# A Gittins Type Index Theorem for Randomly Evolving Graphs

Ernst Presman[1] and Isaac Sonin[2]

[1] Central Economics and Mathematics Inst., Nakhimovskii pr. 47, Moscow, Russia presman@cemi.rssi.ru

[2] Dept. of Mathematics, Univ. of North Carolina at Charlotte, Charlotte, NC, 28223, USA imsonin@email.uncc.edu

**Summary.** We consider the problem which informally can be described as follows. Initially a finite set of independent trials is available. If a Decision Maker (DM) chooses to test a specific trial she receives a reward, and with some probability, the process of testing is terminated or the tested trial becomes unavailable but some random finite set (possibly empty) of new independent trials is added to the set of initial trials, and so on. The total number of potential trials is finite. A DM knows the rewards and transition probabilities depending on the trials. On each step she can either quit (i.e. stop the process of testing), or continue. Her goal is to select an order to test trials and an quitting (stopping) time to maximize the expected total reward. We simplify and generalize some results obtained earlier for similar problems, we prove that an index can be assigned to each possible trial and an optimal strategy uses on each step the trial with maximal index between available ones. We present a recursive procedure with a transparent interpretation to calculate the index. We discuss the connection between introduced index and Gittins index.

*Key words:* Markov Decision Process, Graph, Gittins Index, Priority rules.

## 1 Introduction

The goal of this paper is twofold. First, to generalize the main result and to simplify the proof of the paper by Denardo et al. [3]. In that paper a model of R&D projects is considered. Each stage of a project in the model is represented by an edge of a directed forest. To activate an edge $e$ one needs to pay a certain amount $r(e)$. Each activated edge can pass or fail. The successful completion of a path from a root to a leaf brings certain reward and terminates the activity. In case of failure all edges which follow the failed edge become unavailable. The goal is to maximize the expected reward. The optimal strategy in the model is an index strategy. Each time one should use an edge with the highest value of the index among the available indices. An index for an edge is specified only by the parameters of the directed tree above this edge. We consider more general model where an optimal strategy is also an index strategy. The notion

of the index in both papers is a generalization of the corresponding notion in the model, which we call below a binary elementary (BE) model, studied in early sixties in Mitten (1960) [9].

The second goal of our paper is to show that the index described above is a generalization of the well-known Gittins index (GI). Thus GI, beside the original papers of Gittins [6] and Gittins and Jones [7], has the second root of its origin in the mentioned paper by Mitten [9]. It seems that the proper credit never was given to Mitten and his model.

The strategies of the type, when for selecting an action on each stage it suffices to solve much simpler problem, for example the one-step optimization problem, are called *myopic* or *greedy*. They are very popular and intensively studied though in contrast to model above they usually are not optimal. We call a strategy a Priority Rule (PR) if an index is calculated for each action and an action with the highest value of index among available is selected.

The myopic strategies form a nucleus of developed later so called Multi-armed Bandit (MAB) Theory (for independent (!) arms) (see Gittins [6], Whittle [15], and Berry and Fristedt [1]), where the corresponding strategy is called Gittins index strategy.

The GI index, denoted by $G(x)$, where $x$ is a state of Markov chain, plays an important role in theory of MAB with *independent* arms but it also appears in other problems like the optimal replacement problems. The main result of this theory states that if there are a finite number of independent MC and a decision maker at each moment can engage (test) one of these MC while all other remain frozen then the optimal strategy is to test MC whose state $x^j$ at this moment has the largest value $G^j(x^j)$, where $G^j(x^j)$ is the value of GI of MC $j$ at state $x^j$.

Note also that the same term Multi-armed bandit problem is used also in the classical papers by R. Bellman [2], D. Feldman [4] as well as in the book of Presman and Sonin [10] and in some sections of the book by Berry and Fristedt where arms are *dependent*, i.e. a trial of one arm provides an information about the parameters of other arms also. In this case a myopic Gittins index strategy is not optimal in general.

The traditional *Gittins index* $G(x)$ for a Markov chain (MC) is defined as the maximal value of a discounted expected reward per expected discounted length of a cycle starting from $x$, i.e.

$$G(x) = \sup_{\tau} \frac{E_x \sum_{n=o}^{\tau-1} \beta^n r(Z_n)}{E_x \sum_{n=o}^{\tau-1} \beta^n}, \tag{1}$$

where $\beta$ is a discount factor, $0 < \beta < 1$, $\tau$ is a stopping time, $\tau \geq 1$, $r(\cdot)$ is a reward function, and $Z_n$ is the state of Markov chain at time $n$.

Note, that as usual in the theory of Markov Decision Processes, one can consider the discount factor $\beta$ as a probability of survival of a MC at each step. Formally one can introduce an absorbing state and to introduce new probabilities such that the probability of transition to an absorbing state is

equal to $1-\beta$ and all other transition probabilities are multiplied by the factor $\beta$. Then the denominator in formula (1) multiplied by $(1-\beta)$ is equal to the probability of absorption during the time interval $(0,\tau)$,

$$Q^{\tau}(x) = 1 - E_x\beta^{\tau}. \tag{2}$$

In our paper we will consider the specific Markov decision process on a forest with one absorption state, when probability of absorption $q(A)$ depends on chosen action $A$. We introduce notion of index for control *actions* as follows. For fixed strategy $\pi$ with stopping time $\tau$ and control process $(A_i)$, with $A_0 = e$, we consider the reward $R^{\pi}(e)$, and the probability of absorption $Q^{\pi}(e)$. Following the footsteps of Mitten [9], Granot and Zuckerman [8] and Denardo et al. [3], we define the index

$$\alpha(e) = \sup \frac{R^{\pi}(e)}{Q^{\pi}(e)}\ , \tag{3}$$

where supremum is taken over some set of strategies.

Note that the reward $R^{\pi}(e)$ can be represented in the form

$$R^{\pi}(e) = \mathbf{E}^{\pi}\left[\sum_{i=0}^{\tau-1} r(A_i)\right] = \tilde{\mathbf{E}}^{\pi}\left[\sum_{i=0}^{\tau-1} r(A_i)\prod_{j=0}^{i-1}(1-q(A_j))\right],$$

where $\tilde{E}$ denote the expectation with respect to corresponding Markov chain without absorbing state. The probability of absorption $Q^{\pi}(e)$ can be represented in the same way with $q(\cdot)$ instead of $r(\cdot)$. In case $q(A_i) = 1-\beta$ for all $i$, the denominator in (3) coincides with (2). So, (3) generalizes (1) to the case of Markov decision process with probability of absorption depending on the current state.

In the sequel we consider only the case of finite forest but most of the results can be extended to the case of an infinite forest with some extra conditions.

The plan of our paper is as follows. In Section 2 and 3 we consider correspondingly the BE-model and the model studied in Denardo et al. [3]. In Section 4 we formulate our model and present the main result. In Section 5 we discuss main ideas of the proofs. In Section 6 we present and prove some auxiliary results leaving the proof of one Lemma to the Appendix (Section 9). In Section 6 we give the proof of the main result. In Section 8 we present an algorithm for calculating the index. In Section 9 we discuss connection with Gittins index and some open problems.

## 2 A binary elementary (BE) model of independent trials.

Suppose that there is a finite set of independent Bernoulli trials $e_1, e_2, ..., e_m$, with two possible outcomes in each trial, "continuation" with probability $p_i$,

in $i$-th trial, and "termination", with probability $q_i$. A decision maker (DM) can choose an order in which to conduct (test) the trials. Each trial can be tested only once. The test of $i$-th trial brings a reward $r_i$, and in the case of "continuation" she may continue testing or quit. In the case of "termination" the testing has to be *terminated*. The goal of DM is to select the optimal order to maximize the expected total reward. Such formulation is equivalent to a formulation where DM has to pay an amount $c_i$ in advance, obtains $a_i$ with probability $p_i$, and $b_i$ with probability $q_i$, and $r_i = -c_i + a_i p_i + b_i q_i$.

This problem is a reformulation of a "least cost testing sequencing" problem solved independently by a few authors in 1960 (see Mitten [9]). We call it BE-model (Binary Elementary model). A rather simple proof shows that the optimal strategy has a remarkable simple structure and is based on an index $\alpha$ calculated for *each trial* $e_i$, $\alpha(e_i)$ equal to expected profit divided by probability of termination, i.e.

$$\alpha(e_i) = \frac{r_i}{q_i}. \tag{4}$$

The optimal strategy has the following form: test the trials with positive index in the order of decreasing. If all trials must be tested then all they should be tested in the above order. Mitten analyzed the model when $c_i < 0, a_i = 0$, and $b_i > 0$ but this makes no difference for the analysis of the problem.

## 3 Independent trials on a forest. Binary forest (BF) model.

A model described above was generalized by Granot and Zuckerman [8] in the context of multi-stage R&D models. That paper has many interesting developments but contrary to their claim the Theorem 1 in their paper can be obtained from the Mitten result by transforming semimarkov discounting into absorption probabilities.

This model in turn was recently generalized in a paper by Denardo et al. [3]. The latter model can be described briefly as follows.

