Soltan A. ALIEV, Pavlo R. Shpak, Yaroslav I. YELEYKO

ASSESSMENT AND OPTIMAL POLICIES OF LIMIT KILLED MARKOV DECISION PROCESS

Abstract

We analyzed limit killed Markov decision process on finite time interval for finite and countable models, as well as its assessment and optimal policy. We showed that assessment of limit process is equal to limit of assessments, both for countable and finite model. For finite models we proved that limit point of sequence of optimal policies always exists and each limit point is optimal policy for limit process. For countable models we proved that if limit point of ε -optimal policies exists than this point is δ - optimal policy for limit process, where δ is arbitrary real number greater than ε.

1. Introduction

Bellman's book [6] can be considered as a starting point of Markov decision processes, where ideas of dynamic programming principle were applied to Markov decision processes. Later Markov decision processes are well described in [1]: the definition of Markov decision process was given, as well as definition of optimal policy, assessment of the policy and assessment of process. But in [1] the model does not take into account the risk factor, namely the probability of bankruptcy at some determined moment of time. As a result, we come to the idea of killed Markov decision process where the business can crash with some nonzero probability at every moment of time, with the exception of the initial state. The basic ideas about killing of Markov processes is given in [5]. Also some aspects of Markov decision processes where described by Feinberg and Shwartz [4].

2. Definitions

Let $X_t(t{=}m,\dots,n)$ and let $A_t(t{=}m{+}1,\dots,n)$ be countable or finite.

Remark. All definitions and basic ideas of killed Markov decision process are given accordingly to [2] and [3].

Definition. *A mapping $T : X_t \times A_{t+1} \times X_{t+1} \to P\left(X_t\right)$ which satisfies following conditions:*

$$T\left(x,a,x'\right) = \mathbb{P}\left(x_{t+1} = x \mid x_t = x', a_{t+1} = a\right) \equiv \mathbb{P}$$

is called ***transition function*** *and we we'll it as $p(x|a)$.*

Definition. *The point $x^* \in X_t$ is called* ***killed state****, and $p(x^*|a)$ -* ***probability of kill***

Definition. *The trajectory $l = x_m a_{m+1} x_{m+1} \dots a_t x_t$ n is called way if $t = n$ or $x_t = x^*$.The set of all ways we'll denote as $L = X \times (A \times X)^{n-m}$.*

Definition [Killed Markov decision process]. *A killed Markov decision process on a time interval $[m,n]$ is defined through the following objects:*

1. *Sets $X = \bigcup\limits_{t=m}^{n} X_t$ (spaces of states);*
2. *Sets $A = \bigcup\limits_{t=m+1}^{n} A_t$ (spaces of actions);*
3. *The projection mapping $j : A \to X$, where $j(A_{t+1}) = X_t \backslash \{x^*\}, x^* \in X_t$,*
4. *Probability distribution $p(\cdot|a){=}\mathbb{P}(x_t = x|a_t = a, x_{t-1})$ on X_t with killed states:*

$$\mathbb{P}\left(x_{t+1} = x^* \mid a_t = a\right) \equiv p\left(x^* \mid a\right) \geq 0;;$$

5. *Function $q : A \to \mathbb{R}$ (reward function);*

6. Function $r : X_n \to \mathbb{R}$ (terminal reward);

7. Function c (crash function), defined on killed states:

$$c(x) = \begin{cases} -\sum\limits_{i=m+1}^{t} \sup\limits_{a_i \in A_i} q(a_i) & , \ x = x^* \\ 0 & , \ x \neq x^* \end{cases}$$

$x \in X_t$, x^* - killed state. (function c ensures a total bankruptcy - total loss of accumulated capital or more);

8. Initial distribution μ on X_m.

*A stochastic process defined through (1-8) is called **killed Markov decision process** and is denoted as Z_μ. If the initial distribution μ is concentrated in the point x, we shall write Z_x.*

Definition. *Let $A(x) \subset A$ - is the set of all available actions at state $x \in X$. $\varphi(x) : X \backslash x^* \to A(x)$ is called **simple policy** if $\varphi(x_{t-1}) = a_t$ $\forall t = m+1, \ldots, n$.*

