KILLED MARKOV DECISION PROCESSES FOR COUNTABLE MODELS FOR CRASH FUNCTION ASSESSMENT

Robert Kowal, Ph.D.

Faculty of Management and Administration
University of Humanities and Natural Sciences
Świętokrzyska 21, 25-406 Kielce
e-mail: robertkk7@gmail.com

Received 15 June 2010, Accepted 9 December 2010

Abstract
In this article the killed Markov decision processes for countable models on finite time interval are considered. The existence of a uniform $\varepsilon$-optimal policy is proved. The correctness of the fundamental equation is shown. The optimal control problem is reduced to a similar problem for derived model. Also, the optimality equation and method for simple optimal policies constructing is received. A sufficient condition of simple policies for countable models is proved. The correctness of the Markovian property is shown. Additionally dynamic programming principle is considered.

Keywords: economic system, Markov chains, Markov decision processes, killed Markov decision processes, risk factor, bankruptcy, countable model, derived model, assessment of policy, assessment of process, $e$-optimal policy, dynamic programming.

JEL classification: C10.
Introduction

Markov\(^1\) or similar decision processes arise in many different areas of economics. In particular it is economic work planning of a separate business, an economic sector or entire economies. It is typical for that at the beginning of each period we can build the plan for the next period knowing the last achieved state. At such cases the system development can be described mathematically as deterministic process under the mild assumption that the system state at the end of each period is uniquely defined by the state at the end of a period and by a plan for this period.

But in many cases it is necessary to consider also the influence of such factors like for example meteorological conditions, demographic transition, demand fluctuations, the imperfection of the compound production processes coordination, scientific discoveries and inventions etc. Stochastic models are much 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 states 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 for example by Dynkin and Yushkevich\(^2\). There the definition of Markov decision process is given, the concept of "model" \(Z^\mu\) is presented, the definition of policy \(\pi\) is given, the assessment of policy \(\omega(\pi)\) and \(\nu\) – assessment of process \(Z^\mu\) are defined, the existence of a uniform \(\epsilon\)-optimal policy is proved, the optimality equation and method for simple optimal policies constructing are presented, the sufficient condition of simple policies for countable models is proved, the correctness of the Markovian property is shown and dynamic programming principle is considered.

But the model presented there\(^3\) does not take into account the risk factor, namely the probability of bankruptcy at some determined moment of time. This leads directly to the idea of killed Markov decision process\(^4\), where the business can crash with some nonzero probability at every moment of time, with the exception of the initial state.

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

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. $\forall a \in A_t$ compares with a probability distribution $p(\cdot | a) = P(x_t = x | a_t = a, x_{t-1})$ on $X_t$.

**Definition 1.** Function $p$ which defines the law of transition from $A_t$ to $X_t$ is called transition function.

**Definition 2.** The point $x^*$ is called a killed state, and $p(x^* | a)$ – probability of kill if:

$$P(x_{t+1} = x^* | a_{t+1} = a) = p(x^* | a).$$

**Remark 1.** 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:

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

**Definition 3 (Killed Markov decision process).** A killed Markov decision process on a time interval $[m,n]$ is defined through the following objects:

1. Sets $X_m, \ldots, X_n$ (spaces of states);
2. Sets $A_{m+1}, \ldots, A_n$ (spaces of actions);
3. The projection mapping $j : A \to X$ where $A = \bigcup_{t=m+1}^{n} A_t$, $X = \bigcup_{t=m}^{n} X_t$:

$$j(A_t) = X_{t-1}, (t = m + 1, \ldots, n).$$

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

$$P(x_{t+1} = x^* | a_{t+1} = a) = p(x^* | a) > 0;$$

5. Function $q$ on $A$ (reward function);
6. Function $r$ on $X_n$ (terminal reward);
7. Function $c$ (crash function), defined on killed states:
\[ c(x^*, t) = - \sum_{i=m+1}^{t} \max_{q(i)} q_{i}, \quad t = m+1, \ldots, n \]

(function \( c \) ensures a total bankruptcy - total loss of accumulated capital or more);

8. Initial distribution \( \mu \) on \( X_m \).

A stochastic process defined through (1-8) is called a killed Markov decision process or model and is denoted by \( Z_{\mu}^* \). If the initial distribution \( \mu \) is concentrated in the point \( x \), we shall write \( Z_{x}^* \).

