A Generalized Gittins Index for a Markov Chain and its Recursive Calculation

Isaac M. Sonin

Department of Mathematics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA

Abstract

We discuss the generalization of the classical Gittins Index for a Markov chain and propose a transparent recursive algorithm for its calculation. The foundation for this algorithm is a modified version of the Elimination algorithm proposed earlier by the author to solve the problem of optimal stopping of a Markov chain in discrete time and a finite or countable state space.

MSC: primary 60J22; 62L15; secondary 65C40; 90C40

Key words: Gittins index, Markov chain, Optimal stopping, The Elimination Algorithm

1. Introduction. The goal of this paper is twofold. First, to explain a natural generalization of the classical Gittins Index (GI). This new, Generalized Gittins Index (GGI) in a sense clarifies the "true meaning" of the GI. Second, to present a transparent recursive algorithm to calculate this GGI.

Our algorithm is based on the well-known representation of the GI through a family of stopping problems due to P. Whittle (1980) and on earlier work of the author on the recursive algorithm for the optimal stopping of a Markov chain, the *Elimination algorithm* (EA), described in Sonin (1999a) (see also (1995), (1999b) and (2006)).

We will use the following notation. A pair $M = (X, P)$, where X is a countable state

Email address: imsonin@email.uncc.edu (Isaac M. Sonin).

URL: http://www.math.uncc.edu/∼imsonin (Isaac M. Sonin).

[1] Tel:704–687–2079; fax: 704–687–6415

space, $P = \{p(x,y)\}$ is a transition matrix, is called a *Markov model*. A Markov chain (MC) from a family of MCs defined by a Markov model is denoted by (Z_n). The probabilistic measure for the Markov chain with initial point x and the corresponding expectation are denoted by P_x and E_x, respectively. A tuple $M = (X, P, c, g, \beta)$, where $c(x)$ is a *one step reward (cost)* function, $g(x)$ is a *terminal reward* function, both defined on X, and β is a *discount factor*, $0 < \beta \leq 1$, is called an *Optimal Stopping (OS) model*. The *value function* $v(x)$ for an OS model is defined as $v(x) = \sup_{\tau \geq 0} E_x[\sum_{i=0}^{\tau-1} \beta^i c(Z_i) + \beta^\tau g(Z_\tau)]$, where the sup is taken over all stopping times τ, $\tau \leq \infty$, and $g(Z_\infty) = 0$. We assume that the model M is such that $v(x) < \infty$ for all x. It is well known that function v is a minimal solution of a corresponding *Bellman (optimality) equation* $v = \max(g, c + \beta P v)$, where $Pf(x) = \sum_y p(x,y) f(y)$ is the averaging operator, defined by a transition matrix P. Let us denote by $S = \{x : v(x) = g(x)\}$. If the state space X is finite then the random time $\tau_0 = \min\{n \geq 0 : Z_n \in S\}$ is an optimal stopping time. The set S is called the optimal stopping set. We call an OS model with the terminal reward function $g(x) = 0$ for all x a *reward model*.

Given a reward model $M = (X, P, c, \beta)$, and point $x \in X$ the classical *Gittins index,* $\gamma(x)$, is defined as the maximum of the expected discounted total reward during the interval $[0, \tau)$ per unit of expected discounted time for the Markov chain starting from x, i.e.

$$\gamma(x) = \sup_{\tau > 0} \frac{E_x \sum_{n=0}^{\tau-1} \beta^n c(Z_n)}{E_x \sum_{n=0}^{\tau-1} \beta^n} = (1-\beta) \sup_{\tau > 0} \frac{E_x \sum_{n=0}^{\tau-1} \beta^n c(Z_n)}{1 - E_x \beta^\tau}, \tag{1}$$

where $0 < \beta < 1$, and τ is a stopping time, $\tau > 0$.

The GI index plays an important role in the theory of Multi-armed bandit (MAB) problems with *independent* arms but it also appears naturally in many other problems of stochastic optimization, e.g. in the optimal replacement problems, where in many cases to find an optimal strategy amounts to the calculation of (1). There are a few algorithms to calculate this index (see e.g. Varaiya et al. (1985), Kathehakis and Veinot (1987), Bertsimas and Nino Mora (1996)), Nino Mora (2007) and some generalizations of this index (see e.g. Mandelbaum (1987), El Karoui and Karatzas (1993)), which do not cover our generalization. New interesting results connecting the GI with other problems of stochastic optimization can be found in Bank and El Karoui (2004).

The author would like to thank Robert Anderson, Joseph Quinn and Ernst Presman who read the first version of this paper and made valuable comments, and an anony-

mous referee for helpful suggestions.

