Nestor R. PAROLYA, Yaroslav I. YELEYKO

KILLED MARKOV DECISION PROCESSES ON FINITE TIME INTERVAL FOR COUNTABLE MODELS

Abstract

In this article we consider killed Markov decision processes for countable models on finite time interval. Existence of a uniform ε-optimal policy is proved. We showed correctness of the fundamental equation. Optimal control problem is reduced to a similar problem for derived model. We receive optimality equation and method for simple optimal policies constructing. Sufficient of simple policies for countable models is proved. We show correctness of the Markovian property. Additionally dynamic programming principle is considered.

1. Introduction

Markov decision processes arise in different areas of economics, in particular for economic work planning of separate business, economic sector or entire economics. At the beginning of each period we can build the plan for the next period knowing the last achieved state. The system development can be described mathematically as deterministic process if we assume that the system state at the end of each period is uniquely defined by the state at the end of period and by a plan for this period.

But it is necessary to consider the influence of such factors as meteorological conditions, demographic transition, demand fluctuations, the imperfection of the compound production processes coordination, scientific discoveries and inventions etc. Stochastic models are better able to take into account these factors: if we know the state at the beginning of the period and a plan, we can only calculate the probability distribution for the next period. Therefore, leaving aside the system states in the past periods we come to the idea of Markov decision process ("the future depends not on the past, but only on the present").

The Markov decision processes are well described in [1]: the definition of Markov decision process is given, the concept of "model" Z^{μ} is presented, the definition of policy π is given, the assessment of policy - $\omega(\pi)$ and ν - assessment of process Z^{μ} are defined, the existence of a uniform ε-optimal policy is proved, the optimality equation and method for simple optimal policies constructing are presented, the sufficient of simple policies for countable models is proved, the correctness of the Markovian property is shown and dynamic programming principle is considered.

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 [3].

The concept of killed Markov decision process brings us closer to the real economic system which is not typical without such risk.

2. Killed Markov decision process

Let $X_t(t = m, \dots, n)$ and let $A_t(t = m+1, \dots, n)$ be countable or finite sets and at least one of them is countable. $\forall a \in A_t$ compares with a probability distribution $p(\cdot|a) = \mathbb{P}(x_t = x|a_t = a, x_{t-1})$ on X_t.

Remark. All definitions and basic ideas of killed Markov decision process are given according to [1] and [2].

Definition. *Function p which defines the law of transition from A_t to X_t is called* ***transition function****.*

Definition. *The point $x^* = x_m \in X_t$ is called* ***killed state****, and $p(x^*|a)$ -* ***probability of kill*** *if $\mathbb{P}(x_{t+1} = x^*|a_t = a) = \mathbb{P}(x_{t+1} = x_m|a_t = a) \equiv p(x^*|a), x_m \in X_m$.*

Remark. In other words, the system transits into the initial(home) state when it hits a killed state(process is killed).

From the definition of killed state it follows:

$$\forall a \in A_t \ \exists x^* \in X_t : p(x^*|a) = 1 - \sum_{x \in X_t \setminus x^*} p(x|a) > 0.$$

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

1. Sets $X_m, \dots, X_n$(spaces of states);

2. Sets $A_{m+1}, \dots, A_n$(spaces of actions);

3. The projection mapping $j : A \to X$ where $A = \bigcup_{t=m+1}^{n} A_t$, $X = \bigcup_{t=m}^{n} X_t$: $j(A_t) = X_{t-1} \setminus \{x^*\}, x^* \in X_{t-1}, (t = m+2, \dots, n)$ and $j(A_{m+1}) = X_m$;

4. Probability distribution $p(\cdot|a) = \mathbb{P}(x_t = x|a_t = a, x_{t-1})$ on X_t with killed states

$$\mathbb{P}(x_{t+1} = x^*|a_t = a) = \mathbb{P}(x_{t+1} = x_m|a_t = a) \equiv p(x^*|a) > 0;$$

5. Function q on A (reward function);

6. Function r on X_n (terminal reward);

7. Function c (crash function), defined on killed states $c(x^*) = -\sum_{i=m+1}^{t} \max_{a_i \in A_i} q(a_i)$, $x^* \in X_t, t = m+1, \dots, n$ (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*** *or* ***model*** *and is denoted by Z^*_μ. If the initial distribution μ is concentrated in the point x, we shall write Z^*_x.*

Definition. *The trajectory $l = x_m a_{m+1} x_{m+1} \dots a_n x_n$ is called* ***way****. The set of all ways we'll denote $L = X \times (X \times A)^n$.*

Our goal is to find a decision method which maximizes the mathematical expectation of way l assessment :

$$I(l, x^*) = \sum_{t=m+1}^{n} [q(a_t) + c(x^*_t)] + r(x_n), \tag{2.1}$$

where:

$x^* = (x^*_{m+1}, \dots, x^*_n)$ - vector of killed states;

$l = x_m a_{m+1}, \ldots, a_n x_n$ - way.

