r_t(X_{t-1}, A_t) + R(X_T) \right] \rightarrow \inf_{\pi} \quad (4)$$

i.e. to construct an optimal control strategy.

### 3. DYNAMIC PROGRAMMING APPROACH

The Bellman principle leads to the following equation

$$\begin{cases} v_T(x) = R(x); \\ v_{t-1}(x) = \inf_{a \in \mathbf{A}} \left\{ r_t(x, a) + \int_{\mathbf{X}} v_t(y) p_t(dy | x, a) \right\} \pi_t^*(da | H_{t-1}) \end{cases} \quad (5)$$

called "optimality/Bellman" equation. Its solution  $v_t(x)$  is called Bellman function. (Note, it can take values  $\pm\infty$ ).

Suppose loss functions  $r(\cdot)$  and  $R(\cdot)$  are simultaneously bounded below or above. Then a control strategy  $\pi^*$  is optimal in problem (4) if and only if for all  $t = 1, 2, \dots, T$

$$v_{t-1}(X_{t-1}) = \int_{\mathbf{A}} \left\{ r_t(X_{t-1}, a) + \int_{\mathbf{X}} v_t(y) p_t(dy | X_{t-1}, a) \right\} \pi_t^*(da | H_{t-1}) \quad P_{P_0}^{\pi^*} - a.s. \quad (6)$$

(Here  $H_{t-2} = (X_0, A_1, X_1, \dots, A_{t-1}, X_{t-1})$  is a random history).

$$v^{\pi^*} = \inf_{\pi} v^\pi = \int_{\mathbf{X}} v_0(x) P_0(dx). \quad (7)$$

Suppose a history  $h_t \in \mathbf{H}_t$ ,  $0 \leq \tau \leq T$  is fixed. Then we can consider the controlling process  $A_t$  and the controlled process  $X_t$  as developing on the time interval  $\{\tau + 1, \tau + 2, \dots, T\}$  which is empty if  $\tau = T$ . If a control strategy  $\pi$  (in the initial model) is fixed then one can build the strategic measure on  $\mathbf{H}$ , denoted as  $P_{h_\tau}^\pi$ , similarly to (1), satisfying the "initial condition"  $P_{h_\tau}^\pi(h_\tau \times (\mathbf{A} \times \mathbf{X})^{T-\tau}) = 1$ . The most important case is  $\tau = 0$ ; then we have just  $P_{x_0}^\pi$ . Note that  $P_{x_0}^\pi$

is another denotation for  $P_{P_0}^\pi$  in case  $P_0(\cdot)$  is concentrated at

point  $x_0$ . We introduce  $v_{h_\tau}^\pi = E_{h_\tau}^\pi \left[ \sum_{t=\tau+1}^T r_t(X_{t-1}, A_t) + R(X_T) \right]$

and call a control strategy  $\pi^*$  uniformly optimal if

$$v_{h_\tau}^{\pi^*} = \inf_{\pi} v_{h_\tau}^\pi \text{ for all } h_\tau \in \bigcup_{t=0}^T \mathbf{H}_t.$$

The dynamic programming approach leads to the following statement: if the loss functions are bounded (below or above), then a control strategy  $\pi^*$  is uniformly optimal if and only if equality

$$v_{t-1}(x_{t-1}) = \int_{\mathbf{A}} \left\{ r_t(x_{t-1}, a) + \int_{\mathbf{X}} v_t(y) p_t(dy | x_{t-1}, a) \right\} \pi_t^*(da | h_{t-1}) \quad (8)$$

holds for all  $t = 1, 2, \dots, T$  and  $h_{t-1} \in \mathbf{H}_{t-1}$ . In this case,

$$v_{h_\tau}^{\pi^*} = v_\tau(x_\tau). \quad (9)$$

Note that we use the capital letters for random variables in (4) and (6), so that mathematical expectation and expression  $P_{P_0}^{\pi^*} - a.s.$  make sense. On the opposite, formulae (5), (7), (8), (9) represent equations and statements regarding functions  $r, v$  and so on, for all values of arguments; thus we use small letters for the arguments here.

Very often, the infimum in (5) is provided by a mapping  $a = \varphi_t(x)$ , so that Markov selectors form a sufficient class for solving problem (4). Another general observation: usually, a uniformly optimal strategy is also optimal, but not vice versa. More detailed description of the Dynamic Programming approach is presented in [1-6].

If loss functions  $r(\cdot)$  and  $R(\cdot)$  are not bounded (neither below nor above), the situation becomes more complicated. The following lemma can be helpful.

**Lemma 1.** For any control strategy  $\pi$ ,  $\forall h_t = (x_0, a_1, \dots, x_t) \in \mathbf{H}_t$ ,  $t = 0, 1, \dots, T$ , inequality  $v_{h_t}^\pi \geq v_t(x_t)$  is valid.

In case strategy  $\pi^*$  satisfies equality (8) and  $v_{h_t}^{\pi^*} < +\infty$

for all  $h_t \in \mathbf{H}_t$ ,  $t = 0, 1, \dots, T$ , we have equality

$$v_{h_t}^{\pi^*} \equiv v_t(x_t) = \inf_{\pi} v_{h_t}^\pi,$$

so that  $\pi^*$  is uniformly optimal.

**Corollary 1.**

$$\forall \pi \quad v^\pi \geq \int_{\mathbf{X}} v_0(x_0) P_0(dx_0),$$

so that  $\pi^*$  is optimal if  $v^{\pi^*} = \int_{\mathbf{X}} v_0(x_0) P_0(dx_0)$ .