2. Classical GI and Generalized GI. An important interpretation of the GI, the so called *Retirement Process* formulation was provided by Whittle (1980). Given a reward model $M = (X, P, c(x), \beta), 0 < \beta < 1$, he introduced the family of OS models $M(k) = (X, P, c(x), k, \beta)$, where the terminal reward function $g(x) = k$ for all $x \in X$, k is a real number. Denote $v(x,k)$ the value function for such a model, i.e. $v(x,k) = \sup_{\tau \geq 0} E_x[\sum_{n=0}^{\tau-1} \beta^n c(Z_n) + \beta^\tau k]$, and denote $w(x) = \inf\{k : v(x,k) = k\}$. Since $\beta < 1$, for sufficiently large k it is optimal to stop immediately and $v(x,k) = k$. Thus $w(x) < \infty$. The results of Whittle imply that $v(x,k) = k$ for $k \geq w(x)$, $v(x,k) > k$ for $k < w(x)$, and $\gamma(x) = (1-\beta)w(x)$.

Another interesting interpretation of the GI, the so called *Restart in State* interpretation, was given in Kathehakis and Veinot (1987), though similar ideas of regenerative cycles were used in probability theory a long time ago (see e.g. references in Sonin (1996)). Let us consider a Markov Decision model $M_x = (X, A(y), P, c(y), \beta)$, where a point $x \in X$ is *fixed* and a set of actions $A(y)$ available at y, has two actions - to continue or to return to x and continue from there. In other words, MC (Z_n) starting from a point x after a positive stopping time $\tau > 0$ can be restarted at the same point x, and so on. Let $h(x)$ denote the supremum over all strategies of the expected total reward on the infinite time interval in this model, i.e. $h(x) = \sup_\pi E_x^\pi \sum_{n=0}^{\infty} \beta^n c(Z_n)$, where E_x^π is an expectation with respect to a strategy π. Using the standard results of Markov Decision Processes theory, Kathehakis and Veinot proved that $h(x) = w(x)$ and function $h(x)$ satisfies the equality

$$h(x) = \sup_{\tau > 0} E_x[\sum_{n=0}^{\tau-1} \beta^n c(Z_n) + \beta^\tau h(x)]. \tag{2}$$

Combined with the results of Whittle this implies that $\gamma(x) = (1-\beta)\, h(x) = (1-\beta)w(x)$. We will prove the equality $h(x) = w(x)$ in a more general setting in Theorem 1.

Before introducing the Generalized GI (GGI) let us make the following almost trivial remark. As usual in Markov Decision Processes theory, the optimizations problems, such as described above, with an explicit discount factor β, are equivalent to problems where a state space is complemented by an absorbing point x_* and the new transition probabilities are defined as follows. The probability of entering an absorbing point x_* in one step for any state $y \neq x_*$ (probability of termination) is equal to $1-\beta$ and all other initial transition probabilities are multiplied by β. In other words, β is the probability

of "survival". The model of latter type with a possible *variable* probability of survival $\beta(x)$ plays a crucial role in our subsequent presentation. Thus, to define the GGI $\alpha(x)$ we consider a *reward model with termination* $M = (X, P, c(x), \beta(x))$, where we assume from the beginning that the state space X contains an absorbing point x_*, the function $\beta(x)$ is the probability of "survival" at point x, so $1 - \beta(x) = p(x, x_*)$. Strictly speaking the function $\beta(x)$ is completely specified by a transition matrix P but we include $\beta(x)$ in the tuple M, to stress the presence of x_* and $\beta(x)$. From now on notation E_x, P_x and (Z_n) are referred to such model and survival probabilities $\beta(\cdot)$ now are automatically included under the signs P_x and E_x. We also assume that $c(x_*) = 0$.
Let us denote the numerator in (1), which now equals to $E_x \sum_{n=0}^{\tau-1} c(Z_n)$, by $R^\tau(x)$, and let us denote $Q^\tau(x) = P_x(Z_\tau = x_*)$, the probability of termination on $[0, \tau)$.
The *Generalized GI* (GGI), $\alpha(x)$, for a model with termination is defined as