Definition. *The mapping $\pi : H \to \pi(\cdot|h \in H)$ is called **policy**, where $\pi(\cdot|h \in H)$ - probability distribution on $A(x_{t-1})$ and H - the space of histories up to epoch ($h \in H \Leftrightarrow h = x_m a_{m+1}, \ldots, a_{t-1} x_{t-1}$).*

Definition. *Policy $\pi(\cdot|h)$ is called **Markov policy** if $\pi(\cdot\,|h)$ =$\pi(\cdot\,|x_{t-1})$.*

Assumption. *The reward function q and terminal reward function r have the supremum, $\exists$ $sup_{a\in A}$ $q(a)$ and $\exists$ sup $_{x\in X_n}$ $r(x)$.*

If $p(\cdot|a)$ is the transition function and $\pi(\cdot|h)$ is a policy, then for $\forall\mu$ - initial distribution exists corresponding probability distribution P defined on space L which has such notation:

$$P(l) = \begin{cases} \mu(x_m)\,\pi(a_{m+1} \mid x_m)\,p(x_{m+1} \mid a_{m+1}) \cdot \cdots \cdot \\ \quad \cdot\pi(a_n \mid h_{n-1})\,p(x_n \mid a_n), & , \ \forall t \quad x_t \neq x^* \\ \mu(x_m)\,\pi(a_{m+1} \mid x_m)\,p(x_{m+1} \mid a_{m+1}) \cdot \cdots \cdot \\ \quad \cdot\pi(a_t \mid h_{t-1})\,p(x_t^* \mid a_t), & , \quad x_t = x^* \end{cases}$$

For all function ξ from space L the mathematical expectation of ξ is:

$$E^*(\xi) = \sum_{l\in L} \xi(l)\,\mathrm{P}(l) \quad (2.1.2)$$

The assessment

$$I(l) = \begin{cases} \sum\limits_{t=m+1}^{n} q(a_t) + r(x_n) & , \forall r \quad x_k \neq x^* \\ \sum\limits_{k=m+1}^{t} q(a_t) + c(x_t^*) & , \quad x_t = x^* \end{cases}$$

of the way l is example of such function. And we'll denote its expectation as ω:

$$\omega = \mathrm{E}I(l)$$

Definition [Assessment of policy]. *The value ω is called **assessment of policy** .*

Definition [Assessment of process]. *$\upsilon \equiv sup_\pi\,\omega(\pi)$ is called **assessment of killed Markov decision process** Z_μ^* or **assessment of initial distribution** μ.*

Definition [optimal policy]. *If spaces of states and spaces of actions are finite sets then policy π is called **optimal** if $\omega(\pi) = \upsilon$.*

Definition [ε-optimal policy]. *If spaces of states and spaces of actions are countable or finite sets and at least one of them is countable then policy π is called* ***ε-optimal*** *if $\omega(\pi) \geq v - \varepsilon$.*

3. Formulation of the Problem

Lets consider that we have a family of killed Markov decision processes $\left(Z_\mu^k\right)_{k=0}^\infty$ on a time interval $[m, n]$, where $Z_\mu^k = (X, A, j, p_k, q_k, r_k, c_k, \mu_k)$, and following conditions are hold:

$$\forall t,\ \forall a_t \in A_t,\ \forall x_t \in X_t : \lim_{k\to\infty} p_k\left(x_t \mid a_t\right) \ = p\left(x_t \mid a_t\right),$$

$$\forall t,\ \forall a_t \in A_t : \lim_{k\to\infty} q_k\left(a_t\right) = q\left(a_t\right),$$

$$\forall x_n \in X_n : \lim_{k\to\infty} r_k\left(x_n\right) = r\left(x_n\right),$$

$$\forall x_m \in X_m : \lim_{k\to\infty} \mu_k\left(x_m\right) = \mu\left(x_m\right),$$

$$\forall t : \lim_{k\to\infty} c_k\left(x_t^*\right) = c\left(x_t^*\right),$$

$$\forall k : \sum_{x_m \in\ X_m} \mu_k(x_m) = 1,$$

$$\forall t,\ \forall a_t \in\ A_t,\ \forall k : \sum_{x_t \in X_t} p_k\left(x_t \mid a_t\right) = 1.$$

Our goal is to find assessment and optimal (ε -optimal) policies of limit process $Z_\mu = (X, A, j, p, q, r, c, \mu)$.