**Corollary 2.** If a strategy  $\pi^*$  satisfies equality (8),  $v^{\pi^*} < +\infty$ , and  $v_{h_t}^{\pi^*} < +\infty$  for all  $h_t \in \mathbf{H}_t$ ,  $t = 0, 1, \dots, T$ , then control strategy  $\pi^*$  is optimal and uniformly optimal.

Even if equality (6) or (8) holds, it can happen that strategy  $\pi^*$  is not (uniformly) optimal. The lemma presented

and corollaries provide sufficient conditions of optimality. On the other hand, a control strategy can be optimal even if equalities (6) and (8) are violated.

### 4. COUNTER EXAMPLES

Further, spaces  $\mathbf{X}$  and  $\mathbf{A}$  are countable (or finite).

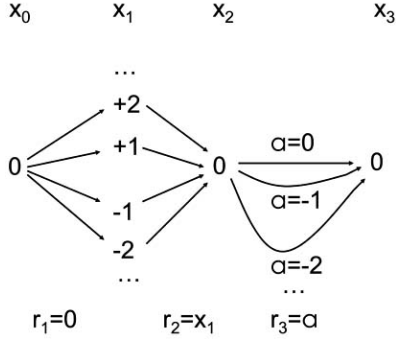
**Example 1.** Only a non-Markov randomized strategy can satisfy equalities (6) and (8) and be optimal and uniformly optimal.

Let  $\mathbf{X} = \{0, \pm 1, \pm 2, \dots\}$ ,  $\mathbf{A} = \{0, -1, -2, \dots\}$ ,  $T = 3$ ,  $P_0(0) = 1$ .

$$p_1(y|x, a) = \begin{cases} \frac{3}{|y|^2 \pi^2}, & \text{if } y \neq 0; \\ 0, & \text{if } y = 0, \end{cases} \quad p_2(0|x, a) = p_3(0|x, a) \equiv 1,$$

$r_1(x, a) \equiv 0$ ,  $r_2(x, a) = x$ ,  $r_3(x, a) = a$ ,  $R(x) = 0$ . (See Fig. (1)).

Since actions  $A_1$  and  $A_2$  play no role, we shall consider only  $A_3$ .



**Fig. (1).** Example 1: only a non-Markov randomized strategy can satisfy equalities (6) and (8) and be optimal and uniformly optimal.

The dynamic programming approach results in the following

$$v_3(x) = 0, \quad v_2(x) = -\infty, \quad v_1(x) = -\infty, \quad v_0(x) = -\infty.$$

Consider Markov control strategy  $\pi^*$  with  $\pi_3^*(0|x_2) = 0$ ,

$$\pi_3^*(a|x_2) = \frac{6}{|a|^2 \pi^2} \quad \text{for } a < 0. \quad \text{Here equalities (8) hold}$$

because

$$\sum_{i=1}^{\infty} \frac{(-i) \times 6}{i^2 \pi^2} = -\infty = v_2(x), \quad x + v_2(0) = -\infty = v_1(x),$$

$$0 + \sum_{|y|=1}^{\infty} \frac{3}{|y|^2 \pi^2} \cdot " - \infty " = -\infty = v_0(x).$$

On the other side, for any Markov strategy  $\pi^m$ ,  $v^{\pi^m} = +\infty$ . Indeed, let  $\hat{a} = \max\{j : \pi_3^m(j|0) > 0\}$ ;  $0 \geq \hat{a} > -\infty$  and consider random variable  $W^+ = (X_1 + A_3)^+$ .

It takes values 1, 2, 3, ... with probabilities not smaller than

$$p_1(-\hat{a} + 1|0, a) \pi_3^m(\hat{a}|0) = \frac{3\pi_3^m(\hat{a}|0)}{|-\hat{a} + 1|^2 \pi^2},$$

$$p_1(-\hat{a} + 2|0, a) \pi_3^m(\hat{a}|0) = \frac{3\pi_3^m(\hat{a}|0)}{|-\hat{a} + 2|^2 \pi^2},$$

$$p_1(-\hat{a} + 3|0, a) \pi_3^m(\hat{a}|0) = \frac{3\pi_3^m(\hat{a}|0)}{|-\hat{a} + 3|^2 \pi^2},$$

...

(The expressions come from trajectories  $(x_0 = 0, x_1 = -\hat{a} + i, a_1, x_2 = 0, a_2 = \hat{a}, x_3 = 0)$ ). That means

$$E_{P_0}^{\pi^m} [W^+] \geq \pi_3^m(\hat{a}|0) \sum_{i=1}^{\infty} \frac{3i}{|-\hat{a} + i|^2 \pi^2} = +\infty$$

and  $v^{\pi^m} = E_{P_0}^{\pi^m} [W] = +\infty$ . In particular  $v^{\pi^*} = +\infty$ .

At the same time, there exist optimal non-Markov strategies providing  $v^{\pi} = -\infty$ . For example put

$$a_3 = \varphi_3(x_1) = \begin{cases} -x_1, & \text{if } x_1 > 0; \\ 0, & \text{if } x_1 < 0. \end{cases} \quad (10)$$

Then  $W = X_1 + A_3 = X_1^- \leq 0$  and  $E_{P_0}^{\varphi} [W] = -\infty$ . Note that  $x_0 = 0$ ; so  $\inf_{\pi} v_{x_0}^{\pi} = \inf_{\pi} v^{\pi} = -\infty$  meaning that no-one Markov control strategy (including  $\pi^*$ ) can be optimal or uniformly optimal.