The decision method is meant to be some *policy.*

3. Policies

Definition. *Let $A(x) \subset A$ is the set of all available actions at state $x \in X$. $\varphi(x) : X \to A(x)$ is called **simple policy** if $\varphi(x_{t-1}) = a_t \ \forall x_t$ - not killed points with probability distributions $p(\cdot|a_t)(m < t \leq n)$ and x_m with the initial distribution μ.*

Remark. When we use simple policy $\varphi(x)$ we get the way $l = x_m a_{m+1}, \ldots, a_n x_n$.

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

Remark. $x_{t-1} \neq x^*$.

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

The next conceptions wont be well-defined without assumption:

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

Definition. *Let $p(\cdot|a)$ is the transition function and let $\pi(\cdot|h)$ is a policy. $\forall \mu$ - initial distribution is compared with probability distribution P^* in space L which has such notation:*

$$P^*(l, x^*) = P^*(x_m a_{m+1}, \ldots, a_n x_n,\ x^*_{m+1}, \ldots, x^*_n) = \mu(x_m)\pi(a_{m+1}|x_m) \times$$

$$\times p(x_{m+1}|a_{m+1}) p(x^*_{m+1}|a_{m+1}) \cdot \ldots \cdot \pi(a_n|h_{n-1}) p(x_n|a_n) p(x^*_n|a_n) \qquad (3.1)$$

Remark. After the definition of measure P^* the way l can be interpreted as stochastic process. Additionally this process is called Markov process if policy π is a Markov policy.

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

$$E^*(\xi) = \sum_{l \in L} \xi(l) P^*(l, x^*) \qquad (3.2)$$

The assessment (2.1) of the way l is example of such function. And we denote its expectation ω:

$$\omega = E^* I(l, x^*) = E^*[\sum_{t=m+1}^{n} [q(a_t) + c(x^*_t)] + r(x_n)] \qquad (3.3)$$

Definition [Assessment of policy]. *The value ω from (3.3) is called **assessment of policy** π and is for a killed Markov decision process Z^*_μ the function of variable π $(\omega = \omega(\pi))$.*

The goal of research is the maximization of function $\omega(\pi)$.

Definition [Assessment of process]. *$\nu \equiv \sup\limits_{\pi} \omega(\pi)$ is called **assessment of killed Markov decision process** Z^*_μ or **assessment of initial distribution** μ.*

Remark. $\nu(x^*) = c(x^*)$.

Definition [ε-optimal policy]. *Killed policy π is called **ε-optimal** for Z^*_μ if $\forall \varepsilon > 0 : \ \omega(\mu, \pi) \geq \nu(\mu) - \varepsilon$.*

Definition [Uniform ε-optimal policy]. *A Killed policy is called **uniform ε-optimal** or **ε-optimal for process** Z^* if π is ε-optimal for Z^*_μ for all μ - initial*

distribution.

4. Existence of uniform ε-optimal policy

Let π_x is ε-optimal policy for process Z_x^*. Its existence follows from the definition of supremum.

We want to build the one killed policy π which is ε-optimal for model Z^* by using a sequence of killed policies π_x.

It's natural to use the policy π_x when x is a starting point. Formally,

$$\bar{\pi}(\cdot|h) = \pi_{x(h)}(\cdot|h) \tag{4.1}$$

where $x(h)$ - the initial state of history h. It's clear that formula (4.1) defines some policy $\bar{\pi}$ and this policy will be ε-optimal. That means $\forall \varepsilon \geq 0 : \omega(x, \bar{\pi}) = \omega(x, \pi_x) \geq \nu(x) - \varepsilon, \forall x \in X_m$.

Proposition [Existence of uniform ε-optimal killed policy]. *Every killed policy $\bar{\pi}$ from (4.1) which is ε-optimal:*

$$\forall \varepsilon \geq 0 : \ \omega(x, \bar{\pi}) \geq \nu(x) - \varepsilon, (x \in X_m)$$

is uniform ε-optimal, that means $\forall \mu, \forall \varepsilon \geq 0 : \sup\limits_{\pi} \omega(\mu, \pi) \leq \omega(\mu, \bar{\pi}) + \varepsilon$.