$$\alpha(x) = \sup_{\tau>0} \frac{R^\tau(x)}{Q^\tau(x)}, \tag{3}$$

i.e. $\alpha(x)$ is the maximum discounted total reward *per chance of termination*. Note that if $\beta(x)$ is a constant β then the denominator in the second equality in (1) coincides with $Q^\tau(x)$ and therefore in this case $\gamma(x) = (1 - \beta)\alpha(x)$.
The crucial point however is that, if $\beta(x)$ is not a constant then the latter equality generally is not true anymore, even if the definition of $\gamma(x)$ is correspondingly modified, i.e. β^n is replaced by $\prod_{i=0}^{n-1} \beta(Z_n)$. Thus, in the general case, the *proportionality of the two indices $\gamma(x)$ and $\alpha(x)$ as functions of x completely disappears.* At the same time, for a reward model with termination we can define in an absolutely similar way as before a (generalized) index $h(x)$, as the value function in a restart in x problem, and a (generalized) index $w(x)$, as $w(x) = \inf\{k : v(x, k) = k\}$, where $v(x, k)$ is a value function in the (generalized) Whittle OS model $M(k) = (X, P, c(x), \beta(x), k)$. In this model we assume that $c(x_*) = g(x_*) = 0, g(x) = k$ for $x \neq x_*$. As Theorem 1 shows below, *the equality $\alpha(x) = w(x) = h(x)$ is preserved* ! This means that the "true meaning" of the Gittins index is given by the expression in (3) and not in (1) !
Now a few words about the origin of definition (3) and some comments. The first publications on GI appeared in the 70s (see Gittins (1979) and references there). But as early as in 1960 the following simple model was analyzed by a few authors simultaneously (see Mitten (1960)). We present it here in a modified form. Suppose that there is a finite set of independent Bernoulli trials $e_1, e_2, ..., e_m$, with probability of success p_i, and

correspondingly with probability of failure q_i, in i-th trial. 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 the i-th trial brings a reward r_i, and in the case of success she may quit or continue testing. In the case of failure the testing has to be *terminated*. The goal of the DM is to select the optimal order to maximize the expected total reward. A rather elementary proof shows that the optimal strategy has a remarkably simple structure and is based on an index α calculated for *each trial* $e_i, \alpha(e_i) = r_i/q_i$, i.e. reward / probability of termination. The optimal strategy has the following form: test the trials with positive index in decreasing order. If all trials must be tested then they should all be tested in the above order. Each trial may be considered as a simple MC with three states, an initial state, success and failure. This problem contains in a nutshell both the simplest form of an index (3) and the main result of Gittins theory, the optimality of an index-based strategy. This model was generalized in the papers of Granot and Zuckerman (1991), Denardo et al. (2004), and Presman and Sonin (2006). All of them are using an index of type (3) though only in the last paper, where new trials may appear in a random fashion, the variable probability of termination is fully considered. These models represent the most general setting in which Gittins theorem remains valid.

Note also that contrary to popular belief the renown Gittins theorem mentioned above holds true for the case of variable $\beta(x)$ and there is no contradiction with a well-known result of Berry and Fristedt (1985) which states that geometric discounting is not only sufficient but is also necessary in the class of discounting sequences $(\beta_1, \beta_2, ...)$. It means that what really matters for the validity of Gittins theorem is stationarity with respect to time but not space.

Theorem 1. *The three indices defined for a reward model with termination* $M = (X, P, c(x), \beta(x))$ *coincide, i.e.* $\alpha(x) = h(x) = w(x)$.

Proof. First, let us prove that in a such model the relations $v(x,k) = k$ for all $k \geq w(x)$, $v(x,k) > k$ for all $k < w(x)$ in the original Whittle's model remains valid. Using our notation $R^\tau(x)$ and $Q^\tau(x)$, we can represent the value function $v(x,k) = \sup_{\tau \geq 0} E_x[\sum_{n=0}^{\tau-1} c(Z_n) + I(Z_\tau \neq x_*)k] = \sup_{\tau \geq 0}[R^\tau(x) + (1 - Q^\tau(x))k]$, where $I(A)$ is a characteristic function of a set A. If $v(x,k) > k$ then there is a stopping time $\tau > 0$ such that $R^\tau(x) + (1 - Q^\tau(x))k > k$. Therefore, $R^\tau(x)/Q^\tau(x) > k$ and if $m < k$ then $R^\tau(x) + (1 - Q^\tau(x))m > m$ and $v(x,m) > m$. Then, using the definition of $w(x)$, we obtain that $v(x,k) > k$ for all $k < w(x)$ and $v(x,k) = k$ for all $k > w(x)$. The equality $v(x, w(x)) = w(x)$ follows from the continuity of function $v(x,k)$ in k. One also may

conclude that if $k < w(x)$ then $\alpha(x) > k$. Since this is true for any $k < w(x)$, we obtain that $\alpha(x) \geq w(x)$. If $k > w(x)$, then, as we proved, $v(x,k) = k$, and therefore by the definition of $v(x,k)$, $k \geq R^\tau(x) + (1 - Q^\tau(x))k$ for any $\tau > 0$. This implies that $k \geq R^\tau(x)/Q^\tau(x)$ and hence $\alpha(x) \leq k$ for any $k > w(x)$. Therefore $\alpha(x) \leq w(x)$. The equality $\alpha(x) = w(x)$ is proved.