Optimal control strategy  $\varphi$  presented satisfies neither equalities (6), nor (8). Indeed,  $v_2(0) = -\infty$ , and, for example, for history  $\hat{h}_2 = (0, a_1, 1, a_2, 0)$  having positive  $P_{P_0}^{\varphi}$ -probability, on the righthand side of (6) and (8) we have

$$r_3(x_2 = 0, a_3 = \varphi_3(1)) + 0 = \varphi_3(1) = -1.$$

Since for this history  $v_{\hat{h}_2}^{\varphi} = -1$  and  $\inf_{\pi} v_{\hat{h}_2}^{\pi} = -\infty$ , optimal control strategy  $\varphi$  is not uniformly optimal. This reasoning is correct for an arbitrary selector, so that non-randomized strategies cannot satisfy equalities (6) and (8) and cannot be uniformly optimal.

Therefore, only a non-Markov randomized strategy can satisfy equalities (6) and (8) and be optimal and uniformly optimal. As an example, take

$$\pi_3(j|x_1) = \begin{cases} \frac{6}{(x_1 + j - 1)^2 \pi^2}, & \text{if } j \leq -x_1 \text{ and } x_1 > 0; \\ \frac{6}{j^2 \pi^2}, & \text{if } j < 0 \text{ and } x_1 < 0; \\ 0 & \text{otherwise.} \end{cases}$$

In the model investigated, for every optimal control strategy  $\pi$  we have  $v_{x_0}^{\pi} = v_0(x_0)$ . It can happen that this statement is false. Consider the following modification of the MDP studied:

$$\mathbf{A} = \{0\}, \quad p_3(y|x, a) = \begin{cases} \frac{6}{|y|^2 \pi^2}, & \text{if } y < 0; \\ 0 & \text{otherwise,} \end{cases} \quad r_3(x, a) = 0, \quad R(x) = x.$$

(See Fig. (2)).

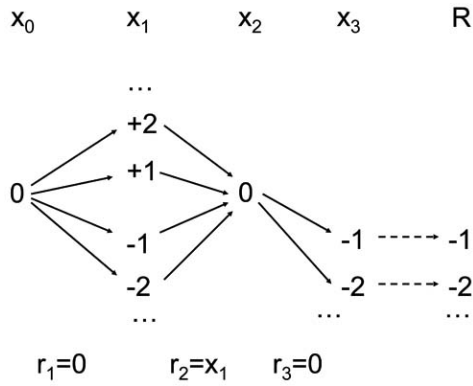


Fig. (2). Example 1:  $v_0(x_0) = -\infty < +\infty = v_{x_0}^\varphi = \inf_{\pi} v_{x_0}^\pi$ .

Actually the process is not controlled and can be interpreted as the previous MDP under a fixed Markov control strategy with distribution  $\pi_3(\cdot|x) = p_3(\cdot|x, a)$ . We know that the total expected loss here equals  $+\infty$ . Thus, in this modified model for the optimal control strategy (which is unique:  $\varphi_t(x) \equiv 0$ ) we have  $v_{x_0}^\varphi = +\infty$ . At the same time, optimality equation (5) still give  $s v_2(x) = -\infty$ ,  $v_1(x) = -\infty$ , and  $v_0(x_0) = -\infty$ .

Another similar example illustrating that  $v_0(x_0) = -\infty$  and  $\inf_{\pi} v_{x_0}^\pi = +\infty$  at some  $x_0$  is presented in [1], 3.2, ex.3.

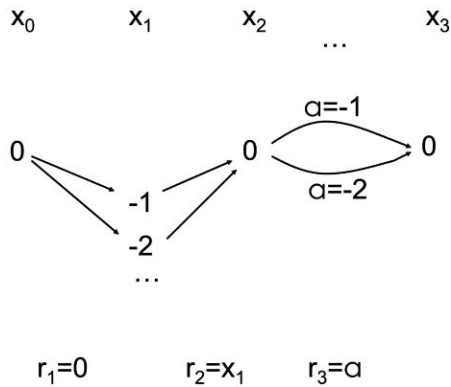


Fig. (3). Example 2: The Bellman principle fails.

**Example 2.** The Bellman principle fails.

Consider the negative modification of Example 1, Fig. (1), i.e. put  $\mathbf{X} = \{0, -1, -2, \dots\}$ ,  $\mathbf{A} = \{-1, -2\}$ ,  $p_1(y|x, a) = \frac{6}{|y|^2 \pi^2}$  (see Fig. (3)). One can easily show that selector  $\varphi_3(x_2) \equiv -1$  gives  $v^\varphi = -\infty$  and so is optimal (like any other control strategy). But this selector does not satisfy equalities (6) and (8) The point is that the Bellman principle is violated: action  $a_3 = -1$  is definitely not optimal at state  $x_2 = 0$ , and nevertheless it is optimal for the whole process on time horizon  $t = 0, 1, 2, 3$ . The very big negative loss  $r_2(x_1, a_2) = x_1$  on the second step improves the performance up to  $-\infty$ .