Proof. From (3.1)-(3.3) it follows that $\forall \pi$:

$$\omega(\mu, \pi) = \sum_{l \in L} I(l, x^*) P^*(l, x^*) = \sum_{X_m} \mu(x) \omega(x, \pi). \tag{4.2}$$

Hence it appears

$$\omega(\mu, \pi) = \sum_{X_m} \mu(x) \omega(x, \pi) \leq \sum_{X_m} \mu(x) \nu(x) \leq \sum_{X_m} \mu(x) [\omega(x, \bar{\pi}) + \varepsilon] = \omega(\mu, \bar{\pi}) + \varepsilon.$$

From received inequalities it follows:

$$\sup_{\pi} \omega(\mu, \pi) \leq \sum_{X_m} \mu(x) \nu(x), \tag{4.3}$$

$$\omega(\mu, \bar{\pi}) \geq \sum_{X_m} \mu(x) \nu(x) - \varepsilon. \tag{4.4}$$

According to arbitrariness of $\varepsilon > 0$ we get now from (4.3) and (4.4)

$$\sup_{\pi} \omega(\mu, \pi) = \sum_{X_m} \mu(x) \nu(x) \leq \omega(\mu, \bar{\pi}) + \varepsilon. \tag{4.5}$$

So policy $\bar{\pi}$ is uniform ε-optimal. ***Proposition 1 is proved.***

Corollary 1. *For all initial distribution μ:*

$$\nu(\mu) = \mu\nu. \tag{4.6}$$

Proof. It follows from $\nu(\mu) = \sum\limits_{X_m} \mu(x) \nu(x) = \mu\nu$.

Remark. Formulas (4.2) and (4.6) allow to reduce the analysis of processes Z_μ^* for all μ to the analysis of processes $Z_x^*, \forall x \in X_m$.

Policy π is built of sequence $\pi_x, (x \in X_m)$ and has following property (1):

For all initial distribution of state $x \in X_m$ the probability distributions in space L which accord with the policies π and π_x from (3.1) are equal.

Definition. *If $\bar{\pi}$ satisfies the property (1) then $\bar{\pi}$ is called* ***combination of policies*** *π_x.*

5. Derived model and fundamental equation

The decision process is a quite number of consecutive steps. The first step is the choice of probability distribution on A_{m+1} which depends on initial state. Since the choice is taken every initial distribution μ on X_m accords with probability distribution $\acute{\mu}$ on X_{m+1}. Now we consider $\acute{\mu}$ as initial distribution in moment of time $m+1$.

As a result, we divide our maximization problem into two problems:

1. We must choose the optimal policy for the next moments of time for every initial distribution on X_{m+1};

2. We must choose the first step according to maximum reward and maximum value of the optimal policy assessment in the next time moments for initial distribution $\acute{\mu}$.

Definition [Derived model]. *The model that builds of model Z^* by deletion X_m and A_{m+1} is called* ***derived model*** *and it denotes $\acute{Z}^*$.*

Proposition [Fundamental equation].

$$\omega(x,\pi) = \sum_{A(x)} \pi(a|x)\Big(q(a) + \acute{\omega}(p_a, \pi_a)\Big), \tag{5.1}$$

where $p_a = p(\cdot|a), \pi_a(\cdot|\acute{h}) = \pi(\cdot|ya\acute{h})$,

$a \in A_{m+1}, y = j(a)$, $\acute{h}$ *- history in model* $\acute{Z}^*$.

Equation(5.1) is called ***fundamental*** *and expresses the assessment ω of random policy π in model Z^* in terms of the assessment $\acute{\omega}$ of some policies in model $\acute{Z}^*$.*

Proof. According to (4.2) we get

$$\acute{\omega}(p_a, \pi_a) = \sum_{X_{m+1}} p(y|a)\acute{\omega}(y, \pi_a) \tag{5.2}$$

Let consider spaces of ways L and $\acute{L}$ in models Z^* and $\acute{Z}^*$. Let P^* is probability distribution on L according to initial state x and policy π and let P_a^* is probability distribution on $\acute{L}$ according to initial distribution p_a and policy π_a.