At initial moment a set of independent trials with two possible outcomes are available. For some of trials the nature of two outcomes is the same as in BE model - "continuation" and "termination". For other trials for both of outcomes one can continue but differently. to pone of outcomes leads to a possibility to continue the process of testing. In the case of one outcome a "continuation" is the same as above, but the second of outcomes adds to the set of available trial a set of new trials, some of them with a similar feature and so on, and so on. Each trial $e$ of the second kind and all trials that "follow" $e$ in one or more steps can be represented by edges of a *directed tree* $T(e)$. A tree corresponding to the trial of the first kind consists of one edge. The total set of potentially available trials is finite and is represented by a union of directed

trees, i.e. by a *directed forest* $F_0$. The trials of the first kind correspond to the leaves of this forest, i.e. to the edges such that no edges follows. All other edges are called stems. The initially available edges are called the *roots* of $F_0$.

If edge $e$ is tested (used) it can *pass* with some probability or *fail* with complimentary probability. These events are independent of similar events for other edges. If an edge $e$ "fails" than $e$ and all edges that follow $e$ are not available any more, but other available edges can be tested. If a stem $e$ passed then it becomes unavailable but all edges that immediately follow $e$ are added to the set of available edges. If a leaf $e$ passed then the testing has to be terminated. An edge $e'$ can be tested only once and only if all edges on the path from one of the roots of $F_0$ to $e'$ "passed" before. The reward on stems (costs) are negative, positive rewards (prizes) are available only on *leaves,* i.e. on edges such that no edge follows. The testing can be conducted *till the termination*, when a prize is obtained, i.e. a leaf is reached and "passed", or till the moment when DM decides to quit, i.e. to stop testing. The goal of a DM is to maximize the expected value of either linear or exponential function of the profit (total reward) over all possible strategies to test edges. We call this model BF-model (Binary Forest model) since the result of each trial has two outcomes.

The main result of paper [3] is that the optimal strategy is based again on an index generalizing (4). This index $\alpha(e)$ is defined as $\alpha(e) = \sup_{\pi} \frac{R^{\pi}(e)}{Q^{\pi}(e)}$, where $R^{\pi}(e)$ and $Q^{\pi}(e)$ are correspondingly the expected total reward and the probability of termination (to obtain a prize) in the linear case and corresponding function in exponential case. The supremum is taken over some class of strategies, which authors call "candidates". The authors also noted that their problem can be described in terms of so called MAB processes and their index is similar to the Gittins index.

We gratefully acknowledge the possibility to read the manuscript of [3] before its publication.

The proof of the main theorem in [3] is complicated and long. Responding to their hope "that someone will devise a simpler proof than theirs" we obtained in the linear case a different, shorter and more transparent inductive proof of this important and interesting result. We found also that our proof covers also more general situation when:

1) a binary result of testing of an edge (a trial) can be replaced by a finite number of outcomes in the spirit of general theory of Markov Decision Processes (MDP);

2) two separate functions, the prize function $b(e) > 0$ for leaves and the cost function $c(e) < 0$ for all other edges are replaced by a general reward function $r(e)$, which can take any finite values (positive, negative or zero) for any edge;

3) the termination when a prize is obtained, is replaced by a possibility of termination with probability depending on the trial tested at any stage.

The last possibility implies also that the discounting with coefficient $\beta, 0 < \beta < 1$ can be considered as a special case of our model since it is equivalent to a termination with a fixed probability $1 - \beta$.

We will consider only the linear function of the profit.

Note also that the optimal strategy in BF-model takes the form of a series of "depth first" searches of paths to leaves. In our model this property is not true generally due to generalization 2.

In the MAB literature the term *arm* is usually understood as a stochastic process which can be engaged again and again. In the BE, BF models and the model presented below each edge can be used only once so we prefer not to use the term arm at all.

## 4 Multiple Forest (MF) Model. Formulation and results.

We present our model in a standard frame of Markov Decision Processes (MDP). A MDP model is given (see e.g. Feinberg and Schwartz [5]) by a tuple $M = (S, A(x), p(y|x,a), L)$, where $S$ is a state space, $x \in S$ represents a state of a system under consideration, $A(x)$ is a set of actions $a$ available at state $x$, $p(y|x,a)$ is a probability that the next state is $y$ if at state $x$ an action $a$ was chosen (transition operator), and $L$ is a functional defined on the *trajectories* of a system.

By $h_n = (x_0, a_0, x_1, \dots, x_{n-1}, a_{n-1}, x_n)$ we denote a trajectory of length $n$, $n \le \infty$, $h_\infty = h$. A general (randomized) strategy $\pi$ in MDP is a sequence $\pi_n(\cdot|h_n), n = 0, 1, 2, ...$ of distributions on action set $A(x_n)$ possibly depending on the whole past history. An initial state $x$ and a strategy $\pi$ define a measure $P_x^\pi$ in the space of infinite trajectories, i.e. the distribution of the state-action process $(X_n, A_n)$, $X_n(h) = x_n, A_n(h) = a_n, n = 0, 1, \dots$. We denote by $E_x^\pi$ the corresponding expectation. If a distribution $\pi_n(\cdot|h_n)$ is a function $\pi(x_n)$ with values in $A(x_n)$, a strategy $\pi$ is a *stationary* (nonrandomized) strategy. A stationary strategy $\pi$ defines the transition probabilities $p(y|x,\pi(x))$ for the (homogeneous) Markov chain $(X_n)$ describing the evolution of the system. The goal of the DM is to maximize the *expected total reward* $R^\pi(x) = E_x^\pi L = E_x^\pi \sum_{i=0}^\infty r(X_i, A_i)$. From the general theory of MDP it follows that for such a functional it suffices to consider only the stationary strategies. The value function $R(x) = \sup_\pi R^\pi(x)$ satisfies the Bellman (optimality) equation $R(x) = \sup\limits_{a \in A(x)} \left[ r(x,a) + \sum\limits_y p(y|x,a) R(y) \right]$.

Let some initial forest $F_0$ be given. We say that edge $e'$ *follows* $e$, if $e$ is on a unique path from a root of a tree to $e'$. Denote by $N(e)$ the edges from $T(e)$ that immediately follow $e$. *Leaves* are edges such that no edge follows. Other edges are *stems*.

The state space $S = \{x\}$ in MF-model consists of absorbing state $x_*$, empty set $\emptyset$, and all subsets of edges of $F_0$ which do not contain any two

edges such that one follows other, i.e. if $e, e' \in x$ for some $x$ and $e \neq e'$ then $T(e) \bigcap T(e') = \emptyset$.

The action set $A(x) = x \cup \{e_*\}$ for $x \neq x_*$, $A(x_*) = e_*$, where $e_*$ is a *quit* action, i.e. at each stage a DM can test any of edges in $x$ or select an action $e_*$ which at the next moment moves a system to $x_*$.

The following parameters are defined for every edge $e$: 1) a number $q(e), 0 \leq q(e) \leq 1$, 2) for each subset $D$ of the set $N(e)$ (including empty set and the full set $N(e)$) a number $p_D(e) \geq 0$ such that $\sum_{D \subset N(e)} p_D(e) = 1 - q(e)$, 3) a reward $r(e)$ such that $r(e_*) = 0$.

The meaning of these parameters is as follows. Edges correspond to trials. If edge $e$ is tested, it becomes unavailable, and with probability $q(e)$ the system moves to the absorbing state $x_*$, and with probability $p_D(e)$ all edges from the set $D$ are added to the set of edges available for testing.

Formally, the transitional probabilities have the following form: $p(x_*|x, e_*) = 1$; if $e \neq e_*$ then $p(y|x, e) = p_D(e)$ for $y = \{x \setminus e\} \cup D$ and $p(x_*|x, e) = q(e)$. Note that the independence of arms (edges $e$) is manifested by the property that $p(y|x, e)$ depends only on $e \in x$, and does not depend on other $e'$ from $x$, and that the "coordinates" of a new state $y$ for edges $e' \neq e$ remain the same.

Given an initial state $x$ and strategy $\pi$, the goal is to maximize the expected total reward, $R^\pi(x) = E_x^\pi \sum_{i=0}^{\infty} r(A_i)$, where $A_i$ is the edge tested at moment $i$.

**Main Problem A:** *Given an initial state $x$, maximize $R^\pi(x)$ over all strategies.*

**Main Problem B:** *Given an initial state $x$, maximize $R^\pi(x)$ over all strategies such that a quit action $e_*$ is available only if $x = \emptyset$,* or *$x = x_*$.*

As we mentioned, the general theory of MDP implies that for these problem the stationary nonrandomized strategies form a sufficient class. Still, stationary strategies may have rather complicated structure. For example, a strategy can test edge $e$ if edges $e, e'$, and $e''$ are available and test edge $e'$ if only edges $e$, and $e'$ are available. We can expect that the optimal strategy will be among stationary strategies having the following simpler structure.

Consider an *ordered* list of different edges $\pi = (e_1, ..., e_k)$. We say that $e_i$ is *senior* than $e_j$ for $\pi$ if $e_i$ is listed earlier i.e. if $i < j$. We denote $\{\pi\} = \{e_1, ..., e_k\}$, i.e. the set of elements of $\pi$. List $\pi$ defines a (nonrandomized) stationary strategy, which we denote also $\pi$, as follows: if there is no available edges, i.e. if $x \cap \{\pi\} = \emptyset$, then $\pi(x) = e_*$, otherwise $\pi(x)$ equals to the most senior element in $x \cap \{\pi\}$. Such strategy is called a *priority rule* (PR).

Note that if $e_i$ is senior than $e_j$, it does not imply that edge $e_i$ for a particular history will be used earlier then $e_j$. It may happens because $e_i$ may be not available when $e_j$ is already available. More than that, it is possible that two different lists define the same PR because the same states have positive probabilities and both lists define the same order for each state that has positive probability.