**Definition 4.** The trajectory \( l = x_m, a_{m+1}, x_{m+1}, \ldots, a_n, x_n \) or \( l = x_m, a_{m+1}, x_{m+1}, \ldots, a_t, x^* \), \( t = m+1, \ldots, n \) is called a way. The set of all ways we’ll denote \( L = X \times (A \times X)^{n-m} \).

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

\[
I(l, x^*) = \begin{cases} 
\sum_{i=m+1}^{n} q(a_i) + r(x_i), & x^* \notin l \\
\sum_{i=m+1}^{t} q(a_i) + c(x^*, t), & x^* = x_t 
\end{cases}
\]

where:

- \( x^* \) - killed state;
- \( l = x_m, a_{m+1}, \ldots, a_n, x_n \) - way.

The decision method is meant to be some policy.

2. Policies

**Definition 5.** 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 \( p(\cdot \mid a_t)(m \lt t \leq n) \) and \( x_m \) with the initial distribution \( \mu \).

**Remark 2.** When we use simple policy \( \varphi(x) \) we get the way \( l = x_m, a_{m+1}, \ldots, a_n, x_n \).

**Definition 6.** The mapping \( \pi : H \to \pi(\cdot \mid h \in H) \) is called killed policy, where \( \pi(\cdot \mid h \in H) \) –
probability distribution on $A(x_{t-1})$ and $H = X \times (A \times X)^{t-m-1}$ – 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 3.** $x_{t-1} \neq x^*$. 

**Definition 7.** Killed policy $\pi(\cdot \mid h)$ is called Markov policy if $\pi(\cdot \mid h) = \pi(\cdot \mid x_{t-1})$.

The next conceptions will not be well-defined without the following assumption:

**Assumption 1.** The reward function $q$ and terminal reward function $r$ have the supremum, $\exists \sup_{a \in A} q(a) \leq M < \infty$ and $\exists \sup_{x \in X} r(x) \leq M < \infty$.

**Definition 8.** Let $p(\cdot \mid a)$ is the transition function and let $\pi(\cdot \mid h)$ is a policy. $\forall \mu$ – initial distribution is compared with probability distribution $P$ in space $L$ which has such notation:

$$P^*(l, x^*) = \begin{cases} \mu(x_m) \pi(a_{m+1} \mid x_m) p(x_{m+1} \mid a_{m+1}) \cdots \pi(a_n \mid x_{n-1}) p(x_n \mid a_n), & x^* \notin l \\ \mu(x_m) \pi(a_{m+1} \mid x_m) p(x_{m+1} \mid a_{m+1}) \cdots \pi(a_i \mid x_{i-1}) p(x^* \mid a_i), & x^* = x_i \end{cases} \quad (2)$$

**Remark 4.** After the definition of measure $P^*$ the way $l$ can be interpreted as stochastic process. Additionally this process is called Markov process if policy $\pi$ is a Markov policy.

For all function $\xi$ from space $L$ the mathematical expectation of $\xi$ is:

$$E^*(\xi) = \sum_{l \in L} \xi(l) P^*(l, x^*) \quad (3)$$

The assessment (1) of the way $l$ is an example of such function. And we denote its expectation $\omega$:

$$\omega = E^I(l, x^*). \quad (4)$$

**Definition 9 (Assessment of policy).** The value $\omega$ from (4) is called assessment of policy $\pi$ and is for killed Markov decision process $Z^*_\mu$ the function of variable $\pi \ (\omega = \omega(\pi))$. 
The goal of this research is the maximization of function $\omega(\pi)$.

**Definition 10 (Assessment of process).** $\nu = \sup \omega(\pi)$ is called assessment of killed Markov decision process $Z^*_\mu$ or assessment of initial distribution $\mu$.