In according to (2.1) and (3.1) $\forall \acute{l} \in \acute{L}$ we get

$$I(xa\acute{l}, x^*) = q(a) + I(\acute{l}, x^*_{-1}) \tag{5.3}$$

$$P^*(xa\acute{l}, x^*) = \pi(a|x)P_a^*(\acute{l}, x^*_{-1}) \tag{5.4}$$

$$a \in A(x), x^*_{-1} = (x^*_{m+2}, \dots, x^*_n), (x^*_{m+1}, x^*_{-1}) = x^*.$$

Under authority of (3.2) and (3.3) we get

$$\omega(x,\pi) = \sum_{L} P^*(l, x^*) I(l, x^*) \tag{5.5}$$

$$\acute{\omega}(p_a, \pi_a) = \sum_{\acute{L}} P_a^*(\acute{l}, x_{-1}^*) I(\acute{l}, x_{-1}^*) \tag{5.6}$$

Measure $P^*(l, x^*)$ is nonzero only for ways which have the starting point x (that's for ways $xa\acute{l}$). That's why by substituting in (5.5) the expression of $I(l, x^*)$ from (5.3) and the expression of $P^*(l, x^*)$ from (5.4), and according to (5.6) we get fundamental equation (5.1). ***Proposition 2 is proved.***

Remark. The fundamental equation is correct even without ***Assumption 1.***

6. Reducing the problem of optimal decision to analogical problem for derived model. From fundamental equation (5.1) it follows the valuation:

$$\omega(x, \pi) \le \sup_{A(x)} [q(a) + \acute{\omega}(p_a, \pi_a)] \le \sup_{A(x)} [q(a) + \acute{\nu}(p_a)] \tag{6.1}$$

$\forall x \in X_m$ and $\forall \pi$ ($\acute{\nu}$ - assessment of model $\acute{Z}^*$).

We'll denote $u(a) = q(a) + \acute{\nu}(p_a), (a \in A_{m+1})$ and call this value - **assessment of action** a.

According to (4.3) and $\nu(x^*) = c(x^*)$ we get $u = U\acute{\nu}$ where operator U transforms functions on not killed states on X to the functions on A and follows the formula:

$$Uf(a) = q(a) + \sum_{y} p(y|a) f(y) + \sum_{y^*} p(y^*|a) c(y^*) \tag{6.2}$$

where y - not killed states, y^* - killed states.

Let operator V transforms functions on A to functions on not killed and not terminal states on X and follows the formula:

$$Vg(x) = \sup_{a \in A(x)} g(a) \tag{6.3}$$

Let write the inequation (6.1) by using operator V:

$$\omega(x, \pi) \le Vu(x).$$

Then we consider $\sup\limits_{\pi}$ of right and left parts of $\omega(x, \pi) \le Vu(x)$ and we get

$$\nu \le Vu. \tag{6.4}$$

Remark. Later we'll show the conditions which assure the equality in (6.4).

Definition [Product of policies]. *Let $\acute{\pi}$ be a killed policy in model $\acute{Z}^*$ and any $x \in X_m$ is compared with some probability distribution $\gamma(\cdot|x)$ on A_{m+1} which is concentrated on $A(x)$. When we choose on the first step an action a and on all other steps we use the killed policy $\acute{\pi}$ then we get killed policy π in model Z^*. This policy is called* ***product of policies*** *γ and $\acute{\pi}$ and is denoted by $\gamma\acute{\pi}$. It has the expression:*

$$\pi(\cdot|h) = \begin{cases} \gamma(\cdot|x) & \text{for } h = x \in X_m, \\ \acute{\pi}(\cdot|\acute{h}) & \text{for } h = xa\acute{h}. \end{cases}$$

Proposition. *Let $\pi = \gamma\acute{\pi}$ is a product of killed policies γ and $\acute{\pi}$. If $\acute{\pi}$ is uniform ε'-optimal for model $\acute{Z}^*$ then:*

$$\nu = Vu. \tag{6.4}$$

Proof. The fundamental equation (5.1) for a product of policies has the following expression:

$$\omega(x,\gamma\acute{\pi})=\sum_{A(x)}\gamma(a|x)\Big(q(a)+\acute{\omega}(p_a,\acute{\pi})\Big) \tag{6.5}$$

Since $\acute{\pi}$ is ε'-optimal (it exists $\forall\ \varepsilon'\geq 0$ according to *Proposition 1.*) we get $\acute{\omega}(p_a,\acute{\pi})\geq\acute{\nu}(p_a)-\varepsilon'$, and according to appearance of u equation (6.5) transforms to

$$\omega(x,\gamma\acute{\pi})\geq\sum_{A(x)}\gamma(a|x)u(a)-\varepsilon'.$$

Let consider the set

$$A_\chi(x)=\{a:a\in A(x),u(a)\geq Vu(x)-\chi\}\ (x\in X_m).$$

$A_\chi(x)$ is nonempty for all $\chi>0$. Let $\gamma(\cdot|x)$ be a probability distribution on $A(x)$ which is concentrated on $A_\chi(x)$.