**Example**. Consider the forest given on Fig.1.

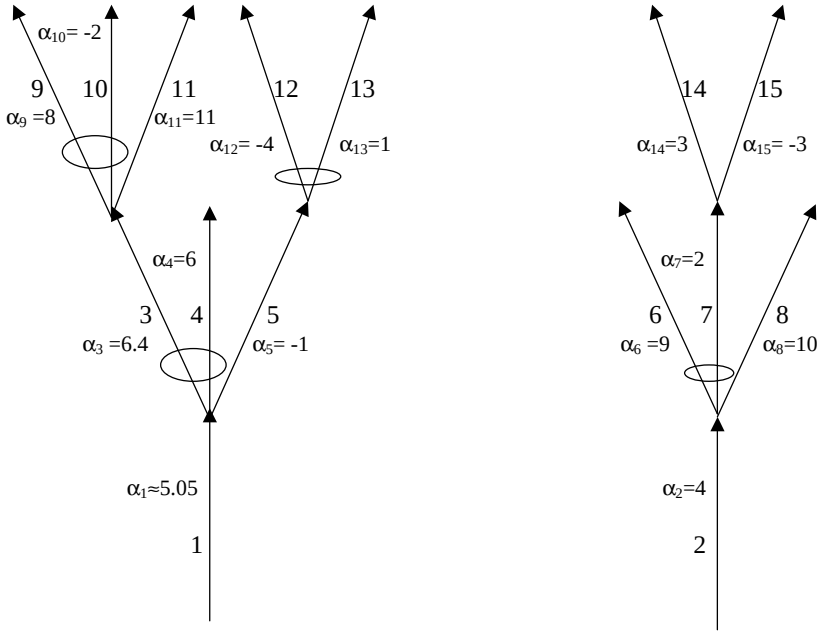


**Fig. 1.** Example of a forest with $\gamma(i) = \alpha_i$.

Edges 1 - 3, 5, 7 are stems, $N(1) = \{3, 4, 5\}$, $N(2) = \{6, 7, 8\}$, $N(3) = \{9, 10, 11\}$, $N(5) = \{12, 13\}$, $N(7) = \{14, 55\}$. Edges 4, 6, 8 - 15 are leafs, so that $N(j) = \emptyset$ for $j = 4, 6, 8 - 15$. $p_{\{3,4\}}(1) > 0$, $p_{\{5\}}(1) > 0$, $p_{\{6,7\}}(2) > 0$, $p_{\{8\}}(2) > 0$, $p_{\{9,10\}}(3) > 0$, $p_{\{11\}}(3) > 0$, $p_{\{12,13\}}(5) > 0$, $p_{\{14\}}(7) > 0$, $p_{\{15\}}(7) > 0$, $p_{\emptyset}(j) > 0$ for all $j = 1, \dots, 15$, $p_D(j) = 0$ for all other subsets of $N(j)$, $j = 1, 2, 3, 5, 7$. Let $\pi_0 = (11, 8, 6, 9, 3, 4, 1, 2, 14, 7, 13, 5, 10, 15, 12)$. Although $11, 8, 6, 3, 9$ are senior then 1 for $\pi_0$, DM will use 1 earlier than these edges because at the initial state $\{1, 2\}$ edge 1 is senior among available. All trajectories of maximal length corresponding to $\pi_0$ and having positive probabilities are given on Fig.2. In each state an exit action $e_*$ is also available so there are also shortened trajectories. In Fig. 2 edges in states are listed in the order of seniority in $\pi_0$.

It follows from Fig. 2 that a list $\pi_1 = (6, 8, 9, 3, 11, 4, 1, 7, 2, 14, 10, 5, 13, 15, 12)$ defines the same PR as $\pi_0$.

Each PR can also be specified as follows. Let $\gamma = \gamma(e)$ be a function defined on edges from $F_0$. Then by definition an edge $e$ is senior than $e'$ if $\gamma(e) > \gamma(e')$. For simplicity we assume that if $e, e' \in x$ for some state $x$ and $e \neq e'$ then $\gamma(e) \neq \gamma(e')$. In opposite case we assume that from the very

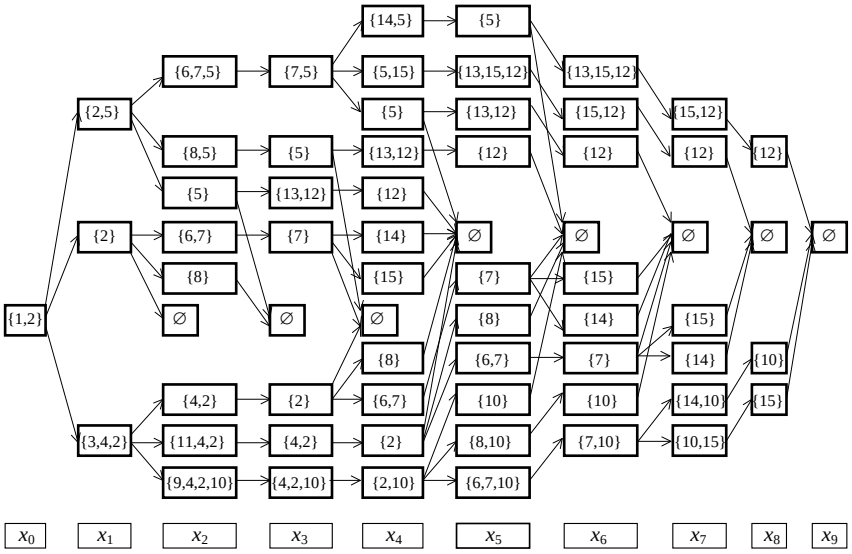


**Fig. 2.** Possible trajectories of maximal length corresponding to $\pi_0$

beginning all edges are numbered and for the edges with equal values of $\gamma(\cdot)$ a senior is with greater initial number. We call a strategy $\pi$ a $(\gamma, c)$-PR if $\{\pi\} = \{e : \gamma(e) \geq c\}$. In other words $\pi$ assigns to use each time the edge with highest value of $\gamma(e)$ among all available with values greater or equal to $c$, and use $e_*$ if there is no available edges with $\gamma(e) \geq c$. The value $c$ is called a *cutoff value*.

Below in Section 8 we consider concrete values of $p,\ q$ and $r$ for all edges in the Example. We show that the PR $\pi_0$ is an optimal strategy in problem B and it corresponds in particular to $\gamma(i) = \alpha_i$, where $\alpha_i$ are given in Fig 1, $\alpha_{11} = 11$, $\alpha_8 = 10$, $\alpha_6 = 9$, $\alpha_9 = 8$, $\alpha_3 = 6.4$, $\alpha_4 = 6$, $\alpha_1 \approx 5,09$, $\alpha_2 = 4$, $\alpha_{14} = 3$ , $\alpha_7 = 2$, $\alpha_{13} = 1$, $\alpha_5 = -1$, $\alpha_{10} = -2$, $\alpha_{15} = -3$, $\alpha_{12} = -4$.

Denote the class of all PRs by $\mathit{\Pi}$.

For any $x \in S$, $x \neq \emptyset$ or $x_*$ let us define $F(x) = \bigcup_{e \in x} T(e)$. Given $x \in S$ and $\pi \in \mathit{\Pi}$ let us define

$$F^{\pi}(x) = \Big\{ e :\ P_x^{\pi}\{A_n = e\} > 0 \text{ for some } n \geq 0 \Big\}. \tag{5}$$

Note that $F^{\pi}(x)$ is also a forest, but some of its leaves can be stems for the initial forest $F_0$. If $x = \{e\}$ then $F^{\pi}(e)$ is a tree and we will denote it $T^{\pi}(e)$. Here and in what follows we use the same notation for a forest $F$ and for the set of edges of $F$. We say that $\pi \in \mathit{\Pi}(x)$ if $\{\pi\} = F^{\pi}(x)$. Given $x \in S$ and $\pi \in \mathit{\Pi}$ we always can assume that $\pi \in \mathit{\Pi}(x)$ eliminating "inaccessible" edges, i.e. such $e \in \{\pi\}$ that $P_x^{\pi}\{A_n = e\} = 0$ for all $n$. If $x = \{e\}$, i.e. $x$ consists only of one edge, we use notation $e$ instead of $\{e\}$, for example we write

$\Pi(e), R^\pi(e)$, $P_e^\pi$ and so on. Thus if $\pi$ is a $(\gamma, c)$-PR and $\pi \in \Pi(e)$ it means that $\{\pi\}$ contains only those edges $e'$ with $\gamma(e') \geq c$ which are accessible from $e$.

For example, PR $\pi_2 = (1, 3, 10)$ in Fig. 1 defines the same PR as $\pi_3 = (1, 3, 10, 12)$ but only $\pi_2 \in \Pi(x)$ for $x = (1, 2)$.