4. Finite Case

Let $X_t(t{=}m, \dots, n)$ and let $A_t(t{=}m{+}1, \dots, n)$ be countable sets.

Theorem 1. *Let $\left(Z_\mu^k\right)_{k=0}^\infty$ be a family of killed Markov decision processes as defined above, then:*

1. $v = \ lim_{k\to\infty} v_k$, where v - assessment of limit process Z_μ and v_k assessments of corresponding processes Z_μ^k.

2. Exists at least one limit point of $(\overline{\pi}_k)_{k=0}^\infty$, where $\overline{\pi}_k$ – optimal policy for corresponding processes Z_μ^k

3. If $\overline{\pi}$ is a limit point of $(\overline{\pi}_k)_{k=0}^\infty$ then $\overline{\pi}$ is optimal policy for limit process Z_μ.

Proof. For $\forall k$ exists optimal policy $\overline{\pi}_k$ of corresponding process Z_μ^k which satisfies following conditions ([3]):

$$\omega_k\left(\mu_k,\ \overline{\pi}_k\right) = \sup_\pi \omega_k\left(\mu_k,\ \pi\right) \ = \ v_k(\mu_k)$$

$$\forall t,\ \forall a_{t+1} \in\ A_{t+1},\ \forall x_t \in\ X_t :\ 0 \leq \overline{\pi}_k\left(a_{t+1} \middle| x_t\right) \ \leq 1.$$

Let's consider a sequence of optimal polices $(\overline{\pi_k})_{k=1}^\infty$. As X and A are finite sets $(\overline{\pi_k})_{k=1}^\infty$ is a subset of finite Cartesian product of compacts, thus exists convergent sub sequence $(\overline{\pi}_{k_s})_{s=1}^\infty$, which satisfies following conditions:

$$\forall t,\ \forall x_t \in X_t : \sum_{a_{t+1}\ \in\ A_t} \overline{\pi}_{k_s}\left(a_{t+1} \middle| x_t\right) = 1 \quad ,$$

$$\forall t,\ \forall a_{t+1} \in\ A_{t+1},\ \forall x_t \in X_t :\ \overline{\pi}_{k_s}\left(a_{t+1} \middle| x_t\right) \ \underset{s\to\infty}{\longrightarrow} \overline{\pi}\left(a_{t+1} \middle| x_t\right)$$

For limit process $Z_\mu = (X, A, j, p, q, r, c, \mu)$ exists optimal policy $\overline{\pi}^*$:

$$\omega(\mu,\ \overline{\pi}^*) = \sup_{\pi} \omega(\mu,\ \pi)$$

We need to show that assessment of policy $\overline{\pi}$ is equal to assessment of limit process Z_μ*:*

$$\omega\left(\mu, \overline{\pi}\right) = \omega\left(\mu, \overline{\pi}^*\right) = \upsilon(\mu)$$

Let's suppose that this is not true:

$$\omega\left(\mu, \overline{\pi}\right) \neq \upsilon(\mu)$$

Let $\omega\left(\mu, \overline{\pi}\right) > \upsilon(\mu)$. As $\upsilon\left(\mu\right) = \omega(\mu,\ \overline{\pi}^*) = \sup_\pi \omega\left(\mu,\ \pi\right) \geq\ \omega\left(\mu,\ \pi'\right)$ for $\forall\pi'$, consequently $\omega\left(\mu, \overline{\pi}\right) \leq \upsilon(\mu)$.