Then

$$\sum_{A(x)}\gamma(a|x)u(a)\geq Vu(x)-\chi.$$

Since $\varepsilon'+\chi\leq\varepsilon$ we get

$$\omega(x,\pi)\geq Vu(x)-\varepsilon,\ (x\in X_m). \tag{6.6}$$

According to (6.4) and (6.6) ***Proposition 3 is proved.***

Corollary. *The assessment ν of model Z^* is expressed in terms of assessment $\acute{\nu}$ of model $\acute{Z}^*$ in the following way:*

$$\nu=Vu,\ u=U\acute{\nu} \tag{6.7}$$

where operators U and V are defined in (6.2) and (6.3);

Corollary. *For all $\chi>0$ exists such $\psi(x):X_m\to A_{m+1}(x)$:*

$$u(\psi(x))\geq\nu(x)-\chi \tag{6.8}$$

Here $\gamma(\cdot|x)$ can be the distribution concentrated in one point $\psi(x)\in A_\chi(x)$.

Corollary. *Let ε' and χ arbitrary nonnegative numbers. If $\acute{\pi}$ uniform ε'-optimal for model $\acute{Z}^*$ and ψ such as in Corollary 3 then killed policy $\psi\acute{\pi}$ is uniform $(\varepsilon'+\chi)$-optimal for model Z^*.*

7. Optimality equation. Method for simple optimal policies constructing. Let assume that in our model Z^* $m=0$. Let consider models $Z_0^*,Z_1^*,\dots,Z_n^*$ where $Z^*=Z_0^*$ and Z_t^* is derived model of Z_{t-1}^*. Let denote the assessments ν and u of model Z_t^* as ν_t and $u_{t+1}(\nu_t$ on X_t, u_{t+1} on $A_{t+1})$. The reward function q and transition function p we denote q_t and p_t.

According to the results of *section 6* we get

$$\nu_{t-1}=Vu_t,\ u_t=U\nu_t\ (1\leq t\leq n) \tag{7.1}$$

where

$$U_tf(a)=q_t(a)+\sum_{y\in X_t}p_t(y|a)f(y)+p_t(y^*|a)c(y^*),\ (a\in A_t,y^*\in X_t),$$

$$V_t g(x) = \sup_{A(x)} g(a),\ (x \in X_{t-1}),$$

and $\nu_n = r$.

Equations (7.1) are called **optimality equations**. Let $T_t = V_t U_t$ then optimality equations transform to

$$\nu_{t-1} = T_t \nu_t. \tag{7.1′}$$

From (7.1),(7.1́) and condition $\nu_n = r$ we calculate $\nu_n, \nu_{n-1}, \ldots, \nu_0$. Then we choose the action $\psi_t(x) : X_{t-1} \to A_t(x)$ for which holds

$$u_t(\psi_t) \geq \nu_{t-1} - \chi_t. \tag{7.2}$$

$\forall t = 1, 2, \ldots, n$ and for all nonnegative $\chi_1, \chi_2, \ldots \chi_n$.

According to *Corollary 3* of *Proposition 3* simple policy $\varphi = \psi_1 \psi_2 \ldots \psi_n$ is uniform ε-optimal for model $Z^* = Z_0^*$ and $\varepsilon = \sum_{i=1}^{n} \chi_i$. Equation (7.2) can be rewritten

$$T_{\psi_t} \nu_t \geq \nu_{t-1} - \chi_t, \tag{7.2′}$$

where operator T_{ψ_t} transforms functions on X_t to functions on X_{t-1} in the following way:

$$T_{\psi_t} f(x) = q_t[\psi_t(x)] + \sum_{X_t} p(y|\psi_t(x)) f(y) + p_t(y^*|a) c(y^*). \tag{7.3}$$

Proposition. *Let π is arbitrary killed policy in derived model Z_k^* $(k = 1, 2, \ldots, n)$ and let $\psi_t : X_{t-1} \to A_t(x)$ $(t = 1, 2, \ldots, k)$ are arbitrary too then*

$$\omega_0(x, \psi_1 \psi_2 \ldots \psi_k \pi) = T_{\psi_1} T_{\psi_2} \ldots T_{\psi_k} \omega_k(x, \pi), \tag{7.4}$$

Proof. It follows from fundamental equation (5.1), formulas (5.2), (7.3) and mathematical induction.

Remark. It follows from (7.4): the result will not change if we'll kill our decision process in moment of time k and take the terminal reward as the assessment of policy π.

Remark. If we can choose ψ_t with $\chi_t = 0$ in (7.2́) $\forall t = 1..n$ then simple policy $\varphi = \psi_1 \ldots \psi_n$ is called uniform optimal.

8. Sufficient of simple policies for countable models

There is the question: shall we lose something by using only simple policies? The previous result can't give the answer. It only makes our losses indefinitely small.