On a set of trajectories $h = (x_0, e_0, x_1, \dots, )$ let us define a stopping time $\tau_* = \tau_*(h) = \min(n : A_n = e_* \text{ or } X_n = x_*)$. Since forest $F_0$ is finite and any PR uses quit action $e_*$ if there is no available actions, we always have $P_x^\pi\{A_{\tau_*} = e_* \text{ or } X_{\tau_*} = x_*\} = 1$, for any $x \in S$ and $\pi \in \Pi(x)$. Thus $\tau_*$ can be described as a random time when either the system runs out of edges in $F^\pi(x)$, and therefore at this moment an action $e_*$ was chosen (a *quit moment*), or at a previous moment some edge $e \neq e*$ from $F^\pi(x)$ was chosen and the transition to $x_*$ has occurred now (at a *termination moment).* For the sake of brevity we call $\tau_*$ an *exit time.* Since $r(e_*) = 0$, we have obviously $R^\pi(x) = E_x^\pi \sum_{i=0}^{\tau_*-1} r(A_i)$. For any initial state $x$ and PR $\pi$ let us define

$$Q^\pi(x) = P_x^\pi\{X_{\tau_*} = x_*\}, \qquad \alpha^\pi(x) = \frac{R^\pi(x)}{Q^\pi(x)}, \tag{6}$$

where $\alpha^\pi(x) = -\infty$ if $Q^\pi(x) = 0$.

Note that the probability of final absorption, i.e. $\lim_n P_x^\pi(X_n = x_*)$ equals to 1 for any PR $\pi$. The value $Q^\pi(x)$ is the probability of *termination*, i.e. probability of transition to $x_*$ without using a quit action $e_*$. Thus $Q^\pi(x) \geq 0$ and $-\infty \leq \alpha^\pi(x) \leq \infty$.

Now we define index $\alpha(e)$ for all $e$. As it was done in [3], we could define it $\alpha(e) = \sup_\pi R^\pi(e)/Q^\pi(e)$ over all $\pi \in \Pi(e)$, but it is more convenient to specify $\alpha(e)$ recursively as follows. For any leaf $e$ we set $\alpha(e) = r(e)/q(e)$ if $q(e) > 0$. If $q(e) = r(e) = 0$ then we set $\alpha(e) = 0$. If $q(e) = 0$, $r(e) > 0$ or $r(e) < 0$ we set $\alpha(e) = +\infty$ (or $-\infty$ correspondingly. For stems we define $\alpha(e)$ as follows. If $\alpha(\cdot)$ is not defined for $e$ but is defined for all other elements of $T(e)$ we set $\alpha(e) = \sup_c \alpha^{\pi_c}$, where $\pi_c \equiv \pi_c(e)$ is a PR which first tests $e$ and after that uses $(\alpha, c)$ -PR from $\Pi(N(e))$. Let us denote by $\pi_*(e)$ the PR where $\alpha(e)$ is attained. We also will call such PR $\alpha$ -optimizer.

**Auxiliary Problem C**$(e)$**:** *For an edge $e$ to find* $\pi_*(e)$ and $\alpha(e)$.

Later we present an algorithm to calculate $\alpha(e)$. It requires no more than $n^2$ operations.

To slightly simplify our proofs sometimes we will assume

**A uniqueness assumptions U:** $\alpha(e) \neq 0$ for all $e$, and if $e \neq e'$ then $\alpha(e) \neq \alpha(e')$.

**Theorem 1**. (a) *An $(\alpha, 0)$-PR is an optimal strategy in the Main Problem A*;

(b) *an $(\alpha, -\infty)$-PR is an optimal strategy in the Main Problem B*;

(c) *an $(\alpha, \alpha(e))$-PR $\pi, \pi \in \Pi(e)$ is an optimal strategy in the Auxiliary Problem C$(e)$.*

*Under the assumption U the optimal strategies in* (a), (b), *and* (c) *are unique.*

If assumption U is not true we can modify the notion of $\alpha$-PR so that statements (a)-(c) of Theorem 1 will still hold.

## 5 One simple idea and three elementary situations.

In this section we describe heuristically the key elements of the proof. There are different proofs of Gittins result (see an interesting paper [14]) but it seems none of them can be immediately applied to our case. At the same time our solution is based on a simple key idea, though its implementation in the case of a random forest is technically cumbersome, and will be presented in the next section. We describe this idea using as illustrations three elementary situations, which can be described as three elementary forests. For the simplicity we will assume that all rewards are positive so a quit action is not at all.

*The first situation* (a) describes in fact the simplest case of Mitten elementary model when there are two interchangeable actions $a_1$ and $a_2$. If used, an action $a_i$ brings a reward $r_i$ and after that with probability $q_i$ the other action becomes unavailable (the process is terminated), with complimentary probability decision process may continue. This situation can be described by a forest consisting of two trees $\{e_1\}$ and $\{e_2\}$. We must compare two PR $\pi_{ij}, i, j = 1, 2, i \neq j$ with corresponding expected rewards $R_{ij}$. In this case it is optimal to use first an action with highest index $\alpha_i = r_i/q_i$. This statement can be checked easily algebraically, but we prefer to demonstrate this as follows.

First, note that the corresponding probability of termination is the same for the both orderings, i.e. we have

$$Q_{12} = q_1 + (1 - q_1)q_2 = q_2 + (1 - q_2)q_1 = Q_{21}. \tag{7}$$

This important property in a general situation is proved in Lemma 1 in Section 6. This property implies that to maximize $R_{ij}$ is the same as to maximize $\alpha_{ij} = R_{ij}/Q_{ij}$. Let us consider

$$\alpha_{12} = \frac{r_1 + (1 - q_1)r_2}{q_1 + (1 - q_1)q_2} = \frac{\alpha_1 q_1 + \alpha_2(1 - q_1)q_2}{q_1 + (1 - q_1)q_2}. \tag{8}$$

It is easy to see that this is a formula for a **center of gravity** of two masses $q_1$ and $(1 - q_1)q_2$ located on a horizontal axis with coordinates $\alpha_1$ and $\alpha_2$. The formula for $\alpha_{21}$ corresponds to a **center of gravity** for masses $(1 - q_2)q_1$ and $q_2$ with the same coordinates $\alpha_1$ and $\alpha_2$. Since the sum of masses is the same for both cases, the center of gravity will have higher value when larger mass will be placed into higher position, i.e.

$$\alpha_{12} > \alpha_{21} \text{ iff } \alpha_1 > \alpha_2. \tag{9}$$

We described situation (a) for two actions but this case implies also that the similar statement is true for any $m$ interchangeable actions, i.e. for BE model. This property for a general situation corresponds to Corollary 2, presented at Section 6.

It is important to observe that the reasoning above does not depend on whether each actions $a_i$ is really one time action or consists of a series of actions. In the latter case we must calculate corresponding quantities $R$ and $Q$ for the whole series.

Let us explain heuristically how the index $\alpha(e)$ should be calculated for the *situation* (b), when some action is followed by a set of actions, i.e. when a forest consists of a tree $T_1 = \{e_0, e_1, e_2, ..., e_m\}$, where $N(e_0) = \{e_1, e_2, ..., e_m\}$, $N(e_i) = \emptyset, i = 1, ..., m$, and $p_0 := p_{N(e_0)}(e_0) = 1 - q(e_0) - p_\emptyset(e_0)$. The indices for the leaves of this tree, $\alpha_i := \alpha(e_i), i = 1, 2, .., m$ are known, $\alpha(e_i) = r_i/q_i$, where $r_i := r(e_i), q_i := q(e_i)$. Without loss of generality we assume that edges are numbered in such a way that $\alpha_1 > \alpha_2 > ... > \alpha_m$.

According to definition, to find $\alpha(e_0)$ we have to choose $k_*$, possibly equal to zero, that maximizes $\alpha^k = R^k/Q^k$, where $R^k$ and $Q^k$ are the reward and termination probability for a PR $\pi_k = (e_0, e_1, e_2, ..., e_k)$. Using notation $\beta_0 = r_0/q_0$, we obtain

$$\alpha^k = \frac{r_0 + r_1 p_0 + r_2 p_0 p_1 + \ldots + r_k \prod_{i=0}^{k-1} p_i}{q_0 + q_1 p_0 + q_2 p_0 p_1 + \ldots + q_k \prod_{i=0}^{k-1} p_i} = \frac{\beta_0 m_0 + \alpha_1 m_1 + \ldots + \alpha_k m_k}{m_0 + m_1 + \ldots + m_k}, \tag{10}$$

where $m_0 = q_0$, $m_i = (p_0 \cdots p_{i-1}) q_i, i = 1, ..., k$. Thus expression $\alpha^k$ also represents a position of a **center of gravity** for a system of masses and to find the value $k$ which brings the maximum value to (10) we can use the following

**Proposition 1**. *Suppose that $m_i$ are the masses and $\alpha_i$ the positions of these masses on the real line, $i = 0, 1, 2, ..., N$, and $\alpha_1 > \alpha_2 > ... > \alpha_N$. Suppose that our goal is to select a subset $J_{\max}$ of a set $\{0, 1, ..., N\}$ which contains a subset $J_0 = \{0\}$ and has the largest possible center of gravity. Then*

*a) $J_{\max}$ can be obtained by adding sequentially masses $m_1, m_2, \cdots$, to a set $J_0 = \{0\}$ till the center of gravity of a system $J_k = \{0, 1, ..., k\}$ will stop to increase;*

*b) $J_{\max} = \{0\} \cup \{i : \alpha_* < \alpha_i\}$, where $\alpha_*$ is the center of gravity of $J_{\max}$. If there are $\alpha_i = \alpha_*$ then $J_{\max}$ is not unique in an obvious way.*

Note that both points of Proposition 1 describe the optimal set: b) describes it in inexplicit form, since $\alpha_*$ *is not known yet,* and a) describes it algorithmically and allows one to calculate $\alpha(e_0)$ in situation b) sequentially step by step.

The proof of Proposition 1 follows from the elementary properties of proportions. (A similar statement was used in a paper by Sonin [11]).