**Remark 5.** $\nu(x^*) = c(x^*, t)$.

**Definition 11 (\(\varepsilon\)-optimal policy).** A Killed policy $\pi$ is called $\varepsilon$-optimal for $Z^*_\mu$ if $\forall \varepsilon > 0: \omega(\mu, \pi) \geq \nu(\mu) - \varepsilon$. If $\varepsilon = 0$ then the policy $\pi$ is called optimal.

**Definition 12 (Uniform \(\varepsilon\)-optimal policy).** A Killed policy is called uniform $\varepsilon$-optimal or $\varepsilon$-optimal for process $Z^*$ if $\pi$ is $\varepsilon$-optimal for $Z^*_\mu$ for all initial distributions $\mu$.

3. Existence of uniform $\varepsilon$-optimal policy

Let $\pi_x$ is $\varepsilon$-optimal policy for process $Z^*_x$. Its existence follows from the definition of supremum. Our aim is to build the one killed policy $\pi$ which is $\varepsilon$-optimal for model $Z^*$ by using a sequence of killed policies $\pi_x$. It's natural to use the policy $\pi_x$ when $x$ is a starting point. Formally,

$$\pi(h) = \pi_x(h), \quad (5)$$

where $x(h)$ — the initial state of history $h$.

It's clear that formula (5) defines some policy $\pi$ and this policy will be $\varepsilon$-optimal. That means $\forall \varepsilon \geq 0: \omega(x, \pi) = \omega(x, \pi_x) \geq \nu(x) - \varepsilon, \forall x \in X_m$.

**Proposition 1 (Existence of uniform $\varepsilon$-optimal killed policy).** Every killed policy $\overline{\pi}$ from (5) which is $\varepsilon$-optimal:

$$\forall \varepsilon \geq 0: \omega(x, \overline{\pi}) \geq \nu(x) - \varepsilon, (x \in X_m), \quad (6)$$

is uniform $\varepsilon$-optimal, that means $\forall \mu, \forall \varepsilon \geq 0: \sup_{\pi} \omega(\mu, \pi) \leq \omega(\mu, \pi) + \varepsilon$. 
Corollary 1. For all initial distribution $\mu$:

$$v(\mu) = \mu v \sum_{x \in X_m} \mu(x) v(x).$$  \hspace{1cm} (7)

**Remark 6.** Formulas (6) and (7) allow to reduce the analysis of processes $Z_\mu^*$ for all $\mu$ to the analysis of processes $Z_x^*, \forall x \in X_m$.

Policy $\pi$ 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 $\pi$ and $\pi_x$ from (2) are equal.*

**Definition 13.** If $\pi$ satisfies the property (1) then $\pi$ is called **combination of policies** $\pi_x$.

**4. 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 $\mu$ on $X_m$ accords with probability distribution $\mu'$ on $X_{m+1}$.

Now we consider $\mu'$ as initial distribution in $(m+1)$ moment of time.

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

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 $\mu'$.

**Definition 14 (Derived model).** The model that builds of model $Z^*$ by deletion $X_m$ and $A_{m+1}$ is called **derived model** and it denotes $Z'^*$.

**Proposition 2 (Fundamental equation).**

$$\omega(x, \pi) = \sum_{a(x)} \pi(a \mid x)(q(a) + \omega'(p_a, \pi_a)),$$  \hspace{1cm} (8)

where:
\[ p_a = p(\cdot | a), \pi_a(\cdot | h') = \pi(\cdot | yah'), \]
\[ a \in A_{m+1}, y = j(a), h' \text{ - history in model } Z^*. \]

Equation (8) is called fundamental and expresses the assessment \( \omega \) of random policy \( \pi \) in model \( Z^* \) in terms of the assessment \( \omega' \) of some policies in model \( Z'' \).

