Open Access

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 traditionnally 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 "$+\infty$"+"$-\infty$" 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 "$+\infty$" and "$-\infty$" 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; $\mathbb{R} = (\text{"}-\infty\text{"},\text{"}+\infty\text{"})$, $\mathbb{R}^* = [-\infty,+\infty]$.

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, $p_t(dx_t \mid x_{t-1}, a_t)$ 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) π is a sequence of measurable stochastic kernels $\pi_t(da_t \mid h_{t-1})$ on $\mathbf{A}$, where $h_{t-1} = (x_0, a_1, x_1, \ldots, a_{t-1}, x_{t-1})$ is a history. A strategy is called Markov if it has the form $\pi_t(da_t \mid h_{t-1}) = \pi_t^m(da_t \mid x_{t-1})$. In case $\forall t\ \pi_t(da \mid h)$ 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 π is fixed, too, then there exists a unique probability measure $P_{P_0}^{\pi}$ on the space of trajectories

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

defined in the usual way:

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

The integral wrt the measure $P_{P_0}^{\pi}$ is denoted by $E_{P_0}^{\pi}$. 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 $h \to x_t$, $h \to a_t$ 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

1874-110X/09 2009 Bentham Open

functions on **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 σ-algebras.

The performance of control strategy π is given by

$$v^{\pi} \stackrel{\Delta}{=} E^{\pi}_{P_0}[W] \stackrel{\Delta}{=} E^{\pi}_{P_0}[W^+] + E^{\pi}_{P_0}[W^-], \tag{2}$$

where $W^+ \stackrel{\Delta}{=} \max\{0, W\}$, $W^- \stackrel{\Delta}{=} \min\{0, W\}$ and

$$"+\infty" + "-\infty" = "+\infty". \tag{3}$$

The aim is to solve problem

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

i.e. to construct an optimal control strategy.

3. DYNAMIC PROGRAMMING APPROACH

The Bellman principle leads to the following equation

$$\begin{cases} v_T(x) = R(x); \\ v_{t-1}(x) = \inf\limits_{a \in \mathbf{A}} \left\{ r_t(x,a) + \int_{\mathbf{X}} v_t(y)\, p_t(dy \mid x, a) \right\}, \quad t = T, T-1, \ldots, 1 \end{cases} \tag{5}$$

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 π^* is optimal in problem (4) if and only if for all $t = 1,2,\ldots,$T

$$v_{t-1}(X_{t-1}) = \int_{\mathbf{A}} \left\{ r_t(X_{t-1}, a) + \int_{\mathbf{X}} v_t(y)\, p_t(dy \mid X_{t-1}, a) \right\} \pi^*_t(da \mid H_{t-1}) \;\; P^{\pi^*}_{P_0} - a.s. \tag{6}$$

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

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

Suppose a history $h_t \in \mathbf{H}_t$, $0 \le \tau \le T$ is fixed. Then we can consider the controlling process A_t and the controlled process X_t as developing on the time interval $\{\tau + 1, \tau + 2, \ldots, T\}$ which is empty if $\tau = T$. If a control strategy π (in the initial model) is fixed then one can build the strategic measure on **H**, denoted as $P^{\pi}_{h_\tau}$, similarly to (1), satisfying the "initial condition" $P^{\pi}_{h_\tau}(h_\tau \times (\mathbf{A} \times \mathbf{X})^{T-\tau}) = 1$. The most important case is $\tau = 0$; then we have just $P^{\pi}_{x_0}$. Note that $P^{\pi}_{x_0}$ is another denotation for $P^{\pi}_{P_0}$ in case $P_0(\cdot)$ is concentrated at point x_0. We introduce $v^{\pi}_{h_\tau} \stackrel{\Delta}{=} E^{\pi}_{h_\tau}\left[\sum_{t=\tau+1}^{T} r_t(X_{t-1}, A_t) + R(X_T)\right]$ and call a control strategy π^* uniformly optimal if

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

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

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

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

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

Note that we use the capital letters for random variables in (4) and (6), so that mathematical expectation and expression $P^{\pi^*}_{P_0} - a.s.$ 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 π, $\forall h_t = (x_0, a_1, \ldots, x_t) \in \mathbf{H}_t$, $t = 0, 1, \ldots, T$, inequality $v^{\pi}_{h_t} \ge v_t(x_t)$ is valid.