The simplest version of situation b) for $m = 1$ gives

$$\alpha^1 > \beta_0 \text{ iff } \alpha_1 > \beta_0. \tag{11}$$

The proof of Theorem 1 in Section 7 is based on the induction with respect to the number of edges, and on Lemma 1, which corresponds to (7), Corollary 1, which corresponds to (9), and Corollary 2, which corresponds to (11). These statements are more general than (7), (9), (11) because each action in Lemma and corollaries consists of some series of actions and after application some action (which corresponds to some PR) the system transits to a random set and the choice of the next action depends on this set.

To illustrate this fact and an algorithm of calculation of $\alpha(e)$ consider the more complicated *situation* c), when in situation b) one of leaves $e_1, e_2, ..., e_m$, let say an edge $e_3$, is replaced by a tree $T(e_3)$. Then the first two steps of our procedure of maximization of center of gravity will be the same. Suppose that the value of $\alpha(e_3)$ is achieved on some PR $\pi = (e_3, v_1, ..., v_k)$ and $\alpha(e_3) = R_3/Q_3$. Then in formula (10) the value $r_3$ should be replaced by $R_3 = \alpha(e_3)Q_3$ and correspondingly the mass $m_3$ will be also modified. After that the set $N(e_3)$ will be added to the set of available edges, where $N(e_3)$ is the set of elements of $T(e_3)$ which does not belong to $\pi$, but follows immediately elements of $\pi$. By the property of $\alpha$ optimizer, all elements of $N(e_3)$ have the values of index less then $\alpha(e_3)$, and on the next step we will choose an edge with maximal value of $\alpha$ in enlarged set of available edges.

## 6 Auxiliary results

To prove Theorem 1 we introduce some new notations and prove some auxiliary statements.

Let $\pi_1$ and $\pi_2$ are PR and $\pi_1 \in \Pi(x)$. Let us define a new PR from $\Pi(x)$ - we denote it $\pi = (\pi_1, \pi_2)$ - which uses first all available edges from $\pi_1$ and after that switches to $\pi_2$, i.e. all edges in the list $\pi_1$ are defined now as senior than all edges in $\pi_2$. The list $\pi$ can be obtained as follows. First, list all elements of $\pi_1$ in their order and after that list those elements of $\pi_2$ - in their order - which does not belong to $\pi_1$ and which are accessible from $x$. We call PR $\pi_2$ a *continuation* of $\pi_1$. The similar meaning has notation $\pi = (\pi_1, \pi_2, \pi_3)$ and so on.

**Remark 1**. *Let $\pi$ be a $(\gamma, c)$-PR and $\pi_1$ be a $(\gamma, c_1)$-PR, where $c_1 > c$. Then obviously $\pi$ can be represented as $\pi = (\pi_1, \pi_2)$, where $\pi_2$ is a $(\gamma, c)$-PR.*

For a PR $\pi = (\pi_1, \pi_2)$ let us define a random time $\sigma = \min(n : X_n = x_*$ or $A_n \in \{\pi_2\})$, i.e. a time of termination or first usage of edges from $\pi_2$. For the sake of brevity we call time $\sigma$ a *time of switching* from $\pi_1$ to $\pi_2$.

**Remark 2**. *Note that for any trajectory $\sigma \leq \tau_*$, but at the same time $P_x^{\pi_1}\{X_{\tau_*} = y\} = P_x^{\pi}\{X_\sigma = y\}$ for any $y$. Equivalently, a moment of termination for $\pi_1$ is a moment of switching from $\pi_1$ to $\pi_2$ in $\pi$.*

Using strong Markov property and the total probabilities formula it is easy to obtain for a $\pi = (\pi_1, \pi_2)$

$$R^{\pi}(x) = E_x^{\pi_1}\left[\sum_{i=0}^{\sigma-1} r_i + R^{\pi_2}(X_\sigma)\right] = R^{\pi_1}(x) + \sum_y P_x^{\pi_1}(X_\sigma = y)R^{\pi_2}(y). \quad (12)$$

**Lemma 1**. *If* $\pi_1, \pi_2 \in \Pi(x)$ *and* $\{\pi_1\} = \{\pi_2\}$, *then*

$$P_x^{\pi_1}\{X_{\tau_*} = y\} = P_x^{\pi_2}\{X_{\tau_*} = y\} \quad (13)$$

*for all* $y \in S$, *and, in particular, for* $y = x_*$, *i.e.* $Q^{\pi_1}(x) = Q^{\pi_2}(x)$.

This lemma is an analog of the simple statement that for a set of independent trials the probability of at least one success does not depend on the order in which these trials are tested. We prove this lemma in an Appendix.

Let us call PRs $\pi_1$ and $\pi_2$ *disjoint* if $\pi_1 \in \Pi(x_1), \pi_2 \in \Pi(x_2)$, and $F(x_1) \cap F(x_2) = \emptyset$.

Let $\pi_1 \in \Pi(x_1)$ and $\pi_2 \in \Pi(x_2)$ are disjoint and $\pi \in \Pi$. Then for any $x$, $x_1 \cup x_2 \subset x$ we can define PRs $\pi_{12} = (\pi_1, \pi_2, \pi)$ and $\pi_{21} = (\pi_2, \pi_1, \pi)$ such that both belong to $\Pi(x)$. Where no confusion is possible we will use shorthand notations $R^{\pi_i}(x) = R_i$, $Q^{\pi_i}(x) = Q_i, \alpha^{\pi_i}(x) = \alpha_i$ and so on.

**Lemma 2**. *Consider two PRs* $\pi_{ij} = (\pi_i, \pi_j, \pi) \in \Pi(x)$, $i, j = 1, 2, i \neq j$, *where* $\pi_1$, $\pi_2$ *are disjoint, and* $\pi_i \in \Pi(x_i)$. *Then for any* $x$, $x_1 \cup x_2 \subset x$

$$R_{ij} = R_i + d_i R_j + R, \quad (14)$$

*where* $d_i = 1 - Q_i$, *and the term* $R$ *is the same for both* $\pi_{12}$ *and* $\pi_{21}$.

Proof. Given PR $\pi_{ij} = (\pi_i, \pi_j, \pi)$ let us define $\sigma_i$ as the switching moment from $(\pi_i, \pi_j)$ to $\pi$. Since $\pi_1$ and $\pi_2$ are disjoint we have $\{(\pi_1, \pi_2)\} = \{(\pi_2, \pi_1)\}$ and therefore by Lemma 1 the distributions $P_x^{\pi_{ij}}\{X_{\sigma_i} = y\}$ coincide. Hence, according to (12) the term $R$ is the same for both $\pi_{12}$ and $\pi_{21}$. The equality in Lemma 3 follows from formula (12) applied to moments $\tau_i$ of switching from $\pi_i$ to $(\pi_j, \pi)$ and the fact that for disjoint PRs the second factor of each term in the sum $\sum_y P_x^{\pi_i}(X_{\tau_i} = y)R^{\pi}(y)$ is the same for all $y$ such that $y \neq x_*$ and $P_x^{\pi_i}(X_{\tau_i} = y) \neq 0$.

Note that any equality for $R$ always implies similar equality for $Q$ because $Q^{\pi} = R^{\pi}$ if all rewards $r(e)$ are put equal $r(e) = q(e)$. Indeed, let us consider a reward function $r'(e, x)$ defined by $r'(e_i, x_{i+1}) = 1$ if $e_i \neq e_*, \quad x_{i+1} = x_*$, and $r'(e_i, x_{i+1}) = 0$ otherwise. Then for such function we have $Q^{\pi}(x) = R^{\pi}(x)$. It remains to note that averaging of such $r'$ gives $r(e_i) = q(e_i)$.

Therefore, we have an equality similar to (14) for $Q$, and hence

$$\alpha_{ij} = \frac{\alpha_i Q_i + \alpha_j d_i Q_j + R}{Q_i + d_i Q_j + Q}. \quad (15)$$

**Corollary 1**. *If under assumptions of Lemma 2* $\alpha_1 > \alpha_2$ *then* $\alpha_{12} > \alpha_{21}$ *(and therefore* $R_{12} > R_{21}$*).*

Proof. This follows from (14) and (15), using equality $Q_1 + (1 - Q_1)Q_2 = Q_2 + (1 - Q_2)Q_1$.

The next lemma shows how the "isolated tail" of a PR $\pi$ contributes to the value of $R^\pi$. If $\pi \in \Pi(x)$ we will omit sometimes the dependence on $x$ of $R$, $Q$ and $\alpha$.

**Lemma 3**. *Let* $\pi_1 \in \Pi(x)$, $\pi_2 \in \Pi(e), e \notin \{\pi_1\}$, $\pi = (\pi_1, \pi_2)$. *Then*

$$R^\pi(x) = R^{\pi_1}(x) + d_1 R^{\pi_2}(e), \tag{16}$$

*where* $d_1 = P_x^{\pi_1}\{e \in X_\sigma\}$.

Proof follows directly from the second equality in (12) and because $R^{\pi_2}(y) = R^{\pi_2}(e)$ for $e \in y$, and $R^{\pi_2}(y) = 0$ if $X_\sigma = y$ and $e \notin y$. Note that the assumption $\pi_2 \in \Pi(e)$ is crucial for validity of (16).