**Example 3.** Uniformly optimal, but not optimal strategy.

We can slightly modify Example 1 Fig. (1): ignore the initial step and put

$$P_0(x) = \begin{cases} \frac{3}{x^2 \pi^2}, & \text{if } x \neq 0; \\ 0 & \text{otherwise.} \end{cases}$$

The numbers of time moments decrease by 1 and  $T = 2$ . We still have that, for any Markov strategy  $\pi^m$ ,  $v^{\pi^m} = +\infty$ , so that all of them are not optimal. Simultaneously, now the non-optimal strategy  $\pi^*$  is uniformly optimal. In the example below, function  $v_t(x)$  is finite.

Let  $\mathbf{X} = \{\pm 1, \pm 2, \dots\}$ ,  $\mathbf{A} = \{0, 1\}$ ,  $T = 1$ ,

$$P_0(x) = \begin{cases} \frac{6}{|x|^2 \pi^2}, & \text{if } x > 0; \\ 0 & \text{otherwise,} \end{cases}$$

$$p_1(y|x, 1) = I\{y = -x\}, \quad p_1(y|x, 0) = \begin{cases} 1/4, & \text{if } y = 2x; \\ 3/4, & \text{if } y = -2x; \\ 0 & \text{otherwise.} \end{cases}$$

$r_1(x, a) = x$ ,  $R(x) = x$ . (See Fig. (4))  $I\{\cdot\}$  stands for the indicator function.

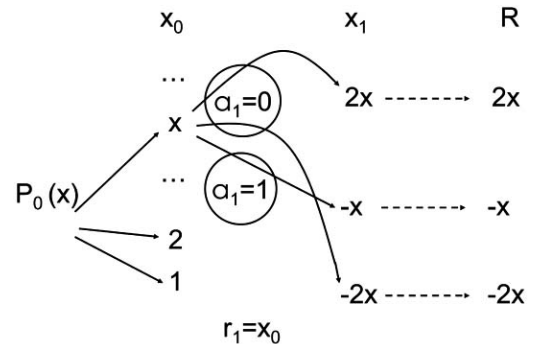


Fig. (4). Example 3.

The dynamic programming approach results in the following:  $v_1(x) = x$ ,  $v_0(x) = 0$ , and the both actions provide the equality in equation (5).

Consider action 1:  $\varphi^1(x) = 1$ . This control strategy  $\varphi^1$  is uniformly optimal since

$$v_{x_0}^{\varphi^1} = 0 = v_0(x_0) = \inf_{\pi} v_{x_0}^\pi.$$

It is also optimal because only trajectories  $(x_0, a_1 = 1, x_1 = -x_0)$  are realized, for which  $W = X_0 - X_0 = 0$ , so that  $v^{\varphi^1} = 0$ .

Consider now action 0:  $\varphi^0(x) = 0$ . This control strategy  $\varphi^0$  is also uniformly optimal since

$$v_{x_0}^{\varphi^0} = x_0 + \frac{1}{4}(2x_0) + \frac{3}{4}(-2x_0) = 0 = v_0(x_0) = \inf_{\pi} v_{x_0}^{\pi}.$$

It also satisfies equality (8). But it is not optimal because

$$\begin{aligned} E_{P_0}^{\varphi^0}[W^+] &= \sum_{i=1}^{\infty} 3i \cdot 1/4 \cdot \frac{6}{i^2 \pi^2} = +\infty, \quad E_{P_0}^{\varphi^0}[W^-] \\ &= \sum_{i=1}^{\infty} (-i) \cdot 3/4 \cdot \frac{6}{i^2 \pi^2} = -\infty, \end{aligned}$$

so that  $v^{\varphi^0} = +\infty > v^{\varphi^1} = 0$ .

This example shows that condition  $v^{\pi^*} < +\infty$  in Corollary 2 is important.

**Example 4.** Conventions about expectation and infinities.

Some authors (rather seldom, actually) suggest the following formula to calculate the performance criterion [8], 4.1:

$$\begin{aligned} v^{\pi} &= E_{P_0}^{\pi} \left[ \sum_{t=1}^T r_t^+(X_{t-1}, A_t) + R^+(X_T) \right] \\ &+ E_{P_0}^{\pi} \left[ \sum_{t=1}^T r_t^-(X_{t-1}, A_t) + R^-(X_T) \right] \end{aligned} \quad (11)$$

still accepting the rule " $+\infty + -\infty = +\infty$ ". (We adjusted the model of maximizing rewards studied in [8] to our basic case of minimizing the losses.) Clearly, in this situation the value of  $v^{\pi}$  can only increase meaning that most of the statements in Examples 1 and 3 still hold. On the other side, in the basic model presented on Fig. (1), any control strategy gives  $v^{\pi} = +\infty$  simply because

$$E_{P_0}^{\pi} [r_2^+(X_1, A_2)] = E_{P_0}^{\pi} [X_1^+] = +\infty.$$

(The same happens to Example 3). Thus, any control strategy can be called optimal! But it seems intuitively clear that selector  $\varphi$  given in (10) is better than many other strategies because it compensates positive values of  $X_1$ . (Similarly, it is natural to call optimal the selector  $\varphi^1$  in Example 3.)

If we accept (11) then it is easy to elaborate an example where optimality equation (5) has a finite solution and nevertheless only a control strategy for which criterion (8) is violated, is optimal.