Let $\omega\left(\mu, \overline{\pi}\right) < \upsilon(\mu)$ and $\omega(\mu,\ \overline{\pi}^*) - \omega\left(\mu, \overline{\pi}\right) =\ \varepsilon > 0$. As

$$\mu_k \underset{k\to\infty}{\longrightarrow} \mu,\ p_k \underset{k\to\infty}{\longrightarrow} p,\ q_k \underset{k\to\infty}{\longrightarrow} q,\ c_k \underset{k\to\infty}{\longrightarrow} c,\ r_k \underset{k\to\infty}{\longrightarrow} r,\ \overline{\pi}_{k_s} \underset{s\to\infty}{\longrightarrow} \overline{\pi}.$$

Then:

$$\forall\ \delta > 0\ \exists\ k_0\ \forall\ k >\ k_0 :\ \ \left|\omega_k\left(\mu_k,\ \overline{\pi}^*\right) - \omega\left(\mu,\ \overline{\pi}^*\right)\right| <\ \delta,$$

$$\forall\ \delta > 0\ \exists\ s_0\ \forall\ s >\ s_0 :\ \ \left|\omega_{k_s}\left(\mu_{k_s},\ \overline{\pi}_{k_s}\right) - \omega\left(\mu,\ \overline{\pi}\right)\right| <\ \delta.$$

And

$$\forall k :\ 0 < \omega\left(\mu, \overline{\pi}^*\right) - \omega\left(\mu, \overline{\pi}\right) =\ \omega\left(\mu, \overline{\pi}^*\right) - \omega_k\left(\mu, \overline{\pi}^*\right) +$$

$$+\ \omega_k\left(\mu, \overline{\pi}^*\right) - \omega_k\left(\mu_k,\ \overline{\pi}_k\right) + \omega_k\left(\mu_k,\ \overline{\pi}_k\right) - \omega\left(\mu, \overline{\pi}\right)$$

Moreover $\overline{\pi}_k$ is optimal policy for process Z_μ^k, thus $\forall k :\ \omega_k\left(\mu, \overline{\pi}^*\right) - \omega_k\left(\mu_k,\ \overline{\pi}_k\right) \leq 0$, as a result:

$$\forall k \quad : \quad 0 < \omega\left(\mu, \overline{\pi}^*\right) - \omega\left(\mu, \overline{\pi}\right) \leq\ \omega\left(\mu, \overline{\pi}^*\right) - \omega_k\left(\mu, \overline{\pi}^*\right) + \omega_k\left(\mu_k,\ \overline{\pi}_k\right) - \omega\left(\mu, \overline{\pi}\right) \leq$$

$$\leq\ \left|\omega_k\left(\mu_k,\ \overline{\pi}^*\right) - \omega\left(\mu,\ \overline{\pi}^*\right)\right| +\ \left|\omega_k\left(\mu_k,\ \overline{\pi}_k\right) - \omega\left(\mu,\ \overline{\pi}\right)\right|$$

$$\forall\ k\ \ > k^* = \max\left\{k_0,\ \ k_{s_o}\right\}\ \ : 2\delta > \left|\omega_k\left(\mu_k,\ \overline{\pi}^*\right) - \omega\left(\mu,\ \overline{\pi}^*\right)\right| +$$

$$+\ \left|\omega_k\left(\mu_k,\ \overline{\pi}_k\right) - \omega_k\left(\mu_k,\ \overline{\pi}\right)\right| \geq \left|\omega\left(\mu, \overline{\pi}^*\right) - \omega\left(\mu, \overline{\pi}\right)\right|\ =\ \omega(\mu,\ \overline{\pi}^*) - \omega\left(\mu, \overline{\pi}\right) =\ \varepsilon >\ 0$$

Which means that $\omega\left(\mu, \overline{\pi}\right) = \omega\left(\mu, \overline{\pi}^*\right) = \upsilon\left(\mu\right)$. Theorem is proved.

5. Countable Case

Let $X_t(t{=}m, \ldots, n)$ and let $A_t(t{=}m{+}1, \ldots, n)$ be countable or finite sets and at least one of them is countable.