In case strategy π^* satisfies equality (8) and $v^{\pi^*}_{h_t} < +\infty$ for all $h_t \in \mathbf{H}_t$, $t = 0, 1, \ldots, T$, we have equality

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

so that π^* is uniformly optimal.

Corollary 1.

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

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

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

Even if equality (6) or (8) holds, it can happen that strategy π^* 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 **X** and **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 $\mathbf{X} = \{0, \pm 1, \pm 2, \ldots\}$, $\mathbf{A} = \{0, -1, -2, \ldots\}$, $T = 3$, $P_0(0) = 1$.

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

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

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

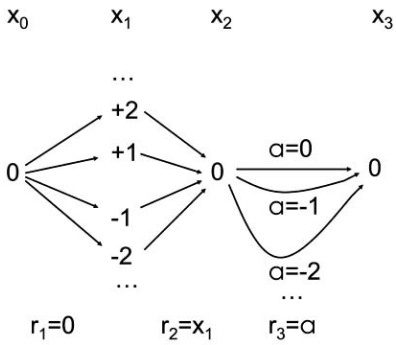

Fig. (1). Example 1: only a non-Markov randomized strategy can satisfy equalities (6) and (8) and be optimal and uniformly optimal.

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 π^* with $\pi_3^*(0 \mid x_2) = 0$, $\pi_3^*(a \mid x_2) = \dfrac{6}{|a|^2 \pi^2}$ for $a < 0$. Here equalities (8) hold because

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

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

On the other side, for any Markov strategy π^m, $v^{\pi^m} = +\infty$. Indeed, let $\hat{a} = \max\{j : \pi_3^m(j \mid 0) > 0\}$; $0 \geq \hat{a} > -\infty$ and consider random variable $W^+ = (X_1 + A_3)^+$. It takes values 1,2,3, ... with probabilities not smaller than

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

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

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

...

(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

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

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

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

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 π^*) can be optimal or uniformly optimal.

Optimal control strategy φ 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 φ 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

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

In the model investigated, for every optimal control strategy π 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:

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

(See Fig. (**2**)).

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

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

Here $\bar{h}_1 = (a_1, x_1, a_2, \ldots, x_T)$. Measure $P^{\pi}_{x_0}(d\bar{h}_1) = q(d\bar{h}_1 \mid x_0)$ can be considered as a (measurable) stochastic kernel on $\mathbf{A}\times\mathbf{X}\times\mathbf{A}\times\ldots\times\mathbf{X}$ given $\mathbf{X}$. If initial distribution P_0 is given then $P^{\pi}_{P_0}(dh) = P^{\pi}_{P_0}(dx_0 \times d\bar{h}_1) = P_0(dx_0) \times q(d\bar{h}_1 \mid x_0)$. Now one can use Lemma 2 for $f(h) = f_2(h) = w(h) = \sum_{t=1}^{T} r_t(x_{t-1}, a_t) + R(x_T)$:

$$v^{\pi} = \int_{\mathbf{H}} w(h) P^{\pi}_{P_0}(dh) \geq \int_{\mathbf{X}} \left[\int_{\mathbf{A}\times\mathbf{X}\times\mathbf{A}\times\ldots\times\mathbf{X}} w(h) q(d\bar{h}_1 \mid x_0) \right] P_0(dx_0)$$

$$= \int_{\mathbf{X}} v^{\pi}_{x_0} P_0(dx_0) \geq \int_{\mathbf{X}} v_0(x_0) P_0(dx_0)$$

and $\int_{\mathbf{X}} v_0(x_0) P_0(dx_0)$ provides the lower boundary for v^{π}, 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 π^* 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 ptinciple 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.