**Remark 7.** The fundamental equation is correct even without Assumption 1.

5. Reducing the problem of optimal decision to analogical problem for derived model

From fundamental equation (8) it follows the valuation:

\[ \omega(x, \pi) \leq \sup_{A(x)}[q(a) + \omega'(p_a, \pi_a)] \leq \sup_{A(x)}[q(a) + \nu'(p_a)], \]
\[ \forall x \in X_m \text{ and } \forall \pi \ (\nu' \text{ – assessment of model } Z''). \]

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

According to (7) and \( \nu(x^*) = c(x^*, t) \) we get \( u = U \nu' \), where operator \( U \) transforms functions on not killed states on \( X \) to the functions on \( A \) and follows the formula:

\[ Uf(a) = q(a) + \sum_y p(y | a)f(y) + \sum_y p(y^* | a)c(y^*), \]

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:

\[ Vg(x) = \sup_{a \in A(x)} g(a). \]

Let write the inequation (9) by using operator \( V \):

\[ \omega(x, \pi) \leq Vu(x). \]

Then we consider \( \sup_{\pi} \) of right and left parts of \( \omega(x, \pi) \leq Vu(x) \) and we get:

\[ \nu \leq Vu. \]
Remark 8. Later we’ll show the conditions which assure the equality in (12).

Definition 15 (Product of policies). Let $\pi'$ be a killed policy in model $Z^*$ and any $x \in X_m$ is compared with some probability distribution $\gamma(\cdot \mid x)$ 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 $\pi'$ then we get killed policy $\pi$ in model $Z^*$. This policy is called product of policies $\gamma$ and $\pi'$ and is denoted by $\gamma \pi'$. It has the expression:

$$\pi(\cdot \mid h) = \begin{cases} 
\gamma(\cdot \mid x) & \text{for } h = x \in X_m, \\
\pi'(\cdot \mid h') & \text{for } h = xah'
\end{cases}$$  \hspace{1cm} (13)

Proposition 3. Let $\pi = \gamma \pi'$ is a product of killed policies $\gamma$ and $\pi'$. If $\pi'$ is uniform $\varepsilon'$-optimal for model $Z^*$ then:

$$\nu = Vu.$$  \hspace{1cm} (14)

Corollary 1. The assessment $\nu$ of model $Z^*$ is expressed in terms of assessment $\nu'$ of model $Z'^*$ in the following way:

$$\nu = Vu, u = U \nu' ,$$  \hspace{1cm} (15)

where operators $U$ and $V$ are defined in (10) and (11).

Corollary 2. For all $\chi > 0$ exists such $\psi(x) : X_m \to A_{m+1}(x)$:

$$u(\psi(x)) \geq \nu(x) - \chi .$$  \hspace{1cm} (16)

Here $\gamma(\cdot \mid x)$ can be the distribution concentrated in one point $\psi(x) \in A_x(x)$.

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

6. Optimality equation. Method for simple optimal policies constructing

Now we assume that in our model $Z^* m = 0$. Let consider models $Z^*_0, Z^*_1, \ldots, Z^*_n$ where $Z^* = Z^*_0$ and $Z^*_i$ is derived model of $Z^*_{i-1}$. Let denote the assessments $\nu$ and $u$ of model $Z^*_i$
as \( \nu_t \) and \( u_{t+1} (\nu_t \text{ on } X_t, u_{t+1} \text{ on } A_{t+1}) \). The reward function \( q \) and transition function \( p \) we denote \( q_t \) and \( p_t \).

According to the results from section 5 we get:
\[
\nu_{t-1} = Vu_t, u_t = U\nu_t \quad (1 \leq t \leq n),
\]
where:
\[
U_x f(a) = q_t(a) + \sum_{y \in X_t} p_t(y | a)f(y) + p_t(y^* | a)c(y^*), \quad (a \in A_t, y^* \in X_t),
\]
\[
V_x g(x) = \sup_{A(x)} (a), \quad (x \in X_{t-1}),
\]
\[
\nu_a = r.
\]

Equations (17) are called **optimality equations**. Let \( T_i = V_i U_i \) then optimality equations transform to:
\[
\nu_{t-1} = T_i\nu_t,
\]
From (17), (17') and condition \( \nu_a = r \) we calculate \( \nu_a, \nu_{a-1}, \ldots, \nu_0 \). Then we choose the action
\[
\psi_t(x) : X_{t-1} \to A_t(x) \quad \text{for which holds:}
\]
\[
u_t(\psi_t) \geq \nu_{t-1} - x_t,
\]
\(\forall t = 1, 2, \ldots, n \) and for all nonnegative \( x_1, x_2, \ldots, x_n \).