According to our remark after Lemma 2, Lemma 3 implies that the formula similar to (12) (with replacement $R$ by $Q$) holds for $Q^\pi$, and hence we have

$$\alpha^\pi = \frac{R_1 + d_1 R_2}{Q_1 + d_1 Q_2} = \frac{\alpha_1 Q_1 + \alpha_2 d_1 Q_2}{Q_1 + d_1 Q_2}. \tag{17}$$

Formula (17) and elementary properties of proportions imply

**Corollary 2**. *Under the assumptions of Lemma 3 either* $\alpha^{\pi_1} = \alpha^{\pi_2} = \alpha^\pi$ *or*

$$\min\{\alpha^{\pi_1}, \alpha^{\pi_2}\} < \alpha^\pi < \max\{\alpha^{\pi_1}, \alpha^{\pi_2}\}. \tag{18}$$

## 7 Proof of Theorem 1

We prove theorem 1 by induction on the number $k$ of edges in the forest $F(x)$ of an initial state $x$. We denote by $|C|$ the number of elements in a finite set $C$. For $k = 1$ the theorem is trivial. Suppose it is proved for all $x$ with $|F(x)| \le k$, and suppose an initial state is $x$ with $|F(x)| = k + 1$. We consider separately two cases: (A) when $|x| > 1$, and (B) when $|x| = 1$. In both cases we will use a well-known Bellman Optimality Principle, a corollary of a Bellman equation for the expected total reward: *if* $\pi$ is an optimal strategy (for the problem A or B) for an initial state $x$, then after the first step it remains optimal for all states that follow $x$. We prove theorem under the Uniqueness assumption U. The proof for the general case is similar.

Case (A). In this case point (c) of the theorem is trivial since each $|T(e)| \le k$ for each $e \in F(x)$ so, it remains to prove (a) and (b). For any $e \in x$ let $\pi_0$ be an $\alpha$-PR (with cutoff value $c = 0$ in Problem A and cutoff value $c = -\infty$ in Problem B). According to the induction assumption it is an optimal PR for any state in $F(x)\backslash e$. So, if $\pi$ is optimal on $F(x)$, and applies $e$ on the first step, by Optimality Principle, PR $(e, \pi_0)$ is also optimal. Let $\alpha_1 = \alpha(e_1) = \max_{e \in x} \alpha(e)$. Let us show that $\pi = (e, \pi_0)$ is not optimal if $\alpha = \alpha(e) < \alpha_1$.

Using the description of $\pi_0$ by point (a) of Theorem 1 and Remark 1 we have $\pi = (e, \nu_1, \pi_1, \nu)$, where $\nu_1$ is an $\alpha$-PR defined on a set $T(e)\setminus e$ with cutoff value $c_1 = \min_{e' \in T(e)\setminus e} \{e' : \alpha(e') > \alpha_1\} > \alpha_1$; PR $\pi_1$ is an $\alpha$-PR with cutoff value $c = \alpha_1$, and $\nu$ is a continuation of $\alpha$-PR (with cutoff value $c(\nu) = 0$ in Problem A and cutoff value $c = -\infty$ in Problem B). Note that it is possible that $\nu_1 = \emptyset$. According to the definitions of $\alpha$-PR and the value $c_1$, all edges used by $\pi_1$ belong to $T(e_1)$.

Note that PRs $\pi_1$ and $\pi_2 = (e, \nu_1)$ are disjoint because they are defined on different trees $T(e_1)$ and $T(e)$, and that $\alpha^{\pi_2}(e) \le \alpha = \alpha(e)$ because PR $(e, \nu_1)$ can be different than $\pi_e$ which gives a solution to the Auxiliary Problem. Let us show that PR $\varphi = (\pi_1, \pi_2, \nu)$ is better than $\pi = (e, \nu_1, \pi_1, \nu) = (\pi_2, \pi_1, \nu)$. According to the induction assumption $\alpha^{\pi_1}(e_1) = \alpha_1$, so $\alpha^{\pi_1}(e_1) = \alpha_1 > \alpha \ge \alpha^{\pi_2}(e)$. Applying Corollary 1 to $\pi_1$ and $\pi_2$ we obtain that $R^{\varphi} > R^{\pi}$, i.e. $\pi$ is not an optimal strategy. It means that an optimal strategy either coincides with $(e_1, \pi_0)$ or appoints to quit from the very beginning.

Case (B). In this case $x$ consists only of one edge and we denote it $e_0$. The first step for any policy is defined uniquely and the resulting state has a forest with no more than $k$ edges, so by the Optimality Principle the points (a) and (b) of the Theorem are trivial but point (c) is trivial for all edges except $e_0$.

Let $\pi_{e_0} = (e_0, \nu)$, where $\pi_{e_0}$ be a solution of an Auxiliary Problem for $e_0$, $\alpha$-PR $\nu \in \mathit{\Pi}(N(e_0))$ and $c$ is a corresponding cutoff value. Let us show that

1) if $e \in F^{\nu}(e_0)$, then $\alpha(e) \ge \alpha(e_0)$,

2) if $e \notin F^{\nu}(e_0)$ and $e \in N(e')$ for some $e'$ which is a leaf of $F^{\nu}(e_0)$ then $\alpha(e) < \alpha(e_0)$.

This will prove that $c$ can be taken equal to $\alpha(e_0)$, i.e. satisfying point (c).

Suppose that 1) is not true and $e \in F^{\nu}(e_0)$ is such that $\alpha(e) < \alpha(e')$ for all $e' \in F^{\nu}(e_0)$, and $\alpha(e) < \alpha(e_0$.) By the definition of $(\alpha, \alpha(e))$-PR all edges that can be used in $\nu$ after $e$ belong to $T(e)$. So, PR $(e_0, \nu)$ can be represented in a form $\pi = (\pi_1, \pi_2)$ where $\pi_2 \in \mathit{\Pi}(e)$ is an $\alpha$-PR. Consequently $\alpha^{\pi_2}(e) \le \alpha(e) < \alpha(e_0) = \alpha^{(e_0,\nu)}$. But Lemma 3 and Corollary 2 applied to PR $(e_0, \nu) = (\pi_1, \pi_2)$ imply that $\alpha^{(e_0,\nu)} < \alpha^{\pi_1}$. This contradicts to the definition of $\pi(e_0)$.

Suppose that 2) is not true and we select $e \in N(e')$ such that $e'$ is a leaf of $F^{\nu}(e_0)$, $\alpha(e) > \alpha(e_0)$ and $e$ is the smallest among such $e$. Let $\pi_2$ is $(\alpha, \alpha(e))$ -PR, $\pi_2 \in \mathit{\Pi}(e)$. Consider PR $\pi = (\pi_1, \pi_2)$, where $\pi_1 = (e_0, \nu)$. Then $\pi$ is a PR with $c = \alpha(e)$. Applying Lemma 1 and Corollary 2 to PR $\pi$ and using that $\alpha^{\pi_1}(e_0) = \alpha(e_0) < \alpha(e) = \alpha^{\pi_2}(e)$ we obtain that $\alpha(\pi) > \alpha^{\pi_1}$. This contradicts to the definition of $\pi_1$.

## 8 A recursive algorithm to calculate $\alpha(e)$ and $\pi_*(e)$.

To formulate the algorithm we first consider the structure of $(\alpha, c)$-PR $\pi^c \in \mathit{\Pi}(x)$ for an initial state $x$. Recall that for any PR $\pi$ and initial state $x$ we

can consider $R^\pi(x)$, $Q^\pi(x)$, $F^\pi(x)$ (or $T^\pi(e)$ if $x$ consists of one edge $e$) (see (5)). We will consider also $N^\pi(x) = N(F^\pi(x))$, where $N(F)$ for any subforest of initial forest $F_0$ denotes the set of all edges that follow immediately "leafs" of $F$, i.e. the set of all edges that do not belong to $F$, but follow immediately elements of $F$. For any $D \subset N^\pi(x)$ (including empty set) we will consider also the probability $p^\pi_D(x) = P^\pi_x\{X_{\tau_*} = D\}$, i.e. the probability that our decision to quit was taken at the state $D$.

**Proposition 2**. *For any $x \in S$ there exist a natural number $k(x)$, non-increasing (decreasing in case of Assumption U) numbers $c_k = c_k(x)$, with $c_0 = +\infty$, and edges $g_k = g_k(x) \in F(x)$, $k = 0, 1, \cdots, k(x)$, such that for $(\alpha, c)$-PR $\pi^c \in \Pi(x)$*

$$\begin{gathered}\pi^c = \pi^{c_k} \text{ for } c_{k+1} < c \le c_k, \quad c_{k+1} = \alpha(g_k),\\ \pi^{c_{k+1}} = (\pi^{c_k}, \pi_*(g_k)), \text{ for } 0 \le k < k(x); \quad \pi^c = \pi^{c_{k(x)}} \text{ for } c \le c_{k(x)},\end{gathered} \tag{19}$$

*where $\pi_*(g_k)$ is $\alpha$-optimizer of $g_k$. Using indices "k" and "*" instead of index $\pi$ for $\pi = \pi^{c_k}$ and $\pi = \pi_*$ correspondingly we get: $\pi^0(x) = (\emptyset)$, $R^0(x) = 0$, $Q^0(x) = 0$, $F^0(x) = (\emptyset)$, $N^0(x) = x$, $p^0_x(x) = 1$ and if $N^k(x) \neq \emptyset$ then*

$$F^{k+1}(x) = F^k(x) \bigcup T_*(g_k), \tag{20}$$

$$N^{k+1}(x) = \big(N^k(x) \setminus g_k\big) \bigcup N_*(g_k), \tag{21}$$

$$R^{k+1}(x) = R^k(x) + R_*(g_k) \sum_{D:\ g_k \in D \subset N^k(x),} p^k_D(x), \tag{22}$$

$$Q^{k+1}(x) = Q^k(x) + Q_*(g_k) \sum_{D:\ g_k \in D \subset N^k(x)} p^k_D(x). \tag{23}$$