To prove $\alpha(x) = h(x)$, note that in a reward model with termination the equality (2) remains true, but now, when $\beta(x)$ is not a discount factor but a "survival" probability, takes the form $h(x) = \sup_{\tau>0} E_x[\sum_{n=0}^{\tau-1} c(Z_n) + I(Z_\tau \neq x_*)h(x)]$. This follows from the fact that the proof of (2) uses only the general properties of Markov Decision Processes equally true for both reward and general reward models. This equality can be rewritten as $h(x) = \sup_{\tau>0}[R^\tau(x) + (1 - Q^\tau(x))h(x)]$. Assuming, as in a classical Gittins case, that $\beta(x) < 1$, and hence $Q^\tau(x) \geq 1 - \beta(x) > 0$, it is easy to see that this is equivalent to the equality $h(x) = \sup_{\tau>0} R^\tau(x)/Q^\tau(x)$, i.e. $h(x) = \alpha(x)$.

3. The Elimination Algorithm for the problem of Optimal Stopping of a Markov chain. We present here just the bare facts necessary for the subsequent discussion and refer the reader to Sonin (1999a), (1999b) and (2006). Let $M_1 = (X_1, P_1)$ be a Markov model, and $D \subset X_1$. If (Z_n) is a MC specified by this model, and (Y_n) be a random sequence obtained by observing (Z_n) during its visits to the set $X_2 = X_1 \setminus D$, then (Y_n) is a MC in X_2 with the following transition probabilities P_2. If matrix P_1 is decomposed as

$$P_1 = \begin{bmatrix} Q_1 & T_1 \\ R_1 & P_1' \end{bmatrix}, \tag{4}$$

where substochastic matrix Q_1 describes the transitions inside of D, P_1' describes the transitions inside of X_2 and so on, then

$$P_2 = P_1' + R_1 U_1 = P_1' + R_1 N_1 T_1. \tag{5}$$

In this formula $U_1 = \{u_1(x,y), x \in D, y \in X_2\}$ is a matrix for the distribution of a MC at the moment of first exit from D (exit probabilities matrix), and $N_1 = \{n_1(x,y), x, y \in D\}$ is a *fundamental matrix* for the substochastic matrix Q_1, i.e. $N_1 = \sum_{n=0}^{\infty} Q_1^n = (I - Q_1)^{-1}$, and $n_1(x,y)$ is the expected number of visits to y before the moment of first exit from D starting at x. Given set D, matrices N_1 and U_1 are related by formula $U_1 = N_1 T_1$.

An important case is when the set D consists of one nonabsorbing point z. In this case formula (5) obviously takes the form

$$p_2(x,\cdot) = p_1(x,\cdot) + p_1(x,z)n_1(z)p_1(z,\cdot), \tag{6}$$

where $n_1(z) = 1/(1-p_1(z,z))$.

According to this formula, each row-vector of the new stochastic matrix P_2 is a linear combination of two rows of P_1 (with the z-column deleted). For a given row of P_2, these two rows are the corresponding row of P_1 and the z^{th} row of P_1. This transformation corresponds formally to one step of the Gaussian elimination method.

Let $M_1 = (X_1, P_1, c_1(x), g(x), \beta_1(x))$ be an OS model with termination, $v_1(x)$ be the value function for this model, and $S = \{x : g(x) = v_1(x)\}$ be the corresponding optimal stopping set. Let us introduce now a *transformation of the cost function* $c_1(x)$ (or any function $f(x)$) defined on X_1 into the cost function $c_2(x)$ defined on X_2, under the transition from model M_1 to model M_2. Given set $D, D \subset X_1$, let τ be the moment of the first *return* to X_2, i.e. $\tau = \min(n \geq 1, Z_n \in X_2)$. Then given function $c_1(x)$ defined for $x \in X_1$, function $c_2(x)$ is defined on $x \in X_2$ as

$$c_2(x) = E_x \sum_{n=0}^{\tau-1} c_1(Z_n) = c_1(x) + \sum_{z\in D} p_1(x,z) \sum_{w\in D} n_1(z,w)c_1(w). \tag{7}$$

In other words the new function $c_2(x)$ represents the expected cost (reward) gained by a MC starting from point $x \in X_2$ up to the moment of first return to X_2. For a function $f(x)$ defined on a set X_1 and a set $G \subset X_1$ denote f_G a column-vector function reduced to a set G. Then formula (7) can be written in matrix form as $c_2 = c_{1,X_2} + R_1 N_1 c_{1,D}$.