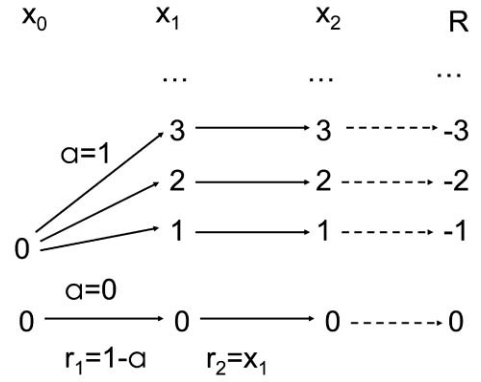
Put  $\mathbf{X} = \{0, 1, 2, \dots\}$ ,  $\mathbf{A} = \{0, 1\}$ ,  $T = 2$ ,  $P_0(0) = 1$ ,

$$p_1(y|x, 0) = I\{y=0\}, \quad p_1(y|x, 1) = \begin{cases} \frac{6}{y^2 \pi^2}, & \text{if } y > 0; \\ 0, & \text{if } y = 0, \end{cases}$$

$$p_2(y|x, a) = I\{y=x\},$$

$$r_1(x, a) = 1 - a, \quad r_2(x, a) = x, \quad R(x) = -x.$$

Since action  $A_2$  plays no role, we shall consider only  $A_1$ . (See Fig. (5)).



**Fig. (5).** Example 4:  $v_1(x)$  is finite, but  $\inf_{\pi} v_{x_0}^{\pi} = v_{x_0}^{\varphi^*} = 1 > v_0(x_0) = 0$ .

The dynamic programming approach results in the following:

$$\begin{aligned} v_2(x) &= R(x) = -x, \quad v_1(x) = x - x = 0, \\ v_0(x) &= \min\{1 + 0, 0 + 0\} = 0, \end{aligned}$$

and action  $a_1 = 1$  provides this minimum. At the same time, for control strategy  $\varphi_1(x_0) = 1$  we have

$$E_{P_0}^{\varphi} [r_2^+(X_1, A_2)] = E_{P_0}^{\varphi} [X_1] = +\infty,$$

so that (11) gives  $v^{\varphi} = +\infty$ . Hence control strategy  $\varphi_1^*(x_0) = 0$  resulting in  $v^{\varphi^*} = 1$  must be called optimal. On the opposite,  $\varphi_1(x_0) = 1$  is optimal if we accept formula

$$v^{\pi} = E_{P_0}^{\pi} [W^+] + E_{P_0}^{\pi} [W^-], \quad (12)$$

where  $W = \sum_{t=1}^T r_t(X_{t-1}, A_t) + R(X_T)$  is the total realized loss. The big loss  $X_1$  on the second step is totally compensated by the final (negative) loss  $R$ .

Now let us discuss the possible conventions about infinity. Basically, if in (12) expression " $+\infty + -\infty$ " appears then random variable  $W$  is said to be not integrable. We have seen in Examples 1, 2, and 3 that convention

$$"+\infty + -\infty = +\infty" \quad (13)$$

leads to the violation of the Bellman principle and to other problems. One can show that all those principal difficulties appear also if we put " $+\infty + -\infty = -\infty$ ". But convention (13) is still better.

Assume for a moment that " $+\infty + -\infty = -\infty$ ". Then in Example 1 Fig. (1), any Markov strategy  $\pi^m$  provides  $v^{\pi} = -\infty$ , so that all of them are equally optimal, as well as all the other control strategies. But again selector  $\varphi$  given by (10) seems better, and we want this to be mathematically confirmed. In a nutshell, if we meet " $+\infty + -\infty$ " in (12), it is better to say that all such strategies are equally bad than to accept that they are equally good.

Lemma 1 and Corollary 1 provided the lower boundary for the performance functional. That will be not the case if " $+\infty + -\infty = -\infty$ ", as the following example shows. (Compare with Example 3).

Let  $\mathbf{X} = \{0, \pm 1, \pm 2, \dots\}$ ,  $\mathbf{A} = \{0, 1\}$ ,  $T = 1$ ,

$$P_0(x) = \begin{cases} \frac{6}{x^2 \pi^2}, & \text{if } x > 0; \\ 0 & \text{otherwise,} \end{cases}$$

$$p_1(y|x, 0) = \begin{cases} 1/4, & \text{if } y = 2x; \\ 3/4, & \text{if } y = -2x \\ 0 & \text{otherwise,} \end{cases} \quad p_1(y|x, 1) = I\{y = -x\},$$

$$r_1(x, a) = x + a, \quad R(x) = x.$$

(See Fig. (6)).

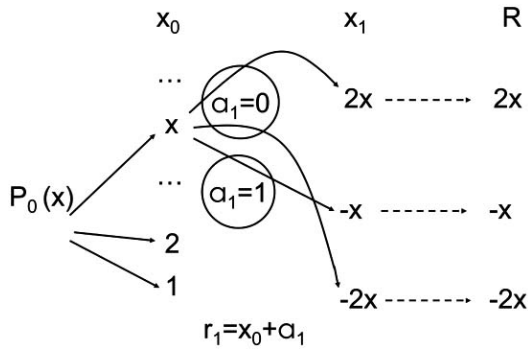


Fig. (6). Example 4: no boundaries for  $v^\pi$ .