*If $D \subset N^{k+1}(x)$ then there exist unique $D_1 \subset N^k(x) \setminus \{g_k\}$ and $D_2 \subset N_*(g_k)$ such that $D = D_1 \bigcup D_2$, and*

$$\text{if } D_1 = \emptyset, D_2 \neq \emptyset, \text{ then } p^{k+1}_D(x) = p^k_{\{g_k\}}(x) p^*_{D_2}(g_k), \tag{24}$$

$$\text{if } D_2 = \emptyset, \text{ then } p^{k+1}_D(x) = p^k_{D_1}(x) + p^k_{\{g_k\} \bigcup D_1}(x) p^*_\emptyset(g_k), \tag{25}$$

$$\text{if } D_1 \neq \emptyset, D_2 \neq \emptyset, \text{ then } p^{k+1}_D(x) = p^k_{\{g_k\} \bigcup D_1}(x) p^*_{D_2}(g_k). \tag{26}$$

Proof. For the sake of simplicity we will prove Proposition 2 under Assumption U. The changes for the general case is straightforward. Let for some $k \ge 0$ we know $c_k$, $\pi^{c_k}$, $R^k(x)$, $Q^k(x)$, $F^k(x)$, $N^k(x)$, and $p^k_D(x)$ for any $D \subset N^k(x)$. The set $N^k(x)$ corresponds to all potentially available edges after application of $\pi^{c_k}$. If $N^k(x) = \emptyset$ then $k = k(x)$ and evidently we obtain the last equality in (19). If $N^k(x) = \emptyset$ then according to the definition of $(\alpha, c)$-PR, all elements of $N^k(x)$ have the value of $\alpha$ less or equal to $c_k$. Consider the edge in $N^k(x)$ with maximal value of $\alpha$ and denote it $g^k$. Denote $c_{k+1} = \alpha(g^k)$. Since there is no edges in $N^k(x)$ with $c_{k+1} < \alpha(e) < c_k$ we have

proved the first equality in (19). According to Remark 1 $\pi^{c_{k+1}}(x) = (\pi^{c_k}, \pi_2)$, where $\pi_2 \in \Pi(g^k)$ is $(\alpha, \alpha(g^k))$-PR. according to statement c) of Theorem 1 this PR coincides with $\pi_*(g^k)$. It proves third equality in (19) and equalities (20), (21. Equalities (22)-(26) are the results of application of total probability formula. It completes the proof of Proposition 2.

Note that if $\alpha(e)$ is known for all $e \in F(x)$ then Proposition 2 gives the algorithm for calculation of optimal value of functional in Main Problems A and B. In case of Problem B it coincides with $R^{k(x)}(x)$, and in case of Problem A it coincides with $R^{k_0}(x)$, where $k_0 = \inf\{k : \alpha(g^{k-1}) > 0\}$.

Now we can formulate algorithm for finding $\alpha(e)$. Recall that we defined $\alpha(e)$ as $r(e)/q(e)$ for leaves, and if $\alpha(e')$ is defined for all $e' \in T(e) \setminus e$ then as a maximum of $R^{\pi_c}(e)/Q^{\pi_c}(e)$ over $c$, where $\pi_c \equiv \pi_c(e)$ is a PR which first tests $e$ and after that uses $(\alpha, c)$ -PR from $\Pi(N(e))$. It is evident that Proposition 2 is valid also for $\pi_c(e)$ with initial values $c_0 = +\infty$, $\pi^0(e) = (e)$, $R^0(e) = r(e)$, $Q^0(e) = q(e)$, $\alpha^0(e) = R^0(e)/Q^0(e)$, $T^0(e) = \{e\}$, $N^0(e) = N(e)$, $p^0_D(e) = p_D(e)$ for all $D \subset N^0(e)$. Define $\alpha^k(e) = R^k(e)/Q^k(e)$. According to Corollary 2 (see also Proposition 1 and (11)) there exists $k_* = k_*(e)$ such that $\alpha^k(e)$ increases for $k < k_*$ and decreases for $k > k_*$ and $k_* = \inf\{k : \alpha(g_k) \le \alpha^k\}$. It means that for finding $\alpha(e)$ we need to conduct calculations (22)-(26) sequentially from $k = 0$ till the time when $\alpha(g_k) < \alpha^k$ and set $\alpha(e) = \alpha^{k_*}$.

Note that if $e \in \pi_*(e')$ for some $e'$, then we do not need to remember all data for $e$. We need remember only the data for $e'$.

Consider now example 1 with

$q(1) = 0.2$, $p_\emptyset(1) = 0.1$ $p_{\{3,4\}}(1) = 0.4$, $p_{\{5\}}(1) = 0.3$, $r(1) = 0.8$;

$q(2) = 0.08$, $p_\emptyset(2) = 0.17$, $p_{\{6,7\}}(2) = 0.5$, $p_{\{8\}}(2) = 0.25$, $r(2) = 0.1$;

$q(3) = 0.1$, $p_\emptyset(3) = 0.24$, $p_{3,\{9,10\}}(3) = 0.5$, $p_{\{11\}}(3) = 0.16$, $r_3 = 0.2$;

$q(4) = 0.3$, $p_\emptyset(4) = 0.7$, $r(4) = 1.8$; $q(6) = 0.04$, $p_\emptyset(6) = 0.96$, $r(6) = 0.36$;

$q(5) = 0.24$, $p_\emptyset(5) = 0.71$, $p_{\{12,13\}}(5) = 0.05$, $r(5) = -0.3$;

$q(7) = 0.05$, $p_\emptyset(7) = 0.45$, $p_{\{14\}}(7) = 0.5$, $p_{\{15\}}(7) = 0.3$, $r(7) = 0.05$;

$q(8) = 0.08$, $p_\emptyset(8) = 0.92$, $r(8) = 0.8$; $q(9) = 0.09$, $p_\emptyset(9) = 0.91$, $r(9) = 0,72$;

$q(10) = 0.7$, $p_\emptyset(10) = 0.3$, $r(10) = -1.4$; $q(11) = 0.5$, $p_\emptyset(11) = 0.5$, $r(11) = 5.5$;

$q(12) = 0.2$, $p_\emptyset(12) = 0.8$, $r(12) = -0.8$; $q(13) = 0.6$, $p_\emptyset(13) = 0.4$, $r(13) = 0.6$;

$q(14) = 0.01$, $p_\emptyset(14) = 0.99$, $r(14) = 0.3$; $q(15) = 0.4$, $p_\emptyset(15) = 0.6$, $r(15) = -1.2$.

For leaves we have:

$$\alpha(4) = \frac{r(4)}{q(4)} = 6,\ \alpha(6) = \frac{r(6)}{q(6)} = 9,\ \alpha(8) = \frac{r(8)}{q(8)} = 10,\ \alpha(9) = \frac{r(9)}{q(9)} = 8,$$

$$\alpha(10) = \frac{r(10)}{q(10)} = -2,\ \alpha(11) = \frac{r(11)}{q(11)} = 11,\ \alpha(12) = \frac{r(12)}{q(12)} = -4,$$

$$\alpha(13)=\frac{r(13)}{q(13)}=1,\ \ \alpha(14)=\frac{r(14)}{q(14)}=3,\ \ \alpha(15)=\frac{r(15)}{q(15)}=-3.$$

To calculate values of $\alpha$ for stems we use the algorithm.

$\alpha^0(3)=\frac{r(3)}{q(3)}=2$. Since $N(3)=\{9,10,11\}$ and $\alpha(11)=11>\alpha(9)=8>\alpha^0(3)>\alpha(10)=-2$, we set $g^0(3)=11$. Since $N(11)=\emptyset$ we have from (21)-(23): $N^1(3)=\{9,10\}$, $R^1(3)=r(3)+p_{\{11\}}(3)r(11)=0.2+0.16*5.5=1.08$, $Q^1_3=q_3+p_{3,\{11\}}q_{11}=0.1+0.16*0.5=0.18$. Using (25) we get: $p^1_{\{9,10\}}(3)=p_{\{9,10\}}(3)=0.5$, $p^1_\emptyset(3)=p_\emptyset(3)+p_{\{g\}}(3)p^*_\emptyset(11)=0.24+0.16*0.5=0.32$, $\alpha^1(3)=\frac{R^1(3)}{Q^1(3)}=\frac{1.08}{0.18}=6$.

Since $N^1(3)=\{9,10\}$ and $\alpha(9)=8>\alpha^1(3)>\alpha(10)=-2$, we set $g^1=9$. Since $N(9)=\emptyset$ we have from (21)-(23): $N^2(3)=\{10\}$, $R^2(3)=R^1(3)+p^1_{\{9,10\}}(3)r(9)=1.08+0.5*0.72=1.44$, $Q^2(3)=Q^1(3)+p^1_{\{9,10\}}(3)q(9)=0.18+0.5*0.09=0.225$. Using (25) we get: $p^2_{\{10\}}(3)=p^1_{\{9,10\}}(3)p_\emptyset(9)=0.5*0.91=0.455$, $p^2_\emptyset(3)=p^1_\emptyset(3)=0.32$, $\alpha^2(3)=\frac{R^2(3)}{Q^2(3)}=\frac{1.44}{0.225}=6.4$.