If the set $D = \{z\}$ then the function $c_1(x)$ is transformed as follows

$$c_2(x) = c_1(x) + p_1(x,z)n_1(z)c_1(z), x \in X_2. \tag{8}$$

The latter formula was obtained earlier in Sheskin (1999) in the context of Markov Decision Processes.

The Elimination algorithm for the OSP of a MC is based on the following three facts.

1. Though in an OSP it may be *difficult* to find the states where it is optimal *to stop*, it is *easy* to find a state (states) where it is optimal *not to stop.* It is optimal not to stop at z if $g(z) < c(z) + Pg(z)$, i.e. the expected reward of doing *one more step* is larger

than the reward from stopping. Generally, it is optimal not to stop at any state where the expected reward of doing some, perhaps random number of steps, is larger than the reward from stopping.

2. After we have found states (state) which are not in the optimal stopping set, we can eliminate them and recalculate the transition matrix using (6) or (5), and recalculate the cost function using (8) or (7). After that in the reduced model we can repeat the first step and so on.

3. Finally, though if $g(z) \geq c(z) + Pg(z)$ at a particular point z, we can not make a conclusion about whether this point belongs to the stopping set or not, but if this inequality is true for *all* points in the state space then we have the following simple and well-known statement

Proposition 1. *Let M be an optimal stopping problem, and $g(x) \geq c(x) + Pg(x)$ for all $x \in X$. Then X is the optimal stopping set in the problem M, and $v(x) = g(x)$ for all $x \in X$.*

The following theorem provides the formal justification for the EA. It was formulated in a slightly different form in Sonin (1995) and proved in Sonin (1999a) for the case when $c(x) = 0$ for all x. (The proof for general $c(x)$ can be found in Sonin (2006)).

Theorem 2. *(Elimination theorem). Let $M_1 = (X_1, P_1, c_1, g)$ be an OS model, $D \subseteq C_1 = \{z \in X_1 : g(z) < c_1(z) + P_1 g(z)\}$. Consider an OS model $M_2 = (X_2, P_2, c_2, g)$ with $X_2 = X_1 \setminus D$, $p_2(x, y)$ defined by* (5), and *c_2 is defined by* (7). *Let S be the optimal stopping set in M_2. Then* a) *S is the optimal stopping set in M_1 also*, b) *$v_1(x) = v_2(x) \equiv v(x)$ for all $x \in X_2$, and for all $z \in D$*

$$v_D = N_1[c_{1,D} + T_1 v_{X_2}]. \tag{9}$$

If the set $D = \{z\}$ then formula (9) can be written as

$$v_1(z) = n_1(z)[c_1(z) + \sum_{y \in X_2} p_1(z, y)v(y)]. \tag{10}$$

For the sake of brevity we call two such OS models M_1 and M_2 *equivalent*.

The EA consists of two stages: reduction and backward stages. The first stage can be described as a sequence of steps where subsets of states that do not belong to the stopping set are eliminated till the stopping set is achieved. The selection of these steps in the countable case is dictated by the structure of the problem and the convenience of

calculation of matrices $U = NT$. The algorithm has an especially simple structure if the state space is finite, and only one state is eliminated at each step.

Finally, on the backward stage, by reversing the steps of the reduction stage, we can calculate recursively the values of $v(x)$ for all $x \in X_1$, using sequentially formula (9) or (10), starting from the equalities $v(x) = g(x)$ for $x \in S = X_k$, where k is the number of iteration where the reduction stage of the algorithm stops.

4. The Gittins Index Elimination (GIE) Algorithm. To apply the EA algorithm to calculate $\alpha(x)$ we need the following statement. Given a reward model with termination $M = (X, P, c(x), \beta(x))$, with $\beta(x) = 1 - p(x, x_*) < 1$, let us define the function $d(x) = c(x)/(1 - \beta(x))$, number $d = \max_{x \in X} d(x)$, and the set $D = \{x : d(x) = d\}$.