$\exists\ sup_{a\in A}\ q(a)$ and $\exists\ sup\ _{x\in X_n} r(x)$

Theorem 2. *Let* $\left(Z_\mu^k\right)_{k=0}^\infty$ *be a family of killed Markov decision processes as defined above. If* $\forall k\ \exists\ sup_{a\in A}\ q_k\left(a\right) \neq \infty$, $\exists\ sup\ _{x\in X_n} r_k(x)$ *and* $\exists\ sup_{a\in A}\ q\left(a\right) \neq \infty$, $\exists\ sup\ _{x\in X_n} r(x)$ *then:*

1. $\upsilon =\ lim_{k\to\infty} \upsilon_k$ *, where* υ *- assessment of limit process* Z_μ *and* υ_k *assessments of corresponding processes* Z_μ^k*.*

2. If exists at limit point of $\left(\overline{\pi}_k\right)_{k=0}^\infty$*, where* $\overline{\pi}_k$ *–* ε *- optimal policy for corresponding processes* Z_μ^k *then this limit point is* δ *- optimal policy for limit process* Z_μ*, where* δ *is arbitrary real number greater than* ε*.*

Proof. As

$$\upsilon_k(\mu_k) = \sup_\pi \omega_k(\mu_k,\ \pi) \ = \sup_\pi \mathrm{E}I(l) \ = \sup_\pi \sum_{l\in L} I(l)\, P_k(l)$$

Let L' be a set of way where $\forall t \ \ x_t \neq x^*$ and $L'' = L\backslash L'$ then

$$\upsilon_k\Big(\mu_k\Big) = \sup_\pi \ \Big(\sum_{l\in L'} I(l)\, P_k(l) + \sum_{l\in L''} I(l)\, P_k(l)\Big) =$$

$$= \sup_\pi \left(\sum_{l\in L'} \left(\sum_{i=m+1}^{n} q_k(a_i) + r_k(x_n)\right) \times\right.$$

$$\times\Big(\mu_k(x_m)\,\pi\Big(a_{m+1} \mid x_m\Big) \ldots \pi(a_n \mid h_{n-1})\, p_k\Big(x_n \mid a_n\Big)\Big) +$$

$$+ \sum_{l\in L''} \Big(\sum_{i=m+1}^{t} q_k(a_i) + c_k\Big(x_t^*\Big)\Big)\times$$

$$\left.\times\Big(\mu_k(x_m)\,\pi\Big(a_{m+1} \mid x_m\Big) \ldots \pi(a_n \mid h_{n-1})\, p_k(x_t^* \mid a_n)\Big)\right)$$

As a result

$$\lim_{k\to\infty} \upsilon_k(\mu_k) = \sup_\pi \left(\sum_{l\in L'} \left(\sum_{i=m+1}^{n} q(a_i) + r(x_n)\right) \times\right.$$

$$\times\left(\mu(x_m)\,\pi(a_{m+1} \mid x_m) \ldots \pi(a_n \mid h_{n-1})\, p(x_n \mid a_n)\right) +$$

$$+ \sum_{l\in L''} \left(\sum_{i=m+1}^{t} q\Big(a_i\Big) + c\Big(x_t^*\Big)\right) \Big(\mu\Big(x_m\Big)\pi\Big(a_{m+1} \Big| x_m\Big)\times$$

$$\left.\times\pi\Big(a_n \Big| h_{n-1}\Big)p\Big(x_t^* \Big| a_n\Big)\Big)\right) \ = \ \upsilon(\mu)$$

In addition if exists at limit point of $(\overline{\pi}_k)_{k=0}^{\infty}$ this means that exists convergent subsequence $(\overline{\pi}_{k_s})_{s=1}^{\infty}$, which satisfies following conditions:

$$\forall t,\ \forall x_t \in X_t : \sum_{a_{t+1}\ \in\ A_t} \overline{\pi}_{k_s}(a_{t+1}|\, x_t) = 1 \quad ,$$

$$\forall t,\ \forall a_{t+1} \in \ A_{t+1},\ \forall x_t \in X_t : \ \overline{\pi}_{k_s}(a_{t+1}|\, x_t) \xrightarrow[s\to\infty]{} \overline{\pi}(a_{t+1}|\, x_t)$$

$$\forall k_s : \ \omega_{k_s}\left(\mu_{k_s},\ \overline{\pi}_{k_s}\right) > \upsilon_{k_s}\left(\mu_{k_s}\right) - \varepsilon$$

So

$$\omega(\mu,\ \overline{\pi}) = \lim_{s\to\infty} \omega_{k_s}\left(\mu_{k_s},\ \overline{\pi}_{k_s}\right) \ \geq \ \upsilon(\mu) - \varepsilon$$

Consequently $\overline{\pi}$ is δ – optimal for limit process $Z_\mu$where δ is arbitrary real number greater than ε. Theorem is proved.