The dynamic programming approach results in the following:

$$v_1(x) = x, \quad v_0(x) = 0,$$

and  $a_1 = 0$  provides the minimum in formula

$$v_0(x_0) = \min_a \{x_0 + a - x_0\} = 0.$$

At the same time, for control strategy  $\varphi_1^0(x_1) = 0$ , we have for  $W = X_0 + X_1$

$$E_{P_0}^{\varphi_0^0} [W^+] = \sum_{i=1}^{\infty} 3i \cdot 1/4 \cdot \frac{6}{i^2 \pi^2} = +\infty,$$

$$E_{P_0}^{\varphi_0^0} [W^-] = \sum_{i=1}^{\infty} (-i) \cdot 3/4 \cdot \frac{6}{i^2 \pi^2} = -\infty,$$

so that  $v^{\varphi_0^0} = -\infty < v_0(x_0) = 0$ . By the way, for selector  $\varphi_1^1(x_1) = 1$  we have

$$W = X_0 + 1 + X_1 = X_0 + 1 - X_0 = 1,$$

and  $v^{\varphi_1^1} = 1 > v_0(x_0) = 0$ . Thus, solution to optimality equation (5) provides no boundaries to the performance functional.

## 5. CONCLUSION

When using the Dynamic Programming approach one must be careful about infinities. Actually, almost all the pathological situations in the presented examples appear only because we meet expressions " $+\infty + -\infty$ " when calculating expectations. That is why people impose the following conditions: for every strategy  $\pi$ ,  $\forall x_0, \forall t$

$$E_{x_0}^\pi [r_t^+(x_{t-1}, a_t)] < +\infty \text{ and } E_{x_0}^\pi [R^+(x_T)] < +\infty$$

or

$$E_{x_0}^\pi [r_t^-(x_{t-1}, a_t)] > -\infty \text{ and } E_{x_0}^\pi [R^-(x_T)] > -\infty.$$

(See e.g. [1], 8.1).

To guarantee this, one can restrict himself to "negative" or "positive" models with

$$r_t(x, a) \leq K, \quad R(x) \leq K, \text{ or with } r_t(x, a) \geq -K, \quad R(x) \geq -K.$$

Another possibility is to consider "contracting" models, where, for some positive function  $v(x)$  and constant  $K$ ,

$$\int_{\mathbf{X}} v(y) p_t(dy|x, a) \leq K v(x) \text{ and } \frac{|r_t(x, a)|}{v(x)} \leq K, \quad \frac{R(x)}{v(x)} \leq K$$

for all  $t, x, a$ .

There is no doubt that one can construct counterexamples also for discounted models and processes with long-run average loss. The classical optimal control theory for them is developed in [1-3, 6], again under conditions which exclude expressions " $+\infty + -\infty$ ".

## APPENDIX

**Lemma 2.** Let  $\mathbf{X}$  and  $\mathbf{Y}$  be two Borel spaces,  $Q(dx)$  be a probability measure on  $\mathbf{X}$  and  $q(dy|x)$  be a (measurable) stochastic kernel on  $\mathbf{Y}$  given  $\mathbf{X}$ . Put  $\Delta$   
 $P(dx \times dy) = Q(dx)q(dy|x)$  be the probability measure on  $\mathbf{X} \times \mathbf{Y}$ . Suppose  $f(x, y) = f_1(x) + f_2(x, y)$ , where

$$f_1: \mathbf{X} \rightarrow \mathbb{R} = (-\infty, +\infty), \quad f_2: \mathbf{X} \times \mathbf{Y} \rightarrow \mathbb{R}^* = [-\infty, +\infty]$$

are measurable functions of their arguments. Then

$$\int_{\mathbf{X} \times \mathbf{Y}} f(x, y) P(dx \times dy) \geq \int_{\mathbf{X}} \left[ f_1(x) + \int_{\mathbf{Y}} f_2(x, y) q(dy|x) \right] Q(dx). \tag{14}$$

In a nutshell, integrating by the compound measure can only enlarge the result, compared with the sequential integrating. Note that according to the proof, inequality in (14) cannot be strict if the expression on the left is finite, i.e. it can be strict only if we have " $+\infty$ " on the lefthand side.

**Proof.** According to the definition and using Lemma 7.11 from [1] we have

$$\int_{\mathbf{X} \times \mathbf{Y}} f(x, y) P(dx \times dy) = \int_{\mathbf{X} \times \mathbf{Y}} f^+(x, y) P(dx \times dy) + \int_{\mathbf{X} \times \mathbf{Y}} f^-(x, y) P(dx \times dy)$$

$$\begin{aligned}
 &= \int_{\mathbf{X}} \left[ \int_{\mathbf{Y}} f^+(x, y) q(dy | x) \right] Q(dx) \\
 &+ \int_{\mathbf{X}} \left[ \int_{\mathbf{Y}} f^-(x, y) q(dy | x) \right] Q(dx) \\
 &\geq \int_{\mathbf{X}} \left[ \int_{\mathbf{Y}} f^+(x, y) q(dy | x) + \int_{\mathbf{Y}} f^-(x, y) q(dy | x) \right] Q(dx) \\
 &= \int_{\mathbf{X}} \left[ \int_{\mathbf{Y}} f(x, y) q(dy | x) \right] Q(dx).
 \end{aligned}$$

If we calculate integral  $\int_{\mathbf{Y}} f(x, y) q(dy | x)$  for a particular value of  $x$  we see that it equals  $f_1(x) + \int_{\mathbf{Y}} f_2(x, y) q(dy | x)$  simply because expression  $\int_{\mathbf{Y}} f_1(x) q(dy | x) = f_1(x)$  is finite. ■

Inequality (14) can be strict indeed as the following example shows. (See Fig. (7)).