Theorem [Sufficient of simple policies]. *Let μ is fixed initial distribution and let π is arbitrary killed policy then exists φ-simple policy such that*

$$\omega(\mu, \pi) \leq \omega(\mu, \varphi). \tag{8.1}$$

Proof. It follows from *Proposition 5* and *Proposition 6*.

Proposition. *$\forall \mu$ and for all killed policy π exists Markov policy θ such that*

$$\omega(\mu, \theta) = \omega(\mu, \pi) \tag{8.2}$$

*(These two policies are called **equivalent**.)*

Proposition. *For all Markov policy θ exists simple policy φ such that*

$$\omega(\mu,\varphi)\geq\omega(\mu,\theta) \tag{8.3}$$

*(we'll say that φ **dominates** θ **uniformly**).*

Proof (*Proposition 5*). Let θ is Markov policy and

$$\theta(a|x)=\mathbb{P}^*\{a_t=a|x_{t-1}=x\}=\frac{\mathbb{P}^*\{x_{t-1}a_t=xa\}}{\mathbb{P}^*\{x_{t-1}=x\}} \tag{8.4}$$

$$(a\in A_t,\quad x\in X_{t-1},\quad m+1\leq t\leq n),$$

where $\mathbb{P}^*$ - measure in space of ways L which compares with initial distribution μ and policy π.

Remark. The expression in a right part of (8.4) makes no sense for $\mathbb{P}^*\{x_{t-1}=x\}=0$. So, for such x(in particular for killed states) we choose instead of $\theta(\cdot|x)$ the arbitrary distribution on $A(x)$.

Let $\mathbb{Q}^*$ denotes probability distribution on space L which compares with initial distribution μ and killed Markov policy θ.

The distribution $\mathbb{Q}^*$ don't match with $\mathbb{P}^*$ in the general case, but it's quite enough for proving (8.2) if any of $x_m,a_{m+1},\ldots,a_n,x_n$ and $x^*_{m+1},x^*_{m+2},\ldots,x^*_n$ has the same probability distribution in relation to measures $\mathbb{P}^*$ and $\mathbb{Q}^*$.

It follows from

$$\omega(\mu,\pi)=\sum_{t=m+1}^{n}\mathbb{P}^*q(a_t)+\sum_{t=m+1}^{n}\mathbb{P}^*c(x^*_t)+\mathbb{P}^*r(x_n),$$

$$\omega(\mu,\theta)=\sum_{t=m+1}^{n}\mathbb{Q}^*q(a_t)+\sum_{t=m+1}^{n}\mathbb{Q}^*c(x^*_t)+\mathbb{Q}^*r(x_n).$$

We shall use the mathematical induction to prove this.

The **basis**: (8.2) holds for x_m because $\mathbb{P}^*=\mathbb{Q}^*=\mu$.

The **induction hypothesis**: let (8.2) holds for x_{t-1}. Let's check it for a_t.

Since θ - is a killed Markov policy then

$$\mathbb{Q}^*\{x_{t-1}a_t=xa\}=\mathbb{Q}^*\{x_{t-1}=x\}\theta(a|x),\quad(a\in A_t,\ x\in X_{t-1}). \tag{8.5}$$

Then from (8.4) and (8.5) we get

$$\mathbb{P}^*\{a_t=a\}=\sum_{x\in X_{t-1}}\mathbb{P}^*\{x_{t-1}a_t=xa\}=\sum_{x\in X_{t-1}}\mathbb{P}^*\{x_{t-1}=x\}\theta(a|x)=$$

$$=\sum_{x\in X_{t-1}}\mathbb{Q}^*\{x_{t-1}=x\}\theta(a|x)=\sum_{x\in X_{t-1}}\mathbb{Q}^*\{x_{t-1}a_t=xa\}=\mathbb{Q}^*\{a_t=a\}.$$

So, our proposition holds for a_t.

The **induction hypothesis**: let (8.2) holds for a_t. Let's show it for x_t.

From the definition of transition function we get

$$\mathbb{P}^*\{a_tx_t=ax\}=\mathbb{P}^*\{a_t=a\}p(x|a), \tag{8.6}$$

$$\mathbb{Q}^*\{a_tx_t=ax\}=\mathbb{Q}^*\{a_t=a\}p(x|a). \tag{8.7}$$

From (8.6) and(8.7) it follows

$$\mathbb{P}^*\{x_t = x\} = \sum_{a\in A_t} \mathbb{P}^*\{a_t x_t = ax\} = \sum_{a\in A_t} \mathbb{P}^*\{a_t = a\}p(x|a) =$$

$$= \sum_{a\in A_t} \mathbb{Q}^*\{a_t = a\}p(x|a) = \sum_{a\in A_t} \mathbb{Q}^*\{a_t x_t = ax\} = \mathbb{Q}^*\{x_t = x\}, \quad (x \in X_t).$$

Proposition 5 is proved.