According to Corollary 3 of Proposition 3 simple policy \( \varphi = \psi_1 \psi_2 \ldots \psi_n \) is uniform \( \varepsilon \)-optimal for model \( Z^* = Z_0^* \) and \( \varepsilon = \sum_{i=1}^n x_i \). Equation (18) can be rewritten in the following form:
\[
T_{\psi_t} \nu_t \geq \nu_{t-1} - x_t,
\]
where operator \( T_{\psi_t} \) transforms functions on \( X_t \) to functions on \( X_{t-1} \) in the following way:
\[
T_{\psi_t} f(x) = q_t[\psi_t(x)] + \sum_{X_t} p(y | \psi_t(x))f(y) + p_t(y^* | a)c(y^*).
\]

**Proposition 4.** Let \( \pi \) is arbitrary killed policy in derived model \( Z^*_1 (k = 1, 2, \ldots, n) \) and let
\[
\psi_t : X_{t-1} \to A_t(x) (t = 1, 2, \ldots, k) \quad \text{are arbitrary too then:}
\]
\[
\omega_k(x, \psi_1 \psi_2 \ldots \psi_k \pi) = T_{\psi_1} T_{\psi_2} \ldots T_{\psi_k} \omega_k(x, \pi),
\]
\(20\)
Remark 9. It follows from (20) that: 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 $\pi$.

Remark 10. If we choose $\psi_i$ with $\chi_i = 0$ in $(18') \forall t = 1..n$ then simple policy $\phi = \psi_1...\psi_n$ is called uniform optimal.

7. Example of killed Markov decision process for finite 3-step model with solution

![Diagram](Fig. 1. Killed Markov decision process for finite 3-step model with solution)

Source: own elaboration.

The numbers in circles are the assessments $\nu_i (t = m,...,n)$ and the black-marked arrows denote the uniform optimal solution (policy).

8. Sufficient condition of simple policies for countable models

There is still 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 1 (Sufficient condition of simple policies).** Let $\mu$ is fixed initial distribution and let $\pi$ is arbitrary killed policy then exists $\phi$-simple policy such that:
\( \omega(\mu, \pi) \leq \omega(\mu, \varphi). \) 

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

**Proposition 5.** \( \forall \mu \) and for all killed policy \( \pi \) exists Markov policy \( \theta \) such that:

\[
\omega(\mu, \theta) = \omega(\mu, \pi).
\]

These two policies are called **equivalent**.

**Proposition 6.** For all Markov policy \( \theta \) exists simple policy \( \varphi \) such that:

\[
\omega(\mu, \varphi) \geq \omega(\mu, \theta).
\]

We say that \( \varphi \) **dominates** \( \theta \) uniformly.

9. **Markovian property**

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

**Proposition 7.** 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 \( P^r_{x, \pi} \) is the probability distribution which compares with initial state \( x \) and killed policy \( \rho \pi \), and analogically \( P^r_{y} \) is the probability distribution on \( L_k \). Then \( \forall \xi = \xi(x, a_{k+1} \ldots x_n) \) on \( L_k \) holds:

\[
E^r_{x, \rho}[E^r_{x, \rho} \xi] = E^r_{x, \rho} \xi.
\]

**Corollary 1 (Markovian property).** Let \( v(y) = P^r_{y} \{ x_k = y \} \) \( y \in X_k \) then \( \forall \mu \) :

\[
E^r_{x, \rho} \xi = E^r_{\mu} \xi.
\]

In particular:

\[
E^r_{\mu} \xi(x, a_{k+1} \ldots x_n) = E^r_{\mu} \xi(x, a_{k+1} \ldots x_n),
\]

(It follows from (24) and \( \sum_{y \in X_k} v(y) P^r_{y} \xi = E^r_{\nu} \xi \).)
The formula (25) shows that the probability distribution for a part of trajectory doesn’t depend on distribution \( \mu \) and policy \( \rho \) on interval \([k,n]\). Namely, the probability forecast of the "future" \( \xi \) depends not on the "past" \( (\mu, \rho) \), but only on the "present" \( (v) \). And that's Markovian property.