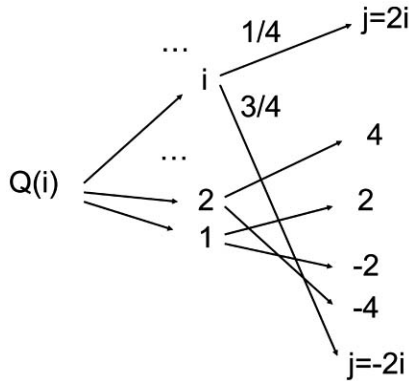


Fig. (7). Strict inequality (14).

Let  $\mathbf{X} = \{1, 2, 3, \dots\}$ ,  $\mathbf{Y} = \{\pm 2, \pm 4, \pm 6, \dots\}$ ,

$$Q(i) = \frac{6}{\pi^2 \cdot i^2}; \quad q(j | i) = \begin{cases} 1/4, & \text{if } j = 2i; \\ 3/4, & \text{if } j = -2i; \\ 0 & \text{otherwise,} \end{cases}$$

$$f_1(i) = i, f_2(i, j) = j.$$

Now in the lefthand part of (14) we have

$$\begin{aligned}
 &\sum_{i \in \mathbf{X}} \left[ \sum_{j \in \mathbf{Y}} (i + j)^+ q(j | i) \right] Q(i) + \sum_{i \in \mathbf{X}} \left[ \sum_{j \in \mathbf{Y}} (i + j)^- q(j | i) \right] Q(i) \\
 &= \sum_{i \in \mathbf{X}} \left( 3i \cdot \frac{1}{4} \right) \frac{6}{\pi^2 \cdot i^2} + \sum_{i \in \mathbf{X}} \left( -i \cdot \frac{3}{4} \right) \frac{6}{\pi^2 \cdot i^2} = "+ \infty" + "- \infty" = +\infty.
 \end{aligned}$$

At the same time, on the right we obviously have zero since

$$i + \sum_{j \in \mathbf{Y}} j \cdot q(j | i) = i + i/2 - 3i/2 = 0.$$

**Proof of Lemma 1.** We use the mathematical induction. If  $t = T$  then  $v_{h_T}^\pi = R(x_T) = v_T(x_T)$ . Suppose Lemma is valid for  $0 < t \leq T$ . Then  $\forall a \in \mathbf{A} \quad \forall x_{t-1} \in \mathbf{X}$

$$v_{t-1}(x_{t-1}) \leq r_t(x_{t-1}, a) + \int_{\mathbf{X}} v_t(y) p_t(dy | x_{t-1}, a).$$

Hence, for any  $h_{t-1} = (x_0, a_1, \dots, x_{t-1}) \in \mathbf{H}_{t-1}$

$$\begin{aligned}
 v_{t-1}(x_{t-1}) &\leq \int_{\mathbf{A}} \left[ r_t(x_{t-1}, a) + \int_{\mathbf{X}} v_t(y) p_t(dy | x_{t-1}, a) \right] \pi_t(da | h_{t-1}) \\
 &\leq \int_{\mathbf{A}} \left[ r_t(x_{t-1}, a_t) + \int_{\mathbf{X}} v_{h_{t-1}, a_t}^\pi p_t(dy | x_{t-1}, a_t) \right] \pi_t(da_t | h_{t-1}).
 \end{aligned}$$

For a fixed  $h_{t-1} = (x_0, a_1, x_1, \dots, x_{t-1}) \in \mathbf{H}_{t-1}$ , measure  $P_{h_{t-1}}^\pi$  on the remaining part of trajectory  $\bar{h}_t = (a_t, x_t, a_{t+1}, \dots, x_T)$  can be disintegrated, so that it is just the product of  $Q(da_t) = \pi_t(da_t | h_{t-1})$  and  $q(dx_t \times d\bar{h}_{t+1} | a_t) = P_{h_{t-1}}^\pi(d\bar{h}_{t+1} | a_t, x_t) \times p_t(dx_t | x_{t-1}, a_t)$ . Note that  $P_{h_{t-1}}^\pi(h | a_t, x_t) = P_{\{h_{t-1}, a_t, x_t\}}^\pi(h)$  according to the definition of a strategic measure. To distinguish  $x_t$  in the lower index which is fixed, from the element of a history (or part of a history)  $h$  that can be arbitrary, we use also denotation  $h = (\bar{x}_0, \bar{a}_1, \dots, \bar{x}_T)$ . Now we can rewrite the last obtained inequality in the form

$$\begin{aligned}
 v_{t-1}(x_{t-1}) &\leq \int_{\mathbf{A}} \left[ r_t(x_{t-1}, a_t) + \int_{\mathbf{X}} \left[ \int_{\mathbf{X} \times \mathbf{A} \times \dots \times \mathbf{X}} \left( \sum_{i=t+1}^T r_i(\bar{x}_{i-1}, \bar{a}_i) + R(\bar{x}_T) \right) \right. \right. \\
 &\quad \left. \left. \times P_{\{h_{t-1}, a_t, x_t\}}^\pi(d\bar{x}_t \times d\bar{h}_{t+1}) \right] p_t(dx_t | x_{t-1}, a_t) \right] \pi_t(da_t | h_{t-1}).
 \end{aligned}$$