Proof.(*Proposition 6.*) For proving this proposition we need the following lemma.

Lemma *Let f is arbitrary function and let ν is arbitrary probability distribution on countable space E.*

If $\nu f < +\infty$ then the set $\Gamma = \{x : f(x) \geq \nu f\}$ has positive measure ν, namely

$$\nu(\Gamma) > 0$$

(See proof in [1]).

According to (4.2) the condition (8.3) is equal to

$$\omega(x, \varphi) \geq \omega(x, \theta), \quad \forall x \in X_m.$$

Let's separate killed Markov policy θ into a product of policies $\theta = \gamma\theta'$ where γ is the restriction of θ to X_m and θ' is the restriction of θ to $X_{m+1} \bigcup X_{m+2} \ldots \bigcup X_n$.

According to fundamental equation (5.1)

$$\omega(x, \theta) = \gamma_x f,$$

where $\gamma_x(\cdot) = \gamma(\cdot|x)$ is probability distribution on $A(x)$, and $f(a) = q(a)+\omega'(p_a, \theta')$, $(a \in A_{m+1})$.

Since ***Lemma 1*** for $\tilde{A}(x) \subset A(x)$ it follows $\gamma_x(\tilde{A}(x)) > 0$ where $\tilde{A}(x) = \{a : f(a) \geq \gamma_x f = \omega(x, \theta)\}$. So, $\tilde{A}(x)$ is nonempty. If $\psi(x)$ is arbitrary point of $\tilde{A}(x)$ then $f(\psi(x)) \geq \omega(x, \theta)$. But since fundamental equation (5.1) we get $f(\psi(x)) = \omega(x, \psi\theta')$ and

$$\omega(x, \psi\theta') \geq \omega(x, \theta).$$

Let's assume that condition (8.3) holds for derived model $\acute{Z}^*$. Then exists a simple policy φ' in $\acute{Z}^*$ which uniformly dominates killed Markov policy θ'. According to fundamental equation (5.1) and our assumption we get

$$\omega(x, \psi\varphi') = q(\psi(x)) + \omega'(p_{\psi(x)}, \varphi') \geq q(\psi(x)) + \omega'(p_{\psi(x)}, \theta') = \omega(x, \psi\theta') \geq \omega(x, \theta).$$

In the model Z^* simple policy $\varphi = \psi\varphi'$ dominates θ uniformly. So, (8.3) holds for model Z^* too.

Proposition 6. is proved.

9. Markovian property

Let $0 < k < n$, let we use killed policy ρ on interval $[0, k]$ and killed policy π on interval $[k, n]$. With the analogical considerations like in ***Definition 15*** we can say that policy $\rho\pi$ is used.

Proposition. *Let L_0 is the space of ways on interval $[0, n]$, let L_k is the space of ways on interval $[k, n]$ and let $P_x^{*\rho\pi}$ is the probability distribution which compares with*

initial state x and killed policy $\rho\pi$, and analogically $P_y^{*\pi}$ is the probability distribution on L_k.

Then $\forall \xi = \xi(x_k a_{k+1} \dots x_n)$ on L_k holds

$$E_x^{*\rho\pi}\xi = E_x^{*\rho}[E_{x_k}^{*\pi}\xi]. \tag{9.1}$$

Proof. $\forall l = y_0 b_1 \dots b_k y_k b_{k+1} \dots y_n$ according to (3.1)

$$P_x^{*\rho\pi}(y_0 b_1 \dots y_n) = P^{*\rho}(cy_k) P_{y_k}^{*\pi}(y_k d), \tag{9.2}$$

where $c = y_0 b_1 \dots b_k$, $d = b_{k+1} \dots y_n$. Any function ξ on the space L_k can be interpreted on L_0 like function which not depends on $x_0 a_1, \dots, a_k$. That's why we multiply the both parts of (9.2) by $\xi(y_k d)$ and summate over all ways

$$E_x^{*\rho\pi}\xi = \sum_{cy_k} P_x^{*\rho}(cy_k) \sum_d P_{y_k}^{*\pi}(y_k d)\xi(y_k d). \tag{9.3}$$

But $P_{y_k}^{*\pi}(yd) = 0$ for $y \neq y_k$ and it follows

$$\sum_d P_{y_k}^{*\pi}(y_k d)\xi(y_k d) = \sum_{yd} P_{y_k}^{*\pi}(yd)\xi(yd) = F(y_k). \tag{9.4}$$