Theorem 3. *Let M be a reward model with termination, number d and set D defined as above. Then $\alpha(x) = d$ for $x \in D$ and $\alpha(x) < d$ for all $x \in X \setminus D$.*

Proof. Let us consider the Whittle OS model $M(k)$ and let $S(k)$ be the optimal stopping set. If $k \geq d$ then the definition of d implies that for this OS model $g(x) - (c(x) + Pg(x)) = k - (c(x) + \beta(x)k) = (1 - \beta(x))(k - d(x)) \geq 0$ for all $x \in X$ and hence by Proposition 1, $S(k) = X$ and $v(x, k) = k$ for *all* $x \in X$. If $x \notin D$ and $d(x) < k < d$ then similarly $v(x, k) = k$ and hence by the definition of $w(x)$, and equality $\alpha(x) = w(x)$ we obtain $\alpha(x) < d$. If $x \in D$ and $k < d$ then $g(x) - (c(x) + Pg(x)) < 0$ for $x \in D$ and therefore $x \notin S(k)$ and $v(x, k) > k$. By the definition of $w(x)$, and equality $\alpha(x) = w(x)$ this implies that $\alpha(x) \geq d$ and therefore $\alpha(x) = d$ for $x \in D$. The theorem is proved.

Now we can describe the GIE algorithm and *prove* that it really calculates the GGI.

Step 1. Given a reward model with termination $M_1 = (X_1, P_1, c_1(x), \beta_1(x))$, calculate defined above the function $d_1(x)$, the number d_1, and the set D_1. By Theorem 3, $\alpha(x) = d_1$ on the set D_1. Without loss of generality we can assume that $D_1 = \{z\}$.

Step 2. Define model $M_2 = (X_2, P_2, c_2(x), \beta_2(x))$, where $X_2 = X_2 \setminus D_1$, stochastic matrix P_2 is obtained by (5) for $D = D_1$, function $c_2(x)$ is obtained by (7), $\beta_2(x) = 1 - p_2(x, x_*)$. Calculate $d_2(x) = c_2(x)/(1 - \beta_2(x))$ on X_2, number $d_2 = \max_{x \in X_2}$, and set $D_2 = \{x : d_2(x) = d_2\}$.

To prove that $\alpha(x) = d_2$ on a set D_2, let us consider Whittle OS model $M_1(k) = (X_1, P_1, c_1(x), \beta_1(x), k)$ with $k < d_1$. Let $S_1(k)$ be an optimal stopping set for this model. Theorem 3 implies that for all such k we have $D_1 \cap S_1(k) = \varnothing$ and hence by Theorem 2 this model is equivalent to a new OS model $M_2(k) = (X_2, P_2, c_2(x), \beta_2(x), k)$, where X_2, P_2, $\beta_2(x)$, $d_2(x)$, d_2, and D_2 are described above. It can be checked using formulas (5) and (7) that $d_2(x) = [k(x)d_1(x) + k'(x, z))d_1(z)]/(k + k')$, where $k, k' > 0$. Hence

$d_2(x) < d_1(x)$ and $d_2 < d_1$. Theorem 3 implies that $\alpha(x) = d_2$ on a set D_2. And so on. If $|X_1|$ is finite, in no more than $|X_1|$ steps $\alpha(x)$ will be calculated for all points. Note also that the elimination of a set D_1 can be performed state by state using (6) and (8) or at once using formulas (5) and (7).

Example 1. Let our reward model has $X_1 = \{1, 2, 3, x_*\}$, with $c_1(1) = 3, c_1(2) = 2, c_1(3) = 1$, and $\beta(x) = .9$ for all $x \neq x_*$ and corresponding transition matrix $P = P_1$ is

$$P_1 = \begin{array}{|c|c|c|c|} \hline .3 & .3 & .3 & .1 \\ \hline .45 & .3 & .15 & .1 \\ \hline .1 & .5 & .3 & .1 \\ \hline 0 & 0 & 0 & 1 \\ \hline \end{array}, \quad P_2 = \begin{array}{|c|c|c|} \hline .4929 & .3429 & .1642 \\ \hline .5429 & .3429 & .1142 \\ \hline 0 & 0 & 1 \\ \hline \end{array}, \quad P_3 = \begin{array}{|c|c|} \hline .7099 & .2901 \\ \hline 0 & 1 \\ \hline \end{array}.$$

Then $d_1 = c_1(1)/(1 - \beta) = 30, D_1 = \{1\}$, and by Theorem 3 $\alpha(1) = 30$. Therefore, we eliminate the state 1 on a first step and, applying formulas (6) and (8), we obtain new transition matrix P_2 and function $c_2(x)$ for a state space $X_2 = \{2, 3, x_*\}$; $c_2(2) = 3.9286, c_2(3) = 1.4286$. Therefore $d_2 = c_2(2)/(1 - \beta_2(2)) = 23.9130 = \alpha(2)$, $D_2 = \{2\}$ and on the second step state 2 is eliminated, and we obtain matrix P_3, and $c_3(3) = 5.6338$. Therefore $\alpha(3) = c_3(3)/(1 - \beta_3(3)) = 19.4175$. (All calculations were rounded up four digits after decimal point.)