Let’s use Markovian property for the intervals \([0,k]\) and \([k,n]\) contribution assessment of killed policy \( \rho \pi \). Instead of \( \xi \) we take 
\[
\xi^* = \left\{ \begin{array}{ll}
\sum_{i=k+1}^{n} q(a_i) + r(x_i), & x^* \notin l \\
\sum_{i=k+1}^{t} q(a_i) + c(x^*, t), & x^* = x_t
\end{array} \right.,
\]
substitute in (25) and get:
\[
\omega(\mu, \rho \pi) = \sum_{i=t}^{k} E_{\mu}^{\rho, \pi} [q(a_i) + c(x^*_i)] + \omega(v, \pi) = \sum_{i=t}^{k} E_{\mu}^{\rho, \pi} [q(a_i) + c(x^*_i)] + \omega(v, \pi). \tag{26}
\]

The summation in (26) express the assessment \( \omega(\mu, \rho) \) of policy \( \rho \) for a zero terminal reward, namely, 
\[
\omega(\mu, \rho \pi) = \omega(\mu, \rho) + \omega(v, \pi).
\]

There is also another interpretation of (26). According to (6) and \( \nu(y) = P_{\mu}^{s, \rho} \{ x_k = y \} \ (y \in X_k) \) we get:
\[
\omega(v, \pi) = \sum_{y} \nu(y) \omega(y, \pi) = E_{\mu}^{\rho, \pi} \omega(x_k, \pi),
\]
\[
\omega(\mu, \rho \pi) = E_{\mu}^{\rho, \pi} \left[ \sum_{i=1}^{k} q(a_i) + \omega(x_i, \pi) \right]. \tag{27}
\]
So, the assessment of killed policy \( \rho \pi \) is equal to the assessment of killed policy \( \rho \) with the terminal reward \( \omega(., \pi) \) in the moment of time \( k \).

10. Dynamic programming principle

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 \( V_{s,t}^*[f] \) - the assessment of the model \( Z_{s,t}^* \) with the terminal reward - \( f \). It's clear that \( V_{s,t}^*[f] = (VU)^{t-s} f = T^{t-s} f \) on \( X \).

Since \( \forall t \in [0,n] \) holds:
\[
V_0^*[r] = V_0^*[v_0^*[r]] \text{ on } X_0 \ (r \text{ on } X_0). \tag{28}
\]
The equation (28) is equivalent to the optimality equations (17) 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 \( \nu^t_t[r] \).

In particular according to (24) it follows if \( \pi'' \) is a uniform \( \varepsilon \) -optimal killed policy for \( Z^n_t \) with terminal reward \( r \) and \( \pi' \) is a uniform \( \varepsilon \) -optimal policy for \( Z^n_0 \) with terminal reward \( \nu_0^n[r] \) then killed policy \( \pi = \pi''\pi' \) has the assessment \( \nu_0^n[r] \) and is uniform \( \varepsilon \) -optimal for model \( Z^n_0 \) (with terminal reward \( r \)).

**Notes**

1 Markov processes are described in Feinberg, Shwartz (2002).
2 Dynkin, Yushkevich (1975).
3 Dynkin, Yushkevich (1975).
4 Some related ideas of this subject appears in Pakes (1997).
5 Elements of dynamic programming one can find in Bellman (1977).

**References**

Dynkin, E.B., Yushkevich, A.A. (1975). *Markov Decision Processes*, Moskwa.

Feinberg, E.A., Shwartz, A. (2002). Introduction. In E.A Feinberg, A. Shwartz (Eds.), *Handbook of Markov Decision Processes*, 1-17, Kluwer.

Pakes, A.G. (1997). Killing and Resurrection of Markov Processes. *Communications in Statistics. Stochastic Models*, 13(2), 255-269.

Bellman, R. (1977). *Programowanie dynamiczne*, Warszawa: Wydawnictwo Naukowe PWN.