By substituting in (9.3) the expression from (9.4) and according to $\sum\limits_{cy_k} P_x^{*\rho}(cy_k)F(y_k) = E_x^{*\rho}F(x_k)$, we get (9.1). ***Proposition 7 is proved.***

Corollary 1 (Markovian property). *Let $\nu(y) = P_\mu^{*\rho}\{x_k = y\}$ $(y \in X_k)$ then $\forall \mu$*

$$E_\mu^{*\rho\pi}\xi = E_\mu^{*\rho}[E_{x_k}^{*\pi}\xi].$$

In particular

$$E_\mu^{*\rho\pi}\xi(x_k a_{k+1} \dots x_n) = E_\nu^{*\pi}\xi(x_k a_{k+1} \dots x_n), \tag{9.5}$$

(It follows form (9.1) and $\sum\limits_{y \in X_k} \nu(y) P_y^{*\pi}\xi = E_\nu^{*\pi}\xi$).

The formula (9.5) shows that the probability distribution for a part of trajectory don't depends on distribution μ and policy ρ on interval $[k, n]$. Namely, the probability forecast of the "future"(ξ) depends not on the "past" (μ, ρ), but only on the "present" (ν). And that's **Markovain property**.

Let's use Markovian property for the intervals $[0, k]$ and $[k, n]$ contribution assessment of killed policy $\rho\pi$. Instead of ξ we take $\xi = \sum\limits_{t=k+1}^{n} [q(a_t) + c(x_t^*)] + r(x_n)$, substitute in (9.5) and get

$$\omega(\mu, \rho\pi) = \sum_{t=1}^{k} E_\mu^{*\rho\pi}[q(a_t) + c(x_t^*)] + \omega(\nu, \pi) = \sum_{t=1}^{k} E_\mu^{*\rho}[q(a_t) + c(x_t^*)] + \omega(\nu, \pi). \tag{9.6}$$

The summation in (8.6) express the assessment $\omega(\mu, \rho)$ of policy ρ for a zero terminal reward, namely, $\omega(\mu, \rho\pi) = \omega(\mu, \rho) + \omega(\nu, \pi)$.

There is also another interpretation of (9.6). According to (4.2) and $\nu(y) = P_\mu^{*\rho}\{x_k = y\}$ $(y \in X_k)$ we get

$$\omega(\nu, \pi) = \sum_y \nu(y)\omega(y, \pi) = E_\mu^{*\rho}\omega(x_k, \pi),$$

$$\omega(\mu, \rho\pi) = E_{\mu}^{*\rho}[\sum_{t=1}^{k} q(a_t) + \omega(x_k, \pi)]. \tag{9.7}$$

So, the assessment of killed policy $\rho\pi$ is equal to the assessment of killed policy ρ with the terminal reward $\omega(\cdot, \pi)$ in the moment of time k.

10. Dynamic programming principle

The ideas of dynamic programming principle for Markov decision processes is given in [4]. Let Z^* is the model on interval $[0, n]$ and let $0 \leq s < t \leq n$. Let's denote $Z^*_{s,t}[f]$ - the model which takes from the model Z^* if $[0, n]$ is restricted to $[s, t]$ and we define the terminal reward f in the moment of time t. We denote $\nu^t_s[f]$- the assessment of the model Z^{*t}_s with the terminal reward - f. It's clear that $\nu^t_s[f] = (VU)^{t-s} f = T^{t-s} f \quad X$.

Since $\forall t \in [0, n]$ holds

$$\nu^n_0[r] = \nu^t_0[\nu^n_t[r]] \text{ on } X_0 \ (r \text{ on } X_n). \tag{10.1}$$

The equation (10.1) is equivalent to the optimality equations (7.1) and condition $\nu^n = r$. It is called ***Dynamic programming principle*** and means: for optimization the decision on the interval $[0, n]$ with terminal reward r we must first optimize the decision on interval $[t, n]$(with such terminal reward) and then optimize the decision on the interval $[0, t]$ with terminal reward $\nu^n_t[r]$.

In particular according to (9.1) it follows if π'' is a uniform ε-optimal killed policy for Z^{*n}_t with terminal reward r and π' is a uniform ε-optimal policy for Z^{*t}_0 with terminal reward $\nu^n_t[r]$ then killed policy $\pi = \pi''\pi'$ has the assessment $\nu^n_0[r]$ and is uniform ε-optimal for model Z^{*n}_0(with terminal reward r).

References

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

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

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

[4]. Bellman R.E. *Dynamic Programming*, .:Izdatelstvo inostrannoj literatury, 1960, 400 p.(Russian).

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

Received February 02, 2010; Revised May 11, 2010.