Note that though we started in this example from a constant survival function $\beta(x)$, after the first step we deal with variable $\beta_i(x)$ for $i > 1$. The classical GI for this model $\gamma(x) = (1 - \beta)\alpha(x) = .1\alpha(x)$.

5. The optimal stopping times and the Representation Identity. For the case when X is finite, we can describe also two optimal stopping times where the value of GGI $\alpha(x)$ is achieved. We omit the proof of both lemmas.

Lemma 1. *Let M be a reward model with termination, and let the sets D_i and the numbers d_i, $i = 1, 2, ...$ be those which occur in the calculation of $\alpha(x)$. Then, if $x \in D_i$, $\tau_1 = \min\{n > 0 : Z_n \notin (D_1 \cup ... \cup D_{i-1})\} \equiv \min\{n > 0 : \alpha(Z_n) \leq \alpha(x) = d_i\}$ and $\tau_1^0 = \min\{n > 0 : Z_n \notin (D_1 \cup ... \cup D_i)\} \equiv \min\{n > 0 : \alpha(Z_n) < \alpha(x) = d_i\}$ are the optimal stopping times.*

Lemma 2. *Let M be a general reward model. Then the following formula (Representation Identity) holds*

$$v(x) \equiv E_x \sum_{n=0}^{\infty} c(Z_n) = E_x \sum_{n=0}^{\infty} [\min_{0 \leq i \leq n} \alpha(Z_i)] I(B_{n+1}), \tag{11}$$

where $\alpha(x)$ is a GGI for this model, and $B_n = \{T = n\} \equiv \{Z_{n-1} \neq x_*, Z_n = x_*\}$.
Note also that using this equality and the sequence of models M_n described in the algorithm, it is possible also to calculate $v(x)$ recursively. This remark can be also extended to the general setting of Gittins theorem.

6. A Brief Comparison with Other Algorithms. The assumption that discount rate is constant is mainly a technical convenience in applied probability models and is not natural in many economic or financial applications. As we described above, our algorithm deals with the more general case of a variable discount rate depending on a state of a MC. It belongs to a wide class of algorithms based on the idea of states elimination. This idea was first applied in 1985 in two independent papers of T. Sheskin, and Grassman, Taksar and Heyman concerned with the calculation of invariant distribution of MC. We refer the reader to our paper Sonin (1999b) where this approach is discussed. Our algorithm has the same computational complexity $O(n^3)$ as the algorithm of Varaya et al or similar algorithm of Bertsimas and Nino-Mora (1996) and Nino-Mora (2007). It has simple and transparent probabilistic interpretation and lends itself naturally to the extension to the case of countable MC. A simple example of such a situation is a random walk on a line $\{x_*, 0, \pm 1, \pm 2, ...\}$ with general transition probabilities $p(i, i+1) = p_i$, $p(i, i-1) = q_i, p(i, i) = r_i, p(i, x_*) = s_i, p_i + q_i + r_i + s_i = 1$, and function $c(i) \geq 0, c(i) \to 0$ as $i \to \pm\infty$. It is easy to check that in this case for any i, the index $\alpha(i)$ is calculated in a finite number of steps. More than that, the computations in this example as in some other examples can be performed in parallel. Finally note that the algorithm based on the idea of using state elimination to calculate Gittins index was briefly, in a few lines described in Tsitsiklis (1994), but the assertion there that for the case of a discrete MC such an algorithm will coincide with the algorithm of Varaya et al and that it is also a special case of the algorithm in Weiss (1988) is not true. Even the calculations for our simple Example 1 will be quite different. This line of study was continued in Katta and Sethuraman (2004) who presented an algorithm similar to our but without rigorous proofs and without a reference to optimal stopping problem. The comparison of computational properties of existing algorithms (for constant β), including our algorithm described in a technical report, is presented in forthcoming paper Nino-Mora (2007).

7. Possible Generalizations and Open Problems. We described our algorithm to calculate the GGI for the case when $\beta(x) < 1$. In a very similar way it can be used for the undiscounted case, $\beta = 1$, assuming that the corresponding GI in (1), $\gamma(x)$, is finite. It can be proved that in this case $\lim_{\beta \to 1} \alpha_\beta(x)(1-\beta) = \gamma(x)$. For example in Example 1, the

corresponding values for γ are: $\gamma(1) = 3, \gamma(2) = 17/7 = 2.428, \gamma(3) = 232/112 = 2.071$. The algorithm described above can be modified to accommodate also the case when the expression $R^\tau(x)$ in the definition of GGI has a form $R^\tau(x) = E_x[\sum_{n=o}^{\tau-1} c(Z_n) + g(Z_\tau)]$, where the function $g(x)$ is a terminal reward at the end of a cycle $[0, \tau)$, but this modification is not quite trivial. We described our algorithm for the case when the state space X is finite. The sequential calculation of $\alpha(x)$ is possible also for the countable case if at each stage the sets D_i are not empty and $\cup_{i=1}^{\infty} D_i = X$. The general description of such situations is an open problem. The other open problems are the analogs of the described algorithm for continuous time and/or space.