But measure  $P_{\{h_{t-1}, a_t, x_t\}}^\pi(d\bar{x}_t)$  is concentrated at point  $x_t$ , so that we omit the bars and obtain

$$\begin{aligned}
 v_{t-1}(x_{t-1}) &\leq \int_{\mathbf{A}} \left[ r_t(x_{t-1}, a_t) + \int_{\mathbf{X}} \left[ \int_{\mathbf{A} \times \mathbf{X} \times \dots \times \mathbf{X}} \left( \sum_{i=t+1}^T r_i(x_{i-1}, a_i) + R(x_T) \right) \right. \right. \\
 &\quad \left. \left. \times P_{\{h_{t-1}, a_t, x_t\}}^\pi(d\bar{h}_{t+1}) \right] p_t(dx_t | x_{t-1}, a_t) \right] \pi_t(da_t | h_{t-1}) \\
 &\leq \int_{\mathbf{A}} \left[ r_t(x_{t-1}, a_t) + \int_{\mathbf{X} \times \mathbf{A} \times \dots \times \mathbf{X}} \left( \sum_{i=t+1}^T r_i(x_{i-1}, a_i) + R(x_T) \right) \right. \\
 &\quad \left. \times q(dx_t \times d\bar{h}_{t+1} | a_t) \right] Q(da_t).
 \end{aligned}$$

In the last inequality and below, we use Lemma 2. Finally

$$\begin{aligned}
 v_{t-1}(x_{t-1}) &\leq \int_{\mathbf{A} \times \mathbf{X} \times \mathbf{A} \times \dots \times \mathbf{X}} \left[ r_t(x_{t-1}, a_t) + \sum_{i=t+1}^T r_i(x_{i-1}, a_i) + R(x_T) \right] P_{h_{t-1}}^\pi(\bar{h}_t) \\
 &= v_{h_{t-1}}^{\pi*}.
 \end{aligned}$$

The last statement of the lemma can be proved in the similar way: when analysing the inequalities starting from  $v_{h_{t-1}}^{\pi*}$  and finishing with  $v_{t-1}(x_{t-1})$ , we see that actually all of them are equalities, according to the remark after Lemma 2. (Value "+ ∞" never appears). ■

**Proof of Corollary 1.** We know that  $\forall \pi$

$$v_{x_0}^\pi = \int_{\mathbf{A} \times \mathbf{X} \times \mathbf{A} \times \dots \times \mathbf{X}} \left[ \sum_{t=1}^T r_t(x_{t-1}, a_t) + R(x_T) \right] P_{x_0}^\pi(d\bar{h}_1) \geq v_0(x_0).$$

Here  $\bar{h}_1 = (a_1, x_1, a_2, \dots, x_T)$ . Measure  $P_{x_0}^\pi(d\bar{h}_1) = q(d\bar{h}_1 | x_0)$  can be considered as a (measurable) stochastic kernel on  $\mathbf{A} \times \mathbf{X} \times \mathbf{A} \times \dots \times \mathbf{X}$  given  $\mathbf{X}$ . If initial distribution  $P_0$  is given then  $P_0^\pi(dh) = P_0^\pi(dx_0 \times d\bar{h}_1) = P_0(dx_0) \times q(d\bar{h}_1 | x_0)$ . Now one can use Lemma 2 for  $f(h) = f_2(h)$

$$\begin{aligned} &= w(h) = \sum_{t=1}^T r_t(x_{t-1}, a_t) + R(x_T) : \\ v^\pi &= \int_{\mathbf{H}} w(h) P_0^\pi(dh) \geq \int_{\mathbf{X}} \left[ \int_{\mathbf{A} \times \mathbf{X} \times \mathbf{A} \times \dots \times \mathbf{X}} w(h) q(d\bar{h}_1 | x_0) \right] P_0(dx_0) \\ &= \int_{\mathbf{X}} v_{x_0}^\pi P_0(dx_0) \geq \int_{\mathbf{X}} v_0(x_0) P_0(dx_0) \end{aligned}$$

and  $\int_{\mathbf{X}} v_0(x_0) P_0(dx_0)$  provides the lower boundary for  $v^\pi$ , indeed. ■

**Proof of Corollary 2.** It is sufficient to note that all inequalities in the proof of Corollary 1 are actually equalities, according to the remark after Lemma 2. Strategy  $\pi^*$  is optimal according to Lemma 1. ■

## REFERENCES

- [1] D. Bertsekas, S. Shreve, *Stochastic Optimal Control*. NY: Academic Press, 1978.
- [2] E.B. Dynkin, A.A. Yushkevich, *Controlled Markov Processes and their Applications*. N.Y.-Berlin: Springer-Verlag, 1979.
- [3] O. Hernandez-Lerma, J.B. Lasserre, *Discrete-Time Markov Control Processes. Basic Optimality Criteria*. NY: Springer, 1996.
- [4] T. Kamihigashi, "On the principle of optimality for nonstationary deterministic dynamic programming," *Int. J. Econ. Theory*, vol. 4, pp. 519-525, 2008.
- [5] A.B. Piunovskiy, *Optimal Control of Random Sequences in Problems with Constraints*. Dordrecht: Kluwer, 1997.
- [6] M.L. Puterman, *Markov Decision Processes*. NY: Wiley, 1994.
- [7] A. Piunovskiy, and X. Mao, "Constrained Markovian decision processes: the dynamic programming approach," *Oper. Res. Lett.*, vol. 27, pp. 119-126, 2000.
- [8] E.A. Feinberg, "Controlled Markov processes with arbitrary numerical criteria," *Theory Probab. Appl.*, vol. 27, pp. 486-503, 1982.

Received: February 11, 2009

Revised: May 18, 2009

Accepted: June 01, 2009

© A.B. Piunovskiy; Licensee *Bentham Open*.

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>), which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.