6. Example

Let's consider a family of killed Markov decision processes $\left(Z_\mu^k\right)_{k=0}^{\infty}$ where Z_μ^k:

From definition of crash function $\mathrm{c}\,(\mathrm{x}) = -\sum\limits_{\mathrm{i=m+1}}^{\mathrm{t}} \sup_{\mathrm{a_i \in A_i}} \mathrm{q(a_i)}$ for $\mathrm{x_t{=}x^*}$ we receive that $\mathrm{c}\,(\mathrm{x}_1^*) = -15$ and $\mathrm{c}\,(\mathrm{x}_2^*) = -30{-}(-1)^{\mathrm{n}}3/\mathrm{n}$ .As a result assessments of states (III), (IV) and (V) would equal to following values:

$$\upsilon_n\,(III) = \max?\{22 + (-1)^{\mathrm{n}}12/5\mathrm{n},\ 22\ +\ 1/2n - (-1)^{\mathrm{n}}1/2n\}$$

$$\upsilon_n\,(IV) = 16 - (-1)^{\mathrm{n}}3/10n$$

$$\upsilon_n\,(V) = 16 - (-1)^{\mathrm{n}}3/4n.$$

Accordingly $\upsilon_n\,(III) = 22 + 12/5\mathrm{n}$ if n is even and $\upsilon_n\,(III) = 22 - 1/\mathrm{n}$ if n is odd number.

Thus is n is even assessments of states (I) and (II) would be as following:

$$\upsilon_n\,(I) = 13 + 41/40n +\ n/n + 1$$

$$\upsilon_n\,(II) = 24 + 193/32n +\ 9/20n^2.$$

And if n is odd:

$$\upsilon_n\,(I) = 13 - 67/120n +\ n/n + 1$$

$$\upsilon_n\,(II) = 24 + 783/160n -\ 17/80n^2.$$

As $\upsilon_n\,(\mu_n) =\ \mu_n\,(I)\,\upsilon_n\,(I) + \mu_n\,(II)\,\upsilon_n\,(II)$

We receive that assessment of limit process equals to:

$$\upsilon\,(\mu) =\ 13\mu\,(I) +\ 24\mu\,(II)\,.$$

References

[1]. Dynkin E.B., Yushkevich A.A. *Markov Decision Processes*, M., 1975, 334 p.(Russian).

[2]. Parolya N. R., Yeleyko Y. I. *Killed Markov decision processes on finite time interval for countable models*, Transactions of NAS of Azerbaijan, 2010, vol. XXX, No 4, pp. 141-152.

[3] Parolya N. R., Yeleyko Y. I. *Killed Markov decision processes on finite time interval for finite models*, Visnyk of the Lviv Univ. Series Mech. Math. 2013. Issue 72, pp. 243-254. (Ukrainian)

[4]. Feinberg E.A., Shwartz A. *Handbook of Markov Decision Processes*, Kluwer, 2002, 565 p.

[5] Pakes A.G., *Killing and Resurrection of Markov Processes*, Stochastic Models, 1997, v.13, I.2, pp.255-269.

[6]. Bellman R.E. *Dynamic Programming*, Princeton University Press, 1957, 400 p.

Soltan A. Aliev
Institute of Mathematics and Mechanics of NAS of Azerbaijan
9, B. Vahabzade str., AZ1141, Baku, Azerbaijan
Tel: (99412) 438 22 44 (off.).

Pavlo R. Shpak, Yaroslav I. Yeleyko
Ivan Franko National University of Lviv
1, Universytetska str., 79000, Lviv, Ukraine
Tel.: (8032)239 45 31 (off.).

Received February 07, 2014; Revised April 10, 2014.