References

[1] Bank, P., El Karoui, N. (2004). A stochastic representation theorem with applications to optimization and obstacle problems. Ann. Probab. **32**, 1B, 1030–1067.

[2] Berry, D. A., Fristedt, B. (1985). Bandit problems, Sequential allocation of experiments. Chapman & Hall, London.

[3] Bertsimas, D., Niño-Mora, J. (1996). Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems. Math. Oper. Res. **21**, 2, 257–306.

[4] Denardo, E., Rothblum, U., Van der Heyden, L. (2004). Index policies for stochastic search in a forest with an application to R&D project management. Math. Oper. Res. **29**, 1, 162–181.

[5] El Karoui, N., Karatzas, I. (1993). General Gittins index in discrete time. Proc. Natl. Acad. Sci. USA, **90**, 1232-1236.

[6] Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. J. Roy. Statist. Soc. Ser. B **41**, 2, 148–177.

[7] Granot, D., Zuckerman, D. (1991). Optimal sequencing and resource allocation in research and development projects, Management Science, **37**, 140-156.

[8] Katehakis, M., Veinott, A. (1987). The multi-armed bandit problem: decomposition and computation. Math. Oper. Res. **12**, 2, 262–268.

[9] Katta, A.-K., Sethuraman, J. (2004). A note on bandits with a twist. SIAM J. Discrete Math. **18**, 1, 110–113.

[10] Mandelbaum, A. (1987). Continuous multi-armed bandits and multiparameter processes. Ann. Probab. **15**, 4, 1527–1556.

[11] Mitten, L. (1960). An Analytic Solution to the Least Cost Testing Sequence Problem. J. of Industr. Eng. **11**, 1, 17.

[12] Niño-Mora, J. (2007). A (2/3) nˆ3 fast pivoting algorithm for the Gittins index and optimal stopping of a Markov chain. Appear in INFORMS Journal on Computing.

[13] Presman, E., Sonin I. (2006). Gittins Type Index Theorem for Randomly Evolving Graphs, In: From Stochastic Calculus to Mathematical Finance. The Shiryaev Festschrift, Springer, Kabanov Y., Lipster, R.; Stoyanov, J. (Eds), pp. 567-588.

[14] Sheskin, T. J. (1999). State reduction in a Markov decision process. Internat. J. Math. Ed. Sci. Tech. **30**, 2, 167–185.

[15] Sonin, I. (1995). Two simple theorems in the problems of optimal stopping, in Proc. INFORMS Appl. Prob. Conf., Atlanta, Georgia, pp. 27-28.

[16] Sonin, I. (1996). Increasing the reliability of a machine reduces the period of its work. J. Appl. Probab. **33**, 1, 217–223.

[17] Sonin, I. (1999a). The Elimination Algorithm for the Problem of Optimal Stopping, Math. Meth. of Oper. Res. **4**, 1, 111-123.

[18] Sonin, I. (1999b). The State Reduction and related algorithms and their applications to the study of Markov chains, graph theory and the Optimal Stopping problem, Advances in Mathematics, **145**, 159-188.

[19] Sonin I. (2006). The Optimal Stopping of Markov Chain and Recursive Solution of Poisson and Bellman Equations, In: From Stochastic Calculus to Mathematical Finance. The Shiryaev Festschrift, Springer, Kabanov Y., Lipster, R.; Stoyanov, J. (Eds), pp. 609-621.

[20] Tsitsiklis, J. N. (1994). A short proof of the Gittins index theorem. Ann. Appl. Probab. **4**, 1, 194–199.

[21] Varaiya, P., Walrand J., Buyukkoc C. (1985). Extensions of the Multiarmed Bandit Problem: IEEE Trans. Autom. Control AC-30, 426-439.

[22] Weiss, G. (1988). Branching Bandit Processes. Probab. Engng. Inform. Sci., **2**, 269-278.

[23] Whittle, P. (1980). Multi-armed Bandits and the Gittins Index. J. Roy. Statist. Soc. Ser. B **42**, 2, 143-149.