Since $N^2(3)=\{10\}$ and $\alpha(10)=-2<\alpha^2(3)=6.4$ we have: $\pi_*(3)=\pi_8(3)=(3,11,9)$, $N_*(3)=N^2(3)=\{10\}$, $R_*(3)=R^2(3)=1,44$, $Q^*(3)=Q^2(3)=0.225$, $p^*_{\{10\}}(3)=p^2_{\{10\}}(3)=0.455$, $p^*_\emptyset(3)=p^2_\emptyset(3)=0.32$, $\alpha(3)=\alpha^2(3)=6.4$.

Calculations for the edges 5,7,1, and 2 are absolutely analogous and we omit them. This calculations give:

$\pi_*(5)=\pi_1(5)=(5,13)$, $N_*(5)=\{12\}$, $R_*(5)=-0.27$, $Q^*(5)=0.27$, $p^*_{\{12\}}(5)=0.02$, $p^*_\emptyset(5)=0.71$, $\alpha(5)=-1$;

$\pi_*(7)=\pi_3(7)=(7,14)$, $N_*(7)=\{15\}$, $R_*(7)=0.2$, $Q^*(7)=0.1$, $p^*_{\{15\}}(7)=0.3$, $p^*_\emptyset(7)=0.6$, $\alpha(7)=2$;

$\pi_*(1)=\pi_{6.4}(1)=(1,3,11,9,4)$, $N_*(1)=\{5,10\}$, $R_*(1)=1,934$, $Q^*(1)=0.383$, $p^*_{\{10\}}(1)=0.1274$, $p^*_{\{5\}}(1)=0.3$, $p^*_\emptyset(1)=0.1896$, $\alpha(1)=\approx 5.05$;

$\pi_*(2)=\pi_9(2)=(2,8,6)$, $N_*(2)=\{7\}$, $R_*(2)=0,48$, $Q^*(2)=0.12$, $p^*_{\{7\}}(2)=0.48$, $p^*_\emptyset(2)=0.4$, $\alpha(2)=4$.

## 9 Connection with Gittins index and Concluding Remarks.

Now we outline how to obtain the proof of the celebrated Gittins result from Theorem 1. Suppose that there is a fixed number $m$ of finite Markov chains with transition probabilities $p_k(i,j), j=1,2,...,m$ and a discount factor $\beta, 0<\beta<1$. Each time a DM can engage one of these MC and a reward $r_k(i)$ is obtained if $k$-th MC was engaged at state $i$. Without loss of generality these MCs have common state space $S=\{1,2,...,N\}$ and we can describe the

possible transitions of these MCs using *infinite forest* $F_0$ which consists of $m$ trees $T_1, ..., T_m$. The set $N(e) = \{e_1, ..., e_N\}$ and partitions of $N(e) = \{e_1\} \cup \{e_2\} \cup ...\{e_N\}$ are the same for each $e \in F_0$. The probability $p(N_j)$ for an edge $e_i \in T_k$ is equal to $\beta p_k(i,j)$, and $q(e) = (1-\beta)$, i.e. we use a standard way to replace a discount by a transition to an absorbing state. The reward $r(e) = r_k(i)$ if $e = e_i \in T_k$. We can prove that for any given $\varepsilon > 0$ we can specify $n$ sufficiently large so that the value function for an initial problem and a problem with finite forest $F_n$ will be different less than in $\varepsilon$. For such finite forest we can apply Theorem 1 where the optimality of PR based on indices all $\alpha_n(e)$ was established. It can be proved also that if if $e = e_i \in T_k$ then $\lim_{n\to\infty} \alpha_n(e) = \alpha_k(i)$, where $\alpha_k(i)$ is the value of the classical Gittins index (GI) for the $k$-th MC at state $i$. This proves the optimality of PR based on GI.

Note also that the value of GI will be obtained as a limit. At the same time there are algorithms that calculate GI for finite case in a finite number of steps, e.g in [13]. A new recursive algorithm to calculate GI even in a more general model is proposed in [12].

Not also that the idea of an infinite forest can be applied to the case of a countable state space under assumption e.g. that the ratio $r(e)/q(e), e \in F$ is bounded by a constant $c$. Note that this assumption holds for the classical Gittins case if Markov chain is finite or $r(e)$ is bounded if it is countable.

## 10 Appendix.

Proof of Lemma 1. We prove lemma 1 by induction on $n = |\{\pi\}|$. For $n = 1$ lemma is trivial. For $n = 2$ we have $\{\pi_i\} = \{e_1, e_2\}$. If $x$ contains only one of these edges then both PRs use this edge on the first step and the other one on the second, so they coincide. Let $e_i \in x$ for $i = 1, 2$, then there are two possible PRs, $\pi_1 = (e_1, e_2)$, and $\pi_2 = (e_1, e_2)$. From the definition of transition probabilities $P_x^{\pi_i}\{X_{\tau_*} = y\} > 0$ only if either $y = x_*$, or $y$ has a form $y_{kQ} = ((x \setminus (e_1, e_2)) \cup N_k(e_1) \cup N_Q(e_1))$ for some $0 \le k \le j(e_1), 0 \le Q \le j(e_2)$, and $P_x^{\pi_i}\{x_{\tau_*} = y_{iQ}\} = p_i(e_1)p_Q(e_2)$ for $i = 1, 2$. For $y = x_*$ we have $P_x^{\pi_i}\{x_{\tau_*} = x_*\} = 1 - \sum_{y \neq e_*} P_x^{\pi_1}\{x_{\tau_*} = y\}$ for $i = 1, 2$. This completes the proof of Lemma 1 for the case $|\{\pi\}| = 2$.

Suppose now that (13) is proved for $n = k, k \ge 2$, and $|\{\pi_i\}| = k + 1$. Given $x \in S$, denote $e_i$ the senior edge among edges in $x$ for a PR $\pi_i$. Then each $\pi_i$ can be represented as $\pi_i = (e_i, \nu_i)$, where $\nu_i$ is a continuation of $\pi_i$ and $|\{\nu_1\}| = k$. Note that if $e_1 = e_2$ then $\{\nu_1\} = \{\nu_2\}$ and lemma 1 holds because the first step for both PRs will be the same and after the first step we can apply an induction assumption to PRs $\nu_i$. Suppose that $e_1 \neq e_2$. Then let us introduce two new PRs $\pi_1' = (e_1, e_2, \nu)$ and $\pi_2' = (e_2, e_1, \nu)$, where $\nu$ is a PR with $\{\nu\} = \{\pi\} \setminus \{e_1, e_2\}$. For two pairs of PRs; $\pi_1$ and $\pi_1'$, and for $\pi_2$ and $\pi_2'$ lemma 1 holds because each pair has the same first edge and we discussed this case earlier. Thus we have to show that Lemma 1 holds for a pair of PRs $\pi_1'$

and $\pi'_2$. This pair of PRs is different only for the first two steps but according to our proof for the case of $n = 2$ the distributions of $X_2$ coincide. After that we can apply an induction assumption. This completes the proof of Lemma 1.

**Acknowledgement.** This work was partly supported by RFBR (grant 03–01–00479).

## References

1. Berry, D.A., Fristedt, B.: Bandit problems. Sequential allocation of experiments. Monographs on Statistics and Applied Probability. Chapman & Hall, London (1985)
2. Bellman, R.: A problem in the sequential design of experiments. Sankhya **16**, 221–229 (1956)
3. Denardo, E.V., Rothblum, U.G., Van der Heyden, L.: Index policies for stochastic search in a forest with an application to R&D project management. Math. Oper. Res. **29**, no. 1, 162–181 (2004)
4. Feldman, D.: Contributions to the "two-armed bandit" problem. Ann. Math. Statist. **33**, 847–856 (1962)
5. Feinberg E., Schwartz A. (eds): Handbook of Markov Decision Processes. Kluwer Acad. Publ. (2002)
6. Gittins, J. C.: A Multi-armed Bandit Allocation indices. Wiley , Ney York (1989)
7. Gittins, J.C., Jones, D.M.: A dynamic allocation index for the sequential design experiments. In: Gani, J., Sarkadi, K., Vince, I. (eds) Progress in Statistics, European Meeting of Statisticians I. North Holland, Amsterdam, 241–266 (1974).
8. Granot, D., Zuckerman, D.,: Optimal sequencing and resource allocation in research and development projects. Management Science **37**, 140–156 (1991)
9. Mitten, L.G.: An Analytic Solution to the Least Cost Testing Sequence Problem. J. of Industr. Eng., **11**, no. 1, 17 (1960)
10. Presman, E.L., Sonin, I.M.: Sequential control with incomplete information. The Bayesian approach to multi-armed bandit problems. Academic Press (1990)
11. Sonin, I.M.: Increasing the reliability of a machine reduces the period of its work. J. Appl. Probab. **33**, no. 1, 217–223 (1996)
12. Sonin, I.M.: A Generalized Gittins Index for Markov Chain and its Recursive Calculation, manuscript (2004)
13. Varaiya, P., Walrand J., Buyukkoc, C.: Extensions of the multiarmed bandit problem: the discounted case. IEEE Trans. Autom. Control **AC-30**, no. 5, 426–439 (1985)
14. Weiss, G.: Branching Bandit Processes. Probability in the Engineering and Information Sciences **2**, 269-278 (1988)
15. Whittle, P.: Arm-acquiring bandits. Annals of Probability **9**, 284-292 (1981)